vol. 191, no. 4 the american naturalist april 2018

Model Adequacy and Microevolutionary Explanations for Stasis in the Fossil Record

Kjetil Lysne Voje,[1,2,*] Jostein Starrfelt,[1] and Lee Hsiang Liow[1,3]

1. Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066 Blindern, NO-0316 Oslo, Norway; 2. Department of Earth Sciences, Natural History Museum, Cromwell Road, SW7 5BD London, United Kingdom; 3. Natural History Museum, University of Oslo, PO Box 1172 Blindern, 0318 Oslo, Norway

Submitted April 7, 2017; Accepted November 3, 2017; Electronically published February 5, 2018

Online enhancements: supplemental material. *Dryad data:* http://dx.doi.org/10.5061/dryad.r5d10.

abstract: Long-term phenotypic stasis is frequently observed in the fossil record, but not readily predicted from microevolutionary theory. To test competing explanations for stasis on macroevolutionary timescales we need reliably estimated parameters from appropriate evolutionary models that adequately describe the evolutionary trait dynamics. Here, we develop tests to assess the adequacy of the most commonly used stasis model in evolutionary biology and apply them to time series of phenotypic traits from fossil lineages. Of the 572 fossil time series we analyzed from the literature, 263 time series showed a better fit to the stasis model relative to alternative models, but only 172 of those fitted the stasis model in both relative and absolute terms. The estimated trait variances from these 172 time series do not correlate with rough proxies of effective population size. Our preliminary investigation of the fixed-optimum hypothesis hence fails to give empirical support to the idea that genetic drift around a constant trait optimum is an explanation for stasis in the fossil record. We argue that optima following stationary processes on the adaptive landscape is a viable hypothesis for stasis that needs further investigation. We end by discussing how investigations of model adequacy can be a valuable approach for increasing our understanding of the dynamics of the adaptive landscape on macroevolutionary timescales.

Keywords: macroevolution, adaptive landscape, paleontology, stabilizing selection, phenotypic evolution.

Introduction

Understanding long-term morphological stasis (i.e., low or no net evolution in a lineage on macroevolutionary timescales; Eldredge and Gould 1972; Gould and Eldredge 1977; Gould 2002) remains a major challenge in evolutionary biology (Bradshaw 1991; Hansen and Houle 2004; Voje 2016). Studies on generational timescales tend to find substantial additive genetic variation in most traits (Houle 1992; Lynch and Walsh 1998; Hansen et al. 2011), which is indicative of a large potential for traits to evolve. This potential is commonly confirmed by the observation of rapid changes in phenotypic traits over a few generations in natural populations (e.g., Hendry and Kinnison 1999; Kinnison and Hendry 2001) and in many artificial selection experiments (e.g., Hill and Caballero 1992). Moreover, population genetic theory predicts faster rates of change in quantitative traits than what is commonly observed in the fossil record (Lynch 1990; Cheetham et al. 1994). Long-term stasis is therefore not readily predicted from microevolutionary theory or empirical insights on short evolutionary timescales. Yet, stasis appears to be a common mode of evolution in the fossil record (Hunt 2007; Hopkins and Lidgard 2012; Hunt et al. 2015; Voje 2016).

Part of the challenge in explaining stasis is to ascertain the validity of various competing hypotheses and to evaluate their relative importance (Hunt and Rabosky 2014). A much-invoked explanation for stasis is genetic and developmental constraints (Eldredge and Gould 1972; Hansen and Houle 2004). For example, stabilizing selection on pleiotropically linked traits may severely reduce the amount of free additive genetic variance available for selection to act on (Hansen 2003; Hansen et al. 2003). Genetic covariation among traits is the basis for understanding multivariate trait evolution (e.g., Lande 1979; Lande and Arnold 1983; Blows 2007), and genetic correlations have been shown to influence evolutionary trajectories on macroevolutionary timescales (e.g., Schluter 1996; Blows and Higgie 2003; Hansen 2003; Hansen et al. 2003; Marroig and Cheverud 2005; Grabowski et al. 2011; Hansen and Voje 2011; Grabowski 2016). Genetic covariances and reduction in the amount of freely available additive genetic variance can reduce the response to selection, but to what extent such constraints can explain morphological stasis on million-year timescales is debated (Hansen 2012). Another hypothesis for stasis invokes homogenizing gene flow

[*] Corresponding author; e-mail: k.l.voje@ibv.uio.no.

ORCIDs: Voje, http://orcid.org/0000-0003-2556-3080; Starrfelt, http://orcid.org/0000-0003-3714-4844; Liow, http://orcid.org/0000-0002-3732-6069.

Am. Nat. 2018. Vol. 191, pp. 509–523. © 2018 by The University of Chicago. 0003-0147/2018/19104-57656$15.00. All rights reserved. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution. For commercial use, contact journalpermissions@press.uchicago.edu. DOI: 10.1086/696265

between locally adapted subpopulations (Lieberman et al. 1995; Lieberman and Dudgeon 1996; Eldredge et al. 2005; Futuyma 2010). Some evidence that evolutionary changes within subpopulations may be swamped by gene flow have been found in studies of brachiopods (Lieberman et al. 1995), but the homogenizing effect of gene flow on macroevolutionary timescales has generally been little explored. Stabilizing selection on the focal trait is the most commonly invoked explanation for long-term stasis (Charlesworth et al. 1982; Smith 1983; Haller and Hendry 2014). The strictest version of the stabilizing-selection hypothesis assumes that stasis is a result of stabilizing selection around a fixed optimum over long timescales (Haller and Hendry 2014). According to the fixed-optimum hypothesis, deviations from the optimum are caused by drift and the size of the deviations from the optimum are thus predicted to be inversely proportional to the effective population size. A second hypothesis also invoking stabilizing selection claims that stasis is the result of a population tracking a fluctuating optimum via directional selection (e.g., Hunt 2007; Hunt and Rabosky 2014; Voje 2016). The role of stabilizing selection in the fluctuating-optimum hypothesis is to keep the population at the optimum when the population has reached the adaptive peak.

As pointed out by Hunt and Rabosky (2014), the fixed-optimum hypothesis yields at least one testable prediction that is not predicted by the three alternative hypotheses: since the effect of genetic drift decreases when the effective population size gets larger, the fixed-optimum hypothesis predicts a negative correlation between effective population size and deviations from the optimum during periods of stasis. In other words, if the fixed-optimum hypothesis is true, a smaller population size predicts on average larger deviations from the optimum compared to a larger population, as long as the average curvature of the adaptive peak is similar. A test of the fixed-optimum hypothesis could therefore be done by a comparison of estimated deviations from a fixed optimum between groups of taxa that differ vastly in their effective population size.

Testing the Fixed-Optimum Hypothesis: Adequate Estimates of Evolutionary Change

There is a long history of research on measuring rates of evolution in the fossil record (e.g., Haldane 1949; Gingerich 1983, 1993, 2001, 2009; Bookstein 1987; Lynch 1990; Sheets and Mitchell 2001; Roopnarine 2003; Hunt 2012; Voje 2016). A fundamental challenge with most rate metrics is their dependence on timescales: lower rates of evolution are typically estimated for data spanning longer time intervals compared to data spanning shorter time intervals (for a discussion, see Hunt 2012). This makes these rates difficult to compare meaningfully across different time intervals and thus among time series of varying durations. In phylogenetic comparative methods, rates of evolution are commonly estimated as model parameters (e.g., Hansen 1997; Butler and King 2004; Hansen et al. 2008; Harmon et al. 2010; Adams 2013; Slater 2013, 2015). Hunt (2012) argued how evolutionary rates could be estimated as model parameters in his models of three canonical modes of evolution in the fossil record, that is, stasis, random walk, and directional change. In the same article, using simulations, he showed that each of these three models could accurately estimate evolutionary rates at different temporal resolutions via maximum likelihood as long as the underlying model of evolution is true. Thus, evolutionary rates in the fossil record can be estimated as model parameters if the model being used is a good descriptor of the data in absolute terms. However, we currently lack tools for evaluating whether a particular model is a good descriptor of fossil time series data. Analyses of modes of evolution in the fossil record have so far been conducted through model selection, which means that researchers typically use an information criterion (e.g., Akaike information criterion; Burnham and Anderson 2004) to select the best model out of a set of candidate models (e.g., Hunt 2007; Monnet et al. 2011; Hopkins and Lidgard 2012; Pearson and Ezard 2014; Hunt et al. 2015; Voje 2016). In this approach, the preferred model will be the one with best relative fit among the candidate models, but that does not guarantee that the preferred model is an adequate description of the data. While the same challenge applies to models within phylogenetic comparative analyses, several statistical procedures have been suggested to test the absolute fit of various models of evolution along a phylogeny (e.g., Garland et al. 1992; Boettiger et al. 2012; Beaulieu et al. 2013; Slater and Pennell 2013; Pennell et al. 2015). To date, there are no tests of adequacy of fitted models to fossil phenotypic time series. To alleviate this deficiency, we construct tests of adequacy to evaluate the absolute fit of Hunt's (2006) stasis model to fossil data. Given a good absolute fit of data to Hunt's model, trait deviations from a fixed optimum among different data sets can then be reliably compared using the same model if these data are analyzed on a comparable scale. A test for a negative correlation between effective population size and deviations from fixed optima during periods of stasis, as predicted by the fixed-optimum hypothesis, can then be applied if measures of effective population size can be obtained.

