Polymorphism Data Assist Estimation of the Nonsynonymous over Synonymous Fixation Rate Ratio ω for Closely Related Species

Carina F. Mugal ,*[1] Verena E. Kutschera ,[1,2,3] Fidel Botero-Castro ,[4] Jochen B.W. Wolf ,[1,4] and Ingemar Kaj[5,*]

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

[2]Science for Life Laboratory, Stockholm University, Stockholm, Sweden

[3]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

[4]Division of Evolutionary Biology, Faculty of Biology, LMU Munich, Planegg-Martinsried, Germany

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

*Corresponding authors: E-mails: carina.mugal@ebc.uu.se; ikaj@math.uu.se.

Associate editor: Sergei Kosakovsky Pond

Abstract

The ratio of nonsynonymous over synonymous sequence divergence, d*N*/d*S*, is a widely used estimate of the nonsynonymous over synonymous fixation rate ratio ω, which measures the extent to which natural selection modulates protein sequence evolution. Its computation is based on a phylogenetic approach and computes sequence divergence of protein-coding DNA between species, traditionally using a single representative DNA sequence per species. This approach ignores the presence of polymorphisms and relies on the indirect assumption that new mutations fix instantaneously, an assumption which is generally violated and reasonable only for distantly related species. The violation of the underlying assumption leads to a time-dependence of sequence divergence, and biased estimates of ω in particular for closely related species, where the contribution of ancestral and lineage-specific polymorphisms to sequence divergence is substantial. We here use a time-dependent Poisson random field model to derive an analytical expression of d*N*/d*S* as a function of divergence time and sample size. We then extend our framework to the estimation of the proportion of adaptive protein evolution α. This mathematical treatment enables us to show that the joint usage of polymorphism and divergence data can assist the inference of selection for closely related species. Moreover, our analytical results provide the basis for a protocol for the estimation of ω and α for closely related species. We illustrate the performance of this protocol by studying a population data set of four corvid species, which involves the estimation of ω and α at different time-scales and for several choices of sample sizes.

Key words: **molecular evolution, codon models, d*N*/d*S*, natural selection, population genetics, Poisson random field model.**

Introduction

The extent to which selection modulates gene sequence evolution has long been a key question in many areas of evolutionary biology. One widely used measure that quantifies the intensity and direction of selection acting on protein-coding sequences is the quantity ω, which represents the ratio of the rate of fixation of nonsynonymous mutations to that of synonymous mutations. The fixation rate ratio ω is estimated in a phylogenetic setting, where the fixation rate is judged from sequence divergence between two or more species (Goldman and Yang 1994; Muse and Gaut 1994). The underlying assumption in phylogenetic models for the equality between ω and the ratio of nonsynonymous to synonymous sequence divergence, denoted as the d*N*/d*S* ratio, is that mutations are fixed instantaneously, such that polymorphism is not observable and each species can be represented by a single-stereotypic sequence. The incorporation of information on the frequency of synonymous and nonsynonymous polymorphisms within a population in a McDonald–Kreitman (MK) framework (McDonald and Kreitman 1991) permits further splitting up ω into a nonadaptive part ω_{na} and an adaptive part ω_{a} (Gossmann et al. 2010, 2012; Galtier 2016). Here, ω_{a} provides information on the rate of adaptive protein evolution and forms the basis for the commonly used measure $\alpha = \omega_{\mathrm{a}}/\omega$, which reflects the proportion of beneficial nonsynonymous substitutions. Whereas initially the estimation of α was based on the assumptions of the neutral theory of molecular evolution (Fay et al. 2001; Bierne and Eyre-Walker 2004), it is now widely recognized that weakly selected mutations significantly contribute to polymorphism and divergence (Ohta 1992; Charlesworth and Eyre-Walker 2008; Ellegren 2008; Hughes 2008). Several extensions of the MK framework therefore account for the segregation of weakly selected mutations (Keightley and Eyre-Walker 2007;

Eyre-Walker and Keightley 2009; Schneider et al. 2011; Galtier 2016; Keightley et al. 2016), and use information on the allele frequency spectrum, which describes the distribution of allele frequencies over a large number of loci, to estimate the distribution of fitness effects (DFE) of new mutations. In this setting, the DFE provides an expected value of ω_{na}, and the difference between ω and ω_{na} is attributed to ω_{a}.

Because of its alleged simplicity and intuitive appeal, the fixation rate ratio ω and derived measures have a strong tradition in evolutionary research. Genome-wide averages of estimates of ω and ω_{a} for a range of species are frequently used to investigate the strength of natural selection and the rate of adaptive evolution in relation to the effective population size, life history traits, and/or demographic history (Ellegren 2008; MacEachern et al. 2009; Nabholz et al. 2013; Weber et al. 2014; Cagan et al. 2016; Figuet et al. 2016; Galtier 2016; Settepani et al. 2016). Notably, this often involves estimation of ω between lineages with different divergence ages, in other words estimation of ω at different time-scales. Besides, at the level of genes and sites, estimates of ω allow for the identification of rapidly evolving genes and provide information on the functional importance of specific sites in proteins (Yang and Nielsen 2000, 2002; Kosakovsky Pond et al. 2011). Classical categories of rapidly evolving genes are genes involved in immune response and reproduction (Heger and Ponting 2007; Lima and McCartney 2013; Lipinska et al. 2016; Weber et al. 2017). Such rapidly evolving genes are often considered candidate genes for hybrid incompatibilities and reproductive isolation promoting speciation (Palumbi 2009; Tang and Presgraves 2009; Lessios 2011). Here, closely related species are best suited to study the mechanisms leading to reproductive isolation and ultimately speciation (Seehausen et al. 2014; Christe et al. 2017). However, estimates of ω based on phylogenetic methodology are found to be time-dependent, and in particular, biased for closely related species (Wolf et al. 2009; Mugal et al. 2014).

If a single copy of a gene sequence is compared between species, segregating polymorphisms represented in the gene copy will contribute to sequence divergence. Although this contribution is negligible if genes carry a large number of differentially fixed mutations, at short evolutionary time-scales differences between gene sequences are largely explained by segregating polymorphisms (Gagnaire et al. 2012; Hart et al. 2018). Moreover, besides segregating polymorphisms, also ancestral polymorphisms contribute to sequence divergence (Peterson and Masel 2009; Charlesworth 2010; Mugal et al. 2014). The contribution of polymorphisms (be it ancestral or segregating) becomes evident if we consider the population genetics of sequence divergence (fig. 1).

Under the assumptions of neutrality and constant population size, sequence divergence $D = 2N\mu + 2N\mu t = 2\theta(1 + t) = 2\theta T$ is proportional to population mutation rate $\theta = N\mu$ and time (reviewed in Edwards and Beerli 2000). Here, N denotes the population size of a haploid population, which under the assumption of additive fitness effects in a diploid organism, is equivalent to a diploid population of size $N/2$, and μ is the mutation rate per generation. Time is scaled in N generations of a haploid population (coalescent units), where T denotes the time of sequence divergence (or time to coalescence) and t denotes the time since the last common ancestor of the two species (or time after speciation). Thus, $2N\mu t$ represents the accumulation of sequence divergence since the last common ancestor of the two species, and $2N\mu$ is the additional sequence divergence between a pair of sequences in the ancestral population. Therefore, phylogenetically derived estimates of substitution rates (i.e., $\hat{u} = D/(2Nt) > \mu$), which are based on models that ignore polymorphisms, will be overestimated (Edwards and Beerli 2000; Phung et al. 2016). The overestimation also extends to estimates of substitution rates for sites evolving under selection. As a consequence, the ratio of nonsynonymous and synonymous sequence divergence is time-dependent (Peterson and Masel 2009; Keightley and Eyre-Walker 2012; Mugal et al. 2014). Importantly, the time-dependence is not resolved by taking the ratio of two divergence estimates, because selection impacts divergence and diversity to a different extent, that is, dN/dS is not equal to pN/pS (Hasegawa et al. 1998; Kryazhimskiy and Plotkin 2008; Mugal et al. 2014). Standard phylogenetic approaches that ignore polymorphisms should therefore not be employed for closely related species, where the contribution of polymorphisms to sequence divergence is substantial.

Keightley and Eyre-Walker (2012) summarized three methodological reasons for an overestimation of ω at narrow time-scales, 1) contribution of ancestral polymorphisms to sequence divergence, 2) misattribution of polymorphisms to sequence divergence, and 3) different rates of fixation of neutral and selected mutations. In addition, biological phenomena have been proposed to contribute to the observed rate acceleration at short evolutionary time-scales (Venditti and Pagel 2010; Ho et al. 2011), where speciation is suggested to act as an active force to promote evolution at the molecular level. However, biased estimates of ω may significantly contribute to the latter observation. In order to differentiate methodological bias from biological phenomena, it is therefore necessary to have a clear analytical understanding of the magnitude and duration of the bias in estimates of ω. Such a mathematical treatment will not only allow to judge what divergence time is large enough to apply current methodology but will further permit developing polymorphism-aware software for unbiased estimation of target parameters ω and α for closely related species. First implementations in this direction do not account for shared ancestral polymorphisms in early stages of speciation (Wilson et al. 2011; Gronau et al. 2013), or have not been designed to model codon sequence evolution (DeMaio et al. 2013). Here, an analytical framework for phylogenetic estimation of ω that considers lineage-specific and shared ancestral polymorphisms permits filling this gap.

New Approaches

To address the problem of dN/dS for closely related species analytically, we apply a time-dependent Poisson random field model of sequence divergence to study codon evolution in protein coding genes. The time-dependent Poisson random

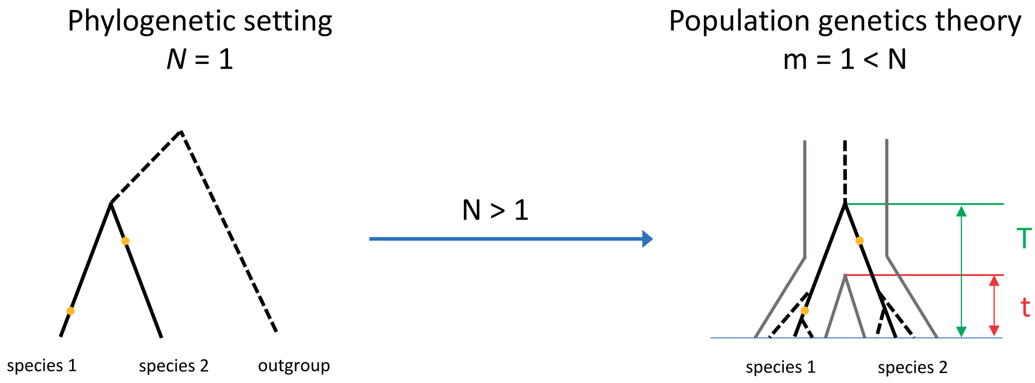

Fig. 1. Phylogenetic settings versus population genetics theory. The left panel depicts a phylogeny of two species together with an outgroup. Each species is represented by a single-stereotypic sequence. Yellow dots represent mutation events in these sequences. Estimation of substitution rates in phylogenetic settings is based on the implicit assumption that mutations are fixed instantaneously, which amounts to $N = 1$. This means that sequence divergence between two sequences, one for each species, is assumed to reflect species divergence. The right panel puts the phylogeny into the context of population genetics theory, and distinguishes between population size $N > 1$ and sample sizes $m_1 = m_2 = 1$ from the two populations. It further illustrates that time of sequence divergence $T \geq$ time of species divergence t due to the presence of ancestral polymorphisms.

field model (Amei and Sawyer 2010; Kaj and Mugal 2016) is an extension of the stationary Poisson random model, which was introduced by Sawyer and Hartl (1992) to describe the population genetics of polymorphism and divergence in distantly related species. However, instead of considering species as independent entities, the time-dependent Poisson random field model considers the fact that species pairs share polymorphisms during early stages of speciation. We use this mathematical framework to express various population functionals and sample functionals in terms of probabilities and expectations related to an underlying diffusion process. This permits us to derive analytical expressions of dN/dS as a function of time and sample size, where the dependence on sample size constitutes a key result of the present study. We then use the dependence on sample size to address the critical question if the sampling of several individuals, that is, polymorphism data, can improve estimation of ω. We further put our results in the context of the MK framework and the estimation of α. Of note, our analytical results show that estimates of ω and α are time-dependent, even though target parameters ω and α themselves are constant. In contrast, changes in demography or selection pressure over time lead to a time-dependence in target parameters ω and α, which are not addressed in the present study.

Model Formulation

Following Kaj and Mugal (2016), we consider an isolation-without-migration speciation event, where at time $t = 0$ an ancestral population of size N is instantaneously replaced by two identical copies of the population. The new branches represent two emerging species both of population size N, which initially are identical and assembled by the ancestral population. From the time of speciation and onward, the two populations evolve independently from each other with no further exchange of genetic material between them. To follow divergence between the species from the onset of speciation, we adopt the standard approach of the Wright–Fisher model with selection for two alleles segregating at one site. The derived allele is considered a new ancestral allele once it is fixed in the population, allowing for additional substitutions in the same site at a later time (Kaj and Mugal 2016). Spatiotemporal scaling is applied to finite population size N, the mutation rate per site and per generation μ, and the selection coefficient on derived alleles s. In this setting, $\theta = NL\mu$ represents the population-scaled mutation intensity for a sequence of L independent nucleotide sites per generation, and $\gamma = Ns$ represents the population-scaled selective pressure on derived alleles. The basic model for selection amounts to assigning each derived allele a fixed selective weight γ (Amei and Sawyer 2010; Kaj and Mugal 2016). We here extend the model by including variation in the strength and direction of selection across sites. Variation in the strength of selection across sites is incorporated by letting each mutation event trigger an independent outcome γ from the distribution of a random variable $\mathcal{V}$, which is commonly referred to as the DFE. Several classes of probability distributions have been suggested for the DFE, in particular, gamma and log-normal distributions for $\mathcal{V} \leq 0$ and weighted mixtures of these on the real line (Eyre-Walker et al. 2006; Loewe and Charlesworth 2006; Eyre-Walker and Keightley 2007; Welch et al. 2008). Without restriction to a specific DFE, we assume that $\mathcal{V}$ is continuous with density function $h_{\mathcal{V}}(v)$.

