Appendix S1: Supplementary materials

Data collection

Endocranial volume (ECV) data were taken from the literature for hominin taxa starting from *Australopithecus* to those specimens whose age ranges overlap with 0.5 Ma (table S1). We used primarily measurements from sources where it was directly stated or we could only conclude that the cranial specimens were measured directly by the author(s). Estimates from secondary sources were excluded unless they represented compilations of measurements that were originally taken by the authors [1,2]. If the same author published different ECV estimates for a given specimen, we included only their most recent estimate. The application of these criteria resulted in a dataset of 191 hominin ECV measurements taken from 94 specimens (median: 2 measurements per specimen; fig. S3a) that have been attributed to 13 different species (using the more speciose taxonomy) (table S1). ECV measurements were log10 transformed prior to all analyses because we were interested in proportional differences (e.g., a doubling from 400 to 800 cm3 was considered more meaningful than an equivalent 400 cm3 increase from 1000 to 1400 cm3).

Age data for the fossil specimens included in this study were also collected from the primary literature. To be considered for this study, the published chronometric date had to satisfy these criteria:

1. The date was published in a primary source such as a peer-reviewed journal article, a peer-reviewed abstract, or an edited volume. A primary source as used here means that the listed authors include the individual(s) who actually determined the chronometric age. Secondary sources were acceptable only if the authors were the same individuals who had originally determined the dates listed.
2. The date was published in approximately the last 30 years. New dating techniques and improvements in existing methods generally render older dates obsolete. However, if no date from more recently than 30 years ago was available for a specimen, then the most recent date, regardless of when published, was used.
3. Each date included in this study must not have been superseded by a more recently published date from the same author or authors. Superseding by the same author may be active or passive; for example, the publication of an explicit statement clarifying that a previously published chronology is no longer valid or the publication of a new date for the same specimen or geologic formation that differs from an older published date, respectively.
4. In the case of controversial chronologies for specimens or geologic formations where multiple authors working independently published contradictory dates, all dates fulfilling the previous three criteria were considered equally valid. As such, the temporal extremes of all valid chronologies were considered the maximum and minimum dates for the specimen. For example, if Specimen X had been dated according to Chronology 1: 0.78 – 0.63 Ma and Chronology 2: 0.80 – 0.69 Ma, then the date range used for Specimen X in this study was 0.80 – 0.63 Ma.

We categorized age uncertainty in two ways: 1) error associated with the analytical technique used to date a particular specimen and 2) error due to stratigraphic uncertainty. The first kind is introduced when specimens are dated by a single age estimate on the geological bed within which the specimen was discovered. For this type of error, we recorded the mean age estimate along with the standard error associated with the dating method. The second kind, error due to stratigraphic uncertainty, is introduced when the age of a specimen is determined by dating geological horizons that bracket the specimen in question. In this case, we recorded the maximum and minimum date out of all possible candidate dates and disregarded the standard error associated with each date (e.g., for a specimen dated using two bracketing geological horizons with dates of 0.9 + 0.02 Ma and 0.8 + 0.02 Ma, we used the age range of 0.9 – 0.8 Ma). This was because of the difficulty in combining different forms of error into a single probability distribution (see “Quantifying ECV and dating error”), and in almost all cases, the error associated with stratigraphic uncertainty far exceeded analytical dating error.

Estimating clade-level evolutionary mode

We explored six evolutionary modes for the clade-level ECV time series: random walk, gradualism, stasis, punctuated equilibrium with two stasis segments, stasis-random walk, and stasis-gradualism [3–5]. The last two modes have not traditionally been tested in the hominin ECV literature, but visual inspection of the time series suggested that these could provide strong fits to the data. The punctuated equilibrium model, in its original formulation, requires lineage splitting [6,7], as opposed to punctuated change within a lineage (which has also been observed in the fossil record [8]). We do not require an association between punctuations and lineage-splitting events in our analyses, so our “punctuation equilibrium” model is merely descriptive and is probably more appropriately named “punctuated clade-level change.” A seventh mode, punctuated equilibrium with three stasis segments, was also tested but received poor model support, so it will not be considered further. All of the evolutionary modes were operationalized as statistical models in R 3.0.3 [9] using the “paleoTS” package [10]. Random walk and gradualism were modeled as having ECV transitions drawn from a normal distribution (i.e., the “step distribution”) where the mean determines directionality and variance determines volatility of the trend [3]. Both models have trait mean and step variance parameters, but the random walk model has a fixed step mean of zero (i.e., ECV increases and decreases are equally likely at each time step) whereas the gradualism model’s step mean is an estimated non-zero parameter (i.e., there is a net positive or negative trend) [3]. Stasis is modeled as normally distributed variation around an optimal ECV, both of which do not change through time [3,11]. Punctuated equilibrium, stasis-random walk, and stasis-gradualism are all permutations of the three simpler models separated by one punctuation, the location of which is estimated as a free parameter [4,5].

Model parameters were estimated via maximum likelihood by using the “Joint” method in *paleoTS*. This involved model parameters being estimated by considering all binned ECV values (see “Examining clade-level evolutionary models while incorporating error”below) within a sequence jointly as a single draw from a multivariate normal distribution. This approach more accurately detects noisy trends than one analyzing differences across adjacent time bins (i.e., first differencing) [12]. For model selection, we used the small-sample, bias-corrected Akaike Information Criterion (AICc) to assess model fit to the data, penalizing for the number of parameters in the model to prevent overfitting [13]. AICc scores for each evolutionary mode were transformed into weights, which range from zero to one (with higher values representing better model support) and sum to one across all modes, to ease interpretability.

Quantifying ECV and dating error

Fossil crania are almost never preserved in a complete, unaltered state, and different researchers have used different techniques to reconstruct and measure the ECV of fragmentary or partial crania. This introduces noise and even possible bias into the measurement process. To estimate this within-specimen, inter-observer error (i.e., variation in different published measurements for the same specimen), we fit a random effects (Type II) ANOVA using the lme() function in the “nlme” R package [14] with log10 ECV as the dependent variable in an intercept-only model, while specimen ID was nested within species as random effects. The varcomp() function from the “ape” R package [15] was used to extract variance components from this model, and the within-specimen ID variance component was treated as inter-observer error. This was done using both a less speciose and a more speciose taxonomy (table S1), which gave variance components of 7.215 x 10-4 (log10cm3)2 and 7.230 x 10-4 (log10cm3)2, respectively (table S2). Though both variance components are very similar, we used the one from the more speciose taxonomy since it was slightly larger and therefore more conservative.

To incorporate dating error into our analyses, we used a Monte Carlo framework where our analyses were iterated 1000 times to explore how variation in dating uncertainty affected the robustness of our results. In each iteration, we resampled each specimen age depending on whether the error was associated with the analytical dating technique or stratigraphic uncertainty (see “Data collection”). For the former, we resampled age estimates from a normal distribution with the mean age estimate as the mean and its standard error as the standard deviation. For error due to stratigraphic uncertainty, we resampled age estimates from a uniform distribution set between the maximum and minimum recorded age estimates. Our choice of the uniform distribution is conservative because it makes no assumptions about sedimentation rates or stratigraphic placement of the fossil specimen.

Examining clade-level evolutionary models while incorporating error

Because each specimen had between one and six published ECV estimates, we used the median estimate to characterize each specimen’s ECV (results were the same whether we used median or mean ECV; r2 = 0.999 between the two measures). Age estimates for each specimen were resampled according to their type of dating error (see “Quantifying ECV and dating error”), and specimens were then placed in 0.2 Ma bins with bin edges at 3.6 Ma, 3.4 Ma, 3.2 Ma, etc. The mean and variance of ECV medians were then calculated for each time bin. To characterize total error associated with estimating mean ECV across specimens within a bin, the variance estimate in each bin was divided by its respective number of specimens to get sampling error [3]. This was then added to the within-specimen, inter-observer error estimate calculated from the random effects ANOVA model (see “Quantifying ECV and dating error”). A Bartlett’s test confirmed that variances across all time bins were not significantly different, so the variances were pooled over all bins, increasing the precision with which population variance was estimated [3]. We used this process to generate the binned time series to which each evolutionary mode model was fit for a single iteration in our analyses (e.g., fig. S1). Model selection produced six AICc weight estimates (one for each model), and the process was iterated 1000 times to produce a distribution of 1000 AICc weight values for each evolutionary mode (see fig. S2 for workflow schematic).

R2 was calculated for the best fit model (gradualism in our case) because AICc scores and AICc weights only provide a measure of *relative* goodness of fit, whereas R2 gives a measure of *absolute* fit. This was done by first estimating the parameters of the gradualism model, which can be viewed as a linear model. The estimated trait mean at the start of the sequence is treated as the intercept, and the mean of the step distribution is treated as the slope. Dates for each ECV specimen are then input into the model to get predicted ECVs. R2 is then calculated as: . Dating error was incorporated by resampling dates of each ECV specimen according to the protocol outlined in “Quantifying ECV and dating error” and then repeating the R2 calculation 1000 times. This gave a distribution 1000 R2 values, the mean of which is the mean estimate, and the standard deviation is the standard error.

