PNAS RESEARCH ARTICLE | EVOLUTION

Analysis of biodiversity data suggests that mammal species are hidden in predictable places

Danielle J. Parsons[a,b], Tara A. Pelletier[c], Jamin G. Wieringa[b,d], Drew J. Duckett[a,b], and Bryan C. Carstens[a,b,1]

Edited by Robert Pringle, Princeton University, Princeton, NJ; received February 19, 2021; accepted February 14, 2022 by Editorial Board Member Simon A. Levin

Research in the biological sciences is hampered by the Linnean shortfall, which describes the number of hidden species that are suspected of existing without formal species description. Using machine learning and species delimitation methods, we built a predictive model that incorporates some 5.0×10^5 data points for 117 species traits, 3.3×10^6 occurrence records, and 9.1×10^5 gene sequences from 4,310 recognized species of mammals. Delimitation results suggest that there are hundreds of undescribed species in class Mammalia. Predictive modeling indicates that most of these hidden species will be found in small-bodied taxa with large ranges characterized by high variability in temperature and precipitation. As demonstrated by a quantitative analysis of the literature, such taxa have long been the focus of taxonomic research. This analysis supports taxonomic hypotheses regarding where undescribed diversity is likely to be found and highlights the need for investment in taxonomic research to overcome the Linnean shortfall.

cryptic species | taxonomy | predictive modeling | species delimitation | class Mammalia

Significance

Only an estimated 1 to 10% of Earth's species have been formally described. This discrepancy between the number of species with a formal taxonomic description and actual number of species (i.e., the Linnean shortfall) hampers research across the biological sciences. To explore whether the Linnean shortfall results from poor taxonomic practice or not enough taxonomic effort, we applied machine-learning techniques to build a predictive model to identify named species that are likely to contain hidden diversity. Results indicate that small-bodied species with large, climatically variable ranges are most likely to contain hidden species. These attributes generally match those identified in the taxonomic literature, indicating that the Linnean shortfall is caused by societal underinvestment in taxonomy rather than poor taxonomic practice.

Species-level taxonomic designations are utilized by conservationists, legislators, ecologists, and evolutionary biologists to manage, conserve, and understand Earth's biodiversity (1). While over 1 million species have been described by taxonomists, this number likely represents 1 to 10% of the actual number of extant species (e.g., refs. 2, 3). Undescribed biodiversity has considerable implications for conservation management and basic research questions in ecology and evolutionary theory (4). Advances in DNA sequencing technology have inspired biologists to develop novel methods of species delimitation that utilize genetic data (5). These methods have led to the discovery of many morphologically cryptic species, defined as two or more distinct species that have been misclassified as a single species due to a lack of diagnosable morphological traits (6). These hidden species may be caused by evolutionary constraints within a clade that inhibit obvious morphological divergence or due to differences in visual, audible, or olfactory cues that humans are unable to detect (4). Whether morphologically cryptic or characterized by unquantified divergent traits, hidden species appear to be present in most metazoan families and biogeographic regions (6). These hidden species represent a "biodiversity wildcard" because so much of what is believed to be known in biology is derived from studies that rely on recognized (i.e., formally described) species as primary units of analysis (4).

Understanding the prevalence of hidden biodiversity over large spatial and taxonomic scales is difficult even in well-studied groups like mammals. In contrast to groups such as arthropods, where taxonomists suspect that the majority of species are currently undescribed (3), most actual species of mammals are thought to be described (7, 8). Even so, and while rates of species description vary across mammalian orders (*SI Appendix*, Fig. S1), new species continue to be discovered (9, 10). It remains unclear whether the observed variation in the magnitude of which species are described across mammalian orders reflects biological reality, because species are more difficult to discover in certain clades, or results from systematic bias in taxonomic practice. Regardless, there are potentially hundreds of actual mammal species that lack formal description, particularly in small-bodied clades such as bats, rodents, and eulipotyphans (9, 11). Previous work (e.g., refs. 9, 10) suggest that hidden species are likely to be discovered in areas of high endemism, with disproportionate numbers expected in insular systems. That such regions are also likely to be a greater conservation risk highlights the need for renewed taxonomic effort (3, 8).

Previous attempts to elucidate large-scale patterns of hidden diversity in mammals have relied on qualitative arguments (e.g., ref. 10) or metaanalysis of the species literature (e.g., ref. 4) and produced conflicting accounts regarding both the extent of hidden diversity and the representation of this diversity across bioregions and mammalian

Author contributions: D.J.P., T.A.P., J.G.W., D.J.D., and B.C.C. designed research; D.J.P., T.A.P., J.G.W., and D.J.D. performed research; D.J.P., T.A.P., J.G.W., and D.J.D. analyzed data; and D.J.P. and B.C.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.P. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: carstens.12@osu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2103400119/-/DCSupplemental.

Published March 28, 2022.

orders (e.g., refs. 6, 7, 12). Since metaanalyses rely on the published literature and as such are potentially subject to systematic bias, we adopt data repurposing here (i.e., the reuse of data that were originally collected for a different purpose) (13) to address a dilemma central to modern taxonomy: Undescribed species are suspected to be common, but understanding this phenomenon is inhibited by a lack of information regarding the attributes that make a clade likely to contain hidden species. We develop a predictive framework to identify the clades that are likely to contain hidden species and identify specific trait complexes that identify where this hidden diversity is likely to be found (Fig. 1).