To study the dynamics of mutations, evolutionary time is measured on the scale of N generations. Therefore, the total mutation intensity per evolutionary time unit is $N\theta$. In the limit of large N and L at most two alleles segregate at one site, hence restricting the dynamics to biallelic states, and the Poisson random field model applies. The resulting allele frequencies of derived alleles over time are described by independent paths $t \mapsto \xi_t$ of Wright–Fisher diffusion processes with selection. For a fixed γ, using the approach in Kaj and Mugal (2016) let $\mathbb{P}_x^\gamma$ and $\mathbb{E}_x^\gamma$ be the law and expectation of the Wright–Fisher diffusion process $(\xi_t)_{t\geq 0}$ with initial frequency $\xi_0 = x$, which solves the stochastic differential equation:

$$d\xi_t = \gamma\xi_t(1-\xi_t)dt + \sqrt{\xi_t(1-\xi_t)}dB_t$$

driven by a Brownian motion B_t. Let τ_0 be the extinction time of a derived allele, τ_1 the fixation time, and $\tau = \min(\tau_0, \tau_1)$ the absorption time. The probability of fixation, $q_\gamma(x)$, is known to be (Kimura 1962):

$$q_\gamma(x) = \mathbb{P}_x^\gamma(\tau_1 < \infty) = \frac{1-e^{-2\gamma x}}{1-e^{-2\gamma}}, \quad \gamma \neq 0, \quad q_0(x) = x.$$

In the limit of large N, the result of applying mutation rate $N\theta$ and fixation probability $q_\gamma(1/N)$ is the scaled effective fixation rate $\theta\omega_\gamma$, where

$$\omega_\gamma = \lim_{N\to\infty} Nq_\gamma(1/N) = \frac{2\gamma}{1-e^{-2\gamma}}, \quad \gamma \neq 0, \quad \omega_0 = 1. \quad (1)$$

Here, ω_γ represents the fixation rate ratio for a site evolving with selective pressure γ versus a neutrally evolving site. For variation in selection across sites, the scaled, effective fixation rate needs to be further averaged over the relevant distribution of $\mathcal{V}$,

$$\mathbb{E}[\omega_\mathcal{V}] = \int_{-\infty}^{\infty} \omega_v h_\mathcal{V}(v)dv.$$

Various other population functionals and sample functionals are conveniently expressed in terms of probabilities and expectations related to the diffusion process, such as the stationary allele frequency spectrum $\omega_\gamma\pi_\gamma(y)$, $0 < y < 1$, which arises in the limit relation:

$$\lim_{N\to\infty} N\mathbb{E}_{1/N}^\gamma \int_0^\tau f(\xi_t)dt = \omega_\gamma \int_0^1 f(y)\pi_\gamma(y)dy,$$

where f is a suitable function with $f(0) = 0$ and

$$\pi_\gamma(y) = \frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}, \quad \gamma \neq 0, \quad \pi_0(y) = \frac{2}{y}.$$

Again, for variation in selection across sites, the limiting functionals are to be averaged over $\mathcal{V}$. In particular, the allele frequency spectrum becomes:

$$\int_0^1 f(y)\mathbb{E}[\omega_\mathcal{V}\pi_\mathcal{V}(y)]dy.$$

To handle time-dependent functionals in the next section, we will also need the probability distribution $\mathbb{P}_x^\gamma(\tau_1 < t) = \mathbb{P}_x^\gamma(\xi_t = 1)$, $t \geq 0$, of the fixation time τ_1, and the conditional probability distribution $\mathbb{P}_x^{*\gamma}(\tau_1 < t) = \mathbb{P}_x^\gamma(\tau_1 < t)/q_\gamma(x)$, $t \geq 0$, given that the fixation time τ_1 is finite.

The Number of Fixed Differences between Two Species

Suppose we have samples of size m_1 and m_2 from two species at time t after speciation. Put,

$$Q_t^{k_1,k_2} = \text{average number of sites with exactly } k_1 \text{ derived alleles in species 1 and } k_2 \text{ derived alleles in species 2.}$$

To analyze $Q_t^{k_1,k_2}$ as a function of time, we distinguish between the ancestral contribution $H_t^{k_1,k_2}$, representing the average number of derived alleles at time t caused by mutations segregating in the ancestral population, and the lineage-specific contributions Z_t^k, representing the average number of derived alleles at time t acquired through lineage-specific mutations for each species (Wakeley and Hey 1997; Chen 2012; Kaj and Mugal 2016). Each site with at least one derived allele present in both populations must arise from an ancestral polymorphism, that is, $Q_t^{k_1,k_2} = H_t^{k_1,k_2}$ whenever $k_1 \geq 1$, $k_2 \geq 1$, whereas sites with derived alleles in one but not the other population carry either ancestral- or lineage-specific mutations,

$$Q_t^{k_1,0} = H_t^{k_1,0} + Z_t^{k_1},\ k_1 \geq 1, \quad Q_t^{0,k_2} = H_t^{0,k_2} + Z_t^{k_2},\ k_2 \geq 1.$$

These and similar quantities are discussed in Amei and Sawyer (2010) and Kaj and Mugal (2016).

In this work, we use the functionals $Q_t^{k_1,k_2}$ to derive explicit solutions of the divergence between species as a function of time after speciation. First of all, the expected total number of pair-wise fixed differences in the sample is given by,

$$D_{\text{pw}}^{m_1,m_2}(t) = Q_t^{m_1,0} + Q_t^{0,m_2} = H_t^{m_1,0} + H_t^{0,m_2} + Z_t^{m_1} + Z_t^{m_2}.$$

In order to get branch-specific estimates of divergence, we consider one species labeled P_1 (sample size m_1) as the primary species of interest, whereas the second species, labeled P_2 (sample size m_2), is used for comparison only. Some additional information is available from an outgroup species, a third species which branched off the ancestral line well before creation of the two species P_1 and P_2. The divergence restricted to the focal species P_1, which we denote $D_\gamma^{m_1,m_2}(t)$, now splits into lineage-specific and ancestral parts, as

$$D_\gamma^{m_1,m_2}(t) = Z_t^{m_1} + H_t^{m_1,0},$$

where the second term also depends on m_2 through

$$H_t^{m_1,0} = \theta\omega_\gamma \int_0^1 \mathbb{E}_y^\gamma[\xi_t^{m_1}]\mathbb{E}_y^\gamma[(1-\xi_t)^{m_2}]\pi_\gamma(y)dy.$$

The first term $Z_t^{m_1}$ only depends on m_1 and is given by,

$$Z_t^{m_1} = \theta\omega_\gamma t - \theta\omega_\gamma \int_0^1 (q_\gamma(y) - y^{m_1})\mathbb{P}_{1-y}^{*\gamma}(\tau_1 \leq t)\pi_\gamma(y)dy. \quad (2)$$

In summary, for a fixed γ and sample size m_1 and m_2 in the two populations, the divergence in population P_1 is given by,

$$D_\gamma^{m_1,m_2}(t) = \theta\omega_\gamma t - \theta\omega_\gamma \int_0^1 (q_\gamma(y) - y^{m_1})\mathbb{P}_{1-y}^{*\gamma}(\tau_1 \leq t)\pi_\gamma(y)dy + \theta\omega_\gamma \int_0^1 \mathbb{E}_y^\gamma[\xi_t^{m_1}]\mathbb{E}_y^\gamma[(1-\xi_t)^{m_2}]\pi_\gamma(y)dy. \quad (3)$$

The functions $\mathbb{P}_x^{*\gamma}(\tau_1 \leq t)$, $\mathbb{E}_x^\gamma[\xi_t^{m_i}]$, and $\mathbb{E}_x^\gamma[(1-\xi_t)^{m_i}]$ appearing in these formulas are explicitly known for the neutral case $\gamma = 0$. The neutral divergence is:

$$D_0^{m_1,m_2}(t) = \theta t - \theta \int_0^1 (y - y^{m_1})\mathbb{P}_{1-y}^{*0}(\tau_1 \leq t)\frac{2}{y}dy + \theta \int_0^1 \mathbb{E}_y^0[\xi_t^{m_1}]\mathbb{E}_y^0[(1-\xi_t)^{m_2}]\frac{2}{y}dy.$$

Supplementary appendix A in Supplementary Material online provides a method for convenient computation of

$D_0^{m_1,m_2}(t)$ and a scheme for approximating $D_\gamma^{m_1,m_2}(t)$ in the case $\gamma \neq 0$.

As an alternative, the divergence may be split into linear and nonlinear parts,

$$D_\gamma^{m_1,m_2}(t) = \theta\omega_\gamma t + \theta F_\gamma^{m_1,m_2}(t),$$

with $F_\gamma^{m_1,m_2}(t)$ representing the affine and nonlinear contributions of lineage-specific and ancestral polymorphisms, given by

$$F_\gamma^{m_1,m_2}(t) = -\omega_\gamma \int_0^1 (q_\gamma(y) - y^{m_1})\mathbb{P}_{1-y}^{*\gamma}(\tau_1 \leq t)\pi_\gamma(y)\mathrm{d}y \qquad (4)$$

$$+\, \omega_\gamma \int_0^1 \mathbb{E}_y^\gamma[\xi_t^{m_1}]\mathbb{E}_y^\gamma[(1-\xi_t)^{m_2}]\pi_\gamma(y)\mathrm{d}y. \qquad (5)$$

Here, the dependence of $F_\gamma^{m_1,m_2}(t)$ on the strength of selection illustrates that contributions of polymorphisms to estimates of divergence differ between neutrally and selected mutations. As before, our representation of divergence, $D_\gamma^{m_1,m_2}(t)$, extends to the case where the strength of selection varies across sites, simply by averaging with respect to the relevant DFE. The overall divergence becomes:

$$D_{\text{DFE}}^{m_1,m_2}(t) = \mathbb{E}[D_\mathcal{V}^{m_1,m_2}(t)] = \theta\mathbb{E}[\omega_\mathcal{V}]t + \theta\mathbb{E}[F_\mathcal{V}^{m_1,m_2}(t)]. \qquad (6)$$

Results and Discussion

Time-Dependence of the Sequence Divergence Rate Ratio d*N*/d*S*

We apply the time-dependent Poisson random field model of sequence divergence to the case of codon evolution in protein coding genes. In this context, each individual is represented by a sequence of L codons, and mutations in the codon sequence are distinguished into either synonymous or nonsynonymous mutations. Following Mugal et al. (2014), we let θ_{syn} denote the synonymous mutation intensity and θ_{non} the nonsynonymous mutation intensity. We apply the common assumption that all synonymous mutations evolve neutrally, while nonsynonymous mutations are subject to natural selection. The extent of selection on nonsynonymous mutations is given by a DFE denoted $\mathcal{V}$. This modeling framework allows us to obtain an analytical description of d*N*/d*S* between two species P_1 and P_2 as a function of divergence time t, where species P_1 and P_2 are represented by a sample of m_1 and m_2 individuals, respectively. Note that in classical phylogenetic approaches the computation of d*N*/d*S* is based on a single protein coding sequence for each species, that is, $m_1 = m_2 = 1$. Since ultimately we are interested to investigate if the sampling of several individuals and the resulting information of allele frequencies of polymorphic sites can assist the estimation of the target parameter ω, we here introduce a more general setting. Consequently, the observed coding sequence (CDS) divergence in species P_1 after its divergence from species P_2 further depends on m_1 and m_2, and in expectation builds up over time according to:

$$D_{\text{CDS}}^{m_1,m_2}(t) = \theta_{\text{syn}}\mathrm{d}S_t + \theta_{\text{non}}\mathrm{d}N_t,$$

where $\mathrm{d}S_t = \mathrm{d}S_t^{m_1,m_2}$ is the scaled synonymous sequence divergence

$$\mathrm{d}S_t = t + F_0^{m_1,m_2}(t),$$

and $\mathrm{d}N_t = \mathrm{d}N_t^{m_1,m_2}$ the scaled nonsynonymous sequence divergence

$$\mathrm{d}N_t = \mathbb{E}[\omega_\mathcal{V}]t + \mathbb{E}[F_\mathcal{V}^{m_1,m_2}(t)].$$

Here, the terms linear in t represent the long-term build-up of lineage-specific fixations that eventually attain nonsynonymous and synonymous linear growth rates $\mathbb{E}[\omega_\mathcal{V}]$ and 1, respectively. The terms $\mathbb{E}[F_\mathcal{V}^{m_1,m_2}(t)]$ and $F_0^{m_1,m_2}(t)$ are introduced in equations (4 and 5) and take into account additional contributions, which are pronounced close to speciation, but fade out as $t \to \infty$.

The branch-specific d*N*/d*S* ratio for species P_1 is now defined as the ratio of the scaled expected synonymous and nonsynonymous divergences,

$$\mathrm{d}N/\mathrm{d}S|_t = \frac{\mathrm{d}N_t}{\mathrm{d}S_t} = \frac{\mathbb{E}[\omega_\mathcal{V}]t + \mathbb{E}[F_\mathcal{V}^{m_1,m_2}(t)]}{t + F_0^{m_1,m_2}(t)}. \qquad (7)$$

Thus, our main results regarding the d*N*/d*S* ratio are based on the analytical representation (eq. 3). The use of equation (3) and the accompanying averaging over $\mathcal{V}$ in equation (6) clarify the shape of $\mathrm{d}N/\mathrm{d}S|_t$ as a function of divergence time, starting from the initial value at $t = 0$,

$$\mathrm{d}N/\mathrm{d}S|_{t=0} = \frac{\int_0^1 y^{m_1}(1-y)^{m_2}\mathbb{E}[\omega_\mathcal{V}\pi_\mathcal{V}(y)]\mathrm{d}y}{\int_0^1 y^{m_1}(1-y)^{m_2}\pi_0(y)\mathrm{d}y},$$

until the limiting d*N*/d*S* ratio as $t \to \infty$, which is

$$\omega = \lim_{t\to\infty} \mathrm{d}N/\mathrm{d}S|_t = \mathbb{E}[\omega_\mathcal{V}] \qquad (8)$$

in accordance with the classical definition of d*N*/d*S* (described in Mugal et al. 2014). We notice that for the case of negative selection, $\mathcal{V} \leq 0$ and $\omega \leq 1$, then $\mathrm{d}N/\mathrm{d}S|_0 \geq \omega$, see equation (16) in supplementary appendix A in Supplementary Material online. Here, $\mathrm{d}N/\mathrm{d}S|_0$ for $m_1 = m_2 = 1$ represents the ratio of nonsynonymous over synonymous heterozygosity sampled in the ancestral population. Inequality $\mathrm{d}N/\mathrm{d}S|_0 \geq \omega$ therefore reflects that slightly deleterious nonsynonymous mutations are more likely to be observed as heterozygous sites than to get fixed in the population.

To enable simple usage of equation (7), we have derived an exact, closed form expression for $\mathrm{d}S_t$ as a function of sample size m_1 and m_2, which is straightforward to compute in terms of known probabilities for Kingman's coalescent process, see equation (20) in supplementary appendix A in Supplementary Material online. Second, we provide an approximation scheme in a series of steps which leads to a computational formula for $F_\mathcal{V}^{m_1,m_2}(t)$, see equation (25) in supplementary appendix A in Supplementary Material online. For a given distribution of $\mathcal{V}$, these results allow us to derive $\mathrm{d}N_t$ and hence $\mathrm{d}N/\mathrm{d}S|_t$ in equation (7) as a function of time and sample size.

