THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals

Citation for published version:
Simmen, MW 2008, 'Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals', *Genomics*, vol. 92, no. 1, pp. 33-40. https://doi.org/10.1016/j.ygeno.2008.03.009

Digital Object Identifier (DOI):
10.1016/j.ygeno.2008.03.009

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genomics

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 29. Apr. 2024

Genomics 92 (2008) 33–40

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals

Martin W. Simmen *

School of Biomedical Sciences, University of Edinburgh, Edinburgh EH8 9XD, UK

ARTICLE INFO

Article history:
Received 6 March 2008
Accepted 26 March 2008
Available online 15 May 2008

Keywords:
Base composition
Dinucleotides
CpG
TpG
DNA methylation
Human
Mouse
Pufferfish
Ciona intestinalis
Evolution

ABSTRACT

In mammalian genomes CpGs occur at one-fifth their expected frequency. This is accepted as resulting from cytosine methylation and deamination of 5-methylcytosine leading to TpG and CpA dinucleotides. The corollary that a CpG deficit should correlate with TpG excess has not hitherto been systematically tested at a genomic level. I analyzed genome sequences (human, chimpanzee, mouse, pufferfish, zebrafish, sea squirt, fruitfly, mosquito, and nematode) to do this and generally to assess the hypothesis that CpG deficit, TpG excess, and other data are accountable in terms of 5-methylcytosine mutation. In all methylated genomes local CpG deficit decreases with higher G + C content. Local TpG surplus, while positively associated with G + C level in mammalian genomes but negatively associated with G + C in nonmammalian methylated genomes, is always explicable in terms of the CpG trend under the methylation model. Covariance of dinucleotide abundances with G + C demonstrates that correlation analyses should control for G + C. Doing this reveals a strong negative correlation between local CpG and TpG abundances in methylated genomes, in accord with the methylation hypothesis. CpG deficit also correlates with CpT excess in mammals, which may reflect enhanced cytosine mutation in the context 5′-YCG-3′. Analyses with repeat-masked sequences show that the results are not attributable to repetitive elements.

© 2008 Elsevier Inc. All rights reserved.

Variations in dinucleotide abundance levels both within and between genomes are striking features. Relative to the frequencies that would be expected from base composition alone, some dinucleotides are overrepresented, others underrepresented. Differing hypotheses have been proposed to account for these features, suggesting that they result from mutational bias, selection effects, DNA structural constraints, mathematical artifact, or some combination of these. Perhaps only for the CpG dinucleotide has a partial consensus emerged. Early experimental studies [1,2] found that in vertebrates the CpG dinucleotide displays the strongest bias, being highly underrepresented. Subsequent direct analysis of complete genome sequences has confirmed this, with the human and mouse both showing approximately fivefold CpG depletion [3,4].

The most convincing explanation of vertebrate CpG depletion is that it is the consequence of the very high likelihood of cytosines in the CpG context to become 5-methylcytosines (5mC), combined with the established fact that 5mC are highly prone to mutating, via spontaneous hydrolytic deamination, to thymine. If endogenous mismatch repair enzymes fail to correct the T/G mispairing, then following the next round of replication the CpG dinucleotide converts to either TpG or CpA if the deamination occurs on the opposite strand [5,6]. Early support for this account came from three observations. First, by combining data on CpG depletion derived from nearest-neighbor analyses with data on methylation levels derived from comparisons of the cleavage patterns obtained by digestion with either MspI or its methylation-sensitive isoschizomer HpaII, it was shown that CpG depletion is strongest in those organisms showing the highest degree of cytosine methylation (vertebrates), negligible in genomes with barely detectable levels of cytosine methylation (e.g., arthropods), and moderate in partially methylated genomes (e.g., echinoderms and tunicates) [7]. Second, a comparison of various animal species found that the magnitude of the CpG deficit was positively correlated with an excess of TpG and CpA [7]. Third, despite its scarcity in the mammalian genome, about one-third of point mutations causing human genetic disorders were found to involve the CpG dinucleotide [8]. More recent evidence for the importance of CpG methylation has come from analysis of the human genome sequence revealing that a disproportionately high fraction of single nucleotide polymorphisms (SNPs) involve CpG → TpG/CpA transitions, e.g., 28% of exonic SNPs are of this type [9].

Research on the quantitative effects of CpG methylation has focused on three issues. First, the deficiency of CpGs is not uniform across the mammalian genome; rather it varies from being approximately fivefold in low G + C content regions to approximately threefold in high G + C regions [2,10]. This trend has been best explained as a consequence of the fact that deamination of 5mC in double-stranded DNA requires transient local strand separation. Regions of high G + C content possess a

* Corresponding author. Fax: +44 131 650 6527.
E-mail address: M.simmen@ed.ac.uk.

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.03.009

higher DNA melting temperature than low G + C content regions; therefore these are less susceptible to strand separation, resulting in a lower rate of deamination and hence a lesser degree of CpG depletion. This explanation has been quantitatively supported both by simulations of dinucleotide evolution [11] and through analysis of human SNP data showing that the specific 5mC deamination rate bears a power law relationship to the G + C content of the region surrounding the CpG SNP (with exponent near to the predicted value of −3) [12], provided that a sufficiently large (> 500bp) region is considered [13]. Additionally, however, Duret and Galtier [14] noted that even under a simplistic model in which the individual rate of 5mC mutation remains constant over regions with different G + C contents, the observed underrepresentation of CpG would still lessen at higher G + C values. Their logic was that depletion of CpGs lowers the observed G + C content, leading to underestimation of the expected number of CpGs and hence to overestimation of the CpG observed/expected ratio, with this bias being stronger in high G + C content regions. Simulations of a corresponding model of dinucleotide evolution using an elevated but constant mutation rate for bases in a CpG doublet and a mutation rate with variable G + C bias for the other nucleotides also generated datasets displaying a correlation between G + C content and CpG observed/expected ratio [14]. Another potential contributory factor in the rise of the relative CpG abundance with increasing G + C content could be the higher gene density in G + C-rich regions due to the CpG islands associated with many genes, although this would produce only a weak effect due to the small size of the CpG island proportion in the mammalian genome [3].

A second issue concerned the relative magnitudes of the CpG deficit and TpG excess. Given that an mCpG mutation causes the loss of two CpGs (taking both strands into account) accompanied by the creation of one TpG and one CpA, it appeared at first sight odd that in mammals CpG is roughly fourfold depleted, whereas TpG and CpA are each only approximately 20% in excess. However, this counting argument does not take dynamics into account. Analysis of appropriate quantitative models of dinucleotide evolution showed that these levels of TpG (CpA) excess are in accord with those expected at equilibrium [11,15]. The reason is essentially that a proportion of the excess TpG and CpA dinucleotides created by CpG depletion are themselves lost via mutation to other dinucleotides over time. It is also formally possible that some TpG changes are lost due to selection, although as such events would likely be restricted to changes occurring in coding sequences or regulatory elements, they would make little contribution to the genome-level data.

Third, TpA dinucleotides also show considerable underrepresentation, not just in vertebrates but throughout eukaryotic genomes [16]. This has been attributed to various factors that could generate weak selection against TpA. UpA is prone to targeting by ribonucleases [17] so may be selected against in mRNAs for reasons of stability. In addition, TpA has the lowest thermodynamic stacking energy of any dinucleotide and is present in key regulatory motifs, which might result in it being selected against in bulk DNA [16]. Data from human gene DNA sequences [18] also show that the observed/expected ratio of TpA decreases at higher G + C levels. It is notable that both of the methylation-based accounts of the variation in CpG depletion with G + C level mentioned above [11,14] also predict this trend, as a secondary statistical consequence of CpG depletion by deamination (see above papers for further details).

