Conditional Approximate Bayesian Computation: A New Approach for Across-Site Dependency in High-Dimensional Mutation–Selection Models

Simon Laurin-Lemay, Nicolas Rodrigue, Nicolas Lartillot, Herve Philippe

Conditional Approximate Bayesian Computation: A New Approach for Across-Site Dependency in High-Dimensional Mutation–Selection Models

Simon Laurin-Lemay,*[,1] Nicolas Rodrigue,[2] Nicolas Lartillot,[3] and Hervé Philippe*[,1,4]

[1]Robert-Cedergren Center for Bioinformatics and Genomics, Department of Biochemistry and Molecular Medicine, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada

[2]Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

[3]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, Lyon, France

[4]Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Écologie Théorique et Expérimentale, UMR CNRS 5321, Moulis, France

***Corresponding authors:** E-mails: simon.laurin-lemay@umontreal.ca; herve.philippe@sete.cnrs.fr.

Associate editor: Rasmus Nielsen

Abstract

A key question in molecular evolutionary biology concerns the relative roles of mutation and selection in shaping genomic data. Moreover, features of mutation and selection are heterogeneous along the genome and over time. Mechanistic codon substitution models based on the mutation–selection framework are promising approaches to separating these effects. In practice, however, several complications arise, since accounting for such heterogeneities often implies handling models of high dimensionality (e.g., amino acid preferences), or leads to across-site dependence (e.g., CpG hypermutability), making the likelihood function intractable. Approximate Bayesian Computation (ABC) could address this latter issue. Here, we propose a new approach, named Conditional ABC (CABC), which combines the sampling efficiency of MCMC and the flexibility of ABC. To illustrate the potential of the CABC approach, we apply it to the study of mammalian CpG hypermutability based on a new mutation-level parameter implying dependence across adjacent sites, combined with site-specific purifying selection on amino-acids captured by a Dirichlet process. Our proof-of-concept of the CABC methodology opens new modeling perspectives. Our application of the method reveals a high level of heterogeneity of CpG hypermutability across loci and mild heterogeneity across taxonomic groups; and finally, we show that CpG hypermutability is an important evolutionary factor in rendering relative synonymous codon usage. All source code is available as a GitHub repository (https://github.com/Simonll/LikelihoodFreePhylogenetics.git).

Key words: **Markov chain Monte Carlo, synonymous substitution, nonsynonymous substitution, posterior predictive, phylogenetics.**

Introduction

The mutational process, a main basis of genetic variability, itself varies according to the environment (e.g., abiotic: Maharjan and Ferenci 2017; biotic Krasovec et al. 2017) and along the genome (Hodgkinson and Eyre-Walker 2011). One example of this mutational heterogeneity is the case of cytosine being much more mutable when followed by guanine in the genomes of vertebrates (Bird 1980), a phenomenon known as CpG hypermutability. As a result, a biased variability is subjected to selective processes, leaving a signal that seems clear in the cases of parallel adaptation (Stoltzfus and McCandlish 2017). Features of selection are probably even more heterogeneous along the genome and over time than those of mutation. Selection acts at multiple levels (e.g., DNA, RNA, protein, cell, tissue, organism, population, community, and ecosystem), and conflicts can exist between levels or because of fluctuations in the environment. The heterogeneity of selection is obvious when examined at a fine scale, for instance within a protein, where each site typically displays a strong preference for a small subset of amino acids (Halpern and Bruno 1998; Lartillot and Philippe 2004; Rodrigue and Lartillot 2012; Tamuri et al. 2012; Rodrigue 2013; Rodrigue and Lartillot 2014; Tamuri et al. 2014; Echave et al. 2016; Hilton et al. 2017; Rodrigue and Lartillot 2017; Wang et al. 2018).

In comparative genomics, these complexities make it difficult to separate the effects of selection from the bias induced by mutational features. Codon usage (CU) in mammals provides a good illustration of this problem. Some authors argue that selection is acting on CU (Yang and Nielsen 2008; Kessler and Dean 2014) to favor efficiency of translation (Drummond and Wilke 2008; Cannarozzi et al. 2010; Tuller et al. 2010), whereas others argue that population sizes are too small to allow selection of such a minor advantage, particularly in Primates (Duret 2002; Pouyet et al. 2016; Laurin-Lemay et al. 2018; Galtier et al. 2018), and therefore that CU is the result of neutral evolution (Ohta 1973). In agreement with the latter view, CU in mammals mostly reflects GC3 content

Article

(Sueoka 1961, 1962; Muto and Osawa 1987; Ermolaeva 2001; Knight et al. 2001; Chen et al. 2004; Li et al. 2015) or isochore structure (Filipski et al. 1973; Bernardi 2000), suggesting that it is determined by the mutational pressure and by fixation biases likely related to GC-biased gene conversion (Duret 2002; Duret and Galtier 2009; Katzman et al. 2011; Glemin et al. 2015).

A promising solution to tease apart mutation and selection in coding sequences is to develop mechanistic codon substitution models (Rodrigue and Philippe 2010) that operate in a mutation–selection framework. Such mutation–selection models have previously been developed to study the role of protein structure (Robinson et al. 2003; Rodrigue et al. 2006, 2005, 2009; Kleinman et al. 2010), codon preference (McVean and Vieira 2001; Nielsen et al. 2007; Rodrigue et al. 2008; Yang and Nielsen 2008; Rodrigue and Philippe 2010; Pouyet et al. 2016), or site-specific amino acid preferences (Halpern and Bruno 1998; Rodrigue et al. 2010; Tamuri et al. 2012; Rodrigue 2013; Tamuri et al. 2014). However, thus far, the main focus has been on the modeling of complex features of selection, whereas simple, homogeneous, parameterization were used for the mutational aspects of the model, often the very simple HKY model (Hasegawa et al. 1985). Yet, violations in the mutational part of the model can easily lead to erroneous detection of selection (e.g., Lartillot 2013; Van den Eynden and Larsson 2017; Laurin-Lemay et al. 2018). In particular, the latter study shows erroneously inferred selection on CU when using simple models on sequence alignments simulated with mild CpG hypermutability, but without any selection on CU.

To take full advantage of the mutation–selection models, it may be necessary to incorporate more complexity (i.e., natural heterogeneity) in both mutation-level and selection-level specifications of the model. However, heterogeneity often implies handling parameter vectors of high dimensionality and across-site dependency, both of which create computational difficulties. High dimensionality can lead to overfitting in a maximum likelihood framework. As for across-site dependency, it leads to intractable likelihood calculations (precluding the use of the pruning algorithm; Felsenstein 1973, 1981). The Bayesian framework, thanks to the use of Markov chain Monte Carlo (MCMC; Metropolis et al. 1953; Hastings 1970), enables the study of rich models accounting for across-site heterogeneity of amino acid profiles, as previously shown in the case of site-specific amino acid preferences (Rodrigue et al. 2010). Approximate Bayesian Computation (ABC) avoids the computation of the likelihood (Pritchard et al. 1999; Beaumont et al. 2002; Marjoram et al. 2003; Sisson et al. 2007), and could be a means of addressing the across-site dependency issue, whether at the level of mutation (e.g., CpG contexts; Pedersen et al. 1998; Jensen and Pedersen 2000; Arndt et al. 2003; Huttley 2004; Hwang and Green 2004; Siepel and Haussler 2004; Arndt and Hwa 2005; Christensen et al. 2005; Christensen 2006; Hobolth et al. 2006; Hobolth 2008; Duret and Arndt 2008; Lindsay et al. 2008; Misawa and Kikuno 2009; Suzuki et al. 2009; Keightley et al. 2011; Misawa 2011; Ying and Huttley 2011; Berard and Gueguen 2012; Huttley and Yap 2012; Lee et al. 2015, 2016) or selection (e.g., on protein structure; Robinson et al. 2003; Rodrigue et al. 2006, 2005, 2009; Kleinman et al. 2010). Unfortunately, the classical rejection sampling (RS) ABC cannot deal with complex models involving parameter vectors of high dimensionality (Kousathanas et al. 2016). Here, we propose a new approach, named Conditional ABC (CABC), which combines the advantages of MCMC and ABC. As a proof-of-concept, we study the across-site dependent hypermutability of CpG, while modelling the high dimensionality of site-specific amino acid selection.

New Approaches: CABC

We consider the general situation where we have a model with parameters (λ, θ) and a data set D under study. The parameter θ represents the (potentially high-dimensional) nuisances. The parameter λ, on the other hand, is our parameter of interest. Computationally, the model is assumed to be intractable by classical MCMC in the generic case, except under a reference value for λ (e.g., $\lambda = 1$). Here, we can think of λ as the relative rate of CpG mutation—a feature that implies across-site dependency—such that, for $\lambda = 1$, the model reduces to the usual site-independent model.

Ideally, we would like to sample from the joint posterior:

$$(\lambda, \theta) \sim p(\lambda, \theta|D), \tag{1}$$

and then conduct inference on λ (e.g., by visualizing the marginal posterior distribution of λ and computing the mean and 95% credible interval). Noting that the joint posterior can be factorized as follows:

$$p(\lambda, \theta|D) = p(\lambda|D, \theta)p(\theta|D), \tag{2}$$

the sampling procedure denoted by equation (1) could equivalently be done in two steps:

$$\theta \sim p(\theta|D), \tag{3a}$$

$$\lambda \sim p(\lambda|D, \theta), \tag{3b}$$

that is, by first sampling θ from its marginal posterior (marginal over λ) and then sampling λ from its conditional posterior (conditional on θ).

Neither of the two sampling steps described by (3a) and (3b) can be performed exactly. Accordingly, the CABC approach proposed here relies on two main approximations. First, the marginal posterior (3a) is approximated by the posterior on θ under the reference model, that is, $p(\theta|D, \lambda = 1)$, using MCMC; we denote this approximated posterior distribution as $p_{MCMC}(\theta|D, \lambda = 1)$. Second, sampling λ conditional on θ (3b) is done by classical ABC, denoted $p_{ABC}(\lambda|D, \theta)$. Provided that the nuisance parameters and λ are weakly correlated under the true posterior, these approximations should be relatively accurate.

In summary, the approach proceeds in two steps:

$$\theta \sim p_{MCMC}(\theta|D, \lambda = 1), \tag{4a}$$

$$\lambda \sim p_{ABC}(\lambda|D, \theta), \tag{4b}$$

or equivalently:

$$(\lambda, \theta) \sim p_{CABC}(\lambda, \theta|D), \quad (5)$$

where:

$$p_{CABC}(\lambda, \theta|D) = p_{ABC}(\lambda|D, \theta)p_{MCMC}(\theta|D, \lambda = 1). \quad (6)$$

Comparing (2) and (6), the two approximations invoked by the CABC are the use of ABC, instead of exact Bayesian inference on λ conditional on θ, and the fact that θ is not from its marginal posterior (marginal on λ), but is instead from its reference posterior (with $\lambda = 1$).