Lower taxonomic level additive partitioning analyses

In order to conduct the partitioning analyses, we needed to assign lineage identities to each species. We operationally define lineage as an evolutionary sequence of populations that temporally succeed each other and therefore may consist of more than one species (in an ancestor-descendant, anagenetic sequence). Because of the debate surrounding the alpha taxonomy and phylogenetic relationships of certain specimens, we adopted two extreme taxonomic schemes that occupy the endpoints of a continuum (i.e., less speciose vs. more speciose; fig. S4). If additive partitioning results are consistent between these two schemes, it can be assumed that the pattern is robust and not influenced by taxonomic philosophy. It should be noted that this is not the same taxonomic dichotomy that was used for the random effects ANOVA model (table S1 & S2). This was because the ANOVA analysis was only concerned with species designations, whereas the additive partitioning analysis is concerned with ancestor-descendant relationships as well.

For the less speciose taxonomy, we used three lineages: eastern African *Australopithecus-Homo*, eastern African *Paranthropus*, and South African *Australopithecus-Paranthropus* (fig. S4). This taxonomy is conservative even for taxonomic philosophies that are considered to lie squarely in the “lumping” category [16]. We put *Australopithecus afarensis* specimens in the lineage leading to *Homo* instead of *Paranthropus* since it conservatively places major ECV increases within a lineage as opposed to during lineage splitting events (which is what the more speciose taxonomy is biased towards). For the more speciose taxonomy, we assigned species to lineages on the basis of morphological similarities *and* spatio-temporal coherence (similar to the paleo-deme paradigm [17], though we are interested in a higher taxonomic level: the species). Thus, OH 9, which shows affinities with Asian *Homo erectus*, is placed into *Homo ergaster* due to its location in eastern Africa. No speciation or major morphological change is assumed to happen within lineages except for the eastern African *Paranthropus* lineage. *Paranthropus aethiopicus* and *Paranthropus boisei* were included in the same anagenetic lineage because even the most ardent splitters acknowledge the plausibility of an ancestor-descendant relationship between these two species [18]. This resulted in a total of eight lineages: *Au. afarensis*, *Australopithecus africanus*, *P. aethiopicus-P. boisei*, *Paranthropus robustus*, early *Homo* (which includes *Homo habilis sensu stricto* and *Homo rudolfensis*), *H. ergaster*, *H. erectus sensu stricto*, and *Homo heidelbergensis* (fig. S4). Details for which specimen was assigned to which lineage in both taxonomic schemes can be seen in the actual dataset.

Previous studies have shown that clade-level changes in morphology (i.e., differences between adjacent bins) can be additively partitioned into anagenetic, origination, and extinction components that together account for the entirety of change observed at the broader clade level [19,20]. Origination is defined as a branching event after which the parent and daughter lineages coexist, and extinction is defined as the termination of a lineage. To do this, we first created 0.3 Ma bins with boundaries at 3.35, 3.05, 2.75, etc. until 0.95 Ma. This binning scheme was a compromise between getting the smallest bin size possible while eliminating any gaps within lineages. To best illustrate the partitioning method [19], we show here how the anagenetic, origination, and extinction components are calculated between two time bins as an example, though these partitions are determined for all between-bin differences within a given time series. The anagenetic component is calculated by first noting which lineages survive between the two bins in question. Let us name all specimen ECVs belonging to these lineages in the earlier time bin as “survivor-earlier” and those belonging to the later time bin as “survivor-later.” The anagenetic component is simply calculated as mean “survivor-later” minus mean “survivor-earlier.” The origination component is determined by subtracting mean “survivor-later” from the mean of *all* ECVs in the later time period. The extinction component is calculated as mean “survivor-earlier” minus the mean of *all* ECVs in the earlier time period. We should note that origination and extinction in this case refer respectively to observed first and last appearances of species’ ECVs and may not represent a species’ true age of origination and extinction [21]. Therefore, we use the phrases “first appearances” and “last appearances” to refer to clade-level change that is driven by the observed appearances and disappearances of ECVs within a given lineage. Clade-level changes between time bins with no through-ranging lineages cannot be partitioned into separate first or last appearances components and are labeled as “First/last appearances” in our figures. The sum of the anagenetic and first and last appearances components between any two adjacent bins is equal to the mean clade-level change between those same bins. This entire procedure was repeated twice using the less speciose and then the more speciose taxonomy. We suggest that patterns seen in both taxonomic schemes reflect signals that are robust to taxonomic designation.

Sensitivity analyses

To test how robust our clade-level analyses were, we conducted a series of sensitivity analyses on our clade-level evolutionary modes analysis. We first tested if the location of bin edges influenced our results by running the same analysis as above but repeating it multiple times by shifting our bin edges +0 Ma, +0.02 Ma, +0.04 Ma, +0.06 Ma, and so on all the way up to +0.18 Ma (fig. S5a). We tested for the influence of bin size by re-running our analyses using bin sizes of 0.10 Ma, 0.15 Ma, 0.20 Ma, 0.25 Ma, and so on all the way up to 0.50 Ma (fig. S5b).

For the lower taxonomic level additive partitioning analyses, we first explored the effect of dating error on our results. We used ECV medians calculated across inter-observer estimates for each specimen (i.e., ignored inter-observer error), randomly resampled age estimates according to the protocol laid out in “Quantifying ECV and dating error,” calculated the additive partitions, and repeated this 1000 times (fig. 4). We then examined the influence of inter-observer ECV error on our results. Instead of using the variance component method of the clade-level analyses, we randomly sampled one published measurement for each specimen, used midpoints for the specimen ages (i.e., ignored dating error), conducted the additive partitioning analysis, and repeated this 1000 times to produce error bars on the partition estimates (fig. S6). We also tested for the effect of bin location on our results by calculating the additive partitions using median ECV and age midpoints but iteratively shifting the bin boundaries used in the main analyses by +0.05 Ma (fig. S8). Finally, we analyzed the effect of bin size on our results by again calculating the partitions using median ECV and age midpoints but using bin sizes of 0.2 Ma, 0.3 Ma, and 0.4 Ma (fig. S9). Bin edges from the bin size sensitivity analysis started from 0.05 Ma and increased by whichever size was being tested. All of these sensitivity analyses were conducted twice, once using a less speciose taxonomy and then again using a more speciose taxonomy.   
  
Comparing within-lineage ECV change to published generational rates of change  
 We used a simple exponential model from Hansen and colleagues [22,23] to investigate whether published estimates of the rate and magnitude of natural selection on generational time scales can be extrapolated to predict observed within-lineage ECV changes on geological time scales. If so, this would parsimoniously suggest that natural selection was consistently operating to increase ECV within hominin populations (as opposed to including periods of stasis or drift). We should note that this model assumes evolvability and the strength of selection are constant across generations, both of which are unlikely to be true (see ref. [24] for an example involving the hominin hip). Like all studies using models, how model results and empirical data differ tells us something about how these assumptions are violated, which is useful information. Thus, this model serves as a useful heuristic for bridging the disparate time scales of microevolutionary studies and the hominin fossil ECV record.

Proportional change in some trait for a given population can be calculated with the equation: , where is the mean-standardized selection gradient (i.e., the percent change in fitness given a 1% change in the trait), is a measure of evolvability quantifying the expected trait response to a unit strength of selection (i.e., ), and *t* is the number of generations [22,23]. Using a database of morphological traits, Hereford and colleagues found median is 0.28 for multi-trait studies that take into account trait covariance [25], which is an appropriate assumption here since hominin brain and body size are known to be highly correlated [26]. Hominins are estimated to have an of 0.006%, as estimated from phenotypic data (M. Grabowski pers. comm.). The bin size in our additive partitioning analyses is 0.3 Ma, which translates to 11,110 generations, assuming a hominin generation time of 27 years [27]. Inputting these numbers into the model, we find hominin ECV is expected to increase via natural selection over 0.3 Ma by 0.08 log10cm3, or a factor of 1.2.

A note on the hominin evolvability estimate

We find within-lineage ECV evolves at a much slower rate than expected given the increase predicted from the microevolutionary exponential model. We obtain this result despite the very small evolvability estimate of hominin ECV (0.006%), which would dampen the amount of change possible given a certain magnitude of selection and number of generations. When Hansen and colleagues analyzed a large compilation of published evolvability estimates, they found the median evolvability for weight and volume traits is 0.94%, where 0.006% is in the 7th percentile [22]. The 0.006% and 0.94% numbers are not strictly commensurate, however, because the former was calculated using data on phenotypic variance, whereas the latter was generated using additive genetic variance. Phenotypic variance is always greater than or equal to additive genetic variance [30] and as a result, evolvability calculated using phenotypic variance is also always larger than or equal to that calculated using additive genetic variance. Therefore, the 0.006% estimate is likely biased upwards relative to the published estimates of Hansen and colleagues [22], further emphasizing the small magnitude of hominin ECV evolvability. Future research should seek to understand why hominin ECV has such unusually low evolvability, as this has implications for setting a maximum limit for how fast ECV could have evolved in the past.