But certain observations seem, at least at first sight, at odds with the methylation hypothesis. CpG depletion is also present in the (unmethylated) mitochondrial genomes of animals [19] and in unmethylated small vertebrate DNA viruses [20,21], raising the prospect of mechanisms for CpG depletion not mediated by 5mC and, by extension, the possibility that these may play a role in vertebrate CpG depletion, too. However, it is striking that the small vertebrate DNA viruses all show TpG (CpA) overrepresentation (observed/expected values ranging from 1.06 to 1.35). Although Shackelton et al. [21] regarded these levels of TpG (CpA) excess as being so small as to make a hypothesis of methylation-mediated CpG deficiency questionable, they are in fact in the range expected from numerical analysis of dinucleotide evolution models (as discussed above). It is therefore tempting to speculate that these small DNA viruses, which rely on cellular machinery for replication, might indeed be subject to a degree of methylation, which in turn influences CpG and TpG (CpA) abundances. Even if this speculation turns out to be false, the relevance of CpG depletion in DNA viruses to vertebrate CpG depletion may be limited, as the (non-5mC-based) mechanism(s) of CpG depletion in such viruses could well be specific to the highly particular life cycles that they possess [21].

More generally, other (non-5mC-based) potential explanations of animal CpG depletion include ones based on the distortion of the DNA backbone that accompanies CpG dinucleotides embedded in particular sequence contexts [22,23] and regional selection arguments [10]. However, to date none of these hypotheses has proven capable of accounting for as much of the data concerning CpG, TpG, and TpA levels in vertebrates as the methylation hypothesis has.

The availability of complete animal genome sequences opens up the prospect of systematic analysis of dinucleotide abundance data and associated hypotheses. The current study presents analyses of the data relevant to evaluating the methylation hypothesis of CpG depletion, using both vertebrate and invertebrate genome sequence data. I first examine the dependence of dinucleotide abundances on the local G + C level. I then test whether local CpG deficit is correlated with local TpG excess and discuss the differences between the current results and those reported previously [24]. Finally, I examine the associations between CpG and other (non-TpG (CpA)) dinucleotides and in particular propose that the methylation hypothesis accounts for some features of the CpT abundance data. A set of parallel analyses using repeat-masked sequences is also discussed.

Results

Overall dinucleotide relative abundances

A selection of animal genome sequences was split into 50-kb segments and analyzed for dinucleotide content using a relative abundance measure (ρ) that reports the ratio of the dinucleotide frequency relative to the frequency expected from random association of the individual

Table 1
G+C content and relative abundance values of dinucleotides in eukaryotic genome sequences

Species	G+C%	ρ_{AT}	ρ_{TA}	ρ_{CG}	ρ_{GC}	$\rho_{AA\|TT}$	$\rho_{CC\|GG}$	$\rho_{TG\|CA}$	$\rho_{TC\|GA}$	$\rho_{CT\|AG}$	$\rho_{GT\|AC}$
Homo sapiens	40 (5)	0.87 (0.04)	0.74 (0.06)	0.22 (0.06)	1.01 (0.04)	1.11 (0.03)	1.23 (0.03)	1.21 (0.04)	0.99 (0.03)	1.17 (0.05)	0.84 (0.03)
Pan troglodytes	40 (4)	0.88 (0.03)	0.74 (0.05)	0.21 (0.06)	1.01 (0.04)	1.11 (0.03)	1.23 (0.02)	1.21 (0.03)	0.99 (0.03)	1.17 (0.04)	0.84 (0.03)
Mus musculus	41 (4)	0.86 (0.05)	0.74 (0.04)	0.18 (0.05)	0.93 (0.04)	1.07 (0.03)	1.19 (0.04)	1.23 (0.04)	1.03 (0.03)	1.22 (0.05)	0.88 (0.03)
Danio rerio	36 (1)	0.92 (0.03)	0.80 (0.04)	0.52 (0.08)	1.17 (0.06)	1.10 (0.03)	1.04 (0.05)	1.26 (0.04)	0.91 (0.03)	0.99 (0.04)	0.97 (0.04)
Takifugu rubripes	45 (2)	0.87 (0.03)	0.66 (0.04)	0.55 (0.11)	1.02 (0.05)	1.13 (0.04)	1.04 (0.03)	1.26 (0.05)	1.01 (0.03)	1.07 (0.03)	0.93 (0.03)
Ciona intestinalis	35 (1)	0.90 (0.03)	0.87 (0.04)	0.85 (0.14)	1.08 (0.07)	1.13 (0.03)	1.09 (0.06)	1.16 (0.05)	0.83 (0.03)	0.87 (0.03)	1.07 (0.03)
Drosophila melanogaster	42 (2)	0.97 (0.04)	0.76 (0.04)	0.93 (0.05)	1.27 (0.07)	1.21 (0.04)	1.05 (0.05)	1.13 (0.04)	0.91 (0.04)	0.89 (0.04)	0.86 (0.03)
Anopheles gambiae	44 (3)	0.92 (0.04)	0.72 (0.05)	1.06 (0.05)	1.14 (0.04)	1.23 (0.05)	0.97 (0.04)	1.13 (0.03)	0.95 (0.03)	0.85 (0.03)	0.97 (0.03)
Caenorhabditis elegans	35 (1)	0.85 (0.04)	0.61 (0.05)	0.99 (0.12)	1.06 (0.09)	1.30 (0.08)	1.06 (0.09)	1.08 (0.06)	1.09 (0.06)	0.89 (0.05)	0.85 (0.06)

Values represent means over the set of 50-kb sequence segments obtained for each genome; SD values in parentheses.

mononucleotides. The standard symmetrized ρ measure, based on counts on both strands, was employed (detailed under Materials and methods). Values of ρ_{XY} of > 1 or < 1 respectively indicate an excess or deficit of dinucleotide *XpY*. Statistical theory (cited in Ref. [25]) indicates that the thresholds for statistically significant over- or underrepresentation are 1.23 and 0.78, respectively. However, these are conservative thresholds [16,20], and the current work takes a more pragmatic approach, regarding ρ_{XY} values above 1.1 or below 0.9 as noteworthy.

The trends apparent in Table 1 broadly support those found in previous surveys, which were often limited to subgenomic scale datasets [16,26]. The CpG dinucleotide shows the most extremal relative abundance, with human, chimpanzee, and mouse having values close to 0.2. The pufferfish and zebrafish also exhibit strong CpG depletion, with $\rho_{CG} \approx 0.5$. The sea squirt, *Ciona intestinalis,* displays moderate depletion, with $\rho_{CG} = 0.85$. In insects and the nematode worm, CpG is present at essentially the expected level: $\rho_{CG} \approx 1$. This pattern is clearly in accord with the cytosine methylation hypothesis, as mammalian genomes are globally methylated, nonvertebrate chordate genomes (such as that of *C. intestinalis*) are partially methylated [27], and insects and other invertebrates show either a complete absence of cytosine methylation or, as in the fruitfly and mosquito, very weak levels limited to early development [28–30]. Why fish genomes, which are also globally methylated [31], show a less dramatic depletion of CpGs than mammalian genomes, is addressed later.

The TpA dinucleotide is clearly underrepresented in all the genomes surveyed, in line with previous studies [16]. As discussed earlier, this is generally thought to be due to structural factors at both the DNA and the RNA level.