The results of our predictive analysis potentially have broad implications. For example, a finding that hidden species are not predictable across taxonomic groups would imply that taxonomists lack the requisite information to recognize and describe hidden species. Alternatively, if hidden species can be predicted, the information needed to recognize hidden species is either available or potentially obtainable, which would aid taxonomic efforts. Furthermore, the results of the predictive analysis may help to identify characteristics of taxa that contain hidden diversity, which would greatly improve the efficiency of taxonomic description. To address this question, we analyze genetic data using species-delimitation methods (14, 15) before applying random forest classification (16) to develop a predictive model. This framework (Fig. 1) uses data from geographic, environmental, morphological, sampling effort, and life history traits as predictor variables to classify taxa as either containing hidden species or not containing hidden species. The use of this categorical response variable allows us to identify predictor variables that indicate where hidden species can be found.

Results

Dataset. We compiled a global dataset of mammalian barcoding gene sequences, occurrence records, and species traits (Fig. 2). A total of 90,759 DNA sequences from the cytochrome oxidase subunit I (COI) and cytochrome *b* (cytb) genes were obtained from 4,310 recognized mammal species available in the National Center for Biotechnology Information genetic sequence database, GenBank. To ensure consistency, we updated the taxonomy associated with all data to reflect that of the Mammal Diversity Database published by the American Society of Mammalogists (Datasets S1 and S2) (17). For all recognized mammal species, we compiled a database of 117 variables, each potentially predictive of hidden species, generated from geographic, environmental, life history, and taxonomic information available on public databases (Dataset S3). We also collected ∼3.3 million global positioning system coordinates from recorded geographic occurrences in order to obtain environmental, climatic, and geographic data for all recognized mammalian species.

Species Delimitation. To identify recognized species that potentially contain hidden diversity, we generated DNA sequence alignments for each family in class Mammalia, determined the optimal model of sequence evolution for each alignment, and performed two methods of automated species delimitation, one based on genetic distance (automated barcode gap discovery [ABGD]) (14) and the other based on an evolutionary model (generalized mixed Yule coalescent [GMYC]) (15). Using a conservative consensus across genes and delimitation methods, our analysis suggests that as many as one-third of the species included in the consensus analysis contain undescribed species (*SI Appendix*, Fig. S2). While preliminary species assignments support previous claims (7) that global mammal diversity remains substantially underdescribed despite a lengthy history of taxonomic effort, a global view of where these hidden species can be found revealed no obvious geographic patterns apart from the observation that regions with high species richness contain a greater number of potentially hidden species (Fig. 3 and *SI Appendix*, Fig. S3 and Table S1). However, our finding that the majority of hidden species are likely to be found in three orders (Chiroptera, Rodentia, and Eulipotyphla) is consistent with recent species descriptions in class Mammalia (17) (*SI Appendix*, Fig. S1) as well as earlier predictions as to which clades were most expected to contain large numbers of undescribed species (e.g., refs. 9–11) (see Dataset S5 for sequence divergence results). While the number of hidden species predicted by our delimitation analyses are congruent with the magnitude of earlier estimates of the number of undescribed mammal species (e.g., refs. 9–11), our delimitation analysis represents the most comprehensive exploration of this question to date.

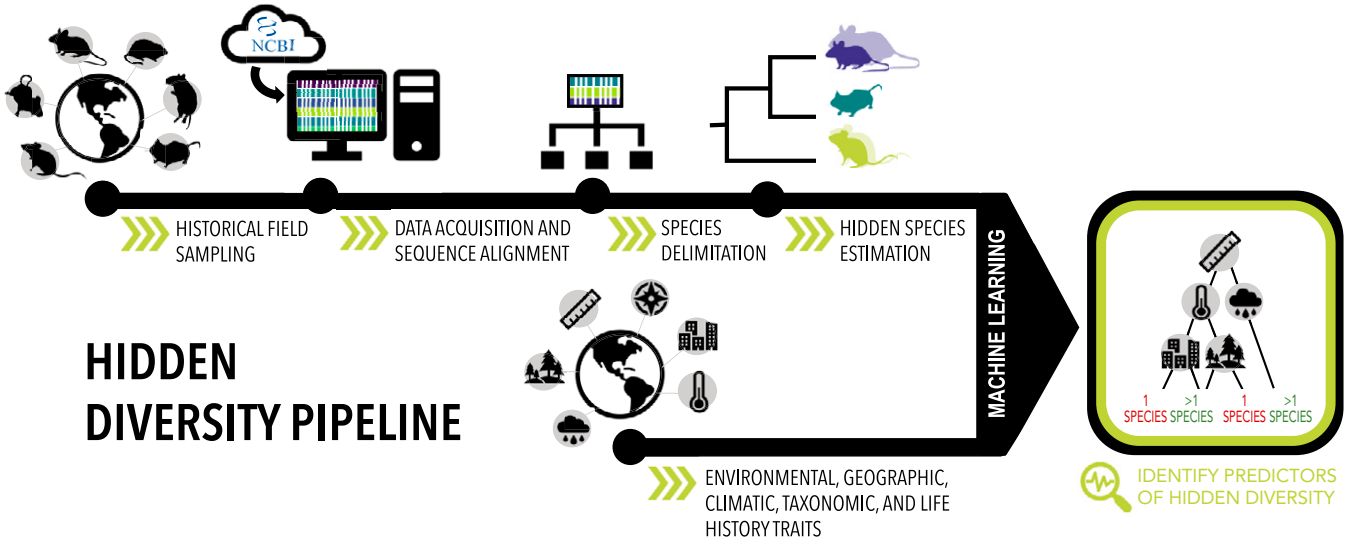

Fig. 1. Predictive modeling workflow. The framework proposed for identifying named mammal species that are likely to contain hidden diversity utilizes barcoding gene sequences and machine learning models built from environmental, geographic, climatic, taxonomic, and life history variables.

