Supporting Information Appendix

Includes:

Supplemental Text
- Updating previous compilations
- Evolutionary models considered in this paper
- Age model error

Supplemental Tables (S1-S5)

Supplemental Figures (Fig. S1 – S9)

Supplemental Text

Updating previous compilations

Paleontological sequence data from two recent compilations (1, 2) were retained, with the following exceptions:

- Unpublished data from Hunt and Brown in (1) was replaced with published versions (3).
- We corrected an error in one sequence from (2) for a sequence published by Bown and Rose (4).
- Two studies computed a large number of principal component scores that were included in the original compilations. As trailing axes are likely to be dominated by sampling error, we discarded all PC scores beyond PC 6 for the data from Hunt (5) and Kim et al. (6).
- The earlier compilations analyzed sequences with at least six sampled populations. We increased this threshold to seven to lessen the difficulty in assessing evolutionary mode in very short sequences.

Evolutionary models considered in this paper

Random Walks and Directional Evolution. Time in these models occurs in discrete increments, at each of which an evolutionary change is drawn independently from a distribution of evolutionary steps. As long as more than a few generations occur between sampling opportunities, the behavior of this model depends only on the mean and variance of the step distribution (7). If the mean of the step distribution is zero, on average traits increase as much as they decrease, and the result is an unbiased random walk (hereafter, random walk). If the mean step is nonzero, the resulting sequence has underlying directionality, although a trend might not be evident in a particular realization if it is weak compared to the variability endowed by the step variance. Directional evolution has three parameters: the ancestral state, the mean step, and the step variance. A random walk is a special case of this directional evolution model in which the mean step is zero.

Stasis. Stasis has been modeled in various ways in the literature. Here we follow Sheets and Mitchell (8) in construing stasis as white noise – uncorrelated, normally distributed variation around a steady mean. This model has two parameters: the mean and the variance around that mean.

This model of stasis has the same mathematical form as sampling error, so it will have a tendency to be preferred whenever true evolutionary changes are quite small, regardless of their form (7, 9). For example, if a trend or random walk results in trait changes so modest that they are swamped by sampling noise, it is likely that the stasis model will win any model comparison. This property is not necessarily a disadvantage, though, as it is consistent with notions of stasis that emphasis the magnitude of change rather than their form (see discussion in 10).

Strict Stasis. This model is a special case of stasis for which the variance around the long-term mean is zero. In other words, this represents the strictest notion of stasis as no real evolutionary

differences among samples; all observed differences in this model must be accounted for by sampling error. This is a useful model because it is of biological interest to distinguish truly static morphologies from the broader definition of stasis that can encompass substantial fluctuations around a steady mean. In addition, this model allows us to avoid ambiguous model interpretation when there are no real evolutionary differences among samples in a sequence. In this case, the stasis model (with zero variance) and the random walk model (with zero step variance) become equivalent, and thus have equal likelihoods and model support. This shared support for two models makes it seem as though there is equal evidence for two different modes of evolution, whereas instead both model fits are special cases that correspond to the same model. Adding strict stasis solves this issue because, even though it has the same likelihood as stasis with a fitted variance parameter of zero, it has one fewer parameter and therefore garners higher model support because of its parsimony advantage.

Punctuations. This model is discussed extensively elsewhere (11), where it is referred to as an "unsampled punctuation." This model involves two segments of stasis with an instantaneous shift in the stasis mean between the two segments. We constrain the stasis variance to be shared across the two segments (see 11), so this model has four parameters: two stasis means (one for each segment), a stasis variance that is shared across segments, and the timing of the shift from one stasis interval to the other.

Mode-shift models. This kind of model, discussed briefly in (11), considers an evolutionary sequence to be composed of two segments, each of which evolves according to its own evolutionary model. Full maximum-likelihood inference (the "Joint" option in the *paleoTS* package) uses the appropriate likelihood functions in (12), applied to each segment separately. These models were named according to the operative models in the two segments of the sequence, separated by a hyphen:

- *Stasis – Random walk* has four parameters: the stasis mean and variance, the step variance (rate) of the random walk, and the timing of the mode shift. The ancestral trait value of the random walk segment was taken to be the stasis mean of the initial segment.
- *Stasis – Directional* has the same parameters as above, plus an additional directionality parameter (the mean step) that determines the strength of the trend.
- *Random walk – Stasis* has five parameters: the ancestral state value and step variance of the random walk, the mean and variance of stasis, and the timing of the mode shift.
- *Directional – Stasis* has the same parameters as *Random walk – Stasis*, plus an additional directionality parameter (the mean step) that determines the strength of the trend.

Age model error

Fitting evolutionary models requires that ages be estimated for the analyzed fossil populations and such age models are always subject to uncertainty. Some kinds of chronological errors will not affect the relative support of different models (11), such as those that are uniformly additive or multiplicative (i.e., all ages are too old or young by the same amount or factor). Thus, the kinds of error that are of most concern are those that alter the *relative* temporal spacing of samples.

Constructing age models involves some degree of age interpolation because it is almost never practical, or even possible, to directly date all samples independently. Rather, one or more tie points of absolute age (or absolute durations between points) are determined, and then an assumption of constant sedimentation rate is used to assign ages to samples between tie points. Given this interpolation, it is a possible that paleontological age models may systematically miss real, significant variations in sedimentation rates. If this is the case for a particular fossil series, it is likely that its age model will be overly smooth because of this undetected variation.

When just one or a few time-series are of interest, it is prudent to perform simulations that explore model fits over plausible alternative age models (11). This is impractical for a large compilation, and so instead we use a series of simulations to explore the potential for undetected variation in sedimentation rate to influence our analyses. Specifically, we simulate trait evolution under simple models of evolutionary mode (random walk, directional change, stasis) and real age models that have substantial variation in sedimentation rates. We then smooth these sedimentation rates so that samples are wrongly assumed to be equally spaced. Finally, we fit the full set of models to the trait data under the correct and incorrect age models. Of particular interest is whether undetected variation in sedimentation rate shifts support towards or away from complex (punctuated and mode-shift) models of evolutionary mode.

