scientific reports

Check for updates

OPEN

Taxonomic assessment of two wild house mouse subspecies using whole-genome sequencing

Raman Akinyanju Lawal[1]✉, Verity L. Mathis[2], Mary E. Barter[1], Jeremy R. Charette[1], Alexis Garretson[1,3] & Beth L. Dumont[1,3]✉

The house mouse species complex (*Mus musculus*) is comprised of three primary subspecies. A large number of secondary subspecies have also been suggested on the basis of divergent morphology and molecular variation at limited numbers of markers. While the phylogenetic relationships among the primary *M. musculus* subspecies are well-defined, relationships among secondary subspecies and between secondary and primary subspecies remain less clear. Here, we integrate de novo genome sequencing of museum-stored specimens of house mice from one secondary subspecies (*M. m. bactrianus*) and publicly available genome sequences of house mice previously characterized as *M. m. helgolandicus*, with whole genome sequences from diverse representatives of the three primary house mouse subspecies. We show that mice assigned to the secondary *M. m. bactrianus* and *M. m. helgolandicus* subspecies are not genetically differentiated from *M. m. castaneus* and *M. m. domesticus*, respectively. Overall, our work suggests that the *M. m. bactrianus* and *M. m. helgolandicus* subspecies are not well-justified taxonomic entities, emphasizing the importance of leveraging whole-genome sequence data to inform subspecies designations. Additionally, our investigation provides tailored experimental procedures for generating whole genome sequences from air-dried mouse skins, along with key genomic resources to inform future genomic studies of wild mouse diversity.

House mice (*M. musculus*) are the premiere mammalian model system for biomedical research and an important natural model system for ecological and evolutionary studies[1,2]. House mice emerged from an ancestral population in the Indian subcontinent less than 3 million years ago[3,4] and subsequently expanded out of this ancestral region, giving rise to three primary subspecies[5–8]. *M. m. domesticus* (DOM) is native to Western Europe, *M. m. musculus* (MUS) is present across Eastern Europe and Siberia, and *M. m. castaneus* (CAS) extends across South and Southeast Asia. Aided by human dispersal in recent history, house mice have subsequently expanded their footprint outside of these native ranges, colonizing all major continents except Antarctica and invading many remote oceanic islands.

Beyond these three primary subspecies, a number of secondary house mouse subspecies have been suggested on the basis of distinct morphology[9–11] and surveys of limited numbers of molecular markers[6,8,11–15]. For example, mice from Yemen and Madagascar have been assigned to *M. m. gentilulus* due to their small body size and distinct mitochondrial lineage[12,13]. Mice from Heligoland, a small German archipelago island in the North Sea, have been characterized as *M. m. helgolandicus* on the basis of their unique skull morphology, distinct mitochondrial D-loop haplotype, and allelic variation at four nuclear loci[11,16]. Similarly, a white belly coat color phenotype[17] and mitochondrial sequence analysis have supported the assignment of house mice from the Indo-Iranian valley to the subspecies *M. m. bactrianus*[15,18]. *M. m. musculus* and *M. m. castaneus* naturally hybridize where their ranges overlap in Japan, and these hybrids have been designated as a distinct subspecies, *M. m. molossinus*[19]. At least six other secondary subspecies of *M. musculus* have been named, including: *M. m. albula, M. m. brevirostris, M. m. homourus, M. m. isatissus, M. m. wagneri,* and *M. m. gansuensis*[20].

Prior studies have leveraged powerful genomic approaches to investigate the evolutionary relationships between the three primary house mouse subspecies, establishing a sister relationship between *M. m. castaneus* and *M. m. musculus*[21–25]. In contrast, the phylogenetic relationships among secondary house mouse subspecies, including their relationships to the primary house mouse subspecies, remain poorly understood. Currently, all secondary subspecies assignments are informed by sparse molecular data, begging the question of whether

[1]The Jackson Laboratory, 600 Main Street, Bar Harbor 04609, USA. [2]Florida Museum of Natural History, University of Florida, 1659 Museum Road, Gainesville, FL 32611, USA. [3]Graduate School of Biomedical Sciences, Tufts University, 136 Harrison Ave, Boston, MA 02111, USA. ✉email: lawalakinyanju@yahoo.com; beth.dumont@jax.org

distinct subspecies labels are truly warranted. Indeed, some subspecies designations have been proposed based on only mitochondrial genetic markers (e.g.,[13]), despite knowledge that mitochondrial inferences provide a strict matrilineal reconstruction of subspecies relationships and may be misleading in the face of sex differences in dispersal or evolutionary history[26]. Subspecies assignments relying on genotypes at limited numbers of nuclear loci are similarly problematic, as the chance sampling of population-private alleles can be misinterpreted as evidence supporting new subspecies designations. Overall, the legitimacy of specific secondary house mouse subspecies remains debatable, with genome-scale investigations standing to provide an ultimate resolution to their taxonomic status.

