Supplementary Materials

Tree of life reveals clock-like speciation and diversification

S. Blair Hedges[a–c,1], Julie Marin[a–c], Michael Suleski[a–c], Madeline Paymer[a–c], and Sudhir Kumar[a–c]

[a]Center for Biodiversity, Temple University, Philadelphia, PA 19122, USA.
[b]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA.
[c]Department of Biology, Temple University, Philadelphia, PA 19122, USA.
[1]To whom correspondence may be addressed. E-mail: sbh@temple.edu

This file includes:

1. Supplementary Materials and Methods

Timetree of life (TTOL) data collection. We synthesized the corpus of scientific literature where the primary research on the timetree of life is published. We first identified and collected all peer-reviewed publications in molecular evolution and phylogenetics that reported estimates of time of divergence among species. These included phylogenetic trees scaled to time (timetrees) and occasionally tables of time estimates and regular text. For collecting timetrees, we began by surveying phylogeny data collections, including TreeBASE and Dryad, which, for the most part, did not produce useable timetrees because such data repositories have not prioritized timetree acquisition (Hedges et al. 2006). Therefore, we conducted automated and manual bibliographic searches on major repositories holding peer-reviewed articles (e.g., Web-of-Science, Google-Scholar, and PubMed) using a series of keywords (Hedges et al. 2006). Because some primary research articles also appear in books and monographs, we manually searched books and monographs when they were not available online. Upon identification of articles containing estimates of molecular divergence times, we adopted a "community contributions" approach and requested timetree data directly from the corresponding authors. Approximately 50% of authors sent data files on request (as of 2013). Our team coded the remaining timetrees in an effort to ensure greater coverage whenever the authors did not respond to our multiple requests. We excluded studies that (i) may have confused paralogy and orthology, (ii) presented a spectrum of different times for the same node, based on different methods, and without stating a consensus or preferred time, (iii) had internal inconsistencies such as major conflict with the geologic or fossil record, (iv) were subsequently revised by the same author, or other authors, in which case the revised study was used, or (v) where times were based on limited evidence or conjecture. Possessing a high or low time estimate, or use of a particular program or method, was not considered as a criterion for inclusion or exclusion.

Using these criteria, we assembled timetree data from 2,274 studies (http://www.timetree.org/reference_list2014.html) that have been published between 1987 and April, 2013. Most (96%) of nodal times used were published in the last decade. In some published, large-scale timetrees (Bininda-Emonds et al. 2007; Fritz et al. 2009; Jetz et al. 2012), authors have included, by interpolation methods, species lacking molecular data. We did not include those dataless species in our TTOL (Fig. 1) and they were not used in analyses. However, we make available, in Newick format, the entire TTOL, smoothed and unsmoothed, separate unsmoothed timetrees of major clades, and smoothed timetrees of birds and mammals containing interpolated species (http://www.biodiversitycenter.org/ttol.html). We included in the TTOL some non-standard taxa (4.5%) lacking a complete Latin name to enhance biological information, mostly from taxonomically poorly-sampled groups. The median age of each is 15 Ma, none is identical (time=0) to any other species, and there are no long series of closely related taxa that suggest extensive sampling within a species. For all of these reasons, this suggested to us that they are valid species that are unidentified or undescribed, and should be included.

The next step was to convert divergence time data into usable and computable objects via standardization and taxonomic representation. This involved mapping of taxa names in the input data with the taxonomic identifiers, which was done automatically by using an in-house program that utilized the NCBI taxa identifiers. We manually resolved all invalid taxa names,

spelling errors, and naming conflicts arising due to non-uniqueness of species and common names. We also conducted automated tests of timetree consistency to ensure that the ancestral nodes were not younger than the descendent nodes, the tree structures were complete as appropriate, and the time units were in millions of years. The standardized representations are stored in a relational database as Computable Timetree Objects (CTOs). The total number of species with divergence time estimates, as found in CTOs, has grown quickly over the last two decades (Supplementary Material Figure 1). The collection of CTOs is peer-reviewed knowledge (beyond primary data) that provides opportunities for enhanced synthesis, discovery, and integration of valuable information produced by the growing community of scientists.

