

CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content

Karl J. Fryxell and Won-Jong Moon

Center for Biomedical Genomics and Informatics, Department of Molecular and Microbiology, George Mason University, Manassas, Virginia

CpG dinucleotides mutate at a high rate because cytosine is vulnerable to deamination, cytosines in CpG dinucleotides are often methylated, and deamination of 5-methylcytosine (5mC) produces thymidine. Previous experiments have shown that DNA melting is the rate-limiting step in cytosine deamination. Here we show, through the analysis of human single-nucleotide polymorphisms (SNPs), that the mutation rate produced by 5mC deamination is highly dependent on local GC content. In fact, linear regression analysis showed that the \log_{10} of the 5mC mutation rates (inferred from SNP frequencies) had slopes of -3 when graphed with respect to the GC content of neighboring sequences. This is the ideal slope that would be expected if the correlation between CpG underrepresentation and GC content had been solely caused by DNA melting. Moreover, this same result was obtained regardless of the SNP locations (all SNPs versus only SNPs in noncoding intergenic regions, excluding CpG islands) and regardless of the lengths over which GC content was calculated (SNP sequences with a modal length of 564 bp versus genomic contigs with a modal length of 163 kb). Several alternative interpretations are discussed.

Introduction

Base substitution rates vary throughout the human genome (Wolfe, Sharp, and Li 1989), but the magnitude and mechanism(s) of these variations remain the subjects of debate (Francino and Ochman 1999; Fryxell and Zuckerkandl 2000; Lander et al. 2001; Arndt, Petrov, and Hwa 2003; Webster, Smith, and Ellegren 2003). One point that is generally accepted is that the CpG dinucleotide mutates at a high rate because cytosine is vulnerable to deamination, cytosines in CpG dinucleotides are often methylated, and deamination of 5-methylcytosine (5mC) produces thymidine (Coulondre et al. 1978; Sved and Bird 1990; Wiebauer et al. 1993; Friedberg, Walker, and Siede 1995; Lander et al. 2001). Deamination of unmethylated cytosine produces uracil (U), which can be removed by uracil glycosylase (Lindahl 1974; Lindahl, Karran, and Wood 1997), but 5mC deamination generates thymine (T), which cannot be processed by this enzyme. The consequence in humans is that the mutation rate from 5mC to T is 10-fold to 50-fold higher than other transitions (Duncan and Miller 1980; Bulmer 1986; Britten et al. 1988; Sved and Bird 1990). More than one third of the germline point mutations that cause human genetic diseases (Cooper and Youssoufian 1988; Cooper and Krawczak 1993), and many of the somatic mutations leading to cancer (Jones et al. 1992; Hollstein et al. 1994) are caused by CpG hypermutability. The evolutionary consequence in humans is that the CpG dinucleotide is statistically underrepresented (Bird 1980) throughout almost the entire human genome (Lander et al. 2001). The extent of CpG underrepresentation is inversely correlated with GC content (Adams and Eason 1984; Bernardi et al. 1985; Bernardi 1995).

Local strand separation is a prerequisite for the deamination of cytosine (or 5mC) in double-stranded DNA

(Fryxell and Zuckerkandl 2000). The reaction mechanism requires the attack of H_3O^+ on the N-3 position, followed by the addition of H_2O to the C-4 position, neither of which are accessible to water in double-stranded DNA (Frederico, Kunkel, and Shaw 1993). The measured activation energies of this reaction are identical in single-stranded and double-stranded DNA, indicating that both reactions progress through the same intermediate; in other words, a single-stranded DNA target (Lindahl and Nyberg 1974; Ehrlich et al. 1986). Moreover, elegant genetic experiments have shown directly that single-base mismatches dramatically accelerate the rate of cytosine deamination in vivo (Frederico, Kunkel, and Shaw 1990, 1993).

Individual internal base pairs within DNA double helices remain paired for only a few milliseconds (Leroy et al. 1988). It is clear that these base pairs open one at a time, far below the melting temperature, and that the lifetimes of individual GC base pairs are about three times longer than those of individual AT base pairs (Leroy et al. 1988). This DNA melting reaction propagates progressively along the double helix for considerable distances in vivo (Skarstad, Baker, and Kornberg 1990). Moreover, the addition of nucleosomes to eukaryotic DNA does not have any measurable effect on the amount of energy required to melt DNA (Almagor and Cole 1989). Thus, it is clear that DNA does melt reversibly (“breathe”) under physiological conditions, and this melting must have an effect on the rates of cytosine deamination. The question then becomes the quantitative extent to which DNA melting (as opposed to other processes such as biased gene conversion, natural selection, transcription-coupled DNA repair, and so on) may be responsible for the correlation between CpG underrepresentation and GC content.

A quantitative relationship between GC content, DNA melting, and rates of cytosine deamination was supported by the reaction rates inferred from human genomic sequence data (Fryxell and Zuckerkandl 2000). However, the mathematical solution of the relevant equations required the assumption of mutational equilibrium, which although plausible for CpG dinucleotides because of their rapid mutation rates (Sved and Bird 1990; Krawczak and

Key words: CpG, GC content, human genome, *Homo sapiens*, SNPs, mutation rates.

E-mail: kfryxell@gmu.edu.

Mol. Biol. Evol. 22(3):650–658. 2005

doi:10.1093/molbev/msi043

Advance Access publication November 10, 2004

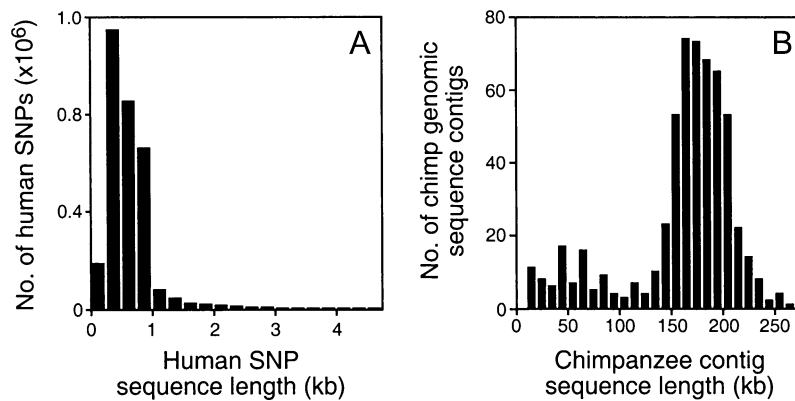


FIG. 1.—Characterization of the DNA sequences used in this study. Both length distributions shown here do not include a small number of much longer sequences (outside the region shown). (A) Length distribution of the 3,102,839 human SNP sequences that were retained for further analysis (see *Methods*). (B) Length distribution of the 483 genomic contigs from *Pan troglodytes* chromosomes 7 or 22 that were used in this study.