In practice, some of the nuisance parameters collectively denoted by θ might be strongly correlated with λ, in which case the approach will be inaccurate. Let us further subdivide the parameterization, by defining:

$$\theta = (\theta_{sc}, \theta_{wc}), \quad (7)$$

where θ_{sc} is strongly correlated, and θ_{wc} weakly correlated, with λ under the joint posterior. Provided that θ_{sc} is sufficiently low-dimensional, we can resample it by ABC, along with λ:

$$\theta_{wc} \sim p_{MCMC}(\theta_{wc}|D, \lambda = 1), \quad (8a)$$

$$(\lambda, \theta_{sc}) \sim p_{ABC}(\lambda, \theta_{sc}|D, \theta_{wc}), \quad (8b)$$

or equivalently:

$$(\lambda, \theta_{sc}, \theta_{wc}) \sim p'_{CABC}(\lambda, \theta_{sc}, \theta_{wc}|D), \quad (9)$$

where:

$$p'_{CABC}(\lambda, \theta_{sc}, \theta_{wc}|D) = p_{ABC}(\lambda, \theta_{sc}|D, \theta_{wc})p_{MCMC}(\theta_{wc}|D, \lambda = 1). \quad (10)$$

Working with (10) instead of (6) will decrease the impact of the approximation implied by using the reference marginal posterior, as opposed to the true marginal posterior, on a smaller component of the parameter vector, although at the cost of an increase in the impact of the approximation entailed by conducting the ABC on a higher dimensional parameter (λ, θ_{sc}). Note that we do not have a theoretical basis from which to establish which nuisance parameters are to be considered as weakly or strongly correlated to λ_{CpG}. This problem is to be addressed empirically, exploiting our knowledge of the underlying biology and modeling system, and ultimately studied through simulations.

To illustrate the potential of the CABC approach, we apply it to the estimation of the well-established hypermutability at CpG sites—which involves dependence across sites—in the context of a complex reference model combining site-independent mutation, handled by a general-time-reversible nucleotide level parameterization, (Lanave et al. 1984), denoted M[GTR], along with purifying selection on amino-acids (i.e., site-specific amino-acid preferences) captured by a Dirichlet process prior (Rodrigue et al. 2010), denoted S[NCatAA*]. In this specific application of CABC, the parameter of interest (denoted above as λ) is λ_{CpG}, the ratio of the mutation rate of transitions at CpG sites to the mutation rate of transitions at nonCpG sites. The reference model (without CpG hypermutation, or equivalently, with $\lambda_{CpG} = 1$) is denoted by M[GTR]-S[NCatAA*], whereas the complete model (with CpG hypermutation) is referred to as M[GTR+ts-CpG]-S[NCatAA*].

The high-dimensional parameter vector of the reference model was partitioned into strongly and weakly correlated components, as discussed above, by reasoning as follows. On one hand, the estimation of the site-specific fitness profiles and of relative branch lengths should be robust to the specific model used for the mutation process (whether or not CpG hypermutation is included). On the other hand, the context-independent component of the mutation process (the GTR process) is expected to be strongly correlated with λ_{CpG} under the true posterior distribution. Accordingly, the high-dimensional amino-acid profiles and the branch lengths were estimated by MCMC under the reference posterior distribution, with $\lambda_{CpG} = 1$ (i.e., were included in θ_{wc}), while the 10 GTR parameters (8 degrees of freedom), as well as three modulator parameters (meant as correcting factors for total tree length, mean nonsynonymous/synonymous rate deviation, and relative position of the root along the branch separating the in- and the out-group, see Materials and Methods for details) were re-estimated at the ABC step, along with λ_{CpG} (i.e., were therefore included in θ_{sc}). We study the approach using simulations, and apply it to 137 real protein-coding genes from 39 mammals (see Materials and Methods for details).

Results and Discussion

Validation of the CABC Procedure

We validated the CABC approach using simulations. We simulated 5,000 alignments, using various values for λ_{CpG} (ranging from 0.5 to 8) combined with empirically estimated parameter values for the reference model. Then, we applied the CABC approach to these simulated alignments, and evaluated the relative mean square error (RMSE) of the approximated posterior and the coverage properties of the posterior credible intervals. For the ABC step, we used two alternative approaches: either a simple ABC RS algorithm (Pritchard et al. 1999), or a more sophisticated approach based on the use of a linear regression model (LRM) for getting closer to the true posterior distribution (Blum and Francois 2010). Note that the ABC step itself relies on simulations, with numerous summary statistics (SS) computed and compared between this step's simulated data sets and the data set under analysis. Only a small percentages of these simulations are retained by the procedure, as controlled by the *tolerance* level. We used 13 SS related to the frequency of certain states and the counts of specific pairwise differences between sequences (see Materials and Methods for details). We explored empirically different sample sizes and tolerance levels.

Table 1. Global Relative Mean Square Error (without λ_{ROOT}) Computed for Different λ_{CpG} Values (1,000 Replicates per λ_{CpG} Value) under Two Tolerance Levels, 10% and 1%, and over 10^5 Simulations.

Method	Tolerance Level	$\lambda_{CpG}=0.5$	$\lambda_{CpG}=1$	$\lambda_{CpG}=2$	$\lambda_{CpG}=4$	$\lambda_{CpG}=8$
RS	10%	34.30 ± 3.09	15.06 ± 3.13	10.25 ± 3.6	9.49 ± 3.89	9.87 ± 4.10
RS	1%	18.20 ± 4.03	10.02 ± 1.73	6.78 ± 1.84	6.06 ± 2.20	6.53 ± 2.55
RS + LRM	10%	0.86 ± 0.19	0.86 ± 0.14	0.89 ± 0.13	0.89 ± 0.12	0.86 ± 0.16
RS + LRM	1%	0.70 ± 0.19	0.69 ± 0.14	0.71 ± 0.13	0.72 ± 0.12	0.71 ± 0.13

RS: rejection sampling algorithm.
LRM: linear regression model.

All the approximate posterior distributions obtained by running the CABC procedure with the RS algorithm alone were inaccurate: the global RMSE ranged from ∼4 to ∼34 depending on the value of λ_{CpG} (tables 1 and 2). The global RMSE decreases when the tolerance level is decreased (from 1% to 0.1% of the simulated samples), but remains high even under the most stringent settings, suggesting that much smaller tolerance levels (implying a much larger total number of simulated samples) would still be needed in order for the simple RS approach to yield a reasonable approximation to the true posterior distribution (Barber et al. 2015).

In contrast, the global RMSE obtained when using the LRM of Blum and Francois (2010) fall under 1 (tables 1 and 2). The accuracy of CABC with LRM rose when the sampling effort is increased and the tolerance level is reduced (tables 1 and 2). The global RMSE remained very similar, around 0.70, when using the best 1% of the 10^5 simulations compared with the best 1% of 10^6 simulations. A reduction of the tolerance level (0.1%), however, decrease RMSE, by ∼20% (table 2). We note that, in spite of performing well here, the behavior of LRM in presence of model violations has been shown to be potentially misleading (Frazier et al. 2017, unpublished data). Given our simulation results, however, all RS results were corrected using the LRM unless stated otherwise.

The RMSE associated with each parameter (fig. 1 and supplementary table S1, Supplementary Material online) appears to be strongly linked to the amount of signal relevant to that parameter. For instance, the RMSE for the transition exchangeabilities (ϱ_{AG} and ϱ_{CT}) were 4 times lower than for the transversion exchangeabilities (ϱ_{AC}, ϱ_{AT}, ϱ_{CG}, and ϱ_{GT}). Similarly, the four nucleotide propensities have the smallest RMSE, being in fact the smallest for the two most frequent nucleotides (C and G). The RMSE for λ_{ROOT} (the correcting factor for the relative position of the root along the branch separating the in- and the out-group) is the highest (> 0.30); indeed the posterior distribution of this parameter is almost identical to its prior distribution, demonstrating that the signal provided by the nonreversibility of the context-dependent mutation process is too tenuous to be captured when analyses are conducted on single genes.

The improvement brought by a sample size of 10^6 and a tolerance level of 0.1% applied mainly to λ_{TBL} and λ_{ω_*} (the correcting factors for total tree length and dN/dS deviation), as well as the transversion exchangeabilities. In contrast, the improvement was minor for λ_{CpG}. Of note, the RMSE for λ_{CpG} is smaller under high rates of CpG hypermutability, reflecting the more abundant empirical signal (i.e., a higher number of CpG hypermutation events) in this regime; thus, when the true λ_{CpG} is equal to 8, the corresponding RMSE (0.0362) is below that observed for transversion exchangeabilities and close to the one obtained for transition exchangeabilities. To explore the idea that more evolutionary signal leads to a decreased in RMSE, we plotted the relation between the total number of expected substitutions and the RMSE computed on λ_{CpG} (supplementary fig. S1, Supplementary Material online). As expected, we found a negative relationship between the amount of evolutionary signal and the RMSE on λ_{CpG}. Moreover, as the evolutionary signal for λ_{CpG} becomes more prominent (panels S1: A–E), the fit of the regression become higher: the r^2 values go from 0.249 (lowest value of λ_{CpG}, 0.5) to 0.797 (highest value of λ_{CpG}, 8). As a result, when applied to real data, CABC will be precise if there is a high rate of transition in the CpG context. In conclusion, the computational burden of 10^6 simulations is mainly useful if one wants to study the effect of CpG on the transversion rates. Otherwise, a less intense sampling effort (10^5), combined with a moderate tolerance level (1%), gives reasonably accurate inference.

The coverage properties (i.e., the frequency at which credible intervals cover the true value) provide another interesting perspective on the statistical properties of CABC. Here, Probability–Probability (P–P) plots are used to investigate the coverage properties for several parameters of interest. On these plots, a straight line along the diagonal indicates that the nominal and true coverage coincide, that is, $1-\alpha$ credible intervals cover the true value at a frequency equal to $1-\alpha$. If this is the case, then credible intervals are true frequentist confidence intervals. Coverage is not necessarily expected to be perfect for all aspects of the model (i.e., nuisance parameters), but is an important property when the intention is to test a null hypothesis (e.g., $\lambda_{CpG}=1$), with a frequentist control of the type I error (rate of false positive).

