# Cytosine Deamination Plays a Primary Role in the Evolution of Mammalian Isochores

*Karl J. Fryxell* and Emile Zuckerkandl†*

*Department of Biology, George Mason University; and †Institute of Molecular Medical Sciences, Stanford, California

DNA melting is rate-limiting for cytosine deamination, from which we infer that the rate of cytosine deamination should decline twofold for each 10% increase in GC content. Analysis of human DNA sequence data confirms that this is the case for 5-methylcytosine. Several lines of evidence further confirm that it is also the case for unmethylated cytosine and that cytosine deamination causes the majority of all C→T and G→A transitions in mammals. Thus, cytosine deamination and DNA base composition each affect the other, forming a positive feedback loop that facilitates divergent genetic drift to high or low GC content. Because a 10°C increase in temperature in vitro increases the rate of cytosine deamination 5.7-fold, cytosine deamination must be highly dependent on body temperature, which is consistent with the dramatic differences between the isochores of warm-blooded versus cold-blooded vertebrates. Because this process involves both DNA melting and positive feedback, it would be expected to spread progressively (in evolutionary time) down the length of the chromosome, which is consistent with the large size of isochores in modern mammals.

## Introduction

Vertebrate chromosomes are composed of DNA segments called ‘‘isochores,’’ which are characterized by a bias in DNA base composition that is maintained over distances of ~0.2–1.3 Mb (Bernardi et al. 1985; Bernardi 1989, 1993*a*, 1993*b*; Bettecken et al. 1992; Beck et al. 1999; Dunham et al. 1999). GC-rich isochores are referred to as ‘‘heavy’’ (H) isochores and account for 35%–50% of the genome in birds and mammals. AT-rich isochores are referred to as ‘‘light’’ (L) isochores. H and L isochores are correlated with (although not identical to) cytological T and G chromosome bands, respectively (Holmquist 1989, 1992; Saccone et al. 1992; Bernardi 1995; Bernardi 2000). Fish and amphibians have neither H isochores nor well-defined cytological chromosome bands (Bernardi et al. 1985; Bernardi 1993*b*, 1995).

Isochore-related biases in base composition are found in all parts of mammalian genes (exons, introns, etc.) and remain relatively consistent along the length of an isochore (Bernardi et al. 1985; Bernardi 1993*b*, 1995). Nevertheless, closely related members of the same gene family often have quite different GC contents (Bernardi et al. 1985; Li and Graur 1991; Ellsworth, Hewett-Emmett, and Li 1994). In mammals, α-globin genes are GC-rich but β-globin genes are AT-rich. In birds, both α- and β-globin genes are GC-rich. Why should base composition be poorly conserved between closely related genes on different chromosomes, or between warm- versus cold-blooded vertebrates, but well conserved between genes from different gene families within the same isochore? These questions have been long-standing puzzles in molecular evolution (Li and Graur 1991; Bernardi 1995).

One conjecture, the ‘‘selectionist’’ hypothesis, holds that natural selection favored a different base composition for each type of isochore (Bernardi et al. 1985, 1988; Bernardi 1993*a*, 1993*b*). To date, the strongest evidence for the selectionist hypothesis has been obtained from noncoding and silent-site substitutions in GC-rich genes in the mammalian major histocompatibility complex (MHC). G/C→A/T mutant alleles (polymorphisms) in the MHC were found to occur in higher numbers, but smaller allelic frequencies, than would be expected if these genes had been in mutational equilibrium (Eyre-Walker 1999). This is an important observation, because allelic diversity and frequencies can help establish rates of mutation and selection. One caveat is that the calculations were based on the ‘‘infinite sites’’ model, which may not apply because of the high rate of simultaneous double-nucleotide substitutions (Averof et al. 2000), because the MHC experiences a high rate of gene conversion (Eyre-Walker 1999), because gene conversion typically converts a continuous tract about 1 kb in length (Curtis et al. 1989), and because a small degree of sequence divergence has a major effect on the frequency and length of gene conversion tracts (Lukacsovich and Waldman 1999). Another caveat is that the calculations assumed mutational equilibrium and unbiased DNA repair, which may not hold either (see *Discussion*). It is notable that direct selection for GC content, in the conventional sense of altered reproductive fitness based solely on base composition, has not been demonstrated, and it remains unclear whether such selection would occur at the level of DNA, RNA, or protein (D’Onofrio et al. 1999; Bernardi 2000). It is also unclear why genes as similar as α- versus β-globins evolved dramatically different base compositions in mammals but not in birds (Bernardi et al. 1985; Li and Graur 1991).

An alternative explanation, the ‘‘mutationist’’ hypothesis, attributes the formation of isochores to regional variations in mutation pressures. This hypothesis was originally motivated by the observation that H isochores tend to replicate earlier in the S phase of the cell cycle, suggesting that some aspect of DNA replication or repair might vary during the S phase (Goldman et al.

Key words: cytosine deamination, 5-methylcytosine deamination, mammalian genes, homeothermy, isochore evolution, computational molecular biology.

Address for correspondence and reprints: Karl J. Fryxell, Department of Biology, MSN 3E1, George Mason University, Fairfax, Virginia 22030. E-mail: kfryxell@gmu.edu.

*Mol. Biol. Evol.* 17(9):1371–1383. 2000

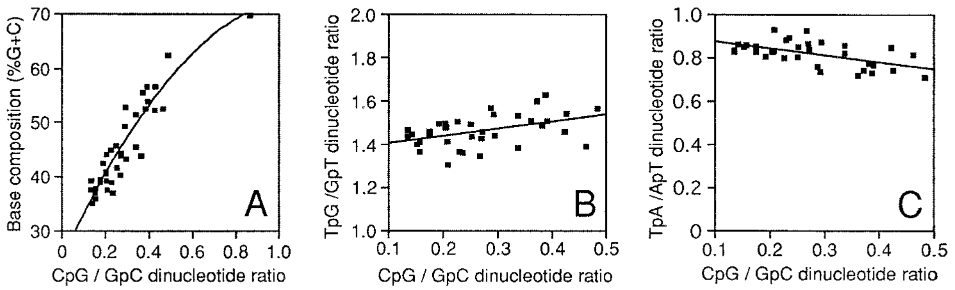

FIG. 1.—Base composition and dinucleotide frequencies in human chromosomal DNA. All human genomic DNA sequences >50 kb in release 96 of GenBank were analyzed (see *Materials and Methods*). *A*, GC content versus the CpG/GpC dinucleotide ratio. The curve shown was obtained by the method of least squares, has a correlation coefficient of 0.856, and corresponds to equation (1). An additional data point at the top, only in *A*, represents the median of 12 human CpG islands (Aïssani and Bernardi 1991*a*). *B*, The TpG/GpT dinucleotide ratio versus the CpG/GpC dinucleotide ratio. The line shown was obtained by the method of least squares and has a correlation coefficient of 0.204. *C*, The TpA/ApT dinucleotide ratio versus the CpG/GpC dinucleotide ratio. The line shown was obtained by the method of least squares and has a correlation coefficient of 0.331.

1984; Leeds, Slabourgh, and Mathews 1985; Filipski 1987; Wolfe, Sharp, and Li 1989; Eyre-Walker 1994; Gu and Li 1994). To date, the strongest evidence for the mutationist hypothesis is that pseudogenes in GC-rich isochores accumulate GC-biased base substitutions, while pseudogenes in AT-rich isochores accumulate AT-biased base substitutions (Francino and Ochman 1999). This is an important observation, because pseudogenes have no known function. If GC-rich pseudogenes were maintained by negative selection acting on GC content, while AT-rich pseudogenes were subject to genetic drift, as has been proposed (Bernardi 2000), then base substitutions would be fixed at a lower rate in GC-rich than in AT-rich pseudogenes. The opposite was observed (Francino and Ochman 1999). One caveat is that selection generally acts on the changing alleles produced by genetic drift through both positive and negative selection, and this may or may not have any overall effect on the base substitution rate. Another caveat is that the pseudogenes were not sequenced in all of the same species, and the divergence dates of these species were only approximately known, so the inferred substitution rates were approximate (Francino and Ochman 1999). The mutationist hypothesis has also failed to explain why constitutive heterochromatin, which is replicated near the end of the cell cycle, is often GC-rich (Bernardi et al. 1988).

The dinucleotide CpG is found in the genomes of birds and mammals at ~¼ of its statistically expected frequency (Jabbari and Bernardi 1998). This underrepresentation is caused by the hypermutability of CpG in humans and other species (Coulondre et al. 1978; Bird 1980; Britten et al. 1988; Cooper and Krawczak 1989; Green et al. 1990; Sved and Bird 1990; Jones et al. 1992; Spruck, Rideout, and Jones 1993), which, in turn, is due to the fact that cytosine is methylated only in CpG dinucleotides (in vertebrates). Both cytosine and 5-methylcytosine undergo high rates of spontaneous hydrolytic deamination, but deamination of 5-methylcytosine produces thymine, and mismatch repair of C→T transitions is less efficient than that of C→U transitions (Coulondre et al. 1978; Razin and Riggs 1980; Ehrlich et al. 1986; Wiebauer et al. 1993).