Cooper 1996; Fryxell and Zuckerkandl 2000), could be questioned because mutational equilibrium does not hold in many other cases (Lander et al. 2001; Webster, Smith, and Ellegren 2003). Here we show directly, through the analysis of human single-nucleotide polymorphisms (SNPs), that the rate of 5mC deamination does depend on local GC content, with a slope (in semilog₁₀ plots) of -3 . This slope was previously predicted, based on the physical chemistry of DNA melting (Fryxell and Zuckerkandl 2000). This, together with several other results discussed below, indicates that DNA melting is indeed the rate-limiting step for the deamination of 5mC *in vivo*.

Methods

Single-Nucleotide Polymorphisms

All human SNPs in the June, 2003 release of the National Cancer Institute SNP database (ftp.ncbi.nih.gov/snp/human/rs_fasta) were downloaded and analyzed. Each entry in this database consists of a DNA sequence, in which the polymorphic base is encoded according to the standard one-letter ambiguity code. We discarded SNPs whose associated sequence data files were 50 bp or less in length. A total of 3,102,839 SNPs passed this test with a modal average length of 564 bp, as shown in figure 1A.

SNPs with a Known Ancestral Base

Next, we searched the November 2003 edition of GenBank (Benson et al. 2002) for chimpanzee genomic DNA sequences with the keyword “complete.” This search yielded 569 genomic DNA sequence files, of which 483 sequences totaling 71.9 Mb had been assigned to *Pan troglodytes* chromosomes 7 or 22 and were retained for further analysis. These chimpanzee chromosome 7 and 22 genomic contigs had a modal average length of 163 kb (fig. 1B). We removed repetitive sequences from the genomic contigs with Repeat Masker (<ftp.genome.washington.edu/cgi-bin/RepeatMasker>), and the masked sequences were used as Blast query sequences (www.ncbi.nlm.nih.gov/SNP/snpblastByCh.html) to search all human SNPs that had been localized to the homologous human chromosomes (7 and 21, respectively). Potentially matching SNPs were

retained only if more than 97% of the nucleotides throughout the entire length of the SNP sequence were identical with the chimpanzee sequence. For this purpose, all nucleotides in repetitive sequences and/or aligned with gaps were counted as mismatches. We also discarded SNPs whose sequences had more than one polymorphic position or more than two possible bases at the polymorphic position. A total of 23,208 human SNPs passed all of these tests and, thus, had an unambiguously orthologous sequence in *Pan troglodytes*.

We chose to discard repetitive sequences because it would be more difficult operationally to establish their orthology with the same level of confidence. Designing such a computer program would require not only performing an additional series of sequence comparisons to establish the sequence context of each element but also programmed attempts to rule out a wide variety of potential historical events such as gene conversion, unequal crossing over, segmental duplications and deletions, and multiple transposon insertions near the same site. The most straightforward solution to such issues would be to focus on single-copy sequences, just as all vertebrate genome projects focus on single-copy sequences in the early stages of constructing DNA sequence scaffolds (Gibson and Muse 2002). Indeed, this was the original motivation for the development of the RepeatMasker program (Lander et al. 2001).

Our list of 23,208 SNPs with a known chimpanzee ortholog was pruned by removing duplicate listings, leaving 18,965 unique SNPs. This list was further pruned by removing SNPs for which the polymorphic base did not match the chimpanzee sequence, leaving 18,801 unique SNPs with a known ancestral base. Finally, this list was pruned down to 4,437 SNPs in which the polymorphic base corresponded to a transition mutation within ancestral CpG or GpC dinucleotides.

These 4,437 transition mutations in CpG or GpC dinucleotides were classified according to the location of the polymorphic base in the annotated human genome sequence. The classification of CpG islands was performed by a computer program, written in PERL by the authors, which applied the original definition of a CpG island (Gardiner-Garden and Frommer 1987) to all of the SNP sequence files. That is, a CpG island was defined as any