Testing the Fixed-Optimum Hypothesis: Proxies for Effective Population Size

An empirical test of the fixed-optimum hypothesis requires data on effective population sizes, a parameter that is challenging to estimate even in extant populations (Wang 2005). In the fossil record, a range of factors such as biased preservation, varying sampling probabilities, time averaging, range shifts of populations, and so on (Patzkowsky and Holland

2012) make estimates of population size measures unreliable, if at all attainable. An alternative approach is to focus on relative differences in this parameter among groups of organisms through proxies, and we discuss population size, body size, and habitat in turn. Body size is a fundamental property of any given organism and frequently shows allometric scaling with various ecological and physiological variables (e.g., Damuth 1981, 1987, 1993; Peters 1983; Schmidt-Nielsen 1984; Charnov 1993; West et al. 1997; Brown and West 2000). How effective population size (N_e) and body size correlates across several orders of magnitude has rarely (if ever) been quantified, but large animals are generally less abundant compared to smaller organisms (Damuth 1981, 1987, 1993), which suggests that larger-sized organisms have lower N_e, all else being equal. A positive (albeit variable) log-linear relationship ($r^2 = 0.43$) between N_e and N was indeed found by Palstra and Fraser (2012) in their analysis of empirical estimates of N_e and N among different species of fish, amphibians, and insects. Life-history parameters also affect effective population size (e.g., Vindenes et al. 2010; Serbezov et al. 2012; Waples et al. 2013). Longer life span and later age at maturity lead to lower N_e and are positively related to body mass across species (e.g., Charnov 1993; Healy et al. 2014), which also suggests that larger-bodied species have smaller effective population sizes. However, variation in other life-history traits (e.g., age structure, fecundity, mating system; Nunney 1991, 1993) may counter these general correlations and reduce the validity of body size as a proxy for effective population sizes, especially among species of comparable sizes. While acknowledging that variation in life-history strategies, phylogenetic history, and other factors may reduce the precision of body size as a proxy for effective population size, organisms that belong to vastly different size classes, such as diatoms and mammoths, are likely to have on average substantially different effective population sizes (for a few examples, see also table 1 in Charlesworth 2009).

Other proxies for effective population size are sizes of habitats. Hunt and Rabosky (2014) suggested comparing fluctuations during stasis in species inhabiting pelagic and benthic marine habitats as a potential test of the fixed-optimum hypothesis, with the underlying assumption being that benthic species on average have smaller effective population sizes compared to pelagic marine planktonic species. We are not aware of any direct estimates of effective population size of species inhabiting these environments to confirm this assumption. In addition, variation in life-history strategies among species may reduce the precision of habitat as proxies for expected differences in effective population size. However, pelagic (open ocean) species are purported to have "enormous population sizes and broad, even global, distributions" (Norris 2000; see also Angel 1993; Gray 1997), features that are implicitly less probable or widespread for species restricted to the benthic zone. Furthermore, holoplanktonic organisms often exhibit little genetic differentiation over large spatial scales compared to many benthic organisms (see Thornhill et al. 2008 and references therein), also suggesting on average larger population sizes in pelagic compared to benthic species. The fact that shallow-marine benthic species in general have smaller effective population sizes compared to pelagic (open ocean) species does not seem like an unrealistic claim, although we acknowledge that it remains to be verified by future research.

Testing Absolute and Relative Fit of Data to a Model of Stasis

One of the main goals of this study is to evaluate the fixed-optimum hypothesis of stasis on macroevolutionary timescales. To do this, we first examine the relative fit of 572 fossil time series to different models of evolutionary modes to detect traits in lineages that evolved in a stasis-like manner, that is, that show trait fluctuations around a constant optimum. We then develop and apply four test statistics to detect which of these time series fit the stasis model in absolute terms. Time series that pass all our adequacy tests are deemed suitable for reliably estimating average deviations from a fixed optimum (the omega in the stasis model). We then partition the time series into categories assumed to reflect substantial differences in effective population size (based on body size or habitat; see previous section) and test for a relationship between these categories and average deviations from fixed optima. The merits of different hypotheses explaining stasis in the fossil record are discussed in light of our results.

Material and Methods

Data

The majority of the fossil time series analyzed in this article overlap with the data analyzed in Voje (2016), which represents a subset of the fossil time series analyzed in Hunt (2007), Hopkins and Lidgard (2012), and Hunt et al. (2015). An additional 85 bryozoan fossil time series from the work of Cheetham et al. (2007) were kindly provided to us by Gene Hunt. In total, we analyzed 572 fossil time series. These data cover a broad range of taxa: mammals ($N = 84$), fish ($N = 9$), brachiopods ($N = 6$), ostracods ($N = 5$), bryozoans ($N = 92$), mollusks ($N = 78$), echinoids ($N = 7$), hemichordates ($N = 1$), trilobites ($N = 7$), conodonts ($N = 22$), foraminiferans ($N = 125$), coccolithophores ($N = 65$), radiolarians ($N = 46$), and diatoms ($N = 25$). A total of 364 size traits, 144 shape traits, and 64 meristic traits were analyzed (for more information on all time series analyzed, see table S1, available online). All time series are measures of a morphological trait (i.e., principal components and discriminant functions were not analyzed) that are either reported on a log scale or where a log transformation of the trait is a meaningful transformation. Log transformation makes changes in traits comparable since a proportional scale is independent of the original scale in which the traits were measured. To further enhance comparability, each sample mean in each time series was divided by the dimensionality of the measurement of the trait to make the trait both dimensionless and scale independent. Similarly, the variances of the trait means in each time series were divided by the square of the dimensionality of the trait.

Time series were divided into three pairs of categories, with the underlying assumption that taxa belonging to different categories within a pair on average have large differences in effective population size: (i) microfossils (ostracods and bryozoans were classified as microfossils) are assumed to have on average larger effective population sizes compared to macrofossils, (ii) mammals are assumed to have on average smaller effective population sizes compared to non-mammals (e.g., invertebrates and unicellular microfossils), and (iii) unicellular microfossils in a marine pelagic habitat are assumed to have on average larger effective population sizes compared to unicellular species living in a shallow-marine benthic habitat. The two first pairs of categories (micro vs. macro, mammals vs. non-mammals) represent different ways of separating the stasis time series data based on differences in overall size, while the last pair of categories (benthic vs. pelagic) is a subset of the total data.

Relative Fit of Data to Different Models of Evolution

Models describing directional change, random walk, and stasis were fit to each time series by maximum likelihood using the fit3models function (specifying joint parameterization) in the paleoTS package, version 0.5-1 (Hunt 2006, 2008), using R, version 3.1.3 (R Development Core Team 2013). The relative fit of each of the three models to each time series was assessed using the Akaike information criterion corrected for small samples (AICc). Time occurs in discrete intervals, and the expected difference between sample means is represented by a normal distribution, with mean (μ) and variance (σ^2) for all three models. In the directional-change model, the mean of the normal distribution is not zero ($\mu \neq 0$) and reflects the direction of evolution of the given trait over time, while σ^2 represents the fluctuations around the directional trend. In the random-walk model, the mean of the step distribution is zero, which means that the expected difference between the ancestor and descendant is normally distributed with a zero mean and a variance of $t\sigma^2$, where t is the number of generations separating the ancestor and descendent. The stasis model differs slightly from the other two models, as it describes a trait that fluctuates with a variance (ω) around an optimal/central phenotype (θ), essentially a white noise process with uncorrelated normally distributed trait values around the fixed mean through time. This makes the interpretation of the rate parameter in the stasis model slightly different from the models of directional change and the random-walk model: while the directional change and random-walk models are models of trait change, the stasis model is a model of trait values around a fixed optimum. Hence, rates of change between discrete time units (e.g., generations) are possible to estimate with the directional-change and random-walk models, while the variance (ω) parameter in the stasis model represents the variance of trait deviations from a fixed mean over time, which is not strictly a description of how the trait changes per discrete time unit. In the context of our study, estimates of ω from the stasis model should therefore not be interpreted as estimates of rates but instead as an estimate of the variance of deviations from a fixed optimum. For a detailed description of the three models, see Hunt (2006).

Relative and Absolute Fit of the Stasis Model (Model Adequacy)

Hunt (2012) showed that ω in the stasis model could be reliably estimated when the underlying evolutionary dynamics are similar to a white noise process with uncorrelated normally distributed trait values around an invariant optimum through time. We therefore need to make sure that the data sets that show a better fit to the stasis model relative to the alternative models (random walk and directional change) also show a good fit in absolute terms to a white noise process. Hence, it is necessary that the following four criteria are fulfilled: (1) the magnitude of deviations around the fixed optimum must not increase or decrease as a function of time, (2) net evolution must be small (i.e., the absolute difference between the first and the last sample mean in the time series must not significantly exceed the expected difference between sample means given the estimated ω), (3) the deviations from the estimated optimum must show randomness in the temporal order (i.e., many successive positive or negative deviations indicate nonrandomness), and (4) the temporal series must show low autocorrelation.