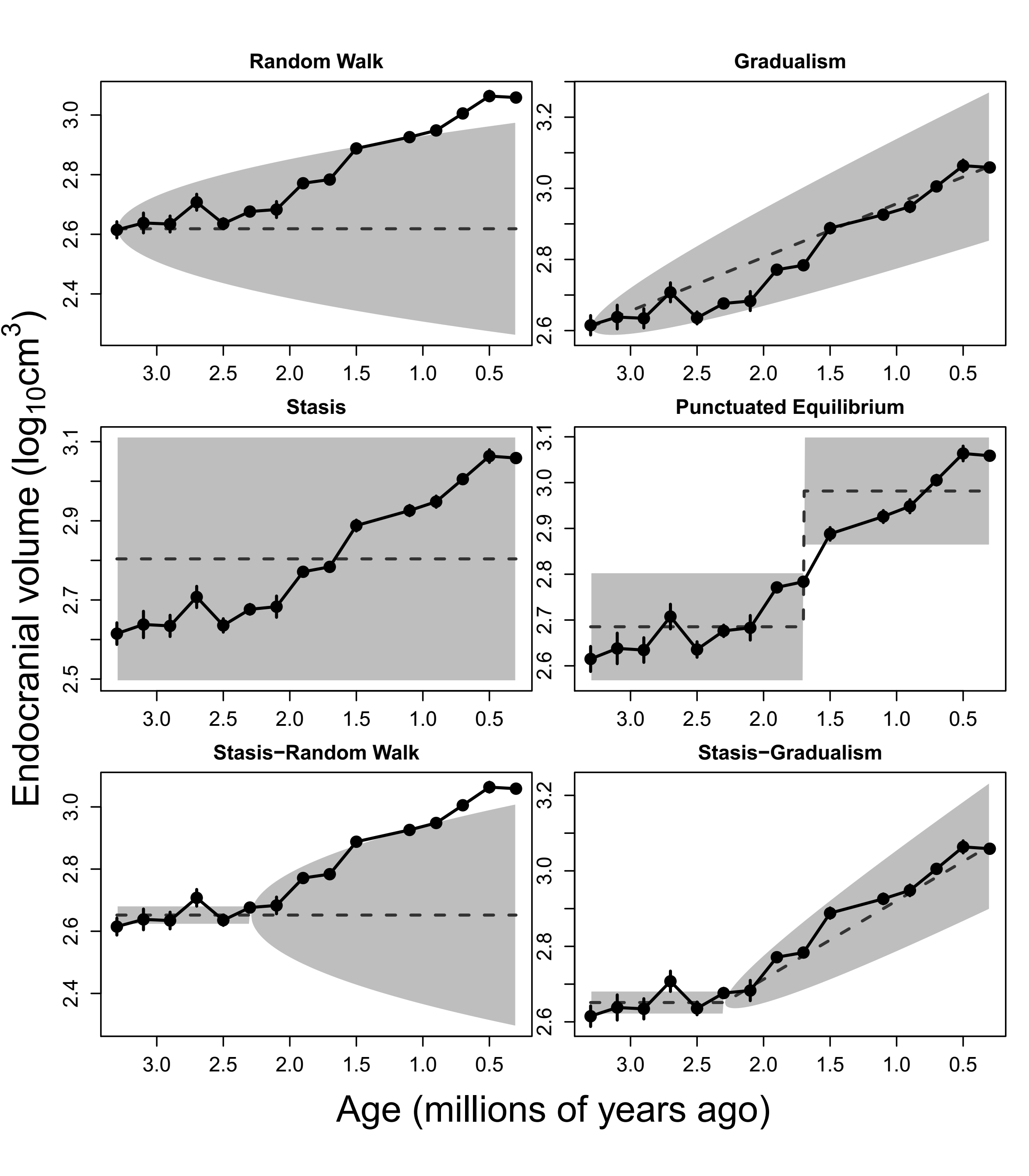


Fig. S1

Hominin ECV model fits for all six evolutionary modes. Binning was done using 0.2 Ma bins and observed (not resampled) age midpoints for plotting purposes only (the main method resamples age estimates before fitting the models, iterating this 1000 times; see fig. S2). Points are mean ECV estimates, and error bars are ± 1 standard error for each time bin. Dotted line indicates the expected evolutionary trajectory for each fitted model, surrounded by the 95% probability envelope in gray. The gradualism model fit is the same as fig. 3a in the main text.

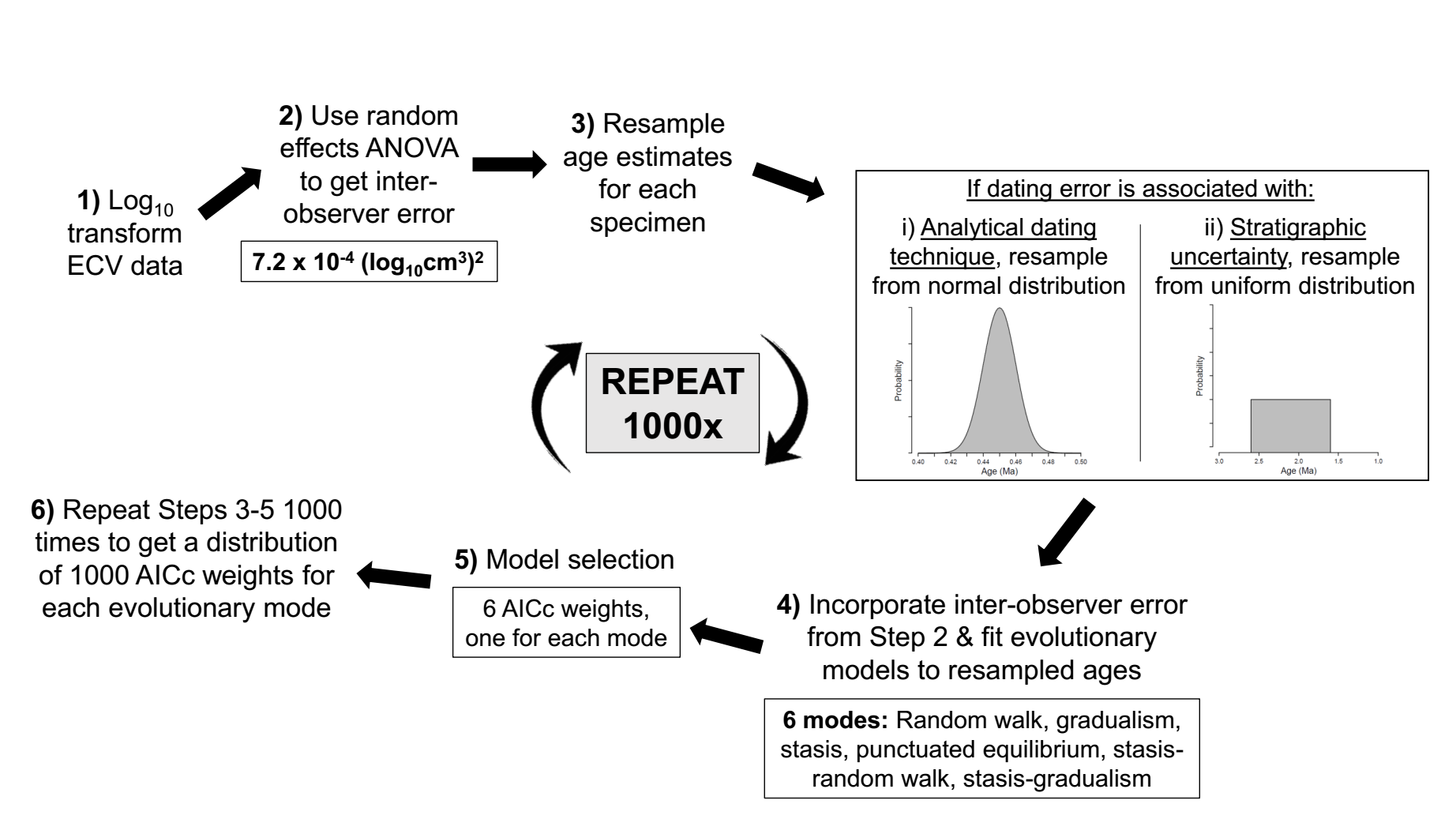


Fig. S2

Flowchart illustrating each step of our method for estimating mode of ECV evolution at the hominin clade level.

