October, 3, 2019

Reconstructing the history of variation in effective population size along phylogenies.

Mathieu Brevet[1], Nicolas Lartillot[2]

[1] *Station d'écologie théorique et expérimentale UMR 5321, 09200 Moulis, France.*

[2] *Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive.*

nicolas.lartillot@univ-lyon1.fr

Running head: A phylogenetic history of N_e

Abstract

The nearly-neutral theory predicts specific relations between effective population size (N_e), and patterns of divergence and polymorphism, which depend on the shape of the distribution of fitness effects (DFE) of new mutations. However, testing these relations is not straightforward since N_e is difficult to estimate in practice. For that reason, indirect proxies for N_e have often been used to test the nearly-neutral theory, although with mixed results. Here, we introduce an integrative comparative framework allowing for an explicit reconstruction of the phylogenetic history of N_e, thus leading to a quantitative test of the nearly-neutral theory and an independent estimation of the shape parameter of the DFE. We applied our method to primates, for which the nearly-neutral predictions were mostly verified. Estimates of the shape parameter were compatible with independent measures based on site frequency spectra. The reconstructed history of N_e in primates seems consistent with current knowledge and shows a clear phylogenetic structure at the super-family level. Altogether, our integrative framework provides a quantitative assessment of the role of N_e in modulating patterns of genetic variation, while giving a synthetic picture of the long-term trends in N_e variation across a group of species.

Introduction

Effective population size (N_e) is a central parameter in population genetics and in molecular evolution, impacting both genetic diversity and the strength of selection (Charlesworth, 2009; Leffler *et al.*, 2012). The influence of N_e on diversity simply reflects the fact that larger populations can store more genetic variation, while the second aspect, efficacy of selection, is driven by the link between N_e and genetic drift: the lower the N_e, the more genetic evolution is influenced by the random sampling of individuals over generations. As a result, long-term trends in N_e are expected to have an important impact on genome evolution (Lynch *et al.*, 2011) and, more generally, on the relative contribution of adaptive and non-adaptive forces in shaping macro-evolutionary patterns.

The nearly-neutral theory proposes a simple conceptual framework for formalizing the role of selection and drift on genetic sequences. According to this theory, genetic sequences are mostly under purifying selection; deleterious mutation are eliminated by selection, whereas neutral and nearly-neutral mutations are subject to genetic drift and can therefore segregate and reach fixation. The inverse of N_e defines the selection threshold under which genetic drift dominates. This results in specific quantitative relations between N_e and key molecular parameters (Ohta, 1995). In particular, species with small N_e are expected to have a higher ratio of nonsynonymous (d_N) to synonymous (d_S) substitution rates and a higher ratio of nonsynonymous (π_N) to synonymous (π_S) nucleotide diversity. Under certain assumptions, these two ratios are linked to N_e through allometric functions in which the scaling coefficient is directly related to the shape of the distribution of fitness effects (DFE) (Kimura, 1979; Welch *et al.*, 2008; Castellano *et al.*, 2018).

The empirical test of these predictions raises the problem that N_e is difficult to measure directly in practice. In principle, N_e could be estimated through demographic and census data. However, the relation between census and effective population size is far from straightforward. Consequently, many studies which have tried to test nearly-neutral theory have used proxies indirectly linked to N_e. In particular, life history traits (LHT, essentially body mass or maximum longevity) are expected to correlate negatively with N_e (Waples *et al.*, 2013). As a result, d_N/d_S or π_N/π_S are predicted to correlate positively with LHT. This has been tested, leading to various outcomes, with both positive and negative results (Eyre-Walker *et al.*, 2002; Popadin *et al.*, 2007; Nikolaev *et al.*, 2007; Lartillot, 2013; Nabholz *et al.*, 2013; Romiguier *et al.*, 2014; Figuet *et al.*, 2016).

More direct estimations of N_e can be obtained from π_S since, in accordance with coalescent theory, $\pi_S = 4N_e u$ (with u referring to the mutation rate per site per generation). Thus, one would predict a negative correlation of d_N/d_S or π_N/π_S with π_S and a positive correlation between LHT and π_S. Such predictions have been tested in several previous studies (Romiguier *et al.*, 2014; Figuet *et al.*, 2016), with encouraging results. However, these more specific tests of the nearly-neutral theory are only qualitative, at least in their current form, in which N_e is indirectly accessed through π_S without any attempt to correct for the confounding effect of the mutation rate u and its variation across species.

In this study, we aim to solve this problem by using a Bayesian integrative approach, in which the joint evolutionary history of a set of molecular and phenotypic traits is explicitely reconstructed along a phylogeny. This method has previously been used to test the predictions of the nearly-neutral theory via indirect proxies of N_e (Lartillot, 2013; Nabholz *et al.*, 2013). Here, we propose an elaboration on this approach, in which the variation in the mutation rate per generation u is globally reconstructed over the phylogeny by combining the relaxed molecular clock of the model with data about generation times. This in turns allows us to tease out N_e and u from the π_S estimates obtained in extant species, thus leading to a complete reconstruction of the phylogenetic history of N_e and of its scaling relations with others traits such as d_N/d_S or π_N/π_S. Using this reconstruction, we can conduct a proper quantitative test of some of the predictions of the nearly-neutral theory and then compare our findings with independent knowledge previously derived from the analysis of site frequency spectra. The approach requires a multiple sequence alignment across a group of species, together with polymorphism data, ideally averaged over many loci to stabilize the estimates, as well as data about life-history traits in extant species and fossil calibrations. Here, we apply it to previously published phylogenetic and transcriptome data (Perelman *et al.*, 2011; Figuet *et al.*, 2016), focussing the analysis on primates, a group for which coding-sequence evolution has been suggested to be globally compatible with a nearly-neutral regime (Eyre-Walker & Keightley, 2009; Galtier, 2016).

