Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides

David N. Cooper,[1] Matthew Mort,[1] Peter D. Stenson,[1] Edward V. Ball[1] and Nadia A. Chuzhanova[2]*

[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK
[2]School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham, NG11 8NS, UK
**Correspondence to*: Tel: +44 2920 744062; Fax: +44 2920 746551; E-mail: CooperDN@cardiff.ac.uk

Date received (in revised form): 28th April 2010

Abstract

The cytosine–guanine (CpG) dinucleotide has long been known to be a hotspot for pathological mutation in the human genome. This hypermutability is related to its role as the major site of cytosine methylation with the attendant risk of spontaneous deamination of 5-methylcytosine (5mC) to yield thymine. Cytosine methylation, however, also occurs in the context of CpNpG sites in the human genome, an unsurprising finding since the intrinsic symmetry of CpNpG renders it capable of supporting a semi-conservative model of replication of the methylation pattern. Recently, it has become clear that significant DNA methylation occurs in a CpHpG context (where H = A, C or T) in a variety of human somatic tissues. If we assume that CpHpG methylation also occurs in the germline, and that 5mC deamination can occur within a CpHpG context, then we might surmise that methylated CpHpG sites could also constitute mutation hotspots causing human genetic disease. To test this postulate, 54,625 missense and nonsense mutations from 2,113 genes causing inherited disease were retrieved from the Human Gene Mutation Database (http://www.hgmd.org). Some 18.2 per cent of these pathological lesions were found to be C → T and G → A transitions located in CpG dinucleotides (compatible with a model of methylation-mediated deamination of 5mC), an approximately ten-fold higher proportion than would have been expected by chance alone. The corresponding proportion for the CpHpG trinucleotide was 9.9 per cent, an approximately two-fold higher proportion than would have been expected by chance. We therefore estimate that ~5 per cent of missense/nonsense mutations causing human inherited disease may be attributable to methylation-mediated deamination of 5mC within a CpHpG context.

Keywords: *CpG dinucleotide, CpNpGp trinucleotide, cytosine methylation, 5-methylcytosine deamination, mutation hotspots, human inherited disease, missense/nonsense mutations*

> Man's yesterday may ne'er be like his morrow;
> Nought may endure but Mutabililty.
>
> Percy Bysshe Shelley (1816) *Mutability*

The first hint that the cytosine–guanine (CpG) dinucleotide might constitute a hotspot for pathological mutations in the human genome came nearly 25 years ago with the finding that two different CGA → TGA (Arg → Term) nonsense mutations had recurred quite independently at different locations in the factor VIII (*F8*) gene causing haemophilia A.[1] The potential generality of this phenomenon was supported by the finding

that 12 of the 34 (35 per cent) single base-pair (bp) substitutions then known to cause human inherited disease were C → T and G → A transitions within CpG dinucleotides.[2] Further studies soon confirmed that the CpG dinucleotide was a mutation hotspot in a variety of different human disease genes, including *PAH*,[3] *F9*,[4] *LDLR*,[5] *RB1*,[6] *HPRT1*[7] and *DMD*.[8] As mutation data accumulated, CGA → TGA transitions were encountered particularly frequently as a cause of human genetic disease; such nonsense mutations are inherently more likely to come to clinical attention than missense mutations.[9,10]

From the outset, it was realised that the hypermutability of the CpG dinucleotide was related to its role as the major site of cytosine methylation in the human genome. The reason traditionally put forward to explain this association has been that, while cytosine spontaneously deaminates to uracil (which is efficiently recognised as a non-DNA base and removed by uracil-DNA glycosylase), the spontaneous deamination of 5-methylcytosine (5mC) yields thymine,[11] thereby creating G•T mismatches whose removal by methyl-CpG binding domain protein 4 (MBD4) and/or thymine DNA glycosylase followed by base excision repair is error prone.[12–16] It remains possible, however, that mCpG transitions are not exclusively caused by the spontaneous deamination of 5-methylcytosine and may also arise through the action of other mechanisms and processes.[17–19] Irrespective of the precise nature of the underlying mechanism, Krawczak *et al.* (1998)[9] estimated that the rate of CG → TG (and CG → CA on the other strand) transitions was five times the base mutation rate. Subsequent estimates of 5mC hypermutability—derived from various studies of polymorphism, disease mutations or evolutionary divergence—have ranged between four-fold and 15-fold.[20–26]