The democratization of DNA sequencing has lowered the cost barrier to genomic investigations across biological disciplines, including phylogenetics and systematics. As a result, sample availability and sample collection have emerged as comparatively greater challenges for many studies. Although house mice are ubiquitous across the globe, many secondary subspecies inhabit small, isolated regions that are not readily accessible. Procedural barriers, including securing necessary permits and customs paperwork, also pose formidable challenges to wild sample collection. As an alternative to live sample collection, many natural history museums harbor large holdings of air-dried animal skins that could provide necessary tissue material for 'omics studies. However, exposure to air, light, and chemicals during long-term storage can lead to extensive DNA degradation and damage, posing technical challenges for DNA isolation, amplification, and sequencing from preserved tissues[27,28]. Protocols for DNA extraction have been developed for historical mammalian specimens (e.g.,[29]), and both genotyping approaches and low coverage whole-genome sequencing methods have been adapted for use on low-quality, archived biospecimens (reviewed in[28,30]). However, to our knowledge, no moderate- to high-coverage whole-genome sequences (WGS) have been generated from archived non-human mammalian tissues, raising uncertainties about the technical feasibility of this approach (but see[31] for exome sequencing and[32] for low coverage WGS of mammalian species).

Here, we combine published WGS from wild mice[33–35] with de novo genome sequencing of a strategic set of museum-preserved specimens to address the validity of two secondary house mouse subspecies designations—*M. m. bactrianus* and *M. m. helgolandicus*. These subspecies labels have been assigned on the basis of subtle morphological differences from the primary house mouse subspecies and divergence at both mitochondrial and limited numbers of nuclear DNA markers[11,15,18]. Our investigations offer a conclusive resolution to the taxonomic status of mice from these subspecies, arguing that these taxonomic assignments are not well justified based on genomic data. Further, our work provides a crucial proof-of-concept demonstration that moderate-coverage whole genome sequences can be readily obtained from archived mammalian tissue.

Results and discussion

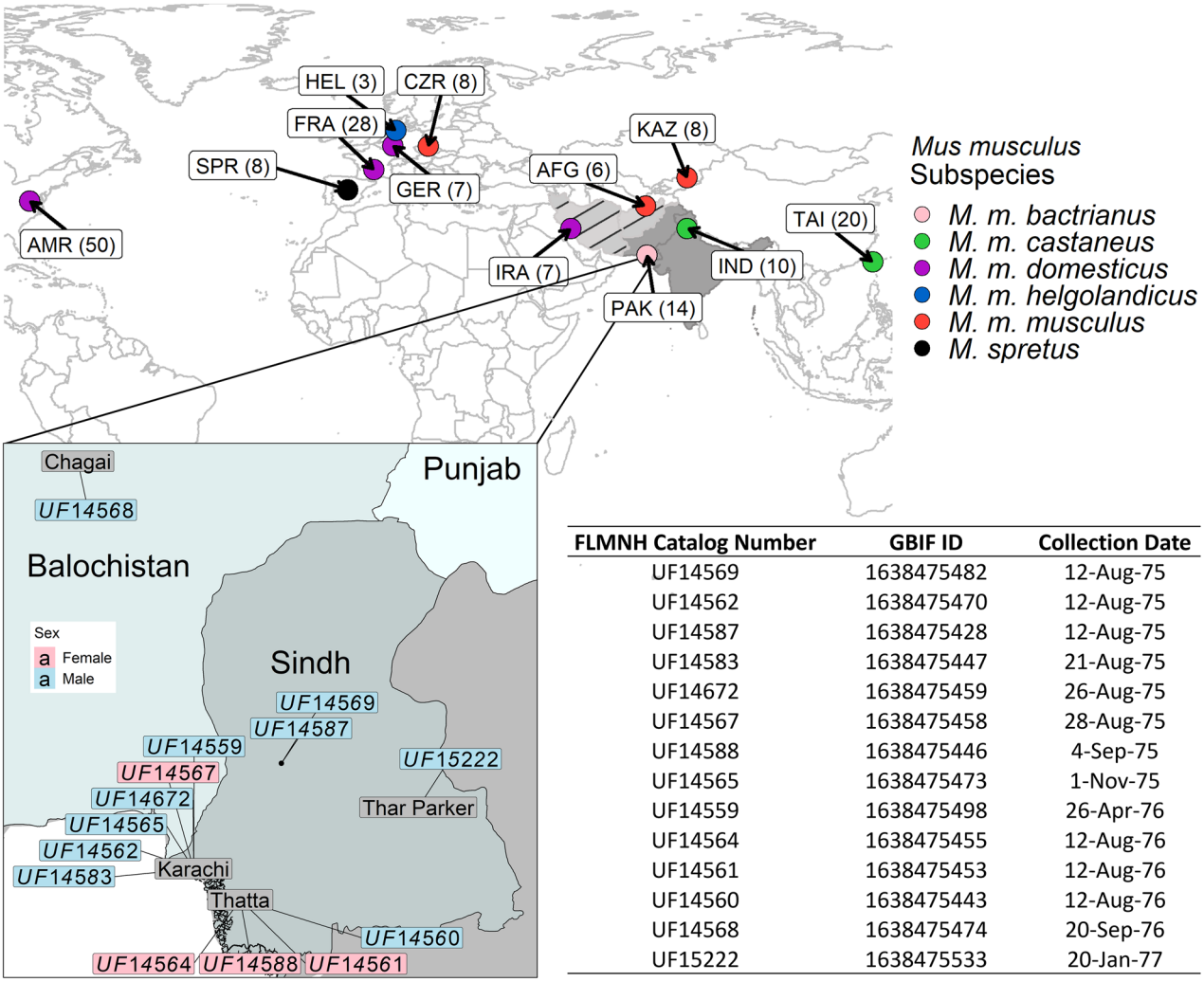
Whole-genome sequencing of museum samples of wild-caught mice from Pakistan. We obtained air-dried skin snips from 14 ~46-year-old museum-preserved specimens of *M. m. bactrianus* initially collected across several counties in Pakistan (PAK) and maintained by the Florida Museum of Natural History (Fig. 1). This geographic sample region overlaps with the presumed ancestral homeland of house mice[6]. We adapted an existing protocol for DNA extraction from museum-preserved mammalian pelts[29], isolating between 0.06 and 1 µg of fragmented DNA (< 500 bp) per specimen. DNA samples were processed for Illumina library preparation and whole-genome sequenced to 17× – 45× coverage using 100 bp paired-end reads (Supplementary Fig. S1a). On average, ~ 99% of sequenced reads from each of these mouse genomes were successfully mapped to the mm10 reference (Supplementary Fig. S1b), suggesting little exogenous contamination of our samples. This represents a dramatic improvement in mapping rate relative to other sequencing studies of archived biospecimens and ancient DNA (~ 20–75%)[28,32]. The proportion of bases supported by at least 5 reads ranges from 66 to 96% across samples (Supplementary Fig. S1c), indicating excellent genomic coverage.

