Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes

Nicolas Lartillot

MBE Advance Access published November 20, 2012

Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes

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, CNRS-Université de Montpellier 2, Montpellier, France

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

Associate editor: Rasmus Nielsen

Abstract

GC-biased gene conversion (gBGC) is a major evolutionary force shaping genomic nucleotide landscapes, distorting the estimation of the strength of selection, and having potentially deleterious effects on genome-wide fitness. Yet, a global quantitative picture, at large evolutionary scale, of the relative strength of gBGC compared with selection and random drift is still lacking. Furthermore, owing to its dependence on the local recombination rate, gBGC results in modulations of the substitution patterns along genomes and across time which, if correctly interpreted, may yield quantitative insights into the long-term evolutionary dynamics of recombination landscapes. Deriving a model of the substitution process at putatively neutral nucleotide positions from population-genetics arguments, and accounting for among-lineage and among-gene effects, we propose a reconstruction of the variation in gBGC intensity at the scale of placental mammals, and of its scaling with body-size and karyotypic traits. Our results are compatible with a simple population genetics model relating gBGC to effective population size and recombination rate. In addition, among-gene variation and phylogenetic patterns of exon-specific levels of gBGC reveal the presence of rugged recombination landscapes, and suggest that short-lived recombination hot-spots are a general feature of placentals. Across placental mammals, variation in gBGC strength spans two orders of magnitude, at its lowest in apes, strongest in lagomorphs, microbats or tenrecs, and near or above the nearly neutral threshold in most other lineages. Combined with among-gene variation, such high levels of biased gene conversion are likely to significantly impact midly selected positions, and to represent a substantial mutation load. Altogether, our analysis suggests a more important role of gBGC in placental genome evolution, compared with what could have been anticipated from studies conducted in anthropoid primates.

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

Introduction

GC-biased gene conversion (gBGC) is increasingly recognized as a major evolutionary force acting on genome evolution (Marais 2003; Galtier and Duret 2007; Berglund et al. 2009; Duret and Galtier 2009; Webster and Hurst 2012). During meiotic recombination, heteroduplex DNA created between homologous chromosomes contains mismatches at heterozygous sites, which might be occasionally repaired. If DNA repair pathways are GC-biased, conversion events will show a systematic preference for converting weak (A or T) alleles into strong (C or G) bases, causing a net over-transmission of GC alleles compared with their AT counterparts. Ultimately, this small asymmetry in the repair machinery creates a recombination-associated segregation and fixation bias, mimicking a genome-wide selective advantage in favor of GC (Nagylaki 1983). If sufficiently strong, biased gene conversion can in principle overwhelm selection, and lead to an increase in frequency, or even to the fixation, of deleterious GC alleles (Bengtsson 2009; Glémin 2010).

Although direct experimental evidence for gBGC is relatively scarce (Birdsell 2002; Mancera et al. 2008), gBGC appears to provide the best explanation for observed intra- and interspecific variation in GC composition in a wide range of organisms (Duret and Galtier 2009), compared with alternative interpretations in terms of mutation pressure or selection (Bernardi 2000; Fryxell and Zuckerkandl 2000).

Within genomes, large regional oscillations in GC content in amniotes (isochores, Bernardi et al. 1985; Bernardi 2000; Eyre-Walker and Hurst 2001) have stimulated much speculation as to the underlying causal mechanisms. Over the last 10 years, a large array of empirical observations in primates and birds have converged to a model conferring a prominent role to gBGC in shaping isochores (Galtier et al. 2001; Montoya-Burgos et al. 2003; Duret et al. 2006; Duret and Galtier 2009). One of the most important arguments in favor of this model is the correlation between local recombination rates and equilibrium GC (Galtier 2004; Meunier and Duret 2004; Duret and Arndt 2008), suggesting that the strong heterogeneity in GC content along genomes primarily reflects, via gBGC, the no less extensive variation in recombination rate.

Linking GC landscapes to recombination landscapes in turn raises the question of what determines the distribution

of recombination. Genome-wide recombination maps obtained using coalescent methods or high-resolution mapping of crossovers in humans have revealed a contrasted picture, with extensive fine-scale variation (hot-spots) superimposed on broad-scale trends in regional recombination rate (Jeffreys et al. 2001; Kong et al. 2002; McVean et al. 2004; Myers et al. 2005; Coop et al. 2008). Hot-spots do not show any correlation between humans and chimpanzees (Ptak et al. 2005; Winckler et al. 2005; Auton et al. 2012), and are variable among human individuals or populations (Coop and Przeworski 2007; Coop et al. 2008; Kong et al. 2010), indicating that they are short lived. In contrast, broad-scale variation in recombination rate is conserved between closely related species, and is linked to karyotypic structure, with a higher regional density of recombination prevailing near telomeres, compared with centromeres (Yu et al. 2001; Myers et al. 2005), and higher overall recombination rates in smaller than in larger chromosomes (Kaback 1996; Coop and Przeworski 2007). The correlation between recombination rate and chromosome size may have a simple mechanistic explanation, namely, that proper orientation and disjunction of homologous chromosomes during meiotic recombination requires at least one crossover event per chromosomal arm, or per chromosome (Dutrillaux 1986; Paliulis and Nicklas 2000; Pardo-Manuel de Villena and Sapienza 2001; Dumas and Britton-Davidian 2002; Fledel-Alon et al. 2009; Hassold et al. 2009). Based on these observations, a simple model of the evolutionary dynamics of recombination landscapes can be proposed, consisting of a fast turnover of recombination hot-spots, on a background of a slow-paced evolution of broad-scale patterns of recombination, themselves partly driven by karyotypic evolution.

In the presence of gBGC, the recombination landscapes thus characterized are mirrored by equivalent genome-wide substitution patterns. At the small scale, clusters of substitutions biased toward GC have been found (Dreszer et al. 2007) and have been interpreted as vestiges of past recombination hot-spots (Berglund et al. 2009; Katzman et al. 2011). Such clusters of GC-biased substitutions have been observed in several species across metazoans (Capra and Pollard 2011). In some cases, they are associated with acceleration of the evolutionary rate (Pollard et al. 2006; Katzman et al. 2010), or elevated d*N*/d*S* in protein-coding genes (Berglund et al. 2009; Galtier et al. 2009; Ratnakumar et al. 2010), suggesting that gBGC episodes caused by local and transient bursts of recombination might result in the fixation of tracks of potentially deleterious mutations. At a larger scale, equilibrium GC composition is higher, and clusters of GC-biased substitutions are more frequent, near telomeres than near centromeres (Arndt et al. 2005; Dreszer et al. 2007). In birds, where the karyotype has been stable over long evolutionary periods (Ellegren 2010), microchromosomes display a higher GC content (Hillier et al. 2004), and a higher substitution rate (Axelsson et al. 2005), compared with macrochromosomes. These correlations are predicted by a model combining gBGC and karyotype-dependent broad-scale variation in recombination rate. They also suggest a possible model for the origin of isochores, as reshuffled pieces of the ancestral karyotype (Duret et al. 2002, 2006).

Between species, variation in genome-wide composition may also be influenced by gBGC, with a higher GC content predicted in species experiencing a stronger gBGC. Among the determinants of between-species variation in gBGC intensity, recombination rate again, but also population size, stand out as two obvious candidates. Concerning recombination, genome-wide average recombination rate shows extensive variation among mammalian species (Coop and Przeworski 2007), determined to a large extent by the underlying karyotypic structure (Pardo-Manuel de Villena and Sapienza 2001). If the recombination rate is higher in smaller than in larger chromosomes within a given genome, for the same reason, it should be higher in species having globally more fragmented karyotypes, (Pardo-Manuel de Villena and Sapienza 2001; Dumas and Britton-Davidian 2002), or smaller genomes. These correlations should directly translate into higher equilibrium GC composition in species with more chromosomes or smaller genome size.