The CpG dinucleotide is underrepresented in L isochores to a greater extent than in H isochores. The reason for this is not well understood, but it is known that there is a general correlation between GC content and CpG/GpC dinucleotide ratios in all mammalian DNA sequences, including exons, introns, CpG islands, mammalian viruses, and long genomic sequences (Bernardi et al. 1985; Aïssani and Bernardi 1991*a*; Bernardi 1993*b*; Jabbari and Bernardi 1998). The simplicity and reproducibility of this correlation may reflect a fundamental process underlying the molecular evolution of isochores (fig. 1). We undertook a series of computer simulations and quantitative DNA sequence analysis to clarify this point. Our initial goal was simply to estimate the relative contribution of 5-methylcytosine deamination to the GC content of human isochores. In order to solve this problem, we found it necessary to analyze the effect of GC content on the rate of 5-methylcytosine deamination, as well as the relation between the GC bias of other base substitutions (excluding 5-methylcytosine deamination) and GC content. Our results show that the deamination of 5-methylcytosine reduces the GC content of the human genome by ~10%. Our results also indicate that the deamination of unmethylated cytosine is primarily responsible for the maintenance of differences in GC content between isochores.

## Materials and Methods

### Simulations of DNA Sequence Evolution

Computational simulations of DNA sequence evolution were performed with sequences 100 kb in length on a personal computer. A variety of initial sequences were used, including random sequences with any specified GC content. In some simulations, nonrandom initial sequences were used, based on tandem repeats of a short sequence (such as CATG). This allowed us to alter the initial dinucleotide frequencies without changing the initial GC content.

Each generation in these simulations corresponded to the time required for the evolutionary fixation of base substitutions in 1% of the sequence (excluding 5-methylcytosine transition mutations). We will refer to this time as a unit evolutionary period (UEP). Equilibrium values were obtained after calculation of 500–1,000 UEPs. Mutations were limited to base substitutions (i.e., insertions, deletions, and duplications were not included in these simulations) and were produced by the computational procedures described below.

### The *5mCt* Function

The *5mCt* (5-methylcytosine transition mutations) function was used to simulate the deamination of 5-methylcytosine. This function was invoked during replication of CpG dinucleotides, when it was triggered by a pseudorandom number generator with a probability that varied between simulations. The probability was varied over the range from 0 to 1 per UEP, which corresponds approximately to the range inferred from DNA sequence studies of the mutability of CpG sequences in human genetic diseases (Britten et al. 1988; Sved and Bird 1990). A *5mCt* value of 0.01 means that there was a probability of 0.01 (per UEP) of a 5-methylcytosine transition mutation on the sense strand (resulting in a CpG→TpG transition mutation), as well as a probability of 0.01 on the antisense strand (resulting in a CpG→CpA transition mutation). These are equally probable, because the two DNA strands are methylated symmetrically (Razin and Riggs 1980).

Random mutations in CpG dinucleotides were also allowed and were produced independently by the *OB* function (see below). However, each base was allowed to mutate not more than once, by any mechanism, per UEP. Mutation of either base in a CpG dinucleotide (by any mechanism) precluded subsequent 5-methylcytosine transition mutations within the same UEP, because the first mutation would prevent subsequent methylation of the other strand (Razin and Riggs 1980; Razin and Cedar 1993).

### The *OB* Function

The *OB* (other base substitutions, besides CpG deamination) function was included to model the effects of random base substitutions. *OB* allows the user to independently specify the transition/transversion ratio and the GC bias of random base substitutions. This was implemented by subdividing the numerical range from 0 to 1 into subsegments whose lengths were proportional to the probability of each of the possible new bases, and then using a pseudorandom number to select the base.

The overall probability of an *OB* mutation was fixed at 0.01 per base per generation, because each generation in these simulations was defined to be a UEP (see above). However, the possibility of mutating each base in the sequence was independently tested with a separate call to a pseudorandom number generator, so that the total number of mutations per generation was subject to stochastic fluctuations, as it is in real organisms. If a mutation was triggered at a particular site, then the new base was selected with an additional pseudorandom number as described above.

### The *MCG* Function

*MCG* (mutations in CpG dinucleotides) is similar to *5mCt* except that, when mutation of a particular CpG dinucleotide is triggered in *MCG*, the actual mutation is executed through *OB*, producing the same spectrum of transitions, transversions, and so on that *OB* uses for other base substitutions. This allowed us to distinguish between effects that are specifically caused by the CpG→TpG transition per se and more general effects that depend only on CpG mutability.

### Human DNA Sequences

All human genomic DNA sequences >50 kb in length from release 96 of GenBank were used for sequence analysis. These comprise 37 sequences, containing a total of 4.3 Mb of sequence data, with the following accession numbers: L29074, L44140, U50871, L43581, U07563, U40455, U52111, U52112, Z72519, L05367, X87344, Z72519, L36092, Z72001, Z73358, L10641, L11910, M26434, M63544, M94081, U01317, U07000, U07562, U47924, Z72004, U51244, X90568, Z70272, Z71182, J03071, L35265, M89651, U03115, U35072, L39891, L47234, and X55448. We calculated the GC content and dinucleotide frequencies of each of these human sequences with the appropriate portions of our computer software described above (i.e., the portions which were also used to record these parameters during theoretical simulations).

## Results

### Effect of CpG Hypermutability on DNA Base Composition

We wrote a computer program for simulating some aspects of the evolution of DNA sequences, including three functions: *5mCt*, *OB*, and a scoring module to record mono- and dinucleotide frequencies and ratios (see *Materials and Methods*). Each generation in these simulations corresponded to the time required for the evolutionary fixation of base substitutions in 1% of the sequence (excluding 5-methylcytosine transition mutations) (i.e., the UEP).

When *5mCt* was initiated in a random sequence, the CpG/GpC dinucleotide ratio declined rapidly for ~5 UEP, equilibrated in ~10 UEP, and remained essentially constant thereafter (fig. 2*A* and *B*). The GC content also declined rapidly for ~10 UEP but then continued to decline slowly for an additional 200 UEP (fig. 2*A–D*). Evidently, the decrease in GC content in these simulations is caused by two processes with different kinetics. Further investigation showed that, in general, the duration of the rapid phase is 2/*5mCt*, where *5mCt* is the probability of a 5-methylcytosine transition mutation per CpG per UEP. The rapid phase of decline in GC content ends when the initial CpG dinucleotides have been eliminated, as expected if the rapid phase simply reflects the declining levels of the CpG dinucleotide. The duration of the slow phase is 2/*P*(*OB*), where *P*(*OB*) is the probability of a random (*OB*) mutation per base per UEP. In other words, the slow phase ends when the effects of

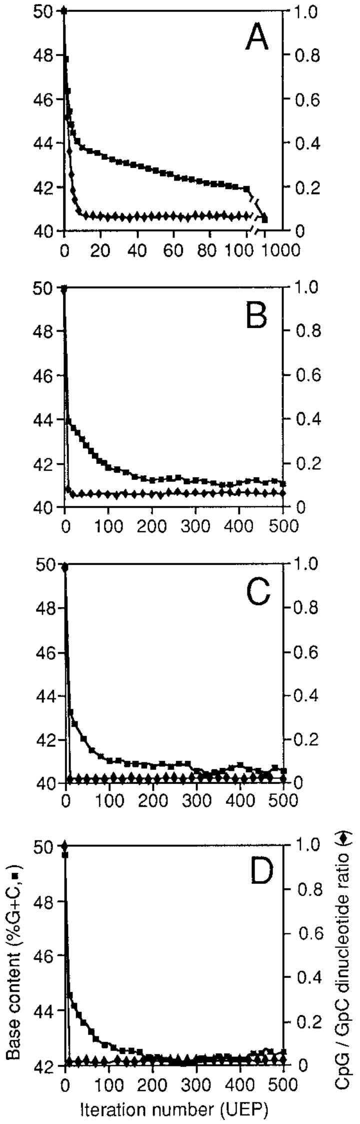

FIG. 2.—The deamination of 5-methylcytosine (*5mCt*) reduces GC content in two distinct kinetic phases. Other base substitutions (*OB*) had a GC bias of 50% (see *Materials and Methods*). The initial sequence had a GC content of 50% with random base order. *A, 5mCt* = 0.2 per CpG per UEP, *OB* transition/transversion ratio = 0.5. *B, 5mCt* = 0.25 per CpG per UEP, *OB* transition/transversion ratio = 1.5. *C, 5mCt* = 1.0 per CpG per UEP, *OB* transition/transversion ratio = 1.0. *D, MCG* = 1.0 per CpG per UEP, *OB* transition/transversion ratio = 1.0.