Assembly of individual timetrees. To enhance our taxonomic coverage in the TTOL we estimated timetrees from two published phylogenetic trees containing large numbers of species (Pyron et al. 2013; Pyron and Wiens 2011), each a composite of public sequence data from many earlier studies and authors. Our calibrations for amphibian families (Supplementary Material Table 1) come directly from published syntheses (Hedges and Kumar 2009). For squamate reptile families we calibrated with mean estimates from two studies (Mulcahy et al. 2012; Vidal and Hedges 2005) for the following crown-group nodes: Squamata (Node 1; 203.9 Ma), Bifurcata (Node 2; 196.0 Ma), Unidentata (Node 3; 187.5 Ma), Episquamata (Node 4; 171.4 Ma), Scinciformata (Node 5; 161.1 Ma), Laterata (Node 6; 158.6 Ma), and Toxicofera (Node 7; 162.9 Ma).

The phylogenetic relationships were assessed with the original molecular alignment and partitions (Pyron and Wiens 2011) using Maximum Likelihood (ML) methods of inference with RAxML 7.2.8 (Stamatakis 2006) and performed 1000 bootstrap replicates to obtain a bootstrap majority rule consensus tree. Trees were visualised with FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). We then spilt the amphibian tree into three parts, Anura (with caudatan *Cryptobranchus alleganiensis* as outgroup), Caudata (with anuran *Ascaphus montanus* as outgroup) and Gymnophiona (with caudatan *Cryptobranchus alleganiensis* as outgroup) for the divergence time estimation. The same method was used to estimate the divergence times of Squamates, Anura, Caudata and Gymnophiona. These trees with branch lengths were then subjected to the RelTime (Tamura et al. 2012) analysis to generate relative times, which were converted into absolute times by multiplication with a global scaling factor obtained by a linear regression, through the origin, between the relative times and the calibrations points. Confidence intervals at each node were extrapolated from the relationship (second order polynomial equation) between the upper and lower of the confidence intervals and node times from similar studies involving reptiles (Vidal et al. 2010) (r^2= 0.99 and 0.99 for upper and lower limits respectively) and amphibians (Hedges and Kumar 2009) (r^2 = 0.99 and 0.97 for upper and lower limits respectively).

Timetree of life (TTOL) analytics and synthesis. The fundamental unit of synthesis across CTOs in our database was the divergence time of a pair of clades (A and B) which have directly descended from an ancestor (X) in the tree of life. Clades A and B contain one or more species (n_A and n_B, respectively), such that every species in A (S_A) and B (S_B) is found in at least one CTO in our dataset. Then, the primary objective is to find the divergence time T_{AB} between clades based on all available timetree data in all studies. To do so, we efficiently scanned the

hierarchical database of CTOs to generate a list of species-pair times ($t_{ab,cd}$) such that $a \in S_A$, $b \in S_B$, c is the CTO, and d is the source publication for c. Using these data, we estimated T_{AB} as follows $T_{AB} = \frac{1}{n_d}\sum_d \left(\frac{1}{n_{cd}}\sum_c MRCA(A,B)\right)$. Here, the outermost summation over studies (d's) is a simple average from all the studies (n_d), where every study contributes exactly one divergence time, which avoided undue influence of one study on the final time estimate for any pair of taxa. The middle summation is over CTOs (c's) such that times from multiple timetrees from the same study (n_{cd}) are given equal weight and an average is computed to represent the time estimate from that study. And, the innermost term produces the timing of the most recent common ancestor (*MRCA*) of all the species in clades A and B that are present in the given timetree (c). In this way, timetree data enables synthesis across studies without taking a traditional phylogeny-partition-based supertree approach, which is not feasible because of the extreme sparseness of the species-studies data.