It has been known for some time that cytosine methylation also occurs in the context of CpNpG sites in mammalian genomes, where N represents any nucleotide.[27,28] Since the intrinsic symmetry of the CpNpG trinucleotide would support a semi-conservative model of replication of the methylation pattern (as with the CpG dinucleotide), it comes as no surprise that both maintenance and *de novo* methylation occurs at CpNpG sites in mammalian cells.[28] In their recent paper on the human methylome, Lister *et al.*[29] reported abundant DNA methylation in CpHpG trinucleotides (where H = A, C or T). Specifically, some 17.3 per cent of 5mC in embryonic stem cells was found to occur within CpApG, CpCpG and CpTpG, with a further 7.2 per cent of 5mC occurring in CpHpH. Although Lister *et al.*[29] suggested that non-CpG methylation is almost entirely lost upon differentiation (a conclusion based solely upon the analysis of foetal lung fibroblasts), others have noted CpNpG methylation within human genes in a variety of different somatic tissues.[30,31] Although the extent of non-CpG methylation in the germline remains unclear, if we were to assume not only that CpHpG methylation occurs in the germline, but also that 5mC deamination can occur within a CpHpG context, then it is very likely that methylated CpHpG sites would constitute mutation hotspots. Indirect evidence that this might indeed be the case has come from a disproportionately high number of C → T and G → A transitions at CpNpG sites in studies of the human *NF1*[32] and *BRCA1*[33] genes. In the light of the above, we have revisited the question of CpG dinucleotide hypermutability and explored the potential contribution that CpHpG transitions might make to human inherited disease.

According to the April 2010 release of the Human Gene Mutation Database (HGMD; http://www.hgmd.org),[34] 56,457 pathological mutations have been reported in a total of 2,242 human genes. A subset of 54,625 pathological missense and nonsense mutations in 2,113 genes, with ±2 bp genomic DNA sequence flanking the site of mutation, was retrieved from the HGMD. The numbers of C → T and G → A mutations in this mutation dataset that were located within either CpG dinucleotides or CpHpG trinucleotides were counted and termed 'mutations in di/trinucleotide' (Table 1). Only these C → T and G → A transitions, found in the context of a CpG dinucleotide or CpHpG trinucleotide, would be compatible with a model of methylation-mediated deamination

Table 1. Numbers of C → T and G → A mutations found in CpG dinucleotides and CpHpG trinucleotides in a dataset of 54,625 missense and nonsense mutations in 2,113 different human genes (HGMD) and the numbers of possible C → T and G → A mutations in CpG dinucleotides and CpHpG trinucleotides within the coding regions of the mutated genes.

Di/trinucleotide	Dataset	Number of mutations in		*p*-value
		in di/trinucleotide	not in di/trinucleotide	
CpG	HGMD	9,947	44,678	$<10^{-230}$
	Possible	292,147	13,269,850	
CpHpG	HGMD	5,402	49,223	$<10^{-230}$
	Possible	610,714	12,951,283	

of 5mC. The remaining mutations in this HGMD dataset that were located in non-CpG or non-CpHpG di/trinucleotides within the genes in question were also counted and termed 'mutations not in di/trinucleotide' (Table 1). Thus, 18.2 per cent of the studied missense/nonsense mutations causing human inherited disease are located in the CpG dinucleotide, while the corresponding proportion for the CpHpG trinucleotide is 9.9 per cent. To assess the significance of these figures, the number of all possible C → T and G → A mutations within either CpG dinucleotides or CpHpG trinucleotides within the coding regions of the mutated genes, termed 'possible mutations in di/trinucleotides', were also counted (Table 1). In parallel, all possible single bp substitutions that occurred in a non-CpG dinucleotide or non-CpHpG trinucleotide context (as well as mutations other than C → T and G → A in CpG and CpHpG) within the coding regions of the mutated genes were counted as 'possible mutations not in di/trinucleotide' (Table 1). A weak positive correlation was noted between the number of CpG mutations in the 2,113 genes analysed and the number of possible CpG mutations in these genes (Pearson's correlation 0.129, $p = 2.45 \times 10^{-9}$), implying that the CpG mutation frequency is influenced to some extent by the frequency of occurrence of the underlying CpG dinucleotide. Unsurprisingly, a significantly greater proportion (approximately ten-fold) of observed pathological missense/nonsense mutations within these genes were C → T and G → A transitions within CpG dinucleotides than would have been expected (by chance alone) for all possible mutations (Table 1; $p < 10^{-230}$). A weak positive correlation (Pearson's correlation 0.251, $p = 1.01 \times 10^{-31}$) was also noted between the number of CpHpG-located mutations and the number of CpHpG trinucleotides in these genes, implying that the CpHpG mutation frequency is influenced to some extent by the frequency of occurrence of the underlying CpHpG trinucleotide. Once again, a greater proportion (approximately two-fold) of observed pathological missense/nonsense mutations within these genes were C → T and G → A transitions within CpHpG trinucleotides than would have been expected by chance alone for all possible mutations (Table 1; $p < 10^{-230}$).