Table 1 also shows that, in addition to an ≈21% excess of TpG (CpA), mammalian genomes also harbor comparable excesses of CpC (GpG) and CpT (ApG). While TpG remains in moderate excess in fish genomes and in the sea squirt, consistent with the interpretation that this stems from deamination of 5mCpG, the high ρ values for CpC and CpT appear limited to mammals. Potential causes of the high mammalian $\rho_{CC|GG}$ and $\rho_{CT|AG}$ values will be discussed later; here I simply observe, as did Karlin and Mrázek [16], that CpC, GpG, CpT, and ApG are the only possible outcomes of one-step cytosine mutations in CpG, apart from TpG and CpA.

The next objective is to examine in detail the association between CpG deficits and TpG (CpA) excess. To do this it is necessary to first consider the influence of the G + C level on the dinucleotide relative abundance values.

Variation of dinucleotide relative abundances with G + C level

Fig. 1 shows the variation of ρ_{CG}, $\rho_{TG|CA}$, and ρ_{TA} with G + C level over the sets of sequence segments for each genome (equivalent plots for all dinucleotides are in Supplementary File 1). Several features are evident:

(i) The three mammalian datasets are highly similar.
(ii) A positive correlation between ρ_{CG} and G + C level is seen, not just in mammals but also in the other chordate genomes subject to cytosine methylation. No such association is evident in the insect and nematode genomes analyzed, in which detectable cytosine methylation is either marginal or absent.
(iii) In the mammalian genomes, the TpG (CpA) excess increases with G + C level.

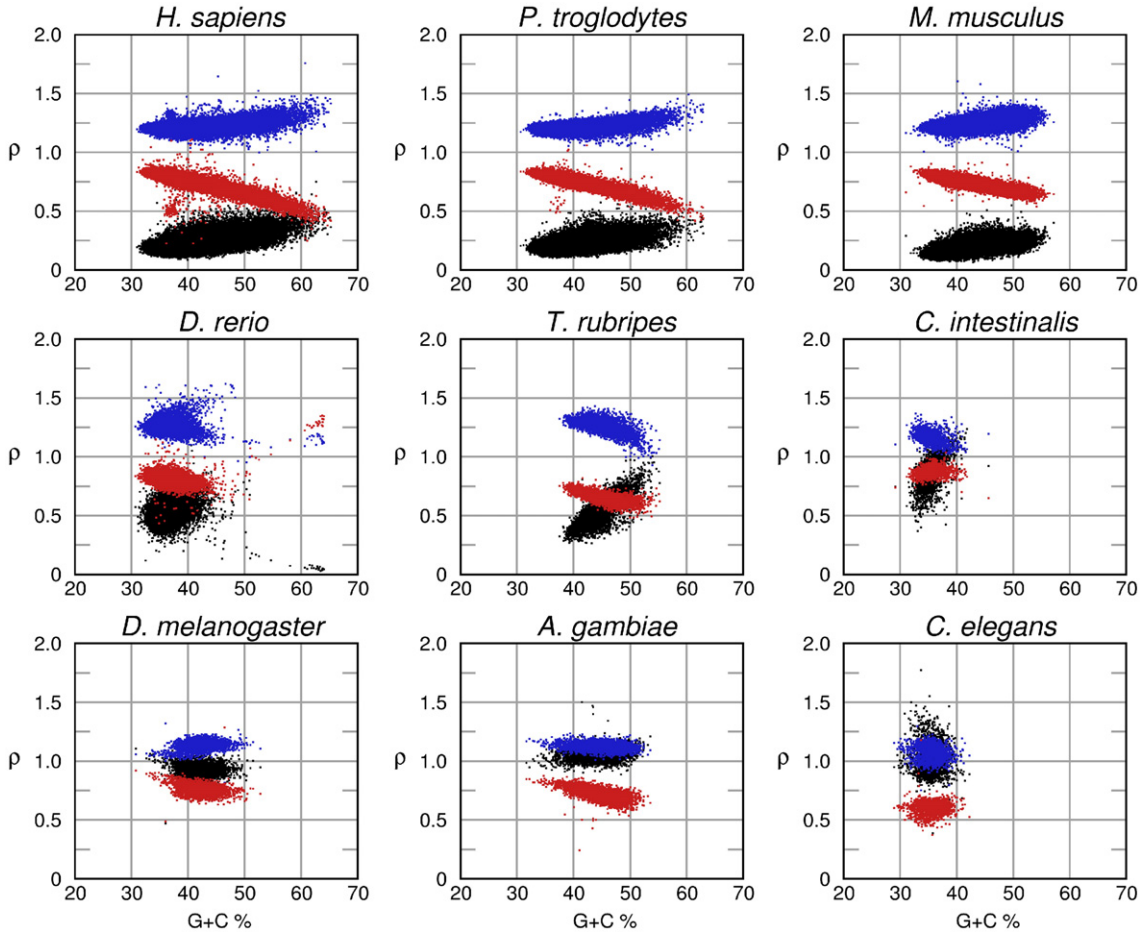

Fig. 1. Scatter plots of selected dinucleotide relative abundance (ρ) values, against G+C % in nine animal genomes. Each point represents the data from a single 50-kb genomic segment. Black dots represent ρ_{CG}, red dots ρ_{TA}, blue dots $\rho_{TG|CA}$. Note that the ρ_{TA} and $\rho_{TG|CA}$ points are plotted over the ρ_{CG} points, so that portions of the ρ_{CG} distributions are obscured for *T. rubripes, C. intestinalis, A. gambiae,* and *C. elegans*.

(iv) The fish and sea squirt profiles differ from the mammalian ones in two ways: they display steeper ρ_{CG} vs G + C trends and they also show negative $\rho_{TG|CA}$ vs G + C trends (for zebrafish this is not immediately clear from the figure but is confirmed by regression analyses, data not shown).

(v) TpA depletion increases with G + C level in all the species surveyed, with the exception of the sea squirt.

Clearly an assessment of the 5mC deamination hypothesis (or any hypothesis concerning what shapes dinucleotide content) that rests on examining associations between the relative abundance of CpG and other doublets must first take into consideration point (ii), i.e., that ρ_{CG} shows a strong dependence on G + C level in methylated genomes. For example, in mammals the greatest $\rho_{TG|CA}$ excess occurs when ρ_{CG} is nearest 1 (see point (iii) above): at first glance this may appear at odds with the 5mC deamination hypothesis. In fact it is not, as shown below by informal argument and also by a simple mathematical model.

The key lies in recalling the definition of the relative abundance measure, ρ_{XY} = observed *XpY* proportion/expected *XpY* proportion. The proportion of CpG dinucleotides expected from random association of mononucleotides grows as the square of the G + C level, whereas the expected proportion of TpG and CpA is a quadratic function of G + C level but with a maximum at G + C = 50%. Depletion of a proportion of the mCpGs by deamination would be accompanied by a concomitant increase in the TpG (CpA) proportion, with the magnitude of this increase growing with G + C level. This, combined with the background of a stationary and then decreasing expected proportion of TpG (CpA) (as G + C moves up through 50%) would therefore cause $\rho_{TG|CA}$ to rise over the G + C range observed in mammals.

Formally, it is possible to predict the key elements of the dependence of $\rho_{TG|CA}$ on G + C level in methylated genomes using a simple mathematical model (see Materials and methods). Figs. 2A and B illustrate the behavior of the model, as applied to human genome data. The model $\hat{\rho}_{TG}$ relative abundance shows a positive gradient with G + C level, in accord with the experimental $\rho_{TG|CA}$ values. The match to data is not perfect, but that is not surprising given the simplicity of the model.