Ancient and museum-stored specimens are known to accumulate C → T and A → G mutations due to post-mortem DNA deamination[36]. This spontaneous damage can lead to sequence biases if not properly accounted for. We determine that, across the PAK genomes, the deamination rate for nucleotide misincorporation is ~ 0.1%, with nearly all damage confined to the first and last 5 bp of a given read (see Methods; Supplementary Fig. S2). By comparison, the frequency of post-mortem DNA damage in a sample of ancient (~ 13 k years old) stickleback fish was 30%[37] and is typically around 50% for Neanderthal DNA[38]. Although there is minimal post-mortem DNA damage in our sequenced PAK genomes, we took the conservative approach of hard clipping the terminal 5 bp on both the 5′ and 3′ ends of every read to eliminate potential artifacts (Supplementary Fig. S1d).

After applying these quality control steps, performing variant calling, and imposing basic hard filters to eliminate low-quality variant calls (see Methods), we identified a total of 81,338,251 autosomal SNPs and 8,802,753 short indels across the 14 PAK genomes. This corresponds to ~ 1 variant every 30 bases relative to the mm10 reference genome.

Re-evaluating phylogenetic relationships among *M. musculus* subspecies. We combined the 14 PAK genome sequences with 3 previously published genome sequences of *M. m. helgolandicus* from Heligoland (HEL) and 152 publicly available whole-genome sequences from wild-caught mice from multiple populations of *M. m. castaneus* (CAS), *M. m. musculus* (MUS), and *M. m. domesticus* (DOM)[33–35]. Using this comprehensive set of 169 whole genome sequences, we evaluated the genetic relationships among these five putative subspecies using a multi-pronged approach.

First, we conducted a principal component analysis to visualize genetic similarities across samples (Fig. 2a). The first two principal components, accounting for 26.14% and 18.21% of the total variance, reveal just three discrete clusters (G1–G3) corresponding to the primary DOM, MUS, and CAS subspecies. Mice from the HEL population group within DOM, while PAK is nearly indistinguishable from the CAS populations (Fig. 2a).



