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

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

1Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden
2Science for Life Laboratory, Stockholm University, Stockholm, Sweden
3Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
4Division of Evolutionary Biology, Faculty of Biology, LMU Munich, Planegg-Martinsried, Germany
5Department 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, dN/dS, 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 dN/dS as a function of divergence time and sample size. We then extend our framework to the estimation of the proportion of adaptive protein evolution ω. 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 ω 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 ω at different time-scales and for several choices of sample sizes.

Key words: molecular evolution, codon models, dN/dS, 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 dN/dS 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 ωna and an adaptive part ωa (Gossmann et al. 2010, 2012; Galtier 2016). Here, ωa provides information on the rate of adaptive protein evolution and forms the basis for the commonly used measure ω = ωa + ωna, which reflects the proportion of beneficial nonsynonymous substitutions. Whereas initially the estimation of ω was based on the assumptions of the neutral theory of molecular evolution (Fay et al. 2001; Biere 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;
population size, sequence divergence

the population genetics of sequence divergence (fig. 1).

Because of its alleged simplicity and intuitive appeal, the fixation rate ratio \( \omega \) and derived measures have a strong tradition in evolutionary research. Genome-wide averages of estimates of \( \omega \) and \( \omega_a \) 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 et al. 2016; Settepani et al. 2016). Notably, this often involves estimation of \( \omega \) between lineages with different divergence ages, in other words estimation of \( \omega \) at different time-scales. Besides, at the level of genes and sites, estimates of \( \omega \) 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; Lipinski 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 (Seehaus et al. 2014; Christie et al. 2017). However, estimates of \( \omega \) 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 \( \mu \) 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 \( pN_s/pS \) (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 \( \omega \) 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 \( \omega \) 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 \( \omega \). Such a mathematical treatment will not only allow to judge what divergence time is large enough to apply current methodology but will furthermore permit developing polymorphism-aware software for unbiased estimation of target parameters \( \omega \) and \( \omega_a \) 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 \( \omega \) 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...
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 $\omega$. We further put our results in the context of the MK framework and the estimation of $\omega$. Of note, our analytical results show that estimates of $\omega$ and $z$ are time-dependent, even though target parameters $\omega$ and $z$ themselves are constant. In contrast, changes in demography or selection pressure over time lead to a time-dependence in target parameters $\omega$ and $z$, 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 $\mu$, 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 $\gamma = NLs$ represents the population-scaled selective pressure on derived alleles. The basic model for selection amounts to assigning each derived allele a fixed selective weight $\gamma$ (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 $\gamma$ from the distribution of a random variable $V$, 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 $V \leq 0$ 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 $V$ is continuous with density function $h_V(\cdot)$.

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 $NL\mu$. 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 $\gamma$, using the approach in Kaj and Mugal (2016) let $P^\gamma$ and $E^\gamma$ 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 \( \tau_0 \) be the extinction time of a derived allele, \( \tau_1 \) the fixation time, and \( \tau = \min(\tau_0, \tau_1) \) the absorption time. The probability of fixation, \( q_1(\xi) \), is known to be (Kimura 1962):

\[
q_1(\xi) = \mathbb{P}_\xi(\tau_1 < \infty) = \frac{1 - e^{-2\gamma \xi}}{1 - e^{-2\gamma}}, \quad \gamma \neq 0, \quad q_0(\xi) = \xi.
\]

In the limit of large \( N \), the result of applying mutation rate \( N\theta \) and fixation probability \( q_1(1/N) \) is the scaled effective fixation rate \( \omega_1 \), where

\[
\omega_1 = \lim_{N \to \infty} N q_1(1/N) = \frac{2\gamma}{1 - e^{-2\gamma}}, \quad \gamma \neq 0, \quad \omega_0 = 1. \quad (1)
\]

Here, \( \omega_1 \) represents the fixation rate ratio for a site evolving with selective pressure \( \gamma \) 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 \( \gamma \):

\[
\mathbb{E}[\omega_1] = \int_0^\infty \omega_1 \nu_1(\nu) d\nu.
\]

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 \( \omega_1 \mu_1(\nu), \, 0 < \nu < 1 \), which arises in the limit relation:

\[
\lim_{N \to \infty} N \mathbb{E}_1[N] \int_0^\tau f(\xi_t) dt \, d\omega_1 \int_0^1 f(y) \pi_1(y) dy,
\]

