Supporting Information

Crampton et al. 10.1073/pnas.1714342115

Data and CONOP Method

The data and methods used to generate and calibrate the CONOP composite have been well-described elsewhere (17, 29, 41), and these details are not repeated here.

Basic HMM

Using the CONOP composite, we fit time series of speciation and extinction probabilities using discrete time HMMs, which (as implemented here) approximate these probability distributions using discrete states. Briefly, HMMs assume that the system being modeled is a process that satisfies the Markov property, meaning that the future state is dependent only on the current state and the between-state transition probabilities. In the case of HMMs, the states are unobserved and are inferred from the data (30). Key outputs from the HMMs are the speciation/extinction probabilities assigned to the discrete states, the transition probabilities relating to transitions between states, and the predicted time series of states. HMMs have seen little use in the paleontological literature, although they have been applied to the analysis of character states in a phylogenetic context (42). HMM fitting, simulations, and predictions are computed here using the package HiddenMarkov (43) in the R language for statistical computing (40); blocks of R code for the fitting of HMMs and for spectral analyses (below) can be downloaded as dataset files that are referenced at relevant places in the SI text.

To replace the CONOP levels that are unevenly spaced in time with pseudolevels that are evenly spaced, we simply shift or "collapse" first and last occurrences in the composite onto the closest pseudolevel. In effect, this step imposes short-duration, uniform-length time bins and assumes that first and last occurrences are at the center of each bin. This protocol degrades the resolution of the composite but on average, results in just a 0.012-My adjustment to the age of each occurrence event and a 0.017-My adjustment to species durations. The 0.05-My interval was chosen conservatively to avoid unrealistic interpolation of pseudolevels between widely spaced levels in the composite but to yield a very high-resolution time series. In this approach, we implicitly assume that pseudolevels without any first or last occurrences are true records of no events for our given sample of species rather than absence of data, an assumption that is consistent with the nature of the CONOP compositing process.

At each pseudolevel, we count the number of speciations or extinctions—encoded as *counts* in the code (Dataset S2)—and the total number of species extant. For extinction calculations, the count of extant species includes the extinguishers (which form part of the pool of species that are exposed to extinction risk) but not originators. Conversely, for speciation calculations, the count of extant species includes the originators but not the extinguishers. Given these data, the R commands given in Dataset S2 are used to estimate the parameters of an HMM with m discrete states (where m is selected by the user) using the Baum–Welch algorithm (44). From this, the most likely sequence of states corresponding to the observed data is predicted using the Viterbi algorithm (45). Here, *Pi* is the $m \times m$ transition probability matrix, *delta* is the probability distribution of the m hidden states at the first time step, *pm* is the vector of m hidden-state values (by default, these are not ordered by magnitude in the output), and *pn* is the vector containing the number of Bernoulli trials (i.e., the number of species living) at each of N time steps as determined from the data. To run the model, one provides initial estimates of *Pi*, *delta*, and *pm*; in practice, these can be generated using random numbers from a normal distribution, with the constraint that *delta* and the rows of *Pi* must sum to one (Dataset S2).

In general, given a family of models with different numbers of discrete states (m), the preferred model is typically selected to minimize Akaike's Information Criterion (AIC). Here, we have tested values of m between 2 and 10. The use of AIC in model selection is, however, indicative only. We find that, in many cases, although the AIC favors four- or five-state models, these yield unstable parameter estimates and/or at least two states that are identical within error, results that suggest ill-determined solutions and overfitting (the measurement of error on parameter estimates is discussed below). In such instances, we use three-state models that produce well-constrained and clearly discriminated state estimates.

From the HMMs, we derive time series of predicted speciation and extinction probability states (given in object *v* in the code), sum these at each pseudolevel to derive turnover, and use this new time series as input to the spectral analysis (below). The HMM predicted turnover probability is highly correlated with the raw turnover probability calculated at each pseudolevel (compare Fig. 4 *H* and *I*), where raw turnover is based on the number of originating/extinguishing species at each pseudolevel divided by the number of species extant (the method of counting extant species is discussed above): the Pearson correlation coefficient is 0.902 ($P < 0.001$). Here and elsewhere, significance of the correlation coefficient is determined using the phase randomized surrogate approach (39) for evaluation of serially correlated data (function surrogateCor in R package Astrochron).

Key results shown in Fig. 2 *C* and *D* and discussed throughout are based on a four-state model of speciation and a three-state model of extinction estimated over the interval 481–419 Ma, with pseudolevels starting at 481.000 Ma. The turnover time series derived from these two models is referred to as the default HMM turnover probability series. Sensitivity tests of different protocols for time series construction are discussed below.

Time Series Analysis

Time series analysis is conducted using the R package Astrochron (46). Our evaluation of band-limited variance in graptoloid HMM turnover probability focuses on candidate Milankovitch grand cycles (~1.2, ~2.4 My). Reliable assessment of astronomical rhythms in stratigraphic data is complicated by a number of issues (47), including inference of a sedimentation rate history to translate space (stratigraphic depth or height) to time (thousands of years) and the presence of stochastic variability in the Earth system, which may overprint the primary quasiperiodic forcing and/or yield stochastic oscillations that are mistaken for an astronomical signal (48, 49). We take a number of steps to address these issues in our analysis as discussed in detail below.

CONOP Time Model

The matter of timescale uncertainty can be especially problematic when testing for high-frequency Milankovitch forcing (quasiperiods spanning 10^4–10^5 y), such as that due to precession, obliquity, and short eccentricity, as these cycles often fall well below the resolution of available independent temporal constraints (e.g., radioisotopic ages) and require inferences that can risk circular reasoning (50–52). In the case of this study, the temporal resolution of the radioisotopic calibration and the long duration of the grand cycles that are under investigation (≥1 My) enable us to avoid astronomical tuning assumptions. The CONOP composite is calibrated using 23 radioiosotopic age data, with typical uncertainties of

±2 My (2σ, total uncertainty) (29). Importantly, however, the composite does not inherit the full magnitude of these errors, because it is calibrated using regression based on the entire ensemble of age data (29). As a result of this age calibration procedure, any graptoloid turnover quasiperiods that match the expected astronomical tempo arise independently of a priori assumptions about astronomical forcing. However, this also implies that, despite the high resolution of sampling associated with the graptoloid results (50,000 y), variable "sedimentation" (i.e., composite calibration inaccuracy) on the shorter term can complicate the assessment of higher-frequency (<0.5-My) Milankovitch cycles; we, therefore, avoid the evaluation of such short period cycles in this study.

Power Spectrum Background Estimation and Significance Testing

A common statistical approach for evaluating astronomical influence on sedimentation is the method of power spectral analysis, which quantifies data variance as a function of frequency (cycles per 1,000 y or cycles per meter) (53). Using this technique, discrete frequency bands that disproportionately contribute to data variance ("spectral power") can be identified and potentially linked to astronomical sources. Concerted effort over the later portion of the past century culminated in the development of reliable statistical approaches for accurate and precise estimates of the power spectrum as embodied in such techniques as the multitaper methods (MTMs) of refs. 36 and 54. While these landmark achievements in spectral analysis technique provide a valuable tool for cyclostratigraphy—allowing a reliable probabilistic assessment of the magnitude of band-limited variance within short and complex data series—they do not in themselves provide a test of the source of the observed cycles. This latter issue of attribution is often addressed by comparing the stratigraphic data spectrum with a "null" stochastic model of the physical system, from which the statistical significance of the peaks can be assessed. When the underlying timescale is highly uncertain, this assessment is commonly supplemented by other methods, such as the "frequency ratio method" or the evaluation of amplitude modulations (51, 55, 56; refs. 50 and 52 discuss astrochronologic testing methods that are independent of spectrum confidence levels).