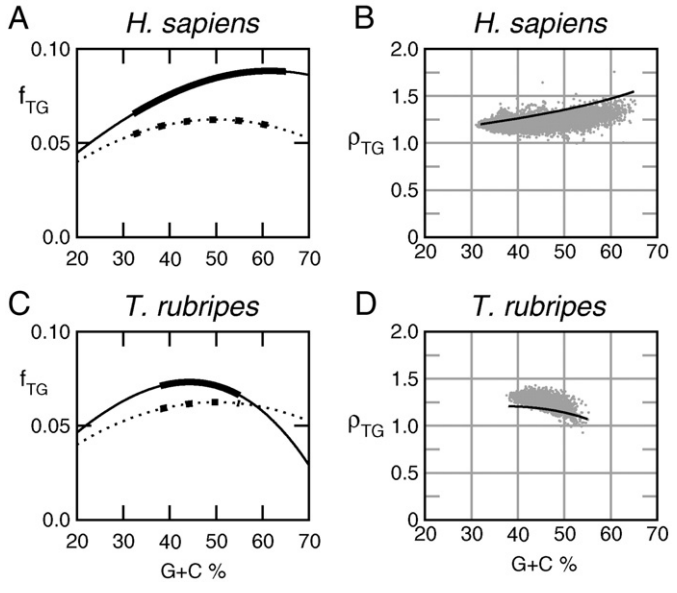

Fig. 2. Analysis of the variation in TpG content with G+C level. (A) Expectations of the proportion of TpG dinucleotides as a function of G+C level. The dotted curve plots the expectation when methylation is absent. The solid curve plots the expectation under the simple model of methylation described in the text, using κ fitted to the human ρ_{CG} data. Thickened portions of the curves indicate the G+C range relevant to the human data. (B) Model TpG relative abundance, $\hat{\rho}_{TG}$, computed under the methylation model for human (solid line). For comparison the observed human $\rho_{TG|CA}$ data values are replotted from Fig. 1 (gray dots). (C and D) Same as (A) and (B) but for *T. rubripes* data.

This analysis also explains the observation (point (iv) above) that in the fish and sea squirt results the association between $\rho_{TG|CA}$ and G + C level has flipped sign to become negative. In these genomes ρ_{CG} rises particularly steeply (for reasons as yet unknown) over the fairly narrow band of G + C values found in the sets of 50-kb genomic segments. Under the methylation model, this rapid decrease in the degree of CpG depletion leads to a situation with two opposing tendencies. On one hand, as discussed above, an increasing G + C level produces a statistical tendency for greater numbers of CpGs and hence greater numbers of TpGs resulting from mCpG deaminations. But on the other hand, as the degree of CpG depletion decreases sharply as G + C increases, this will logically tend to decrease the supply of TpGs arising from mCpG deaminations. The net effect is that, in the low–moderate G + C range found in these genomes, the proportion of TpG (CpA) tends to peak at G + C levels below 50%. This is illustrated in Figs. 2C and D, which show the behavior of the model as applied to pufferfish data. The modeled $\hat{\rho}_{TG}$ relative abundance falls as G + C level increases, as also seen in the observed data.

Correlation between CpG deficit and TpG (CpA) excess

A central premise of the methylation hypothesis is that in methylated genomes depletion of CpGs is accompanied by the formation of TpG or CpA dinucleotides. A negative correlation is therefore predicted between $\rho_{TG|CA}$ and ρ_{CG}. Here I examine the human data in detail and then summarize for the other species studied. Fig. 3A shows the result of a naive scatter plot of $\rho_{TG|CA}$ and ρ_{CG}: the two quantities are essentially uncorrelated (R^2 = 0.01). However, this is misleading, because as discussed above, both quantities are influenced by the local G + C level. To give an adequate test, the effect of G + C content needs to be controlled for. This can be done by looking at the correlation between $\rho_{TG|CA}$ and ρ_{CG} within a set of sequences that have similar G + C contents. To do this the sequences are split into distinct subsets, each corresponding to a specific 2.5%-wide interval of the G + C range. Fig. 3B illustrates the result by showing the scatters for sequences lying in two typical G + C % intervals. Both these subsets show strong negative correlations between $\rho_{TG|CA}$ and ρ_{CG}; the other 2.5% G + C-wide intervals covering the range from 30% G + C up to 62.5% G + C also show negative *R* values (range −0.33 to −0.73) (unpublished data). This demonstrates that there is a robust negative correlation between CpG deficit and TpG (CpA) excess in local (50-kb) regions within the human genome. The overall scale of the correlation can also be quantified by a process of mean-centering the scatters associated with each G + C % interval and then aggregating (see Materials and methods). The outcome, depicted in Fig. 3C, essentially reveals the underlying correlation between $\rho_{TG|CA}$ and ρ_{CG}, after removing the effect of G + C level: again, a strong negative correlation is evident ($R = -0.57$, $p < 0.001$).

Equivalent analyses for the other species showing CpG depletion show similar results, the correlation coefficients from the mean-centered analyses being chimpanzee (−0.60), mouse (−0.41), zebrafish (−0.53), pufferfish (−0.76), and sea squirt (−0.82). It is notable that the correlations are always negative, irrespective of whether $\rho_{TG|CA}$ is positively (mammals) or negatively (fish and sea squirt) associated with G + C level, as discussed earlier. Thus these genome-scale analyses show that local CpG deficit is consistently associated with a local excess of TpG (CpA) in animal genomes subject to cytosine methylation.

Equivalent correlation analyses were also performed for *Drosophila, Anopheles,* and *Caenorhabditis elegans*. These also show negative correlation between the relative abundances of CpG and TpG (CpA), with the *R* values from the mean-centered analyses being −0.51, −0.45, and −0.81, respectively. Given the lack of any appreciable CpG deficit in these species (see Table 1 and Fig. 1), the biological significance of these correlations is questionable (see Discussion).

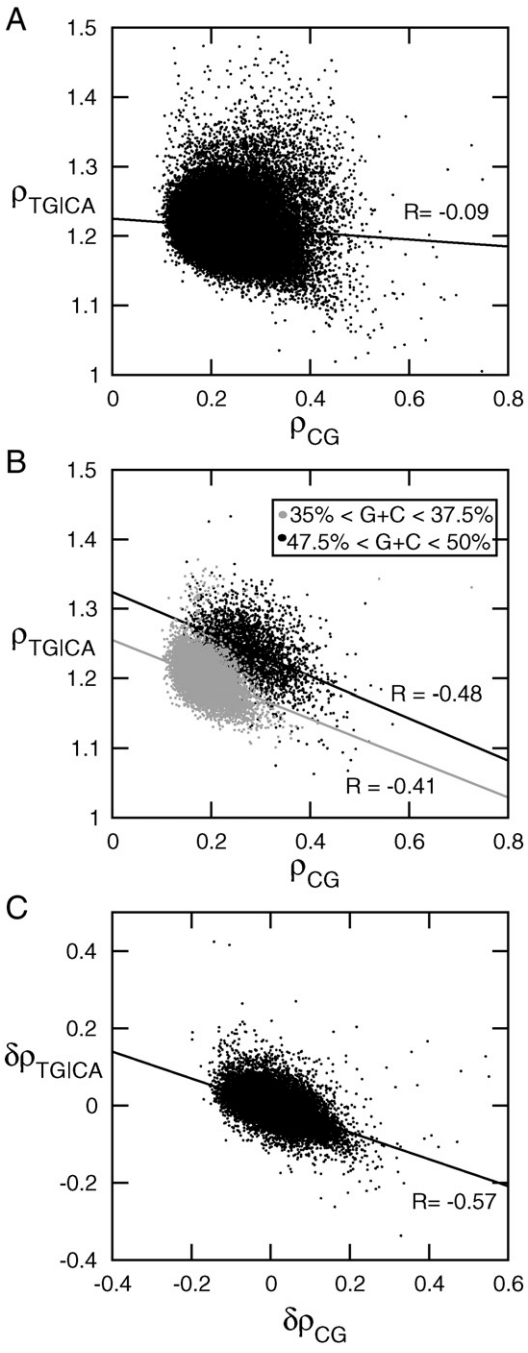

Fig. 3. Correlation of ρ_{CG} and $\rho_{TG|CA}$ in human. Each point represents the data from a single 50-kb genomic segment. Regression lines are shown, *R* values are correlation coefficients. (A) Data from all 56,759 genomic segments. (B) Subsets of the data from two distinct 2.5%-wide intervals of the G+C range. (C) Aggregation of all the data after splitting into 2.5%-wide G+C content intervals and mean-centering the data from each subset. Each data subset is plotted with its centroid placed at the origin, with each individual data point's *x* and *y* coordinate values representing the deviations (denoted $\delta\rho$) from the centroid of each data subset.

