Supplementary Information for

Analysis of biodiversity data suggest that species are hidden in predictable places.

Danielle J. Parsons\textsuperscript{1,2}, Tara A. Pelletier\textsuperscript{3}, Jamin G. Wieringa\textsuperscript{3,4}, Drew J. Duckett\textsuperscript{1,2}, Bryan C. Carstens\textsuperscript{1,2*}.

Bryan C. Carstens

Email: carstens.12@osu.edu

This PDF file includes:

- Supplementary text
- Figures S1 to S8
- Tables S1 to S8
- Legends for Datasets S1 to S7
- SI References (1-45)

Other supplementary materials for this manuscript include the following:

- Datasets S1 to S7
Extended methods

Genetic data quality control, cleaning, and processing. After gathering sequence information from NCBI, we followed the basic genetic preprocessing pipeline outlined by Upham et al. 2019 (1). We manually checked each family level alignment for common errors that could potentially skew our results, such as contamination, taxonomic mismatches, and inclusion of duplicate sequences and/or extinct species. All alignments were visually inspected for gaps. Problematic sequences causing severe gaps or misalignment that could not be resolved through reverse complement or manual alignment were discarded if other representatives of the species were present in the alignment. Sequence ends containing no variable regions were trimmed to maximize computational efficiency.

Taxonomic reconciliation was performed to standardize both the genetic and trait data. The first step in this process was to update the taxonomy of downloaded NCBI DNA sequences to that of the Mammal Diversity Database (MDD). The framework for the choices made during this process are outlined in supplemental Figure S6, and species-level information regarding synonyms, subspecies, etc. is provided in the attached mammal diversity database Data S1 (i.e., S1, column “notes”). We generated an updated supplemental dataset containing additional information on each taxonomic change made. This new dataset contains a list of each genetic sequence used in the analyses, as well as its original NCBI species name and the final name it was classified as in this research.

Second, we reconciled the PanTHERIA taxonomy with the MDD taxonomy being used by matching PanTHERIA species names to those listed in the MDD. PanTHERIA species names that did not have a match in the MDD were either reconciled or discarded, according to the framework shown in supplemental Figure S6. Each of these changes is documented in a secondary supplemental database, Data S3.

To check for any effect of nuclear DNA of mitochondrial origin (NUMTs; 2-3), we checked for premature stop codons in the raw sequences from three rodent taxa. Rodents were used as examples because NUMTs have previously been found in multiple rodent clades (4-6). Representative clades included Capromyidae, Sciurinae, and Microtus. For each clade, unaligned COI and cytb sequences were aligned with MACSE, which assesses premature stop codons and frameshift mutations while aligning sequences (7). MACSE was run using reference sequences from Homo sapiens (GenBank NC_012920), following Delsuc and Ranwez (8) and using the MACSE_BARCODE pipeline (https://github.com/ranwez/MACSE_V2_PIPELINES). Despite indicating a few frameshift and insertion mutations, which may have been corrected in the original alignments through manual checking, MACSE found no premature stop codons in any gene or taxon. Additionally, the number of sequences retained in MACSE alignments was very similar to the number retained in the original alignments (see table S4), indicating that NUMTs likely had little effect on our analyses.

Generalized Mixed Yule Coalescent (GMYC) model for species delimitation (9). Gene trees were estimated for each alignment using Bayesian inference in BEASTv2.5.0 (10) because this method produces more accurate results than maximum likelihood estimation when using the GMYC (11-12). The accuracy of the GMYC increases as data are added such that analyzing large clades is a useful strategy for circumventing known shortcomings of the GMYC related to sampling biases, such as singleton species and incomplete species sampling (11,13-15). With the
exception of the models of sequence evolution, the default settings from BEAUTi v2.5.2 were applied and a Yule process tree prior was used (16). BEAST analyses were conducted for a minimum of 1x10^8 generations (sampling every 1000 generations) and increased as necessary to achieve acceptable ESS values (>200) and trace plots. Multiple runs were performed, and the results were combined, if convergence was not reached in the initial analysis. We excluded alignments from the GMYC analysis that did not reach adequate ESS values after three runs (Table S3; groups marked with *). Convergence and ESS values were assessed and optimal burnin values were determined using Tracer v1.5 (17). Independent runs were combined using LogCombiner v.2.5.0 with a 10-30% burnin, and resampling frequency of 10000*(number of runs performed). The maximum clade credibility tree was obtained through TreeAnnotator v2.5.0 using the BEAST posterior distribution of gene trees. We used the GMYC model from the ‘splits’ R package (18) implemented in R v3.6.3 (19) to estimate the number of species present in the genetic data, per family or subgroup. We used the single threshold implementation of the GMYC method, as it has been found to largely outperform the multiple threshold implementation, as the latter is prone to oversplit taxa (20-21).

Automated Barcode Gap Discovery (ABGD) method for species delimitation (22). We used ABGD (22) to estimate the number of genetic partitions for each family alignment. This method is efficient because it does not rely on a gene tree estimate and has worked well in metazoan taxa but does not perform quite as well when there are few sequences per species in the dataset (e.g., 22). Pairwise genetic distance matrices used as input for ABGD were calculated for each alignment under the previously determined best-fit model of sequence evolution (23) using PAUP* (24). The Pmax parameter was set to 0.01, as this produced the number of partitions that most closely matched those of empirical estimates in a test of accuracy for this method (22). After several exploratory analyses for each family and locus, we present results that are largely based on the default values for all other parameters. However, the value of X, which estimates the relative gap width, was adjusted for multiple alignments containing a large number of sequences (Table S5). In such cases, we evaluated multiple X values from 0.1-1.5 to determine the most consistent and reasonable result.

Species delimitation consensus. Delimitation results generated from both GMYC and ABGD analysis of the genes COI and cytb were used to estimate the number of hidden species suggested by the genetic data. For each recognized species, we compared estimates of hidden diversity predicted by the different delimitation analyses to evaluate credibility of the results and generate a conservative estimate of predicted hidden diversity. Empirical studies have shown that the tree-based GMYC tends to oversplit species, while the distance-based ABGD tends to undersplit species (14-15). As such, high levels of dissimilarity between the two could indicate the influence of underlying population structure present within the group. 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 (see Fig. S7 for consensus assignment pipeline). For species with results from multiple delimitation analyses, we excluded any species with conflicting results (i.e., species with results identifying them as both hidden and not hidden, depending on the delimitation method or gene used). Species in which all delimitation results agreed were included in the final consensus. Species with only a single delimitation result were also included in the consensus and assigned as either hidden or not hidden according to said result.
**Geographic, environmental, taxonomic, and life history data.** We explored a large number of geographic, environmental, morphological, taxonomic, and life history variables to determine whether any of these traits could be used to predict the presence of hidden diversity. These data were coded using the described MDD taxonomy. To build a dataset of environmental and sampling effort variables, we downloaded all geographic coordinates for class Mammalia from the Global Biodiversity Information Facility (GBIF; see Data S7 for list of GBIF download DOIs) and used these to extract data from several GIS layers (below). Only those occurrences with no known coordinate issues were selected, and we excluded Fossil, Machine Observation, and Unknown Occurrence as the basis of record from the search to avoid inaccurate points to the best of our ability. Of the GBIF occurrence records downloaded, 10,197 records were found to be explicitly linked to a particular DNA sequence in our dataset (Fig. S8). We then used the R package ‘CoordinateCleaner’ (25) to automatically flag and remove points with potential errors based on outliers (5*interquartile range), country centroids, latitude and longitude equivalence, empty values, and proximity to GBIF headquarters and known biodiversity institutions. The curated occurrence records were then matched to MDD species present in our database.