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

\[
\pi_1(y) = \frac{1 - e^{-2\gamma(1-y)}}{\gamma y(1-y)}, \quad \gamma \neq 0, \quad \pi_0(y) = \frac{1}{y}.
\]

Again, for variation in selection across sites, the limiting functionals are to be averaged over \( \nu_1 \). In particular, the allele frequency spectrum becomes:

\[
\int_0^1 f(y) \mathbb{E}[\omega_1 \nu_1(\nu_1) f(y) dy.
\]

To handle time-dependent functionals in the next section, we will also need the probability distribution \( \mathbb{P}_\xi(\tau_1 < t) = \mathbb{P}_\xi(\xi_t = 1), \, t \geq 0 \), of the fixation time \( \tau_1 \), and the conditional probability distribution \( \mathbb{P}_\xi(\tau_1 < t) = \mathbb{P}_\xi(\xi_t < t)/q_1(\xi), \, t \geq 0 \), given that the fixation time \( \tau_1 \) 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,

\[
Q_{k_1,k_2} = \text{average number of sites with exactly } k_1 \text{ derived alleles in species } 1 \text{ and } k_2 \text{ derived alleles in species } 2.
\]

To analyze \( Q_{k_1,k_2} \) as a function of time, we distinguish between the ancestral contribution \( H_{k_1,k_2} \), representing the average number of derived alleles at time \( t \) caused by mutations segregating in the ancestral population, and the lineage-specific contributions \( Z^\gamma_t \), 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, \( Q^1_{k_1,k_2} = H_{k_1,k_2}^1 \) 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,

\[
Q^0_{k_1,k_2} = H_{k_1,k_2}^0 + Z^\gamma_t, \quad k_1 \geq 1, \quad Q^0_{1,k_2} = H_{1,k_2}^0 + Z^\gamma_t, \quad k_2 \geq 1.
\]

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

In this work, we use the functionals \( Q^1_{k_1,k_2} \) 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,

\[
D_{Q^m,m^0}^m(t) = Q^m_{0,0} + Q^0_{0,m} = H_{m,0}^m + H_{1,m}^0 + Z_{m}^x_1 + Z_{m}^y_2.
\]

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 \( D_{y,m}^m(t) \), now splits into lineage-specific and ancestral parts, as

\[
D_{y,m}^m(t) = Z_{y}^m + H_{t,1}^y,
\]

where the second term also depends on \( m_2 \) through

\[
H_{t,1}^y = \omega_1 \int_0^1 \mathbb{E}_1^y[(1 - \xi_t)^{m_2}] \pi_1(y) dy.
\]

The first term \( Z_{y}^m \) only depends on \( m_1 \) and is given by,

\[
Z_{y}^m = \omega_1 \tau_t - \omega_1 \int_0^1 (q(y) - y^m) \mathbb{E}_1^y[(1 - \xi_t)^{m_2}] \pi_1(y) dy. \quad (2)
\]

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

\[
D_{y,m}^m(t) = \omega_1 \tau_t - \omega_1 \int_0^1 (q(y) - y^m) \mathbb{E}_1^y[(1 - \xi_t)^{m_2}] \pi_1(y) dy + \omega_1 \int_0^1 \mathbb{E}_1^y[(1 - \xi_t)^{m_2}] \pi_1(y) dy. \quad (3)
\]

The functions \( \mathbb{E}_1^y[(1 - \xi_t)^{m_2}] \) appearing in these formulas are explicitly known for the neutral case \( \gamma = 0 \). The neutral divergence is:

\[
D_{y,m}^0(t) = \omega_1 \tau_t - \omega_1 \int_0^1 (y - y^m) \mathbb{E}_1^0[(1 - \xi_t)^{m_2}] \pi_1(y) dy + \int_0^1 \mathbb{E}_1^0[(1 - \xi_t)^{m_2}] \pi_1(y) dy.
\]

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

263
Here, the dependence of \( F_{m,t} \) functions of lineage-specific and ancestral polymorphisms, given by \( F_{m,t} \) with nonlinear parts, of selection on nonsynonymous mutations is given by a DFE its divergence from species \( 2, \) see equation (25) in supplementary appendix A in Supplementary Material online. Here, \( dN/dS|_{t=0} \) for \( m_1 = m_2 = 1 \) represents the ratio of nonsynonymous over synonymous heterozygosity sampled in the ancestral population. Inequality \( dN/dS|_{t=0} \geq \omega \) 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 \( dS_t \), 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 \( F_{m_1,m_2}^{\text{syn}}(t) \), see equation (25) in supplementary appendix A in Supplementary Material online. For a given distribution of \( V \), these results allow us to derive \( dN_t \) and hence \( dN/dS|_{t=0} \) in equation (7) as a function of time and sample size.