Correlations between CpG and non-TpG (CpA) dinucleotide abundances

Depletion of CpG through 5mC → T transitions affects not just the counts of CpG and TpG (CpA) but also the counts of other dinucleotides through alterations in the neighboring doublets in the N-C-G-N context. By proper accounting of all possible cases, the statistical expectations for these second-order changes to counts can be deduced. Consider for example the dinucleotide TpC and its reverse complement GpA. With probability f_T the CpG will be embedded in a TCG triplet. Independently, the CpG will be embedded in a CGA triplet with probability f_A. As the deamination event can occur on either strand, the net change to the symmetrized count n_{TC} (which tracks TpC number on the forward strand plus reverse complement, see Materials and methods) is expected to equal $-0.5(f_A + f_T)$. In principle such changes should often induce correlations between ρ_{CG} and the ρ value for the dinucleotide concerned. Table 2 shows the statistical expectations and the sign of the theoretical correlations with ρ_{CG}. Unsurprisingly, the expectations are functions of G + C level. Taking a G + C level of 50% to get a working estimate shows these changes are approximately fourfold smaller than the first-order change to the TpG (CpA) count; intuitively this is because each flanking site could be one of the four bases.

For comparison, Table 2 also gives the experimentally derived correlation coefficients from analyses of mean-centered datasets from human and mouse (chimpanzee gives similar results), performed in the same manner as described above for TpG (CpA). There is clearly no systematic matching between the observed and the theoretical correlations, which suggests that these second-order effects are generally too slight to be robustly detected.

Table 2 also shows that, for human, although TpG (CpA) has the strongest correlation with CpG, other dinucleotides have correlations with CpG almost as strong, notably CpT and ApA. For CpT the correlation is negative, i.e., local CpT excess correlates with local CpG deficit. Although this is in line with the sign of the predicted second-order effect, it would be inconsistent to attribute the correlation to this, given the lack of systematic evidence for these second-order effects in the other dinucleotides. An alternative speculation (discussed below) is that the CpT data reflect variability in the rate of CpG mutation depending on the trinucleotide within which CpG is embedded.

Robustness to masking of repetitive sequences

All of the analyses described above were performed on unmasked genomic sequences. To assess the possibility that some of the results may be attributable to the fractions of repetitive DNA in the genomes concerned, all the analyses were redone using repeat-masked versions of the same genomic sequences (see Materials and methods). The overall dinucleotide relative abundances for the repeat-masked genomes (Supplementary File 2) are very similar to the equivalent values for unmasked sequence: for human, the mean absolute difference between the masked and the unmasked ρ values over the set of 10 possible dinucleotides was only 0.014, with the maximum absolute difference being 0.026 (the equivalent values over the complete set of nine genomes being 0.015 and 0.07, respectively).

Table 2
Theoretical consequences of a single CpG → TpG event and comparison with human and mouse data

Dinucleotide *XpY*[a]	Expected change in count of *XpY*[b,c]	Sign of predicted correlation between ρ_{CG} and ρ_{XY}	Observed correlation *R* in human[d,e]	Observed correlation *R* in mouse[e]
CpG	−2			
TpG (CpA)	+1	−	−0.57 (−0.54)	−0.41 (−0.57)
CpT (ApG)	$0.5f_{G+C}$	−	−0.56 (−0.66)	−0.41 (−0.48)
TpC (GpA)	$-0.5f_{A+T}$	+	−0.44 (−0.44)	−0.44 (−0.45)
GpT (ApC)	$0.5(f_{G+C} - f_{A+T})$	Varies	+0.13 (+0.25)	+0.19 (+0.24)
ApA (TpT)	$0.5f_{A+T}$	−	+0.55 (+0.57)	+0.51 (+0.63)
CpC (GpG)	$-0.5f_{G+C}$	+	−0.16 (−0.12)	−0.21 (−0.19)
GpC	$-0.5f_{G+C}$	+	+0.48 (+0.33)	+0.44 (+0.39)
ApT	$0.5f_{A+T}$	−	−0.17 (−0.09)	−0.33 (−0.36)
TpA	0	0	+0.25 (+0.33)	+0.21 (+0.28)

[a] Reverse complement dinucleotide given in parentheses.
[b] Using strand-symmetrized count n_{XY}, as defined under Materials and methods.
[c] f_{G+C} denotes $f_C + f_G$, i.e., the G + C content expressed as a fraction between 0 and 1.
[d] From scatter plot aggregating all mean-centered G + C interval subsets.
[e] Values in parentheses are from analyses using repeat-masked sequences (see main text).

Scatter plots showing ρ_{CG}, $\rho_{TG|CA}$, and ρ_{TA} vs G + C level over the sets of repeat-masked sequence segments for each genome (Supplementary File 3) are also very similar to the equivalent plots for unmasked sequence (Fig. 1). Only two minor differences were found. First, in the repeat-masked versions the data distributions are slightly more dispersed; this will be due to the fact that the repeat-masked 50-kb segments can contain far fewer than 50,000 valid bases, thus making extremal values more likely. Second, several "outlier" clusters evident in the unmasked plots (human ρ_{TA} for G + C ≈37%, and zebrafish ρ_{CG}, ρ_{TA}, and $\rho_{TG|CA}$ for G + C ≥ 55%) are absent when repeat-masked data are analyzed, indicating that these features were attributable to features of repetitive DNA. The comparison of observed $\rho_{TG|CA}$ values with $\hat{\rho}_{TG}$ values derived on a semiempirical basis under a methylation model (Fig. 2) showed little change when performed with repeat-masked sequences (data not shown).

Similarly, removal of the repeat-derived portions of the genomes has only minor quantitative effects on the correlations measured between relative dinucleotide abundances. A strong negative correlation is again found between ρ_{CG} and $\rho_{TG|CA}$ in human after removing the effect of G + C level ($R = -0.54$, $p < 0.001$), as shown in Fig. 4, which is the repeat-masked equivalent of Fig. 3C. The correlations between CpG and other non-TpG/CpA dinucleotide abundances for repeat-masked human and mouse data are shown in Table 2 alongside the equivalent unmasked values—no major differences are apparent.

In summary, the results obtained are robust to masking of repetitive sequences, indicating that the data features found are not attributable to the compositional characteristics of repetitive sequences.

Discussion and conclusions

Three key points emerge from the examination of how the G+C content is associated with trends in the dinucleotide abundances (Fig. 1). A positive association between ρ_{CG} and G+C level is observed in all the genomes subject to significant cytosine methylation but not in the others. This is consistent with the evidence that such an association is due to DNA melting constituting the rate-limiting step in 5mC deamination *in vivo* [11]. In contrast, the $\rho_{TG|CA}$ values show a positive association with G+C level in mammals, but a negative association in fish and a urochordate. Importantly, both of these $\rho_{TG|CA}$ trends are explicable in terms of the corresponding CpG trends (Fig. 2), illuminating an intimate connection between CpG deficit and TpG excess, as would be expected under a methylation-driven process. Finally, the existence of these associations with G+C level demonstrates the necessity of taking them into account when analyzing correlations between abundances of dinucleotides.