Our GIS data layers consisted of elevation (26), the 19 BIOCLIM layers at 1km resolution pertaining to temperature and precipitation available from the WorldClim database (27), population density (28), gross domestic product (29), light pollution (30), protected areas (31), and anthropogenic biomes (32). In addition, we included GlobCover by the European Space Agency (33), which describes 23 land cover variables at 260m x 300m resolution. We used the following R packages to extract information from these layers on a species-by-species basis: ‘geosphere’ (34), ‘raster’ (35), ‘rgdal’ (36), and ‘plyr’ (37). For each species, we used unique GPS points to draw a polygon and calculated the area contained by that polygon as a proxy for occurrence area. We then sampled 1000 points at regular intervals from this polygon to calculate the mean, range, and variance from each of the 26 data layers as a proxy for the environmental variability that occurred within the occupied range of each species. To further characterize the extent of species occurrences, we extracted the minimum, maximum, median, and length of both latitude and longitude from the GPS points of each species.

In addition to the GIS layers described in the previous paragraph, we included several morphological and life history traits gathered from the PanTHERIA database (38), which contains traits for 4,629 mammal species: adult body mass (g), diet breadth, habitat breadth, terrestriality, trophic level, litter size, actual evapotranspiration, and potential evapotranspiration. These variables were included because they are quantified across mammalian taxa and thus had the least amount of missing data. The R package ‘mice’ (39) was used to impute missing values using multivariate imputation by chain equations. Predictive mean matching was used to impute numeric values (adult body mass, litter size, actual evapotranspiration, and potential evapotranspiration), polytomous logistic regression was used to impute unordered factors (terrestriality and trophic level), and proportional odds models were used to impute ordered factors (diet breadth and habitat breadth). As the inclusion of phylogenetic information has been shown to improve estimation of missing trait data in mammals (e.g., 40), we imputed data within each order separately, treating family assignment as a predictor for imputation, when applicable. Imputation was repeated 10 times and results were averaged to produce the final imputed values. We also included several range-related geographic traits from the PanTHERIA database (38) to complement the values previously
extracted from available species occurrence data. These include range area (km$^2$), maximum latitude of range, minimum latitude of range, mid-range latitude of range, maximum longitude of range, minimum longitude of range, mid-range longitude of range, mean population density (n/km$^2$), population density minimum (n/km$^2$), 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. We performed a literature search using the R package ‘wosr’ (41) to query Web of Science and estimate publication numbers through time on a species-by-species basis. For each species, we calculated the total number of publications that included the species name from the year 1900 to the present (2020), as well as for each 20-year interval during that period. The number of publications including the keywords ‘synonymize’ and ‘revision’ were used to estimate the extent of synonymization and revision for each species, respectively. Similarly, the number of publications including ‘sp. nov’ were used to estimate the extent of novel species description. We excluded a small number of taxa that returned an exceptionally large number of publications (500 species-level publications or 10,000 genus-level publications), as these were either model species, such as Mus musculus, or those of agricultural importance (excluded species listed in Table S6). Finally, this process was repeated at the level of genus, family, and order. The resulting dataset contains information from 160 variables for over 6000 currently recognized species of mammals. Detailed information for each variable can be found in Data S6. Recognized species with data missing from any of the above-mentioned variables were excluded from the subsequent predictive modeling and variable importance analysis.

Extended Results

Genetic sequence dataset. To identify potentially hidden lineages in mammals and examine large scale patterns of hidden diversity, we compiled a global dataset of mammalian barcoding gene sequences (Fig. 2). A total of 90759 mitochondrial DNA sequences from 4310 mammal species were obtained from the NCBI genetic sequence database, GenBank. Of these sequences, roughly three fourths belong to the gene cytb (n=68426), while the remainder belong to COI (n=22333). In total, we obtained genetic data for approximately 70% of all currently recognized mammalian species. All 27 mammalian orders are represented, with 23 orders containing sequences from both COI and cytb and the remaining four orders having only sequences from cytb (Fig. 2A). The proportion of sequences belonging to each order in the database (Fig. 2B; blue bars) largely reflects underlying patterns of currently recognized species diversity in class Mammalia (Fig. 2B; gray bars).

Species delimitation (GMYC and ABGD). To identify potentially hidden lineages and evaluate the extent of hidden diversity in mammals we used our genetic database to generate DNA sequence alignments for each mammalian family. We then determined the best model of sequence evolution individually for each alignment and performed two methods of automated species delimitation to assign barcoding sequences to preliminary species. When compared to taxonomic designations of records in the database, preliminary species assignments from both delimitation methods revealed significant levels of hidden diversity (Fig. 3; Table S1), supporting previous claims that, despite a lengthy history of taxonomic effort, global mammal diversity remains significantly underestimated (e.g., 42).
Using the distance-based ABGD model, analysis of the cytb dataset identified 1295 species (roughly 31%) of the 4177 recognized species evaluated as potentially containing hidden diversity. Analysis of the same gene under the tree-based GMYC model identified 1713 species (roughly 46%) of the 3730 recognized species evaluated as potentially containing hidden diversity. There was substantial overlap in the species represented in both the COI and cytb datasets since both markers have seen wide use at the species level. Consequently, delimitation results suggest general agreement between the level of hidden diversity predicted by COI and that predicted by cytb. Under the distance-based ABGD model, analysis of the COI dataset identified 571 species (roughly 25%) of the 2259 species evaluated as potentially containing hidden diversity, while analysis of COI using GMYC identified 813 species (roughly 36%) of the 2230 recognized species evaluated as potentially containing hidden diversity.

While estimates of undescribed diversity are relatively consistent across analyses (between approximately 25-45%; Fig. S2), delimitation of cytb resulted in slightly elevated estimates of hidden diversity compared to COI, which likely reflects the increased coverage of cytb sequences in the database. Estimates of hidden diversity from both genes were significantly higher under the model-based GMYC, potentially reflecting underlying population structure detected by the model-based approach. To account for differences between delimitation analyses and availability of sequences, we generated a conservative estimate of mammalian hidden diversity using a consensus of delimitation results for species in which results from all analyses agreed. A total of 3015 recognized species were included in the final consensus model, of which 807 species (roughly 27%) were identified as potentially containing hidden diversity.

Our analysis of genetic data from 4310 recognized mammal species representing over 70% of class Mammalia suggests that approximately 30% of recognized mammal species could potentially harbor undescribed diversity. These results were not evenly distributed across mammal clades (Table S1) or geographic regions (Fig. S3), with several clades being predicted to contain substantially higher levels of hidden diversity than the number of species recognized by current taxonomy. Tropical regions contain more hidden species because these areas generally contain the highest species richness in Class Mammalia (Fig. S4) and SE Asia contains the greatest number of hidden species relative to its species richness (Fig. S5).

