THERYA, 2026, Vol. 17(2):201-212 DOI: 10.12933/therya.2026.6275 ISSN 2007-3364

Status and shortfalls in the availability of genetic sequences and associated metadata for Mexican mammals

Luis D. Verde Arregoitia[1*], Martín Y. Cabrera Garrido[1], Julia Saravia[2,3], Lillian Draper Parker[4], and Fabricio Villalobos[1]

[1]Red de Biología Evolutiva, Instituto de Ecología A.C, Xalapa, Veracruz 91073, Mexico. Email: martin.cabrera@posgrado.ecologia.edu.mx (MYCG); fabricio.villalobos@inecol.mx (FV)
[2]Instituto de Ciencias Marinas y Limnológicas, Universidad Austral de Chile, Valdivia, Chile
[3]Millenium Institute Biodiversity of Antarctic and Subantarctic Ecosystems, BASE. E-mail: saraviajulia@gmail.com (JS)
[4]Laboratories of Molecular Anthropology and Microbiome Research, Department of Anthropology, University of Oklahoma, Norman, OK, USA. E-mail: lilly.parker@gmail.com (LDP)
[*]Corresponding author: luis@liomys.mx

Genetic information represents an important dimension of biodiversity beyond species or ecosystems, and genetic sequences are essential for all themes in biodiversity research. Genetic data are not exempt from gaps and biases, particularly in the tropics. Even for well-studied groups like mammals, many species remain understudied, and the completeness of underlying metadata for existing sequences is largely unknown. We quantified genetic data availability for 548 species of Mexican land mammals in the NCBI Nucleotide database. We evaluated the status and completeness of metadata regarding dates, locations, and voucher information associated with sequences, noting endemic species. We located genetic data for 90% of species, including sequences for 77% of Mexican endemics. Availability was skewed, with most species having very few sequences. In our sample, 316 of 496 species (64%) had at least one record with location, date, and voucher information, yet metadata quality was highly variable, and most records lacked key spatiotemporal details. We identified highly biased coverage in nucleotide sequence availability and relatively poor metadata quality for Mexican land mammals. Despite decades of sampling, decreasing costs, and advances in museomics and sequencing, major gaps and disparities remain. At a time of open science and open data, we insist on a research culture where sampling, sequencing, and adequate metadata recording happen concurrently.

Keywords: bioinformatics, entrez, GenBank, molecular, reproducibility

La información genética representa una dimensión clave de la diversidad más allá de especies o ecosistemas, y las secuencias son esenciales para investigar sobre biodiversidad. Los datos genéticos presentan vacíos y sesgos, especialmente en regiones tropicales. Incluso en grupos bien estudiados como los mamíferos, muchas especies permanecen poco estudiadas, y se desconoce la integridad de los metadatos asociados a sus secuencias. En el presente trabajo, cuantificamos la disponibilidad de datos genéticos para 548 especies de mamíferos terrestres mexicanos en la base de datos 'Nucleotide' de NCBI, evaluando la completitud de los metadatos sobre fechas, localidades e información de ejemplares asociada a las secuencias, tomando en cuenta la situación particular de las especies endémicas. Encontramos datos genéticos para el 90% de las especies, incluyendo 77% de las endémicas, con un sesgo en la disponibilidad, ya que la mayoría de las especies contaba con muy pocas secuencias. Encontramos que 316 de 496 especies (64 %) tenían al menos un registro con localidad, fecha, y ejemplar, aunque la calidad de los metadatos fue variable y la mayoría carecía de datos espaciotemporales. A pesar de décadas de muestreo, costos decrecientes y avances en museómica y secuenciación, persisten los vacíos. Identificamos un sesgo en la disponibilidad de secuencias nucleotídicas y una calidad relativamente deficiente en los metadatos para los mamíferos de México. En una era de ciencia y datos abiertos, insistimos en una cultura de investigación donde muestreo, secuenciación y registro adecuado de metadatos ocurran en paralelo.

Palabras clave: bioinformática, entrez, GenBank, molecular, reproducibilidad

Globally, DNA sequence data have been generated, collected, and aggregated for over 35 years. This endeavor has been supported and promoted by various national and international initiatives, including specialized infrastructure to collate and share this important information online (Arita *et al.* 2021). These collections of open genetic resources have an unquestionable role in answering research questions in phylogenetics, evolution, ecology, and conservation (Blaxter *et al.* 2022). For example, the GenBank (Benson *et al.* 2013) sequence database is a prime example of globally-centralized and standardized repositories of nucleotide sequence data meant for archiving, reuse, and re-analysis. GenBank is one of multiple primary repositories in the International Nucleotide Sequence Database Collaboration (INSDC), alongside nodes such as the European Nucleotide Archive (ENA) and the DNA DataBank of Japan (DDBJ); through international collaboration, all members synchronize and share their data to form a unified global resource. GenBank is produced and maintained by the National Center for Biotechnology Information (NCBI; part of the National Institutes of Health in the USA). GenBank is public and freely and openly accessible, and it has become a key component of modern biodiversity research (Leray *et al.* 2019).

Sequence data alone is incredibly valuable, but it is also a sensible scientific practice to record and share supporting metadata related to sample provenance, such

as spatiotemporal information regarding sampling location and date, to achieve its full potential (Dunnum *et al.* 2020). GenBank originated as and remains a public archive of genetic sequences and annotations and increased over time in the amount and nature of secondary information that can be recorded. As an archive for sequences produced across multiple disciplines (e.g., biomedical research, synthetic biology, human health, and molecular function), its metadata focuses on documenting the sequences compactly. However, these limited metadata meant to at least record a sampling location, remain useful for biodiversity science. Since 2005, GenBank has had the option for recording location data (country, locality, latitude, and longitude) for sequences in a dedicated metadata field, which can connect sequence data to the geographical locations where samples were collected (Gratton *et al.* 2017). Linking sequences to the museum vouchers sampled for sequencing (when applicable) is also possible. Vouchers in this context are the crucial link between nucleotide sequences, the taxonomic identity of the referred sequence, and physical specimens that are preserved, cataloged, and curated following standardized protocols (Buckner *et al.* 2021). When sequences are derived from physical specimens in scientific collections, these specimens can and should provide additional individual-level information such as sex, age, or body size.