*OB* on base composition have equilibrated. *OB* substitutions do include A/T→G/C base substitutions, but these will contribute less to the long-term equilibration of base composition if the CpG's they create are short-lived. That is, *5mCt* acts as a ''CpG sink'' that biases the equilibration between A/T→G/C and G/C→A/T substitutions produced by *OB*. The slow phase in decline of GC content is not mediated by the production or destruction of TpG or CpA dinucleotides, as shown by computer simulations with alternative procedures to maintain low levels of CpG, which produced the same slow phase (fig. 2*D*) without high levels of TpG or CpA. Because TpG and CpA have GC contents of 50%, and the *OB* function had a GC bias of 50% in the simulations in question, the net base composition of these dinucleotides must have been unchanged by random point mutation, and hence their mutational decay could not contribute any net change to the GC content in these simulations. The kinetics and magnitude of the slow phase were also independent of the *OB* transition/transversion ratio (fig. 2*B* and *C* and additional data not shown).

In previous attempts to simulate the evolutionary effects of 5-methylcytosine deamination, Sved and Bird (1990) demonstrated the rapid phase but not the slow phase in decline of GC content. Their differing results are attributable to their use of several approximations, particularly their assumption that all dinucleotide frequencies are independent variables (which is not the case; see below). Cooper and Krawczak (1993) used a different set of equations, and these did appear to produce a two-phase decline in GC content, but Cooper and Krawczak assumed that the two-phase decline was caused by their starting sequence and did not investigate further.

To determine the cumulative effect of *5mCt* on the equilibrium GC content, additional computer simulations were continued for 500 UEP, and the rates of *5mCt* were varied between simulations. The results showed a systematic relationship between the GC content and the CpG/GpC dinucleotide ratio at equilibrium, which was well fit by a quadratic equation (fig. 3). This equilibrium relationship was not significantly affected by the initial GC content (fig. 3*B*), the initial frequency of CpG dinucleotides (fig. 3*A* and *C*), the transition/transversion ratio (fig. 3*A–C*), or even which procedure was used to maintain low levels of CpG (fig. 3*D*). Thus, the effect of *5mCt* on GC content was not an artifact of any of these parameters.

The effect of *5mCt* on GC content increased dramatically as the GC bias of *OB* was increased (fig. 4). This is attributable to the fact that a higher GC bias of *OB* increases the rate at which random base substitutions produce new CpG dinucleotides, and thus more 5-methylcytosine deamination events occur (fig. 4*C*). At maximal *5mCt* rates, ~⅓ of each increase in the GC bias of *OB* was offset by an increase in effectiveness of *5mCt*, so that each increment in *OB* lifted the curve and increased its curvature (fig. 4*A*).

## TpA and TpG Dinucleotide Levels

The TpA/ApT dinucleotide ratios in our simulations were <1.0 (fig. 4*D*) and declined in proportion to the CpG/GpC dinucleotide ratio (fig. 5). TpA is underrepresented in vertebrate DNA sequences (Bulmer 1987; Karlin and Mrázek 1996). The relation between CpG/GpC and TpA/ApT in our simulations did not depend on the initial base composition, the initial frequency of CpG or TpA dinucleotides, or the transition/transversion ratio (fig. 5*A–C*). However, alternative computational procedures to maintain low levels of CpG dinucleotides (*MCG*; see *Materials and Methods*) did not reproduce the effect of *5mCt* on the TpA/ApT ratio (fig. 5*D*), even

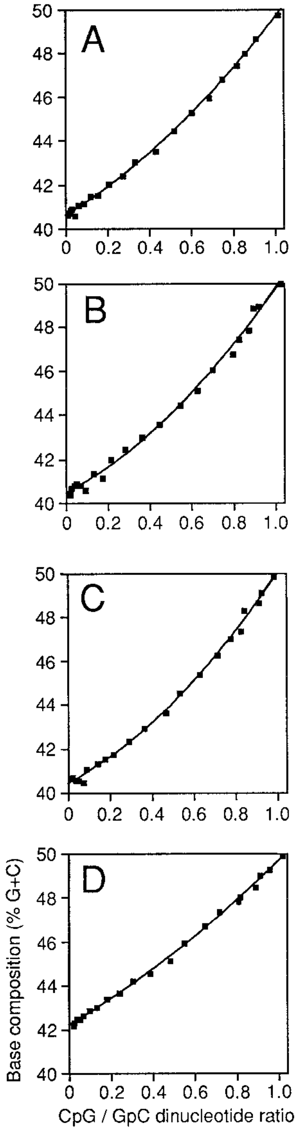

FIG. 3.—The effect of 5-methylcytosine deamination (*5mCt*) on GC content at equilibrium. *OB* (other base substitutions, besides *5mCt*) had a GC bias of 50% (see *Materials and Methods*). Unless stated otherwise, the initial sequence had a GC content of 50%, the initial base order was randomized, and *5mCt* was used to reduce CpG frequencies. *A, OB* transition/transversion ratio = 1.5. *B, OB* transition/transversion ratio = 1.0, initial GC content = 0% (similar results were obtained with an *OB* transition/transversion ratio of 0.5, or with an initial GC content of 50% or 100%). *C, OB* transition/transversion ratio = 1.0, initial sequence = tandem CATG repeats (i.e., the CpG/GpC ratio was initially 0). *D, OB* transition/transversion ratio = 0.5; *MCG* was used rather than *5mCt* to reduce the frequency of CpG dinucleotides.

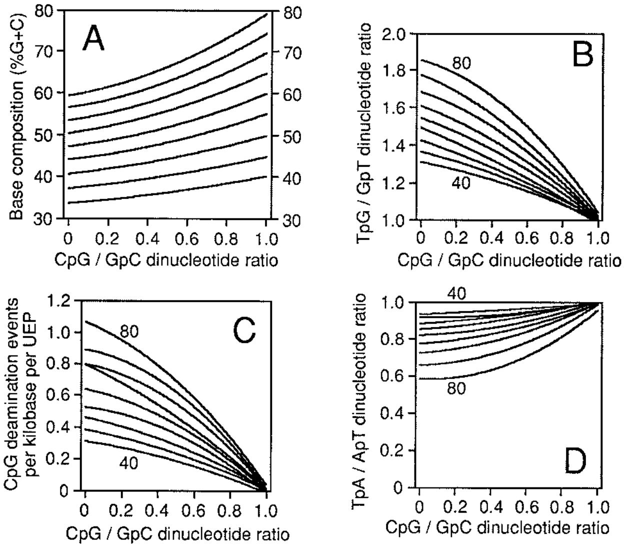

FIG. 4.—Computer simulation of the combined effects of 5-methylcytosine deamination (*5mCt*) and the GC bias of other base substitutions (*OB*). Each curve represents a least-squares fit of a quadratic equation to 21 data points, where each data point represents the equilibrium values obtained in a computer simulation with selected rates of *5mCt* and GC biases of *OB.* Individual data points were too numerous to be shown here. Within each curve, the rate of *5mCt* varied and the GC bias of *OB* was held constant. The GC bias of *OB* varied between curves from 40% to 80% in 5% increments. The *OB* transition/transversion ratio was 1.5 for all graphs shown here (similar results were obtained with a transition/transversion ratio of 0.5). *A,* Constant *OB* curves plotted with respect to GC content. The GC bias of *OB* was 40% in the bottom curve and is labeled on the right of the figure. Correlation coefficients in this graph were all >0.99. *B,* Constant *OB* curves plotted with respect to the TpG/GpT ratio. The GC bias of OB was 40% in the bottom curve and 80% in the top curve (see labels). Correlation coefficients in this graph averaged >0.99. *C,* Constant *OB* curves plotted with respect to the absolute CpG deamination rate (per kb per UEP) at equilibrium. The GC bias of OB was 40% in the bottom curve and 80% in the top curve (see labels). Correlation coefficients in this graph averaged >0.97. *D,* Constant *OB* curves plotted with respect to the equilibrium TpA/ApT ratio. The GC bias of OB was 80% in the bottom curve and 40% in the top curve (see labels). Correlation coefficients in this graph averaged 0.94.

though these procedures did reproduce the slow phase of decline in GC content (fig. 2*B*). That is, the relation between CpG/GpC and TpA/ApT was specifically caused by increased levels of TpG and CpA, but the slow phase of decline in GC content was not. This conclusion was confirmed by detailed studies of the variation of dinucleotide levels (with time) during individual simulations (not shown), as well as variation between simulations with varying GC bias of *OB* (fig. 4*B* and *D*), all of which confirmed that an inverse relation between TpG/GpT and TpA/ApT held in all cases.