FLMNH Catalog Number	GBIF ID	Collection Date
UF14569	1638475482	12-Aug-75
UF14562	1638475470	12-Aug-75
UF14587	1638475428	12-Aug-75
UF14583	1638475447	21-Aug-75
UF14672	1638475459	26-Aug-75
UF14567	1638475458	28-Aug-75
UF14588	1638475446	4-Sep-75
UF14565	1638475473	1-Nov-75
UF14559	1638475498	26-Apr-76
UF14564	1638475455	12-Aug-76
UF14561	1638475453	12-Aug-76
UF14560	1638475443	12-Aug-76
UF14568	1638475474	20-Sep-76
UF15222	1638475533	20-Jan-77

Figure 1. Map showing the approximate geographic sampling locations of house mouse populations profiled by whole-genome sequencing. The house mouse ancestral region extends from India "IND" and Pakistan "PAK" (dark grey area) and may include the broader region from Iran "IRA" and Afghanistan "AFG" (light grey area with black markings). All genome sequences except for PAK were retrieved from public databases[33–35]. The additional sequences include populations from America (AMR), France (FRA), Germany (GER), Heligoland (HEL), Kazakhstan (KAZ), and Czech Republic (CZR), Taiwan (TAI), and *M. spretus* from Spain (SPR). Each population sample sizes are indicated in the bracket. The enlarged plot inset shows the locations of PAK samples collected across four counties in the Sindh region of Pakistan. Florida Museum of Natural History (FMNH) Catalog numbers, Global Biodiversity Information Facility (GBIF) identifiers[41], and collection dates for each PAK sample are presented in the table.

Additional approaches based on average pairwise distance among individuals (Fig. 2b), allele sharing distance (Fig. 2c), and co-ancestry (Fig. 2d) also support three discrete phylogenetic units aligning to the three primary *M. musculus* subspecies. These findings are in agreement with a recent study that similarly found support for only three primary *M. musculus* clades using an independent set of whole genome sequences from mice across the globe[25].

Wild mice may experience substantial gene flow between populations and subspecies[39], which could obscure subspecies relationships. To evaluate the possible influence of gene flow on our inferred subspecies groupings, we used TreeMix to assess the robustness of our findings under multiple distinct migration models[40] (see Methods). We recovered identical genealogical relationships both in the absence of gene flow and under different migration scenarios ($p < 1 \times 10^{-300}$; Fig. 3). Importantly, no tested migration model altered the composition of the three core subspecies groups or offered support for the *M. m. helgolandicus* or *M. m. bactrianus* subspecies designations.

We find no compelling genetic evidence to justify a unique subspecies designation for mice from HEL or PAK. These conclusions contradict earlier findings based on limited numbers of genomic markers[11] or single loci[15], underscoring the power and importance of leveraging whole-genome data to inform subspecies designations.

Our investigations suggest that PAK and HEL are populations of *M. m. castaneus* and *M. m. domesticus*, respectively. Combining these populations with their modified subspecies groupings, we next partitioned the genome into 13,122 blocks, constructed phylogenetic trees from each partition, and estimated the percentage