Accurate documentation and usable metadata are crucial for enhancing the utility and reproducibility of molecular datasets (Galvao Elias *et al.* 2024). However, recording spatiotemporal and voucher metadata was optional when uploading sequences to databases like GenBank until late 2024 (NCBI Staff 2023). Even when this information adds context, value, and usability to primary molecular data, these data may not be shared—even when it was recorded. Such reluctance can be due to a variety of reasons relating to technical workflows, journal or funding body requirements (or lack thereof), or research culture (Sidlauskas *et al.* 2010). These issues are pervasive and not limited to metadata for genetic sequences, and work is underway to understand and lower the different barriers to data-sharing (Pope *et al.* 2015; Verde Arregoitia *et al.* 2020).

Existing studies have already revealed important gaps in metadata for nucleotide sequence data. In 2013, most sequences on GenBank lacked fundamental data such as geographical and vouchering information (Marques *et al.* 2013). Even when geographic information is provided for sequence data, it is often not precise enough for repurposing in spatially explicit analyses, mainly because coordinates or unambiguous locality details are not provided (Pope *et al.* 2015). In addition, missing temporal information also affects subsequent analyses. For example, a 2020 study of global genetic diversity in time and space had to focus on a small subset of sequences relative to all data available on GenBank because of lacking collection years in the metadata (Millette *et al.* 2020). Missing metadata for existing sequences undermines the lasting reuse and integration of public data and depletes massive amounts of resources in terms of funds, time, and wasted opportunities such as unique samples destroyed and expensive reagents consumed. Even when substantial efforts have been made to cross-reference available repositories of geographic and genetic information and ultimately bring spatial context to sequences (Pelletier *et al.* 2022), this is technically and computationally challenging. Cross-referencing massive batches of data from disparate sources involves interacting with application programming interfaces (APIs) and working efficiently with big data, and ultimately this process would need to be repeated periodically.

In addition to gaps in metadata for existing nucleotide sequences, there is great variation in the taxa or places that are better represented in public databases of genetic information (Pitogo 2025). A recent evaluation of public sequence data found that one quarter of terrestrial vertebrates worldwide had no genetic data and that mammals were the most sampled group, with 81% of species represented with at least one gene (Šmíd 2022). This study found that unsampled mammals tend to be taxa from Equatorial Africa, New Guinea, Sumatra, and inland Australia. Taxonomically, the least-sequenced families are Abrocomidae and Dasyproctidae (both Rodentia), with only 40% of their species sampled, followed by Gliridae (48%) and other less speciose families, including Tragulidae and Nycteridae, at ~50% completeness (Šmíd 2022).

In general, the species that remain unsampled are important to investigate, since they are not a random subset of diversity. Entire clades, regions, or countries may be altogether missing molecular data (Lim 2012). These unsampled taxa follow known gaps in biodiversity knowledge in which highly diverse tropical regions are less studied (Hughes *et al.* 2021). Research effort is also biased towards larger-bodied and more widely distributed species (Moura *et al.* 2024). It is also possible for regions or taxa to have high-quality genetic or genomic data, but without adequate metadata. This disconnect confounds the true nature of knowledge gaps and reduces the value of existing sequences, potentially compromising the results and inferences made with incomplete information.

Despite their high genetic sampling relative to other vertebrate groups, mammals are a group worth investigating in regard to the availability of nucleotide sequence data, as well as the status of the underlying metadata for existing sequences. In particular, terrestrial mammals that occur in Mexico represent an important opportunity for study given the high species richness and endemism (Vázquez and Gaston 2004) and worrying conservation status (Zamora-Gutierrez *et al.* 2019; Kennerley *et al.* 2021). Mexico is a megadiverse country and a tropical diversity hotspot for multiple vertebrate groups (Arita 1997; Nieves Delgado 2024), in addition to also having a rich cultural diversity (Flores-Santiago *et al.* 2024). Like most megadiverse countries, Mexico faces scientific challenges related to limited resources and funding in a context of

rapid biodiversity loss (Vilaça *et al.* 2024). Incomplete or biased data can misguide conservation, policy, and restoration efforts, since biodiversity conservation depends on the expansion of taxonomy and systematics research (Ruedas and Gardner 2025), which in turn relies heavily on molecular information (Hoban *et al.* 2024). It is therefore a pressing matter to quantify which species have sequences and which do not, gauge the size of the knowledge gap, and identify hidden shortfalls in quality that ultimately reduce the potential for existing sequences to be reused in future research.

Here, we assessed the status and availability of nucleotide sequence information and associated metadata for the land mammals of Mexico in the NCBI-GenBank Nucleotide database. After taxonomic harmonization, we ran a species-by-species search for any existing sequences and examined the underlying metadata. For each sequence, we checked whether the metadata included date, location, and associated voucher and, for those with location, if the sample was from Mexico. We found that ~10% of Mexican land mammals lack public nucleotide sequence data, while many others have very few sequences, and that many Mexican endemic species are critically undersampled. When evaluating metadata quality, we determined that the associated metadata for a significant proportion of existing sequences remains largely incomplete. Concerted efforts are urgently needed to increase sampling and sequencing and to improve metadata quality during deposition.

Materials and methods

We focused on extant, non-marine mammals (i.e., 'land mammals', which includes bats) known to occur in Mexico using the taxonomic scheme from the American Society of Mammalogists' Mammal Diversity Database (MDD) version 2.1. This database uses distributional data to track the native distribution of species in countries and continents. Accordingly, we filtered species on the 'countryDistribution' field, keeping only records that contained 'Mexico'. Of 590 species that met this condition, we excluded 38 cetaceans but retained pinnipeds except for the Galapagos fur seal *Arctocephalus galapagoensis* which is only recorded in Mexico from transient individuals (Tamayo-Millán *et al.* 2021). We also excluded the pocket mouse *Chaetodipus lineatus* which is listed in MDD 2.1, but following recent work, we considered this taxon a synonym of *Chaetodipus nelsoni* (Light *et al.* 2025). The taxonomic scheme that we followed also includes two extinct or possibly extinct species of rice rats (*Oryzomys nelsoni* and *Oryzomys peninsulae*), which we also excluded as they lack fundamental data and are absent from the NCBI taxonomy.

For the remaining 548 nonmarine species of Mexican mammals, we harmonized the scientific names using functions from the R package taxize (Chamberlain and Szocs 2013), matching and standardizing the taxon names to the taxonomy used by NCBI. This workflow matched the species names against the Global Names Verifier service (https://gni.globalnames.org/). We also used details on recent taxonomic revisions provided in the Mammal Diversity Database to find the corresponding identifiers for taxa in the NCBI taxonomy. In MDD v2.1, several taxa are recognized as distinct species because of recent taxonomic revisions, including species splits and the elevation of previously recognized subspecies to species status. This information is not typically reflected in the underlying NCBI taxonomy, so there is no straightforward way to link these taxa with an existing NCBI identifier. Additionally, linking existing sequences that ultimately correspond to newly recognized species involves extensive 'detective work' in the supporting data and figures of the publications linked to the taxonomic changes. For example, in many cases the updated identities of existing sequences are only reported secondarily in dendrograms, maps, appendices, or lists of specimens examined. We omitted 52 species in this situation from our search of sequences and metadata but provide this list as supplementary data.

