Why Time Matters: Codon Evolution and the Temporal Dynamics of d*N*/d*S*

Carina F. Mugal,[1] Jochen B.W. Wolf,[1] and Ingemar Kaj*,[2]

[1]Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

[2]Department of Mathematics, Uppsala University, Uppsala, Sweden

***Corresponding author:** E-mail: ikaj@math.uu.se.

Associate editor: John H. McDonald

Abstract

The ratio of divergence at nonsynonymous and synonymous sites, d*N*/d*S*, is a widely used measure in evolutionary genetic studies to investigate the extent to which selection modulates gene sequence evolution. Originally tailored to codon sequences of distantly related lineages, d*N*/d*S* represents the ratio of fixed nonsynonymous to synonymous differences. The impact of ancestral and lineage-specific polymorphisms on d*N*/d*S*, which we here show to be substantial for closely related lineages, is generally neglected in estimation techniques of d*N*/d*S*. To address this issue, we formulate a codon model that is firmly anchored in population genetic theory, derive analytical expressions for the d*N*/d*S* measure by Poisson random field approximation in a Markovian framework and validate the derivations by simulations. In good agreement, simulations and analytical derivations demonstrate that d*N*/d*S* is biased by polymorphisms at short time scales and that it can take substantial time for the expected value to settle at its time limit where only fixed differences are considered. We further show that in any attempt to estimate the d*N*/d*S* ratio from empirical data the effect of the intrinsic fluctuations of a ratio of stochastic variables, can even under neutrality yield extreme values of d*N*/d*S* at short time scales or in regions of low mutation rate. Taken together, our results have significant implications for the interpretation of d*N*/d*S* estimates, the McDonald–Kreitman test and other related statistics, in particular for closely related lineages.

***Key words:* population genetics of d*N*/d*S*, codon evolution, genomic signatures of natural selection, Poisson random field approximation.**

Introduction

The extent to which selection promotes evolutionary change has long been a key question in the evolutionary sciences. Although at the phenotypic level the importance of selection is widely recognized, its role in modulating evolution at the molecular level remains debated (Nei et al. 2010). One popular indicator of selection acting on protein-coding DNA sequences is the d*N*/d*S* ratio. Because of its alleged simplicity and intuitive appeal, this measure has a strong tradition in evolutionary research, notably for the identification of genes with a history of positive selection (Nielsen 2005). In short, the d*N*/d*S* ratio quantifies the mode and strength of selection by comparing synonymous substitution rates (d*S*)—assumed to be neutral—with nonsynonymous substitution rates (d*N*), which are exposed to selection as they change the amino acid composition of a protein. Unity of the ratio is generally taken to indicate neutrality, values exceeding unity are interpreted as selection promoting change (positive selection), and values less than one are usually taken as an indication for selection suppressing protein change (purifying selection).

Originally the d*N*/d*S* ratio was developed in a phylogenetics context, and its estimation was based on codon sequences of distantly related lineages, where it is reasonable to assume that d*N*/d*S* represents the ratio of fixed nonsynonymous to synonymous differences between lineages (Miyata et al. 1980; Li et al. 1985; Nei and Gojobori 1986; Goldman and Yang 1994; Muse and Gaut 1994). The d*N*/d*S* ratio can then be approximated as a deterministic function of population size *N* and the selection coefficient *s* (Nielsen and Yang 2003; Kryazhimskiy and Plotkin 2008). However, recent empirical (Wolf et al. 2009) and theoretical work (Kryazhimskiy and Plotkin 2008; Peterson and Masel 2009) has challenged whether d*N*/d*S* appropriately reflects the outcome of selection across all relevant evolutionary time scales. As soon as we leave the realm of phylogenetics where single stereotypic genomes are compared and enter the realm of population genetics, the d*N*/d*S* ratio is no longer based on only fixed nonsynonymous versus synonymous differences. Segregating polymorphisms can substantially alter estimates of divergence and consequently estimates of d*N*/d*S* (Peterson and Masel 2009; Charlesworth 2010). As a consequence, both recently arisen lineage-specific variants as well as shared ancestral polymorphisms need to be taken into account. Kryazhimskiy and Plotkin (2008) theoretically investigated the two most extreme cases in timescale considering 1) the pure phylogenetics context, where the d*N*/d*S* ratio is based on fixed differences between distantly related lineages and 2) the pure population genetics context, where the d*N*/d*S* ratio is based on segregating polymorphisms within

© The Author 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

doi:10.1093/molbev/mst192 Advance Access publication October 14, 2013

one panmictic population of conspecific individuals. In a phylogenetics context, codon evolution is modeled as a Markov process, which indirectly assumes that fixation of a mutation in the population occurs instantaneously (Goldman and Yang 1994; Muse and Gaut 1994). In a population genetics context, simulations of sequence evolution in the presence of selection are often based on Wright–Fisher sampling (Wright 1931). Under the assumptions that 1) codons evolve independently of each other under free recombination, and 2) polymorphic codon positions are not allowed to mutate further, allowing for a maximum biallelic state, the Markov model of codon evolution can be viewed as a time limit of a Wright–Fisher population process. That is, the jump rates of the Markov model of codon evolution can be interpreted in terms of the fixation probability of a Wright–Fisher population process with selection (Nielsen and Yang 2003). However, for closely related lineages the assumption that divergence time is large enough to view the Markov model of codon evolution as a time limit of a Wright–Fisher population process will be violated. This violation gives rise to a gap between pure population genetics and phylogenetic modeling approaches, neither of which can adequately capture the evolutionary relevant, temporal dynamics of the dN/dS ratio.

To fill this gap, we anchor the dN/dS ratio firmly in population genetics theory and develop a codon substitution model that allows us to describe the continuous dynamics of dN/dS across evolutionary time starting from a single panmictic population followed by a speciation event eventually resulting in deep phylogenetic divergence (fig. 1). We derive analytical expressions for the dN/dS measure in a Poisson Random Field framework integrating the relative contributions of ancestral polymorphisms, lineage-specific polymorphisms, and fixed differences through time. Our analysis shows under which evolutionary conditions, namely population size, selection coefficient, and mutation rate, polymorphisms influence the expectation for dN/dS at any given time point after speciation. The comprehensive mathematical description of dN/dS based on a ratio of stochastic variables further allows to estimate the associated variation and reveals that for recently diverged lineages stochastic forces acting on dN/dS are not negligible. In that, we provide a null model making apparent the inherent biases in the estimation of dN/dS generating false positive inference in the study of adaptive evolution. The results do not merely affect estimation of the dN/dS ratio itself, but are of likewise importance for related statistics such as the McDonald–Kreitman test (McDonald and Kreitman 1991) or the α-estimate (Smith and Eyre-Walker 2002). We finally advocate that a combination of divergence and polymorphism data be used to estimate true dN/dS ratios and associated confidence intervals for closely related lineages, something that appears to be feasible in light of the current progress in sequencing technology.

Results

Review of the Classical Definition of dN/dS

Over long time scales, selection is generally inferred from evolutionary change between divergent lineages that arose after a distant population split or speciation event. Each lineage is then represented by one stereotypic genome sequence, where sequence comparison allows quantifying evolutionary change at orthologous positions. Here,

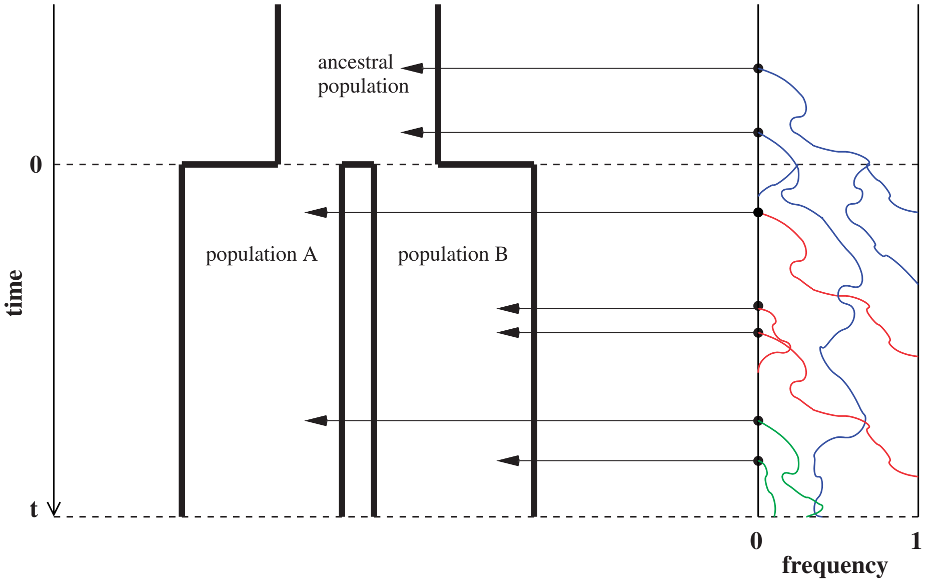

Fig. 1. Scheme of the evolutionary model—Speciation occurs instantaneously at time 0 and the two populations evolve separately and do not interbreed until present time t. Mutations are depicted by black dots in the right part of the graph, where the arrows from right to left point to the population in which the mutation happened. The right part of the graph shows the path to absorption (fixation or extinction) of these mutations. The blue lines show paths of mutations that occurred before speciation, where paths evolve separately after speciation (ancestral polymorphism). At time 0, these mutations constitute shared polymorphisms. The red lines show paths to absorption of lineage-specific mutations that occurred and got absorbed in one of the two populations between $[0, t]$ (fixed differences). The green lines show paths of lineage-specific mutations, which at present time t are still segregating in the respective population (lineage-specific polymorphism).

protein-coding sequences offer the great possibility that they allow to contrast nonsynonymous and synonymous changes, which yields intuitive insight into the mode and strength of selection. The problem of estimating the strength of selection then essentially becomes a problem of estimating substitution rates for two classes of changes. This has been addressed by two sets of methods, heuristic counting methods (Miyata et al. 1980; Nei and Gojobori 1986) and maximum likelihood based approaches (Goldman and Yang 1994; Muse and Gaut 1994), the latter modeling the substitution process as a continuous-time Markov process with 61 possible states corresponding to the 61 sense codons. Under an infinite sites model and under the assumptions of free recombination and instantaneous fixation of novel mutations, dN/dS can be recaptured based on Kimura's expression for the probability of fixation of newly arising variants of frequency $1/N$ (Sawyer and Hartl 1992). To apply Kimura's expression for the probability of fixation under a finite sites model, we have to make the additional assumption that polymorphic codon positions are not allowed to mutate further, such that there are never more than two alleles segregating at the same codon position (Nielsen and Yang 2003). It is then considered that synonymous mutations evolve in the population under neutral reproduction. Nonsynonymous mutations are influenced by selective forces in such a way that the fitness of the ancestral to the derived alleles are 1 to $1 + s$, where all nonsynonymous mutations have the same selection coefficient. Each time a derived allele becomes fixed fitness is reassigned such that the derived allele is considered as the new ancestral state and fitness is set to 1. The fitness of any potential new mutation (including back mutation) is set to $1 + s$. Under this model, selection acts as a mechanism which promotes ($s > 0$) or prevents ($s < 0$) changes of codons involving amino acid replacements, and does not represent a preference for or against specific codon types (for discussion see Nielsen and Yang [2003]). If the chance of an immediate back mutation is small enough to be neglected, then for nonsynonymous mutations the fixation probability is given by (Kimura 1962),

$$q_\gamma = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} = \frac{1 - e^{-2\gamma/N}}{1 - e^{-2\gamma}} \approx \frac{1}{N}\frac{2\gamma}{1 - e^{-2\gamma}}, \qquad (1)$$

where N is the (effective) population size, s is the selection coefficient, and $\gamma = Ns$ is the population-scaled selection coefficient. By taking $\gamma \to 0$, we recover the neutral case of a synonymous mutation, for which the probability that a novel mutation gets fixed in the population is $q_0 = 1/N$. Note that derivations of the probability of fixation are based on a haploid population of size N. Under the assumption of additive fitness effects in a diploid organism, these derivations are equivalent to a diploid population of size $N/2$. The expected numbers of nonsynonymous and synonymous substitutions per generation scale with q_γ and q_0, and hence dN/dS is interpreted as the ratio of these given by

$$\omega_\gamma = \frac{q_\gamma}{q_0} \approx \frac{2\gamma}{1 - e^{-2\gamma}}, \quad \gamma \neq 0, \qquad \omega_0 = 1. \qquad (2)$$

Here, ω_γ corresponds to the ω typically estimated from data using software packages such as PAML (Yang 2007). One objection to bear in mind is that fixation effects were assumed to be instantaneous, where sequence divergence is equal to the number of mutations that occurred and got fixed between two populations after population split. In practice, however, the sequence divergence of two populations is measured based on the number of differences observed at divergence time t between two sequences each sampled from one of the two distinct populations. Thus, in addition to mutations that occurred and got fixed after population split also shared ancestral and newly arisen lineage-specific polymorphisms will contribute to the total divergence (fig. 1).

Definition of dN/dS in a Population Genetics-Phylogenetics Framework

We will now drop the assumption of instantaneous fixation and formulate an explicit codon substitution model that allows to describe the expectation of nonsynonymous and synonymous divergence at any point in time. Following the standard approach, dN/dS is derived from amino acid sequence divergence between two divergent lineages (or populations). We make the simplifying assumption that speciation follows an isolation-without-migration model as illustrated in figure 1. We thus consider two independent populations both of size N where each element is a sequence of L codons or $3L$ nucleotide sites. Let t denote the population-scaled evolutionary divergence time between the two populations at present time, where Nt generations have passed since population divergence time at $t = 0$. Mutation events occur at a rate $\mu > 0$ per individual (nucleotide) site and generation. Whenever a nucleotide is hit by mutation, a target nucleotide is chosen according to a 4×4 Markov chain transition probability matrix **H**. Note that **H** can be viewed as any commonly used nucleotide substitution model, such as the Jukes–Cantor model. The fate in the population of this newly introduced derived type nucleotide over subsequent generations is extinction or fixation, determined by a standard Wright–Fisher reproduction mechanism, which furthermore distinguishes between nonsynonymous and synonymous changes. We assume that codons evolve independently (free recombination) and that μ is sufficiently small to allow for a scenario where each new mutation will only affect monomorphic codon sites. Hence, in this model, codon sites will be at most biallelic.