Concerning population size, in finite populations, random drift will interact with biased gene conversion to the same extent as it does with selection (Nagylaki 1983), resulting in a less pronounced bias toward GC in smaller populations. In mammals, effective population size is itself negatively correlated with body mass. This correlation has been used in empirical tests of the nearly neutral theory, where a positive correlation between the ratio of nonsynonymous over synonymous substitution rates (d*N*/d*S*) and body mass was observed, and interpreted in terms of an underlying negative correlation of d*N*/d*S* with population size (Eyre-Walker et al. 2002; Popadin et al. 2007; Nikolaev et al. 2007; Lartillot and Poujol 2011). In the present context, a negative correlation between population size and life-history would translate into a negative correlation between body-size and the population-level intensity of gBGC.

In a comparative analysis including all available complete mammalian genomes (Romiguier et al. 2010), a negative correlation was found between GC content and body mass, and between GC content and genome size (*C* value) in mammals. Similarly, the equilibrium GC was found to be negatively correlated with body mass, and positively correlated with the number of chromosomes, in placental mammals (Lartillot 2012). Such correlations are most easily interpreted in terms of an indirect influence of effective population size and genome-wide recombination rate on equilibrium nucleotide composition, via gBGC, thus providing a global validation of the predictions of the gBGC model at the scale of mammals.

On the other hand, in part because they were meant as a first global and a priori agnostic evaluation of the correlates of the evolutionary dynamics of GC landscape in mammals, these two recent comparative analyses (Romiguier et al. 2010; Lartillot 2012) were mostly phenomenological and qualitative. Now that a working model linking recombination and nucleotide landscapes via the population genetics of biased gene conversion has found some empirical support, developing mechanistic models of the substitution process

explicitly defined in terms of mutation rates and fixation biases would make it possible to leverage a more quantitative estimation of the global intensity and distribution of gBGC among lineages and along genomes.

In this direction, we now introduce an integrated Bayesian model for reconstructing the evolutionary history of gBGC, and for estimating the correlation of gBGC with life-history and karyotypic traits. The framework uses mechanistic arguments based on population genetics theory to make a quantitative connection between mutation, gBGC, and the substitution process at putatively neutral sites (Lipatov et al. 2006; Duret and Arndt 2008; Harrison and Charlesworth 2011). Substitution patterns are thereby made explicitly dependent on the parameters of the population genetics environment, more specifically, on the scaled conversion coefficient $B = 4Nb$, where N is the effective population size, and b is the conversion bias. Variation in B among lineages is then modeled by a Brownian process covarying with karyotypic formula and life-history traits (Lartillot and Poujol 2011), thus indirectly capturing correlations of gBGC with global recombination rate and effective population size. Simultaneously, B is modulated across loci, so as to capture variation within genomes. Fitting the model onto empirical data using a Markov chain Monte Carlo algorithm, we find globally high levels of gBGC across placentals, modulated by karyotype and population size, and uncover a general pattern of rugged and rapidly evolving recombination landscapes across the entire group.

Materials and Methods

Mutation-Conversion Model

From a population-genetics perspective, gBGC is formally equivalent to a selective advantage of strong (GC) alleles over weak (AT) alleles (Nagylaki 1983). Accordingly, the average strength of gBGC can be captured through a pseudoselection coefficient b, equal to the net probability of biased conversion per position and per generation. Since biased gene conversion is intimately connected to recombination, this selection coefficient can be assumed to be proportional to the recombination rate: $b = b_0 r$. If N is the population size, in the absence of selection, and defining the scaled coefficient $B = 4Nb$, the relative (i.e., relative to neutral) fixation probability of a strong allele S against a weak background W (P_{WS}) can be approximated by:

$$2NP_{WS} = \frac{B}{1 - e^{-B}},$$

and conversely,

$$2NP_{SW} = \frac{-B}{1 - e^{B}}$$

for the relative fixation probability a weak variant in a strong background (P_{SW}). Weak to weak or strong to strong substitutions behave neutrally ($2NP_{WW} = 2NP_{SS} = 1$).

Based on these mechanistic insights, we can express substitution rates as the product of mutation rates and relative fixation probabilities (Lipatov et al. 2006; Duret and Arndt 2008; Harrison and Charlesworth 2011), leading to the following instant rate matrix for the nucleotide substitution process:

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & \mu_{AC}\frac{B}{1-e^{-B}} & \mu_{AG}\frac{B}{1-e^{-B}} & \mu_{AT} \\ C & \mu_{CA}\frac{-B}{1-e^{B}} & - & \mu_{CG} & \mu_{CT}\frac{-B}{1-e^{B}} \\ G & \mu_{GA}\frac{-B}{1-e^{B}} & \mu_{GC} & - & \mu_{GT}\frac{-B}{1-e^{B}} \\ T & \mu_{TA} & \mu_{TC}\frac{B}{1-e^{-B}} & \mu_{TG}\frac{B}{1-e^{-B}} & - \end{array}.$$

The mutation rates μ_{XY} are assumed to be strand symmetric and site independent. An implicit assumption behind the mutation-selection derivation is a low-mutation approximation, such that there is no significant interference between mutants cosegregating in the population (Sella and Hirsh 2005). This is a reasonable approximation given that the per base scaled mutation rate $4N\mu$ is small, and that the recombination in mammalian nuclear genomes effectively uncouples fixation events at neighboring loci. With the constraint that $\mu_{A:T>T:A} = 1$, the mutation process represents a set of five free parameters. For some analyses, the mutation rates are fixed to the empirical values reported in Lynch (2010), or to the empirical estimates of the substitution rates in weakly recombining regions (data used in Duret and Arndt 2008, kindly provided by Laurent Duret), in both cases renormalized so that $\mu_{A:T>T:A} = 1$. These empirical estimates are reported in table 2 for reference.

Structure of the Estimation Model

The substitution process thus defined is then modulated among lineages and across loci, under the following assumptions: 1) the mutation spectrum (relative mutation rates between bases) is constant across placentals, and uniform along the genome, 2) the absolute mutation rate is variable among lineages, but uniform along the genome, and 3) the biased-conversion strength B is variable both among lineages and between loci. Although assumption 3 is reasonable, and is supported by current empirical knowledge, assumptions 1 and 2 certainly deserve further scrutiny (see discussion).

In the following, all diffusion and substitution processes are defined as a function of the evolutionary time t running along the branches of a time-calibrated phylogeny, with the implicit convention that the processes split into two independent subprocesses at each cladogenetic event (node) of the phylogenetic tree. Note that the divergence times are not fixed a priori, but are co-estimated with all other parameters of the model (as in Lartillot and Delsuc 2012).

The absolute mutation rate $r(t)$ is modeled as a log-normal Brownian diffusion process of variance parameter ν, as in relaxed molecular clock models for estimating divergence times (Thorne et al. 1998; Rannala and Yang 2007). Concerning the genome-wide average conversion strength $B(t)$, assuming a fixed number of recombination events per chromosome and per generation (Dumas and

Britton-Davidian 2002), the average recombination rate for a given genome is roughly proportional to n/C, where n is the number of chromosomes, and C is the size of the genome (the C value being used here as a proxy). Alternatively, assuming a fixed number of recombination events per chromosomal arm and per generation (Dutrillaux 1986; Pardo-Manuel de Villena and Sapienza 2001) leads to a proportionality with f/C, where f is the fundamental number (i.e., the number of arms). Like most life-history traits, population size can be assumed to scale allometrically with body mass: $N \sim M^{\gamma_M}$, where $\gamma_M < 0$ (population size decreases when mass increases). Combining these equations leads to the following scaling expectation:

$$B \sim \frac{M^{\gamma_M} n}{C}.$$

This suggests to model the variations of B as a Brownian log-normal diffusion process correlated with M, C, and n:

$$\mathrm{d}\ln B(t) = \gamma_M \mathrm{d}\ln M(t) + \gamma_C \mathrm{d}\ln C(t) + \gamma_n \mathrm{d}\ln n(t) + \sigma \mathrm{d}W(t), \quad (1)$$

with $M(t)$, $C(t)$, and $n(t)$ themselves modeled as a trivariate log-normal Brownian diffusion process, of covariance matrix Σ, and $W(t)$ being a standard Brownian motion (a similar derivation can be made with the fundamental number $f(t)$ instead of the number of chromosomes). Here, we have chosen to leave the allometric coefficients unconstrained, and to estimate them from empirical data. The resulting estimates should then be consistent with the following relations: $\gamma_M < 0$, $\gamma_C = -1$, and $\gamma_n = 1$. The covariance between M, C, and n, represented by the symmetric matrix Σ, is also estimated from the data.