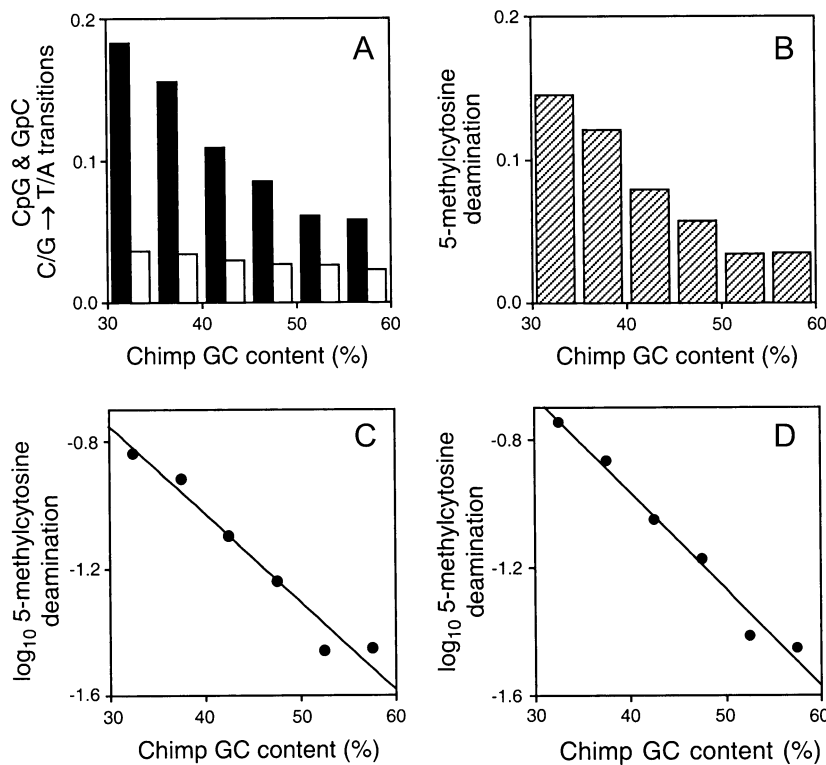


FIG. 2.—Rates of 5mC deamination, based on human SNPs with a known ancestral base, plotted with respect to the GC content of neighboring sequences (the orthologous chimpanzee genomic contig [see *Methods*]). (A) Black bars show the rate of CpG → TpG or CpG → CpA base substitutions per CpG dinucleotide in human SNP sequences with a known ancestral base. White bars show the corresponding rate of GpC → GpT or GpC → ApC base substitutions per GpC dinucleotide in the same set of SNP sequences. (B) Striped bars show the mathematical difference between adjacent black and gray bars (inferred 5mC deamination rate). (C) A semilog₁₀ plot of the 5mC deamination rates from figure 2B. The linear regression line is shown (equation $Y = 0.1 - 2.8X$, $r^2 = 0.96$). (D) A semilog₁₀ plot of the 5mC deamination rates, calculated as in (C) but using only SNPs in intergenic noncoding DNA sequences (see *Methods* and figure 3A). The linear regression line is shown (equation $Y = 0.2 - 3.0X$, $r^2 = 0.98$).

contiguous sequence at least 200 bp in length, for which the GC content is more than 50% and the CpG/GpC ratio is greater than 0.6 (Gardiner-Garden and Frommer 1987). If a polymorphic base was found to be inside a CpG island by these criteria, then the location of that SNP was classified as a “CpG island,” regardless of its annotated location (see below). For the record, the majority of the SNPs in CpG islands were annotated as residing within intergenic noncoding DNA (88) or introns (35), as expected (Bird 1986; Aïssani and Bernardi 1991). Smaller numbers of SNPs in CpG islands were located in 5′ untranslated regions (10), protein-coding sequences (11), 3′ untranslated regions (one), or pseudogenes (zero).

The classification of all other SNPs (transitions in CpG or GpC dinucleotides) was performed by a second PERL computer program written by the authors, which parsed the SNP locations within annotated GenBank files into 5′ untranslated regions (260 SNPs), protein-coding sequences (187 SNPs), introns (1,222 SNPs), 3′ untranslated regions (52 SNPs), pseudogenes (three SNPs), and intergenic noncoding DNA sequences (excluding CpG islands) (2,568 SNPs). The following input data files were used to classify annotated SNP locations on human chromosome 21 (accession numbers with version number): NT_011512.9, NT_011515.10, NT_029490.3, NT_030187.1, and NT_030188.3. The following input data files were used to classify annotated SNP locations on human chromosome 7 (accession numbers with version number):

NT_007741.12, NT_007758.10, NT_007819.14, NT_007914.13, NT_007933.13, NT_023603.5, NT_023629.12, NT_029998.6, NT_030008.6, NT_033968.5, NT_034885.3, NT_077528.1, NT_078033.1, NT_079514.1, NT_079515.1, NT_079516.1, NT_079590.1, NT_079591.1, NT_079592.1, NT_079593.1, NT_079594.1, NT_079595.1, NT_079596.1, and NT_079597.1.

Results

SNPs in CpG and GpC Dinucleotides

SNPs that were produced by a CpG → TpG mutation can be distinguished from SNPs that were caused by the reverse mutation (TpG → CpG) by using the chimpanzee as an out-group to determine which of the two polymorphic bases was ancestral. We found 18,801 SNPs in human single-copy DNA that had a known ancestral base in the chimpanzee (see *Methods*). From these, we further found that 4,437 SNPs were produced by transition mutations in ancestral CpG or GpC dinucleotides (fig. 2A). Because the GpC dinucleotide is not methylated in mammalian genomes (Razin and Riggs 1980), it follows that the mathematical difference (CpG transition rate minus GpC transition rate) represents the rate of methylation-dependent transitions; in other words, mutations that result from the deamination of 5mC. This difference was exponentially dependent on local GC content (fig. 2B). A semilog₁₀ plot of the rate of 5mC deamination