All simulated time series were assumed to span 1 Myr. Sampling times were generated by sampling ages uniformly from 0 to 1 Myr and sorting from smallest to largest. This produces age models with substantial variation in temporal spacing; most points are rather close together with fewer larger gaps (point-to-point durations are exponentially distributed under this approach). These true age models were retained, and smoothed age models were also generated by spacing points evenly over the 1 Myr duration of the sequence. This smoothing is probably unrealistically severe for many sedimentological conditions, biasing against our methods/assumptions in the primary analyses of this paper. However, it is valuable to exaggerate this effect in order to get an upper bound on its likely role. We generated trait sequences according to (i) a random walk with unit step variance, (ii) a moderately strong trend (mean step = 2, unit step variance), and (iii) stasis (stasis variance = 1). We simulated sequences of 21, 41, and 61 samples, with 500 replicates for each combination of generating model and sequence length.

For each realized sequence, we fit the suite of nine models to the trait data using the true and (incorrectly) smoothed age models. We are most concerned that undetected variation in sedimentation rates might unduly favor complex evolutionary models, so we summarized the aggregate support for all complex models as their summed Akaike weights and compared this value across both age models. Figure S9 plots the difference between this summed support between the true and smoothed age models, with positive values indicating that support for complex models is spuriously increased by smoothing the spacing between samples.

There is a tendency for age model error caused by undetected variation in sedimentation rate to shift support to more complex models of evolution, but the effect is not large. The median increase in summed Akaike weight across mode-shift and punctuational models ranges from 2.5% to 7.8% across different null models and sequence lengths (Fig. S9). Importantly, this

effect does not vary much with sequence length. Thus, this source of age model error is not likely to account for the increasing dominance of complex models in longer sequences.

Supplemental References

1. Hunt G (2007) The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. *Proc. Natl. Acad. Sci. USA* 104(47):18404-18408.
2. Hopkins MJ & Lidgard S (2012) Evolutionary mode routinely varies among morphological traits within fossil species lineages. *Proc. Natl. Acad. of Sci. USA* 109(50):20520-20525.
3. Hunt G, Wicaksono S, Brown JE, & Macleod GK (2010) Climate-driven body size trends in the ostracod fauna of the deep Indian Ocean. *Palaeontology* 53(6):1255-1268.
4. Bown TM & Rose KD (1986) Patterns of dental evolution in Early Eocene anaptomorphine primates (Omomyidae) from the Bighorn Basin, Wyoming. *Paleontological Society Memoir* 23:162 p.
5. Hunt G (2007) Evolutionary divergence in directions of high phenotypic variance in the ostracode genus *Poseidonamicus*. *Evolution* 61(7):1560-1576.
6. Kim K, Sheets HD, & Mitchell CE (2009) Geographic and stratigraphic change in the morphology of *Triarthrus beckii* (Green) (Trilobita): a test of the Plus ça change model of evolution. *Lethaia* 42(1):108-125.
7. Hunt G (2006) Fitting and comparing models of phyletic evolution: random walks and beyond. *Paleobiology* 32(4):578-601.
8. Sheets HD & Mitchell CE (2001) Why the null matters: statistical tests, random walks and evolution. *Genetica* 112-113:105-125.
9. Hannisdal B (2006) Phenotypic evolution in the fossil record: Numerical experiments. *J. Geol.* 114(2):133-153.
10. Hunt G & Rabosky DL (2014) Phenotypic evolution in fossil species: pattern and process. *Annu. Rev. Earth Planet. Sci.* 42:in press.
11. Hunt G (2008) Gradual or pulsed evolution: when should punctuational explanations be preferred? *Paleobiology* 34(3):360-377.
12. Hunt G (2008) Evolutionary patterns within fossil lineages: model-based assessment of modes, rates, punctuations and process. *From Evolution to Geobiology: Research Questions Driving Paleontology at the Start of a New Century,* The Paleontological Society Papers, eds Bambach RK & Kelley PH (The Paleontological Society, New Haven, Connecticut), pp 117-131.
13. Theriot EC, Fritz SC, Whitlock C, & Conley DJ (2006) Late Quaternary rapid morphological evolution of an endemic diatom in Yellowstone Lake, Wyoming. *Paleobiology* 32(1):38-54.
14. Knappertsbusch M (2000) Morphologic evolution of the coccolithophorid *Calcidiscus leptoporus* from the Early Miocene to recent. *J. .Paleont.* 74(4):712-730.
15. Kucera M & Malmgren BA (1998) Differences between evolution of mean form and evolution of new morphotypes: an example from late Cretaceous planktonic foraminifera. *Paleobiology* 24(1):49-63.

16. Wildenborg AFB (1991) Evolutionary aspects of the miogypsinids in the Oligo-Miocene carbonates near Mineo (Sicily). *Utrecht Micropaleontological Bulletins* 41:1-139.

Supplemental Tables

Table S1. Post-hoc significant differences among *env* and *taxon* variables values for logistic model predicting whether Strict Stasis will be the best-supported model. The environment variable (*env*) takes on values of lacustrine, marine-deep, marine-pelagic, marine-shelf, and terrestrial. The *taxon* variable has possible values of macroinvertebrate (invert), benthic microfossil (bm), planktonic microfossil (pm) and vertebrate (vert). Significance codes: ***, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$

Variable	**Difference**	**Estimate**	**P-value**
env	marine.deep – lacustrine	5.4778	0.00442 **
	marine.pelag – lacustrine	4.7036	0.01118 *
	marine.shelf – lacustrine	4.5704	0.0132 *
	terrestrial – lacustrine	1.7487	0.46269
	marine.pelag – marine.deep	-0.7742	0.6871
	marine.shelf – marine.deep	-0.9074	0.35331
	terrestrial – marine.deep	-3.729	0.01773 *
	marine.shelf – marine.pelag	-0.1332	0.99559
	terrestrial – marine.pelag	-2.9549	0.05515
	terrestrial – marine.shelf	-2.8217	0.06277
taxon	invert – bm	1.6126	< 0.001 ***
	pm – bm	-0.9824	0.33927
	vert – bm	3.7171	0.00572 **
	pm – invert	-2.595	< 0.001 ***
	vert – invert	2.1045	0.17943
	vert – pm	4.6995	< 0.001 ***

Table S2. Post-hoc significant differences among *env* and *taxon* variables values for logistic model predicting whether Stasis (including strict stasis) will be the best-supported model. Conventions follow Table S1.