The underrepresentation of TpA may be understood as follows. It is known that TpG and CpA are overrepresented in mammalian DNA sequences because they are produced by deamination of 5-methylcytosine in CpG dinucleotides (Bulmer 1987; Sved and Bird 1990; Karlin and Mrázek 1996). By definition, TpG overrepresentation means that any particular T has an increased likelihood of being followed by a G, and therefore a decreased likelihood of being followed by any of the other possible bases (A, C, or T). Similarly, CpA overrepresentation means any given A has an increased likelihood of being preceded by a C, and therefore a decreased likelihood of being preceded by any of the other possible bases (T, G, or A). The abundances of TpA and CpG are reduced by both statistical effects and thus will show a greater reduction than dinucleotides that are affected by only one of these statistical effects. This is the first demonstration that CpG underrepresentation and

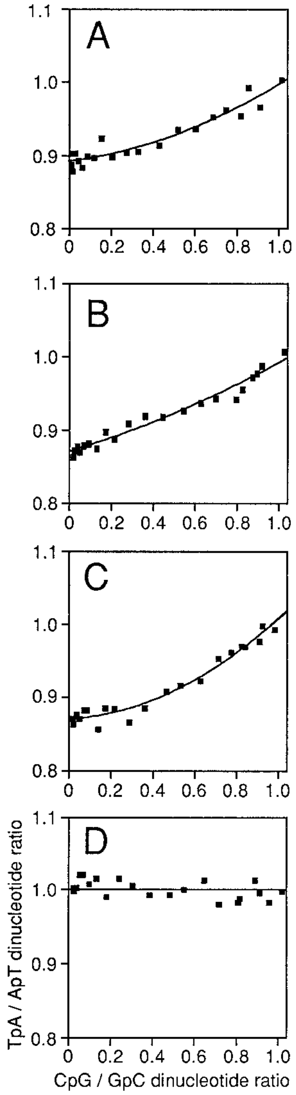

FIG. 5.—Deamination of 5-methylcytosine causes underrepresentation of the TpA dinucleotide. Unless stated otherwise below, the GC bias of *OB* was 50%, the initial GC content was 50%, the initial base order was randomized, and *5mCt* was used to reduce CpG frequencies (see *Materials and Methods*). *A, OB* transition/transversion ratio = 1.5. *B,* GC content of the initial sequence = 0% with random base order; *OB* transition/transversion ratio = 1.0 (an initial GC content of 50% or 100% produced similar results, as did an *OB* transition/transversion ratio of 0.5). *C,* Initial sequence = perfect tandem CATG repeats (i.e., the initial CpG/GpC and TpA/ApT ratios were 0); *OB* transition/transversion ratio = 1.0 (a transition/transversion ratio of 1.5 gave similar results). *D, MCG* was used to reduce the frequency of CpG dinucleotides; *OB* transition/transversion ratio = 0.5.

TpA underrepresentation are caused by the same process.

If the CpG/GpC ratio was held constant in our simulations, then the equilibrium TpG/GpT ratio increased as the GC bias of *OB* was increased (fig. 4*B*). Increasing the GC bias of *OB* increases the rate at which CpG dinucleotides are created, which increases the number of 5-methylcytosine deamination events (fig. 4*C*) and hence increases the rate at which TpG dinucleotides are created (fig. 4*B*). If the GC bias of *OB* was held constant, then TpG/GpT ratios were inversely proportional to the CpG/GpC ratio (fig. 4*B*), because the CpG/GpC ratio is inversely proportional to the number of deamination events per kilobase at equilibrium (fig. 4*C*). That is, plots of TpG/GpT versus CpG/GpC have a negative slope if the GC bias of *OB* is held constant (e.g., for the curve in fig. 4*B* with the GC bias of *OB* = 50%, the slope is significantly less than 0; $P < 0.001$ by the *t*-test), but they have a positive slope for human DNA sequences (fig. 1*B*; $P < 0.01$). Conversely, plots of TpA/ApT have a positive slope if the GC bias of *OB* is held constant (fig. 5*A*; $P < 0.001$), but they have a negative slope for human DNA sequences (fig. 1*C*; $P < 0.001$). These observations can be explained only if the GC bias of *OB* varies along with the rate of *5mCt* in human DNA. In other words, the human DNA curve in figure 1*A* is essentially equivalent to tracing a path that crosses all of the constant *OB* curves in figure 4*A*. It is immaterial to our analysis (at this point) whether variation in the GC bias of *OB* is caused by natural selection or mutation pressure—we simply observe that the GC bias of *OB* does vary.

## The Relation Between DNA Base Composition and CpG Mutability

In order to determine the rate of 5-methylcytosine transitions (*5mCt*) in human chromosomal DNA, we selected points at intervals of 5% GC content in figure 1*A* and obtained the consensus value of CpG/GpC at this point from the best fit equation:

$$\%\text{GC content} = 24.583 + 86.942[\text{CpG/GpC}] - 39.297[\text{CpG/GpC}]^2. \quad (1)$$

The CpG/GpC ratio was chosen because this dinucleotide ratio responds specifically to the deamination of 5-methylcytosine. Other types of mutations do not affect the CpG/GpC ratio because they occur equally at GpC dinucleotides. In fact, varying the GC bias of *OB* in computer simulations did not affect CpG/GpC (i.e., the curves in fig. 6*B* are all nearly horizontal lines). The human values of GC content and CpG/GpC ratio from equation (1) were used to solve for the corresponding value of *5mCt* by linear interpolation between the family of equations illustrated in figure 6*A*. The resulting *5mCt* values are shown in fig. 6*C* and were best fit by the following exponential equation:

$$5mCt = 69(10^{-3.0B}), \quad (2)$$

where *5mCt* is the rate of 5-methylcytosine transitions per 100 CpG per UEP (see *Materials and Methods*) and *B* is the GC content in vivo (expressed as a decimal fraction).

## Rates of Cytosine Deamination In Vitro

We found three published measurements of the rate of cytosine deamination in native, double-stranded DNA (Lindahl and Nyberg 1974; Frederico, Kunkel, and Shaw 1990). When graphed on an Arrhenius plot, these fall on a perfectly straight line (fig. 6*D*) that corresponds to the following equation:

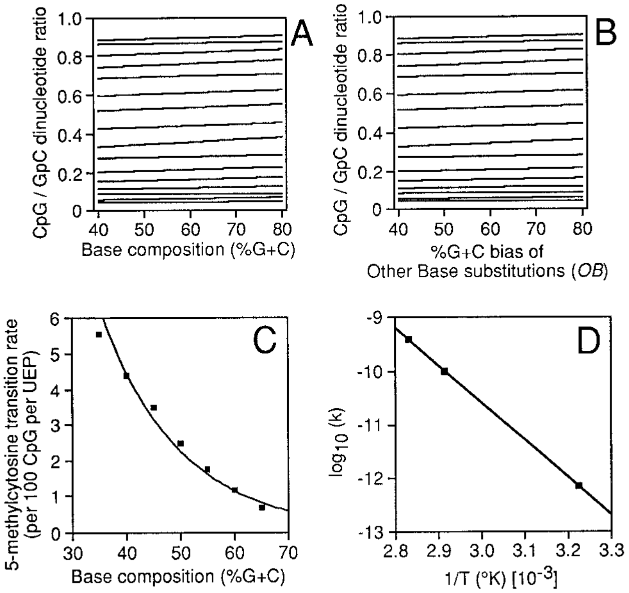

FIG. 6.—Rates of cytosine deamination. *A,* Constant *5mCt* curves plotted with respect to the equilibrium GC content. *5mCt* was 25 (per 100 CpG per UEP) in the bottom curve, was less by a factor of $\sqrt{1/2}$ in each successively higher curve, and was 0.14 in the top curve. Each curve represents a least-squares fit of a linear equation to nine data points (correlation coefficients averaged 0.39), where each data point represented the endpoint of a computer simulation. The GC bias of *OB* varied within each curve, but the *OB* rate (0.01 per UEP) and transition/transversion ratio (1.5) remained constant. *B,* Constant *5mCt* curves plotted with respect to the GC bias of *OB.* Other details were as in *A. C,* Rates of 5-methylcytosine deamination (*5mCt*) as a function of base composition in the human genome. Values of *5mCt* in human DNA were obtained by linear interpolation from the equations in figures 1 and 4 as described in the text. The best-fit curve shown here was obtained by the method of least squares, has a correlation coefficient of 0.98, and corresponds to equation (2). *D,* Arrhenius plot of cytosine deamination in double-stranded (nondenatured) DNA (Lindahl and Nyberg 1974; Frederico, Kunkel, and Shaw 1990). The best-fit line was obtained by the method of least squares, has a correlation coefficient of 1.000, and corresponds to equation (3).

$$\log[k] = 10.425 - 7004.2/T, \qquad (3)$$

where $k$ is the rate constant in $s^{-1}$ and $T$ is the temperature in °K. This equation indicates that a 10°C increase in temperature increases the rate of cytosine deamination in double-stranded DNA by $k_{37°C}/k_{27°C} = (7.0 \times 10^{-13}/s)/(1.23 \times 10^{-13}/s) = 5.7$-fold.