The residual variation in B embodied by $\sigma W(t)$ accounts for at least three distinct factors: residual variation in population size, given body mass, residual variation in recombination rate, given the number of chromosomes, and possibly also variation in the intrinsic strength of gBGC (b_0) among lineages. Without any direct information about effective population size and recombination rate, it is not possible to disentangle these contributions.

Modulation across loci of the scaled conversion coefficient is modeled by the use of i.i.d. branch- and locus-specific gamma-distributed modulators, acting multiplicatively on the scaled selection coefficient B. Gamma-distributed locus-specific multiplicative offsets are also defined, to account for the fact that some genes might be selectively, or at any rate permanently, maintained in highly or weakly recombining environments. More specifically, for locus $i = 1, \ldots, L$, and branch $j = 1, \ldots, 2P - 2$, where L is the number of independent loci (exons) and P is the number of taxa, multipliers η_{ij} are introduced, distributed as:

$$\eta_{ij} \sim \mathrm{Gamma}(\alpha_m, \alpha_m),$$

where α_m is the shape parameter, tuning the variance of the strength of gBGC across loci. In addition, locus-specific multiplicative offsets β_i are defined, for $i = 1, \ldots, L$, which are distributed according to another gamma distribution:

$$\beta_i \sim \mathrm{Gamma}(\alpha_o, \alpha_o),$$

of shape parameter α_o. The strength of gBGC acting on branch j, and for gene i is then equal to:

$$B_{ji} = \beta_i \eta_{ij} \bar{B}_j,$$

where $\bar{B}_j$ is the global strength of gBGC on branch j, such as defined by the Brownian multivariate process introduced above (eq. 1). In practice, $\bar{B}_j$ is approximated by the arithmetic average of the instant values of the process at both ends of the branch (Lartillot and Poujol 2011).

Given the assumption of strand-symmetry, the nucleotide frequency distribution for locus i at the root of the tree has only one degree of freedom, the frequency of $G + C$ (denoted below as π_i). For convenience, this distribution is parameterized in terms of an apparent B_i^{root} parameter, such that

$$\pi_i = \frac{e^{B_i^{\mathrm{root}}}}{1 + e^{B_i^{\mathrm{root}}}}. \quad (2)$$

In practice, the apparent scaled conversion parameters $(B_i^{\mathrm{root}})_{i=1\ldots L}$ integrate the effect of the mutation bias, as well as the complex history of the variation of gBGC before the last common ancestor of placentals. As such, they should not be interpreted as bona fide estimates of the strength of gBGC in the lineage leading to this ancestor, but should be understood as phenomenological parameters whose presence is necessary for correctly adjusting the model to the empirical data. These locus-specific parameters are drawn from a gamma distribution of shape parameters α_r.

Priors

The following priors were used: on the scaling coefficients γ_M, γ_C, γ_n, a normal distribution of mean 0 and variance 1; on the covariance matrix Σ, an inverse Wishart prior of parameter Σ_0 and with three degrees of freedom, where $\Sigma_0 = \mathrm{Diag}(\kappa_1, \kappa_2, \kappa_3)$ is a diagonal matrix; on $\ln B(0)$, $\ln M(0)$, $\ln C(0)$, and $\ln n(0)$, the values of the scaled conversion coefficient and the quantitative traits at the root of the tree, independent normals of mean 0 and variance 10. A truncated log-uniform prior, of support $[10^{-3}, 10^3]$, was imposed on κ_1, κ_2, and κ_3, the shape parameters α_m, α_o, α_r, the mutation rates μ_{XY} and the parameter σ.

On divergence times, we imposed a uniform prior on relative ages (relative to the age of the root). The prior on the absolute age of the root T_0 is an exponential of mean 150 Myr. Concerning the relaxed molecular clock, we put a uniform prior in $[-10, 10]$ on $\ln r(0)$, and a truncated log-uniform prior on ν.

MCMC Sampling and Posterior Mean Estimates

The MCMC sampler was run for a total of 330,000 cycles, performing a complex series of MCMC updates of the variables of the model for each cycle. One point was saved every 30 cycles. A burnin of 1,000 points was discarded, and

posterior expectations were estimated on the remaining 10,000 points. In the case of simulations, the sampler was run for 75,000 cycles, saving every 30 cycles, discarding 500 points and computing expectations on the remaining 2,000 points. The sampler relies on data augmentation procedures (Lartillot 2006; Mateiu and Rannala 2006; Lartillot and Poujol 2011). Specifically, detailed substitution mappings are sampled conditional on the data and the current parameter values. All Metropolis-Hastings updates are then performed based on a likelihood evaluated conditional on the current data augmentation. The substitution mapping is resampled each time before starting a new cycle. Two independent runs were performed under each settings. Convergence and mixing of the MCMC were first assessed visually, and then quantified using convergence diagnostics based on empirical autocorrelations (Lartillot et al. 2009).

Correlation coefficients between B and body-size ($k = 1$), C value ($k = 2$), and number of chromosomes ($k = 3$) are given by:

$$r_k = \frac{\sum_{l=1}^{3} \Sigma_{kl}\gamma_l}{\sqrt{\Sigma_{kk}\nu_B}},$$

where

$$\nu_B = \sigma^2 + \sum_{l=1}^{3}\sum_{m=1}^{3} \Sigma_{lm}\gamma_l\gamma_m.$$

The coefficient r_k is averaged over the samples obtained from the MCMC sampler.

Data

The taxon-rich dataset, introduced in Lartillot and Delsuc (2012), gathers sequences of 17 single-exon nuclear genes from 73 placental taxa. To build the exon-rich datasets, all single-exon multiple sequence alignments of the Orthomam version 6 database (Ranwez et al. 2007) were downloaded, and filtered from their ambiguously aligned codon positions using Gblocks (Castresana 2000), with the default options of the program. Alignments were then selected based on the criterion that at least 30 (resp. 25) out of the 33 placental taxa should be represented in the alignment, which resulted in a subset of 180 (resp. 1,874) exons.

For each taxon and exon sampling, two datasets were constructed, by selecting either all of the 4-fold degenerate third positions, or only the 4-fold degenerate third positions neither preceded by a C nor followed by a G in at least one taxon. In the CpG-filtered case, exons with less than 5 positions (10 in the 1,874 exon dataset) were eliminated. This resulted in a total of 16 exons out of 17 for the taxon-rich dataset, 167 out of 180 exons and 1,329 out of 1,874 exons for the exon-rich datasets. From the 1,329 exons, 30 jackknife replicates, of 100 exons each, were randomly sampled.

Estimates of adult body mass were obtained from the AnAge database (de Magalhaes and Costa 2009), and compiled as described in Lartillot and Delsuc (2012) in the case of the taxon-rich datasets. Karyotypic data and C-values were obtained from Romiguier et al. (2010), the Animal Genome Size Database (Gregory et al. 2007), and from the Atlas of Mammalian Chromosomes (O'Brien et al. 2006).