Results

Life-history traits do not reflect effective population size in primates

Using an integrative comparative approach, based on a multivariate log-Brownian covariant model for rates and traits (Coevol, Lartillot & Poujol, 2011), we first tested the predictions of the nearly-neutral theory in primates, taking life history traits (age of sexual maturity, body mass, longevity and generation time, hereafter abbreviated as LHT) as tentative proxies for effective population size N_e. This first analysis gave limited insight into the problem. First, π_N/π_S does not correlate with any of the LHT. In the case of d_N/d_S, a significant positive correlation was observed only with longevity and generation time, whereas no correlation was seen with body mass. Finally, concerning π_S, only a marginally significant correlation with body mass was observed, which is surprisingly positive.

On the other hand, the correlations among molecular quantities are in agreement with the nearly-neutral predictions. Thus, d_N/d_S and π_N/π_S are positively correlated with each other, and both correlate negatively with π_S ($r^2 = 0.741$ for π_N/π_S and 0.449 for dN/dS). The somewhat weaker correlation of d_N/d_S with π_S could be due either to the presence of a minor fraction of adaptive substitutions or, alternatively, to a discrepancy between the short-term effects reflected in both π_S and π_N/π_S and long-term trends captured by dN/dS.

The simplest interpretation of these contrasted results is that the nearly-neutral model is essentially valid for primates, except that there is just no clear correlation between effective population size and body size or other related life-history traits in this group. Possibly, the phylogenetic scale might be too small to show sufficient variation in LHT that would be interpretable in terms of variation in N_e. Alternatively, N_e might be driven by other life-history characters (in particular, the mating systems), which may not directly correlate with body size. Of note, even in those cases where the estimated correlation of dN/dS or π_N/π_S with LHTs were in agreement with the predictions of the nearly-neutral theory (Eyre-Walker *et al.*, 2002; Popadin *et al.*, 2007; Nikolaev *et al.*, 2007; Lartillot, 2013; Nabholz *et al.*, 2013; Romiguier *et al.*, 2014; Figuet *et al.*, 2016), the reported correlation strengths were often weak, weaker than the correlations found here and elsewhere directly between π_S and π_N/π_S and dN/dS (Romiguier *et al.*, 2014; Figuet *et al.*, 2016). The use of LHTs as an indirect proxy of N_e therefore appears to typically result in a substantial loss of power. In contrast, the more direct use of polymorphism data at the exome level seems promising.

Teasing apart substitution rates, divergence times and effective population size

The correlation patterns shown by the three molecular quantities π_S, d_N/d_S and π_N/π_S are sufficiently clearcut to lend themselves to a more direct and more quantitative formalization of the underlying population-genetic mechanisms. In order to achieve this, an explicit estimate of the key parameter N_e, and of its variation across species, is first necessary. In this direction, a first simple but fundamental equation relates π_S with N_e:

$$\pi_S \quad = \quad 4\,N_e\,u. \tag{1}$$

In order to estimate N_e from equation 1, an estimation of u is also required. Here, it can be obtained by noting that:

$$u \;=\; r\,\tau, \tag{2}$$

where r is the mutation rate per site and per year and τ the generation time. Assuming that synonymous mutations are neutral, we can identify the mutation rate with the synonymous substitution rate dS, thus leading to:

$$u \;=\; dS\,\tau. \tag{3}$$

Finally, combining equations 1 and 3 and taking the logarithm gives:

$$\ln N_e \;=\; \ln \pi_S - \ln d_S - \ln \tau - \ln 4. \tag{4}$$

This expression suggests to operate a linear transformation on the three variables $\ln \pi_S$, $\ln d_S$ and $\ln \tau$, all of which are jointly reconstructed across the tree by the Bayesian integrative framework used here, so as to obtain a direct phylogenetic reconstruction of $\ln N_e$. In addition, since the transformation is linear, the correlation patterns between $\ln N_e$ and all other variables can be recovered by applying elementary matrix algebra to the covariance matrix estimated under the initial parameterization (see Materials and Methods).

The results of this linearly-transformed correlation analysis are gathered in Table 1. As predicted by the nearly-neutral theory, π_N/π_S and d_N/d_S are negatively correlated with N_e. Importantly, these two ratios correlate more strongly with our reconstructed N_e than with the originally observed π_S ($r^2 = 0.767$ and $r^2 = 0.501$ with N_e, against $r^2 = 0.741$ and $r^2 = 0.449$ with π_S) – as expected if the correlation with π_S is mediated by N_e. This also suggests that the approach has successfully teased out u and N_e from π_S, giving more confidence about the reconstructed history of N_e returned by the method (see below).