We used the above method of estimating divergence times (T's) of clade pairs to build the Super Timetree (STT), along with a procedure for testing and updating topological partitions to ensure the highest degree of consistency with individual timetrees in every study. For simplicity, we started with the topology at the NCBI taxonomy browser. This choice was made because it contains all the species for which molecular data have been reported (and thus used in building molecular timetrees). Also, it is rather conservative, which means that it has a large number of polytomies to avoid potential biases associated with the use of incorrect topological resolutions. We began by first resolving each polytomy locally using a Hierarchical Average Linkage (HAL) method illustrated in Supplementary Material Figure 2. The application of HAL method to the timetrees from 2274 studies resulted in 40,918 resolutions (92% of the 44,502 total nodes; 3,584 polytomies) that, along with partitions in the initial tree, represent phylogenetic hypotheses in the STT.

These hypotheses were then evaluated by estimating the number of timetree topologies in the database that showed concordance and discordance with each partition. An overwhelming majority of partitions in the STT represented the majority-rule consensus of timetrees (90.5%) when partitions in the study trees could be used to test the local topological configurations in the STT (25,186 partitions). For some STT partitions the number of discordances outnumbered concordances because of short inter-branches and/or the limitations on the number of studies that results in time estimates with large variances. So we tested alternative topological configurations for every HAL resolution and adopted the topology that minimizes the discordances with the topological data in the timetrees. This made the final STT highly concordant with the study trees; 98.0% of the testable partitions in the STT now represented the majority-rule consensus. The average study tree concordance over all the STT partitions tested was 98.6%.

In some cases, we found that the descendant nodes were older than the ancestral node time resulting in negative branch lengths, which were carefully examined and some were resolved by examining online taxonomies. In general, this issue was fixed by changing the heights of the conflicting nodes to an average of the two nodes' ages. If an ancestral node had multiple older descendants, we identified the descendant supported by the most studies and used its age in the average. In this instance, the ages of all the conflicting descendants, regardless of age, were changed to this average. Similarly, if descendant node ages conflicted

with multiple ancestors, all of these ancestors' ages were changed to the average. To ensure that these node heights were fixed consistently without creating additional negative branches, we recursively searched for negative branches beginning from the root of the tree.

At the end of this process, a STT was produced which has 50,632 species and 3,584 remaining polytomies (Fig. 1). Each study time at each node in the tree has its associated uncertainties, reflected in combined among-study confidence intervals at each node (Supplementary Material Table 2); nodes with only a single study show the confidence interval of the node from of that study. Coefficients of variation based on mean times among studies are ~30% for recent nodes (< 5 Ma) and decline to less than 10% for ancient nodes (> 1000 Ma) in TTOL. Standard errors (among studies) for nodes with 10 or more studies averaged about 5% of the mean.

A recent study of birds (Jarvis et al., 2014) was published too late to be included in our TTOL. The time estimates in that study, for splits among orders of birds, are generally younger than in previous studies and in our TTOL. We note that the single maximum calibration used in that study, and probably of importance in establishing times throughout their timetree, was chosen arbitrarily to correspond to a geological boundary. Further analysis of that data set is warranted.

We did not use interpolated species (species lacking molecular data) in any diversification analysis, and they are not included in the TTOL (Fig. 1). However, they are used widely in the field of biodiversity analysis and we wished to make them available for the research community. Also, since they brought the number of species in a group up to the full described species count, they facilitated our analyses of clade age versus size, in facilitating counts of clade size for genera and families of birds and mammals. To add interpolated species to a timetree of a group (e.g., birds, mammals), we started by first locating the species missing from the tree by searching the online taxonomic databases. Then we identified the node representing the genus and attached the species directly to it. Where the genus node was not available, we searched the tree for another species of the same genus and created a new genus node containing the interpolated taxa and all other representatives of the genus in the tree. We chose not to use interpolated species if the tree did not contain other representatives of the same genus, although this was rare.