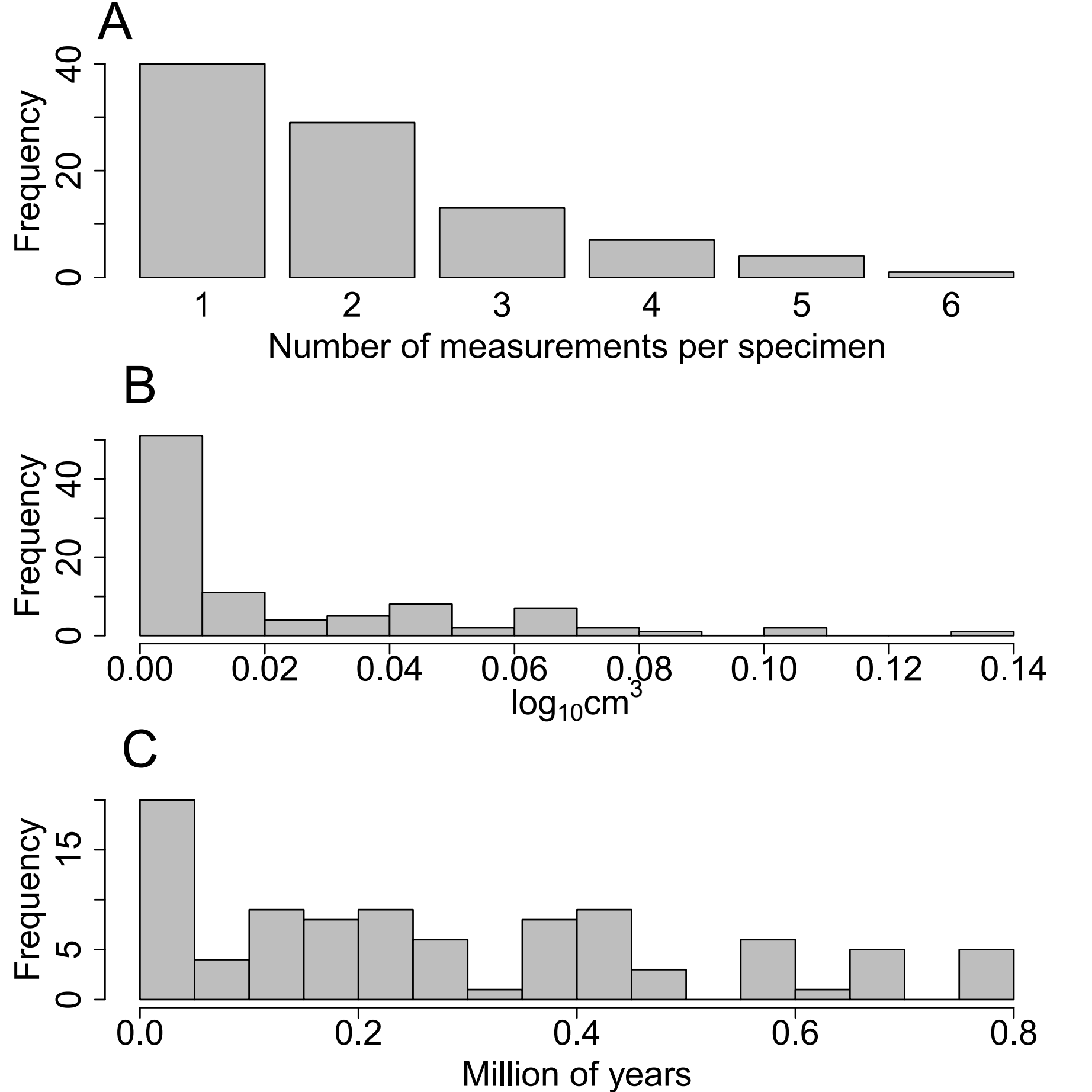
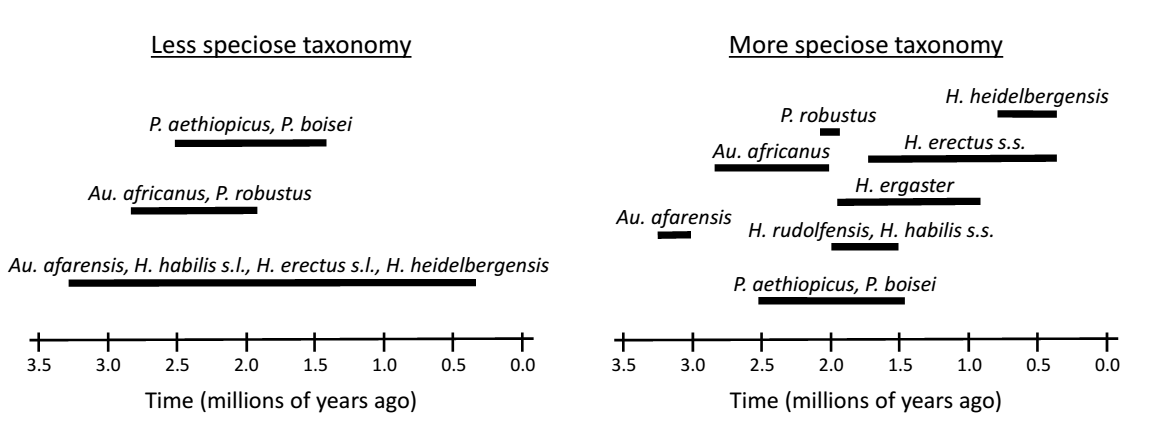


Fig. S3  
Descriptive frequency distribution plots. (A) displays the number of measurements per specimen (median: 2 measurements). 73% of specimens were measured only once or twice. (B) shows ECV error ranges per specimen (median=0.0077 log10cm3). The inflated frequency of low error estimates is due to single-measurement specimens, which by definition have zero error (n=40). (C) provides the age error ranges per specimen (median=0.25 Ma). If a specimen was dated by a single method so the age error was normally distributed (i.e., error associated with the analytical dating technique; see “Quantifying ECV and dating error”), the range was calculated as the mean ± 3 standard deviations.



**Fig. S4.**Schematic illustrating which species were grouped into which lineages for the additive partitioning analysis. Because this analysis required assigning specimens to lineages, we conducted our analysis using one version of a less speciose and one version of a more speciose taxonomy to see if our results may be influenced by taxonomic philosophy. The first and last appearance dates of each lineage (using midpoint ages, and ignoring dating error) represent, respectively, the earliest and most recent specimen on which ECV could be measured, not the earliest and latest dates for the species as a whole. Our analysis makes no phylogenetic assumptions about how lineages are related to each other. “*s.l.*” stands for *sensu lato*, whereas “*s.s.*” stands for *sensu stricto*.