It is commonplace in cyclostratigraphy to identify power spectrum peaks for which the null model can be rejected with a high degree of confidence (e.g., 95% confidence level or a P value of 0.05) and then to link the statistically significant cycles to astronomical periods. A wide range of null models has been used to approximate the spectral background in cyclostratigraphic studies (53), but the most common parameterizations are the autoregressive lag-1 (AR1) model (57) and the power law model (58), in which case the noise model parameters are estimated directly from the data. These stochastic models often provide a reasonable approximation to the spectral background and are appealing because they have a physical basis that is rooted in an understanding of modern climate and depositional systems (59–61).

While the theory underlying power spectrum significance testing is well-established, prior work has highlighted a number of essential challenges that can hinder its utility in cyclostratigraphy. Among these issues is the problem of accurate background estimation (48, 49, 53, 62–64). Although AR1 and power law stochastic models are often used in cyclostratigraphic assessments, there is no guarantee that they are always appropriate representations of the spectral background, especially when data filtering and/or detrending have been applied (48, 49, 64). Furthermore, if data series do, in fact, preserve substantial spectral power associated with astronomical cycles, these cyclic variations can serve to bias the estimates of the stochastic noise models, with consequences for the assigned statistical significance (48, 62). It should also be noted that noise parameters—such as the AR1 coefficient (ρ) and power law slope (β)—are of fundamental paleoclimatic interest in their own right, as they can yield insight into the underlying physics of climate and deposition (37, 61, 65, 66); thus, their accurate determination is paramount.

Another major issue is the problem of multiple testing: typical power spectral analyses involve hundreds to thousands of null hypothesis tests at different frequencies, which result in an excess occurrence of false positives beyond that expected from the assigned confidence levels (49, 67, 68). Whereas statistical procedures exist to account for this issue, the commonly used Bonferroni correction (49, 69) is known to be overly conservative (67, 70), making it exceedingly difficult to reject the null hypothesis when it is false (in statistical parlance, this is termed low "statistical power," which should not be confused with spectral power) (48). We nonetheless use the Bonferroni correction among other approaches, because we want to be as cautious as possible in ruling out false positives. Taken together, the problems of appropriate background estimation and multiple testing have created an apparent conundrum regarding the use of spectral analysis for astrochronologic testing (71), where it may seem that the choice is between two dueling philosophies: an overly pessimistic approach that is ill-suited to cyclostratigraphy or an overly optimistic approach that is prone to finding astronomical cycles even when they are not there. The problem is exacerbated by the fact that complex procedures are often utilized during data processing and analysis—with algorithms that are not always transparent (i.e., "open source") or appropriate—the statistical consequences of which are not necessarily obvious (48, 64).

In an effort to address these challenges, this study adopts a multifaceted strategy for cyclostratigraphic evaluation composed of three essential elements: (*i*) a complementary analysis of stratigraphic data and stochastic surrogates (via Monte Carlo simulation), which allows consideration of the suitability of particular noise models for background estimation given the specific data processing procedures used; (*ii*) the application of "multiple comparison" procedures that protect against inflated false-positive rates, while also being sensitive to the inherent limitations of cyclostratigraphic data; and (*iii*) the use of time–frequency approaches to evaluate the assumption of signal stationarity that underlies power spectrum estimation and significance testing.

To facilitate replication of our results and broader application of these methods, we provide a series of functions in the Astrochron software package for R (46) (version 0.8): testBackground, multiTest, confAdjust, pwrLawFit, mtmPL, pwrLaw, makeNoise. Details concerning these functions, including example applications, are outlined in the Astrochron package. We introduce the key functions in the following section, as they are applied to evaluate the graptoloid HMM turnover probability series.

Spectral Analysis of the Graptoloid HMM Results

The objective of our analysis is to test the graptoloid HMM turnover probability time series for significant band-limited variance at frequencies consistent with the predicted grand cycles, which in the Late Cenozoic, have periods of ∼1.2 and ∼2.4 My. As a first step, we evaluate the suitability of the spectral analysis procedures provided in Astrochron with a series of surrogate simulations that have the same sampling characteristics as the data (a time series duration of 61.96 My and a sampling interval 0.05 My) using both AR1 models and power law models. In all cases, the noise surrogates are generated using the same noise model parameters (ρ or β) as observed in the HMM turnover probability time series. The specific spectral methods examined include the conventional AR1 MTM approach (36, 57) ("MTM-AR1"), the robust red noise MTM approach (62) ("MTM-ML96"), the MTM-based LOWSPEC approach (48), power law assessments using the MTM spectrum ("MTM-PL"), power law assessments using a 25% cosine-tapered periodogram (72) ("Periodogram-PL"), and conventional AR1 assessments using a 25% cosine-tapered periodogram (57) ("Periodogram-AR1"). The MTM-ML96 approach is updated as in ref. 73 to remove

problematic median smoothing edge effects, as the original methodology (also used in the popular SSA-MTM Toolkit) has a high propensity to generate spurious significant cycles in the lower range of the spectrum that we are investigating (48). Estimation of power law-based confidence levels follows the bias correction approach of ref. 72. As is commonplace in cyclostratigraphy, a linear trend is removed from the time series before spectral analysis to decrease bias from trends/cycles that are longer than the length of the data series (53). The surrogate simulations allow us to assess any false-positive inflation associated with this procedure.

The simulations can be reproduced as shown in Dataset S3.

The results of the surrogate simulations are presented in Fig. S1 and Table S1, indicating correct reconstruction of false-positive rates for all methods with the exception of the MTM-ML96 approach (Fig. S1*D*), which exhibits elevated false-positive rates. It can similarly be shown that, when AR1-based null models are used to analyze the power law surrogates and vice versa, the false-positive rates deviate from the expected values. Two examples of this issue are illustrated in Fig. S1 *G* and *H*. These simulations can be reproduced as shown in Dataset S4.

Based on these surrogate simulations, we identify four background estimation approaches to assess for the graptoloid HMM turnover probability series. Two of these approaches yield correct false-positive rates for AR1 noise (LOWSPEC and Periodogram-AR1) (Table S1), and two yield correct false-positive rates for power law noise (MTM-PL and Periodogram-PL) (Table S1). LOWSPEC is selected in preference to the conventional MTM-AR1 approach, because it exhibits higher statistical power (table 3 in ref. 48). The periodogram-based approaches are selected due to their finer frequency resolution and the lower statistical dependence of power estimates from nearby frequencies, but these features come at the cost of greater bias and inconsistency (a consistent estimate is one in which the variance of the power spectrum—the "error bar" on the estimate—decreases to zero as the number of samples in the time series increases to infinity). Finally, these simulations do not guarantee that an AR1 or power law background is appropriate for any particular dataset, which should be judged independently. As will be shown below, both types of noise background are plausible for the HMM turnover probability series; thus, the final interpretations should be resilient to either null model.

Spectral analyses of the HMM turnover probability series can be reproduced as shown in Dataset S5.