At constant temperature, single-stranded DNA undergoes cytosine deamination ~143-fold more rapidly than double-stranded DNA (Frederico, Kunkel, and Shaw 1990). This dramatic difference is due to the fact that the deamination of cytosine (or 5-methylcytosine) in double-stranded DNA requires temporary, local strand separation (melting). The requirement for DNA melting has been confirmed not only by the reaction mechanism (which 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) and activation energies (which are identical in single-stranded and double-stranded DNA, indicating that the reaction intermediates have the same, single-stranded, conformation), but also by elegant genetic experiments in vivo (which have proven that single-base mismatches dramatically accelerate the rate of cytosine deamination; see Lindahl and Nyberg 1974; Ehrlich et al. 1986; Frederico, Kunkel, and Shaw 1990, 1993). Thus, a decrease in the DNA melting temperature ($T_M$) by 10°C will have the same effect on the rate of cytosine deamination as an increase of 10°C in temperature. Given that a 10% decrease in GC content reduces $T_M$ by 4.1°C (Wahl, Berger, and Kimmel 1987), it follows that a 10% change in GC content will change the rate of cytosine deamination by $k_{37°C}/k_{32.9°C} = (7.0 \times 10^{-13}/s)/(3.46 \times 10^{-13}/s) = 2.0$-fold. This corresponds to the following equation:

$$D = C(10^{-3.0B}), \qquad (4)$$

where $D$ is the rate of cytosine deamination in vivo, $C$ is an arbitrary constant, and $B$ is the DNA base composition (expressed as a decimal fraction). Note that equation (4), which we derived from physical and chemical studies of DNA melting and cytosine deamination, is formally identical to equation (2), which we derived from human DNA sequence data. Thus, the correlation between GC content and the CpG/GpC ratio in the mammalian genome follows directly from the physics and chemistry of DNA melting and cytosine deamination. The residual differences between isochores attributable to differential DNA methylation are negligible, and therefore DNA methylation must be relatively uniform on an isochore scale in the germ line, which is consistent with available data (when embryos and adults of both sexes are included [Razin and Cedar 1993; Sasaki, Allen, and Surani 1993; Rubin et al. 1994]).

We note that Eason and colleagues previously suggested that high GC content might help protect CpG's against cytosine deamination (Adams et al. 1987), although the sequence data available at that time were insufficient to support their hypothesis (Gardiner-Garden and Frommer 1987).

## Computer Simulation of Human Isochores

To separate the effects of *5mCt* (on GC content) from the effects of *OB,* we selected points at intervals of 5% GC content in figure 1*A,* obtained the value of CpG/GpC at these points from equation (1) as before, and then obtained the corresponding GC bias of *OB* by linear interpolation between the family of curves in figure 4*A.* The value of *5mCt* at this point was also checked by linear interpolation between the family of curves in figure 6*B.* The resulting values of the GC bias of *OB* and the rate of *5mCt* were well fit by a linear equation (fig. 7*D*):

$$OB = 78 - 6.9(5mCt), \qquad (5)$$

where *5mCt* is the rate of 5-methylcytosine transitions (per 100 CpG per UEP) and *OB* is the GC bias of all other base substitutions. Equation (5) was used as the basis of computer simulations in which the GC bias of *OB* and the rate of *5mCt* were covaried between simulations and the GC content was allowed to reach equilibrium within each simulation. These simulations reproduced the covariance of GC content and dinucleotide

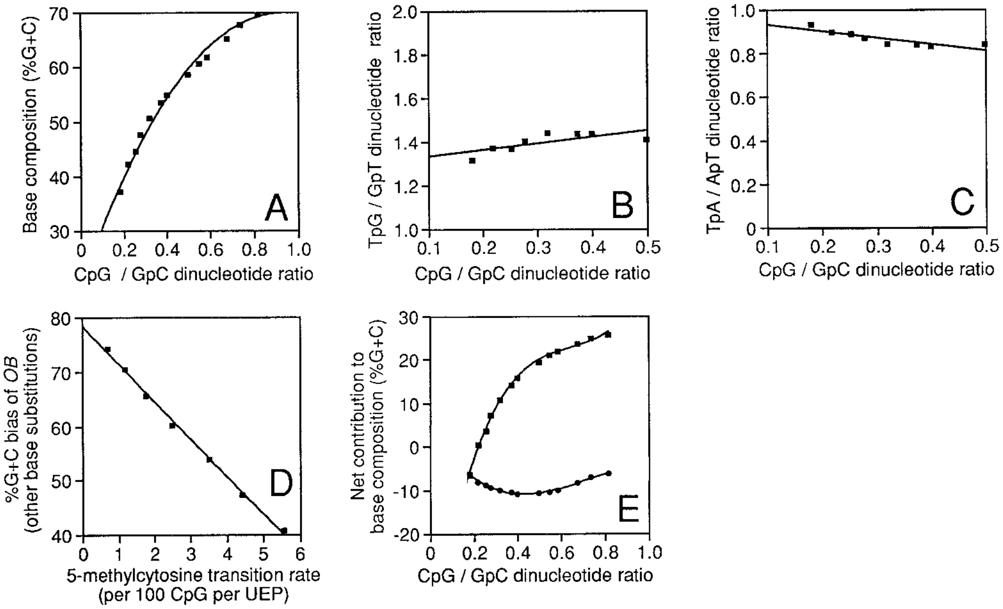

FIG. 7.—Computer simulation of human isochores. The rate of *5mCt* varied between simulations, the corresponding GC bias of *OB* was determined by equation (5) (this is illustrated in *D*), and the GC content was allowed to reach equilibrium within each simulation. *A,* Results of model simulations: plot of GC content as a function of the CpG/GpC ratio (compare with fig. 1*A*). *B,* Plot of TpG/GpT as a function of CpG/GpC (compare with fig. 1*B*). *C,* Plot of TpA/ApT as a function of CpG/GpC (compare with fig. 1*C*). *D,* The basis of this model: inverse linear covariance of the GC bias of *OB* with the rate of *5mCt* in the human genome. Values of the GC bias of *OB* and the rate of *5mCt* were derived from human DNA sequence data as described in the text. The best-fit line was obtained by the method of least squares, has a correlation coefficient of 0.997, and corresponds to equation (5). *E,* Analysis of the model, showing the net contributions of *5mCt* (●) and *OB* (■) calculated from model simulations, based on the following equation: %GC content = 50 + $5mCt_{\mathrm{contrib}}$ + $OB_{\mathrm{contrib}}$. $5mCt_{\mathrm{contrib}}$ was calculated from the equation $5mCt_{\mathrm{contrib}}$ = [GC bias of *OB*] − [equilibrium GC content], because the GC content in these simulations equilibrates to the GC bias of *OB* in the absence of 5-methylcytosine deamination (see fig. 4*A*).

frequencies in the human genome (compare fig. 7*A* with fig. 1*A*). Moreover, plots of TpG/GpT from these isochore simulations have a positive slope (fig. 7*B*; $P < 0.05$ by the *t*-test), and plots of TpA/ApT have a negative slope (fig. 7*C*; $P < 0.01$), as they do in human DNA sequences (see above and fig. 1*B* and *C*). In both cases, these slopes are caused by the inverse linear covariance of *OB* and *5mCt.* Increasing the GC bias of *OB* increases the rate at which CpG dinucleotides are created and hence the rate at which TpG dinucleotides are created, while TpA levels follow (inversely) those of TpG (see previous sections). The ability of equation (5) to recreate all of these correlations confirms that it is a good approximation to the processes that maintain isochores in human DNA.

Analysis of the model shown in figure 7 indicates that *5mCt* reduces the GC content in human DNA by 10%–11% in sequences with CpG/GpC ratios of 0.3–0.6 (fig. 7*E*). At lower CpG/GpC ratios, the *5mCt* contribution decreases slightly, due to a lower rate of creation of CpG dinucleotides (produced by the lower GC bias of *OB*). At higher CpG/GpC ratios, the *5mCt* contribution also decreases slightly, due to the lower rate of *5mCt* (produced primarily by the increasing GC content and secondarily by hypomethylation of CpG islands; see *Discussion*).

## The Deamination of Unmethylated Cytosine

The linear relationship between the GC bias of *OB* and the rate of *5mCt* (eq. 5 and fig. 7*D*) could be explained if the factors affecting the deamination of unmethylated cytosine were similar to those affecting the deamination of 5-methylcytosine. We note that this is the case in vitro (Lindahl and Nyberg 1974; Ehrlich et al. 1986; Frederico, Kunkel, and Shaw 1990, 1993). Recall that

$$TB = v/(u + v), \qquad (6)$$

where *TB* is the total bias of all base substitutions (including *OB* and *5mCt*), *u* is the overall rate of G/C→A/T substitutions, and *v* is the overall rate of A/T→G/C substitutions (Sueoka 1988). Because DNA melting is rate-limiting for cytosine deamination, the $5mCt = 0$ intercept of equation (5) should correspond to the minimal rate of cytosine deamination. Because cytosine is by far the most unstable base in vitro (Lindahl and Nyberg 1974; Ehrlich et al. 1986; Frederico, Kunkel, and Shaw 1990, 1993), the minimal rate of cytosine deamination should correspond to the minimal value of *u* (which we will call $u_{\mathrm{min}}$). In the GC-rich globin pseudogenes of higher primates (Francino and Ochman 1999), $u = 33.84\%$ and $v = 45.58\%$. Assuming these rates are similar at $5mCt = 0$ (where $OB = TB$, by definition) and combining equations (5) and (6), we obtain $(100\%)(45.58)/(u_{\mathrm{min}} + 45.58) = 78\%$, and hence $u_{\mathrm{min}} \approx 12.9\%$.