As an additional result, we obtain the \( dN/dS \) ratio explicitly as a function of time in the limit of large sample size (supplementary appendix B in Supplementary Material online).
The relation (eq. 10) shows that the fixation rate ratio can provide a basis for future method development for ratio can be explained by differences in the time to fixation point estimation of served CDS divergence in the sample. As evident inequation with fixed (but unknown) divergence time illustrates that the intrinsic time-dependence of the \( d_N \) fixation rate ratio of polymorphism and divergence data can improve the estimation of the x Polymorphism Data Assist Estimation of m

In particular, letting \( m_1 \rightarrow \infty \) for fixed \( m_2 = 1 \), the large sample limit function \( dN/dS|_{t}^{\infty} \) for \( t > 0 \) is:

\[
\frac{dN}{dS}|_{t}^{\infty} = \frac{E[(\omega_V(t - \int_{0}^{1} p_V^0(\tau_1 < t)E[p_V^0(\xi_1)p_V(y)dy])]}{t - \int_{0}^{1} p_V^0(\tau_1 < t)2dy}.
\]

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

\[
P_V^y(\tau_1 < t) = P_V^0(\tau_1 < t)q(y) \approx P_V^0(\tau_1 < t)q(y),
\]

and use an explicit summation formula for the neutral conditional probability \( P_V^0(\tau_1 < t) \), 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, \( dN/dS|_{t}^{\infty} \) does not converge to \( \omega \) for large sample sizes as long as \( t \) is fixed. By taking \( t \rightarrow \infty \) in equation (9), we recover again the limit \( \omega \). A similar but slightly different expression results if we take both sample sizes large, \( m_1, m_2 \rightarrow \infty \). A further interpretation arises by rewriting equation (9) in the form:

\[
\omega = \frac{dN}{dS}|_{t}^{\infty} (1 - \frac{1}{t}B_t) + \frac{1}{t}C_t
\]

with

\[
B_t = \int_{0}^{1} p_V^0(\tau_1 < t)2dy,
\]

\[
C_t = E\left[\int_{0}^{1} \omega_V(p_V^0(\tau_1 < t)E[p_V^0(\xi_1)p_V(y)dy])\right].
\]

The relation (eq. 10) shows that the fixation rate ratio \( \omega \) is equal to the present ratio \( \frac{dN}{dS}|_{t}^{\infty} \) 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 \( p_V^0(\tau_1 < t) \) and \( p_V^y(\tau_1 < t) \), 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 \( \omega \).

Polymorphism Data Assist Estimation of \( \omega \)

We are interested to investigate if the joint usage of polymorphism and divergence data can improve the estimation of the fixation rate ratio \( \omega \) for closely related species. More precisely, the task is point estimation of \( \omega = E[\omega_V] \) 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 \( \omega \) with observed divergence data \( \frac{dN}{dS}|_{t}^{\infty} \) or its large sample version \( \frac{dN}{dS}|_{t}^{\infty} \) remains open as long as the divergence time \( t \) is unknown. To help understand the bias, we study the time-dependent function \( \frac{dN}{dS}|_{t}^{\infty} \) 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 \( P_F \), for which we want to estimate \( \omega \), 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 \( V \) 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 \( V \) is:

\[
h_V(v) = \frac{(-v)^{a-1}e^{-v/b}}{b^a\Gamma(a)}, \quad v \leq 0,
\]

where \( a > 0 \) is the shape parameter and \( b > 0 \) the scale parameter, hence the expected value \( E(V) = -ab \). Values of \( V \) 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 \( V \) below some value \( v_- \) < 0, on the other hand, represent a fraction \( P(V < v_-) \) of strongly deleterious mutations, which in practice are removed instantaneously (Welch 2006; Eyre-Walker and Keightley 2007).