To better ascertain the significance of these results, a multiple test correction is applied to the spectrum confidence levels using the Bonferroni method (Astrochron function confAdjust), with a special modification to make it more appropriate for cyclostratigraphy. The magnitude of the inflated false-positive rate problem is diminished in this study due to the use of the CONOP timescale and our focus on narrow frequency bands encompassing the present day grand cycles. Our study design obviates the need to blindly prospect for cycles across the entire spectrum, which would require a much larger adjustment to the confidence levels. Nonetheless, a multiple test correction is warranted to account for the bandwidth resolution of the spectral estimate and to allow for variable sedimentation due to composite calibration inaccuracy. It should be noted that variable sedimentation serves to diminish statistical significance by defocusing periodic cycles and also through the need to prospect across a band of frequencies, which necessitates a multiple test correction. We establish two distinct types of Bonferroni correction for our assessment, a narrow-band correction suitable for evaluating the predicted grand cycles and a global Bonferroni correction for other regions of the spectrum. For the narrow-band correction, we investigate frequencies within the bandwidth resolution of the 2π MTM spectrum, which includes all periods from 2.23 to 2.60 My and from 1.16 to 1.25 My. Although the periodogram has a finer bandwidth resolution, we apply the same frequency bands to allow for sedimentation rate instability. The Bonferroni multiple test correction can be reproduced as shown in Dataset S6.

This analysis (Fig. S2 and Table S2) provides a strong indication of a potential grand cycle at ~2.6 My that is distinguishable from the null models (LOWSPEC > 95% confidence level; Periodogram-AR1 > 95% confidence level; Periodogram-PL > 90% confidence level) and indicates a reasonable fit between the background spectra and the models. The reported confidence levels are anticipated to be underestimates, however, as the Bonferroni correction is known to be overly conservative (70). A number of improved approaches have been used extensively for multiple corrections (74), foremost of which is the false discovery rate (FDR) (75), which controls the expected proportion of falsely rejected hypotheses. Prior work documents the enhanced statistical power of the FDR compared with the Bonferroni-style procedures (75). We thus utilize the FDR as well as a number of other established multiple correction approaches to evaluate the significance of the grand cycles (Astrochron function multiTest) (Table S2). Before interpreting these results, however, it must be noted that red noise serves to reduce the statistical independence of power estimates at adjacent frequencies (76) and furthermore, that the MTM-based spectral power estimates are not independent within the investigated bandwidths surrounding each predicted grand cycle (36). Thus, it is necessary to conduct surrogate simulations to determine the expected false-positive rates for each multiple correction procedure given the analysis-specific parameterizations of the noise models. This analysis is conducted with the Astrochron function testBackground (see above and Datasets S3 and S4) using the option "multi=T." Results of these simulations indicate that, for the analysis of the entire default HMM turnover probability series (419–481 Ma), most of the multiple correction methods yield acceptable false-positive rates that can be interpreted at face value (green in Tables S1 and S2). Notable exceptions include (*i*) the FDR-BY method (false discovery rate approach of Benjamini and Yakutieli; explained in Tables S1 and S2), which is too conservative and thus, overestimates the *P* value (that is, it underestimates the confidence level) (gray in Tables S1 and S2), and (*ii*) most of the multiple correction approaches applied to the MTM-AR1 approach, which also are too conservative. Note that this assessment is separate from the issue of statistical power, which is generally expected to be greater for the non-Bonferroni methods, giving them a stronger ability to identify cycles when present.

Whereas there has been much effort in this study to establish accurate and meaningful *P* values that reflect their true statistical meaning, the *P* value threshold selected for designation of "significant" cycles in cyclostratigraphic research (e.g., a *P* value of 0.01, 0.05, 0.1) is ultimately a subjective one (67, 77). Recently, there has been much debate among academic statisticians about the utility of *P* values as a governing criterion for determining the significance of results and more generally, about how to improve reproducibility in science (78). Here, we adopt the philosophy that *P* values provide one useful measure of the magnitude of the statistical evidence for a hypothesis—which should be considered in combination with other statistical and nonstatistical criteria (e.g., geological and astronomical constraints)—and that the ability of other investigators to reproduce the results and evaluate other parameterizations is essential; the latter objective is facilitated by incorporation of these procedures in the software Astrochron.

The multiple test corrections can be reproduced as shown in Dataset S7. Note that the assessments of statistical significance conducted above have avoided the common practice of zero padding (53), as this approach serves to inflate false-positive rates associated with the nominal confidence-level estimates (72) and also implicates a greater number of frequencies to assess for the multiple test corrections. Having identified a significant cycle, however, we can now more accurately estimate its period using

zero padding to refine the frequency grid (77), indicating a duration of ∼2.6 My, as shown in Dataset S8.

Underlying these spectral analyses is the general assumption of signal stationarity, which is oftentimes a problem in stratigraphic assessments due to sedimentation rate variability, changes in the Earth system response to astronomical forcing (e.g., the Mid-Pleistocene transition) (79), and changes in the sensitivity of sedimentation to climate (e.g., due to basin evolution). For example, when a stratigraphic record is analyzed in its entirety, strong cyclic variability in one portion of the record may be obscured in the power spectrum due to its poor expression in another portion of the record, with the consequence being a low measure of statistical significance. Thus, while a significant 2.6-My signal has been identified in our spectral analyses, further assessment of its stratigraphic variability is essential as well as examination of the possibility that a strong 1.2-My signal emerges in portions of the time series. To evaluate the temporal evolution of the potential grand cycles, we use three methods: EPSA (37), EHA (38), and bandpass filtering (19).

EPSA and EHA utilize three 2π prolate tapers with a 20-My moving window; a linear trend is removed from each window before spectral analysis. For the sake of plotting, the EPSA and EHA results are normalized, such that the maximum power or amplitude in each 20-My window is unity. This approach facilitates a prompt assessment of changes in the relative strength of individual frequency components, particularly when there are large changes in the dynamic range of the spectra across the analyzed interval. Analysis of the graptoloid HMM turnover probability time series reveals dominant ∼2.6-My power for most of the study interval, which transitions to ∼1.3-My power in the older interval (Figs. 2*D* and 3 and Fig. S3). It is also notable that the EHA results—which allow a higher-frequency resolution than EPSA—show well-defined peaks. The EHA and EPSA results can be reproduced as shown in Dataset S9.

The EPSA and EHA results motivate an assessment of the statistical significance of the strong ∼1.3-My power observed in the older interval. We focus on the segment from 460 to 466 Ma, but first, a series of surrogate simulations (Table S1) is conducted to evaluate the suitability of the spectral methods given these data (similar to those outlined above but not reproduced here). The surrogate simulations support use of the same general protocols noted above, but caution is warranted with the LOWSPEC AR1-based approach, since it yields inflated false-positive rates given these parameterizations (∼2.1, 6.6, and 12.1% false-positive rates for the 99, 95, and 90% confidence levels). It should be noted that, with this shorter dataset, the LOWSPEC method is near its lower limit of applicability (at least 100 data points are required). Evaluation of the multiple test correction procedures for this shorter dataset indicates that most methods are either too conservative (gray in Table S1) or prone to high false-positive rates (red in Table S1). The green and gray *P* values in Tables S1 and S2 highlight methods that can be used in our assessment, with the caveat that the gray values are overestimates (the confidence level is underestimated by the method).

The analysis of the HMM turnover probability series between 460 and 466 Ma can be reproduced as shown in Dataset S10.

The results reveal a significant cycle of ∼1.2-My duration (Fig. S4 and Table S2). Using the EHA and EPSA results, which use a finer-frequency grid, it is apparent that the periodicity is ∼1.3 My.

Having identified statistically significant cycles at ∼2.6 and ∼1.3 My in the graptoloid default HMM turnover probability series, we now seek a more explicit quantification of the fraction of the variance attributable to each and their evolution through time. This is conducted by integration of the individual power spectra that make up the EPSA results (Fig. 3). For this assessment, evaluation of the fraction of variance associated with the ∼2.6-My cycle is determined by integrating from 0.21 to 0.52 cycles per 1 My, while evaluation of the ∼1.3-My cycle integrates from 0.53 to 0.90 cycles per 1 My. Finally, bandpass filtering utilizes the Taner method (19); note that, due to the smoothing inherent in the MTM power spectral analysis, it is necessary to use a larger bandwidth for power spectrum integration than is required in the bandpass filtering. The spectrum integration and bandpass filtering can be reproduced as shown in Dataset S11.