Variable	**Difference**	**Estimate**	**P-value**
env	marine.deep – lacustrine	2.549	0.03961 *
	marine.pelag – lacustrine	1.4196	0.36652
	marine.shelf – lacustrine	2.2357	0.02407 *
	terrestrial – lacustrine	-0.5064	0.94329
	marine.pelag – marine.deep	-1.1294	0.26155
	marine.shelf – marine.deep	-0.3133	0.95474
	terrestrial – marine.deep	-3.0554	0.00141 **
	marine.shelf – marine.pelag	0.8161	0.14939
	terrestrial – marine.pelag	-1.926	0.04047 *
	terrestrial – marine.shelf	-2.7421	< 0.001 ***
taxon	invert – bm	1.8388	<0.001 ***
	pm – bm	1.0022	0.16
	vert – bm	2.8095	<0.001 ***
	pm – invert	-0.8366	0.1058
	vert – invert	0.9707	0.36
	vert – pm	1.8073	0.0542

Table S3. Post-hoc significant differences among *env* and *taxon* variables values for logistic model predicting whether Directional evolution will be the best-supported model. Conventions follow Table S1. Contrasts for the environment variable not shown because this variable was not retained by the stepwise AIC procedure.

Variable	**Difference**	**Estimate**	**P-value**
taxon	invert – bm	-0.8178	0.195
	pm – bm	-0.5265	0.545
	vert – bm	0.1509	0.988
	pm – invert	0.2914	0.817
	vert – invert	0.9688	0.121
	vert – pm	0.6774	0.368

Table S4. Post-hoc significant differences among *env* and *taxon* variables values for logistic model predicting whether any of the complex models will be the best-supported model. Conventions follow Table S1.

Variable	**Difference**	**Estimate**	**P-value**
env	marine.deep – lacustrine	-1.807	0.43301
	marine.pelag – lacustrine	-15.8814	1
	marine.shelf – lacustrine	-1.1714	0.4024
	terrestrial – lacustrine	0.8891	0.71845
	marine.pelag – marine.deep	-14.0744	1
	marine.shelf – marine.deep	0.6357	0.93397
	terrestrial – marine.deep	2.6961	0.06197
	marine.shelf – marine.pelag	14.7101	1
	terrestrial – marine.pelag	16.7705	1
	terrestrial – marine.shelf	2.0604	0.00981 **
taxon	invert – bm	-1.1914	0.04847 *
	pm – bm	14.5462	1
	vert – bm	-2.8895	0.00183 **
	pm – invert	15.7376	1
	vert – invert	-1.6981	0.05362
	vert – pm	-17.4357	1

Table S5. Frequency with which each model is the best supported under the "Joint" and "AD" approaches to fitting the models, which correspond to maximum-likelihood (ML) and restricted maximum-likelihood (REML), respectively. Number of estimated parameters for each model in parentheses; where number is different for the "AD" approach, this is given in italics (the "AD" approach does not require a parameter estimating the trait mean at the beginning of the series, 12). Conventions follow Table 1.

Model category	**Model**	**Joint (ML)**	**AD (REML)**
Simple	Strict Stasis (1)	124	144
	Stasis (2)	147	153
	Random Walk (2, *1*)	201	190
	Directional (3, *2*)	62	27
Complex	Punctuation (4)	67	92
	Stasis-RW (4)	19	30
	Stasis-Dir (5)	16	8
	RW-Stasis (5, *4*)	59	59
	Dir-Stasis (6, *5*)	14	6

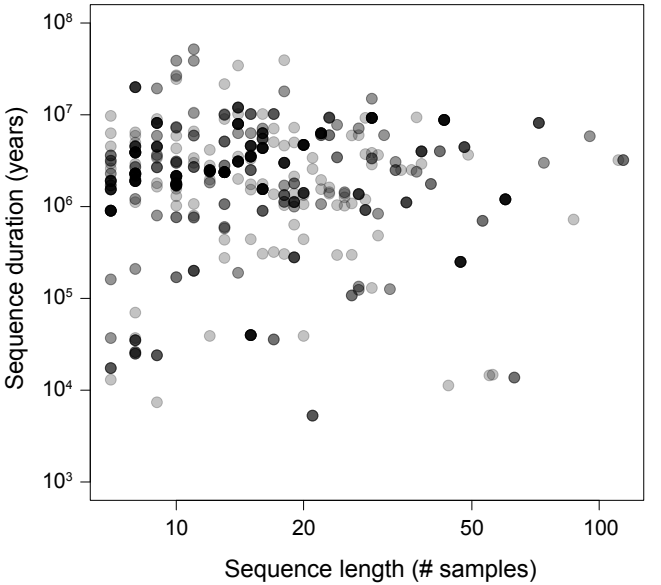

Figure S1. Sequence duration plotted with respect to sequence length for the 709 time-series of trait evolution captured in the fossil record. Points are semi-transparent; darker shades represent overplotting of multiple points. Both variables plotted on a logarithmic scale.