Estimating the shape parameter of the distribution of fitness effects

The quantitative scaling behavior of π_N/π_S and d_N/d_S as a function of N_e can be further investigated and interpreted in the light of an explicit mathematical model of the nearly-neutral regime. Such mathematical models, which are routinely used in modern Mac-Donald Kreitman tests (Charlesworth & Eyre-Walker, 2008; Eyre-Walker & Keightley, 2009; Halligan *et al.*, 2010; Galtier, 2016), formalize how demography modulates the detailed patterns of polymorphism and divergence. In turn, these modulations depend on the structure of the distribution of fitness effects (DFE) over non-synonymous mutations (Eyre-Walker & Keightley, 2007). Mathematically, the DFE is often modelled as a gamma distribution. The shape parameter of this distribution (usually denoted as β) is classically estimated based on empirical synonymous and non-synonymous site frequency spectra. Typical estimates of the shape parameter are of the order of 0.2 in humans (Boyko *et al.*, 2008; Eyre-Walker *et al.*, 2006), thus suggesting a strongly leptokurtic distribution, with the majority of mutations having either very small or very large fitness effects.

Here, we approach the problem from a different direction. Instead of the site frequency spectrum within a given species, we use the interspecific allometry of dN/dS and π_N/π_S with N_e to estimate the shape parameter of the DFE. To this aim, we make use a theoretical result (Kimura, 1979; Welch *et al.*, 2008), showing that, when β is small, both π_N/π_S and d_N/d_S scale as a function of N_e as a power-law, with a scaling exponent equal to β:

$$\pi_N/\pi_S = \kappa_1 N_e^{-\beta}, \tag{5}$$

$$dN/dS = \kappa_2 N_e^{-\beta}. \tag{6}$$

In the present case, an estimation of this scaling parameter can easily be obtained, by just computing the slope of the correlation between $\ln N_e$ and $\ln \pi_N/\pi_S$ or $\ln dN/dS$, which is then predicted to be equal to $-\beta$:

$$\ln dN/dS = -\beta \ln N_e + \ln \kappa_1, \tag{7}$$

$$\ln \pi_N/\pi_S = -\beta \ln N_e + \ln \kappa_2. \tag{8}$$

Of note, this specific relation with the shape of the DFE was used recently for analysing the impact of the variation in N_e along the genome in *Drosophila* (Castellano *et al.*, 2018).

As shown in Table 2, the resulting two estimates of β (based on dN/dS and π_N/π_S) are congruent with each other, with point estimates at 0.1 and 0.15, respectively, and credible intervals ranging from 0.01 to 0.23. Thus, they are compatible with (although a bit lower than) previously reported independent estimates obtained from site frequency spectra (also reported in Table 2). This important result further consolidates both our phylogenetic reconstruction of N_e and the idea of an essentially nearly-neutral regime in primates.

A mechanistic nearly-neutral phylogenetic codon model

Since all of the results presented thus far are compatible with a nearly-neutral regime, we decided to construct a mechanistic version of the model directly from first principles. Thus far, the whole set of variables of interest (dS, dN/dS, π_S, π_N/π_S and generation-time τ) were jointly reconstructed along the phylogeny, as a multivariate log-normal Brownian process with 5 degrees of freedom. Here instead, only three degrees of freedom are considered (which could be taken as u, N_e and τ), and then the empirically measurable molecular quantities dS, dN/dS, π_S and π_N/π_S are obtained as deterministic functions of these three fundamental variables, according to the equations introduced above (3, 4 and 7). The three structural parameters β, κ_1 and κ_2 involved in these equations (which represent the constraints induced by the assumed DFE), are also estimated. In practice, the model is more conveniently expressed, via a log-linear change of variables, in terms of π_S, π_N/π_S and τ as the three free variables, since these are directly observed in extant primates – from which dS,

dN/dS and N_e are then derived deterministically (see Materials and Methods).

As a result of the reduction in the dimensionality of the Brownian process, this mechanistic model is more constrained than the previous version explored above (which is more phenomenological in spirit) and is thus expected to have more statistical power. Indeed, compared to its phenomenological counterpart, the mechanistic model yields a more focussed estimate of β (Table 2), with a posterior median at 0.23 and a credible interval equal to $(0.19, 0.27)$. This is also closer to independent estimates obtained from site frequency spectra (Kimura, 1979; Welch *et al.*, 2008).

Phylogenetic Reconstruction of N_e

The marginal reconstruction of the history of N_e along the phylogeny of primates returned by the mechanistic model introduced in the last section is shown in Figure 1. More detailed information, with credible intervals, is given in Table 3 for several species of interest and key ancestors along the phylogeny.

N_e estimates for the four extant Hominidae (*Homo, Pan, Gorilla and Pongo*) are globally congruent with independent estimates based on other coalescent-based approaches (Prado-Martinez *et al.*, 2013). In particular, for Humans, N_e is estimated to be between 13 000 and 24 000. Concerning the successive last common ancestors along the hominid subtree, our estimation is also consistent with the independent-locus multi-species coalescent (Rannala & Yang, 2003), giving both similar point estimates and comparable credible intervals (in the range of 25 000 to 100 000 for the three ancestors). On the other hand, the pSMC approach (Li & Durbin, 2011), such as used in Prado-Martinez *et al.* (2013), tends to give systematically lower estimates for extant species and systematically higher estimates for ancestors, compared to both our approach (table 3) and the independent-locus multi-species coalescent (Rannala & Yang, 2003).

Zooming out over the entire primate phylogeny, we observe that, starting with a point estimate at around 100 000 in the last common ancestor of primates, N_e then goes down in Haplorrhini, stabilizing at around 80 000 in Cercopithecidae, 40 000 in Hominoidae (going further down more specifically in Humans), and 40 000 in Platyrrhini (old-world monkeys). Conversely, in Strepsirrhini, N_e tends to show higher values, staying at 100 000 in Lemuroidae and going up to 160 000 to 200 000 in Lorisiformes. Finally, large effective sizes are estimated for the two isolated species *Tarsius* (300 000) and *Daubentonia* (greater than 10^6) – although these latter estimates may not be so reliable, owing to the very long branches leading to these two species.