In each polymorphic site, the pair of ancestral and derived codon will be either synonymous or nonsynonymous. Mutations leading to synonymous changes evolve in the population under neutral reproduction, whereas mutations leading to nonsynonymous changes are influenced by selection such that the fitness of the ancestral to the derived alleles are 1 to $1 + s$. Following practice in much of the population genetics literature including theoretical studies of dN/dS (Sawyer and Hartl 1992), we consider the diffusion approximation scaling regime of large N and small s, where the population-scaled selection coefficient $\gamma = Ns$ reflects the total (signed) selection pressure per site and generation. Similarly, the constant $\theta = 3\mu LN$ measures total mutation pressure per codon sequence and generation. In the Materials and Methods, we show that based on the fundamental parameters **H**, γ, and θ together with the knowledge of the genetic code, it is possible in the framework of our model to derive the proportions of mutation events leading to synonymous and nonsynonymous codon pairs. Hence, we can write

$$\theta = \theta_{\text{syn}} + \theta_{\text{non}} + \theta_{\text{stop}},$$

to distinguish synonymous and nonsynonymous changes, as well as sorting out an intensity θ_{stop} for events that lead to stop codons.

Now, to obtain d*N*/d*S*, we consider the sequence divergence between two sequences sampled from two distinct populations or lineages. Sequence divergence can be split into the two contributions of nonsynonymous and synonymous divergence and their ratio can be used to quantify the impact of selection acting on the entire coding sequence. To this end, we let $D^{\text{non}}(t)$ denote nonsynonymous divergence and $D^{\text{syn}}(t)$ synonymous divergence at time *t*, and write $D(t) = D^{\text{non}}(t) + D^{\text{syn}}(t)$ for the total divergence. Sequence divergence should naturally be proportional to sequence length and increase over time essentially with the same rate as that of substitutions occurring in either population from the time of population split and onward. This is indeed a property of our model, in which $D^{\text{non}}(t)$ and $D^{\text{syn}}(t)$ are independent and have approximate Poisson distributions with expected values

$$\mathbb{E}D^{\text{non}}(t) = 2\theta_{\text{non}}\, d^{\text{non}}(t) \quad \text{and} \quad \mathbb{E}D^{\text{syn}}(t) = 2\theta_{\text{syn}}\, d^{\text{syn}}(t),$$

where d^{non} and d^{syn} are functions of γ and *t*. The factor 2 arises from the fact that we consider divergence between two populations.

As a first measure of d*N*/d*S* we take the ratio of expected values

$$\mathrm{d}N/\mathrm{d}S = \frac{\mathbb{E}D^{\text{non}}(t)/\theta_{\text{non}}}{\mathbb{E}D^{\text{syn}}(t)/\theta_{\text{syn}}} = \frac{d^{\text{non}}(t)}{d^{\text{syn}}(t)}, \tag{3}$$

where the normalization by θ_{non} and θ_{syn} accounts for the difference in mutation pressure for nonsynonymous and synonymous changes, respectively. The ratio in equation (3) represents a measure of the average rates of nonsynonymous to synonymous divergence. Similar to previous work (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008; Peterson and Masel 2009), our aim is to understand the relation between d*N*/d*S* and natural selection as a function of evolutionary time. Our contribution here is to provide more detailed expressions than reported earlier for $d^{\text{non}}(t)$ and $d^{\text{syn}}(t)$ across all relevant evolutionary timescales with special attention to small *t*. At the same time, we are cautious about the use of equation (3) as a single d*N*/d*S* measure. After all, upon accepting the underlying model assumption that divergence is the result of random sampling from random populations of random sequences, it is restrictive in the end to only compare two expected values. To initiate a discussion of alternative measures of d*N*/d*S*, perhaps more suitable to reflect the various fluctuations involved, we will compare in the next subsection the ratio of the expected values $\mathbb{E}D^{\text{non}}(t)$ and $\mathbb{E}D^{\text{syn}}(t)$ in equation (3) with the expected value of the ratio of the independent random variables $D^{\text{non}}(t)$ and $D^{\text{syn}}(t)$, see equation (7) later.

To incorporate the contribution of polymorphism and thus expand the classical definition of d*N*/d*S* we consider three levels of mathematical modeling which are described in detail in the Materials and Methods. In short, we begin with a full codon substitution model (the phylogenetics component) embedded in a population genetics framework represented by a discrete Markov chain. Because of the complexity of the model, analytical insight is limited. In a second step, we therefore resort to an analytically more tractable continuous time approximation. This allows us to find the codon equilibrium distribution and to extract the typical rates of synonymous and nonsynonymous mutations. For standard mutation models, the latter can be derived explicitly. Finally, in a third step, we argue that key properties including the rate of divergence over time are captured well by approximate Poisson distributions. This approach is reminiscent of the Poisson's random fields model, which has been used for similar purposes earlier (Sawyer and Hartl 1992). The main assumptions for the model parameters are that *N* and *L* are both large while the ratio $2\theta \log(N)/L$, which represents the fraction of polymorphic sites in the sequence, is kept sufficiently small. From this, we derive detailed results for the d*N*/d*S* ratio, in particular with regards to dependence on the selection parameter γ and divergence time *t*.

As a consequence of the Poisson approximation, we can treat nonsynonymous as well as synonymous divergence as the sum of three independent Poisson distributed components, arising from divergence due to fixation of new mutations since population divergence, lineage-specific polymorphisms, and shared ancestral polymorphisms. This basically corresponds to sampling two sequences, one sequence from each population, aligning them and counting the number of synonymous and nonsynonymous differences. We then distinguish three cases how these differences could have arrived. First, we distinguish mutations which occurred before or after population split. Second, for mutations that occurred after population split we make the further distinction whether the mutant is already fixed in its population or still segregating. The first case of mutations, which occurred before population split, are referred to as ancestral divergence (blue lines in fig. 1). Fixed differences due to mutations that occurred after population split are referred to as fixed divergence (red lines in fig. 1). Finally, mutations, which occurred after population split and are still segregating, are referred to as polymorphic divergence (green lines in fig. 1). Accordingly, the mean divergence splits into three types of contributions, and we can write

$$\begin{aligned} d^{\text{non}}(t) &= d^{\text{non}}_{\text{fix}}(t) + d^{\text{non}}_{\text{pol}}(t) + d^{\text{non}}_{\text{anc}}(t), \\ d^{\text{syn}}(t) &= d^{\text{syn}}_{\text{fix}}(t) + d^{\text{syn}}_{\text{pol}}(t) + d^{\text{syn}}_{\text{anc}}(t), \end{aligned} \tag{4}$$

where $d^{\text{non}}_{\text{fix}}(t)$, $d^{\text{syn}}_{\text{fix}}(t)$ represent divergence due to fixation of new mutations since population divergence, $d^{\text{non}}_{\text{pol}}(t)$, $d^{\text{syn}}_{\text{pol}}(t)$

are contributions from sampling of lineage-specific polymorphic sites, and $d^{\mathrm{non}}_{\mathrm{anc}}(t), d^{\mathrm{syn}}_{\mathrm{anc}}(t)$ take into account the effect of ancestral polymorphisms which existed at $t = 0$. In the Materials and Methods, we present in detail approximation formulas for all of these, using three auxiliary functions which we denote by $G_\gamma(t)$, $H_\gamma(t)$, and $J_\gamma(t)$, Briefly, $t - G_\gamma(t)$ scales with the average number of fixations up to time t and $H_\gamma(t)$ scales with the average number of lineage-specific polymorphisms that get sampled for its derived allele at time t, in both cases referring to mutations that occurred after population split. The function $J_\gamma(t)$ scales with the average number of sampled differences at t, which originate from mutations in the ancestral population.

With time-explicit derivations of all terms in equation (4) established, we may sum up the expected divergence from each contribution, be it ancestral, polymorphic, or fixed, and compare nonsynonymous and synonymous terms, writing

$$\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{total}} = \frac{d^{\mathrm{non}}(t)}{d^{\mathrm{syn}}(t)} = \omega_\gamma \frac{t - G_\gamma(t) + H_\gamma(t) + J_\gamma(t)}{t - G_0(t) + \min(t,2) + 1}. \tag{5}$$

The asymptotic limits are given by ω_γ as $t \to \infty$ and by $(\omega_\gamma - 1)/\gamma$ as $t \to 0$.

To help interpret the various contributions in equation (5), we proceed to look at the separate dN/dS ratios for each type, which we denote by $dN/dS\,|_{\mathrm{fixation}}$, $dN/dS\,|_{\mathrm{polymorphic}}$, and $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{ancestral}}$. The ratio $d^{\mathrm{non}}_{\mathrm{fix}}(t)/d^{\mathrm{syn}}_{\mathrm{fix}}(t)$ is insensitive to t and quickly converges to ω_γ with increasing t. Hence, in agreement with equation (3), the contribution to dN/dS due to fixations of lineage-specific mutations is

$$\begin{aligned}\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{fixation}} &= d^{\mathrm{non}}_{\mathrm{fix}}(t)/d^{\mathrm{syn}}_{\mathrm{fix}}(t)\\ &= \frac{\omega_\gamma(t - G_\gamma(t))}{t - G_0(t)} \to \omega_\gamma = \frac{2\gamma}{1 - \mathrm{e}^{-2\gamma}}.\end{aligned}$$

The slight deviation of $d^{\mathrm{non}}_{\mathrm{fix}}(t)/d^{\mathrm{syn}}_{\mathrm{fix}}(t)$ from ω_γ is essentially due to our definition of fixed differences, which are based on lineage-specific mutations only. In fact, fixed differences could arise due to lineage-specific mutations as well as due to shared ancestral polymorphisms. However, once the sum of the various distributions to divergence estimates is computed (as in $dN/dS\,|_{\mathrm{total}}$), both kinds of fixed differences are considered. For divergence attributed to lineage-specific polymorphisms, we find that $d^{\mathrm{non}}_{\mathrm{pol}}(t)/d^{\mathrm{syn}}_{\mathrm{pol}}(t)$, which equals 1 at $t = 0$, quickly approaches a limiting value ρ_γ, such that

$$\begin{aligned}\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{polymorphic}} &= d^{\mathrm{non}}_{\mathrm{pol}}(t)/d^{\mathrm{syn}}_{\mathrm{pol}}(t)\\ &= \frac{\omega_\gamma H_\gamma(t)}{\min(t,2)} \to \rho_\gamma \approx \frac{\mathrm{e}^{-2\gamma} - 1 + 2\gamma + 2\gamma^2}{2\gamma(1 - \mathrm{e}^{-2\gamma})}, \quad t \to \infty.\end{aligned}$$

Finally, the ancestral contribution as time evolves has a limiting ratio ϕ_γ such that

$$\begin{aligned}\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{ancestral}} &= d^{\mathrm{non}}_{\mathrm{anc}}(t)/d^{\mathrm{syn}}_{\mathrm{anc}}(t)\\ &= \omega_\gamma J_\gamma(t) \to \phi_\gamma \approx \omega_\gamma\Big(1 - \frac{\gamma}{3} - \frac{\gamma^2}{18}\Big) \approx 1 + \frac{2\gamma}{3} - \frac{\gamma^2}{18}.\end{aligned}$$

At $t = 0$, the ancestral ratio takes into account the effect of selection when we sample two individuals from the single population, which forms the common ancestry at population split. In equation (26) of the Materials and Methods, we show that $J_\gamma(0) = (\omega_\gamma - 1)/\gamma\omega_\gamma$, so that we have the initial ratio

$$d^{\mathrm{non}}_{\mathrm{anc}}(0)/d^{\mathrm{syn}}_{\mathrm{anc}}(0) = \frac{\omega_\gamma - 1}{\gamma} \approx 1 + \frac{\gamma}{3} - \frac{\gamma^3}{45}. \tag{6}$$

The time dependence of the various ratios is illustrated in the top panel of figure 2 for $\gamma = -1$. The three separate ratios $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{fixation}}$, $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{polymorphic}}$, $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{ancestral}}$ are plotted together with the ratio of total expectations $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{total}}$, as well as the limiting value ω_γ. In the bottom panel of figure 2, the limiting long time ratios ω_γ, ρ_γ, and ϕ_γ as well as the initial ancestral contribution $(\omega_\gamma - 1)/\gamma$ are plotted together as functions of γ.

The interesting observation is that when combining the effects of all three contributions by forming $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{total}}$, the ancestral and polymorphic divergence influence the total divergence ratio over a substantial time period before settling down at ω_γ. With an increasing number of fixation events and hence actual lineage-specific nucleotide substitutions building up differences between the two populations over a considerable time span, it is of course the rate of linear increase in $d^{\mathrm{non}}_{\mathrm{fix}}(t)$ in comparison with that of $d^{\mathrm{syn}}_{\mathrm{fix}}(t)$, which will ultimately decide the asymptotic dN/dS ratio. But as evident in figure 2, ancestral and lineage-specific polymorphisms which also generate differences between the observed sequences seek out their own preferred balance of nonsynonymous to synonymous change. As long as ancestral and polymorphic differences measure up on the scale of fixations, the limiting numbers ρ_γ and ϕ_γ influence the total ratio. The ancestral initial value, which is manifestly different from ω_γ, ensures that the transition to fixation asymptotics is clearly visible on the evolutionary time scale. In summary, this clearly indicates that, indeed, dN/dS is naturally a function of time. For the case $\gamma < 0$ of negative selection, dN/dS decreases from its initial ratio $(\omega_\gamma - 1)/\gamma < 1$ to $\omega_\gamma < 1$. If $\gamma > 0$, then dN/dS increases from its initial ratio $(\omega_\gamma - 1)/\gamma > 1$ to $\omega_\gamma > 1$. Figure 3 illustrates the time dependence of $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{total}}$ for five values of γ.

We conclude this section with additional remarks on the relation of our results to previous results in the literature. The work of Kryazhimskiy and Plotkin (2008) is concerned with the relationship between selection and dN/dS values measured from two sequences sampled from a single population. In this situation, differences between the sequences reflect segregating polymorphisms and not fixed differences. The authors offer a theoretical foundation of the dN/dS ratio for single population data and demonstrate that the frequent use of equation (2) in the context of intraspecific sequence data lacks proper justification and is inappropriate. The desired dN/dS ratio addressed in (Kryazhimskiy and Plotkin 2008) is closely related to $\mathrm{d}N/\mathrm{d}S\,|_{\mathrm{ancestral}}$ at $t = 0$ in our model, that is, the dN/dS ratio based on two sequences sampled from a single (ancestral) population existing prior to speciation.