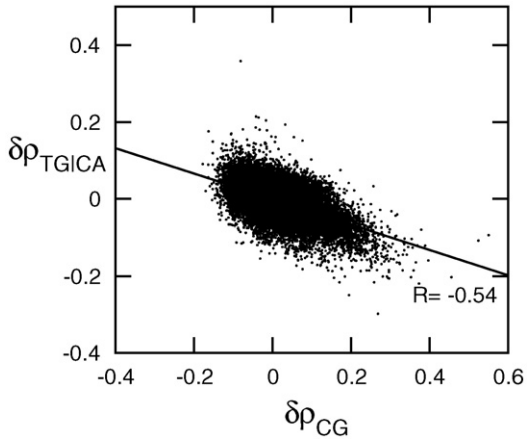

Fig. 4. Correlation of ρ_{CG} and $\rho_{TG|CA}$ in repeat-masked human sequence after splitting the data into 2.5%-wide G+C intervals and mean-centering the data from each subset. Each point represents the data from a single 50-kb genomic segment (56,516 segments). Other details are as described for Fig. 3C.

Direct examination of the (ρ_{CG}, $\rho_{TG|CA}$) data, conditioned on G+C value, reveals a strong negative correlation between CpG deficit and TpG excess in methylated genomes (Fig. 3), contradicting the claim of Jabbari and Bernardi [24] that such a correlation was absent. Those authors reached their conclusion indirectly, through interpretations of plots of CpG and TpG (CpA) proportions against G+C level. That previous work was probably unable to detect the correlation because it neither took the covariation of both abundances with G+C content directly into account nor employed a dinucleotide abundance measure suitably normalized with respect to the expected counts. In agreement with earlier analyses based on physical approaches [7], TpG (CpA) is also found to be the doublet showing the strongest correlation with CpG in human.

A perhaps surprising result is that a negative correlation is also found between ρ_{CG} and $\rho_{TG|CA}$ in the unmethylated nematode genome and very lightly methylated arthropod genomes. Furthermore even the low levels of 5-methylcytosine that are detected in *Drosophila melanogaster* and other drosophilids, as well as in *Anopheles* [29,30], cannot explain this correlation as, surprisingly, the 5mC sites are found mainly in non-CpG dinucleotides. There is, however, evidence that the negative correlation in the nematode is not robust. This comes from equivalent correlation analyses conducted using alternative relative abundance measures, namely CpG/GpC count ratio and TpG/GpT count ratio. No correlation is found between CpG and TpG abundances for *C. elegans* using this alternative measure ($R = 0.01$), whereas for the other eight species the R values are broadly similar to their values under the original abundance measure (data not shown).

In itself, this does not throw any light on whether the methylation model adequately accounts for the data in methylated genomes, but it does demonstrate that the existence of a negative correlation alone between ρ_{CG} and $\rho_{TG|CA}$ cannot be adduced as conclusive evidence for the methylation hypothesis. Instead, I contend that strong evidence for the role of methylation in shaping dinucleotide content within a particular genome lies in the clustering of four indicative features:

(1) global CpG deficit;
(2) global TpG (CpA) excess;
(3) positive correlation between local CpG relative abundance and G+C level;
(4) negative correlation between local relative abundances of CpG and TpG (CpA).

The mammalian, fish, and sea squirt genomes display this clustering.

The observation that fish show less CpG depletion than mammals was also made by Jabbari and Bernardi [24]. Those authors also implied that this demonstrated the lack of a correlation between methylation and CpG depletion, based on their earlier claim that fish have a higher methylation level than mammals [32]. The measure of methylation level reported was the percentage of bases found to be 5mC. Given that in vertebrates 5mC is almost exclusively found in CpG doublets, the 5mC percentage will inevitably be highly dependent on the level of CpG depletion, as well as the per-CpG-site degree of methylation. The twofold difference between the 5mC levels of 1.70 and 0.88% for fishes and mammals, respectively [32], can most simply be explained by the approximately twofold difference in CpG proportions between mammals and fish (see ρ_{CG} data in Table 1), coupled with a common per-site degree of methylation. Evidence for the latter is the observation that on average 60–90% of CpGs in mammalian and fish genomes are methylated [27]. The claim that the fish–mammal comparison demonstrates the lack of a correlation between CpG methylation and CpG depletion is therefore based on an inappropriate measure of methylation level.

Nevertheless, this still leaves open the question as to why fish genomes show less depletion of CpGs than mammalian genomes. At least two, nonexclusive, explanations are possible. First, it may flow from the higher body temperature in mammals, as analysis of the

kinetics of deamination of 5-methylcytosine in double-stranded DNA [33] shows that a ΔT of −10 °C implies a threefold drop in deamination rate. Second, the repair enzymes that attempt to correct the T/G mismatches arising from deamination events [33,34] may perhaps have differing efficiencies in fish and mammals. A further open question concerning fish genomes is why they display such a steep rise in ρ_{CG} over the narrow range of G+C level found within them.

Earlier analyses using partial genome sequence found that CpT (ApG) and CpC (GpG) were also overrepresented in human by essentially the same degree as TpG (CpA) [26]. The current study both confirms that these observations hold true in the complete human sequence (Table 1) and shows for the first time that CpT excess is correlated with CpG deficit in human and mouse (shown by the strong negative correlation between ρ_{CG} and $\rho_{CT|AG}$ in Table 2). I propose that both of these features of the CpT data may be explicable in terms of sequence context effects on the mutation rate of the mCpG dinucleotide.

Many studies have already demonstrated specifically enhanced rates of cytosine mutation in the context 5′-YCG-3′, Y being a pyrimidine. In a study of the sequence context of point mutations in primate pseudogenes, TCG → TTG and CCG → CTG were found to be the most statistically overrepresented triplet transitions [35]. Analysis of point mutation hot spots in human hereditary disorders also shows a clear bias toward YCG → YTG transitions [36] as do studies in transgenic mice [37]. Genome-wide surveys of single-nucleotide polymorphisms also reveal this strong bias, in both human [38] and mouse [39]. The doublet products of such transitions, CpT (ApG) and TpT (ApA), would therefore be expected to be overrepresented in mammalian genomes. Interestingly, the study of Blake et al. [35] also showed that while transversions are rarer than transitions, the most statistically overrepresented triplet transversion was CGG → CTG (129% more common than expected), which would further lift the abundance of CpT (ApG). Thus the mammalian CpT data are consistent with the known enhanced rate of cytosine mutation in the triplet context 5′-YCG-3′.

Some aspects of vertebrate dinucleotide landscapes, however, are not fully explained by the methylation hypothesis. For example, although they lie below the equivalent values for heavily methylated genomes, the TpG observed/expected ratios for the nematode, fruitfly, and mosquito are all still greater than 1 (1.08, 1.13, and 1.13, respectively) and quite close to the value of 1.16 found in the partially methylated sea squirt genome. As mentioned above, the discovery of low levels of 5mC in drosophilids as well as in *Anopheles* [29,30] does not provide an explanation for the insect data, as the methylation is mainly in non-CpG sites. It remains possible therefore that there is some weak non-methylation-dependent process acting to enhance TpG (CpA) levels slightly across animal genomes.

