Modeling the Impact of DNA Methylation on the Evolution of *BRCA1* in Mammals

Gavin A. Huttley

Centre for Bioinformation Science, John Curtin School of Medical Research and Mathematical Sciences Institute, Australian National University, Australia

The modified base 5-methylcytosine (^{m}C) plays an important functional role in the biology of mammals as an epigenetic modification and appears to exert a striking impact on the molecular evolution of mammal genomes. The collective epigenetic functions of ^{m}C revolve around its effect on gene transcription, while the influence of this modified base on the evolution of mammal genomes derives from the greatly elevated spontaneous mutation rate of ^{m}C to T. In mammals, ^{m}C occurs at the dinucleotides CpG, CpA, and CpT. As a step toward a comprehensive statistical examination of the role of ^{m}C in mammal molecular evolution, we have developed novel Markov models of codon substitution that incorporate dinucleotide-level terms relevant to ^{m}C mutation. We apply these models to two data sets of aligned *BRCA1* exon 11 sequences from bats and primates. In all cases, terms specific to mutations that affect the dinucleotides CpG, CpA, and CpT significantly improved model fit. For the CpG-specific terms, both transition and transversion substitution rates were elevated. These rates differed between the data sets. Bats exhibited a lower relative rate of substitutions at CpG-containing codons. Transition substitutions were significantly less than 1 at CpA-containing codons but greater than 1 at CpT-containing codons. The inclusion of interaction terms in the codon models to represent possible confounding with the effect of natural selection were supported for codons that contained CpG and CpT, but not CpA. From the results, we infer that mutation of ^{m}C is a probable factor that affects *BRCA1* codons containing the dinucleotide CpG, a possible factor for CpA-containing codons, and an unlikely factor that affects CpT-containing codons. The confounding of estimated terms with the effect of natural selection indicate this confounding must be addressed for comparisons between different coding and noncoding regions.

Introduction

The modified nucleotide 5-methylcytosine (^{m}C) occupies a unique place in mammal biology. On the one hand, this base fulfills complex functional roles that are subjected to scrutiny by natural selection, whereas on the other hand, it exhibits a striking propensity for mutation. The potential impact of changes to this mutation-selection balance on the genomic distribution of substitution rates, although commonly appreciated, is yet to be addressed in a statistically comprehensive manner. As a step toward this objective, we present novel Markov process models of codon substitution. These models can be applied within a maximum-likelihood framework to examine the possible impact of ^{m}C on the molecular evolution of protein-coding sequences. Epigenetic modifications, such as ^{m}C, dictate which genes are active at each stage of organismal development and, thus, constitute an additional type of biological information that is not directly recorded in raw sequence data. The significance of ^{m}C is such that altered levels (overabundance or dearth) of this nucleotide in the genomic sequence has a profound impact on mammal development that can manifest as disease (e.g., Reik and Walter 2001*a*). Such phenotypic effects arise from disruption of any one of the many roles played by ^{m}C—genetic imprinting, chromatin modeling, or X-chromosome inactivation.

How ^{m}C exerts its biological effect on expression is complex, but the process revolves around an effect on the level of gene transcription. This is achieved through changes in the methylation status of regulatory sequences that lie in promoter regions and CpG islands. CpG islands, for instance, tend to be hypomethylated, which allows the transcription of flanking genes. Yet, not all alleles with an unmethylated promoter will be expressed. Methylation of an antisense promoter for a maternal allele can suppress translation of an unmethylated paternal allele (Reik and Walter 2001*b*). The biological functions of ^{m}C, and the operation of natural selection, have acted to shape a complex spatial and temporal pattern of methylation in mammal genomes. The scale of this heterogeneity extends to differences between the sexes in which the male germline exhibiting a higher incidence of ^{m}C than the female germline (Monk, Boubelik, and Lehnert 1987). In addition to the regulatory regions of genes, ^{m}C is also known to occur within gene exons (e.g., Rideout et al. 1990).

Methylation of cytosine occurs at characteristic dinucleotides by a mechanism that remains unclear. ^{m}C occurs predominantly at the dinucleotide CpG. This dinucleotide is strand symmetric (the same on both DNA strands), and it is typically, but not always, the case that both the plus and minus DNA-strand Cs are methylated (Tomatsu et al. 2002). ^{m}C has been reported to also occur at the non-CpG dinucleotides CpA and CpT in murine cell lines, oocytes, sperm, and embryos (Ramsahoye et al. 2000; Haines, Rodenhiser, and Ainsworth 2001).

Another mechanism by which ^{m}C influences phenotype arises because of its mutation propensity, first identified from studies of mutation hot spots in *Escherichia coli* (Coulondre et al. 1978). This work revealed that ^{m}C spontaneously deaminates to T at a high rate relative to other mutation types. The inference that high mutability was also a property of ^{m}C in mammals derives in part from the demonstration by Cooper and Youssoufian (1988) that human disease–causing nucleotide polymorphisms were frequently located within CpGs. Since that study, numerous others have strengthened the association between coding

Key words: BRCA1, codon substitution model, molecular evolution, DNA methylation, CpG, maximum-likelihood.

E-mail: gavin.huttley@anu.edu.au.

Mol. Biol. Evol. 21(9):1760–1768. 2004
doi:10.1093/molbev/msh187
Advance Access publication June 9, 2004

region methylation (e.g., Rideout et al. 1990; Tomatsu et al. 2002) and mutation hot spots responsible for human disease (e.g., Rousseau et al. 1994; Girard et al. 2001; Trappe et al. 2001). Given the chemical process underlying ^{m}C mutation, a natural expectation is that ^{m}C-bearing dinucleotides should exhibit just an elevated rate of transitions. Both transitions and transversions at CpG dinucleotides are accelerated in the factor IX gene (Sommer, Scaringe, and Hill 2001), however, which suggests that the repair of the T/G mismatch is error prone.

