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.

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 under-representation 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}	$\rho_{AA\vert TT}$	$\rho_{CC\vert GG}$	$\rho_{TG\vert CA}$	$\rho_{TC\vert GA}$	$\rho_{CT\vert AG}$	$\rho_{GT\vert AC}$
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.