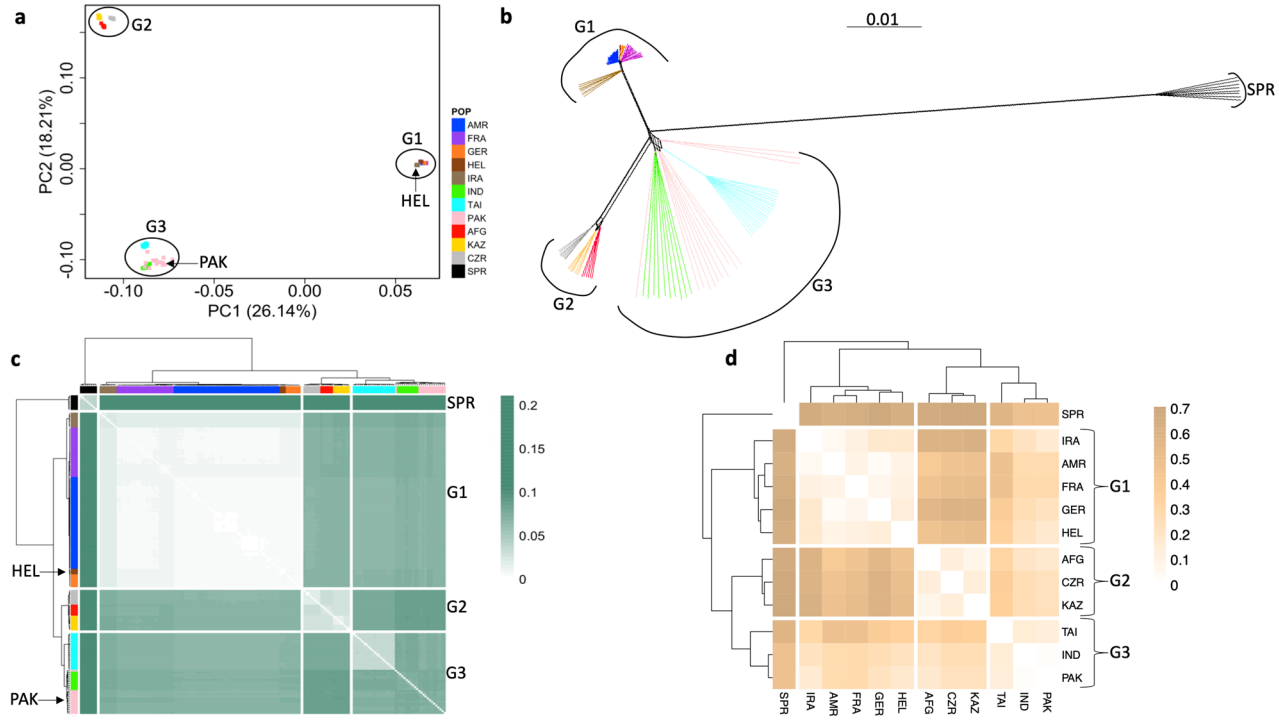


Figure 2. Genetic relationships among five putative *M. musculus* subspecies reveal three taxonomic groups (G1–G3). Genetic relationships were assessed via (**a**) principal component analysis, (**b**) a phylogenetic tree constructed from a pairwise distance matrix, (**c**) allele sharing distance, and (**d**) co-ancestry based on F_{ST}. The color legend on the right side of panel (**a**) is also applicable to panels (**b**) and (**c**). G1 groups populations belonging to the DOM subspecies: America (AMR), France (FRA), Germany (GER), Heligoland (HEL), and Iran (IRA). G2 groups populations of MUS: Afghanistan (AFG), Kazakhstan (KAZ), and Czech Republic (CZR). G3 groups CAS populations: India (IND), Taiwan (TAI), and Pakistan (PAK). *M. spretus* (SPR) is used as an outgroup.

of the genome supporting each of the three possible topologies relating the three primary *M. musculus* subspecies. Overall, the topology placing CAS and MUS as sister is the most abundant in the genome, capturing a total weight of 34.20% (autosomes), 36.70% (X chromosome), and 44.20% (MT) (Fig. 4). This finding validates prior conclusions about house mouse subspecies phylogenetic relationships based on representative inbred strain genomes[21,22] and smaller samples of wild mice[23,25]. Notably, our use of a broader set of wild house mouse genomes sampled from the presumed ancestral region offers a more comprehensive survey of ancestral house mouse diversity, leading to increased power to detect alleles that are still segregating across multiple subspecies due to incomplete lineage sorting. In turn, this improved ancestral sampling is expected to allow more accurate estimates of the extent of phylogenetic discordance across the *M. musculus* genome, as well as the proportional representation of each topology.

Conclusions

Using both new and published whole-genome sequences from diverse wild house mice, we addressed support for two secondary *M. musculus* subspecies assignments: *M. m. bactrianus* and *M. m. helgolandicus*. We showed that mice from Pakistan previously assigned to *M. m. bactrianus* are genetically indistinguishable from *M. m. castaneus* mice. Similarly, mice assigned to the *M. m. helgolandicus* subspecies exhibit no meaningful genome-wide pattern of divergence from *M. m. domesticus*. While these subspecies may harbor distinct morphological adaptations[11], the adoption of a strict genetic species concept argues that *M. m. bactrianus* and *M. m. helgolandicus* are not well-justified taxonomic groups. Instead, mice previously assigned to these subspecies appear to capture population-level genetic diversity within the primary house mouse subspecies. Our work motivates additional genomic investigations into whether other secondary house mouse subspecies designations are warranted.

In addition to providing novel insights into house mouse taxonomy, our work yields newly sequenced house mouse genomes that will serve as key genetic resources for future investigations into wild mouse demographic history and diversity. In particular, genome sequencing of wild mice sampled from the cradle of *M. musculus* evolution will enable studies of the impact of ancestral variation on contemporary patterns of global wild mouse diversity. Further, our work has established the feasibility of whole genome sequencing of archived mammalian tissue and represents an important advance for the emerging discipline of museomics. Broader application of this methodology to additional museum samples will offer a facile approach for strategically expanding genomic catalogs of wild mouse diversity, and potentially population genomic and reference genome sequencing of other mammalian species.

