Supplementary Text for "Death and Taxa"

Peter D Smits

April 26, 2015

1 Materials and Methods

1.1 Species occurrence and covariate information

Fossil occurrence information was downloaded from the Paleobiology Database (PBDB; http://paleodb.org/). Occurrence, taxonomic, stratigraphic, and biological information was downloaded for all North American mammals. This data set was filtered so that only occurrences identified to the species-level, excluding all "sp."-s. All aquatic and volant taxa were also excluded. Additionally, all occurrences without latitude and longitude information were excluded from the sample.

Species dietary and locomotor category assignments were done using the assignments in the PBDB, which were then reassigned into coarser categories (Table S2). This was done to improve interpretability, increase sample size per category, and make results comparable to previous studies (*1, 2*).

All individual fossil occurrences were assigned to 2 My bins ranging through the entire Cenozoic. Taxon duration was measured as the number of 2 My bins from the first occurrence to the last occurrence, inclusive. This bin size was chosen because it approximately reflects the resolution of the North American Cenozoic mammal fossil record (*3–5*). Species originating in the youngest cohort, 0-2 My, were excluded from analysis because every species duration would be both left and right censored, which is illogical.

1.1.1 Body size

Species body size estimates in grams were sourced from a large selection of primary literature and database compilations. Databases used include the PBDB, PanTHERIA (*6*), and the Neogene Old World Mammal database (NOW; http://www.helsinki.fi/science/now/). Major sources of additional compiled body size estimates include (*7–12*). These were then supplemented with an additional literature search to try and fill in the remaining gaps. In many cases, species body mass was estimated using various published regression equations based on tooth or skull measurements (Table S3). If multiple specimens were measured, I used the mean of specimen measures as the species mean. The See Supplementary Table S4 for a complete list of individual measures and sources.

1.1.2 Biogeographic network

Species geographic extent was measured as the mean of the relative number of bioprovinces occupied by a species for each 2 My bin the species was present. Bioprovinces were identified using a network-theoretic approach that has previously been applied to paleontological data (*13, 14*). This approach relies on defining a biogeographic bipartite network of taxa and localities. In this study, taxa were defined as species and localities were grid cells from a regular lattice on a global equal-area cylinder map projection. The regular lattice was defined as a 70 x 34 global grid where each cell corresponds to approximately 250000 km^2. This network is considered bipartite because taxa are connected to localities based on their occurrence but taxa are not connected to taxa nor are localities connected to localities.

A biogeographic network was constructed for each of the 2 My bins used in this study. Emergent bioprovinces were then identified using the map equation (*15,16*) as has been done before (*13, 14, 17*). These bioprovinces correspond to taxa and localities that are more interconnected with each other than with other nodes.

The map projection and regular lattice were made using shape files from http://www.naturalearthdata.com/ and the `raster` package for R (*18*). The map equation and other network related analysis was done using the `igraph` package for R (*19*).

1.1.3 Supertree

As there is no single, combined formal phylogenetic hypothesis of all Cenozoic fossils mammals from North America, it was necessary to construct a semi-formal supertree. This was done by combining taxonomic information for all the observed species and a few published phylogenies.

The initial taxonomic classification of the observed species was based on the associated taxonomic information from the PBDB. This information was then updated using the Encyclopedia of Life (http://eol.org/) which collects and collates taxonomic information in a single database. This was done programatically using the `taxize` package for R (*20*). Finally, this taxonomic information was further updated using a published taxonomy of fossil mammals (*21, 22*).

This taxonomy serves as an initial phylogenetic hypothesis which was then combined with a selection of species-level phylogenies (*10, 23*) in order to better constrain a minimum estimate of the actual phylogenetic relationships of the species. The supertree was inferred via matrix representation parsimony implemented in the `phytools` package for R (*24*). While four most parsimonious trees were found, I selected a single of these for use in analysis.

Polytomies were resolved in order of species first appearance in order to minize stratigraphic gaps. The resulting tree was then time scaled using the `paleotree` package via the "minimum branch length" approach with a minimum length of 0.1 My (*25*). The minimum length is necessary to avoid zero-length branches which cause the phylogenetic covariance matrix not to be positive definite, which is important for computation (see below). While other time

scaling approaches are possible (*26, 27*) this method was chosen for its simplicity and not requiring additional information about diversification rates which are the interest of this study.

1.2 Survival model

Presented here is the model development process used to formulate the two survival models used in this study.

First, define y as a vector of length n where the ith element is the duration of species i, where $i = 1, \cdots, n$.

