Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution

Tomotaka Matsumoto,* Hiroshi Akashi,*,†,1 and Ziheng Yang‡,§,1

*Division of Evolutionary Genetics, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan, †Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Mishima, Shizuoka 411-8540, Japan, ‡Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, and §Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, United Kingdom

ORCID IDs: 0000-0002-3456-7553 (T.M.); 0000-0001-5325-6999 (H.A.); 0000-0003-3351-7981 (Z.Y.)

ABSTRACT Inference of gene sequences in ancestral species has been widely used to test hypotheses concerning the process of molecular sequence evolution. However, the approach may produce spurious results, mainly because using the single best reconstruction while ignoring the suboptimal ones creates systematic biases. Here we implement methods to correct for such biases and use computer simulation to evaluate their performance when the substitution process is nonstationary. The methods we evaluated include parsimony and likelihood using the single best reconstruction (SBR), averaging over reconstructions weighted by the posterior probabilities (AWP), and a new method called expected Markov counting (EMC) that produces maximum-likelihood estimates of substitution counts for any branch under a nonstationary Markov model. We simulated base composition evolution on a phylogeny for six species, with different selective pressures on G+C content among lineages, and compared the counts of nucleotide substitutions recorded during simulation with the inference by different methods. We found that large systematic biases resulted from (i) the use of parsimony or likelihood with SBR, (ii) the use of a stationary model when the substitution process is nonstationary, and (iii) the use of the Hasegawa-Kishino-Yano (HKY) model, which is too simple to adequately describe the substitution process. The nonstationary general time reversible (GTR) model, used with AWP or EMC, accurately recovered the substitution counts, even in cases of complex parameter fluctuations. We discuss model complexity and the compromise between bias and variance and suggest that the new methods may be useful for studying complex patterns of nucleotide substitution in large genomic data sets.

KEYWORDS ancestral reconstruction; codon usage; GC content; nonstationary models; nucleotide substitution; stochastic mapping

IF the gene or genomic sequences in extinct ancestral species were known, they could be compared with sequences from modern species to identify the changes that have occurred on every branch of the phylogeny and at every site of the gene sequence. Such detailed information would be extremely valuable for making inferences concerning the processes and causes of molecular sequence evolution. As ancestral sequences are unknown but can be inferred in a phylogenetic analysis of modern sequences using phylogenetic methods such as parsimony (Fitch 1971; Hartigan 1973) and likelihood (*i.e.*, Bayesian) (Yang *et al.* 1995; Koshi and Goldstein 1996), it has appeared very natural to molecular evolutionists to use such reconstructed ancestral sequences as if they were real observed data. This approach has been extremely popular throughout the history of the field of molecular evolution. In the early 1960s, Dayhoff *et al.* (1965) used parsimony-like ideas to count amino acid changes on the phylogeny to construct the famous PAM matrices of amino acid substitution. The same approach was used to construct the JTT matrix (Jones *et al.* 1992) and to estimate relative rates of nucleotide substitution (Gojobori *et al.* 1982). Other uses of the approach have included counting synonymous and nonsynonymous substitutions on the phylogeny to infer adaptive protein evolution

doi: 10.1534/genetics.115.177386
Manuscript received February 9, 2015; accepted for publication April 28, 2015; published Early Online May 6, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177386/-/DC1.
[1]Corresponding authors: Division of Evolutionary Genetics, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan. E-mail: hiakashi@lab.nig.ac.jp; and Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, United Kingdom. E-mail: z.yang@ucl.ac.uk

affecting particular lineages (Messier and Stewart 1997; Zhang *et al.* 1997) or sites (Fitch *et al.* 1991; Suzuki and Gojobori 1999), inferring changes in nucleotide or amino acid compositions (Duret *et al.* 2002; Gaucher *et al.* 2003, 2008; Khelifi *et al.* 2006; Groussin and Gouy 2011; Aoki *et al.* 2013), and detecting coevolving nucleotides or amino acids (*e.g.*, Shindyalov *et al.* 1994; Tuffery and Darlu 2000; Osada and Akashi 2012; Liao *et al.* 2014). In analysis of population data or data from closely related species, use of an outgroup species to "polarize" the changes to identify the ancestral and derived nucleotide states (*e.g.*, Lohse and Barton 2011) is based on the same idea. Ancestral reconstruction has also had a long history of application in studies of codon usage in *Drosophila* (*e.g.*, Akashi 1995; Eanes *et al.* 1996; Kilman 1999; Begun 2001; Takano 2001; Bauer Dumont *et al.* 2004; Comeron 2005; Presgraves 2005; Gardiner *et al.* 2008; Haddrill *et al.* 2008; Bauer Dumont *et al.* 2009; Terekhanova *et al.* 2013).

However, reconstructed ancestral sequences are pseudodata and may involve random errors and systematic biases. The reconstruction bias in the case of the parsimony method has been discussed (Collins *et al.* 1994; Perna and Kocher 1995; Eyre-Walker 1998). The bias exists also with the likelihood (Bayesian) method (Yang *et al.* 1995; Yang 2006, pp. 126–128), as the problem lies with the use of the single best reconstruction (the one that requires the minimum number of changes by parsimony or that has the highest posterior probability by likelihood) rather than with the reconstruction method. As an example, Jordan *et al.* (2005) used parsimony to polarize amino acid substitutions on three-taxon trees and found that common amino acids were even more common in ancestors. This "universal trend" of amino acid gain and loss appears to be an artifact of ancestral reconstruction (Goldstein and Pollock 2006). While reconstruction bias is more serious for more divergent sequences, it has been noted to be substantial in the analysis of human mitochondrial D-loop sequences and even in analysis of population data (Hernandez *et al.* 2007).

One approach to reducing the reconstruction bias, referred to below as averaging weighted by posterior probabilities (AWP), is to average over multiple reconstructions, with their posterior probabilities (PPs) calculated in the likelihood (Bayesian) method used as weights (Yang 2006, pp. 126–128; see also Krishnan *et al.* 2004; Dutheil *et al.* 2005; Akashi *et al.* 2007). If the likelihood model is too simplistic, the weights will be incorrect, but the approach may still be less biased than the use of the single best reconstruction. In a computer simulation, Akashi *et al.* (2007) examined the AWP approach, with the PPs calculated under the Hasegawa-Kishino-Yano (HKY) model (Hasegawa *et al.* 1985), to count substitutions between preferred (common) and unpreferred (rare) codons in protein-coding genes. AWP was found to have much smaller bias than parsimony, but the bias was still unacceptably high under complex scenarios of nonstationary base composition evolution, apparently because the stationary HKY model is too simplistic to describe adequately the substitution process.

This simulation and further simulation experiments we conducted later have motivated us to develop more powerful methods to correct for the reconstruction bias and to count substitutions along branches under complex nonstationary models. As a result we have extended the AWP method to stationary and nonstationary general time reversible (GTR) model of nucleotide substitution (Tavaré 1986; Yang 1994; Zharkikh 1994). We have also implemented a new method for counting different types of nucleotide substitutions along a branch on the phylogeny, taking into account the changes in base compositions over time.

There has been much effort in developing nonstationary, nonhomogeneous, or nonreversible models of nucleotide or amino acid substitution for use in inference of phylogenetic relationships among distant species, in both the maximum-likelihood (Yang and Roberts 1995; Galtier and Gouy 1998; Dutheil and Boussau 2008; Jayaswal *et al.* 2011; Groussin *et al.* 2013; Gueguen *et al.* 2013; Jayaswal *et al.* 2014) and Bayesian (Foster 2004; Blanquart and Lartillot 2006, 2008) frameworks. Here our focus is on estimation of substitution rates and counting of substitutions to study the process of sequence evolution, with the phylogeny assumed known. We are conducting an analysis of genome sequences from the *Drosophila melanogaster* subgroup to infer the relative roles of mutation and selection driving the evolution of synonymous sites and intron base compositions. Previous studies have highlighted the importance of natural selection affecting synonymous codon usage and the fluctuating selective strength across lineages in this species group (*e.g.*, Nielsen *et al.* 2007; Duret and Arndt 2008; Singh *et al.* 2009). To tease apart the effects of mutation and selection and to estimate the changing selective strengths, one needs reliable methods for counting substitutions on the branches of the tree. The main objective of the present study is thus to compare the different inference methods by simulation to inform us of the suitable methods in our future data analysis. However, our results should apply more broadly to inference of lineage-specific evolution when evolutionary forces are fluctuating.

In this article we first describe our implementation of the methods of data analysis, including the new method for estimating the expected substitution counts along a branch under a nonstationary model. We then report an extensive computer simulation study designed to evaluate the different methods under both simple stationary and complex nonstationary scenarios of sequence evolution.

Theory and Methods

Substitution models assumed in data analysis

We use the (stationary and homogeneous) HKY model of nucleotide substitution (Hasegawa *et al.* 1985), as well as two versions of nonstationary models implemented by Yang and Roberts (1995), referred to below as HKY-NH and HKY-NH$_b$ (Table 1). Here "NH" stands for "nonhomogeneous," although those models are nonstationary (with the base compositions changing over time) but time homogeneous (with the substitution rate matrix constant among lineages). The stationary HKY model involves $2s - 3$ branch lengths on the unrooted tree for s species and four free parameters in the substitution

Table 1 Summary of models used in data analysis in this study

Model	Assumptions	No. parameters
HKY	Homogeneous and stationary process	$(2s - 3) + 4$
HKY-NH	Nonstationary, with one set of base frequency parameters for every branch on the rooted tree and the same κ for all branches	$(2s - 2) + (2s - 1) \times 3 + 1$
HKY-NH$_b$	Nonstationary, with one set of base frequency parameters and one κ parameter for every branch on the rooted tree	$(2s - 2) + (2s - 1) \times 3 + (2s - 2)$
GTR	Homogeneous and stationary process	$(2s - 3) + 8$
GTR-NH	Nonstationary, with one set of base frequency parameters for every branch on the rooted tree and the same exchangeability parameters (a, b, c, d, e) for all branches	$(2s - 2) + (2s - 1) \times 3 + 5$
GTR-NH$_b$	Nonstationary model, with one set of base frequency parameters and one set of exchangeability parameters (a, b, c, d, e) for every branch on the rooted tree	$(2s - 2) + (2s - 1) \times 3 + (2s - 2) \times 5$

Under the stationary models (HKY and GTR), unrooted trees are used so that there are $(2s - 3)$ branch lengths in the unrooted tree for s species. Under the nonstationary models (the NH and NH$_b$ models), rooted trees are used so that there are $(2s - 2)$ branch lengths.

rate matrix (the transition/transversion rate ratio κ and three base frequencies). In HKY-NH, each branch on the rooted tree is assigned a set of base frequency parameters, plus a set of initial base frequency parameters at the root, while the same κ is assumed for all branches on the tree. The model involves the following parameters: $(2s - 2)$ branch lengths on the rooted tree, $(2s - 1) \times 3$ base frequency parameters, and one κ parameter. The HKY-NH$_b$ model allows each branch to have its own κ parameter, with $(2s - 2) + (2s - 1) \times 3 + (2s - 2)$ parameters in total. Note that the nonhomogeneous model studied by Galtier and Gouy (1998; see also Dutheil and Boussau 2008) is a special case of the HKY-NH model and uses the GC content rather than the base compositions as a parameter.

In this study we have implemented the GTR versions of those models, referred to as GTR, GTR-NH, and GTR-NH$_b$. The GTR model has the instantaneous rate matrix

$$Q = \{q_{ij}\} = \begin{bmatrix} . & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & . & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & . & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & . \end{bmatrix}, \qquad (1)$$

where q_{ij} is the substitution rate from nucleotides i to j, with the nucleotides ordered T, C, A, and G (Yang 1994; Zharkikh 1994; see also Tavaré 1986). We refer to π_T, π_C, π_A, and π_G (with the sum to be 1) as the *frequency parameters* and to a, b, c, d, and e as the *rate parameters* (Yang 1994) or *exchangeability parameters* (Whelan and Goldman 2001). The diagonals of the matrix are determined by the requirement that each row sums to 0. As the data depend on Qt but not Q and t separately, we fix $f = 1$ to avoid nonidentifiability, and then the exchangeability parameters are relative. As is common practice, the matrix Q is multiplied by a factor so that the average rate is $-\sum_i \pi_i q_{ii} = 1$, and then time t is measured by distance (the expected number of nucleotide substitutions per site when the process is stationary).

The stationary GTR model has $(2s - 3) + 8$ parameters since there are 8 free parameters in the rate matrix (Equation 1). GTR-NH involves a set of frequency parameters for each branch on the rooted tree, but the exchangeability parameters (a, b, c, d, and e) are shared among branches, with $(2s - 2) + (2s - 1) \times 3 + 5$ parameters. GTR-NH$_b$ assigns an independent rate matrix for each branch and involves $(2s - 2) + (2s - 1) \times 3 + (2s - 2) \times 5$ parameters.

All the nonstationary models (NH and NH$_b$) are implemented in the BASEML program in the PAML package (Yang 2007). One option assigns a set of frequency parameters for every branch. This is the NH$_b$ models used in this simulation study. Another option allows the user to specify the number of sets of frequency parameters and to assign every branch to a particular set. This may be useful if all lineages in a clade or subtree have similar base compositions and share similar substitution patterns. Calculation of the likelihood function follows Felsenstein's (1981) pruning algorithm (see Yang 1995), and the optimizer in PAML (Yang 2007) is used to estimate the model parameters including the base frequency parameters by maximizing the log-likelihood function. We note that GTR-NH$_b$ is the same model as the nstGTR model of Zou *et al.* (2012), who implemented the model by using branch lengths estimated under the more general Barry–Hartigan model (Barry and Hartigan 1987) and had to deal with the nonidentifiability problems under that model.