We develop four test statistics reflecting the criteria above to investigate whether the stasis model adequately describes a particular fossil time series. Criterion 1 is necessary since the estimated variance parameter in the stasis model is assumed to be constant over time. A correlation in the deviations from the optimum over time will therefore indicate a violation of this assumption and thus a bad model fit in absolute terms. The zero-slope test represents the slope of the least squares regression of the size of deviations (their absolute value) from the optimal phenotype as a function of time. A slope of zero is expected in a true stasis time series, as there should be no relationship between time and the magnitude of deviations from the optimum. A positive or negative slope indicates a tendency for the trait to show larger or smaller deviations from the optimum as a function of time (heteroscedasticity), which is a violation of our first criterion. The second criterion reflects an essential part of the general (verbal) definition of stasis, that a trait shows little net change over time. The net-evolution test represents the absolute difference in trait value between the first and the last sample mean in the time series. Criteria 3 and 4 reflect the nature of a white noise process. There should be no tendency for a trait to successively deviate from the optimum in the same direction, and a runs test is applied to the sign of the residuals (i.e., $\theta -$ trait value) to identify series that have nonrandom patterns in the sign of deviations. For a time series of length n, the number of runs (one run is a sequence of consecutive numbers with the same sign) is approximately normal with mean $\mu = 2(n_+n_-)/n+1$ and variance $(\mu-1)(\mu-2)/(n-1)$, where n_+ and n_- are the number of residuals above and below the optimum, respectively. The mean and variance are used to calculate the standard/Z score implemented as the test statistic. Last, since trait values are seen as random draws from a normal distribution, they should exhibit low levels of autocorrelation. We therefore apply an autocorrelation test, which is the correlation of the first $n-1$ observations with the last $n-1$.

Criteria 1–4 do not need to be fulfilled for a fossil time series to show a relative better fit to Hunt's (2006) stasis model compared to alternative models, which is why a better fit to the stasis model in relative terms is no guarantee that a fossil time series will fulfill these properties. Hence, if all four criteria are met for a particular fossil series, we consider the stasis model a sufficiently good descriptor of the data in both relative and absolute terms. Furthermore, for fossil time series that meet all four criteria, the estimated variance parameter ω could be considered an adequate descriptor of the magnitude of deviations from the fixed optimum.

The procedure for our model-adequacy tests on time series that is best explained by the stasis model in relative terms is a parametric bootstrapping approach that follows these steps (fig. 1): (1) we fit the stasis model (Hunt 2006) to an observed fossil time series under consideration and estimate the ω and θ using maximum likelihood. (2) We calculate each of the four test statistics on the observed time series. (3) Using the ω and θ parameters estimated from the observed time series (step 1), we simulate 1,000 new stasis time series (using the sim.Stasis function in the R pack-

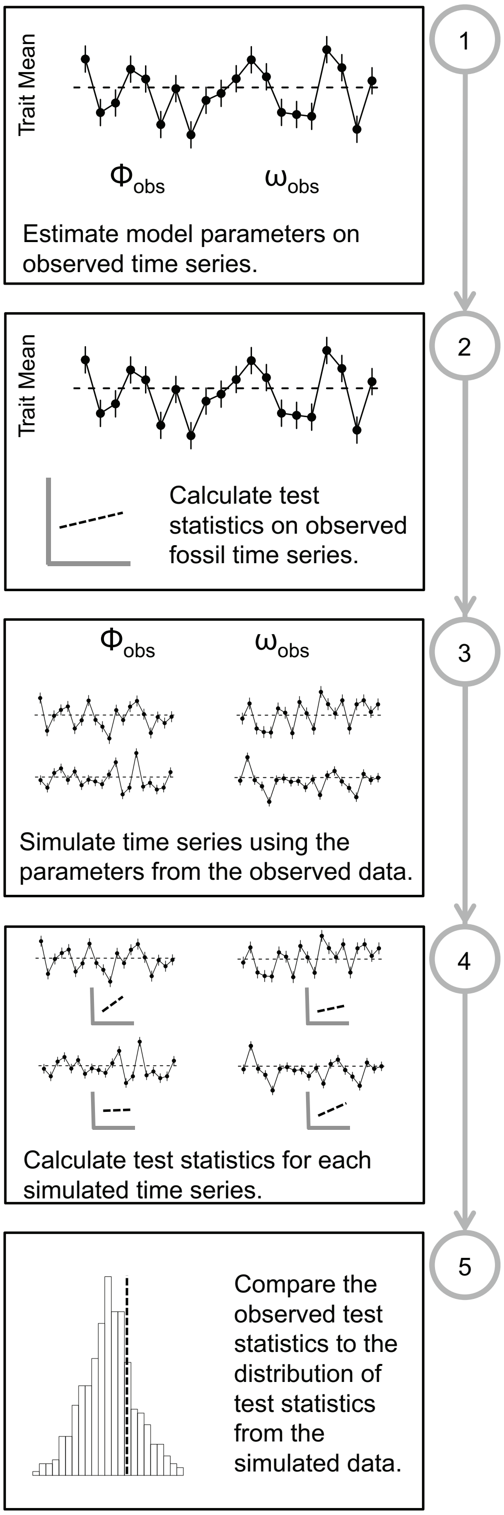

1
Trait Mean
Φ_{obs}
ω_{obs}
Estimate model parameters on observed time series.
2
Trait Mean
Calculate test statistics on observed fossil time series.
3
Φ_{obs}
ω_{obs}
Simulate time series using the parameters from the observed data.
4
Calculate test statistics for each simulated time series.
5
Compare the observed test statistics to the distribution of test statistics from the simulated data.

Figure 1: Stepwise representation of the approach for assessing model adequacy for data sets showing a better relative fit to the stasis model than the models portraying random walk and directional change according to their AICc score. Step 1: fit the stasis model to the fossil time series and estimate the parameters θ and ω. Step 2: calculate the four test statistics on the observed data. Step 3: use the estimated θ and ω parameters to simulate 1,000 stasis data sets of the same length as the observed data. Step 4: calculate the four test statistics on each of the 1,000 simulated data sets. Step 5: compare the test statistic estimated from the observed data to the test statistics estimated from the simulated data sets. If the observed test statistic lies outside the 2.5% most extreme test statistics from the simulated data, the data are considered inadequate for our purpose.

age PaleoTS, version 0.5-1; Hunt 2006) with the same number of trait means as the focal time series. (4) We calculate the four test statistics on each of the simulated time series. (5) Last, we compare each of the test statistics from the observed data to the distribution of test statistics calculated on the 1,000 simulated time series. The stasis model is deemed unsuitable as a descriptor of a particular observed time series if one of the four observed test statistics is in the lower or upper 2.5 percentiles in the distribution of test statistics from the simulated time series. Inflated type I error rates can easily be introduced when applying several tests to the same data. This is problematic if the goal of applying these test statistics is to evaluate whether a specific model should be accepted or rejected. However, our goal is not acceptance versus rejection but to evaluate the suitability of individual data sets in order to reliably estimate model parameters. We are accordingly not correcting for multiple testing. A similar approach was also used by a recent study on model adequacy in phylogenetic comparative methods (Pennell et al. 2015).

Simulations of Model-Adequacy Tests

We assessed the type I error rates of our model-adequacy test statistics by performing simulations. Many of our test statistics have well-known statistical properties, but we perform the simulations to assess the effect of varying lengths of fossil sequences and variation in other underlying properties of the data. The simulations follow the procedure described above and shown in figure 1, except that step 1 is a simulated time series with known parameter values. All simulations of stasis time series were done using the sim.Stasis function in the PaleoTS package, version 0.5-1 (Hunt 2006). For all simulated time series, θ was set to 1 and the within-species variance was set to 0.05. We varied both sequence length (10, 20, 40, 80) and ω (0.1, 0.2, 0.4, 0.6), which covers most of the observed variation in these parameters in the empirical fossil time series we analyze (see table S1). For each combination of ω and sequence length, one stasis time series was simulated and the four test statistics were calculated (step 2). We then used the estimated parameters from this observed stasis time series to simulate 1,000 new stasis sequences (step 3) to obtain distributions for each test statistic (step 4), which were then used to investigate the frequency of type I error for these simulated data (step 5). This procedure was repeated 500 times for each combination of ω and sequence length.

Testing for a Relationship between Trait Variance and Effective Population Size

The estimated variance parameters (ω) for data that passed all of the four adequacy tests described above were used as data to investigate whether groups of species hypothesized to have vastly different effective population size differ in their trait variance during periods of stasis. Before doing this, however, we tested for a relationship between the variance parameter (ω) and the interval length of the fossil time series, as our tests depend on these two variables being independent: it is difficult to rely on estimates of ω if they correlate with interval length since estimates of ω should be independent of the duration of phenotypic stasis in a given lineage. Next, under the assumption that taxa that on average differ in body mass by several orders of magnitude also have very different effective population sizes, we investigated whether microfossils have a smaller trait variance during stasis compared to macrofossils and whether mammals have a larger trait variance than non-mammals (see table 1). Similarly, assuming that unicellular microfossils that inhabit the (shallow-marine) benthic zone on average have substantially smaller effective population sizes compared to unicellular microfossils that inhabit the (open ocean) pelagic zone, we investigated whether such planktonic microfossils have smaller trait variances compared to benthic microfossils. Tests of differences in trait variance during stasis between these pairs of categories were done using mixed-effects models, as implemented in the R package lme4, version 1.1.13 (Bates et al. 2015). Study and species were included as random factors to control for nonindependence in the data. R code to run all analyses and all the analyzed time series are deposited in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.r5d10 (Voje et al. 2017).[1]

Table 1: Estimated average trait variance around the fixed optimum for different categories of fossil lineages

Model and fixed effect (*N*)	Trait variance ω (SE)
1:	
Microfossils (122)	.0108 (.0045)
Macrofossils (50)	.0047 (.0029)
2:	
Non-mammals (154)	.0093 (.0026)
Mammals (18)	.0060 (.0077)
3:	
Benthic (17)	.0015 (.0153)
Planktonic (41)	.0141 (.0093)

Model 1: microfossils are represented by foraminiferans, coccolithophores, radiolarians, diatoms, ostracods, and bryozoans, while macrofossils are represented by mammals, mollusks, trilobites, and conodonts. Model 2: the non-mammals category consists of all taxa represented in our data with the exception of mammals. Model 3: a comparison of benthic and planktonic unicellular microfossils. $N =$ number of analyzed traits.