The simplest survival model where durations are assumed to follow an exponential distribution with a single "rate" or inverse-scale parameter λ (*28*). This is written out as

$$\begin{aligned} p(y|\lambda) &= \lambda \exp(-\lambda y) \\ y &\sim \text{Exp}(\lambda). \end{aligned} \tag{1}$$

The exponential distribution corresponds to situations where extinction risk is independent of age. To understand this, we need to define two functions: the survival function $S(t)$ and the hazard function $h(t)$.

$S(t)$ corresponds to the probability that a species having existed for t 2 My bins will not have gone extinct while $h(t)$ corresponds to the instantaneous extinction rate for some taxon age t (*28*). For an exponential model, $S(t)$ is defined

$$S(t) = \exp(-\lambda t) \tag{2}$$

and $h(t)$ is defined

$$h(t) = \lambda \tag{3}$$

The choice of the exponential distribution corresponds directly to the Law of Constant Extinction (*29*) as the right side of Eq. 3 does not depend on species age t.

The current sampling statement (Eq. 1) assumes that all species share the same rate parameter with no variation. To allow for variation in λ associated with relevant covariate information like species body size, λ is reparameterized as $\lambda_i = \exp(\sum \beta^T \mathbf{X}_i)$ with i indexing a given observation and its covariates, β is a vector of regression coefficients, and $\mathbf{X}$ is a matrix of covariates. This is a standard regression formulation, where one column of $\mathbf{X}$ is all 1-s and its corresponding β coefficient is the intercept. This approach is essentially a generalized linear model (GLM) approach where instead of normally distributed errors there are exponentially distributed errors (*28*).

To relax the assumption of age-independent extinction of the Law of Constant Extinction we substitute the Weibull distribution for the exponential (*28*). The Weibull distribution has a shape parameter α and scale parameter σ. Conceptually, σ is the inverse of λ. α modifies the impact of taxon age on extinction risk. When $\alpha > 1$ then $h(t)$ is a monotonically increasing function,

but when $\alpha < 1$ then $h(t)$ is a monotonically decreasing function. When $\alpha = 1$ then the Weibull distribution is equivalent to the exponential.

The Weibull distribution and sampling statement were defined

$$p(y|\alpha, \sigma) = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{y}{\sigma}\right)^{\alpha}\right)$$
$$y \sim \text{Weibull}(\alpha, \sigma). \tag{4}$$

The corresponding $S(t)$ and $h(t)$ functions are defined

$$S(t) = \exp\left(-\left(\frac{t}{\sigma}\right)^{\alpha}\right) \tag{5}$$

$$h(t) = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1}. \tag{6}$$

To allow for σ to vary with a given observation's covariate information it is reparameterized in a similar fashion to λ with a few key differences. Because $\sigma = 1/\lambda$ in order to preserve the interpretation of β, while taking α into account, σ is reparameterized as

$$\sigma_i = \exp\left(\frac{-(\beta_0}{\alpha}\right) \tag{7}$$

where β_0 is the intercept term.

The model described here was the final model at the end of a continuous model development framework where the sampling and prior distributions were iteratively modified to best reflect theory, knowledge of the data, the inclusion of important covariates, and the fit to the data. This follows the approach described in (*30*) and (*31*). A survival model was fit in a Bayesian context where species duration were assumed to be drawn from either an exponential or Wiebull distribution (Eq. 4) with shape α and scale σ parameters. α was assumed constant, which is standard practice in survival analysis (*28*). α was given a weakly informative half-Cauchy (C^+) prior. σ was reparameterized as an exponentiated regression model (Eq. 7). This was further expanded (Eq. 8) to allow for two hierarchical factors as discussed below. This is written

$$\sigma_i = \exp\left(\frac{-(h_i + \eta_{j[i]} + \sum \beta^T \mathbf{X}_i)}{\alpha}\right) \tag{8}$$

where equivalent statement for the exponential distribution is defined

$$\lambda_i = \exp\left(h_i + \eta_{j[i]} + \sum \beta^T \mathbf{X}_i)\right). \tag{9}$$

$\mathbf{X}$ is an $n \times K$ matrix of species-level covariates. Three of the covariates of interest are the logit of mean relative occupancy, and the logarithm of body size (g). The discrete covariate index variables of dietary and locomotor category were transformed into $n \times (k-1)$ matrices where each column is an indicator variable (0/1) for that species's category, k being the number of categories of

the index variable (3 and 4, respectively). Only $k-1$ indicator variables are necessary as the intercept takes on the remaining value. Finally, a vector of 1-s was included in the matrix $\mathbf{X}$ whose corresponding β coefficient is the intercept, making K equal eight.

β is the vector of regression coefficients. The intercept term was given a weak normal prior, $\beta_0 \sim \mathcal{N}(0, 10)$ while all of these other coefficients were slightly more informative priors, e.g. $\beta_{mass} \sim \mathcal{N}(0, 5)$. These priors were chosen because it is expected that the effect size of each variable on duration will be small, as is generally the case with binary covariates (*30*).