These joint functional and mutation properties of ^{m}C imply an important role for this modification in shaping the evolution of protein-coding sequences. Although numerous statistical approaches can be used to evaluate this hypothesis, we will consider three here. The influence of mutation-affecting CpGs on coding-sequence evolution can be evaluated by log-linear modeling, in which dinucleotide incidence is considered in terms of nucleotide incidence (Huttley et al. 2000*b*). This approach does not lend itself to examination of more than two species, as the shared ancestry in the sample violates the assumption of independence. Another approach uses correlation between the rate of substitution and measures of sequence composition. Tsunoyama, Bellgard, and Gojobori (2001) adopted this approach to examine the influence of mutation of CpG dinucleotides on rodent sequence evolution. Tests of the correlation coefficient between the estimated proportion of synonymous substitutions and the average frequency of CpG dinucleotides from nonoverlapping windows in paired homologous sequences was significant for many of the genes studied. This correlation approach also does not readily lend itself to examination of trees with more than two taxa. The third approach of explicitly representing CpG mutation in the substitution model has no such limitation if codon substitutions are represented as a Markov process. This representation allows exploitation of the phylogeny-based maximum-likelihood framework for statistical inference. Such a model is that of Pedersen, Wiuf, and Christiansen (1998), who extended the Muse and Gaut (1994) codon model to consider CpG mutation. The affect of CpGs on codon substitution was represented by adding to the Muse and Gaut parameterisation a single parameter (λ). Mutations that generated a CpG were assigned $1/\lambda$, whereas those that destroyed a CpG were assigned λ (Pedersen, Wiuf, and Christiansen 1998). Under this parameterization, a value of λ greater than 1 indicates a higher rate of mutation of CpGs than background. This model was found to significantly improve the description of HIV sequence evolution. Pedersen and Jensen (2001) extended this model to account for dependence between adjacent codons and established a codon-position effect on CpG substitutions, which likely arises from the differential influence of natural selection on the three codon positions. This latter model is presently restricted to two sequences.

In consideration of the recent rediscovery of methylation at non-CpG dinucleotides and the putative elevation of both transition and transversion mutations at CpGs, the previous codon model parameterizations (Pedersen, Wiuf, and Christiansen 1998; Pedersen and Jensen 2001) incompletely capture the potential effects of methylation on codon substitutions. Estimates of CpG mutation rates from these models may also be confounded by the influence of natural selection, precluding direct comparison between genes experiencing different selective constraints. Specifically, the subset of codons that contain the dinucleotide CpG may experience a different selective regime to the non–CpG containing codons. As this regime can differ between different proteins, a comparison among genes potentially becomes problematic. That natural selection influences the rate of substitution at CpGs is demonstrated by the codon-position effect.

Here, we assess the influence of substitutions pertinent to ^{m}C mutation at the dinucleotides CpG, CpA, and CpT on coding-sequence evolution. We evaluate the influence of methylation on codon substitutions indirectly by devising several novel models of codon substitution that includes terms for specific dinucleotides and terms for assessing the influence of natural selection on their estimation. We apply these models to DNA sequences of the tumor suppressor BRCA1 to assess whether terms for mutation that affects the candidate ^{m}C-containing dinucleotides significantly improve model fit, whether both transitions and transversions are elevated at CpG dinucleotide, and whether these parameter estimates are confounded by the influence of natural selection.

Methods

Statistics

The model terms and their definitions, which we will describe now, are also summarized in Table 1. Our baseline parameterization is a modified Goldman and Yang (1994) codon substitution model (Yang 1998), which uses the parameterization of Hasegawa, Kishino, and Yano (1985) to represent the underlying nucleotide substitution process. Following the notation of Yang (1998), the instantaneous rate matrix is termed Q, and individual entries in this matrix, termed q_{ij}, correspond to the relative rate of change from codon i to codon j. The q_{ij} are defined as

$$q_{ij,i\neq j} = \begin{cases} 0, & \text{more than one nucleotide difference} \\ \pi_j, & \text{synonymous transversion} \\ \pi_j K, & \text{synonymous transition} \\ \pi_j R, & \text{nonsynonymous transversion} \\ \pi_j KR, & \text{nonsynonymous transition} \end{cases}$$

In this and all other models, we follow the convention of estimating π_j as the averaged frequency of codon j in the alignment. This parameterization will be hereafter referred to as Y98.

Before defining the new model terms, we point out in advance that all of them confound forward and backward substitution rates. This property stems from constraints imposed to achieve reversibility. To be reversible, a Markov process must realize the condition $\pi_i q_{ij} = \pi_j q_{ji}$. Because q_{ij} includes π_j, the π's cancel, and any other terms in q_{ij} must also be in q_{ji}.

To evaluate the affect of mutation at ^{m}C, we seek to accommodate the different incidence of methylation at the

Table 1
Substitution Model Terms

Term	Definition
K	Transition substitution rate
R	Nonsynonymous substitution rate
G	CpG substitution rate
A	CpA substitution rate
T	CpT substitution rate
G.K	CpG transition substitutions
A.K	CpA transition substitutions
T.K	CpT transition substitutions
A.K.R	CpA transition nonsynonymous substitutions
G.R	CpG nonsynonymous substitutions
T.K.R	CpT transition nonsynonymous substitutions

three dinucleotides (Haines, Rodenhiser, and Ainsworth 2001) and an elevated mutation rate for both transitions and transversions (Sommer, Scaringe, and Hill 2001). We deal with the different incidence by parameterizing the properties of substitution at each of the dinucleotides CpG, CpA, and CpT separately.

The elevation of both transition and transversion mutations at CpGs suggests that repair of the T/G mismatch may involve a complete replacement. In this case, the impact of ^{m}CpG mutations should be adequately accounted for by a single term, ***G***, added to the Y98 model when two codons have a single difference that lies within a CpG dinucleotide. Given that CpG is strand symmetric (the same on both DNA strands), changes to either the C or G are considered. As each codon consists of two overlapping dinucleotides, ***G*** is applied to instantaneous changes of the form CGN↔DGN, CGN↔CHN, NCG↔NDG, and NCG↔NCH, where D and H are IUPAC ambiguity symbols corresponding to A/G/T and A/C/T respectively. For reversibility of the Markov process, ***G*** is assigned to both mutation directions between the codons. These considerations result in the following transition matrix:

$$q_{ij,i\neq j} = \begin{cases} 0, & \text{more than one change} \\ \pi_j, & \text{synonymous transversion} \\ \pi_j G, & \text{synonymous transversion involving CpG} \\ \pi_j K, & \text{synonymous transition} \\ \pi_j KG, & \text{synonymous transition involving CpG} \\ \pi_j R, & \text{nonsynonymous transversion} \\ \pi_j RG, & \text{nonsynonymous transversion involving CpG} \\ \pi_j KR, & \text{nonsynonymous transition} \\ \pi_j KRG, & \text{nonsynonymous transition involving CpG} \end{cases}$$

If the repair process has a tendency to replace just one of the mismatched nucleotides, rather than both as implied above, we expect an elevated repair of T/G to T/A (and T/G to C/G, which results in no change). In this case, transitions in CpG-containing codons will be further elevated beyond that represented by the ***G*** and ***K*** terms. We assess this possibility by applying an interaction term (***G.K***) whenever two codons differ by a CpG (***G***) transition (***K***) substitution (CpG↔TpG or CpG ↔CpA). If deamination of ^{m}C is an important factor in mutation of codons that contain the CpG dinucleotide, then either or both the ***G*** and ***G.K*** terms should be significantly greater than 1, depending on the nature of repair. The transition matrices for this and all subsequent models described in this article are presented in the Supplementary Material online.