Results

Simulations to Evaluate Adequacy Tests

Error rates for the four test statistics in the simulation procedure show that type I errors are centered around 0.05, in-

[1] Code that appears in the *American Naturalist* is provided as a convenience to the readers. It has not necessarily been tested as part of the peer review.

dependent of the length of the time series or the size of the trait fluctuations (ω) during stasis (fig. 2). This indicates that our test statistics for evaluating model adequacy work as intended. The proportion of simulated data sets that violates at least one of the test statistics is between 0.15 and 0.20, which indicates that most of the simulated time series that are deemed inadequate violate only one of the four test statistics.

Model Fit and Model Adequacy

Of the 572 analyzed fossil time series, 263 fitted the stasis model best, 253 fitted the random-walk model best, and 56 fitted the directional-change model best, based on AICc. Of the 263 time series that best fitted the stasis model, 91 (34.6%) failed at least one of the four test statistics, 28 showed a relationship between the magnitude of deviations from the

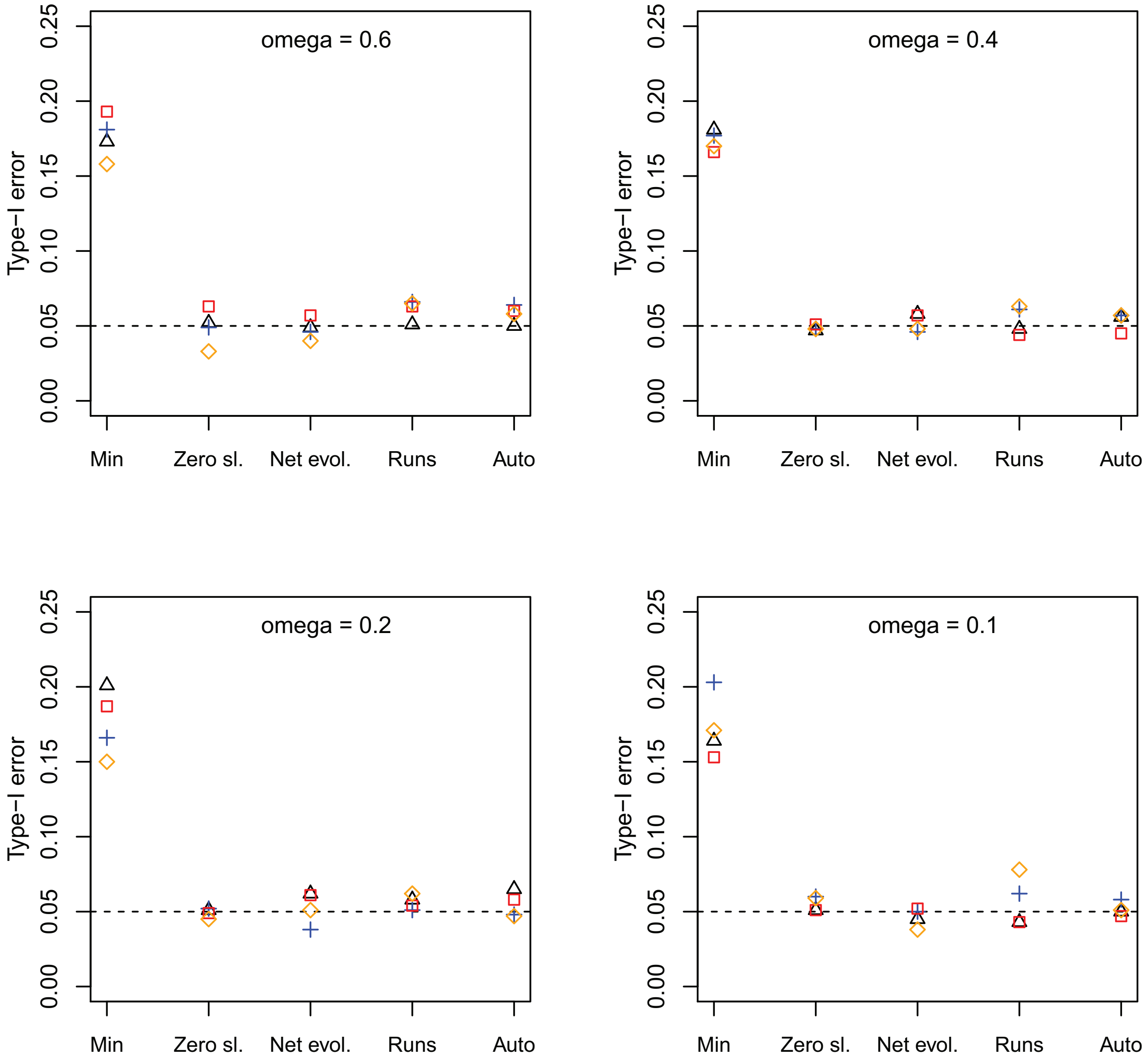


Figure 2: Type I error rates of adequacy tests for simulated stasis time series. Five hundred "true" stasis time series were simulated for a given sequence length (orange diamond = 10, red square = 20, blue cross = 40, black triangle = 80) and a given size of the omega parameter. For each of these "true" stasis time series, 1,000 time series were simulated to check whether the test statistics estimated on the "true" data fall within the 95% of the observed test statistics conducted on the simulated data. Type I error rates for the four test statistics are around the expected 0.05 threshold. Min, which lies between 0.15 and 0.20, represents the number of simulated time series that deviated from the 95% distribution in at least one of the four test statistics. Auto = test for autocorrelation in the data; Net evol. = test for larger amounts of net evolution than expected; Runs = test for nonrandom patterns in the sign of deviations from the optimum; Zero sl. = test for a relationship between time and the magnitude of deviations from the optimum.

optimum with time (zero slope), 25 showed a larger net evolution than expected, 36 had a larger or smaller number of runs than expected, and 33 showed a level of autocorrelation exceeding the expectation. Figure 3 shows examples of data sets violating (and not violating) the four test statistics. Detailed results for all adequacy tests of the 263 time series fitting the stasis model are reported in table S1. Importantly, the estimated trait variance (ω) from the 172 time series that did not fail any of our tests showed no relationship with interval length (fig. 4): the ordinary least squares slope of ω as a function of time in millions of years was -0.0000128 ($\pm$ 0.0008329), $R^2 < 0.00001\%$, $P = .988$.

Relationship between Trait Variance and Effective Population Size

We did not find evidence for differences in the magnitude of trait variance around a fixed optimum for groups of taxa assumed to have large differences in relative effective population sizes based on rough proxies (fig. 5; table 1). The average deviation from the optimum is larger in microfossils compared to macrofossils and larger in non-mammals compared to mammals. These differences are in the opposite direction predicted from the fixed-optimum hypothesis. The difference in average trait variance in unicellular benthic lineages compared to unicellular planktonic lineages is also in the opposite direction from the prediction following from the fixed-optimum hypothesis. Confidence intervals show extensive overlap in all three comparisons.

Discussion

A well-known challenge in comparative studies of the fossil record is the difficulty of obtaining comparable estimates of trait change across data that cover different time intervals (e.g., Gingerich 1993, 2001, 2009; Hunt 2012; Voje 2016). Hunt (2012) showed that reliable estimates of trait change can be obtained as model parameters when trait dynamics in the fossil record are closely matched by a particular model of evolution. Here, we developed tests to ensure that a model for stasis in the fossil record adequately describes time series data. This allowed estimates of the variance in 172 traits during periods of long-term stasis and enabled investigations of predictions from competing hypotheses explaining stasis on macroevolutionary timescales.

Explanations for Stasis in the Fossil Record

We do not find support for the hypothesis that stasis in the fossil record generally is the result of stabilizing selection around a constant phenotypic optimum. The fixed-optimum hypothesis predicts that the size of deviations from the optimum is negatively correlated with effective population size, since genetic drift is what causes phenotypes to diverge from the optimum. A species with a large effective population size is accordingly predicted to show smaller trait variance during periods of stasis relative to a species with smaller effective population size, as long as the curvatures of the adaptive peaks are similar. However, when comparing groups of species that are hypothesized to have large differences in effective population size based on their body size or habitat, we did not find any indication that effective population size predicted trait variances during stasis. In fact, the differences in trait variance were in the opposite direction from that predicted by the fixed-optimum hypothesis. We acknowledge, however, that the categories we used when sorting our data may not represent strong predictors of an association between effective population size and drift, as these categories consist of organisms with very different biology and life histories. Our investigation should therefore be interpreted as a first step in investigating the fixed-optimum hypothesis that we hope will inspire further work in understanding long-term stasis in the fossil record. More detailed knowledge on relationships between effective population size and variables such as habitat, population density, body size, and so on may allow for more robust evaluations of the hypothesis in the future. A stasis model containing a parameter interpretable as an evolutionary rate on a generational scale may also allow estimates of quantitative genetic parameters (e.g., Estes and Arnold 2007; Hansen 2012), including estimates of effective population size, which can then be evaluated as biologically plausible or not in relation to the fixed-optimum hypothesis.