After the maximum-likelihood estimates (MLEs) of parameters are generated, joint ancestral reconstruction (*i.e.*, assignment of nucleotide states to all internal nodes on the tree at a site) is conducted by calculating the posterior (conditional) probabilities for the reconstructions given the data using the MLEs of the parameters (Yang *et al.* 1995, equation 2). The dynamic programming algorithm of Pupko *et al.* (2000) is implemented, which calculates the best joint reconstruction and its posterior probability. While Yang (1995) and Pupko *et al.* (2000) considered stationary models, the algorithm applies to nonstationary models as well. The BASEML program implements a modified version of the algorithm so that suboptimal reconstructions, with PP $\geq$ 0.1%, are also listed. This is

achieved by storing in computer memory not only the best state but also the second or third best states during each stage on the Viterbi path. However, for a highly variable site, the algorithm may miss many of the reconstructions with PP > 0.1%, even though it is very likely to find the second and third best reconstructions. The sum of the posterior probabilities over the listed reconstructions will indicate whether some likely reconstructions are missed.

Note that joint ancestral reconstruction (Yang *et al.* 1995, equation 2) is used here because our interest is on the substitution counts along branches of the tree. For any given site, the joint reconstruction evaluates assigned character states to all ancestral nodes in the tree and accounts for the strong correlation among the nodes. In this case, marginal reconstruction (Yang *et al.* 1995, equation 4) is not the appropriate method (Yang 2006, p. 121).

Methods for counting nucleotide substitutions

For each of the six models (HKY, HKY-NH, HKY-NH$_b$, GTR, GTR-NH, and GTR-NH$_b$), we consider several methods to count the 12 nucleotide substitutions on every branch of the phylogeny. The substitution counts are then used to calculate a substitution skew index that is diagnostic of the relative roles of mutation and selection affecting synonymous codon usage or base composition evolution. The index is described later. Here we describe the counting methods, which use either the MLEs of parameters under the model or the ancestral reconstructions with their posterior probabilities calculated at the MLEs of model parameters.

i. Counting using single best reconstruction (SBR). For each site, the best reconstruction (that is, the reconstruction with the highest posterior probability) is generated using the algorithm of Pupko *et al.* (2000). If the nucleotides (let them be i and j) at the start and end of a branch are different according to this best reconstruction, we count an $i \rightarrow j$ change on the branch. The counts are summed over sites in the sequence to generate the substitution counts for the whole branch. Note that this method ignores the suboptimal reconstructions and also ignores possible multiple hits within the branch. In this study, we do not distinguish this SBR method from maximum parsimony (MP).

ii. AWP. In this method, we average over ancestral reconstructions generated by the BASEML program, with their posterior probabilities used as weights. Consider the tree of Figure 1, which has six species so that a reconstruction at a site may be represented as $y_7y_8y_9y_{10}y_{11}$. Suppose the best two reconstructions at a site are $y_7y_8y_9y_{10}y_{11}$ = CCCCC (with PP = 0.50) and CCTTT (with PP = 0.48), with other reconstructions having negligible probabilities (Yang *et al.* 1995, equation 2). Then for branches 7–9 (branch *tyeo*), we count 0.49 C → T changes since 0.48/(0.50 + 0.48) = 0.49. The SBR method would count 0 changes for every internal branch at this site. Similarly the counts are summed over sites in the sequence to generate the substitution counts for the whole branch. Note that the posterior probabilities for the ancestral reconstructions are calculated using the MLEs of parameters. This approach, known as empirical Bayes (EB), ignores the sampling errors in the parameter estimates. As our simulated data sets are large (see below), the sampling errors are expected to be small. See *Discussion* below for a discussion of the different ancestral reconstruction methods.

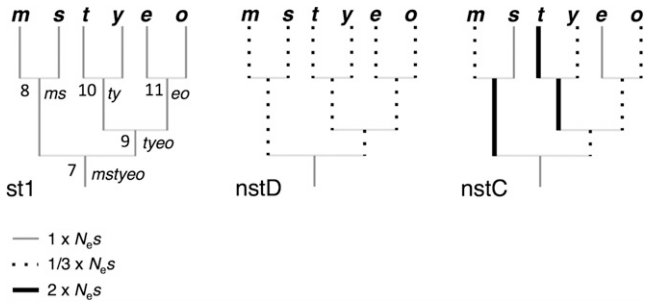

Figure 1 Three schemes of natural selection driving base composition evolution illustrated in phylogenetic trees for six *Drosophila* species with relative branch lengths used to simulate sequence data. Scheme st1 represents a stationary scenario, in which selection has been homogeneous and the base compositions and codon usage have been stationary throughout the tree. Scheme nstD is a simple nonstationary scenario, with relaxed selection in all lineages since the root, which may represent a population size reduction in all lineages. Scheme nstC represents a complex nonstationary scenario, in which some lineages experienced strengthened selection and others weakened selection.

iii. Expected Markov counting (EMC). The EMC method counts nucleotide substitutions along a branch by taking expectations over the Markov-chain substitution process, accounting for the fact that the base compositions may be changing. Let $E\{X_{ij}(t)\}$ be the expected number of $i \rightarrow j$ substitutions per site over time interval $(0, t)$ when the initial distribution of the Markov chain at time 0, $\pi^{(0)} = \left\{\pi_T^{(0)}, \pi_C^{(0)}, \pi_A^{(0)}, \pi_G^{(0)}\right\}$, differs from the stationary distribution, $\pi = \{\pi_T, \pi_C, \pi_A, \pi_G\}$. Here π and $\pi^{(0)}$ are row vectors, and times 0 and t represent the start and end of the branch, which has length t. The rate matrix $Q = \{q_{ij}\}$ for the Markov chain is constant over time (or over the branch). Let the spectral decomposition of Q be

$$Q = U\Lambda V \tag{2}$$

or equivalently

$$q_{ij} = \sum_{\alpha=1}^{4} u_{i\alpha} v_{\alpha j} \lambda_\alpha, \tag{3}$$

where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ is a diagonal matrix, with the diagonal elements to be the eigenvalues and the off-diagonal elements to be zero, and where columns of U are the right eigenvectors, with $V = U^{-1}$. The matrix of transition probabilities over time t is then

$$P(t) = e^{Qt} = Ue^{\Lambda t}V = U\ \text{diag}\left\{e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t}\right\}V. \tag{4}$$

Note that $\lambda_1 = 0$ and $\lambda_2, \lambda_3, \lambda_4 < 0$. The first column of U has all entries to be 1 and the first row of V is π; that is,

$u_{i1} = 1$ and $v_{1i} = \pi_i$ for all i. The matrices Λ, U, and V are available analytically for HKY (Hasegawa *et al.* 1985) and can be calculated numerically for GTR (*e.g.*, Yang 2006, pp. 68 and 69).

Break the interval $(0, t)$ into N segments, each of width $\Delta t = t/N$. Let $t_h = h\Delta t = ht/N$, $h = 1, \ldots, N$. The expected distribution (base compositions) at time t_h is

$$\begin{aligned}\boldsymbol{\pi}^{(h)} &= \left\{\pi_T^{(h)}, \pi_C^{(h)}, \pi_A^{(h)}, \pi_G^{(h)}\right\} \\ &= \boldsymbol{\pi}^{(0)}P(t_h) = \boldsymbol{\pi}^{(0)}e^{Qt_h} = \boldsymbol{\pi}^{(0)}Ue^{\Lambda t_h}V \\ &= \boldsymbol{\pi}^{(0)}U\ \mathrm{diag}\left\{e^{\lambda_1 t_h}, e^{\lambda_2 t_h}, e^{\lambda_3 t_h}, e^{\lambda_4 t_h}\right\}V.\end{aligned} \tag{5}$$

Within each time segment (t_{h-1}, t_h), the distribution is nearly constant and equal to $\boldsymbol{\pi}^{(h)}$. Thus

$$\begin{aligned}E\{X_{ij}(t)\} &\simeq \sum_{h=1}^{N}\pi_i^{(h)}q_{ij}\Delta t \\ &= q_{ij}\Delta t \times \sum_{h=1}^{N}\left(\boldsymbol{\pi}^{(0)}U\ \mathrm{diag}\left\{e^{\lambda_1 t_h},\ e^{\lambda_2 t_h},\ e^{\lambda_3 t_h},\ e^{\lambda_4 t_h}\right\}V\right)_i \\ &= q_{ij}\Delta t \times \sum_{h=1}^{N}\sum_{k=1}^{4}\pi_k^{(0)}\sum_{\alpha=1}^{4}u_{k\alpha}v_{\alpha i}e^{\lambda_\alpha t_h} \\ &= q_{ij}\Delta t \times \sum_{k=1}^{4}\sum_{\alpha=1}^{4}\pi_k^{(0)}u_{k\alpha}v_{\alpha i}\sum_{h=1}^{N}e^{\lambda_\alpha t_h} \\ &\simeq q_{ij} \times \sum_{k=1}^{4}\sum_{\alpha=1}^{4}\pi_k^{(0)}u_{k\alpha}v_{\alpha i}\int_0^t e^{\lambda_\alpha s}ds \\ &= q_{ij} \times \sum_{k=1}^{4}\pi_k^{(0)}\left[u_{k1}v_{1i}t + \sum_{\alpha=2}^{4}\left(u_{k\alpha}v_{\alpha i} \times \frac{1}{\lambda_\alpha}(e^{\lambda_\alpha t} - 1)\right)\right] \\ &= \pi_i q_{ij}t + q_{ij}\sum_{k=1}^{4}\sum_{\alpha=2}^{4}\left[\pi_k^{(0)}u_{k\alpha}v_{\alpha i} \times \frac{1}{\lambda_\alpha}(e^{\lambda_\alpha t} - 1)\right] \\ &= \pi_i q_{ij}t + q_{ij}c_i.\end{aligned} \tag{6}$$

Here $(a)_i$ represents the ith element in vector a. Note that if the process is stationary, with $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$, the correction term $q_{ij}c_i$ vanishes with $c_i = 0$ and $E\{X_{ij}(t)\} = \pi_i q_{ij}t$. Also while $\pi_i q_{ij} = \pi_j q_{ji}$, the counts generated by Equation 6 are not symmetrical if the process is nonstationary.

Thus under a nonstationary model of nucleotide substitution,

$$v = \sum_{i\neq j}E\{X_{ij}(t)\} = -\sum_i(\pi_i q_{ii}t + q_{ii}c_i) \tag{7}$$

is a more accurate definition of branch length than t is since v accounts for the changing base compositions and the changing substitution rates over time. This is the same branch-length definition of Zou *et al.* (2012, equation 6), who derived it using Minin and Suchard's (2008b) expected count of substitutions conditioned on the nucleotide states at the two ends of the branch. The derivation here appears to be simpler.

As defined in Equation 6, $E\{X_{ij}(t)\}$ is a function of the model parameters. For example, for the GTR-NH$_b$ model, the parameters include the base frequencies at the root, the exchange rates (a, b, c, d, e, with $f = 1$ fixed) for every branch, and the branch lengths. Given those parameters, one can use Equation 6 to calculate the 12 expected substitution counts for every branch on the tree. Note that given the distribution at the root and the rate matrices for all branches, the distribution at every internal node of the tree can be calculated using a preorder tree traversal (Equation 5), with the distribution for ancestral nodes calculated before those for descendent nodes. Using the invariance property of MLEs, we replace the parameters in Equation 6 by their MLEs to produce the MLE of $E\{X_{ij}(t)\}$, the expected number of $i \to j$ substitutions per site along the branch. Note also that under a stationary model (such as HKY or GTR), $\widehat{\pi}_i\widehat{q}_{ij}\widehat{t}$ (where the caret indicates the MLE of the parameters) is the MLE of $E\{X_{ij}(t)\}$.

In summary, under each model, we first perform ML optimization to generate the MLEs of model parameters. We then use the methods described above to produce the substitution counts for every branch on the tree. For SBR and AWP, the MLEs of parameters are used to calculate the posterior probabilities for ancestral reconstructions (Yang *et al.* 1995, equation 2), and then either the SBR is used or multiple reconstructions are averaged (with their posterior probabilities as weights, AWP) to generate the counts. For EMC, we use Equation 6, with parameters replaced by their MLEs, to calculate the expected substitution counts for the branch.

Simulation model of base composition evolution

The simulation considers weak fixation biases that affect base composition. Such a model appears to capture the main features of synonymous codon usage in *Drosophila* (Kliman and Hey 1993; Akashi 1994, 1995, 1996; Moriyama and Powell 1997; Duret and Mouchiroud 1999; McVean and Vieira 2001; Vicario *et al.* 2008; Singh *et al.* 2009; Poh *et al.* 2012; Campos *et al.* 2013). Codon usage, or base composition at synonymous sites, appears to be under weak selection so that, in general, GC-ending codons are preferred and AT-ending codons are unpreferred. However, the strength or efficacy of selection relative to mutation biases may vary among lineages, resulting in a nonstationary substitution process. Inference of changes in selective strength and in mutation bias on the different lineages requires reliable counts of different types of nucleotide substitutions, and we evaluate the performance of the different methods in recovering the actual counts recorded during the simulation. We note that biased gene conversion may affect the nucleotide fixation process in a manner similar to weak selection (Nagylaki 1983), resulting in substitution biases throughout the genome and thus affecting GC content variation among genes (reviewed in Marais 2003; Duret and Galtier 2009). However, gene conversion is not considered in our simulation.