Geographic, environmental, taxonomic, and life history data. If hidden species are distributed more or less randomly across various taxonomic groups, it would imply that the factors that prevent taxonomic recognition and description of hidden species are idiosyncratic and thus difficult to address. However, the uneven distribution of hidden diversity indicated by the results of our delimitation analyses suggest it might be possible to predict which clades harbor hidden species. To accomplish this goal, we applied random forest classification to develop a predictive model for our estimates of mammalian hidden diversity, using use a combination of geographic, environmental, morphological, taxonomic, and life history traits as predictive variables. We built individual random forest models using the results of each delimitation analysis, as well as our final consensus estimation of mammalian hidden diversity.

For all recognized species of mammals, we added to our genetic dataset a series of potentially predictive variables generated from geographic, environmental, life history, and taxonomic information. We collected and analyzed information from publicly available trait databases, GIS layers, and approximately 3.3 million GPS coordinates from recorded geographic occurrences to
obtain species-level information for over 6000 mammalian species. From this, we produced an updated dataset which includes estimates of potential hidden diversity, statistics on genetic and geographic record availability, and data from 160 different predictive variables (Data S4).

**Predictive models.** We generated random forest classification models using species in which information for all of the above-mentioned variables was able to be obtained. Of the 4310 recognized species for which we generated estimates of hidden diversity, between 1332 and 2127 were included in subsequent machine learning models (Table S2A). For each of the resulting models, we calculated the following standard evaluation metrics to assess overall performance: i) model accuracy: the proportion of species correctly identified as either containing hidden species or not containing hidden species over the total number of species, ii) positive predictive value: the proportion of species correctly identified as containing hidden species over the total number of species predicted as containing hidden species iii) negative predictive value: the proportion of species correctly identified as not containing hidden species over the total number of species identified as not containing hidden species, and iv) model error: the proportion of species incorrectly identified over the total number of species (Table S2B). Classification models based on single gene delimitation results were able to predict potential hidden species with between 64-74% accuracy. Classification resulting solely from tree-based (GMYC) delimitation of the gene COI resulted in the model with highest error and lowest predictive value for hidden species (36% and 56%, respectively). This model also contained the fewest total species and the highest proportion of species identified as containing hidden diversity, suggesting that the elevated prediction error rates likely result from false positives in the model-based delimitation method that are caused by misdiagnosis of population genetic structure as species level diversity. While analyses based on the ABGD results were more accurate in their prediction, with average model accuracy of 71% (compared to an average of 65% in GMYC based models), our optimal results were found when using the conservative delimitation consensus model, which was able to predict hidden species of mammals with approximately 80% accuracy (81% predictive value for hidden species and 78% model accuracy; Table S2B).

**Predictors of hidden diversity in mammals.** Our machine learning models were able to identify potential hidden species status with up to 80% accuracy (Table S2B), indicating that we can make a reasonable prediction of which clades are likely to contain hidden species in class Mammalia. For many of the predictor variables used, the range of values observed in the species identified as potentially containing hidden diversity and those that were not identified as hidden are largely overlapping (Fig. 4B), making it challenging to predict the likelihood of hidden diversity based on any single variable. However, machine learning can be used to develop trait profiles, or complexes of biological features, that when taken together, can better predict a specific biological outcome (e.g., 43). We evaluated the individual decisions shaping the thousands of decision trees in our random forest models to identify specific trait complexes that distinguish taxa harboring potentially undescribed diversity. Predictors found to be consistently important across all models include adult body mass, range size, recent sampling effort, and several climatic variables representing the variation in precipitation and temperature values present across a species known range of occurrences. For the highest performing predictive model based on a strict consensus of delimitation results, adult body mass was the most important predictive variable, measured by both MDA and Gini (29% and 33%, respectively). Species identified as potentially containing hidden diversity have, on average,
smaller adult body mass than those not predicted to contain hidden diversity (average values 36 g and 229 g for species predicted and not predicted to contain hidden diversity, respectively; Fig. 4A). Following adult body mass, the area of both the described range and that of the known range of species occurrences were found to be the next most important variables in predicting the presence of potentially hidden lineages. Both measurements of range were found to be larger, on average, for potentially hidden species than for species not identified as containing hidden diversity (average described range values 3,274,062 km² and 1,259,940 km² for species predicted and not predicted to contain hidden diversity, respectively; Fig. 4B). Other variables of importance include recent sampling effort (which is, on average, higher for species identified as hidden), precipitation range of the warmest quarter (which is, on average, larger for species identified as hidden), and range of isothermality, or the extent of day to night temperature oscillation relative to the annual summer to winter oscillations (which is also, on average, larger for species identified as potentially hidden than for those not identified as containing hidden diversity).

**Model sensitivity testing.** To address the concern that a disproportionate amount of species classified as containing hidden diversity found in the largest groups of mammals (specifically rodents, bats, and shrews) could skew the results of our final random forest model, we performed a sensitivity analysis to evaluate the effect of excluding these hyper-diverse groups. The three largest mammalian orders, Rodentia, Eulipotyphla, and Chiroptera collectively make up just over half of the total number of hidden species in our model (Table S7). To determine what affect the inclusion of these species rich orders has on our random forest analysis, we systematically excluded each order from our final dataset and reran the random forest model. Each of these models are described in Figure S9. As our original consensus model was able to classify hidden species with approximately 80% accuracy, these results indicate that the removal of each large order results in a fairly large decrease in overall model accuracy. Despite the decrease in accuracy, variable importance values remain relatively stable across the models, with adult body mass being assigned the highest variable importance in each model, and geographic range size being assigned the second highest variable importance in all models except that excluding the order Chiroptera.

These results support our argument that the complex of traits originally identified by our consensus model as important predictors of mammalian hidden diversity are not likely to be simply an artifact of bias introduced by large, hyper-diverse groups.