Figure 2 illustrates the behavior of \( \frac{dN}{dS}|_{t}^{\infty} \) 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 \( \omega = E[\omega_V] \) is overestimated by \( \frac{dN}{dS}|_{t}^{\infty} \) at any fixed time \( t \). The bias is larger for the smaller sample sizes and varies over time until it fades out as \( t \rightarrow \infty \). 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 \( E[\omega_V] \) using \( \frac{dN}{dS}|_{t}^{\infty} \) 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 \( \frac{dN}{dS}|_{t}^{\infty} \geq E[\omega_V] \) (blue curve in fig. 2B), and not the constant function \( E[\omega_V] \).

The inequality \( \frac{dN}{dS}|_{t}^{\infty} \geq E[\omega_V] \) 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
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:

$$S_m(t) = \Theta \int_0^t \left( \sum_{k=1}^{m-1} \frac{1}{k} \right) du,$$

with stationary limit $S_m = 20 N_s$, as $t \to \infty$, see Kaj and Mugal (2016), in particular, equation (21). It follows that the function $P_{AP}$ defined by:

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

measures the proportion of ancestral polymorphisms at time $t$ in the sample. The dynamics over time with $P_{AP}(0) = 1$ at the time of speciation and convergence toward 0 as $t \to \infty$ is fairly insensitive to sample size. For the larger sample sizes, <10% of ancestral polymorphisms remain at time $t = 2$, $P_{AP}(2) = 0.095$, and <1% at time $t = 5$, $P_{AP}(5) = 0.005$. The influence of ancestral polymorphisms on $dN/dS$ $\propto$ wears off at a similar rate and $dN/dS$ $\propto$ approaches $\Theta$ 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 following relation:

$$H_{11}^{m_1,0} \to \Theta \Theta \int_0^t q_V(y)(1 - q_V(y))\pi_V(y) dy, \quad t \to \infty,$$

which scales with population size. The limit is the same if we take first $m_1 \to \infty$ and then $t \to \infty$. 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 over-estimation is thus more complex and time-dependent.

In practice, this suggests that for sufficiently large divergence time where $P_{AP}$ small, polymorphism data allow to improve estimates of $\Theta$. 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 $dN/dS$ depends strongly on sample size for small $t$, while this dependence fades out as $t \to \infty$. The dependence on sample size can therefore be used as an indicator for estimation accuracy, when polymorphism data are used to make inference on $\Theta$. 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 $\Theta$ 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 $V$ on the negative real line. As a result, the scaled nonsynonymous sequence divergence is replaced by $dN_t = c^+ t + E[\Theta V] + E[F_{V}^{m_1,0}(t)]$, and the limiting $dN/dS$ ratio is replaced by:

$$\omega = E[\Theta V] + c^+,$$

where $c^+$ represents the rate of adaptive evolution. Second, we relax the assumption of instantaneous fixation of advantageous mutations and instead let $V$ 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
advantageous mutations $p$. The limiting $dN/dS$ ratio is represented by:

$$\omega = E[\omega_V] = (1-p)E[\omega_V|\mathcal{V} \leq 0] + pE[\omega_V|\mathcal{V} > 0].$$

Figure 4 illustrates the behavior of $dN/dS$ 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 4A shows the behavior of $dN/dS$ under the first model and $c^+=0.10$. Figure 4B shows the behavior of $dN/dS$ under the second model, $c=0.2$ and $p=0.1$. Parameters were chosen such that $\omega = 0.33$ 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. 4A), we also observe a strong effect of the parameter $c^+$ on the $dN/dS$ 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 $dN/dS$ at narrow time-scales. For smaller sample sizes, segregating polymorphisms attenuate the impact of the parameter $c^+$, and overestimation is less pronounced and $dN/dS < \omega$ 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, $dN/dS|_{t\rightarrow\infty}$ converges toward $\omega$ 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 $\omega$. In addition, under the scenario that mutations are mostly weakly advantageous and do not fix instantaneously (fig. 4B), we again observe that the accuracy of the estimates systematically improves with increasing sample size. We therefore conclude that $dN/dS$ 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 $\omega$, 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 $\omega$ at narrow time-scales.
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 $\omega$ 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 $\omega$ 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 $\omega$ 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.
Alternatively, if we relax the assumption of instantaneous substitutions that are advantageous, the ratio of the expected number of advantageous fixations and instead allow for a DFE extending on the positive and negative time interval is time-dependent even though the rate of adaptation 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 \( \omega_a = c^+ \). We discuss point estimation of \( \omega_a \) and \( \alpha_\infty \) in the MK framework, that is splitting up \( \omega \) into \( \omega_n \) and \( \omega_a \) 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 \( \omega_{na} = \mathbb{E}[\omega_{V}] \) is also known. The standard estimation procedure to draw inference on \( \omega_a \) and \( \alpha_\infty \) is motivated by equation (12), and amounts to replacing \( \mathbb{E}[\omega_{V}] \) by \( \omega_{na} \) (Keightley and Eyre-Walker 2007), putting