The coverage properties were poor for all parameters when using RS alone (supplementary fig. S2, Supplementary Material online). As for RMSE, the use of LRM greatly improved the concordance between nominal and true coverage (supplementary fig. S3, Supplementary Material online), while the increase in sample size from 10^5 to 10^6 allowed a minor improvement (supplementary fig. S4, Supplementary Material online). The coverage properties were good for all parameters but λ_{ω_*}, λ_{TBL}, and λ_{ROOT}, and to a lesser extent for nucleotide exchangeabilities when $\lambda_{CpG}=8$. The poor coverage of ϱ_{AG} and ϱ_{CT} when λ_{CpG} is >1 (supplementary fig. S4, Supplementary Material online) could be explained by the rise of the uncertainty since a great amount of mutational

Table 2. Global Relative Mean Square Error (without λ_{ROOT}) Computed for Different λ_{CpG} Values (1,000 Replicates per λ_{CpG} Value) under Two Tolerance Levels, 1% and 0.1%, and over 10^6 Simulations.

Method	Tolerance Level	$\lambda_{CpG}=0.5$	$\lambda_{CpG}=1$	$\lambda_{CpG}=2$	$\lambda_{CpG}=4$	$\lambda_{CpG}=8$
RS	1%	18.21 ± 3.86	10.01 ± 1.66	6.78 ± 1.84	6.05 ± 2.18	6.54 ± 2.56
RS	0.1%	8.22 ± 2.98	6.27 ± 1.35	4.50 ± 0.76	3.73 ± 0.95	4.07 ± 1.28
RS + LRM	1%	0.69 ± 0.18	0.69 ± 0.14	0.71 ± 0.13	0.71 ± 0.11	0.70 ± 0.12
RS + LRM	0.1%	0.53 ± 0.17	0.52 ± 0.14	0.54 ± 0.13	0.54 ± 0.11	0.54 ± 0.11

RS: rejection sampling algorithm.
LRM: linear regression model.

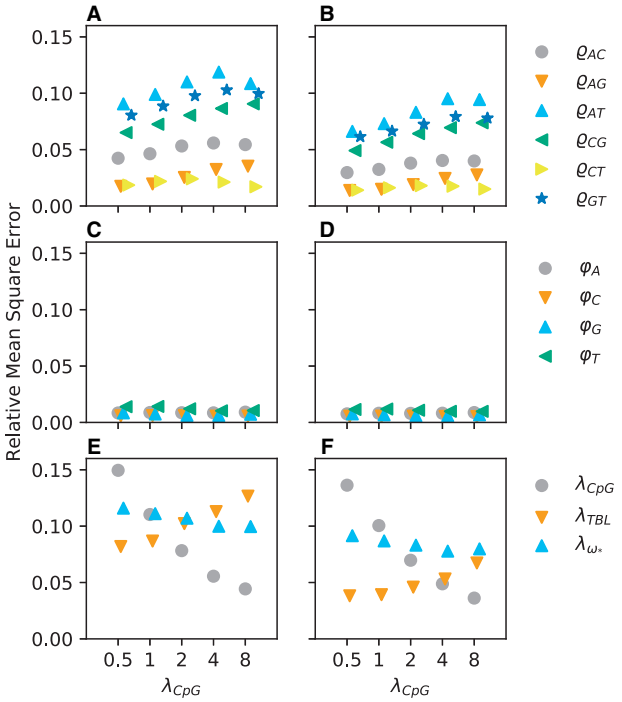

Fig. 1. Relative mean square error (mean over 1,000 replicates) under different λ_{CpG} values (*x* axes). Two tolerance levels, 1% (left panels) and 0.1% (right panels) over 10^6 simulations were used. Parameter values were corrected using linear regression model. (*A* and *B*) Mean RMSE of the six nucleotide exchangeabilities (ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT}, and ϱ_{GT}). (*C* and *D*) Mean RMSE of the four nucleotide propensities (φ_A, φ_C, φ_G, and φ_T). (*E* and *F*) Mean RMSE of λ_{CpG}, λ_{TBL}, and λ_{ω_*}.

signal related to the GTR component is transferred to the λ_{CpG}. Importantly, our parameter of interest, λ_{CpG}, had excellent coverage properties (fig. 2), which is of prime importance to test the hypothesis that λ_{CpG} is >1.

We further characterized the properties of CABC by transferring the parameters of the GTR mutation model from θ_{sc} to θ_{wc} at the ABC step. The expectation is that CABC will be inaccurate, because the hypermutability of CpG will lead to an artifactual increase in the transition/transversion ratio and the A + T content inferred under the reference model. To investigate this point, we used simulations made with a $\lambda_{CpG} = 8$ and a sample size of 10^5. Indeed, under these new settings, the RMSE on λ_{CpG} was much increased (with a 2-fold increase of the RMSE). Similarly, coverage was poor for λ_{CpG}, as well as for all the GTR parameters (supplementary fig. S5, Supplementary Material online). This is in sharp contrast to the case where the GTR parameters are re-estimated (i.e., within θ_{sc}, supplementary fig. S3, Supplementary Material online): in this case, λ_{CpG} and nucleotide propensities (except φ_G) are well estimated. The estimation of relative nucleotide exchangeabilities is equally poor in the two cases, suggesting that these parameters might not be strongly correlated with λ_{CpG} (see below), but probably just impacted by the lack of signal under $\lambda_{CpG} = 8$ for the GTR component, as previously explained. The correcting factors, λ_{TBL} and λ_{ω_*}, are more accurately inferred when the GTR parameters are not themselves re-estimated (supplementary figs. S3 and S5, Supplementary Material online).

Finally, we evaluated the impact of the estimation of the large number of nuisance parameters represented by branch lengths and site-specific amino-acids profiles on the overall accuracy of CABC, by running the entire procedure with all these parameters fixed to their true values. Granting perfect knowledge about these nuisances is expected to improve the accuracy of the estimation of all other parameters. However, if the improvement turns out to be minor, this will show that 1) in itself, uncertainty about these nuisance parameters is not detrimental, and 2) our approximation based on estimating these nuisances under the reference model (and not under the target model) does not compromise the overall quality of the inference. We used simulations made with a $\lambda_{CpG} = 8$ and a sample size of 10^5.

The RMSE for all parameters were very similar to the results obtained under the standard validation procedure (supplementary table S1, Supplementary Material online). For instance, the estimation of λ_{CpG} was only weakly impacted by the use of the true branch lengths and the true site-specific amino acid preferences. The parameter most impacted was λ_{ω_*}: its RMSE decreased from 0.100 to 0.053 (supplementary table S1, Supplementary Material online), accounting for 67% of the reduction of the global RMSE. The P–P plots (supplementary fig. S6, Supplementary Material online) are in agreement with RSME and are very similar to the case where we drew θ_{wc} from $p(\theta_{wc}|D, \lambda_{CpG} = 1)$ (supplementary fig. S3, Supplementary Material online).

In conclusion, the CABC procedure is reasonably accurate as long as the parameters included in θ_{wc} are indeed weakly influenced by λ_{CpG}. In particular, the accuracy suggested by our simulation study is largely sufficient to test the hypothesis that λ_{CpG} is equal to 1, with a good control of type I error, and even to study the impact of the CpG hypermutability on the GTR parameters (with a somewhat greater uncertainty concerning the four transversion exchangeabilities). From here, all

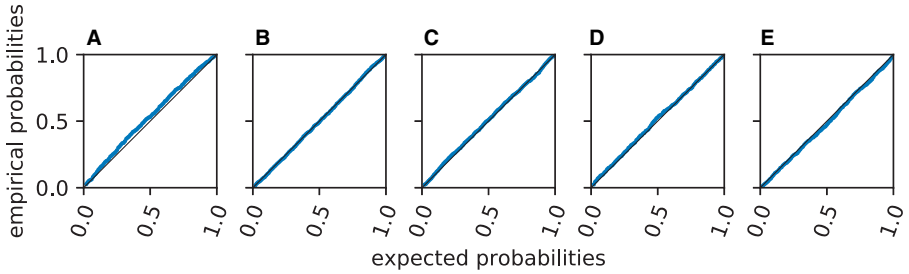

FIG. 2. P–P plots of the λ_{CpG} recovered from the analyses of simulated alignments generated under λ_{CpG} values (0.5, 1.0, 2.0, 4.0, 8.0), corresponding respectively to (*A*–*E*). Empirical probabilities were obtained using rejection sampling (the best 0.1% of 10^6 simulations) corrected with a linear regression model. The frequency at which the true values of λ_{CpG} within each credibility intervals is uniformly distributed (two sided Kolmogorov–Smirnov test: $p = 0.848$, $p = 1$, $p = 0.999$, $p = 0.996$, and $p = 1$ respectively). A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

the results we present below are obtained with the LRM (see Materials and Methods) made on the 0.1% best of 10^6 simulations.

Estimation of the Mutation Rate in the CpG Context Using CABC

We applied the CABC to approximate the posterior distribution of λ_{CpG} for a sample of 137 mammalian genes from 39 species (fig. 3). In agreement with previous observations (Hodgkinson and Eyre-Walker 2011), CABC always inferred a posterior mean transition rate in the CpG context greater than one, with an average value of 7.45; none of the 137 genes included $\lambda_{CpG} = 1$ within their 99% credible intervals (supplementary table S2, Supplementary Material online). The bimodal shape of the marginal distribution (fig. 3) is due to two genes (*ARNT* and *KIAA0100*) for which the transition rate in the CpG context obtains a posterior mean of 18.8 and 19.5, respectively (supplementary table S2, Supplementary Material online). A total of 16 genes displayed posterior mean values for $\lambda_{CpG} > 10$, that is, outside the prior belief [1/10, 10]. Values outside the prior were reached through the use of the LRM approach. To further explore this result, new CABC analyses were conducted over the 137 genes with a broader prior (log-uniform over [1/50, 50]), using the same sampling scheme and tolerance level. The impact of the prior on the estimation of λ_{CpG} was minor (supplementary fig. S7, Supplementary Material online), as indicated by the fact that the posterior means are highly correlated between the two alternative prior settings ($R^2 = 0.98$). Of note, the use of a narrow prior (over [1/10, 10]), leads to an underestimation of λ_{CpG}, making our approach conservative in the evaluation of hypermutability in the CpG context.