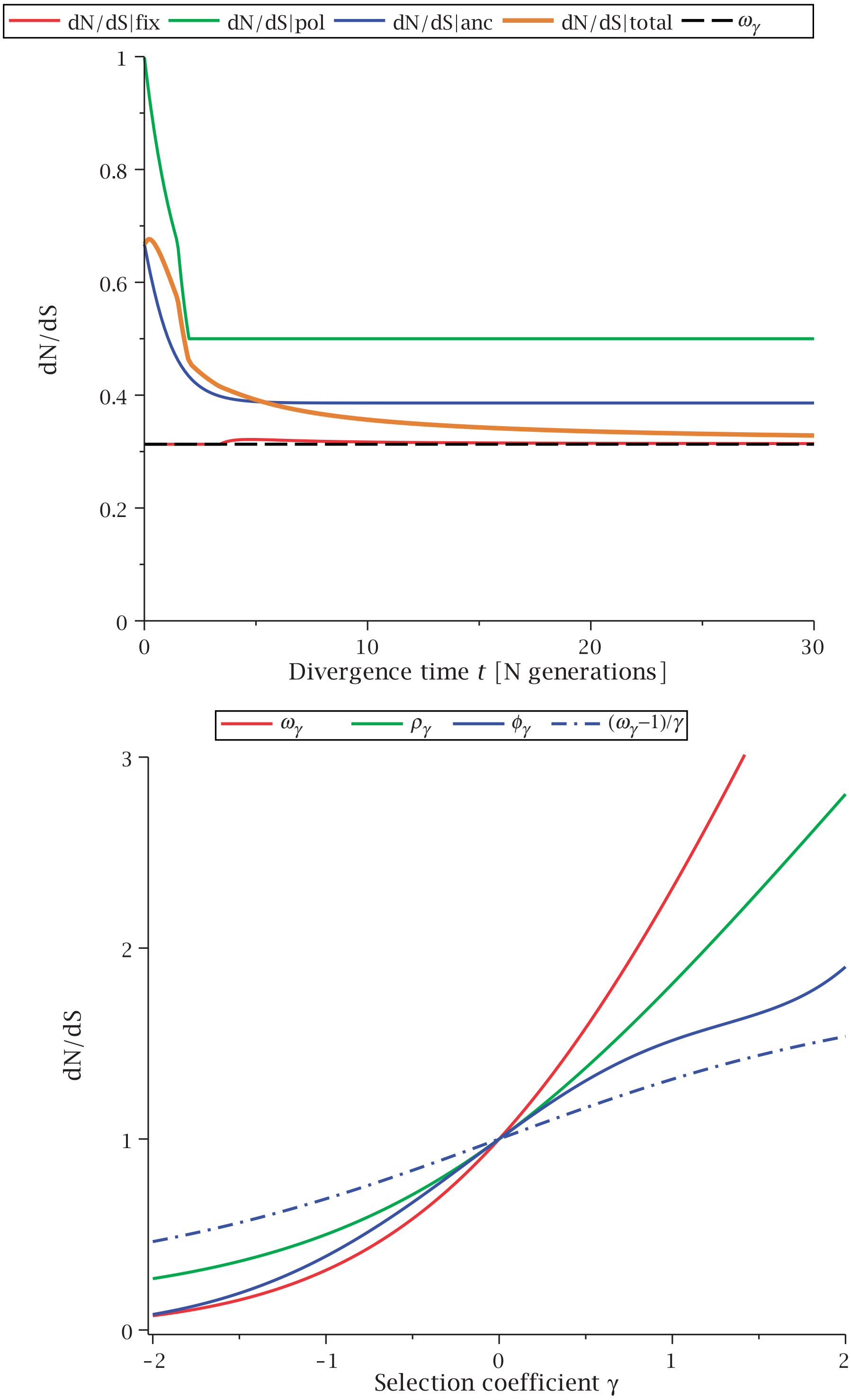

Fig. 2. (Top) dN/dS ratios for fixation (red), polymorphic (green), ancestral (blue), and total (gold line) divergence for $\gamma = -1$ as a function of divergence time t; (bottom) the limiting values ω_γ for dN/dS $|_{\text{fixation}}$ (red), ρ_γ for dN/dS $|_{\text{polymorphic}}$ (green), ϕ_γ for dN/dS $|_{\text{ancestral}}$ (solid blue), and the initial ancestral contribution $(\omega_\gamma - 1)/\gamma$ (dashed blue) as a function of γ.

In fact, the initial ratio found to be $(\omega_\gamma - 1)/\gamma$ in equation (6) is a stationary dN/dS ratio for sequences sampled in a single population. However, the corresponding quantity $\omega_{\text{pop}}(\gamma,\theta)$ obtained in (Kryazhimskiy and Plotkin 2008), equation (5), depends not only on γ but also on θ. This fact can be traced back to the derivation of $\omega(\gamma,\theta)_{\text{pop}}$, which is based on a Wright–Fisher approximation that allows for back and forth mutations during the segregating phase, further assuming that θ is sufficiently small. It is then natural to interpret the $\theta \rightarrow 0$ limit of $\omega(\gamma,\theta)_{\text{pop}}$ as a generic dN/dS ratio within populations, and straightforward to check that $\omega(\gamma,0)_{\text{pop}} = (\omega_\gamma - 1)/\gamma$ in complete agreement with equation (6).

The work by Sawyer and Hartl (1992) form the theoretical basis for the McDonald–Kreitman test (McDonald and

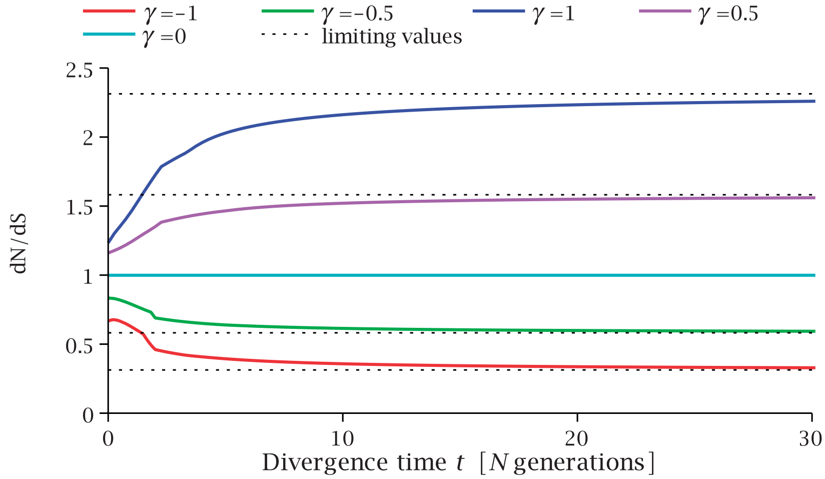

Fig. 3. Total dN/dS-ratios for the five values $\gamma = -1$ (red line), $\gamma = -0.5$ (green line), $\gamma = 0$ (turquoise line), $\gamma = 0.5$ (purple line), and $\gamma = 1$ (blue line).

Kreitman 1991). These authors provide sampling formulas for fixed nonsynonymous and synonymous differences and for nonsynonymous and synonymous polymorphisms in a model setting, which is rather close to the model advocated in the present work. Regarding mutations, our rates θ^{syn}, θ^{non}, which are derived from the codon substitution model, can be considered equivalent to the mutation rate parameters μ_s and μ_r used by Sawyer and Hartl. Turning to the expected number of fixed differences Sawyer and Hartl assume linearity in time, whereas we get refined expressions for the number of fixed differences as we distinguish between fixed differences originated from shared ancestral polymorphisms and fixed differences originated from lineage-specific mutations. Our expressions are $d^{\text{syn}}_{\text{fix}}(t) + d^{\text{syn}}_{\text{anc}}(t) = t - G_0(t) + 1$ and $d^{\text{non}}_{\text{fix}}(t) + d^{\text{non}}_{\text{anc}}(t) = \omega_\gamma(t - G_\gamma(t) + J_\gamma(t))$ compared with t and $\omega_\gamma t$ in (Sawyer and Hartl 1992), equation (13). However, the assumption of linearity in time is well justified for distantly related lineages and critical only for closely related lineages, where ancestral fixed differences measure up on the scale of lineage-specific fixed differences. Turning to the sampling formulas for nonsynonymous and synonymous polymorphisms Sawyer and Hartl consider arbitrary samples of size n and m from two species. In our settings, $m = n = 1$ as typical in phylogenetic approaches. Their results (Sawyer and Hartl 1992), equations (17) and (18), with $m = 1$ correspond to

$$d^{\text{syn}}_{\text{pol}}(t) = 2, \quad d^{\text{non}}_{\text{pol}}(t) = \omega_\gamma \int_0^1 \frac{1 - e^{-2\gamma x}}{\gamma x} dx$$

where we derive the time dependent functions min(t,2) and $\omega_\gamma H_\gamma(t)$, see equations (21) and (22). The differences between our results and the results by Sawyer and Hartl arise from the fact that Sawyer and Hartl assume that the number of lineage-specific segregating sites has reached its equilibrium. Although this assumption is well justified for distantly related lineages, it is clearly violated for more closely related lineages. Our results better capture the reality of divergence between evolutionary young lineages, and converge to the results by Sawyer and Hartl for $t \to \infty$.

Statistical Properties of dN/dS

The previous section has treated the time dependence of the dN/dS ratio, measured as the ratio of expected values of two independent Poisson random variables. However, when the ratio of nonsynonymous to synonymous divergence is estimated from sequence data, we in fact do not know their expected values but rather carry out single observations of $D^{\text{non}}(t)$ and $D^{\text{syn}}(t)$ and consider the ratio of these. Regardless of the estimation procedure, for example, counting methods or maximum likelihood approaches, an estimation based on sequence data will always just reflect a single observation or measurement. This is important to notice, as the ratio of expected values is in general not equal to the expected value of a ratio. We are therefore interested in the statistical properties of the expected value of a ratio of two Poisson random variables rather than in a ratio of expected values. Proper statistical inference therefore must take into account the natural fluctuations of such a ratio of random variables. This leads us to studying the ratio of nonsynonymous to synonymous divergence in a population genetics framework as the scaled ratio of Poisson variables

$$\frac{D^{\text{non}}(t)/\theta_{\text{non}}}{D^{\text{syn}}(t)/\theta_{\text{syn}}} \quad \text{on} \quad D^{\text{syn}}(t) > 0.$$

Thus, we define a new measure of dN/dS as the conditional expectation

$$\begin{aligned} dN/dS(t) &= \mathbb{E}\left(\frac{D^{\text{non}}(t)/\theta_{\text{non}}}{D^{\text{syn}}(t)/\theta_{\text{syn}}} \mid D^{\text{syn}} > 0\right) \\ &= \frac{d^{\text{non}}(t)}{d^{\text{syn}}(t)} C(2\theta_{\text{syn}} d^{\text{syn}}(t)), \end{aligned} \tag{7}$$

which is based on a series approximation of the expected value of a ratio of two random variables. In the Materials and Methods, we introduce function C (eq. 28) and provide a detailed derivation of equation (7) and show that this function can be easily computed numerically. Note, however, that unlike the ratio $\mathrm{d}N/\mathrm{d}S\,|_{\text{total}}$, $\mathrm{d}N/\mathrm{d}S(t)$ depends on the mutation pressure θ. In the limit $\theta \to \infty$ (which could be reached by an infinitely long codon sequence $L \to \infty$), we recover the previous ratio of expected values, as $\lim_{\theta\to\infty} \mathrm{d}N/\mathrm{d}S(t) = \mathrm{d}N/\mathrm{d}S\,|_{\text{total}}$. To visualize the difference between the two measures, the upper panel of figure 4 illustrates the general shape of the curve $\mathrm{d}N/\mathrm{d}S(t)$ up to time $t = 100$ after population split for the case $\gamma = -1$ with the Jukes–Cantor mutation model and four different values of θ. Also shown in the same graph is $\mathrm{d}N/\mathrm{d}S\,|_{\text{total}}$, that is, the ratio of expectations $d^{\text{non}}(t)/d^{\text{syn}}(t)$, and the limit $\omega_\gamma \approx 0.313$ as $t \to \infty$. The distinct change in appearance of the dN/dS curves with varying values of θ is somewhat similar to what the effect would be of changing the time scale from t to θt. Of course this comes natural since lowering the overall mutation pressure in the model would cause the system to run on a slower time scale. It is important to note that the striking deviation in figure 4 of the expected ratio $\mathrm{d}N/\mathrm{d}S(t)$ from the ratio of expectations $d^{\text{non}}(t)/d^{\text{syn}}(t)$, is not directly an effect of the selection mechanism. On the contrary, the lower panel of figure 4 shows the corresponding set of curves for the neutral case $\gamma = 0$ for which we have $d^{\text{non}}(t) = d^{\text{syn}}(t)$.

To provide insight into the shape of the dN/dS curves and their intrinsic random variations, we further estimate upper and lower confidence bands $I_+(\alpha)$ and $I_-(\alpha)$ for dN/dS defined as a ratio of two Poisson random variables, such that

$$\mathbb{P}\left(I_-(\alpha/2) \le \frac{D^{\text{non}}(t)/\theta_{\text{non}}}{D^{\text{syn}}(t)/\theta_{\text{syn}}} \le I_+(\alpha/2)\right) \approx 1-\alpha.$$

In the Materials and Methods, we obtain

$$\begin{aligned} I_-(\alpha/2) &= \left(\left(p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{\hat{n}}}\right)^{-1} - 1\right)\frac{\pi^{\text{syn}}}{\pi^{\text{non}}}, \\ I_+(\alpha/2) &= \left(\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{\hat{n}}}\right)^{-1} - 1\right)\frac{\pi^{\text{syn}}}{\pi^{\text{non}}}, \end{aligned} \tag{8}$$

where z_α is the α-quantile of the standard normal distribution, π^{non} and π^{syn} represent the proportion of nonsynonymous and synonymous changes, respectively, and

$$p = \frac{\pi^{\text{syn}} d^{\text{syn}}(t)}{\hat{n}}, \qquad \hat{n} = \pi^{\text{syn}} d^{\text{syn}}(t) + \pi^{\text{non}} d^{\text{non}}(t).$$

To cross-validate these derivations, we next explore the variation in dN/dS by 100 independent simulation runs with parameter settings $N = 500$, $L = 2000$, $\gamma = -1$, and $\theta = 1$. We keep track of the precise numbers of nonsynonymous and synonymous differences between two populations from population split at $t = 0$ up to evolutionary time $t = 100$, and then plot the ratio of these differences at a