Regression coefficients are not directly comparable without first standardizing the input variables to have equal standard deviations. This is accomplished by subtracting the mean of the covariate from all values and then dividing by the standard deviation, resulting in a variable with mean of zero and a standard deviation of one. This linear transform greatly improves the interpretability of the coefficients as expected change in mean duration given a difference of one standard deviation in the covariate (*32*). Additionally, this makes the intercept directly interpretable as the estimate of mean (transformed) σ (Eq. 7). However, because the expected standard deviation for a random binary variable is 0.5, in order to make comparisons between the binary and continuous variables, the continuous inputs must instead be divided by twice their standard deviation (*33*).

1.2.1 Hierarchical effects

The two hierarchical effects of interest in this study are origination cohort and shared evolutionary history, or phylogeny. Hierarchical modeling can be considered an intermediate between complete and no pooling of groups (*30*), where complete pooling is when the differences between groups are ignored and no pooling is where different groups are analyzed separately. By allowing for partial pooling, we are modeling the appropriate compromise between these two extremes, allowing for better and potentially more informative overall inference. This is done by having all of the groups share the same normal prior with mean 0 and a scale parameter estimated from the data, which then acts as an indicator of the amount of pooling. A scale of 0 and ∞ indicate complete and no pooling, respectively. The choice of mean 0 allows for the individual group estimates to be interpreted as deviations from the intercept. Hierarchical modeling is analogous to mixed-effects modeling (*30*).

Origination cohort is defined as the group of species which all originated during the same 2 My temporal bin. Because the most recent temporal bin, 0-2 My, was excluded, there are 32 different cohorts. The effect of origination cohort j was modeled with each group being a sample from a common cohort effect, η, which was considered normally distributed with mean 0, and standard deviation σ_c. The value of σ_c was then estimated from the data itself, corresponding to the amount of pooling in the individual estimates of η_j. This approach is a conceptual and statistical unification between dynamic and cohort survival analysis in paleontology (*34–38*), with σ_c acting as a measure of compromise between these two end members.

$$\eta_j \sim \mathcal{N}(0, \sigma_c)$$
$$\sigma_c \sim \mathrm{C}^+(0, 2.5)$$

The choice of the half-Cauchy prior on σ_c follows (*39*).

The impact of shared evolutionary history, or phylogeny, was modeling as an individual effect where each observation, i, is modeled as a multivariate normal, h, where the covariance matrix Σ is known up to a constant, σ_p^2 (*40, 41*). This is written

$$h \sim \text{multivariate } \mathcal{N}(0, \mathbf{\Sigma})$$
$$\mathbf{\Sigma} = \sigma_p^2 \mathbf{V}_{phy}$$
$$\sigma_p \sim \mathrm{C}^+(0, 2.5).$$

$\mathbf{V}_{phy}$ is the phylogenetic covariance matrix defined as an $n \times n$ matrix where the diagonal elements are the distance from root to tip, in branch length, for each observation and the off-diagonal elements are the amount of shared history, measured in branch length, between observations i and j. σ_p was given a weakly informative half-Cauchy hyperprior. Note that because the phylogeny used here is primarily based on taxonomy, estimates of σ_p represent minimum estimates (*40, 41*). Improved phylogenetic estimates of all fossil Cenozoic mammals would greatly improve this estimate.

1.2.2 Censored observations

An important part of survival analysis is the inclusion of censored observations where the failure time has not been observed (*28, 42*). The most common censored observation is right censored, where the point of extinction had not yet been observed in the period of study, such as taxa that are still present in the most recent time bin (0-2 My). Left censored observations, on the other hand, correspond to observations that went extinct any time between 0 and some known point. In order to account for the minimum resolution of the fossil record encountered here, taxa that occurred in only a single time bin were left censored.

Censored data are modeled using the survival function of the distribution, $S(t)$, defined earlier for the Weibull distribution (Eq. 5) with σ defined as above (Eq. 8). $S(t)$ is the probability that an observation will survive longer than a given time t. The likelihood of uncensored observations is evaluated as normal using Equation 4 while right censored observations are evaluated at $S(t)$ and left censored observations are evaluated at $1 - S(t)$. Note, $1 - S(t)$ is equivalent to the cumulative distribution function and $S(t)$ is equivalent to the complementary cumulative distribution function (*31*).

The full likelihood for both uncensored and both right and left censored observations is written

$$L \propto \prod_{i \in C} \text{Weibull}(y_i|\alpha,\sigma) \prod_{j \in R} S(y_j|\alpha,\sigma) \prod_{k \in L} \left(1 - S(y_k|\alpha,\sigma)\right),$$

where C is the set of uncensored observations, R is the set of right censored observations, and L is the set of left censored observations.