With a final list of scientific names for the land mammals of Mexico and their corresponding NCBI unique identifiers when we could locate one, we used the R package rentrez v 1.2.4 (Winter 2017) to search the GenBank nucleotide database and then download and parse the metadata for each result. These search parameters were meant to return all available sequence data for a given taxon, from short fragments to complete genomes, including full/partial gene sequences, mRNA, tRNA, rRNA, genomic DNA, cDNA, expressed sequence tags, whole genome shotgun, and transcriptome shotgun assemblies (Sayers *et al.* 2019). We tailored our searches to exclude computationally generated (predicted) mRNA sequences and viral cRNA sequences associated with our target species because of internal host-pathogen links in the databases, because these two types of results are not informative for species-level biodiversity studies for our focal species. All programmatic queries of the entrez API were performed in batches between December 4 and December 9, 2025.

We searched for all available sequences for our study taxa, but when parsing the metadata for each result, we focused on three key pieces of information: the location of genetic sampling, the date, and the associated vouchers (i.e., permanently preserved specimens stored and maintained in an accessible collection) when applicable. These metadata are of central importance to most studies in ecology and evolution and the most likely to influence the reuse or repurposing of the nucleotide sequence data (Pope *et al.* 2015). We considered sequences produced from any type of sample but checked for voucher information given the importance of linking sequences to physical specimens. Sequences are also produced from other sources such as tissue or blood samples, scats, hair, or environmental DNA, and sample source or type was not a criterion in our searches.

In an age of high-throughput or massively parallel sequencing, it is commonplace to sometimes find tens of

thousands of sequences for a given species, which may be misleading for our purposes if these are small DNA or RNA fragments obtained from a single sample or specimen and uploaded separately to GenBank. These results may misrepresent the research interest for a species or region. Fortunately, entries derived from high-throughput methods in the Nucleotide database are usually linked with entries in the NCBI BioSample database, which keeps a record of biological isolates with unique physical properties. BioSample entries generally relate to multiple sequences, and a single BioSample may link to tens of thousands of DNA or RNA fragments. We attempted to link all our results with entries in BioSample so we could deduplicate observations by keeping only one sequence at random from a group of sequences from the same BioSample. This way we could make better comparisons of research interest and sequence availability across species by not having a small number of samples with tens of thousands of derived sequences mislead our inferences. All our reported numbers of sequences per species refer to unique BioSample entries and not raw fragment counts.

Additionally, we examined the GenBank accession numbers in our results using regular expressions to detect sequential patterns such as identifiers with the same prefixes and consecutive numerical suffixes (e.g., BV389079, BV389080, BV389081 and so on until BV399201 where "BV" is the prefix). These accession identifiers indicate batch submissions from a single study or project and likely reflect assembly sequences from a single sample rather than extensive species-level sampling efforts. We checked the accessions in our results for batches greater than 1000 sequences, which were then checked manually on the NCBI web portal to determine if these were Genome or Transcriptome assembly sequences. Assembly sequences were also deduplicated by keeping only one sequence from a batch (chosen at random) as long as these entries could be linked to a single study and all shared the same metadata.

For each accession in our final deduplicated dataset, we retrieved the full GenBank record in XML format. These records were parsed to extract the molecule type, organelle annotation (to infer genomic location), and gene name(s) from the gene qualifier of annotated features. When no explicit gene qualifier was present, we used the feature key itself as a descriptor (e.g., D-loop, rRNA, tRNA). We retained multi-gene records, which were primarily mitogenomes and multi-locus amplicons, as delimited strings. Lastly, we standardized all gene names to uppercase and unified common synonyms (e.g., COI/CO1 to COX1; CYT-B/cytochrome B to CYTB, etc.).

Results

Our final list of land mammals that occur in Mexico included 548 species, of which 203 are endemic. After taxonomic harmonization and manual standardization, we worked with 496 species. This set of 496 species spanned 12 orders and 39 families. Similar to global patterns, rodents (249 species, 50.2%) and bats (141 species, 28.4%) accounted for nearly 80% of all species. The best represented families in these two orders were Cricetidae (146 spp.), Phyllostomidae (58 spp.), Vespertilionidae (44 spp.), Heteromyidae (42 spp.), and Sciuridae (34 spp.). The remaining diversity includes various orders: Carnivora (40 spp.), Eulipotyphla (29 spp.), Lagomorpha (11 spp.), and few representatives each from Artiodactyla, Didelphimorphia, Primates, Cingulata, Pilosa, and Perissodactyla. Of these 496 species, 157 are endemic to Mexico. Rodents make up most of the endemism (119 species, 75.8%), followed by Chiroptera (17 spp.) and Eulipotyphla (15 spp.). For families, Cricetidae accounts for more than half of all endemics (82 spp.) and other rodent families Geomyidae (13 spp.), Heteromyidae (13 spp.), and Sciuridae (11 spp.) also contribute substantially. Non-rodent endemics are sparse, including mainly shrews (14 spp.), and small numbers of lagomorphs, carnivorans, and didelphids.

These 496 species all had at least one record in the Nucleotide database, meaning that the remaining 52 species or approximately 10% of species in our chosen taxonomic scheme, have no nucleotide sequence data, or none that are readily accessible. In total, we located 98,327 sequences after deduplicating entries linked to the same BioSample. We reduced this total further to 79,419 by deduplicating batch submissions linked to a Whole Genome Shotgun assembly for *Canis latrans* (10,123 sequences) and a Transcriptome Shotgun Assembly for *Desmodus rotundus* (8,787 sequences).

The 52 species with no sequences under our methods and search criteria spanned 5 orders (Rodentia: 22 species, Eulipotyphla: 17 species, Chiroptera: 5 species, Didelphimorphia: 4 species, and Lagomorpha: 4 species) and 9 families (Didelphidae, Leporidae, Cricetidae, Heteromyidae, Dasyproctidae, Sciuridae, Soricidae, Vespertilionidae, and Phyllostomidae), and was mainly represented by shrews and rodents. These unsampled species represent a combination of species in groups that are in constant taxonomic flux and well-recognized taxa that have simply not been sequenced yet. Roughly half of the species without sequences were rodents, and a third were shrews (*Sorex, Cryptotis, Notiosorex*). A small number of bats, rabbits, and opossum species also lacked sequences under our methods, mainly resulting from a lack of information to link existing sequences with now-outdated identities with their new taxonomic assignments. For a limited number of species, we found mentions of molecular data used in publications (e.g., studies reported amplifying gene fragments in their methods and using the resulting sequences in their analyses) without corresponding data on GenBank, the respective supplementary materials, or any other public platform or archive.