A birth–death (BD) polytomy resolution approach was applied to our TTOL (Kuhn et al. 2011), and the resulting tree was used for the diversification rate analyses. We used the stand-alone input file generator to produce a BD script, carried out with R, and split the tree into 35 subtrees (backbone and plants, eubacteria, insects, etc.). With this script a uniform prior is employed for both the diversification rate ($\lambda - \mu$) and extinction fraction (μ/λ) parameters. We recorded samples every 1000 iterations (MCMC) and divided the analysis into seven independent runs of 5 million iterations. A burnin period of 500,000 was applied to reach stationarity and trees were resampled at lower frequency (2,000 to 10,000) to avoid memory issues. The smoothed trees were then reassembled together to a smoothed TTOL. This polytomy resolution method was also applied to the several groups where interpolated species were added (birds, mammals, and squamates). Although smoothing is necessary to remove the bias of artificial polytomies on diversification, it can introduce bias in the resulting diversification model supported (Kuhn et al. 2011) and will resolve polytomies in a random

pattern, thus restricting the usefulness of smoothed trees. For this reason we compared our results using smoothed and unsmoothed trees.

Estimation of variances for divergence times in the super timetree. Estimation variances for divergence times obtained from molecular data were expressed in form of 95% confidence intervals. To generate this information, we scanned the supplementary information in source studies and collected confidence intervals by parsing the newick tree files, which contained 95% highest priority density estimates from Bayesian analyses. We also collected information tables in publications, which provided estimates of confidence intervals for a subset of nodes in the respective timetrees. In addition, many studies provided estimates of standard error in tables or timetrees, which were converted into a confidence interval assuming a normal distribution. All of these confidence intervals mapped to 21,901 nodes in the super timetree. In this mapping, the confidence interval for the mean divergence time for the node was scaled such that the ratio of the confidence interval size to the node age in the source tree was preserved. We explored the relationship of the node age with the ratio of the confidence interval with normalized by node age and found that a power regression model closely fit the data within studies and over all studies, so we estimated the lower and upper bounds of confidence intervals for all other nodes in the super timetree through predictive modeling.

Linnaean rank analyses. To estimate branch times of Linnaean ranks, we assessed the bootstrap modes (Hedges and Shah 2003) of their branching times. The bootstrap standard error was calculated, and 500 (or twice 50 iterations for groups with more than 2000 values) were performed. We also obtained the confidence interval and the half-range mode. The results are summarized in Supplementary Material Table 3.

Branch length distribution. We plotted the cumulative frequency of branch lengths of the TTOL (non-smoothed tree) with bins of 0.1 Myr. Then we fitted eight models (Supplementary Material Table 14) and selected the best model using AIC scores. The exponential model (y ~ I(exp(1)^(a + b * x))) was selected as the best fit to the branch length distribution of eukaryotes.

Branch times from diversification rate. Branch times of the clades were estimated using the splitting rate (λ) obtained with the slope method (see below) and the coalescent method, with the formula: $\frac{1}{2*\lambda}$ corresponding to the expected length of a random interior edge length under the Yule process (Steel and Mooers 2010) (because we wanted true branch times unaffected by extinction). The slope method (Ricklefs et al. 2007) was used on 11 clades (the same 10 non-nested groups over the major Linnaean groups used for both coalescent and gamma analyses, and the eukaryotes; Supplementary Material Table 4) to estimate the parameters λ and μ. To assess the slope and the intercept values (required in the following formulas) we applied a linear model (computed in R) using a sliding window (5 or 10 My steps) along the log linear through time plot (LTT plot). Then the bootstrap-mode method (Hedges and Shah 2003) was applied in order to obtain the modal slope and intercept of each plot. According to this method, the slope corresponds to the diversification rate (λ – μ), and (a) is the difference between the intercept (I) and the actual number of lineages (N) (a= (log(N)-I)). Then the parameters λ and μ are obtained as follows: $\lambda = \frac{\lambda-\mu}{\exp(-a)}$ and $\mu = \lambda - (\lambda - \mu)$.

Time-to-speciation analyses. In addition to our species-level TTOL data collection described above, we collected a separate data set on TTS from published molecular timetrees that included timed nodes among populations and closely related species of three major groups: vertebrates, arthropods, and plants (Supplementary Material Tables 11–13). To be included, species were required to be monophyletic and with no taxonomic confusion or possible cryptic species included. The methodology used is illustrated in Fig. 6a and described in the text. We defined intervals of time ('speciation intervals') between crown and stem ages of each species which would presumably contain the TTS (Supplementary Material Table 11). For each major group we constructed histograms of these intervals (Supplementary Material Table 12) and determined the mode and confidence interval of each (Fig. 6b) using the bootstrap mode (Hedges and Shah 2003). For each major group, we also constructed similar histograms of divergences among populations, within each species (Fig. 6b; Supplementary Material Table 13).