The reconstruction of N_e shown in figure 1 mirrors the patterns of dN/dS estimated over the tree (supplementary figure 1). Thus, dN/dS starts in the range of 0.25 in the early primate lineages. On the side of Haplorrhini, it goes up to 0.30 in Simiiformes (monkeys), stabilizes around this value in Catarrhini (old-world monkeys), goes further up to 0.32 in Platyrrhini (new-world monkeys). On the other branch (Strepsirrhini), the dN/dS is around the value of 0.25 in Lemuroidae, versus 0.22 in Lorisidae. The two species *Daubentonia* and *Tarsius* show the lowest dN/dS values (around 0.20).

In order to get some insight about how much dN/dS and π_N/π_S inform the reconstructed history of N_e, an alternative version of the model was also explored, in which all time-dependent variables are assumed to evolve independently from each other. Under this control, N_e is thus only informed at the tips by $\pi_S = 4N_e u$ and by the estimates of u implied by the relaxed clock and data about generation times. Compared to the mechanistic version just presented, this uncoupled model gives a globally similar result in terms of point estimation (supplementary figures 2), except for Lemuroidae, which show globally higher N_e estimates than Lorisiformes (whereas the converse was true under the model informed by dN/dS) and a lower N_e in the last common ancestor of primates (40 000, versus 100 000 under the integrative model), as well as in the two species *Daubentonia* and *Tarsius*. The deviations between the two models give some insight about the respective roles of π_S and dN/dS in informing the reconstruction. In the case of Strepsirrhini, π_S is globally higher in two of the three Lemuroidae for which polymorphism data are available (*Propithecus* and *Varecia*) than in the two Lorisiformes *Galago* and *Nyctecibus*. In contrast, dN/dS is globally lower in Lorisiformes than in Lemuroidae. Thus, the two sources of molecular information tend to locally contradict each other in this case. Concerning *Daubentonia* and *Tarsius*, no information about π_S is available, but the dN/dS is particularly low for these two species, which explains their particularly high N_e estimates specifically under the reconstruction informed by dN/dS. Finally, the reconstruction under the phenomenological version of the model gives an intermediate picture (supplementary figure 3), showing the patterns that are in common between the mechanistic and the uncoupled versions of the model such as just described (i.e. a higher N_e in Strepsirrhini, a lower N_e in Haplorrhini and more particularly in Platyrrhini), while striking a compromise between the opposite signals of π_S and dN/dS in Strepsirrhini by estimating a similar range of effective population sizes in Lorisiformes and Lemuroidae.

Interestingly, under the uncoupled model, there is a substantial uncertainty about the estimation of the instant values of N_e across all nodes of the tree: the 95% credible intervals span one order of magnitude on average. This uncertainty is somewhat reduced under the reconstructions relying on the additional information contributed by dN/dS and π_N/π_S, quantitatively, by 30% under the phenomenological, and by 50% under the mechanistic covariant models. In the end, there is thus on average a factor 5 between the lower and the upper bound of the 95% credible intervals on N_e estimates under the most constrained (mechanistic) model. Concerning the deep branches of the tree, most of this reduction in uncertainty is primarily contributed by dN/dS – which thus gives an idea of how much information can be extracted from multiple sequence alignments about very ancient population genetic regimes.

Discussion

In population genetics, effective population size (N_e) plays two different roles. First, N_e directly determines how much genetic diversity can be maintained in a given population. Second, its inverse, $1/N_e$, defines the relative strength of random drift, compared to selection. Quantitatively, the first role of N_e is reflected in the synonymous nucleotide diversity $\pi_S = 4N_e u$. However, as can be seen from this equation, diversity also depends on the mutation rate u per generation. A first key result obtained here, through an integrative dating and comparative analysis, is a separate estimation of N_e and u from π_S, thus leading to a quantitative reconstruction of the history of N_e along the phylogeny of primates.

As for the second role of N_e, it is reflected in the two ratios of non-synonymous over synonymous polymorphism (π_N/π_S) and divergence (d_N/d_S). Under relatively mild asumptions, the nearly-neutral theory predicts that these two quantities should decrease with N_e, according to a power-law with a scaling coefficient equal to the shape parameter of the distribution of fitness effects (Kimura, 1979; Welch *et al.*, 2008). Another key result obtained in the present work is a test of the validity of those predictions of the nearly-neutral theory, leading to an estimate of the scaling coefficient of the order or 0.2, thus compatible with previous estimates based on site frequency spectra (Boyko *et al.*, 2008; Eyre-Walker *et al.*, 2006). Altogether, our results therefore confirm previous findings from empirical population genetics and suggest that primate coding sequences are essentially evolving under a nearly-neutral regime.

Our model-based inference is potentially subject to several sources of bias. First, it depends on a sufficiently accurate estimation of the mutation rate u, which is itself obtained via the dating component of the model. Molecular dating is known to be sensitive to the exact assumptions of the dating model, in particular the prior on divergence times and fossil calibrations (Inoue *et al.*, 2010). Another potential issue is that short-term N_e (such as reflected by π_S) may be strongly dependent on recent demographic events (Charlesworth, 2009) and may thus not be identical with long-term N_e (such as reflected by d_N/d_S). This might be one of the reasons why d_N/d_S shows a weaker correlation with π_S than π_N/π_S (Table 1). A possible improvement of our model in this direction would consist in allowing for an additional level of variability at the leaves, representing the mismatch between long- and short-term N_e.