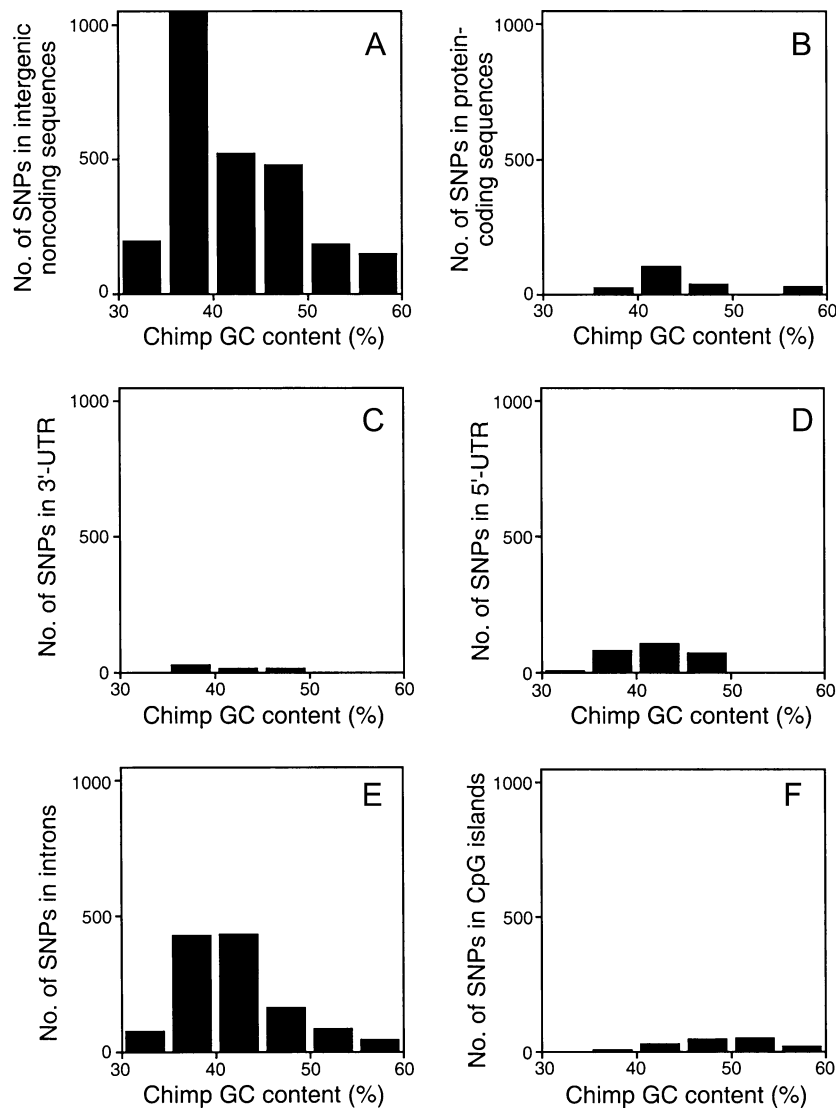


FIG. 3.—The distribution of the SNPs used in figure 2, with respect to annotated location and the GC content of neighboring sequences (the orthologous chimpanzee genomic contig [see *Methods*]). (A) Intergenic noncoding DNA, (B) protein-coding sequences, (C) 3' untranslated regions (3'-UTR), (D) 5' untranslated regions (5'-UTR), (E) introns, (F) CpG islands. One additional category (pseudogenes) is not shown here, because we found only three SNPs in pseudogenes (that corresponded to transitions in ancestral CpG or GpC dinucleotides), and such small numbers were not visible when this figure was drawn to scale.

(with respect to the GC content of the homologous genomic region) had a slope of -2.8 (fig. 2C), with 95% confidence limits of -2.0 to -3.5 . This is in good agreement with the slope of -3.0 that was predicted based on DNA melting as a function of base composition (Fryxell and Zuckerkandl 2000).

The Distribution of CpG and GpC SNPs with Respect to Functional Domains

The potential influence of differential methylation of CpG islands (Bird 1986; Gardiner-Garden and Frommer 1987; Aïssani and Bernardi 1991) and differential DNA repair of transcribed sequences (Green et al. 2003) was addressed by repeating this analysis using only transitions in ancestral CpG and GpC dinucleotides that also

happened to be located within intergenic noncoding DNA sequences (i.e., excluding all exons, introns, and CpG islands). To accomplish this, each of our 4,437 SNPs representing transition mutations in CpG or GpC dinucleotides were further classified according to the location of the polymorphic base in the annotated human genome sequence (see *Methods*). The numbers of SNPs in each category are listed in *Methods*, and their distribution with respect to the GC content of neighboring sequences is shown in figure 3.

Semilog plots of 5mC deamination in intergenic noncoding DNAs (excluding CpG islands) produced results that were indistinguishable from the total data set (compare figure 2C versus D). More specifically, a semi-log₁₀ plot of the rate of 5mC deamination in intergenic noncoding DNA (with respect to the GC content of the

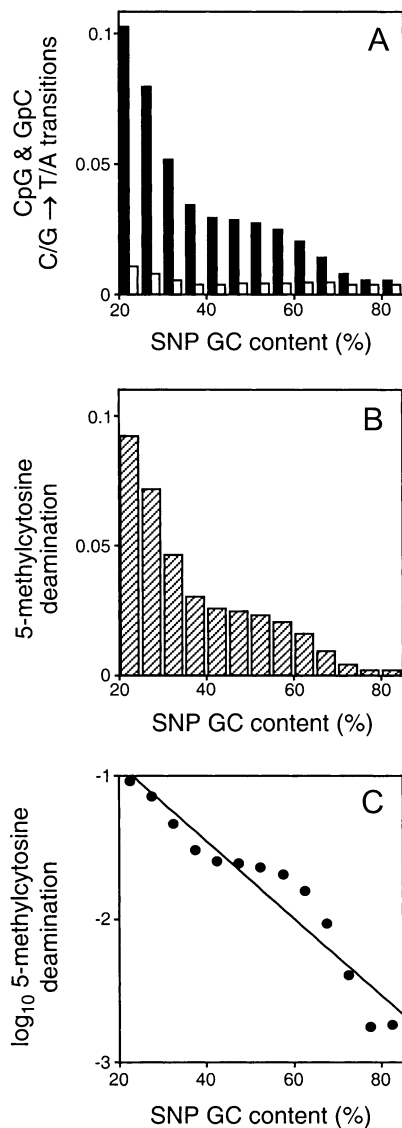


FIG. 4.—Rates of CpG turnover, based on SNP frequencies in the human genome. See equations 1 and 2 in *Results*. (A) The values of CpG_{trans} (black bars) and GpC_{trans} (white bars) with respect to SNP GC content. (B) The mathematical difference $CpG_{trans} - GpC_{trans}$ (striped bars), as a function of SNP GC content. (C) A semilog₁₀ plot of $CpG_{trans} - GpC_{trans} = 5mC$ deamination rate, as a function of SNP GC content. The linear regression line is shown (equation $Y = -0.4 - 2.7X$, $r^2 = 0.91$).

homologous genomic region) had a slope of -3.0 (fig. 2D), with 95% confidence limits of -2.4 to -3.6 . This agrees exactly with the slope of -3.0 that was predicted based on DNA melting as a function of base composition (Fryxell and Zuckerkandl 2000). Because figure 2D did not include SNPs from exons, introns, or CpG islands, it follows that this slope was not generated by transcription-coupled DNA repair (Green et al. 2003), transcription-induced cytosine deamination (Francino and Ochman 2001), or the demethylation of CpG islands during part of the life cycle (Frank et al. 1991). We will defer the issue of the accuracy of the annotation of the human genome until the *Discussion* section.