Species interval simulations. To test the robustness of our approach for estimating TTS, we used simulations. A birth-death tree was simulated using the function 'sim.bd.taxa' (TreeSim in R). The number of tips was set at 1000, the birth rate at 0.2, and the death rate at 0.1. We choose 2.0 Myr as the known ('true') TTS for the simulations. The nodes younger than 2 Myr were considered as population nodes and the nodes older as species nodes. We sampled randomly 3.5 population nodes, corresponding to the mean number of population lineages (per species) in our data set, and one species node between 2.0 and 50.0 Ma, to obtain 2000 species intervals. To measure interval density we sampled 0.05 Mya bins between 0 and 50 Ma, counting the number of species intervals in each bin. Then we applied two modifications to this distribution at the same time to simulate 1) higher degree of undersampling and 2) noise. To simulate a higher degree of undersampling we increased the probability to sample older species nodes by multiplying the occurrence of each node by its rank (younger to older). The opposite design was applied to population nodes, younger nodes had a higher probability to be sampled. The resulting distribution still showed a very distinct peak (and a mode) at 2 Myr. To add noise to this distribution we added species intervals that did not include the true TTS, either lower or higher.

2. Supplementary Material Figures

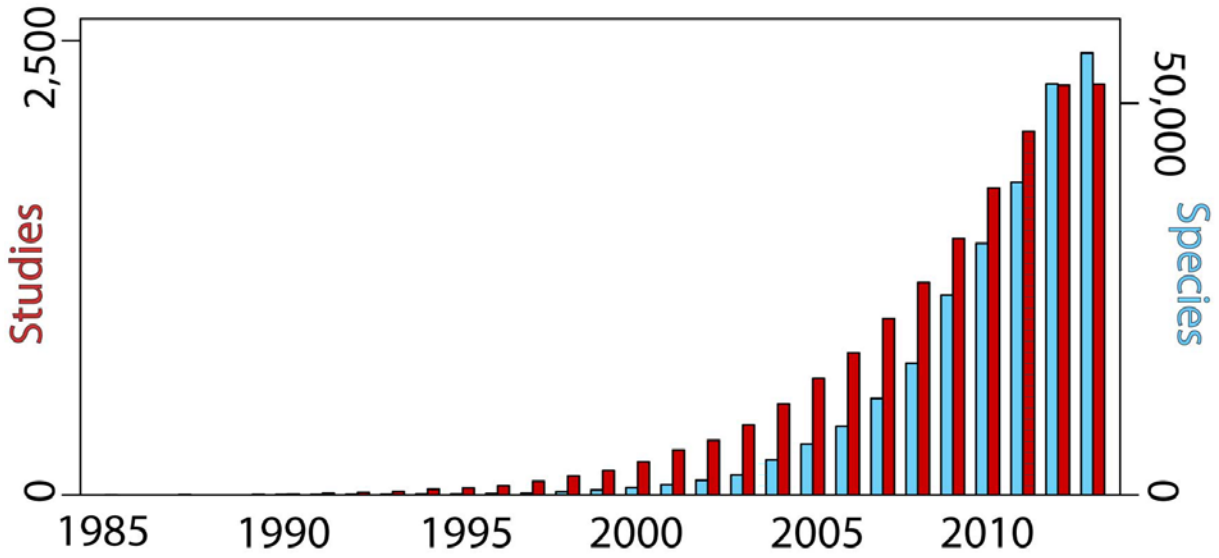

Figure S1: Cumulative growth of knowledge bearing on the timetree of life (TTOL), by year of publication (1987 to April, 2013). Published studies (red) and species (blue) in the TTOL presented in Fig. 2.