We then applied CABC to investigate whether the value of λ_{CpG} is homogeneous across the placental tree. We subdivided our data sets into three clades: Glires (7 species), Laurasiatheria (14 species), and Primates (12 species). The gene-specific estimates of λ_{CpG} obtained for each of these three clades (supplementary fig. S8, Supplementary Material online) are well correlated with the ones obtained for placentals ($r^2 = 0.94$, 0.96, and 0.96, respectively), indicating that the hypermutability in the CpG context is relatively well conserved along the placental tree. However, the slope of the

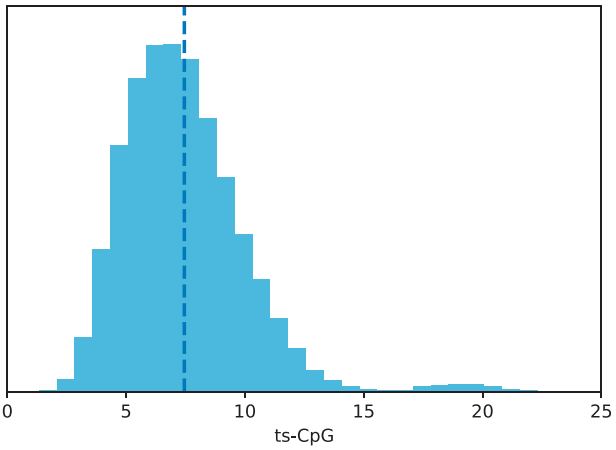

FIG. 3. Aggregation of posterior distributions of λ_{CpG} recovered from 137 mammalian genes using the CABC methodology. Rejection sampling (the best 0.1% of 10^6 simulations) with linear regression model were used to approximate posteriors. The vertical blue dash line represents the mean λ_{CpG} value (7.45) over all posterior values pooled.

regression (passing through the origin) is below one (0.96 and 0.91) for Glires and Laurasiatheria, respectively, and greater than one (1.14) for Primates. The higher level of CpG transition rate in Primates is congruent with the results of Keightley et al. (2011), although heterogeneity across clades is less marked here. This could be due to the fact that the analysis of Keightley et al. (2011) is based on pairs of species, whereas the present analysis relies on the information contributed by 39 placental species considered simultaneously.

Finally, we looked at the effect of taking into account CpG hypermutability on the other parameters of the mutation process of the model, M[GTR+ts-CpG]. Overall, these parameters were slightly affected by the inclusion of the λ_{CpG} parameter, which is understandable given the relative rarity of CpG in mammalian protein coding sequences (with the mean observed/expected CpG ratio of 0.41). However, comparison of the values of $\varphi_G + \varphi_C$ (fig. 4) shows that the CpG hypermutability has a complex effect, strongly dependent on the gene. This is congruent with our assumption that the GTR parameters are strongly correlated with λ_{CpG} and should be re-estimated at the ABC step. On average, a tenuous increase in $\varphi_G + \varphi_C$ is observed (fig. 4), which is expected

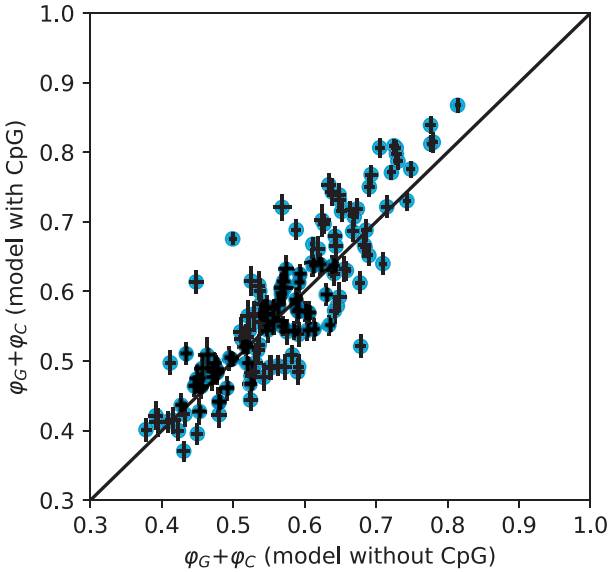

FIG. 4. Comparison of the $\varphi_G + \varphi_C$ posterior mean estimates under the models without (x axis) and with CpG hypermutation (y axis) recovered from the analysis of the 137 mammalian gene alignments. A diagonal line is added (black) to appreciate any deviation between both models estimate. The error bars correspond to the standard deviations computed from each posterior.

since the hypermutability of CpG tends to decrease G + C content.

Posterior Predictive Checks to Analyze the Effect of CpG Hypermutability on Some Sequence Characteristics

Instead of looking at the GTR parameters, a more sensible approach is to examine the predictions made by both models, including and not including CpG hypermutability. First, we compared the GC content observed at the third codon positions (GC3) in empirical data to the GC3 predicted by the two models (fig. 5). The model not including CpG hypermutability overpredicts GC3, as previously noticed by Mugal et al. (2015). The model including CpG hypermutability gets closer to the observed GC3, but with a small underprediction especially for high values of GC3. The inclusion of the CpG hypermutability by CABC therefore allows to improve the prediction of GC3, a widely used measure to estimate mutational pressure (Sueoka 1961, 1962; Muto and Osawa 1987; Ermolaeva 2001; Knight et al. 2001; Chen et al. 2004; Li et al. 2015).

Second, during the simulations conducted to generate the posterior predictive alignments, we computed statistics on the substitution histories over the tree (table 3). The number of substitutions is higher for the model including CpG hypermutability compared with the reference model (6,342 ± 3,202 vs. 5,214 ± 2,511). Focusing on the relative frequencies of substitution types, the model including CpG hypermutability predicted more transitions relative to transversions (77.3% vs. 71.3% for the reference model). The C->T and G->A transition rates show the sharpest increase (+14.2% and +10.7%), in agreement with the increased transition rate at CpG sites implied by CpG hypermutability, but

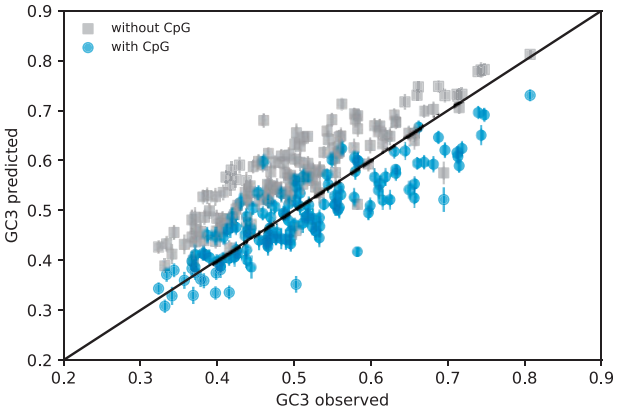

FIG. 5. Comparison of the ability of the models without (gray squares) and with CpG hypermutation (blue circles) to predict the GC3 content of the 137 mammalian gene alignments using posterior predictive simulations. The observed GC3 is plot against the mean predictions (y axis) from both models. A diagonal line is added (black) to appreciate any deviation between observations and the predictions. The error bars correspond to the standard deviations computed from models predictions.

Table 3. Comparison of the Proportions of Substitution Types Recovered from the Posterior Predictive Simulations (Mean over 137 Mammalian Gene Analyses).

Type of Substitutions	Without CpG	With CpG
ts	71.33 ± 3.38	77.29 ± 2.91
A>G	17.16 ± 2.1	17.33 ± 2.05
G>A	17.04 ± 2.15	18.87 ± 2.05
C>T	18.49 ± 2.01	21.12 ± 2.25
T>C	18.64 ± 2.08	19.97 ± 2.17
tv	28.67 ± 3.38	22.71 ± 2.91
A>C	4.70 ± 0.93	4.68 ± 0.88
C>A	4.74 ± 0.92	3.29 ± 0.77
A>T	2.53 ± 0.7	2.53 ± 0.7
T>A	2.59 ± 0.72	2.46 ± 0.68
C>G	3.70 ± 1.05	2.49 ± 0.75
G>C	3.62 ± 1.06	2.23 ± 0.71
G>T	3.52 ± 1.08	1.92 ± 0.49
T>G	3.28 ± 0.69	3.10 ± 0.59

ts for transitions; tv for transversions.

the T->C (+7.1%) and A->G (+1.0%) transition rates also show a nonnegligible increase. The pattern is similarly complex for transversions, with an important decrease for G->T (−54.5%), G->C (−61.6%), C->G (−67.3%), and C->A (−69.4%), the other types of transversions being relatively unaffected. The complexity of the impact of the CpG hypermutability on the relative frequencies of the 12 types of substitutions is difficult to interpret, being the result of an interplay between the mutation process, the genetic code and selection on amino-acids.

Third, we exclusively looked at the substitutions in the CpG context (table 4), which should be easier to interpret. Unsurprisingly, the number of CpG->TpG or ->CpA transitions among all substitutions were much more frequent (from 234 ± 133 to 584 ± 318) than other substitution types. When analyzed with respect to the position of CpG within codons, it appears that only CpG->CpA at positions 2-3 and

Table 4. Comparison of the Proportion of Transition Substitutions within CpG Context Recovered from the Posterior Predictive Simulations (Mean over 137 Mammalian Gene Analyses).

Codon Position	Substitution Types	Without CpG	With CpG
1-2	CpG>TpG	0.14 ± 0.14	0.23 ± 0.22
2-3	CpG>TpG	0.54 ± 0.32	0.65 ± 0.42
3-1	CpG>TpG	4.33 ± 1.05	8.27 ± 2.14
1-2	CpG>CpA	0.45 ± 0.36	0.75 ± 0.57
2-3	CpG>CpA	3.19 ± 0.79	6.36 ± 1.44
3-1	CpG>CpA	0.78 ± 0.42	0.88 ± 0.54
Synonymous	CpG>TpG + CpG>CpA	7.52 ± 1.47	14.63 ± 2.85
Nonsynonymous	CpG>TpG + CpG>CpA	1.91 ± 1.02	2.51 ± 1.47

CpG->TpG at positions 3-1 drastically increase under the model with CpG hypermutability. This is entirely expected since most of these substitutions are synonymous. In fact, the proportion of nonsynonymous substitutions (transitions) at CpG sites only increases from 1.9% to 2.5% whereas the synonymous transitions jump from 7.5% to 14.6%. This is congruent with the analysis of thousands of genes between human and chimpanzee showing that ~14% of the substitutions (synonymous or nonsynonymous) are related to CpG hypermutability (Misawa and Kikuno 2009). Table 4 shows that selection at the amino-acid level severely filters the effect of CpG hypermutability (Stoltzfus and McCandlish 2017), but suggests that CU might be affected (see below).