The Turnover of CpG and GpC Dinucleotides Throughout the Human Genome

The analysis in figure 2 was limited to single-copy DNA sequences on two chromosomes (see *Methods*) and further required inferences based on sequence comparisons to the chimpanzee. We, therefore, sought to confirm whether this same pattern of CpG turnover could be discerned on other human chromosomes and without requiring inferences from another species. To do this, we analyzed SNP frequencies directly, without reference to any out-group. This approach was based on the fact that deamination of 5mC produces a CpG \rightarrow TpG (or CpG \rightarrow CpA) base substitution, which would be represented in a SNP database as a YG or CR dinucleotide. Of course, the reverse substitution, TpG \rightarrow CpG or CpA \rightarrow CpG, could also produce a YG or CR dinucleotide. Consider the following equation:

$$CpG_{trans} = (YG + CR)/CG \quad (1)$$

where Y and R represent C/T and G/A polymorphisms, respectively. YG , CR , and CG represent the number of occurrences of each of those dinucleotides in a SNP database. Thus, equation 1 provides an estimate of the rate of CpG turnover. More specifically, equation 1 indicates the rate at which transition base substitutions cause SNPs that create or destroy CpG dinucleotides. We can also readily calculate the comparable turnover rate for GpC dinucleotides:

$$GpC_{trans} = (GY + RC)/GC \quad (2)$$

In this case, the mathematical difference $CpG_{trans} - GpC_{trans}$ represents the additional turnover in CpG dinucleotides that is attributable to DNA methylation. This formula can be readily applied to the entire human genome. In fact, computational analysis of 3,102,839 human SNPs longer than 50 bp showed that the difference ($CpG_{trans} - GpC_{trans}$) varied over a 50-fold range and was strongly dependent on the GC content of the neighboring DNA sequence (fig. 4A and B). Linear regression analysis of a semilog₁₀ plot gave a best-fit slope of -2.7 (95% confidence limits of -2.1 to -3.2 [see figure 4C]), which was essentially identical to the slope obtained with all SNPs that have a known ancestral base (-2.8 [see figure 2C]). Again, both are in good agreement with the theoretical slope of -3.0 predicted from DNA melting as a function of base composition (Fryxell and Zuckerkandl 2000).

This analysis of all human SNPs were plotted with respect to the GC contents of the human SNP sequences (fig. 4C), whose modal length was 564 bp (fig. 1A). In contrast, our previous analysis of 5mC deamination based on human SNPs with a known ancestral base was plotted with respect to the GC content of the orthologous chimpanzee genomic contig (fig. 2C), whose modal length was 163 kb (fig. 1B). Both methods of analysis yielded essentially the same result (see above), notwithstanding a 290-fold difference in the lengths of the neighboring sequences over which GC content was measured. Moreover, one method used the GC content of the orthologous chimpanzee genomic contig (because it was a convenient sequence of appropriate length), but the other method used

the GC content of the human SNP sequences themselves (because the larger sample size averaged out the variations in shorter sequences). This shows that neither the human-chimp species difference nor the varying lengths over which GC content was measured had a significant effect on the slope of semilog plots of 5mC deamination.

Discussion

Does Differential DNA Methylation Cause the Correlation Between CpG Underrepresentation and GC Content?

CpG islands are hypomethylated (Bird 1986; Gardiner-Garden and Frommer 1987; Aïssani and Bernardi 1991). More specifically, CpG islands are methylated during a subset of the life cycle (Frank et al. 1991). Mathematical analysis indicates that the hypomethylation of CpG islands reduces their CpG mutation rates approximately twofold (with the remainder of the rate reduction being attributable to their GC content [see Fryxell and Zuckerkandl {2000}]). Sequence analysis of the human genome has shown that CpG islands account for 19.8 Mb of the 3,289 Mb sequence (Lander et al. 2001), or 0.6% of the total. Obviously, the inclusion or exclusion of CpG islands should have little effect on the overall correlation between CpG mutation rates and GC content, and we confirmed that was the case (compare figure 2C versus D).

Another question is whether there is any systematic relationship between GC content and DNA methylation patterns in the human genome. The levels of 5mC have been measured in compositional DNA fractions from human cells (Caccio et al. 1997), and the results showed a positive correlation between 5mC levels (which ranged from 0.4% to 0.8% of total bases) and GC content (which ranged from 35% to 50%). Over this same range of GC contents in noncoding DNA (introns), CpG dinucleotides ranged from approximately 0.5% to 1.5% of total dinucleotides (Aïssani and Bernardi 1991). Thus, the majority of all CpG dinucleotides in noncoding DNA are methylated in both AT-rich and GC-rich fractions.

Does Transcription Cause the Correlation Between CpG Underrepresentation and GC Content?

Separate semilog plots of 5mC deamination in protein-coding sequences and introns had lower best-fit slopes (-1.6 and -1.9 , respectively), although the smaller sizes of these samples increased their 95% confidence limits to the point where these differences from the theoretical slope of -3.0 (Fryxell and Zuckerkandl 2000) were not statistically significant (not shown). In any case, it is evident that transcription-induced cytosine deamination (which tends to increase 5mC deamination [see Francino and Ochman {2001}]) and transcription-coupled DNA repair (which tends to decrease inherited transitions in CpG dinucleotides [see Green et al. {2003}]), when taken together, had a relatively minor net effect on SNP frequencies in CpG and GpC dinucleotides. In fact, we found that SNPs in protein-coding sequences represented 4.5% of the total transitions in CpG and GpC dinucleotides (see *Methods* and figure 3C). This is threefold larger than

the 1.5% of the human genome that consists of protein-coding sequences (Lander et al. 2001; Waterston et al. 2002). However, we found that the computational partition of our human genomic contigs (accession numbers listed in *Methods*) into protein-coding versus noncoding sequences yielded a total of 15,795,547 bp of protein-coding sequences and 292,679,549 bp of noncoding sequences (disregarding possible overlaps between contigs). In other words, protein-coding sequences accounted for approximately 5.1% of the DNA in our sample, which agrees rather well with the 4.5% of the total transitions in CpG and GpC dinucleotides (see *Methods* and figure 3C). This implies that purifying selection has caused a rather small reduction in the number (not the abundance) of SNPs in protein-coding sequences and is consistent with the observation that the number of synonymous SNPs is approximately equal to the number of nonsynonymous SNPs in human protein-coding sequences (Bentley 2003).

The distribution of SNPs was also proportional to DNA lengths in introns, which represented 28% of the total transitions in CpG and GpC dinucleotides (see *Methods* and figure 3B) and also represented approximately 28% of the annotated human genome sequence (based on estimates of about 30,000 genes with an average of 30 kb of intron sequences per gene [see Lander et al. {2001} and Waterston et al. {2002}]). Thus, transcription-coupled DNA repair does not substantially bias the numbers of these SNPs. We will discuss the accuracy of the annotations on which these classifications were based in a separate section (see below).