The combination of approaches applied here provides four complementary perspectives on the evolution of variance within frequency bands defined by the prospective grand cycles: normalized EPSA results, normalized EHA results, power spectrum integrations, and bandpass filter outputs. Note that the "fraction of variance" trends present in the integrated EPSA results need not always match the bandpass filter results, since the total variance (in each 20-My spectrum) changes throughout the interval.

We note that astronomical theory suggests that the ratio between the eccentricity and obliquity grand cycles should most likely be 2:1 or 1:1 and can flip back and forth between these ratios (80). Our findings of apparent 2.6- and 1.3-My grand cycles in the Paleozoic are consistent with this inference and with geological evidence available at this time.

The multifaceted approach that we have utilized to evaluate potential grand cycles in the HMM turnover probability series has been guided by an understanding of statistical theory ("statistical reasoning"), astronomical physics ("astronomical reasoning"), and the nature of the stratigraphic record ("stratigraphical reasoning"). Given the complexity of both the scientific problem and the analytical procedures, a wide range of parameterizations is possible. We encourage the community to further evaluate these conclusions, a process that will be facilitated by the Astrochron software and the functions that accompany this study (46).

Tests for Sensitivity and Bias in Spectral Analysis of Turnover

HMMs are somewhat sensitive to the low diversities and extreme speciation/extinction probabilities at each end of the time series and also to changes in the starting point of the time series. For these reasons and to test sensitivity of our results based on the default HMM turnover probability series, we modify the approach described above in two ways. First, we fit the HMMs using the well-constrained interval 475–425 Ma, which avoids the low diversities at each end of the graptoloid clade history, although ultimately, we predict the time series of states using these model parameters applied to the interval 481–419 Ma (see below). Second, we fit 1,000 three-state models with small offsets on the starting time of the pseudolevels and use a model-averaging approach for inference (cf. ref. 81). The starting offsets used are one-tenth of the pseudolevel spacing (i.e., offsets drawn from the list 0.005, 0.01 … 0.04, 0.045 My). During model averaging and for any given starting offset, we first identify the minimum Akaike's Information Criterion (AIC_{min}) and then ignore any models with $AIC_{model} - AIC_{min} > 2$ (i.e., $\Delta AIC > 2$). This ΔAIC threshold of two captures models that have "substantial support" (81) in the family of models for any given offset; here, we assume that these models are equally likely. From these 1,000 repeated models, we calculate median values of the estimated-state probabilities (output *pm* vectors) and input these values to a constrained HMM (explained below) to derive estimates of the transition probabilities (output *Pi*) and then the predicted, model averaged time series of states for the full interval 481–419 Ma (i.e., using *pn* for the full interval).

To run an HMM that is constrained to honor a supplied vector of median-state probabilities, *pm.med*, the R code given above is modified as shown in Dataset S12.

Importantly, we find that Milankovitch grand cycles emerge from the model average time series (Fig. S5*B*). The grand cycles are also expressed in the raw turnover time series based on raw speciation and extinction probabilities at each pseudolevel, although the signal is degraded somewhat by noise (Fig. S5*C* from the time series illustrated in Fig. 4*H*). Furthermore, the grand

cycles are also observed in the time series of HMM speciation and extinction states analyzed separately (Fig. S5 *D* and *E* from the time series illustrated in Fig. 4 *C* and *F*).

It is useful to evaluate uncertainties on the HMM parameter estimates. To do this, we use an approach based on simulation that captures both model-fitting and model selection uncertainties. Starting with observed parameter estimates for a particular HMM, a data time series is simulated to satisfy both the Markov property and the parameters of the observed model. The resulting simulated sequence of level by level extinction or speciation probabilities is the same length as the observed data. From this simulated dataset, the HMM parameters are reestimated using the Baum–Welch algorithm; differences between observed and reestimated parameters are attributed to model-fitting uncertainty (stochastic fluctuations of some workers). This procedure is repeated 1,000 times using the observed parameter estimates from the family of repeated, best-fit models described above, thus capturing uncertainty related both to model selection and to the existence of multiple equally likely models. Median absolute deviations derived from resulting distributions of reestimated parameters are used as overall uncertainty estimates. The resulting CIs are "unconditional confidence intervals" sensu (81), inasmuch as they reflect model selection uncertainty and are not conditional on a single model. Data simulation uses the functions in package HiddenMarkov in R, shown in Dataset S13, where *in.Pi*, *in.delta*, and *in.pm* are supplied parameters from a given HMM and *pn* is the observed number of extant species at each pseudolevel as above. *Seed* is simply a supplied seed for the random number generator; using a particular value of *seed* will allow exact reproduction of a given simulation (Dataset S13).

As an aside, we experimented with estimation of uncertainties using the bootstrap and jackknife, using both species and pseudolevels as units of resampling. Both methods, however, yield biased estimates of HMM parameters as indicated by the fact that resulting uncertainty bounds commonly do not bracket observed parameter estimates. We attribute this bias to serial dependence in the data that violates the assumption, inherent to both approaches, of invariance under permutation. This problem persists even when a stationary, block bootstrap design is used (82). In addition, using species as the unit of resampling—intuitively the most appealing unit—results in upward bias of HMM-state estimates because of the artificial clustering of first and last occurrences that is a consequence of repeated sampling of a particular species when resampling with replacement.

For the analyses described above, we have assumed time invariance of the HMMs—that it is appropriate to fit a single HMM across the entire time series. In general, it is desirable to fit a single model to the data, because this results in better statistical power and makes subsequent spectral analysis more straightforward. To test the assumption of time invariance, we use the median-state probabilities from the repeated models, described previously, to constrain 1,000 independent HMMs for each of the intervals 475–447 and 447–425 Ma and look for consistency in the resultant, freely varying transition probabilities. If the assumption of time invariance is justified, we would expect to see stability of transition probabilities from one interval to the next. As shown in Fig. S6, this assumption is apparently justified for both speciation and extinction probabilities; the only inconsistencies are for state 3 for both speciation and extinction in the "Ordovician," which is represented by just 4% of the time in both cases and is thus poorly constrained for that interval.

Finally, we wish to assess possible impacts of pseudospeciation or pseudoextinction on the expression of the grand cycles, where these phenomena result from within-lineage, gradual phyletic evolution from one species to another (discussion in ref. 17). To test this, we use a modification of the method of ref. 17 to identify hypothetical phyletic lineages, combining the ranges of candidate, congeneric species for which the last occurrence of the putative "ancestor" is within three CONOP levels of the first occurrence of the putative "descendant." During this process, zero-range species are retained given that they may contribute to a lineage (although any remaining zero-range species are subsequently eliminated before the fitting of HMMs). In cases where there are multiple potential descendants, the one with smallest gap or overlap with the ancestral range is selected; if there are two or more with equal gap or overlap, one is chosen at random. The algorithm works recursively until no further joins are possible within the target genus. This procedure reduces the number of species by 27% of the starting full dataset and is a highly conservative and stringent test of possible phyletic evolution within the graptoloids. The time series of HMM-derived turnover probability derived from this modified dataset reveals Milankovitch grand cycles (Fig. S5*F*). The result shown is based on three-state models for both speciation and extinction estimated over the interval 481–419 Ma with pseudolevels starting at 481.000 Ma.

Data Availability

Data on which this study is based are available in Dataset S1. CONOP 9 is available at www.geobiodiversity.com/Download.aspx. Program manuals and other programs in the CONOP "family" are available from ref. 83.