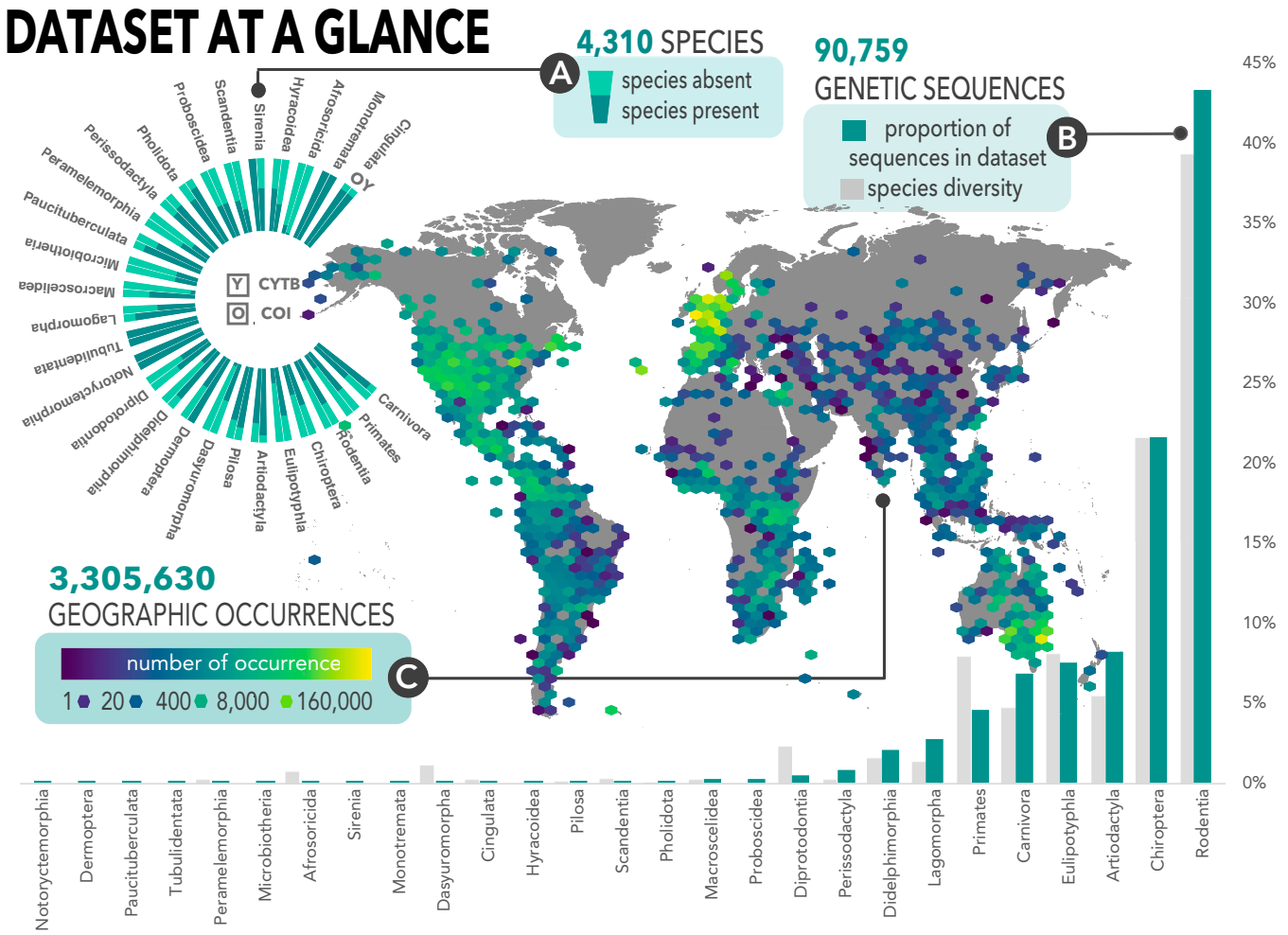

Fig. 2. Scope of the dataset. Genetic sequences for ~70% of currently recognized mammalian species were obtained. All mammalian orders are represented, with 23 orders containing sequences from both COI and cytb and 4 having only sequences from cytb. (*A*) Circle plots reflect species representation for the COI and cytb genes in each order. Dark bars represent the species present in the dataset and light bars represent species for which no genetic data are available. (*B*) Blue bars represent the proportion of the sequence database represented by each order, and gray bars represent the proportion of recognized species in each order. (*C*) A total of 3,205,630 geographic occurrence records were obtained for species present in the genetic database.

Predictors of Hidden Diversity in Mammals. The uneven distribution of hidden diversity indicated by our delimitation results implies that it may be possible to predict which clades harbor hidden species. To accomplish this goal, random forest classification was utilized to develop a predictive model using the geographic, environmental, morphological, taxonomic, and life history traits as predictive variables. For each random forest analysis, we used 80% of our data to train the model and reserved the remaining 20% as a test set to evaluate prediction accuracy of the resulting model. The random forest analysis based on this consensus estimate was able to predict hidden species of mammals with >80% accuracy (*SI Appendix*, Table S2B), a relatively high value for an analysis of this type (18).