**Significance of trait differences.** A series of statistical tests were performed to evaluate the significance of trait differences observed between species predicted to contain hidden diversity and those not predicted to contain hidden diversity. The dataset used for testing contained the same species used in our final (consensus) random forest model. Species were grouped based on whether they were predicted to contain hidden diversity, median variable values were calculated for each group, and the resulting group medians were compared using a Kruskal-Wallis test. Results for each of the variables tested are shown in Table S6. In all cases, there is a statistically-supported difference in the median values, supporting the claim that each of the traits identified by our random forest analysis does in fact differ between hidden and non-hidden species. This table has now been added to the Supplemental Materials (Table S8).
Inclusion of phylogenetic information as a predictive trait. To determine whether phylogenetic information could be incorporated into the model as a predictive trait, we added phylogenetic information in the form of order and family designations to the dataset used in our consensus random forest analysis and evaluated the results, as compared to our original model. Results of this model are provided in Figure S8. By including phylogenetic information as a predictive trait, we find that the overall model accuracy decreases only slightly. However, we lose a great deal in interpretability because, while we see a similar pattern in terms of which variables are identified as important predictors, the overall amount of importance has decreased for both MDA and Gini because the algorithm is forced to allocate this importance across these variables and the related phylogenetic data. Even Rodents, which contain 20% of the species in the database and are predicted to contain 45% of the hidden species, have their taxonomic order variable as relatively unimportant. As a major goal of this study is to evaluate the importance of particular species traits in predicting whether a given taxa is likely to contain hidden diversity, it is counterproductive to include traits that will prohibit us from doing so.
**Fig. S1.** Magnitude of species description rates (initial) across orders of mammals. Magnitudes of primary species description have varied over the past 200 years for the 5 most speciose mammalian orders (shown in color according to key in upper left). The remaining 22 mammal orders (shown in gray) have experienced a comparatively lower magnitude of description rates. Data from the ASM Mammal Diversity Database (44). Animal silhouettes from PhyloPic (45).
Fig. S2. Species delimitation results. Total number of species evaluated in each individual delimitation analysis performed as well as the consensus, with the blue bar (bottom) representing the number of species not identified as potentially containing hidden diversity and the green bar (top) representing the number of species identified as potentially containing hidden diversity.
**Fig. S3. Geographic distribution of species delimitation results.** Geographic spread of hidden diversity estimated from (A) strict consensus of delimitation results, (B) ABGD delimitation of COI, (C) GMYC delimitation of COI, (D) ABGD delimitation of cytb, and (E) GMYC delimitation of cytb. In each map, blue points represent species not identified as potentially containing hidden diversity and green points represent species identified as potentially containing hidden diversity. Occurrences were determined as the median latitude and longitude of species occurrence records in the dataset.
Fig. S4. Geographic distribution of species delimitation results. Geographic spread of hidden diversity estimated from strict consensus of delimitation results. Maps (A-B) represent all species contained in our consensus model, (C-D) represent species in the order Rodentia, (E-F) represent species in the order Chiroptera, and (G-H) represent species in the order Eulipotyphla. Map layers represent species occurrence area, which was calculated from GBIF occurrence. The value of ‘sp’ corresponds to the number of species that occur in a specific location on the map.
Fig. S5. Geographic spread of hidden diversity estimated as the proportion of hidden species over not hidden species. (A) Layers on the map, measured as H/NH, indicate the number of hidden species divided by the number of not hidden species occurring in a specific area of the map. (B-C) Each point on the scatterplot represents a single grid cell from map A and displays the cell’s proportion of hidden diversity (H/NH value) and its corresponding latitude and longitude, respectively.
Fig. S6. Initial data processing pipeline. Framework for initial processing of genetic sequence data and updating of species names to Mammal Diversity Database taxonomy (see Datasets S2-S3 for a list of the transformations used to update taxonomy).
Fig. S7. Delimitation consensus assignment process. Framework for determining consensus of species delimitation results for the strict consensus model.
Fig. S8. GBIF occurrence records directly linked to NCBI DNA sequences. (A) 10,197 of the GBIF occurrence records downloaded for this research were found to be directly linked to specific NCBI DNA sequences used in our dataset, with 8,140 belonging to the gene COI, and 2,057 belonging to the gene cytb. (B) Of the occurrences linked to cytb sequences, 1,549 records were from preserved specimens and 508 records were from material samples. (C) Of the occurrences linked to COI sequences, 8,065 records were from preserved specimens and 75 records were from material samples.
**Fig. S9. Exclusion of hyper-diverse orders.** Random Forest analysis based on subsets of the consensus dataset purposefully excluding the most diverse orders (Chiroptera, Eulipotyphla, and Rodentia). For each data subset overall model accuracy is given, along with the most important predictive variables (judged by mean decrease in accuracy [MDA]).
Fig. S10. Results of strict consensus model including phylogenetic information. Results from the consensus random forest classification model after the inclusion of phylogenetic information (specifically, order and family designations). Predictive variable importance is evaluated by both Mean Decrease in Accuracy (MDA) and Mean Decrease in Gini (Gini). Overall model accuracy is given at the top of the figure.
Table S1. Taxonomic results of species delimitation analysis. Species delimitation results for each of the individual species delimitation analyses as well as the strict consensus, including number of recognized species, number of predicted species, and the ratio of predicted (p) to recognized (r) species.

| ORDER                  | ACGD COI | ACGD cytb | GMYC COI | GMYC cytb | Consensus |
|------------------------|----------|-----------|----------|-----------|-----------|
|                        | recognized | predicted | ratio (p/r) | recognized | predicted | ratio (p/r) | recognized | predicted | ratio (p/r) |
| Ahosia arctica        | 2        | 1         | 1.33      | 0         | 0         | 0          | 2          | 1         | 1          |
| Artiodactyla          | 253      | 310       | 1.23      | 305       | 439       | 1.44       | 246        | 407       | 1.65       |
| Carnivora             | 172      | 207       | 1.22      | 260       | 343       | 1.29       | 173        | 254       | 1.48       | 228       | 620       | 2.72       | 175        | 223        | 1.28       |
| Chiroptera            | 572      | 936       | 1.64      | 817       | 1319      | 1.61       | 571        | 1264      | 2.22       | 830       | 2034      | 2.45       | 596        | 1061       | 1.78       |
| Osteichthyes          | 20       | 22        | 1.11      | 20        | 23        | 1.13      | 20         | 23        | 1.13       | 20        | 23        | 1.13       | 20         | 23         | 1.13       |
| Dasyuriformes         | 4        | 4         | 1.00      | 4         | 5         | 1.25      | 4          | 5         | 1.25       | 4          | 5         | 1.25       | 4          | 5          | 1.25       |
| Dermoptera            | 1        | 1         | 1.00      | 1         | 2         | 1.00      | 0          | 2         | 0          | 0         | 2         | 0          | 0         | 2          | 0          |
| Didelphimorphidae     | 31       | 51        | 1.65      | 90        | 213       | 2.37      | 31         | 81        | 2.61       | 90        | 478       | 5.31       | 58         | 194.75     | 3.36       |
| Diprotodontida        | 69       | 73        | 1.11      | 86        | 103       | 1.22      | 63         | 72        | 1.14       | 78        | 106       | 1.37       | 74         | 83         | 1.1       |
| Eulipotyphla          | 109      | 142       | 1.32      | 323       | 515       | 1.60      | 108        | 167       | 1.55       | 222       | 687       | 3.06       | 237        | 383        | 1.61       |
| Hapalidae             | 2        | 2         | 1.00      | 3         | 10        | 3.33      | 2          | 7         | 3.50       | 2         | 7         | 3.50       | 2          | 7          | 3.50       |
| Lagenomorphida        | 52       | 77        | 1.48      | 75        | 136       | 1.83      | 53         | 128       | 2.39       | 29         | 136       | 4.66       | 56         | 105.58     | 1.89       |
| Macroscelidea         | 2        | 2         | 1.00      | 13        | 15        | 1.15      | 2          | 2         | 1.00       | 2         | 2         | 1.00       | 2          | 2          | 1.00       |
| Microbiotheria        | 0        | 0         | 0.00      | 0         | 1         | 10.00     | 0          | 0         | 0.00       | 0         | 1         | 10.00      | 0          | 0          | 0.00       |
| Monotremata           | 2        | 2         | 1.00      | 3         | 3         | 1.00      | 0          | 2         | 0.50       | 0         | 2         | 0.50       | 0          | 2          | 0.50       |
| Notoryctemorphida     | 2        | 2         | 1.00      | 2         | 2         | 1.00      | 0          | 2         | 0.50       | 0         | 2         | 0.50       | 0          | 2          | 0.50       |
| Paucituberculata      | 5        | 5         | 1.00      | 6         | 6         | 1.00      | 5          | 5         | 1.00       | 6         | 6         | 1.00       | 5          | 5          | 1.00       |
| Peramlemorphida       | 6        | 6         | 1.00      | 9         | 12        | 1.33      | 6          | 6         | 1.00       | 9         | 12        | 1.33       | 6          | 6          | 1.00       |
| Perissodactyla        | 18       | 21        | 1.17      | 17        | 23        | 1.35      | 17         | 44        | 2.59       | 18         | 45        | 2.50       | 18         | 45         | 2.50       |
| Pholidota             | 7        | 14        | 2.00      | 2         | 6         | 3.00      | 7          | 14        | 2.00       | 2         | 6         | 3.00       | 7          | 14         | 2.00       |
| Pilosa                | 9        | 22        | 2.44      | 10        | 23        | 2.30      | 9          | 20        | 2.22       | 10        | 26        | 2.60       | 9          | 20         | 2.22       |
| Primates              | 185      | 227       | 1.23      | 330       | 492       | 1.50      | 185        | 213       | 1.15       | 332       | 631       | 1.92       | 219        | 282.75     | 1.29       |
| Proboscidea           | 3        | 3         | 1.00      | 3         | 1         | 1.00      | 3          | 1         | 1.00       | 3         | 1         | 1.00       | 3          | 1          | 1.00       |
| Rodentia              | 721      | 1110      | 1.54      | 1725      | 3007      | 1.74      | 713        | 1466      | 2.06       | 1506      | 4004      | 2.66       | 1392       | 2457.75    | 1.86       |
| Scandentia            | 8        | 23        | 2.88      | 10        | 18        | 1.80      | 8          | 31        | 3.88      | 10        | 19        | 1.90       | 8          | 20         | 2.50       |
| Sirenia               | 3        | 3         | 1.00      | 3         | 1         | 1.00      | 3          | 3         | 1.00       | 3         | 3         | 1.00       | 3          | 3          | 1.00       |
| Tubulidentata         | 1        | 2         | 2.00      | 2         | 3         | 1.50      | 1          | 3         | 3.00      | 1         | 3         | 3.00       | 1          | 3          | 3.00       |
Table S2. Results of random forest predictive models. (A) Specifics of species used in each random forest classification model, including total number of species, number of species identified by delimitation as potentially containing hidden diversity, and number of species not identified by delimitation analysis as potentially containing hidden diversity. (B) Predictive model evaluation metrics, including model accuracy, the 95% confidence interval for model accuracy, positive predictive value (prediction value for species identified as hidden), and negative predictive value (prediction value for species not identified as hidden).