As an additional result, we obtain the d*N*/d*S* ratio explicitly as a function of time in the limit of large sample size (supplementary appendix B in Supplementary Material online).

In particular, letting $m_1 \to \infty$ for fixed $m_2 = 1$, the large sample limit function $\mathrm{d}N/\mathrm{d}S|_t^{\infty}$ for $t > 0$ is:

$$\mathrm{d}N/\mathrm{d}S|_t^{\infty} = \frac{\mathbb{E}[\omega_{\mathcal{V}}(t - \int_0^1 \mathbb{P}_y^{\mathcal{V}}(\tau_1 < t)\mathbb{E}_y^{\mathcal{V}}(\xi_t)\pi_{\mathcal{V}}(y)\mathrm{d}y)]}{t - \int_0^1 \mathbb{P}_y^0(\tau_1 < t)2\mathrm{d}y}. \tag{9}$$

In the absence of advantageous nonsynonymous mutations, that is, $\mathcal{V} \leq 0$, then $\mathrm{d}N/\mathrm{d}S|_t^{\infty} \geq \omega$ (supplementary appendix A in Supplementary Material online). The expression (eq. 9) becomes computationally feasible if we apply the approximation:

$$\mathbb{P}_y^{\gamma}(\tau_1 < t) = \mathbb{P}_y^{*\gamma}(\tau_1 < t)q_{\gamma}(y) \approx \mathbb{P}_y^{*0}(\tau_1 < t)q_{\gamma}(y),$$

and use an explicit summation formula for the neutral conditional probability $\mathbb{P}_y^{*0}(\tau_1 < t)$, see equation (7) in Kaj and Mugal (2016). The limit function in equation (9) captures the intrinsic time-dependence of the dN/dS ratio and shows that the time-dependent effect due to the contribution of polymorphisms remains no matter how many individuals are sampled. Thus, $\mathrm{d}N/\mathrm{d}S|_t$ does not converge to ω for large sample sizes as long as t is fixed. By taking $t \to \infty$ in equation (9), we recover again the limit ω. A similar but slightly different expression results if we take both sample sizes large, $m_1, m_2 \to \infty$. A further interpretation arises by rewriting equation (9) in the form:

$$\omega = \mathrm{d}N/\mathrm{d}S|_t^{\infty}(1 - \frac{1}{t}B_t) + \frac{1}{t}C_t \tag{10}$$

with

$$B_t = \int_0^1 \mathbb{P}_y^0(\tau_1 < t)2\mathrm{d}y,$$

$$C_t = \mathbb{E}[\int_0^1 \omega_{\mathcal{V}}\mathbb{P}_y^{\mathcal{V}}(\tau_1 < t)\mathbb{E}_y^{\mathcal{V}}(\xi_t)\pi_{\mathcal{V}}(y)\mathrm{d}y].$$

The relation (eq. 10) shows that the fixation rate ratio ω is equal to the present ratio $\mathrm{d}N/\mathrm{d}S|_t^{\infty}$ adjusted by means of the neutral factor $1 - B_t/t$ and the selective update term C_t/t. Here, the dependence of the neutral and selective update term on $\mathbb{P}_y^0(\tau_1 < t)$ and $\mathbb{P}_y^{\gamma}(\tau_1 < t)$, respectively, clearly illustrates that the intrinsic time-dependence of the dN/dS ratio can be explained by differences in the time to fixation between neutrally and selected mutations. Besides, equation (10) can provide a basis for future method development for point estimation of ω.

Polymorphism Data Assist Estimation of ω

We are interested to investigate if the joint usage of polymorphism and divergence data can improve the estimation of the fixation rate ratio ω for closely related species. More precisely, the task is point estimation of $\omega = \mathbb{E}[\omega_{\mathcal{V}}]$ for a pair of species with fixed (but unknown) divergence time t using the observed CDS divergence in the sample. As evident in equation (10), the matching of ω with observed divergence data $\mathrm{d}N/\mathrm{d}S|_t$ or its large sample version $\mathrm{d}N/\mathrm{d}S|_t^{\infty}$ remains open as long as the divergence time t is unknown. To help understand the bias, we study the time-dependent function $\mathrm{d}N/\mathrm{d}S|_t$ for different choices of sample sizes, where we distinguish between 1) a symmetric and 2) an asymmetric sampling scenario (supplementary appendix C in Supplementary Material online). Here, symmetric sampling means that $m_1 = m_2 > 1$, in other words polymorphism data are available for both species. Asymmetric sampling means that $m_1 > m_2 = 1$, in other words polymorphism data are only available for the focal species P_1, for which we want to estimate ω, but not for the reference species. We first investigate the case of absence of advantageous nonsynonymous mutations, by assuming that solely purifying selection acts on nonsynonymous mutations as described by a distribution of $\mathcal{V}$ on the negative real line. A common choice of the DFE in this situation is the gamma-distribution (Eyre-Walker et al. 2006; Loewe and Charlesworth 2006; Charlesworth and Eyre-Walker 2008), such that the density of $\mathcal{V}$ is:

$$h_{\mathcal{V}}(v) = \frac{(-v)^{a-1}e^{v/b}}{b^a\Gamma(a)}, \quad v \leq 0, \tag{11}$$

where $a > 0$ is the shape parameter and $b > 0$ the scale parameter, hence the expected value $E(\mathcal{V}) = -ab$. Values of $\mathcal{V}$ near 0 represent mutations in the weak selection regime, which will segregate as polymorphisms until they either reach fixation or are removed from the population. Large negative values of $\mathcal{V}$ below some value $v_- < 0$, on the other hand, represent a fraction $P(\mathcal{V} \leq v_-)$ of strongly deleterious mutations, which in practice are removed instantaneously (Welch 2006; Eyre-Walker and Keightley 2007).

Figure 2 illustrates the behavior of $\mathrm{d}N/\mathrm{d}S|_t$ as a function of time and sample size for the symmetric and the asymmetric sampling scenario for a DFE on nonsynonymous mutations estimated for the EGP African human data set, $a = 0.15$, $ab =$ 2,500, (Keightley and Eyre-Walker 2007). For both scenarios and all sample sizes, the asymptotic ratio $\omega = \mathbb{E}[\omega_{\mathcal{V}}]$ is overestimated by $\mathrm{d}N/\mathrm{d}S|_t$ at any fixed time t. The bias is larger for the smaller sample sizes and varies over time until it fades out as $t \to \infty$. The shape of the curves in figure 2 with a maximum in the vicinity of $t = 2$ reflects the contribution of segregating polymorphisms to estimates of sequence divergence. The maximum in the vicinity of $t = 2$ is especially transparent in small samples and is caused partly by slightly deleterious mutations segregating at low-frequency. The larger the sample size, the smaller the probability to observe such low-frequency derived alleles as fixed differences between samples. Hence, the estimation of $\mathbb{E}[\omega_{\mathcal{V}}]$ using $\mathrm{d}N/\mathrm{d}S|_t$ at some (unknown) t improves with increasing sample size. As we have seen previously in equation (9), however, the relevant large sample approximation is the function $\mathrm{d}N/\mathrm{d}S|_t^{\infty} \geq \mathbb{E}[\omega_{\mathcal{V}}]$ (blue curve in fig. 2B), and not the constant function $\mathbb{E}[\omega_{\mathcal{V}}]$.

The inequality $\mathrm{d}N/\mathrm{d}S|_t^{\infty} \geq \mathbb{E}[\omega_{\mathcal{V}}]$ is related to the fact that polymorphisms of the pair of species are not independent of each other, in particular, for small t. Instead, a large proportion of segregating polymorphisms is amounted to ancestral polymorphisms present in the common ancestor of the two species if divergence time t is small. Figure 3A illustrates the proportion of ancestral polymorphisms to

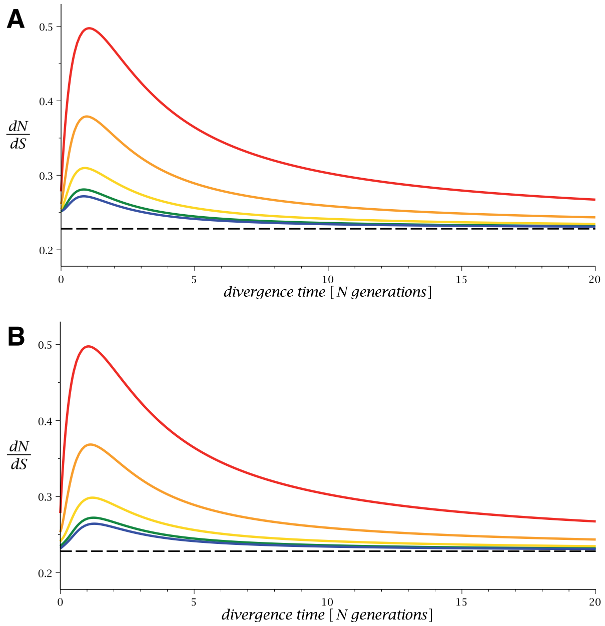

FIG. 2. dN/dS as a function of divergence time for a DFE based on human data. (A) Shows a symmetric ($m_1 = m_2$) and (B) an asymmetric sampling scenario $m_2 = 1$, for sample sizes $m_1 = 1$ (red), $m_1 = 2$ (orange), $m_1 = 4$ (gold), $m_1 = 8$ (green), and $m_1 = 16$ representative for the large sample approximation (blue). The black dashed line indicates the target parameter ω.

the total number of segregating sites over time under neutral evolution. Here, we quote a result that under neutral evolution the limiting expected number of segregating sites at time t after speciation in a sample of size m equals:

$$S_m(t) = \theta\int_0^t(\mathbb{E}_m^0(A_u) - \mathbb{P}_m^0(A_u = 1))\mathrm{d}u,$$

with stationary limit $S_m = 2\theta\sum_{k=1}^{m-1}\frac{1}{k}$ as $t \longrightarrow \infty$, see Kaj and Mugal (2016), in particular, equation (21). It follows that the function P_{AP} defined by:

$$P_{\mathrm{AP}}(t) = 1 - \frac{S_m(t)}{S_m}, \quad t \geq 0,$$

measures the proportion of ancestral polymorphisms at time t in the sample. The dynamics over time with $P_{\mathrm{AP}}(0) = 1$ at the time of speciation and convergence toward 0 as $t \longrightarrow \infty$ is fairly insensitive to sample size. For the larger sample sizes, $<10\%$ of ancestral polymorphisms remain at time $t = 2$, $P_{\mathrm{AP}}(2) = 0.095$, and $<1\%$ at time $t = 5$, $P_{\mathrm{AP}}(5) = 0.005$. The influence of ancestral polymorphisms on $\mathrm{d}N/\mathrm{d}S|_t^\infty$ wears off at a similar rate and $\mathrm{d}N/\mathrm{d}S|_t^\infty$ approaches $\mathbb{E}[\omega_{\mathcal{V}}]$ as P_{AP} approaches 0, that is, as lineage sorting is complete. Roughly, we can distinguish two cases, distantly related species, for which ancestral polymorphisms are completely sorted (commonly referred to as complete lineage sorting; fig. 3B), and closely related species, for which ancestral polymorphisms contribute to segregating polymorphisms (incomplete lineage sorting; fig. 3C). In case of complete lineage sorting (and taking $m_2 = 1$), the contribution of ancestral polymorphisms to sequence divergence is quantified by the limiting constant in the following relation:

$$H_t^{m_1,0} \longrightarrow \theta\mathbb{E}[\omega_{\mathcal{V}}\int_0^1 q_{\mathcal{V}}(y)(1 - q_{\mathcal{V}}(y))\pi_{\mathcal{V}}(y)\mathrm{d}y], \quad t \longrightarrow \infty,$$

which scales with population size. The limit is the same if we take first $m_1 \longrightarrow \infty$ and then $t \longrightarrow \infty$. However, in case of incomplete lineage sorting, the misattribution of polymorphisms and the contribution of ancestral polymorphisms to sequence divergence are interrelated and the resulting overestimation is thus more complex and time-dependent.

In practice, this suggests that for sufficiently large divergence time where P_{AP} is small, polymorphism data allow to improve estimates of $\mathbb{E}[\omega_{\mathcal{V}}]$. The improvements perform similarly well for both sampling scenarios discussed in figure 2, which suggests that the improvement depends primarily on the sample size of the focal species P_1. In other words, it seems sufficient to include polymorphism data only for the focal species. Polymorphism data for reference species seem not to be necessary. Moreover, our results have emphasized that $\mathrm{d}N/\mathrm{d}S|_t$ depends strongly on sample size for small t, while this dependence fades out as $t \longrightarrow \infty$. The dependence on sample size can therefore be used as an indicator for estimation accuracy, when polymorphism data are used to make inference on ω. Strong dependence on sample size, which even shows for sample sizes $m \geq 2$, suggests that divergence time is relatively recent and resulting estimates of ω should be interpreted with caution. When in doubt about completion of lineage sorting between the focal and reference species, it seems therefore advisable to estimate and compare dN/dS for different choices of sample sizes.