Regarding the ApA data, the mutation analyses mentioned above [35–37] imply that ApA will be overabundant in mammalian genomes, and this is confirmed here (see Table 1). However, the reason(s) the local relative abundances of ApA and CpG should be positively correlated in the human genome (Table 2) is unclear. Poly(A) tracts are frequently found in human *Alu* elements, which occur at higher densities in G+C-rich regions [3]. However, the correlation persists when the repeat elements are disregarded (Table 2), so *Alu* content cannot be the cause. It is quite likely that structural factors play a greater role in influencing the statistics and distribution of ApA than do secondary effects of 5mC mutation (which would predict a negative correlation between ApA and CpG). For example, arrays of ApAs are believed to be important in establishing the regular arrangement of nucleosomes in eukaryotic chromatin due to the high intrinsic bendability of ApA [40]. Furthermore, polypurine and polypyrimidine tracts display markedly low mutation rates [35], which may also contribute to the observed overrepresentation of ApA (TpT) and CpC (GpG) across all eukaryotes (see Table 1), not just those with cytosine methyltransferase activity.

A fully satisfactory quantitative understanding of animal dinucleotide content may also need to take into account the fact that mutagenesis at mCpG sites can occur, not just via spontaneous deamination leading to C → T transitions, but also via other processes including endogenous and exogenous mutagens that are known to produce products other than TpG (CpA) [34]. Such substitutions are observed in various tumors, but their influence in normal evolutionary processes is currently unclear. Furthermore, there is also the structural issue that the presence of mCpG can in some local sequence contexts distort the DNA backbone leading to an intrinsically enhanced susceptibility to mutation, not necessarily by deamination [22]. Future analyses of CpG mutation spectra using SNP data could help to assess the impact of these processes, as could further detailed structural studies. In addition, improved understanding of the relative importance of various processes driving G+C content evolution—e.g., selection pressure [41] or biased gene conversion [42]—will also provide insights into dinucleotide content evolution.

These points notwithstanding, the hypothesis of methylation-mediated CpG depletion, which is founded in established biochemical mechanisms, already has the virtue of providing a parsimonious explanation of many features of the dinucleotide data in animal species showing global or partial methylation. Mutational pressure on methylated cytosines therefore appears likely to have been a major force in the evolution of dinucleotide content in vertebrates.

Materials and methods

DNA sequences

Sequence data were obtained for the following species from the Ensembl database (release 37, ftp.ensembl.org): *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), *Danio rerio* (zebrafish), *Takifugu rubripes* (pufferfish), *Ciona intestinalis* (sea squirt), *Drosophila melanogaster* (fruitfly), *Anopheles gambiae* (mosquito), and *Caenorhabditis elegans* (nematode worm). For *Takifugu*, all available genomic scaffold sequences were used. For the other species, all the assigned chromosomal sequences were used. Each chromosomal (or scaffold) sequence was then split into segments, and the base composition and dinucleotide relative abundances were computed within each segment. A segment size of 50 kb was used, as in prior studies of genome composition [15,22,23]. Those segments with over 10% ambiguous bases were discarded from further analysis. The final numbers of 50-kb segments and total number of bases analyzed are human, 56,759, 2.84 Gb; chimpanzee, 38,180, 1.85 Gb; mouse, 48,261, 2.40 Gb; zebrafish, 22,354, 1.11 Gb; pufferfish, 5406, 268 Mb; sea squirt, 1599, 79 Mb; fruitfly, 2484, 124 Mb; mosquito, 4273, 212 Mb; and nematode, 2002, 100 Mb.

Repeat-masked versions of the same genomic sequence data were also obtained from Ensembl. As above, a segment size of 50 kb was used, with segments having over 10% ambiguous bases being discarded. In addition, segments having under 5000 bases neither marked as being in a repeat nor ambiguous were also discarded from further analysis—this was done to prevent the possibility of very heavily masked segments with low valid counts contributing extremal data values. The final numbers of 50-kb segments and total number of nonrepeat bases analyzed are human, 56,516, 1.49 Gb; chimpanzee, 38,142, 1.00 Gb; mouse, 48,255, 1.41 Gb; zebrafish, 22,305, 558 Mb; pufferfish, 5406, 246 Mb; sea squirt, 1599, 63 Mb; fruitfly, 2484, 118 Mb; mosquito, 4273, 185 Mb; and nematode, 2002, 90 Mb.

Dinucleotide relative abundance

I measured the abundance of dinucleotides relative to their abundances expected on the basis of the observed mononucleotide composition. As the focus was on genome-wide patterns, rather than just the patterns pertaining to coding sequences, it was appropriate to use the standard symmetrized approach, introduced by Burge et al. [43] and subsequently employed in many studies [16,20,26,44], which takes into account abundances on both strands. Specifically, let n_X denote the count of nucleotide X (A, C, G, or T) and n_{XY} the count of dinucleotide XpY in the sequence (of N bases) concatenated with its reverse complement. The corresponding proportions are denoted f_X and f_{XY}. The expected proportion of dinucleotide XpY is then $f_X f_Y$ and the relative abundance value given by $\rho_{XY}=f_{XY}/f_X f_Y$. Values of $\rho_{XY}>1$ or <1 indicate an excess or deficit of XpY, respectively. The effect of using symmetrized ratios is that, for doublets that are not self-complementary, the relative abundance of both the doublet and its reverse complement is assessed in a single value, for example, $\rho_{TG}=\rho_{CA}$ and, for simplicity, I let it be denoted $\rho_{TG|CA}$ (where "TG|CA" has the sense "TpG or CpA").

Mathematical model of $\rho_{TG|CA}$ dependence on G+C level

The proportion of TpG dinucleotides expected in the absence of any methylation effects, $f_{\text{null TG}}$, is equal to $\theta(1-\theta)/4$, where θ is the G+C proportion. For CpG, the

expected proportion in the absence of methylation effects is equal to $\theta^2/4$. The effect of methylation is taken to be the loss of a proportion (κ) of the CpG doublets, with half of these creating TpG doublets (for simplicity only one strand will be considered here), leading to a fraction, $\kappa\theta^2/8$, of all dinucleotides being TpGs created via 5mC mutation. The overall proportion of TpG under this methylation model is therefore $\theta^2(\kappa - 2)/8 + \theta/4$. It follows that the ratio of the TpG proportion expected under the methylation model to that expected ($f_{\text{null TG}}$) in the absence of methylation equals $(1 - \theta + \kappa\theta/2)/(1 - \theta)$. This ratio, here denoted $\hat{\rho}_{TG}$, is the model equivalent of the observed relative abundance $\rho_{TG|CA}$. I set an appropriate κ from the observed level of CpG depletion in each genome: $\kappa = 1 - \rho_{CG}$. To take account of ρ_{CG}'s dependence on G+C level (θ), I model it as a linear function: $\rho_{CG} = \alpha\theta + \beta$, where α and β are the regression coefficients obtained from the experimental ρ_{CG} datasets (as in Fig. 1).

Data analysis

To assess the association between a pair of dinucleotide relative abundance values (ρ_1 and ρ_2) in a particular genome, removing any influence of G+C level, the complete set of (ρ_1, ρ_2) points (one per sequence segment) is first partitioned into subsets according to the G+C level of each sequence. Subsets are created for each 2.5%-wide interval of the G+C range observed for the particular species under study, e.g., for human from 25 to 65%. Data for the subset of sequence segments lying in each G+C interval can then be plotted on a ρ_1 vs ρ_2 scatter plot, if desired, and the correlation computed. To analyze the global association between ρ_1 and ρ_2, independent of any G+C level effect, each subset of (ρ_1, ρ_2) points from sequences lying in a particular G+C interval is first mean-centered, i.e., the coordinate frame is shifted so that the centroid lies at the origin. After the data from each G+C interval are individually mean-centered, all the data subsets can be aggregated and visualized on a scatter plot and the correlation computed. Pearson correlation coefficients, denoted *R*, are used. Comparison studies using narrower, 1%-wide G+C intervals gave very similar results, e.g., over the nine genomes studied, the maximum deviation between *R* values computed with 2.5%- or 1%-wide G+C intervals for correlation between ρ_{CG} and $\rho_{TG|CA}$ was only 0.02 (mean absolute deviation 0.007).