Fourth, we investigated the dinucleotide frequencies related to CpG hypermutability (CpG, TpG, and CpA) and, as negative controls, the other dinucleotides involving the same pairs of nucleotides (GpC, GpT, and ApC). For all codon positions (i.e., 1-2, 2-3, 3-1) the negative controls are similarly, and accurately, predicted by both models, with or without CpG hypermutability (supplementary figs. S9–11*D–F*, Supplementary Material online). In contrast, introducing CpG hypermutability severely impacted the prediction of CpG, TpG, and CpA dinucleotide frequencies (supplementary figs. S10 and 11*A–C*, Supplementary Material online), except at codon position 1-2 (supplementary fig. S9*A–C*, Supplementary Material online). This is expected because almost all substitutions at these positions are nonsynonymous, hence almost exclusively predicted by the selection part of the model, which is identical between the two models. At codon positions 2-3 and 3-1, the CpG frequency is always better predicted by the model that includes CpG hypermutability (supplementary figs. S10 and 11*A*, Supplementary Material online) and are in fact very close to the observed values (mean *Z*-score of −0.58 and 0.01, respectively). The frequency of TpG at codon position 3-1, and of CpA at position 2-3, are both better predicted by the model with CpG hypermutability (supplementary figs. S11*C* and S10*B*, Supplementary Material online). As noticed for the predicted substitutions (table 4), the mutational results of CpG hypermutability are synonymous events at the codon level. For TpG (CpA) frequency at codon position 3-1 (2-3), the predictions of the two models are virtually identical because these products of CpG hypermutability are nonsynonymous. Including the CpG hypermutability therefore allows to improve the prediction of dinucleotide frequencies almost exclusively in a synonymous context. In contrast, in a nonsynonymous context, the model without CpG hypermutability appears to yield globally correct predictions.

Fifth, we compared the amino acid frequencies predicted by the two models. We did not observe major differences (supplementary fig. S12, Supplementary Material online), again, probably because this characteristic is mainly modelled by the selection part, which is shared by the two models. However, it is known that the mutational process has an impact on amino acid frequencies, through variation in GC content in mitogenomes (Foster et al. 1997), or differences between the leading and lagging strands (Rocha et al. 1999). The case of arginine constitutes a good illustration of this specific point. The frequency of arginine is overpredicted by the reference model, and underpredicted by the model including CpG hypermutability. Strikingly, arginine is the only amino acid encoded by codons having a CpG at position 1-2 (CGN, and also by codons AGR). If the selective advantage of arginine at a given position is not sufficiently strong, the high mutational pressure away from CpG can easily lead to the replacement of arginine by a less favorable amino acid. The case of arginine also demonstrates that site-specific amino-acid preferences might in fact be correlated with λ_{CpG} under the posterior, something which was ignored in our analysis, by pre-estimating amino-acid fitness parameters under the reference model (without CpG) without any subsequent correction. In this respect, one possible improvement of our approach would be to globally modulate the site-specific amino acid fitness profiles using a vector of 20 correcting factors that would be estimated at the ABC step. However, the pattern shown in supplementary figure S12, Supplementary Material online is complex. For instance, it is not clear why the model including CpG hypermutability overpredicts the frequencies of isoleucine (codons AUH) and methionine (codon AUG).

Posterior Predictive Checks and the CU Bias in Mammals

We have shown that the modelling of CpG hypermutability has a major impact on the ability to predict synonymous aspects of mammalian protein coding gene evolution. It is therefore particularly interesting to examine its effect on CU. We used posterior predictive checks to study the entropy of relative synonymous codon usage (RSCU, fig. 6) and of relative codon frequencies (RFC, supplementary fig. S13, Supplementary Material online). The results obtained with these two alternative statistics were similar and we will focus on RSCU, which is commonly used in empirical analyses of CU (e.g., Pouyet et al. 2017). The model with CpG hypermutability more accurately predicts the CU entropy observed on the empirical alignments, compared with the reference model (fig. 6): the mean *Z*-scores are −4.84 and −3.16, respectively. Since the entropy is maximal under equal use of each synonymous codon, the predicted RSCU are generally more homogeneous than the observed RSCU. A large proportion of the genes (41.6% and 20.4% for the models without and with λ_{CpG} respectively) yields very poor predictions of the entropy of RSCU with *Z*-scores under −5 (fig. 6). This suggests that

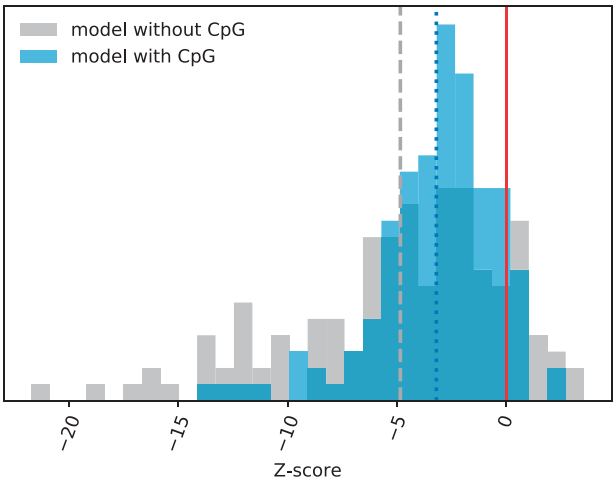

FIG. 6. Distribution of Z-scores computed from RSCU (without stop, methionine and tryptophan codons) entropy predicted under the models without (gray) and with (blue) CpG hypermutation. The vertical dashed (gray) and dotted (blue) lines represent the mean Z-scores obtained under each model respectively (i.e., without and with CpG hypermutation). The vertical solid line (red) represents the zero value.

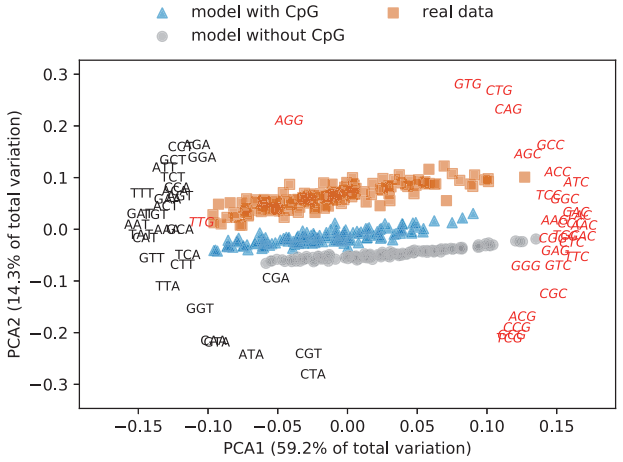

FIG. 7. Principal component analysis of the RSCU (without stop, methionine and tryptophan codons) recovered from the 137 mammalian gene alignments and from the mean RSCU predicted under models without and with CpG hypermutation. G/C-ending codons are annotated in red (italic) whereas A/T-ending codons are annotated in black.

other important determinants, such as splicing enhancer, or mRNA structure, are still missing to our modeling strategy.

To better understand the mutational or selective forces determining the small entropy of RSCU observed in mammalian protein coding genes, we performed a principal component analysis of the RSCU predicted by the two models, along with the RSCU observed in the empirical alignments (fig. 7). The first axis of the PCA explains most of the variance (59.2%), and is related to the GC3 content (the r^2 between the first axis and the GC3 of the real alignments is equal to 0.984). This is congruent with similar analyses based on a larger number of genes but restricted to *Homo sapiens* (e.g., Pouyet et al. 2017). The model without CpG hypermutation is slightly shifted to the right (fig. 7), in agreement with its GC3 overprediction (fig. 5). In contrast, the predictions of the model including CpG are comparable to real data on the first axis. The second axis explains 14.3% of the variance and strongly discriminates the real data from the predictions of the reference model, the predictions of the model with CpG hypermutability being in-between. The model that includes CpG hypermutability is, as expected from previous results, closer to the real data.

All the G/C ending codons (in red) but TTG and AGG are located to the right, in agreement with the correlation between the first axis and GC3 content. The second axis, driven by the difference between observed and predicted data, is more complex to interpret. The codons ending by CpG are all located in the lower right corner, indicating that CpG hypermutability contributes to this axis. Including λ_{CpG} is indeed necessary to explain why codons TCG, GCG, CCG, and ACG are unpreferred in humans with a RSCU of 0.05, 0.11, 0.11, 0.11, respectively (Nakamura et al. 2000), whereas G ending codons are always otherwise preferred. However, the synonymous products of transitions of these codons (NCA) do not meaningfully contribute to the second axis. In contrast, three codons ending by G (GTG, CTG, and CAG, up right corner) heavily contribute to this axis but do not seem to be linked to CpG. Codons CTA (Leucine), ATA (Isoleucine), and GTA (Valine) are also major drivers of the second axis. They are overpredicted by both models. A deficit of TpA could be due to the hypermutability of this dinucleotide (Milholland et al. 2017) or to selection against the attachment to transcription or termination factors (Burge et al. 1992). Codons for arginine are also separated on the second axis, AGR clearly on the upper part and CGN on the lower one (very weakly for CGG). This is likely related to CpG hypermutability, which will erode CGN codons towards TGN and CAN. The possibly complicated evolutionary path between CGN and AGR codons to conserve functionally important arginines could be responsible for the surprising position of codon AGG along the first axis. In summary, the PCA of RSCU (fig. 7) demonstrates that including CpG hypermutability into the mutation–selection model leads to an improved prediction of CU but that other characteristics (e.g., TpA) are poorly predicted, requiring future additions in the mutation and/or selection part(s) of the model.

Conclusions and Future Directions

We have proposed a new approach, CABC, that combines MCMC and ABC to simultaneously handle high-dimensional parameter vectors and site-interdependent substitution processes. We have shown that this approach allows accurate estimation of the level of transition hypermutability in the CpG context. Our analysis confirms that CpG hypermutability is prevalent in mammals and variable among loci. This proof of concept of the CABC methodology opens new perspectives towards improved mutation–selection models better able to tease apart the relative role of these two evolutionary forces.

We used a simple implementation for ABC, where *SS* were manually selected and the posterior distribution was approximated with the RS algorithm followed by the use of a LRM. This appears to be sufficient to accurately estimate the rate of CpG hypermutation, λ_{CpG}, although some biases and/or inaccuracies were observed for other parameters (e.g. λ_{TBL} and λ_{ω_*}). From there, the method could be improved in several respects. For instance, RS could be replaced by MCMC (Marjoram et al. 2003) or by sequential Monte Carlo (Sisson et al. 2007). Similarly, LRM could be replaced by other regression models (e.g., random forest; Raynal et al. 2017, unpublished work), in the hope of getting closer to the true posterior and potentially reducing the computation burden. The choice of *SS* could also be reconsidered, for instance by computing the number of substitutions by maximum parsimony instead of simply counting the number of observed pairwise differences, which might improve the estimation of λ_{TBL}. Perhaps more importantly, the choice of *SS* could be performed automatically (Prangle, Fearnhead, et al. 2014). The Random Forest ABC (Pudlo et al. 2016) may be particularly well suited for sequence data, for which hundreds of *SS* can in principle be contemplated. Finally, one specific aspect of strategy that was adopted here, that is, introducing modulator parameters, which are estimated at the ABC step, to correct for the fact that most nuisance parameters are sampled under the reference posterior distribution (i.e., under $\lambda_{CpG} = 1$), could be generalized to other aspects of the model, in particular, to amino-acid frequencies across the proteome (as illustrated by the case of arginine, supplementary fig. S12, Supplementary Material online).