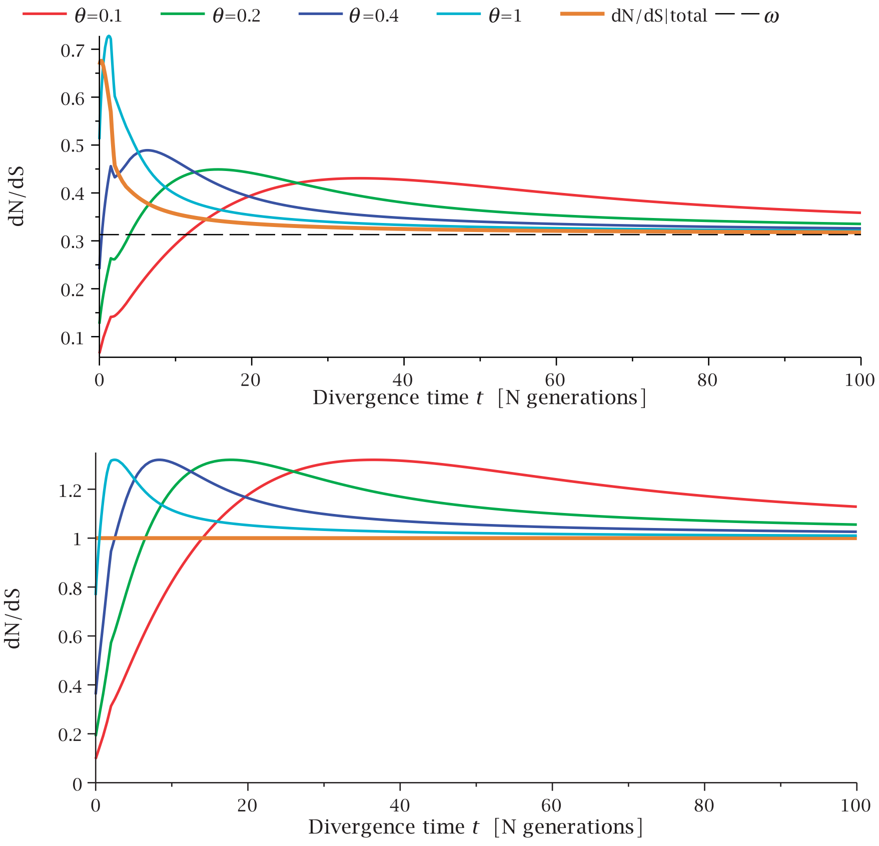

FIG. 4. $\mathrm{d}N/\mathrm{d}S(t)$ for θ equal to 0.1 (red), 0.2 (green), 0.4 (blue), 1 (cyan) compared with the ratio $\mathrm{d}N/\mathrm{d}S\,|_{\text{total}}$ (gold). (Top) $\mathrm{d}N/\mathrm{d}S(t)$ for $\gamma = -1$ and (bottom) $\mathrm{d}N/\mathrm{d}S(t)$ for $\gamma = 0$. The limiting value ω_γ is indicated by the dashed black line, which in the lower panel is identical to unity and hidden by the golden line.

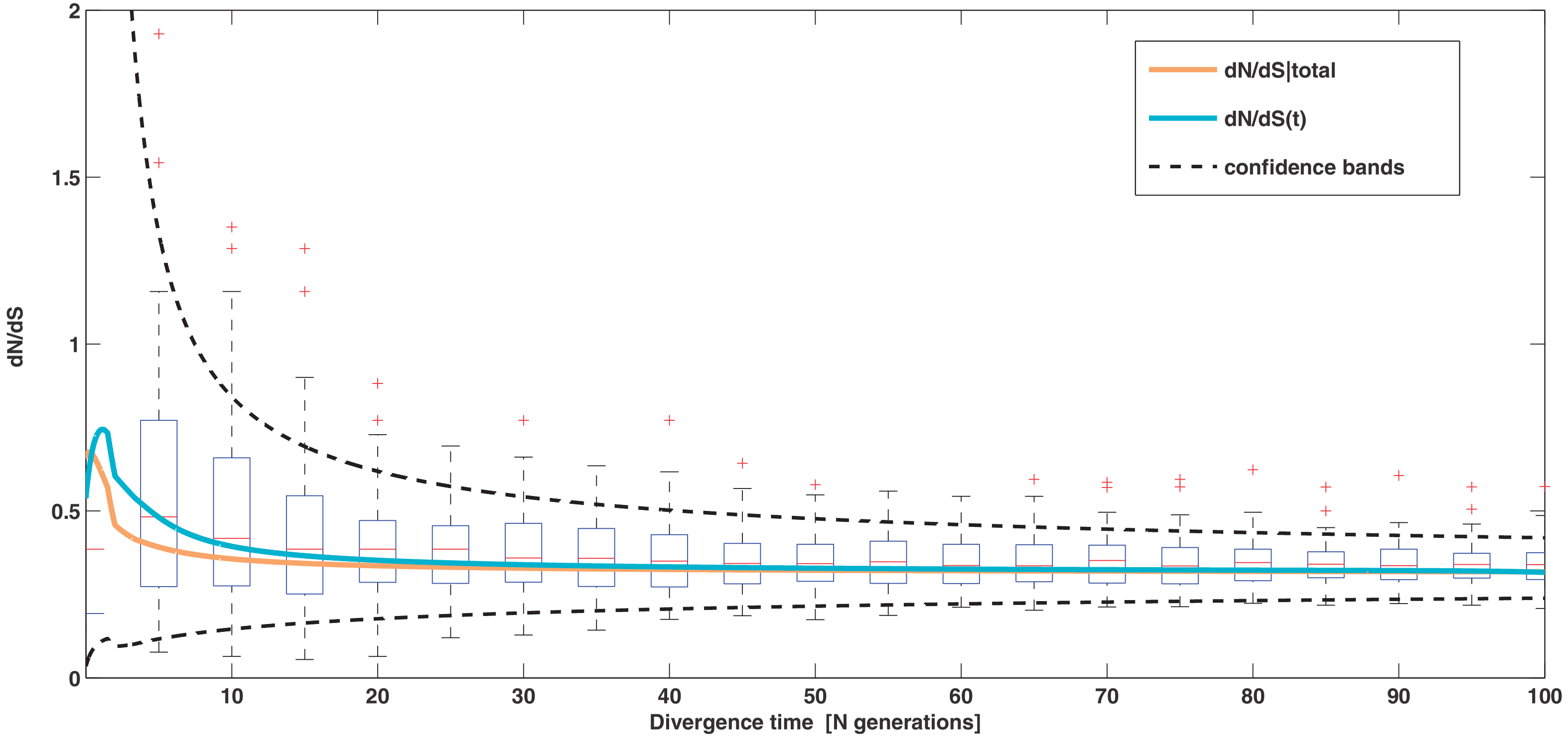

Fig. 5. Joint simulated dN/dS for two populations with $N = 500$ and $L = 2{,}000$ shown from population split at $t = 0$ up to evolutionary time $t = 100$, with an accumulated ancestral population since $t = -20$. Selection and mutation pressure are set to $\gamma = -1$ and $\theta = 1$. Also shown in the graph are $dN/dS\,|_{\text{total}}$ (golden), the dN/dS-ratio in (7) (cyan) and confidence bands given by (8) (black dashed lines).

resolution of every 5N generations. The distribution and stochastic fluctuations in $D^{\text{non}}(t)/D^{\text{syn}}(t)$ are visualized in figure 5. Also shown in figure 5 are dN/dS$|_{\text{total}}$, the dN/dS-ratio in equation (7) as well as the shape of the confidence bands (eq. 8). The confidence bands represent a 5% chance of seeing larger fluctuations at a specific point in time and do not provide joint confidence for the entire function over time. For the present simulation of 100 runs, 14.3% of the data points fell outside of the region between upper and lower bands, where 9.7% fell above the upper band and 4.6% below the lower band. Moreover, note that during initial divergence extremely high dN/dS values that would be commonly taken as evidence for positive selection are frequently obtained even under negative selection pressure.

Discussion

Value and Limitations of the Model

The standard phylogenetic model of codon evolution and the estimation of dN/dS was introduced in a pure phylogenetics context in 1994 by two independent publications (Goldman and Yang 1994; Muse and Gaut 1994). The observations that dN/dS can be influenced by mutation rate (Wyckoff et al. 2005), branch length (Wolf et al. 2009), and polymorphisms motivated theoretical studies on the temporal dynamics of the measure. We here pick three previous studies of relevance and briefly discuss them in relation to our current approach. The first study by Rocha et al. (2006) describes the time dependence of dN/dS for closely related taxa, starting with a clonal population which over time becomes more diverse. As in our modeling approach their simulation study applies Wright–Fisher sampling in a population of fixed size where each generation is subject to mutation. However, instead of incorporating a full codon model each mutation is simply set to be synonymous with probability 1/4 and nonsynonymous with probability 3/4. The number of synonymous mutations is assumed to increase linearly in time, while nonsynonymous mutations are sampled with a selective weight. Importantly, the number of accumulated nonsynonymous mutations is assumed to reach a limiting value over time which means that possible fixations are not taken into account. Although this may not be critical for short time periods, for long timescales this leads to the inappropriate property that dN/dS $\rightarrow$ 0 as $t \rightarrow \infty$. Another model developed for similar purposes by Peterson and Masel (2009), is refined in several ways. As in our study, Peterson and Masel consider divergence between two populations after population split from a common ancestor and derive estimates of the expected divergence as function of divergence time. They include the effects of recent fixations and shared ancestral polymorphisms, but neglect the effect of lineage-specific polymorphisms. Their study was motivated by earlier studies on the effect of ancestral polymorphisms on estimates of mutation rate for closely related lineages, related to the apparent mutation rate acceleration (Ho and Larson 2006; Balbi and Feil 2007). A third study closely related to ours is the study by Kryazhimskiy and Plotkin (2008). Their emphasis is on the comparison of two extreme cases, where dN/dS is estimated from sequences of 1) conspecific individuals and 2) distantly related lineages. We expand on their approach as we study the continuous transition between these two cases. Ideally, our description of dN/dS would show the same initial value as the one described by Kryazhimskiy and Plotkin for conspecific individuals. At first glance, this is not the case. Kryazhimskiy and Plotkin use a mutation model that allows for back and forth mutation between the ancestral and the derived allele during the time of segregation, which in the setting of codon evolution seems to be inappropriate and yields different results. However, if we no longer allow for back and forth mutation by letting $\theta \rightarrow 0$ in Kryazhimskiy and Plotkin, their dN/dS measure converges to our measure based on a single population prior to speciation. Hence our analysis of dN/dS with regards to the single population prior to speciation is consistent with that of Kryazhimskiy and Plotkin for conspecific individuals not allowing for back and forth mutations. The second extreme case investigated by Kryazhimskiy and Plotkin also reflecting the classical definition of dN/dS as introduced in the pure phylogenetics context (Goldman and Yang 1994; Muse and Gaut 1994) is in full agreement with our description of the limiting value of dN/dS for $t \rightarrow \infty$.

A further novelty of our work in comparison with Rocha et al. (2006), Peterson and Masel (2009), Kryazhimskiy and Plotkin (2008), or any other study investigating the temporal dynamics of dN/dS, is that we incorporate a full codon substitution matrix into our model, and consider selection for or against changes in the codon sequence. This is in contrast to the other works where estimates of dN/dS are based on a comparison of sites evolving under selective pressure versus neutrally evolving sites. The slightly more complicated, population genetic Markov model of codon sequence evolution seems a natural choice as it closely mimics biological reality. Moreover, our model allows to capture the dynamics of dN/dS at any point in time and expands its inferential value beyond mere phylogenetic considerations. In addition, the incorporation of a nucleotide substitution matrix and the resulting codon substitution matrix in a Markovian framework, should make it possible to specifically consider processes such as GC-biased gene conversion that are known to mimic the signature of selection (Berglund et al. 2009). Several other expansions of our model are conceivable. For its basic formulation, we restricted our model to instantaneous speciation not allowing for the occurrence of gene flow during the onset of divergence. We expect that under such an isolation-with-migration scenario the bias introduced by polymorphisms will extend for even longer times and would certainly be worth exploring. Besides, other less stringent model assumptions such as site-specific variation in selection strength or selection on synonymous changes via codon usage bias might be relevant to consider.

Implications for Empirical Evolutionary Genetics Studies

The dN/dS measure is commonly used to 1) disclose evolutionary processes across species (Wright and Andolfatto 2008;

Ellegren 2009) or 2) to identify genes under positive selection for an evolutionary lineage of interest (Clark et al. 2003; Bustamante et al. 2005). We here demonstrated that dN/dS is biased for the comparison of evolutionary young lineages when using the standard (phylogenetic) model. Is this time dependence of dN/dS at all relevant for the paramater space empirical work is generally dealing with? Let us first consider the former case, where genome-wide mean dN/dS is used as a proxy for average selection pressure in specific lineages that are then related to life history traits of remnant species (Nikolaev et al. 2007; Wright and Andolfatto 2008). According to our model, we expect a clear upward bias of dN/dS estimates for short branches, as has been indicated by empirical evidence (Wolf et al. 2009). As branch length and life history traits such as body size or generation time are known to covary (Gillooly et al. 2005; Bromham 2009), this artifact may lead to erroneous conclusions. But how closely do lineages need to be related for this to be of concern? Let us consider the case of human–chimp divergence as an example. For a realistic value of $\gamma = -1$ and considering a large enough sequence length L that dN/dS can be approximated by the ratio of expected values, our results suggest that dN/dS is on average upward biased by approximately 46% 2 N_e generations after speciation, and still by approximately 14% 10 N_e generations after the split. Assuming 5 million years for the split time between human and chimp from a common ancestor, an overall generation time of 20 years for the human lineage and a minimum effective population size of 14,000, we obtain an estimated time to the most common recent ancestor of approximately 18 N_e generations. At first sight, this suggests only a mild contribution of polymorphisms to dN/dS of the human lineage. Eighteen N_e generations, however, are an overestimate for two reasons. First, ancestral population sizes have been larger than current human effective population size. Assuming an average effective population size of 45,000 (Prüfer et al. 2012) split time would be 5 N_e rather than 18 N_e generations, which falls squarely within the critical range of an upward biased dN/dS. Second, our model does not allow for migration after speciation, which will extend the influence of polymorphisms over longer time frames. These considerations are qualitatively consistent with evidence from Prüfer et al. (2012) suggesting that approximately 3% of genetic variation in the human genome are cases of incomplete lineage sorting with respect to bonobo or chimp. We thus conclude that for human–chimp and lineages with similar or even shorter divergence histories, polymorphisms are an issue and need to be considered for correct inference of selection pressure. With some knowledge on divergence time and effective population sizes, our model can in principle be used to rescale dN/dS accordingly and correct for the bias.