The accuracy of the random forest prediction implies that there may be specific trait complexes that identify recognized taxa that potentially contain hidden species. Important predictors include adult body mass, range size, and climatic variables representing the temporal and spatial range of precipitation across the species range (Fig. 4 and Dataset S6). Species identified as potentially containing hidden diversity have, on average, smaller adult body mass (45 g to 135 g). Taxa identified as potentially containing hidden species also tended to have larger geographic ranges (trait "range area"; 1,270,799 km^2 to 378,072 km^2). Two climatic variables were important: Precipitation range of the warmest quarter and isothermality. On average, species identified as containing hidden diversity inhabit regions that have more precipitation in the warmest quarter of the year than do species that do not contain hidden diversity. The other variable, the range of isothermality, which quantifies the extent of day-to-night temperature oscillation relative to the annual summer-to-winter oscillations, is also larger for species identified as potentially containing hidden diversity. Intriguingly, two variables that partially quantify sampling effort, the geographic dispersion of species occurrence records and the number of recent publications that reference the species epithet, both for the named species that is predicted to contain hidden species, were also identified as important (Fig. 4).

Discussion

Predictors of Hidden Species in Mammals. As products of natural selection acting over thousands of generations, organismal traits are vitally important to taxonomic classification. However, we found that such traits are generally not predictive of hidden diversity in mammals. The notable exception is body mass, as hidden species are significantly more likely to be found in smaller-bodied taxa than larger-bodied taxa (Kruskal–Wallis test; $\chi^2 = 49.18$; $P = 2.3 \times 10^{-12}$). It seems possible that this trait is predictive because subtle morphological differences among congeners are more difficult to quantify in smaller

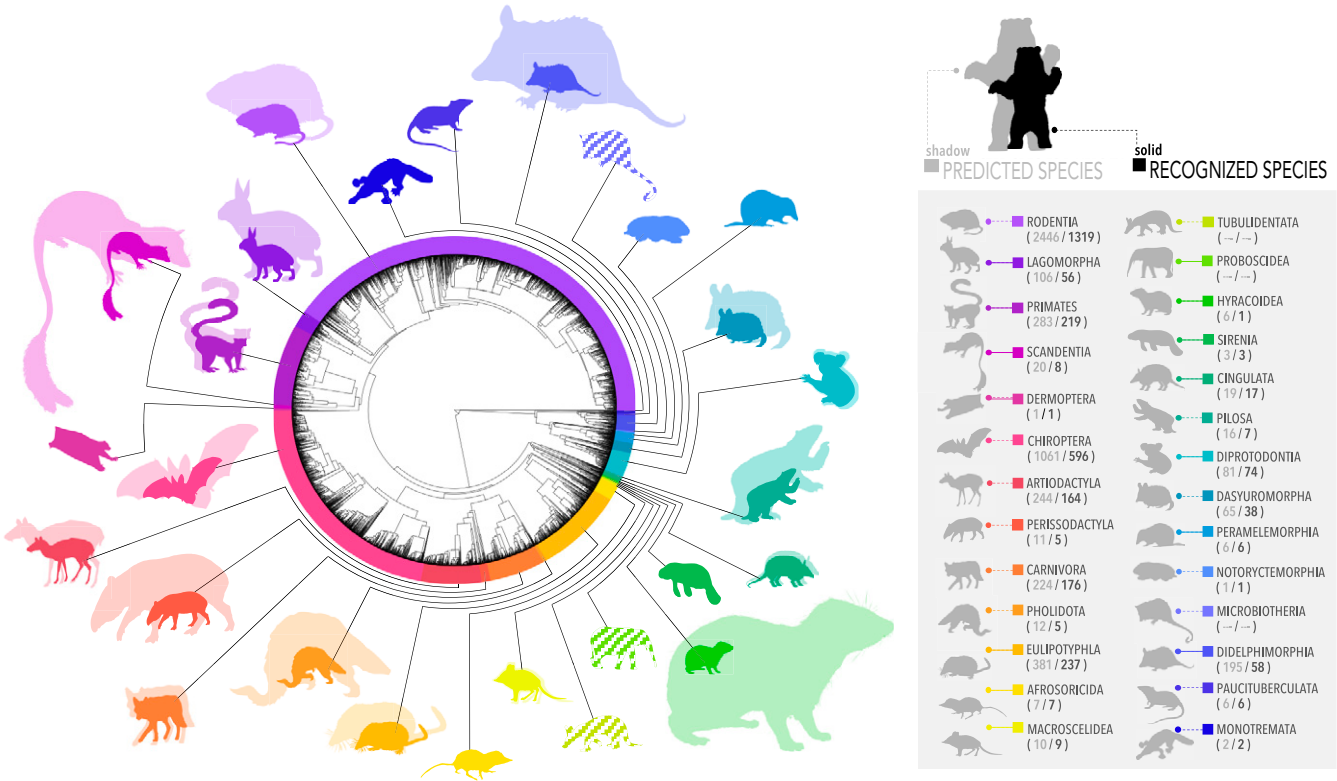

Fig. 3. Consensus results of species delimitation analyses. Phylogenetic distribution of hidden diversity estimated from strict consensus of delimitation results (*SI Appendix*, Table S1). Each silhouette represents a mammalian order with its shadow reflecting the ratio of predicted species to recognized species. Striped silhouettes represent orders with conflicting delimitation results that were not included in the predictive analysis. Phylogeny was adapted from ref. 31.

mammals than they would be in larger species. The variable "range area," which is taken from the PanTHERIA database (19) and measures the estimated size of the geographic range of each species on the basis of species descriptions in the taxonomic literature, is nearly as predictive as body size (Fig. 4). Hidden species are more likely to be found in recognized taxa with large ranges than those with small ranges (Kruskal–Wallis test; $\chi^2 = 98.35$; $P < 2.2 \times 10^{-16}$). Other predictors include climatic variables such as the range of isothermality, which describes the oscillations in daily and annual temperature, and precipitation during the warmest quarter. Hidden species are more likely to be found in taxa whose geographic range includes high isothermality and high precipitation range in the warmest quarter of the year (*SI Appendix*, Table S7). These climatic variables potentially are predictive because together they identify recognized taxa with geographic ranges expected to contain high variance in local habitat conditions, which may lead to the formation of genetic structure, either via habitat instability or local adaptation. However, since they are also characteristic of portions of the wet tropics, it is possible that the latitudinal gradient in mammalian species richness contributes to the identification of these variables as predictive. While this positive relationship (i.e., species richness and the number of hidden species) is expected since any hidden species as defined here exists within some named species, further analysis demonstrates that Southeast Asia contains the greatest proportion of hidden species relative to its species richness (*SI Appendix*, Fig. S5).