Both neontologists (e.g., Hansen and Houle 2004) and paleontologists (e.g., Lieberman and Dudgeon 1996; Gould 2002) have been skeptical toward the fixed-optimum hypothesis. It has been argued that it is difficult to reconcile long-term stability of fitness optima with the observation of continuous change in species' environments (e.g., Hansen 2012), although niche tracking can buffer against environmental changes and keep fitness optima stable (e.g., Gould 2002; Eldredge 2003; Eldredge et al. 2005; Brett et al. 2007). The fixed-optimum hypothesis also seems at odds with the infrequent detection of stabilizing selection relative to directional selection in natural populations (Kingsolver et al. 2012; Morrissey and Hadfield 2012; Morrissey 2016). However, a recent article by Haller and Hendry (2014) argues how empirical selection estimates showing frequent directional selection and the idea that most traits are under stabilizing selection around a constant optimum can be reconciled: if the adaptive landscape is not sharply peaked, a population can be expected to fluctuate stochastically around the peak, which would lead to frequent directional selection being detected on short timescales. Some of the estimated variance parameters among the 172 investigated traits are very small or even zero, making it difficult to exclude the possibility that stasis in some lineages may be explained by the fixed-optimum hypothesis. For example,

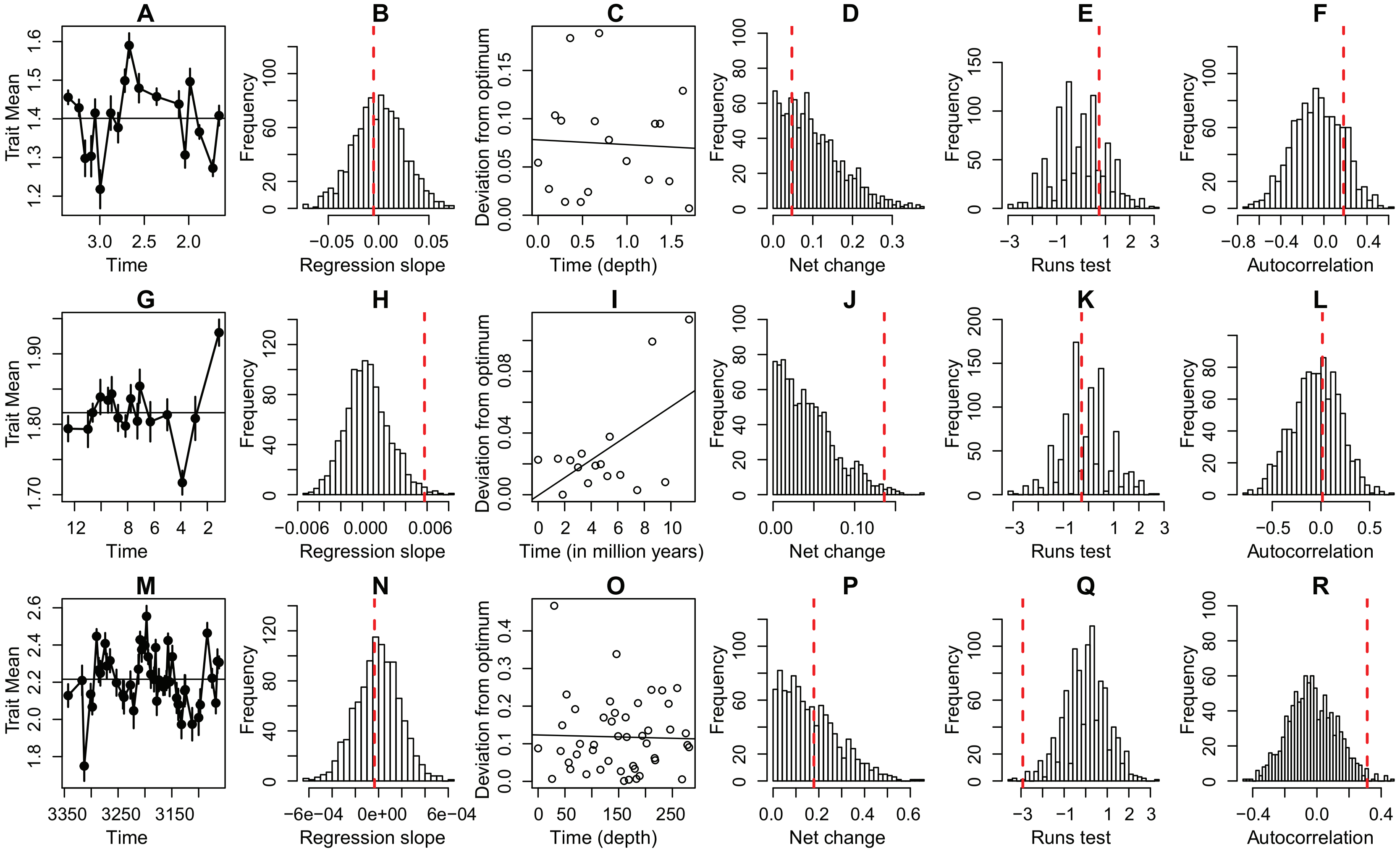


Figure 3: Examples of model-adequacy tests. The first column shows the time series, and the other columns show test statistics. The histograms show the distribution of test statistics calculated on the simulated stasis time series, and the red dashed vertical lines are the values of the test statistics from the observed fossil time series, while the third column shows the absolute deviation from the optimum as a function of time. *A* (first row) shows how the height of the hyaline area in *Rhizosolenia bergonii* (a diatom) changes over time. In this case, all test statistics of the observed stasis time series (red dashed vertical lines) are in the middle of the distributions of the four simulated test statistics, which indicate that this time series fulfills our criteria for being an adequate stasis time series. *G* (second row) shows how the length of the coccolithophore *Watznaueria* aff. *communis* changes over time. The test statistics show that the deviations from the optimum increase with time (*H*, *I*) and that the observed net evolution falls outside the expected range (i.e., falls outside the 95% density of the 1,000 simulated time series; *J*). *M* (lower row) shows a measure of the test length of the foraminiferan *Afrobolivina afra*, which has a number of runs (*Q*) and level of autocorrelation (*R*) that fall outside the 95% density of the expected distribution.

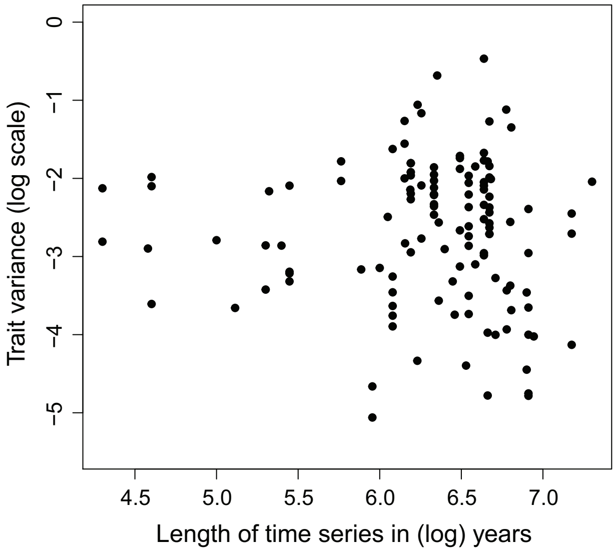


Figure 4: No relationship between the estimated trait variance (ω) and the length of time series in millions of years. The ordinary least squares regression has a slope of -0.0000128 ($\pm$ 0.0008329), and the relationship has an R^2 value of <0.00001% ($P = .9878$). Values are log transformed to increase clarity for the plot, which means that the 32 traits showing zero trait variance have been omitted from the plot (but not from the regression analysis).

certain traits may serve functions that remain fairly constant over millions of years. In a study of phenotypic evolution in canids, Slater (2015) found that the trait dynamics of the relative size of the grinding surface of lower teeth showed a faster rate of adaptation toward more stable fitness peaks compared to the estimated trait dynamics of body mass. Slater (2015) interpreted this difference in dynamics between the two traits based on their difference in functional role: the properties and availability of different types of food (e.g., vertebrate flesh and insect chitin) have likely remained relatively constant over millions of years and have been only marginally influenced by biotic and abiotic changes in the environment. The grinding surface of teeth used to eat food within these diet categories has consequently not needed to change much, a situation that is less likely to hold true for body mass, as this trait influences multiple aspects of the general ecology of species (e.g., Schmidt-Nielsen 1984). A proper test of whether traits with very specific functional roles show different trait dynamics compared with other types of traits deserves attention in future studies. We also note that in our analyses, traits with small variance parameters are almost always associated with lineages where the sample variance is large compared to the differences among sample means, which indicates that highly variable trait observations during the time period used to estimate the sample mean can in fact lead to low estimates of trait variance over time. Stabilizing selection on a trait that is close to an invariant fitness optimum is therefore not supported by our results as a general explanation for stasis in the fossil record.