Finally, the dN/dS may contain a fraction of adaptative substitutions, susceptible to distort the relation between dN/dS and N_e. Although a relatively minor problem in the case of primates (Eyre-Walker & Keightley, 2009; Galtier, 2016), adaptive substitutions might be a far more important issue when applying the method to other phylogenetic groups. Here also, the model could be further elaborated, by explicitly including an adpative component to the total dN/dS. Quite interestingly, the resulting model could then be seen as an integrative multi-species version of the Mac-Donald Kreitman test, returning an estimate of the history of the adaptive substitution rate over the phylogeny – which could then be compared with independent estimates based on pairs of sister species (Charlesworth & Eyre-Walker, 2008; Eyre-Walker & Keightley, 2009; Halligan *et al.*, 2010; Galtier, 2016).

Thus, overall, further developments are needed, and there are several reasons to be cautious. Nevertheless, the phylogenetic history of N_e obtained here provides a first synthetic picture about the evolution of N_e at the scale of an entire mammalian order. What this image suggests in the present case (Figure 1) is that N_e stays within a 10-fold range across primates (roughly between 20 000 and 200 000), with a simple structure at the sub-order / super-family level (higher values in Strepsirrhini and lower values in Haplorrhini, particularly in Hominoidae and, less expectedly, in Platyrrhini). Such broad-scale reconstructions, as opposed to focussed estimates in isolated species using coalescent-based methods, are potentially useful in several respects. First, they provide a basis for further testing some of the key ideas about the role of genetic drift in genome evolution

(Lynch *et al.*, 2011). Second, the integrative framework could be augmented with trait-dependent diversification models (Fitzjohn, 2010), so as to examine the role of N_e in speciation and extinction patterns. Finally, the underlying codon model could be further elaborated (Rodrigue *et al.*, 2010; De Maio *et al.*, 2013), so as to achieve a more complete integration of polymorphism and divergence in model-based molecular evolutionary studies (Hernandez *et al.*, 2011).

Materials & Methods

Coding sequence data, phylogenetic tree and fossil calibration

The coding sequences were taken from Perelman *et al.* (2011) and modified. It consists in a modified subset, codon compliant, based on 54 nuclear autosomal genes in 61 species of primates, and of a total length 15.9 kb. We used the tree topology published by Perelman et al. (itself based on a maximum likelihood analysis), as well as the eight fossil calibrations that were used in this previous study to estimate divergence times. These calibrations were encoded as hard constraints on the molecular dating analysis.

Life History Traits

We used four life history traits (LHT) in this study. Adult body mass (as a proxy for body mass, 16 missing values), maximum recorded lifespan (ML, as a proxy for longevity, 19 missing values) and female age of sexual maturity (ASM, 26 missing values) were obtained from the AnAge database (de Magalhaes & Costa, 2009). Generation time (τ) was calculated from maximum longevity and age at maturity following a method detailed by UICN (Pacifici *et al.*, 2013):

$$\tau = ML \times 0.29 + ASM.$$

Estimation of Polymorphism (π_S and π_N/π_S)

The estimates of the synonymous nucleotide diversity π_S and the ratio of non-synonymous over synonymous diversity π_N/π_S and π_S of 9 primate species were obtained from Figuet *et al.* (2016). For each species, estimates were based on 4 individuals. We matched the polymorphism data reported by Figuet et al for the three species *Pan troglodytes*, *Propithecus vereauxi coquereli* and *Eulemur mongoz*, to *Pan paniscus*, *Propithecus verreauxi* and *Eulemur rufus*, respectively, from the Perelman et al multiple sequence alignment. Of note, π_S and π_N/π_S were estimated in Figuet et al on different subset of genes / contigs, so as to avoid the artifactual correlations that would be induced between these two parameters by shared data sampling error.

Models

General Principles of the integrative strategy (Coevol)

Coevol is a Bayesian inference software program based on a comparative approach applied to molecular data (Lartillot & Poujol, 2011). The principal aims of this program are to estimate ancestral continuous traits and to determine the correlations between the different molecular parameters along a phylogeny. Coevol follows a generative modeling strategy to describe the evolution of continuous traits along a phylogenetic tree. The joint evolutionary process followed by traits is modeled as a log-Brownian process. This process is parameterized by a variance-covariance matrix, which thus captures the correlations between traits corrected for phylogenetic inertia. Sequence evolution is described by a codon model (with a separate evolution of d_S and d_N/d_S along the tree). The model is conditioned on data obtained in current species (multiple sequence alignments and quantitative traits), with fossil calibrations, and samples from the joint posterior distribution are obtained by Markov Chain Monte-Carlo (MCMC). The analysis returns an estimation of correlation patterns between traits (covariance matrix) and a reconstruction of the history of the traits along the phylogeny.