The second, more prominent application of dN/dS is the quest for genes under positive selection in specific lineages. Naturally, much effort has been devoted to isolate the genes (or gene classes) under adaptive selection in the human lineage (Clark et al. 2003; Bustamante et al. 2005). Within the context of our model, we can only discuss potential implications for approaches inferring selection for genes, and do not consider possible time dependencies of models inferring selection for single codon sites. Positive selection on genes or functional subsets of genes is generally inferred by comparing the likelihood of dN/dS being larger than in a neutral or nearly neutral scenario (Nielsen and Yang 1998) making use of software applications such as PAML that are based on the continuous Markov process with instantaneous fixation described earlier. These likelihood-based approaches used for inference on selection do not incorporate the contribution of ancestral or lineage-specific polymorphism and we may expect increased false positive detection for evolutionary young lineages, and, in particular, for genes where polymorphic sites substantially contribute to divergence. Judging from our results, we may predict which genes will be most severely affected. Looking at the per-gene level, we cannot any longer assume sequence length L to be large enough that dN/dS can be approximated by the ratio of expected values. Instead, we have to look at the statistical properties of the ratio of two Poisson random variables. Here, our results suggest that estimates of dN/dS will in particular be biased by polymorphisms if 1) the mutation pressure is low or 2) sequence length is short. Moreover, not only the expected value of dN/dS tends to be biased for such genes but also the random error or the intrinsic fluctuations in the estimate are particularly strong for the same set of genes, as indicated by wide confidence bands at shorter time scales. As a consequence the interpretation of dN/dS needs caution, and likelihood ratio tests are necessary to account for the random error. However, likelihood ratio tests can only account for the random error, but not for the systematic bias in the expected value caused by polymorphisms. This bias is expected to be strongest in genes with low divergence, which can either be due to recent divergence time, low mutation pressure, or short sequence length. Hence, the systematic bias caused by polymorphisms may at least partly explain the common observation that genes with low divergence are preferably found to be under positive selection, as has been indicated previously (Wolf et al. 2009).

Future Perspectives

We have here introduced an analytical model to illustrate the time dependence of dN/dS and aspects of the effects of estimating dN/dS as a ratio of two Poisson random variables. Our approach expands existing models on codon evolution and integrates the contribution of polymorphisms to amino acid sequence divergence. Although not explicitly formulated for this purpose, we hope that our model may provide the basis for a refinement of the underlying theory of the widely used McDonald–Kreitman test and might improve the inference on the mode and strength of selection for closely related lineages by jointly using polymorphism and divergence data. The 1000 human genome project (1000 Genomes Project Consortium 2012) and emerging population genomic studies in genetic nonmodel organisms (Ellegren et al. 2012) demonstrate that the necessary population genomic data sets will soon be readily available for a growing number of species.

Materials and Methods

A Stochastic Model of Codon Evolution

In this section, we introduce a detailed population genetics Markov model of codon sequence evolution and use it for two main purposes. First, we find natural equilibrium rates of synonymous and nonsynonymous mutations and show how to obtain them from standard assumptions of nucleotide mutation. These rates provide a reference for the volume fractions of the two types of mutations among the polymorphic sites and represent in the model an estimate of the number of synonymous and nonsynonymous sites found in data. Our second main purpose of introducing the model is to keep a sufficiently detailed record of all polymorphic sites over time to later help analyzing the rate of divergence between two populations over time. It is crucial to distinguish the contributions to sequence divergence attributed to ancestral, polymorphic, and fixed differences. This is what will enable us to count synonymous and nonsynonymous divergence taking into account all three of these mechanisms and in the end to estimate d*N*/d*S*.

A single population consists of N individuals. Each individual is represented by a sequence of nucleotide sites of length $3L$ structured as L consecutive codon nucleotide triplets. Random mutation based on standard assumptions acts on each nucleotide in a triple and the genetic code allows us to distinguish synonymous and nonsynonymous mutations. The fate of a mutant allele is extinction or fixation determined by Wright–Fisher reproduction acting independently on the L sites with the 64 codon states at each site. Although new alleles which originate from synonymous codon transitions evolve under neutral conditions of population reproduction, the evolution of mutant codon alleles that are nonsynonymous with respect to the ancestral codon are affected by selective sampling. The chance to see two or more mutations at the same site overlap in time will be so small that for our purposes is justified to study the approximative biallelic model.

In the following, we will provide a detailed account of the assumptions for the codon mutation model and for reproduction with selection weights. This level of detail is necessary to introduce the appropriate notation and prepare for the analytical description of d*N*/d*S* through time.

A Markov Model for Codon Mutations

We begin by fixing the numbering of nucleotides $A = 1$, $C = 2$, $G = 3$, $T = 4$ and an ordered list $S' = \{u_1, \ldots, u_{64}\} = \{111, 112, 113, 114, 121, 122, \ldots, 443, 444\}$, which gives an enumeration of the 64 codon types. Here, $S_0 = \{411, 413, 431\}$ is the subset of stop codons. We write S for the remaining elements, the sense codons, so that $S' = S \cup S_0$. By applying the biological code, we associate to each sense codon $u \in S$ one of the 20 existing amino acids. The change of a nucleotide affects the first, second, or third position of the corresponding codon and causes a transition from the original codon to one of eight or nine possible target codons. If codon u changes to codon v in this manner the mutation is said to be synonymous if u and v are coding for the same amino acid and nonsynonymous if the amino acids are different. To record this information, we introduce for each pair of sense codons $u,v \in S$, $u \neq v$, the indicator variables

$$J(u,v) = \begin{cases} 0 & \text{if } u \text{ and } v \text{ are synonymous,} \\ 1 & \text{if } u \text{ and } v \text{ are nonsynonymous.} \end{cases}$$

Mutations involving stop codons will happen with positive probability but will be regarded immediately extinct.

We assume that mutation occurs uniformly and independently over nucleotide sites with mutation rate $\mu > 0$ per site and per generation. Writing θ for the total mutation rate per generation, we have $\theta = 3LN\mu$. A codon site is said to be clonal when all individuals share the same nucleotide triplet and is said to be polymorphic if not. For the type of model studied here, typically the number of polymorphic sites will be small in comparison with the length L and hence the number of polymorphic sites with more than two alleles will be even smaller. Applying the criteria that N is not too large in relation to L, see supplementary text equation (14) (Supplementary Material online), mutation is assumed to be suppressed in already polymorphic sites. Hence, all polymorphic sites are biallelic in the sense that one ancestral and one derived codon coexist with frequencies summing to one. Mutation is reactivated at extinction or fixation of the derived codon.

To find the rates of synonymous and nonsynonymous mutation events, we introduce a Markov chain of codon mutations. At the level of nucleotides, given that a mutation occurs at a site in one sequence of the population the nucleotide changes from i to j, $i, j \in \{1,2,3,4\} = \{A,C,G,T\}$, according to a transition probability matrix $\mathbf{H} = (h_{ij})$ with zero diagonal elements, strictly positive nondiagonal elements and row sums equal to one. With probability one-third the affected site is the first, second, or third position of a codon. Thus, taking into account only one-site mutations, the nucleotide transitions in $\mathbf{H}$ generate a corresponding Markov chain of codon mutations on the state space S' given by a 64×64 transition probability matrix $\mathbf{M}'$. Then to account for stop codons, we replace $\mathbf{M}' = (m'_{uv})$ with the modified 64×64 mutation probability matrix $\mathbf{M} = (m_{uv})$ obtained by retaining all jumps to the states S_0. More precisely, if $m'_{uv} > 0$ for some $v \in S_0$, we put $m_{uv} = 0$ and $m_{uu} = \sum_{v \in S_0} m'_{uv}$.

Now, we are in position to mark each mutation event synonymous, nonsynonymous, or stopped by decomposing the mutation matrix $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{M}^{\text{syn}} + \mathbf{M}^{\text{non}} + \mathbf{M}^{\text{stop}},$$

where $\mathbf{M}^{\text{syn}}$ collects all nondiagonal elements m_{uv} for which the pair u, v is synonymous ($J(u,v) = 0$), the elements of $\mathbf{M}^{\text{non}}$ represent nonsynonymous changes ($J(u,v) = 1$) and $\mathbf{M}^{\text{stop}}$ stores the diagonal elements m_{uv}, $u = v$, of $\mathbf{M}$. Let $\mathbf{1}$ be a 64-column vector of only ones and let a and b denote the 64-column vectors

$$a = \mathbf{M}^{\text{syn}}\mathbf{1}, \quad b = \mathbf{M}^{\text{non}}\mathbf{1}. \tag{9}$$

In these vectors, the kth elements a_k and b_k are the conditional probabilities to obtain synonymous and nonsynonymous derived codons, given that a mutation occurs in codon u_k.

If we focus on a single codon site at a given generation, the chance to see a mutation is proportional to $3\mu N$ (which we may assume is much less than one). Hence, mutation events occur over time according to the transition probability matrix $\mathbf{M}_\mu = (1 - 3\mu N)I + 3\mu N\mathbf{M}$. Adding up the L sites of a sequence, it follows by independence of the mutation mechanism that the number of mutation events in a given generation is approximately Poisson distributed with mean θ. But only a small fraction of these events result in actual nucleotide substitutions, as we will see next by adding reproduction and selection to the model.

Discrete Time Wright–Fisher Model with Selection

For the reproductive dynamics of the model, we make the simplifying assumption that there is free recombination, that is, no linkage, between sites of a sequence. Each new generation is obtained from the previous generation by Wright–Fisher sampling acting on codons such that all codon sites develop independently of each other. Hence, a clonal site remains clonal until a newly mutated codon allele enters in one individual of the population. At this instance, the site becomes polymorphic and remains so over a period of time during which the frequency of the derived codon evolves according to a Wright–Fisher Markov chain until absorption. If the underlying mutation event is synonymous, then reproduction is neutral whereas if the mutation is nonsynonymous then the derived and ancestral codons are sampled with the selective weights $1 + s$ and 1, respectively. Typically, we consider selection to act deleteriously, prohibiting nonsynonymous changes by letting the selection parameter s be negative, $-1 < s < 0$. This is, however, no restriction as the model covers positive selection, $s > 0$, as well. At the time of absorption, the site becomes clonal.

To summarize the dynamics of codon evolution in the population, we keep track of L triplets $W^i_n = (A^i_n, B^i_n, X^i_n) \in S \times S \times [0,N]$, $i = 1, \ldots, L$. In each generation n, $A^i_n \in S$ is the type of the ancestral codon at site i, B^i_n is the type of the derived codon if i is polymorphic and equal to A^i_n if i is clonal, and X^i_n is the number of individuals with the derived codon allele at site i. By construction, the components $(W^i_n)_{n\geq 0}$, $i = 1, \ldots, L$ are independent and identically distributed discrete time Markov chains with state space $S \times S \times \{0, \ldots, N\}$. The one-step transitions of the single codon site chain are as follows. For mutation, jumps $(u,u,0) \to (u,v,1)$ are governed by the transition matrix $\mathbf{M}_\mu$ and occur with probability $3\mu N m_{uv}$, $v \neq u$. For reproduction, we let Y denote a random variable with the binomial distribution Bin (N,p), where $p = p(u,v,x)$ is the sampling probability

$$p = \frac{x(1 + sJ(u,v))}{N + sxJ(u,v)}.$$

Then, the jumps and corresponding transition probabilities of Wright–Fisher sampling are given by

$$(u,v,x) \to \begin{cases} (u,v,y) & \text{with prob.} \quad P(Y = y),\ 1 \leq y \leq N - 1 \\ (u,u,0) & \quad -\text{''}- \qquad P(Y = 0) \\ (v,v,0) & \quad -\text{''}- \qquad P(Y = N). \end{cases}$$

Continuous Time Approximation of the Full Model

Next, we consider large population size N and apply to each discrete time single site Markov chain (W^i_n), $1 \leq i \leq L$, the standard scheme of approximation $(A^i_{[Nt]}, B^i_{[Nt]}, N^{-1}X^i_{[Nt]})$ under the change of time and scale given by

$$n \mapsto Nt, \quad s \mapsto \gamma/N. \tag{10}$$

For the third component, it is well known that the derived allele frequency in the Wright–Fisher model with selection and no mutation converges as $N \to \infty$ to a diffusion process with absorbing boundaries in 0 and 1 given by the solution of the stochastic differential equation

$$\mathrm{d}\xi_t = \gamma\xi_t(1 - \xi_t)\,\mathrm{d}t + \sqrt{\xi_t(1 - \xi_t)}\,\mathrm{d}B_t,$$

where (B_t) is Brownian motion. Here, ξ_0 is the initial fraction of derived alleles. In our case $\xi_0 = 1/N \to 0$, which suggests that derived alleles would go extinct immediately. In our approach, however, the large population size scaling (eq. 10) is balanced against a total mutation intensity of order $\theta N/L$ per codon site. Thus, we replace the frequency $N^{-1}X^i_{[Nt]}$ by a process $(\mathcal{X}^i_t)$ where each nonzero excursion follows a path of (ξ_t), with $\xi_0 = 1/N > 0$ (the same approximation is used in Evans et al. [2007]). Then during the clonal periods, for which $\mathcal{X}^i_t = 0$, the first two components in $(A^i_{[Nt]}, B^i_{[Nt]}, 0)$ are continuous time Markov chains $(\mathcal{A}^i_t, \mathcal{B}^i_t, 0)$ (forced to have $\mathcal{A}^i_t = \mathcal{B}^i_t$) in holding until the next jump of $\mathcal{B}^i$. As the jump probability per generation is $3\mu N$, it follows that the jump rate per N generations is $3\mu N^2 = \theta N/L$. Hence, the generator matrix of $\mathcal{B}^i$ equals $\mathbf{M}_{N\mu} - I$. Each jump of $\mathcal{B}^i_t$ leaves $(\mathcal{A}^i_t)$ unaffected but initiates a diffusion path (ξ_t) embedded in $\mathcal{X}^i_t$. If the faith of (ξ_t) is extinction then $\mathcal{B}^i_t$ returns to its previous value stored in $(\mathcal{A}^i_t)$. If instead the path (ξ_t) gets fixed then $\mathcal{A}^i_t$ attains the current value of $\mathcal{B}^i_t$. More explicitly, we are approximating the discrete time Markov chain (W^i_n) with a continuous time Markov process $(\mathcal{W}^i_t) = (\mathcal{A}^i_t, \mathcal{B}^i_t, \mathcal{X}^i_t)$ with state space $(S,S,[0,1])$. The state space is a mixture of two jump coordinates and one continuous state coordinate and the process has the specific feature of holding and jumping from the boundary. We provide background information on diffusion processes with holding and jumping boundary in the supplementary text (Supplementary Material online). In our case, the boundary consists of all points $(u,u,0)$, $u \in \mathcal{S}$. If the current state of $\mathcal{W}^i_t$ is $(u,u,0)$ then after an exponential holding time of rate $\theta N/L$ the process $(\mathcal{B}^i_t)$ jumps to state v with probability m_{uv}. At the instance of such a jump $(\mathcal{W}^i_t)$ begins tracing a path of $(u,v,(\xi_t))$, with

$$d\xi_t = \gamma J(u,v)\xi_t(1 - \xi_t)\,dt + \sqrt{\xi_t(1 - \xi_t)}\,dB_t, \quad \xi_0 = 1/N,$$

until time of absorption. If the absorption event is fixation, the process jumps to $(v,v,0)$, if it is extinction, the jump is to $(u,u,0)$. Then again, the jump intensities apply until the next (ξ_t)-excursion takes place, and so on. The summation measure process

$$\Lambda_t = \sum_{i=1}^{L} \delta_{\mathcal{W}_t^i}, \tag{11}$$

now models the codon distribution in the entire sequence of length L across the population and over time, and provides a site frequency spectrum for the ensemble of clonal and polymorphic sites in the sequence. Each term features a record $(\mathcal{A}_t^i, \mathcal{B}_t^i)$ of ancestral and derived codon types plus a process $(\mathcal{X}_t^i)$, which alternates between dormant periods and active sessions. In each cycle, the dormant period has exponential length with intensity $\theta N/L$ and the active session consists of a Wright–Fisher diffusion corresponding to the absorption time of a Wright–Fisher diffusion with initial value $1/N$ running until absorption in 0 or 1.