In addition to the organismal trait and climatic variables, two variables associated with sampling effort have high predictive value. Hidden species are predicted to exist in taxa with a high number of occurrence records in genera who have been more represented in recent publications in the taxonomic literature. To a large extent, both variables reflect sampling effort. Our finding that these sampling effort variables (i.e., "recent publications" and "occurrence area") have predictive value implies that the very taxa that have been the focus of previous research are more likely to contain hidden species. We interpret this as evidence that taxonomists working in mammal systems are generally aware of our finding that small-bodied mammals with large geographic ranges are likely to contain hidden species. Most species descriptions in mammals since 1992 have occurred in systems matching the general characteristics (i.e., small body size, large ranges, high isothermality) identified by our predictive model (9). Although our results are consistent with earlier efforts (e.g., refs. 11, 17), they provide quantitative evidence that taxonomists are actively researching the clades where undescribed species are likely to be found.

Broader Implications. The identification of sampling effort variables as predictive has broad implications beyond class Mammalia. An accurate implementation of the random forest algorithm for machine learning requires that observations are balanced across the response variables (i.e., here that there is a substantial proportion of recognized taxa that contain and do not contain hidden species). Since mammals are a well-studied clade, there exist sufficient data in public databases to conduct this analysis and to build a predictive model with high accuracy. However, we know of no reason to think that the taxonomists working in class Mammalia are outliers in terms of their abilities to identify and describe species. Rather, we suspect that there have been more taxonomists working in mammals in proportion to the actual species diversity in the clade than in hyperdiverse clades such as beetles or mites.

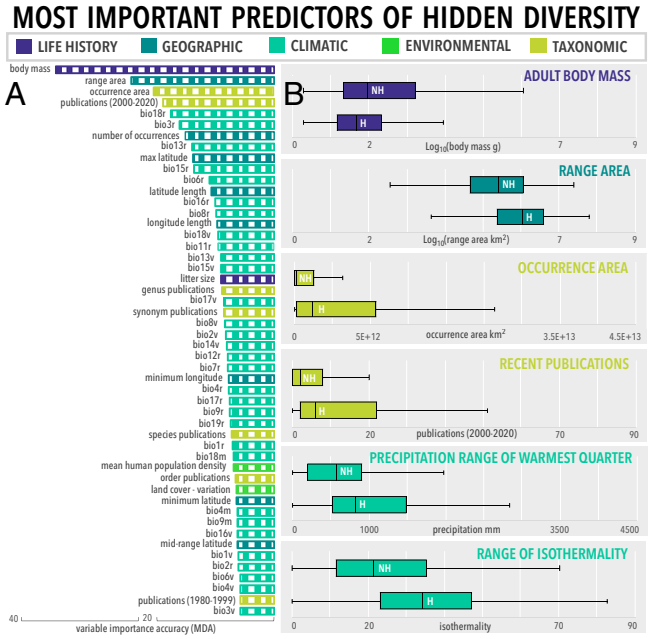

Fig. 4. Important predictors of hidden species in mammals. (*A*) From *Top* to *Bottom*, the 50 most important predictive variables (judged by MDA), for the consensus random forest classification model. In both plots, variables are color coded by life history, geographic, climatic, taxonomic, and environmental. (*B*) Boxplots representing values of the top predictive variables for species included in the consensus model. Values from species identified as hidden are shown at the *Bottom* of each plot (labeled "H"), and values from species not identified as hidden are shown *Above* (labeled "NH"). Outliers are excluded from boxplots.

While the popular conception of species discovery consists of scientific expeditions to remote regions, specimens from many undescribed species are already present in natural history collections (20). Our results support earlier conclusions (4, 7, 9) that taxonomic work is far from complete even in well-studied clades such as class Mammalia. By extrapolation, our predictive model supports calls for a substantial investment in taxonomy across the tree of life (e.g., ref. 21). A broad-scale analysis of available sequence data from barcoding genes can provide researchers predictions about the clades that are likely to contain undescribed species, and these results broadly justify the allocation of resources to conduct additional taxonomic work (22, 23). For example, the $>6.0 \times 10^6$ sequences from barcoding genes in $>2.0 \times 10^5$ recognized insect species suggest that a study similar to ours could be conducted in class Insecta.

The delimitation methods used here rely on genetic species concepts that are based either on genetic distance among putative species (14) or on a coalescent model (15). Such methods are similar in spirit to earlier proposals that identified hidden species by calculating sequence divergence in mitochondrial genes, for example our results are broadly congruent with the range of ~3 to 7% suggested by one previous study (11). However, mitochondrial data are most appropriate for initial surveys (e.g., ref. 24) and should not form the sole basis for species description in mammals, despite the routine use of similar concepts in other clades (25). Single-locus methods for species delimitation are practical in analyses with hundreds of species and thousands of sequences (24) and thus represent an important part of the species discovery pipeline (5) even if analyses based on single-locus genetic markers can fail to recognize cryptic species under certain conditions (26, 27). The widespread availability of barcoding gene data makes them particularly cost effective for a large-scale assessment of hidden biodiversity that can supplement taxonomic description (5). The predictive modeling approach developed here is not intended to replace traditional species description, which requires considerable expertise, but to aid in such efforts by identifying clades where such work is most needed.