Sequence availability. The number of sequences per species varied greatly. The median number of sequences was 43, and these totals ranged from 1 to 3296. The distribution of total sequences per species was highly

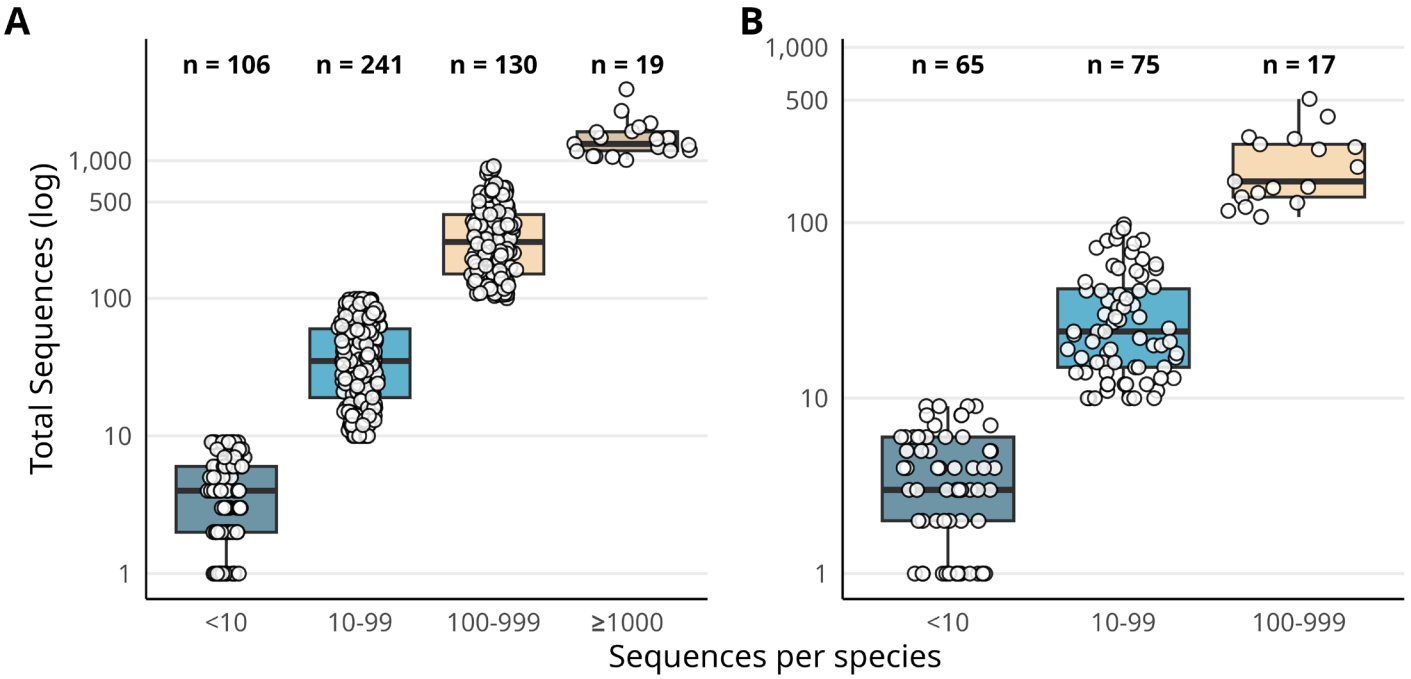

Figure 1. Distribution of nucleotide sequence counts per species, binned using cutpoints at 10, 100, 500, and 1,000 sequences. (A) All 496 species. (B) Mexican endemic species only. Box plots show medians and interquartile ranges; individual species are overlaid as points. Sample sizes per bin are indicated above boxes.

skewed—70% of species had fewer than 100 sequences, and a small remainder (~4%) made up the long tail of taxa with >1000 sequences (Figure 1A). Mexican endemics showed a similar skew: the median number of sequences was 13 (min 1, max 508) and 90% of species had fewer than 100 sequences (Figure 1B). No endemic species had more than 1000 sequences, although the deer mouse *Peromyscus mexicanus* had 508 sequences (driven mainly by 368 COX1 barcoding submissions), followed by the harvest mouse *Reithrodontomys sumichrasti* (403 sequences, mainly CTYB and COX1 but also a full mitogenome and nuclear markers) and the rice rat *Casiomys chapmani* (308 sequences in which nuclear markers such as CD14, RBP3, FUT4, DXS254E, NUP160, PRKC1 notably outnumber mitochondrial ones).

In general, there was a massive disparity in total sequences per species: a few widespread or abundant species received far more attention, while the bottom tier of species with only one or two sequences are mainly rare and endemic rodents and shrews. The species with the most hits in our searches were wide-ranging taxa that have historically attracted greater research interest relative to other mammals, including the white-tailed deer *Odocoileus virginianus,* the short-tailed bat *Carollia perspicillata*, the vampire bat *Desmodus rotundus,* and the deer mouse *Peromyscus leucopus* (Table 1). On the other hand, some species with only one sequence included threatened endemics such as the Zempoaltepec vole *Microtus umbrosus* and the surprising result of only finding one hit (a partial

Table 1. Top 20 mammal species by nucleotide sequence availability, ranked by total number of sequences.

Rank	Species	Family	Total Sequences	Rank	Species	Family	Total Sequences
1	*Odocoileus virginianus*	Cervidae	3,296	11	*Procyon lotor*	Procyonidae	1,299
2	*Carollia perspicillata*	Phyllostomidae	2,294	12	*Odocoileus hemionus*	Cervidae	1,251
3	*Desmodus rotundus*	Phyllostomidae	1,868	13	*Ateles geoffroyi*	Atelidae	1,189
4	*Peromyscus leucopus*	Cricetidae	1,749	14	*Phoca vitulina*	Phocidae	1,182
5	*Canis latrans*	Canidae	1,627	15	*Tamiasciurus douglasii*	Sciuridae	1,172
6	*Eptesicus fuscus*	Vespertilionidae	1,613	16	*Ursus americanus*	Ursidae	1,080
7	*Bos bison*	Bovidae	1,466	17	*Peromyscus truei*	Cricetidae	1,077
8	*Artibeus jamaicensis*	Phyllostomidae	1,461	18	*Artibeus lituratus*	Phyllostomidae	1,062
9	*Taxidea taxus*	Mustelidae	1,434	19	*Molossus molossus*	Molossidae	1,010
10	*Dasypus mexicanus*	Dasypodidae	1,321	20	*Sorex monticolus*	Soricidae	912

Table 2. Mammal species with a single nucleotide sequence (n = 21 species). Congeneric species are grouped for space. Flag indicates species endemic to Mexico (n = 16 of 21 species).