Natural selection may confound estimates of the dinucleotide substitution terms. The term ***G*** identifies a subset of codons and, thus, amino acids. If those amino acids are subjected to natural selection differently from the non-CpG–containing codons, estimation of ***G*** will be affected. Under Y98, the term ***R*** represents the mean affect of natural selection on nonsynonymous changes for all alignment positions and all codons. We, therefore, assess the possible confounding of ***G*** with ***R*** by including an interaction term ***G.R*** when two codons differ by an instantaneous nonsynonymous change at a CpG dinucleotide (for transition matrix see Supplementary Material online).

For the non-CpG dinucleotides CpA and CpT, only mutations that affect the cytosine at either the plus or minus strand dinucleotides are considered pertinent to mutation of ^{m}C. For the dinucleotide CpA, whose minus strand complement is TpG, we add the term ***A*** when two codons differed at a single base as CpA↔DpA or TpG↔TpH. The complement of CpT is ApG, and the term ***T*** was included when two codons differed at a single base as CpT ↔DpT or ApG↔ApH. Both transition and nonsynonymous interaction terms are employed for the CpA and CpT substitutions as described for CpG dinucleotides above (for transition matrices see Supplementary Material online).

The support for each parameterization was evaluated by calculating the log-likelihood of the parameterization given the data in the conventional way. Specifically, the log-likelihood of the parameterization was calculated as the conditional probability of observing the alignment given the parameterization using the likelihood formulation of Felsenstein (1981). In adopting this formulation we also adopt its assumptions concerning independence between sites, compositional stationarity, and, as indicated above, reversibility of the Markov process. The different parameterizations were then compared by use of likelihood ratio (LR) tests. The LR is calculated as $\mathrm{LR} = 2(\ln L_u - \ln L_c)$, where $\ln L_u$ and $\ln L_c$ are log-likelihoods for the unconstrained (more parameter rich) and constrained models, respectively. Nested parameterizations allow estimation of the probability of the observed or a larger LR statistic by use of the χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the unconstrained and constrained models. Although the accuracy of the χ^2 approximation is sensitive to the amount of data, robust estimates can be obtained from alignments as short as 125 sites (Ota et al. 2000). The χ^2 approximation is also conservative if terms are estimated as 0 (Ota et al. 2000), which may occur because of the nonnegativity constraint on parameters. This situation arose for only one branch in the primate data set (described below). Because adjusting the probabilities had no effect on the interpretation, we report only the unadjusted probabilities.

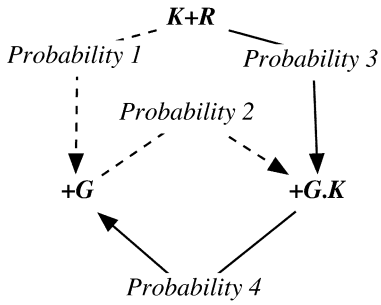

FIG. 1.—Flow diagram illustrating alternative orders of term fitting. The two possible orders of LR tests to evaluate the contribution of the terms ***G*** and ***G.K*** to the baseline ***K+R*** parameterization are represented by the solid and dashed lines. Arrows indicate the term-fitting direction. The ***+term*** notation indicates that ***term*** has been added to the parameterization at the arrows origin. Each *Probability* label corresponds to the results from the LR test that compares the two parameterizations connected by the arrow.

Modeling Notation and Approach

We introduce the notation that will be used throughout the remainder of this article to distinguish between the different model parameterizations. In our notation, the baseline Y98 model is represented as ***K+R***, and a model that includes the additional term ***G*** is ***K+R+G***. We point out that the + symbol represents the inclusion of the terms in the parameterization, not how the terms are treated in calculating the substitution probabilities. Additionally, note that permutations of terms represent equivalent parameterizations (e.g., ***K+R+G*** is the same as ***G+K+R***.)

The modeling approach adopted was to initially examine the contribution of dinucleotide-specific terms separately, and then consider joint models. Because the independent and interaction terms overlap, we considered alternative orders of term fitting. We illustrate this procedure for the CpG-specific substitution terms ***G*** and ***G.K*** in figure 1. To assess the contributions of the ***G*** and ***G.K*** terms, four parameterizations were used: Y98 (***K+R***), ***K+R+G***, ***K+R+G.K***, and ***K+R+G+G.K***. The two possible orders by which the ***G*** and ***G.K*** terms can be added to the ***K+R*** parameterization are indicated by the solid and dashed lines in figure 1. The ***+term*** notation in the figure indicates that ***term*** has been added to the parameterization at the arrows origin. Hence, the parameterizations represented by ***+G*** are ***K+R+G.K+G*** following solid lines and ***K+R+G*** following the dashed line. Following the dashed line of figure 1, the two LR tests performed compare ***K+R*** versus ***K+R+G*** and ***K+R+G*** versus ***K+R+G+G.K***, which results in *Probability 1* and *Probability 2*, respectively. Following the solid line in figure 1, the two LR tests compare ***K+R*** versus ***K+R+G.K*** and ***K+R+G.K*** versus ***K+R+G+G.K***, which results in *Probability 3* and *Probability 4*, respectively. If the term ***G***, for example, significantly improves model fit for both orders (*Probability 1* < 0.05 and *Probability 4* < 0.05), the effect of ***G*** is robust, and we conclude that it significantly improves the model fit.

Software

All models were implemented by the PyEvolve version 0.8 statistical molecular evolutionary modeling toolkit (Butterfield et al. 2004) that is available from http://cbis.anu.edu.au/software. Settings of the simulated annealing numerical optimizer were temperature reduction factor of 0.6, five iterations before a temperature reduction, and 20 cycles before step size was modified. The substitution models reported here will be made available as part of the standard PyEvolve distribution on acceptance of this article.