The CABC approach will make it possible to develop complex mutation–selection models handling several of the well identified and complex features of mutation and selection processes. Concerning mutation, context-dependent effects are clearly understudied in molecular evolution, mostly for computational reasons, and despite the fact that the prevalence of such effects is widely recognized (Siepel and Haussler 2004; Nevarez et al. 2010; Seplyarskiy et al. 2017; Guo et al. 2018). In addition to CpG hypermutability, TpA hypermutation (Milholland et al. 2017) or more complex context-dependent mutational pattern, such as inferred from the large number of de novo mutations discovered through the sequencing of trios (e.g., Francioli et al. 2015; Wong et al. 2016; Jonsson et al. 2017), could be further investigated. Concerning selection, the perspectives are broader, including, among other things, selection against mono-nucleotide repeats, mRNA secondary structure, motif for RNA binding proteins (e.g., splicing enhancers) and obviously protein structure. Such improved models should have a broad applicability in molecular evolutionary studies, by making it possible to tease apart the role of mutation, purifying, and diversifying selection in the evolution of genomic sequences.

Materials and Methods

Data Sets and Tree Topology

All the 137 mammalian gene alignments used in this work as well as the mammalian tree were recovered from Laurin-Lemay et al. (2018), and both are available via the GitHub repository (https://github.com/Simonll/LikelihoodFreePhylogenetics/; last accessed September 12, 2018).

Codon Substitution Models

To mechanistically disentangle mutation from selection processes, we used the codon substitution model of Rodrigue et al. (2010) with the modification of Rodrigue and Lartillot (2017), as implemented in Phylobayes-MPI (Lartillot et al. 2013; Rodrigue and Lartillot 2014). Let us briefly recall the parameterization of this reference model, denoted as M[GTR]-S[NCatAA*]. Branch lengths are free parameters, while the tree topology is kept fixed. The mutational part of the model, M[GTR], is modelled with the general-time-reversible approach (Lanave et al. 1984) using 10 parameters (8 degrees of freedom) and assumes a point mutation process between codon *a* to codon *b*. Stop codons are prohibited (i.e., have zero probability). The mutational process act identically on all codon positions (1, 2, and 3), whereas codons *a* and *b* differ at the *c*th position. The nucleotide propensities, are defined as $\varphi = (\varphi_n)_{1 \leq n \leq 4}$, with $\sum_{n=1}^{4} \varphi_n = 1$ and the nucleotide exchangeabilities are defined as $\varrho = (\varrho_{mn})_{1 \leq m,n \leq 4}$, with $\sum_{1 \leq m < n \leq 4} \varrho_{mn} = 1$. The selective part of the model, S[NCatAA*], acts at the amino acid level. The amino acid relative scaled fitness *profiles*, or NCatAA from S[NCatAA*], are elements of a Dirichlet process (Rodrigue et al. 2010). The Dirichlet process is a nonparametric method, controlled by a few hyper-parameters, which allows to approximate any unknown distribution (Ferguson 1973). Here the dimensionality of the latent variables is huge (e.g., the number of profiles times 20 amino acids) noting that there is in average $\sim 70.80 \pm 22.42$ profiles to deal with when working with the mammalian gene alignments studied here. The *K* profiles are defined as vectors $\psi = (\psi_l^{(k)})_{1 \leq l \leq 20, 1 \leq k \leq K}$. Site specific allocation of the *K* profiles is specified for the length of the gene, *N*, via the vector $z = (z_i)_{1 \leq i \leq N}$. Therefore, the scaled selection coefficient for nonsynonymous events, $S_{ab}^{(i)}$, is obtained as in Yang and Nielsen (2008):

$$S_{ab}^{(i)} = \ln\left(\frac{\psi_{f(b)}^{(z_i)}}{\psi_{f(a)}^{(z_i)}}\right), \tag{11}$$

where $f(a)$ returns an index, from 1 to 20, of the amino acid encoded by codon *a*. The value of $S_{ab}^{(i)}$ in turn defines a fixation factor, denoted $h(S_{ab}^{(i)})$, and calculated as

$$h(S_{ab}^{(i)}) = \frac{S_{ab}^{(i)}}{1 - e^{-S_{ab}^{(i)}}}. \tag{12}$$

A deviation parameter, ω_*, or $*$ from S[NCatAA*], was recently introduced by Rodrigue and Lartillot (2017) to capture the excess or the deficit of nonsynonymous rates with respect to the purifying selection modelled by the amino acid fitness profiles, corresponding to Darwinian selection or other forms of purifying selection (e.g., the secondary structure of mRNA or the 3D structure of protein), respectively.

The substitution rate matrix Q of the reference model has entries of the form:

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if syn.,} \\ \varrho_{a_c b_c} \varphi_{b_c} h(S_{ab}^{(i)}) \omega_*, & \text{if non} - \text{syn.} \end{cases} \quad (13)$$

To capture the transition mutation rate in the CpG context (i.e., CpG > TpG or CpG > CpA), we extended the mutation component M[GTR] of the reference model by including an across-site dependent parameter, λ_{CpG}. The Q matrix has now entries of the form:

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if syn. tr. or ts. non} - \text{CpG,} \\ \varrho_{a_c b_c} \varphi_{b_c} \lambda_{CpG}, & \text{if syn. ts. CpG,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}^{(i)}), & \text{if non} - \text{syn. tr. or ts. non} - \text{CpG,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}^{(i)}) \lambda_{CpG}, & \text{if non} - \text{syn. ts. CpG.} \end{cases} \quad (14)$$

Overview of CABC

The hypermutability of C in the CpG context introduces across-site dependency, making the computation of the likelihood intractable. CABC eschews this difficulty by inferring the high-dimensional parameter vectors using a standard MCMC with the model M[GTR]-S[NCatAA*], that is, with $\lambda_{CpG} = 1$, and then by using ABC to infer the posterior distribution of the model with CpG hypermutation, assuming the values of the high-dimensional parameter vectors previously estimated. More precisely, the parameter vectors θ_{sc} and θ_{wc} of equation (9) consist of the propensities and exchangeabilities of the GTR matrix (φ and ϱ) plus three modulators (λ_{ω_*}, λ_{TBL}, and λ_{ROOT}) and of the branch lengths plus the amino acid fitness profiles, respectively. As explained above, the parameters of θ_{wc} are of high dimensionality and cannot be accurately inferred by ABC, whereas the parameters of θ_{sc} are strongly correlated with the site-interdependent parameter λ_{CpG} and therefore cannot be inferred by MCMC. As a result, θ_{wc} is first obtained using MCMC assuming $\lambda_{CpG} = 1$ and λ_{CpG} and θ_{sc} are obtained using ABC conditional on θ_{wc} as formulated CABC equation (9). Priors are defined for λ_{CpG} and θ_{sc} before running the ABC step.

MCMC Part of CABC

We applied the reference model implemented in PhyloBayes-MPI on the 137 alignments, composed of 39 placentals, available from Laurin-Lemay et al. (2018). All the analyses were carried under fixed topology previously obtained by Laurin-Lemay et al. (2018). The analyses were also conducted on subparts of the mammalian tree: Glires (7 species), Laurasiatheria (14 species), and Primates (12 species). Convergence was first visually assessed using two independent chains, and then by computing the effective size of the parameters. The priors used under the reference model are listed in Rodrigue and Lartillot (2014). Parameters from θ_{wc} are drawn from the posterior distribution estimated under the reference model using MCMC (i.e., assuming $\lambda_{CpG} = 1$).

ABC Part of CABC

In addition to the 10 GTR parameters, which are expected to be strongly correlated to λ_{CpG}, θ_{sc} includes λ_{TBL}, a modulator serving as a multiplicative parameter for every branch length. As the mutation–selection equilibrium was disrupted by the new parameterization, the tree length (Total Branch Length or TBL), which is measured in the number of mutations per site, will likely increase because of the additional mutations proposed at CpG sites (when $\lambda_{CpG} > 1$). Also included in θ_{sc} is λ_{ω_*}, a multiplicative modulator to ω_*, to respond to changes in nonsynonymous rates that might emerge when accommodating CpG hypermutability. It is difficult to anticipate the value of λ_{ω_*}, given the potentially complex interplay between parameter values that could be produced from modeling CpG hypermutation. Finally, θ_{sc} includes λ_{ROOT}, a multiplicative parameter to fix the exact position of the root of the tree between the in- and out-group. Since we used a model, M[GTR+ts-CpG]-S[NCatAA*], that makes the process nontime-reversible, the position of the root, in our case on the branch separating Afrotheria (set as the out-group from Xenarthra + Euarchontoglires + Laurasiatheria), influences the output. It is difficult to know whether the phylogenetic signal will be sufficient to precisely estimate λ_{ROOT}.

Priors used are noninformative except for the nucleotide exchangeabilities, or ϱ. We informed the model that the transition rates are in average two times higher than transversions (Wakeley 1996) to reduce the dimensionality of the ABC search. The use of noninformative priors makes the rejection of the null hypothesis more reliable at the expense of computational time.

$$\lambda_{CpG} \sim \log_{10} \text{Uniform}[0.1, 10]$$

$$\varrho_{ts} \sim \text{Gamma}[\alpha = 1, \beta = 1]$$

$$\varrho_{tr} \sim \text{Gamma}[\alpha = 2, \beta = 1]$$

$$\varphi \sim \text{Dirichlet}[1, 1, 1, 1]$$

$$\lambda_{\omega_*} \sim \log_2 \text{Uniform}[0.5, 2]$$

$$\lambda_{TBL} \sim \log_2 \text{Uniform}[0.5, 2]$$

$$\lambda_{ROOT} \sim \text{Uniform}[0, 1]$$

The simulator developed in Laurin-Lemay et al. (2018) allows one to generate sequence alignments from the model with CpG across-site dependency along a phylogenetic tree.

It was modified to work in parallel and to compute distances between the vectors of SS recovered from simulated and true alignments. Concretely, the simulator program generates a reference table of SS along with the parameter values (SS, λ_{CpG}, θ_{sc}, θ_{wc}), ordered by increasing distance values. The ABC rejection sampling algorithm (RS; Pritchard et al. 1999) was implemented into the simulation package. One can run the RS for a defined number of simulations (i.e., sampling size) and select the best simulations (i.e., tolerance level) on the basis of the distances computed for each simulation (see below). The selected simulations correspond to the *RS table* used to approximate the posterior distribution. The program is accessible via the GitHub repository (https://github.com/Simonll/LikelihoodFreePhylogenetics/; last accessed September 12, 2018). The two steps procedure (MCMC followed by CABC) developed here takes for a single gene analysis ~10 h on an AMD Opteron 6172 using 12 cores (for 100,000 simulations).