Next, we study the behavior of dN/dS in the presence of advantageous mutations. To incorporate such a distribution into our settings, we consider two different models. First, following earlier approaches (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009), we assume that advantageous alleles do not contribute to polymorphisms but instead are immediately fixed upon arrival. In this setting, strongly advantageous mutations arrive at rate $c^+\mu$ per site. All remaining mutations, of total intensity $\theta(1 - c^+/N) \sim \theta$, are represented as before by a distribution of $\mathcal{V}$ on the negative real line. As a result, the scaled nonsynonymous sequence divergence is replaced by $\mathrm{d}N_t = c^+t + \mathbb{E}[\omega_{\mathcal{V}}]t + \mathbb{E}[F_{\mathcal{V}}^{m_1,m_2}(t)]$, and the limiting d$N$/d$S$ ratio is replaced by:

$$\omega = \mathbb{E}[\omega_{\mathcal{V}}] + c^+,$$

where c^+ represents the rate of adaptive evolution. Second, we relax the assumption of instantaneous fixation of advantageous mutations and instead let $\mathcal{V}$ be a continuous random variable on the real line. A common choice for the DFE is to represent deleterious mutations by a gamma distribution on the negative real line, and advantageous mutations by an exponential distribution (Galtier 2016; Tataru et al. 2017). This model requires two additional parameters, the mean of the exponential distribution c, and the frequency of

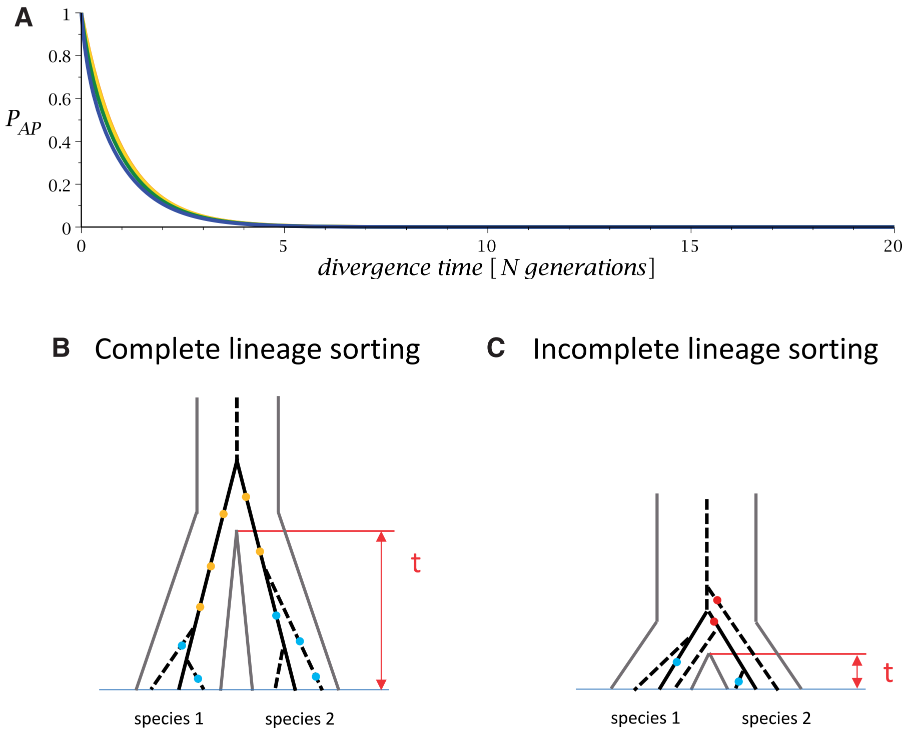

Fig. 3. (*A*) The proportion of ancestral polymorphisms P_{AP} to the total number of segregating sites over time in a sample of $m = 2$ (orange), $m = 4$ (gold), $m = 8$ (green), $m = \infty$ (blue). (*B*) Illustration of complete lineage sorting. Each species is represented by three individuals. Dots denote mutations accumulating over time. Yellow dots indicate mutations that represent fixed differences between the two species. Blue dots indicate lineage-specific polymorphisms. (*C*) Illustration of incomplete lineage sorting. Blue dots indicate lineage-specific polymorphisms. Red dots indicate ancestral polymorphisms that contribute to segregating polymorphisms at present.

advantageous mutations *p*. The limiting d*N*/d*S* ratio is represented by:

$$\omega = \mathbb{E}[\omega_{\mathcal{V}}] = (1-p)\mathbb{E}[\omega_{\mathcal{V}}|\mathcal{V} \le 0] + p\mathbb{E}[\omega_{\mathcal{V}}|\mathcal{V} > 0].$$

Figure 4 illustrates the behavior of d*N*/d*S* as a function of time and sample size for the same human DFE on deleterious nonsynonymous mutations as above, but now in the presence of advantageous nonsynonymous mutations. Given the similarity in performance of both sampling scenarios, we here and in the following only focus on the asymmetric sampling scenario. Figure 4*A* shows the behavior of d*N*/d*S* under the first model and $c^{+} = 0.10$. Figure 4*B* shows the behavior of d*N*/d*S* under the second model, $c = 0.2$ and $p = 0.1$. Parameters were chosen such that $\omega \approx 0.33$ for both models. Although we still observe a strong dependence on sample size for small *t*, under the assumption of instantaneous fixation of advantageous mutations (fig. 4*A*), we also observe a strong effect of the parameter c^{+} on the d*N*/d*S* ratio in large samples and for small *t* in the range $t < 2$. This reflects the fact that at very narrow divergence primarily advantageous mutations will have had the time to sweep to fixation, while neutral and slightly deleterious mutations will need more time until eventual fixation (Maynard Smith and Haigh 1974; Kaplan et al. 1989). This observation confirms the claim by Keightley and Eyre-Walker (2012), that different rates of fixation of neutral and selected mutations lead to an overestimation of d*N*/d*S* at narrow time-scales. For smaller sample sizes, segregating polymorphisms attenuate the impact of the parameter c^{+}, and overestimation is less pronounced and $\mathrm{d}N/\mathrm{d}S < \omega$ close to 0. Thus, the accuracy of the estimates no longer systematically improves with increasing sample size. However, despite the presence of strongly advantageous nonsynonymous mutations, $\mathrm{d}N/\mathrm{d}S|_t^\infty$ converges toward ω as P_{AP} converges toward 0. Moreover, the dependence on sample size remains for small *t*, and thus serves as an indicator for the accuracy of estimates of ω. In addition, under the scenario that mutations are mostly weakly advantageous and do not fix instantaneously (fig. 4*B*), we again observe that the accuracy of the estimates systematically improves with increasing sample size. We therefore conclude that d*N*/d*S* is overestimated at narrow time-scales due to 1) contribution of (lineage-specific and ancestral) polymorphisms to sequence divergence, and 2) different rates of fixation of neutral and selected mutations. Incorporation of polymorphism data allows to improve (1), but does not allow to improve the overestimation caused by (2). To address (2), the dependence on sample size serves as an indicator for the accuracy of estimates of ω, regardless of the model for advantageous mutations.

For cases where the majority of nonsynonymous substitutions are advantageous and $\omega > 1$, the contribution of polymorphisms to sequence divergence and different rates of fixation of neutral and selected mutations, lead to an underestimation of ω at narrow time-scales

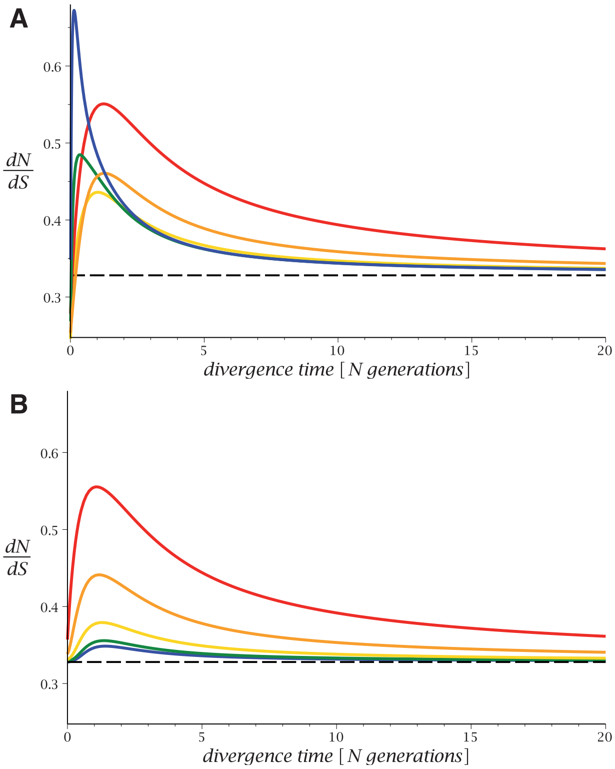

FIG. 4. dN/dS as a function of divergence time for a DFE on the negative real line based on human data and additional advantageous mutations with (A) a model assuming instantaneous fixation of advantageous mutations and $c^+ = 0.10$ and (B) a fraction of $p = 0.10$ advantageous mutations modeled by an exponential distribution with a mean of $c = 0.20$. Sample sizes are $m_1 = 1$ (red), $m_1 = 2$ (orange), $m_1 = 4$ (gold), $m_1 = 8$ (green), $m_1 = 16$ representative for the large sample approximation (blue), and $m_2 = 1$ (asymmetric case). The black dashed line indicates the target parameter ω.

(supplementary fig. 1, Supplementary Material online). Although such scenarios occasionally happen for gene-by-gene estimates (small L), they are not common on a genome-wide scale (large L). However, estimates of ω for a single gene (small L) are in addition to the time-dependence further affected by the large stochasticity of a ratio of small numbers, addressed in Mugal et al. (2014). We therefore here do not discuss in detail extreme scenarios where the majority of nonsynonymous substitutions are advantageous and $\omega > 1$, but instead focus on scenarios common for genome-wide estimates of ω.

The Impact of Linkage among Sites on the Time-Dependence of dN/dS

Like most phylogenetic methodology, the basic Poisson random field model assumes free recombination among sites. Physical linkage between sites of genome-wide data violates this assumption leading to correlated ancestral histories between neighboring sites and a reduction of local N_e. Although reduction in N_e does not affect the fixation rate of neutrally evolving sites itself, estimates of divergence will be affected by linkage, in particular, at short evolutionary time-scales (Phung et al. 2016). Moreover, interference between selected sites (Hill–Robertson interference) has been shown to affect the efficacy of selection (Hill and Robertson 1966), which further has been suggested to affect dN/dS ratios (Campos et al. 2012). To address the impact of linkage between sites, we performed forward simulations based on the SLiM software package (Haller and Messer 2019).

SLiM is based upon a Wright–Fisher model, that is, generations are nonoverlapping and discrete, the probability of an individual being chosen as a parent for a child in the next generation is proportional to the individual's fitness, individuals are diploid, and offspring are generated by recombination of parental chromosomes with the addition of new mutations. We first implemented simulations based on free recombination in order to explore the agreement between our analytical results which provides expected values based on a Poisson random field model, and stochastic simulations based on assumptions specified within SLiM. To adjust stochastic simulations to meet the assumptions of our Poisson random field model, which treats sequences as a collection of independent sites, we specified a recombination rate of $r = 0.5$ between neighboring sites and further implemented additive fitness effects. These adjustments lead to a model which in the limit is equivalent to our Poisson random field model. Figure 5A shows results for human DFE parameters in the absence of advantageous mutations, which are in good agreement with analytical results (fig. 2B). We observe an overestimation of ω at narrow time-scales, which wears off for larger t, and is less pronounced for larger sample sizes. In addition, we observe that stochasticity influences dN/dS in particular at narrow time-scales, which is expected given that the number of fixed differences is low between closely related species and is another reason for caution in interpreting estimates of ω at narrow time-scales (Mugal et al. 2014). Figure 5B shows results for the same DFE parameters, but in the presence of linkage. As expected, Hill–Robertson interference reduces the efficacy of selection and elevates the asymptotic limit of ω above its expectation under free recombination. Nevertheless, the time-dependence and sample size-dependence are preserved even in the presence of linkage. Results in the presence of advantageous mutations are provided in supplementary figure 2, Supplementary Material online, and are again in good agreement with analytical results. This suggests that our analytical results are robust to violation of the assumption of free recombination and relevant for biologically plausible parameters.

Analytical Results on the Rate of Adaptive Evolution

Given that dN/dS depends on time and sample size, naturally all measures of the mode and strength of selection that rely on phylogenetic estimates of ω will also show this dependence. One such measure is the rate of adaptive evolution ω_a, which measures the asymptotic contribution of advantageous mutations to ω. Starting from equation (8), ω is split up into a nonadaptive and an adaptive part, $\omega_{na} + \omega_a = \omega$. For the case of instantaneous fixation of advantageous mutations and the restriction of $\mathcal{V} \leq 0$, this is:

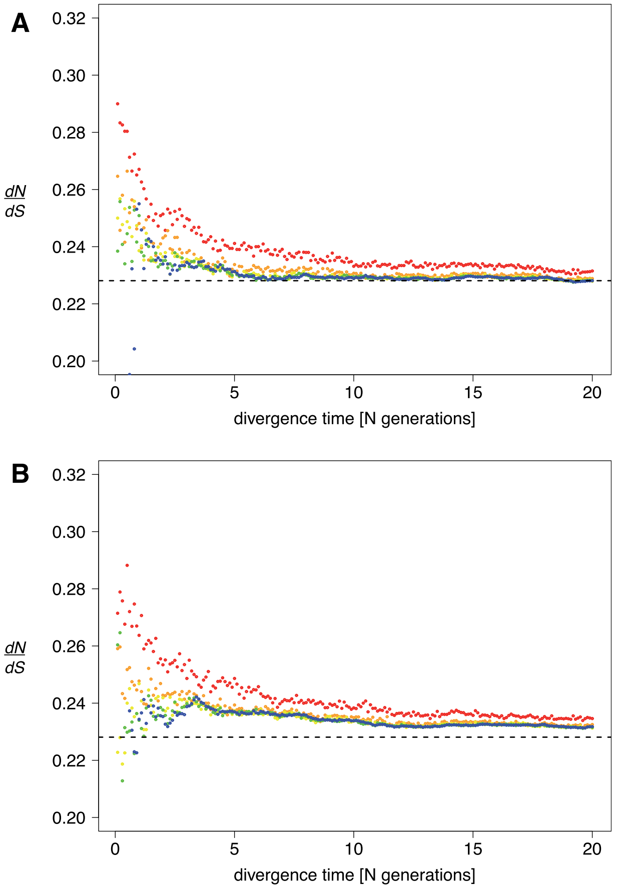

Fig. 5. Mean dN/dS as a function of divergence time for different choices of sample size based on individual-based forward simulations for a DFE based on human data. Mean dN/dS is reported every 0.1N generations. (A) Shows a results under the assumption of free recombination $r = 0.5$ between neighboring sites, for sample sizes $m_1 = 1$ (red), $m_1 = 2$ (orange), $m_1 = 4$ (gold), $m_1 = 8$ (green), $m_1 = 16$ (blue), and $m_1 = N$ (dark blue). Note that results for $m_1 = 16$ (blue), and $m_1 = N$ (dark blue) are identical and overlap each other. The black dashed line indicates the target parameter ω. (B) Shows corresponding results in the presence of linkage between sites and a recombination rate of the same order of magnitude as mutation rate, that is, $r = 10^{-7}$. The black dashed line indicates the target parameter ω under the assumption of free recombination $r = 0.5$.

$$\omega_{\mathrm{na}} = \mathbb{E}[\omega_{\mathcal{V}}], \quad \omega_{\mathrm{a}} = c^{+}.$$

Alternatively, if we relax the assumption of instantaneous fixations and instead allow for a DFE extending on the positive real line, this becomes:

$$\omega_{\mathrm{na}} = (1-p)\mathbb{E}[\omega_{\mathcal{V}} | \mathcal{V} \leq 0], \quad \omega_{\mathrm{a}} = p\mathbb{E}[\omega_{\mathcal{V}} | \mathcal{V} > 0].$$

In our setting, these quantities naturally extend to separating time-dependent ratios $\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{na}}$ and $\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{a}}$. The long-term rate of nonadaptive and adaptive evolution are the limiting values:

$$\lim_{t\to\infty} \mathrm{d}N/\mathrm{d}S|_t^{\mathrm{na}} = \omega_{\mathrm{na}}, \quad \lim_{t\to\infty} \mathrm{d}N/\mathrm{d}S|_t^{\mathrm{a}} = \omega_{\mathrm{a}}.$$

A related measure is the proportion of nonsynonymous substitutions that are advantageous, α, which is given by the ratio of the expected number of advantageous nonsynonymous substitutions and the expected number of all nonsynonymous substitutions. The quantity α hence arises as the function of time and sample size defined by the ratio:

$$\alpha_t = \frac{\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{a}}}{\mathrm{d}N/\mathrm{d}S|_t} = 1 - \frac{\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{na}}}{\mathrm{d}N/\mathrm{d}S|_t}. \quad (12)$$

The asymptotic limit of α_t as $t \to \infty$ is:

$$\alpha_\infty = \frac{\omega_{\mathrm{a}}}{\omega} = 1 - \frac{\omega_{\mathrm{na}}}{\omega}.$$

Here, α_∞ corresponds to the target parameter commonly referred to as α or α_{TRUE}, which can be estimated from data using software packages such as DFEalpha (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). However, comparison between two species restricts the contribution of advantageous mutations to sequence divergence to a finite time interval $[0, t]$, from species divergence to present time. In a finite time interval, different rates of fixation of neutral and advantageous mutations will affect the contribution of advantageous mutations to sequence divergence. As a consequence, the proportion of advantageous fixations that contribute to sequence divergence in the finite time-interval is time-dependent even though the rate of adaptation ω_{a} is constant over time.