The general lack of support for the fixed-optimum explanation of stasis indicates that other explanations for stasis in the fossil record need to be evaluated. The fluctuating-optimum hypothesis claims that stasis is the result of populations tracking the bounded movements of an optimum over macroevolutionary timescales (Hunt 2007; Uyeda et al. 2011; Arnold 2014; Voje 2016). This hypothesis does not predict any relationship between effective population size and deviations from the optimum and is therefore not in conflict with the results presented in this study. We stress, however, that further tests of this hypothesis are needed. For example, the fluctuating-optimum hypothesis predicts that morphological stasis should overlap in time with bounded fluctuations

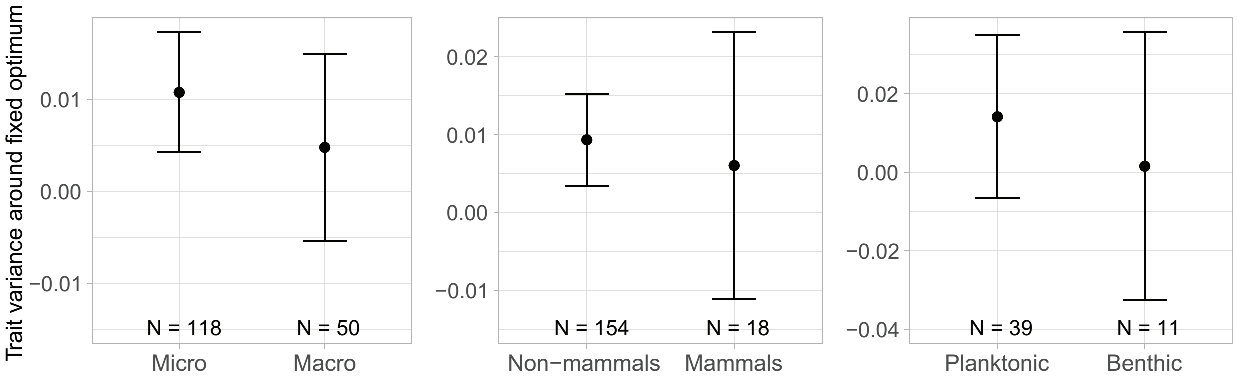


Figure 5: Comparisons of predicted trait variance around a fixed optimum for different size and habitat categories of fossil lineages. Taxa defined as microfossils in the left panel are foraminiferans, coccolithophores, radiolarians, diatoms, bryozoans, and brachiopods, while mammals, mollusks, echinoids, and conodonts represent macrofossils. Non-mammals are represented by all other taxa in the data, except mammals. The right panel shows unicellular microfossils (foraminiferans, coccolithophores, radiolarians, and diatoms) divided into the habitats benthic (shallow marine) and planktonic (open ocean). $N =$ number of traits analyzed. Error bars represent 95% confidence intervals. Values are given in table 1.

of the environmental variables that act as agents of selection on the trait (Hunt and Rabosky 2014). This prediction has been explored by Hunt et al. (2015); assuming a general reaction norm for how changes in temperature lead to changes in body size, they tested whether a proxy for temperature change on macroevolutionary timescales produced a pattern of morphological change similar to the blunderbuss pattern shown by Uyeda et al. (2011). The blunderbuss pattern of evolution represents bounded morphological change on time spans less than about 1 myr, while larger changes in morphology happen on timescales above the threshold of about 1 myr. Hunt et al.'s (2015) model was able to recreate the blunderbuss pattern to some extent, which suggests that bounded variation in environmental components may be an important explanation for long-term stationarity in morphology (i.e., stasis). The work by Hunt et al. (2015) shows how the fluctuating-optimum hypothesis can be tested. We suggest that future investigations of this hypothesis could explore clade-specific models for how an environmental variable (e.g., temperature) affects the evolution of a given trait.

The results of our study are not relevant for evaluating constraints, defined broadly as any mechanism that biases, limits, or prevents an evolutionary response to selection (Arnold 1992; Houle 2001), as a potential explanation for stasis. For instance, pleiotropy is widespread (Walsh and Blows 2009), and a specific trait can have reduced ability to evolve when its pleiotropically linked traits are under stabilizing selection (Hansen 2003; Hansen et al. 2003). The timescales at which such multivariate genetic correlation can constrain traits are unclear, however, as it is possible to envision scenarios where pleiotropic constraints have long-lasting effects but yet decay rather fast (e.g., Futuyma 2010; Hansen 2012). Genetic constraints can be neither accepted nor rejected as a general explanation of stasis in the fossil record given the current state of the field (Hansen 2012).

Envisioning long-term stasis as the result of homogenizing gene flow between locally adapted subpopulations (e.g., Lieberman et al. 1995; Futuyma 2010) is an explanation of a slightly different nature than the genetic constraint and the two stabilizing-selection hypotheses. While fixed or variable optima and genetic constraints can be linked to particular traits, their adaptive landscapes, and trait covariance structure, homogenizing gene flow is not an explanation at the level of traits but at the level of a species and its ecology. The fact that a species is spatially distributed and selection varies over space does not seem commensurate with the observation that several time series from the same sequences of population can show different modes of evolution (Hopkins and Lidgard 2012; this study; table S1). If it is only the structure of a metapopulation that gives rise to observed stasis in the fossil record, this explanation would suggest that all traits of a particular lineage should show stasis, which is not the case. Additionally, for a trait showing stasis, the homogenizing gene flow explanation would also need to invoke some degree of stationarity of the optima in each subpopulation, essentially including aspects of the fluctuating-optimum hypothesis with a spatial component.

The Importance of Model Adequacy in the Study of the Fossil Record

All models are simplifications of the phenomena we want to investigate. Models are hence useful only if they capture key properties of the phenomenon we are interested in. Comparing how alternative models fit a particular data set allows for a relative evaluation of how each model explains the data, but this approach does not guarantee that the best model out of the alternatives describes the data particularly well (e.g., Pennell et al. 2015). While tests of absolute fit of models within phylogenetic comparative methods continue to be developed and investigated (e.g., Slater and Pennell 2013; Pennell et al. 2015; Chira and Thomas 2016), this has not been a focus for studies fitting models to fossil sequence data. We have argued that development of test statistics to evaluate model adequacy for studies of the fossil record is important if the goal is to reliably estimate model parameters of interest. Given our goal of testing the fixed-optimum hypothesis, we developed adequacy tests only for the stasis model, but adequacy tests can also be developed for alternative models of evolution, such as the models of random walk and directional change used as alternative models in our study. For example, the random-walk model, like the stasis model, assumes no autocorrelation in consecutive sample means, no change in step size as a function of time, and no long runs of consecutive steps from one side of the normal distribution. Similar statistics can also be applied to the directional-trend model. Adequacy tests for different models of trait dynamics could pave the way to identify data sets that can be used for reliable and model-based evolutionary rate comparisons.

One-third of the data sets showing a better relative fit to the stasis model than to the random-walk and directional-trend models failed at least one of our four tests of model adequacy. This illustrates the point that a relative fit is no guarantee that the model is a good description of a particular data set. A series of seminal papers in the 1980s and 1990s by Cheetham and colleagues (Cheetham 1986; Cheetham et al. 1993, 1994; Jackson and Cheetham 1999) consolidated the bryozoan genus *Metrarabdotos* as a textbook example of a punctuated equilibrium-like mode of evolution, where evolutionary changes happen predominantly during rapid speciation events while lineages experience stasis after their first appearance in the fossil record. We reanalyzed 85 time series of trait change in eight different *Metrarabdotos* lineages in our study. Of these, 71 showed a relative better fit to the stasis model, and 21 of these time series failed one or more of the adequacy tests. In other words, 41% of the *Metrarabdotos* data we analyzed was poorly described by the stasis model. Model adequacy is likely an important tool for developing a more nuanced view of trait dynamics in the fossil record.

Test failures might have both biological and nonbiological causes (Pennell et al. 2015). For example, many sample means in fossil time series have been calculated based on rather modest sample sizes. The resolution of a fossil sequence may also vary substantially through time, and strong time-averaging effects for estimated sample means may also be present in many data sets. Such factors can contribute to poor fit of data to simple process models. On the other hand, biological explanations may potentially underlie failures to pass our adequacy tests (Pennell et al. 2015). For example, a particular trait may evolve according to one model of evolution for a period of time and then evolve according to a different model, and such heterogeneous evolutionary trait dynamics can be captured by fitting different models to different parts of a time series (Hunt 2008; Hunt et al. 2015). Furthermore, assuming populations have the ability to track optima on the adaptive landscape with an insignificant time lag on million-year timescales, patterns of trait change in the fossil record can be interpreted as descriptions of how the adaptive landscape itself changes (Hunt 2007; Uyeda et al. 2011; Hansen 2012; Hunt and Rabosky 2014; Voje 2016). Hence, failing a particular adequacy test can suggest how the adaptive landscape changed on macroevolutionary timescales. Twenty-eight traits showed an increased or decreased trait variance with time (the zero-slope test). Several scenarios involving changes in the adaptive landscape can potentially explain such patterns. For example, a change in the curvature of the adaptive landscape over time might alter the range of morphologies with (more or less) equal fitness: a flattening of the adaptive landscape over time may increase the permissible disparity during the same time period. An alternative explanation is that the environmental variable influencing the trait shows an increase or decrease in its fluctuations over time, causing a similar increase or decrease in the magnitude of fluctuations of the optimum on the adaptive landscape (Hunt et al. 2015). Twenty-five of the 263 fossil sequences that showed a relative better fit to stasis also showed a larger net evolution than expected (table S1). This means that the first or the last (or both) sample mean in the time series deviates substantially from the optimum in the stasis model, which might indicate a shift in the optimum in the adaptive landscape either in the beginning or at the very end of the fossil sequence. Again, given a hypothesized driver of the phenotypic evolution in such a time series, one could predict that this driver should also show an abrupt change in the same time period. Failures of the runs test or the autocorrelation test may also suggest that the optimum is not constant during the stasis period. About 25% of the times series that failed the runs test had a larger number of consecutive observations both above or below the optimum than expected from a white noise process, which might indicate that a model where the optimum switches between two selective regimes could fit the data well.