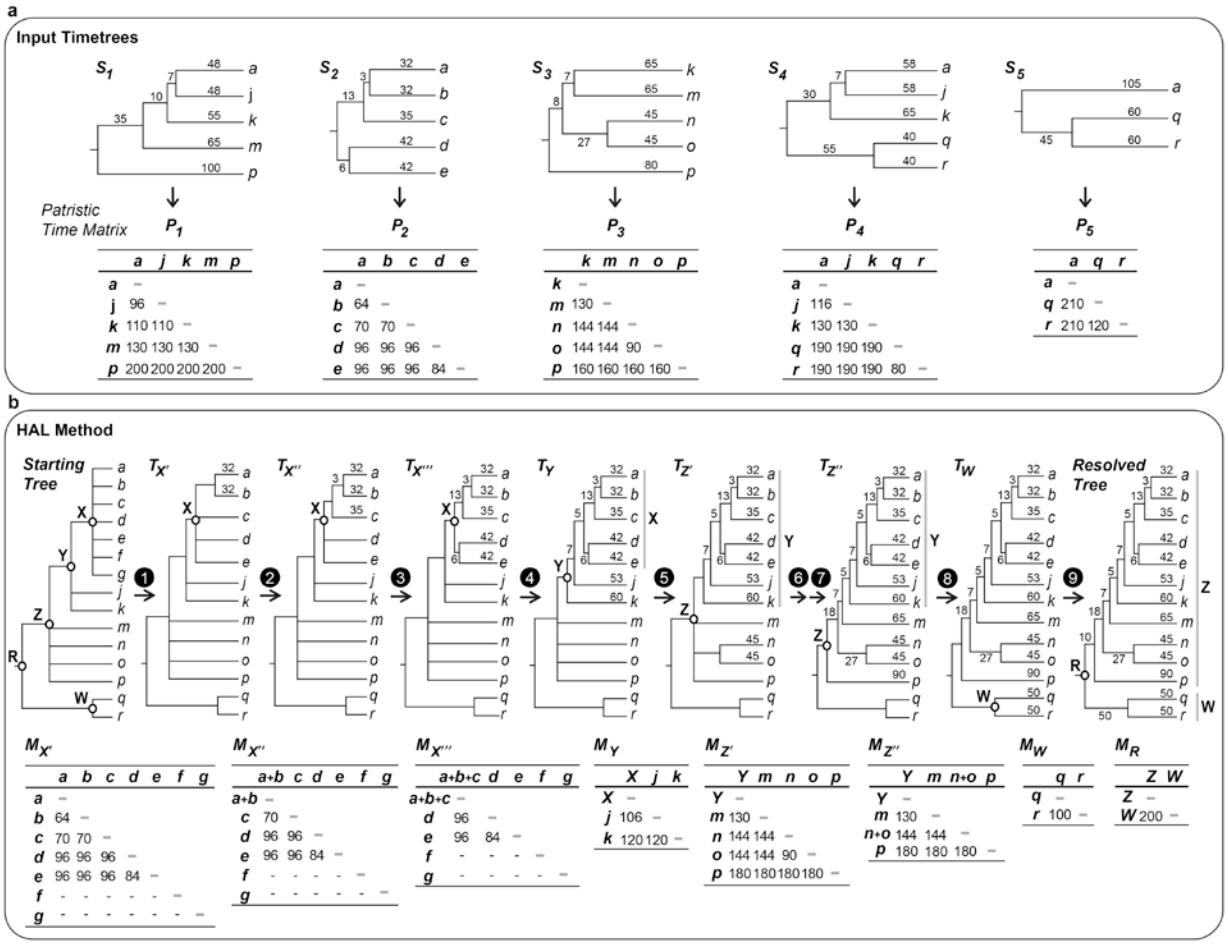

Figure S2: Illustration of method for assembling the super timetree (STT). (*a*) (Five Input time trees and associated patristic time matrices) and (*b*) (Application of Hierarchical Average Linkage, HAL, method to resolve polytomies). The input to HAL method is assumed to consist of five timetrees from different studies (S_1 – S_5), which are applied for resolving three polytomies (X, Y, and Z) in the Starting Tree and for estimating all the node times in the Resolved Tree. First, each timetree is expressed in form of a Patristic Time Matrix (*P*), where the divergence time between a pair of species (*i* and *j*) is obtained by summing the branch lengths required to traverse branches from species *i* to species *j* in the given timetree. A given *P* completely describes the corresponding *S* and has a one-to-one relationship with it.