Ex-post log-linear transformation of the correlation analysis

As a first step, we ran the original model (such as defined by the current version of Coevol), which we call the *phenomenological* model in the following. We then used the outputs from these runs to estimate N_e and its correlation patterns with other traits. At any given time, the multivariate Brownian process is structured as follows:

$$\begin{cases} X_1 &= \ln dS \\ X_2 &= \ln dN/dS \\ X_3 &= \ln \tau \\ X_4 &= \ln \pi_S \\ X_5 &= \ln \pi_N/\pi_S \\ \dots \end{cases}$$

Where τ is the generation time and entries X_i for $i > 5$ correspond to all other LHT. Using equation 4, which gives $\ln N_e$ as a linear combination of the components of the Brownian process, we can define the following linear change of variables:

$$X = \begin{pmatrix} \ln dS \\ \ln dN/dS \\ \ln \tau \\ \ln \pi_S \end{pmatrix} \quad \rightarrow \quad Y = \begin{pmatrix} \ln dS \\ \ln dN/dS \\ \ln \tau \\ \ln N_e \end{pmatrix}$$

where

$$Y_4 = X_4 - X_1 - X_3 + K,$$

where K is a numerical constant (depending on the absolute time scale). So, if we define the matrix A:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 1 \end{pmatrix}$$

then $Y = AX + K$. Finally, since X follows a Brownian process (parameterized by a variance-covariance matrix Σ_X), according to elementary multivariate normal theory, Y follows a Brownian process parameterized by a variance-covariance matrix $\Sigma_Y = A \times \Sigma_X \times A^{-1}$. In practice, we added a new method in Coevol to read the output and apply the linear transformation (from X and Σ_X to Y and Σ_Y) on each sample from the posterior distribution. This allows us to produce a reconstruction of N_e (posterior mean, credible intervals) and of the correlation matrix Σ_Y.

Mechanistic Nearly-Neutral Model

This alternative model uses the original Coevol framework but introduces additional constraints, such that some of the parameters are deduced through deterministic relations implying other Brownian dependent parameters. Specifically, the Brownian free variables are now:

$$\begin{cases} X_1 &= \ln \pi_S \\ X_2 &= \ln \pi_N / \pi_S \\ X_3 &= \ln \tau \end{cases}$$

Then, using equations 3, 4 and 7, the other variables of interest can be expressed as deterministic functions:

$$\begin{cases} \ln N_e &= -1/\beta \left(\ln \pi_N / \pi_S + \ln \kappa_2 \right), \\ \ln dS &= \ln \pi_S - \ln 4N_e - \ln \tau, \\ \ln dN/dS &= -\beta \ln N_e + \ln \kappa_1. \end{cases}$$

This model has three structural free parameters, β, κ_1 and κ_2, which were each endowed with a normal prior, of mean 0 and variance 1.

Uncoupled Model

The *uncoupled* model, already implemented in Coevol, is similar to the phenomenological version of the model, except that the variables of interest (dS, dN/dS, π_S, π_N/π_S and τ) are modelled as independent Brownian processes along the tree. Equivalently, we use a multivariate Brownian model with a diagonal covariance matrix (see Lartillot and Poujol, 2011, for details).

All these additional models and outputs were written in C++, as addition to the Coevol software programming environment. The original source code, as well as the modifications introduced for the present study, are available on github: `https://github.com/bayesiancook/coevol`.

Markov Chain Monte-Carlo (MCMC) and post-analysis

At least two independent chains were run under each model configuration. Convergence of the chains was first checked visually (a burnin of approximately 10% of the total run was chosen) and quantified using Coevol's program Tracecomp (effective sample size greater than 200 and maximum discrepancy smaller than 0.2 across all pairs of runs and across all statistics). We used the posterior median as the point estimate. The statistical support for correlations is assessed in terms of the posterior probability of a positive or a negative correlation. Correlation slopes are estimated using the major axis method. The slope is estimated for each covariance matrix sampled from the distribution, which then gives a sample from the marginal posterior distribution over the slope.

Software and data availability

Coevol (Lartillot & Poujol, 2011) is an open source program available from github.com/bayesiancook/coevol.git. All models and data used here are accessible through the branch coevolNe.

Authors contribution

All modifications of Coevol and new models made in this study are attributable to MB, who also gathered and formatted the data and conducted all analyses, in the context of an internship (master Biosciences of École Normale Supérieure de Lyon). MB and NL both contributed to the writing of the manuscript.

Competing interests

The Authors have no competing interests.

Acknowledgements

We wish to thank Emeric Figuet and Nicolas Galtier for sharing polymorphism data (PopPhyl project), Emmanuel Douzery and Frederic Delsuc for editing the multiple sequence alignment of Perelman et al to make it codon-compliant, and Laurent Duret and Thibault Latrille for their input on this work and their comments on the manuscript. Phylogenetic analyses were conducted using the computing facilities of the CC LBBE/PRABI.

Funding

French National Research Agency, Grant ANR-15-CE12-0010-01 / DASIRE.

References

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J. *et al.* 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.*, **4**(5), e1000 083.

Castellano, D., James, J. & Eyre-Walker, A. 2018 Nearly Neutral Evolution across the Drosophila melanogaster Genome. *Mol. Biol. Evol.*, **35**(11), 2685–2694.

Charlesworth, B. 2009 Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, **10**(3), 195–205.

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

de Magalhaes, J. & Costa, J. 2009 A database of vertebrate longevity records and their relation to other life-history traits. *J. Evol. Biol.*, **22**, 1770–1774.

De Maio, N., Schlötterer, C. & Kosiol, C. 2013 Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.*, **30**(10), 2249–2262.

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

Eyre-Walker, A. & Keightley, P. D. 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., Keightley, P. D., Smith, N. G. C. & Gaffney, D. 2002 Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.*, **19**(12), 2142–2149.