1.2.3 Estimation

Parameter posteriors were approximated using a Markov-chain Monte Carlo (MCMC) routine implemented in the Stan programming language (*43*). Stan implements a Hamiltonian Monte Carlo using a No-U-Turn sampler (*44*). Posterior approximation was done using four parallel MCMC chains run for 30000 steps, thinned to every thirtieth sample, split evenly between warm-up and sampling. Convergence was evaluated using the scale reduction factor, $\hat{R}$. Values of $\hat{R}$ close to 1, or less than or equal to 1.1, indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed (*31*).

In order to speed up the posterior approximation, a custom multivariate normal sampler was used to estimate the unknown constant term in the covariance matrix. This is necessary because inverting and solving the complete covariance matrix on every iteration is a memory intense procedure. The custom sampler limits the necessary number of operations and matrix inversions per iteration.

1.3 Posterior predictive checks

The most basic assessment of model fit is that simulated data generated using the fitted model should be similar to the observed. This is the idea behind posterior predictive checks. Using the covariates from each of the observed durations, and randomly drawn parameter estimates from their marginal posteriors, a simulated data set y^{rep} was generated. This process was repeated 1000 times and the distribution of y^{rep} was compared with the observed (*31*).

An example posterior predictive check used in this study is a graphical comparison between the Kaplan-Meier (K-M) survival curve estimated from the observed data and the K-M survival curves estimated from 1000 simulation sets. K-M survival curves are non-parametric estimates of $S(t)$ (*28*). Other posterior predictive checks used here include comparison of the mean and quantiles of the observed durations to the distributions of the same quantities from the simulations, and inspection of the deviance residuals, defined below.

In standard linear regression, residuals are defined as $r_i = y_i - y_i^{est}$. For the model used here, this definition is inadequate. The equivalent values for survival analysis are deviance residuals. To define how deviance residuals are calculated, we first define the cumulative hazard function (*28*). Given $S(t)$ (Eq. 5), we define the cumulative hazard function as

$$\Lambda(t) = -log\left(S\left(t\right)\right).$$

Next, we define martingale residuals m as

$$m_i = I_i - \Lambda(t_i).$$

I is the inclusion vector of length n, where $I_i = 1$ means the observation is completely observed and $I_i = 0$ means the observation is censored. Martingale residuals have a mean of 0, range between 1 and $-\infty$, and can be viewed as the difference between the observed number of deaths between 0 and t_i and the expected number of deaths based on the model. However, martingale residuals are asymmetrically distributed, and can not be interpreted in the same manner as standard residuals.

The solution to this is to use the deviance residuals, D. This is defined as a function of martingale residuals and takes the form

$$D_i = \text{sign}(m_i)\sqrt{-2[m_i + I_i log(I_i - m_i)]}.$$

Deviance residuals have a mean of 0 and a standard deviation of 1 by definition.

1.4 Variance partitioning

There are three different variance components in this model: sample σ_y^2, cohort σ_c^2, and phylogenetic σ_p^2. The sample variance, σ_y^2, is similar to the residual variance from a normal linear regression. Partitioning the variance between these sources allows the relative amount of unexplained variance of the sample to be compared. However, the Weibull based model used here (Eq. 4) does not include an estimate of the sample variance, σ_y^2. Partitioning the variance between these three components was approximated via a simulation approach modified from (*45*).

The procedure is as follows:

1. Simulate w (50,000) values of η; $\eta \sim \mathcal{N}(0, \sigma_c)$.
2. For a given value of $\beta^T\mathbf{X}$, calculate σ^{c*} (Eq. 7) for all w simulations, holding h constant at 0.
3. Calculate υ_c, the Weibull variance (Eq. 10) of each element of σ^{c*} with α drawn from the posterior estimate.
4. Simulate w values of h; $h \sim \mathcal{N}(0, \sigma_p)$.
5. For a given value of $\beta^T\mathbf{X}$, calculate σ^{p*} (Eq. 7) for all w simulations, holding η constant at 0.
6. Calculate υ_p, the Weibull variance (Eq. 10) of each element of σ^{p*} with α drawn from the posterior estimate.
7. $\sigma_{y*}^2 = \frac{1}{2}\left(\left(\frac{1}{w}\sum_i^w \upsilon_{pi}\right) + \left(\frac{1}{w}\sum_j^w \upsilon_{cj}\right)\right)$.
8. $\sigma_{c*}^2 = var(\upsilon_c)$ and $\sigma_{p*}^2 = var(\upsilon_p)$.