The Codon Equilibrium Distribution

The first component $(\mathcal{A}_t^i)$ of each $(\mathcal{W}_t^i)$ has its jumps restricted to times of actual nucleotide substitution events. Synonymous mutations get fixed with probability $1/N$ and nonsynonymous mutations fix with probability ω_γ/N, $\gamma \neq 0$, with ω_γ introduced in equation (2) (the formal background is given in eq. 2 of the supplementary text, Supplementary Material online). So, if we single out only substitution events then the Markov chain transition probabilities reduce to $(\mathbf{M}^{\mathrm{syn}} + \omega_\gamma \mathbf{M}^{\mathrm{non}})/N$. Thus, we consider a continuous time Markov chain with infinitesimal generator

$$\begin{gathered} \mathbf{Q}_\gamma = 3\mu N\,(\mathbf{M}^{\mathrm{syn}} + \omega_\gamma \mathbf{M}^{\mathrm{non}} - \mathbf{V}^{\mathrm{diag}}), \\ (\mathbf{V}^{\mathrm{diag}})_{kk} = a_k + \omega_\gamma b_k \end{gathered} \tag{12}$$

where the total jump intensities stored in the diagonal matrix $\mathbf{V}^{\mathrm{diag}}$ were introduced in equation (9). Here, the stop codons S_0 are transient states. We conclude that the continuous time approximation $\mathcal{A}_t^i$ behaves as the Markov chain with generator $\mathbf{Q}_\gamma$, except that each holding time is prolonged by the fixation time of the corresponding diffusion path in $\mathcal{X}_t^i$. To obtain a steady state codon distribution, however, one should restrict to the clonal population, which is given precisely by the Markov generator $\mathbf{Q}_\gamma$. Hence, we define η to be the unique stationary distribution which satisfies $\eta\mathbf{Q}_\gamma = \mathbf{0}$ and which provides a steady state for the irreducible Markov chain restricted to S and with $\eta_k = 0$ for $u_k \in S_0$. Now for large N, we have the interpretation that substitutions occur according to $\mathbf{Q}_\gamma$ and the typical codon frequencies observed in a clonal site in equilibrium is given by η. Furthermore, in this equilibrium, we can measure the proportions of synonymous and nonsynonymous events among all mutations. For example, whenever a mutation event occurs according to $\mathbf{M}$ it is synonymous with probability given by the scalar product $\langle \eta, a \rangle = \eta\, \mathbf{M}^{\mathrm{syn}} \mathbf{1}$. Consequently, we introduce the probability distribution $\pi = (\pi_{\mathrm{syn}}, \pi_{\mathrm{non}}, \pi_{\mathrm{stop}})$, by

$$\pi_{\mathrm{syn}} = \eta\, \mathbf{M}^{\mathrm{syn}} \mathbf{1}, \quad \pi_{\mathrm{non}} = \eta\, \mathbf{M}^{\mathrm{non}} \mathbf{1}, \quad \pi_{\mathrm{stop}} = \eta\, \mathbf{M}^{\mathrm{stop}} \mathbf{1}. \tag{13}$$

In conclusion, the typical rates at which mutation events are synonymous, nonsynonymous, or inert are obtained as the weighted mutation rates

$$\mu_{\mathrm{syn}} = \mu\, \pi_{\mathrm{syn}}, \quad \mu_{\mathrm{non}} = \mu\, \pi_{\mathrm{non}}, \quad \mu_{\mathrm{stop}} = \mu\, \pi_{\mathrm{stop}},$$

and the conditional distribution of synonymous and nonsynonymous mutations given that a nonstop codon transition occurs is

$$p_{\mathrm{syn}} = \frac{\pi_{\mathrm{syn}}}{\pi_{\mathrm{syn}} + \pi_{\mathrm{non}}}, \quad p_{\mathrm{non}} = \frac{\pi_{\mathrm{non}}}{\pi_{\mathrm{syn}} + \pi_{\mathrm{non}}}.$$

As a consequence, the resulting synonymous and nonsynonymous mutation intensities for the population of sequences are given by

$$\theta_{\mathrm{syn}} = \pi_{\mathrm{syn}}\theta = 3LN\mu_{\mathrm{syn}}, \quad \theta_{\mathrm{non}} = \pi_{\mathrm{non}}\theta = 3LN\mu_{\mathrm{non}}. \tag{14}$$

Mutation Rates for Standard Models

The probability distributions π and $(p_{\mathrm{syn}}, p_{\mathrm{non}})$, and hence the rates $\theta_{\mathrm{syn}}, \theta_{\mathrm{non}}$ can be found explicitly for standard mutation matrices $\mathbf{H}$. The Kimura model with $\alpha + 2\beta = 1$, takes into account a mutation ratio $\kappa = \alpha/\beta$ of transitions versus transversions. For this model, the uniform distribution on nonstop codons given by $\eta_k = 1/61$, $u_k \in S$, is stationary for $\mathbf{M}$ and hence $\mathbf{Q}_\gamma$. This holds not only for the neutral case $\gamma = 0$ but also in general. Indeed, $\mathbf{Q}_\gamma$ is doubly stochastic for any γ, and by equation (13),

$$\pi = \left(\frac{12}{61} + \frac{26}{183}\alpha, \frac{46}{61} - \frac{22}{183}\alpha, \frac{3}{61} - \frac{4}{183}\alpha \right),$$
$$p_{\mathrm{syn}} = \frac{36 + 26\alpha}{174 + 4\alpha}, \quad p_{\mathrm{non}} = \frac{138 - 22\alpha}{174 + 4\alpha}.$$

A special case of the Kimura model, which we have used for the simulation study in this work, is $\alpha = 1/2$, $\beta = 1/4$, $\kappa = 2$. Then

$$\pi = \left(\frac{49}{183}, \frac{127}{183}, \frac{7}{183} \right), \quad p_{\mathrm{syn}} = \frac{49}{176} = 0.2784,$$
$$p_{\mathrm{non}} = \frac{127}{176} = 0.7216.$$

Our model is flexible and allows for any other nucleotide substitution pattern, including asymmetric versions with nonzero diagonal elements h_{ii}. In general, if $\rho = (\rho_1\, \rho_2\, \rho_3\, \rho_4)$ is a steady state for $\mathbf{H}$, hence representing the typical fractions of nucleotides in the population, the corresponding steady state of codons will be

$$\eta_u = c \cdot \rho_i \rho_j \rho_k, \quad u = (ijk), \quad c^{-1} = \sum_u \rho_i \rho_j \rho_k,$$

and the vector π is again found from equation (13). Commonly used versions of the Goldman–Yang model (Goldman and Yang 1994), fall in this category.

Poisson Random Field Approximation

In equation (11), letting the initial times and paths of all active sessions form points of a Poisson random point measure leads to what has been called a Poisson random fields model (Sawyer and Hartl 1992). Although we do not pursue this line of argument formally, our approach is similar in spirit. The quantities of primary interest in this work, which measure divergence of sampled sequences between populations, are recognized as random functionals of Λ_t. These functionals count specific events that occur along the site processes with probabilities proportional to θ/L. With L independent sites, this leads to an approximate Poisson number of events over the total length of the sequence with mean proportional to θ. To find the expected values of the divergence functionals, we apply the site frequency spectrum, which arises by superposing Wright–Fisher diffusions starting in $1/N$ at Poisson times with rate θN (Evans et al. 2007).

Recall from equation (12) that the typical codon frequencies are given by the steady state $\{\eta_u, u \in \mathcal{S} \setminus \mathcal{S}_0\}$. Conditional on the first component of Λ_t in site i being $\mathcal{A}_t^i = u$, then $\mathcal{B}_t^i$ typically makes a large number of jumps from u to a sense codon neighbor v of u and then back to u, before one of these excursions eventually fixes and results in a change of state of $\mathcal{A}_t^i$. Given such a u, $\mathcal{B}_t^i$ settles in a steady state with probabilities m_{uv}. Summing again over u with the stationary weight, we recover $\sum_u \eta_u m_{uv} = \eta_v$. Therefore, it is reasonable to assume that mutations occur along the time axis as an approximate Poisson process with intensity $N\theta$ and each event sparks the excursion of a Wright–Fisher diffusion (ξ_s) with $\xi_0 = 1/N$. The fractions of mutation events, which lead to synonymous and nonsynonymous codon pairs are obtained by the weighted summations

$$\sum_{u,v \in S \setminus S_0} \eta_u m_{uv}^{\mathrm{syn}} = \pi_{\mathrm{syn}}, \qquad \sum_{u,v \in S \setminus S_0} \eta_u m_{uv}^{\mathrm{non}} = \pi_{\mathrm{non}},$$

and the corresponding mutation intensities are given by $N\theta_{\mathrm{syn}}$ and $N\theta_{\mathrm{non}}$. A key feature here is that nonsynonymous and synonymous codon mutations evolve independently of each other.

Rate of Divergence over Time

In this section, we consider a population split where a population of size N has been running indefinitely from the past and is replaced at time $t = 0$ instantaneously by two identical copies of the population. The new branches represent two emerging species both of population size N with initially the same number of clonal and polymorphic sites and with identical codon frequencies. From the splitting time and onwards each of the two populations evolve independently according to the same mechanisms of mutation and selection. To follow the onset of divergence between the species, we sample randomly one individual in each population. Let $D(t)$ denote the number of nucleotide differences between these two sequences at time $t \geq 0$. We will analyze three types of differences which contribute to the total divergence $D(t)$, by writing

$$D(t) = D_{\mathrm{fix}}(t) + D_{\mathrm{pol}}(t) + D_{\mathrm{anc}}(t),$$

where

$D_{\mathrm{fix}}(t) =$ sequence divergence at t from mutations during [0, t] fixed uring [0, t],
$D_{\mathrm{pol}}(t) =$ number of derived alleles sampled from lineage-specific polymorphic sites at t and
$D_{\mathrm{anc}}(t) =$ sequence divergence at t attributed to ancestral polymorphisms existing at t = 0.

Here, $D_{\mathrm{fix}}(t)$ and $D_{\mathrm{pol}}(t)$ are sums of two independent contributions, one from each population, whereas $D_{\mathrm{anc}}(t)$ involves the joint initial state at $t = 0$. The dominant source of divergence between two sequences which is visible after a longer time span t, is the fixation of new alleles from recent mutations in each population during $(0,t)$. The growth in the number of substitutions and the subsequent growth of $D_{\mathrm{fix}}(t)$ is essentially linear in t. This is the same mechanism, which is responsible for the mixing of codons and the appearance of a steady state of codon frequencies in the long run. The additional contributions to the total divergence $D(t)$ are bounded as functions of t but are important to understand how the linear growth regime is attained after population split.

We will analyze the three types of divergence by relating the components of $D(t)$ to suitable functionals of Λ_t in equation (11), extended to cover a common ancestry for $t \leq 0$ and two independent species populations for $t > 0$. At any given time, some of the L codons in the model are likely to be polymorphic and hence exempt from mutation events. But as the number of polymorphic sites is typically much smaller than L, we will apply the approximation that the total mutation intensity is $\theta = 3\mu NL$ per sequence and generation, hence $N\theta$ per sequence and time units t. It is the independent Poisson mutation processes in our model that drive the various contributions to $D(t)$. In particular, nucleotide substitutions count into $D_{\mathrm{fix}}(t)$, which therefore has a Poisson distribution. Sampled differences, both ancestral and present polymorphic, arise at most one in each codon site. By equations (15) and (16) of the supplementary text (Supplementary Material online), the probability to see one of these differences at a given site after sampling sequences in two populations is proportional to the mutation intensity per codon and time unit, namely θ/L. Summing over L codons this gives a binomial, hence approximately Poisson, number of differences with mean proportional to θ. But then $D_{\mathrm{pol}}(t)$ and $D_{\mathrm{anc}}(t)$, and therefore $D(t)$ itself have approximate Poisson distributions. Our next focus will be to find the corresponding expected values.