\[
\hat{\alpha}_t = 1 - \frac{\mathbb{E}[\omega_{V}]}{\mathbb{E}[\omega_{V}]} - \frac{\mathbb{E}[\omega_{V}]}{\mathbb{E}[\omega_{V}] - \mathbb{E}[\omega_{V}]]}. 
\]

Since \( \hat{\omega}_a(t) \) is derived from \( \mathbb{E}[\omega_{V}] \) by subtraction of a constant value, \( \hat{\omega}_a(t) \) shows the same dependence on sample size and divergence time as \( \mathbb{E}[\omega_{V}] \). The quantity \( \hat{z}_t \) is a time-dependent equivalent of \( \alpha \) 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 \( \alpha_t \) and \( \hat{\alpha}_t \) as functions of time and sample size, where the differences between \( \alpha_t \) and \( \hat{\alpha}_t \) arise due to the fact that computation of \( \hat{\alpha}_t \) is based on replacing \( \mathbb{E}[\omega_{V}] \) by \( \omega_{na} \). More specifically, figure 6A illustrates how the shape of \( \alpha_t \) varies with
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 $\tilde{x}_a$ falls below its asymptotic limit $x_\infty$. 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 $\tilde{x}_a$ stays above its asymptotic limit $x_\infty$ for larger sample sizes. Of note, $\tilde{x}_a$ for $m_1 \to \infty$ (blue line in fig. 6A) represents the actual proportion of advantageous nonsynonymous fixations in the focal species after split from the reference species. Besides, the fact that $\tilde{x}_a$ is close to $x_\infty$ 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 $\tilde{x}_a$ in relation to $x_a$ and $x_\infty$ is illustrated in figure 6B for the example of the EGP African human DFE, (Keightley and Eyre-Walker 2007), and $\omega_a = 0.10$. The significant deviation between $\tilde{x}_a$ and $x_a$ is pronounced for the smaller sample sizes but gradually decreases with larger samples. As for $dN/dS|_t$, we observe that $\tilde{x}_a$ overestimates $x_\infty$ 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 $\tilde{x}_a$, specifically the large sample approximation,

$$\tilde{x}_a = 1 - \frac{\omega_{na}}{dN/dS|_t},$$

is a relevant estimate of $x_a$. 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 $x_a$ at narrow time-scales. The potential relevance of $\tilde{x}_a$ as a point estimate of $x_\infty$, and with it the potential relevance of $\omega_a(t) = \tilde{x}_a \cdot dN/dS|_t$ as a point estimate of $\omega_a$, 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 $V$ of:

---

Fig. 6. (A) $x_a$ and (B) $\tilde{x}_a$ as functions of divergence time for a DFE based on human data with $\omega_a = c^+ = 0.10$. Dashed lines are used to represent $x_a$, solid lines to represent $\tilde{x}_a$. 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 $x_\infty$. (C and D) Comparison of different estimates of $\tilde{x}_a$ for $\omega_a = 0$ and $\omega_a = 0.10$, respectively. $\tilde{x}_a$ for $m_1 = 1$ (red) and $m_1 = 16$ representative for the large sample approximation (blue), and $x_{\text{cor}}$ (turquoise), together with $\tilde{x}_a$ for $m_1 = 16$ representative for the large sample approximation (blue dashed line). The black dashed line indicates the target parameter $x_\infty$. Note that in panel (C) the blue dashed line and the black dashed line take the constant value of 0.
\[
\lim_{N \to \infty} N E_{1/N}[\xi_i u_i] = \omega N \int_0^1 q(y)dy = 2 \int_0^1 q(y)dy,
\]

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 \(dN/dS\) by \(dN^\text{cor}/dS^\text{cor}\), where

\[
dN^\text{cor} = dN - 2 \int_0^1 \mathbb{E}(q(y)/y)dy, \quad dS^\text{cor} = dS - 2.
\]

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,

\[
dN/dS|_{t=} \leq dN/dS|_{t} \text{ if } \int_0^1 \mathbb{E}(q(y)/y)dy \geq dN/dS|_{t},
\]

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

\[
\int_0^1 \mathbb{E}(q(y)/y)dy \geq \max_{\xi \geq 0} dN/dS|_{t}.
\]

For our set of DFE parameters, however, the above criterion is not satisfied as the left hand side computes to \(\sim 0.26\), whereas the maximum of \(dN/dS\) typically is larger, compare figure 2.