The SS are key to capturing the relevant information in the data (Fu and Li 1997; Tavare et al. 1997; Weiss and von Haeseler 1998; Pritchard et al. 1999). Preliminary analyses were performed to select among >200 possible SS those that are the most useful in discriminating different values of λ_{CpG} and θ_{sc}. Thirteen SS were selected to summarize the alignments. First, we used the relative dinucleotide frequency of CpG, TpG, and CpA (SS_{C3pG1}, SS_{T3pG1}, SS_{C3pA1}) at the third and first positions of two adjacent codons, mainly in order to fit the λ_{CpG} parameter. Second, the frequency of four nucleotides at the GC3 (SS_{A3}, SS_{C3}, SS_{G3}, SS_{T3}) was considered, mainly to fit the nucleotide propensities, or φ. Third, the sum over all the possible pairs of sequences of the absolute numbers of differences were computed at the nucleotide level for each possible unordered pair of nucleotides, leading to six SS ($SS_{A<>C}$, $SS_{A<>G}$, $SS_{A<>T}$, $SS_{C<>G}$, $SS_{C<>T}$, $SS_{G<>T}$); they should mainly allow to fit the exchangeability parameters (ϱ), but also λ_{TBL}. Fourth, we also computed the sum over all the possible pairs of sequences of the absolute number of all nonsynonymous differences indiscriminately (SS_{NS}), with the aim to fit λ_{TBL} and λ_{ω_*}. We did not find any SS informative for λ_{ROOT}. In this study, the ordering of simulations was achieved by using the squared Euclidean distance. All the 13 SS were log base 2 transformed to avoid over representing SS with large values (e.g., $SS_{A<>C} \sim 10^5$ while SS_{C3pG1} are $\sim 10^{-2}$) when applying the distance function.

Two sampling sizes (10^5 or 10^6 simulations) were investigated under the RS algorithm. To approximate the posterior distribution of λ_{CpG} and θ_{sc}, we selected the best simulations following different tolerance levels: we kept the best 10% or 1% for the sampling size of 10^5 or the best 1% or 0.1% for 10^6. Given the large combinatorics of parameter values for λ_{CpG} and θ_{sc}, it is likely that the RS algorithm would require a much larger sampling size to accurately infer the posterior distribution (Barber et al. 2015). To get closer to the true posterior distribution, we modeled the relationship between the parameter values sampled during the CABC (i.e., λ_{CpG}, θ_{sc}) as the response variables and the SS present in the RS table of the best simulations as the predictors of a regression model as introduced by Beaumont et al. (2002). More specifically, we applied the nonparametric weighted multiple linear regression model (previously identified as LRM), which also accounts for heteroskedasticity, as proposed by Blum and Francois (2010) and available in the ABC package (Csilléry et al. 2012) from R CRAN (R Core Team 2017). The weighted scheme, done for each entry of the RS table, are obtained by applying an Epanechnikov kernel to the Euclidean distances computed. In other words, the weights are maximal for the entries with the smallest distances, and minimal for the biggest distances. This ensures that the LRM optimizes its parameters from the best samples present in the RS table.

Validation of CABC

To validate the new CABC method we analyzed alignments simulated using known parameter values. To ensure the realism of the simulated alignments, we drew the parameter values from the posteriors obtained under the reference model (10 genes) along with five values of λ_{CpG} (0.5, 1, 2, 4, and 8). The same mammalian tree topology was used for the validation and the analyses, taken from Laurin-Lemay et al. (2018). The 10 genes (see supplementary table S3, Supplementary Material online for details) were selected among the 137 used in Laurin-Lemay et al. (2018) to represent the variation of the GC content found within mammalian genomes (supplementary table S3, Supplementary Material online) and to have a sequence length of ~1000 codons (a compromise between the amount of evolutionary signal and the computational burden). More specifically, for each gene, 100 sets of parameter values were drawn from the posterior distribution and used for five simulation sets (i.e., the five values of λ_{CpG}). This leads to a total of 5,000 ($5 \times 10 \times 100$) DNA sequence alignments to benchmark the CABC.

This validation framework enables us to investigate the reliability of inferences conducted in this study, as a function of the various settings of our approximation methods. Specifically, we explored the number of simulations (i.e., 10^5 or 10^6), the tolerance level to be applied (10%, 1%, or 0.1%), as well as the use of regression models. The tolerance levels were chosen to have RS table of at least 1,000 points.

Two standard methods were used to evaluate the accuracy of CABC. First, we quantified estimation error for each parameter fitted under the CABC procedure by using the RMSE as used by Beaumont et al. (2002):

$$RMSE_i = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{\theta}_i - \theta_{ij}}{\hat{\theta}_i}\right)^2, \qquad (16)$$

where $RMSE_i$ corresponds to the average error computed for the parameter i (e.g., λ_{CpG}). The RMSE is obtained by averaging the relative squared discrepancy between the true parameter value ($\hat{\theta}_i$) used for generating the simulated alignment and the N parameter values (θ_{ij}) from the approximated posterior recovered under the CABC procedure when analyzing that very same simulated alignment. Note that the error is calculated relative to the scale of the true parameter value. A global RMSE can be obtained by averaging the total error

computed independently for each parameter ($RMSE_i$) over all the analyses of a validation set (i.e., defined upon the five λ_{CpG} values).

We also investigated the coverage property of each parameter (Cook et al. 2006; Fearnhead and Prangle 2012; Prangle, Blum, et al. 2014) fitted under the CABC using the P–P plots. The coverage was investigated with a set of 99 credibility intervals ($1-\alpha$), where α is ranging from 1 to 99%, and increased by steps of 1%. More precisely we computed the frequency for which the true parameter value was found within each credibility interval (1,000 replicates per λ_{CpG} values) and compared those frequencies to the expected ones ($1-\alpha$). In other words, when a 95% credibility interval ($1 - 0.05$) is used, we should recover the true value within this credibility interval 95% of the times. Conformity between the coverage recovered was assessed using a two sided Kolmogorov–Smirnov test available from SciPy (Eric et al. 2018). The rejection of the null hypothesis (i.e., the coverage is expected to be uniform along all credible intervals tested) would demonstrate that there is a bias in the CABC analyses.

We also evaluated the impact of the two approximations of CABC on its accuracy. To study the choice of the parameters to be included in θ_{sc}, we transferred the GTR parameters into θ_{wc}. The strongly correlated (to λ_{CpG}) parameter vector, θ_{sc}, is now only composed of λ_{TBL}, λ_{omega_*}, and λ_{ROOT}. The second approximation (that the parameters contained in θ_{wc} might not be uncorrelated to λ_{CpG}) was investigated by fixing the values of the θ_{wc} parameters to the true values. In other words, instead of drawing θ_{wc} from the posterior distribution of the simulated alignments under the reference model, we took the θ_{wc} values that have been used to generate the simulated alignments.

Application to Mammalian Protein Coding Genes

We would like to evaluate the ability of CABC to estimate hypermutability in the CpG context in the cases of mammalian protein coding genes. All the 137 genes of Laurin-Lemay et al. (2018) were analyzed with CABC using a sampling size of 10^6 and a tolerance level of 0.1%. The topology of Laurin-Lemay et al. (2018) was used for all the genes. We then carried out the hypothesis testing related to CpG hypermutability (i.e., $\lambda_{CpG} > 1$), for credibility intervals of 95% and 99%. We further investigated the impact of the prior on λ_{CpG} parameter by comparing our results to the ones obtained by using a broader prior on λ_{CpG} (i.e., [1/50, 50]). We also explored the heterogeneity of CpG hypermutability over the placental tree by analyzing three clades independently. For each analysis (i.e., Glires, Laurasiatheria, and Primates) we sampled the root position using λ_{ROOT} parameter on the branch connecting *Dipodomys* and the rest of the Glires (7 species), on the branch connecting *Mustela* and the rest of Laurasiatheria (14 species), and on the branch connecting *Callithrix* and the rest of the Primates (12 species).

Posterior Predictive Checks

Posterior predictive analysis is a powerful framework to evaluate model properties (Gelman et al. 2013). Ten replicates were generated per posterior sample under the model without and with CpG hypermutability. First, we compared the model predictions on the basis of substitution histories generated over simulations. Specifically, we quantified the total number of substitutions and the proportion of each substitution types as defined by each unique pair of nucleotide substitution (e.g., A–C) or by their effect at the amino acid level (synonymous vs. nonsynonymous). We also tracked substitutions related to the CpG context (i.e., CpG to TpG and CpG to CpA) from all codon position (1-2, 2-3, and 3-1). We computed various *SS* from simulated alignments to compare model fit using *Z*-scores. Among the key features investigated, we looked at the GC3 content, the entropy of the RSCU (relative synonymous CU), the entropy of the RCF (relative codon frequencies), the relative dinucleotide frequencies for codon positions 1-2, 2-3, and 3-1, as well as the amino acid frequencies. We also performed a principal component analysis using the VEGAN package (Oksanen et al. 2017) from R CRAN (R Core Team 2017) on the matrix of the RSCU recovered from the true alignments and from alignments generated by both models.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the French Laboratory of Excellence project entitled TULIP (ANR-10-LABX-41; ANR-11-IDEX-0002-02), and by the Natural Sciences and Engineering Research Council of Canada. Computations were made on the supercomputer Mammouth-parallel from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the Ministère de l'Économie, de la Science et de l'Innovation du Québec-Nature et technologies (FRQ-NT). S.L.L. is the recipient of a Fonds de la Recherche en Santé du Québec (FRSQ) Graduate Scholarship.

References

Arndt P, Burge C, Hwa T. 2003. DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol*. 10(3–4):313–322.

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

Barber S, Voss J, Webster M. 2015. The rate of convergence for approximate Bayesian computation. *Electron J Stat*. 9:80–105.

Beaumont M, Zhang W, Balding D. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Berard J, Gueguen L. 2012. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Syst Biol*. 61:510–521.

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

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

Blum M, Francois O. 2010. Non-linear regression models for approximate Bayesian computation. *Stat Comput*. 20:63–73.