Data

We used DNA sequences of exon 11 from the tumor suppressor BRCA1 from primates (plus flying lemur and tree shrew) and bats. These analyses will be referred to as the primate and bat analyses, respectively. The sequences were aligned by application of ClustalW (Thompson, Higgins, and Gibson 1994) and adjusted manually. Gaps in the alignment were treated as ambiguities by recoding as N's. The lengths of the alignments were 2,880 nucleotides for the primate sequences and 2,837 nucleotides for the bats. The common name, species name, and accession numbers were as follows: (primates) chimpanzee, *Pan troglodytes*, AF207822; flying lemur, *Cynocephalus variegatus*, AF019081; galago, *Otolemur crassicaudatus*, AF019080; gorilla, *Gorilla gorilla*, AF019076; howler monkey, *Alouatta seniculus*; AF019079; human, *Homo sapiens*, NM_007306; orangutan, *Pongo pygmaeus*, AF019077; rhesus macaque, *Macaca mulatta*, AF019078; tree shrew, *Tupaia tana*, AF284006; (bats) flying fox, *Pteropus rayneri*, AF203751; tomb bat, *Taphozous sp*., AF203748; Daubenton's bat, *Myotis daubentoni*, AF203746; false vampire bat, *Megaderma lyra*, AF203749; leaf-nosed bat, *Hipposideros commersoni*, AF203752; free-tailed bat, *Tadarida brasiliensis*, AF203747; dog-faced bat, *Cynopterus sphinx*, AF203750; round-eared bat, *Tonatia bidens*, AF203745. The primate phylogenetic tree was the same as that used previously (Huttley et al. 2000*a*), with the following exceptions: the rodents and bush baby sequences are not included, and the tree shrew was included according to the phylogeny of Murphy et al. (2001). We note that the position of the flying lemur as an outgroup in that tree is also strongly supported by other analyses (Killian et al. 2001; Waddell, Kishino, and Ota 2001; Schmitz et al. 2002). The bat phylogenetic tree was that of Murphy et al. (2001). Both alignments and trees are available from the author.

Results

Primate Analysis

The LR tests support an important role of mutation at all the putatively methylated dinucleotides in the evolution of primate *BRCA1*. Codons that contain CpG dinucleotides exhibit the greatest substitution rate. The results from modeling CpG mutation are presented in figure 2*b*, and parameter and ln*L* estimates are presented in table 2. In the case of CpGs, both the ***G*** and the ***G.K*** terms significantly improved model fit, irrespective of the order in which they were included in the model. The maximum-likelihood estimates of ***G*** ($\sim$9.8) and ***G.K*** ($\sim$1.9) were both greater

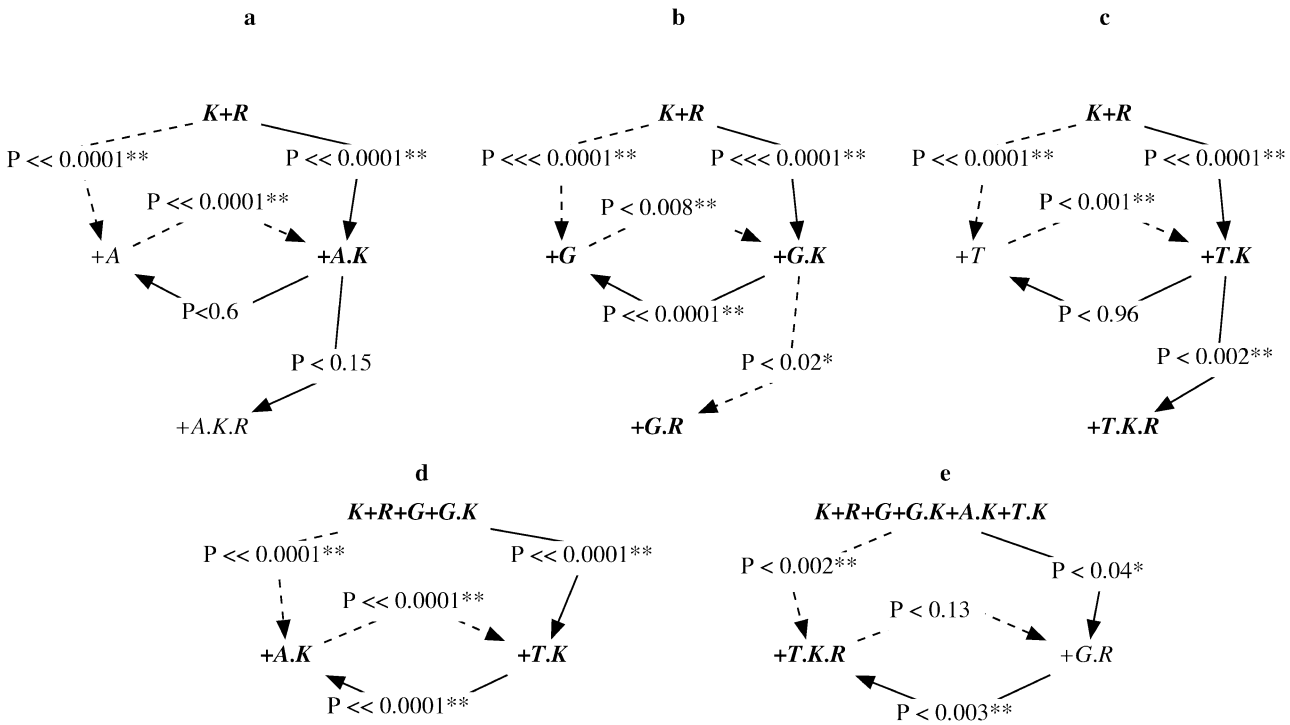

FIG. 2.—Flow diagram of nested model parameterizations and LR test results from analysis of the primate data set. Arrows represent the order of term fitting, and, thus, the direction of the LR test. Thick lines indicate the determined appropriate order of term fitting, whereas terms in bold were significant irrespective of order. (*a*) Testing for CpA/TpG terms. (*b*) Testing for CpG terms. (*c*) Testing for CpT/ApG terms. (*d*) Testing for substitution terms selected in *a*–*c*. (*e*) Testing the effect of selection on the substitution processes. *P* is probability of observing the LR given the df (determined by the statistics presented in table 2) from the χ^2 approximation. The symbols * and ** indicate significant at the 0.05 and 0.01 levels, respectively.

than 1 (table 2), consistent with the expectation that they derive from mutation of ^{m}C. To assess the nature of the interaction between CpG mutation and natural selection, we added the ***G.R*** term. This term also significantly improved model fit, and its estimated value of approximately 0.5 indicates that the mean nonsynonymous substitution rate for CpG-containing codons is significantly lower than the average codon substitution rate for *BRCA1*. Including the ***G.R*** term also markedly increased the estimated value of ***G*** (~15.8).