The simulated values of h were drawn from a univariate normal distribution because each simulated value is in isolation, so there is no concern of phylogenetic autocorrelation. The chosen value for $\beta^T \mathbf{X}$ was a draw from the posterior estimate of the intercept. Because input variables were standardized prior to model fitting, the intercept corresponds to the estimated effect on survival of the sample mean.

Weibull variance is calculated as

$$var(x) = \sigma^2 \left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^2 \right), \tag{10}$$

where Γ is the gamma function.

The variance partitioning coefficients are then calculated, for example, as $VPC_{phylo} = \frac{\sigma^2_{p*}}{\sigma^2_{y*}+\sigma^2_{c*}+\sigma^2_{p*}}$ and similarly for the other components.

I used variance partitioning coefficients (VPC) to estimate the relative importance of the different variance components (*30*). Phylogenetic heritability, h^2_p (*40, 41*), is identical to the VPC of the phylogenetic effect. Additionally, because phylogenetic effect was estimated using a principally taxonomy based tree the estimates derived here can be considered minimum estimates of the phylogenetic effect.

2 Results of posterior predictive checks

With all marginal posterior estimates having converged ($\hat{R} < 1.1$) it is possible to examine the quality of model fit (Table S1). If the model is an adequate descriptor of the observed data, then relatively confident inference can be made (*31*).

Visual examination of the deviance residuals from twelve different sets of posterior predictive simulations indicates few systematic problems (Fig. S1). The only concern is that the residuals are slightly skewed,however this bias appears very small. This is confirmed by comparing the K-M estimate of the empirical survival function to 100 estimated survival functions from posterior simulations (Fig. 1, main text).

Comparisons of the observed 25th, 50th, and 75th quantiles, and mean durations to the results from the posterior predictive simulations indicate adequate model fit (Fig. S2). Because all the different posterior predictive checks seem to agree, the inferred model appears adequate at capturing the observed variation.

3 Concerns surrounding estimates of α

The estimate of the Weibull shape parameter, α, is greater than 1 meaning that extinction risk is expected to increase with taxon age (Table S1). As the value of α is between 1 and 1.5, extinction risk for a given species only gradually increases with age (Fig. S3). There are three possible explanations for this result:

1) older taxa being out competed by younger taxa (*46*); or 2) this is an artifact of the minimum resolution of the fossil record (*47*).

An additional concern is that there may be an upward bias in estimates of α at this sample size, similar to that for scale parameters (*31*). The plausibility of third possibility in this example can be explored in simulation. I simulated from 10, 100, 1000, and 10000 samples from a Weibull($alpha = 1.3$, $\sigma = 1$) 100 times each. For each of these simulated datasets, I then estimated the values of α and σ in a Bayesian context, giving each of them weak C^+ distributions (e.g. $\alpha \sim C^+(2.5)$). The means of the marginal posterior estimate for the simulated datasets were then compared to the known values. The results from these simulations demonstrate that the estimates of α in the above analyses (Table S1) should not be overly biased based on my sample size of approximately 2000 species.

The model used in this analysis, however, is unable to distinguish between the remaining two hypotheses (*46, 47*). Further work on how to better constrain estimates α is necessary. A possibly is somehow incoporating these hypotheses as prior information.

4 Supplementary figures

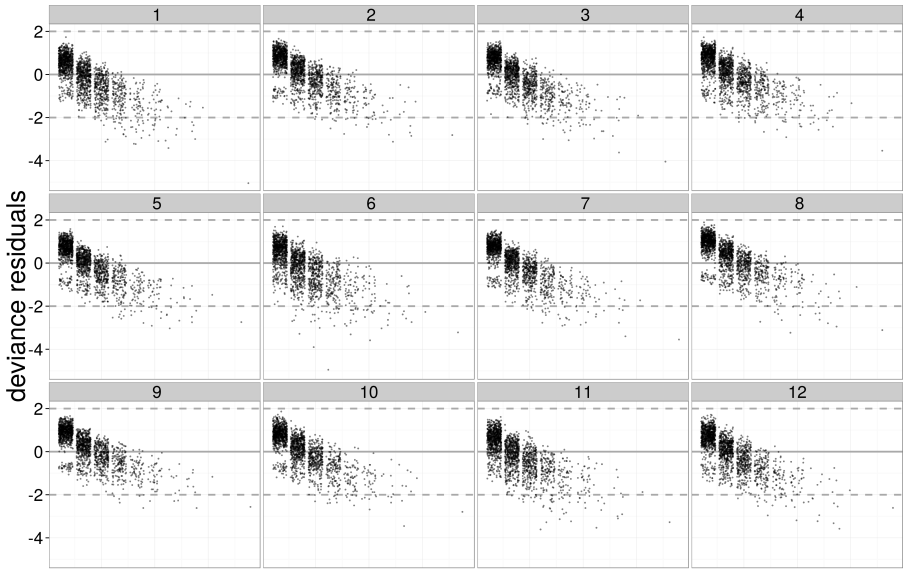

Figure S1: Deviance residuals from the fitted survival model. Each graph depicts the residuals from single draws from the posterior distributions of all estimated parameters. Positive values indicate an underestimate of the observed duration, while negative values indicate an overestimate of the observed duration. Twelve different examples are provided here to indicate the lack of individual observation based biases.