Results

Validation by Simulations

Simulation experiments were conducted, assuming a nonreversible mutation process with context-dependent CpG hypermutation (Arndt and Hwa 2005), a Brownian evolution of genome-wide strength of gBGC along a tree of 33 species (based on the phylogeny of the exon-rich dataset), modulated across 100 loci (exons) by a superposition of two discrete jump processes representing slow karyotypic evolution combined with fast hot-spot turnover. Mutation rates were as in Lynch (2010), except for the CpG deamination rate, which was set equal to either 15 times or 50 times the transversion rate between A and T (Arndt et al. 2003). Current empirical knowledge of recombination rates in humans and chimps suggest that between 50% of the total recombination occurs in less than 10% of the genome, and that hotspots occur on average every 200 kb or less in the human genome (McVean et al. 2004). Accordingly, we tested parameter configurations such that 1% to 10% of the genome would be in hot-spots at any time, with recombination rates in hotspots ranging from 1 to 100 times the background recombination rate elsewhere in the genome, and a hot-spot mean duration between 0.3 and 3 Myr (see supplementary methods, Supplementary Material online).

The reconstructed phylogenetic history of the scaled conversion coefficient $B = 4Nb$ (where N is the effective population size, and b is the conversion bias) appears to be qualitatively reliable over a wide range of conditions (supplementary table S1 and fig. S1, Supplementary Material online), even assuming strong CpG effects and a large variance induced by hot-spot turnover. The reconstructed history of B is relatively insensitive to the specific assumptions of the underlying site-independent mutation model (mutation rates known a priori or inferred from the data) and appear to be more reliable when excluding positions in a CpG context: on average across simulations and across all nodes along the phylogeny, the 95% credibility intervals contain the true (simulated) values in 53% of the cases when positions in a CpG context are included, and 80% when CpG positions are excluded (supplementary table S1, Supplementary Material online). In several instances, large deviations due to particularly strong gBGC episodes at one locus in one lineage appear to result in locally distorted genome-wide estimates of B (supplementary fig. S1, Supplementary Material online). This observation suggests that jackknife procedures should be used to average out the impact of outliers.

Mutation rates, on the other hand, appear to be more difficult to estimate than the scaled conversion coefficient, with credibility intervals encompassing the true value in only 30% of the cases (supplementary tables S2 and S3, Supplementary Material online).

Phylogenetic Covariance

The model was applied to the reconstruction of the history of biased gene conversion at the scale of placentals. Two datasets were analyzed, representing complementary solutions to the problem of finding a good tradeoff between gene and taxon sampling: a taxon-rich data set, made of 17 single-exon genes in 73 taxa, and a exon-rich dataset, made of 180 exons in 33 species. In addition, a set of 1,876 exons was considered, from which a set of 30 jackknife replicates of 100 exons were derived and analyzed.

On the taxon-rich dataset, the mechanistic model provides a significant support (posterior probability greater than 0.95) for a negative correlation of gBGC with body mass (table 1). The posterior mean estimate of the scaling coefficient, $\gamma_M \simeq -0.12$, suggests a relatively weak dependence. Concerning karyotype, no significant correlation is seen between gBGC strength and C value (table 1). In contrast, a significant correlation is found between gBGC and number of chromosomes ($P > 0.99$), interestingly, with a slope γ_n of the order of 1 (table 1). A correlation is also found between gBGC and the number of chromosomal arms, although less strong, and only marginally significant ($P = 0.92$), suggesting that, at the level of placentals, the number of chromosomes is a better predictor of the global recombination rate, or at least of the global gene conversion rate, than the number of chromosomal arms. All these correlations were robust to leave-one-out jackknife resampling of the genes (supplementary table S4, Supplementary Material online), suggesting that the correlations obtained here are not due to some extreme substitution behavior displayed by a single outlying exon.

The exon-rich dataset also gives a significant negative relation between gBGC and body mass (slope $\gamma_M \simeq -0.17$, table 1), but no significant correlation with karyotype. The lack of correlation with karyotype in the case of the exon-rich dataset is likely due to the poor taxon-sampling afforded by currently available complete genomes. Reducing the taxon-rich dataset to the same taxon sampling as the exon-rich dataset leads to a loss of significance for both correlations, either with body mass or with karyotype (data not shown). The lack of significance for the relation between C value and B, even under the taxon-rich dataset, could be due to the fact that the C value is a poor proxy of genome size. Note that, although a correlation between C value and GC content was found previously in mammals (Romiguier et al. 2010), significance was marginal when the analysis was restricted to placentals.

Taken together, the scaling coefficients estimated by our mechanistic phylogenetic covariance model are compatible with a simple model in which the average strength of gBGC in a genome is proportional to the product of population size and genome-wide recombination rate, itself roughly proportional to the number of chromosomes. The correlation, on the other hand, is relatively weak (with no more than 20% of the variance explained by body mass and karyotype, table 1), presumably because population size is only partially explained by body mass, and genome-wide recombination rate is only partially predicted by the karyotypic structure. Other factors may also be involved in the residual variation, such as varying levels of inbreeding between species (Glémin 2010), or species-specific molecular mechanisms involved in the regulation of recombination and gene conversion.

Variation among Loci

The significant correlation between the genome-wide level of gBGC and the number of chromosomes is consistent with a role of karyotype in determining broad-scale recombination rates. The distribution of gBGC across exons, on the other hand, and the modulation of such exon-specific effects along the phylogeny, might represent an interesting and complementary source of information, reflecting more local patterns of recombination rates, such as recombination hot spots. In the present context, a regime of active but short-lived hot spots would translate into a large variance between exons, combined with an absence of correlation of the distribution of gBGC across exons between closely related branches of the phylogenetic tree.

Concerning the variance between exons, and assuming a simple proportionality relation between gBGC and the local recombination rate, the relative standard deviation (RSD or coefficient of variation) of gBGC between exons should reflect the RSD of the fine-scale variation in recombination rate, and can therefore be directly compared with current quantitative estimates obtained by coalescent methods. In humans, broad scale variation, including karyotypic effects, accounts for a low RSD ($\sim$0.5). In contrast, fine-scale variation in recombination rate represents a RSD of about 2 (Myers et al. 2005).

In the present case, the coefficient of variation of gBGC across genes for the placental exon-rich dataset is as high as 2.3 (2.2 excluding positions potentially in a CpG context.) Fitting the model separately on subclades reveals the presence of some heterogeneity, with a larger RSD in Euarchonta (2.6) and Laurasiatheria (3.0) than in Glires (1.6) or in Atlantogenata (1.6). Since the RSD is computed based on branchwise averages of B, part of the differences between the four subclades could be due to differential taxon sampling (with poor sampling in particular in Atlantogenata), or to differing rates of evolution of recombination landscapes. In all cases, however, when compared with estimates obtained in humans, the RSD measured here is too large to be explained only by karyotypic effects, and is instead of the order of the fine-scale variation observed in humans and chimps, thus representing indirect evidence in favor of a general hot spot regime across placentals.

Concerning the rate of turnover, visual inspection of the exon-specific reconstructions of the variation in B along the phylogeny (supplementary fig. S2, Supplementary Material online) clearly suggests a pattern of rare and short-lived episodes of strong BGC at the level of single exons. As a way of obtaining a rough measure of the regime followed by the turnover process, we defined, for a given exon, a hot branch as being such that the local value of B is at least five times higher than the average value of B over the tree for that exon. We then estimated that exons have on average between 1 and 2 hot branches (with less than 15% of the

Table 1. Bayesian Estimates (posterior mean, 0.95 credibility intervals, and correlation) of Scaling Coefficients.