Conclusions. Our work supports published syntheses, which suggest that many mammal species are undescribed (9–11) and demonstrates that an accurate prediction can be made regarding where hidden species are likely to be found. This finding is attributable to the analytical framework developed here. While our results contradict previous research, which argued that hidden species were idiosyncratic in their characteristics (6), it is broadly congruent with suggestions that species-rich regions at low latitudes contain the bulk of the hidden diversity in class Mammalia (e.g., ref. 7) (*SI Appendix*, Figs. S4 and S5). Previous syntheses on this topic relied on metaanalysis, which is appropriate for literature synthesis (28) but of limited utility for prediction. Automated data repurposing, where existing data are reanalyzed with the goal of fostering new insight (12), can leverage the massive amount of information present across hundreds of databases in the biological sciences. Predictive analyses complement investigations that are conducted on a localized scale (29). Our results indicate that traditional taxonomic research is effective, as taxonomists have long suspected that hidden species in mammals are likely to be found in small-bodied species with large geographic ranges (e.g., refs. 9, 17). Our finding that two variables, which quantify sampling effort are among the most predictive in our analysis (Fig. 4), suggest that the Linnean shortfall in mammals can be overcome with a greater investment in taxonomic research. If one is willing to accept that class Mammalia has historically received a disproportionate amount of taxonomic effort relative to the actual number of species as compared to groups such as arthropods, then more broadly our results indicate that the Linnean shortfall can be addressed across the tree of life with concerted effort and increased funding. Our study reinforces existing calls for a greater investment in taxonomic research (21, 30), particularly in understudied and undescribed taxa facing quiet extinction (31). It suggests that hidden species exist in predictable places, waiting for formal description.

Materials and Methods

Genetic Data Acquisition and Record Processing. We downloaded all available mammalian DNA sequences for the mitochondrial genes cytochrome-c oxidase I (COI) and cytochrome-b (cytb) from the NIH genetic sequence database, GenBank, after which we followed the basic genetic preprocessing pipeline outlined by Upham et al. (32). For each gene, we grouped sequences by species and then manually checked all species records for errors (e.g., subspecies, duplicates, extinct species, etc.; see *SI Appendix*, Fig. S4 for initial data processing pipeline). To ensure standardization across groups, we updated all sequence taxonomy to reflect that of the Mammal Diversity Database (MDD) published by the American Society of Mammologists (17) (Dataset S1 for MDD taxonomy and Datasets S2 and S3 for list of taxonomic updates).

Multiple Sequence Alignment and Sequence Evolution Models. Following taxonomic standardization, we grouped sequence records for each gene by family and generated multiple sequence alignments for COI and cytb independently using MUSCLE v3.5 (33). We then visually inspected each family-level alignment for gaps and removed problematic sequences causing severe gaps or misalignment that could not be resolved through reverse complement or manual correction. In order to maximize downstream computational efficiency, we trimmed sequence ends that contained no variation and split alignments containing over 2,000 sequences into subgroups of related taxa using high-level taxonomy (i.e., subfamily, tribe, genus; see *SI Appendix*, Table S3 for final groupings). Finally, we determined best-fit models of nucleotide substitution for each alignment using jModelTest2 (34).

Species Delimitation and Hidden Species Estimation. We used two different methods of species delimitation to estimate levels of hidden diversity within groups of phylogenetically related mammals. Species delimitation was first done using a likelihood-based method under the GMYC (15). We generated maximum clade credibility trees for each alignment using BEASTv2.5.0 (35). We then used the resulting trees as input for the GMYC model from the "splits" R package (36) implemented in R v3.6.3 (37). Species delimitation was then repeated using the distance-based delimitation method, ABGD (14). Pairwise genetic distance matrices used as input for ABGD were calculated for each alignment under the previously determined best-fit model of sequence evolution (38) using PAUP* (39). We used the delimitation results generated from both GMYC and ABGD analysis of the genes COI and cytb to estimate the number of hidden species suggested by the genetic data. To measure general agreement between delimitation methods and genes while accounting for variation in the underlying analyses and sequence availability, we generated a conservative estimate of mammalian hidden diversity using a consensus of delimitation results for species in which results from all analyses agreed. Additional information regarding consensus model and species delimitation methods can be found in *SI Appendix*.

Predictor variables. For each recognized species, we explored a large number of geographic, environmental, morphological, sampling effort, and life history variables to determine whether any of these traits could be used to predict the presence of hidden diversity in mammals. We first downloaded all geographic coordinates for class Mammalia from the Global Biodiversity Information Facility (GBIF) (see Dataset S6 for a list of GBIF DOIs) and used these to extract data from several geographic information system (GIS) layers, including elevation (40), the 19 BIOCLIM layers at 1-km resolution pertaining to temperature and precipitation available from the WorldClim database (41), population density (42), gross domestic product (43), light pollution (44), protected areas (45), anthropogenic biomes (46), and GlobCover by the European Space Agency (47). We used the following R packages to extract information from these layers on a species-by-species basis: "geosphere" (48), "raster" (49), "rgdal" (50), and "plyr" (51). In addition to data generated from occurrence records and GIS layers, we included several morphological, geographic, and life history traits gathered from the PanTHERIA database (51), including adult body mass (in grams), diet breadth, habitat breadth, terrestriality, trophic level, litter size, actual evapotranspiration, potential evapotranspiration, geographic range area (in square kilometers), maximum latitude of range, minimum latitude of range, midrange latitude of range, maximum longitude of range, minimum longitude of range, midrange longitude of range, mean population density (number per square kilometer), population density minimum (number per square kilometer), and population density (change). Finally, we generated a set of variables from the taxonomic species description literature to act as a proxy for sampling effort by performing a literature search using the R package "wosr" (52) to query Web of Science and estimate various publication-related metrics on a species-by-species basis. The dataset building process is described in full detail in *SI Appendix*.