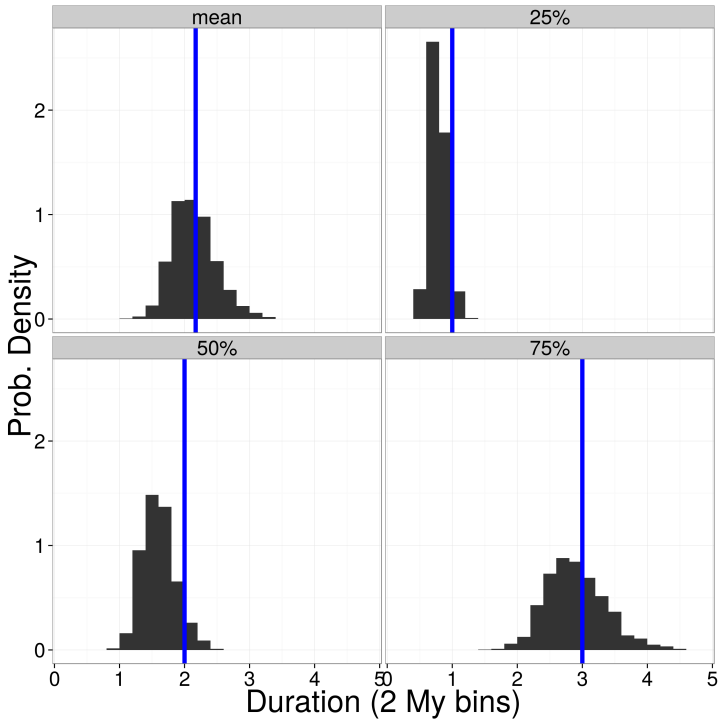

Figure S2: The results of additional posterior predictive checks for four summaries of the observed durations, as labeled. Blue vertical lines indicate the observed value. None of the observed values are significantly different from the posterior predictive distributions.

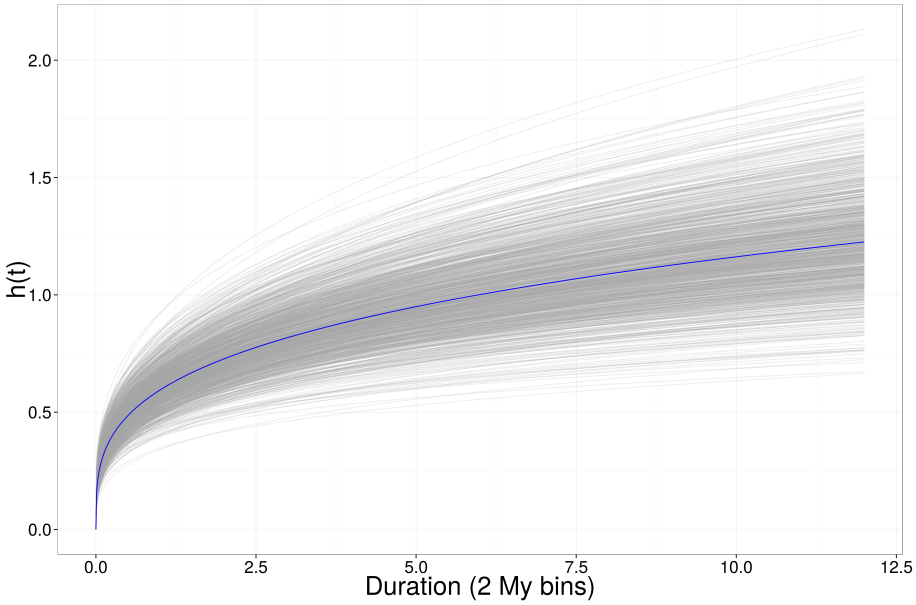

Figure S3: 100 estimates of the hazard function ($h(t)$ for the observed species mean (grey), along with the median estimated hazard function. $h(t)$ is an estimate of the rate at which a species of age t is expected to go extinct. Hazard functions were estimated from random draws from the estimated posterior distributions and evaluated with all covariate information set to 0, which corresponds to the expected duration of the mean species.