Conclusion

Our preliminary investigation of the fixed-optimum hypothesis fails to find empirical support for the claim that genetic drift around a constant adaptive peak on the adaptive landscape is a general explanation for stasis in the fossil record. The common observation that traits from a single lineage often follow different models of evolution also questions homogenizing gene flow between locally adapted subpopulations as a common explanation for long-term morphological stasis. Stasis may instead reflect stationary optima on the adaptive landscape, but models of how the adaptive landscape change on macroevolutionary timescales are needed to investigate this hypothesis. The recent explosion of evolutionary models in comparative phylogenetic approaches has increased the toolbox for testing and investigating alternative dynamics of the adaptive landscape on million-year timescales (e.g., Butler and King 2004; Hansen et al. 2008; Harmon et al. 2010; Bartoszek et al. 2012; Beaulieu et al. 2012; Thomas and Freckleton 2012; Ingram and Mahler 2013; Slater and Pennell 2013; Pennell et al. 2014; Uyeda and Harmon 2014; Khabbazian et al. 2016; Caetano and Harmon 2017), and the usefulness of model-adequacy tests within comparative methods has recently been stressed (Pennell et al. 2015). Likewise, a more detailed investigation of the absolute fit of fossil time series to specific evolutionary models can result in additional knowledge of the evolutionary dynamics of the trait of interest and may inspire new models of trait dynamics on macroevolutionary timescales that have not yet been considered.

Acknowledgments

We thank Gene Hunt for providing the raw data of bryozoan fossil time series from the work of Alan H. Cheetham and colleagues. Suggestions and input from associate editor Scott J. Steppan and reviewers Matthew W. Pennell and Graham J. Slater greatly improved the manuscript. K.L.V. was supported by grant 249961 from the Norwegian Research Council, and J.S. was supported by grant 235073 from the Norwegian Research Council to L.H.L.

Literature Cited

Adams, D. C. 2013. Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. Systematic Biology 62:181–192.

Angel, M. V. 1993. Biodiversity of the pelagic ocean. Conservation Biology 7:760–772.

Arnold, S. J. 1992. Constraints on phenotypic evolution. American Naturalist 140(suppl.):S85–S107.

———. 2014. Phenotypic evolution: the ongoing synthesis. American Naturalist 183:729–746.

Bartoszek, K., J. Pienaar, P. Mostad, S. Andersson, and T. F. Hansen. 2012. A phylogenetic comparative method for studying multivariate adaptation. Journal of Theoretical Biology 314:204–215.

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67: 1–48.

Beaulieu, J. M., D. C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Systematic Biology 62:725–737.

Blows, M. W. 2007. A tale of two matrices: multivariate approaches in evolutionary biology. Journal of Evolutionary Biology 20:1–8.

Blows, M. W., and M. Higgie. 2003. Genetic constraints on the evolution of mate recognition under natural selection. American Naturalist 161:240–253.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? measuring the power of comparative methods. Evolution 66: 2240–2251.

Bookstein, F. L. 1987. Random walk and the existence of evolutionary rates. Paleobiology 13:446–464.

Bradshaw, A. D. 1991. The Croonian lecture, 1991: genostasis and the limits to evolution. Philosophical Transactions of the Royal Society B 333:289–305.

Brett, C. E., A. J. W. Hendy, A. J. Bartholomew, J. R. Bonelli, and P. I. McLaughlin. 2007. Response of shallow marine biotas to sea-level fluctuations: a review of faunal replacement and the process of habitat tracking. Palaios 22:228–244.

Brown, J. H., and G. B. West. 2000. Scaling in biology. Oxford University Press, Oxford.

Burnham, K., and D. Anderson. 2004. Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. American Naturalist 164:683–695.

Caetano, D. S., and L. J. Harmon. 2017. ratematrix: an R package for studying evolutionary integration among several traits on phylogenetic trees. Methods in Ecology and Evolution 8:1920–1927.

Charlesworth, B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10:195–205.

Charlesworth, B., R. Lande, and M. Slatkin. 1982. A neo-Darwinian commentary on macroevolution. Evolution 36:474–498.

Charnov, E. L. 1993. Life history invariants. Oxford University Press, Oxford.

Cheetham, A. H. 1986. Tempo of evolution in a Neogene bryozoan: rates of morphologic change within and across species boundaries. Paleobiology 12:190–202.

Cheetham, A. H., J. B. C. Jackson, and L.-A. C. Hayek. 1993. Quantitative genetics of bryozoan phenotypic evolution. I. Rate tests for random change versus selection in differentiation of living species. Evolution 47:1526–1538.

———. 1994. Quantitative genetics of bryozoan phenotypic evolution. II. Analysis of selection and random change in fossil species using reconstructed genetic parameters. Evolution 48:360–375.

Cheetham, A. H., J. Sanner, and J. B. C. Jackson. 2007. *Metrarabdotos* and related genera (Bryozoa: Cheilostomata) in the late Paleogene and Neogene of tropical America. Journal of Paleontology 81:1–91.

Chira, A. M., and G. H. Thomas. 2016. The impact of rate heterogeneity on inference of phylogenetic models of trait evolution. Journal of Evolutionary Biology 29:2502–2518.

Damuth, J. 1981. Population density and body size in mammals. Nature 290:699–700.

———. 1987. Interspecific allometry of population density in mammals and other animals: the independence of body mass and population energy-use. Biological Journal of the Linnean Society 31: 193–246.

———. 1993. Cope's rule, the island rule and the scaling of mammalian population density. Nature 365:748–750.

Eldredge, N. 2003. The sloshing bucket: how the physical realm controls evolution. Pages 3–30 *in* J. Crutchfield and P. Schuster, eds. Evolutionary dynamics: exploring the interplay of selection, accident, neutrality, and function. Oxford University Press, New York.

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Pages 82–115 *in* T. J. M. Schopf and J. M. Thomas, eds. Models in paleobiology. Freeman Cooper, San Francisco.

Eldredge, N., J. N. Thompson, P. M. Brakefield, S. Gavrilets, D. Jablonski, J. B. C. Jackson, R. E. Lenski, B. S. Lieberman, M. A. McPeek, and W. Miller. 2005. The dynamics of evolutionary stasis. Paleobiology 31:133–145.

Estes, S., and S. J. Arnold. 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. American Naturalist 169:227–244. doi:10.1086/510633.

Futuyma, D. J. 2010. Evolutionary constraint and ecological consequences. Evolution 64:1865–1884.

Garland, T., Jr., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology 41:18–32.

Gingerich, P. D. 1983. Rates of evolution: effects of time and temporal scaling. Science 222:159–161.

———. 1993. Quantification and comparison of evolutionary rates. American Journal of Science 293A:453–478.

———. 2001. Rates of evolution on the time scale of the evolutionary process. Genetica 112/113:127–144.

———. 2009. Rates of evolution. Annual Review of Ecology, Evolution, and Systematics 40:657–675.

Gould, S. J. 2002. The structure of evolutionary theory. Harvard University Press, Cambridge, MA.

Gould, S. J., and N. Eldredge. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology 3:115–151.

Grabowski, M. 2016. Bigger brains led to bigger bodies? the correlated evolution of human brain and body size. Current Anthropology 57: 174–196.

Grabowski, M., J. D. Polk, and C. C. Roseman. 2011. Divergent patterns of integration and reduced constraint in the human hip and the origins of bipedalism. Evolution 65:1336–1356.

Gray, J. S. 1997. Marine biodiversity: patterns, threats and conservation needs. Biodiversity and Conservation 6:153–175.

Haldane, J. B. S. 1949. Suggestions as to quantitative measurement of rates of evolution. Evolution 3:51–56.

Haller, B. C., and A. P. Hendry. 2014. Solving the paradox of stasis: squashed stabilizing selection and the limits of detection. Evolution 68:483–500.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

———. 2003. Is modularity necessary for evolvability? remarks on the relationship between pleiotropy and evolvability. Biosystems 69: 83–94.

———. 2012. Adaptive landscapes and macroevolutionary dynamics. Pages 205–226 *in* E. I. Svensson and R. Calsbeek, eds. The adaptive landscape in evolutionary biology. Oxford University Press, Oxford.

Hansen, T. F., and D. Houle. 2004. Evolvability, stabilizing selection, and the problem of stasis. Pages 130–150 *in* M. Pigliucci and K. Preston, eds. Phenotypic integration. Oxford University Press, Oxford.

Hansen, T. F., C. Pélabon, W. S. Armbruster, and M. L. Carlson. 2003. Evolvability and genetic constraint in *Dalechampia* blossoms: components of variance and measures of evolvability. Journal of Evolutionary Biology 16:754–766.

Hansen, T. F., C. Pélabon, and D. Houle. 2011. Heritability is not evolvability. Evolutionary Biology 38:258–277.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965–1977.

Hansen, T. F., and K. L. Voje. 2011. Deviation from the line of least resistance does not exclude genetic constraints: a comment on Berner et al. (2010). Evolution 65:1821–1822.

Harmon, L. J., J. B. Losos, T. J. Davies, R. G. Gillespie, J. L. Gittleman, W. B. Jennings, K. H. Kozak, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64: 2385–2396.

Healy K., T. Guillerme, S. Finlay, A. Kane, S. B. Kelly, D. McClean, D. J. Kelly, I. Donohue, A. L. Jackson, and N. Cooper. 2014. Ecology and mode-of-life explain lifespan variation in birds and mammals. Proceedings of the Royal Society B 281:20140298.

Hendry, A. P., and M. T. Kinnison. 1999. Perspective: the pace of modern life: measuring rates of contemporary microevolution. Evolution 53:1637–1653.

