**Phylogenetic Clustering of Origination and Extinction in North American Mammals through the Cenozoic**

GREGORY JAMES SMITH\*, SIMON A. F. DARROCH

*Gregory James Smith [gregory.j.smith@vanderbilt.edu] Department of Earth and*

*Environmental Science, Vanderbilt University, Nashville, Tennessee 37235*

*Simon A.F. Darroch [simon.a.darroch @vanderbilt.edu] Department of Earth and Environmental Science, Vanderbilt University, Nashville, Tennessee 37235*

\*corresponding author

**Abstract**

**Introduction**

The causes that underpin extinction events in the past are a topic of interest to paleontologists. Not all extinctions result from the same causes. For example, organisms can face pressures from a variety of sources, including abiotic (climate, environmental) or biotic (competition, disease) stressors. Assuming organisms who share a recent common ancestor also have similar ecologic tolerances, climatic stressors that lead to extinctions would preferentially eliminate organisms that are closely related. On the other hand, organisms that are not closely related are unlikely to have similar ecologic tolerances; thus, biotic stressors (such as competition) are likely to cause extinction without discrimination, and extinctions should be random.

To investigate the extent to which mammalian extinctions and originations were clustered throughout the Cenozoic, we investigated the manner in which extinction and origination events were correlated with taxonomic relatedness in the past. Specifically, we used mammalian occurrence data from the paleobiology database (<https://paleobiodb.org/>) to address the following questions: 1) To what extent have recent (i.e., Cenozoic) extinctions structured the composition of the modern biosphere?, 2) What is the relationship between extinction magnitude and phylogenetic clustering in the Cenozoic, and, 3) How did different extinction drivers compare with degree of phylogenetic clustering in the Cenozoic?

**Methods**

We downloaded fossil occurrence data for the late Cretaceous and Cenozoic from the Paleobiology Database (<https://paleobiodb.org/>) in April 2018, using the taxon name ‘mammalia’ and the parameters: time interval range = 70 Ma to 0 Ma , continent = North America. We removed any occurrences of aquatic organisms (Orders Sirenia, Cetacea and Desmostylia) as well as those occurrences with uncertain genus or family IDs. These occurrence data were used to define the first occurrence datum (FAD), last occurrence datum (LAD) and longevity for each mammalian genus. We then partitioned data by the North American Land Mammal Ages (NALMAs; see Supporting information, Table S1) because they delineate relatively stable community assemblages; the dates for all NALMAs are based on Woodburne (2004). We do not assign taxa to the newly-defined Santarosean or Saintaugustinean NALMAs (Barnosky et al. 2014) and have kept the minimum age of the Rancholabrean as 0.0117 Ma because the interval boundaries used in the Paleobiology Database do not yet follow these definitions.

For each NALMA, we used the FAD and LAD data to identify the four fundamental classes of taxa as defined by Foote (2000). Briefly summarized here, the classes are: (1) taxa confined to the interval, i.e., taxa whose FAD and LAD are both within the interval; (2) taxa that exist prior to the interval and make their last appearance during the interval; (3) taxa that make their first appearance during the interval and exist into the subsequent interval; and (4) taxa that range through the entire interval and exist in both the previous and subsequent interval (Figure 1). We use the same terminology as Foote (2000) in identifying the numbers of each of these taxa (NFL, NbL, Nbt, and NFt, respectively), including the total number of taxa crossing the bottom of an interval (Nb = NbL + Nbt) and the top of the interval (Nt = NFt + Nbt), and use these values to calculate the number of originations (No = NFL+ NFt) and extinctions (Ne = NFL + NbL) in each interval. The per-capita rate of origination (*p*) and extinction (*q*) for the duration of each interval (Δt) are defined as:

In addition to calculating origination and extinction rates for each NALMA, we computed two indices of phylogenetic clustering. The first index, RCL, is Pearson’s correlation coefficient between two matrices. These are the same values as those first calculated in Roy et al. (2009), and we follow Krug and Patzkowsky (2015) in calculating RCL values of extinction and origination for each interval. Each of the matrices used in the calculation of RCL are *n* by *n* binary matrices (with the lower triangles, including the diagonals, removed), where *n* is equal to the number of genera that appear in the NALMA. In the matrix MTAX, the cross-product (i.e., the cell at the intersection of a row and a column) takes on a value of 1 when both genera belong to the same taxonomic family, and 0 otherwise. In MEXT, the cross-product takes on a value of 1 when both genera go extinct at the end of the interval, and a 0 otherwise; for MORIG, the cross-product takes on a value of 1 when both genera originate at the beginning of the interval, and a 0 otherwise. RCL for extinctions is the correlation between MTAX and MEXT, while RCL for originations is the correlation between MTAX and MORIG. As RCL is a correlation coefficient, the value can take on any value from +1 to –1; a value of +1 suggests extinctions or originations are perfectly clustered (i.e., genera originating/going extinct are more closely-related than would be expected by chance), a value of –1 suggests extinctions or originations are perfectly even (i.e., the numbers of genera originating/going extinct are perfectly dispersed amongst families), and a value of 0 suggests extinctions or originations are perfectly random. To assess the significance of these results, we kept the number of extinctions or originations in each interval the same, randomized the specific genera originating or going extinct, and recomputed RCL 5000 times to generate a null distribution. We then plotted the 95% confidence limits of this null distribution to identify excursions of RCL above or below the range of this null model, representing significantly clustered or even extinction and origination events, respectively.

We also computed Moran’s I (here, as an index of clustering, ICL), a metric used in assessing patterns of spatial clustering and evenness (Moran 1950). This metric has been used to estimate taxonomic patterns of extinction risk in extant vertebrates (Gittleman and Kot 1990; Lockwood et al. 2002) and was also computed by Roy et al. (2009) when assessing phylogenetic patterns of extinction in bivalves since the late Jurassic. Our calculation of ICL uses the same matrices as used in the calculation of RCL, but rather than calculated as a correlation coefficient, the calculation is:

Where wij is equivalent to MTAX and zi and zj refer to the normalized row and column sums (n = number of rows/columns) of MEXT or MORIG, respectively. Like RCL, ICL takes on more positive values when extinctions or originations are clustered, negative values when they are even, and a value close to 0 when extinctions or originations are random with respect to taxonomic membership. As for RCL, we ran a Monte Carlo randomization of 5000 iterations to assess the significance of positive or negative excursions of ICL.

All calculations were done in Rstudio, and the code is available on GitHub (URL here). We assessed the degree of correlation between our indices of clustering (RCL and ICL) and various metrics of diversity, extinction or speciation using Spearman’s Rank Correlation tests. Plots were made using the “ggplot2” package and were combined in Adobe Illustrator prior to publication.

**Results**

**Discussion**

**Conclusion**

**Acknowledgements**

We thank David Furbish for his extensive assistance with developing and interpreting the Monte Carlo used in assessing the validity of positive or negative excursions of our indices of phylogenetic clustering, and for his encouragement and assistance throughout the development of this project and the writing of this publication. We also thank Gene Hunt for his insight and help with writing the R code for Moran’s I and the creation of the Mtax, Mext and Morig matrices. We thank Stephanie Birnbaum and Brandt Gibson for their roles in the initial development of this project, which took a number of twists and turns and ultimately changed directions entirely.

**References Cited**

Gittleman, J.L. and Kot, M. (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology* 39(3): 227 – 241.

Krug, A.Z. and Patzkowsky, M.E. (2015) Phylogenetic clustering of origination and extinction across the Late Ordovician mass extinction. *PLoS ONE* 10(12): e0144354

Lockwood, J.L., Russell, G.R., Gittleman, J.L., Daehler, C.C., McKinney, M.L. and Purvis, A. (2002) A metric for analyzing taxonomic patterns of extinction risk. *Conservation Biology* 16(4): 1137 – 1142.

Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika* 37(1/2): 17 – 23.

Roy, K., Hunt, G. and Jablonski, D. (2009) Phylogenetic conservatism of extinctions in marine bivalves. *Science* 325(5941): 733 – 737.

**Figure Captions**

**Fig. 1** – Rates of (A) extinction and (B) origination in North American mammals throughout the Cenozoic, and the degree to which these events were clustered using (C, D) the index of relative clustering, RCL, and (E, F) Moran’s *I*.

**Table Captions**