|                      | ABDG COI | ABDG cyt b | GMYC COI | GMYC cyt b | Consensus |
|----------------------|----------|------------|----------|------------|-----------|
| Total Species        | 1355     | 2127       | 1332     | 1898       | 1376      |
| Hidden Species       | 379      | 800        | 515      | 1043       | 478       |
| Not Hidden Species   | 976      | 1327       | 817      | 855        | 898       |
| B MODEL EVALUATION   |          |            |          |            |           |
| Model Accuracy       | 0.737    | 0.68       | 0.6429   | 0.6517     | 0.781     |
| Accuracy (95% CI)    | (0.6802, 0.7885) | (0.6333, 0.7241) | (0.5821, 0.7004) | (0.6014, 0.6996) | (0.7273, 0.8285) |
| Pos Predictive Value | 0.56667  | 0.6304     | 0.5571   | 0.6624     | 0.807     |
| Neg Predictive Value | 0.75833  | 0.6937     | 0.6735   | 0.6345     | 0.7742    |
Table S3. Final alignment subgroups. The final subgroups used for alignment of sequences belonging to each mammalian order are shown. Groups marked with * were not used for GMYC analysis of cyt b, due to nonconvergence.
| Family | Subfamily | Order |
|--------|-----------|-------|
| Chiroptidae | Eptesicoidea | Chiroptera |
| Ctenodactylidae | | |
| Ctenomyidae | | |
| Cuniculidae | | |
| Dasyproctidae | | |
| Didelphimorphia | | |
| Dipodomys | | |
| Dipotamidae | Cervidae | |
| Echimyidae | Delphidae | |
| Erethizontidae | Eschrichtiidae | |
| Geomyidae | Giraffidae | |
| Gliridae | Haplotragidae | |
| Heterocephalidae | Hyaenidae | |
| Heteromyidae | Hystrixidae | |
| Muridae | Deomyinae | Monotremata |
| | Gerbillinae | |
| | Lophiomyinae | |
| Murinae | Apodemus1 | Phocenidae |
| | Apodemus2 | Physteridae | |
| | Agomys+ | Plataniidae | |
| | Ratony+ | Postotragidae | |
| | Bullimus+ | Suidae | |
| | Grammomys+ | Tayassuidae | |
| | Hapalomys | Tragulidae | |
| | Hylocomys+ | Ziphidae | |
| | Malcomys | Cingulata | Chlamyphoridae |
| | Mastomys+ | Dasyproctidae | |
| | Mus | Dasyuromorphia | Dasyuridae |
| | Niviventer+ | Myrmecobiidae | |
| | Pragomys | Thylacinidae | |
| | Rattus | Dermoptera | Cynocephalidae |
| | Tokudaia | Didelphimorphia | Didelphidae |
| Otomyinae | Diprotodontia | Acrobatidae |
| | Nesomyidae | Barmyidae | |
| | Octodontidae | Hysiprymnontidae | |
| | Pedetidae | Macropodidae | |
| | Petrodromidae | Petauridae | |
| | Platycnemidomyidae | Phalangeridae | |
| | Sciuridae | Callotrichiinae | Potoroidae |
| | | Marmotina | Pseudochilidae |
| | | Proticerini | Tarsipedidae |
| | | Ratufinae | Vombatidae |
| | | Sciurillinae | Hyracoidae | Procaviidae |
| | | Sciurinae | Lagomorpha | Leporidae |
| | | Sciurotamias | Ochotonidae | |
| | | Sperophillina | Macroscelidea | Macroscelidae |
| | | Tamina | Microbiotheria | Microbiotheriidae |
| | | Xenini | Paucituberculata | Caenolestidae |
| | | Sminthidae | Peramelemorphia | Peramelidae |
| | | Spalacidae | Thylacomyidae | |
| | | Thylogaleomyidae | Perissodactyla | Equidae |
| | | Zapedidae | Rhinocerotidae | |
| | | Zerknerellidae | Tapiroidea | |
| Scandentia | Tupaiidae | Pholidota | Manidae |
| | | | Ptilinae | Bradypodidae |
| | | | | | Cynocephalidae |
| | | | | | Megalonychidae |
| | | | | | Myrmecophagidae |
| | | | | | Proboscidea | Elephantidae |
| | | | | | Chrysochloridae | Monotrema | Ornithorhynchidae |
| | | | | | | Tachyloceridae |
| Notoryctemorphia | Notoryctidae | |
Table S4. MACSE analysis of NUMTs. Results of MACSE analysis for the effects of NUMTs are show

| CLADE    | GENE | # TAXA IN ALIGNMENT (ORIGINAL) | # TAXA IN ALIGNMENT (MACSE) |
|----------|------|--------------------------------|----------------------------|
| Capromyidae | COI   | 19                             | 18                         |
| Capromyidae | cytb  | 63                             | 63                         |
| Sciurinae  | COI   | 75                             | 74                         |
| Sciurinae  | cytb  | 791                            | 800                        |
| Microtus   | COI   | 150                            | 148                        |
| Microtus   | cytb  | 2790                           | 2763                       |
Table S5. Alternative parameter values for ABGD relative gap width (X). Relative gap width values used for ABGD analysis of alignments in which X had to be adjusted for delimitation. All other alignments were delimited using the default value of X=1.5.