We begin by resolving the polytomy X, which requires the computation of pairwise divergence times of its descendants (*a* to *g*) using all relevant *P* matrices. This is accomplished using the procedure mentioned in the Methods section and results in $M_{X'}$. The taxon pair with the smallest divergence time in this matrix is chosen and grouped into a composite clade (*a*+*b*), and the pairwise divergence times are computed between this new clade and the rest of the species (*c* to *g*) to produce a new $M_{X''}$ with one fewer taxon. We repeatedly select the taxon pairs with the smallest divergence time (steps ❶, ❷, and ❸) and compute time matrices until all the clades have been put together in a local topology ($T^{X'''}$). The relationships of five species (*a*-*e*) are now resolved as reflected in the input trees, with two other species (*f* and *g*)

automatically pruned because none of the input timetrees contained any information about their divergence times. Next, we resolve the polytomy Y consisting of three taxa (X, *j*, and *k*). Now, M_Y is the new pairwise time matrix and the application of HAL results in T^Y, where the relationships of X with *j* and *k* are now resolved (step ❹). Next, we build M_Z first and pair *m* and *n*, which has the smallest divergence time (step ❺). This is followed by two more rounds of pairings (steps ❻ and ❼) to generate the tree T^Z in which all polytomies are resolved. Ultimately, divergence times for all other nodes in the tree (W and R) are estimated and the final tree is obtained. As mentioned in the methods section, this is followed by an examination of the degree of topological concordance of each node (partition) in the final tree with the input timetrees. In this example, all of nodes are consistent with the input timetrees. In general, however, we rearrange all partition where discordances outnumber concordances to ensure that partitions in the final tree reflect the consensus topological configuration in the input timetrees (see Supplementary Methods: Timetree of life (TTOL) analytics and synthesis).

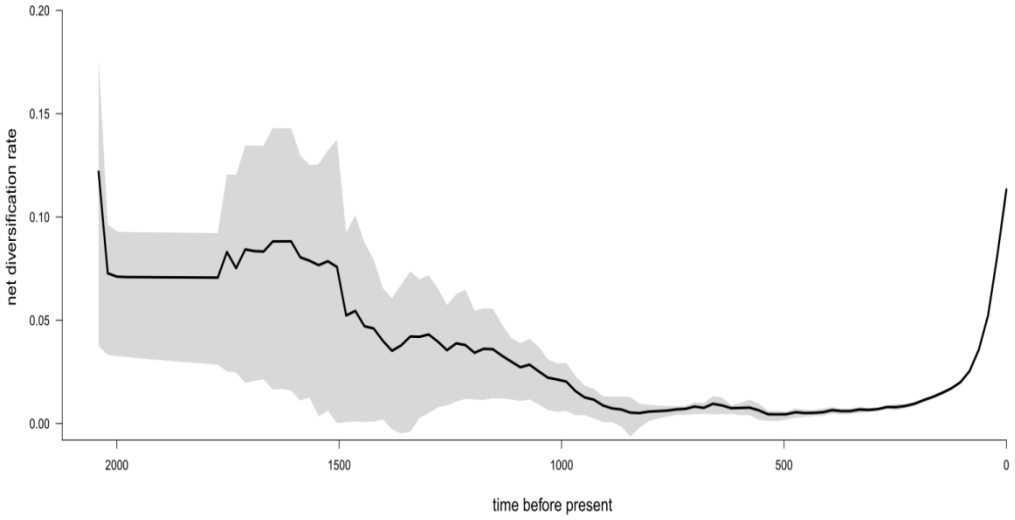

Figure S3: Diversification rate plot (mean) through time of eukaryotes (BAMM analysis). The confidence interval (0.95) is represented in grey.