Acknowledgments

I thank Susan Tweedie and Adrian Bird for discussions and comments and Mike Shipston for additional comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.03.009.

References

[1] M.N. Swartz, A. Kornberg, T.A. Trautner, Enzymatic synthesis of deoxyribonucleic acid. 11. Further studies on nearest neighbor base sequences in deoxyribonucleic acids, J. Biol. Chem. 237 (1962) 1961–1967.
[2] G.J. Russell, P.M.B. Walker, R.A. Elton, J.H. Subaksharpe, Doublet frequency-analysis of fractionated vertebrate nuclear-DNA, J. Mol. Biol. 108 (1976) 1–20.
[3] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[4] Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[5] W. Salser, Globin messenger-RNA sequences—analysis of base-pairing and evolutionary implications, Cold Spring Harbor Symp. Quant. Biol. 42 (1977) 985–1002.
[6] C. Coulondre, J.H. Miller, P.J. Farabaugh, W. Gilbert, Molecular-basis of base substitution hotspots in Escherichia-coli, Nature 274 (1978) 775–780.
[7] A.P. Bird, DNA methylation and the frequency of CpG in animal DNA, Nucleic Acids Res. 8 (1980) 1499–1504.
[8] D. Cooper, H. Youssoufian, The CpG dinucleotide and human genetic disease, Hum. Genet. 78 (1988) 151–155.
[9] C. Jiang, Z. Zhao, Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms, Genomics 88 (2006) 527–534.
[10] B. Aissani, G. Bernardi, CpG islands, genes and isochores in the genomes of vertebrates, Gene 106 (1991) 185–195.
[11] K.J. Fryxell, E. Zuckerkandl, Cytosine deamination plays a primary role in the evolution of mammalian isochores, Mol. Biol. Evol. 17 (2000) 1371–1383.
[12] K.J. Fryxell, W.J. Moon, CpG mutation rates in the human genome are highly dependent on local GC content, Mol. Biol. Evol. 22 (2005) 650–658.
[13] Z. Zhao, C. Jiang, Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions, Mol. Biol. Evol. 24 (2007) 23–25.
[14] L. Duret, N. Galtier, The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact, Mol. Biol. Evol. 17 (2000) 1620–1625.
[15] J. Sved, A. Bird, The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 4692–4696.
[16] S. Karlin, J. Mrázek, Compositional differences within and between eukaryotic genomes, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 10227–10232.
[17] E. Beutler, T. Gelbart, J.H. Han, J.A. Koziol, B. Beutler, Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage, Proc. Natl. Acad. Sci. U. S. A. 86 (1989) 192–196.
[18] R. Hanai, A. Wada, The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome, J. Mol. Evol. 27 (1988) 321–325.
[19] L.R. Cardon, C. Burge, D.A. Clayton, S. Karlin, Pervasive CpG suppression in animal mitochondrial genomes, Proc. Natl. Acad. Sci. U. S. A. 91 (1994) 3799–3803.
[20] S. Karlin, W. Doerfler, L.R. Cardon, Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. 68 (1994) 2889–2897.
[21] L.A. Shackelton, C.R. Parrish, E.C. Holmes, Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses, J. Mol. Evol. 62 (2006) 551–563.
[22] A. Lefebvre, O. Mauffret, E. Lescot, B. Hartmann, S. Fermandjian, Solution structure of the CpG containing d(CTTCGAAG)2 oligonucleotide: NMR data and energy calculations are compatible with a BI/BII equilibrium at CpG, Biochemistry 35 (1996) 12560–12569.
[23] Y. Wang, F.C.C. Leung, DNA structure constraint is probably a fundamental factor inducing CpG deficiency in bacteria, Bioinformatics 20 (2004) 3336–3345.
[24] K. Jabbari, G. Bernardi, Cytosine methylation and CpG, TpG (CpA) and TpA frequencies, Gene 333 (2004) 143–149.
[25] S. Karlin, L.R. Cardon, Computational DNA sequence analysis, Annu. Rev. Microbiol. 48 (1994) 619–654.
[26] A.J. Gentles, S. Karlin, Genome-scale compositional comparisons in eukaryotes, Genome Res. 11 (2001) 540–546.
[27] S. Tweedie, J. Charlton, V. Clark, A. Bird, Methylation of genomes and genes at the invertebrate–vertebrate boundary, Mol. Cell. Biol. 17 (1997) 1469–1475.
[28] A. Regev, M.J. Lamb, E. Jablonka, The role of DNA methylation in invertebrates: developmental regulation or genome defense? Mol. Biol. Evol. 15 (1998) 880–891.
[29] F. Lyko, B.H. Ramsahoye, R. Jaenisch, DNA methylation in Drosophila melanogaster, Nature 408 (2000) 538–540.
[30] J. Marhold, N. Rothe, A. Pauli, C. Mund, K. Kuehle, B. Brueckner, F. Lyko, Conservation of DNA methylation in dipteran insects, Insect Mol. Biol. 13 (2004) 117–123.
[31] D. Macleod, V.H. Clark, A. Bird, Absence of genome-wide changes in DNA methylation during development of the zebrafish, Nat. Genet. 23 (1999) 139–140.
[32] K. Jabbari, S. Cacciò, J.P. Païs de Barros, J. Desgrès, G. Bernardi, Evolutionary changes in CpG and methylation levels in the genome of vertebrates, Gene 205 (1997) 109–118.
[33] J.-C. Shen, W.M. Rideout III, P.A. Jones, The rate of hydrolytic deamination of 5-methylcytosine in double stranded DNA, Nucleic Acids Res. 22 (1994) 972–976.
[34] G. Pfeifer, Mutagenesis at methylated CpG sequences, Curr. Top. Microbiol. Immunol. 301 (2006) 259–281.
[35] R.D. Blake, S.T. Hess, J. Nicholson-Tuell, The influence of nearest neighbors on the rate and pattern of spontaneous point mutations, J. Mol. Evol. 34 (1992) 189–200.
[36] J. Ollila, I. Lappalainen, M. Vihinen, Sequence specificity in CpG mutation hotspots, FEBS Lett. 396 (1996) 119–122.
[37] H. Ikehata, M. Takatsu, Y. Saito, T. Ono, Distribution of spontaneous CpG-associated G:C→A:T mutations in the lacZ gene of Muta(TM) mice: effects of CpG methylation, the sequence context of CpG sites, and severity of mutations on the activity of the lacZ gene product, Environ. Mol. Mutagen. 36 (2000) 301–311.
[38] Z. Zhao, E. Boerwinkle, Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome, Genome Res. 12 (2002) 1679–1686.
[39] F. Zhang, Z. Zhao, The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs, Genomics 84 (2004) 785–795.
[40] J. Widom, Role of DNA sequence in nucleosome stability and dynamics, Q. Rev. Biophys. 34 (2001) 269–324.
[41] G. Bernardi, The neoselectionist theory of genome evolution, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8385–8390.
[42] N. Galtier, L. Duret, Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution, Trends Genet. 23 (2007) 273–277.
[43] C. Burge, A.M. Campbell, S. Karlin, Over- and under-representation of short oligonucleotides in DNA sequences, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 1358–1362.
[44] S. Karlin, I. Ladunga, Comparisons of eukaryotic genomic sequences, Proc. Natl. Acad. Sci. U. S. A. 91 (1994) 12832–12836.