Expected Divergence after Population Split

We begin with divergence based on fixed differences. Our model associates with the Wright–Fisher diffusion (ξ_t) its fixation time τ_1, extinction time τ_0 and absorption time $\tau = \min(\tau_0, \tau_1)$. The total number of mutation events in $[0,t)$ is Poisson with mean $N\theta t$ and conditional on this number the events are uniformly distributed on $[0,t)$. One such mutation occurring at time s results in a fixation if $s + \tau_1 < t$. Summing over both populations and taking into account the fractions of synonymous and nonsynonymous events, this gives

$$\begin{aligned}\mathbb{E}D_{\mathrm{fix}}(t) &= 2N\theta t \cdot \frac{1}{t}\int_0^t \mathbb{P}_{1/N}(s + \tau_1 < t)\,\mathrm{d}s \\ &= 2N\theta_{\mathrm{syn}} \int_0^t \mathbb{P}^0_{1/N}(\tau_1 < s)\,\mathrm{d}s \\ &\quad + 2N\theta_{\mathrm{non}} \int_0^t \mathbb{P}^{\gamma}_{1/N}(\tau_1 < s)\,\mathrm{d}s\end{aligned}$$

Rewrite as $\mathbb{P}^{\gamma}_{1/N}(\tau_1 < t) = P^{*\gamma}_{1/N}(\tau_1 < t)\, q_\gamma(1/N)$, where $\mathbb{P}^{*\gamma}_{x}(\tau_1 < t)$ is the conditional probability given fixation $(\tau_1 < \infty)$ and $q_\gamma(x) = \mathbb{P}^{\gamma}_{x}(\tau_1 < \tau_0)$ denotes the fixation probability of a mutated allele which emerges with frequency x in the population. By equation (2), $q_\gamma(1/N)\sim\omega_\gamma/N$. Hence, in the large N limit,

$$\mathbb{E}D_{\mathrm{fix}}(t) = 2\theta_{\mathrm{syn}} d^{\mathrm{syn}}_{\mathrm{fix}}(t) + 2\theta_{\mathrm{non}}\omega_\gamma d^{\mathrm{non}}_{\mathrm{fix}}(t)$$

with

$$d^{\mathrm{syn}}_{\mathrm{fix}}(t) = \int_0^t \mathbb{P}^{*0}_0(\tau_1 < s)\,ds, \quad d^{\mathrm{non}}_{\mathrm{fix}}(t) = \int_0^t \mathbb{P}^{*\gamma}_0(\tau_1 < s)\,ds.$$

Equivalently,

$$d^{\mathrm{syn}}_{\mathrm{fix}}(t) = t - \mathbb{E}^{*0}_0(\min(\tau_1,t)), \quad d^{\mathrm{non}}_{\mathrm{fix}}(t) = t - \mathbb{E}^{*\gamma}_0(\min(\tau_1,t)),$$

which reveals the deviation from linear growth of $\mathbb{E}D_{\mathrm{fix}}(t)$. For $\gamma = 0$,

$$\mathbb{E}^{*0}_0(\min(\tau_1,t)) = G_0(t),$$

with the function G_0 defined in equation (12) of the supplementary text (Supplementary Material online). One option for the general selective case $\gamma \neq 0$ would be to apply a spectral representation for the transition density of the Wright–Fisher model with selection, and rely on numerical computations of the corresponding eigenvalue/eigenvector problem. To keep things simpler while retaining a reasonable degree of accuracy, instead we propose at this place the approximation

$$\mathbb{E}^{*\gamma}_0(\min(\tau_1,t)) \approx G_\gamma(t), \quad G_\gamma(t) = \min(G_0(t),\mathbb{E}^{*\gamma}_0(\tau_1)).$$

By applying a known integral expression for $\mathbb{E}^{*\gamma}_x(\tau_1)$, see Karlin and Taylor (1981), Ch. 15, (9.9), and expanding the resulting integral in γ, we obtain

$$\begin{aligned}\mathbb{E}^{*\gamma}_0(\tau_1) &= \int_0^1 \frac{(1-e^{-2\gamma y})(1-e^{-2\gamma(1-y)})}{y(1-y)\gamma(1-e^{-2\gamma})}\,dy \\ &= 2 - \frac{1}{9}\gamma^2 + \frac{7}{675}\gamma^4 + O(\gamma^6)\end{aligned} \tag{15}$$

and hence

$$G_\gamma(t) \approx \min(G_0(t), 2-\gamma^2/9).$$

In summary,

$$\begin{aligned}\mathbb{E}D_{\mathrm{fix}}(t) &\approx 2\theta_{\mathrm{syn}}(t - G_0(t)) \\ &\quad + 2\theta_{\mathrm{non}}\omega_\gamma(t - \min(G_0(t), 2-\gamma^2/9)).\end{aligned} \tag{16}$$

Next, we turn to divergence based on lineage-specific polymorphisms. As discussed earlier, consider a mutation at a uniformly distributed time s in $[0,t)$. The corresponding derived allele exists at time t if $s+\tau > t$, in which case $0 < \xi^s_t < 1$. In each population, we have sampled one particular individual. The probability that this individual carries the derived allele is approximately ξ^s_t. Hence, the probability that this lineage-specific polymorphism contributes to estimates of divergence is

$$\begin{aligned}\frac{1}{t}\int_0^t \mathbb{E}_{1/N}(\xi^s_t, s+\tau > t)\,ds &= \frac{1}{t}\int_0^t \mathbb{E}_{1/N}(\xi^0_r, \tau > r)\,dr \\ &= \frac{1}{t}\mathbb{E}_{1/N}\Big[\int_0^{\min(t,\tau)} \xi_r\,dr\Big].\end{aligned}$$

Now summing over all mutation events in $[0,t]$ in both populations and splitting neutral synonymous ones from selective nonsynonymous mutations, we find

$$\mathbb{E}D_{\mathrm{pol}}(t) = 2\theta_{\mathrm{syn}} d^{\mathrm{syn}}_{\mathrm{pol}}(t) + 2\theta_{\mathrm{non}} d^{\mathrm{non}}_{\mathrm{pol}}(t), \tag{17}$$

where

$$d^{\mathrm{syn}}_{\mathrm{pol}}(t) = \lim_{N\to\infty} N\mathbb{E}^0_{1/N}\Big[\int_0^{\min(t,\tau)} \xi_s\,ds\Big],$$

$$d^{\mathrm{non}}_{\mathrm{pol}}(t) = \lim_{N\to\infty} N\mathbb{E}^\gamma_{1/N}\Big[\int_0^{\min(t,\tau)} \xi_s\,ds\Big].$$

To continue estimating $d^{\mathrm{syn}}_{\mathrm{pol}}(t)$ and $d^{\mathrm{non}}_{\mathrm{pol}}(t)$, it is a useful fact that the functional $\mathbb{E}^\gamma_x[\int_0^\tau \xi_s\,ds]$ has an explicit representation as an integral over the corresponding Green's function (for details see supplementary text, Supplementary Material online). Indeed, for $\gamma = 0$, we have $N\mathbb{E}^0_{1/N}[\int_0^\tau \xi_s\,ds] \to 2$ and for $\gamma \neq 0$,

$$\lim_{N\to\infty} N\mathbb{E}^\gamma_{1/N}\Big[\int_0^\tau \xi_s\,ds\Big] = \omega_\gamma \int_0^1 \frac{1-e^{-2\gamma y}}{\gamma y}\,dy, \tag{18}$$

as N tends to infinity. To accommodate the behavior for small t, we apply the approximation

$$\mathbb{E}^\gamma_x\Big[\int_0^{\min(t,\tau)} \xi_s\,ds\Big] \approx \min\Big(\int_0^t \mathbb{E}^\gamma_x[\xi_s]\,ds, \mathbb{E}^\gamma_x\Big[\int_0^\tau \xi_s\,ds\Big]\Big). \tag{19}$$

Here $\mathbb{E}^0_x[\xi_s] = x$ for $\gamma = 0$, so for large N

$$\begin{aligned}d^{\mathrm{syn}}_{\mathrm{pol}}(t) &\approx N\mathbb{E}^0_{1/N}\Big[\int_0^{\min(t,\tau)} \xi_s\,ds\Big] \\ &\approx \min\Big(N\int_0^t \mathbb{E}^\gamma_{1/N}[\xi_s]\,ds, N\mathbb{E}^\gamma_{1/N}\Big[\int_0^\tau \xi_s\,ds\Big]\Big) \approx \min(t,2).\end{aligned}$$

More generally, by conditioning,

$$\mathbb{E}^\gamma_x[\xi_t] = \mathbb{E}^\gamma_x[\xi_t \mid \tau_1 < \tau_0]\,q_\gamma(x) + \mathbb{E}^\gamma_x[\xi_t \mid \tau_0 < \tau_1]\,(1-q_\gamma(x)).$$

The conditional expectations on the right hand side appear to be well approximated by those for the neutral case $\gamma = 0$, which can be derived explicitly,

$$\begin{aligned}\mathbb{E}^\gamma_x[\xi_t \mid \tau_1 < \tau_0] &\approx \mathbb{E}^0_x[\xi_t \mid \tau_1 < \tau_0] = 1-(1-x)e^{-t}, \\ \mathbb{E}^\gamma_x[\xi_t \mid \tau_0 < \tau_1] &\approx \mathbb{E}^0_x[\xi_t \mid \tau_0 < \tau_1] = xe^{-t}.\end{aligned}$$

Thus,

$$\mathbb{E}^\gamma_x[\xi_t] \approx q_\gamma(x)(1-e^{-t}) + xe^{-t} \tag{20}$$

and

$$N\int_0^t \mathbb{E}^\gamma_{1/N}[\xi_s]\,ds \approx \omega_\gamma t + (1-\omega_\gamma)(1-e^{-t}).$$

and if we plug this and equation (18) into equation (19) it follows

$$d^{\mathrm{non}}_{\mathrm{pol}}(t) \approx N\mathbb{E}^{\gamma}_{1/N}\Big[\int_0^{\min(t,\tau)} \xi_s \,\mathrm{d}s\Big]$$

$$\approx \min\Big(\omega_\gamma t + (1-\omega_\gamma)(1-\mathrm{e}^{-t}), \omega_\gamma \int_0^1 \frac{1-\mathrm{e}^{-2\gamma y}}{\gamma y}\mathrm{d}y\Big).$$

Thus, if we apply the integral approximation

$$\int_0^1 \frac{1-\mathrm{e}^{-2\gamma y}}{\gamma y} dy \approx \frac{\mathrm{e}^{-2\gamma} - 1 + 2\gamma + 2\gamma^2}{2\gamma^2},$$

which should be sufficiently accurate at least in the range $|\gamma| \le 2$, and put

$$H_\gamma(t) = \min\Big(t + \frac{1-\omega_\gamma}{\omega_\gamma}(1-\mathrm{e}^{-t}), \frac{\mathrm{e}^{-2\gamma} - 1 + 2\gamma + 2\gamma^2}{2\gamma^2}\Big), \tag{21}$$

then $d^{\mathrm{non}}_{\mathrm{fix}}(t) \approx \omega_\gamma H_\gamma(t)$ and we may sum up the contribution to divergence based on polymorphic sites as

$$\mathbb{E}_x D_{\mathrm{pol}}(t) \approx 2\theta_{\mathrm{syn}} \min(t,2) + 2\theta_{\mathrm{non}}\omega_\gamma H_\gamma(t). \tag{22}$$

The ancestral contribution to divergence between the two populations has its origin in the common history of mutation events that has occurred at times $s < 0$. Of all polymorphic sites which exist at the time of population split $t = 0$, those for which the sampled individuals at a later time are different in the two populations add to ancestral divergence. This includes polymorphic and fixed or extinct states as long as one is ancestral and the other derived. A derived allele starting at $s < 0$ exists with frequency $\xi^s_0 > 0$ at time $t = 0$, if $s + \tau > 0$. Hence, conditionally given ξ^s_0, the chance to see such a difference is

$$\mathbb{E}_{\xi^s_0}[\xi^{(1)}_t(1-\xi^{(2)}_t) + \xi^{(2)}_t(1-\xi^{(1)}_t)]$$

where each population is indicated with an additional upper index. By independence after time $t = 0$ this is

$$2m^\gamma_t(\xi^s_0)(1-m^\gamma_t(\xi^s_0)), \qquad m^\gamma_t(x) = \mathbb{E}^\gamma_x[\xi_t], \quad m^0_t(x) = x.$$

To average over all states at $t = 0$, we note that the number of mutation events over a finite interval $(-K,0)$ is Poisson with mean $N\theta K$. Hence, conditional on the number of events the mutation times are uniformly distributed in $(-K,0)$. The probability that one of these events contributes to ancestral divergence is

$$\frac{1}{K}\int_{-K}^{0} \mathbb{E}_x[1_{\{s+\tau>0\}} 2m^\gamma_t(\xi^s_0)(1-m^\gamma_t(\xi^s_0))]\,\mathrm{d}s$$

$$= \frac{1}{K}\int_0^K \mathbb{E}_x[1_{\{r<\tau\}} 2m^\gamma_t(\xi^0_r)(1-m^\gamma_t(\xi^0_r))]\,\mathrm{d}r.$$

Letting $K \to \infty$, and separating synonymous and nonsynonymous events,

$$\mathbb{E}D_{\mathrm{anc}}(t) = 2\theta_{\mathrm{syn}} N\mathbb{E}^0_{1/N}\int_0^\tau \xi_s(1-\xi_s)\,\mathrm{d}s$$

$$+ 2\theta_{\mathrm{non}} N\mathbb{E}^\gamma_{1/N}\int_0^\tau m^\gamma_t(\xi_s)(1-m^\gamma_t(\xi_s))\,\mathrm{d}s. \tag{23}$$

Again, we rely on being able to evaluate functionals of the Wright–Fisher process of the type $\mathbb{E}^0_x\int_0^\tau g(\xi_s)\,\mathrm{d}s$ for sufficiently nice functions g. First (eq. 5 of the supplementary text, Supplementary Material online),

$$N\mathbb{E}^0_{1/N}\int_0^\tau \xi_s(1-\xi_s)\,\mathrm{d}s \to 1.$$

For the second term in equation (23), we use again the approximation (20) of $m^\gamma_t(x)$. Then

$$N\mathbb{E}^0_{1/N}\int_0^\tau m^\gamma_t(\xi_s)(1-m^\gamma_t(\xi_s))\,\mathrm{d}s \to \omega_\gamma J_\gamma(t),$$

$$J_\gamma(t) = \int_0^1 \frac{1-\mathrm{e}^{-2\gamma(1-y)}}{\gamma y(1-y)} m^\gamma_t(y)(1-m^\gamma_t(y))\,\mathrm{d}y,$$

with

$$m^\gamma_t(y)(1-m^\gamma_t(y))$$

$$\approx (1-\mathrm{e}^{-t})^2 q_\gamma(y)(1-q_\gamma(y)) + \mathrm{e}^{-2t}y(1-y)$$

$$+\ \mathrm{e}^{-t}(1-\mathrm{e}^{-t})q_\gamma(y)(1-y) + \mathrm{e}^{-t}(1-\mathrm{e}^{-t})(1-q_\gamma(y))y.$$

By expanding the resulting integrals in a series up to second order in γ,

$$J_\gamma(t) \approx 1 - \frac{\gamma}{3}(1+\mathrm{e}^{-t}) - \frac{\gamma^2}{18}(1 - 10\mathrm{e}^{-t} + 3\mathrm{e}^{-2t}). \tag{24}$$

Hence, we conclude from equation (23) that the expected divergence attributed to ancestral polymorphisms in this model is

$$\mathbb{E}D_{\mathrm{anc}}(t) = 2\theta_{\mathrm{syn}} + 2\theta_{\mathrm{non}}\omega_\gamma J_\gamma(t). \tag{25}$$

At $t = 0$, as $m^\gamma_0(x) = x$,

$$J_\gamma(0) = \int_0^1 \frac{1-\mathrm{e}^{-2\gamma(1-y)}}{\gamma}\mathrm{d}y = \frac{\omega_\gamma - 1}{\gamma\omega_\gamma} \tag{26}$$

so that

$$\mathbb{E}D_{\mathrm{anc}}(0) = 2\theta_{\mathrm{syn}} + 2\theta_{\mathrm{non}}\frac{\omega_\gamma - 1}{\gamma}$$

and we have the initial ratio

$$\frac{\mathbb{E}D^{\mathrm{non}}_{\mathrm{anc}}(0)/\theta_{\mathrm{non}}}{\mathbb{E}D^{\mathrm{syn}}_{\mathrm{anc}}(0)/\theta_{\mathrm{syn}}} \approx \frac{\omega_\gamma - 1}{\gamma} \approx 1 + \frac{1}{3}\gamma - \frac{1}{45}\gamma^3.$$

Now, we are in position to study synonymous and nonsynonymous divergence separately. By summing up the three parts of divergence in equations (16), (22), and (25), that is, fixed differences and divergence attributed to lineage-specific and ancestral polymorphisms, we find for any fixed t that $D^{\mathrm{syn}}(t)$ and $D^{\mathrm{non}}(t)$ have approximate Poisson distributions with expected values

$$\mathbb{E}D^{\mathrm{syn}}(t) = 2\theta_{\mathrm{syn}} d^{\mathrm{syn}}(t), \quad \mathbb{E}D^{\mathrm{non}}(t) = 2\theta_{\mathrm{non}} d^{\mathrm{non}}(t), \tag{27}$$

such that

$$d^{\mathrm{syn}}(t) = t - G_0(t) + \min(t,2) + 1$$

and

$$d^{\mathrm{non}}(t) = \omega_\gamma\big(t - \min(G_0(t), 2 - \gamma^2/9) + H_\gamma(t) + J_\gamma(t)\big)$$

with G_0 defined in the supplementary text (Supplementary Material online), H_γ in equation (21) and J_γ in equation (24).