Species with 1 Sequence		
Species[1]	**Family**	**Order**
Cryptotis 🇲🇽 *(2 spp.): griseoventris, soricinus*	Soricidae	Eulipotyphla
Cryptotis goodwini	Soricidae	Eulipotyphla
Ictidomys mexicanus 🇲🇽	Sciuridae	Rodentia
Microtus 🇲🇽 *(3 spp.): oaxacensis, quasiater, umbrosus*	Cricetidae	Rodentia
Neotoma insularis 🇲🇽	Cricetidae	Rodentia
Peromyscus 🇲🇽 *(4 spp.): madrensis, pembertoni, polius, sagax*	Cricetidae	Rodentia
Procyon pygmaeus 🇲🇽	Procyonidae	Carnivora
Reithrodontomys hirsutus 🇲🇽	Cricetidae	Rodentia
Sciurus alleni 🇲🇽	Sciuridae	Rodentia
Sciurus aureogaster	Sciuridae	Rodentia
Sigmodon zanjonensis	Cricetidae	Rodentia
Sorex emarginatus 🇲🇽	Soricidae	Eulipotyphla
Sorex chiapensis	Soricidae	Eulipotyphla
Trachops coffini	Phyllostomidae	Chiroptera
Tylomys tumbalensis 🇲🇽	Cricetidae	Rodentia
[1] 🇲🇽 indicates Mexican endemic species		

cytochrome b sequence) for the ubiquitous Mexican gray squirrel *Sciurus aureogaster* (Table 2).

Sequence types and genetic markers. After deduplication, we fetched and parsed a total of 79,419 accessions. The majority of sequences were genomic DNA (75,898; 95.6%), with smaller proportions of mRNA (2,651; 3.3%) and transcribed or unassigned RNA (692; 0.9%). Mitochondrial sequences accounted for 61.4% of records; the remaining 38.6% were derived from the nuclear genome. By sequence type, 60.1% of records were classified as targeted single-locus or multi-locus gene sequences, 19.1% as mitochondrial control region/D-loop sequences, and 6.4% as nuclear introns. Smaller proportions included mRNA (3.4%), rRNA (2.7%), microsatellites (2.2%), mobile elements (1.0%), and complete mitogenomes (1.0%). The remaining 3.7% could not be assigned to a specific category.

We found gene annotations for 93.4% of records (74,151), of which 7.5% (5,920) contained multiple genes. For single-gene records, the two most common markers were CYTB (n = 12,212) and COX1 (n = 12,119), accounting for almost one third of all annotated records. The D-loop/control region was the next most frequently represented (8,464 records),

followed by nuclear markers including PRNP (2,672), RAG2 (770), FGB (568), RAG1 (462), and IRBP (460).

Metadata. We found that date, location, and voucher data for available sequences were rarely present and highly variable across species. Non-trivial proportions of species had incomplete information on collection date, location, or associated vouchers: 34% of species have no collection dates at all, 44% lack geographic coordinates, and 8% have no associated voucher specimens for any of their sequences.

We found an important gap in temporal information. Only 19% (15,102 of 79,419) of sequences had usable dates recorded in the appropriate metadata field. Sequences with dates were distributed unevenly across species. A concerning 170 species (34%) had no collection dates whatsoever. Most species (356 or 72%) had minimal temporal metadata (dates for < 25% of their sequences), and only 65 species (13%) had ≥ 50% temporal coverage. Parsing the sampling year (the most common time unit shared across sequences), we found that the collection dates span 125 years from 1900 to 2025, although most sequences originate from samples collected between 1980 and 2020. Sequences from historical specimens (pre-1980) are minimal (0.4% of all dated records).

Despite the importance of spatial context for gene sequences, 59% of sequences lack information in the 'country' field (in which more detailed locality information may be stored and not just the country), and 85% lack geographic coordinates. More sequences have locality information (median = 38% of records per species) than coordinates (median = 1.3% per species), and we found that 218 species have no coordinate data for their sequences.

Of all the sequences, roughly 42% (33,075 of 79,419) listed a voucher specimen in the metadata, and most species (455 of 496 or 92%) had at least one sequence with an associated voucher. On average, 60% of a species' sequences had voucher information, and the orders Rodentia and Eulipotyphla had the highest proportions of sequences with vouchers, with 67 and 86%, respectively. Bats (Order Chiroptera) had a lower proportion (43%) of sequences with vouchers despite their high diversity.

Discussion

Genetic data have been generated at an exponential rate for many years now (Pope *et al.* 2015). However, such data is highly taxonomically and geographically biased, and significant gaps remain. We found that a non-trivial proportion (~10%) of Mexican land mammals lack public nucleotide sequence data altogether, and another large fraction is critically undersampled. In general, when sequences do exist, the associated metadata describing vouchers, dates, and locations for existing sequences remains largely incomplete.

Of the 548 species of terrestrial mammals that occur in Mexico, we could not locate usable genetic data for 52, and of the remaining 496, one in five species (21%) have critically low representation (>10 sequences, a rough minimum for achieving usable estimates of diversity, structure, gene flow, or inbreeding, Scaketti *et al.* 2025) in the GenBank Nucleotide database. These low numbers of sequences for a large number of species, especially those endemic to Mexico, are likely insufficient to capture population-level variation, thus limiting the statistical power of the common methods used in evolution, systematics, or conservation (Paz-Vinas *et al.* 2025). These substantial gaps remain despite the boost in sequencing efforts derived from the Mexican Barcode of Life project (MEXBOL), part of a global initiative for establishing a system for species identification and discovery through mitochondrial gene sequences (Alvarez Castañeda *et al.* 2012). Other more recent and ongoing initiatives to generate high-quality genomic data for particular groups such as the Bat1k Project (Teeling *et al.* 2018) include concrete plans to sequence species that occur in Mexico, which should eventually help fill gaps in genetic knowledge.

