Interaction between Selection and Biased Gene Conversion in Mammalian Protein-Coding Sequence Evolution Revealed by a Phylogenetic Covariance Analysis

Nicolas Lartillot

MBE Advance Access published October 25, 2012

Interaction between Selection and Biased Gene Conversion in Mammalian Protein-Coding Sequence Evolution Revealed by a Phylogenetic Covariance Analysis

Nicolas Lartillot*[,1,2]

[1]Centre Robert-Cedergren pour la Bioinformatique, Département de Biochimie, Université de Montréal, Québec, Canada

[2]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, Montpellier, France

***Corresponding author:** E-mail: nicolas.lartillot@umontreal.ca.

Associate editor: Asger Hobolth

Abstract

According to the nearly-neutral model, variation in long-term effective population size among species should result in correlated variation in the ratio of nonsynonymous over synonymous substitution rates (d*N*/d*S*). Previous empirical investigations in mammals have been consistent with this prediction, suggesting an important role for nearly-neutral effects on protein-coding sequence evolution. GC-biased gene conversion (gBGC), on the other hand, is increasingly recognized as a major evolutionary force shaping genome nucleotide composition. When sufficiently strong compared with random drift, gBGC may significantly interfere with a nearly-neutral regime and impact d*N*/d*S* in a complex manner. Here, we investigate the phylogenetic correlations between d*N*/d*S*, the equilibrium GC composition (GC*), and several life-history and karyotypic traits in placental mammals. We show that the equilibrium GC composition decreases with body mass and increases with the number of chromosomes, suggesting a modulation of the strength of biased gene conversion due to changes in effective population size and genome-wide recombination rate. The variation in d*N*/d*S* is complex and only partially fits the prediction of the nearly-neutral theory. However, specifically restricting estimation of the d*N*/d*S* ratio on GC-conservative transversions, which are immune from gBGC, results in correlations that are more compatible with a nearly-neutral interpretation. Our investigation indicates the presence of complex interactions between selection and biased gene conversion and suggests that further mechanistic development is warranted, to tease out mutation, selection, drift, and conversion.

Key words: nearly-neutral, GC-biased gene conversion, comparative method, independent contrasts, Bayesian inference.

Introduction

Understanding the forces shaping genome evolution has been the focus of intense research since molecular data first became available. In this context, protein-coding sequences have been especially investigated, not only because of the prominent role of proteins in cellular processes but also because the degeneracy of the genetic code makes it possible to contrast the fate of synonymous and nonsynonymous mutations, thus yielding important insights about the relative strength of selection and random drift.

Although it is not yet totally clear which evolutionary regime best describes the substitution process in protein-coding sequences, the nearly-neutral model (Ohta 1974, 1995; Kimura 1979, 1983) can be considered as an important step toward answering this question. According to the nearly-neutral model, most mutations are deleterious (Ohta 1974) or merely compensate previous deleterious substitutions (Charlesworth and Eyre-Walker 2007), whereas adaptive substitutions represent a small fraction of the observed polymorphism and divergence. These hypotheses lead to a number of predictions concerning the long-term evolutionary behavior of protein-coding sequences. At least in mammals, which have relatively small effective population sizes, synonymous substitutions should be essentially neutral and should therefore reflect the mutation rate and its variation among lineages. The ratio of nonsynonymous over synonymous substitutions (d*N*/d*S*), on the other hand, should be inversely related to the effective population size. In effect, this ratio can be considered as a measure of the width of the selective sieve, decreasing when effective population size (and therefore selection stringency) increases.

Comparative analyses of mitochondrial and nuclear protein-coding genes have in part confirmed these predictions. In mammals, selection on synonymous substitutions appears to be negligible (Duret et al. 2002) or weak (Yang and Nielsen 2008). The synonymous substitution rate correlates negatively with generation time, longevity, or mass (Li and Tanimura 1987; Nabholz et al. 2008; Welch, Bininda-Emonds, et al. 2008; Lanfear et al. 2010). Life-history traits are all strongly correlated with each other, making it difficult to identify the primary cause of variation in the synonymous substitution rate. However, these empirical observations are globally consistent with the idea that rate variation is fundamentally caused by changes in the mutation rate. Finally, a positive correlation of d*N*/d*S* with body size in mitochondrial genomes (Popadin et al. 2007) or generation time in nuclear genomes

(Eyre-Walker et al. 2002; Nikolaev et al. 2007) has been found, which is consistent with a decrease of the strength of purifying selection in species with smaller effective population sizes. The relation between the width of the selective sieve and effective population size has also been modeled more directly, in the context of mutation-selection models (Nielsen and Yang 2003). Such models explicitly rely on mechanistic arguments derived from population genetics to derive the dependency of d*N*/d*S* on effective population size in a manner dependent on the assumed distribution of selective effects (Welch, Eyre-Walker, et al. 2008). In principle, further elaborating such models would make it possible to test hypotheses about the population genetics mechanisms underlying protein evolution using comparative data or to reconstruct the patterns of population size variation at large evolutionary scale.

Beyond the question of its general validity, the neutral model can also be considered as a convenient null model, specifying what the default behavior of genetic sequences should be in the absence of adaptive events (Kreitman 1996). Departures from the null expectation, in particular in the form of an elevated d*N*/d*S* in protein-coding genes, might then be considered as significant evidence in favor of adaptive phenomena (Yang 2002). The relative importance of positive selection is still a matter of debate. Positive selection appears to be less pervasive in vertebrates than in invertebrates, perhaps owing to smaller population sizes in the former (Eyre-Walker and Keightley 2009). On the other hand, methods relying on phylogenetic models or on comparisons between divergence and polymorphism have suggested that a significant number of genes in mammalian genomes might be under positive selection (e.g., Kosiol et al. 2008; Halligan et al. 2010).

The issue of discriminating between purifying or positive selection becomes more complicated in the presence of GC-biased gene conversion (gBGC, Galtier et al. 2001; Duret et al. 2006; Galtier and Duret 2007). Mechanistically related to recombination and repair, gBGC causes a segregation and fixation bias in favor of G and C over A and T. It is increasingly recognized as a major evolutionary force shaping genomic compositional landscapes (Arndt et al. 2005; Galtier et al. 2006; Duret and Arndt 2008; Duret and Galtier 2009). At least in some mammals and birds, the high heterogeneity in the recombination rate within genomes results in large-scale variation in the intensity of biased gene conversion, with higher intensity in regions undergoing more frequent recombination (Meunier and Duret 2004; Arndt et al. 2005; Webster et al. 2005, 2006; Duret and Arndt 2008). In the regions where it is strongest, gBGC can result in episodes of accelerated fixation of mutations toward G and C (Dreszer et al. 2007; Katzman et al. 2011), thus leading to an increase in the overall substitution rate and an increase in the d*N*/d*S* ratio (Berglund et al. 2009; Galtier et al. 2009; Ratnakumar et al. 2010). In principle, biased gene conversion can therefore confound the diagnostic of which selective regime is acting on protein-coding sequences based on a measure of d*N*/d*S*.

Variation in genome-wide GC content between species may also be due to changes in the strength of gBGC. First, because it is formally equivalent to positive selection in favor of G or C (Nagylaki 1983; Glémin 2010), gBGC is expected to be stronger in species with a larger effective population size. Second, gBGC correlates positively with the recombination rate (Webster et al. 2006; Duret and Arndt 2008). For proper disjunction, at least one recombination event has to take place per chromosome during meiosis, and therefore, the recombination rate per base pair should be negatively correlated with chromosome size. As a consequence, a higher recombination rate is observed in mammalian species with a larger number of (generally smaller) chromosomes (Pardo-Manuel de Villena and Sapienza 2001). Altogether, under the combined effect of random drift and recombination rate, equilibrium GC content is thus expected to be higher in species with either larger populations or smaller chromosomes.