From the data presented in Table 1, we estimate that ~11.8 per cent of the 9,947 CpG mutations (ie 1,176) occurred within this dinucleotide by chance alone and hence would not necessarily have originated via the methylation-mediated deamination of 5mC. In a similar vein, we estimate that ~46 per cent (2,460) of the CpHpG mutations (5,402) occurred within these trinucleotides by chance alone and hence may not have originated via methylation-mediated deamination of 5mC. The other side of this particular coin, however, is that the remaining 54 per cent of the 5,402 observed CpHpG mutations in HGMD (ie the excess 2,842 over expectation, or ~5 per cent of all the missense/nonsense mutations analysed) may well be attributable to methylation-mediated deamination of 5mC within a CpHpG context. As far as we are aware, this is the first (albeit crude) estimate of the potential impact of CpHpG mutations on human inherited disease.

Table 2. Numbers of C → T and G → A mutations found in CpG dinucleotides and CpHpG trinucleotides in a dataset of 1,766 regulatory mutations of 191 gene promoters (HGMD) and the numbers of possible C → T and G → A mutations in CpG dinucleotides and CpHpG trinucleotides.

Di/trinucleotide	Dataset	Number of mutations in		*p*-value
		in di/trinucleotide	not in di/trinucleotide	
CpG	HGMD	94	1,672	6.03×10^{-9}
	Possible	1,940	64,213	
CpHpG	HGMD	54	1,712	0.011
	Possible	2,838	63,315	

A similar analysis was performed for 1,766 regulatory mutations (identified in the promoters of 191 human genes) retrieved from the HGMD. The numbers of actual and possible CpG and CpHpG mutations were counted as before, using the promoter sequences of each gene. In order to determine the total numbers of possible CpG/CpHpG and non-CpG/CpHpG mutations, the wild-type promoter sequences for each gene (total length, 22,051 bp) were used (Table 2). As with the missense/nonsense mutations, an approximately two-fold higher proportion of observed pathological regulatory mutations within these genes were C → T and G → A transitions within CpG dinucleotides than would have been expected by chance alone for all possible mutations (Table 2; $p = 6.03 \times 10^{-9}$). We estimate that ~55 per cent of the 94 CpG mutations (ie ~52) probably occurred within these dinucleotides by chance alone rather than via methylation-mediated deamination of 5mC. By contrast, a *lower* than expected proportion of C → T and G → A regulatory mutations located in CpHpG trinucleotides was observed ($p = 0.011$). The absence of any excess of C → T and G → A mutations located in CpHpG trinucleotides indicates that most, if not all, the promoter CpHpG mutations probably occurred by chance alone, making it unnecessary to invoke methylation-mediated deamination of 5mC to account for them. Since neither CpG nor CpHpG were found to be under-represented in the examined promoter regions as compared to the coding regions, we surmise that the reduced (CpG) or absent (CpHpG) preponderance of C → T and G → A promoter mutations in the methylatable di/trinucleotides may be due to the relative paucity of cytosine methylation within the promoter regions[35] that would render unmethylated CpG and CpHpG di/trinucleotides no more mutable than any other di/trinucleotide.