| ORDER       | ALIGNMENT         | cytb | COI |
|-------------|-------------------|------|-----|
| Carnivora   | Felidae           | X=0.5|     |
|             | Procynionidae     | X=1.0|     |
| Primates    | Atelidae          | X=0.5|     |
|             | Cebidae           | X=1.0|     |
|             | Cercopithecidae   | X=1.0|     |
|             | Daubentoniidae    | X=1.0|     |
|             | Indriidae         | X=0.5|     |
|             | Pithecidae        | X=1.0|     |
|             | Tarsiidae         | X=0.5|     |
| Rodentia    | Abrocomidae       | X=0.5|     |
|             | Anomaluridae      | X=0.5|     |
|             | Baiomynini        | X=0.5|     |
|             | Castoridae        | X=1.0|     |
|             | Chinchillidae     | X=0.5|     |
|             | Deomyinae         | X=0.5|     |
|             | Geomyidae         | X=1.0|     |
|             | Microtus          | X=0.25| |
|             | Myodes            | X=0.5|     |
|             | Myodini           | X=0.5|     |
|             | Cricetinae        | X=0.5|     |
|             | Reithrodontomyini | X=0.001| |
|             | Ichthyomyini      | X=1.0|     |
|             | Ctenomyidae       | X=0.5|     |
| Chiroptera  | Nycteridae        | X=0.5|     |
|             | Carollinae        | X=0.5| X=0.5|
|             | Cynopterinae      | X=0.5|     |
| Eulipotyphla| Not_Crocidura     | X=0.5|     |
|             | Erinaceidae       | X=1.0|     |
|             | Nectogalini       | X=1.0|     |
|             | Soricini          | X=0.5|     |
| Artiodactyla| Cervidae          | X=0.5|     |
|             | Delphinidae       | X=0.5|     |
|             | Ziphiiidae        | X=0.1|     |
| Dasyuromorpha| Dasyuridae       | X=1.0|     |
| Diprotodontia| Macropodidae     | X=0.5|     |
| Peramelemorpha| Peramelidae     | X=0.5|     |
Table S6. Species removed due to publication bias. Taxa excluded from random forest analysis due to the large number of publications returned in our Web of Science query at either species, genus, family, or order level.

| SPECIES NAME  | NUMBER OF PUBLICATIONS |   |   |   |
|---------------|------------------------|---|---|---|
|               | species | genus | family | order |
| Axis axis     | 225     | 498453| 1087 | 1730 |
| Axis calaminensis | 0      | 498453| 1087 | 1730 |
| Axis kuhlii   | 3       | 498453| 1087 | 1730 |
| Axis porcinus | 38      | 498453| 1087 | 1730 |
| Bos frontalis | 262     | 12909 | 1588 | 1730 |
| Bos grunniens | 603     | 12909 | 1588 | 1730 |
| Bos javanicus | 138     | 12909 | 1588 | 1730 |
| Bos mutus     | 37      | 12909 | 1588 | 1730 |
| Bos sauveli   | 18      | 12909 | 1588 | 1730 |
| Bos taurus    | 5667    | 12909 | 1588 | 1730 |
| Canis adustus | 25      | 17841 | 898  | 3366 |
| Canis anthus  | 11      | 17841 | 898  | 3366 |
| Canis aureus  | 261     | 17841 | 898  | 3366 |
| Canis latrans | 1391    | 17841 | 898  | 3366 |
| Canis lupus   | 3642    | 17841 | 898  | 3366 |
| Canis lycaon  | 39      | 17841 | 898  | 3366 |
| Canis mesomelas | 136    | 17841 | 898  | 3366 |
| Canis rufus   | 116     | 17841 | 898  | 3366 |
| Canis simensis | 62     | 17841 | 898  | 3366 |
| Castor canadensis | 732    | 11609 | 103  | 8496 |
| Castor fiber  | 481     | 11609 | 103  | 8496 |
| Cervus elaphus | 5676   | 7368  | 1087 | 1730 |
| Fossa fossana | 10      | 30513 | 31   | 3366 |
| Ia io         | 10      | 52560 | 1880 | 6576 |
| Macaca arctoides | 385   | 21630 | 439  | 99895 |
| Macaca assamensis | 90    | 21630 | 439  | 99895 |
| Macaca cyclopis | 86     | 21630 | 439  | 99895 |
| Macaca fascicularis | 5999  | 21630 | 439  | 99895 |
| Macaca fuscata | 1542   | 21630 | 439  | 99895 |
| Macaca hecki  | 6       | 21630 | 439  | 99895 |
| Macaca leonina | 48     | 21630 | 439  | 99895 |
| Macaca leucogenys | 6     | 21630 | 439  | 99895 |
| Macaca maura  | 16      | 21630 | 439  | 99895 |
| Macaca mulatta | 9959   | 21630 | 439  | 99895 |
| Macaca munzala | 16     | 21630 | 439  | 99895 |
| Macaca nemestrina | 1329  | 21630 | 439  | 99895 |
| Macaca nigra  | 148     | 21630 | 439  | 99895 |
| Macaca nigrescens | 3     | 21630 | 439  | 99895 |
| Macaca ochreata | 10     | 21630 | 439  | 99895 |
| Macaca pagensis | 2      | 21630 | 439  | 99895 |
| Macaca radiata | 527    | 21630 | 439  | 99895 |
| Macaca siberu  | 7       | 21630 | 439  | 99895 |
| Macaca silenus | 132     | 21630 | 439  | 99895 |
| Macaca sinica  | 48      | 21630 | 439  | 99895 |
| Macaca sylvanus | 416    | 21630 | 439  | 99895 |
| Macaca thibetana | 131    | 21630 | 439  | 99895 |
| Macaca tonkeana | 154    | 21630 | 439  | 99895 |
| Mops bakarii  | 0       | 10711 | 428  | 6576 |
| SPECIES NAME | NUMBER OF PUBLICATIONS | NUMBER | NUMBER | NUMBER |
|-------------|------------------------|--------|--------|--------|
| Mops brachypterus | 0 | 10711 | 428 | 6576 |
| Mops condlurus | 32 | 10711 | 428 | 6576 |
| Mops conicus | 0 | 10711 | 428 | 6576 |
| Mops demonstrator | 0 | 10711 | 428 | 6576 |
| Mops leonis | 0 | 10711 | 428 | 6576 |
| Mops leucostigma | 4 | 10711 | 428 | 6576 |
| Mops midas | 4 | 10711 | 428 | 6576 |
| Mops mops | 9 | 10711 | 428 | 6576 |
| Mops nanulus | 0 | 10711 | 428 | 6576 |
| Mops niagarae | 0 | 10711 | 428 | 6576 |
| Mops nivei | 0 | 10711 | 428 | 6576 |
| Mops petersoni | 0 | 10711 | 428 | 6576 |
| Mops sarasinorum | 0 | 10711 | 428 | 6576 |
| Mops spurrelli | 0 | 10711 | 428 | 6576 |
| Mops thersites | 2 | 10711 | 428 | 6576 |
| Mops trevori | 0 | 10711 | 428 | 6576 |
| Mus baoulei | 2 | 19486 | 2384 | 8496 |
| Mus booduga | 119 | 19486 | 2384 | 8496 |
| Mus bufo | 2 | 19486 | 2384 | 8496 |
| Mus callewaerti | 1 | 19486 | 2384 | 8496 |
| Mus caroli | 91 | 19486 | 2384 | 8496 |
| Mus cervicolor | 35 | 19486 | 2384 | 8496 |
| Mus cookii | 5 | 19486 | 2384 | 8496 |
| Mus crociduroides | 1 | 19486 | 2384 | 8496 |
| Mus cypriacus | 3 | 19486 | 2384 | 8496 |
| Mus famulus | 6 | 19486 | 2384 | 8496 |
| Mus fernandoni | 1 | 19486 | 2384 | 8496 |
| Mus fragilicauda | 4 | 19486 | 2384 | 8496 |
| Mus goundae | 1 | 19486 | 2384 | 8496 |
| Mus haussa | 0 | 19486 | 2384 | 8496 |
| Mus imberbis | 1 | 19486 | 2384 | 8496 |
| Mus indutus | 4 | 19486 | 2384 | 8496 |
| Mus lepidoides | 1 | 19486 | 2384 | 8496 |
| Mus macedonicus | 36 | 19486 | 2384 | 8496 |
| Mus mahomet | 8 | 19486 | 2384 | 8496 |
| Mus mattheyi | 6 | 19486 | 2384 | 8496 |
| Mus mayori | 2 | 19486 | 2384 | 8496 |
| Mus minutoides | 64 | 19486 | 2384 | 8496 |
| Mus muscooides | 5 | 19486 | 2384 | 8496 |
| Mus musculus | 7844 | 19486 | 2384 | 8496 |
| Mus neavei | 0 | 19486 | 2384 | 8496 |
| Mus nitidulus | 2 | 19486 | 2384 | 8496 |
| Mus orangiae | 0 | 19486 | 2384 | 8496 |
| Mus oubangui | 1 | 19486 | 2384 | 8496 |
| Mus pahari | 23 | 19486 | 2384 | 8496 |
| Mus phillipsi | 1 | 19486 | 2384 | 8496 |
| Mus platythrix | 29 | 19486 | 2384 | 8496 |
| Mus saxicola | 6 | 19486 | 2384 | 8496 |
| Mus setulosus | 3 | 19486 | 2384 | 8496 |
| SPECIES NAME      | NUMBER OF PUBLICATIONS | NUMBER | NUMBER | NUMBER |
|------------------|------------------------|--------|--------|--------|
| Mus setzeri      | 2                      | 19486  | 2384   | 8496   |
| Mus shortridgei  | 1                      | 19486  | 2384   | 8496   |
| Mus sorella      | 0                      | 19486  | 2384   | 8496   |
| Mus spicilegus   | 118                    | 19486  | 2384   | 8496   |
| Mus spretus      | 547                    | 19486  | 2384   | 8496   |
| Mus tenellus     | 3                      | 19486  | 2384   | 8496   |
| Mus terricolor   | 26                     | 19486  | 2384   | 8496   |
| Mus triton       | 4                      | 19486  | 2384   | 8496   |
| Mus vulcani      | 0                      | 19486  | 2384   | 8496   |
| Pan paniscus     | 1190                   | 82259  | 319    | 99895  |
| Pan troglodytes  | 6356                   | 82259  | 319    | 99895  |
| Rattus norvegicus| 5313                   | 8578   | 2384   | 8496   |
| Sus ahoenobarbus | 2                      | 16886  | 367    | 1730   |
| Sus barbatus     | 33                     | 16886  | 367    | 1730   |
| Sus cebifrons    | 16                     | 16886  | 367    | 1730   |
| Sus celebensis   | 13                     | 16886  | 367    | 1730   |
| Sus oliveri      | 10                     | 16886  | 367    | 1730   |
| Sus philippensis | 4                      | 16886  | 367    | 1730   |
| Sus scrofa       | 6189                   | 16886  | 367    | 1730   |
| Sus scrofa       | 6189                   | 16886  | 367    | 1730   |
| Sus verrucosus   | 12                     | 16886  | 367    | 1730   |
**Table S7. Proportion of hidden diversity represented by mammalian orders.** Information based on strict consensus of delimitation methods is displayed for each of the five most species rich orders of mammals. The columns ‘Not Hidden’ and ‘Hidden’ represent the number of species in each order classified as either not hidden or hidden, respectively. The column ‘Total Species’ represents the number of species from each order present in the consensus model. The column ‘Proportion of Model Hidden Diversity’ represents the proportion of the model’s total hidden diversity contained in each order.