Akashi *et al.* (2007) simulated codon usage evolution in different codon-degeneracy classes. Here we used 777,600

third positions of fourfold degenerate codons, generated in that simulation. The scenario is equivalent to nucleotide-level selection on GC content. The selection intensity is measured by *Ns*, where *N* is the (effective) chromosomal population size and *s* the selection coefficient. G- and C-ending codons are preferred codons with relative fitness 1, while A- and T-ending codons are unpreferred with fitness $1 - s$, so that the selection coefficient is *s* for any *up* (unpreferred → preferred) mutation, $-s$ for a *pu* mutation, and 0 for *uu* and *pp* mutations. Synonymous codon substitution is simulated using a discrete Markov chain. The transition probability, over one generation, from any codon to any other synonymous codon is given by the number of mutations per generation in the population multiplied by the fixation probability (Fisher 1930),

$$p_{\mathrm{up}} = N\mu \times \frac{2s}{1 - e^{-2Ns}} \tag{8}$$

for any *up* substitution (such as T → C),

$$p_{\mathrm{pu}} = N\mu \times \frac{-2s}{1 - e^{2Ns}} = N\mu \times \frac{2s}{e^{2Ns} - 1} \tag{9}$$

for any *pu* substitution (such as C → T), and

$$p_{\mathrm{pp}} = p_{\mathrm{uu}} = N\mu \times \frac{1}{N} = \mu \tag{10}$$

for a neutral substitution (such as T → A or C → G). Here μ is the mutation rate from one nucleotide to another. When this substitution process reaches equilibrium, the GC content is given by the relationship $\mathrm{GC}/(1 - \mathrm{GC}) = p_{\mathrm{up}}/p_{\mathrm{pu}}$ or

$$\mathrm{GC} = \frac{e^{2Ns}}{1 + e^{2Ns}} \quad \text{and} \quad Ns = \frac{1}{2}\log\frac{\mathrm{GC}}{1 - \mathrm{GC}} \tag{11}$$

(Li 1987; Bulmer 1991).

Phylogenetic tree, selection scheme, and simulation of sequence alignments

The model phylogeny used in the simulation is shown in Figure 1. The tree was based on the relationships and approximate branch lengths of six species in the *D. melanogaster* subgroup: *D. melanogaster*, *D. simulans*, *D. teissieri*, *D. yakuba*, *D. erecta*, and *D. orena* (Ko *et al.* 2003). Those six species are referred to as *m*, *s*, *t*, *y*, *e*, and *o* and the five ancestral species as *ms*, *ty*, *eo*, *tyeo*, and *mstyeo*. All data were simulated using this tree and the analyses assumed knowledge of the true tree. The effective population size (used to determine fixation probabilities) was set to $N = 5000$. We assumed equal mutation rates between any two nucleotides, with $\mu = 2 \times 10^{-5}$ per site per generation (with the total mutation rate per site per generation, 3μ). The branch lengths were 3600 generations for branches *m*, *s*, *t*, *y*, *e*, *o*, *ty*, and *eo*; 5400 generations for *ms*; and 1800 generations for *tyeo*. Here a "generation" represents one step in the forward simulation of the discrete Markov chain that is small enough so that multiple substitutions at the same site are negligible. Those three branch lengths corresponded to ~5%, 7.5%, and 2.5% of synonymous sequence divergence for the selective scheme corresponding to an equilibrium GC content of 70% (selection scheme st1; see below). "Fixations" were instantaneous and we did not account for polymorphism.

Sequences at the root of the tree were generated through a burn-in phase. Sequences were initialized with random nucleotide composition and then evolved under a fixed selection regime (fixed *Ns*) until the nucleotide frequencies reached equilibrium. We considered six GC contents for the root sequence, referred to as $\mathrm{GC}_{\mathrm{initial}}$: 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95, which corresponded to $Ns = 0$, 0.20, 0.43, 0.69, 1.10, and 1.47, respectively (Equation 11). In other words, we used the correct selective coefficient *Ns* to run the burn-in until the root sequence reached the desired GC content. This is equivalent to sampling sites for the root sequence in proportion to the expected frequencies given by Equation 11 and noting the corresponding *Ns* value.

After the burn-in, the root sequence was allowed to evolve over generations along branches of the tree according to the transition probabilities of Equations 8–10 under a specified selection scheme. We considered several selection schemes and report results for three of them (Figure 1). In the stationary scheme st1, all branches of the tree evolved at the same selective strength (the same *Ns*) as during the burn-in, so that base compositions did not change over lineages on the tree. We also considered a variation to scheme st1 in which all branch lengths were doubled. This is called st2x and the increased sequence divergence was useful to reveal estimation errors in some methods. The second scheme was a simple nonstationary scheme (nstD), in which all lineages experienced relaxed selection, with *Ns* for all branches one-third of the burn-in value. In such a scenario, the GC content decreased in all lineages. The third scheme was a complex nonstationary scheme (nstC), in which some lineages experienced strengthened selection with *Ns* set to twice the burn-in value, while some lineages experienced relaxed selection with *Ns* one-third of the burn-in value. For example, for the case where $\mathrm{GC}_{\mathrm{initial}} = 0.7$ in the root sequence, the three possible *Ns* values assigned to branches on the tree (Figure 1, nstC scheme) are 0.141 ($1/3 \times Ns$), 0.424 ($1 \times Ns$) and 0.847 ($2 \times Ns$), which correspond to equilibrium GC contents of 0.57, 0.7, and 0.84 (Equation 11). Thus while the initial GC content in the root sequence was 70%, GC was decreasing toward 0.57 along the dashed branches because of relaxed selection and was increasing toward 0.84 along the thick branches because of strengthened selection. Figure 2 shows the substitution probabilities per generation for those three *Ns* values.

Simulation of the evolutionary process along branches of the tree under the specified selection scheme led to sequences for the six modern species at the tips of the tree. Those sequences constitute the data to be analyzed using different methods implemented in the BASEML program (see above). In addition, the simulation generated the

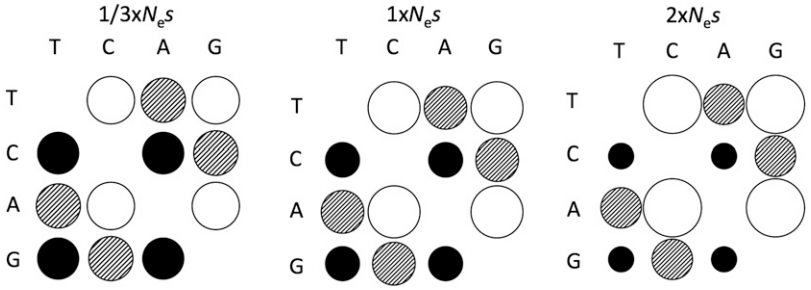

Figure 2 Transition rates (q_{ij}) from nucleotides *i* to *j*, represented by the area of the circles for three selection strengths: 1/3 × *Ns*, 1 × *Ns*, and 2 × *Ns*, for an intermediate GC bias, with $GC_{initial}$ = 0.7 in the sequence at the root (*Ns* = 0.424). The same scale is used for all three plots. Color patterns indicate mutation classes, for advantageous (*up*, open circles), deleterious (*pu*, solid circles), and neutral (*pp* and *uu*, hatched circles) substitutions. Note that there are only three distinct rates in the rate matrix under our simulation model.

sequences at the five internal nodes, as well as the counts of all 12 types of substitutions for every branch on the tree. Those actual counts were used for comparison with the inferred counts.

Analysis of the simulated data

The BASEML program in PAML 4.8 (Yang 2007) was used to conduct all analyses by the methods described above. Branch lengths and parameters in the substitution models (such as the exchangeability and frequency parameters) were estimated by maximum likelihood (ML). The true tree topology was always used. As the parameter-rich nonstationary models pose difficult numerical optimization problems, we ran the same analysis 10 times, using random starting values, and used the results that correspond to the highest log likelihood. While this strategy did not guarantee the success of finding the global maximum, it seemed to have worked well in our analysis of the simulated data. Each run took a few seconds. In this case, data generation required more computation than data analysis.

We implemented the MP reconstruction by using the SBR under the Jukes–Cantor (JC) model (Jukes and Cantor 1969). Note that MP and likelihood under JC rank the reconstructions in exactly the same order if all branches on the tree have the same length (Yang *et al.* 1995, equation 1) and should be very similar if branch lengths are different but small (Akashi *et al.* 2007).

We do not evaluate inference on the two branches around the root in the rooted tree. Under the stationary models HKY and GTR, the root of the tree is unidentifiable as both models are time reversible. Under nonstationary models, the root of the tree is identifiable, but we expect the data to contain little information to root the tree in this way (Yang and Roberts 1995; Huelsenbeck *et al.* 2002).

We grouped the 12 nucleotide substitutions into four types, *up* (T, A → C, G), *pu* (C, G → T, A), *uu* (T ↔ A), and *pp* (C ↔ G), and calculated a *up–pu* skew index

$$d_{\mathrm{up,pu}} = \frac{\mathrm{up} - \mathrm{pu}}{\mathrm{up} + \mathrm{pu}}, \tag{12}$$

where up and pu are the counts of *up* and *pu* substitutions. This is a measure of the direction and strength of base composition/codon usage evolution. As discussed in Akashi *et al.* (2007), if base composition changes on a lineage reflect lineage-specific selection intensity, $d_{\mathrm{up,pu}}$ should vary almost linearly as a function of base composition bias. On the other hand, if base composition changes reflect variable mutation rate, $d_{\mathrm{up,pu}}$ would uniformly increase or decrease independent of initial base composition bias. Thus, $d_{\mathrm{up,pu}}$ can be used to test for causes(s) of fluctuations in base composition. We compare inferred counts of substitutions and the $d_{\mathrm{up,pu}}$ index with the actual values from our simulations to determine the accuracy of ancestral inference, using the different methods.

Results

Inference when the substitution process is stationary

We first consider calculation of the $d_{\mathrm{up,pu}}$ index when the data are simulated under a homogeneous selection scheme, st2x. The $d_{\mathrm{up,pu}}$ values for lineage *m* calculated using several different methods are plotted against the initial GC content in the root sequence in Figure 3A and B. Results for other lineages on the tree show similar patterns. Also the results for the stationary scheme st1 are similar, although the pattern is less pronounced due to the lower sequence divergences than under scheme st2x. Under scheme st2x, selection has been homogeneous and there is no change in base compositions along any branch, so that $d_{\mathrm{up,pu}}$ should be 0 whatever the initial GC content. Nevertheless, MP infers a negative trend that is very similar to the predicted trend under relaxed selection. However, this trend is spurious and is caused by the well-known bias of MP, which ignores the suboptimal reconstructions that require more changes than the most parsimonious reconstruction. As a result, the common nucleotides (C and G) or the preferred codons are inferred to be even more common in the ancestors, generating an artifactual trend of decreasing GC bias and weakened selection.

The AWP method uses posterior probabilities as weights to average over multiple ancestral reconstructions when counting changes. AWP under the stationary HKY model performed better than parsimony (Figure 3A). However, at high GC bias it was not reliable. In Figure 3C, we show the actual substitution rates between nucleotides in the simulation model as well as the estimates under HKY. Overall HKY overcounted *pu* and undercounted *up* changes because the model assigns the same transversion rate parameter (in the notation of Hasegawa *et al.* 1985) for *uu* and *pu* substitutions and for *pp* and *up* substitutions. The model is not

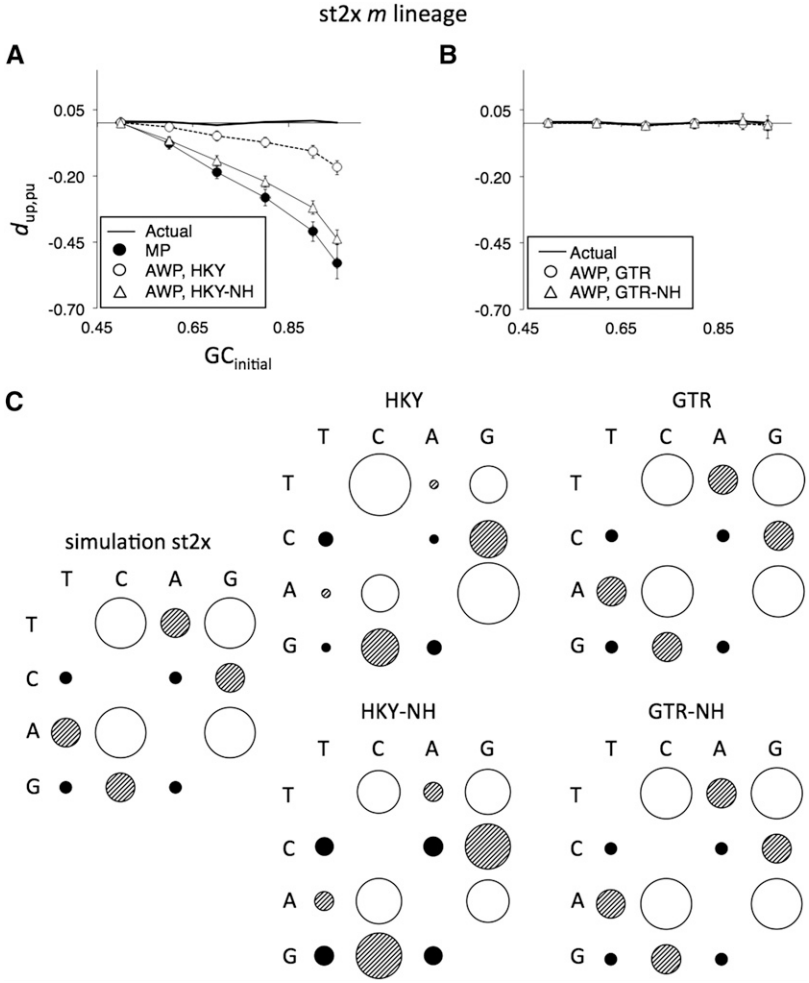

Figure 3 (A and B) Actual and inferred $d_{up,pu}$ for the *m* lineage when the data are generated under the stationary scheme st2x. Shown are the averages and 95% confidence intervals over 300 bootstrap replicate data sets, generated by bootstrap resampling of the 777,600 sites in the simulated data set. Results are obtained using the MP and AWP methods, under stationary and nonstationary HKY model (A), and stationary and nonstationary GTR model (B). The substitution rates in C are for the high GC bias ($GC_{initial}$ = 0.95).

complex enough to describe the true pattern of nucleotide substitution adequately. The impact of model violation is particularly pronounced at the high GC bias, with $GC_{initial}$ = 0.95 (Figure 3A). Those results are consistent with the findings of Akashi *et al.* (2007), who used simulation to compare MP with AWP under HKY.