We found that ~61% of all the sequences for Mexican land mammals target the mitochondrial genome. Two markers in particular (CYTB and COX1) accounted for almost one third of all sequences. This pattern reflects the historic use of mitochondrial phylogenetics and barcoding, but it also means that for many species, sequences are effectively limited to a single genomic compartment and a narrow set of loci. Nuclear markers had a substantial representation of ~38%, covering a heterogeneous assortment of markers that more likely reflect the priorities of different research groups. Nuclear coverage is uneven across taxa: some species may have several nuclear loci sequenced while others have none. Collectively, uneven taxonomic sampling across species is compounded by uneven genomic representation, which limits comparative studies unless these gaps are addressed.

With our approach of first linking taxa from our list of species to NCBI identifiers, we had to exclude 52 species that could not be linked to the NCBI reference taxonomy. This was either because the species have truly not been sequenced or because of recent taxonomic changes, and the distinction is not always clear. This is partly because of the dynamic nature of mammalian systematics and because our chosen taxonomic scheme (MDD) is constantly updated and recognizes taxonomic changes unrelated to molecular data. To help alleviate these issues with taxonomic identities, we call for authors of publications that describe novel taxa or suggest changes to species identities based on molecular data to make explicit links between existing sequences and their assignments in both existing and new taxonomic arrangements.

For Mexican land mammals, we found a skewed distribution of the total number of sequences (regardless of gene or type of sequence) per species, which indicates gaps and biases in research effort. This skew is likely driven in part by the particular agendas of individuals or research groups and by a positive feedback loop in which more data attracts more studies and more funding, which in turn

increases the disparity in sequences between well-studied species relative to undersampled taxa. With this exercise, we were able to identify that not only rare endemics but also widespread and abundant species, including the Mexican gray squirrel (*Sciurus aureogaster*), are being overlooked in molecular studies.

Incomplete metadata for sequences undermines their value and reuse in studies that need spatially and/or temporally explicit information. This includes phylogeographic studies or any efforts to calculate genetic diversity for any kind of spatial unit. Without date information it is also not possible to compare changes in genetic diversity or structure through time or work in modern ecological timescales, which is important for studies of genetic diversity in relation to recent land-use changes. In addition to undersampled taxa, the metadata of existing sequences may lack dates, locations, vouchers, or different combinations of these important fields.

We found numerous sequences that in their current version on GenBank are essentially disconnected from physical specimens, sampling locations, and temporal references. In a striking case, the blackish deer mouse *Peromyscus furvus* had 270 sequences (mainly mitochondrial markers CYTB and ND3 and a smaller complement of nuclear markers consistent with Sanger sequencing-based phylogeography or systematics studies), but none had any usable information about location (no coordinates, country, or locality) or date, and only 6 sequences had an associated voucher. This pattern is unlike other deer mouse species, which generally had high proportions of sequences with linked vouchers, as is often the case for rodents. Two bat species showed a similar problematic pattern: *Vampyrum spectrum* and *Macrotus californicus* both had fewer than 400 sequences but a less than 3% voucher rate and near-zero location/date data.

A key inference from this work is to state the need for further molecular sampling of Mexican mammals, either from new field work or using existing materials in natural history collections. We identified numerous species with either zero or critically low numbers of sequences (>10), which are mainly rodents and shrews and include multiple endemic taxa. Their low genetic representation on GenBank could make them priority species for sequencing, but other factors may be taken into consideration as well, including conservation status (prioritize threatened taxa), phylogenetic position (prioritize taxa in unresolved clades), or geographic range (prioritize missing parts of species' ranges).

It is worth mentioning that we aimed to quantify sequence availability and metadata completeness. In doing so, we deduplicated multiple reads from high-throughput sequencing methods from the same NCBI BioSample identifier and from shotgun assemblies and excluded predicted sequences and those of the parasites or pathogens associated with our species of interest. However, we did not filter our data by type of molecule, and thus our summary statistics include direct-submission entries from mRNA converted to cDNA and amplified for sequencing. These individual uploads may lack a BioSample identifier, either because they predate the introduction of the BioSample database in 2011 (Barrett *et al.* 2012) or because the submitting researchers are explicitly treating this information as sequence-only data not linked to a physical sample. These mRNA sequences may contribute to the skewed sequence counts and could be examined in more detail in future studies. Similarly, the resulting sequences may be refined further to only include molecular markers compatible with taxonomic, systematics, and biodiversity conservation studies, rather than sequences derived from physiological or immunological research.

Without question, we support existing calls for those involved in generating and submitting sequence data to record and carefully upload metadata such as collection date and location, at minimum, when depositing sequences. In a promising change, as of late 2024, databases in the International Nucleotide Sequence Database Collaboration (including GenBank) require submitters to include minimum information about sample locations and dates, which should ultimately improve the quality and completeness of metadata for future submissions. In parallel, research communities, journals, and funding agencies are requiring deposition of vouchers and sharing specimen information for publications built on biological materials (Colella *et al.* 2021). These changes in culture and policy should ultimately improve metadata quality, addressing part of the shortfalls that we identified for Mexican mammals. We encourage submitters and data managers to follow the increasingly detailed and user-friendly documentation for the relevant databases and to follow best practice guides (Renner *et al.* 2024), considering how to ensure that the sequences submitted can be reused in the future.

Policy changes are promising, but they do not solve the issue of incomplete metadata for hundreds of thousands of existing sequences. Although it is possible for the original depositors of sequences to directly update the metadata on their uploads, this is additional work, and updates like these are rare (Deck *et al.* 2017; Crandall *et al.* 2023). For case studies like ours that focused on a discrete and manageable set of species, it may be possible to take advantage of the voucher identifiers provided for a substantial number of species to enrich the metadata. Leveraging recent initiatives to digitize and publish information from biocollections into online platforms (Guralnick *et al.* 2016), dates and locations may be parsed and linked with nucleotide sequences. Numerous species had good voucher coverage but poor spatiotemporal metadata, such as *Peromyscus truei, Sturnira parvidens, Chaetodipus intermedius, Megascapheus umbrinus,* and *Peromyscus hylocetes*, which all have >250 sequences and >90% vouchering information but no sampling dates at all. Similarly, various rodents (many of which are endemic), including *Heteromys nelsoni, Megascapheus sheldoni, Megascapheus atrovarius,*

Chaetodipus goldmani, *Peromyscus truei*, and *Chaetodipus intermedius* all have high proportions of sequences with vouchers but no locations.