Figure 6C and D show the match between the corrected estimate of \(\omega\) of Keightley and Eyre-Walker (2012) (referred to as \(\omega^\text{cor}\)) and \(\omega\), together with \(\tilde{\omega}\) for \(m_1 = 1\) (i.e., the standard uncorrected estimate of \(\omega\)), \(\tilde{\omega}\), for \(m_1 \to \infty\) (our proposed correction), and \(\omega_{\infty}\) for \(m_1 \to \infty\), for \(\epsilon^+ = 0\) and \(\epsilon^+ = 0.10\). This again illustrates that our proposed correction allows to correct for the misattribution of polymorphisms to sequence divergence, and that \(\tilde{\omega}\) for \(m_1 \to \infty\) is a relevant estimate of \(\omega_{\infty}\). 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 \(\omega_{\infty}\), and \(\omega_{\infty}\) at narrow time-scales.

A Protocol for the Estimation of \(\omega\) and \(\alpha\) for Closely Related Species

Our analytical results suggest that estimation of \(\omega\) for closely related species can be improved by incorporating information on segregating polymorphism. To obtain branch-specific estimates of \(dN/dS|_{\infty} = \omega\) 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 \(\omega\). 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 \(\omega\) small.

During the early phase of species divergence where ancestral polymorphisms contribute substantially to the amount of segregating polymorphisms, estimates of \(\omega\) 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, \(P_{DA}\) 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 \(\omega\)?

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 \(\omega\).

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 \(\theta\), \(t = D/(2\theta) - 1\) (Wiu 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 \(m_1 \geq 2\) 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 \(\omega\) 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 \(\omega\). The dependence on sample size thus constitutes a second criterion for the expected accuracy of estimates of \(\omega\). 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
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 \( \text{dN/dS} \) on sample size and further allows to compute an estimate of \( t = D/(2L) - 1 \) 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 \( \omega \) 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

\( \text{dN/dS} \) is estimated for all sets of triple alignments. For the computation of \( \text{dN/dS} \), a standard protocol can be followed such as Yang (2007). In case \( \text{dN/dS} \) values show a strong dependence on sample size, we recommend repeating the analysis with a more divergent reference species. As soon as \( \text{dN/dS} \) 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.

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 inappropriate and if \( \text{dN/dS} \) represents a reliable estimate of \( \omega \). 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 \( \text{dN/dS} \) 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 \( \omega \) 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 \( \omega \)

Based on these results, we propose the following step-by-step protocol for the estimation of \( \omega \) 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,
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 million years (My) of divergence to the outgroup is sufficiently large for lineage sorting between the focal and reference species. These results are in good agreement with our analytical observation that the contribution of polymorphisms biases estimates of \( \omega \) 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).

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 \( \omega \) 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).

Given the relationships (eq. 13), we expect to observe equivalent time and sample size dependence for estimates of \( \omega \) and \( x \) 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 DFEalpa 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 \( \mathbb{E}[\omega_t] = 0.056 \) as estimate of \( \omega_{na} \). Next, we computed \( \omega_{a} \) and \( x \) with the DFEalpa method (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009), which essentially is equivalent to solving for \( \omega_{a}(t) = dN/dS - \mathbb{E}[\omega_t] \). Thus, we naturally observed the same time and sample size dependence for estimates of \( \omega \), as for \( dN/dS \) (supplementary table 1).