To What Extent Do the Present Results Depend on the Accuracy of the Human Genome Annotation?

Comprehensive comparisons of expressed sequence tag (EST) sequences from various species suggests that the list of identified protein-coding transcripts in the human genome is nearing completion (Nekrutenko 2004). However, some authors have suggested that additional functional transcription units (presumably noncoding) may cover at least half the human genome (Sémon and Duret 2004). This assertion was based on regional differences in transposon abundance and orientation. These may represent the fossil traces of transcripts that were functional during the historical periods when those transposons were active (Arndt, Petrov, and Hwa 2003) or may even represent the sort of transposon insertion biases that have been observed in *Drosophila* (Rio 2002). In any case, our results remained essentially unchanged, regardless of whether all known human exons, introns, and CpG islands were included or excluded (fig. 2C and D). Therefore, it is unlikely that the inclusion or exclusion of additional genes (i.e., in future revisions of the human genome annotation) would significantly affect our results.

Is the Correlation Between CpG Underrepresentation and GC Content a Mathematical Artifact?

It has been proposed that the correlation of the CpG/GpC dinucleotide ratio with GC content was a mathematical artifact (Duret and Galtier 2000). However, the

computer simulations used to support this hypothesis assumed that the rate of 5mC deamination was unaffected by GC content. This (somewhat circular) assumption was refuted by our counts of human SNP frequencies (figs. 2 and 4). Moreover, Duret and Galtier (2000) used simulated sequences that were limited to 3 bp in length and further limited because only the middle base was allowed to mutate (Duret and Galtier 2000). Comprehensive simulations of 100-kb sequences, in which all bases were allowed to mutate, produced significantly different results (Fryxell and Zuckerkandl 2000) that were consistent with the 5mC deamination rates shown here.

Is the Correlation Between CpG Underrepresentation and GC Content a Fossil Remnant of Premammalian Mutational Pressures?

Another alternative hypothesis was that regional differences in mutation biases (including CpG mutation rates) were correlated with local GC content in ancient vertebrates, but this correlation disappeared at the time of the radiation of eutherian mammals (Arndt, Petrov, and Hwa 2003). We do not address the question of mutation bias in general, but we do show here that the current frequencies of CpG \rightarrow TpG mutations are strongly correlated with local GC content in the present-day human population (figs. 2 and 4). This discrepancy is probably caused by the different methods used to measure regional mutation rates. Arndt, Petrov, and Hwa (2003) utilized sequence comparisons between non-LTR retrotransposons, which are incapable of self excision and, thus, can provide useful phylogenetic markers (Batzer et al. 1996; Moran and Gilbert 2002). However, exhaustive analysis has shown that the average human genome contains about 80 to 100 copies of *LI* elements that have transposition activity in cell culture assays (Brouha et al. 2003). Moreover, the majority of full-length transposed daughter elements in these assays are also initially capable of replicative transposition (DeBerardinis et al. 1998). Although six “hot” *LI* elements account for the majority of the current *LI* transposition events, it is also true that three of these elements belong to the same subclass (Ta-1d). Moreover, the three youngest *LI* subclasses (Ta-1d, Ta-nd, and Ta-0) were represented by 12, seven, and 10 active *LI* elements, respectively, located at widely separated chromosomal sites (Brouha et al. 2003; Lutz et al. 2003). Thus, the hypothesis that each transposon subclass corresponds to a single active copy at a single chromosomal site (Deininger et al. 1992) is no longer tenable. In fact, even highly diverged older *Alu* subfamilies and noncanonical *LI* elements do continue to transpose (Brouha et al. 2003; Johanning et al. 2003; Lutz et al. 2003). Inactive transposons also exchange information without transposition, by gene conversion and unequal crossing over (Belmaaza et al. 1990; Graur and Li 2000; Hardies et al. 2000; Roy et al. 2000). The molecular consequences of gene conversion have been clarified in *Drosophila*, where continuous gene conversion tracts were found to be 0.2 to 3.1 kb in length (Curtis et al. 1989). Such events would frequently pass undetected in mammalian retrotransposons because they would completely erase the diagnostic positions in

most *Alu* and many *LI* elements. In other words, several molecular processes do cause retrotransposons to inherit base substitutions that originally occurred at other chromosomal sites.

Does Biased Gene Conversion Cause the Correlation Between CpG Underrepresentation and GC Content?

Another alternative hypothesis is that regional variations in mutation bias were influenced by regional variation in the rates of recombination, gene conversion, and biased DNA mismatch repair (Galtier et al. 2001). The rationale for this “biased gene conversion” hypothesis is closely related to cytosine deamination because the biases in mammalian DNA mismatch repair are primarily attributable to enzymatic adaptations that facilitate the repair of the U:G and T:G base pair mismatches resulting from cytosine deamination (Brown and Jiricny 1988; Friedberg, Walker, and Siede 1995; Lieb and Bhagwat 1996; Fryxell and Zuckerkandl 2000). Mammalian recombination rates are positively correlated with GC content, although the correlation is rather weak (Galtier et al. 2001).

Mammals possess several thymine glycosylase enzymes, which appear to be specialized for the repair of 5mC deamination events (Hendrich et al. 1999). These thymine glycosylases are less efficient than the uracil glycosylases that repair GpC deamination events (Gallinari and Jiricny 1996; Kavli et al. 2002). Another difference is that uracil glycosylases must repair the DNA before it is replicated, whereas thymine glycosylases could continue to repair a mutation in subsequent generations, including the heteroduplex DNAs formed during gene conversion events. Thus, the “repair” of 5mC deamination events (TpG) might be more responsive to the rates of gene conversion than the “delayed repair” of GpC deamination events (GpT) because of the additional enzymes that are specifically adapted to repair the former.