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.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H. & Galtier, N. 2016 Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Mol. Biol. Evol.*

Fitzjohn, R. G. 2010 Quantitative traits and diversification. *Syst. Biol.*, **59**(6), 619–633.

Galtier, N. 2016 Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLoS Genet.*, **12**(1), e1005 774.

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B. & Keightley, P. D. 2010 Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet.*, **6**(1), e1000 825.

Hernandez, R. D., Andolfatto, P. & Przeworski, M. 2011 A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.*, **7**(12).

Inoue, J., Donoghue, P. C. J. & Yang, Z. 2010 The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.*, **59**(1), 74–89.

Kimura, M. 1979 Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl. Acad. Sci. USA*, **76**(7), 3440–3444.

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.

Lartillot, N. & Poujol, R. 2011 A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, **28**(1), 729–744.

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P. & Przeworski, M. 2012 Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.*, **10**(9), e1001 388.

Li, H. & Durbin, R. 2011 Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357), 493–496.

Lynch, M., Bobay, L.-M., Catania, F., Gout, J.-F. & Rho, M. 2011 The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet*, **12**, 347–366.

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.

Nikolaev, S. I., Montoya-Burgos, J. I., Popadin, K., Parand, L., Margulies, E. H., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program & Antonarakis, S. E. 2007 Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc. Natl. Acad. Sci. USA*, **104**(51), 20 443–20 448.

Ohta, T. 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.*, **40**, 56–63.

Pacifici, M., Santini, L., Di Marco, M. & Baisero, D. 2013 Generation length for mammals. *Nature.*

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., Kessing, B., Pontius, J., Roelke, M. *et al.* 2011 A Molecular Phylogeny of Living Primates. *PLoS Genet.*, **7**(3), e1001 342.

Popadin, K., Polishchuk, L., Mamirova, L., Knorre, D. & Gunbin, K. 2007 Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc. Natl. Acad. Sci. USA*, **104**(33), 13 390.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D. *et al.* 2013 Great ape genetic diversity and population history. *Nature*, **499**(7459), 471–475.

Rannala, B. & Yang, Z. 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**(4), 1645–1656.

Rodrigue, N., Philippe, H. & Lartillot, N. 2010 Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA*, **107**(10), 4629–4634.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L. *et al.* 2014 Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*.

Waples, R. S., Luikart, G., Faulkner, J. R. & Tallmon, D. A. 2013 Simple life-history traits explain key effective population size ratios across diverse taxa. *Proc. Roy. Soc. London Ser. A*, **280**(1768), 20131 339.

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

Tables

Table 1. Correlation coefficients between dS, dN/dS, πS, $\pi N/\pi S$, N_e and life-history traits.

cov.	d_S	d_N/d_S	Mat.	Mass	Longev.	π_S	π_N/π_S	Gen.	N_e
d_S	0.973	0.0825	-0.394	-1.38**	-0.213	-0.841*	0.241*	-0.243	-1.57**
d_N/d_S	...	0.095	0.0543	0.069	0.134**	-0.25**	0.0701*	0.119**	-0.45**
Maturity	...	...	0.876	1.38**	0.382**	0.203	0.0129	0.455**	0.141
Mass	...	...	...	5.19	0.886**	1.14*	-0.231	0.955**	1.56
Longevity	...	...	...	...	0.518	-0.224	0.0607	0.489**	-0.5
π_S	...	...	...	...	...	1.51	-0.40**	-0.148	2.49**
π_N/π_S	...	...	...	...	...	...	0.139	0.0517	-0.69**
Gen. Time	...	...	...	...	...	...	...	0.479	-0.384
N_e	...	...	...	...	...	...	...	...	4.45

r^2	d_S	d_N/d_S	Mat.	Mass	Longev.	π_S	π_N/π_S	Gen.	N_e
d_S	...	0.0724	0.179	0.377**	0.0894	0.489*	0.428*	0.126	0.557**
d_N/d_S	...	...	0.0449	0.0139	0.383**	0.449**	0.394*	0.331**	0.501**
Maturity	...	...	...	0.411**	0.316**	0.0262	0.0025	0.487**	0.00377
Mass	...	...	...	...	0.286**	0.154*	0.0635	0.36**	0.0967
Longevity	...	...	...	...	...	0.0676	0.0581	0.964**	0.105
π_S	...	...	...	...	...	...	0.741**	0.0331	0.929**
π_N/π_S	...	...	...	...	...	...	...	0.0462	0.767**
Gen. Time	...	...	...	...	...	...	...	...	0.0671

Asterisks indicate strength of support (* $pp > 0.95$, ** $pp > 0.975$).

Table 2. Alternative estimates of the shape parameter β of the distribution of fitness effects.

method	point estimate	credible/confidence interval
phenomenological model (dN/dS)	0.11	(0.01, 0.21)
phenomenological model ($\pi N/\pi S$)	0.16	(0.07, 0.23)
mechanistic model	0.25	(0.19, 0.28)
Eyre-Walker et al	0.23	(0.19, 0.27)
Boyko et al	0.18	(0.16, 0.21)

Table 3. Estimates of effective population size ($\times 10^{-3}$, posterior median and 95% credible interval) for several extant and ancestral species.