Although two specific examples of non-CpG methylation altering the binding of transcription factors to promoter elements within human genes have so far been reported,[36,37] the functional role of most non-CpG methylation in the human genome is still unclear. Irrespective of the functionality or otherwise of this specific type of post-synthetic DNA modification in the human genome, it would appear that methylation of the CpHpG trinucleotide may leave a significant imprint on the spectrum of missense/nonsense mutations causing human genetic disease.

References

1. Youssoufian, H., Kazazian, H.H. Jr., Phillips, D.G., Aronis, S. *et al.* (1986), 'Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots', *Nature* Vol. 324, pp. 380–382.
2. Cooper, D.N. and Youssoufian, H. (1988), 'The CpG dinucleotide and human genetic disease', *Hum. Genet.* Vol. 78, pp. 151–155.
3. Abadie, V., Lyonnet, S., Maurin, N., Berthelon, M. *et al.* (1989), 'CpG dinucleotides are mutation hot spots in phenylketonuria', *Genomics* Vol. 5, pp. 936–939.
4. Koeberl, D.D., Bottema, C.D., Ketterling, R.P., Bridge, P.J. *et al.* (1990), 'Mutations causing hemophilia B: Direct estimate of the underlying rates of spontaneous germ-line transitions, transversions, and deletions in a human gene', *Am. J. Hum. Genet.* Vol. 47, pp. 202–217.
5. Rideout, W.M., 3rd, Coetzee, G.A., Olumi, A.F. and Jones, P.A. (1990), '5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes', *Science* Vol. 249, pp. 1288–1290.
6. Mancini, D., Singh, S., Ainsworth, P. and Rodenhiser, D. (1997), 'Constitutively methylated CpG dinucleotides as mutation hot spots in the retinoblastoma gene (*RB1*)', *Am. J. Hum. Genet.* Vol. 61, pp. 80–87.