Estimates of \( x \) were strongly sample size dependent with \( C. \) brachyrhynchos as reference and settled at \( \sim 0.54-0.56 \) 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 \( C. \) dauuricus, 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

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

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>( C. ) brachyrhynchos</th>
<th>( C. ) frugilegus</th>
<th>( C. ) dauuricus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.22</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>0.17</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>8</td>
<td>0.16</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>16</td>
<td>0.16</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>32</td>
<td>0.15</td>
<td>0.12</td>
<td>0.13</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>( C. ) brachyrhynchos</th>
<th>( C. ) frugilegus</th>
<th>( C. ) dauuricus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.74</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>2</td>
<td>0.67</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>8</td>
<td>0.65</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>16</td>
<td>0.65</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>32</td>
<td>0.62</td>
<td>0.54</td>
<td>0.56</td>
</tr>
</tbody>
</table>
information from polymorphisms and use dependence on sample size to obtain unbiased estimates of \( \omega \) and related measures.

Note, however, that unbiased estimation of \( \omega \) does not guard against incorrect estimation of \( \omega_a \) and \( \alpha \). The latter parameters further depend on the inferred DFE which is sensitive to demographic perturbation (Rouselle et al. 2018). The estimates of \( \omega_a \) (0.07) and \( \alpha \) (0.55) obtained under the assumption of constant population size are likely an over-estimation owing to population expansion in \( C. (c) \) 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 \( \omega_a \) and \( \alpha \) may thus reflect biases in the estimation of the DFE, and stress the importance of interpreting point estimates of \( \omega_a \) and \( \alpha \) 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 \( \omega \), and should therefore only be applied to closely related species if the time-dependence of \( \text{dN/dS} \) has been taken into account.

Conclusion

Our analytical results show that phylogenetic estimates of \( \omega \) 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 \( \alpha \) within the MK framework, which rely on phylogenetic estimates of \( \omega \) are also time-dependent. Expression of \( \text{dN/dS} \) as a function of sample size further shows that exclusion of polymorphic sites allows to reduce bias in the estimation of \( \omega \) and \( \alpha \) with respect to the contribution of lineage-specific and ancestral polymorphisms to sequence divergence. Moreover, while one cannot systematically improve estimation of \( \omega \) and \( \alpha \) 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 \( \omega \) and \( \alpha \), 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 \( \omega \) and \( \alpha \) 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 \( \omega \) 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 \( \text{dS} \) 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 \( \omega \) 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 \( \text{dN/dS} \) has also been observed in closely related bacteria (Rocha et al. 2006). Moreover, the \( \text{dN/dS} \) 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 \( \text{dN/dS} \) 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 \( \text{dN/dS} \) is one among several other biases in estimation of \( \omega \). For example, it has been shown that the assumption of stationary base composition biases estimates of \( \omega \) if base composition is nonstationary (Guéguen and Duret 2018). In addition, GC-biased gene conversion has been shown to influence \( \text{dN/dS} \) across a wide range of species (Ratnakumar et al. 2010; Lartillot 2013; Bolivar et al. 2019). Moreover, selection on codon usage and exonic splicing 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 multinationucleotide 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 accordan 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 5N 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 \( \mu \) 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 \( L = 10^3 \), 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 \( \alpha = 0.2 \). 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, and a scenario in the presence 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 dN/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 1C. brachyrhynchos, 1C. frugilegus, and 1C. 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 1C. (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 16C. (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 1C. brachyrhynchos, 1C. frugilegus, 1C. dauuricus individual, and for 16, 8, 4, and 1C. (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 1C. (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) (BLASTp version 2.5.0; Camacho et al. 2009), followed by orthology prediction using orthoAgoge 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. dauricus, and C. (corone) spp. (for 16, 8, 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 sampling 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 dN/dS**

We used the YN98+F3X4 model implemented in PAML (Yang 2007) to compute branch-specific maximum likelihood estimates of dN/dS 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 m1 = 1, 2, 8, 16, or 32, one reference species and the outgroup T. guttata (fig. 7).


This setting allows to compute branch-specific estimates of dN/dS for the C. (corone) spp. for three different branch lengths, C. (corone) spp. after the split from C. brachyrhynchos, C. (corone) spp. after the split from C. frugilegus and C. (corone) spp. after the split from C. dauricus, and five different sample sizes each (m1 = 1, 2, 8, 16, or 32). We retrieved bootstrap confidence intervals for estimates of dN/dS 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. dauricus 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. dauricus, 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. (cor) 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 $\xi$

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 $\xi$ using DFEalpha v.2.16 (Eyre-Walker and Keightley 2009) separately for dN/dS estimates of each sample size ($m = 1, 2, 8, 12, \text{or} 16$), 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


Downloaded from https://academic.oup.com/mbe/article-abstract/37/1/260/5556818 by guest on 26 March 2020