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.

Table 1: 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.

1 Methods

1.1 Species 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.

Species dietary and locomotor category assignments were done using the assignments in initial the PBDB which were then reasigned into coarser categories (Table **??**). This was done to improve interpretability, increase sample size per category, and make these results comparable to previous studies [**?** **?**].

Fossil occurrences were assigned to 2 My bins ranging through the entire Cenozoic. Taxon duration was measured as the number of bins from the first bin of occurrence to the last bin of occurrence, inclusive.

Species body size estimates were sourced from a large selection of primary literature and compilations, principally the PBDB, PanTHERIA [**?**], the Neogene Old World Mammal database (Now; `http://www.helsinki.fi/science/now/`), and other large scale data collection efforts [**?** **?** **?** **?** **?**]. In many cases, species body mass was estimated from anatomical dimensions such as tooth size. These estimates were made using a variety of published regression equations. See Appendix: Data for a complete list of individual sources and equations.

1.1.1 Bioprovince occupancy

For each 2 My time bin, a bipartite biogeographic network was created between species occurrences and spatial units. Spatial units were defined was 2x2 latitude–longitude grid cells from an azimuthal equal-area map projection. In these bipartite networks, taxa can only be linked to localities and *vice versa*. Taxa are not linked to each other, nor are localities linked. Emergent bioprovinces within the biogeographic occurrence network were identified using the map equation [**?** **?**]. A bioprovince is a set of species–locality connections that are more interconnected within the group than without. This was done for each bin's biogeographic network using the `igraph` package for R [**?** **?**]. The relative number of bioprovinces occupied per time bin was then determined for each species.

1.1.2 Semi-formal supertree

Because there exists no 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 taxonomic information from the PBDB served as the basis for additional revision. The taxonomy of many species was 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 [**?**]. This was additionally correct using various published phylogenies and taxonomies of fossil mammals [**?** **?** **?**], producing a tree that was a series of nested polytomies.

Polytomies were resolved resolved with respect to the order of their appearance. 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 [**?**]. The minimum length is necessary to avoid zero-length branches which cause the phylogenetic covariance matrix not be positive definite, which is an imporant covenience for compuation (see below). While other time scaling approaches are possible [**?** **?**] this method was chosen for it's simplicity and not requiring additional information about diversification rates which are of interest in this study.

1.2 Survival model

I implemented a fully Bayesian model of taxon durations. For the sampling distribution or likelihood, I assumed a parametric survival model where the observations followed a Weibull distribution (Eq. **??**) with shape α and scale σ defined as in a regression model with $\mathbf{X}$ being a matrix of predictor variables (Eq. **??**).

$$
\begin{aligned}
p(y_i|\alpha,\sigma) &= \text{Weibull}(y_i|\alpha,\sigma) \\
&= \frac{\alpha}{\sigma}\left(\frac{y_i}{\sigma}\right)^{\alpha-1}\exp\left(-\left(\frac{y_i}{\sigma}\right)^{\alpha}\right) \qquad (1) \\
\sigma &= \exp\left(\frac{-(h_i + r_{j[i]} + \sum \beta^T \mathbf{X}_i)}{\alpha}\right) \qquad (2)
\end{aligned}
$$

These predictors are as follows. Log of mean occupancy and log body size (g) were used as continuous predictors. For modeling discrete predictors in a regression model, the vector of states is transformed into a $n \times (k-1)$ matrix where each column is a series of 1's and 0's indicating the observed's category, k being the number of categories. These are sometimes called 'dummy variables'. Only $k-1$ columns are necessary as the intercept takes on the remaining value. This was done for dietary and locomotor category. In total, this is 5 binary predictors. Finally, a 1 was included in the predictor matrix $\mathbf{X}$, the corresponding β coefficient being the intercept.

All β coefficients were given a different, diffuse informative normally distributed prior. 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 predictors. In all cases, posterior inference was not effected by changes to this choice of prior.

Note that regression coefficients (β-s) are not directly comparable without standardizing input variables to have equal standard deviations. This linear transform allows for the coefficients to indicate expected change in duration given change per one standard deviation of the covariate [**?**]. However, because the expected standard deviation for a binary variable is 0.5, in order to make comparisons between the binary and continuous variables, the continuous inputs must be divided by twice their standard deviation [**?**]. The above model was fit with both unstandardardized and standardized inputs.