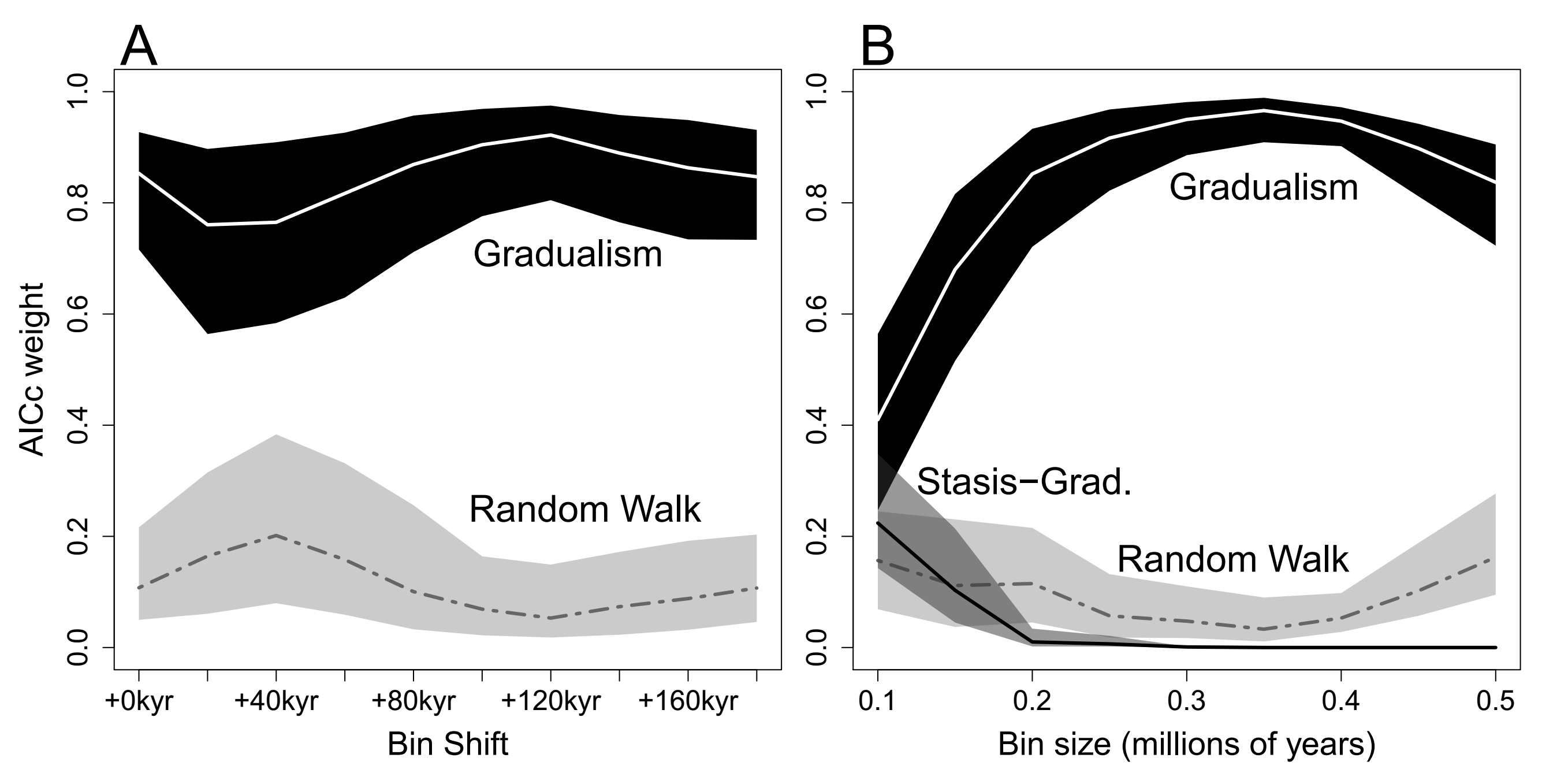
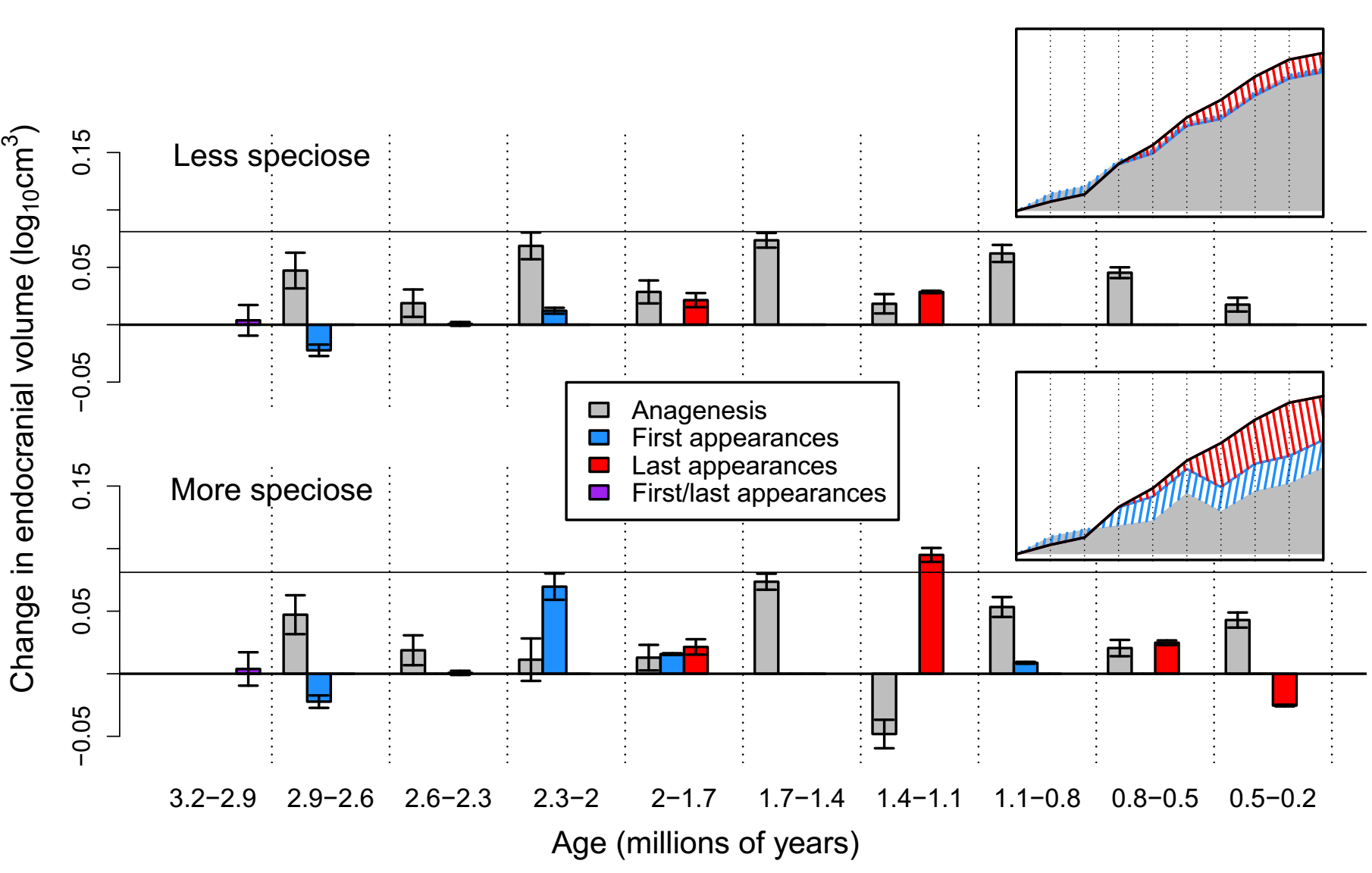
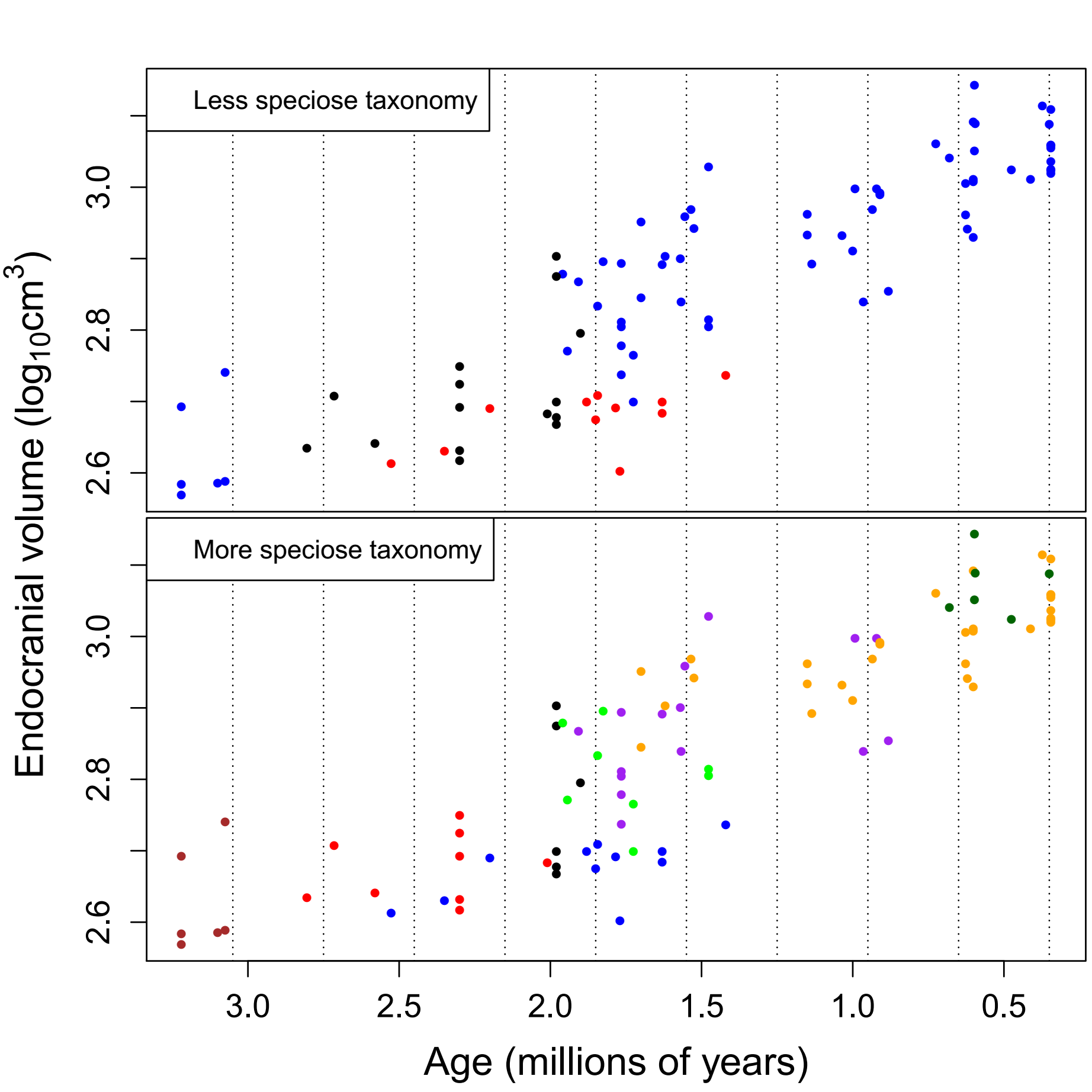


Fig. S5

Sensitivity analyses for the clade-level evolutionary mode analysis testing the effect of (A) bin location and (B) bin size. Lines denote the median AICc weight values of the top models, and confidence intervals mark the 1st and 3rd quartiles. Analyses were rerun while (A) iteratively shifting the bin location by +0.02 Ma up to +0.18 Ma and then (B) iteratively increasing bin size by 0.05 Ma from 0.1 to 0.5 Ma. Minimum number of bin samples per segment needed to be set to three for the bin size sensitivity analyses. Results show that arbitrarily selected bin attributes do not affect gradualism being the dominant model unless bin size is below 0.2 Ma. However, bin sizes this small result in a very low number of fossil hominin cranial specimens per bin (median: 3), which gives a very large sampling error and is considered technically unsound [3].