The role of gBGC in mammalian genome evolution has been extensively investigated in a comparative framework (Duret et al. 2002; Capra and Pollard 2011; Clément and Arndt 2011; Escobar et al. 2011). In a recent study of a large set of protein-coding sequence alignments from 33 placentals (Romiguier et al. 2010), the genome-wide patterns of GC content were found to display contrasted patterns in different lineages, with some species displaying marked increase of the GC content and others conforming to the erosion pattern initially observed in primates or murid rodents. In addition, the overall GC content was found to correlate with body mass and genome size, thus partially confirming the theoretical predictions. On the other hand, no significant correlation was found with the number of chromosomes.

Altogether, both purifying selection and gBGC appear to undergo systematic changes at large phylogenetic scale. Such variation is amenable to comparative analysis, thereby offering the opportunity to test hypotheses about evolutionary mechanisms and to reconstruct the evolution of genetic systems and life-history patterns (Lartillot and Delsuc 2012). On the other hand, both purifying selection and gBGC are susceptible to correlate with the same features of mammalian life history, because of the underlying dependence of both variables on effective population size. Finally, because gBGC mimics positive selection, there is a risk that the correlation patterns of d*N*/d*S* across mammals would also be significantly affected by the confounding effects of gBGC. The combined effects of gBGC and purifying selection across species differing in their genetic system and life-history traits are likely to be complex. Teasing them apart, however, could reveal the correlates of each variable and might lead to a better understanding of their interactions.

In this study, we investigate the interplay between selection and biased gene conversion in mammalian protein-coding sequence evolution. Extending a phylogenetic covariance framework introduced previously (Lartillot and Poujol 2011; Lartillot and Delsuc 2012), so as to analyze the correlations between the synonymous substitution rate, the equilibrium GC content, and specialized versions of the d*N*/d*S* ratio, with life-history traits, and genomic variables, we show

that biased gene conversion is a likely determinant of the observed variation in equilibrium nucleotide composition among placental mammals and has a significant influence on the substitution patterns observed in protein sequences and on the correlations estimated by phylogenetic covariance methods.

Materials and Methods

Model

The phylogenetic covariance model was introduced by Lartillot and Poujol (2011). Briefly, it is a generalization of comparative methods based on the principle of phylogenetically independent contrasts (Felsenstein 1985) to the problem of estimating correlations between quantitative characters and parameters of the substitution process. Substitution parameters and quantitative characters are jointly represented by the components of a multivariate Brownian diffusion process $X(t)$ running along the lineages of a time-valued phylogeny. The Brownian process is parameterized by a covariance matrix Σ to be estimated. Inference is done by conditioning the model on a combination of a codon sequence alignment and a matrix of quantitative characters on the same set of taxa, using fossil constraints to calibrate divergence times. Samples from the posterior distribution over the parameters of the model are obtained using Monte Carlo methods, thus providing estimates in terms of marginal expectations, and confidence in terms of posterior probabilities.

The simplest version of the model considered here (similar to the one defined in Lartillot and Poujol 2011) relies on a codon substitution matrix Q, itself based on a time-reversible process between nucleotides defined as follows:

$$R = \begin{pmatrix} - & \rho_{AC}\frac{\gamma}{2} & \rho_{AG}\frac{\gamma}{2} & \rho_{AT}\frac{1-\gamma}{2} \\ \rho_{AC}\frac{1-\gamma}{2} & - & \rho_{CG}\frac{\gamma}{2} & \rho_{CT}\frac{1-\gamma}{2} \\ \rho_{AG}\frac{1-\gamma}{2} & \rho_{CG}\frac{\gamma}{2} & - & \rho_{GT}\frac{1-\gamma}{2} \\ \rho_{AT}\frac{1-\gamma}{2} & \rho_{CT}\frac{\gamma}{2} & \rho_{GT}\frac{\gamma}{2} & - \end{pmatrix},$$

where γ is the equilibrium GC frequency, and ρ_{XY} is the relative exchangeability between nucleotides X and Y. The matrix R is normalized, so that rates are measured in expected number of substitutions per position. The rate of substitution between any pair of codons (b_1, b_2) differing only at one position and with respective nucleotides n_1 and n_2 at that position is then defined as follows (Muse and Gaut 1994):

$$\begin{aligned} Q_{b_1b_2} &= r\,R_{n_1n_2}, && \text{if } b_1 \text{ to } b_2 \text{ is synonymous,} \\ Q_{b_1b_2} &= r\,\omega\,R_{n_1n_2}, && \text{if } b_1 \text{to } b_2 \text{ is nonsynonymous,} \end{aligned}$$

where r is thus the rate of synonymous substitution. The model allows for correlations between the instant substitution rate $dS = r(t)$ and the ratio of synonymous over nonsynonymous substitutions $dN/dS = \omega(t)$, with L quantitative characters. The multivariate Brownian diffusion process is therefore of dimension $M = L + 2$:

$$\begin{aligned} X_1(t) &= \ln r(t), \\ X_2(t) &= \ln \omega(t), \\ X_{l+2}(t) &= \ln C_l(t) \quad l = 1, \ldots, L. \end{aligned}$$

A first variant of this model was obtained by introducing a time-dependent equilibrium GC composition [$GC^* = \gamma(t)$]. The Brownian process of this new model is of dimension $M = L + 3$:

$$\begin{aligned} X_1(t) &= \ln r(t), \\ X_2(t) &= \ln \omega(t), \\ X_3(t) &= \ln \frac{\gamma(t)}{1-\gamma(t)}, \\ X_{l+3}(t) &= \ln C_l(t) \quad l = 1, \ldots, L. \end{aligned}$$

A logit transformation is used for $\gamma(t)$ so as to transform a variable in (0,1) into a variable over the entire real line. In this model, as in other nonhomogeneous models (Galtier and Gouy 1998; Boussau et al. 2008), an important distinction should be made between the instant value of the equilibrium GC composition at the root of the tree, $\gamma(0) = e^{X_3(0)}/(1+e^{X_3(0)})$, and the GC content of the sequence at the root of the tree, γ_{root}. This latter variable captures all the (potentially complex) history of compositional variation along the lineage leading to the rooting point of the phylogeny. Here, $\gamma(0)$ and γ_{root} are two independent parameters of the model.

A second adaptation of the model introduces time-dependent modulators of the ratio of transition over GC-conservative transversion rates (κ_{ts}) and of the ratio of non-GC-conservative over GC-conservative transversion rates (κ_{tvgc}), such that the nucleotide matrix is now:

$$R = \begin{pmatrix} - & \rho_{AC}\kappa_{tvgc}\frac{\gamma}{2} & \rho_{AG}\kappa_{ts}\frac{\gamma}{2} & \rho_{AT}\frac{1-\gamma}{2} \\ \rho_{AC}\kappa_{tvgc}\frac{1-\gamma}{2} & - & \rho_{CG}\frac{\gamma}{2} & \rho_{CT}\kappa_{ts}\frac{1-\gamma}{2} \\ \rho_{AG}\kappa_{ts}\frac{1-\gamma}{2} & \rho_{CG}\frac{\gamma}{2} & - & \rho_{GT}\kappa_{tvgc}\frac{1-\gamma}{2} \\ \rho_{AT}\frac{1-\gamma}{2} & \rho_{CT}\kappa_{ts}\frac{\gamma}{2} & \rho_{GT}\kappa_{tvgc}\frac{\gamma}{2} & - \end{pmatrix},$$

where $\gamma = GC^*$ is also allowed to vary among lineages.