An elevated substitution rate that affected CpA dinucleotides was also apparent, consistent with an ^{m}C effect. The significance of terms that involve CpA mutations was, however, sensitive to order of inclusion in the models. The LR test that compared the ***K+R+A.K*** and ***K+R+A+A.K*** parameterizations was not significant (fig. 2*a*). The ***A.K*** term was significant in both directions, however, and its estimated value of approximately 2.4 indicated substitution rate acceleration of C and G nucleotides in codons that contained CpA/TpG dinucleotides. The improvement conferred to the Y98 model by ***A***, therefore, derives from the CpA transitions represented by ***A.K***. To assess the impact of natural selection on ***A.K*** estimation, we included the interaction term ***A.K.R*** to the ***K+R+A.K*** model. The ***A.K.R*** (~0.7) term did not significantly improve fit.

The rate of substitution affecting codons containing CpT/ApG was significantly less than 1, inconsistent with an ^{m}C effect. As for CpA, only the transition interaction term, ***T.K***, proved robust to the order of inclusion (fig. 2*c*). The estimated value of ***T.K*** (~0.6) indicated a lower rate of substitution that affected the C and G nucleotides in codons that contained CpT/ApG. This suppression did not appear to derive from the confounding influence of natural selection. Inclusion of the interaction term ***T.K.R*** (~1.7) further reduced the estimated value of ***T.K*** (~0.4). We infer from the magnitude of the ***T.K.R*** term, and its robustness to order of term fitting, illustrated in figure 1*e*, that the rate of nonsynonymous substitutions that affect codons that contain CpT/ApG dinucleotides is significantly greater than the average nonsynonymous substitution rate estimated for all codons.

Joint modeling supported five of the six terms identified above. In figure 2*d* and *e*, we illustrate the effect of model-fitting order on the significance of terms when they were combined into increasingly parameter-rich models. Regardless of order, both the ***A.K*** and the ***T.K*** terms significantly improved model fit, and their estimated values in the ***A.K+G+G.K+K+R+T.K*** parameterization of approximately 1.7 and approximately 0.7 were consistent with their previously estimated values. Although inclusion of the ***T.K.R*** term was robust to model fit and maintained its elevation above 1 (see table 2), the ***G.R*** term was not significant when added last. The latter probably reflects differences in statistical power of the generally low frequency of codons that contain this dinucleotide in *BRCA1* (average frequency 0.012) relative to those that contain CpT/ApG (average frequency 0.33). We note here that the frequency of codons that contain CpA/TpG is also high (average frequency of 0.22).

Table 2
Estimated Statistics from Primate *BRCA1*

Model	ln*L*	nfp	***A***	***A.K***	***A.K.R***	***G***	***G.K***	***G.R***	***K***	***R***	***T***	***T.K***	***T.K.R***
K+R	−9465.94	17	—	—	—	—	—	—	3.62	0.70	—	—	—
G.K+K+R	−9377.43	18	—	—	—	—	18.38	—	3.18	0.65	—	—	—
A+K+R	−9452.75	18	1.59	—	—	—	—	—	3.85	0.68	—	—	—
K+R+T.K	−9444.65	18	—	—	—	—	—	—	4.22	0.66	—	0.58	—
K+R+T	−9450.24	18	—	—	—	—	—	—	3.78	0.68	0.67	—	—
A.K+K+R	−9437.00	18	—	2.39	—	—	—	—	3.22	0.67	—	—	—
G+K+R	−9356.98	18	—	—	—	14.38	—	—	3.81	0.68	—	—	—
G+G.K+K+R	−9353.46	19	—	—	—	9.80	1.93	—	3.55	0.66	—	—	—
K+R+T.K+T.K.R	−9439.56	19	—	—	—	—	—	—	4.11	0.58	—	0.43	1.70
A+A.K+K+R	−9436.84	19	0.92	2.58	—	—	—	—	3.17	0.67	—	—	—
A.K+A.K.R+K+R	−9435.94	19	—	3.06	0.71	—	—	—	3.25	0.69	—	—	—
K+R+T+T.K	−9444.65	19	—	—	—	—	—	—	4.23	0.66	1.01	0.58	—
G+G.K+G.R+K+R	−9350.42	20	—	—	—	15.81	2.00	0.48	3.61	0.70	—	—	—
G+G.K+K+R+T.K	−9340.81	20	—	—	—	9.70	1.67	—	4.04	0.63	—	0.65	—
A.K+G+G.K+K+R	−9341.43	20	—	1.90	—	9.63	1.84	—	3.31	0.64	—	—	—
A.K+G+G.K+K+R+T.K	−9333.14	21	—	1.68	—	9.58	1.64	—	3.74	0.62	—	0.69	—
A.K+G+G.K+G.R+K+R+T.K	−9330.96	22	—	1.66	—	14.40	1.74	0.53	3.78	0.65	—	0.70	—
A.K+G+G.K+K+R+T.K+T.K.R	−9327.77	22	—	1.70	—	9.13	1.73	—	3.62	0.55	—	0.51	1.74
A.K+G+G.K+G.R+K+R+T.K+T.K.R	−9326.60	23	—	1.69	—	12.28	1.81	0.63	3.67	0.57	—	0.53	1.66

NOTE.—ln*L* is log-likelihood; nfp is number of free parameters. The order of terms in the each parameterization is alphabetical. See table 1 for definitions of terms.

Bat Analysis

The results from the analysis of bats substantially mirrored results from the primates, with some exceptions involving the terms ***G.K***, ***G.R*** and ***T.K.R*** (table 3). In the case of the ***G.K*** and ***T.K.R*** terms, the differences are small—the LR tests in table 3 still support the nominal significance of these terms, albeit at the higher 0.1 level. The maximum-likelihood parameter estimates (tables 2 and 4) are also consistent; their respective estimates from primates and bats being approximately 2 and approximately 1.4 for ***G.K***, and approximately 1.7 and approximately 1.30 for ***T.K.R***. The bat data set does not, however, support the significance of the ***G.R*** term. Whereas the relative magnitude of ***G.R*** estimated between the two data sets is consistent (bats $\sim$ 0.75, primates $\sim$0.48), the estimate of ***G*** from bats is approximately one half that estimated for primates ($\sim$8 and $\sim$14). A test of the hypothesis that ***G*** for bats equals ***G*** from primates versus the alternate hypothesis that ***G*** is different was strongly rejected (LR = 38.64, df = 1, $P < 10^{-9}$). A lesser value of ***G*** indicates a smaller relative number of substitutions involving CpG-containing codons in the bat data set and, thus, a smaller number of CpG-related nonsynonymous substitutions. The ***G.R*** discrepancy, therefore, plausibly reflects differences in statistical power.