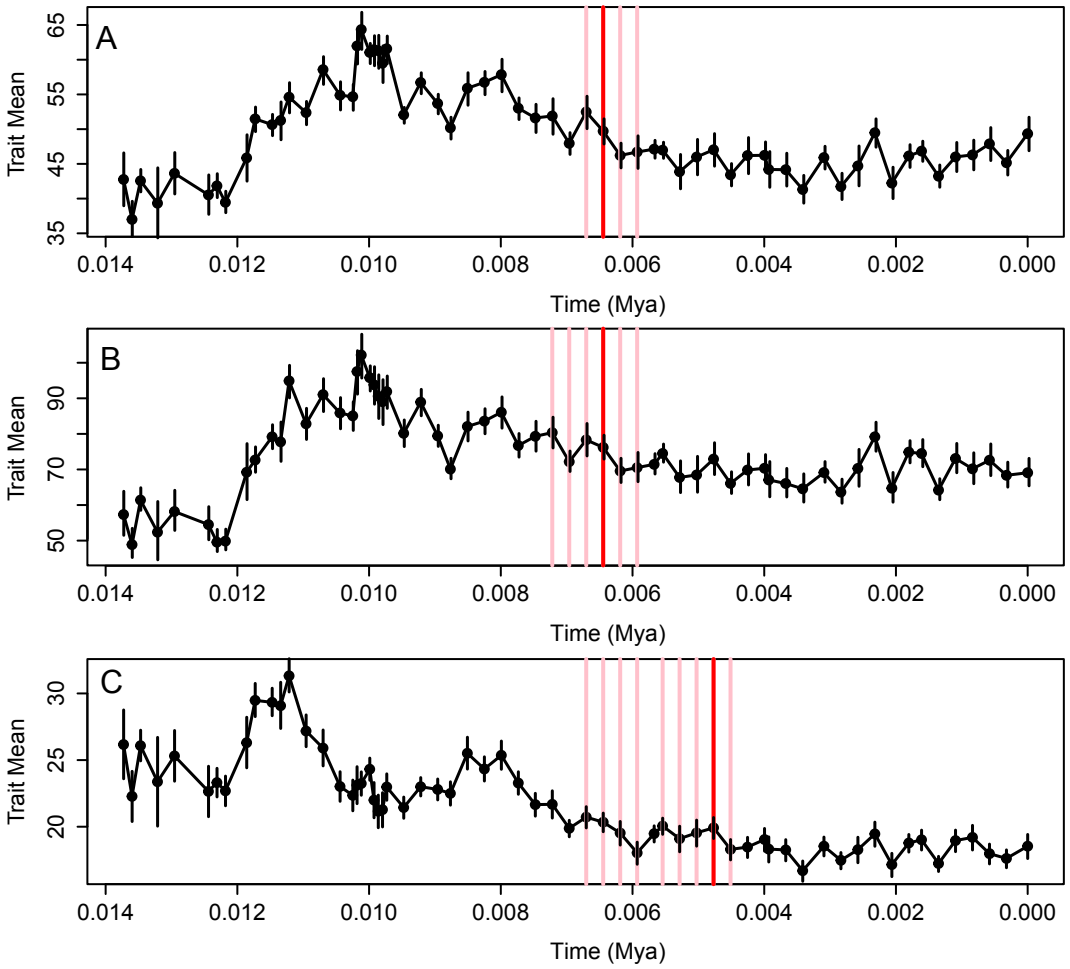

Figure S2. Example of a species lineage where all measured traits show the same mode of evolution (here, a shift from a random walk to stasis) and the shift in evolutionary dynamics coincides in time. Error bars represent 95% confidence intevals. Red line shows maximum-likelihood solution and pink lines show all solutions within 1.92 log-likelihood units of the maximum-likelihood solution (corresponding to a 95% confidence interval around the placement of the shift point). Time in millions of years ago. Data from *Stephanodiscus yellowstonensis* (13). (A) Diameter of diatom test. (B) Number of costae. (C) Number of spines.