At the codon level, three distinct time-dependent ratios of nonsynonymous over synonymous substitutions are introduced, so that the rate of substitution between any pair of codons (b_1, b_2) differing only at one position and with respective nucleotides n_1 and n_2 at that position is as follows:

$$\begin{aligned} Q_{b_1b_2} &= r\,R_{n_1n_2}, && \text{if } b_1 \text{ to } b_2 \text{ is synonymous,} \\ Q_{b_1b_2} &= r\,\omega_{ts}\,R_{n_1n_2}, && \text{if } b_1 \text{to } b_2 \text{is a nonsynonymous transition,} \\ Q_{b_1b_2} &= r\,\omega_{tv0}\,R_{n_1n_2}, && \text{if } b_1 \text{ to } b_2 \text{ is a nonsynonymous GC-conservative transversion,} \\ Q_{b_1b_2} &= r\,\omega_{tvgc}\,R_{n_1n_2}, && \text{if } b_1 \text{ to } b_2 \text{ is a nonsynonymous and non GC-conservative transversion.} \end{aligned}$$

The overall Brownian process is now of dimension $M = L + 7$:

$$\begin{aligned}
X_1(t) &= \ln r(t),\\
X_2(t) &= \ln \kappa_{ts}(t),\\
X_3(t) &= \ln \kappa_{tvgc}(t),\\
X_4(t) &= \ln \omega_{ts}(t),\\
X_5(t) &= \ln \omega_{tv0}(t),\\
X_6(t) &= \ln \omega_{tvgc}(t),\\
X_7(t) &= \ln \frac{\gamma(t)}{1-\gamma(t)},\\
X_{l+7}(t) &= \ln C_l(t) \quad l = 1, \ldots, L.
\end{aligned}$$

To avoid nonidentifiability between the time-independent exchangeability parameters (ρ) and the time-dependent modulators (κ), we impose the constraint that $\kappa_{ts}(0) = \kappa_{tvgc}(0) = 1$ at the root of the tree.

All other aspects of the model, including the priors, are as in Lartillot and Poujol (2011), except for a uniform prior on γ_{root}. In addition, the prior on the covariance matrix and on divergence times are as in Lartillot and Delsuc (2012): on the covariance matrix Σ, an inverse Wishart prior of parameter $\Sigma_0 = \mathrm{Diag}(\eta_1, \ldots, \eta_M)$, and with M degrees of freedom, where $\eta_m, m = 1M$, are themselves from a truncated log-uniform prior, on $[10^{-3}, 10^3]$, and on divergence times, a birth–death prior with parameters λ (birth rate), μ (death rate), and θ (sampling fraction at $t = 0$). To ensure identifiability, and as in Rannala and Yang (2007), we set $p_1 = \lambda - \mu$ and $p_2 = \lambda\theta$. We impose an exponential prior of mean 10^{-3} on both p_1 and p_2 (with time being measured, such that the root-to-tip distance is equal to 1).

Under each condition, the Markov chain Monte Carlo (MCMC) sampler was run for a total of 150,000 cycles. Saving every 30 cycles and discarding the first 1,000 points allowed estimation of posterior averages on the remaining 4,000 points. Two independent runs were performed under each of the settings. Convergence and mixing of the MCMC were first assessed visually, and then quantified by measuring, for several key statistics (log likelihood, log prior, mean substitution rate over the tree, mean omega over the tree, entries of the covariance matrix, and root age), the effective sample sizes, and the discrepancy between the credibility intervals obtained from the two independent runs (Lartillot et al. 2009). The estimated effective sample size was greater than 100, and discrepancy between two independent chains was smaller than 0.2 for all statistics. In the case of the entries of the covariance matrix, which are the key variables of interest in the present context, the effective sample sizes were larger than 800, and the discrepancy was smaller than 0.1. Globally across models, estimated correlation coefficients differed by at most 0.01 and posterior probabilities by at most 0.02 between two independent runs.

Multiple regression is performed as in Lartillot and Poujol (2011). Briefly, the covariance between traits k and l (entries k and l of the Brownian process), when controlling for variation in trait m, is (e.g., Mardia et al. 1979, p. 170):

$$\Sigma_{kl;m} = \Sigma_{kl} - \frac{\Sigma_{km}\Sigma_{lm}}{\Sigma_{mm}}.$$

This function of Σ is averaged over the posterior distribution, and significance is assessed by computing the posterior probability that it is negative (or positive).

Data and Fossil Constraints

The codon sequence alignment used in this study is a previously introduced concatenation of 17 nuclear protein-coding genes in 73 placental taxa (Lartillot and Delsuc 2012). All analyses were performed under a tree topology reflecting the classical multigene phylogenetic studies dividing placental mammals into four major groups (Madsen et al. 2001; Murphy et al. 2001): Afrotheria, Xenarthra, Euarchontoglires, and Laurasiatheria. The tree was rooted, such that Afrotheria and Xenarthra form a monophyletic sister group of Boreoeutheria (Murphy et al. 2007). For calibrating the tree, we used the nine fossil constraints defined by Springer et al. (2003). Information about life-history traits was obtained from the AnAge database (de Magalhaes and Costa 2009) and compiled as described in Lartillot and Delsuc (2012). Karyotypic data and C values were obtained from the Animal Genome Size Database (Gregory et al. 2007) and from the Atlas of Mammalian Chromosomes (O'Brien et al. 2006).

Results

Correlation Analyses

We conducted a first covariance analysis based on a model correlating dS, dN/dS, and the equilibrium GC (GC*) with three life-history traits: female maturity, adult body mass and maximum recorded life span (taken as proxies for generation time, body size, and longevity), and two genomic variables, the C value (as a proxy for genome size) and the number of chromosomes ($2n$, table 1).

In the following, all correlations are between changes in the variables under investigation, as is also the case in other methods related to the phylogenetically independent contrasts (Felsenstein 1985; Martins and Hansen 1997; Paradis and Claude 2002; Garland et al. 2005). However, for simplicity, we refer to these correlations as if they were directly between the values of the variables. For instance, the correlation between the variation in dS and the variation in body mass along the lineages (both on the logarithmic scale) is more simply referred to as the correlation between dS and body mass.

The correlation of dS with life-history traits is described elsewhere (Lartillot and Delsuc 2012). Briefly, in accordance with previous analyses, we find a significant negative correlation of dS with all three life-history traits, primarily due to a longevity or generation time effect (table 1). The dN/dS ratio displays a marginally significant positive correlation with longevity (correlation coefficient $R = 0.27$, posterior probability of a positive correlation pp = 0.96) and with female age at sexual

Table 1. Covariance Analysis (Posterior Mean Correlation Coefficients) under the (dS, ω, GC*) Parameterization.