Model	Body Mass		*C* value		No. of Chromosomes		No. of arms	
	γ_M	r_M	γ_C	r_C	γ_n	r_n	γ_a	r_a
txCpG$^+$	−0.12** (−0.21, −0.03)	−0.32	−0.55 (−1.69, 0.64)	−0.16	1.28** (0.48, 2.10)	0.33	0.98 (−0.13, 2.07)	0.11
txCpG$^-$	−0.05* (−0.11, −0.00)	−0.30	−0.35 (−1.15, 0.48)	−0.18	0.67** (0.18, 1.20)	0.38	0.61 (−0.08, 1.32)	0.23
exCpG$^+$	−0.17** (−0.29, −0.05)	−0.49	−0.15 (−1.53, 1.22)	−0.09	0.06 (−0.85, 1.00)	−0.12	−0.31 (−1.40, 0.80)	−0.21
exCpG$^-$	−0.17** (−0.31, −0.03)	−0.43	−0.08 (−1.47, 1.38)	−0.06	0.08 (−0.98, 1.19)	−0.09	−0.25 (−1.48, 1.04)	−0.17

NOTE.—tx, taxon-rich data set; ex, exon-rich data set; CpG$^+$, all 4-fold degenerate positions; CpG$^-$, 4-fold degenerate positions not in a CpG context.
*Posterior probability of a positive regression coefficient >0.95 or <0.05.
**Posterior probability of a positive regression coefficient >0.975 or <0.025.

Table 2. Bayesian Estimates (posterior mean and 0.95 credibility intervals) of Mutation (mut.) and Substitution (subst.) Rates.

Model	A:T > G:C	G:C > A:T	G:C > T:A	A:T > C:G	G:C > C:G
Dir. est.[a]	3.33	7.45	2.00	1.18	2.30
Dir. est.[b]	3.99	5.75	2.00	1.12	1.77
txCpG$^+$ mut.	2.91 (2.64, 3.19)	4.45 (3.99, 4.94)	0.91 (0.80, 1.06)	1.17 (1.05, 1.31)	0.96 (0.86, 1.06)
txCpG$^-$ mut.	3.74 (2.42, 5.48)	7.98 (4.95, 12.71)	1.62 (0.95, 2.67)	1.53 (0.95, 2.31)	1.60 (1.11, 2.24)
exCpG$^+$ mut.	2.34 (2.27, 2.44)	4.86 (4.67, 5.05)	0.93 (0.88, 0.99)	1.02 (0.98, 1.07)	1.02 (0.98, 1.06)
exCpG$^+$ subst.	4.11 (3.96, 4.27)	3.62 (3.48, 3.76)	0.70 (0.66, 0.74)	1.86 (1.79, 1.95)	1.26 (1.21, 1.31)
exCpG$^-$ mut.	5.05 (4.43, 5.75)	6.96 (6.14, 7.88)	1.47 (1.26, 1.70)	2.35 (2.05, 2.70)	2.54 (2.25, 2.86)
exCpG$^-$ subst.	8.58 (7.67, 9.63)	4.90 (4.38, 5.51)	1.04 (0.91, 1.19)	4.00 (3.49, 4.53)	2.64 (2.34, 2.98)

NOTE.—All rates are relative to A:T > T:A. tx, taxon-rich data set; ex, exon-rich data set; CpG$^+$, all 4-fold degenerate positions; CpG$^-$, 4-fold degenerate positions not in a CpG context.
[a]Direct estimate from Lynch (2010).
[b]Estimate from Duret and Arndt (2008).

exons have more than two hot branches), indicating that strong gBGC episodes are rare. In addition, focusing on exons with exactly two hot branches, we observe that, in most cases (≥95%), hot branches are not successive along the phylogeny, suggesting that hot episodes are short-lived. Given the low time resolution afforded by the present taxonomic sampling, this would mean that hot spots typically last for less than a few million years.

Altogether, our observations therefore suggest that the pattern of strong and short-lived hotspots previously observed in humans and chimpanzees is a general feature of placentals.

Estimated Mutation Rates

When compared with independent estimates of the mutation spectrum in humans, obtained either from a compilation of reported human genetic variants (Lynch 2010), or based on patterns of point substitutions between humans, chimps and orangutans in low recombining regions (Duret and Arndt 2008), the mutation parameters estimated using the present method (table 2) are qualitatively reasonable at the level of transitions. Thus, according to the model, transitions display a clear bias toward AT at the level of the mutation process, which is in contrast to the raw substitution process, for which transition rates tend to be biased in favor of GC (in particular when CpG positions are excluded, table 2). Such a discrepancy between mutation and substitution patterns is an expected consequence of mild to strong levels of gBGC, and appears to be interpreted as such by the model.

More surprisingly, however, non-GC-conservative transversions are estimated to be GC-biased, whereas direct empirical estimates suggest that transversion mutations, like transitions, are AT-biased (table 2). The reason for this qualitatively incorrect estimation is unclear. Note that this is not observed in simulation experiments, despite the relatively large error in the estimated mutation rates also observed in this latter context (supplementary tables S2 and S3, Supplementary Material online), suggesting more fundamental violations of some of the model's assumptions by empirical data. Excluding CpG positions from the analysis results in relative mutation rates that are in closer agreement with direct estimates, although this does not appear to improve the estimation of the non-GC-conservative transversion rates.

The Macro-Evolutionary History of gBGC in Placental Mammals

The marginal reconstruction of the genome-wide intensity of gBGC along the phylogeny is displayed in figure 1 in the case of the exon-rich dataset, and in figure 2 in the case of the taxon-rich dataset. Globally, biased gene conversion varies over approximately a 50-fold range across placentals, going from weak levels to values largely above the nearly-neutral threshold. The intensity of gBGC is weakest in hominoids ($B \sim 0.1$), weak in the elephant and the sloth, moderate ($B \sim 1$), everywhere else, and strong ($B \sim 3$–5) in the bat *Myotis*, the shrew *Sorex*, the tenrec *Echinops*, and the lagomorphs *Oryctolagus* and *Ochotona* (fig. 1).