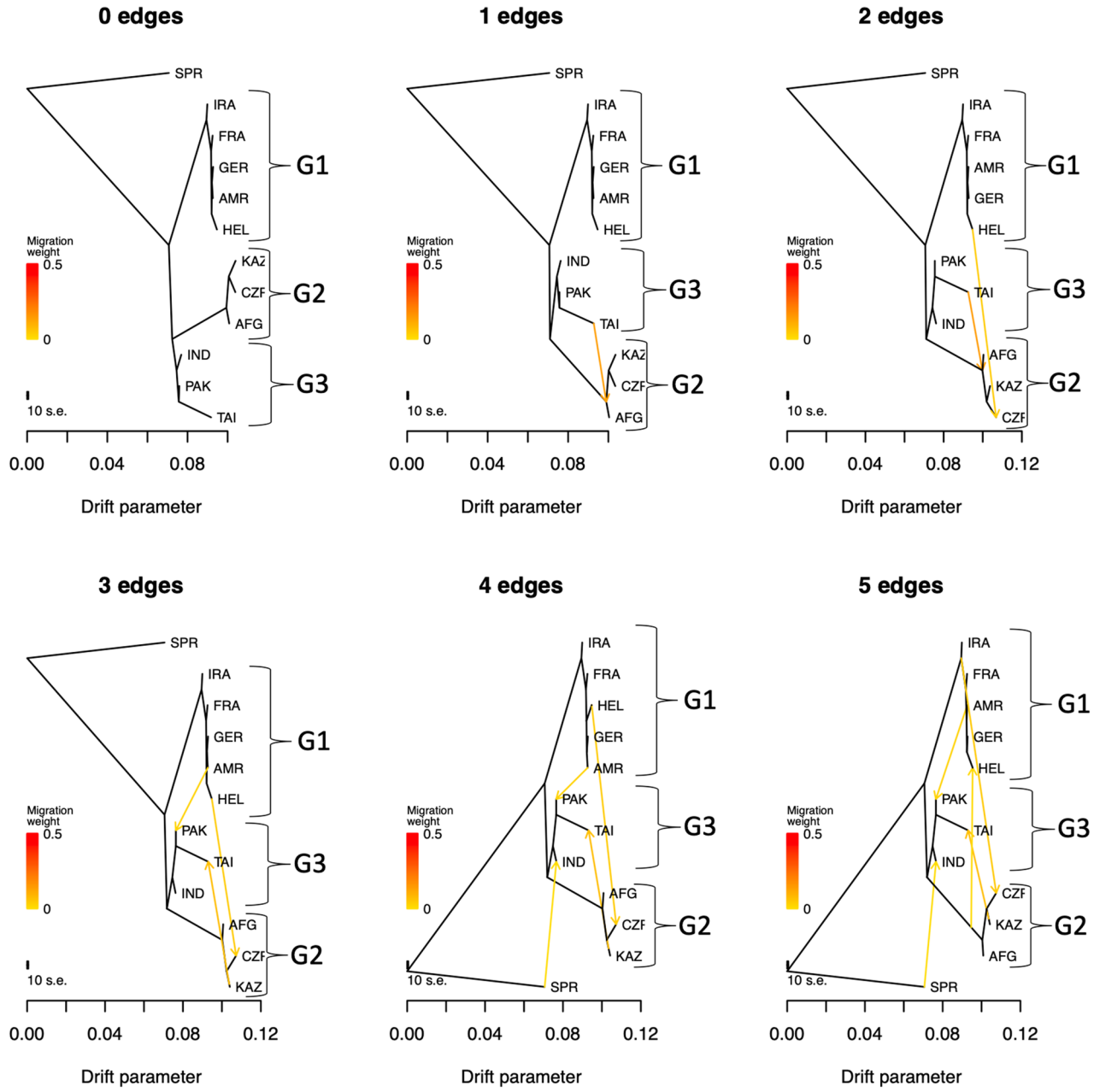


Figure 3. Phylogenetic relationships from TreeMix between house mouse populations. Treemix models the effect of different numbers of populations experiencing gene flow and different pairwise migration rates between populations. Zero edges correspond to the absence of gene flow. Under all considered scenarios, HEL is embedded within the taxonomic group G1 (*M. m. domesticus*) and PAK in G3 (corresponding to *M. m. castaneus* populations). *M. spretus* (SPR) was used as an outgroup.

Materials and methods

Museum sample collection. We destructively sampled the skins of 14 M. *m. bactrianus* mice housed in the Florida Museum of Natural History Mammalogy Collection (http://specifyportal.flmnh.ufl.edu/mammals/). These specimens were collected and preserved between 1975 and 1977 across four counties in Pakistan 'PAK' (see Fig. 1 inset). Further details can be found in the GBIF.org (25 April 2022) GBIF occurrence download (https://doi.org/10.15468/dl.xuksm3)[41]. Skin snips were obtained by removing a small (approx. 5 × 5 mm) section of skin from the ventral side of the study skin, sterilizing instruments in between samples.