Table 3
Probabilities from LR Tests Discordant Between the Primate and Bat Data Sets

Null Terms	Unique Alternate Terms	Probability	
		Primates	Bats
G+K+R	***G.K***	0.008	0.120
G+G.K+K+R	***G.R***	0.014	0.254
A.K+G+G.K+K+R+T.K	***G.R***	0.037	0.494
A.K+G+G.K+K+R+T.K	***T.K.R***	0.001	0.070
A.K+G+G.K+G.R+K+R+T.K	***T.K.R***	0.003	0.086

NOTE.—Null terms are the Null hypothesis parameterization (the order of terms in each parameterization is alphabetical). Unique Alternate terms are the additional terms in the Alternate hypothesis parameterization. Probability is the result of the LR test that compares the Null and Alternate parameterizations.

Discussion

Incorporating dinucleotide effects in codon substitution models markedly improved model fit, with a nominal increase in model complexity. Five of the nine dinucleotide terms considered were highly significant in the primate analysis, and three were highly significant in the bat analysis. The significant terms support a distinct rate of mutation for each of the three putatively methylated dinucleotides in *BRCA1* but rate elevation only for codons that contain CpA and CpG. The magnitude of the rate of substitution was greatest for codons that contain CpG. In our analyses the estimation of the dinucleotide terms was confounded by the operation of natural selection, which indicates that comparisons of these terms between genes will require incorporating interaction terms with the nonsynonymous substitution rate.

As all terms in the models presented here confound forward and backward substitution processes, the detected effects may theoretically arise from either substitution direction. The evidence from pedigree studies, however, overwhelmingly identifies loss of ^{m}C as the predominant direction of mutation (e.g., Sommer, Scaringe, and Hill 2001). The ***A.K***, ***G***, and ***T.K*** terms will, therefore, underestimate the rate of ^{m}C-driven dinucleotide loss, but should still exceed 1 if methylation of their dinucleotides is common. In the following discussion we assume that loss of ^{m}C is the dominant direction of substitution.

CpG-containing codons exhibited an elevated rate of substitutions consistent with a contribution from ^{m}C mutation. As the spontaneous mutation of ^{m}CpG leads to TpG, we naturally expect only an elevated transition rate at

Table 4
Estimated Statistics from Bat *BRCA1*

Model	ln*L*	nfp	***A***	***A.K***	***A.K.R***	***G***	***G.K***	***G.R***	***K***	***R***	***T***	***T.K***	***T.K.R***
K+R	−10629.06	15	—	—	—	—	—	—	4.14	0.70	—	—	—
G.K+K+R	−10543.64	16	—	—	—	—	8.18	—	3.75	0.68	—	—	—
A+K+R	−10619.52	16	1.43	—	—	—	—	—	4.34	0.68	—	—	—
K+R+T.K	−10603.32	16	—	—	—	—	—	—	4.76	0.66	—	0.58	—
K+R+T	−10610.84	16	—	—	—	—	—	—	4.30	0.69	0.67	—	—
A.K+K+R	−10610.29	16	—	1.91	—	—	—	—	3.84	0.68	—	—	—
G+K+R	−10517.92	16	—	—	—	7.58	—	—	4.36	0.70	—	—	—
G+G.K+K+R	−10516.71	17	—	—	—	6.08	1.39	—	4.19	0.70	—	—	—
K+R+T.K+T.K.R	−10601.41	17	—	—	—	—	—	—	4.71	0.62	—	0.49	1.35
A+A.K+K+R	−10610.25	17	0.96	1.99	—	—	—	—	3.81	0.68	—	—	—
A.K+A.K.R+K+R	−10609.56	17	—	2.34	0.76	—	—	—	3.86	0.70	—	—	—
K+R+T+T.K	−10603.22	17	—	—	—	—	—	—	4.82	0.66	1.06	0.55	—
G+G.K+G.R+K+R	−10516.06	18	—	—	—	7.41	1.40	0.76	4.22	0.71	—	—	—
G+G.K+K+R+T.K	−10500.25	18	—	—	—	5.99	1.24	—	4.72	0.66	—	0.64	—
A.K+G+G.K+K+R	−10501.19	18	—	1.86	—	6.04	1.43	—	3.91	0.67	—	—	—
A.K+G+G.K+K+R+T.K	−10490.18	19	—	1.65	—	5.97	1.29	—	4.38	0.65	—	0.68	—
A.K+G+G.K+G.R+K+R+T.K	−10489.94	20	—	1.65	—	6.70	1.31	0.84	4.39	0.66	—	0.69	—
A.K+G+G.K+K+R+T.K+T.K.R	−10488.54	20	—	1.67	—	5.88	1.32	—	4.33	0.62	—	0.59	1.32
A.K+G+G.K+G.R+K+R+T.K+T.K.R	−10488.47	21	—	1.66	—	6.26	1.32	0.91	4.34	0.62	—	0.59	1.31

NOTE.—ln*L* is log-likelihood; nfp is number of free parameters. The order of terms in the each parameterization is alphabetical. See table 1 for definitions of terms.

these dinucleotides. The studies on factor IX mutations (Sommer, Scaringe, and Hill 2001) suggested, however, that transversions would also be elevated, which indicates that the repair process in some instances replaces the mismatched nucleotide pair T/G, rather than just correcting to either T/A or C/G. The replacement of both nucleotides of the mismatched pair should elevate both transitions and transversions at CpG-containing codons, an effect represented by the term ***G*** in our parameterization. If the repair system replaces only one nucleotide of the mismatched pair, only the rate of transitions is expected to increase, an effect represented by the ***G.K*** term. Both terms significantly improved model fit for the primate data set, whereas only the ***G*** term was consistently significant for the bat data set. The discrepancy between the data sets with respect to the ***G.K*** term plausibly derives from a lower relative rate of substitution that affects CpG-containing codons in the bats (see tables 2 and 4). The effect of the lower relative rate, coupled with the low frequency of CpG-containing codons, is a reduced statistical power to detect the ***G.K*** interaction parameter. These results, therefore, support both a complete replacement of the mismatched pair and a correction of mismatched T/G to T/A.