5 Supplementary tables

Table S1: Marginal posterior estimates for the praameters of interested based on 1000 posterior samples. The intercept can be interpreted as the estimate for the mean observed species. The other values are the effect of a trait on the expected species duration as expressed as deviation from the mean. The categorical variables are binary index variables where an observation is of that category or not. $\hat{R}$ values of less than 1.1 indicate approximate chain convergence for the posterior samples.

	mean	sd	2.5%	25%	50%	75%	97.5%	$\hat{R}$
alpha	1.29	0.03	1.23	1.27	1.29	1.31	1.36	1.00
intercept	-0.78	0.14	-1.05	-0.87	-0.78	-0.68	-0.51	1.00
logit(occupancy)	-0.53	0.08	-0.69	-0.59	-0.53	-0.48	-0.38	1.00
log(size)	-0.05	0.05	-0.14	-0.08	-0.05	-0.01	0.05	1.00
ground dwelling	-0.28	0.10	-0.47	-0.34	-0.28	-0.21	-0.09	1.00
scansorial	-0.22	0.11	-0.43	-0.29	-0.22	-0.14	-0.00	1.00
herbivore	0.09	0.09	-0.09	0.03	0.09	0.14	0.27	1.00
insectivore	0.10	0.11	-0.11	0.03	0.10	0.17	0.31	1.00
omnivore	-0.12	0.11	-0.33	-0.19	-0.12	-0.05	0.09	1.00
sd cohort	0.33	0.06	0.23	0.29	0.33	0.37	0.48	1.00
sd phylogeny	0.11	0.05	0.03	0.07	0.10	0.14	0.23	1.03

Table S2: Species trait assignments in this study are a coarser version of the information available in the PBDB. Information was coarsened to improve per category sample size and uniformity and followed this table.

This study		PBDB categories
Diet	Carnivore	Carnivore
	Herbivore	Browser, folivore, granivore, grazer, herbivore.
	Insectivore	Insectivore.
	Omnivore	Frugivore, omnivore.
Locomotor	Arboreal	Arboreal.
	Ground dwelling	Fossorial, ground dwelling, semifossorial, saltatorial.
	Scansorial	Scansorial.

Group	Equation	log(Measurement)	Source
General	$\log(m) = 1.827x + 1.81$	lower m1 area	(*48*)
General	$\log(m) = 2.9677x - 5.6712$	mandible length	(*49*)
General	$\log(m) = 3.68x - 3.83$	skull length	(*50*)
Carnivores	$\log(m) = 2.97x + 1.681$	lower m1 length	(*51*)
Insectivores	$\log(m) = 1.628x + 1.726$	lower m1 area	(*52*)
Insectivores	$\log(m) = 1.714x + 0.886$	upper M1 area	(*52*)
Lagomorph	$\log(m) = 2.671x - 2.671$	lower toothrow area	(*12*)
Lagomorph	$\log(m) = 4.468x - 3.002$	lower m1 length	(*12*)
Marsupials	$\log(m) = 3.284x + 1.83$	upper M1 length	(*53*)
Marsupials	$\log(m) = 1.733x + 1.571$	upper M1 area	(*53*)
Rodentia	$\log(m) = 1.767x + 2.172$	lower m1 area	(*48*)
Ungulates	$\log(m) = 1.516x + 3.757$	lower m1 area	(*54*)
Ungulates	$\log(m) = 3.076x + 2.366$	lower m2 length	(*54*)
Ungulates	$\log(m) = 1.518x + 2.792$	lower m2 area	(*54*)
Ungulates	$\log(m) = 3.113x - 1.374$	lower toothrow length	(*54*)

Table S3: CAPTION

References

1. J. Jernvall, M. Fortelius, *American Naturalist* **164**, 614 (2004).
2. S. A. Price, S. S. B. Hopkins, K. K. Smith, V. L. Roth, *Proceedings of the National Academy of Sciences* **109**, 7008 (2012).
3. J. Alroy, *Speciation and patterns of diversity*, R. K. Butlin, J. R. Bridle, D. Schluter, eds. (Cambridge University Press, Cambridge, 2009), pp. 302–323.
4. J. Alroy, P. L. Koch, J. C. Zachos, *Paleobiology* **26**, 259 (2000).
5. J. D. Marcot, *Paleobiology* **40**, 237 (2014).
6. K. E. Jones, *et al.*, *Ecology* **90**, 2648 (2009).
7. B. W. Brook, D. M. J. S. Bowman, *Journal of Biogeography* **31**, 517 (2004).
8. M. Freudenthal, E. Martín-suárez, *Scripta Geologica* **145**, 1 (2013).
9. R. T. McKenna, Potential for Speciation in Mammals Following Vast , Late Miocene Volcanic Interruptions in the Pacific Northwest, Masters, Portland State University (2011).
10. P. Raia, F. Carotenuto, F. Passaro, D. Fulgione, M. Fortelius, *The American naturalist* **179**, 328 (2012).
11. F. A. Smith, *et al.*, *The American Naturalist* **163**, 672 (2004).
12. S. Tomiya, *The American Naturalist* **182**, 196 (2013).
13. C. A. Sidor, *et al.*, *Proceedings of the National Academy of Sciences* **110**, 8129 (2013).
14. D. A. Vilhena, *et al.*, *Scientific Reports* **3**, 1790 (2013).
15. M. Rosvall, C. T. Bergstrom, *Proceedings of the National Academy of Sciences* **105**, 1118 (2008).
16. M. Rosvall, D. Axelsson, C. Bergstrom, *The European Physical Journal Special Topics* **178**, 13 (2009).
17. D. A. Vilhena, Boundaries and dynamics of biomes, Ph.D. thesis, University of Washington (2013).
18. R. J. Hijmans, *raster: Geographic data analysis and modeling* (2015). R package version 2.3-24.
19. G. Csardi, T. Nepusz, *InterJournal* **Complex Systems**, 1695 (2006).
20. S. Chamberlain, E. Szocs, *F1000Research* (2013).