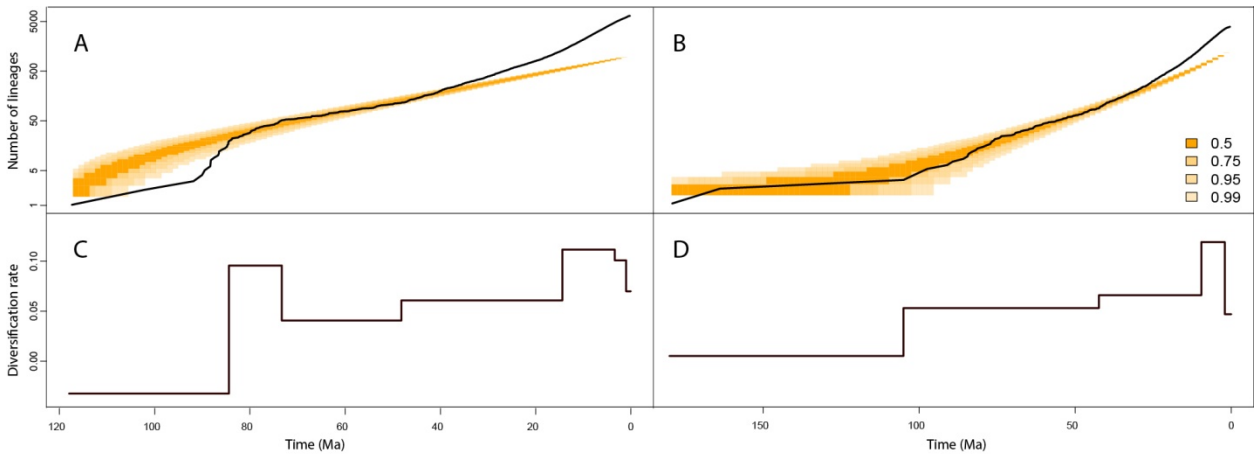

Figure S4: Lineages-through-time (LTT) and rate diversification plots of birds and mammals. (*a*) (birds; 5,363 sp.) and (*b*) (mammals; 9,879 sp.): log-transformed LTT plots; the colors represents the confidence intervals for the different p-values of the distribution of LTT plots (see methods). (*c*) (birds) and (*d*) (mammals): diversification rate plots (result from the TreePar analysis; 6 shifts for birds and 4 shifts for mammals).

References

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. Nature 446(7135):507-512.

Fritz SA, Bininda-Emonds ORP, Purvis A. 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecology Letters 12(6):538-549.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23):2971-2972.

Hedges SB, Kumar S. 2009. The timetree of life. Oxford: Oxford University Press. p. 551.

Hedges SB, Shah P. 2003. Comparison of mode estimation methods and application in molecular clock analysis. BMC Bioinformatics 4:31.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. Nature 491(7424):444-448.

Jarvis, ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320-1331.

Kuhn TS, Mooers AO, Thomas GH. 2011. A simple polytomy resolver for dated phylogenies. Methods in Ecology and Evolution 2(5):427-436.

Mulcahy DG, Noonan BP, Moss T, Townsend TM, Reeder TW, Sites JW, Wiens JJ. 2012. Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. Molecular Phylogenetics and Evolution 65(3):974-991.

Pyron RA, Burbrink FT, Wiens JJ. 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evolutionary Biology 13:93.

Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Molecular Phylogenetics and Evolution 61(2):543-583.

Ricklefs RE, Losos JB, Townsend TM. 2007. Evolutionary diversification of clades of squamate reptiles. Journal of Evolutionary Biology 20(5):1751-1762.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Steel M, Mooers A. 2010. The expected length of pendant and interior edges of a Yule tree. Applied Mathematical Letters 23:1315-1319.

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. Proceedings of the National Academy of Sciences of the United States of America 109(47):19333-8.

Vidal N, Hedges SB. 2005. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. Comptes Rendus Biologies 328(10-11):1000-1008.

Vidal N, Marin J, Morini M, Donnellan S, Branch WR, Thomas R, Vences M, Wynn A, Cruaud C, Hedges SB. 2010. Blindsnake evolutionary tree reveals long history on Gondwana. Biology Letters 6:558-561.