Statistical Properties of Divergence and Poisson Ratios

The codon evolution model introduced above allows us to study in detail the accumulation of fixed differences and the variation of sampled differences between sequences from two independent populations after the split from an ancestral population. We continue in this direction by studying the ratio of nonsynonymous to synonymous divergence d*N*/d*S* in a population genetics framework as a ratio of Poisson random variables.

Intrinsic Variation of Poisson Ratios

Considering the ratio of scaled Poisson variables

$$\frac{D^{\text{non}}(t)/\theta_{\text{non}}}{D^{\text{syn}}(t)/\theta_{\text{syn}}} \quad \text{on} \quad D^{\text{syn}}(t) > 0,$$

enables us to study d*N*/d*S* as a function of divergence time by defining

$$\mathrm{d}N/\mathrm{d}S(t) = \mathbb{E}\left(\frac{D^{\text{non}}(t)}{D^{\text{syn}}(t)} \mid D^{\text{syn}}(t) > 0\right)\frac{\theta_{\text{syn}}}{\theta_{\text{non}}}.$$

To compute this conditional expectation, we note that if Y is a Poisson random variable with mean m then

$$\begin{aligned}\mathbb{E}(1/Y \mid Y > 0) &= \sum_{k=1}^{\infty}\frac{1}{k}P(Y = k \mid Y > 0)\\ &= \sum_{k=1}^{\infty}\frac{1}{k}\frac{m^k e^{-m}}{k!(1-e^{-m})}.\end{aligned}$$

Hence, if we introduce the function

$$C(m) = m\,\mathbb{E}(1/Y \mid Y > 0) = \sum_{k=1}^{\infty}\frac{m^{k+1}}{k\cdot k!}\frac{e^{-m}}{1-e^{-m}}, \quad m > 0, \tag{28}$$

it follows

$$\begin{aligned}\mathrm{d}N/\mathrm{d}S(t) &= \theta_{\text{non}}^{-1}\mathbb{E}D^{\text{non}}(t)\,\theta_{\text{syn}}\mathbb{E}[D^{\text{syn}}(t)^{-1} \mid D^{\text{syn}} > 0]\\ &= \frac{d^{\text{non}}(t)}{d^{\text{syn}}(t)}\,C(2\theta_{\text{syn}}d^{\text{syn}}(t)).\end{aligned} \tag{29}$$

Thus, we have two alternative methods to estimate d*N*/d*S* based on computing suitable expected values. The previous ratio of expected values $d^{\text{non}}(t)/d^{\text{syn}}(t)$ and the new expected value of a ratio. Our main motivation for introducing equation (29) is that current practice in empirical work on d*N*/d*S* does not per se seek to estimate the expected number of synonymous divergence or expected number of nonsynonymous divergence but rather aim at determining single observations of these quantities. Then forming the ratio of these observations leads to equation (29). It is noteworthy that the new estimate is simply a multiplicative perturbation of the previous ratio and that the prefactor $C(2\theta_{\text{syn}}d^{\text{syn}}(t))$ depends on both θ and t. However, the dependence is restricted to variations in the product $2\theta_{\text{syn}}d^{\text{syn}}(t)$. If $t > 2$, this is $2\pi^{\text{syn}}\theta(1+t)$ and if we take in addition the Jukes–Cantor mutation model, then the prefactor is approximately $C(\theta(1+t)/2)$. For example, if $\theta(1+t) \approx 8$ then $\mathrm{d}N/\mathrm{d}S(t)$ is more than 30% larger than $\mathrm{d}N/\mathrm{d}S(t)_{\text{total}}$ and if $\theta(1+t) \approx 20$ then $\mathrm{d}N/\mathrm{d}S(t)$ is more than 10% larger than the reference ratio. This said, one should also notice that for sufficiently large θ or t, the two estimates coincide. In fact, $C(t)\sim 1$ as $t \to \infty$ and hence $\mathrm{d}N/\mathrm{d}S(t) \approx d^{\text{non}}(t)/d^{\text{syn}}(t)$ as $\theta_{\text{syn}}d_{\text{syn}}(t) \to \infty$. For small t,

$$\mathrm{d}N/\mathrm{d}S(t) \to J_\gamma(0)\,C(2\theta_{\text{syn}}), \quad t \to 0.$$

Confidence Intervals

Suppose we have fixed θ and H, and determined the vector π. Considering γ as an unknown parameter we suppose k and $n - k$ are observations of the independent Poisson random variables $D^{\text{syn}}(t)$ and $D^{\text{non}}(t)$ at some unknown time t. Then, given the sum n, $D^{\text{syn}}(t)$ has a binomial distribution $\mathrm{B}in(n,p(t))$, where

$$p(t) = \frac{2\theta^{\text{syn}}d^{\text{syn}}(t)}{2\theta^{\text{syn}}d^{\text{syn}}(t) + 2\theta^{\text{non}}d^{\text{non}}(t)} = \left(1 + \frac{\pi^{\text{non}}}{\pi^{\text{syn}}}\frac{d^{\text{non}}(t)}{d^{\text{syn}}(t)}\right)^{-1}. \tag{30}$$

Now apply a confidence interval $[a,b]$ for $p(t)$. It follows that $[\pi^{\text{syn}}(1/b - 1)/\pi^{\text{non}}, \pi^{\text{syn}}(1/a - 1)/\pi^{\text{non}}]$ is a confidence interval for $d^{\text{non}}(t)/d^{\text{syn}}(t)$. If we apply specifically a 95% interval of the so-called Wilson type for $p(t)$, based on normal approximation of the binomial distribution, then with $z = 1.96$ the limits a and b are given by

$$\left(\frac{k}{n} + \frac{z^2}{2n} \pm z\sqrt{\frac{k(n-k)}{n^3} + \frac{z^2}{4n^2}}\right)\Big/\left(1 + \frac{z^2}{n}\right).$$

The lower und upper limits of the observed confidence interval for $d^{\text{non}}(t)/d^{\text{syn}}(t)$ become

$$\frac{\pi^{\text{syn}}}{\pi^{\text{non}}}\left(\frac{1 + \frac{z^2}{n}}{\frac{k}{n} + \frac{z^2}{2n} \pm z\sqrt{\frac{k(n-k)}{n^3} + \frac{z^2}{4n^2}}} - 1\right).$$

For example, if the upper limit would attain a value less than one we have evidence with a 5% degree of significance to reject a null hypothesis of neutral evolution in favor of negative selection.

Now we change the perspective and consider γ and hence $d^{\text{non}}(t)$ as known. We want to estimate the variation in $D^{\text{non}}(t)/D^{\text{syn}}(t)$. Given that the total number of observed differences at time t is n, then with $p = p(t)$

$$\begin{aligned}&\mathbb{P}\left(p - z\sqrt{\frac{p(1-p)}{n}} \le \frac{D^{\text{syn}}(t)}{n} \le p + z\sqrt{\frac{p(1-p)}{n}}\right)\\ &= \mathbb{P}\left(\left(p + z\sqrt{\frac{p(1-p)}{n}}\right)^{-1} - 1 \le \frac{D^{\text{non}}(t)}{D^{\text{syn}}(t)}\right.\\ &\qquad\left.\le \left(p - z\sqrt{\frac{p(1-p)}{n}}\right)^{-1} - 1\right) = 0.95.\end{aligned}$$

Hence, we may take p as in (30) and estimate n with $\hat{n} = 2\theta(\pi^{\mathrm{syn}} d^{\mathrm{syn}}(t) + \pi^{\mathrm{non}} d^{\mathrm{non}}(t))$ to get a 95% "confidence band" for $\frac{D^{\mathrm{non}}(t)/\theta^{\mathrm{non}}}{D^{\mathrm{syn}}(t)/\theta^{\mathrm{syn}}}$ as

$$\left(\left(p \pm z\sqrt{\frac{p(1-p)}{n}}\right)^{-1} - 1\right)\frac{\pi^{\mathrm{syn}}}{\pi^{\mathrm{non}}}. \qquad (31)$$

Simulation of Codon Evolution

The accuracy of our analytical results of course depends on the degree to which the various approximations that we applied during their derivation have distorted the properties of the original model. In particular, our results were derived as large population size approximations based not only on the rescaling of time and mutation and selection parameters but also on Poisson approximations that ignored some of the subtler dependency structures in the model. Hence, one must ask if the d*N*/d*S* ratios derived here correctly capture the relation of nonsynonymous to synonymous changes over discrete generations as it evolves in the initial modeling setup. Furthermore, one would like to know whether the confidence bands in equation (31) derived with similar approximation methods reflect the true statistical variation in the original model, or not.

For the purpose of validating our analytical results, and hence providing evidence that our d*N*/d*S* ratios with reasonable accuracy indeed capture both the average behavior and the variation of nonsynonymous to synonymous divergence, we carried out a simulation study of the discrete time Wright–Fisher model with selection as introduced previously in the Materials and Methods. The code was written in Matlab (R2012b) and simulates the Markov chain $(W_n^1, \ldots, W_n^L)_{n\geq 0}$ with the following choice of parameters: $N = 500$, $L = 2000$, $\mu = (1/3) \times 10^{-6}$, $s = -2 \times 10^{-3}$, and the mutation matrix $\mathbf{H}$ chosen to be the Kimura matrix with $\alpha = 1/2$, $\beta = 1/4$. As a consequence, we have the scaled parameters $\theta = 1$ and $\gamma = -1$. The initial codon distribution was chosen to be clonal according to arbitrary codon usage. In other words, each independent component W_n^i in the codon sequence is given the initial distribution $W_0^i = (u,u,0)$ with u arbitrarily selected. Then, a single population was generated during a burn-in period of 10,000 generations (20 time units) to move toward equilibrium codon usage. The particular configuration of codons and its polymorphic states constitutes the distribution of shared ancestral polymorphisms. Then, two populations evolve in parallel but otherwise independent over 50,000 generations (100 time units). During the generation of these data, the code keeps track of the accumulated number of fixations as well as the resulting ancestral and polymorphic divergence if one sequence had been sampled from each population at that particular time. With the additional knowledge of the type of each fixed or sampled difference—nonsynonymous or synonymous—one obtains the simulated d*N*/d*S* ratio as a function over discrete time.

Supplementary Material

Supplementary text is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a European Research Council grant ERC-2009-AdG-249976-NEXTGENMOLECOL to Hans Ellegren.

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.

Balbi KJ, Feil EJ. 2007. The rise and fall of deleterious mutation. *Res Microbiol.* 158:779–786.

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

Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biol Lett.* 5:401–404.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Charlesworth D. 2010. Don't forget the ancestral polymorphisms. *Heredity* 105:509–510.

Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–1963.

Ellegren H. 2009. A selection model of molecular evolution incorporating the effective population size. *Evolution* 63:301–305.

Ellegren H, Smeds L, Burri R, et al. (12 co-authors). 2012. The genomic landscape of species divergence in Ficedula flycatchers. *Nature* 491: 756–760.

Evans SE, Shvets Y, Slatkin M. 2007. Non-equilibrium theory of the allele frequency spectrum. *Theor Popul Biol.* 71:109–119.

Gillooly JF, Allen AP, West GB, Brown JH. 2005. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc Natl Acad Sci U S A.* 102:140–145.

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

Ho SYW, Larson G. 2006. Molecular clocks: when times are a-changin'. *Trends Genet.* 22:79–83.

Karlin S, Taylor HM. 1981. A second course in stochastic processes. New York: Academic Press, Inc..

Kimura M. 1962. On probability of fixation of mutant genes in a population. *Genetics* 47:713–719.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of d*N*/d*S*. *PLoS Genet.* 4:e1000304.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* 2:150–174.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.

Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci U S A.* 77:7328–7332.

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

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418–426.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genom Hum Genet.* 22: 265–289.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet.* 39:197–218.

Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.

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

Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, Antonarakis SE. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A.* 104: 20443–20448.

Peterson GI, Masel J. 2009. Quantitative prediction of molecular clock and K(a)/K(s) at short timescales. *Mol Biol Evol.* 26:2595–2603.

Prüfer K, Munch K, Hellmann I, et al. (41 co-authors). 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of d*N*/d*S* are time dependent for closely related bacterial genomes. *J Theor Biol.* 239:226–235.

Sawyer SA, Hartl DL. 1992. Population-genetics of polymorphism and divergence. *Genetics* 132:1161–1176.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.

Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous (d(*N*)) and synonymous (d(*S*)) substitution rates affects inference of selection. *Genome Biol Evol.* 1: 308–319.

Wright S. 1931. Evolution in mendelian populations. *Genetics* 16: 97–159.

Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu Rev Ecol Evol Syst.* 39:193–213.

Wyckoff GJ, Malcolm CM, Vallender EJ, Lahn BT. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet.* 21: 381–385.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.