We also used AWP under the nonstationary HKY-NH model to calculate $d_{up,pu}$ even though the true process is stationary. The other nonstationary model HKY-NH$_b$, which uses an independent κ for every branch, produced very similar results to those of HKY-NH, and both are considerably *more* biased than the stationary HKY model (Figure 3A). The "wrong" parameters added in the HKY-NH model actually caused greater biases than the stationary HKY. For the case of high GC bias with $GC_{initial}$ = 0.95, parameter estimates under HKY-NH suggested base frequency changes, with GC = 0.971 at the root, 0.958 at node *ms*, and 0.95 at the tip *m* (Figure 1), when in fact the process has been stationary and all nodes had GC = 0.95.

The $d_{up,pu}$ index for lineage *m* calculated using the AWP method under the stationary GTR and the nonstationary GTR-NH models for scheme st2x is plotted in Figure 3B. GTR is general enough to accurately describe the codon usage selection model, so that both GTR and GTR-NH are correct. Both models performed equally well and gave a high accuracy in the $d_{up,pu}$ calculation. The estimated substitution rates under the two models (Figure 3C) also match closely the true rates. The counts of different nucleotide substitutions inferred by the two models also matched closely the actual counts. The good performance of those two models may be expected given the match between the simulation and inference models and the large sample size. Note, however, that the root of the tree is not identifiable under the stationary models HKY and GTR and the inference of the root is highly unreliable under the nonstationary models HKY-NH and GTR-NH.

Inference when the substitution process is nonstationary

Next, we examine the performance of the stationary models HKY and GTR and the nonstationary models HKY-NH and GTR-NH when the data are simulated under the simple nonstationary scheme nstD. In this scheme, selection is weakened and the GC content has been decreasing along all branches on the tree. The results for lineage *m* are shown in Figure 4. Those for other lineages are similar.

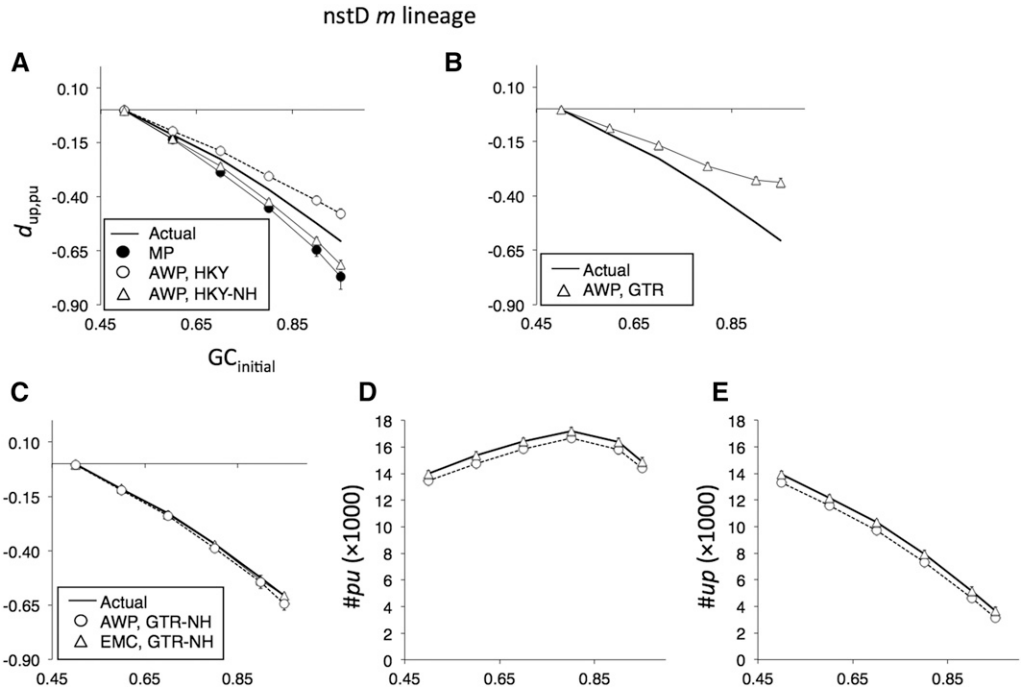

Figure 4 Actual and inferred $d_{up,pu}$ index and *pu* and *up* substitution counts for lineage *m* under the simple nonstationary scheme nstD. Shown are the averages and 95% confidence intervals over 300 bootstrap replicate data sets, generated by bootstrap resampling of the 777,600 sites in the simulated data set. (A–C) The $d_{up,pu}$ index is calculated using various methods. (D and E) The *pu* and *up* substitution counts are calculated using the AWP and EMC methods under the GTR-NH model.

The nonstationary model HKY-NH accommodates the base composition changes but the rate matrix is too restrictive to describe the substitution process. This model mismatch caused undercounting of *up* changes and underestimation of $d_{up,pu}$ (Figure 4A), showing bias in the same direction as when the stationary HKY is applied to the stationary scheme st2x (Figure 3A). For those data, two component assumptions of the stationary HKY model are violated. First, HKY does not accommodate the base composition changes. Second, the rate matrix of HKY is too restrictive to describe the real substitution process. Violations of the two assumptions had opposite effects, as the bias in HKY is in the opposite direction to the bias in HKY-NH (Figure 4A). Furthermore, the violation of the stationarity assumption had a greater impact on calculation of $d_{up,pu}$ than the mismatch between HKY and GTR.

The homogeneous GTR model does not accommodate the changing base compositions and overcounted the *up* changes and undercounted the *pu* changes, showing similar but slightly larger bias than HKY (Figure 4B). The larger bias in GTR than in HKY reflects a cancellation of errors in HKY that does not occur in GTR.

Results obtained under the nonstationary GTR-NH model using both the AWP and EMC methods for selection scheme nstD are shown in Figure 4, C–E. The results for GTR-NH_b are very similar to those for GTR-NH. For those data, GTR-NH matches the simulation model, and both AWP and EMC performed very well. Nevertheless, AWP showed small biases and undercounted both *pu* and *up* substitutions slightly (Figure 4, C and D), leading to a slight underestimation of $d_{up,pu}$ at high GC. This appears to be due to the failure of AWP to correct for multiple hits within the same branch. The EMC method accounts for multiple hits and produced accurate counts of substitution numbers as well as $d_{up,pu}$ (Figure 4, D and E).

It is noteworthy that under selection scheme nstD, all six modern sequences have the same base compositions, because they all started from the same base compositions at the root and have since been drifting in the same direction over the same time period. While the substitution process is nonstationary, this nonstationarity is not detectable using tests that examine the homogeneity of base compositions among modern sequences such as the matched-pairs test (Tavaré 1986; Ababneh *et al.* 2006). However, the nonstationarity can be detected from a phylogenetic analysis of the modern sequences, using the likelihood-ratio test (LRT) to compare the stationary and nonstationary models. For example, at the extreme GC bias ($GC_{initial} = 0.95$), the LRT statistic is $2\Delta\ell = 24{,}668$ (d.f. = 31) for comparing models HKY and HKY-NH and is $2\Delta\ell = 35{,}282$ (d.f. = 31) for comparing GTR and GTR-NH (Supporting Information, Table S1). In both cases, the test provides overwhelming evidence that the substitution process is nonstationary. Even though all modern sequences have the same base compositions, the asymmetrical nucleotide substitutions (that is, different numbers of $i \rightarrow j$ and $j \rightarrow i$ changes) on the branches of the tree will make a stationary model fit the data poorly, allowing it to be rejected when compared with a nonstationary model. At any rate it is important to note that while heterogeneous base compositions among sequences in the alignment necessarily mean a nonstationary substitution process, homogeneous base compositions among modern sequences do not necessarily mean a stationary process.

Results for the complex nonstationary scheme nstC are shown in Figure 5, where we calculate $d_{up,pu}$ as well as *up* and *pu* substitution counts under the nonstationary models

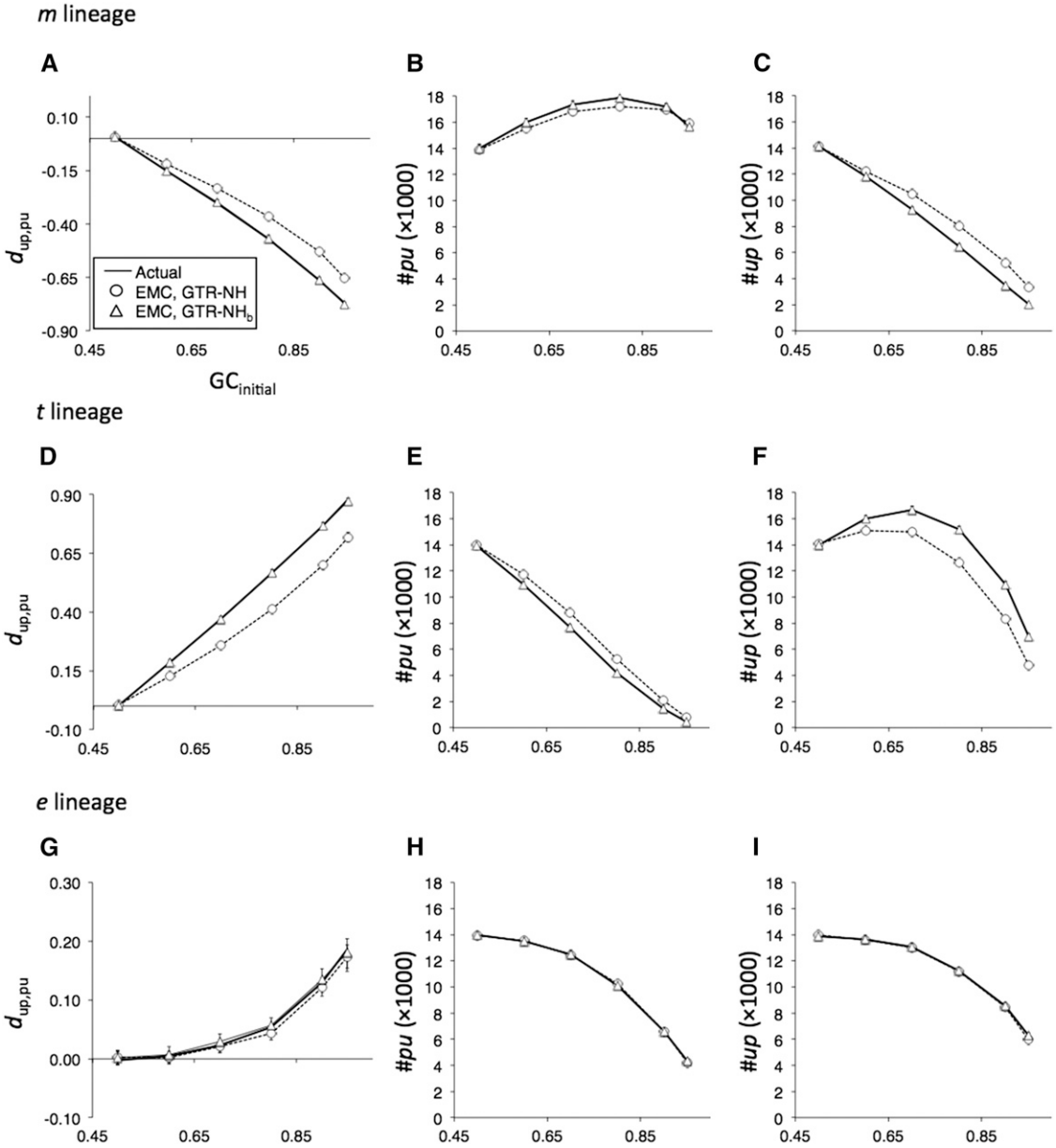

Figure 5 (A–I) Inferred $d_{up,pu}$ index and substitution counts for the *m* (A–C), *t* (D–F), and *e* (G–I) lineages obtained using the EMC method under the GTR-NH and GTR-NH$_b$ models for the complex nonstationary scheme nstC. Shown are the averages and 95% confidence intervals calculated using 300 bootstrap replicates, which were generated by bootstrap resampling, with replacement, the 777,600 sites in the simulated data set. Note that the confidence intervals may be too small to be visible in some plots.

GTR-NH and GTR-NH$_b$. As selection varies among lineages, we show results for three representative lineages on the tree: *m*, *t*, and *e*. Because of the different scaled selection coefficients (*Ns*) applied to different branches (Figure 1), both the exchangeability parameters and base frequency parameters differ among branches so that the GTR-NH model is violated. As a result, for lineage *m* (and lineages *y* and *o*) experiencing relaxed selection, GTR-NH undercounted the *pu* substitutions, overcounted the *up* substitutions, and overestimated $d_{up,pu}$. In contrast, for lineages undergoing strengthened selection (lineages *t*), the bias is in the opposite direction. Finally, for lineages that experienced moderate levels of selection (lineage *e*), the bias is small. This complex pattern of biases can be understood by examining the MLEs of parameters in the model (Table 2 and Table S2). While in the true selection model, the exchangeability parameters (*a*, *b*, *c*, *d*, *e*) are different among branches, the GTR-NH model fits one set of exchangeability parameters, which will be a kind of average over all branches. Thus for lineages experiencing relaxed selection (*e.g.*, lineage *m*), the fitted exchangeability parameter *b* is too large and *e* is too small, while for lineages experiencing strengthened selection (*e.g.*, lineage *t*), the fitted *b* is too small and *e* is too large. The discrepancies are particularly pronounced at high GC contents, causing serious biases in the analysis. The GTR-NH

model produced biased estimates of branch lengths (Figure 6A) and biased estimates of equilibrium base compositions in the lineages on the tree (Figure S1), but those biases are in general small, in comparison with the poor estimates of the exchangeability parameters (b and e) (Table 2). Thus we suggest that the biases in $d_{up,pu}$ calculation in the GTR-NH model were mainly caused by the poor estimates of the exchangeability parameters.

All those biases disappeared when the nonstationary model GTR-NH$_b$ was used (Figure 5 and Table 2). This model uses a whole GTR rate matrix for each branch and fits the complex simulation model adequately. The substitution counts as well as the exchangeability parameters are all estimated with high accuracy (Figure 5 and Table 2). The model also produced accurate estimates of the branch lengths (Figure 6B) and equilibrium base compositions (Figure S1).