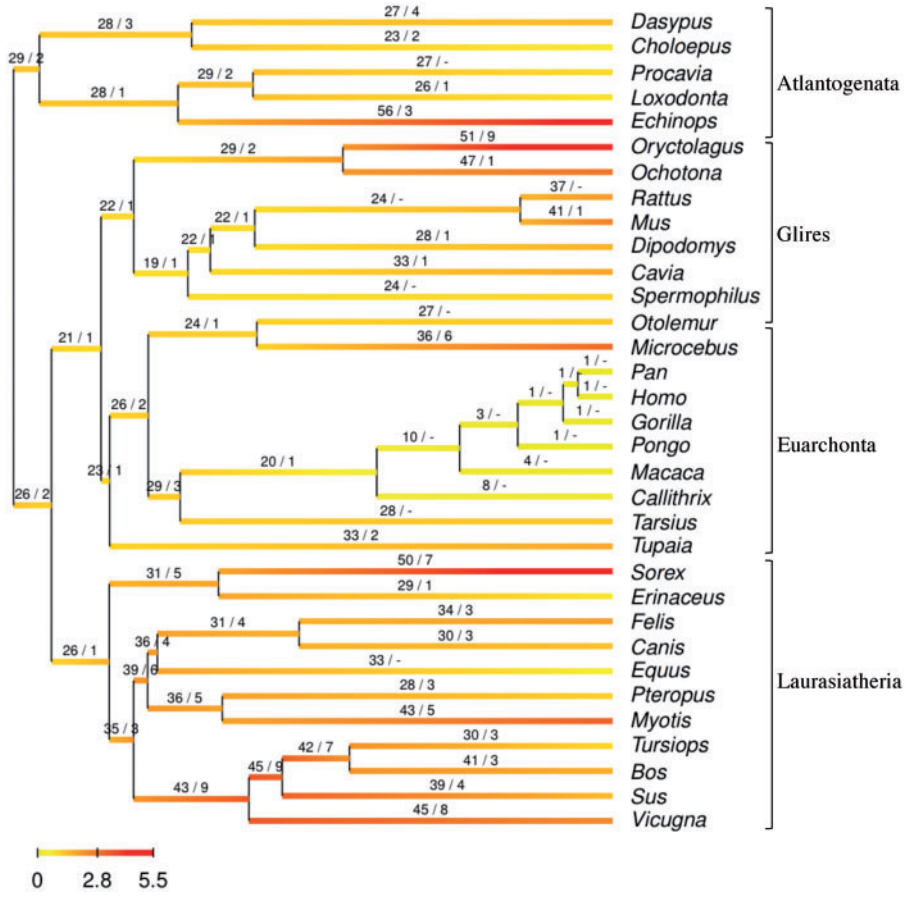

Fig. 1. Reconstructed (posterior mean) phylogenetic history of $B = 4Nb$ on the exon-rich dataset. The two numbers above each branch indicate the percentage of exons under $B > 1$ and $B > 10$ on average along the branch. Dashes indicate proportions smaller than 1%.

The reconstruction is robust, whether CpG positions are included in the analysis, and whether the mutation rates are estimated from the data, or are constrained to empirical values obtained from independent sources (supplementary fig. S3, Supplementary Material online). Reconstructing the history of gBGC separately on each subclade, Euarchonta, Glires, Laurasia, and Atlantogenata, thereby separately estimating the mutation and variance parameters on each subgroup, leads to very similar results (supplementary fig. S4, Supplementary Material online), compared with those obtained at the scale of Placentalia.

The reconstruction on the taxon-rich dataset (fig. 2 and supplementary fig. S5, Supplementary Material online) is globally congruent with that obtained under the exon-rich dataset, except for several localized differences, in particular for dogs, lemurs, and the tarsier, which are inferred to have a stronger genome-wide gBGC under the taxon-rich than under the exon-rich dataset. Exon-jackknife resampling on both data sets leads to remarkably stable reconstructions (supplementary figs. S6 and S7, Supplementary Material online).

The correlation between gBGC and body-size is apparent from the fact that, among extant lineages, the highest levels of gBGC are seen only in small mammals (figs. 1 and 2). The contrast between closely related taxa also suggests a body-size effect, with larger species often having a weaker gBGC than their smaller cousins, for example, micro- versus macrobats, or the extreme case of the tenrec versus the elephant. On the other hand, lagomorphs are globally larger than rodents, yet have a stronger gBGC. Karyotypic effects are less clear, except for a slightly stronger gBGC in the lama *Vicugna* ($2n = 74$) and the cow *Bos* ($2n = 60$), than in the dolphin *Tursiops* (44) and the pig *Sus* (38), and similarly, a stronger conversion bias in the tarsier *Tarsius* ($2n = 80$), the lemurs *Microcebus* ($2n = 62$), and *Otolemur* ($2n = 66$) than in the marmoset *Callithrix* ($2n = 44$, figs. 1 and 2).

For most of the tree, the intensity of gBGC is around or above $4Nb = 1$. This, however, represents the genome-wide average. Given the substantial variance in B among exons, even such relatively mild values of B imply that a significant fraction of the exons are under strong gBGC. This point finds confirmation upon estimating the fraction of the exons inferred under mild ($1 < B < 5$), or strong ($B > 10$) gBGC on each branch (fig. 1). Hominoids (apes) are the only group for which most ($\geq$99%) of the 180 exons are inferred under weak gBGC ($B < 1$). All other placental mammals have more than 5% of their exons under mild or strong gBGC, and in some lineages, in particular in Laurasiatheria and Glires, this proportion can represent more than half of the exons. Exons under strong gBGC ($B > 10$) represent more than 1% of the genes in approximately half of the lineages, and between 5% and 10% of the exons in lineages displaying high levels of

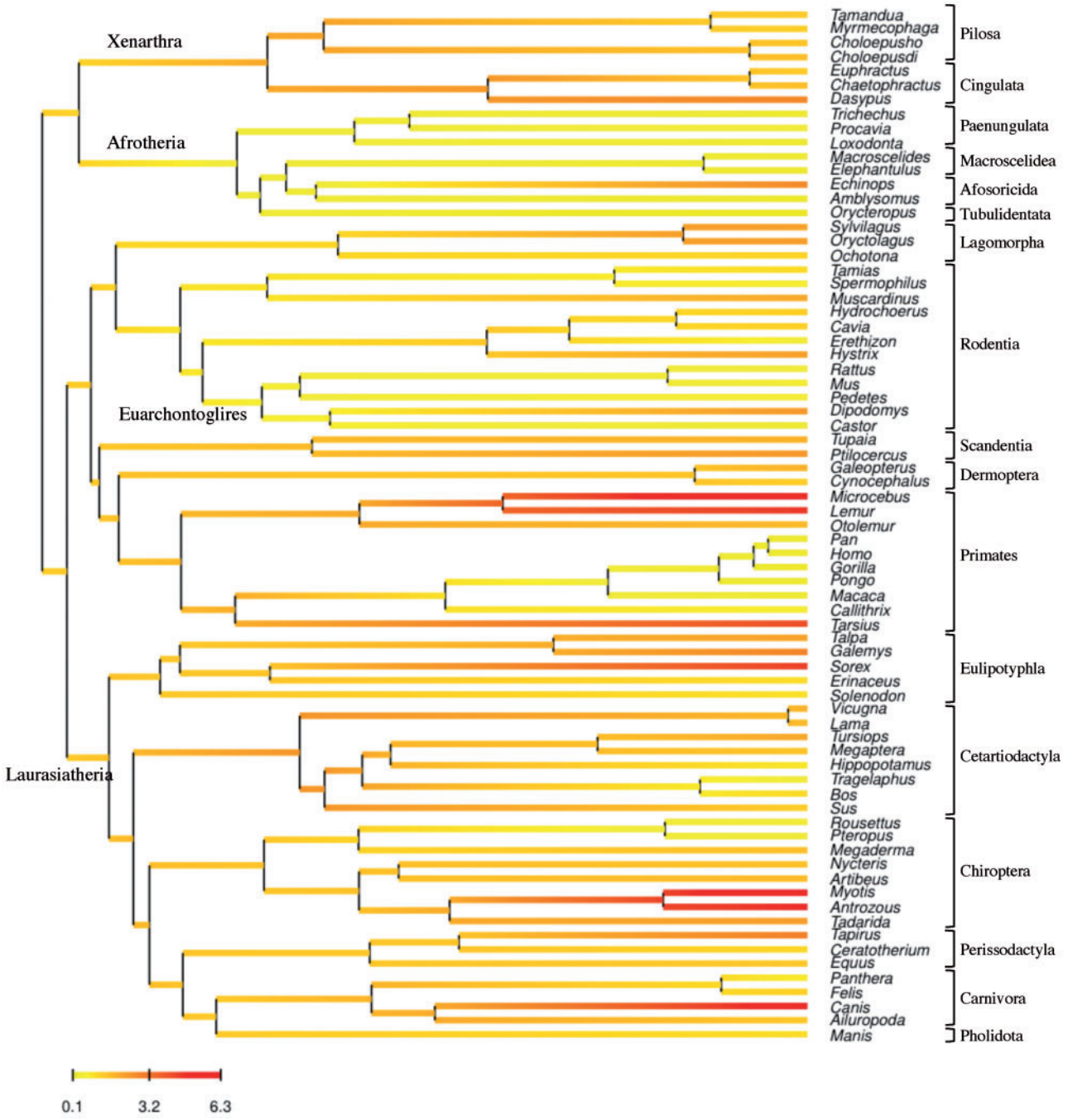

FIG. 2. Reconstructed (posterior mean) phylogenetic history of $B = 4Nb$ on the taxon-rich dataset.

genome-wide gBGC ($B = 3$ or more). Importantly, all these proportions are estimated based on branch averages, which are more stable than node estimates, but are such that short-term variation has been averaged out to some extent. Therefore, they provide a conservative estimate of the proportion of exons experiencing a peak of strong gBGC at any instant.