In any case, we found that the frequency of human SNPs caused by transitions in GpC dinucleotides showed little variation with GC content (white bars in figure 2A). However, replotting this data on a semilog scale gave a best-fit slope of -0.8 (with 95% confidence limits from -0.6 to -1.0 [data not shown]). In a similar fashion, we also replotted the turnover of GpC dinucleotides throughout the human genome (white bars in figure 4A) on a semilog scale, which gave a best-fit slope of -0.5 (with 95% confidence limits from -0.1 to -0.8 [data not shown]). Because both slopes were significantly different from zero, we cannot rule out the possibility that our results may be caused by biased gene conversion (Galtier et al. 2001). However, it seems more likely that the GpC slopes reflect regional differences in the deamination of unmethylated cytosines (Fryxell and Zuckerkandl 2000), together with finite rates of uracil repair.

Conclusions

DNA melting is the rate-limiting step for 5mC deamination (Lindahl and Nyberg 1974; Frederico, Kunkel, and Shaw 1990, 1993; Fryxell and Zuckerkandl 2000). Although this fact is well established, it does not necessarily

follow that differential DNA melting is solely responsible for the correlation between CpG underrepresentation and GC content in the human genome. If it were solely responsible, then plots of the \log_{10} CpG mutation rate versus GC content (expressed as a decimal fraction) should ideally have a slope of -3.0 (Fryxell and Zuckerkandl 2000). Based on an analysis of SNP frequencies in the human genome, we show that the best-fit slope is actually -2.7 for all SNPs and -3.0 for intergenic noncoding DNA (excluding CpG islands). This shows that the slope of -3 could not be specifically caused by exons, introns, differential methylation of CpG islands, transcription-coupled DNA repair, or transcription-induced cytosine deamination. Nor could regional differences in DNA methylation explain our results, because the majority of the CpG dinucleotides throughout the human genome are methylated (see above). By comparing alternative methods of analysis, we further show that our results neither depended on the DNA lengths over which the GC content of neighboring sequences was measured nor were they an artifact of the chromosomes or sequence comparisons used to identify the ancestral base.

We did observe a small but significant correlation between GC content and the rates of transitions in GpC dinucleotides inferred from human SNPs. This is consistent with the biased gene conversion hypothesis (Galtier et al. 2001), which implies that deamination events in both CpG and GpC dinucleotides would be “corrected” at higher rates by biased gene conversion in GC-rich sequences with higher rates of recombination. However, it is also consistent with the reaction mechanism of cytosine deamination itself (Fryxell and Zuckerkandl 2000). That is, unmethylated cytosines in GpC dinucleotides undergo deamination at lower rates in GC-rich sequences because of reduced DNA melting. This explanation is both simpler and more precise because it successfully predicts the slope of the 5mC deamination mutation rate with respect to neighboring GC content.

Acknowledgments

We thank Megan V. Chapman and Ramya Sundararajan for technical assistance with the identification of orthologous chimpanzee sequences.

Literature Cited

- Adams, R. L., and R. Eason. 1984. Increased GC content of DNA stabilizes methyl CpG dinucleotides. *Nucleic Acids Res.* **12**:5869–5877.
- Aïssani, B., and G. Bernardi. 1991. CpG islands, genes and isochores in the genome of vertebrates. *Gene* **106**:185–195.
- Almagor, M., and R. D. Cole. 1989. In physiological salt conditions, the core proteins of the nucleosomes in large chromatin fragments denature at 73°C and the DNA unstacks at 85°C. *J. Biol. Chem.* **264**:6515–6519.
- Arndt, P. F., D. A. Petrov, and T. Hwa. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **20**:1887–1896.
- Batzer, M. A., S. S. Arcot, J. W. Phinney et al. (16 co-authors). 1996. Genetic variation of recent *Alu* insertions in human populations. *J. Mol. Evol.* **42**:22–29.
- Belmaaza, A., J. C. Wallenburg, S. Brouillette, N. Gusew, and P. Chartrand. 1990. Genetic exchange between endogenous and exogenous LINE-1 repetitive elements in mouse cells. *Nucl. Acids Res.* **18**:6385–6391.
- Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. 2002. GenBank. *Nucleic Acids Res.* **30**:17–20.
- Bentley, D. R. 2003. DNA sequence variation of *Homo sapiens*. Cold Spring Harbor Symp. Quant. Biol. **68**:55–63.
- Bernardi, G. 1995. The human genome: organization and evolutionary history. *Annu. Rev. Genet.* **29**:445–476.
- Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228**:953–958.
- Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**:1499–1504.
- . 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321**:209–213.
- Britten, R. J., W. F. Baron, D. B. Stout, and E. H. Davidson. 1988. Sources and evolution of human *Alu* repeated sequences. *Proc. Natl. Acad. Sci. USA* **85**:4770–4774.
- Brouha, B., J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, and H. H. Kazazian. 2003. Hot *L1*s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. USA* **100**:5280–5285.
- Brown, T. C., and J. Jiricny. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**:705–711.
- Bulmer, M. 1986. Neighboring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.* **3**:322–329.
- Caccio, S., K. Jabbari, G. Matassi, F. Guermonprez, J. Desgres, and G. Bernardi. 1997. Methylation patterns in the isochores of vertebrate genomes. *Gene* **205**:119–124.
- Cooper, D. N., and M. Krawczak. 1993. Human gene mutation. BIOS Scientific Publishers Limited, Oxford.
- Cooper, D. N., and H. Youssoufian. 1988. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78**:151–155.
- Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**:775–780.
- Curtis, D., S. H. Clark, A. Chovnick, and W. Bender. 1989. Molecular analysis of recombination events in *Drosophila*. *Genetics* **122**:653–662.
- DeBerardinis, R. J., J. L. Goodier, E. M. Ostertag, and H. H. J. Kazazian. 1998. Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. *Nat. Genet.* **20**:288–290.
- Deininger, P. L., M. A. Batzer, C. A. Hutchison, and M. H. Edgell. 1992. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**:307–311.
- Duncan, B. K., and J. H. Miller. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**:560–561.
- Duret, L., and N. Galtier. 2000. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.* **17**:1620–1625.
- Ehrlich, M., K. F. Norris, R. Y.-H. Wang, K. C. Kuo, and C. W. Gehrke. 1986. DNA cytosine methylation and heat-induced deamination. *Biosci. Rep.* **6**:387–393.
- Francino, M. P., and H. Ochman. 1999. Isochores result from mutation not selection. *Nature* **400**:30–31.
- . 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol. Biol. Evol.* **18**:1147–1150.
- Frank, D., I. Keshet, M. Shani, A. Levine, A. Razin, and H. Cedar. 1991. Demethylation of CpG islands in embryonic cells. *Nature* **351**:239–241.