Fig. S6  
Sensitivity analysis investigating how inter-observer ECV measurement error affects the additive partitioning of clade-level changes into anagenesis and observed first and last appearances components. Ages separated by hyphens indicate over which two age bins the ECV transition is measured. “First/last appearances” represents macroevolutionary change that cannot be partitioned into separate first or last appearances components (see “Materials and Methods”). Error bars are ± 1 SE calculated by randomly sampling single ECV measurements for each specimen (see “Materials and Methods”). Dating error was ignored, and only age midpoint data were used. The horizontal black line represents the expected amount of within-lineage ECV increase over 0.3 Ma given our knowledge of how quickly natural selection operates from microevolutionary studies (see “Materials and Methods”). Insets depict the cumulative effect of each component (excluding “First/last appearances”) on the net clade-level trend (black line). Vertical dotted lines in the inset correspond to the vertical dotted lines in the main figure. Anagenesis accounts for 90% and 52% of clade-level change in the less and more speciose taxonomy, respectively. The panel on the top was created using the less speciose taxonomy, and the one on the bottom was created using the more speciose taxonomy.

  
Fig. S7  
Scatter plots of hominin specimen ECVs through time. Points represent each individual specimen’s median age and ECV measurement (error bars are removed here for clarity) and are color-coded by lineage. Vertical dotted lines represent the bin edges used in the main partitioning analysis. Top plot represents a less speciose taxonomy (blue: eastern African *Australopithecus*-*Homo*; red: eastern African *Paranthropus*; black: South African *Australopithecus*-*Paranthropus*). Bottom plot represents the more speciose taxonomy (brown: *Australopithecus* *afarensis*; blue: eastern African *Paranthropus*; red: *Australopithecus africanus*; black: *Paranthropus robustus*; green: early *Homo* (*Homo habilis* & *Homo rudolfensis*); purple: *Homo ergaster* (i.e., eastern African *Homo erectus*); orange: *Homo erectus sensu stricto* (i.e., Asian *Homo erectus*); dark green: *Homo heidelbergensis*).

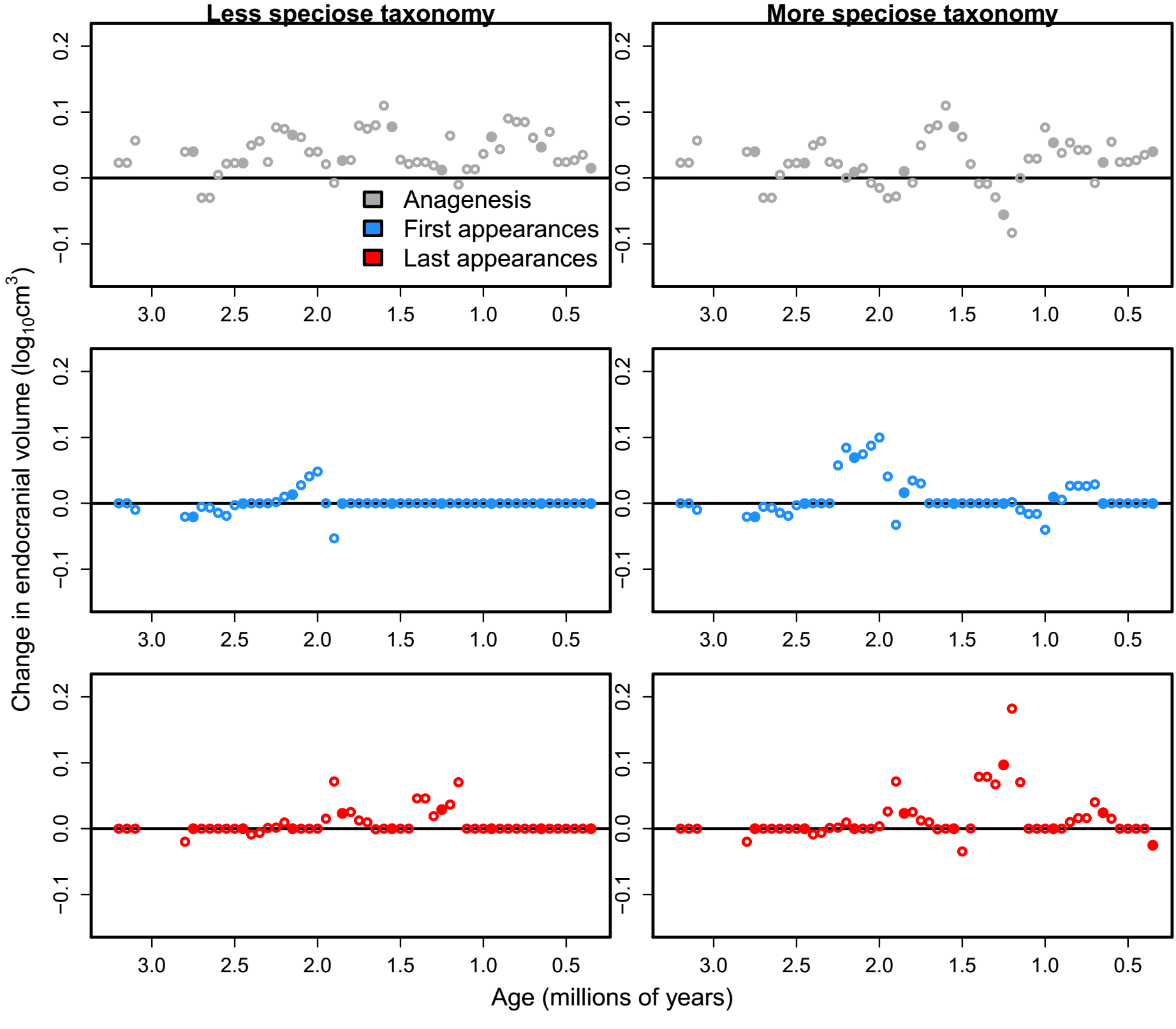
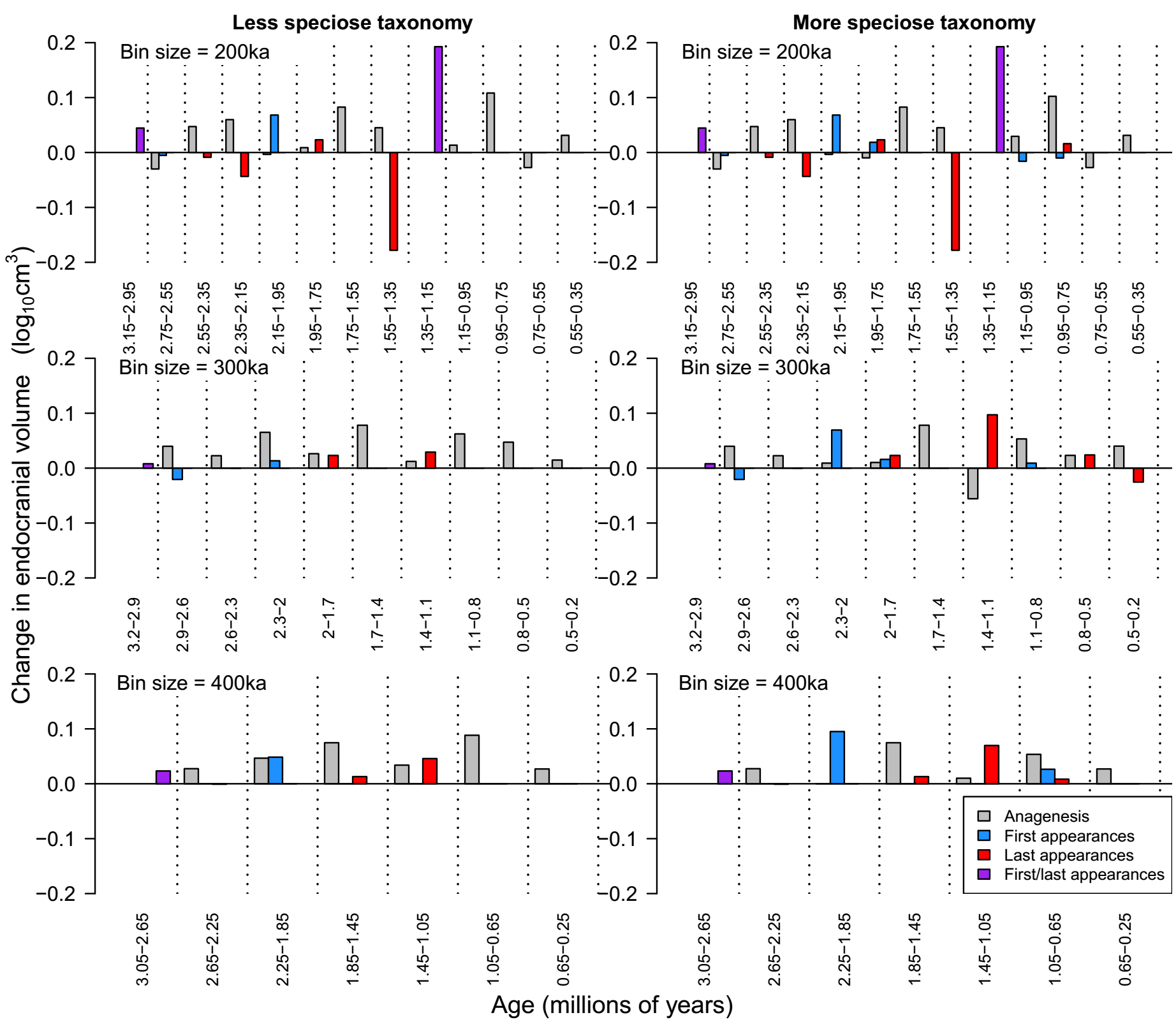
Fig. S8  
Sensitivity analysis looking at how bin location affects the additive partitioning results. Analyses were calculated using ECV medians and age midpoints while iteratively shifting bins (of 0.3 Ma size) by 0.05 Ma. Filled points represent the original bin configuration used in our main analyses (figs. 4 & S8), while open points are the results of shifting bin location. The gap in points at c.3 Ma represents clade-level changes that could not be partitioned into anagenesis or first and last appearances components (see “Materials and Methods”). Panels on the left were created using the less speciose taxonomy while those on the right were created using the more speciose one. The filled points are representative of the overall pattern created by the open points, demonstrating our results are robust to bin location.