Hill, W. G., and A. Caballero. 1992. Artificial selection experiments. Annual Review of Ecology and Systematics 23:287–310.

Hopkins, M. J., and S. Lidgard. 2012. Evolutionary mode routinely varies among morphological traits within fossil species lineages. Proceedings of the National Academy of Sciences of the USA 109:20520–20525.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics 130:195–204.

———. 2001. Characters as the units of evolutionary change. Pages 109–140 *in* G. P. Wagner, ed. The character concept in evolutionary biology. Academic Press, San Diego, CA.

Hunt, G. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology 32:578–601.

———. 2007. Evolutionary divergence in directions of high phenotypic variance in the ostracode genus *Poseidonamicus*. Evolution 61: 1560–1576.

———. 2008. Gradual or pulsed evolution: when should punctuational explanations be preferred? Paleobiology 34:360–377.

———. 2012. Measuring rates of phenotypic evolution and the inseparability of tempo and mode. Paleobiology 38:351–373.

Hunt, G., M. J. Hopkins, and S. Lidgard. 2015. Simple versus complex models of trait evolution and stasis as a response to environmental change. Proceedings of the National Academy of Sciences of the USA 112:4885–4890.

Hunt, G., and D. L. Rabosky. 2014. Phenotypic evolution in fossil species: pattern and process. Annual Review of Earth and Planetary Sciences 42:421–441.

Ingram, T., and D. L. Mahler. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. Methods in Ecology and Evolution 4:416–425.

Jackson, J. B. C., and A. H. Cheetham. 1999. Tempo and mode of speciation in the sea. Trends in Ecology and Evolution 14:72–77.

Khabbazian, M., R. Kriebel, K. Rohe, and C. Ané. 2016. Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. Methods in Ecology and Evolution 7:811–824.

Kingsolver, J. G., S. E. Diamond, A. M. Siepielski, and S. M. Carlson. 2012. Synthetic analyses of phenotypic selection in natural populations: lessons, limitations and future directions. Evolutionary Ecology 26:1101–1118.

Kinnison, M. T., and A. P. Hendry. 2001. The pace of modern life. II. From rates of contemporary microevolution to pattern and process. Genetica 112/113:145–164.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain : body size allometry. Evolution 33:402–416.

Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Lieberman, B. S., C. E. Brett, and N. Eldredge. 1995. Patterns and processes of stasis in two species lineages of brachiopods from the Middle Devonian of New York State. Paleobiology 21:15–27.

Lieberman, B. S., and S. Dudgeon. 1996. An evaluation of stabilizing selection as a mechanism for stasis. Palaeogeography 127:229–238.

Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. American Naturalist 136:727–741.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, MA.

Marroig, G., and J. M. Cheverud. 2005. Size as line of least evolutionary resistance: diet and adaptive morphological radiation in New World monkeys. Evolution 59:1128–1142.

Monnet, C., K. De Baets, and C. Klug. 2011. Parallel evolution controlled by adaptation and covariation in ammonoid cephalopods. BMC Evolutionary Biology 11:115.

Morrissey, M. B. 2016. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. Journal of Evolutionary Biology 29:1882–1904.

Morrissey, M. B., and J. D. Hadfield. 2012. Directional selection in temporally replicated studies is remarkably consistent. Evolution 66:435–442.

Norris, R. D. 2000. Pelagic species diversity, biogeography, and evolution. Paleobiology 26:236–258.

Nunney, L. 1991. The influence of age structure and fecundity on effective population size. Proceedings of the Royal Society B 246:71–76.

———. 1993. The influence of mating system and overlapping generations on effective population size. Evolution 47:1329–1341.

Palstra, F. P., and D. J. Fraser. 2012. Effective/census population size ratio estimation: a compendium and appraisal. Ecology and Evolution 2:2357–2365.

Patzkowsky, M. E., and S. M. Holland. 2012. Stratigraphic paleobiology. University of Chicago Press, Chicago.

Pearson, P. N., and T. H. G. Ezard. 2014. Evolution and speciation in the Eocene planktonic foraminifer *Turborotalia*. Paleobiology 40: 130–143.

Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmond. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216–2218.

Pennell, M. W., R. G. FitzJohn, W. K. Cornwell, and L. J. Harmon. 2015. Model adequacy and the macroevolution of angiosperm functional traits. American Naturalist 186:E33–E50.

Peters, R. H. 1983. The ecological implications of body size. Cambridge University Press, New York.

R Development Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Roopnarine, P. D. 2003. Analysis of rates of morphologic evolution. Annual Review of Ecology, Evolution, and Systematics 34:605–632.

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766–1774.

Schmidt-Nielsen, K. 1984. Scaling—why is animal size so important? Cambridge University Press, New York.

Serbezov D., P. E. Jorde, L. Bernatchez, E. M. Olsen, and L. A. Vøllestad. 2012. Life history and demographic determinants of effective/census size ratios as exemplified by brown trout (*Salmo trutta*). Evolutionary Applications 5:607–618.

Sheets, H. D., and C. E. Mitchell. 2001. Why the null matters: statistical tests, random walks and evolution. Genetica 8:105–125.

Slater, G. J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. Methods in Ecology and Evolution 4:734–744.

———. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. Proceedings of the National Academy of Sciences of the USA 112:4897–4902.

Slater, G. J., and M. W. Pennell. 2013. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Systematic Biology 63:293–308.

Smith, J. M. 1983. The genetics of stasis and punctuation. Annual Review of Genetics 17:11–25.

Thomas, G. H., and R. P. Freckleton. 2012. MOTMOT: models of trait macroevolution on trees. Methods in Ecology and Evolution 3:145–151.

Thornhill D. J., A. R. Mahon, J. L. Norenburg, and K. M. Halanych. 2008. Open-ocean barriers to dispersal: a test case with the Antarctic Polar Front and the ribbon worm *Parborlasia corrugatus* (Nemertea: Lineidae). Molecular Ecology 17:5104–5117.

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. Proceedings of the National Academy of Sciences of the USA 108:15908–15913.

Uyeda, J. C., and L. J. Harmon. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Systematic Biology 63:902–918.

Vindenes, Y., A. M. Lee, S. Engen, and B. E. Sæther. 2010. Fixation of slightly beneficial mutations: effects of life history. Evolution 64: 1063–1075.

Voje, K. L. 2016. Tempo does not correlate with mode in the fossil record. Evolution 70:2678–2689.

Voje, K. L., J. Starrfelt, and L. H. Liow. 2017. Data from: Model adequacy and microevolutionary explanations for stasis in the fossil record. American Naturalist, Dryad Digital Repository, http://dx.doi.org/10.5061/dryad.r5d10.

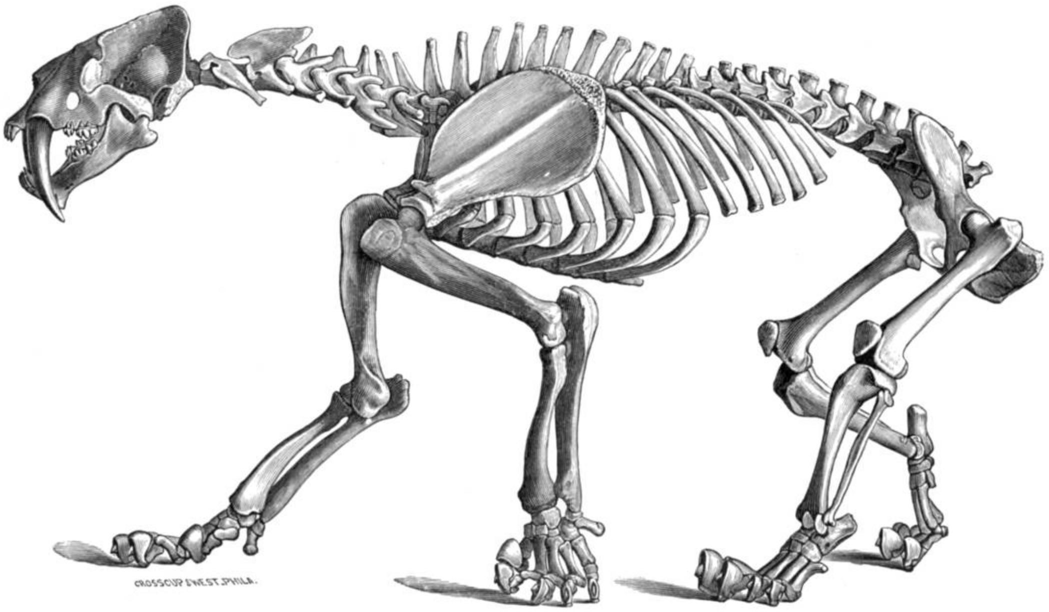
Walsh, B., and M. W. Blows. 2009. Abundant genetic variation + strong selection = multivariate genetic constraints: a geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics 40:41–59.

Wang, J. 2005. Biological sciences estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society B 360:1395–409.

Waples, R. S., G. Luikart, J. R. Faulkner, and D. A. Tallmon. 2013. Simple life-history traits explain key effective population size ratios across diverse taxa. Proceedings of the Royal Society B 280:20131339.

West, G. B., J. H. Brown, and B. J. Enquist. 1997. A general model for the origin of allometric scaling laws in biology. Science 276:122–126.

Associate Editor: Scott J. Steppan
Editor: Judith L. Bronstein



"*Smilodon necator* Gervais . . . The species is about the size of the lion, and of the most formidable character." From "On the Extinct Cats of America" by E. D. Cope (*The American Naturalist*, 1880, 14:833–858).