Altogether, the overall picture emerging from this analysis is one in which the fixation bias caused by GC-biased transmission distortion is widespread, mild on average, but globally above the nearly neutral threshold, and with a significant percentage of genes under strong gBGC in many lineages. Our findings therefore indicate that gBGC is likely to have a strong impact on the molecular evolutionary regime experienced by a large fraction of protein coding genes in placental mammals.

Discussion

gBGC combines several key properties making it a potentially significant player in evolutionary genomics. Because of its mechanistic connection with recombination, gBGC may represent a useful source of information about the evolution of recombination landscapes. In addition, gBGC has the potential to significantly distort the evolutionary process, both at neutral and selected sites, compared with the expectation under the neutral and the nearly neutral theory. In these

two respects, our estimation of the overall intensity and distribution of gBGC in mammalian genomes provides new and important insights, both about the determinants of the evolution of recombination landscapes, and concerning the overall quantitative impact of gBGC on genome evolution.

Karyotype and Hot Spots: The Evolution of Recombination Landscapes

Our main results concerning recombination landscapes and their evolution can be summarized as follows. First, phylogenetic correlations emphasize the role of karyotype in setting the genome-wide recombination rate. The distribution of gBGC across exons, on the other hand, and the modulation of these exon-specific effects over the phylogeny, reveals the existence of a large variance and a fast turnover of fine-scale recombination rates within genomes, suggesting that recombination hot-spots are a generic feature of placentals.

The role of karyotype in setting the recombination rate of a genome is already well documented (Kaback 1996; Yu et al. 2001; Myers et al. 2005; Coop and Przeworski 2007). There have been contradictory reports as to whether this mechanistic constraint should translate into a correlation of the recombination rate with the number of chromosomes, or the number of chromosomal arms. The classical view is that a minimum of one crossover per arm is needed (Dutrillaux 1986; Pardo-Manuel de Villena and Sapienza 2001), yet, what we find here is a better correlation of gBGC with the number of chromosomes than with the number of arms (table 1). A possible explanation could be that the constraint of one crossover per arm refers to the minimum possible crossover rate. The average recombination rate, on the other hand, could follow distinct trends. For instance, observations in closely related mouse species have suggested that Robertsonian fusions, which conserve the number of arms but reduce the number of chromosomes, result in a reduction of the average number of additional crossovers beyond the minimum imposed by the one-per-arm rule (Dumas and Britton-Davidian 2002). Such effects would mediate a correlation of gBGC with the number of chromosomes, rather than with the number of arms.

Concerning fine-scale variation in recombination rate, our observations reveal a variance in gBGC strength across exons compatible with what is known about hotspot intensity in humans, and generalize previous observations reporting the presence of clusters of GC-biased substitutions in human, mouse and dog (Capra and Pollard 2011) to all placental orders. On the other hand, the lack of resolution, both in time and along the genomes, still prevent us from drawing a detailed and quantitatively refined picture of the hot spot process in placentals. Since gBGC episodes occur on a time-scale that is shorter than, or of the order of, the average duration between successive cladogenetic events along the phylogenetic tree, they can manifest themselves only indirectly, either in terms of their contribution to the observed variance across loci, or alternatively, in terms of correlated substitution patterns, typically, clusters of GC-biased substitutions (Dreszer et al. 2007; Capra and Pollard 2011). This suggests that, ultimately, composite-likelihood methods (Deng and Moore 2009; Varin et al. 2011), or other approaches capturing correlated patterns across sites, could be used for translating the statistical properties of clusters into quantitative mechanistic insights about hot-spot dynamics.

Mutation Rates versus Neutral Substitution Rates

In its present form, the model introduced here attempts to jointly estimate mutation rates and conversion bias. This raises the question of the identifiability of these two quantities. A fundamental idea is that B and mutation rates can be jointly estimated as long as the variation in B across exons has a sufficient variance, so that a significant fraction of exons are under negligible gBGC at any time. Assuming this is the case, the model can then rely on the exons displaying the most AT-biased substitution patterns on each branch to estimate the mutation rates. Current empirical knowledge indicate that the distribution of recombination rates in mammalian genomes is indeed characterized by a large relative variance (Arnheim et al. 2007), suggesting in turn that the assumption of a leptokurtic distribution for B across loci is reasonable.

In several respects, however, the mutation rate estimates obtained by the model (table 2) are questionable, in particular concerning transversions. Thus, it is not clear whether mutation rates can be robustly estimated in practice. More fundamentally, the model makes several important assumptions about mutation rates, such as ignoring context-dependent effects, and assuming that the mutation spectrum is constant among lineages, and uniform genome-wide. Such simplifying assumptions certainly deserve closer scrutiny. Variation between species in mutation bias, and not just in B, could contribute to the observed variation in substitution patterns across the phylogeny. If this were the case, then the present analysis would erroneously ascribe a variation in mutation bias to variation in B. On the other hand, both simulations and empirical analyses (with rates fixed or estimated, with or without CpG positions, with mutation rates estimated globally or separately in each subclade), indicate that the reconstruction of B is robust to the details of the underlying mutation process. The reasons for the differential sensitivity of B and the mutation rates to the model's assumptions and approximations are not clear.

Concerning variation along the genome, we note that a significant part of the empirical evidence thus far cited in favor of variable mutation rates and patterns along genomes in fact comes from estimated substitution processes at putatively neutral sites (Lercher and Hurst 2002; Hellmann et al. 2005; Walser and Furano 2010; Johnson and Hellmann 2011), thus potentially mistaking variation in gBGC intensity for variation in mutation rates (Duret 2009). Our estimation of both mutation and substitution rates (table 2) suggests that, in the presence of gBGC, the distinction between mutation and neutral substitution rates is not merely conceptual, and should be taken into account before ascribing empirically observed variation to mutational causes.

Nevertheless, further work will be needed for relaxing the assumption of uniform and constant mutation patterns. A potential problem in this direction is how to preserve identifiability of the parameters of the model. Possibly, the model could be calibrated using sequences from weakly recombining regions (as in Duret and Arndt 2008). Calibrating the model in this way would make it possible to relax the assumption of constant mutation rates among lineages, although it would still assume uniformity along genomes. Relaxing this latter assumption would probably require to actually model the mechanistic causes for among-gene variation, while explicitly correlating local mutation rates with specific information about the genomic environment of each gene. Such methodological developments appear to be challenging. Ultimately, however, they would make it possible to test specific hypotheses about the extent and correlates of mutation rate variation among lineages and within genomes, in a statistical context where the confounding effects of biased gene conversion would have been explicitly modeled.

Implications for Molecular Evolution

According to the reconstruction of figure 1, in most placentals, gBGC is above the nearly neutral threshold ($B > 1$) for up to half of the protein coding genes, and is even strong ($B > 10$) for 1% to 10% of the genes, depending on the lineage. This finding has potentially important implications for molecular evolution.