Processes other than ^{m}C mutation possibly contributed to the significant effect of the CpG-related terms. As the ***G*** term incorporates both transition and transversion changes that affect the C and G nucleotides, any difference (increase or decrease) of the substitution rate that affects these nucleotides could give rise to a ***G*** effect. Arguing against such confounding with another source of mutation is the consistent magnitude (> 1) of the estimates between the two data sets (tables 2 and 4), the highly significant support for the ***G.K*** term from the primate analysis (fig. 2), and numerous intraspecific studies that establish the high rate of CpG mutation (e.g., Rideout et al. 1990; Rousseau et al. 1994; Templeton et al. 2000; Sommer, Scaringe, and Hill 2001). We further note that Markov process substitution models applied to noncoding DNA also revealed elevation of transitions (Butterfield et al. 2004; Siepel and Haussler 2004) and transversions (Siepel and Haussler 2004) at CpGs.

Deamination of methylated cytosine may account for the elevated rate of mutation at codons that include CpA/TpG dinucleotides. Although methylation of dinucleotides other than CpG is well established for prokaryotes and other organisms, this possibility has largely been overlooked in mammals until recently (see Haines, Rodenhiser, and Ainsworth [2001]). In the *Nf1* gene of mouse, the most abundantly methylated non-CpG dinucleotide was CpA. If *BRCA1* is similarly methylated, we should detect an elevated rate of substitution in codons that contain this dinucleotide on either strand. This conjecture was supported by the significantly improved fit of a model that contained the ***A.K*** term and this term's estimated value of approximately 1.7 (see tables 2 and 4). The contribution of non–^{m}C-related mutation processes to this term cannot be eliminated by this analysis or reference to empirical studies (given their paucity). Only a direct examination (such as sodium bisulfite sequencing) of germline DNA will establish the occurrence of ^{m}CpA in *BRCA1* and affirm ^{m}C's likely contribution.

If ^{m}C mutation is truly being detected by the ***A.K*** term, then the lower relative rate compared with that estimated for the CpG terms may result from differences in the extent of time these dinucleotides are methylated in the germline. The developmental pattern of non-CpG methylation is distinct from that of CpG. For the *Nf1* gene, the former was not detected in somatic murine cells, occurred with a maternal germline bias, and disappeared at a point after the two-cell embryo stage (Haines, Rodenhiser, and Ainsworth 2001). We note, however, that no germline bias in non-CpG methylation was evident for the *ADA* locus (Haines, Rodenhiser, and Ainsworth 2001). Nonetheless, these patterns differ from the male-germline bias observed for ^{m}CpG (Monk, Boubelik, and Lehnert 1987). Methylation of CpG also extends to somatic cells, where it is an

important modulator of gene expression. In addition to these developmental differences in methylation of CpG and non-CpG dinucleotides, the proportions of the available non-CpG dinucleotides that become methylated also appears to be lower (Haines, Rodenhiser, and Ainsworth 2001). A lower ^{m}C incidence at CpA relative to CpG dinucleotides would reduce the rate of ^{m}C mutation. One implication of this condition is reduced statistical power to detect the general (transition and transversion) effects for mutation of ^{m}CpA compared with ^{m}CpG for the same sample. A joint examination of other genes will be necessary to attain sufficient statistical power to address whether both transitions and transversions are elevated at codons that contain the CpA/TpG dinucleotides.

The patterns of substitution that affect CpT-containing codons were distinct. In contrast to the elevated rates estimated for CpG-containing codons and CpA-containing codons, the ***T.K*** term that described substitution at CpT/ApG codons was significantly less than 1.0. This finding indicates a reduced rate of substitution relative to the average codon substitution rate. Inclusion of the ***T.K.R*** interaction term, which addresses the confounding influence of natural selection on estimation of ***T.K***, further reduced the estimate of ***T.K*** (~0.43). Consequently, we conclude that these analyses do not support ^{m}C mutation as the basis for the significant CpT/ApG mutation rate in *BRCA1*. The pattern of substitution affecting these dinucleotides in other coding and noncoding regions will establish whether this depression is specific to *BRCA1*.

The significance of the dinucleotide and nonsynonymous substitution interaction terms indicates that these terms are confounded and that comparisons of substitution propensities between genes will need to address this confounding. Although the primate and bat analyses were discordant in support for the ***G.R*** term, this difference plausibly reflects different statistical power of the reduced rate of CpG mutation in the bats, as discussed above. The sensitivity of the test for significance of the ***G.R*** term in the primate analysis of the joint models again probably reflects reduced statistical power casued by the low frequency of codons that contain CpG in *BRCA1* (~1.2%). ***T.K.R*** did not, however, exhibit such sensitivity—both the primate and the bat analyses were concordant in supporting a significant confounding of ***T.K*** and natural selection. The significant effects provided by the terms presumably derive from both the distinctive properties of their amino acids and their positions within the functional protein. As different protein-coding genes are unlikely to have the same selective regime, comparison of ***G***, ***G.K*** or ***T.K*** across genes is not valid unless such confounding effects are modeled. Furthermore, we suggest this possible confounding should be considered for any term that identifies a codon subset, including ***A.K***, despite the lack of support for a significant effect of the ***A.K.R*** term. Similarly, the effect of selection on protein-coding genes will need to be addressed for comparisons of genes with noncoding regions.

We have not accounted for all ^{m}C mutations, because of the affect of neighboring codons. Methylatable dinucleotides at the first and third codon positions can be created if a neighboring codon ends or begins with an appropriate nucleotide; for example, the codon sequence GACGAC creates a CpG dinucleotide. This omission arises because codons are treated independently. Alternative procedures have been described that incorporate, to some degree, dependence on neighboring states (Pedersen, Wiuf, and Christiansen 1998; Siepel and Haussler 2004). Parameter estimates did not differ markedly, however, between dinucleotide and trinucleotide substitution models with and without dependence on preceding states (Siepel and Haussler 2004). From this observation, we suggest that the parameter estimates we obtained are likely to be robust.