Burge C, Campbell A, Karlin S. 1992. Over-representation and under-representation of short oligonucleotides in DNA-sequences. *Proc Natl Acad Sci USA*. 89:1358–1362.
Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* 141(2):355–367.
Chen S, Lee W, Hottes A, Shapiro L, McAdams H. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA*. 101(10):3480–3485.
Christensen O, Hobolth A, Jensen J. 2005. Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *J Comput Biol*. 12:1166–1182.
Christensen O. 2006. Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Stat Appl Genet Mol Biol*. 5:1–29.
Cook S, Gelman A, Rubin D. 2006. Validation of software for Bayesian models using posterior quantiles. *J Comput Graph Stat*. 15:675–692.
Csilléry K, François O, Blum M. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol*. 3:475–479.
Drummond D, Wilke C. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
Duret L, Arndt P. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.
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. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*. 12(6): 640–649.
Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*. 17(2): 109–121.
Eric J, Travis O, Pearu P, et al. 2018. SciPy: Open source scientific tools for Python. http://www.scipy.org/ (last accessed: September 12, 2018)
Ermolaeva M. 2001. Synonymous codon usage in bacteria. *Curr Issues Mol Biol*. 3(4): 91–97.
Fearnhead P, Prangle D. 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J R Stat Soc Series B Stat Methodol*. 74:419–474.
Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*. 25:471–492.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.
Ferguson T. 1973. A Bayesian analysis of some nonparametric problems. *Ann Stat*. 1:209–230.
Filipski J, Thiery J, Bernardi G. 1973. Analysis of bovine genome by cs2so4-ag+ density gradient centrifugation. *J Mol Biol*. 80:177–197.
Foster P, Jermiin L, Hickey D. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol*. 44:282–288.
Francioli L, Polak P, Koren A, Menelaou A, Chun S, Renkens I, van Duijn C, Swertz M, Wijmenga C, van Ommen G, et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*. 47(7): 822.
Frazier D, Robert C, Rousseau J. 2017. Model Misspecification in ABC: Consequences and Diagnostics. *1708.01974v1 [math.ST]*. https://arxiv.org/pdf/1708.01974.
Fu YX, Li WH. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol Biol Evol*. 14:195–199.
Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Mol Biol Evol*. 35:1092–1103.
Gelman A, Carlin J, Stern H, Rubin D. 2013. Bayesian data analysis. 3rd ed. Boca Raton (FL): Chapman and Hall/CRC.
Glemin S, Arndt P, Messer P, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res*. 25:1215–1228.
Guo Y, Chang M, Huang W, Ooi W, Xing M, Tan P, Skanderup A. 2018. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat Commun*. 9(1):1520.
Halpern A, Bruno W. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*. 15:910–917.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. *J Mol Evol*. 22:160–174.
Hastings W. 1970. Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97.
Hilton S, Doud M, Bloom J. 2017. phydms: software for phylogenetic analyses informed by deep mutational scanning. *Peerj* 5:e3657.
Hobolth A, Nielsen R, Wang Y, Wu F, Tanksley S. 2006. CpG plus CpNpG analysis of protein-coding sequences from tomato. *Mol Biol Evol*. 23:1318–1323.
Hobolth A. 2008. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *J Comput Graph Stat*. 17:138–162.
Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*. 12(11): 756–766.
Huttley G, Yap V. 2012. Robust estimation of natural selection using parametric codon models. In: Cannarozzi GM, Schneider A, editors. Codon evolution: mechanisms and models, book section 8. Oxford: OUP.
Huttley G. 2004. Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. *Mol Biol Evol*. 21:1760–1768.
Hwang D, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*. 101:13994–14001.
Jensen J, Pedersen A. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv App Prob*. 32:499–517.
Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson M, Hjorleifsson K, Eggertsson H, Gudjonsson S, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* 549(7673):519.
Katzman S, Capra J, Haussler D, Pollard K. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol*. 3:614–626.
Keightley P, Eory L, Halligan D, Kirkpatrick M. 2011. Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate bayesian computation. *Genetics* 187:1153–U268.
Kessler M, Dean MD. 2014. Effective population size does not predict codon usage bias in mammals. *Ecol Evol*. 4:3887–3900.
Kleinman C, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol*. 27:1546–1560.
Knight R, Freeland S, Landweber L. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*. 2:research0010.
Kousathanas A, Leuenberger C, Helfer J, Quinodoz M, Foll M, Wegmann D. 2016. Likelihood-free inference in high-dimensional models. *Genetics* 203(2): 893.
Krasovec R, Richards H, Gifford D, Hatcher C, Faulkner K, Belavkin R, Channon A, Aston E, McBain A, Knight C. 2017. Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biol*. 15(8):e2002731.
Lanave C, Preparata G, Sacone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol*. 20(1): 86–93.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095–1109.
Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*. 62:611–615.

Lartillot N. 2013. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol Biol Evol.* 30:489–502.
Laurin-Lemay S, Philippe H, Rodrigue N. 2018. Multiple factors confounding phylogenetic detection of selection on codon usage. *Mol Biol Evol*. 35(6):1463–1472.
Lee H, Kishino H, Rodrigue N, Thorne J. 2016. Grouping substitution types into different relaxed molecular clocks. *Proc Natl Acad Sci USA*. 371(1699). pii: 20150141.
Lee H, Rodrigue N, Thorne J. 2015. Relaxing the molecular clock to different degrees for different substitution types. *Mol Biol Evol*. 32:1948–1961.
Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-content of synonymous codons profoundly influences amino acid usage. *G3 (Bethesda)* 5:2027–2036.
Lindsay H, Yap V, Ying H, Huttley G. 2008. Pitfalls of the most commonly used models of context dependent substitution. *Biol Direct* 3:52.
Maharjan R, Ferenci T. 2017. A shifting mutational landscape in 6 nutritional states: stress-induced mutagenesis as a series of distinct stress input-mutation output relationships. *PLoS Biol.* 15(6): e2001477.
Marjoram P, Molitor J, Plagnol V, Tavare S. 2003. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*. 100(26): 15324–15328.
McVean G, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics* 157:245–257.
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys.* 21:1087–1092.
Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. 2017. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun.* 8:15183.
Misawa K, Kikuno R. 2009. Evaluation of the effect of CpG hypermutability on human codon substitution. *Gene* 431:18–22.
Misawa K. 2011. A codon substitution model that incorporates the effect of the GC contents, the gene density and the density of CpG islands of human chromosomes. *BMC Genomics.* 12:397.
Mugal C, Arndt P, Holm L, Ellegren H. 2015. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. *G3 (Bethesda)* 5:441–447.
Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA*. 84:166–169.
Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. 28:292.
Nevarez P, DeBoever C, Freeland B, Quitt M, Bush E. 2010. Context dependent substitution biases vary within the human genome. *BMC Bioinformatics.* 11(1): 462.
Nielsen R, Bauer DuMont V, Hubisz M, Aquadro C. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. *Mol Biol Ecol.* 24:228–235.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.
Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara R, Simpson G, Solymos P, et al. 2017. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan (last accessed: September 12, 2018).
Pedersen A, Wiuf C, Christiansen F. 1998. A codon-based model designed to describe lentiviral evolution. *Mol Biol Evol.* 15(8):1069–1081.
Pouyet F, Bailly-Bechet M, Mouchiroud D, Gueguen L. 2016. SENCA: a multilayered codon model to study the origins and dynamics of codon usage. *Genome Biol Evol.* 8:2427–2441.
Pouyet F, Mouchiroud D, Duret L, Semon M. 2017. Recombination, meiotic expression and human codon usage. *Elife* 6. pii: e27344.
Prangle D, Blum M, et al. 2014. Diagnostic tools for approximate Bayesian computation using the coverage property. *Aust N Z J Stat.* 56(4): 309–329.
Prangle D, Fearnhead P, et al. 2014. Semi-automatic selection of summary statistics for ABC model choice. *Stat Appl Genet Mol Biol.* 13(1): 67–82.
Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*. 16:1791–1798.
Pudlo P, Marin J, Estoup A, Cornuet J, Gautier M, Robert C. 2016. Reliable ABC model choice via random forests. *Bioinformatics* 32(6): 859–866.
R Core Team 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing.
Raynal L, Marin J, Pudlo P, Ribatet M, Robert C, Estoup A. 2017. ABC random forests for Bayesian parameter inference. *1605.05537v4 [stat.ME].* https://arxiv.org/pdf/1605.05537.
Robinson D, Jones D, Kishino H, Goldman N, Thorne J. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*. 20:1692–1704.
Rocha E, Danchin A, Viari A. 1999. Universal replication biases in bacteria. *Mol Microbiol*. 32(1): 11–16.
Rodrigue N, Kleinman C, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol*. 26:1663–1676.
Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–217.
Rodrigue N, Lartillot N, Philippe H. 2008. Bayesian comparisons of codon substitution models. *Genetics* 180(3):1579–1591.
Rodrigue N, Lartillot N. 2012. Monte Carlo computational approaches in Bayesian codon substitution modeling. In: Cannarozzi GM and Schneider A, editors. Codon evolution: mechanisms and models, book section 4, Oxford: OUP. p. 45–59.
Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation–selection models within the PhyloBayes-MPI package. *Bioinformatics* 30:1020–1021.
Rodrigue N, Lartillot N. 2017. Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation–selection codon substitution model. *Mol Biol Evol.* 34:204–214.
Rodrigue N, Philippe H, Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol.* 23:1762–1775.
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.
Rodrigue N, Philippe H. 2010. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet.* 26:248–252.
Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* 193:557–564.
Seplyarskiy V, Andrianova M, Bazykin G. 2017. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Res.* 27:175–184.
Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol.* 21:468–488.
Sisson S, Fan Y, Tanaka MM. 2007. Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*. 104(6): 1760–1765.
Stoltzfus A, McCandlish D. 2017. Mutational biases influence parallel adaptation. *Mol Biol Evol*. 34(9): 2163–2172.
Sueoka N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci USA*. 47:1141.
Sueoka N. 1962. On the genetic basis of variation and hetrogeneity of DNA base composition. *Proc Natl Acad Sci USA*. 48:582–592.
Suzuki Y, Gojobori T, Kumar S. 2009. Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection

operating at the amino acid sequence level. *Mol Biol Evol*. 26:2275–2284.
Tamuri A, dos Reis M, Goldstein R. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation–selection models. *Genetics* 190:1101–1115.
Tamuri AU, Goldman N, dos Reis M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* 197:257–271.
Tavare S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518.
Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141(2): 344–354.
Van den Eynden J, Larsson E. 2017. Mutational signatures are critical for proper estimation of purifying selection pressures in cancer somatic mutation data when using the dN/dS metric. *Front Genet*. 8:74.
Wakeley J. 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol Evol*. 11:158–163.
Wang H, Minh Q, Susko E, Roger A. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol*. 67:216–235.
Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. *Genetics* 149:1539–1546.
Wong W, Solomon B, Bodian D, Kothiyal P, Eley G, Huddleston K, Baker R, Thach D, Iyer R, Vockley J, et al. 2016. New observations on maternal age effect on germline de novo mutations. *Nat Commun*. 7:10486.
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.
Ying H, Huttley G. 2011. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome Biol Evol*. 3:938–949.