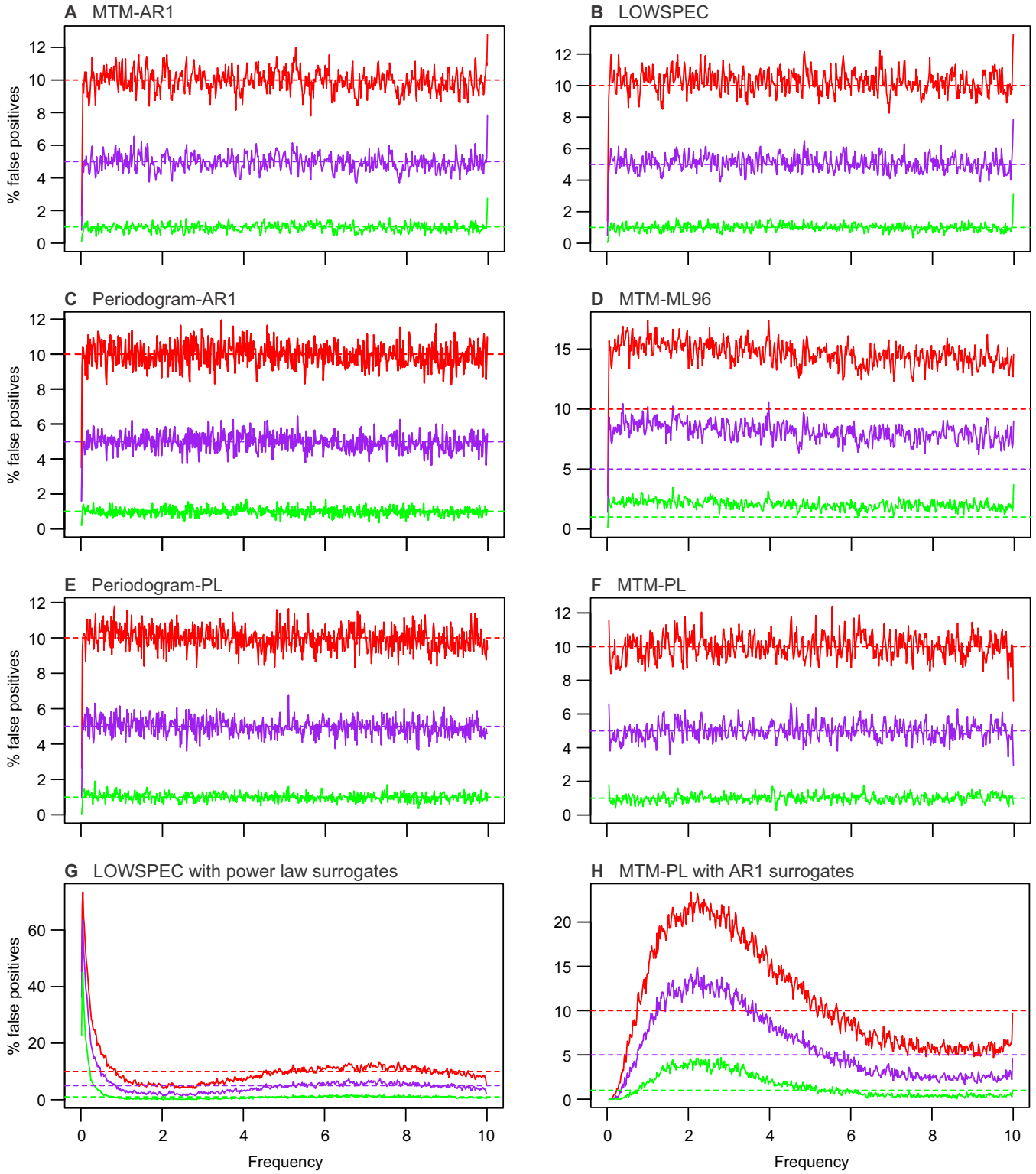

Fig. S1. An evaluation of false-positive rates associated with six spectral analysis procedures provided in Astrochron using surrogate simulations that have the same sampling characteristics as the data (a time series duration of 61.96 My and a sampling interval 0.05 My). False-positive rates shown for the 90% (red), 95% (purple), and 99% (green) confidence levels. The noise surrogates are generated using the same noise model parameters ($\rho = 0.3082017$ or $\beta = 0.3938128$) as observed in the HMM turnover probability time series. These plots illustrate results from the 2,000 simulations reported in Table S1. (*A*) Conventional AR1 MTM approach (function mtm) analysis of AR1 surrogates. (*B*) LOWSPEC (48) (function lowspec) analysis of AR1 surrogates. (*C*) Conventional AR1 approach with 25% cosine-tapered periodogram (function periodogram) analysis of AR1 surrogates. (*D*) Mann and Lees (62) robust red noise MTM (function mtmML96) analysis of AR1 surrogates. (*E*) Power law fit to a 25% cosine-tapered periodogram (function periodogram) analysis of power law surrogates. (*F*) Power law fit to an MTM spectrum (function mtmPL) analysis of power law surrogates. (*G*) LOWSPEC analysis of power law surrogates. (*H*) Power law fit to an MTM spectrum analysis of AR1 surrogates. All MTM-based methods use three 2π prolate tapers.