The total rate of C→A and G→T transversions in GC-rich globin pseudogenes is 6.9% (Francino and Ochman 1999). Assuming a similar rate at $5mCt = 0$, it follows that $u_{\mathrm{min(transitions)}} = u_{\mathrm{min}} - u_{\mathrm{min(transversions)}} \approx$

12.9% − 6.9% ≈ 6.0%. That is, we postulate that C→T and G→A transition mutations are caused by two distinct biochemical pathways. The first pathway requires cytosine deamination and doubles in rate for each 10% decline in GC content. The second pathway(s) does not require cytosine deamination and occurs at a relative rate of ~6%, which is comparable to the rate of C→A plus G→T transversions (in primates).

Continuing with our example, we have $u_{\min} = 12.9\%$ and $u = 33.84\% = u_{\min} + u_{\mathrm{d}}$ (where $u_{\mathrm{d}}$ is the deamination-dependent component of $u$) in a group of related pseudogenes with a GC content of 59% (Francino and Ochman 1999). This implies that $u_{\mathrm{d}59} = 33.84\% - 12.9\% = 20.9\%$. We can test this hypothesis, because the values of $u$ and $v$ were also measured in AT-rich globin pseudogenes (GC content = 43%; see Francino and Ochman 1999), in which cytosine deamination should occur 3.0-fold more rapidly (based on eqs. 2 and 4, $10^{-3.0(0.43)})/10^{-3.0(0.59)} = 3.0$). The values of $u_{43}$ and $v_{43}$ can then be predicted based on the following formulas:

$$u_{43} = [(u_{\min} + 3u_{\mathrm{d}59})/(u_{\min} + 3u_{\mathrm{d}59} + v_{59} + w_{59})] \times 100\% = 53\% \quad (7)$$

$$v_{43} = [v_{59}/(u_{\min} + 3u_{\mathrm{d}59} + v_{59} + w_{59})] \times 100\% = 32\%, \quad (8)$$

where $w_{59}$ is defined as the sum of all G/C→C/G plus all A/T→T/A substitutions in pseudogenes with a GC content of 59% (Francino and Ochman 1999). Our hypothesis predicts relative substitution rates of 53% and 32% (eqs. 7 and 8), which is in good agreement with the observed values of 51% and 34%, respectively (Francino and Ochman 1999). We use relative substitution rates in this calculation because the absolute substitution frequencies observed at high and low GC contents were not strictly comparable to each other (the pseudogenes were not sequenced in the same species; see Francino and Ochman 1999).

Our cytosine deamination hypothesis is also consistent with the fact that C→T and G→A transitions account for most of the variation in relative base substitution rates between isochores (Francino and Ochman 1999) and that C→T and G→A transitions occur at higher rates than other base substitutions in mammals (Li and Graur 1991; Krawczak and Cooper 1996). Moreover, our predictions of $u$ and $v$ as a function of base composition were derived from equations (2)–(5), which were based on sequence data and the biochemical properties of cytosine. These equations were sufficient to reproduce the relation between base composition and dinucleotide frequencies in human isochores (figs. 1 and 7). They indicate that most of the difference in GC content between human isochores is attributable to the deamination of unmethylated cytosine (fig. 7*E*).

## Discussion

### Cytosine Deamination and GC Content Form a Positive Feedback Loop

Our results immediately suggest solutions to three of the puzzles posed by the mosaic genome of birds and mammals (Bernardi et al. 1985). The first puzzle is why closely related genes on different chromosomes should often have dramatically different GC contents (Li and Graur 1991). The answer is that cytosine deamination and GC content form a positive feedback loop, such that an increase (or decrease) in GC content causes the mutation pressure to shift to a proportionately higher (or lower) GC bias (see eqs. 2–8).

All of the elements of this positive feedback loop are well established (C→T transitions affect the GC content, GC content affects DNA melting, DNA melting is rate-limiting for cytosine deamination, and cytosine deamination causes C→T transitions). We were simply the first to recognize how these elements fit together into a positive feedback loop and to analyze its overall magnitude during mammalian evolution. We did so in three different ways: (1) from computer simulation and quantitative analysis of long human DNA sequences, (2) from basic biochemical considerations, and (3) from pseudogene base substitution rates. These three lines of analysis are in good agreement with each other and indicate that the positive feedback between cytosine deamination and GC content is substantial enough to account for the evolutionary maintenance of mammalian isochores.

This positive feedback loop implies an evolutionary pattern of divergent genetic drift to high or low GC contents. But after these high or low GC contents had evolved, they would tend to be conserved in daughter species (as they have been in the α- and β-globin gene clusters of mammals) because their GC content would be maintained by a strong mutational bias. Evolutionary stability of GC content would be further reinforced by interactions along the length of the chromosome (see below). Nevertheless, distantly related phyla that did not share this history would be free to adopt dramatically different GC contents, even in orthologous genes (as they did in the α- and β-globin gene clusters of birds; Bernardi et al. 1985). In other words, the observed evolutionary metastability of GC content of orthologous genes is consistent with a positive feedback loop between cytosine deamination and GC content.

### Rates of Cytosine Deamination, as well as GC Content, Are Likely to Spread Along the Chromosome

The second puzzle is why a particular bias in GC content should be maintained over long stretches of chromosomal DNA, including all sequence elements along the way (introns, exons, flanking sequences, intergenic regions, and so on; see Bernardi 1995). It is known that the DNA double helix undergoes progressive and reversible strand separation (‘DNA breathing’), starting within AT-rich regions, spreading along the chromosome for distances that depend on local base composition, and resulting in temporary single-stranded ‘bubbles’ in reproducible locations (Inman 1966; Wetmur and Davidson 1968). In *Escherichia coli*, DNA breathing has been proven to propagate for considerable distances under physiological conditions (Skarstad, Baker, and Kornberg 1990). In eukaryotes, nuclease-hypersensitive sites often exhibit sensitivity to single-strand–specific nucleases that are specifically caused by DNA breathing (Umek and Kowalski 1990; Agustin et al. 1997), which can further lead to cruciform and triple-helical conformations in some cases (Soyfer and Potaman 1996; Agustin et al. 1997). From these and other results, it seems clear that nucleosomes inhibit but do not prevent DNA breathing, and nucleosome phasing is random (i.e., variable) throughout most of the mammalian genome (Nelson, Albright, and Garrard 1979; Widlak, Gaynor, and Garrard 1997), such that all mammalian DNA sequences are likely to breathe on an evolutionary timescale.

Because DNA breathing is based on progressive and reversible strand separation, any sequence placed adjacent to a GC-rich domain will undergo less breathing simply because of its location, and should therefore undergo a reduced rate of cytosine deamination, causing it to become more GC-rich. The converse would hold for sequences adjacent to an AT-rich domain. In other words, the observed correlation between 5-methylcytosine deamination and base composition (fig. 1) implies a specific biochemical mechanism (figs. 6 and 7), and this mechanism would cause any bias in base composition to gradually spread along the chromosome, eventually resulting in large domains with relatively uniform base compositions, which are the rule in mammalian genomes (Bernardi et al. 1985; Beck et al. 1999; Dunham et al. 1999).

Spreading of GC content along the chromosome would be expected to continue for considerable periods of time. In our simulations, equilibration of GC content required 200 UEP, which would correspond to roughly 500 Myr for a typical mammalian pseudogene evolving at $4 \times 10^{-9}$ substitutions per base pair per year (Li and Graur 1991). In mammalian genomes, chromosomal translocations and inversions have been fixed at intervals of 5–10 Myr or less (O'Brien et al. 1999). We would therefore expect that most of these rearrangements joined different isochores recently enough that a relatively sharp isochore boundary would still remain.

Another type of isochore boundary may exist in the human MHC, where a relatively sharp isochore boundary is associated with a boundary of DNA replication timing and with long polypurine/polypyrimidine tracts (Tenzen et al. 1997). Long polypurine/polypyrimidine tracts tend to form triple-helical structures that can pause or stop DNA polymerases (Soyfer and Potaman 1996). DNA triple helices are also likely to be associated with discontinuities in DNA breathing, because the structure of a triple helix stabilizes (holds closed) the nearby double-helical region on one side but destabilizes the adjacent double helix on the other side (i.e., forces the two strands apart; see Soyfer and Potaman 1996). Triple-helical structures may help to explain the connection between isochore boundaries and DNA replication timing (Tenzen et al. 1997; Bernardi 2000).