R	ω	GC*	Maturity	Mass	Longevity	*C* Value	No. of Chromosomes
dS	−0.16	0.43*	−0.51*	−0.62*	−0.67*	0.00	0.18
ω	—	−0.37*	0.27**	0.13	0.27**	−0.08	−0.09
GC*	—	—	−0.06	−0.33*	−0.21	−0.06	0.35*
Maturity	—	—	—	0.6*	0.68*	0.03	−0.03
Mass	—	—	—	—	0.74*	0.13	0.06
Longevity	—	—	—	—	—	−0.04	0.05
C value	—	—	—	—	—	—	0.03

*Posterior probability of a positive covariance > 0.975 or < 0.025.
**Posterior probability of a positive covariance > 0.95 or < 0.05.

maturity ($R = 0.27$, pp = 0.95). This positive correlation could be interpreted as an indirect negative correlation between d*N*/d*S* and effective population size, and as such, it would agree with the nearly-neutral model, that is, a stronger purifying selection in species with larger effective population sizes (Ohta 1974; Kimura 1979; Eyre-Walker et al. 2002; Nikolaev et al. 2007; Popadin et al. 2007). On the other hand, no correlation is seen with body mass ($R = 0.13$, pp = 0.82). Under a model assuming no variation in GC* (thus correlating only d*S* and d*N*/d*S* with life-history traits), the results are slightly different, in that d*N*/d*S* displays a significant positive correlation with maturity (pp = 0.98) and a marginally significant correlation with longevity (pp = 0.95), however, again, no correlation with body mass (pp = 0.67). Controlling for longevity and maturity even suggests a weak negative correlation between d*N*/d*S* and body mass ($R = -0.26$, pp = 0.92) under both models. Although a lack of correlation with body mass in the case of nuclear sequences was already reported (Nikolaev et al. 2007), this is surprising, given that body size is often considered as the most reliable predictor of population size in mammals. Empirical support for the nearly-neutral model in mammals, based on mitochondrial protein-coding sequences, relied on positive correlations of d*N*/d*S* and body size in cytochrome B (Popadin et al. 2007; Lartillot and Poujol 2011).

As for GC*, it correlates positively with the number of chromosomes ($R = 0.35$) and with d*S* ($R = 0.43$) and negatively with d*N*/d*S* ($R = -0.37$) and body mass ($R = -0.33$). All four correlations were significant (pp > 0.99) and remained so after controlling for all other variables (not shown). A weak correlation is also observed between GC* and longevity ($R = -0.21$, pp = 0.92, table 1). The correlation between GC* and the number of chromosomes is in agreement with the prediction that recombination rate is indirectly a major determinant of GC*, strongly suggesting a role of biased gene conversion in driving genomic GC content (Meunier and Duret 2004; Duret and Arndt 2008). The positive correlation between d*S* and GC* could be interpreted in two different ways. Species with short generation times (hence a higher d*S*) also have larger population sizes and are therefore under stronger gBGC (hence a higher GC*). Thus, the correlation could be indirectly mediated by the correlation between generation time and population size. Not totally in accordance with this interpretation, however, is the absence of correlation of GC* with maturity. An alternative explanation would be that gBGC is the direct cause of the increase in the rate of synonymous substitution.

Finally, the negative correlation of GC* with d*N*/d*S* and body size could be interpreted in terms of an indirect correlation with effective population size, with stronger gBGC being expected in larger populations. This interpretation, however, would imply symmetrical correlation patterns for GC* and d*N*/d*S* with life-history traits, yet, as already mentioned, d*N*/d*S* does not correlate with body size, and conversely, unlike d*N*/d*S*, GC* correlates only marginally with longevity and not at all with maturity. Thus, in spite of the fact that d*N*/d*S* and GC* are both expected to be primarily influenced by effective population size, the correlation patterns observed for these two variables with life-history traits are not congruent.

Teasing Out gBGC and Selection

A possible explanation for the discrepancies observed between the correlation patterns of GC* and d*N*/d*S* would be the presence of confounding effects acting on substitution patterns. Saturation is one possible candidate for such confounding factors. Alternatively, gBGC itself could significantly affect substitution rates. The effects of gBGC on d*S* and d*N*/d*S* are likely to be complex. First, gBGC is expected to change the relative rate at which substitutions occur, depending on whether they are GC conservative or not. Second, for a given type of substitution, GC-conservative, toward or from GC, gBGC may differentially affect synonymous and nonsynonymous substitutions, in a way that will depend on the intensity of gBGC and on the distribution of selective effects on nonsynonymous mutations (Galtier et al. 2009).

Importantly, in the case of non-GC-conservative mutations, the synonymous substitution rate is not anymore simply equal to the mutation rate in the presence of gBGC, and therefore, the ratio of nonsynonymous over synonymous substitutions is not easily interpretable. On the other hand, the d*N*/d*S* computed on GC-conservative transversions is still interpretable as the width of the selective sieve. Therefore, restricting the estimation of d*N*/d*S* using exclusively GC-conservative transversions would provide a method for insulating the measure of the strength of selection classically

Table 2. Covariance Analysis (Posterior Mean Correlation Coefficients) under the Complete Parameterization.

R	κ_{ts}	κ_{tvgc}	ω_{ts}	ω_{tv0}	ω_{tvgc}	GC*	Maturity	Mass	Longevity	*C* Value	No. of Chromosomes
dS	−0.62*	0.02	0.15	−0.45*	−0.18	0.36*	−0.4*	−0.58*	−0.63*	0.04	0.25**
κ_{ts}	—	0.39**	−0.55*	0.15	−0.17	−0.10	0.03	0.25	0.26	−0.18	−0.22
κ_{tvgc}	—	—	−0.6*	−0.30	−0.72*	0.72*	−0.08	−0.23	−0.13	−0.28	−0.03
ω_{ts}	—	—	—	0.53*	0.74*	−0.39*	0.26	0.15	0.24	0.00	−0.02
ω_{tv0}	—	—	—	—	0.65*	−0.50*	0.46*	0.45*	0.59*	−0.05	−0.22
ω_{tvgc}	—	—	—	—	—	−0.74*	0.14	0.22	0.26	0.02	−0.22
GC*	—	—	—	—	—	—	−0.04	−0.33*	−0.19	−0.17	0.34*
Maturity	—	—	—	—	—	—	—	0.58*	0.67*	0.05	−0.03
Mass	—	—	—	—	—	—	—	—	0.73*	0.17	0.05
Longevity	—	—	—	—	—	—	—	—	—	−0.03	0.04
C value	—	—	—	—	—	—	—	—	—	—	0.02

*Posterior probability of a positive covariance > 0.975 or < 0.025.
**Posterior probability of a positive covariance > 0.95 or < 0.05.

associated with the d*N*/d*S* ratio from the potentially confounding effects of biased gene conversion.

On the basis of these observations, we devised a variant of the phylogenetic covariance model introduced earlier, by dividing substitutions in three different pools: transitions (C:G to T:A), GC-conservative transversions (C:G to G:C and A:T to T:A), and non-GC-conservative transversions (C:G to A:T). Variation in dS and d*N*/d*S* are estimated separately on each pool (see Materials and Methods). The model also allows for variation in GC* among lineages.

The model so extended reveals several interesting patterns (table 2). First, the synonymous substitution rate (now based on GC-conservative transversions) still correlates negatively with maturity, mass, and longevity (pp > 0.99). The relative transition rate (κ_{ts}), on the other hand, displays a weak positive correlation with body mass and longevity (pp = 0.92). A possible explanation would be that transitions, in particular at CpG sites, may be less replication dependent than transversions (Taylor et al. 2006), making the transition rate less dependent on generation time than the transversion rate. Alternatively, transitions could be saturated at the scale of placentals and be more saturated in faster evolving lineages. This would result in the relative transition rate to appear artifactually lower in fast evolving species. The relative non-GC conservative over GC-conservative transversion rate (κ_{tvgc}) does not display any significant correlation with life-history or genomic variables.

Concerning nonsynonymous substitutions, in the absence of gBGC, or any other confounding effect, all three d*N*/d*S*, restricted to transitions, GC-conservative, or non-GC-conservative transversions, should display the same correlation patterns with other variables. This is not what is observed, however. Although the two non-GC-conservative d*N*/d*S* do not display significant correlations with life-history traits nor with genomic variables, the d*N*/d*S* restricted to GC-conservative transversions correlates positively with all three life-history traits ($R > 0.45$ and pp ≥ 0.99 in all three cases). In particular, unlike the uncorrected d*N*/d*S* but like the GC* and the mitochondrial d*N*/d*S*, the GC-conservative d*N*/d*S* displays the positive correlation with body size that is expected assuming a nearly-neutral model with body size as a proxy for population size.