We found nucleotide sequence data for 90% of species of Mexican land mammals. However, this seemingly high proportion conceals gaps, biases, and incomplete metadata. These issues compromise the use we can give to these molecular data, especially when published sequences lack spatiotemporal context. Moving forward, we need to fill taxonomic gaps by generating new sequences with special care in recording the metadata that gives them biological meaning. Thanks to preserved specimens, it may be possible to enrich sequences with their missing context, but we insist on a shift in research culture towards rigorous and open recording of metadata. This will maximize the value of public molecular data in a way that can truly help us understand and conserve the mammals of this megadiverse country.

Acknowledgments

We dedicate this work to Livia S. León Paniagua, in appreciation for the mentorship and guidance that was instrumental for the personal and professional development of those of us fortunate enough to have been her students. We also wish to acknowledge Livia's contributions to the study of mammals in Mexico and towards establishing a healthy, friendly, and productive collaborative network. We would also like to thank Maximiliano M. Maronna for his active role in understanding and closing gaps in genetic sequence data and metadata that inspired this work and for feedback on this manuscript. Financial support: ANID Fondecyt Postdoctorado Grant Number 3230234 and SECIHTI scholarship 778388.

Declaration of Artificial Intelligence use

Initial data exploration was assisted by Databot, an LLM-based coding assistant integrated within Positron (version 2025.10.1, Posit PBC), using the Claude Sonnet 4.5 model (Anthropic PCB) served through a GitHub Copilot. All AI-generated R code was reviewed and validated before use.

Author contributions

Luis Darcy Verde Arregoitia and Fabricio Villalobos conceptualized the ideas and aims of the manuscript. Luis Darcy Verde Arregoitia and Martín Y. Cabrera Garrido collected the data. Luis Darcy Verde Arregoitia analyzed the data. Julia Saravia, Lillian Draper Parker, and Luis Darcy Verde Arregoitia interpreted the results. All authors contributed substantially to the final manuscript text and gave approval for submission and publication.

Supplementary data

Supplementary datasets including the species examined and their corresponding NCBI identifiers, plus the processed results for each sequence located, are available on OSF (https://doi.org/10.17605/OSF.IO/K4QNP).

Literature cited

Alvarez Castañeda ST, Lorenzo C, Rios Mendoza EP, Cortes-Calva P, Gutierrez ME, Ortega J, *et al*. 2012. DNA barcoding of mammals in Mexico: Implications for biodiversity. *The Open Zoology Journal* 5:18–26. https://doi.org/10.2174/1874336601205010018

Arita HT. 1997. The non-volant mammal fauna of Mexico: species richness in a megadiverse country. *Biodiversity and Conservation* 6:787–795. https://doi.org/10.1023/B:BIOC.0000010402.08813.ab

Arita M, Karsch-Mizrachi I, Cochrane G, on behalf of the International Nucleotide Sequence Database Collaboration. 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Research* 49(D1):D121–D124. https://doi.org/10.1093/nar/gkaa967

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, *et al*. 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research* 40(Database issue):D57–D63. https://doi.org/10.1093/nar/gkr1163

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, *et al*. 2013. GenBank. *Nucleic Acids Research 41*(Database issue):D36–D42. https://doi.org/10.1093/nar/gks1195

Blaxter M, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, *et al*. 2022. Why sequence all eukaryotes? *Proceedings of the National Academy of Sciences* 119:e2115636118. https://doi.org/10.1073/pnas.2115636118

Buckner JC, Sanders RC, Faircloth BC, and Chakrabarty P. 2021. The critical importance of vouchers in genomics. *eLife* 10:e68264. https://doi.org/10.7554/eLife.68264

Chamberlain S, and Szocs E. 2013. taxize - taxonomic search and retrieval in R. F1000Research.

Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ, and Mclean BS. 2021. The Open-Specimen Movement. *BioScience* 71:405–414. https://doi.org/10.1093/biosci/biaa146

Crandall ED, Toczydlowski RH, Liggins L, Holmes AE, Ghoojaei M, Gaither MR, *et al*. 2023. Importance of timely metadata curation to the global surveillance of genetic diversity. *Conservation Biology* 37:e14061. https://doi.org/10.1111/cobi.14061

Deck J, Gaither MR, Ewing R, Bird CE, Davies N, Meyer C, *et al.* 2017. The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLOS Biology* 15:e2002925. https://doi.org/10.1371/journal.pbio.2002925

Dunnum J, Malaney J, Cook J, Dunnum J, Malaney J, Cook J. 2020. Sustained impact of holistic specimens for mammalogy and parasitology in South America: Sydney Anderson's legacy. *Therya* 11:347–358. https://doi.org/10.12933/therya-20-1011

Flores-Santiago I, Baena ML, Delfín-Alfonso CA, Silva-Rivera E, and Pérez-Chacón JL. 2024. Perception and uses about mammals in México: a literature review. *Ethnobiology and Conservation* 13:25. https://doi.org/10.15451/ec2024-08-13.22-1-12

Galvao Elias S, Cervieri Guterres D, Weingart Barreto R, and Mario Martins do Vale H. 2024. GeneConnector: Unlocking the full potential of Genbank metadata. *IEEE Latin America Transactions* 22:99–105. https://doi.org/10.1109/TLA.2024.10412034

Gratton P, Marta S, Bocksberger G, Winter M, Trucchi E, and Kühl H. 2017. A world of sequences: Can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography 44: 475–486.* https://doi.org/10.1111/jbi.12786

Guralnick RP, Zermoglio PF, Wieczorek J, LaFrance R, Bloom D, and Russell L. 2016. The importance of digitized biocollections as a source of trait data and a new VertNet resource. *Database* 2016:baw158. https://doi.org/10.1093/database/baw158

Hoban S, Paz-Vinas I, Shaw RE, Castillo-Reina L, da Silva JM, DeWoody JA, *et al*. 2024. DNA-based studies and genetic diversity indicator assessments are complementary approaches to conserving evolutionary potential. *Conservation Genetics* 25:1147–1153. https://doi.org/10.1007/s10592-024-01632-8

Hughes AC, Orr MC, Ma K, Costello MJ, Waller J, Provoost P, *et al*. 2021. Sampling biases shape our view of the natural world. *Ecography* 44:1259–1269. https://doi.org/10.1111/ecog.05926

Kennerley RJ, Lacher Jr. TE, Hudson MA, Long B, McCay SD, Roach NS, *et al*. 2021. Global patterns of extinction risk and conservation needs for Rodentia and Eulipotyphla. *Diversity and Distributions* 27:1792–1806. https://doi.org/10.1111/ddi.13368