### Positive Feedback Between Cytosine Deamination and GC Content Is Effectively Limited to Warm-Blooded Vertebrates

The third puzzle is why all of this should happen in warm-blooded vertebrates but not in cold-blooded vertebrates (Bernardi et al. 1985). The answer is that the rate of cytosine deamination is strongly temperature-dependent. Given a typical body temperature of 20°C in fish and amphibians versus 37°C in mammals, cytosine deamination should occur 20.6-fold more slowly in fish and amphibians (based on eq. 3, $k_{37°C}/k_{20°C} = (7.0 \times 10^{-13}/s)/(0.34 \times 10^{-13}/s) = 20.6$). This indicates that positive feedback between cytosine deamination and GC content is insignificant in fish and amphibians, which is consistent with the lack of distinct classes of isochores in fish and amphibians (Bernardi et al. 1985). Reptiles are intermediate between cold-blooded vertebrates (i.e., fish and amphibians) and homeothermic vertebrates (i.e., birds and mammals) in terms of body temperature (Seebacher, Grigg, and Beard 1999), remaining levels of 5-methylcytosine (Jabbari et al. 1997), presence of GC-rich isochore structures (Hughes, Zelus, and Mouchiroud 1999), and presence of cytological chromosome bands (Schmid and Guttenbach 1988). Thus, the evolution of GC-rich isochores may have begun when early vertebrates adopted a terrestrial lifestyle.

Increased body temperature must have increased cytosine deamination (which would increase the genetic load [Krawczak and Cooper 1996]), in response to which natural selection presumably favored more efficient repair of G:U and G:T mismatched base pairs (Wiebauer et al. 1993). In fact, studies of DNA mismatch repair have shown that G:T mismatched base pairs are repaired with far higher efficiency, and far higher GC bias, than any other mismatched DNA base pair in cultured mammalian cells (Brown and Jiricny 1988). The G:T mismatch was repaired to a G:C base pair 24-fold more often than to an A:T pair (Brown and Jiricny 1988). If the majority of G:T mismatches in mammals are produced by deamination of 5-methylcytosine, then biased G:T repair would be adaptive (more precisely, unbiased repair would be mutagenic).

In contrast, G:T mismatch repair in cold-blooded vertebrates is unbiased. G:T mismatches in *Xenopus* are equally likely to be repaired to a G:C pair or an A:T pair and are repaired with somewhat below average efficiency (Varlet, Radman, and Brooks 1990). The lack of biased repair in *Xenopus* is consistent with our estimate that the rate of 5-methylcytosine deamination is ~20.6-fold lower in cold-blooded vertebrates than in mammals (see above). Moreover, CpG/GpC ratios in birds and mammals average 0.26, as compared with 0.36 in fish and amphibians (Jabbari et al. 1997). This corresponds to a 1.7-fold difference in *5mCt* (calculated by linear interpolation between the constant *5mCt* equations in fig. 6*B*) and indicates that G:T mismatch repair is about $20.6/1.7 \approx 12$-fold more efficient in warm-blooded vertebrates. This estimate is in good agreement with the previously cited studies, which demonstrated a 16-fold difference in G:T repair bias between mammals and

*Xenopus* (when unrepaired G:T mismatches were taken into account).

The efficient and strongly biased repair of G:T mismatches in mammals must reduce the rate of C→T transitions caused by misincorporation of thymidine, as well as G→A transitions on the complementary strand, both of which would reduce the value of $u_{min}$ (eqs. 7 and 8). Biased (incorrect) repair of G:T mismatches not caused by cytosine deamination would also tend to increase the background rate of A→G and T→C transitions, which would increase the value of $v$ in mammals (eqs. 7 and 8). All four of these effects will cause spontaneous mutations to become GC-biased in mammals if the rate of cytosine deamination is reduced. In contrast, the unbiased repair of G:T mismatches in *Xenopus* will prevent spontaneous mutations from having a GC bias of >50%, regardless of the rate of cytosine deamination. In other words, natural selection for the biased repair of G:T mismatches is likely to have been an essential prerequisite for the evolution of GC-rich isochores. We note that the evolution of homeothermy was accompanied by a pronounced increase in the GC content of ~⅓ of the genome, as well as a slight decrease in the GC content of the remaining ~⅔ of the genome (Bernardi et al. 1985; Cross et al. 1991; Ellsworth, Hewett-Emmett, and Li 1994), so that the total genomic GC content remained approximately the same (Jabbari et al. 1997).

## CpG Islands and CpG Mutability

CpG islands are relatively short (~500 bp) GC-rich sequences that are often associated with constitutively expressed promoters and are enzymatically demethylated during a particular stage of embryonic development (Gardiner-Garden and Frommer 1987; Aïssani and Bernardi 1991*a*, 1991*b*; Cross et al. 1991; Cedar and Verdine 1999). The demethylation of constitutive promoters is important for their function, also occurs in cold-blooded vertebrates, and is presumably maintained by natural selection (Cross et al. 1991; Cedar and Verdine 1999). We estimate that the average CpG island experiences an approximately twofold reduction in *5mCt* as a result of net hypomethylation over the entire life cycle in the germ line and embryos of both sexes. This estimate was derived as follows: equation (4) corresponds to the ideal case of uniform DNA methylation, while equation (5) also fits the median values of CpG islands (see below). Predicted CpG/GpC ratios derived from equation (4) (not shown) and equation (5) (fig. 7*A*) agree precisely with each other over the range of GC contents from 37% to 61% (which includes all five isochore classes in human DNA [Bernardi et al. 1985; Bernardi 1993*b*]), but they diverge at a GC content of 70% (which corresponds to the average CpG island; see the legend to fig. 1*A*). Thus, the difference between these equations is attributable to the influence of hypomethylation on CpG islands. Given the correspondence between the CpG/GpC ratio and *5mCt* (fig. 6*A* and *B*), the difference in CpG/GpC ratios predicted by these equations can be restated in terms of *5mCt*, and in those terms it corresponds to a twofold difference in *5mCt* at a GC content of 70%.

The reason equation (5) is able to provide an approximate fit to the average values of CpG islands is that the GC bias of *OB* equilibrates in proportion to any change in the rate of *5mCt*, including changes in *5mCt* caused by DNA hypomethylation. In the constitutive promoters of early mammals, an approximately twofold reduction in *5mCt* would have increased the GC content of these promoters by ~5% (i.e., half of the average contribution of *5mCt* in fig. 7*E*), which, in turn, would cause further reductions in cytosine deamination, increases in GC content, and so on, ultimately resulting in the dramatically GC-rich CpG islands of modern mammals (Aïssani and Bernardi 1991*a*, 1991*b*; Antequera and Bird 1993).

It is clear that CpG hypermutability causes a substantial fraction of the genetic load in mammals (Krawczak and Cooper 1996). This is equivalent to saying that CpG dinucleotides in coding sequences are maintained by natural selection, which on an evolutionary timescale would effectively reduce *5mCt* within exons and hence increase their GC content (fig. 7*E*). Human exons in all isochores average about 6% higher GC content than their associated introns (Eyre-Walker 1999). This could be accomplished by an approximately twofold reduction in *5mCt* (i.e., half of the average contribution of *5mCt* in fig. 7*E*), which is consistent with the higher CpG/GpC ratios observed in exons than in introns (Bernardi 1995). Since exons are more GC-rich than the surrounding DNA, the tendency of base compositions to spread along the chromosome would make GC-rich isochores more likely to form in regions that happened to have high gene density, particularly if these genes also contained CpG islands and/or amino acid compositions high in GC-rich codons. All of these characteristics are influenced by natural selection and correlated with GC-rich isochores in mammals (Bernardi 1995; D'Onofrio et al. 1999). Thus, mutational pressures and natural selection were both intimately interconnected with the evolution of isochore structures in the mammalian genome.

## Acknowledgments

We thank T. Gojobori and H. Watanabe for assistance in obtaining human genomic sequences. We also thank E. Mayr, G. Bernardi, W.-H. Li, C. W. Schmid, T. V. Jordan, F. E. Hall, O. Clay, and P. A. Fryxell for helpful comments. Preliminary summaries of some of these results were presented at the 1997 and 1999 meetings of the International Society of Molecular Evolution in Costa Rica. This work was supported by the sponsors of the ISME meetings and by grants to K.J.F. and E.Z. from the National Science Foundation.

## LITERATURE CITED

ADAMS, R. L. P., T. DAVIS, A. RINALDI, and R. EASON. 1987. CpG deficiency: dinucleotide distributions and nucleosome positioning. Eur. J. Biochem. **165**:107–116.

AGUSTIN, A., J. E. PEREZ-ORTIN, C. J. BENHAM, and M. DEL OLMO. 1997. Analysis of the structure of a natural alternating d(TA)-n sequence in yeast. Yeast **13**:313–326.

AÏSSANI, B., and G. BERNARDI. 1991*a*. CpG islands, genes and isochores in the genome of vertebrates. Gene **106**:185–195.

———. 1991*b*. CpG islands: features and distribution in the genome of vertebrates. Gene **106**:173–183.

ANTEQUERA, F., and A. BIRD. 1993. CpG islands. Pp. 169–185 *in* J. P. JOST and H. P. SALUZ, eds. DNA methylation: molecular biology and biological significance. Birkhäuser Verlag, Basel, Switzerland.