Long-Term Trends in Population Size Variation

The variation in the global or GC-conservative d*N*/d*S* over the phylogeny might provide an interesting first glimpse into the variation in effective population size at the scale of placentals. The equilibrium GC can also be reconstructed over the entire tree (fig. 2 and supplementary fig. S6, Supplementary Material online), so that the two substitution variables can be jointly examined.

Defining an effective population size at such a large evolutionary scale certainly requires some qualification (see Discussion). Nevertheless, the global picture emerging from this reconstruction displays several interesting features (fig. 1 and supplementary fig. S5, Supplementary Material online). Groups of large and long-living mammals such as cetartiodactyls and more particularly cetaceans, or, most strikingly, anthropoid primates, display elevated values of GC-conservative d*N*/d*S*. In afrotherians, the elephant *Loxodonta* and the manatee *Trichechus* have higher d*N*/d*S* than the hyrax *Procavia* or the elephant shrews *Elephantulus* and *Macroscelides*. The relative values of d*N*/d*S* between closely related groups with different body sizes are also instructive. Thus, the large hystricomorphs have higher d*N*/d*S* than other smaller rodents. Similarly, megachiropterans have larger d*N*/d*S* compared with microchiropterans. On the other hand, carnivores, and particularly Caniformia, have a low d*N*/d*S*. This is perhaps surprising, both in terms of body size and given the high position of carnivores in food webs.

The reconstructed d*N*/d*S* provides us with information about more ancient lineages as well. For instance, lineages within cetartiodactyls display a pattern of convergent evolution toward larger values of d*N*/d*S*, indicating a convergent decrease in the long-term effective population size over a time span of 60 My. A similar pattern is observed in the case of anthropoid primates, over the 30 My since their last common ancestor. Interestingly, a pronounced slowdown of the synonymous substitution rate in these two groups also

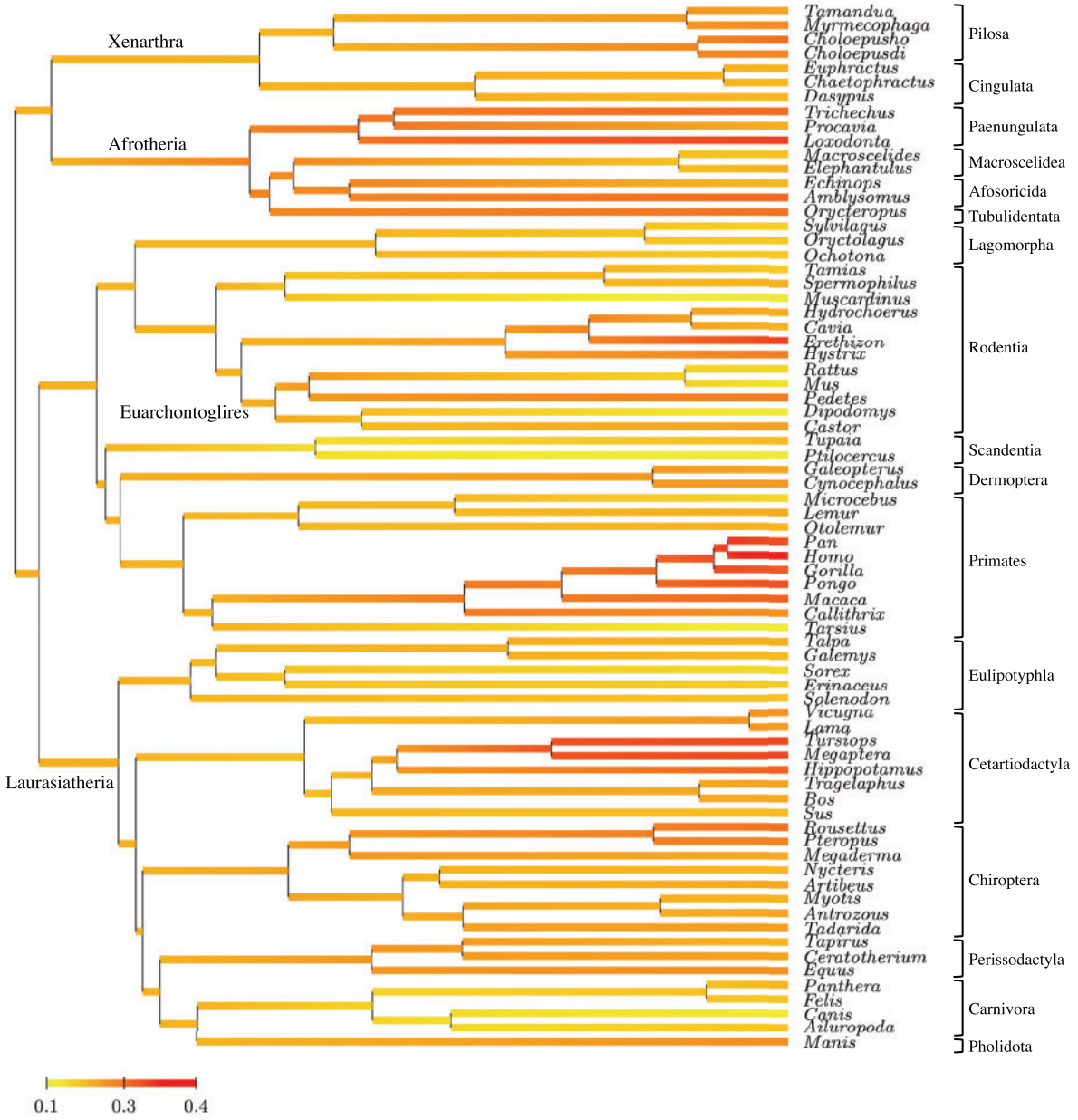

Fig. 1. Posterior mean reconstruction of the GC-conservative d*N*/d*S* along the placental phylogeny. For credibility intervals, see supplementary figure S5, Supplementary Material online.

supports a scenario consisting of a convergent increase in body size in both cases (Lartillot and Delsuc 2012). In contrast, afrotherians seem to have had large values of d*N*/d*S* since early in their evolution, followed by a secondary decrease in elephant shrews and the hyrax.

It is instructive to compare the picture offered by the GC-conservative d*N*/d*S* (fig. 1 and supplementary fig. S5, Supplementary Material online) with what would have been inferred based on the global d*N*/d*S* (supplementary figs. S1 and S7, Supplementary Material online). The two are partially concordant, but there are several differences. For instance, among primates, anthropoids do not have an especially high global d*N*/d*S*, which is in sharp contrast with their GC-conservative d*N*/d*S*, one of the highest among placentals. Conversely, xenarthrans, in particular *Dasypus*, and some microchiropteres (*Antrozous* and *Myotis*) have a high global d*N*/d*S* but a moderate GC-conservative d*N*/d*S*.

Discussion

The Many Possible Causes of d*N*/d*S* Variation

There certainly are reasons to remain cautious when interpreting the fine-grained structure of nonsynonymous versus synonymous substitution patterns across placentals. The variation in d*N*/d*S* observed among mammals is small and potentially influenced by many confounding factors. Nevertheless, what the present analysis at least shows is that the influence of gBGC on d*N*/d*S* should not be ignored in comparative studies. In the absence of biased gene conversion, the various measures of d*N*/d*S* obtained by restricting its

estimation on any particular subtype of substitutions (fig. 1 and figs. S2, S3, S5, S8, and S9, Supplementary Material online) should have resulted in similar correlation patterns with all other variables. What we have obtained in this analysis contradicts this prediction. Global patterns of nonsynonymous substitutions, across proteins and lineages, therefore seem to be potentially distorted by the presence of gBGC, and this urges caution when interpreting such patterns in terms of strength of selection.