First, such high levels of gBGC are sufficiently strong to overwhelm the subtle modulations of the substitution patterns as a function of population size predicted by the classical nearly neutral theory, all of which depend on fitness effects that are in the range of $4Ns \sim 1$ (Ohta 1972; Kimura 1979, 1983). In particular, levels of gBGC of the order of $4Nb$ between 1 and 10 are sufficient to create substantial distortions of the ratio dN/dS of non-synonymous over synonymous substitutions, such that genes experiencing a gBGC episode may be mistakenly diagnosed as being under positive selection. This has already been observed in primates (Galtier et al. 2009), despite of the relatively weak gBGC inferred in this group (fig. 1). Here, we have seen that as much as 50% of the exons can be under mild to strong gBGC in certain placental lineages, with an average of approximately 35%, therefore suggesting that dN/dS does not provide a reliable measure of selection stringency for more than one third of the exons in mammalian genomes. Caution may also be needed when interpreting global lineage specific variation in dN/dS genome-wide solely in terms of variation in effective population size (Lartillot 2012).

In addition to causing substantial deviation from the classical nearly neutral regime, the levels of gBGC estimated here also raise the question of the extent of the mutation load contributed by gBGC in placentals (Bengtsson 2009; Glémin 2010). In the absence of gBGC, the limit to selection imposed by random drift defines a threshold around $4Ns \sim 1$ below which selection is inefficient. The presence of gBGC, however, results in another limit, such that all mutations with $4Ns < 4Nb$ are dominated by gBGC. For exons under strong gBGC ($4Nb > 10$), the threshold above which selection is effective is therefore substantially raised, compared with what would be achieved in the absence of biased conversion.

On the other hand, upon an increase of N, both Nb and Ns will increase to the same extent, and therefore this will not result in the fixation of increasingly deleterious mutations. In contrast, if b itself increases, and if $4Nb > 1$, then this will result in the fixation of mutations that, taken individually, are more deleterious. Since our method gives an estimate only of the compound parameter $B = 4Nb$, it it difficult to know which of two parameters, N or b, is responsible for the variation in B observed in figures 1 and 2. The fact that the lineages inferred to have experienced the strongest gBGC all correspond to small mammals would seem to suggest that the increase in B is primarily due to an increase in population size in their case. Most probably, then, the strong gBGC observed in these groups does not result in a particularly overwhelming fixation load, but merely offsets the advantage normally conferred by larger population sizes in the face of slightly deleterious mutations.

Ultimately, an independent reconstruction of the evolution of long-term trends in effective population size along the phylogeny would greatly help sorting out the relative contribution of N and b. It would also provide an independent validation of the scaling relation between population size and body mass obtained using the present method (table 1). Estimating the phylogenetic variation in N could be done by reconstructing the dN/dS, or any other molecular evolutionary quantity expected to be dependent on population size in a predictable manner, using mitochondrial sequences, which are immune from gBGC. Alternatively, further theoretical developments could be pursued, in the aim of integrating gBGC and selection into an extended theory applicable to mammalian nuclear genomes (Kostka et al. 2012). In this direction, combining mutation-selection models (Halpern and Bruno 1998; Rodrigue et al. 2010; Tamuri et al. 2012) with the mutation-conversion model used here, seems to represent a promising avenue of research.

Supplementary Material

Supplementary methods, tables S1–S4, and figures S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Frédéric Delsuc, Nicolas Rodrigue, Mathieu Groussin, and Hervé Philippe for their useful comments on the manuscript, and Laurent Duret for kindly providing substitution rates estimates. Computational resources were provided by Calcul Québec and Compute Canada and the Canadian Foundation for Innovation. This work was supported by the Natural Science and Engineering Research Council of Canada.

References

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* 21: 2322–2328.

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.

Arnheim N, Calabrese P, Tiemann-Boege I. 2007. Mammalian meiotic recombination hot-spots. *Annu Rev Genet.* 41:369–399.

Auton A, Fledel-Alon A, Pfeifer S, et al. (23 co-authors). 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* 336: 193–198.

Axelsson E, Webster MT, Smith NGC, Burt DW, Ellegren H. 2005. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Res.* 15:120–125.

Bengtsson BO. 2009. The effect of biased conversion on the mutation load. *Genet Res.* 55:183–187.

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

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

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol.* 19:1181–1197.

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

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.

Coop G, Przeworski M. 2007. An evolutionary view of human recombination. *Nat Rev Genet.* 8:23–34.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319: 1395–1398.

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.

Deng L, Moore DF. 2009. Composite likelihood modeling of neighboring site correlations of DNA sequence substitution rates. *Stat Appl Genet Mol Biol.* 8:article 6.

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.

Dumas D, Britton-Davidian J. 2002. Chromosomal rearrangements and evolution of recombination: comparison of chiasma distribution patterns in standard and Robertsonian populations of the house mouse. *Genetics* 162:1355.

Duret L. 2009. Mutation patterns in the human genome: more variable than expected. *PLoS Biol.* 7:e1000028.

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.

Dutrillaux B. 1986. Role of chromosomes in evolution: a new interpretation. *Annales de Génétique.* 29:69–75.

Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol Evol.* 25:283–291.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet.* 2:549–555.

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.

Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C, Coop G, Przeworski M. 2009. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet.* 5:e1000658.

Fryxell K, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol.* 17: 1371–1383.

Galtier N. 2004. Recombination, GC-content and the human pseudoautosomal boundary paradox. *Trends Genet.* 20:347–349.

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

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.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol.* 15:910–917.

Harrison RJ, Charlesworth B. 2011. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu* stricto group of yeasts. *Mol Biol Evol.* 28:117–129.

Hassold T, Hansen T, Hunt P, VandeVoort C. 2009. Cytological studies of recombination in rhesus males. *Cytogenet Genome Res.* 124: 132–138.

Hellmann I, Prüfer K, Ji H, Zody MC, Pääbo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* 15: 1222–1231.

Hillier LW, Miller W, Birney E, et al. (175 co-authors). 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695–716.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet.* 29:217–222.

Johnson PLF, Hellmann I. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol Evol.* 3: 842–850.

Kaback DB. 1996. Chromosome-size dependent control of meiotic recombination in humans. *Nat Genet.* 13:20–21.

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.

Katzman S, Kern AD, Pollard KS, Salama SR, Haussler D. 2010. GC-biased evolution near human accelerated regions. *PLoS Genet.* 6:e1000960.

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: Cambridge University Press.

Kong A, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. *Nat Genet.* 31:241–247.

Kong A, Thorleifsson G, Gudbjartsson DF, et al. (15 co-authors). 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467:1099–1103.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol.* 29:1047–1057.

Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J Comput Biol.* 13:1701–1722.

Lartillot N. 2012. Interaction between selection and biased gene conversion in mammalian protein coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol Biol Evol.* Advance Access published September 29, 2012, doi:10.1093/molbev/mss231.

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

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* 18: 337–340.

Lipatov M, Arndt PF, Hwa T, Petrov DA. 2006. A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. *J Mol Evol.* 62:168–175.

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

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19:330–338.

Mateiu L, Rannala B. 2006. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Syst Biol.* 55:259–269.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.

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

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet.* 19: 128–130.

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

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

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. 1972. Population size and rate of evolution. *J Mol Evol.* 1: 305–314.

Paliulis LV, Nicklas RB. 2000. The reduction of chromosome number in meiosis is determined by properties built into the chromosomes. *J Cell Biol.* 150:1223–1232.

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

Pollard KS, Salama SR, King B, et al. (13 co-authors). 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* 2: e168.

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.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet.* 37: 429–434.

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

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol.* 7:241.

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.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A.* 107:4629–4634.

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.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A.* 102:9541–9546.

Tamuri AU, Dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190:1101–1115.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15: 1647–1657.

Varin C, Reid N, Firth D. 2011. An overview of composite likelihood methods. *Statistica Sinica* 21:5–42.

Walser JC, Furano AV. 2010. The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Res.* 20: 875–882.

Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet.* 28:101–109.

Winckler W, Myers SR, Richter DJ, et al. (11 co-authors). 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308:107–111.

Yu A, Zhao C, Fan Y, et al. (11 co-authors). 2001. Comparison of human genetic and sequence-based physical maps. *Nature* 409:951–953.