Predictive Modeling and Variable Importance. Machine learning was used to identify traits that predict the presence of hidden lineages within currently recognized species. We used random forest analysis, an approach that utilizes multiple decision trees to predict the response (presence or absence of hidden diversity) based on many potential predictor variables (**16, 18**). In order to avoid bias induced by correlation among predictor variables, each individual decision tree consists of a subset of the data and a random ordering of variables at the nodes. The importance of each variable is determined by either measuring the quality of the prediction after the removal of each variable in the predictive function (measured by mean decrease in accuracy [MDA]) or by summing over the number of splits that include each variable across all trees, weighted by the number of samples being split (measured by Gini importance [Gini]). We used the R package "caret" (**53**) to build a random forest classification model for each of our four delimitation results (ABGD COI, ABGD cytb, GMYC COI, and GMYC cytb) as well as our strict consensus model. We explored each random forest model using a series of the independent variables mentioned above (see Dataset S3 for a list of variables used in final machine learning models). For each analysis we generated 1,000 decision trees, using 80% of our data as a training set and reserving the remaining 20% as a test set to evaluate prediction accuracy of the resulting model. Within each model, we used 10-fold cross-validation with five repeats to tune the parameter mtry, the number of variables randomly sampled as candidates at each split, by optimizing model specificity, sensitivity, and area under the receiver operating characteristic curve. The optimal mtry value was then used to generate the final model, from which variable importance measurements (MDA and Gini) were extracted (Dataset S6). We then applied this model to the remaining test set data to evaluate model accuracy, positive predictive value, negative predictive value, and model error.

Data Availability. Genetic and geographic data have been deposited in Dryad (https://doi.org/10.5061/dryad.b2rbnzshp). Code related to this manuscript, including DNA sequence alignments, analysis files, trait data, and machine learning input files have been deposited in GitHub (https://github.com/parsons463/HiddenDiversity). All other data are available in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. We thank Carl von Linné and all taxonomists who have worked to describe biological diversity. We thank L. N. Barrow, E. M. Fonseca, F. M. Lanna, M. L. Smith, M. C. T. Thomé, T. Yuri, R. Norris, H. Klompen, M. Daly, and J. Freudenstein for their discussion of this work. We thank editors and reviewers for their comments, which improved this manuscript. Support for this work was provided by the NSF (DBI-1661029 and DBI-1910623 to B.C.C. and DBI-1911293 to T.A.P.) and the Ohio Supercomputing Center (PAA1174).

Author affiliations: [a]Museum of Biological Diversity, The Ohio State University, Columbus, OH 43212; [b]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43212; [c]Department of Biology, Radford University, Radford, VA 24142; and [d]Department of Evolution, Ecology, and Organismal Biology, School of Environment and Natural Resources, Ohio Biodiversity Conservation Partnership, The Ohio State University, Columbus, OH 43210

1. F. E. Zachos, *Species Concepts in Biology: Historical Development, Theoretical Foundations and Practical Relevance* (Springer US, 2016).
2. V. H. Heywood, R. T. Watson, U. N. E. Programme, *Global Biodiversity Assessment* (Cambridge University Press, Cambridge; New York, 1995).
3. M. J. Costello, R. M. May, N. E. Stork, Can we name Earth's species before they go extinct? *Science* **339**, 413–416 (2013).
4. D. Bickford *et al.*, Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22**, 148–155 (2007).
5. B. C. Carstens, T. A. Pelletier, N. M. Reid, J. D. Satler, How to fail at species delimitation. *Mol. Ecol.* **22**, 4369–4383 (2013).
6. M. Pfenninger, K. Schwenk, Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evol. Biol.* **7**, 121 (2007).
7. G. Ceballos, P. R. Ehrlich, Discoveries of new mammal species and their implications for conservation and ecosystem services. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3841–3846 (2009).
8. M. A. Titley, J. L. Snaddon, E. C. Turner, Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. *PLoS One* **12**, e0189577 (2017).
9. D. M. Reeder, K. M. Helgen, D. E. Wilson, Global trends and biases in new mammal species discoveries. *Occas. Pap. Mus. Texas Tech Univ.* **269**, 1–36 (2007).
10. C. J. Burgin, J. P. Colella, P. L. Kahn, N. S. Upham, How many species of mammals are there? *J. Mammal.* **99**, 1–14 (2018).
11. R. J. Baker, R. D. Bradley, Speciation in mammals and the genetic species concept. *J. Mammal.* **87**, 643–662 (2006).
12. R. Poulin, G. Pérez-Ponce de León, Global analysis reveals that cryptic diversity is linked with habitat but not mode of life. *J. Evol. Biol.* **30**, 641–649 (2017).
13. B. Sidlauskas *et al.*, Linking big: The continuing promise of evolutionary synthesis. *Evolution* **64**, 871–880 (2010).
14. N. Puillandre, A. Lambert, S. Brouillet, G. Achaz, ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol. Ecol.* **21**, 1864–1877 (2012).
15. J. Pons *et al.*, Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* **55**, 595–609 (2006).
16. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
17. American Society of Mammalogists, Mammal Diversity Database (2021). https://www.mammaldiversity.org. Accessed 5 February 2021.
18. G. Biau, Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063–1095 (2012).
19. K. E. Jones *et al.*, PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648 (2009).
20. D. P. Bebber *et al.*, Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22169–22171 (2010).
21. H. C. J. Godfray, Challenges for taxonomy. *Nature* **417**, 17–19 (2002).
22. J. D. S. Witt, D. L. Threloff, P. D. N. Hebert, DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: Implications for desert spring conservation. *Mol. Ecol.* **15**, 3073–3082 (2006).
23. L. L. Knowles, B. C. Carstens, Delimiting species without monophyletic gene trees. *Syst. Biol.* **56**, 887–895 (2007).
24. J. A. Esselstyn, B. J. Evans, J. L. Sedlock, F. A. Anwarali Khan, L. R. Heaney, Single-locus species delimitation: A test of the mixed Yule-coalescent model, with an empirical application to Philippine round-leaf bats. *Proc. Biol. Sci.* **279**, 3678–3686 (2012).
25. F. M. Cohan, Systematics: The cohesive nature of bacterial species taxa. *Curr. Biol.* **29**, R169–R172 (2019).