Fig. S9  
Sensitivity analyses testing the effect of bin size on the additive partitioning results. Analyses were conducted using ECV medians and age midpoints while iteratively increasing bin size from 0.2 to 0.3 to 0.4 Ma. Plots in the left column were created using a less speciose taxonomy while those on the right were created using the more speciose taxonomy. As bin size is increased, temporal resolution decreases, so the patterns are averaged and are of lower magnitude. Despite the drastic smoothing, many patterns still hold such as first and last appearance peaks at around 2 and 1.3 Ma, respectively, and anagenetic peaks at around 1.5, 0.9, and 0.5 Ma. It is interesting that the patterns between taxonomic schemes within a given bin size are more similar than are patterns between different bin sizes within a taxonomic scheme. This suggests that choice of taxonomic designation will not radically change our results in any discernible way, and temporal resolution and scale may be more important in interpreting hominin macroevolutionary patterns, at least with regard to ECV.

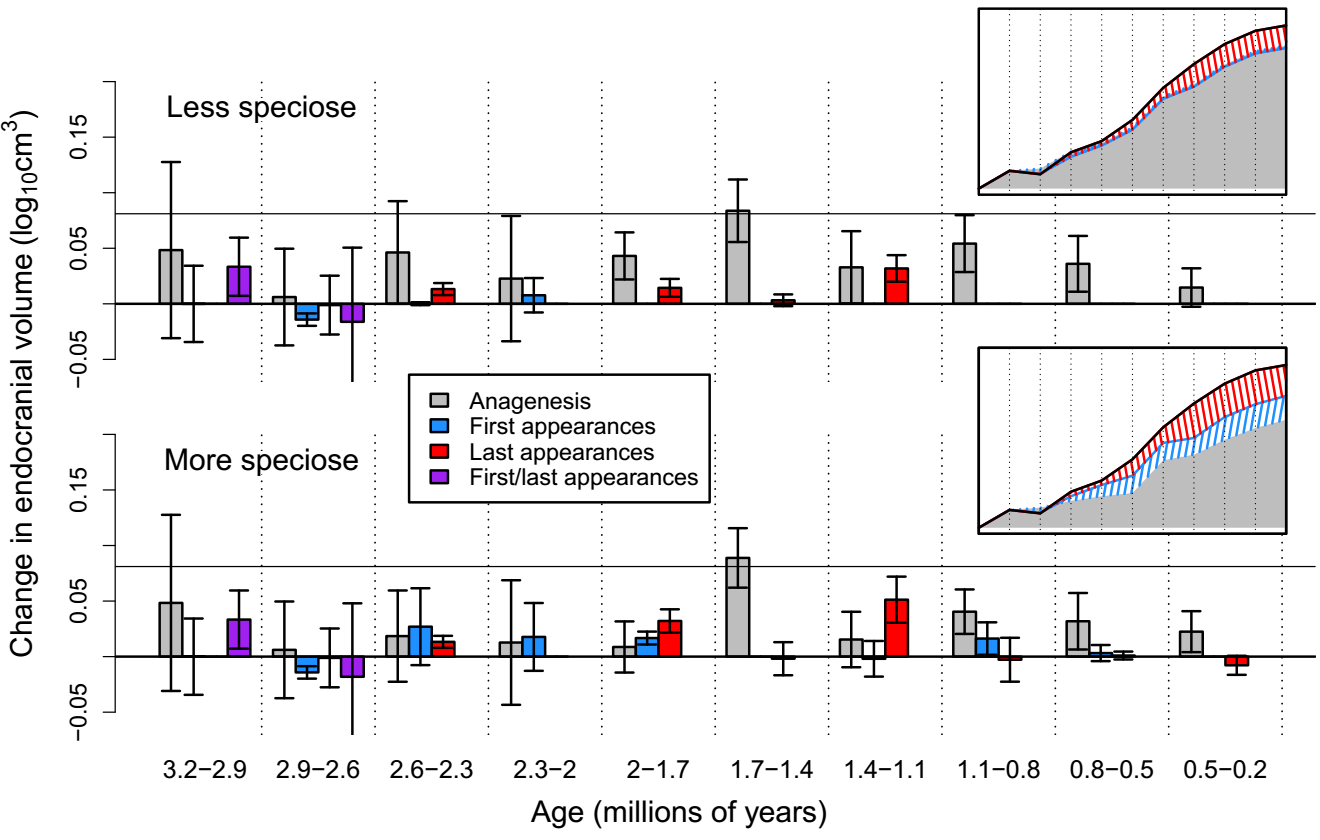
Fig. S10  
Sensitivity analysis investigating how the removal of especially incomplete or damaged specimens (n = 10; see main text) affects the additive partitioning of clade-level changes into anagenesis and observed first and last appearances components. Error bars are ± 1 SE calculated by randomly resampling age estimates. Except for the removal of especially incomplete or damaged specimens, this is the same exact analysis that produced figure 4. The fact that the patterns from these two figures are virtually identical demonstrates that the inclusion or exclusion of these specimens does not affect our results.

Table S1.  
Summary table of the hominin species included in our analysis and their respective more and less speciose species designations (for the random effects ANOVA), number of measured specimens, total number of measurements, and median number of measurements per specimen (from different researchers). It should be noted that a “lumper” might be more likely to label *Paranthropus boisei* as *Australopithecus boisei*, but for the purposes of our analyses, which species are aggregated is more important than the actual species name.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| More speciose taxonomy | Less speciose taxonomy | Number of measured  specimens | Number of total measurements | Median number of measurements per specimen |
| *Australopithecus afarensis* | Same | 6 | 9 | 1.5 |
| *Australopithecus africanus* | Same | 9 | 28 | 3 |
| *Australopithecus garhi* | Same | 1 | 1 | 1 |
| *Australopithecus sediba* | Same | 1 | 1 | 1 |
| *Paranthropus aethiopicus* | Same | 3 | 3 | 1 |
| *Paranthropus robustus* | Same | 6 | 9 | 1 |
| *Paranthropus boisei* | Same | 8 | 17 | 2 |
| *Homo habilis sensu stricto* | *Homo habilis sensu lato* | 6 | 18 | 3 |
| *Homo rudolfensis* | *Homo habilis sensu lato* | 2 | 4 | 2 |
| *Homo ergaster* | *Homo erectus sensu lato* | 5 | 8 | 2 |
| *Homo erectus sensu stricto* | *Homo erectus sensu lato* | 36 | 78 | 2 |
| *Homo georgicus* | *Homo erectus sensu lato* | 5 | 8 | 2 |
| *Homo heidelbergensis* | Same | 6 | 7 | 1 |
| **TOTAL** |  | 94 | 191 | 2 |

**Table S2**

Extracted variance components from a random effects(Type II)ANOVA used to quantify within-specimen, inter-observer ECV error. Log10 ECV was modeled as the dependent variable in an intercept only model, while “specimen ID” was nested within “species” as random effects. The model was fit by restricted maximum likelihood, and variance components were extracted using the R code: ape::varcomp(nlme::lme(ecv ~ 1, random= ~1 | species / ID, data=mydata, na.action=na.omit)). The residual variance component not accounted for by the model (i.e., “Within ID”) was used to quantify within-specimen, inter-observer error in our main analyses. We used the “Within ID” variance component from the more speciose taxonomy ANOVA model (in bold) because it was larger than that of the less speciose taxonomy and therefore was the more conservative choice. However, “Within ID” variance components do not differ appreciably between the two taxonomic treatments, so our choice of species designations does not affect our clade-level results.

|  |  |  |
| --- | --- | --- |
| **Variance component** | **Less speciose taxonomy** | **More speciose taxonomy** |
| Species | 2.253 x 10-2 | 2.027 x 10-2 |
| Specimen ID | 5.368 x 10-3 | 3.373 x 10-3 |
| Within ID | 7.215 x 10-4 | **7.230 x 10-4** |