AVEROF, M., A. ROKAS, K. H. WOLFE, and P. M. SHARP. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science **287**:1283–1286.

BECK, S., D. GERAGHTY, H. INOKO et al. (29 co-authors). 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature **401**:921–923.

BERNARDI, G. 1989. The isochore organization of the vertebrate genome. Annu. Rev. Genet. **23**:637–661.

———. 1993*a*. The isochore organization of the human genome and its evolutionary history—a review. Gene **135**:57–66.

———. 1993*b*. The vertebrate genome: isochores and evolution. Mol. Biol. Evol. **10**:186–204.

———. 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. **29**:445–476.

———. 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241**:3–17.

BERNARDI, G., D. MOUCHIROUD, C. GAUTIER, and G. BERNARDI. 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. J. Mol. Evol. **28**:7–18.

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.

BETTECKEN, T., B. AÏSSANI, C. R. MÜLLER, and G. BERNARDI. 1992. Compositional mapping of the human dystrophin gene. Gene **122**:329–335.

BIRD, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

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.

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. 1987. A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol. Biol. Evol. **4**:395–405.

CEDAR, H., and G. L. VERDINE. 1999. The amazing demethylase. Nature **397**:568–569.

COOPER, D. N., and M. KRAWCZAK. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181–188.

———. 1993. Human gene mutation. BIOS Scientific Publishers, Oxford, England.

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.

CROSS, S., P. KOVARIK, J. SCHMIDTKE, and A. BIRD. 1991. Non-methylated islands in fish genomes are GC-poor. Nucleic Acids Res. **19**:1469–1474.

CURTIS, D., S. H. CLARK, A. CHOVNICK, and W. BENDER. 1989. Molecular analysis of recombination events in *Drosophila.* Genetics **122**:653–662.

D'ONOFRIO, G., K. JABBARI, H. MUSTO, F. ALVAREZ-VALIN, S. CRUVEILLER, and G. BERNARDI. 1999. Evolutionary genomics of vertebrates and its implications. Ann. N.Y. Acad. Sci. **870**:81–94.

DUNHAM, I., N. SHIMIZU, B. A. ROE, and S. CHISSOE. 1999. The DNA sequence of human chromosome 22. Nature **402**: 489–495.

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.

ELLSWORTH, D. L., D. HEWETT-EMMETT, and W.-H. LI. 1994. Evolution of base composition in the insulin and insulin-like growth factor genes. Mol. Biol. Evol. **11**:875–885.

EYRE-WALKER, A. 1994. DNA mismatch repair and synonymous codon evolution in mammals. Mol. Biol. Evol. **11**: 88–98.

———. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152**:675–683.

FILIPSKI, J. 1987. Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **217**:184–186.

FRANCINO, M. P., and H. OCHMAN. 1999. Isochores result from mutation not selection. Nature **400**:30–31.

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.

GARDINER-GARDEN, M., and M. FROMMER. 1987. CpG islands in vertebrate genomes. J. Mol. Biol. **196**:261–282.

GOLDMAN, M. A., G. P. HOLMQUIST, M. C. GRAY, L. A. CASTON, and A. NAG. 1984. Replication timing of mammalian genes and middle repetitive sequences. Science **224**:686–692.

GREEN, P. M., A. J. MONTANDON, D. R. BENTLEY, R. LJUNG, I. M. NILSSON, and F. GIANNELLI. 1990. The incidence and distribution of CpG→TpG transitions in the coagulation factor IX gene. A fresh look at CpG mutational hotspots. Nucleic Acids Res. **18**:3227–3231.

GU, X., and W. H. LI. 1994. A model for the correlation of mutation rate with GC content and the origin of GC-rich isochores. J. Mol. Evol. **38**:468–475.

HOLMQUIST, G. P. 1989. Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. **28**:469–486.

———. 1992. Chromosome bands, their chromatin flavors, and their functional features. Am. J. Hum. Genet. **51**:17–37.

HUGHES, S., D. ZELUS, and D. MOUCHIROUD. 1999. Warm-blooded isochore structure in the Nile crocodile and turtle. Mol. Biol. Evol. **16**:1521–1527.

INMAN, R. B. 1966. A denaturation map of the lambda phage DNA molecule determined by electron microscopy. J. Mol. Biol. **18**:464–476.

JABBARI, K., and G. BERNARDI. 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene **224**:123–128.

JABBARI, K., S. CACCIO, J. P. PAIS DE BARROS, J. DESGRES, and G. BERNARDI. 1997. Evolutionary changes in CpG and methylation levels in the genome of vertebrates. Gene **205**: 109–118.

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.

Karlin, S., and J. Mrázek. 1996. What drives codon choices in human genes? J. Mol. Biol. **262**:459–472.

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, England.

Leeds, J. M., M. B. Slabourgh, and C. K. Mathews. 1985. DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and cytoplasm of mammalian cells. Mol. Cell. Biol. **5**:3443–3450.

Li, W.-H., and D. Graur. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

Lindahl, T., and B. Nyberg. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry **13**:3405–3410.

Lukacsovich, T., and A. S. Waldman. 1999. Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence. Genetics **151**:1559–1568.

Nelson, P. P., S. C. Albright, and W. T. Garrard. 1979. Nucleosome arrangement with regard to DNA base composition. J. Biol. Chem. **254**:9194–9199.

O'Brien, S. J., M. Menotti-Raymond, W. J. Murphy, W. G. Nash, J. Wienberg, R. Stanyon, N. G. Copeland, N. A. Jenkins, J. E. Womack, and J. A. Marshall-Graves. 1999. The promise of comparative genomics in mammals. Science **286**:458–481.

Razin, A., and H. Cedar. 1993. DNA methylation and embryogenesis. Pp. 343–357 *in* J. P. Jost and H. P. Saluz, eds. DNA methylation: molecular biology and biological significance. Birkhäuser Verlag, Basel, Switzerland.

Razin, A., and A. D. Riggs. 1980. DNA methylation and gene function. Science **210**:604–610.

Rubin, C. M., C. A. VandeVoort, R. L. Teplitz, and C. W. Schmid. 1994. *Alu* repeated DNAs are differentially methylated in primate germ cells. Nucleic Acids Res. **22**:5121–5127.

Saccone, S., A. De Sario, G. Della Valle, and G. Bernardi. 1992. The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA **89**:4913–4917.

Sasaki, H., N. D. Allen, and M. A. Surani. 1993. DNA methylation and genomic imprinting in mammals. Pp. 469–486 *in* J. P. Jost and H. P. Saluz, eds. DNA methylation: molecular biology and biological significance. Birkhäuser Verlag, Basel, Switzerland.

Schmid, M., and M. Guttenbach. 1988. Evolutionary diversity of reverse (R) fluorescent chromosome bands in vertebrates. Chromosoma **97**:101–114.

Seebacher, F., G. C. Grigg, and L. A. Beard. 1999. Crocodiles as dinosaurs: behavioural thermoregulation in very large ectotherms leads to high and stable body temperatures. J. Exp. Biol. **202**:77–86.

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.

Soyfer, V. N., and V. N. Potaman. 1996. Triple-helical nucleic acids. Springer-Verlag, New York.

Spruck, C. H. III, W. M. Rideout III, and P. A. Jones. 1993. DNA methylation and cancer. Pp. 487–509 *in* J. P. Jost and H. P. Saluz, eds. DNA methylation: molecular biology and biological significance. Birkhäuser Verlag, Basel, Switzerland.

Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85**:2653–2657.

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.

Tenzen, T., T. Yamagata, T. Fukagawa, K. Sugaya, A. Ando, H. Inoko, T. Gojobori, A. Fujiyama, K. Okumura, and T. Ikemura. 1997. Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. Mol. Cell. Biol. **17**: 4043–4050.

Umek, R. M., and D. Kowalski. 1990. Thermal energy suppresses mutational defects in DNA unwinding at a yeast replication origin. Proc. Natl. Acad. Sci. USA **87**:2486–2490.

Varlet, I., M. Radman, and P. Brooks. 1990. DNA mismatch repair in *Xenopus* egg extracts: repair efficiency and DNA repair synthesis for all single base-pair mismatches. Proc. Natl. Acad. Sci. USA **87**:7883–7887.

Wahl, G. M., S. L. Berger, and A. R. Kimmel. 1987. Molecular hybridization of immobilized nucleic acids: theoretical concepts and practical considerations. Methods Enzymol. **152**:399–407.

Wetmur, J. G., and N. Davidson. 1968. Kinetics of renaturation of DNA. J. Mol. Biol. **31**:349–370.

Widlak, P., R. B. Gaynor, and W. T. Garrard. 1997. *In vitro* chromatin assembly of the HIV-1 promoter: ATP-dependent polar repositioning of nucleosomes by Sp1 and NF-kappa-B. J. Biol. Chem. **272**:17654–17661.

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.

Howard Ochman, reviewing editor

Accepted May 22, 2000