| Order       | Not Hidden | Hidden | Total Species | Proportion of Model Hidden Diversity |
|-------------|------------|--------|---------------|-------------------------------------|
| Rodentia    | 296        | 215    | 511           | 0.454545455                         |
| Primates    | 90         | 18     | 108           | 0.038054968                         |
| Eulipotyphla| 60         | 41     | 101           | 0.086680761                         |
| Chiroptera  | 167        | 117    | 284           | 0.247357294                         |
| Carnivora   | 81         | 24     | 105           | 0.050739958                         |
Table S8. Results of statistical comparison between trait values for taxa containing hidden / no hidden species based on a Kruskal-Wallis test. Shown are counts of the number of taxa in each category, the median trait values, the $c^2$ statistic, and $p$-value for each test.

| CONSENSUS MODEL                  | Count | Median       | Kruskal-Wallis Test |
|----------------------------------|-------|--------------|---------------------|
|                                  | Not Hidden | Hidden | Not Hidden | Hidden | chi-squared | p-value |
| Adult Body Mass (g)              | 898   | 478     | 135        | 45     | 49.181     | 2.33E-12 |
| Range Area (km^2)                | 898   | 478     | 378072     | 1270779| 98.347     | < 2.2E-16|
| Occurrence Area (km^2)           | 898   | 478     | 8.525E+11  | 4.17E+12| 88.042     | < 2.2E-16|
| Recent Publications              | 898   | 478     | 5          | 9      | 41.184     | 1.39E-10 |
| bio18r (precipitation mm)        | 898   | 478     | 651        | 844    | 74.006     | < 2.2E-16|
| bio3r (isothermality)            | 898   | 478     | 25.8       | 37.3   | 89.285     | < 2.2E-16|
Dataset S1 (separate file). Standardized taxonomy. The version of the American Society of Mammalogists Mammal Diversity Database used to standardize taxonomy.

Dataset S2. (separate file). NCBI taxonomic name changes. Changes made to NCBI and Pantheria taxonomy used to standardize data to that used by the MDD (see dataset S1; 44).

Dataset S3. (separate file). PanTHERIA taxonomic name changes. Changes made to PanTHERIA taxonomy used to standardize data to that used by the MDD (see dataset S1; 16).

Dataset S4. (separate file). Predictor variable information. Detailed information on each predictor variable used in this study.

Dataset S5. (separate file). Minimum expected sequence divergence between species. Shown for all alignment subgroups (see Table S3) are the name of the taxon, the gene used in the analysis, the number of OTUs in the alignment, the sequence length, the threshold value of the GMYC analysis (measured in units of substitutions per site from the tip of the tree) and minimum % sequence divergence. For some alignments, marked with ‘n/a’ the minimum % sequence divergence was not calculated due to the low taxon sampling, a factor that is likely to produce erroneous results in a GMYC analysis (e.g., 11).