Implications for Inferences of the Rate of Adaptive Evolution

In this section, we assume, as in Keightley and Eyre-Walker (2007), that the DFE is restricted to the negative real line, $\mathcal{V} \leq 0$, and that all advantageous mutations are strongly advantageous mutations subject to instantaneous fixation, hence $\omega_{\mathrm{a}} = c^{+}$. We discuss point estimation of ω_{a} and α_∞ in the MK framework, that is splitting up ω into ω_{na} and ω_{a} under the assumption that the DFE of deleterious mutations $\mathcal{V} \leq 0$ is known from other sources (e.g., estimated based on observations of the allele frequency spectra of nonsynonymous and synonymous polymorphisms; Eyre-Walker et al. 2006). Since the distribution of $\mathcal{V}$ is known, the expected value $\omega_{\mathrm{na}} = \mathbb{E}[\omega_{\mathcal{V}}]$ is also known. The standard estimation procedure to draw inference on ω_{a} and α_∞ is motivated by equation (12), and amounts to replacing $\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{na}}$ by ω_{na} (Keightley and Eyre-Walker 2007), putting

$$\hat{\alpha}_t = 1 - \frac{\mathbb{E}[\omega_{\mathcal{V}}]}{\mathrm{d}N/\mathrm{d}S|_t}, \quad \hat{\omega}_{\mathrm{a}}(t) = \mathrm{d}N/\mathrm{d}S|_t - \mathbb{E}[\omega_{\mathcal{V}}]. \quad (13)$$

Since $\hat{\omega}_{\mathrm{a}}(t)$ is derived from $\mathrm{d}N/\mathrm{d}S|_t$ by subtraction of a constant value, $\hat{\omega}_{\mathrm{a}}(t)$ shows the same dependence on sample size and divergence time as $\mathrm{d}N/\mathrm{d}S|_t$. The quantity $\hat{\alpha}_t$ is a time-dependent equivalent of α of Keightley and Eyre-Walker (2007), and represents an estimate of the proportion of strongly advantageous nonsynonymous fixations as a function of time.

Figure 6A and B, respectively, illustrate α_t and $\hat{\alpha}_t$ as functions of time and sample size, where the differences between α_t and $\hat{\alpha}_t$ arise due to the fact that computation of $\hat{\alpha}_t$ is based on replacing $\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{na}}$ by ω_{na}. More specifically, figure 6A illustrates how the shape of α_t varies with

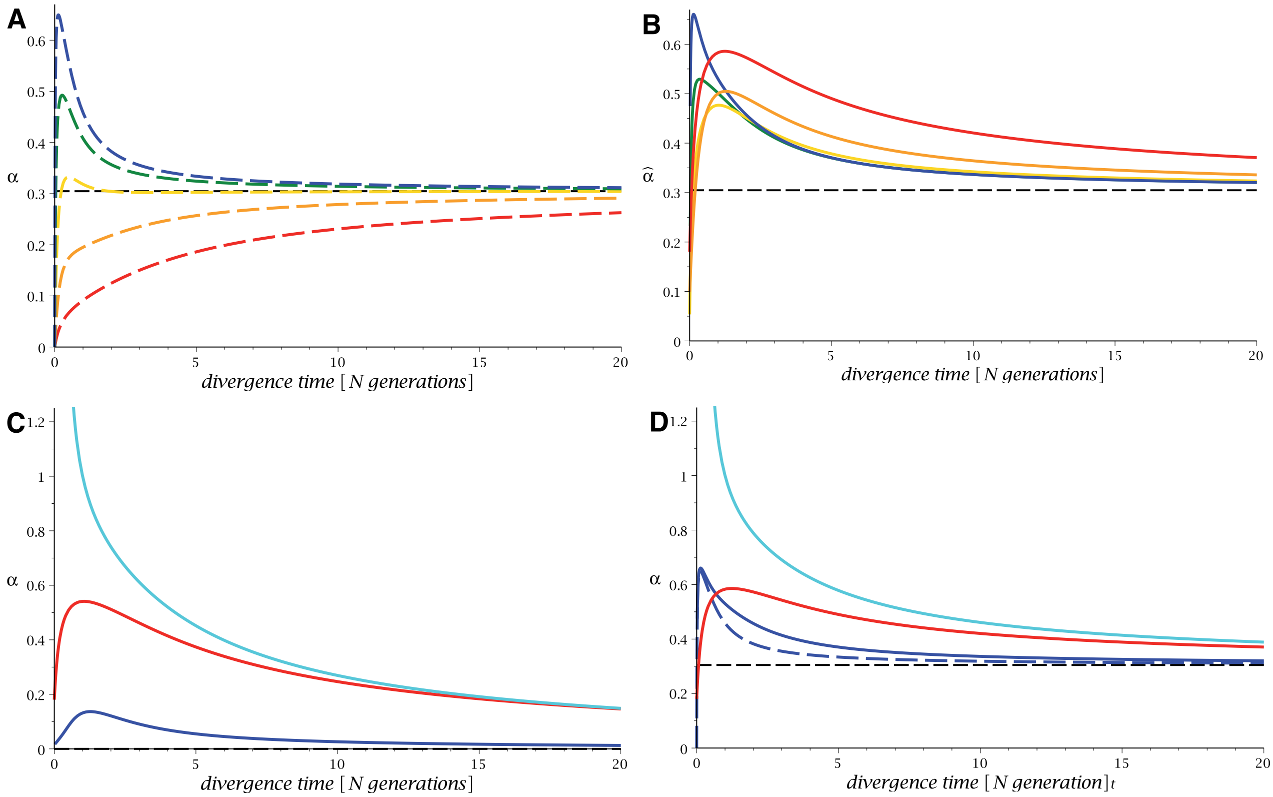

Fig. 6. (*A*) α_t and (*B*) $\hat{\alpha}_t$ as functions of divergence time for a DFE based on human data with $\omega_a = c^+ = 0.10$. Dashed lines are used to represent α_t, solid lines to represent $\hat{\alpha}_t$. Sample sizes are $m_1 = 1$ (red), $m_1 = 2$ (orange), $m_1 = 4$ (gold), $m_1 = 8$ (green), $m_1 = 16$ representative for the large sample approximation (blue). The black dashed line indicates α_∞. (*C* and *D*) Comparison of different estimates of α for $\omega_a = 0$ and $\omega_a = 0.10$, respectively. $\hat{\alpha}_t$ for $m_1 = 1$ (red) and $m_1 = 16$ representative for the large sample approximation (blue), and α_t^{cor} (turquoise), together with α_t for $m_1 = 16$ representative for the large sample approximation (blue dashed line). The black dashed line indicates the target parameter α_∞. Note that in panel (*C*) the blue dashed line and the black dashed line take the constant value of 0.

sample size m_1 for the asymmetric sampling scheme and $\omega_a = c^+ = 0.10$. For small sample sizes, slightly deleterious mutations segregating at low frequencies contribute significantly to the observed nonsynonymous divergence in the sample and hence α_t falls below its asymptotic limit α_∞. However, the larger the sample size, the fewer low-frequency segregating polymorphisms are observed as fixed in the sample. Instead, and in accordance with the fact that after short divergence time only the most advantageous mutations have had time to go to fixation, we observe that α_t stays above its asymptotic limit α_∞ for larger sample sizes. Of note, α_t for $m_1 \to \infty$ (blue line in fig. 6*A*) represents the actual proportion of advantageous nonsynonymous fixations in the focal species after split from the reference species. Besides, the fact that α_t is close to α_∞ for $m_1 = 4$ is not a general phenomenon, but will differ for different scenarios of selection, that is, depends on the DFE.

The bias of $\hat{\alpha}_t$ in relation to α_t and α_∞ is illustrated in figure 6*B* for the example of the EGP African human DFE, (Keightley and Eyre-Walker 2007), and $\omega_a = 0.10$. The significant deviation between $\hat{\alpha}_t$ and α_t is pronounced for the smaller sample sizes but gradually decreases with larger samples. As for $dN/dS|_t$, we observe that $\hat{\alpha}_t$ overestimates α_∞ for small divergence time, where larger sample size shows faster convergence with time. Besides, we highlight that if the sample size is sufficiently large, then $\hat{\alpha}_t$, specifically the large sample approximation,

$$\hat{\alpha}_t = 1 - \frac{\omega_{na}}{dN/dS|_t^\infty},$$

is a relevant estimate of α_t. Thus, consistent with the estimation of dN/dS incorporation of polymorphism data allows to correct for the misattribution of polymorphisms to sequence divergence. However, different rates of fixation of neutral and selected mutations lead to an overestimation of the target parameter α_∞ at narrow time-scales. The potential relevance of $\hat{\alpha}_t$ as a point estimate of α_∞, and with it the potential relevance of $\hat{\omega}_a(t) = \hat{\alpha}_t \cdot dN/dS|_t$ as a point estimate of ω_a, relies further on additional knowledge of t, and is only justified for large t or small P_{AP}.

Keightley and Eyre-Walker (2012) propose an alternative correction method to correct for the misattribution of polymorphisms to sequence divergence for small divergence time in the phylogenetic setting of $m_1 = m_2 = 1$. Specifically, it is suggested to subtract the stationary contribution of segregating polymorphisms from the estimates of dN_t and dS_t. In our framework, the quantity to be subtracted from dN_t is the average over $\mathcal{V}$ of:

$$\lim_{N\to\infty} N\mathbb{E}^{\mathcal{V}}_{1/N}\left[\int_0^{\tau} \xi_u \mathrm{d}u\right] = \omega_{\mathcal{V}} \int_0^1 y\pi_{\mathcal{V}}(y)\mathrm{d}y = 2\int_0^1 \frac{q_{\mathcal{V}}(y)}{q_0(y)}\mathrm{d}y,$$

which represents the observation of derived alleles in steady state in a sample of size $m_1 = m_2 = 1$. The suggested correction in Keightley and Eyre-Walker (2012) therefore amounts to replace $\mathrm{d}N/\mathrm{d}S|_t$ by $\mathrm{d}N_t^{\mathrm{cor}}/\mathrm{d}S_t^{\mathrm{cor}}$, where

$$\mathrm{d}N_t^{\mathrm{cor}} = \mathrm{d}N_t - 2\int_0^1 \frac{\mathbb{E}[q_{\mathcal{V}}(y)]}{q_0(y)}\mathrm{d}y, \quad \mathrm{d}S_t^{\mathrm{cor}} = \mathrm{d}S_t - 2.$$

However, this approach relies on the simplifying assumption that allele frequency spectra of the two species are in steady state, that is, that lineage-sorting is complete and the two allele frequency spectra are independent of each other. This assumption is violated in particular for closely related species (Amei and Sawyer 2010; Kaj and Mugal 2016). The correction therefore only performs well under certain conditions,

$$\mathrm{d}N/\mathrm{d}S|_t^{\mathrm{cor}} \le \mathrm{d}N/\mathrm{d}S|_t \quad \text{if} \quad \int_0^1 \mathbb{E}[q_{\mathcal{V}}(y)/y]\mathrm{d}y \ge \mathrm{d}N/\mathrm{d}S|_t,$$

and hence a validation of the correction method if we have the inequality,

$$\int_0^1 \mathbb{E}[q_{\mathcal{V}}(y)/y]\mathrm{d}y \ge \max_{t\ge 0} \mathrm{d}N/\mathrm{d}S|_t.$$

For our set of DFE parameters, however, the above criterion is not satisfied as the left hand side computes to ~0.26, whereas the maximum of $\mathrm{d}N/\mathrm{d}S|_t$ typically is larger, compare figure 2.

Figure 6C and D show the match between the corrected estimate of α of Keightley and Eyre-Walker (2012) (referred to as α_t^{cor}) and α_∞ together with $\hat{\alpha}_t$ for $m_1 = 1$ (i.e., the standard uncorrected estimate of α), $\hat{\alpha}_t$ for $m_1 \to \infty$ (our proposed correction), and α_t for $m_1 \to \infty$, for $c^+ = 0$. and $c^+ = 0.10$. This again illustrates that our proposed correction allows to correct for the misattribution of polymorphisms to sequence divergence, and that $\hat{\alpha}_t$ for $m_1 \to \infty$ is a relevant estimate of α_t. However, due to the different rates of fixation of neutral and selected mutations more elaborate correction approaches will be necessary to allow for point estimation of the target parameters α_∞ and ω_a at narrow time-scales.

A Protocol for the Estimation of ω and α for Closely Related Species

Our analytical results suggest that estimation of ω for closely related species can be improved by incorporating information on segregating polymorphism. To obtain branch-specific estimates of $\mathrm{d}N/\mathrm{d}S|_\infty = \omega$ for a species of interest, to which we refer as the focal species, we first need to select an appropriate reference species and an outgroup. Often, we are interested in the prevalence of natural selection and the contribution of adaptive evolution in the recent history of the focal species, which motivates the choice of a closely related reference species. As we have seen, however, observed dN/dS values based on the comparison of closely related species provide unreliable estimates of the mode and strength of selection represented by ω. This leaves us with the conflicting situation that we are interested in selecting a closely related species as reference and at the same time keep the bias in estimates of ω small.