Saturation might also confound the estimation of the strength of selection in protein-coding sequences. Saturation of transitions is visible but appears to be weak on a saturation plot (supplementary fig. S4, Supplementary Material online). On the other hand, it is known that transitions resulting from the deamination of methylated cytosines in a CpG context occur at a very high rate, 10–50 times higher than for other types of point substitutions (Nachman and Crowell 2000; Arndt et al. 2003). Therefore, it is possible that saturation specifically affecting methylated cytosines is responsible for deviations in the estimated d*N*/d*S*. Further investigation of this possibility is warranted, although this would require more complex codon-models with context-dependent effects.

Apart from biased gene conversion and saturation, positive selection is another important factor potentially influencing d*N*/d*S*. The relatively low value of d*N*/d*S* over the tree (between 0.1 and 0.4, fig. 1 and supplementary fig. S5, Supplementary Material online) suggests that selection at the coding level is mostly purifying. In addition, at least under mutation-limited models, positive selection would normally result in a positive correlation between d*N*/d*S* and effective population size (Kimura 1983) and therefore a negative, instead of a positive, correlation between d*N*/d*S* and life-history traits. On the other hand, it is possible to argue that positive selection might result in the accumulation of adaptive nonsynonymous substitutions at a rate that would be partially decoupled from the mutation rate (Gillespie 1991) and that would be essentially determined by the fluctuations of the selective environment. If this was the case, then variation in the mutation rate, and therefore in d*S*, among lineages could result in apparent variation in d*N*/d*S* that would not be related to changes in effective population size but would merely reflect variation in the denominator of the ratio. In a similar spirit, one could imagine that, to remain adapted to a constantly changing environment, species with long generation times may accumulate more adaptive substitutions per generation, compared with species with short generation times. This, again, would result in a positive correlation of d*N*/d*S* with generation time and, therefore, with body size. Discriminating between such adaptive hypotheses and the more classical nearly-neutral interpretation appears to be difficult in the present context.

The good complementarity between GC* (fig. 2 and supplementary fig. S6, Supplementary Material online) and the GC-conservative d*N*/d*S* (fig. 1 and supplementary fig. S5, Supplementary Material online), although it may simply reflect the opposite responses of the two variables to population size, suggests a possible artifact on d*N*/d*S*. As an estimator of the stringency of purifying selection, the GC-conservative d*N*/d*S* could also suffer from some bias that would be dependent on GC* or on the GC content. This would create spurious correlations between GC-conservative d*N*/d*S* and GC*. It is not totally clear how a bias dependent on GC would specifically affect the GC-conservative d*N*/d*S*, however. At first sight, GC conservativeness should instead lead to a cancelling out of potential GC-compositional effects. In addition, the correlation of the GC-conservative d*N*/d*S* with life-history traits remains significant after controlling for variation in GC*. On the other hand, it is not easy to control for the GC content in the context of the present method, and thus, one cannot completely exclude the possibility of a compositional effect.

Altogether, the picture emerging from the correlation patterns between d*N*/d*S* and life-history traits and other molecular evolutionary quantities is complex. The many possible biases that could potentially affect d*N*/d*S*, and its various specializations, can certainly not be totally ruled out, even by carefully controlling for the relevant variables. Nevertheless, even if alternative interpretations can be entertained, an overall nearly-neutral interpretation, partially confounded by biased gene conversion, seems to be consistent with the observations gathered in the present analysis.

The Impact of Biased Gene Conversion on Protein-Coding Sequence Evolution

In contrast to the ratio of nonsynonymous over synonymous substitution rates, the equilibrium GC appears to display relatively robust and easily interpretable correlation patterns. The most important result, namely simultaneous correlations of GC* with body size and chromosome number, provides convincing empirical evidence in favor of a role of gBGC in explaining compositional variation among placental mammals. From a mechanistic perspective, these observations strongly suggest a role in genome-wide recombination rate and effective population size in determining the level of gBGC in placental genomes. On the other hand, the correlation with body size and chromosome number is relatively modest, with each variable explaining approximately 10% of the variation in GC* (table 1). Part of the residual variation could be caused by changes in the population genetics environment (in particular, by the level of inbreeding) or in the genetic system (mutation or repair bias).

A point that has not been considered in our analysis thus far is whether variation in the mutation bias, in addition to contributing to the residual variation in GC*, may also partly explain the correlation of GC* with body size and karyotype. Concerning karyotype, there is no known mechanism that would result in correlation between the direction of the mutation pressure and the number of chromosomes. In the case of body mass, on the other hand, a possibility would be that cytosine deamination, which is responsible for most of the AT bias (Lynch 2010b), may be less replication dependent than other types of mutation processes (Chen et al. 2010) and may therefore be proportionally more important in species with long generation times. This would result in a negative correlation of GC* with generation time that could indirectly