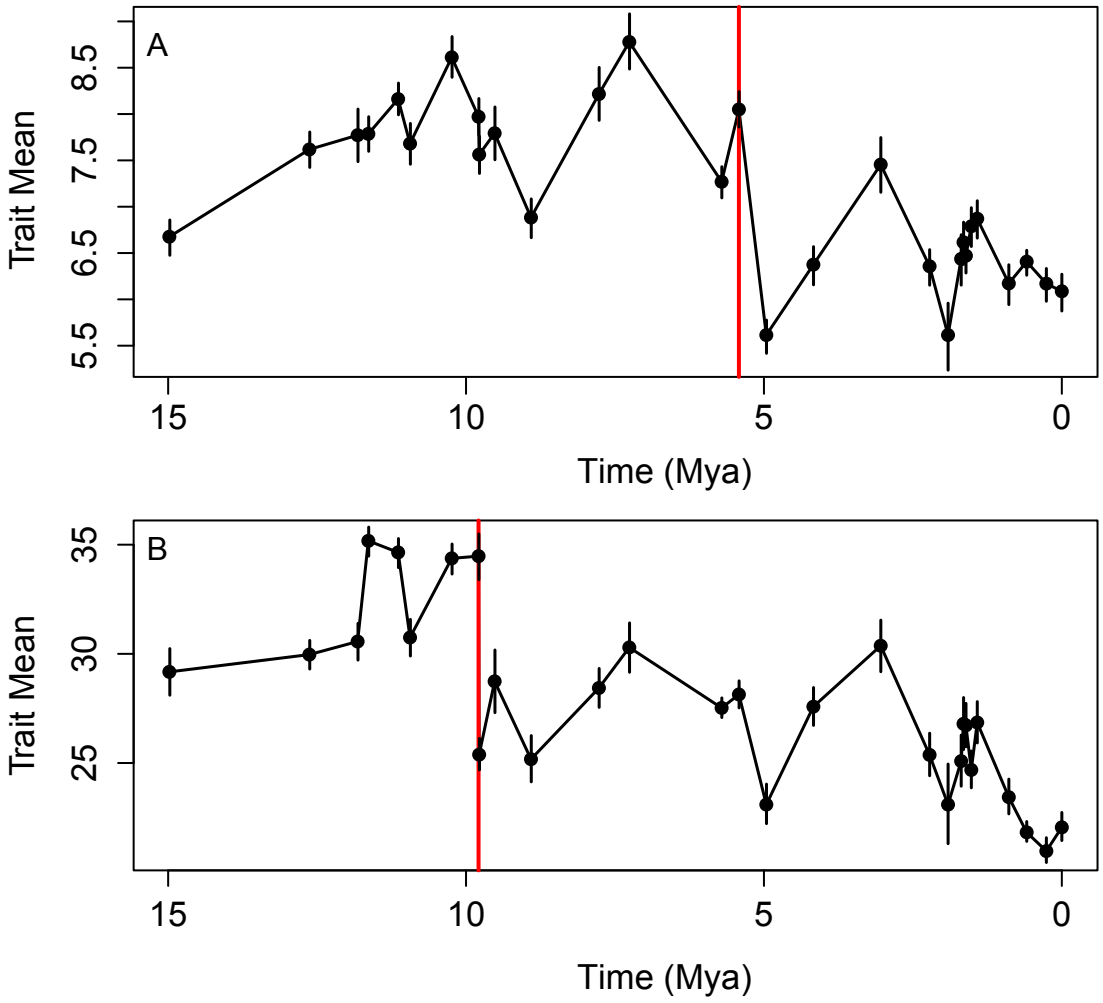

Figure S3. Example of a species lineage where measured traits show the same mode of evolution (here, punctuation bracketed by two intervals of stasis) but the timing of the shift in evolutionary dynamics is different. Error bars are 95% confidence intervals. Red line shows maximum-likelihood solution (no other shift points are in the 95% confidence interval around the placement of the shift point). Time in millions of years ago. Data from *Calcidiscus leptoporus* (14). (A) Diameter of distal shield. (B) Number of elements in distal shield.

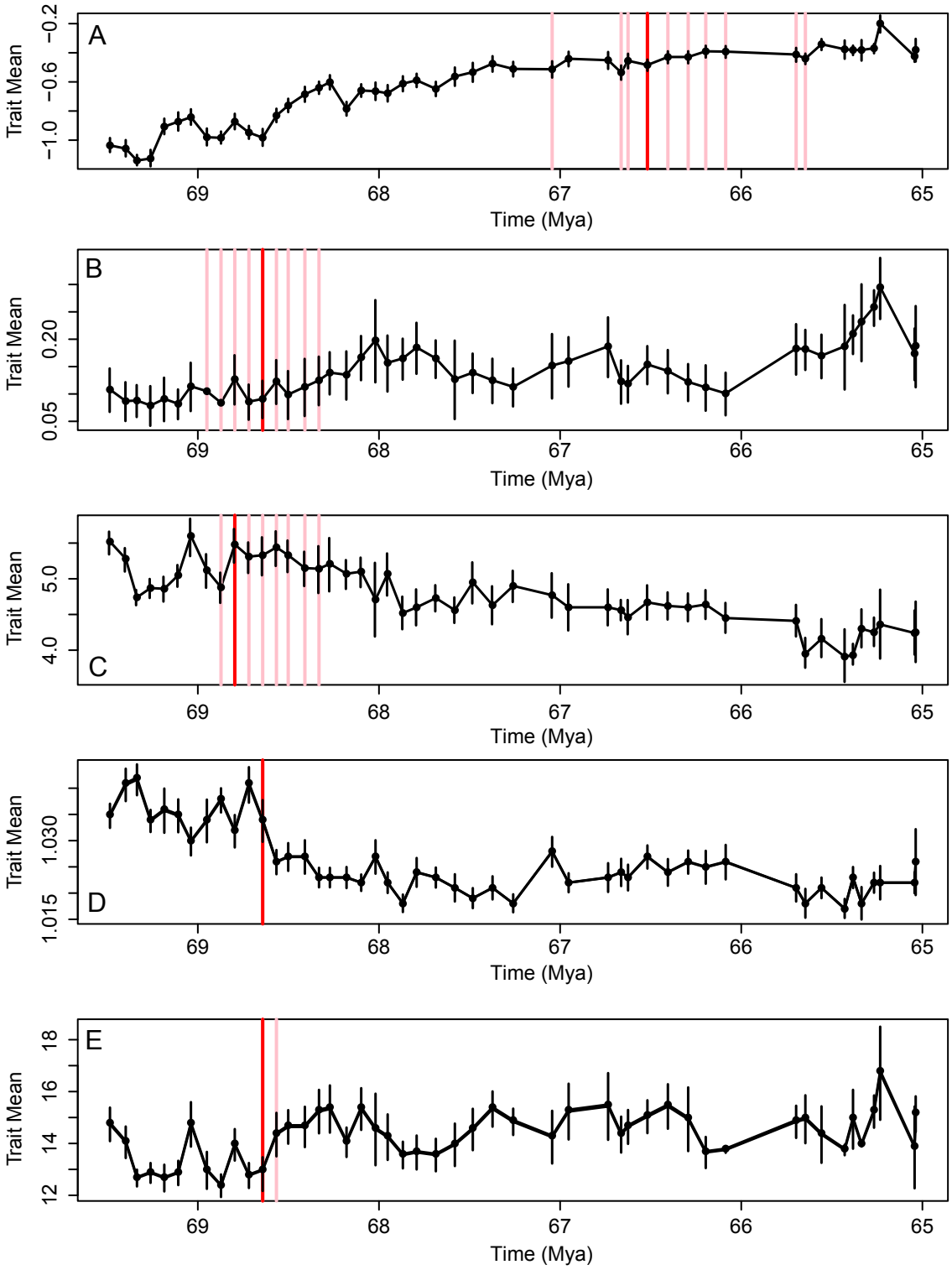

Figure S4. Example of a species lineage where most measured traits show coordinated timing of shifts in evolutionary dynamics even though the pattern of evolution varies. Error bars are 95% confidence intervals. Red line shows maximum-likelihood solution and pink lines show all solutions within 1.92 log-likelihood units of the maximum-likelihood solution (corresponding to a 95% confidence interval around the placement of the shift point). Time in millions of years ago. Data from *Contusotruncata sp.* (15). (A) Conicity of test, best-fit by a random walk shifting to stasis. (B) Mean test size, best-fit by stasis followed by a random walk. (C) Mean number of chambers in the last whorl, best-fit by stasis followed by a random walk. (D) Mean spiral roundness of test, best-fit by a punctuation separating two intervals of stasis. (E) Mean number of chambers, best-fit by a punctuation separating two intervals of stasis.