DNA extraction protocol for museum-stored samples. DNA was isolated from desiccated skin samples following a previously published protocol[29] with minor modifications. Briefly, the skin samples were scraped with a sterilized scalpel to remove possible contaminants. Samples were transferred to a 2 ml tube, washed three times with sterile water, three times with 70% ethanol, three times with sterile water, and then cut into small pieces. Samples were hydrated before digestion by incubating for 24 h in 1 mL of TE (10 mM Tris; 1 mM EDTA,

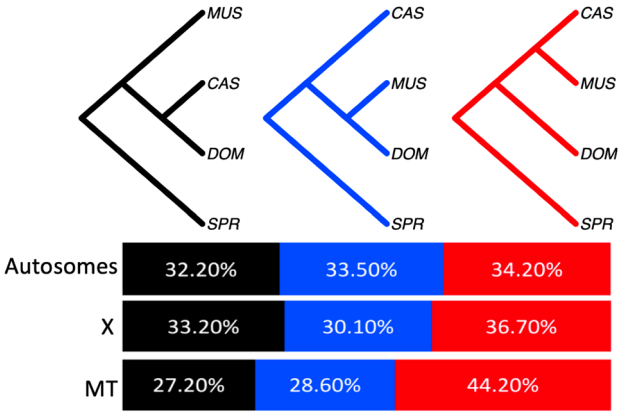


Figure 4. The percentage of the autosomal, X, and mitochondrial (MT) genome supporting each of the three possible topological relationships relating the three primary house mouse subspecies. Percentages correspond to the representation of each topology across 13,122 genomic regions. *M. m. castaneus* (CAS), *M. m. musculus* (MUS), *M. m. domesticus* (DOM), and outgroup *M. spretus* (SPR).

pH 7.6), washing with 70% ethanol and sterile water, and hydrating again in TE solution for a further 24 h. Samples were digested in TNE solution (10 mM Tris HCl, pH 8; 400 mM NaCl; 2 mM EDTA, pH 8.0) plus SDS 1% and Proteinase K (0.58 mg/ml final conc.) at 55 °C for 24–36 h until the tissue was completely digested. The DNA was extracted with one volume of phenol:chloroform:isoamilic alcohol (25:24:1), rotated at 20 rpm for 10 min, and centrifuged for 10 min at 4000 rcf, after which the supernatant solution was transferred to another tube. The DNA was precipitated by adding two volumes of 100% ethanol, gently inverting the tube, and maintaining the solution at − 20 °C for 16 h. The samples were centrifuged for 2 min at 3000 rcf before discarding the ethanol and resuspending the pellet in 50 µL of TE.

DNA concentration and quality (size) were assessed using the Nanodrop 2000 spectrophotometer (Thermo Scientific), the Qubit 3.0 dsDNA BR Assay (Thermo Scientific), and the D5000 DNA ScreenTape Analysis Assay (Agilent Technologies). DNA fragment sizes ranged from 76 to 431 bp. Only samples with DNA concentration > 63 ng/µl were used for genome sequencing.

Genomic DNA library preparation. Whole-genome libraries were constructed using the KAPA HyperPrep Kit (Roche Sequencing and Life Science) according to the manufacturer's protocols. No fragmentation or sizing was done on the samples before proceeding with ligation of Illumina-specific barcoded adapters and PCR amplification. The quality and concentration of the libraries were assessed using the D5000 ScreenTape (Agilent Technologies) and the KAPA Library Quantification Kit (Roche Sequencing and Life Science) according to the manufacturers' instructions.

Libraries were pooled and sequenced on the NovaSeq 6000 (Illumina) using the S4 Reagent Kit (Illumina) and100 bp paired-end reads. We targeted 30X coverage per sample, with the amount of generated data ranging from 33 to 112 Gb across samples (see Supplementary Fig. S3).

Evaluating museum-stored DNA for post-mortem damage. The long-term storage of museum specimens is associated with DNA degradation by deamination, leading to an excessive accumulation of cytosine to uracil (read by sequencer as thymine) changes[36]. In downstream analyses, such post-mortem DNA damage can lead to biases and incorrect data interpretation. We evaluated the PAK genome sequences derived from the museum-stored samples using the Bayesian approach implemented in mapDamage 2.0, a program designed to track and quantify DNA damage patterns[42]. Specifically, we focused our attention on the unusual accumulation of C to T and A to G mutations at the 5′ and 3′ termini as they represent the signatures of post-mortem deamination. Across the 14 PAK samples, the frequency of post-mortem DNA damage was estimated to be no greater than 0.1% of bases in each genome (Supplementary Fig. S2) and error signals were restricted to the 5 bp within the 5′ and 3′ termini of reads. To eliminate potential biases and errors in our data, we trimmed the first and last 5 bases from each read using the "trimBam" option in "BamUtil"[43].

Additional wild mouse genome sequences. We retrieved 155 previously published genomes belonging to *M. m. domesticus* (America, AMR = 50, France, FRA = 28, Germany, GER = 7, Iran = 7), *M. m. helgolandicus* (HEL, n = 3), *M. m. castaneus* (India, IND = 10, Taiwan, TAI = 20), *M. m. musculus* (Afghanistan, AFG = 6; Czech Republic, CZR = 8; Kazakhstan, KAZ = 8), and *M. spretus* (Spain, SPR = 8)[33–35].