Similarly the LRT allows one to reject the simpler GTR and GTR-NH models in comparison with the more complex and more realistic GTR-NH$_b$ model. For example, at the extreme GC bias ($GC_{initial}$ = 0.95), the test statistic is $2\Delta\ell$ = 73,836 (d.f. = 76) for the GTR *vs.* GTR-NH$_b$ comparison and is $2\Delta\ell$ = 9290 (d.f. = 45) for the GTR-NH *vs.* GTR-NH$_b$ comparison (Table S1). The simpler models are rejected by a big margin.

Our simulation model assumes that base composition bias is caused by selection, with stronger selection acting on genes with highly biased GC content. Strengthened or weakened selection is modeled by multiplying the selective coefficient (Ns) across all genes in the genome. This may represent a scenario in which the selective coefficient s is not changing, but the population size changes along some branches of the tree, which affects all genes in the genome. For example, the $2 \times Ns$ scheme may represent a doubling of the population size.

The effect of data size and model complexity

We simulated very large data sets with long sequences to evaluate ancestral inference methods for counting substitutions along every branch of the phylogeny in analysis of the genome-wide trend in base composition evolution. Sampling errors in parameter estimates are small enough to be ignored in such large data sets. However, in small or intermediate-sized data sets, parameter-rich models tend to suffer from large variances due to random sampling errors. If both simple and complex models are adequate in describing the evolutionary scenario, the simple model should in general be preferred as its parameter estimates tend to have smaller variances. To quantify this bias–variance tradeoff, we constructed a selection scenario in which GC content is stationary and the substitution process matches the HKY model, so that all the models considered in this study (HKY, HKY-NH, GTR, GTR-NH, and GTR-NH$_b$) are correct. We simulated a large alignment of 777,600 sites and generated bootstrap replicate data sets of smaller sizes, by sampling sites with replacement, to study the variation among data sets. Figure 7 shows "confidence intervals" for $d_{up,pu}$ in the m lineage for the AWP method at different sample sizes. At reduced sequence lengths, the confidence intervals became much larger, with the parameter-rich models GTR-NH and GTR-NH$_b$ showing the widest intervals. The large variances in parameter estimates in the small data sets may become a major concern if the nonstationary models such as GTR-NH$_b$ are applied to a single short gene. We advise caution in such an analysis. We suggest that the NH$_b$ models may be useful if it is appropriate to pool genes or genomic regions into one analysis. Pooling may be appropriate in analysis of synonymous sites or noncoding regions (as is the focus here) but less so in analysis of coding genes or protein sequences because of the heterogeneous selection pressures among sites in a protein or among different proteins.

Figure 7 suggests that the absolute differences among the five models become very small when the sequence length reaches 5000, so that at such data sizes, the cost of using the parameter-rich GTR-NH$_b$ model has decreased to insignificant levels. Note, however, that this cutoff is specific to the scenarios we simulated here and may not apply to other data sets. In particular, the information content in the sequence data depends on the sequence divergence levels (branch lengths). The sequences generated in our simulation are highly similar and thus lack information about the parameters (see *Theory and Methods*). At higher divergences (*i.e.*, with more informative data sets), the required sample size may be much smaller (with 1000 or 500 sites, say). Conversely, if the sequences are even more similar, one will need even larger samples for the sampling errors in the MLEs to become negligible.

As discussed earlier, the LRT allows us to test for the goodness of fit of the model. In Table S1, we show the log-likelihood values under different models for data simulated under various selection schemes for two levels of GC bias ($GC_{initial}$ = 0.7 and 0.95). For those data, the LRT always chose the simplest model that fits the data adequately, without either underfitting or overfitting. We note that for short sequences (1000 sites), the LRT may not have much power in rejecting inadequate models. For example, at the moderate base composition bias with $GC_{initial}$ = 0.7 for the complex nonstationary scheme nstC, significant test results were observed only between HKY and HKY-NH while GTR-NH$_b$ was not favored in comparison with HKY-NH (Table S3). In this case the differences between HKY-NH and GTR-NH$_b$ were still considerable, even though the effect is much smaller than at the extreme GC bias of $GC_{initial}$ = 0.95. With the small sample size the LRT did not have sufficient power and model selection for accurate ancestral inference is challenging. It is clear that to make reliable inference of the substitution pattern that is changing on the tree requires a large amount of data.

Discussion

Errors in ancestral reconstruction in AWP and stochastic mapping methods

The AWP method uses posterior probabilities calculated using the MLEs of parameters (*i.e.*, empirical Bayes) as weights to average over multiple ancestral reconstructions. If the model

Table 2 True exchangeability parameters *b* and *e* in the GTR model in the nstC simulation scheme and their 95% confidence intervals under the GTR-NH and GTR-NH$_b$ models

	True values		GTR-NH		GTR-NH$_b$	
Lineage	*b*	*e*	*b*	*e*	*b*	*e*
			$GC_{initial}$ = 0.6 (*Ns* = 0.20)			
m	1.071	0.935			1.021–1.127	0.888–0.978
t	1.541	0.685	1.178–1.220	0.820–0.847	1.504–1.660	0.662–0.720
e	1.233	0.822			1.201–1.309	0.803–0.872
			$GC_{initial}$ = 0.7 (*Ns* = 0.43)			
m	1.156	0.871			1.129–1.239	0.830–0.927
t	2.623	0.482	1.415–1.473	0.680–0.701	2.435–2.720	0.449–0.497
e	1.574	0.674			1.435–1.592	0.630–0.693
			$GC_{initial}$ = 0.8 (*Ns* = 0.69)			
m	1.271	0.801			1.227–1.388	0.782–0.873
t	5.410	0.338	1.803–1.861	0.545–0.560	4.948–5.686	0.316–0.343
e	2.164	0.541			2.036–2.309	0.522–0.573
			$GC_{initial}$ = 0.9 (*Ns* = 1.10)			
m	1.475	0.709			1.470–1.760	0.692–0.795
t	18.205	0.225	2.374–2.484	0.405–0.419	15.961–19.156	0.217–0.241
e	3.641	0.405			3.474–4.049	0.389–0.424
			$GC_{initial}$ = 0.95 (*Ns* = 1.47)			
m	1.700	0.637			1.608–2.071	0.610–0.758
t	61.130	0.169	2.822–3.098	0.334–0.352	49.374–65.159	0.165–0.182
e	6.113	0.322			5.288–6.416	0.296–0.338

The confidence intervals for *b* and *e* were calculated using 300 bootstrap replicates, generated by bootstrap resampling with replacement, of the 777,600 sites in the simulated data set. Under GTR-NH, all lineages have the same exchangeability parameters. The scheme nstC leads to different *b* and *e* values (see Equation 1) in different lineages. Other exchangeability parameters (*a*, *c*, *d*) have the true value 1 and their estimates under both models are correct (see Table S2).

and parameter values used are exactly correct, and if all possible reconstructions are considered in the averaging, the AWP method will recover the correct base compositions in the ancestral nodes. Note that AWP is very similar to the approach of sampling ancestral reconstructions according to their posterior probabilities (Williams *et al.* 2006; Goldstein *et al.* 2015); indeed, averaging over all possible ancestral reconstructions (AWP) is equivalent to sampling if the number of samples is infinite. Used to count substitutions along a branch, the AWP method may suffer from two sources of errors. The first is possible errors in the calculated posterior probabilities for reconstructions because the parameter estimates are in error. The errors in parameter estimates may be systematic, caused by violations of model assumptions, or random, caused by the limited sequence length. In our simulation, the unrealistic nature of the assumed HKY model relative to the actual GTR substitution rate matrix is seen to cause considerable systematic biases in substitution counts and in the $d_{up,pu}$ index (Figure 2 and Figure 3). When the true process is stationary, HKY creates a spurious nonstationary pattern, similar to the use of the MP method (Figure 3). For estimating the pattern of nucleotide substitution, GTR is preferred to HKY (Yang 1994). Furthermore, when the selection regime is fluctuating on the tree but the data are analyzed under a stationary model (HKY or GTR), the incorrect stationarity assumption is seen to have an even greater impact on the substitution counts, whatever method of inference is used (AWP or EMC) (Figure 4 and Figure 5).

The second source of errors in the AWP method is its failure to correct for multiple hits within a branch. When a reconstruction assigns nucleotides *i* and *j* at the two ends of the branch, this is counted as an $i \rightarrow j$ substitution. The method is very similar to the simple *p* distance, which treats the raw proportion of different sites between two sequences as an estimate of the number of substitutions per site, ignoring the possibility for multiple hits. This error is more serious for longer branches. It should be possible to correct for this bias, using a standard multiple-hit correction under the GTR model (*e.g.*, Gu and Li 1996; Yang and Kumar 1996) or using the expected counts in stochastic mapping (Hobolth and Jensen 2005; Minin and Suchard 2008b; Tataru and Hobolth 2011). This is not pursued here, partly because the bias due to failure to correct for multiple hits within the branch is avoided by the EMC method. In our simulation, the bias caused by multiple hits within the branch is not so important (Figure 2B and Figure 4, C–E) as the large systematic errors caused by the violations of assumptions discussed above (Figure 3 and Figure 4).

Minin and Suchard (2008a) and O'Brien *et al.* (2009) have advocated stochastic mapping as a general approach for producing substitution counts. Stochastic mapping calculates expected character changes at every site conditioned on the ancestral reconstructions and then averages over ancestral reconstructions weighted by their posterior probabilities (*e.g.*, O'Brien *et al.* 2009). This is equivalent to the AWP method except for its correction for multiple hits within the branch. O'Brien *et al.* (2009) have argued that stochastic mapping, even if implemented under a simplistic model, may generate substitution counts that are suitable for testing hypotheses and diagnosing model violations. The authors demonstrated the utility of stochastic mapping through calculation of sequence distances and estimation of

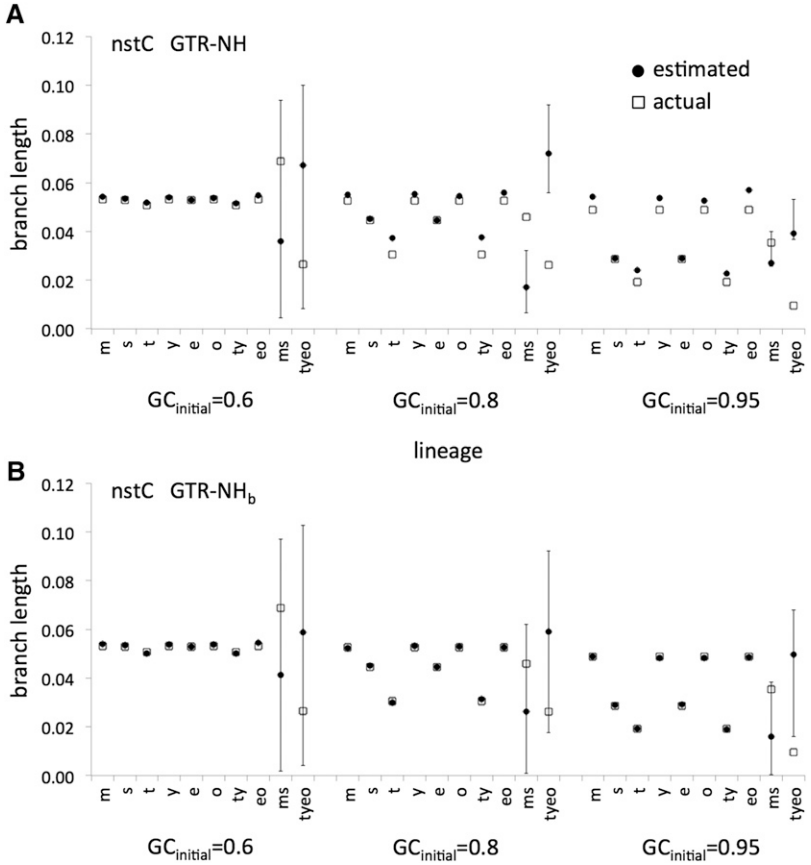

Figure 6 (A and B) Actual and estimated branch lengths under the GTR-NH (A) and GTR-NH$_b$ (B) models for data simulated under the nstC scheme. Three $GC_{initial}$ levels (0.6, 0.8, and 0.95) are used. Actual branch length was calculated as the observed number of substitutions on the branch divided by the number of sites. Estimated branch length was calculated as the expected number of substitutions per site when the base compositions are in equilibrium for the substitution rate matrix (Yang and Roberts 1995), but the new definition of Equation 7 gave similar values. The two branch lengths around the root of the tree should be ignored. Shown are the averages and 95% confidence intervals over 300 bootstrap replicate data sets, generated by bootstrap resampling of the 777,600 sites in the simulated data set.

the nonsynonymous/synonymous rate ratio (ω). However, to study complex nonstationary patterns of nucleotide substitution, stochastic mapping under simple and unrealistic models may not be accurate enough, because the posterior probabilities for the ancestral reconstructions calculated under simplistic and wrong models (such as HKY) may be systematically biased. In our simulations, AWP (or indeed even SBR) under the simple stationary HKY model did indicate the nonstationary nature of nucleotide substitution (Figure 4A), but it also indicated a spurious nonstationary trend when the true substitution process is stationary (Figure 3A). For the purpose of inferring complex substitution patterns, we advocate parametric likelihood models that accommodate the main features of the substitution process, with ML used for parameter estimation and LRT for hypothesis testing.

Limitation of the EMC method

The EMC method estimates the expected number of nucleotide substitutions along the branch by taking into account the nonstationary nature of the Markov substitution process. It corrects for multiple hits within a branch and more importantly accounts for the changing base compositions over time. Implemented under the nonstationary GTR-NH and GTR-NH$_b$ models, it produced highly accurate ancestral inference in our simulation. The nonstationary GTR models appear complex enough to accommodate a wide range of evolutionary scenarios and may be useful for studying the complex nucleotide substitution process in genomic data sets. While the rate matrix under the GTR model assumes time reversibility and symmetrical substitution counts (with $\pi_i q_{ij} = \pi_j q_{ji}$) at equilibrium, asymmetry in the substitution process can be accommodated by allowing the base compositions to drift over time, thus allowing the model to be used to estimate substitution counts when the substitution process is nonstationary (see Figure 3C).