During the early phase of species divergence where ancestral polymorphisms contribute substantially to the amount of segregating polymorphisms, estimates of ω are biased even for large sample sizes, since neutral and selected mutations differ in their rates of fixation (figs. 2 and 4). This phase will last until incomplete lineage sorting becomes negligible, that is, P_{AP} approaches 0, and divergence times are large enough that mutations in the weak selection regime have had sufficient time for eventual fixation. Naturally, the question arises which amount of time is large enough to retrieve reliable estimates of ω?

In practice, we generally do not know the separation time between two diverging populations in coalescent units (CUs). Even where accurate information on split times in years is available, the conversion to CUs is impeded by the difficulty of estimating the appropriate effective population size (N_e) in real-world populations (Palstra and Fraser 2012). Census population sizes usually differ widely from the underlying effective population size (Palstra and Ruzzante 2008) and estimates of N_e derived from genetic data differ in their parametrization and may not be appropriate in the context of protein evolution (Lanfear et al. 2014; Platt et al. 2018). The conversion is further complicated by empirical uncertainty in measures of generation time (Tremblay and Vézina 2000). In light of these complications, we propose two ad hoc criteria providing information on the suitability of the reference species for inference of ω.

Species Differentiation

The relationship $D = 2\theta(1 + t)$ under neutrality and $m_1 = m_2 = 1$ allows to obtain an estimate of t in CUs based on an estimate of divergence D and population mutation rate θ, $t = D/(2\theta) - 1$ (Wiuf et al. 2004). For codon sequences, this requires an estimate of synonymous divergence per site and an estimate of synonymous heterozygosity within the focal population. This can be readily computed based on a sample size of $m_1 \ge 2$ and $m_2 = 1$. Note, however, that equality $D = 2\theta(1 + t)$ is based on the assumptions of speciation without gene flow and constant population size. Migration and population demography will therefore impact estimates of t. Nevertheless, we suggest caution in interpreting phylogenetic estimates of ω if $t < 5$.

Sample Size Dependence

Based on our analytical results, we find that dN/dS shows a strong dependence on sample size at short evolutionary timescales. The dependence on sample size wears off at larger divergence time, as the estimate approaches the target parameter ω. The dependence on sample size thus constitutes a second criterium for the expected accuracy of estimates of ω. Specifically, we propose to compute and compare dN/dS for different choices of sample sizes ranging from 1 to 16 chromosomal copies (or 8 diploid individuals). Large differences in

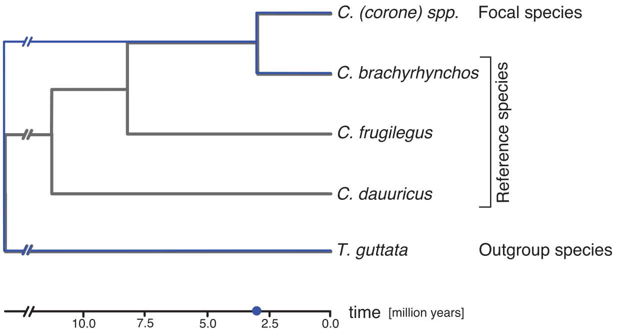

FIG. 7. Phylogeny and time-axis of the corvid species analyzed in the study together with zebra finch as outgroup. The phylogenetic reconstruction for all species is shown in black including the focal species *Corvus* (*corone*) spp., three respective reference species and one outgroup. A corresponding time-axis in My is given below (Jønsson et al. 2016). Time-dependence of d*N*/d*S* estimates are evaluated using increasingly divergent reference species. This is exemplified in the blue phylogeny for the most recent split between the focal species *C.* (*corone*) spp. and its sister species *C. brachyrhynchos* plus the outgroup *Taeniopygia guttata* (zebra finch). The blue dot on the time-axis corresponds to the split time between *C.* (*corone*) spp. and *C. brachyrhynchos*.

d*N*/d*S* values for different sample sizes $m_1 \geq 2$ suggest that the choice of reference species was inappropriate. If the d*N*/d*S* ratios appear robust to variation in sample size, estimation of ω is meaningful. Then estimates of the largest sample size should be considered as most accurate.

To summarize, estimates of *t* based on species differentiation can serve as first measure of the suitability of the reference species. Ultimately, the dependence on sample size allows to judge if the choice of reference species was appropriate and if d*N*/d*S* represents a reliable estimate of ω. Of note, it is sufficient to include polymorphism data for the focal species. Polymorphism data for the reference species do not lead to an additional improvement. If the d*N*/d*S* ratios appear robust to variation in sample size for $m_1 \geq 2$, removing segregating polymorphisms from the estimation is expected to significantly reduce the bias introduced by polymorphisms. In case a strong dependence on sample size is observed, the reference species should instead be replaced by a more distant species. If such data are not available, alternative approaches to estimate ω that do not rely on divergence estimates, but solely rely on variation within one species should be considered. Here, the recent method by Tataru et al. (2017) appears to be well-suited.

Step-by-Step Protocol for the Estimation of ω

Based on these results, we propose the following step-by-step protocol for the estimation of ω between closely related species.

1 Population Sample, Choice of the Reference Species, and Outgroup

We recommend the compilation of resequencing data for a minimum of 16 sequence copies of the focal species, that is, 8 diploid individuals or 16 haploid individuals. For the reference species, a single-sequence copy is sufficient. This permits evaluation of the dependence of d*N*/d*S* on sample size and further allows to compute an estimate of $t = D/(2\theta) - 1$ between the focal and the reference species. Only if species differentiation is reasonably large (roughly $t > 5$), the reference will be considered appropriate to infer ω based on standard phylogenetic approaches. The outgroup should be chosen close enough to avoid homoplasy, but distant enough such that the probability of incomplete lineage sorting is negligible.

2 Data Preparation

Once an appropriate reference species is chosen, we recommend to generate consensus sequences for different sample sizes of the focal species, where all polymorphic nucleotide sites within the sample are masked (printed as "N"). Each set of consensus sequences is then aligned to orthologs of the reference and the outgroup species. This results in several sets of triple alignments differing by sample size of the focal species.

3 Assessment of Sample Size Dependence

d*N*/d*S* is estimated for all sets of triple alignments. For the computation of d*N*/d*S*, a standard protocol can be followed such as Yang (2007). In case d*N*/d*S* values show a strong dependence on sample size, we recommend repeating the analysis with a more divergent reference species. As soon as d*N*/d*S* values appear robust to variation in sample size for $m_1 \geq 2$, the estimate based on the largest sample size should be considered most accurate.

The estimation of α based on the MK framework should only be considered if d*N*/d*S* estimates appear robust to variation in sample size. For the estimation itself, a standard protocol can be followed such as Keightley and Eyre-Walker (2007), Eyre-Walker and Keightley (2009) and Galtier (2016), where the d*N*/d*S* estimate based on the largest sample size should be considered. We do not recommend to perform a correction for ancestral polymorphisms following Keightley and Eyre-Walker (2012), since the correction only works under specific conditions. Moreover, if d*N*/d*S* estimates show a strong dependence on sample size and no other appropriate reference species is available, the estimation of the full DFE and α based on the recent method by Tataru et al. (2017) provides a relevant alternative approach.

Application to a Data Set of Corvid Species

We apply the proposed protocol to estimate the mode and strength of selection in species of the crow family, specifically in the evolutionary lineage leading to the *C.* (*corone*) spp. species complex (Vijay et al. 2016) (fig. 7).

Population Sample, Choice of the Reference Species, and Outgroup

Genome-wide resequencing data from 118 individuals of the focal species *C.* (*corone*) spp. (= *Corvus* (*corone*) *corone*/*cornix*/*orientalis*/*pectoralis*) was mapped to the reference genome and each individual was genotyped (Poelstra et al. 2014; Vijay et al. 2016). For the purpose of this study, this

Table 1. Branch-Specific Estimates of ω for the *Corvus* (*corone*) Species for Three Different Branch-Lengths and for Five Different Sample Sizes $m_1 = 1, 2, 8, 16, 32$.

Sample Size	dN/dS		
	C. brachyrhynchos	*C. frugilegus*	*C. dauuricus*
1	0.22	0.13	0.13
2	0.17	0.13	0.13
8	0.16	0.12	0.13
16	0.16	0.12	0.13
32	0.15	0.12	0.13

Table 2. Branch-Specific Estimates of α for the *Corvus* (*corone*) Species for Three Different Branch-Lengths and for Five Different Sample Sizes $m_1 = 1, 2, 8, 16, 32$.

Sample Size	$\hat{\alpha}$		
	C. brachyrhynchos	*C. frugilegus*	*C. dauuricus*
1	0.74	0.57	0.57
2	0.67	0.55	0.56
8	0.65	0.54	0.56
16	0.65	0.54	0.56
32	0.62	0.54	0.56

sample was subset to sample sizes of $m_1 = 1, 2, 8, 16, 32$ chromosomal copies (see Materials and Methods). In addition to population sampling of the focal species, we generated whole-genome sequencing data for three different corvid reference species with increasing divergence: these included the closely related sister species, the American crow *C. brachyrhynchos* (Vijay et al. 2016), the rook *C. frugilegus* and the Daurian jackdaw *C. dauuricus* (fig. 7). As outgroup, we used the distantly related zebra finch. Phylogenetic analyses place the separation between corvids (Corvoidea) and zebra finch (Passerida) at over 50 My (Jetz et al. 2012). Vijay et al. (2017) estimated 40 to 125×$2N_e$ generations as time to the most recent common ancestor of these two major songbird clades assuming a range of generation times and effective population sizes for species within these clades. Under the premise that $2N_e$ represents the coalescent effective population size this would reflect a separation time of $>$ 40 CUs (Lynch and Conery 2003), with the implication that reciprocal monophyly is expected to be reached for essentially all loci (Rosenberg 2003). Even though some phylogenies suggest earlier split times (Jarvis et al. 2014), it is safe to assume that divergence to the outgroup is sufficiently large for lineage sorting to be completed.

Incomplete lineage sorting between the focal and reference species may, however, be substantial. As suggested earlier, we calculated $t = D/(2\theta) - 1$ (Wiuf et al. 2004) based on synonymous divergence and heterozygosity. Split time between *C.* (*corone*) spp. and *C. brachyrhynchos* was found to be <1 CU ($t = 0.25$). Between *C.* (*corone*) spp. and *C. frugilegus* and between *C.* (*corone*) spp. and *C. dauuricus*, we observed $t > 5$ ($t = 6.13$ and $t = 11.82$, respectively). This suggests that lineage sorting between *C.* (*corone*) spp. and *C. frugilegus* and between *C.* (*corone*) spp. and *C. dauuricus* should be sufficiently advanced to reliably estimate ω. Estimation of ω using *C. brachyrhynchos* as the reference species is, on the other hand, expected to show strong dependence on sample size, and to be not reliable.

Sample Size Dependence

The theoretically predicted time-dependence of dN/dS of the focal species was reflected by phylogenetic distance to the reference. Estimates using *C. brachyrhynchos* as reference were almost twice as high as for *C. frugilegus* and *C. dauuricus* (table 1). Moreover, for *C. brachyrhynchos* as reference, dN/dS showed a strong sample size dependence. Estimates were highest when a single chromosome was sampled and decreased with increasing sample size. Sample size dependence was substantially less pronounced for the more distant reference species. These results are in good agreement with our analytical observation that the contribution of polymorphisms biases estimates of ω upward at short time-scales and that dependence on sample size is most pronounced at short time-scales (figs. 2 and 4). In addition, in agreement with Mugal et al. (2014), variance in estimation is particularly pronounced for the closest species comparison, but negligible for the more distant comparisons (supplementary table 1, Supplementary Material online).

Given the relationships (eq. 13), we expect to observe equivalent time and sample size dependence for estimates of ω_a and α based on the MK framework. To obtain empirical estimates of these quantities, we computed the DFE using all 16 individuals of the *C.* (*corone*) species with the DFEalpha method (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009) which approximates the DFE of nonadaptive mutations using a gamma distribution. Under the assumption of constant population size (one-epoch model), we obtained $a = 0.5742$, $b = 126.76$ as shape and scale parameters of the gamma distribution, respectively, corresponding to $\mathbb{E}[\omega_\mathcal{V}] = 0.056$ as estimate of ω_{na}. Next, we computed $\hat{\omega}_a$ and $\hat{\alpha}$ with the DFEalpha method (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009), which essentially is equivalent to solving for $\hat{\omega}_a(t) = \mathrm{d}N/\mathrm{d}S|_t - \mathbb{E}[\omega_\mathcal{V}]$. Thus, we naturally observed the same time and sample size dependence for estimates of ω_a as for dN/dS (supplementary table 1, Supplementary Material online).

Estimates of α were strongly sample size dependent with *C. brachyrhynchos* as reference and settled at ~0.54–0.56 for $m_1 > 2$ using *C. frugilegus* and *C. dauuricus* as reference (table 2). The method proposed by Keightley and Eyre-Walker (2012) to correct for the bias introduced by segregating polymorphism assuming $m_1 = m_2 = 1$ for the estimation based on *C. brachyrhynchos* yielded 0.33. For *C. frugilegus* and *C. daururicus*, we obtained an unrealistic value >1 and 0.72. Thus, the correction appeared sensitive to the chosen reference species, and no clear trend for the correction could be observed. This is in line with our analytical findings that deviations from the underlying assumptions of the correction may lead to erratic estimates. We therefore advocate the use of population samples to choose an appropriate reference, incorporate information from polymorphisms and use dependence on sample size to obtain unbiased estimates of ω and related measures.

Note, however, that unbiased estimation of ω does not guard against incorrect estimation of ω_a and α. The latter parameters further depend on the inferred DFE which is sensitive to demographic perturbation (Rousselle et al. 2018). The estimates of ω_a (0.07) and α (0.55) obtained under the assumption of constant population size are likely an overestimation owing to population expansion in *C.* (*corone*) spp. (Poelstra et al. 2013; Vijay et al. 2016). The two-epoch model aiming to correct for demographic perturbation yielded, however, negative values for both parameters and was not supported by a likelihood-ratio test (supplementary table 1, Supplementary Material online). Unrealistic values of $\hat{\omega}_a$ and $\hat{\alpha}$ may thus reflect biases in the estimation of the DFE, and stress the importance of interpreting point estimates of ω_a and α with care (Eyre-Walker and Keightley 2009). Here, the asymptotic MK test, which is a heuristic method that circumvents demographic inference from synonymous polymorphism data, constitutes a relevant alternative approach (Messer and Petrov 2013). However, like the MK test also, the asymptotic MK test relies on a phylogenetic estimate of ω, and should therefore only be applied to closely related species if the time-dependence of d*N*/d*S* has been taken into account.