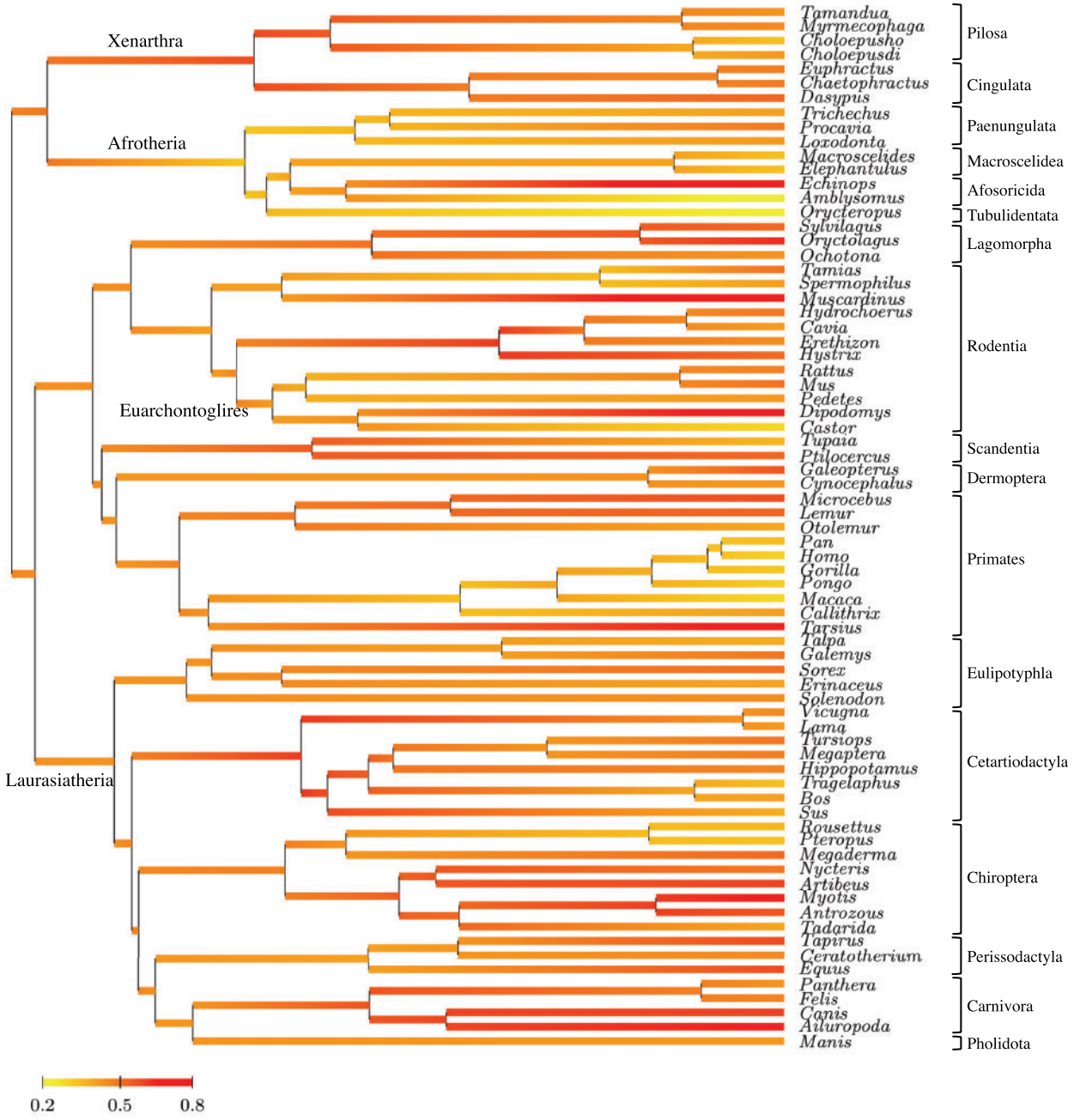

Fig. 2. Posterior mean reconstruction of GC* along the placental phylogeny. For credibility intervals, see supplementary figure S6, Supplementary Material online.

translate into a negative correlation of GC* with body mass. However, in the present case, no correlation is seen between GC* and maximum recorded life span or female age at maturity in spite of the fact that, these two variables are more sensitive to generation time effects than to body size. Alternatively, mutation patterns could be dependent on the metabolic rate (Martin and Palumbi 1993), such that variation in the metabolic rate, itself correlating negatively with body size, would result in a correlation of the AT bias with body mass. It is not clear, however, how much metabolic rate impacts the mutation processes acting on the nuclear genomes (Lanfear et al. 2010). Altogether, gBGC appears to be a more satisfactory explanation than variation in the mutation bias, as it can explain the correlation of GC* simultaneously with karyotype and body size.

In a previous investigation, and in contrast to the present analysis, no significant correlation was found between genomic GC content and number of chromosomes (Romiguier et al. 2010). There are two possible explanations for this difference. First, in the present case, we rely on a larger number of taxa (73 vs. 33). The power of phylogenetic covariance analyses is sensitive to taxonomic sampling, and therefore, previous negative results might simply reflect a lack of statistical power. Alternatively, as noted by Romiguier et al. (2010),

there is a conceptual difference between the GC content and the equilibrium GC*. The GC content integrates the effect of past variation in gBGC over a potentially long time scale. In contrast, the equilibrium GC is an indicator of the current intensity of gBGC.

On the other hand, because gBGC is directly proportional to the local recombination rate, the very large variance in the fine-scale recombination rate within genomes (Myers et al. 2005) should result in a strong heterogeneity across genes in the strength of gBGC. However, the present analysis does not model variation among genes. Importantly, we have checked that the correlation observed between GC* and body mass and number of chromosomes is robust to leave-one-out jackknife resampling of the genes (supplementary table S1, Supplementary Material online), thus excluding the possibility that the correlation patterns observed here would be dominated by some isolated large deviation in GC* affecting one particular gene. Ultimately, however, a proper analysis of the strength of gBGC in a comparative context should simultaneously account for variation in GC* across genomes and among lineages and rely on a large array of sequences at the genome-wide scale.

More fundamentally, the phylogenetic covariance analysis undertaken in this article was mostly empirical. In the long term, mechanistic developments will be of utmost importance to get a more detailed and quantitative view of the overall phenomenon. The interactions between biased gene conversion and selection might be difficult to model correctly. On the other hand, this interaction raises the interesting possibility that some of the parameters that were confounded under a classical nearly-neutral model, such as the shape parameter of the distribution of selective effects and the variation in effective population size, might be separately identifiable in the presence of gBGC.

Toward More Elaborate Codon Models?

In the long term, an important question is whether the model extensions proposed here (i.e., variation in GC* or d*N*/d*S* ratios restricted to particular types of nucleotide substitutions) could be incorporated in codon models in general. In particular, the idea of restricting the estimation of d*N*/d*S* on GC-conservative transversions could be used in the context of models aimed at detecting positive selection, in a site- or branch-specific manner (Yang 2002). Previous investigations have provided strong arguments suggesting that some of the genes displaying increased d*N*/d*S* in specific lineages in primates or mammals might in fact have undergone episodes of strong gBGC (Berglund et al. 2009; Galtier et al. 2009; Ratnakumar et al. 2010). These analyses allowed for an estimation of the global fraction of genes for which gBGC was a likely cause. Codon models based on GC-conservative substitutions, in contrast, would allow assessment of the role of gBGC on a gene-by-gene basis.

Proposing such generalizations of current codon models, however, raises the question of being able to routinely measure the relative fit of alternative parameterizations. Measuring model fit is still a delicate issue in Bayesian inference. In particular, numerical evaluation of Bayes factors is challenging (Lartillot and Philippe 2006; Xie et al. 2011) and would require much additional work in the present context to be done correctly. In the present case, qualitative insights about the relative merits of the models can be gained from indirect observations. For instance, the significant correlation between GC* and body mass and number of chromosomes is in itself an indication of the existence of a substantial variation in GC* among lineages, thus giving qualitative support in favor of models with time-dependent GC* over stationary models. Similarly, obtaining significant yet mutually incompatible correlation patterns for the various specializations of d*N*/d*S* or d*S* also implies that the model with one d*S* and one d*N*/d*S* should be rejected in favor of models allowing for subspecializations of d*S* and d*N*/d*S*. On the other hand, these observations do not imply that the more complex models specifically proposed here have a better overall fit than the simpler models. The reason is that such complex models, even if they detect meaningful patterns, may not rely on an overall adequate parameterization and may go too far in the direction of reducing bias and increasing variance. Additional work is needed to address these complex issues in a more systematic manner.

Ancestral Reconstruction

Arguments suggesting a fundamental role for nonadaptive evolutionary forces in genome evolution have stimulated a series of empirical investigations seeking to determine how changes in the width of the selective sieve might influence high-level features of genetic systems (Lynch 2010a), genomes (Lynch and Conery 2003; Lynch et al. 2006; Whitney et al. 2011), and proteomes (Fernández and Lynch 2011). In this perspective, phylogenetic tools for reconstructing large-scale variation in effective population size, and for performing correlations with other fundamental traits pertaining to life history or genome architecture, would represent an essential methodological progress for further empirical testing and theoretical elaboration. Following pioneering work (Nielsen and Yang 2003), the integrated probabilistic approach pursued in this study represents an encouraging step in this direction.

Reconstructing long-term variation in effective population size is certainly ambitious. Apart from the methodological problems potentially encountered, as mentioned earlier, it is not yet totally clear how much variation in d*N*/d*S* among lineages reflects mostly nearly-neutral effects, and how much it is confounded by adaptive phenomena. In addition, some clarification may be needed about what the concept of effective population size is supposed to mean at such a large evolutionary scale. In the short time scale (of the order of N_e generations, where N_e is the effective population size), the harmonic mean over short-term fluctuations is generally considered as the relevant population size. In the present case, however, time intervals between successive cladogenetic events are often much longer than the average time between successive speciations. Considering that speciations themselves probably represent nontrivial events in terms of changes in population size and structure, it is not clear whether the reconstructed d*N*/d*S* represents any meaningful

average, of whether it reflects the depth of the most serious bottlenecks experienced by the lineage, possibly, at the speciation events themselves.