Dataset S6. (separate file). Variable importance results. Complete results of variable importance (MDA and Gini) analysis for all variables used in each random forest model as well as specific model parameters used.

Dataset S7. (separate file). GBIF DOIs. Complete list of DOIs for all GBIF downloads used in this study.
SI References

1. 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 Biology*. 17, (2019).

2. J. V. Lopez, N. Yuhki, R. Masuda, W. Modi, S. J. O’Brien, Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of molecular evolution*. 39, 174-190 (1994).

3. E. Hazkani-Covo, M. Zeller, W. Martin, Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet*, 6 (2010).

4. D. A. Triant, J. A. DeWoody, Molecular analyses of mitochondrial pseudogenes within the nuclear genome of arvicoline rodents. *Genetica*, 132, 21-33 (2008).

5. M. Galan, M. Pages, Cosson, Next-generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. *PloS one*, 7 (2012).

6. M. Pages, P. Chevret, M. Gros-Balthazard, S. Hughes, J. A. Alcover, R. Hutterer, C. Hänni, Paleogenetic analyses reveal unsuspected phylogenetic affinities between mice and the extinct *Malpaisomys insularis*, an endemic rodent of the Canaries. *PLoS One*, 7, (2012).

7. V. Ranwez, E. J. Douzery, C. Cambon, N. Chantret, F. Delsuc, MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular biology and evolution*, 35, 2582-2584 (2018).

8. F. Delsuc, V. Ranwez, Accurate alignment of (meta) barcoding data sets using MACSE. *Phylogenetics in the Genomic Era*, (2020).

9. J. Pons, T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, A. P. Vogler, Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol*. 55, 595–609 (2006).

10. R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. Du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C. H. Wu, D. Xie, C. Zhang, T. Stadler, A. J. Drummond, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol*. 15, 1–28 (2019).

11. 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).

12. C. Q. Tang, A. M. Humphreys, D. Fontaneto, T. G. Barraclough, Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods Ecol. Evol*. 5, 1086–1094 (2014).

13. G. Talavera, V. Dincă, R. Vila, Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. *Methods Ecol. Evol*. 4, 1101–1110 (2013).

14. S. Dellicour, J.-F. Flot, Delimiting Species-Poor Data Sets using Single Molecular Markers: A Study of Barcode Gaps, Haplowebs and GMYC. *Syst. Biol*. 64, 900–908 (2015).

15. D. Ahrens, T. Fujisawa, H.-J. Krammer, J. Eberle, S. Fabrizi, A. P. Vogler, Rarity and Incomplete Sampling in DNA-Based Species Delimitation. *Syst. Biol*. 65, 478–494
16. M. T. Monaghan, R. Wild, M. Elliot, T. Fujisawa, M. Balke, D. J. G. Inward, D. C. Lees, R. Ranaivosolo, P. Eggleton, T. G. Barraclough, A. P. Vogler, Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Syst. Biol.* **58**, 298–311 (2009).

17. A. Rambaut, A. J. Drummond, Tracer v1.5 (2013), (available at http://beast.bio.ed.ac.uk/Tracer).

18. T. Ezard, T. Fujisawa, T. G. Barraclough, SPLITS: species’ limits by threshold statistics (2009), (available at http://r-forge.r-project.org/projects/splits/).

19. R. C. Team, R: A language and environment for statistical computing (2020), (available at https://www.r-project.org/).

20. 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. R. Soc. B Biol. Sci.* **279**, 3678–3686 (2012).

21. M. Kekkonen, P. D. N. Hebert, DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Mol. Ecol. Resour.* **14**, 706–715 (2014).

22. N. Puillandre, A. Lambert, S. Brouillet, G. Achaz, ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol. Ecol.* **21**, 1864–1877 (2012).

23. A. J. Barley, R. C. Thomson, Assessing the performance of DNA barcoding using posterior predictive simulations. *Mol. Ecol.* **25**, 1944–1957 (2016).

24. D. L. Swofford, PAUP (phylogenetic analysis using parsimony). *Encycl. Genet. Genomics, Proteomics Informatics*, 1455–1455 (2008).

25. A. Zizka, D. Silvestro, T. Andermann, J. Azevedo, C. Duarte Ritter, D. Edler, H. Farooq, A. Herdean, M. Ariza, R. Scharn, S. Svantesson, N. Wengström, V. Zizka, A. Antonelli, CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10**, 744–751 (2019).

26. ASTER GDEM validation team, Aster Global Digital Elevation Model Version 2 (2011), (available at https://asterweb.jpl.nasa.gov/gdem.asp).

27. 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** (2005), pp. 1965–1978.

28. CIESIN, Socioeconomic Data And Applications Center (SEDAC) Gridded Populations of the World (GPW) (2016), (available at https://sedac.ciesin.columbia.edu/data/collection/gpw-v4).

29. DECRG, World Bank Development Economics Research Group (DECRG) Gross Domestic Product (2010), (available at https://datacatalog.worldbank.org/dataset/gross-domestic-product-2010/resource/added7173-a15f-4cee-8f07-0ad76ae389b0).

30. NOAA, NOAA Night Light Development Index (NLDI) (2006) (available at https://www.ngdc.noaa.gov/eog/dmsp/download_nldi.html).

31. European Environment Agency, World Database on Protected Areas (2012), (available at https://www.eea.europa.eu/data-and-maps/figures/overview-of-protected-areas-as).

32. 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.
33. European Space Agency, ESA GlobCover project (2009), (available at http://due.esrin.esa.int/page_globcover.php. 489 ).
34. R. J. Hijmans, Geosphere: Spherical Trigonometry (2016), (available at https://cran.r-project.org/package=geosphere. ).
35. R. . Hijmans, raster: Geographic Data Analysis and Modeling (2016), (available at https://cran.r-project.org/package=raster ).
36. R. Bivand, T. Keitt, B. Rowlingson, rgdal: Bindings for the Geospatial Data Abstraction Library (2017), (available at https://cran.r-project.org/package=rgdal).
37. H. Wickham, The Split-Apply-Combine Strategy for Data Analysis. J. Stat. Software; Vol 1, Issue 1 (2011), doi:10.18637/jss.v040.i01.
38. K. E. Jones, J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, W. Sechrest, E. H. Boakes, C. Carbone, C. Connolly, M. J. Cutts, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. P. Bininda-Emonds, J. L. Gittleman, G. M. Mace, A. Purvis, PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology. 90, 2648 (2009).
39. S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R. J. Stat. Software; Vol 1, Issue 3 (2011) (available at https://www.jstatsoft.org/v045/i03).
40. C. Penone, A. D. Davidson, K. T. Shoemaker, M. Di Marco, C. Rondinini, T. M. Brooks, B. E. Young, C. H. Graham, G. C. Costa, Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol. Evol. 5, 961–970 (2014).
41. C. Baker, wosr: Clients to the “Web of Science” and “InCites” APIs (2018), (available at https://cran.r-project.org/package=wosr).
42. 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).
43. B. A. Han, J. P. Schmidt, S. E. Bowden, J. M. Drake, Rodent reservoirs of future zoonotic diseases. Proc. Natl. Acad. Sci.12, 7039–44 (2015).
44. Mammal Diversity Database, (2021), (available at https://www.mammaldiversity.org).
45. M. Keesey, PhyloPic, (available at http://phylopic.org).