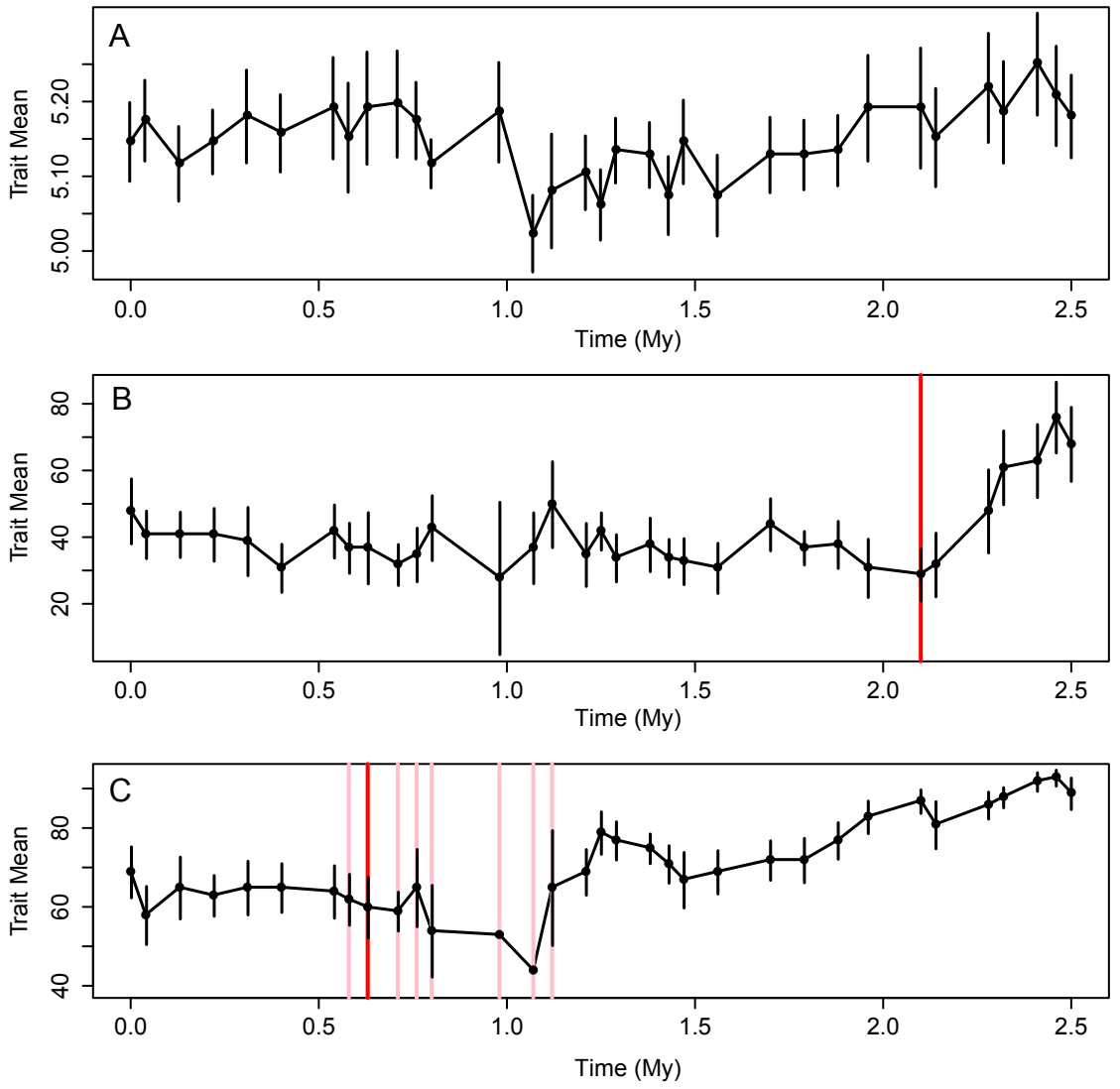

Figure S5. Example of a species lineage where both evolutionary mode and timing of shifts in evolutionary mode (where they occur) vary across measured traits. Error bars are 95% confidence intervals. Red line shows maximum-likelihood solution and pink lines show all solutions within 1.92 log-likelihood units of the maximum-likelihood solution (corresponding to a 95% confidence interval around the placement of the shift point). Time in millions of years. Data from *Miogypsina sp.* (16). (A) Diameter of protoconch, best-fit model is a random walk. (B) Orientation of nepiont, best-fit model is stasis followed by directional evolution. (C) Degree of symmetry of the protocochial spirals, best-fit model is stasis followed by a random walk.