species	phenomenological	mechanistic	uncoupled	coalescent[a]	m.sp. coal[b]	coalHMM[c]
Homo	19 (13,26)	14 (9,23)	13 (9,21)	(13,16)		8
Pan	44 (36,55)	72 (50,106)	66 (43,109)	(31,69)		30
Gorilla	68 (26,163)	106 (34,327)	41 (18,100)	(28.57)		21
Pongo	38 (15,91)	47 (19,124)	35 (11, 116)	(42,85)		19
Homo - *Pan*	39 (24,64)	44 (25,82)	34 (21, 56)		(10,47)	50
Homo - *Gorilla*	43 (24,72)	40 (25,99)	36 (21,63)		(27.61)	47
Homo - *Pongo*	44 (21,81)	54 (26,127)	47 (21,108)			125
Hominoidae	55 (19,69)	77 (34,206)	58 (24, 149)			
Cercopithecidae	79 (44,145)	84 (40,171)	72 (34, 152)			
Catarrhini	74 (37,142)	76 (32,189)	58 (25,141)			
Platyrrhini	37 (19,69)	30 (15,61)	32 (13,82)			
Simiiformes	59 (28,123)	39 (14,100)	36 (14,98)			
Haplorrhini	90 (39,230)	33 (8,111)	33 (11,121)			
Lorisiformes	143 (68,348)	50 (13,151)	50 (19,135)			
Lemuroidae	91 (41,195)	77 (27,205)	126 (50,322)			
Strepsirrhini	102 (43,262)	28 (59,100)	33 (12,109)			
Primates	93 (38,229)	29 (66,99)	32 (11,115)			

[a] from Prado-Martinez et al, 2013, table 1

[b] from Rannala and Yang, 2003

[c] from Prado-Martinez et al, 2013, figure 2

Figure

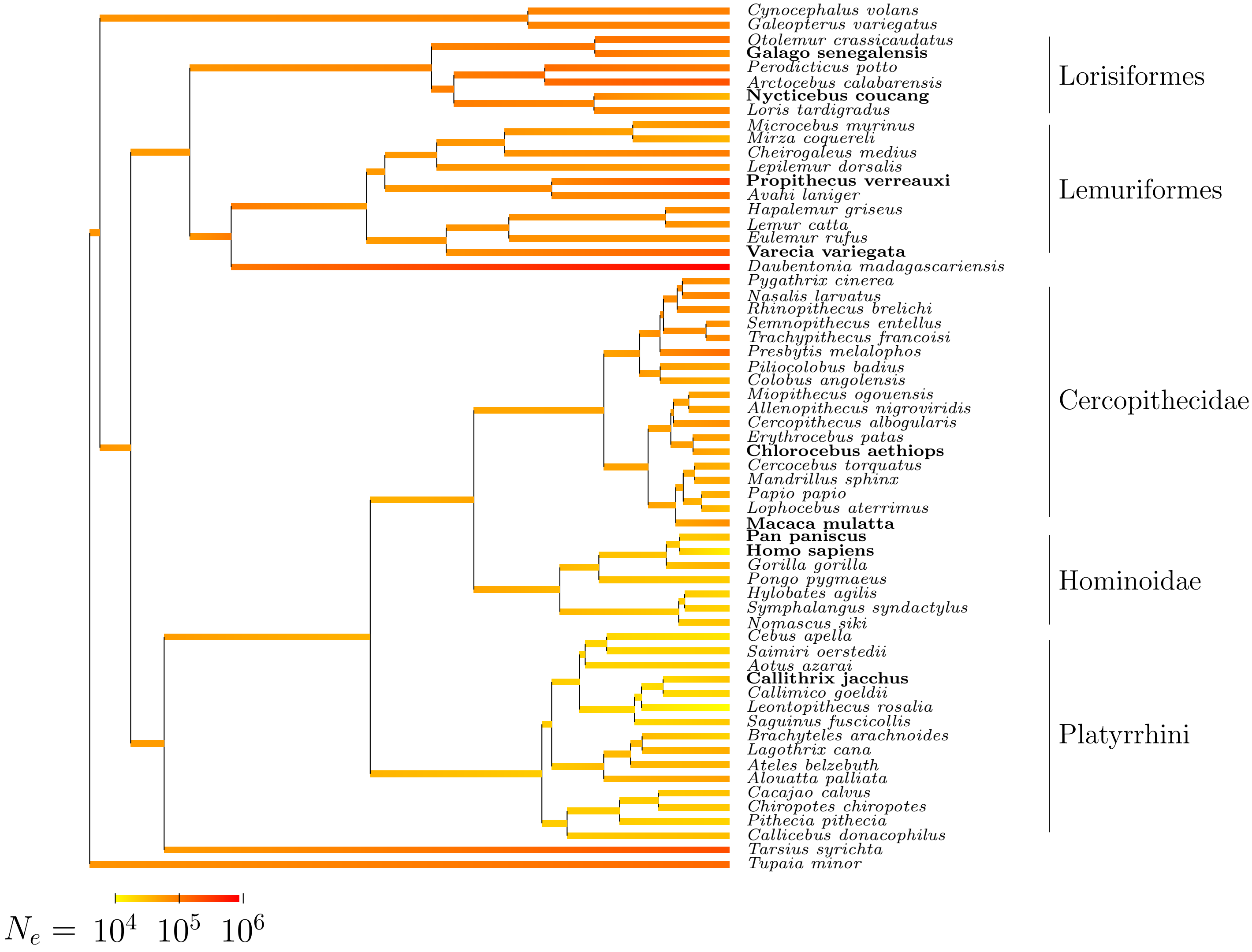

Figure 1. Reconstructed phylogenetic history of N_e (posterior median estimate) under the mechanistic nearly-neutral model. Species for which transcriptome-wide polymorphism data were used are indicated in bold face.