Conclusion

Our analytical results show that phylogenetic estimates of ω are time-dependent due to 1) contribution of (lineage-specific and ancestral) polymorphisms to sequence divergence, and 2) different rates of fixation of neutral and selected mutations. As a natural consequence, estimates of α within the MK framework, which rely on phylogenetic estimates of ω are also time-dependent. Expression of d*N*/d*S* as a function of sample size further shows that exclusion of polymorphic sites allows to reduce bias in the estimation of ω and α with respect to the contribution of lineage-specific and ancestral polymorphisms to sequence divergence. Moreover, while one cannot systematically improve estimation of ω and α with respect to the different rates of fixation of neutral and selected mutations by simple exclusion of polymorphic sites, polymorphism data can be used to assess the accuracy in estimates. Based on these findings, we suggest a best-practice protocol for the estimation of ω and α, and illustrate the performance of this protocol by studying 11,035 genes in a genome data set of four crow species. This data set involves the estimation of branch-specific ω and α at three different time-scales and for several choices of sample sizes. In summary, our results highlight that polymorphism data can be a useful source of information and guide inferences of the strength and direction of selection for closely related species.

As a rule of thumb, we suggest caution in interpreting phylogenetic estimates of ω for $t < 5$. Here, estimates of t based on species differentiation can serve as first measure of the suitability of the reference species (Wiuf et al. 2004). A prominent species comparison that likely falls within the critical range of $t < 5$, is the comparison of chimpanzee with human. Using estimates of d*S* and heterozygosity from Chimpanzee Sequencing and Analysis Consortium (2005), we observe an estimate of $t \in (3, 6)$. In line with this, larger estimates of ω have been observed for branches between human and chimpanzee than for branches between more divergent vertebrates (Kosiol et al. 2008; Wolf et al. 2009). A similar time-dependence of d*N*/d*S* has also been observed in closely related bacteria (Rocha et al. 2006). Moreover, the d*N*/d*S* test statistic and the MK test are frequently applied to closely related species (subspecies) in the context of speciation research (Brand et al. 2015; Weber et al. 2017; Schirrmann et al. 2018) and transition of sexual systems (Shimizu and Tsuchimatsu 2015), where generally t is bound to be small. The time-dependence of d*N*/d*S* should therefore be considered a problem of broad interest, and we hope that our work increases the awareness and understanding of it, and thereby stimulates future research in developing polymorphism-aware phylogenetic approaches.

However, we would like to conclude by emphasizing that the time-dependence of d*N*/d*S* is one among several other biases in estimation of ω. For example, it has been shown that the assumption of stationary base composition biases estimates of ω if base composition is nonstationary (Guéguen and Duret 2018). In addition, GC-biased gene conversion has been shown to influence d*N*/d*S* across a wide range of species (Ratnakumar et al. 2010; Lartillot 2013; Bolívar et al. 2019). Moreover, selection on codon usage and exonic splice regulation impose selective constraint on synonymous mutations (Hershberg and Petrov 2008; Savisaar and Hurst 2018), biasing estimates of selection on amino acid changes (Matsumoto et al. 2016). Another bias is multinucleotide mutations (Venkat et al. 2018), which are found to be common in eukaryotes (Schrider et al. 2011) but are not treated within our framework. Considering each of the biases independently is of value for general conceptual understanding. However, more integrative models that consider several relevant biases will be necessary in the future in order to evaluate how these biases interact and influence inference of natural selection.

Materials and Methods

Evaluation and Visualization of Analytical Results

Evaluation and visualization of analytical results were performed with the software Maple, release 18.00, Maplesoft, Waterloo Maple Inc., Waterloo ON, Canada.

Forward Simulations

We performed forward simulations based on the SLiM software package version 3.2.1 (Haller and Messer 2019). We first implemented simulations based on free recombination in order to explore the agreement between our analytical results, and stochastic simulations based on accordant model assumptions. To meet the model assumptions underlying our analytical results, we therefore first chose a recombination rate of $r = 0.5$ between neighboring sites. Additive fitness effects were implemented, and population size was set to $N = 1{,}000$ (500 diploid individuals). We implemented an instantaneous speciation event after a burn in of 5 *N* generations and then tracked the number of nonsynonymous and synonymous fixed differences between the two populations as a function of time and sample size for 20 *N* generations. Parameters of mutation rate per site and generation μ and sequence length *L* were adjusted to keep the number of multiple mutation hits at a single site at low frequency over the entire simulation of 25 *N* generations, with $\mu = 10^{-7}$ and $L = 10^5$, such that the number of fixed differences between population samples corresponds to sequence divergence. Synonymous mutations were assumed to evolve neutrally, and selection on nonsynonymous mutations was specified by a DFE for deleterious and advantageous mutations. The DFE of deleterious mutations was modeled by a gamma distribution on the negative real line, with shape parameter $a = 0.15$ and a mean of $ab = 2{,}500$. The DFE of advantageous mutations was modeled by an exponential distribution with a mean of $c = 0.2$. Two independent runs of simulations were performed for a scenario in the absence of advantageous mutations, and a scenario in the presence of advantageous mutations. For the first scenario, the frequency of advantageous mutations was set to 0, and to 0.1 for the second scenario. We ran 1,000 replicates for each scenario and computed the mean d*N*/dS ratio as a function of time and sample size. In addition, we repeated the two independent runs of simulations for invoking linkage between sites. To do so, recombination rate was set to the same order of the same order of magnitude as mutation rate, that is, $r = 10^{-7}$. Thus, in total, four independent runs of simulations were performed. SLiM config files are available online (github.com/carinafm/dNdS_study).

Data and Taxonomy

Whole-genome resequencing data for 118 individuals from across the range of *Corvus* (*corone*) spp. species complex were collated from Poelstra et al. (2014) and Vijay et al. (2016), as well as of six *C. brachyrhynchos* individuals, four *C. frugilegus* and four *C. dauuricus* individuals. Details on sampling and data generation are available in Poelstra et al. (2014) and Vijay et al. (2016). Sample information including accession numbers of the Sequence Read Archive (SRA) can be found in supplementary table 2, Supplementary Material online.

Taxonomic status within the *C. corone* species group (hereafter *C.* (*corone*) spp.) is currently under debate (Parkin et al. 2003; Haring et al. 2007, 2012; Poelstra et al. 2014; Vijay et al. 2016). For the purpose of the present study, *C.* (*corone*) corone, *C.* (*c.*) cornix, and *C.* (*c.*) orientalis were considered as a single species (Vijay et al. 2016). In addition, CDSs and peptide sequences from the *Taeniopygia guttata* (zebra finch) genome version 3.2.4 were downloaded from Ensembl release 86 (accession GCF_000151805.1).

Sequence Alignment

Raw reads were trimmed using trimmomatic version 0.32 (Bolger et al. 2014) to remove Illumina adapter sequences and were subsequently mapped for each species to the *C.* (*c.*) cornix reference assembly (Poelstra et al. 2014) (accession GCF_000738735.1) using the Burrows–Wheeler Aligner BWA-MEM version 0.7.13 (Li and Durbin 2009). Duplicates were removed and the resulting bam files were sorted using samtools version 1.3 (Li et al. 2009). A consensus sequence in fasta format was generated for 1 *C. brachyrhynchos*, 1 *C. frugilegus*, and 1 *C. dauuricus* individual, as well as for four different sample sizes (16, 8, 4, and 1 individuals, respectively, or $m = 32$, 16, 8, and 2 sequence copies) of *C.* (*corone*) spp. Additionally, a consensus sequence was generated from the unmasked VCF of 1 *C.* (*corone*) spp. individual where, for heterozygous sites, one base was randomly chosen to construct a haploid sequence, that is, the sequence for $m = 1$. The 16 *C.* (*corone*) spp. individuals were chosen to represent different populations from across the distribution range to reflect the diversity of the species complex (supplementary table 2, Supplementary Material online).

The consensus sequences were generated as follows. For each consensus sequence, variants were called among the respective individuals using samtools mpileup version 1.3 and bcftools version 1.3 (Li et al. 2009), that is, producing one VCF file each for 1 *C. brachyrhynchos*, 1 *C. frugilegus*, 1 *C. dauuricus* individual, and for 16, 8, 4, and 1 *C.* (*corone*) spp. individuals, respectively. VCF files generated with samtools and bcftools contain one record per variant. This file format allows for the removal of indels while preserving one record at each affected site in the VCF file, which makes it possible to preserve congruence of reference annotations across the consensus sequences. We used GATK SelectVariants version 3.4.0 (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013) to remove indels from the VCF files and to generate a masking file for sites with a total coverage below 3×. The positions of variable sites within each data set were added to the masking files. From each VCF file, one genome-wide consensus sequence in fasta format was generated using bcftools, printing sites from the corresponding masking file as N. The consensus sequence of 1 haploid *C.* (*corone*) spp. sequence was generated from the VCF file of 1 *C.* (*corone*) spp. using GATK FastaAlternateReferenceMaker version 3.4.0 (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013) and the script convert_ambiguity_codes.py (github.com/markravinet; last accessed May 5, 2017). Sites with a coverage below 3× were removed from the haploid consensus sequence using BEDtools version 2.26.0 (Quinlan and Hall 2010).

Next, unique one-to-one orthologs between *C.* (*c.*) *cornix* (the reference assembly) and the outgroup *T. guttata* were identified. Specifically, the longest CDSs were extracted from the NCBI *Corvus cornix* Annotation Release 100 (assembly accession GCF_000738735.1) (Poelstra et al. 2014) to avoid technical duplicates, and then translated into peptide sequences using the gffread utility from the Cufflinks package. *Taeniopygia guttata* (accession GCF_000151805.1, Ensembl release 86) and *C.* (*c.*) *cornix* peptide sequences were compared with each other using an all-against-all protein BLASTp (e-value 1e-6) (BLAST+ version 2.5.0; Camacho et al. 2009), followed by orthology prediction using orthAgogue version 1.0.2 (Ekseth et al. 2014) based on the BLASTp scores.

The resulting 11,590 unique one-to-one orthologs were subsequently further filtered. First, the corresponding CDS was extracted from each of the genome-wide *Corvus* consensus sequences and genes with >50% masked sites in any of the *Corvus* consensus sequences were removed. Additionally, genes harboring at least two sites with >95% heterozygous genotypes in the VCF file of 16 *C.* (*corone*) spp. individuals were excluded to avoid analysis of duplicated genome regions.

For each of the 11,035 one-to-one orthologous genes that fulfilled all filtering criteria, one fasta file was generated containing the *T. guttata* ortholog plus the corresponding *C. brachyrhynchos*, *C. frugilegus*, *C. dauuricus*, and *C.* (*corone*) spp. (for 16, 8, 4, and 1 individuals) ortholog, respectively. Multiple sequence alignments were performed for each gene individually in PRANK version 150803 (Löytynoja 2013). The script pamlCleaner.py from github.com/RAWWiberg/ThCh6, last accessed January 31, 2017 was used to check if sequence lengths were multiples of three, and to remove stop codons. Next, the alignments of all 11,035 genes were concatenated to a single alignment. About 100 bootstrap replicates of the alignment were generated by drawing codons with replacement until the total length of the alignment was reached. Finally, 15 different subdata sets, each containing three species, were extracted for further analyses from the concatenated alignment and from each of the 100 bootstrap replicates. In-house code utilized for bioinformatics analysis is available online (github.com/verku/dNdS_study).

Estimation of d*N*/d*S*

We used the YN98+F3X4 model implemented in PAML (Yang 2007) to compute branch-specific maximum likelihood estimates of d*N*/d*S* for the *C.* (*corone*) spp. based on 15 combinations of triple alignments each including a consensus sequence of the *C.* (*corone*) spp. either based on $m_1 = 1, 2, 8,$ 16, or 32, one reference species and the outgroup *T. guttata* (fig. 7),

- *T. guttata*—*C. brachyrhynchos*—*C.* (*corone*) spp.
- *T. guttata*—*C. frugilegus*—*C.* (*corone*) spp.
- *T. guttata*—*C. dauuricus*—*C.* (*corone*) spp.

This setting allows to compute branch-specific estimates of d*N*/d*S* for the *C.* (*corone*) spp. for three different branch lengths, *C.* (*corone*) spp. after the split from *C. brachyrhynchos*, *C.* (*corone*) ssp. after the split from *C. frugilegus* and *C.* (*corone*) spp. after the split from *C. dauuricus*, and five different sample sizes each ($m_1 = 1, 2, 8, 16,$ or 32). We retrieved bootstrap confidence intervals for estimates of d*N*/d*S* for a bootstrap sample size of 100 based on resampling of the alignments as described earlier.

Variant Calling and Computation of the SFS

To be specific that we estimate the allele frequency spectrum of polymorphic nucleotide sites, we here use the otherwise equivalent term site frequency spectrum (SFS). For the computation of the SFS for *C.* (*corone*) spp. and for analyses of shared polymorphism among the different *Corvus* species, VCF files of all available 118 *C.* (*corone*) spp. individuals, as well as of 6 *C. brachyrhynchos*, 4 *C. frugilegus*, and 4 *C. dauuricus* individuals from Poelstra et al. (2014) and Vijay et al. (2016) were generated separately for each species using the GATK pipeline version 3.4.0 (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). This pipeline allows the use of machine learning algorithms for base score quality recalibration (BQSR) to improve variant discovery and genotype calls, and variant quality score recalibration (VQSR) to balance sensitivity and specificity of variant calls.

For *C.* (*corone*) spp., bam files were generated and processed by Vijay et al. (2016). For *C. brachyrhynchos*, *C. frugilegus*, and *C. dauuricus*, bam files were processed as follows, Readgroup information was assigned, bam files were sorted and duplicate read-pairs were marked using Picard version 1.141. Then, bam files were merged per individual and duplicates were marked a second time. To improve alignments in regions of insertion–deletion (indel) polymorphism, local realignment was performed per species using GATK's indel realigner. Next, BQSR was applied to each bam file. A truth set of variant sites was generated for each species using an iterative approach. First, we performed an initial round of variant calling on the uncalibrated bam files using three different variant discovery tools: GATK's HaplotypeCaller in variant discovery mode, samtools v1.3 (Li et al. 2009) and FreeBayes v1.0.2 (Garrison and Marth 2012). The intersection of variant sites among the methods was extracted and used as set of known sites for a first round of BQSR. Next, we called variants a second time exclusively in HaplotypeCaller. The highest quality variants (~10–15%) were extracted and used as set of known variants for a second round of BQSR.