- Frederico, L. A., T. A. Kunkel, and B. R. Shaw. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**:2532–2537.
- . 1993. Cytosine deamination in mismatched base pairs. *Biochemistry* **32**:6523–6530.
- Friedberg, E. C., G. C. Walker, and W. Siede. 1995. DNA Repair and mutagenesis. American Society for Microbiology, Washington, DC.
- Fryxell, K. J., and E. Zuckerkandl. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**:1371–1383.
- Gallinari, P., and J. Jiricny. 1996. A new class of uracil-DNA glycosylases related to human thymine-DNA glycosylase. *Nature* **383**:735–738.
- Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**:907–911.
- Gardiner-Garden, M., and M. Frommer. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**:261–282.
- Gibson, G., and S. V. Muse. 2002. A primer of genome science. Sinauer Associates, Sunderland, Mass.
- Graur, D., and W.-H. Li. 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.
- Green, P., B. Ewing, W. Miller, P. J. Thomas, NISC Comparative Sequencing Program, and E. D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**:514–517.
- Hardies, S. C., L. Wang, L. Zhou, Y. Zhao, N. C. Casavant, and S. Huang. 2000. LINE-1 (L1) lineages in the mouse. *Mol. Biol. Evol.* **17**:616–628.
- Hendrich, B., U. Hardeland, H.-H. Ng, J. Jiricny, and A. Bird. 1999. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**:301–304.
- Hollstein, M., K. Rice, M. S. Greenblatt, T. Soussi, R. Fuchs, T. Sorlie, E. Hovig, B. Smith-Sorensen, R. Montesano, and C. C. Harris. 1994. Database of p53 gene somatic mutations in human tumors and cell lines. *Nucleic Acids Res.* **22**:3551–3555.
- Johanning, K., C. A. Stevenson, O. O. Oyeneran, Y. M. Gozal, A. M. Roy-Engel, J. Jurka, and P. L. Deininger. 2003. Potential for retroposition by old *Alu* subfamilies. *J. Mol. Evol.* **56**:658–664.
- Jones, P. A., W. M. Rideout III, J.-C. Shen, C. H. Spruck, and Y. C. Tsai. 1992. Methylation, mutation and cancer. *Bioessays* **14**:33–36.
- Kavli, B., O. Sundheim, M. Akbari, M. Otterlei, H. Nilsen, F. Skorpen, P. A. Aas, L. Hagen, H. E. Krokan, and G. Slupphaug. 2002. hUNG2 is the major repair enzyme for removal of uracil from U:A matches, U:G mismatches, and U in single-stranded DNA, with hSMUG1 as a broad specificity backup. *J. Biol. Chem.* **277**:39926–39936.
- Krawczak, M., and D. N. Cooper. 1996. Mutational processes in pathology and evolution. Pp. 1–33 in M. Jackson, T. Strachan, and G. Dover, eds. Human genome evolution. BIOS Scientific Publishers, Oxford.
- Lander, E. S., L. M. Linton, B. Birren et al. (249 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
- Leroy, J. L., M. Kochoyan, T. Huynh-Dinh, and M. Gueron. 1988. Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *J. Mol. Biol.* **200**:223–238.
- Lieb, M., and A. S. Bhagwat. 1996. Very short patch repair: reducing the cost of cytosine methylation. *Mol. Microbiol.* **20**:467–473.
- Lindahl, T. 1974. An N-glycosidase from *Escherichia coli* that releases free uracil from DNA containing deaminated cytosine residues. *Proc. Natl. Acad. Sci. USA* **71**:3649–3653.
- Lindahl, T., P. Karran, and R. D. Wood. 1997. DNA excision repair pathways. *Curr. Opin. Genet. Dev.* **7**:158–169.
- Lindahl, T., and B. Nyberg. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**:3405–3410.
- Lutz, S. M., B. J. Vincent, H. H. J. Kazazian, M. A. Batzer, and J. V. Moran. 2003. Allelic heterogeneity in LINE-1 retrotransposition activity. *Am. J. Hum. Genet.* **73**:1431–1437.
- Moran, J. V., and N. Gilbert. 2002. Mammalian LINE-1 retrotransposons and related elements. Pp. 836–869 in N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. ASM Press, Washington, DC.
- Nekrutenko, A. 2004. Reconciling the numbers: ESTs versus protein-coding genes. *Mol. Biol. Evol.* **21**:1278–1282.
- Razin, A., and A. D. Riggs. 1980. DNA methylation and gene function. *Science* **210**:604–610.
- Rio, D. C. 2002. P transposable elements in *Drosophila melanogaster*. Pp. 484–518 in N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. ASM Press, Washington, DC.
- Roy, A. M., M. L. Carroll, S. V. Nguyen, A.-H. Salem, M. Oldridge, A. O. M. Wilkie, M. A. Batzer, and P. L. Deininger. 2000. Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**:1485–1495.
- Sémon, M., and L. Duret. 2004. Evidence that functional transcription units cover at least half the human genome. *Trends Genet.* **20**:229–232.
- Skarstad, K., T. A. Baker, and A. Kornberg. 1990. Strand separation required for initiation of replication at the chromosomal origin of *Escherichia coli* is facilitated by a distant RNA DNA hybrid. *EMBO J.* **9**:2341–2348.
- Sved, J., and A. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. USA* **87**:4692–4696.
- Waterston, R. H., K. Lindblad-Toh, E. Birney et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520–562.
- Webster, M. T., N. G. Smith, and H. Ellegren. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Mol. Biol. Evol.* **20**:278–286.
- Wiebauer, K., P. Neddermann, M. Hughes, and J. Jiricny. 1993. The repair of 5-methylcytosine deamination damage. Pp. 510–522 in J. P. Jost, and H. P. Saluz, eds. DNA Methylation: molecular biology and biological significance. Birkhäuser Verlag, Basel, Switzerland.
- Wolfe, K. H., P. M. Sharp, and W.-H. Li. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**:283–285.

Takashi Gojobori, Associate Editor

Accepted October 26, 2004