Nevertheless, we note here a few limitations of the EMC method or the nonstationary models. First, those models involve many parameters so that the random sampling errors in the parameter estimates may be a major concern if the analyzed data set is small (see discussion above). Second, the models are nucleotide based. To calculate the probabilities of ancestral codon states one should either apply those models to analyze the fourfold degenerate sites at the third codon position only (Akashi *et al.* 2007 and this study) or implement nonstationary codon models to explicitly model mutational bias and weak selection on codon usage (Nielsen *et al.* 2007). Third, the models assume independent substitutions among sites, an assumption that appears to be seriously violated in mammalian genome data (see Arndt *et al.* 2003).

Methods and biases of ancestral sequence reconstruction

There has been much discussion in the literature about the accuracy of ancestral sequence reconstruction and its

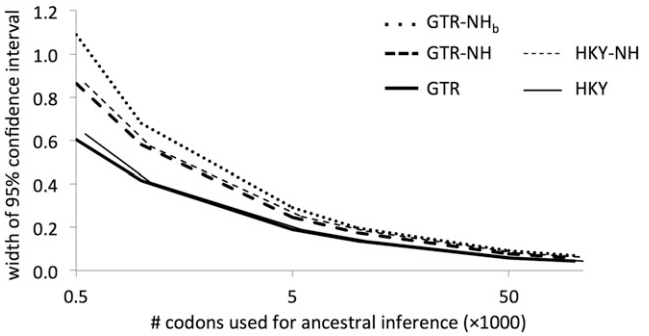

Figure 7 Bootstrap confidence intervals for $d_{up,pu}$ at different sequence lengths, illustrating the cost of parameter-rich models. A data set of the 777,600 sites was generated under a model consistent with HKY, with $GC_{initial} = 0.95$ (that is, $\pi_T = \pi_A = 0.025$, $\pi_C = \pi_G = 0.475$, $\kappa = 1$). Then 300 bootstrap replicate data sets of specified sequence length were generated by resampling sites with replacement and were analyzed using the AWP method under five models: HKY, HKY-NH, GTR, GTR-NH, and GTR-NH$_b$.

applicability to study the properties and functions of ancestral proteins (Jordan *et al.* 2005; Goldstein and Pollock 2006; Williams *et al.* 2006). A perceived contradiction exists in that on one hand the accuracy of the reconstructed ancestral sequences (measured by the proportion of correctly reconstructed sites) is typically very high (*e.g.*, Yang *et al.* 1995; Chang *et al.* 2002b; Hanson-Smith *et al.* 2010) while on the other the amino acid compositions and thermostability of the reconstructed proteins appear to be seriously biased (*e.g.*, Krishnan *et al.* 2004; Williams *et al.* 2006). Note that the most commonly used reconstruction methods have been MP and likelihood with SBR (Yang *et al.* 1995; Koshi and Goldstein 1996). Our simulation has focused on estimating base compositions at ancestral nodes and counting substitutions along the branches, but we have also generated results concerning the accuracy of ancestral sequence reconstruction, which may shed light on the controversy.

It may be useful to first clarify the statistical nature of the inference problem as this is frequently misrepresented in the literature. In all Markov models of character evolution used in phylogenetics, the ancestral states are random variables and do not occur in the likelihood function (the probability of the modern sequences given the model parameters), which averages over and integrates out the ancestral states (Felsenstein 1981). Thus the "maximum-likelihood" method for ancestral reconstruction is a misnomer. There have been attempts to estimate ancestral character states by maximizing what was thought to be the "likelihood function," but those do not constitute valid statistical methods (see Yang 2006, p. 124). Instead the proper method is to calculate the conditional (posterior) probabilities of ancestral states given the data and model parameters. This is the EB method (Yang *et al.* 1995; Koshi and Goldstein 1996). EB is considered a likelihood method. Note that likelihood is a general methodology of using the likelihood function for statistical inference (and thus includes EB for estimating random variables) while maximum likelihood is a specific method for estimating parameters by maximizing the likelihood function (Edwards 1972); EB is a likelihood method but not a maximum-likelihood method. The full or hierarchical Bayesian (FB) method (Huelsenbeck and Bollback 2001) differs from EB in that it accommodates uncertainties in model parameters (such as branch lengths and the exchangeability and frequency parameters) by assigning priors on them and integrating over them, typically achieved using a Markov chain Monte Carlo (MCMC) algorithm. The difference between EB and FB is small and unimportant in analysis of large data sets (as in this study), but FB arguably has an advantage in analysis of small data sets (but see Hanson-Smith *et al.* 2010).

With both EB and FB, one can either use the SBR or average over the ancestral reconstructions, using the posterior probabilities as weights (AWP), or sample ancestral states using the posterior probabilities. We do not distinguish between the latter two approaches, because averaging (if over all possible ancestral states) and sampling (if the sample size or the length of the MCMC is infinite) are equivalent. Note that an MCMC implementation of FB produces posterior probabilities for ancestral states, which can be used to generate the *maximum a posteriori* (MAP) reconstruction and to apply SBR. Note that SBR is not a unique feature of likelihood (EB) and can also be applied to the Bayesian (FB) method, and similarly sampling or averaging is not a unique feature of Bayesian (FB) and can be applied to likelihood as well. There is essentially just one valid statistical method for ancestral state reconstruction (that is, Bayesian, including both EB and FB) even though one can use the calculated posterior probabilities differently (SBR or AWP).

Among all the factors discussed here and elsewhere (*e.g.*, Goldstein and Pollock 2006), the use of the SBR while ignoring the suboptimal reconstructions is the most important in producing serious biases if we are interested in certain features of the whole ancestral sequence. This bias is easy to explain. Suppose we want to reconstruct the ancestor for three sequences with high GC contents. At a site with data TCC (T, C, and C in the three sequences at an alignment site), the most likely state for the ancestor is C, whatever method we use (*e.g.*, MP or EB) or whatever substitution model we use in the EB calculation. However, counting a C at every TCC site will lead to an overcount of C when we consider the whole sequence. The bias is opposite at the TTC site, but as the sequence is GC rich, there will be many more TCC sites than TTC sites. Overall we will infer even higher frequencies of the common nucleotides (C) in the ancestors than in modern sequences, suggesting a trend of common to rare changes. This bias is due to the use of SBR. It exists for parsimony, likelihood (EB), and Bayesian (FB) approaches and exists even if the correct substitution model is assumed in calculating the posterior probabilities.

Our simulation showed the same patterns as observed by others: high sequence reconstruction accuracy by SBR and strong bias in compositions. Many previous studies have observed high accuracy of ancestral sequence reconstruction

by SBR, with >90% or 95% of sites in the sequence correctly reconstructed (*e.g.*, Yang *et al.* 1995; Williams *et al.* 2006). In our simulation, the probability for correct joint reconstruction at the variable sites (that is, for correctly reconstructing all six nodes in the tree of Figure 1 to a variable site) is >95% (Table S4). The accuracy will be even higher if the constant sites are included in the calculation or if only the node is considered. This high accuracy is mainly due to the high similarity of the sequences. At the same time, SBR produced large systematic biases and spurious trends of base composition evolution. Figure S2 shows the performance of SBR implemented under the true GTR-NH$_b$ model for data simulated under the nstC scheme. SBR created large biases, in sharp contrast to highly accurate results obtained from using the AWP and EMC methods under the same model (Figure 5). The poor performance of SBR is very similar to the finding of Williams *et al.* (2006), who used computer simulation to evaluate the accuracy of ancestral sequence reconstruction. Note that those authors' ML method is our EB with SBR while their Bayesian inference (BI) is our EB with sampling (*i.e.*, AWP). Williams *et al.* (2006) found that ancestral sequences reconstructed using MP and SBR had even higher thermostability than the true proteins generated in the simulation, while AWP did not show similar bias. The authors convincingly demonstrated that the notion that errors in ancestral reconstruction should lead to less stable ancestral proteins (Thornton 2004) does not hold.

Nevertheless, there does not appear to be a contradiction between the high accuracy of sequence reconstruction by SBR or parsimony and the large biases in composition reconstruction. Sequence reconstruction accuracy appears to be invariably measured on a per-site basis, but accumulation of small reconstruction errors at individual sites may lead to a substantial bias when one considers a property of the whole sequence, such as the base compositions, the thermostability, and function of a protein. Even though the accuracy of reconstructing a single site by SBR is very high, the probability for correct reconstruction of the whole sequence, calculated by multiplying the probabilities over individual sites, is typically vanishingly small. In other words, the probability of correctly reconstructing the whole sequence is very low. Note that both accuracy measures are calculated in PAML (Yang 1997). Ignoring suboptimal reconstructions in SBR leads to systematic biases in the substitution counts along the branches of the tree (Akashi *et al.* 2007 and this study) and in base compositions in the ancestral sequences (Collins *et al.* 1994; Perna and Kocher 1995; Eyre-Walker 1998, and this study), and it is not surprising that it may also lead to biases in the functional properties of the reconstructed protein. We suggest that caution be exercised if SBR or MP methods are used to reconstruct ancestral sequences and that the biases caused by those methods be carefully considered.

In the case where it is possible to calculate the property of interest of the ancestral protein, the averaging or sampling method (AWP) may be effective to correct for the bias in ancestral reconstruction (SBR). To study complex patterns of nucleotide substitution, the use of realistic substitution models also becomes important. Indeed, our simulation suggests that the AWP and EMC methods implemented under the nonstationary models can produce highly reliable substitution counts along branches and reliable base compositions at internal nodes on the tree, even if SBR applied to the same data generates large biases. For studies that synthesize ancestral proteins and examine their biochemical properties in the laboratory (Chang *et al.* 2002a; Thornton 2004), the EMC method is not directly applicable since it does not generate whole sequences, and the sampling method (AWP) increases the experimental cost considerably, since many ancestral proteins have to be examined and averaged over.

Implications to study of synonymous codon usage

While our analysis has used "empirical" models of nucleotide substitution such as HKY and GTR that describe nucleotide substitution rates without considering the underlying biological factors that influence rates, our simulation uses a "mechanistic" model that explicitly considers the population genetic process of mutation and selection. With a combined analysis of multiple genes with different codon usage or base compositions, we envisage that the estimated substitution rates from the empirical models may be converted to evolutionary parameters of mutation and selection that characterize the forces of gene sequence evolution. Furthermore, even though the mutation model in the simulation assumed an equal mutation rate between any two nucleotides, mutation bias such as the transition/transversion rate ratio can be accommodated in a straightforward manner in the GTR-NH model.

In general, suppose the mutation process can be described using a GTR mutation model, with the rate of mutation from nucleotides i to j given as $\mu_{ij} = a_{ij}\pi_j^*$, with $a_{ij} = a_{ji}$ for all $i \neq j$. Here π_j^* reflects mutation bias; if π_C^* is large, mutations are biased toward C. Suppose weak selection operates on base composition/codon usage, such that different nucleotides have different fitness $F_i = 2Nf_i$ for nucleotide i, and the $i \rightarrow j$ mutation has the scaled selection coefficient $S_{ij} = 2Ns_{ij} = F_j - F_i$. The $i \rightarrow j$ substitution rate per generation is then

$$q_{ij} = N\mu_{ij} \times \frac{2s_{ij}}{1 - e^{2Ns_{ij}}} = \mu_{ij} \times \frac{2S_{ij}}{1 - e^{-S_{ij}}} = \left[a_{ij} \times \frac{F_j - F_i}{e^{F_j} - e^{F_i}}\right] \times \left(\pi_j^* e^{F_j}\right). \quad (13)$$

Here the quantity in the brackets is symmetrical for $i \rightarrow j$ and $j \rightarrow i$, while the quantity in the parentheses depends on j but not on i. Thus the substitution process specified by the rate matrix $Q = \{q_{ij}\}$ is time reversible, with the stationary distribution given as

$$\pi_j \propto \pi_j^* e^{F_j}. \quad (14)$$

Here the proportionality constant is chosen to ensure that π_j sum to one. This clearly reflects the effects of both the mutation bias (π_j^*) and selection (e^{F_j}). The derivation here is very similar to equation 4 of Yang and Nielsen (2008).

Suppose we use the HKY mutation model with transition/transversion rate ratio κ and mutation bias parameters $\pi_T^*, \pi_C^*, \pi_A^*$, and π_G^* and suppose the fitness parameters for the nucleotides are F_T, F_C, F_A, and F_G. By matching the mechanistic model of Equation 13 with the empirical model of Equation 1, we obtain the exchangeability parameters a, b, c, d, and e (with $f = 1$) of Equation 1 as

$$\begin{aligned} a &= \frac{F_T - F_C}{e^{F_T} - e^{F_C}} \times \frac{e^{F_A} - e^{F_G}}{F_A - F_G}, \\ b &= \frac{1}{\kappa}\frac{F_T - F_A}{e^{F_T} - e^{F_A}} \times \frac{e^{F_A} - e^{F_G}}{F_A - F_G}, \\ c &= \frac{1}{\kappa}\frac{F_T - F_G}{e^{F_T} - e^{F_G}} \times \frac{e^{F_A} - e^{F_G}}{F_A - F_G}, \\ d &= \frac{1}{\kappa}\frac{F_C - F_A}{e^{F_C} - e^{F_A}} \times \frac{e^{F_A} - e^{F_G}}{F_A - F_G}, \\ e &= \frac{1}{\kappa}\frac{F_C - F_G}{e^{F_C} - e^{F_G}} \times \frac{e^{F_A} - e^{F_G}}{F_A - F_G}. \end{aligned} \tag{15}$$

The stationary distribution is given by Equation 14. The simulation model of this study (Equations 8–10) is a special case of the above, with $\kappa = 1$ and $\pi_T^* = \pi_C^* = \pi_A^* = \pi_G^*$ in the mutation model and $F_T = F_A = 0$ and $F_C = F_G = 2Ns$ for selection on base compositions. The stationary distribution is given by $\pi_T = \pi_A = (1/2)(1 - \text{GC})$ and $\pi_C = \pi_G = (1/2)\text{GC}$ (Equation 11), and the exchangeability parameters are given by Equation 15 as $a = c = d = 1$, $b = (e^{2Ns} - 1)/2Ns$, and $e = (1 - e^{-2Ns})/2Ns$. Note that when $x \to y$, $(e^x - e^y)/(x - y) \to e^x$.