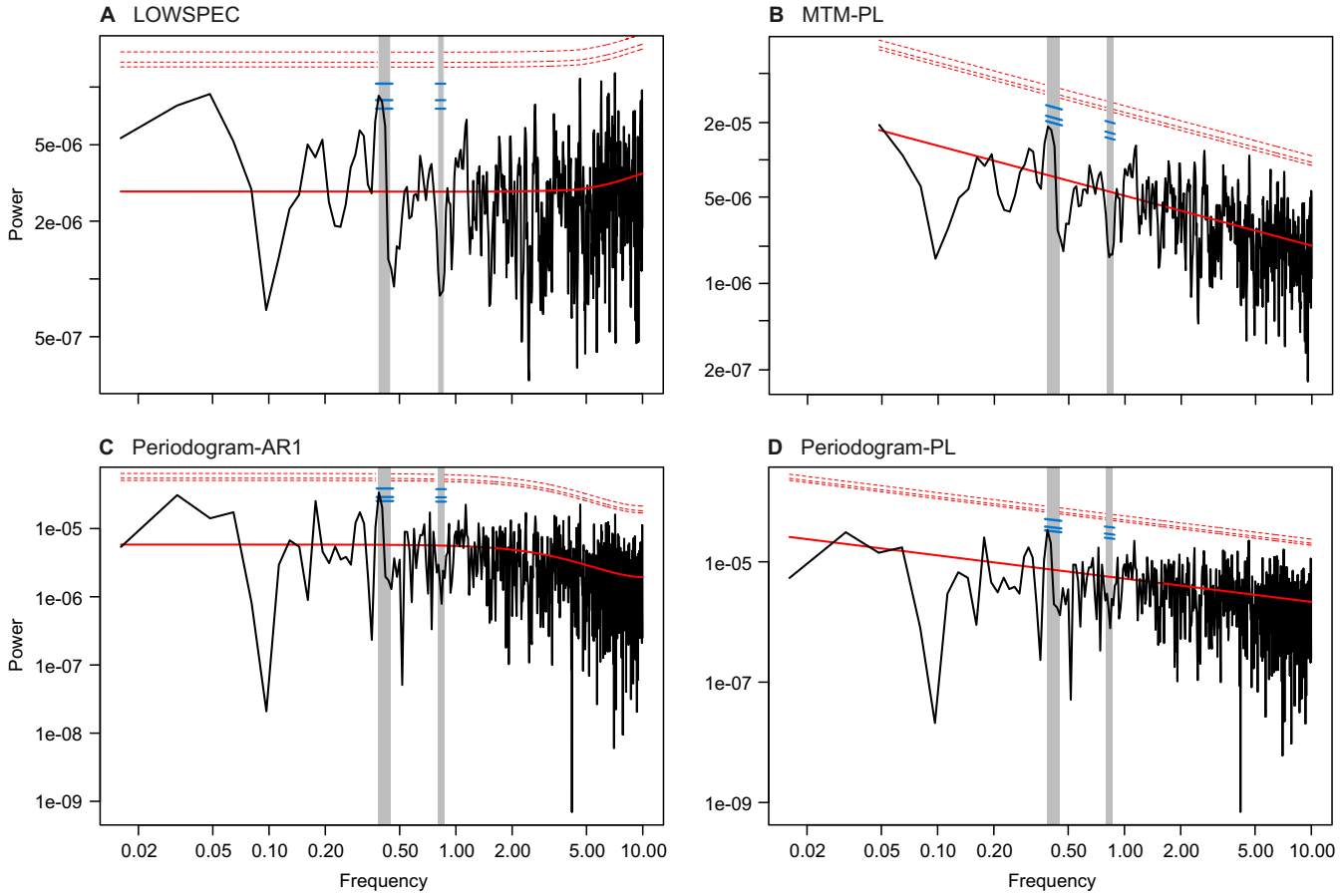

Fig. S2. Power spectral analysis and background estimation for the graptoloid default HMM turnover probability series (419–481 Ma) using four different approaches. All plots are log frequency vs. log power. Gray boxes identify the bands investigated for potential ∼1.2- and ∼2.4-My grand cycles; the solid red lines are the estimated spectral background; the dotted red lines are the Bonferroni-corrected 90, 95, and 99% global confidence levels; and the solid blue lines are the Bonferroni-corrected 90, 95, and 99% confidence levels for the target frequency bands (for clarity, these lines have been extended beyond the gray boxes). (*A*) LOWSPEC. (*B*) Power law fit to an MTM spectrum. (*C*) Conventional AR1 approach with 25% cosine-tapered periodogram. (*D*) Power law fit to a 25% cosine-tapered periodogram. MTM-based methods use three 2π prolate tapers. Additional discussion is in the text.

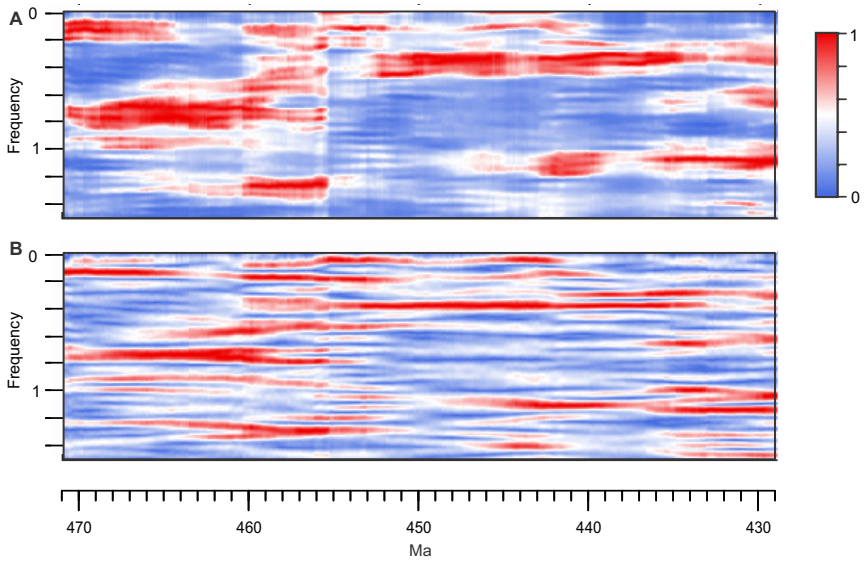

Fig. S3. EPSA and EHA of the default HMM species turnover probability series (compare with Fig. 2*D*) using three 2π prolate tapers and a 20-My moving window. The maximum power or amplitude in each window is normalized to unity. (*A*) Normalized EPSA power results. (*B*) Normalized EHA amplitude results.