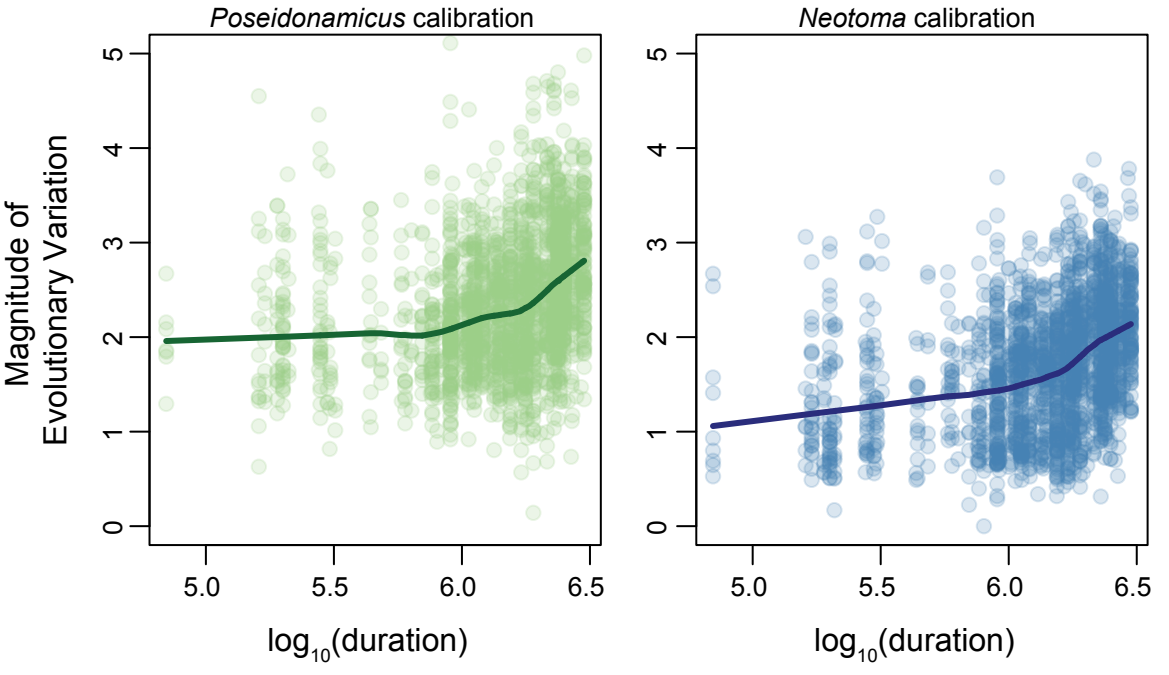

Figure S6. Results of simulations from the temperature-tracking model, calibrated by *Poseidonamics major* (left) and *Neotoma cinerea* (right). Vertical axis is magnitude of evolutionary variation measured as the standard deviation of samples in a sequence, with the contribution from measurement error removed (see Methods). Horizontal axis is sequence duration, in years and on a $\log_{10}$ scale (a value of 6 is 1 Myr). Points are semi-transparent and lines represent locally weighted (lowess) regressions with a smoother span = 0.5.

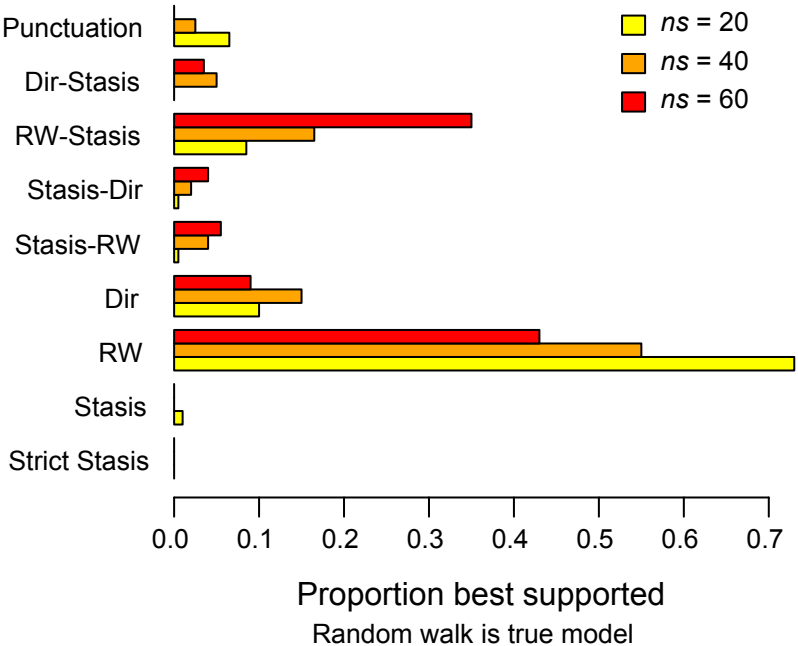

Figure S7. Results of simulations suggesting complex models can be unduly favored by AICc when simple models are true. Shown is the distribution of best-supported models for 200 simulated random walks under three sequence lengths: *ns* = 20, *ns* = 40, and *ns* = 60 (in yellow, orange, and red, respectively). Mode-shift models are increasingly favored as the number of samples in a sequence increase.