7. O'Neill, J.P. and Finette, B.A. (1998), 'Transition mutations at CpG dinucleotides are the most frequent *in vivo* spontaneous single-based substitution mutation in the human *HPRT* gene', *Environ. Mol. Mutagen.* Vol. 32, pp. 188–191.
8. Buzin, C.H., Feng, J., Yan, J., Scaringe, W. *et al.* (2005), 'Mutation rates in the dystrophin gene: A hotspot of mutation at a CpG dinucleotide', *Hum. Mutat.* Vol. 25, pp. 177–188.
9. Krawczak, M., Ball, E.V. and Cooper, D.N. (1998), 'Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes', *Am. J. Hum. Genet.* Vol. 63, pp. 474–488.
10. Mort, M., Ivanov, D., Cooper, D.N. and Chuzhanova, N.A. (2008), 'A meta-analysis of nonsense mutations causing human genetic disease', *Hum. Mutat.* Vol. 29, pp. 1037–1047.
11. Shen, J.C., Rideout, W.M., 3rd and Jones, P.A. (1994), 'The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA', *Nucleic Acids Res.* Vol. 22, pp. 972–976.
12. Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J. *et al.* (1999), 'The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites', *Nature* Vol. 401, pp. 301–304.
13. Waters, T.R. and Swann, P.F. (2000), 'Thymine-DNA glycosylase and G to A transition mutations at CpG sites', *Mutat. Res.* Vol. 462, pp. 137–147.
14. Walsh, C.P. and Xu, G.L. (2006), 'Cytosine methylation and DNA repair', *Curr. Top. Microbiol. Immunol.* Vol. 301, pp. 283–315.
15. Cortázar, D., Kunz, C., Saito, Y., Steinacher, R. *et al.* (2007), 'The enigmatic thymine DNA glycosylase', *DNA Repair* Vol. 6, pp. 489–504.
16. Boland, M.J. and Christman, J.K. (2008), 'Characterization of Dnmt3b:thymine-DNA glycosylase interaction and stimulation of thymine glycosylase-mediated repair by DNA methyltransferase(s) and RNA', *J. Mol. Biol.* Vol. 379, pp. 492–504.
17. Shen, J.C., Rideout, W.M., 3rd and Jones, P.A. (1992), 'High frequency mutagenesis by a DNA methyltransferase', *Cell* Vol. 71, pp. 1073–1080.
18. Zhang, X. and Mathews, C.K. (1994), 'Effect of DNA cytosine methylation upon deamination-induced mutagenesis in a natural target sequence in duplex DNA', *J. Biol. Chem.* Vol. 269, pp. 7066–7069.
19. Pfeifer, G.P. (2006), 'Mutagenesis at methylated CpG sequences', *Curr. Top. Microbiol. Immunol.* Vol. 301, pp. 259–281.
20. Nachman, M.W. and Crowell, S.L. (2000), 'Estimate of the mutation rate per nucleotide in humans', *Genetics* Vol. 156, pp. 297–304.
21. Kondrashov, A.S. (2003), 'Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases', *Hum. Mutat.* Vol. 21, pp. 12–27.
22. Tomso, D.J. and Bell, D.A. (2003), 'Sequence context at human single nucleotide polymorphisms: Overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands', *J. Mol. Biol.* Vol. 327, pp. 303–308.
23. Jiang, C. and Zhao, Z. (2006), 'Directionality of point mutation and 5-methylcytosine deamination rates in the chimpanzee genome', *BMC Genomics* Vol. 7, p. 316.
24. Elango, N., Kim, S.H., Vigoda, E. and Yi, S.V. (2008), 'Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation', *PLoS Comput. Biol.* Vol. 4, p. e1000015.
25. Misawa, K. and Kikuno, R.F. (2009), 'Evaluation of the effect of CpG hypermutability on human codon substitution', *Gene* Vol. 431, pp. 18–22.
26. Li, J.B., Gao, Y., Aach, J., Zhang, K. *et al.* (2009), 'Multiplex padlock targeted sequencing reveals human hypermutable CpG variations', *Genome Res.* Vol. 19, pp. 1606–1615.
27. Woodcock, D.M., Crowther, P.J. and Diver, W.P. (1987), 'The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide', *Biochem. Biophys. Res. Commun.* Vol. 145, pp. 888–894.
28. Clark, S.J., Harrison, J. and Frommer, M. (1995), 'CpNpG methylation in mammalian cells', *Nat. Genet.* Vol. 10, pp. 20–27.
29. Lister, R., Pelizzda, M., Dowen, R.H. and Hawkins, R.D. *et al.* (2009), 'Human DNA methylomes at base resolution show widespread epigenomic differences', *Nature* Vol. 462, pp. 315–322.
30. Lee, J., Jang, S.J., Benoit, N., Hoque, M.O. *et al.* (2010), 'Presence of 5-methylcytosine in CpNpG trinucleotides in the human genome', *Genomics*, Vol. 96, pp. 67–72.
31. Laurent, L., Wong, E., Li, G., Huynh, T. *et al.* (2010), 'Dynamic changes in the human methylome during differentiation', *Genome Res.* Vol. 20, pp. 320–331.
32. Rodenhiser, D.I., Andrews, J.D., Mancini, D.N., Jung, J.H. *et al.* (1997), 'Homonucleotide tracts, short repeats and CpG/CpNpG motifs are frequent sites for heterogeneous mutations in the neurofibromatosis type 1 (*NF1*) tumour-suppressor gene', *Mutat. Res.* Vol. 373, pp. 185–195.
33. Cheung, L.W., Lee, Y.F., Ng, T.W., Ching, W.K. *et al.* (2007), 'CpG/CpNpG motifs in the coding region are preferred sites for mutagenesis in the breast cancer susceptibility genes', *FEBS Lett.* Vol. 581, pp. 4668–4674.
34. Stenson, P.D., Mort, M., Ball, E.V., Howells, K. *et al.* (2009), 'The Human Gene Mutation Database: 2008 update', *Genome Med.* Vol. 1, p. 13.
35. Illingworth, R.S. and Bird, A.P. (2009), 'CpG islands — "A rough guide"', *FEBS Lett.* Vol. 583, pp. 1713–1720.
36. Clark, S.J., Harrison, J. and Molloy, P.L. (1997), 'Sp1 binding is inhibited by (m)Cp(m)CpG methylation', *Gene* Vol. 195, pp. 67–71.
37. Inoue, S. and Oishi, M. (2005), 'Effects of methylation of non-CpG sequence in the promoter region on the expression of human synaptotagmin XI (syt11)', *Gene* Vol. 348, pp. 123–134.