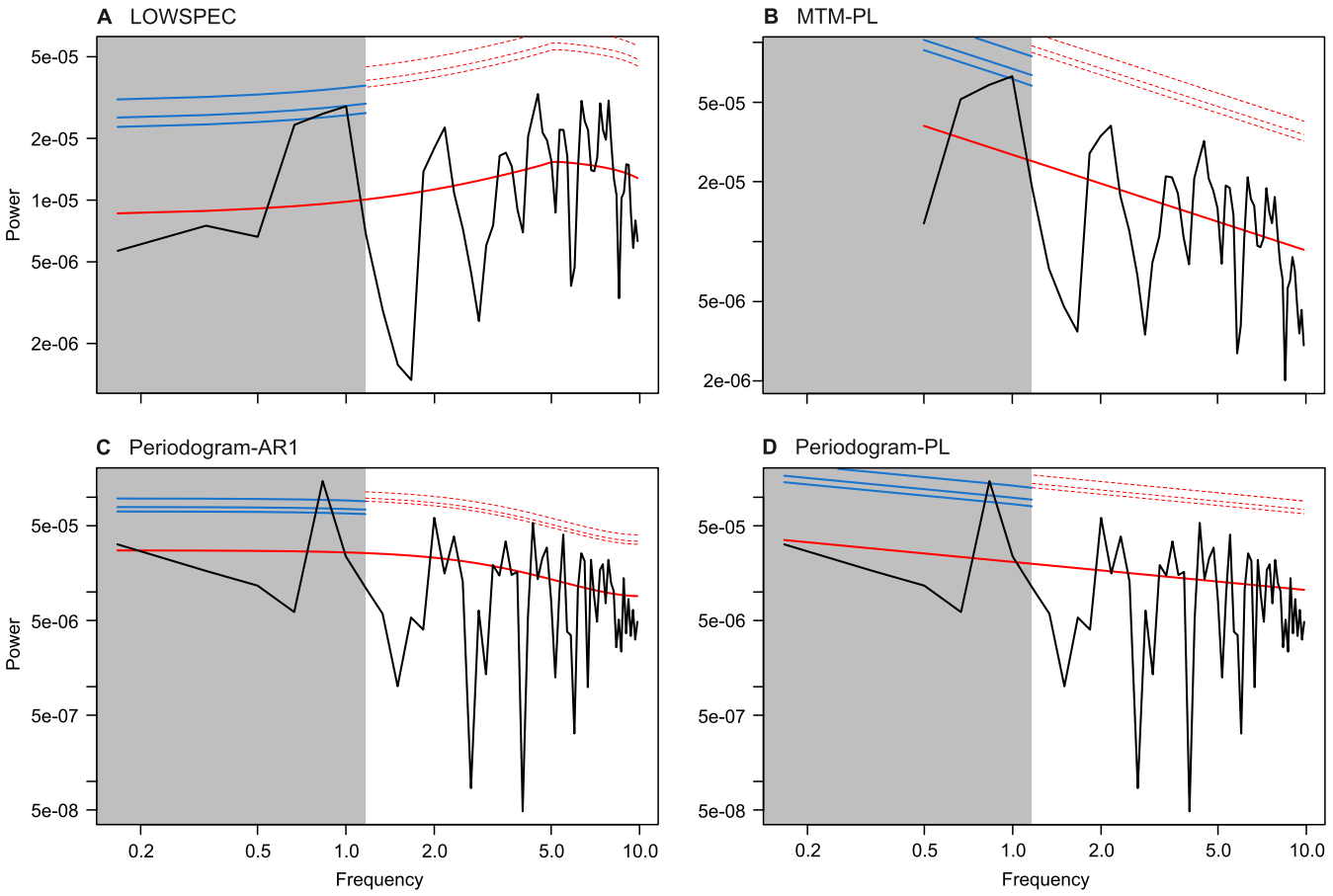

Fig. S4. Power spectral analysis and background estimation for the graptoloid default HMM species turnover probability series spanning 460–466 Ma using four different approaches. All plots are log frequency vs. log power. Gray boxes identify the bands investigated for potential ~1.2- and ~2.4-My grand cycles; the solid red lines are the estimated spectral background; the dotted red lines are the Bonferroni-corrected 90, 95, and 99% global confidence levels; and the solid blue lines are the Bonferroni-corrected 90, 95, and 99% confidence levels for the target frequency bands. (*A*) LOWSPEC. (*B*) Power law fit to an MTM spectrum. (*C*) Conventional AR1 approach with 25% cosine-tapered periodogram. (*D*) Power law fit to a 25% cosine-tapered periodogram. MTM-based methods use three 2π prolate tapers. *Supporting Information* has further discussion.

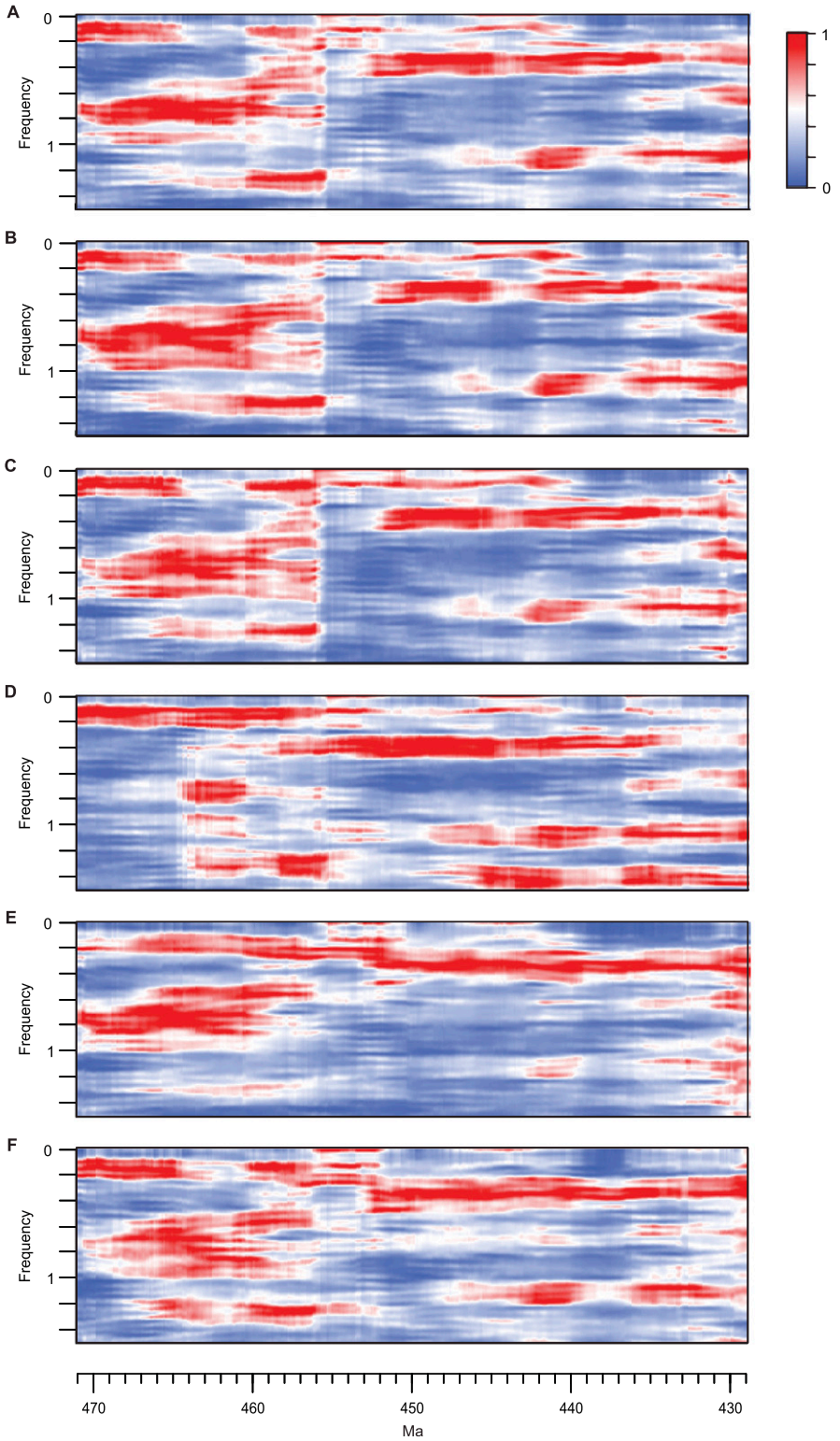

Fig. S5. EPSA for different time series of species turnover, speciation, and extinction to test for sensitivity of results to time series construction. All EPSAs utilize three 2π prolate tapers, with a 20-My moving window; the maximum power in each window is normalized to unity. *Supporting Information* has a detailed explanation of each time series. (*A*) Default HMM species turnover probability series based on single HMMs for speciation and extinction; this is the key result shown in Fig. 2*D*. (*B*) Turnover based on model-averaged, three-state HMMs for speciation and extinction. (*C*) Turnover based on raw speciation and extinction probabilities at each pseudolevel for the same pseudolevels as used in *A*. (*D*) HMM speciation. This is the model used in the construction of the default HMM turnover probability series of our key result. (*E*) HMM extinction. This is the model used in the construction of the default HMM turnover probability series of our key result. (*F*) Turnover based on single three-state HMMs for speciation and extinction derived from a dataset in which the effects of speculative, widespread, gradual phyletic evolution have been removed.

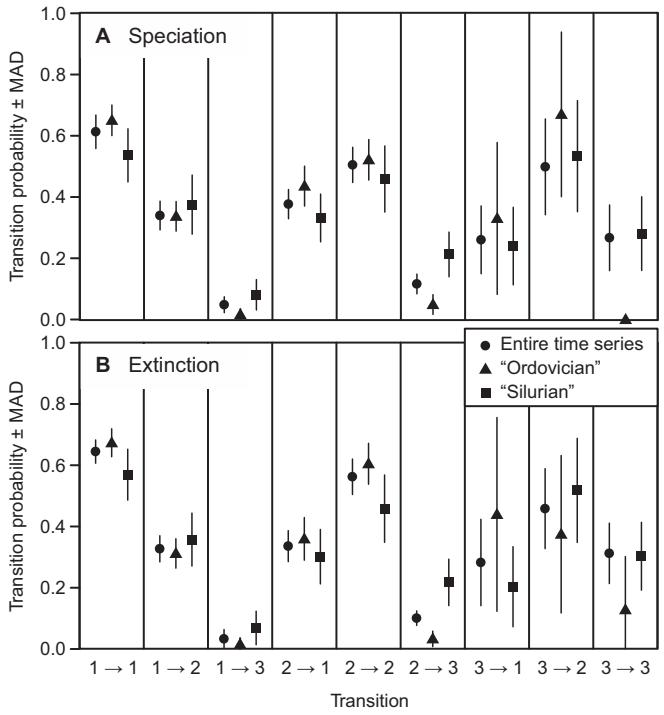

Fig. S6. Median transition probabilities for three-state HMMs fitted to the entire well-sampled part of the time series (475–425 Ma) compared with the subintervals from 475 to 447 Ma ("Ordovician") and from 447 to 425 Ma ("Silurian"). The consistency in transition probability estimates between the subintervals and entire time series suggests that it is appropriate to fit a single HMM to the entire time series. (*A*) Speciation. (*B*) Extinction. Each median is based on 1,000 replicates. For the subintervals, HMMs were constrained to honor the median-state values identified for the entire time series, but transition probabilities were free to vary. Unconditional CIs are median absolute deviations (MADs) derived by simulation. *Supporting Information* has additional explanation.