26. M. J. Hickerson, C. P. Meyer, C. Moritz, DNA barcoding will often fail to discover new animal species over broad parameter space. *Syst. Biol.* **55**, 729–739 (2006).
27. T. Fujisawa, T. G. Barraclough, Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: A revised method and evaluation on simulated data sets. *Syst. Biol.* **62**, 707–724 (2013).
28. J. Gurevitch, J. Koricheva, S. Nakagawa, G. Stewart, Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
29. T. A. Pelletier, B. C. Carstens, D. C. Tank, J. Sullivan, A. Espíndola, Predicting plant conservation priorities on a global scale. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 13027–13032 (2018).
30. A. V. Suarez, N. D. Tsutsui, The value of museum collections for research and society. *Bioscience* **54**, 66–74 (2004).
31. N. Eisenhauer, A. Bonn, C. A Guerra, Recognizing the quiet extinction of invertebrates. *Nat. Commun.* **10**, 50 (2019).
32. N. S. Upham, J. A. Esselstyn, W. Jetz, Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* **17**, e3000494 (2019).
33. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
34. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).
35. R. Bouckaert *et al.*, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
36. T. Ezard, T. Fujisawa, T. G. Barraclough, SPLITS: Species' limits by threshold statistics (2009). https://r-forge.r-project.org/projects/splits/. Accessed 1 December 2020.
37. R. C. Team, R: A language and environment for statistical computing (2020). https://www.r-project.org/. Accessed 1 December 2020.
38. A. J. Barley, R. C. Thomson, Assessing the performance of DNA barcoding using posterior predictive simulations. *Mol. Ecol.* **25**, 1944–1957 (2016).
39. D. L. Swofford, "PAUP (phylogenetic analysis using parsimony)," in *Encyclopedia of Genetics, Genomics, Proteomics, and Informatics*, G. P. Rédei, Ed. (Springer Press, 2008), pp. 1455–1455.
40. ASTER GDEM Validation Team, Aster global digital elevation model version 2 (2011). https://asterweb.jpl.nasa.gov/gdem.asp. Accessed 1 December 2019.
41. R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
42. CIESIN, Socioeconomic Data And Applications Center (SEDAC) Gridded Populations of the World (GPW) (2016). https://sedac.ciesin.columbia.edu/data/collection/gpw-v4. Accessed 1 December 2019.
43. DECRG, World Bank Development Economics Research Group (DECRG) Gross Domestic Product (2010). https://datacatalog.worldbank.org/dataset/gross-domestic-product-2010/resource/addfd173-a15f-4cee-8f07-0ad76ae389b0. Accessed 1 December 2019.
44. NOAA, NOAA Night Light Development Index (NLDI) (2006). https://www.ngdc.noaa.gov/eog/dmsp/download_nldi.html. Accessed 1 December 2019.
45. European Environment Agency, World database on orotected areas (2012). https://www.eea.europa.eu/data-and-maps/figures/overview-of-protected-areas-as. Accessed 1 December 2019.
46. E. C. Ellis, K. Klein Goldewijk, S. Siebert, D. Lightman, N. Ramankutty, Anthropogenic transformation of the biomes, 1700 to 2000. *Glob. Ecol. Biogeogr.* **19**, 589–606 (2010).
47. European Space Agency, ESA GlobCover project (2009). due.esrin.esa.int/page_globcover.php. Accessed 1 December 2019.
48. R. J. Hijmans, Geosphere: Spherical trigonometry (2016). https://cran.r-project.org/package=geosphere. Accessed 1 December 2020.
49. R. J. Hijmans, raster: Geographic data analysis and modeling (2016). https://cran.r-project.org/package=raster. Accessed 1 December 2020.
50. R. Bivand, T. Keitt, B. Rowlingson, rgdal: Bindings for the Geospatial Data Abstraction Library (2017). https://cran.r-project.org/package=rgdal. Accessed 1 December 2020.
51. H. Wickham, The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).
52. C. Baker, wosr: Clients to the "Web of Science" and "InCites" APIs (2018). https://cran.r-project.org/package=wosr. Accessed 1 December 2020.
53. M. Kuhn, Caret package. *J. Stat. Softw.* **28**, 1–26 (2008).