In the above we have formulated the model at the nucleotide level. While nonsynonymous mutations are under stronger selection (mostly purifying selection) than synonymous mutations, the two types of mutations operate on different timescales, so that one can use closely related species to study synonymous codon usage while essentially ignoring nonsynonymous mutations. Alternatively one can formulate the model at the level of codons and analyze protein-coding gene sequences directly (Nielsen *et al.* 2007; Yang and Nielsen 2008).

While the mutation and selection parameters are confounded in Equations 13 and 14 when a single gene is analyzed, they will be estimable if we analyze multiple genes (especially genes of different base compositions) simultaneously and if we assume that the mutation parameters are shared among genes. Another strategy is to analyze the protein-coding genes together with the noncoding regions, assuming shared mutation parameters. We are implementing nonstationary nucleotide and codon models for combined analysis of protein-coding genes to estimate the mutation and selection parameters. The results from this simulation study support the feasibility of the approach.

Program availability

Maximum-likelihood estimation and ancestral reconstruction (that is, the SBR, AWP, and EMC methods) under the nonstationary models GTR-NH and GTR-NH$_b$, as well as under other nonstationary models based on HKY and F84, are implemented in the BASEML program in the PAML package (version 4.8). Programs for processing BASEML output to implement the AWP method are available from the authors upon request.

Acknowledgments

We are grateful to D. Pollock, R. Goldstein, T. Ohta, N. Osada, N. Mishra and K Kawashima, and two anonymous reviewers for their comments that helped to improve this manuscript. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (to Z.Y.) and by an award from the National Institute of Genetics Collaborative Research Program (2012-A16, 2013-A18, 2014-A16).

Literature Cited

Ababneh, F., L. Jermiin, C. Ma, and J. Robinson, 2006 Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22: 1225–1231.

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136: 927–935.

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139: 1067–1076.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144: 1297–1307.

Akashi, H., P. Goel, and A. John, 2007 Ancestral state inference and the study of codon bias evolution: implications for molecular evolutionary analysis of the *Drosophila melanogaster* subgroup. PLoS ONE 2: e1065.

Aoki, S., M. Ito, and W. Iwasaki, 2013 From beta- to alphaproteobacteria: the origin and evolution of rhizobial nodulation genes *nodIJ*. Mol. Biol. Evol. 30: 2494–2508.

Arndt, P. F., D. A. Petrov, and T. Hwa, 2003 Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol. Biol. Evol. 20: 1887–1896.

Barry, D., and J. A. Hartigan, 1987 Statistical analysis of hominoid molecular evolution. Stat. Sci. 2: 191–210.

Bauer DuMont, V., J. C. Fay, P. P. Calabrese, and C. F. Aquadro, 2004 DNA variability and divergence at the notch locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. Genetics 167: 171–185.

Bauer DuMont, V., N. D. Singh, M. H. Wright, and C. F. Aquadro, 2009 Locus-specific decoupling of base composition evolution at synonymous sites and introns along the Drosophila melanogaster and Drosophila sechellia lineages. Genome Biol. Evol. 1: 67–74.

Begun, D. J., 2001 The frequency distribution of nucleotide variation in *Drosophila simulans*. Mol. Biol. Evol. 18: 1343–1352.

Blanquart, S., and N. Lartillot, 2006 A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23: 2058–2071.

Blanquart, S., and N. Lartillot, 2008 A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25: 842–858.

Bulmer, M. G., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907.

Campos, J. L., K. Zeng, D. J. Parker, B. Charlesworth, and P. R. Haddrill, 2013 Codon usage bias and effective population sizes on the X chromosome *vs.* the autosomes in Drosophila melanogaster. Mol. Biol. Evol. 30: 811–823.

Chang, B. S., M. A. Kazmi, and T. P. Sakmar, 2002a Synthetic gene technology: applications to ancestral gene reconstruction and structure-function studies of receptors. Methods Enzymol. 343: 274–294.

Chang, B. S., K. Jonsson, M. A. Kazmi, M. J. Donoghue, and T. P. Sakmar, 2002b Recreating a functional ancestral archosaur visual pigment. Mol. Biol. Evol. 19: 1483–1489.

Collins, T. M., P. H. Wimberger, and G. J. P. Naylor, 1994 Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst. Biol. 43: 482–496.

Comeron, J. M., 2005 Intragenic Hill-Robertson interference influences selection intensity on synonymous mutations in *Drosophila*. Mol. Biol. Evol. 22: 2519–2530.

Dayhoff, M. O., R. V. Eck, M. A. Chang, and M. R. Sochard, 1965 *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.

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

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

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, *Arabidopsis*. Proc. Natl. Acad. Sci. USA 96: 4482–4487.

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

Dutheil, J., and B. Boussau, 2008 Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol. 8: 255.

Dutheil, J., T. Pupko, A. Jean-Marie, and N. Galtier, 2005 A model-based approach for detecting coevolving positions in a molecule. Mol. Biol. Evol. 22: 1919–1928.

Eanes, W. F., M. Kirchner, J. Yoon, C. H. Biermann, I. N. Wang *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the G6pd locus in *Drosophila melanogaster* and *D. simulans*. Genetics 144: 1027–1041.

Edwards, A. W. F., 1972 *Likelihood*. Cambridge University Press, Cambridge, UK.

Eyre-Walker, A., 1998 Problems with parsimony in sequences of biased base composition. J. Mol. Evol. 47: 686–690.

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

Fisher, R., 1930 The distribution of gene ratios for rare mutations. Proc. R. Soc. Edinb. **50:** 205–220.

Fitch, W. M., 1971 Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20: 406–416.

Fitch, W. M., J. M. Leiter, X. Q. Li, and P. Palese, 1991 Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA 88: 4270–4274.

Foster, P. G., 2004 Modeling compositional heterogeneity. Syst. Biol. 53: 485–495.

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

Gardiner, A., D. Barker, R. K. Butlin, W. C. Jordan, and M. G. Ritchie, 2008 Evolution of a complex locus: exon gain, loss and divergence at the *Gr39a* locus in *Drosophila*. PLoS ONE 3: e1513.

Gaucher, E. A., J. M. Thomson, M. F. Burgan, and S. A. Benner, 2003 Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425: 285–288.

Gaucher, E. A., S. Govindarajan, and O. K. Ganesh, 2008 Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451: 704–707.

Gojobori, T., W. H. Li, and D. Graur, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18: 360–369.

Goldstein, R. A., and D. D. Pollock, 2006 Observations of amino acid gain and loss during protein evolution are explained by statistical bias. Mol. Biol. Evol. 23: 1444–1449.

Goldstein, R. A., S. T. Pollard, S. D. Shah, and D. D. Pollock, 2015 Nonadaptive amino acid convergence rates decrease over time. Mol. Biol. Evol. (in press).

Groussin, M., and M. Gouy, 2011 Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. Mol. Biol. Evol. 28: 2661–2674.

Groussin, M., B. Boussau, and M. Gouy, 2013 A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. 62: 523–538.

Gu, X., and W.-H. Li, 1996 A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc. Natl. Acad. Sci. USA 93: 4671–4676.

Gueguen, L., S. Gaillard, B. Boussau, M. Gouy, M. Groussin *et al.*, 2013 Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30: 1745–1750.

Haddrill, P. R., D. Bachtrog, and P. Andolfatto, 2008 Positive and negative selection on noncoding DNA in *Drosophila simulans*. Mol. Biol. Evol. 25: 1825–1834.

Hanson-Smith, V., B. Kolaczkowski, and J. W. Thornton, 2010 Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol. Biol. Evol. 27: 1988–1999.

Hartigan, J. A., 1973 Minimum evolution fits to a given tree. Biometrics 29: 53–65.

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

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24: 1792–1800.

Hobolth, A., and J. L. Jensen, 2005 Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Stat. Appl. Gen. Mol. Biol. 4: Article 18.

Huelsenbeck, J. P., and J. P. Bollback, 2001 Empirical and hierarchical Bayesian estimation of ancestral states. Syst. Biol. 50: 351–366.

Huelsenbeck, J. P., J. P. Bollback, and A. M. Levine, 2002 Inferring the root of a phylogenetic tree. Syst. Biol. 51: 32–43.

Jayaswal, V., L. S. Jermiin, L. Poladian, and J. Robinson, 2011 Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst. Biol. 60: 74–86.

Jayaswal, V., T. K. F. Wong, J. Robinson, L. Poladian, and L. S. Jermiin, 2014 Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63: 726–742.

Jones, D. T., W. R. Taylor, and J. M. Thornton, 1992 The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8: 275–282.

Jordan, I. K., F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin *et al.*, 2005 A universal trend of amino acid gain and loss in protein evolution. Nature 433: 633–638.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Khelifi, A., J. Meunier, L. Duret, and D. Mouchiroud, 2006 GC content evolution of the human and mouse genomes: insights from the study of processed pseudogenes in regions of different recombination rates. J. Mol. Evol. 62: 745–752.

Kilman, R. M., 1999 Recent selection on synonymous codon usage in *Drosophila*. J. Mol. Biol. 49: 343–351.

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol. Biol. Evol. 10: 1239–1258.
Ko, W. Y., R. M. David, and H. Akashi, 2003 Molecular phylogeny of the Drosophila melanogaster species subgroup. J. Mol. Evol. 57: 562–573.
Koshi, J. M., and R. A. Goldstein, 1996 Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. 42: 313–320.
Krishnan, N. M., H. Seligmann, C. B. Stewart, A. P. De Koning, and D. D. Pollock, 2004 Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. Mol. Biol. Evol. 21: 1871–1883.
Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24: 337–345.
Liao, H. X., R. Lynch, T. Zhou, F. Gao, S. M. Alam *et al.*, 2014 Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature 496: 469–476.
Lohse, K., and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics 189: 977–987.
Marais, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19: 330–338.
McVean, G. A., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics 157: 245–257.
Messier, W., and C.-B. Stewart, 1997 Episodic adaptive evolution of primate lysozymes. Nature 385: 151–154.
Minin, V. N., and M. A. Suchard, 2008a Fast, accurate and simulation-free stochastic mapping. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363: 3985–3995.
Minin, V. N., and M. A. Suchard, 2008b Counting labeled transitions in continuous-time Markov models of evolution. J. Math. Biol. 56: 391–412.
Moriyama, E. N., and J. R. Powell, 1997 Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. 45: 514–523.
Nagylaki, T., 1983 Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. USA 80: 6278–6281.
Nielsen, R., V. L. Bauer DuMont, M. J. Hubisz, and C. F. Aquadro, 2007 Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol. Biol. Evol. 24: 228–235.
O'Brien, J. D., V. N. Minin, and M. A. Suchard, 2009 Learning to count: robust estimates for labeled distances between molecular sequences. Mol. Biol. Evol. 26: 801–814.
Osada, N., and H. Akashi, 2012 Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome c oxidase complex. Mol. Biol. Evol. 29: 337–346.
Perna, N. T., and T. D. Kocher, 1995 Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. 12: 359–361.
Poh, Y. P., C. T. Ting, H. W. Fu, C. H. Langley, and D. J. Begun, 2012 Population genomic analysis of base composition evolution in Drosophila melanogaster. Genome Biol. Evol. 4: 1245–1255.
Presgraves, D. C., 2005 Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr. Biol. 15: 1651–1656.
Pupko, T., I. Pe'er, R. Shamir, and D. Graur, 2000 A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. 17: 890–896.
Shindyalov, I. N., N. A. Kolchanov, and C. Sander, 1994 Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. 7: 349–358.
Singh, N. D., P. F. Arndt, A. G. Clark, and C. F. Aquadro, 2009 Strong evidence for lineage and sequence specificity of substitution rates and patterns in Drosophila. Mol. Biol. Evol. 26: 1591–1605.
Suzuki, Y., and T. Gojobori, 1999 A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16: 1315–1328.
Takano, T., 2001 Local changes in GC/AT substitutions biases and in crossover frequencies on Drosophila chromosome. Mol. Biol. Evol. 18: 606–619.
Tataru, P., and A. Hobolth, 2011 Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. BMC Bioinformatics 12: 465.
Tavaré, S., 1986 Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math Life Sci. 17: 57–86.
Terekhanova, N. V., G. A. Bazykin, A. Neverov, A. S. Kondrashov, and V. B. Seplyarskiy, 2013 Prevalence of multinucleotide replacements in evolution of primates and Drosophila. Mol. Biol. Evol. 30: 1315–1325.
Thornton, J., 2004 Resurrecting ancient genes: experimental analysis of extinct molecules. Nat. Rev. Genet. 5: 366–375.
Tuffery, P., and P. Darlu, 2000 Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol. Biol. Evol. 17: 1753–1759.
Vicario, S., C. E. Mason, K. P. White, and J. R. Powell, 2008 Developmental stage and level of codon usage bias in Drosophila. Mol. Biol. Evol. 25: 2269–2277.
Whelan, S., and N. Goldman, 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18: 691–699.
Williams, P. D., D. D. Pollock, B. P. Blackburne, and R. A. Goldstein, 2006 Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput. Biol. 2: e69.
Yang, Z., 1994 Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39: 105–111.
Yang, Z., 1995 On the general reversible Markov-process model of nucleotide substitution: a reply to Saccone *et al.* J. Mol. Evol. 41: 254–255.
Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13: 555–556.
Yang, Z., 2006 *Computational Molecular Evolution*. Oxford University Press, Oxford.
Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591.
Yang, Z., and S. Kumar, 1996 Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13: 650–659.
Yang, Z., and R. Nielsen, 2008 Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25: 568–579.
Yang, Z., and D. Roberts, 1995 On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12: 451–458.
Yang, Z., S. Kumar, and M. Nei, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641–1650.
Zhang, J., S. Kumar, and M. Nei, 1997 Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol. Biol. Evol. 14: 1335–1338.
Zharkikh, A., 1994 Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39: 315–329.
Zou, L., E. Susko, C. Field, and A. J. Roger, 2012 Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry-Hartigan model. Syst. Biol. 61: 927–940.