One effect of assuming independence is reduced statistical power to detect the finer partitions of evolutionary events. For at least the ***G*** term, the strength of support suggests a reduction in power may not be a concern for other mammal coding genes. The reduced power may, however, be pertinent to the interaction terms and the non-CpG dinucleotides considered here. One alternative modeling approach that overcomes this problem is to parameterize substitutions that affect C at the nucleotide level. We implemented such a model, which represents all substitutions that affect C by the term ***E*** (see Supplementary Material online for transition matrix) and applied it to the primate data set. Although adding the ***E*** term to the baseline Y98 model significantly improved fit ($LR = 6.8$, $df = 1$, $P < 0.01$), its effect was considerably smaller than that obtained from the dinucleotide terms. This difference arises in part because the rate of substitutions that affect the C of CpT/ApG dinucleotides opposes the rate of substitution that affect C of CpG and CpA/TpG dinucleotides.

The cause of the significant difference in the relative rate of substitutions that affect CpG-containing codons in bats compared with primates cannot be completely resolved from the analyses here. It seems unlikely, however, that this discordance derives from differences in natural selection affecting *BRCA1* as the model that included the ***G.R*** term did not cause the estimated values of ***G*** to be more similar. Instead, it seems more likely that the lower relative rate reflects either a lower incidence of ^{m}C in *BRCA1* or better repair of ^{m}C mutations in the bat species sampled here. Resolving these possibilities will require either a direct examination of ^{m}C incidence in *BRCA1* or analyses of other protein-coding sequences from these taxa.

The modeling presented here demonstrates new approaches to dissecting the impact of the modified base ^{m}C on the evolution of protein-coding regions in mammal genomes. At least CpG-containing condon, and possibly CpA-containing codons, may be modified in *BRCA1*, evidenced by their elevated substitution rate. We have also taken advantage of the design of the PyEvolve toolkit to examine interactions between terms. The value of such an effort is indicated by the significant influence of natural selection (represented by the nonsynonymous substitution term ***R***) on estimation of the dinucleotide-specific codon exchangeability terms for *BRCA1*. This effect emphasizes the necessity of explicitly accounting for such confounding when comparing different genomic sequences. Whether these dinucleotide effects are evident for other genes and

whether they exhibit heterogeneity across the genome will be the subject of future articles.

Supplementary Material

Transition matrices for all substitution models are found in Huttley03-0621-supplementary.pdf.

Acknowledgments

We thank the associate editor and three anonymous reviewers for their constructive suggestions. This research used facilities provided by the Australian Partnership for Advanced Computing.

Literature Cited

Butterfield, A., V. Vedagiri, E. Lang, C. Lawrence, M. J. Wakefield, A. Isaev, and G. A. Huttley. 2004. PyEvolve: a toolkit for statistical modelling of molecular evolution. BMC Bioinformat. **5**:1.

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.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

Girard, M., P. Couvert, A. Carrie, M. Tardieu, J. Chelly, C. Beldjord, and T. Bienvenu. 2001. Parental origin of de novo MECP2 mutations in Rett syndrome. Eur. J. Hum. Genet. **9**:231–236.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Haines, T. R., D. I. Rodenhiser, and P. J. Ainsworth. 2001. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. Dev. Biol. **240**:585–598.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

Huttley, G. A., S. Easteal, M. C. Southey, G. G. Giles, M. R. E. McCredie, J. L. Hopper, and D. J. Venter. 2000*a*. Adaptive evolution of the tumor suppressor *BRCA1* in humans and chimpanzees. Nat. Genet. **24**:410–413.

Huttley, G. A., I. B. Jakobsen, S. R. Wilson, and S. Easteal. 2000*b*. How important is DNA replication for mutagenesis? Mol. Biol. Evol. **17**:929–937.

Killian, J. K., T. R. Buckley, N. Stewart, B. L. Munday, and R. L. Jirtle. 2001. Marsupials and Eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. Mammal. Genome **12**:513–517.

Monk, M., M. Boubelik, and S. Lehnert. 1987. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development **99**:371–382.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294**:2348–2351.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. **17**:798–803.

Pedersen, A. K., C. Wiuf, and F. B. Christiansen. 1998. A codon-based model designed to describe lentiviral evolution. Mol. Biol. Evol. **15**:1069–1081.

Pedersen, A. M., and J. L. Jensen. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. **18**:763–776.

Ramsahoye, B. H., D. Biniszkiewicz, F. Lyko, V. Clark, A. P. Bird, and R. Jaenisch. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. USA **97**:5237–5242.

Reik, W., and J. Walter. 2001*a*. Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. **2**:21–32.

———. 2001*b*. Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote. Nat. Genet. **27**:255–256.

Rideout, W. M., 3rd, G. A. Coetzee, A. F. Olumi, and P. A. Jones. 1990. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science **249**:1288–1290.

Rousseau, F., J. Bonaventure, L. Legeai-Mallet, A. Pelet, J. M. Rozet, P. Maroteaux, M. Le Merrer, and A. Munnich. 1994. Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. Nature **371**:252–254.

Schmitz, J., M. Ohme, B. Suryobroto, and H. Zischler. 2002. The colugo (Cynocephalus variegatus, Dermoptera): the primates' gliding sister? Mol. Biol. Evol. **19**:2308–2312.

Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **21**:468–488.

Sommer, S. S., W. A. Scaringe, and K. A. Hill. 2001. Human germline mutation in the factor IX gene. Mut. Res. **487**:1–17.

Templeton, A. R., A. G. Clark, K. M. Weiss, D. A. Nickerson, E. Boerwinkle, and C. F. Sing. 2000. Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am. J. Hum. Genet. **66**:69–83.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Tomatsu, S., K. O. Orii, M. R. Islam, G. N. Shah, J. H. Grubb, K. Sukegawa, Y. Suzuki, T. Orii, N. Kondo, and W. S. Sly. 2002. Methylation patterns of the human beta-glucuronidase gene locus: boundaries of methylation and general implications for frequent point mutations at CpG dinucleotides. Genomics **79**:363–375.

Trappe, R., F. Laccone, J. Cobilanschi, M. Meins, P. Huppke, F. Hanefeld, and W. Engel. 2001. MECP2 mutations in sporadic cases of Rett syndrome are almost exclusively of paternal origin. Am. J. Hum. Genet. **68**:1093–1101.

Tsunoyama, K., M. I. Bellgard, and T. Gojobori. 2001. Intragenic variation of synonymous substitution rates is caused by nonrandom mutations at methylated CpG. J. Mol. Evol. **53**:456–464.

Waddell, P. J., H. Kishino, and R. Ota. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Inform Ser Workshop Genome Inform **12**:141–154.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

Mark Springer, Associate Editor

Accepted June 2, 2004