21. C. M. Janis, G. F. Gunnell, M. D. Uhen, *Evolution of Tertiary mammals of North America. Vol. 2. Small mammals, xenarthrans, and marine mammals* (Cambridge University Press, Cambridge, 2008).

22. C. M. Janis, K. M. Scott, L. L. Jacobs, *Evolution of Tertiary mammals of North America. Vol. 1. Terrestrial carnivores, ungulates, and ungulatelike mammals* (Cambridge University Press, Cambridge, 1998).

23. O. R. P. Bininda-Emonds, *et al.*, *Nature* **446**, 507 (2007).

24. L. J. Revell, *Methods in Ecology and Evolution* **3**, 217 (2012).

25. D. W. Bapst, *Methods in Ecology and Evolution* **3**, 803 (2012).

26. D. W. Bapst, *Methods in Ecology and Evolution* **4**, 724 (2013).

27. M. M. Hedman, *Paleobiology* **36**, 16 (2010).

28. J. P. Klein, M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data* (Springer, New York, 2003), second edn.

29. L. Van Valen, *Evolutionary Theory* **1**, 1 (1973).

30. A. Gelman, J. Hill, *Data Analysis using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, New York, NY, 2007).

31. A. Gelman, *et al.*, *Bayesian data analysis* (Chapman and Hall, Boca Raton, FL, 2013), third edn.

32. H. Schielzeth, *Methods in Ecology and Evolution* **1**, 103 (2010).

33. A. Gelman, *Statistics in Medicine* pp. 2865–2873 (2008).

34. M. Foote, *Paleobiology* **14**, 258 (1988).

35. D. M. Raup, *Paleobiology* **4**, 1 (1978).

36. D. M. Raup, *Paleobiology* **1**, 82 (1975).

37. L. Van Valen, *Evolutionary Theory* **4**, 129 (1979).

38. T. K. Baumiller, *Paleobiology* **19**, 304 (1993).

39. A. Gelman, *Bayesian Analysis* **1**, 515 (2006).

40. M. Lynch, *Evolution* **45**, 1065 (1991).

41. E. A. Housworth, P. Martins, M. Lynch, *The American Naturalist* **163**, 84 (2004).

42. J. G. Ibrahim, M.-H. Chen, D. Sinha, *Bayesian Survival Analysis* (Springer, New York, 2001).

43. Stan Development Team, Stan: A c++ library for probability and sampling, version 2.5.0 (2014).

44. M. D. Hoffman, A. Gelman, *arXiv* **1111** (2011).

45. H. Goldstein, W. Browne, J. Rasbash, *Understanding Statistics* **1**, 1 (2002).

46. P. J. Wagner, G. F. Estabrook, *Proceedings of the National Academy of Sciences* **111**, 16419 (2014).

47. J. J. Sepkoski, *Paleobiology* **1**, 343 (1975).

48. S. Legendre, *Paleovertebrata* **16**, 191 (1986).

49. J. R. Foster, *PaleoBios* **28**, 114 (2009).

50. Z.-X. Luo, A. W. Crompton, A.-L. Sun, *Science* **292**, 1535 (2001).

51. B. Van Valkenburgh, *Body size in mammalian palebiology: estimation and biological implications*, J. Damuth, B. J. Macfadden, eds. (Cambridge University Press, Cambridge, 1990), pp. 181–205.

52. J. I. Bloch, K. D. Rose, P. D. Gingerich, *Journal of Mammalogy* **79**, 804 (1998).

53. C. L. Gordon, *Journal of Mammalian Evolution* p. 21 (2003).

54. M. Mendoza, C. M. Janis, P. Palmqvist, *Journal of Zoology* **270**, 90 (2006).