The impact of origination cohort (j) was included as a random effect η in the parameterization of σ. Origination cohort is defined as the taxa which all originated during the same temporal bin. The most recent temporal bin, 0-2 Mya, was excluded, leaving 32 different cohorts. Each cohort was considered an exchangeable sample of a shared general "cohort effect." The individual cohort effects were estimated in a hierarchical framework where the between cohort variance constrained the individual cohort effect estimate. This is done by giving η a normally distributed prior centered at 0 with scale τ which is then estimated from the data. τ is given a diffuse half-Cauchy hyperprior following **?**].

The impact of shared evolutionary history, or phylogeny, was also included as an multivariate normal random effect h. The covariance matrix of h was assumed known, up to a constant υ, from the phylogenetic covariance matrix ($\mathbf{V}_{phy}$) [**? ?**]. $\mathbf{V}_{phy}$ was calculated as the amount of shared branch length between taxa. The constant υ was given a diffuse half-Cauchy prior.

The mixed effects approach used here for estimating phylogenetic effect partitions variance into phylogenetic and nonphylogenetic components. The percent of total variance described by

phylogeny is called the phylogenetic heritability, h^2_{phy} [**?**], which was calculated as $h^2_{phy} = \frac{\upsilon}{\upsilon+\sigma_e}$. σ_e is the variance of the residuals as calculated from the posterior predictive checks (below) and is the nonphylogenetic changes associated uniquely with a species.

The shape parameter α was assumed constant, as is standard practice in survival analysis, and was given a diffuse half-Cauchy prior.

An important part of survival analysis is the inclusion of "censored" observations [**?** **?**]. These are observations where the failure time has not occurred during the period of interest. In this way both duration and event information are modeled simultaneously, as opposed to just duration or event as in linear and logistic regression, respectively.

The most common censored observation is right censored, where the point of extinction had not yet been observed in the period of study. In this case, this means taxa that are still extant. For each right censored observation, the log probability is incremented by the complementary cumulative density function evaluated at the observed duration.

$$P(T > t) = 1 - F(t|\theta) \tag{3}$$

Left censored observations, on the other hand, correspond to observations that went extinct any time between 0 and some known point. In this study, taxa occurring in only a single time bin were left censored. Because of the minimum resolution of the record, we cannot observe if these taxa went extinct in less than that single bin or not. For each left censored observation, the log probability is incremented by the cumulative density function evaluated at the observed duration.

$$P(T < t) = F(t|\theta) \tag{4}$$

Below is a summary of the priors used for each estimated parameter

$$\begin{aligned}
\tau &\sim \text{half Cauchy}(2.5) \\
\eta &\sim \text{Normal}(0, \tau) \\
\upsilon &\sim \text{half Cauchy}(2.5) \\
h &\sim \text{MultiNormal}(0, \upsilon \times \mathbf{V}_{phy}) \\
\beta &\sim \text{Normal}(0, 10) \\
\alpha &\sim \text{half Cauchy}(2.5)
\end{aligned}$$

The parameter posteriors were approximated using a Markov-chain Monte Carlo (MCMC) routine implemented in the Stan programming language [**?**]. Stan implements a Hamiltonian Monte Carlo using a No-U-Turn sampler [**?**]. Posterior approximation was done using four parallel MCMC chains. Chain 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 [**?**].

Both models with and without phylogenetic effects were estimated. Because inverting a large matrix is a memory intense procedure and because the phylogenetic covariance matrix is only assumed known up to a constant, every iteration of the MCMC would involve solving a very large matrix which is not ideal. In order to speed up the MCMC routine, this aspect of the model had to be reparameterized for efficiency purposes. One way of doing this is by using a Cholesky parameterized version of the multivariate normal distribution with a Cholesky decomposed covariance matrix, however because of the size of the covariance matrix this speed up is only minor. Instead, a custom multivariate sampler was used (see Appendix: Code).

For the model without phylogenetic effect the four MCMC chains ran for 2000 steps, with the first 1000 used as warm-up and the last 1000 as samples from the posterior. Because of the added complexity of estimating the phylogenetic effect, all four chains were run 10000 steps thinned to every tenth sample split evenly between warm-up and sampling.

1.2.1 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 predictors 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 repeated 1000 times and the distribution of y^{rep} was compared with the observed y [**?**]. This was done both graphically and numerically.

An example posterior predictive check used in this study is a graphical comparison of the Kaplan-Meier (K-M) survival curve estimated from the observed data with the survival curves from 1000 simulations. K-M survival curves are non-parametric estimates of the function $S(t)$ or the probability of a species going extinct given that it has lived to time t [**?**]. Other posterior predictive checks included comparison of the mean and quantiles of the observed durations to the distributions of the same quantities from the simulations.

The model described above 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 to fit the data.