Variant calling was conducted for each individual bam file separately using GATK's HaplotypeCaller in ERC mode. Next, data from all individuals from one species were jointly genotyped using GenotypeGVCFs in GATK, generating one VCF file per species. VQSR implemented in GATK was used to filter the called variants. VQSR uses a set of known variant sites to estimate a calibrated probability that each call in the raw VCF file is a true variant or a technical artifact. Without previous knowledge of true variants, variants that had the $10 - 15\%$ highest quality scores were used to generate an error model for each of the four species. Finally, only those variants that passed the 99.0 tranche in GATK's VQSR were kept.

Site categories were determined based on the NCBI *Corvus cornix* Annotation Release 100 (assembly accession GCF_000738735.1) using the script NewAnnotateRef.py (Williamson et al. 2014), and 0- and 4-fold degenerate sites from autosomal scaffolds were extracted from the VCF file. Repeat content based on an updated repeat annotation of the *C.* (*c.*) cornix reference assembly (Weissensteiner et al. 2017) and known contaminations (Poelstra et al. 2014) were removed. Since mutation rate might differ between CpG-sites and other sites in the genome due to the hypermutability of CpG-sites (Nachman and Crowell 2000; Suzuki et al. 2009), we further identified CpG, TpG, and CpA sites (fixed sites and sites polymorphic for CpG/TpG and CpG/CpA) in the resequencing data and excluded these sites from the VCF file using a custom script. We used a cutoff for minor allele frequencies (MAF) of 0.005, which excludes all singletons from the VCF file of all available 118 *C.* (*corone*) spp. individuals in order to exclude potential sequencing errors. For further analysis, we downsampled the MAF-filtered VCF file and the unfiltered VCF file, that is, without exclusion of singletons, to the same 16 *C.* (*corone*) spp. individuals that were used for the generation of the consensus sequence. Finally, folded SFS were computed separately for 0- and 4-fold degenerate sites using the scripts vcfSummarizer.py and bootstrapRegions.py (Williamson et al. 2014). In-house code utilized for bioinformatics analysis is available online (github.com/verku/dNdS_study).

Estimation of the DFE and α

We estimated the DFE of deleterious 0-fold degenerate sites with DFEalpha v.2.16 (Keightley and Eyre-Walker 2007) using 4-fold degenerate sites as neutral reference. The SD for DFE parameter estimates were obtained from the SFS from 200 randomly sampled nonoverlapping 10 kb windows (with replacement) using the script bootstrapRegions.py (Williamson et al. 2014). Next, we proceeded to estimate α using DFEalpha v.2.16 (Eyre-Walker and Keightley 2009) separately for d*N*/d*S* estimates of each sample size ($m = 1$, 2, 8, 12, or 16), and for each of the three different branch lengths, without invoking the correction for ancestral polymorphism, and with invoking it for $m = 1$. The DFE was estimated for the MAF-filtered and -unfiltered SFS. Results for the MAF-filtered data are reported in the main text. Results for the unfiltered data are provided in supplementary table 1, Supplementary Material online.

Supplementary Material

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

Acknowledgments

The authors thank Sylvain Glémin for valuable discussions, Pascal Milesi for generous advice on running forward simulations, Rebekka Müller for proof-reading equations, R. Axel Wiberg for sharing scripts and Saurabh Pophaly for help with data upload. The authors thank the UPPMAX Next-Generation Sequencing Cluster and Storage (UPPNEX) project, funded by the Knut and Alice Wallenberg Foundation and the Swedish National Infrastructure for Computing, for access to computational resources, and Douglas Scofield for computational assistance. C.F.M. is funded by grants to Hans Ellegren from the Swedish Research Council (2013/08271) and Knut and Alice Wallenberg Foundation. V.E.K. was supported by the Sven and Lilly Lawskis Foundation (to V.E.K. and J.B.W.W.), F.B.C. by the European Research Council (grant number ERCStG-336536 to J.B.W.W.). Further funding for data generation was provided by the European Research Council (grant number ERCStG-336536 to J.B.W.W.) and Knut and Alice Wallenberg Foundation (to J.B.W.W.).

References

Amei A, Sawyer S. 2010. A time-dependent Poisson random field model for polymorphism within and between two related biological species. *Ann Appl Probab.* 20(5):1663–1696.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol.* 21(7):1350–1360.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Bolívar P, Guéguen L, Duret L, Ellegren H, Mugal CF. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol.* 20(1):5.

Brand P, Ramírez SR, Leese F, Quezada-Euan JJG, Tollrian R, Eltz T. 2015. Rapid evolution of chemosensory receptor genes in a pair of sibling species of orchid bees (*Apidae: Euglossini*). *BMC Evol Biol.* 15(1):176.

Cagan A, Theunert C, Laayouni H, Santpere G, Pybus M, Casals F, Prufer K, Navarro A, Marques-Bonet T, Bertranpetit J, et al. 2016. Natural selection in the great apes. *Mol Biol Evol.* 33(12):3268–3283.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. *Genome Biol Evol.* 4(3):278–288.

Charlesworth D. 2010. Don't forget the ancestral polymorphisms. *Heredity* 105(6):509–510.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol.* 25(6):1007–1015.

Chen H. 2012. The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theor Popul Biol.* 81(2):179–195.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Christe C, Stölting KN, Paris M, Fraïsse C, Bierne N, Lexer C. 2017. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol Ecol.* 26(1):59–76.

DeMaio N, Schlötterer C, Kosiol C. 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol Biol Evol.* 30:2249–2262.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498.

Edwards SV, Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54(6):1839–1854.

Ekseth OK, Kuiper M, Mironov V. 2014. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* 30(5):734–736.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. *Mol Ecol.* 17(21):4586–4596.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet.* 8(8):610–618.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26(9):2097–2108.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2):891–900.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.

Figuet E, Nabholz B, Bonneau M, Carrio EM, Nadachowska-Brzyska K, Ellegren H, Galtier N. 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Mol Biol Evol.* 33(6):1517–1527.

Gagnaire PA, Normandeau E, Bernatchez L. 2012. Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American eels. *Mol Biol Evol.* 29(10):2909–2919.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet.* 12(1):e1005774.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv.: 1207.3907v2.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11(5):725–736.

Gossmann T, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol.* 27(8):1822–1832.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol.* 4(5):658–667.

Gronau I, Arbiza L, Mohammed J, Siepel A. 2013. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol Biol Evol.* 30(5):1159–1171.

Guéguen L, Duret L. 2018. Unbiased estimate of synonymous and nonsynonymous substitution rates with nonstationary base composition. *Mol Biol Evol.* 35(3):734–742.

Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol.* 36(3):632–637.

Haring E, Däubl B, Pinsker W, Kryukov A, Gamauf A. 2012. Genetic divergences and intraspecific variation in corvids of the genus *Corvus* (Aves: Passeriformes: Corvidae) – a first survey based on museum specimens. *J Zool Syst Evol Res.* 50(3):230–246.

Haring E, Gamauf A, Kryukov A. 2007. Phylogeographic patterns in widespread corvid birds. *Mol Phylogenet Evol.* 45(3):840–862.

Hart MW, Stover DA, Guerra V, Mozaffari SV, Ober C, Mugal CF, Kaj I. 2018. Positive selection on human gamete-recognition genes. *PeerJ* 6:e4259.

Hasegawa M, Cao Y, Yang ZH. 1998. Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. *Mol Biol Evol.* 15(11):1499–1505.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. *Genome Res.* 17(12):1837–1849.

Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet.* 42:287–299.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8(3):269–294.

Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. *Mol Ecol.* 20(15):3087–3101.

Hughes AL. 2008. Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann N Y Acad Sci.* 1133:162–179.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1320–1331.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. *Nature* 491(7424):444–448.

Jønsson KA, Fabre PH, Kennedy JD, Holt BG, Borregaard MK, Rahbek C, Fjeldså J. 2016. A supermatrix phylogeny of corvoid passerine birds (Aves: Corvides). *Mol Phylogenet Evol.* 94(Pt A):87–94.

Kaj I, Mugal CF. 2016. The non-equilibrium allele frequency spectrum in a Poisson random field framework. *Theor Popul Biol.* 111:51–64.

Kaplan NL, Hudson RR, Langley CH. 1989. The hitchhiking effect revisited. *Genetics* 123(4):887–899.

Keightley PD, Campos JL, Booker TR, Charlesworth B. 2016. Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. *Genetics* 203(2):975–984.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.

Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J Mol Evol.* 74(1–2):61–68.

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

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol.* 28(11):3033–3043.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4(8):e1000144.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet.* 4(12):e1000304.

Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. *Trends Ecol Evol.* 29(1):33–41.

Lartillot N. 2013. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol Biol Evol.* 30(2):356–368.

Lessios HA. 2011. Speciation genes in free-spawning marine invertebrates. *Integr Comp Biol.* 51(3):456–465.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Lima TG, McCartney MA. 2013. Adaptive evolution of M3 Lysin-A CandiYear gamete recognition protein in the *Mytilus edulis* species complex. *Mol Biol Evol.* 30(12):2688–2698.

Lipinska AP, Van Damme EJM, De Clerck O. 2016. Molecular evolution of candidate male reproductive genes in the brown algal model *Ectocarpus*. *BMC Evol Biol.* 16(1):5.

Loewe L, Charlesworth B. 2006. Inferring the distribution of mutational effects on fitness in *Drosophila*. *Biol Lett.* 2(3):426–430.

Löytynoja A. 2013. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Multiple sequence alignment methods. Totowa, New Jersey: Humana Press. p. 155–170.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302(5649):1401–1404.

MacEachern S, McEwan J, McCulloch A, Mather A, Savin K, Goddard M. 2009. Molecular evolution of the Bovini tribe (Bovidae, Bovinae): is there evidence of rapid evolution or reduced selective constraint in domestic cattle? *BMC Genomics* 10:179.

Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. 2016. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. *Mol Biol Evol.* 33(6):1580–1589.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res.* 23(1):23–35.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297–1303.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. *Proc Natl Acad Sci U S A.* 110(21):8615–8620.

Mugal CF, Wolf JBW, Kaj I. 2014. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol Biol Evol.* 31(1):212–231.

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(5):715–724.

Nabholz B, Uwimana N, Lartillot N. 2013. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome Biol Evol.* 5(7):1273–1290.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23(1):263–286.

Palstra FP, Fraser DJ. 2012. Effective/census population size ratio estimation: a compendium and appraisal. *Ecol Evol*. 2(9):2357–2365.

Palstra FP, Ruzzante DE. 2008. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Mol Ecol*. 17(15):3428–3447.

Palumbi SR. 2009. Speciation and the evolution of gamete recognition genes: pattern and process. *Heredity* 102(1):66–76.

Parkin DT, Collinson M, Helbig AJ, Knox AG, Sangster G. 2003. The taxonomic status of carrion and hooded crows. *Brit Birds*. 96:274–290.

Peterson GI, Masel J. 2009. Quantitative prediction of molecular clock and K(a)/K(s) at short timescales. *Mol Biol Evol*. 26(11):2595–2603.

Phung TN, Huber CD, Lohmueller KE. 2016. Determining the effect of natural selection on linked neutral divergence across species. *PLoS Genet*. 12(8):e1006199.

Platt A, Weber CC, Liberles DA. 2018. Protein evolution depends on multiple distinct population size parameters. *BMC Evol Biol*. 18(1):17.

Poelstra JW, Ellegren H, Wolf J. 2013. An extensive candidate gene approach to speciation: diversity, divergence and linkage disequilibrium in candidate pigmentation genes across the European crow hybrid zone. *Heredity* 111(6):467–473.

Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, Baglione V, Unneberg P, Wikelski M, Grabherr MG, et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* 344(6190):1410–1414.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci*. 365(1552):2571–2580.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol*. 239(2):226–235.

Rosenberg NA. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 57(7):1465–1477.

Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. *Biol Lett*. 14(5):20180055.

Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res*. 28(10):1442–1454.

Sawyer SA, Hartl DL. 1992. Population-genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176.

Schirrmann MK, Zoller S, Croll D, Stukenbrock EH, Leuchtmann A, Fior S. 2018. Genomewide signatures of selection in *Epichloë* reveal candidate genes for host specialization. *Mol Ecol*. 27(15):3070–3086.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189(4):1427–1437.

Schrider DR, Hourmozdi J, Hahn MW. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol*. 21(12):1051–1054.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre GP, Bank C, Brannstrom A, et al. 2014. Genomics and the origin of species. *Nat Rev Genet*. 15(3):176–192.

Settepani V, Bechsgaard J, Bilde T. 2016. Phylogenetic analysis suggests that sociality is associated with reduced effectiveness of selection. *Ecol Evol*. 6(2):469–477.

Shimizu KK, Tsuchimatsu T. 2015. Evolution of selfing: recurrent patterns in molecular adaptation. *Annu Rev Ecol Evol Syst*. 46(1):593–622.

Suzuki Y, Gojobori T, Kumar S. 2009. Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. *Mol Biol Evol*. 26(10):2275–2284.

Tang SW, Presgraves DC. 2009. Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. *Science* 323(5915):779–782.

Tataru P, Mollion M, Glemin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* 207(3):1103–1119.

Tremblay M, Vézina H. 2000. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am J Hum Genet*. 66(2):651–658.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 43:11.10.1–11.10.33.

Venditti C, Pagel M. 2010. Speciation as an active force in promoting genetic evolution. *Trends Ecol Evol*. 25(1):14–20.

Venkat A, Hahn MW, Thornton JW. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat Ecol Evol*. 2(8):1280–1288.

Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf J. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nat Commun*. 7(1):13195.

Vijay N, Weissensteiner MH, Burri R, Kawakami T, Ellegren H, Wolf J. 2017. Genome-wide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. *Mol Ecol*. 26(16):4284–4295.

Wakeley J, Hey J. 1997. Estimating ancestral population parameters. *Genetics* 145(3):847–855.

Weber AA, Abi-Rached L, Galtier N, Bernard A, Montoya-Burgos JI, Chenuil A. 2017. Positive selection on sperm ion channels in a brooding brittle star: consequence of life-history traits evolution. *Mol Ecol*. 26(14):3744–3759.

Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol*. 15(12):542.

Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW. 2017. Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res*. 27(5):697–708.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* 173(2):821–837.

Welch JJ, Eyre-Walker A, Waxman D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J Mol Evol*. 67(4):418–426.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet*. 10(9):e1004622.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet*. 7(12):e1002395.

Wiuf C, Zhao K, Innan H, Nordborg M. 2004. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* 168(4):2363–2372.

Wolf JBW, Kunstner 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.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*. 17(1):32–43.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19(6):908–917.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.