Table S3  
Model selection results for our clade-level evolutionary mode analyses. 1000 Monte Carlo iterations (see fig*.* S2) generated a distribution of 1000 raw AICc scores and 1000 AICc weights for each evolutionary mode. Medians and 1st and 3rd quartiles were calculated for the AICc weights as opposed to mean ± standard error because of the skewed nature of the distributions (since they are bounded between zero and one). “k” is the number of free parameters estimated in each evolutionary model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evolutionary mode** | **k** | **Raw AICc mean ± one standard error** | **AICc weight median (1st & 3rd quartiles)** |
| Random walk | 2 | -45.83 **±** 3.79 | 0.06 (0.02, 0.12) |
| Gradualism | 3 | -51.62 **±** 5.94 | 0.90 (0.80, 0.95) |
| Stasis | 2 | -8.90 **±** 0.97 | 0.00 (0.00, 0.00) |
| Punctuated equilibrium | 4 | -29.66 **±** 2.77 | 0.00 (0.00, 0.00) |
| Stasis-random walk | 4 | -39.06 **±** 4.75 | 0.00 (0.00, 0.01) |
| Stasis-gradualism | 5 | -43.08 **±** 6.77 | 0.02 (0.00, 0.06) |

Table S4  
1000 Monte Carlo iterations resampling ages using 0.2 Ma bins and then fitting the gradualism model generated sampling distributions of parameter estimates and R2 values, the means of which were the mean estimates and the standard deviations of which were the standard error. The mean of the step distribution determines directionality, and the variance determines volatility of the trend [3]. The mean of the step distribution can be reliably used as a time-invariant rate metric when the underlying model is actually gradualism and the step mean is much larger than the step variance, as in this case [31]. R2 was calculated for the best fit model (i.e., gradualism) because AICc scores and AICc weights only provide a measure of relative goodness of fit, whereas R2 gives a measure of absolute fit.

|  |  |
| --- | --- |
| **Parameter** | **Mean ± standard error** |
| Initial ECV | 2.609 **±** 0.004 |
| Mean of step distribution | 0.146 **±** 0.005 |
| Variance of step distribution | 0.004 **±** 0.003 |
| R2: 0.676 **±** 0.025 | |

Table S5  
Model selection results for our clade-level evolutionary mode sensitivity analysis with especially incomplete or damaged specimens removed (n = 10; see main text). Except for the removal of those specimens, this is the same exact analysis that produced Table S3. The fact that the numbers from these two tables are virtually identical demonstrates that the inclusion of these specimens does not affect our results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evolutionary mode** | **k** | **Raw AICc mean ± one standard error** | **AICc weight median (1st & 3rd quartiles)** |
| Random walk | 2 | -43.92 **±** 4.60 | 0.06 (0.02, 0.13) |
| Gradualism | 3 | -49.33 **±** 6.84 | 0.90 (0.79, 0.96) |
| Stasis | 2 | -8.35 **±** 1.01 | 0.00 (0.00, 0.00) |
| Punctuated equilibrium | 4 | -29.16 **±** 3.26 | 0.00 (0.00, 0.00) |
| Stasis-random walk | 4 | -36.57 **±** 5.20 | 0.00 (0.00, 0.01) |
| Stasis-gradualism | 5 | -40.34 **±** 6.98 | 0.01 (0.00, 0.05) |

Additional Data table S1 (separate file)

Excel spreadsheet with the raw data used for all analyses. Each row is a separate specimen along with its ID. Columns include the “lumper’s” and “splitter’s” taxonomy used in the random effects ANOVA to get inter-observer error (“lump.taxon” and “split.taxon”), the less and more speciose lineages used in the lower-taxonomic additive partitioning analyses (“lump.part” and “split.part”), region where each specimen comes from which aided in the allocation of specimens to lineages (“region”), grade for coding points in fig. 2 (“grade”), which specimens were excluded for the damaged specimens sensitivity analysis (“reliab.sens”), ECV replicate measurements from different researchers (“ecv1” to “ecv6”), and the various dates for each specimen (“min.date”, “max.date”, “mean.date”, and “sd.date”) and their respective age error distribution (“date.dist”).

**References**

1. Tobias PV. 1971 *The Brain in Hominid Evolution*. New York: Columbia University Press.

2. Holloway RL, Broadfield DC, Yuan MS. 2004 *The Human Fossil Record, Brain Endocasts: The Paleoneurological Evidence, Volume 3*. Volume 3. New York: Wiley-Liss.

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

4. Hunt G. 2008 Gradual or pulsed evolution: When should punctuational explanations be preferred? *Paleobiology* **34**, 360–377.

5. Hunt G, Hopkins MJ, Lidgard S. 2015 Simple versus complex models of trait evolution and stasis as a response to environmental change. *Proc. Natl. Acad. Sci.* **112**, 4885–4890. (doi:10.1073/pnas.1403662111)

6. Eldredge N, Gould SJ. 1972 Punctuated equilibria: an alternative to phyletic gradualism. In *Models in Paleobiology* (ed TJM Schopf), pp. 82–115. San Francisco: Freeman, Cooper and Company.

7. Hunt G, Rabosky DL. 2014 Phenotypic Evolution in Fossil Species: Pattern and Process. *Annu. Rev. Earth Planet. Sci.* **42**, 421–441. (doi:10.1146/annurev-earth-040809-152524)

8. Jablonski D. 2007 Scale and hierarchy in macroevolution. *Palaeontology* **50**, 87–109. (doi:10.1111/j.1475-4983.2006.00615.x)

9. R Core Team. 2015 *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. See http://www.R-project.org/.

10. Hunt G. 2015 *paleoTS: Analyze Paleontological Time-Series. R package version 0.5-1.* See cran.r-project.org/web/packages/paleoTS/.

11. Sheets HD, Mitchell CE. 2001 Why the null matters: statistical tests, random walks and evolution. *Genetica* **112**–**113**, 105–125.

12. Hunt G. 2008 Evolutionary patterns within fossil lineages: Model-based assessment of modes, rates, punctuations and process. In *From Evolution to Geobiology: Research Questions Driving Paleontology at the Start of a New Century. Paleontological Society Papers* (eds PH Kelley, RK Bambach), pp. 117–131. Pittsburgh: Paleontological Society.

13. Burnham KP, Anderson DR. 2002 *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd ed. New York: Springer.

14. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. In press. *nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-120*. See http://CRAN.R-project.org/package=nlme.

15. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290.

16. White TD. 2009 Human Origins and Evolution: Cold Spring Harbor, Deja Vu. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 335–344. (doi:10.1101/sqb.2009.74.016)

17. Howell FC. 1999 Paleo-Demes, Species Clades, and Extinctions in the Pleistocene Hominin Record. *J. Anthropol. Res.* **55**, 191–243.

18. Wood B, Grabowski M. 2015 Macroevolution in and Around the Hominin Clade. In *Macroevolution* (eds E Serrelli, N Gontier), pp. 345–376. Springer International Publishing. (doi:10.1007/978-3-319-15045-1\_11)

19. Rego BL, Wang SC, Altiner D, Payne JL. 2012 Within- and among-genus components of size evolution during mass extinction, recovery, and background intervals: a case study of Late Permian through Late Triassic foraminifera. *Paleobiology* **38**, 627–643. (doi:10.1666/11040.1)

20. Simpson C. 2013 Species selection and the macroevolution of coral coloniality and photosymbiosis. *Evolution* **67**, 1607–1621. (doi:10.1111/evo.12083)

21. Wang SC, Marshall CR. 2016 Estimating times of extinction in the fossil record: Table 1. *Biol. Lett.* **12**, 20150989. (doi:10.1098/rsbl.2015.0989)

22. Hansen TF, Pélabon C, Houle D. 2011 Heritability is not Evolvability. *Evol. Biol.* **38**, 258–277. (doi:10.1007/s11692-011-9127-6)

23. Hansen TF. 2012 Adaptive landscapes and macroevolutionary dynamics. In *The Adaptive Landscape in Evolutionary Biology* (eds EI Svensson, R Calsbeek), pp. 205–226. Oxford: Oxford University Press.

24. Grabowski M, Roseman CC. 2015 Complex and changing patterns of natural selection explain the evolution of the human hip. *J. Hum. Evol.* **85**, 94–110. (doi:10.1016/j.jhevol.2015.05.008)

25. Hereford J, Hansen TF, Houle D, Fenster C. 2004 Comparing strengths of directional selection: how strong is strong? *Evolution* **58**, 2133–2143. (doi:10.1554/04-147)

26. Grabowski M. 2016 Bigger Brains Led to Bigger Bodies?: The Correlated Evolution of Human Brain and Body Size. *Curr. Anthropol.* **57**, 174–196. (doi:10.1086/685655)

27. Langergraber KE *et al.* 2012 Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci.* **109**, 15716–15721. (doi:10.1073/pnas.1211740109)

28. Gould SJ. 2002 *The structure of evolutionary theory*. Cambridge, Mass: Belknap Press of Harvard University Press.

29. Jablonski D. 2008 Species selection: Theory and data. *Annu. Rev. Ecol. Evol. Syst.* **39**, 501–524.

30. Gillespie JH. 2004 *Population genetics: a concise guide*. 2nd ed. Baltimore, Md: Johns Hopkins University Press.

31. Hunt G. 2012 Measuring rates of phenotypic evolution and the inseparability of tempo and mode. *Paleobiology* **38**, 351–373. (doi:10.1666/11047.1)