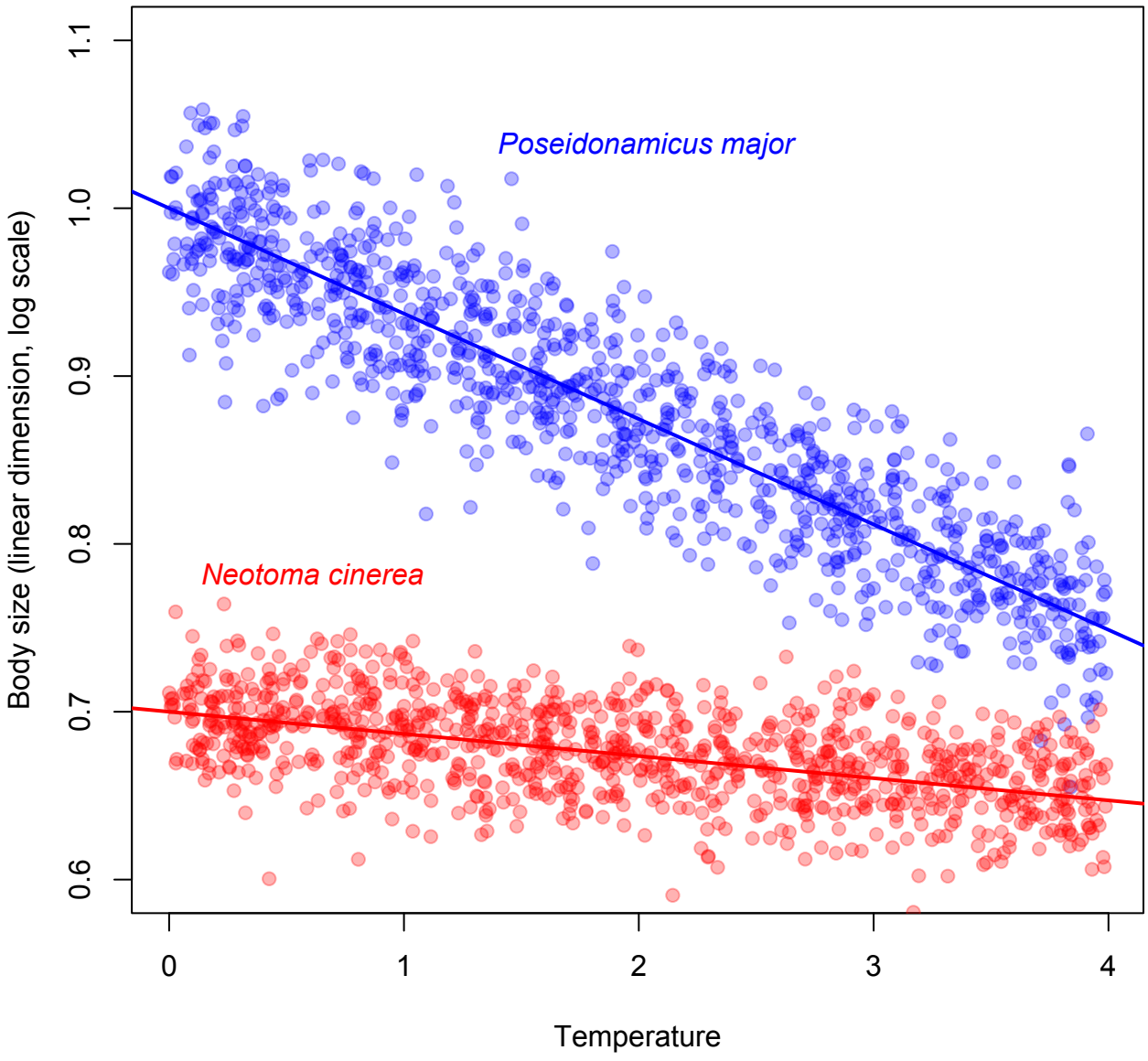

Figure S8. Illustration of two calibrations of the temperature-tracking model based on instances of Bergmann's rule, an inverse relationship between body size and temperature. In blue is the deep-sea ostracode *Poseidonamics major* and in red is the packrat *Neotoma cinerea*. The lines capture the slopes of the temperature-size relationships (intercepts do not matter and are arbitrary here). Points illustrate the magnitude of residual variance (generated from 1,000 uniformly distributed temperatures).

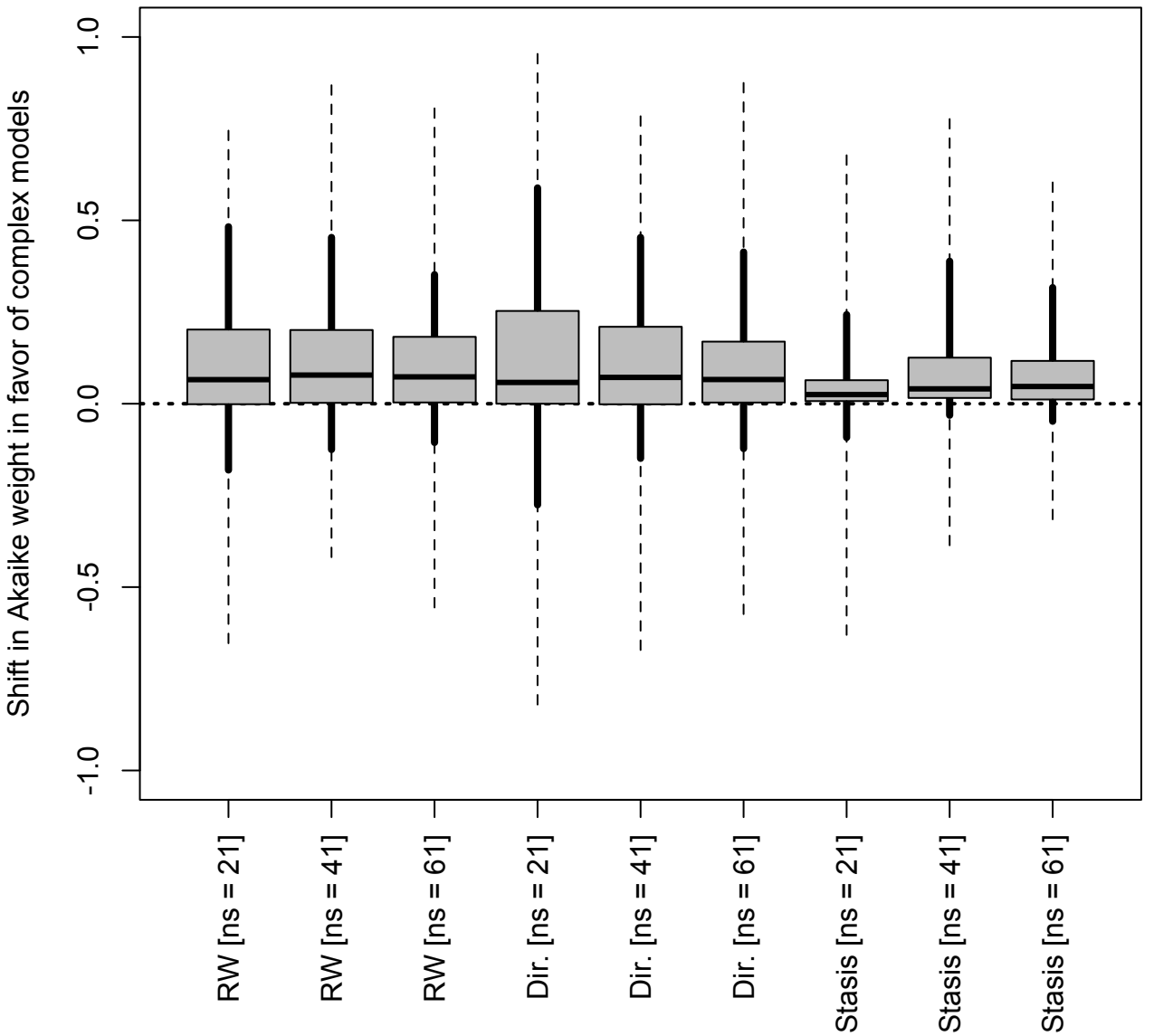

Figure S9. Boxplots summarizing the effects of age model error caused by undetected variation in sedimentation rate. The quantity plotted is the difference in the aggregate support for complex evolutionary models (summed Akaike weights for punctuation and mode-shift models). Median is horizontal bar; box is interquartile range, thick vertical lines delimit 90% probability region and thin bars give total range over 500 simulated replicates. Null models are random walk (RW), directional evolution (Dir), and stasis, with the number of samples (ns) in the sequence enclosed in brackets. Positive values indicate that statistical support for complex models is increased by the presence of undetected variation in sedimentation rate.