Communicating editor: M. W. Hahn

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177386/-/DC1

Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution

Tomotaka Matsumoto, Hiroshi Akashi, and Ziheng Yang

DOI: 10.1534/genetics.115.177386

Supplementary Material for Matsumoto, Akashi, & Yang, "Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution"

Table S1. Log likelihood values for different models in data simulated under various selection schemes with initial GC content at 70% and 95%

Model (p)	st	st2x	nstD	nstC
GC = 0.7				
HKY (13)	0	0	0	0
HKY-NH (44)	4,419	5,172	12,334	14,144
HKY-NH$_b$ (53)	4,476	5,215	12,344	14,384
GTR (17)	**8,867**	**9,384**	8,325	9,322
GTR-NH (48)	8,881	9,401	**12,809**	17,696
GTR-NH$_b$ (93)	8,906	9,418	12,836	**19,629**
GC = 0.95				
HKY (13)	0	0	0	0
HKY-NH (44)	10,049	12,592	69,603	60,849
HKY-NH$_b$ (53)	10,376	12,724	69,161	64,071
GTR (17)	**22,469**	**22,765**	53,786	37,509
GTR-NH (48)	22,486	22,784	**71,427**	69,782
GTR-NH$_b$ (93)	22,510	22,805	71,449	**74,427**

Note.— p is the number of parameters in the model. The log likelihood value (ℓ) for HKY is set to 0, while those for other models are shown as differences from HKY. The simplest correct models are highlighted in bold; these are also the chosen models by the LRT.

Table S2. True rate parameters *a*, *c* and *d* in the GTR model in the nstC simulation scheme and their 95% confidence intervals under the GTR-NH and GTR-NH$_b$ models

	True values			GTR-NH			GTR-NH$_b$		
Lineage	*a*	*c*	*d*	*a*	*c*	*d*	*a*	*c*	*d*
$GC_{initial}$ = 0.6									
m	1	1	1				0.954 - 1.049	0.992 - 1.090	0.960 - 1.044
t	1	1	1	0.972 - 1.010	0.989 - 1.019	0.974 - 1.005	0.956 - 1.046	0.954 - 1.040	0.963 - 1.050
e	1	1	1				0.933 - 1.046	0.978 - 1.071	0.934 - 1.027
$GC_{initial}$ = 0.7									
m	1	1	1				0.984 - 1.094	0.999 - 1.095	0.968 - 1.056
t	1	1	1	0.987 - 1.019	0.995 - 1.026	0.991 - 1.026	0.967 - 1.075	0.935 - 1.008	0.929 - 1.025
e	1	1	1				0.911 - 1.020	0.929 - 1.033	0.976 - 1.066
$GC_{initial}$ = 0.8									
m	1	1	1				0.984 - 1.125	0.983 - 1.089	0.976 - 1.064
t	1	1	1	0.989 - 1.030	0.985 - 1.024	0.984 - 1.013	0.933 - 1.044	0.928 - 1.027	0.900 - 1.002
e	1	1	1				0.946 - 1.068	0.954 - 1.066	0.977 - 1.079
$GC_{initial}$ = 0.9									
m	1	1	1				0.931 - 1.109	0.962 - 1.127	0.986 - 1.091
t	1	1	1	0.980 - 1.022	0.985 - 1.025	0.997 - 1.030	0.945 - 1.078	0.988 - 1.028	0.941 - 1.077
e	1	1	1				0.939 - 1.065	0.978 - 1.097	0.976 - 1.097
$GC_{initial}$ = 0.95									
m	1	1	1				0.983 - 1.213	0.887 - 1.136	0.992 - 1.096
t	1	1	1	0.952 - 1.048	0.961 - 1.021	0.933 - 1.003	0.927 - 1.064	0.933 - 1.071	0.894 - 1.057
e	1	1	1				0.892 - 1.086	0.917 - 1.079	0.891 - 1.028

Note.— The confidence intervals are calculated using 300 bootstrap replicates, generated by bootstrap resampling with replacement, of the 777,600 sites in the simulated dataset. The scheme nstC have the same values of *a*, *c* and *d* in all lineage, and they are estimated correctly both under GTR-NH and GTR-NH$_b$.

Table S3. Log likelihood differences in likelihood ratio tests of models for data of different sizes simulated under the nstC scheme with different initial GC contents

Model	1,000 sites	10,000 sites	100,000 sites
GC = 0.7			
HKY vs HKY-NH	41**	204**	1885**
HKY-NH vs GTR- NH_b	26	77**	752**
GC = 0.95			
HKY vs HKY-NH	104**	819**	8262**
HKY-NH vs GTR - NH_b	39**	157**	1477**

** Significant at the 1% level.

Table S4. Joint reconstruction accuracy at variable sites (and at all sites, in parentheses) for data simulated under the nstC scheme

Model	$GC_{initial} = 0.7$	$GC_{initial} = 0.95$
MP	0.914 (0.965)	0.911 (0.974)
HKY	0.962 (0.981)	0.952 (0.984)
HKY-NH	0.972 (0.984)	0.959 (0.986)
HKY-NH_b	0.972 (0.985)	0.966 (0.988)
GTR	0.970 (0.984)	0.957 (0.985)
GTR-NH	0.976 (0.986)	0.970 (0.989)
GTR-NH_b	0.979 (0.987)	0.983 (0.992)

Note.— Accuracy is measured by the proportion of sites at which the single best joint reconstruction with the highest posterior probability matches the ancestral states recorded in the simulation.

Supplementary figure 1

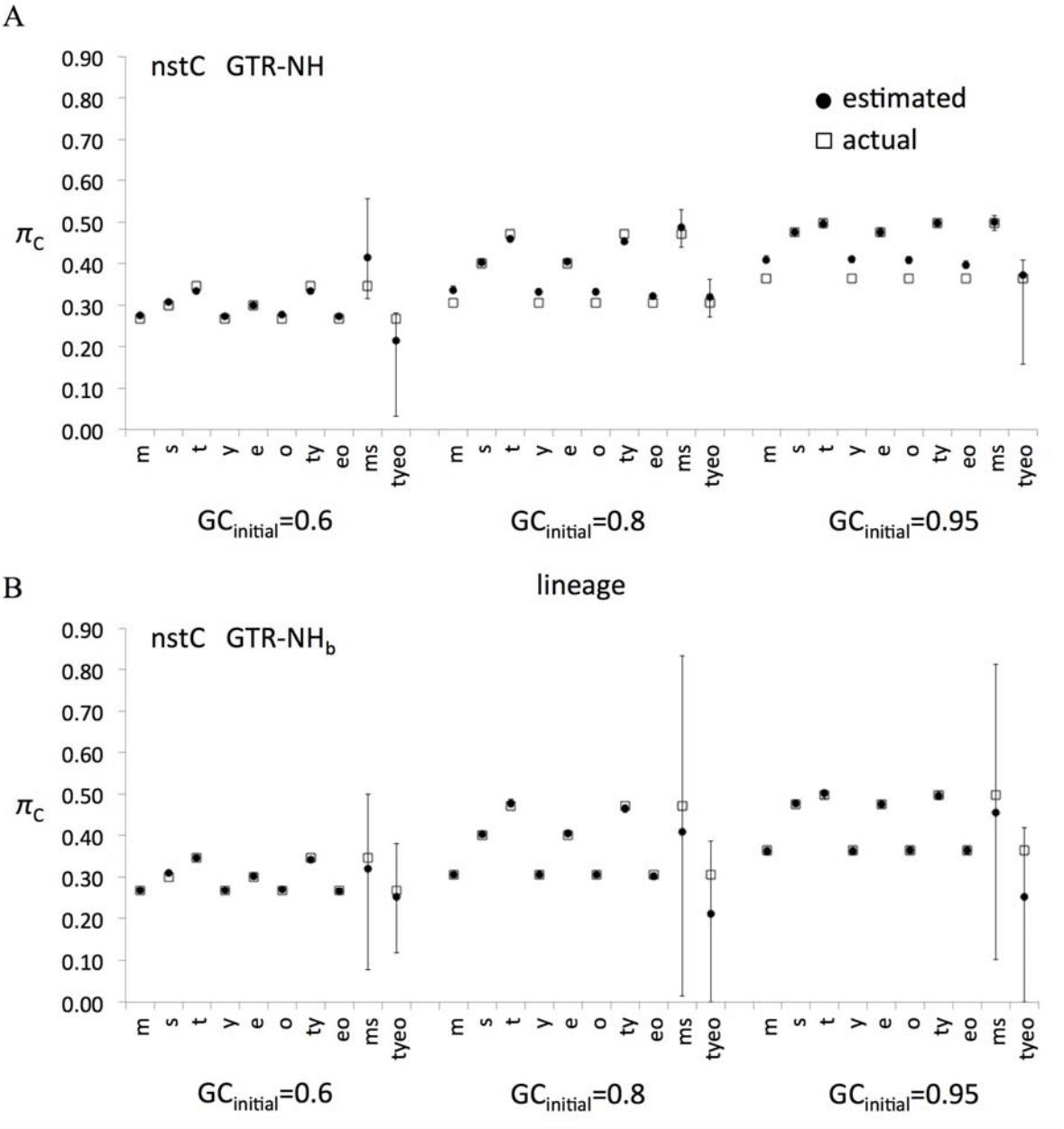

Figure S1. Actual and estimated equilibrium frequencies of nucleotide C (π_C) at nodes on the tree for selection scheme nstC under (A) the GTR-NH and (B) the GTR-NH_b models. Three initial GC biases (with $GC_{initial}$ = 0.6, 0.8 and 0.95) were considered. Shown are the averages and 95% confidence intervals over 300 bootstrap replicate datasets generated by bootstrap resampling of the 777,600 sites in the simulated dataset. Note that the results for the two branches around the root (ms and tyeo) are highly unreliable.

Supplementary figure 2

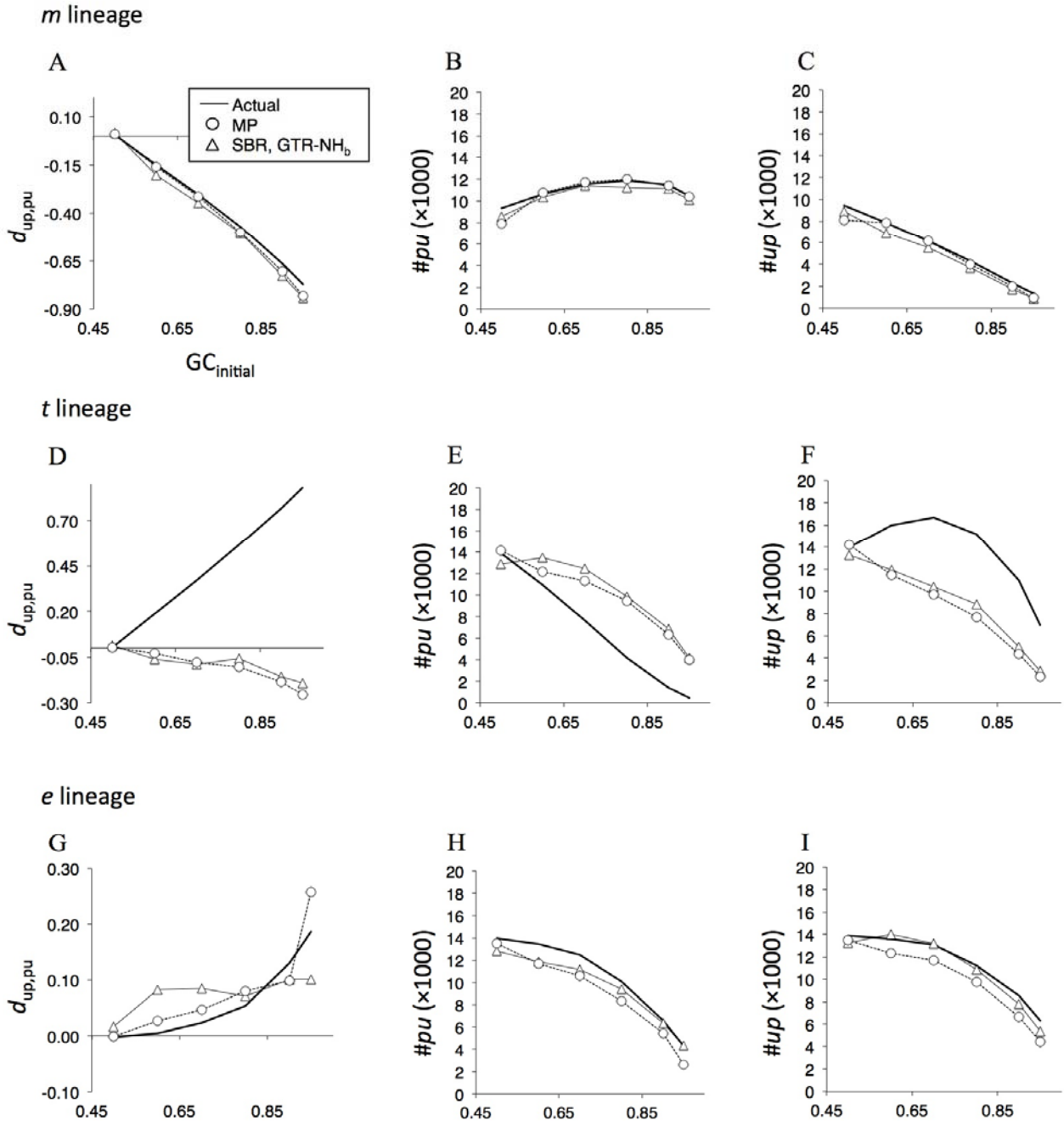

Figure S2. Inferred $d_{up,pu}$ index and substitution counts for lineages *m*, *t* and *e* under the selection scheme nstC obtained using the single best reconstruction (SBR) method using ML under the GTR-NH*b* model. Results for the MP method are shown as well.