The present analysis suggests that dN/dS might be sensitive to various distortions and artifacts. On the other hand, the reconstruction obtained here, based on GC-conservative substitutions (fig. 1 and supplementary fig. S5, Supplementary Material online), appears to correlate reasonably well with variation in life-history traits such as body size, suggesting that the reconstructed dN/dS might reflect meaningful long-term trends in effective population size or at least in life-history evolution. In addition, it is still possible to imagine methodological adaptations that would overcome the weaknesses uncovered in this study. For instance, the ratio of radical over conservative amino acid replacement (Sainudiin et al. 2005; Popadin et al. 2007) could be explored as an alternative to dN/dS. Intraspecific polymorphism could in principle be used to calibrate the reconstruction, either by estimates of population size at the leaves of the tree, or by providing an estimate of the distribution of fitness effects across the positions of the multiple alignment (Keightley and Eyre-Walker 2007, 2010; Boyko et al. 2008).

Our analysis also suggests that GC* could provide useful and more robust information about past molecular evolutionary regimes, compared with dN/dS. In particular, the coupling between GC* and effective population size created by gBGC can in principle be measured in neutral regions of the genome, thus offering the possibility of detecting correlates of effective population size that are less likely to be confounded by adaptive effects, compared with dN/dS. On the other hand, recombination and effective population size are confounded in GC*, and thus, ancestral reconstruction based on GC* might require further model elaboration, and additional empirical knowledge, to tease apart the respective contributions of recombination and population size. In principle, karyotype is reconstructable at the scale of placentals (Bourque et al. 2004; Murphy et al. 2005; Chauve and Tannier 2008; Zhao and Bourque 2009), providing a first entry point into ancestral reconstruction of recombination rates. However, other factors might significantly contribute to the variation in recombination rate, not to mention possible lineage-specific variation in the strength of the conversion bias. Altogether, it seems that reconstruction of variables such as dN/dS and its variants will be necessary to provide an independent variable, against which GC* and karyotype could be regressed, so as to better understand how recombination, effective population size, and other factors jointly contribute to genome evolution.

Supplementary Material

Supplementary table S1 and figures S1–S9 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The author thanks Frédéric Delsuc, Laurent Duret, Hervé Philippe, Konrad Scheffler and two anonymous reviewers for their useful comments on the manuscript. Computational resources were provided by Calcul Québec and Compute Canada and the Canadian Foundation for Innovation. This work was funded by the Natural Science and Engineering Research Council of Canada.

References

Arndt PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J Mol Evol.* 60: 748–763.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol.* 20:1887–1896.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol.* 7:e26.

Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* 14:507–516.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456: 942–945.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.

Capra JA, Pollard KS. 2011. Substitution patterns are GC-biased in divergent sequences across the metazoans. *Genome Biol Evol.* 3: 516–527.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc Natl Acad Sci U S A.* 104:16992–16997.

Chauve C, Tannier E. 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol.* 4:e1000234.

Chen CL, Rappailles A, Duquenne L, et al. (11 co-authors). 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* 20:447–457.

Clément Y, Arndt PF. 2011. Substitution patterns are under different influences in primates and rodents. *Genome Biol Evol.* 3: 236–245.

de Magalhaes J, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol.* 22:1770–1774.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res.* 17:1420–1430.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4:e1000071.

Duret L, Eyre-Walker A, Galtier N. 2006. A new perspective on isochore evolution. *Gene* 385:71–74.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10:285–311.

Duret L, Sémon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162:1837–1847.

Escobar JS, Glémin S, Galtier N. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Mol Biol Evol.* 28:2561–2575.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26:2097–2108.

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol.* 19:2142–2149.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat.* 125:1–15.

Fernández A, Lynch M. 2011. Non-adaptive origins of interactome complexity. *Nature* 474:502–505.

Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics* 172:221–228.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23:273–277.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25:1–5.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.* 15: 871–879.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159:907–911.

Garland T, Bennett AF, Rezende EL. 2005. Phylogenetic approaches in comparative physiology. *J Exp Biol.* 208:3015–3035.

Gillespie JH. 1991. The causes of molecular evolution. Oxford: Oxford University Press.

Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. *Genetics* 185: 939–959.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. *Nucleic Acids Res.* 35:D332–D338.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet.* 6:e1000825.

Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol.* 3: 614–626.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251–2261.

Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philos Trans R Soc Lond B Biol Sci.* 365:1187–1193.

Kimura M. 1979. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc Natl Acad Sci U S A.* 76: 3440–3444.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge (UK): Cambridge University Press.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4:e1000144.

Kreitman M. 1996. The neutral theory is dead. Long live the neutral theory. *Bioessays* 18:678–683.

Lanfear R, Welch JJ, Bromham L. 2010. Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol Evol.* 25:495–503.

Lartillot N, Delsuc F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution* 66:1773–1787.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol.* 55:195–207.

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol.* 28:729–744.

Li WH, Tanimura M. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* 326:93–96.

Lynch M. 2010a. Evolution of the mutation rate. *Trends Genet.* 26: 345–352.

Lynch M. 2010b. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A.* 107:961–968.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–1404.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science* 311: 1727–1730.

Madsen O, Scally M, Douady C, Kao D, DeBry R, Adkins R, Amrine H, Stanhope M, de Jong W, Springer M. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409: 610–614.

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis. San Diego (CA): Academic Press.

Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A.* 90: 4087–4091.

Martins E, Hansen T. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat.* 149: 646–667.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol.* 21:984–990.

Murphy WJ, Eizirik E, O'Brien SJ, et al. (11 co-authors). 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294:2348–2351.

Murphy WJ, Larkin DM, Everts-van der Wind A, et al. (25 co-authors). 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309:613–617.

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* 17:413–421.

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

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310:321–324.

Nabholz B, Glémin S, Galtier N. 2008. Strong variations of mitochondrial mutation rate across mammals—the longevity hypothesis. *Mol Biol Evol.* 25:120–130.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A.* 80:6278–6281.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol.* 20:1231–1239.

Nikolaev S, Montoya-Burgos J, Popadin K, Parand L, Margulies E, Program N, Antonarakis S. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A.* 104:20443–20448.

O'Brien SJ, Menninger JC, Nash WE. 2006. Atlas of mammalian chromosomes. New York: Wiley-Liss.

Ohta T. 1974. Mutational pressure as the main cause of molecular evolution and polymorphisms. *Nature* 252:351–354.

Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol.* 40: 56–63.

Paradis E, Claude J. 2002. Analysis of comparative data using generalized estimating equations. *J Theor Biol.* 218:175–185.

Pardo-Manuel de Villena F, Sapienza C. 2001. Recombination is proportional to the number of chromosome arms in mammals. *Mamm Genome.* 12:318–322.

Popadin K, Polishchuk L, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A.* 104:13390.

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol.* 56:453–466.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci.* 365:2571–2580.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20: 1001–1009.

Sainudiin R, Wong WSW, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Evol.* 60:315–326.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A.* 100:1056–1061.

Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. 2006. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol.* 23:565–573.

Webster M, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol.* 23: 1203–1216.

Webster MT, Smith NGC, Hultin-Rosenberg L, Arndt PF, Ellegren H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol Biol Evol.* 22: 1468–1474.

Welch JJ, Bininda-Emonds ORP, Bromham L. 2008. Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol Biol.* 8:53.

Welch JJ, Eyre-Walker A, Waxman D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J Mol Evol.* 67:418–426.

Whitney KD, Boussau B, Baack EJ, Garland T. 2011. Drift and genome complexity revisited. *PLoS Genet.* 7:e1002092.

Xie W, Lewis P, Fan Y, Kuo L, Chen MH. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol.* 60:150–160.

Yang Z. 2002. Inference of selection from multiple species alignments. *Curr Opin Genet Dev.* 12:688–694.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol.* 25:568–579.

Zhao H, Bourque G. 2009. Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* 19:934–942.