Table S1. False-positive rates for six spectral background estimation approaches to evaluate their suitability for analysis of the entire HMM turnover probability series (419–481 Ma) and the 460- to 466-Ma interval

Confidence level	Complete record: 419–481 Ma						460- to 466-Ma interval					
	Periodogram-AR1	MTM-AR1	MTM-ML96	LOWSPEC	MTM-PL	Periodogram-PL	Periodogram-AR1	MTM-AR1	MTM-ML96	LOWSPEC	MTM-PL	Periodogram-PL
90% CL*	10.02	10.02	14.54	10.18	10.05	10.02	10.17	8.47	11.86	11.86	8.77	10.17
95% CL*	5.01	4.85	7.92	5.01	4.86	5.01	5.08	3.39	6.78	6.78	3.51	5.08
99% CL*	0.97	0.97	1.94	0.97	0.97	0.97	0.00†	0.00†	0.00†	0.00†	0.00†	0.00†
Multiple test corrections												
90% CL FDR-BH‡,§	10.90	11.25	104.7	17.80	17.70	14.20	7.45	7.35	47.45	80.00	13.30	21.10
95% CL FDR-BH‡,§	4.90	4.50	43.15	9.30	9.30	6.60	3.35	2.40	23.85	46.40	5.00	10.45
99% CL FDR-BH‡,§	0.80	0.85	9.5	2.00	1.30	1.15	0.35	0.35	7.5	16.60	1.05	2.40
90% CL FDR-BY‡,¶	1.10	1.25	13.25	2.40	1.95	1.80	1.15	0.70	13.75	26.50	2.30	4.00
95% CL FDR-BY‡,¶	0.60	0.50	7.95	1.55	0.65	0.70	0.45	0.40	7.55	17.05	1.05	2.45
99% CL FDR-BY‡,¶	0.15	0.00	1.6	0.10	0.05	0.25	0.00	0.05	2.8	6.85	0.30	0.90
90% CL Hommel‡,#	9.00	6.10	40.2	11.10	11.05	11.60	6.70	3.70	25.05	42.10	6.55	14.45
95% CL Hommel‡,#	4.30	3.05	22.75	5.70	5.25	5.40	3.15	1.15	16.9	28.60	3.65	8.40
99% CL Hommel‡,#	0.80	0.55	6.95	1.50	0.65	1.10	0.35	0.30	5.45	11.40	0.70	2.20
90% CL Hochberg‡,‖	9.00	6.10	40.2	11.10	11.05	11.60	6.70	3.70	24.75	41.75	6.50	14.40
95% CL Hochberg‡,‖	4.30	3.05	22.75	5.70	5.25	5.40	3.10	1.15	16.9	28.40	3.65	8.40
99% CL Hochberg‡,‖	0.80	0.55	6.95	1.50	0.65	1.10	0.35	0.30	5.45	11.40	0.70	2.20
90% CL Holm‡,**	9.00	6.10	40.2	11.10	11.05	11.60	6.70	3.70	24.75	41.75	6.50	14.40
95% CL Holm‡,**	4.30	3.05	22.75	5.70	5.25	5.40	3.10	1.15	16.9	28.40	3.65	8.40
99% CL Holm‡,**	0.80	0.55	6.95	1.50	0.65	1.10	0.35	0.30	5.45	11.40	0.70	2.20
90% CL Bonferroni‡	9.00	6.10	40.15	11.10	11.05	11.60	6.70	3.70	24.6	41.35	6.50	14.35
95% CL Bonferroni‡	4.30	3.05	22.75	5.60	5.25	5.40	3.10	1.15	16.6	28.20	3.60	8.40
99% CL Bonferroni‡	0.80	0.55	6.9	1.50	0.65	1.10	0.35	0.30	5.45	11.30	0.70	2.20

The simulations utilize the same parameterizations as the HMM data and use 2,000 spectra. Green highlights methods that yield the correct false-positive rate, gray identifies methods that are overly conservative, and red indicates methods that have elevated false-positive rates. Methods highlighted in red should be avoided. CL, confidence level; FDR-BH, false discovery rate of ref. 75; FDR-BY, false discovery rate of ref. 84.

*Median percentage of false-positive frequencies per spectrum.

†The expected median percentage of false-positive frequencies per spectrum is zero at this CL, as only 58 frequencies are investigated per spectrum.

‡Total percentage of false-positive frequencies per 2,000 spectra.

§Benjamini and Hochberg (75).

¶Benjamini and Yekutieli (84).

#Hommel (85).

‖Hochberg (86).

**Holm (87).

Table S2. Corrected *P* values for the ∼2.6-My cycle (419- to 481-Ma interval) and ∼1.3-Ma cycle (460- to 466-Ma interval) associated with spectral analysis of the HMM turnover probability series

Background estimation	Frequency (cycles per 1 My)	Period (My)	Uncorrected *P* value	FDR-BH*	FDR-BY†	Hommel‡	Hochberg§	Holm¶	Bonferroni
419–481 Ma									
LOWSPEC	0.387	2.583	0.004	***0.027***	***0.075***	***0.030***	***0.034***	***0.034***	***0.034***
Periodogram-AR1	0.387	2.583	0.003	***0.024***	***0.064***	***0.024***	***0.024***	***0.024***	***0.024***
MTM-PL	0.387	2.583	0.022	0.112	0.303	0.151	0.173	0.173	0.173
Periodogram-PL	0.387	2.583	0.012	***0.094***	0.254	***0.094***	***0.094***	***0.094***	***0.094***
460–466 Ma									
LOWSPEC	1.000	1.000	0.008	***0.039***	0.102	***0.045***	***0.053***	***0.053***	***0.053***
Periodogram-AR1	0.833	1.200	0.004	***0.023***	***0.056***	***0.023***	***0.023***	***0.023***	***0.023***
MTM-PL	1.000	1.000	0.021	***0.086***	0.178	***0.086***	***0.086***	***0.086***	***0.086***
Periodogram-PL	0.833	1.200	0.001	***0.007***	***0.017***	***0.007***	***0.007***	***0.007***	***0.007***

Numbers in bold italics are *P* values < 0.1, representing the 90% confidence level. Green highlights methods for which *P* values are reported correctly, gray identifies *P* values that are overestimates (the true *P* value is less), and red indicates *P* values that are underreported (the true *P* value is greater) (Table S1). Red *P* values should be avoided.

*Benjamini and Hochberg (75).

†Benjamini and Yekutieli (84).

‡Hommel (85).

§Hochberg (86).

¶Holm (87).

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (DOCX)
Dataset S3 (DOCX)
Dataset S4 (DOCX)
Dataset S5 (DOCX)
Dataset S6 (DOCX)
Dataset S7 (DOCX)
Dataset S8 (DOCX)
Dataset S9 (DOCX)
Dataset S10 (DOCX)
Dataset S11 (DOCX)
Dataset S12 (DOCX)
Dataset S13 (DOCX)