Leray M, Knowlton N, Ho S-L, Nguyen BN, and Machida RJ. 2019. GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences* 116:22651–22656. https://doi.org/10.1073/pnas.1911714116

Light JE, Siciliano-Martina L, Morley L, Castellanos AA, LaMonica L, and Hafner DJ. 2025. Taxonomic status of *Chaetodipus lineatus* and phylogeography of the *Chaetodipus nelsoni* species group. *Journal of Mammalogy* 106:1187–1198. https://doi.org/10.1093/jmammal/gyaf052

Lim BK. 2012. Preliminary Assessment of Neotropical Mammal DNA Barcodes: An Underestimation of Biodiversity. *The Open Zoology Journal* 5:10–17. https://doi.org/10.2174/1874336601205010010

Marques AC, Maronna MM, and Collins AG. 2013. Putting GenBank Data on the Map. *Science* 341(6152):1341–1341. https://doi.org/10.1126/science.341.6152.1341-a

Millette KL, Fugère V, Debyser C, Greiner A, Chain FJJ, and Gonzalez A. 2020. No consistent effects of humans on animal genetic diversity worldwide. *Ecology Letters* 23:55–67. https://doi.org/10.1111/ele.13394

Moura MR, Ceron K, Guedes JJM, Chen-Zhao R, Sica YV, Hart J, *et al*. 2024. A phylogeny-informed characterisation of global tetrapod traits addresses data gaps and biases. *PLOS Biology* 22:e3002658. https://doi.org/10.1371/journal.pbio.3002658

NCBI Staff. 2023. Coming Soon! Including Sample Location and Collection Date and Time for Sequences Submitted to GenBank and SRA. NCBI Insights. [accessed 14 Jan 2026] https://ncbiinsights.ncbi.nlm.nih.gov/2023/05/01/sequences-genbank-sra/

Nieves Delgado A. 2024. "México es un país megadiverso": Biocultural Heritage and Exceptionality in Mexican Ethnobiology. History of Anthropology Review. [accessed 5 Mar 2026] https://histanthro.org/notes/mexico-megadiverso/

Paz-Vinas I, Vandergast AG, Schmidt C, Leigh DM, Blanchet S, Clark RD, *et al*. 2025. Sparse genetic data limit biodiversity assessments in protected areas globally. *Frontiers in Ecology and the Environment* 23:e2867. https://doi.org/10.1002/fee.2867

Pelletier TA, Parsons DJ, Decker SK, Crouch S, Franz E, and Ohrstrom J. 2022. phylogatR: Phylogeographic data aggregation and repurposing. *Molecular Ecology Resources* 22:2830–2842. https://doi.org/10.1111/1755-0998.13673

Pitogo KME. 2025. Gaps and biases in vertebrate wildlife genetics from a global biodiversity hotspot. *Environmental Conservation* 52:127–138. https://doi.org/10.1017/S0376892925000141

Pope LC, Liggins L, Keyse J, Carvalho SB, and Riginos C. 2015. Not the time or the place: the missing spatio-temporal link in publicly available genetic data. *Molecular Ecology* 24:3802–3809. https://doi.org/10.1111/mec.13254

Renner SS, Scherz MD, Schoch CL, Gottschling M, and Vences M. 2024. Improving the gold standard in NCBI GenBank and related databases: DNA sequences from type specimens and type strains. *Systematic Biology* 73:486–494. https://doi.org/10.1093/sysbio/syad068

Ruedas LA, and Gardner SL. 2025. Biodiversity conservation depends on the expansion of taxonomy and systematics research. *Journal of Mammalogy* 106:1495–1512. https://doi.org/10.1093/jmammal/gyaf067

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank. *Nucleic Acids Research* 47(D1):D94–D99. https://doi.org/10.1093/nar/gky989

Scaketti M, Sujii PS, Alves-Pereira A, Schwarcz KD, Francisconi AF, Moro MS, *et al.* 2025. Sample Size Impact (SaSii): An R script for estimating optimal sample sizes in population genetics and population genomics studies. *PLOS One* 20:e0316634. https://doi.org/10.1371/journal.pone.0316634

Sidlauskas B, Ganapathy G, Hazkani-Covo E, Jenkins KP, Lapp H, McCall LW, *et al*. 2010. Linking big: The continuing promise of evolutionary synthesis. *Evolution* 64:871–880. https://doi.org/10.1111/j.1558-5646.2009.00892.x

Šmíd J. 2022. Geographic and taxonomic biases in the vertebrate tree of life. *Journal of Biogeography* 49:2120–2129. https://doi.org/10.1111/jbi.14491

Tamayo-Millán CJ, Ahumada-Sempoal MÁ, Cortés-Gómez A, Chacón-Romo Leroux IM, Bermúdez-Díaz D, *et al*. 2021. Molecular identification of the first Galapagos fur seal (*Arctocephalus galapagoensis*) reported on the central coast of Oaxaca. *Ciencias Marinas* 47:201–209. https://doi.org/10.7773/cm.v47i3.3184

Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, *et al*. 2018. Bat biology, genomes, and the Bat1K Project: To generate chromosome-level genomes for all living bat species. *Annual Review of Animal Biosciences* 6:23–46. https://doi.org/10.1146/annurev-animal-022516-022811

Vázquez L-B, and Gaston KJ. 2004. Rarity, commonness, and patterns of species richness: the mammals of Mexico. *Global Ecology and Biogeography* 13:535–542. https://doi.org/10.1111/j.1466-822X.2004.00126.x

Verde Arregoitia LD, Teta P, and D'Elía G. 2020. Patterns in research and data sharing for the study of form and function in caviomorph rodents. *Journal of Mammalogy* 101:604–612. https://doi.org/10.1093/jmammal/gyaa002

Vilaça ST, Vidal AF, Pavan AC, Silva BM, Carvalho CS, Povill C, *et al*. 2024. Leveraging genomes to support conservation and bioeconomy policies in a megadiverse country. *Cell Genomics* 13:4. https://doi.org/10.1016/j.xgen.2024.100678

Winter DJ. 2017. rentrez: an R package for the NCBI eUtils API. *The R Journal* 9:520–526.

Zamora-Gutierrez V, Amano T, and Jones KE. 2019. Spatial and taxonomic biases in bat records: Drivers and conservation implications in a megadiverse country. *Ecology and Evolution* 9:14130–14141. https://doi.org/10.1002/ece3.5848

Associated editors: Giovani Hernández Canchola and Pablo Colunga Salas
Submitted: January 16, 2026; Reviewed: March 3, 2026
Accepted: April 11, 2026; Published online: May 29, 2026