Sequence alignment and variant calling. For the newly sequenced 14 PAK samples, we trimmed Illumina adapters using *cutadapt*[44]. The clean reads were mapped to mm10 reference genome using the default parameters in BWA version 0.7.15[45]. Data from the 14 PAK samples were processed simultaneously with the 155 publicly available genome sequences to generate an integrated call set. We followed the standard Genome Analy-

sis Toolkit (GATK; version 4.2) pipeline for subsequent pre-processing before variant calling[46,47]. We performed variant calling for each sample separately using the "-ERC GVCF" mode in the "HaplotypeCaller". Samples were then jointly genotyped using the "GenotypeGVCFs" GATK function and trained with previously ascertained mouse variants[21] using both the "VariantRecalibrator" and "ApplyVQSR" option of GATK. For the latter, the truth sensitivity level to initiate filtration was set to the default (i.e., 99). We filtered variants to exclude sites with missing alleles using VCFtools version 0.1.16[48]. All downstream analyses focus on autosomal bi-allelic single nucleotide variants.

Analyses of population genetic structure. Principal component analysis was performed on all 169 wild mouse genomes using Plink version 2.0[49]. To construct a distance matrix tree, we first reformatted the variant file using "bcftools view file.vcf | bcftools query -f'%CHROM\t%POS[\t%TGT]\n' | sed -e 's/\./N/g'" with BCFtools[50]. We then used the python script "distMat.py" obtained from https://github.com/simonhmartin/genomics_general to generate the tree matrix[51]. The tree was viewed using SplitsTree version 4.17.0[52]. We assessed the robustness of this topology to gene flow using TreeMix version 1.13[40], allowing 0–5 migrations between any population pair in our dataset.

To calculate the allele sharing distance, we used the default option of "asd" (https://github.com/szpiech/asd)[53] and viewed the data using the R package "pheatmap"[54]. We estimated the co-ancestry based Fst using the python program "popgenWindows.py" accessed from https://github.com/simonhmartin/genomics_general[51], and visualized results using the pheatmap R package[54].

Inference of the dominant subspecies topology. We summarized the relationships among samples by building phylogenetic trees across 13,122 unique genomic regions, each defined by a fixed window of 50 SNPs. Trees were built for each window using the script "phyml_sliding_windows.py" accessed from https://github.com/simonhmartin/genomics_general. The output from the tree was used as input and weight assigned to each topology using *Twisst*—Topology Weighting by Iterative Sampling of Sub-Trees—based on the following options:—method complete,—abortCutoff 1000—backupMethod fixed—iterations 400[55]. Topologies were viewed in R using the package "APE" version 5.5[56].

Data availability

The raw fastq reads of the newly sequenced 14 wild house mice from Pakistan have been deposited in the NCBI Short Read Archive under the BioProject accession PRJNA851025. https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA851025.

Received: 28 June 2022; Accepted: 29 November 2022
Published online: 02 December 2022

References

1. Phifer-Rixey, M. & Nachman, M. W. The natural history of model organisms: Insights into mammalian biology from the wild house mouse *Mus musculus*. *Elife* **4**, e05959 (2015).
2. Boursot, P., Auffray, J.-C., Britton-Davidian, J. & Bonhomme, F. The evolution of house mice. *Ann. Rev. Ecol. Syst.* **24**, 119–152 (1993).
3. Suzuki, H., Shimada, T., Terashima, M., Tsuchiya, K. & Aplin, K. Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. *Mol. Phylogenet. Evol.* **33**, 626–646 (2004).
4. Suzuki, H., Aplin, K. & Pialek, J. Phylogeny and biogeography of the genus *Mus* in Eurasia. *Evol. House Mouse* **3**, 35 (2012).
5. Bonhomme, F. & Searle, J. B. House mouse phylogeography. *Evol. House Mouse* **3**, 278 (2012).
6. Suzuki, H. *et al.* Evolutionary and dispersal history of Eurasian house mice *Mus musculus* clarified by more extensive geographic sampling of mitochondrial DNA. *Heredity* **111**, 375–390 (2013).
7. Geraldes, A., Basset, P., Smith, K. L. & Nachman, M. W. Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Mol. Ecol.* **20**, 4722–4736 (2011).
8. Boursot, P. *et al.* Origin and radiation of the house mouse: mitochondrial DNA phylogeny. *J. Evol. Biol.* **9**, 391–415 (1996).
9. Mayr, E. *Principles of Systematic Zoology* (Scientific Publishers, 2015).
10. Mayr, E. *Animal Species and Evolution* Vol. 797 (Belknap Press of Harvard University Press Cambridge, 1963).
11. Babiker, H. & Tautz, D. Molecular and phenotypic distinction of the very recently evolved insular subspecies *Mus musculus* helgolandicus ZIMMERMANN, 1953. *BMC Evol. Biol.* **15**, 1–14 (2015).
12. Prager, E. M., Orrego, C. & Sage, R. D. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150**, 835–861 (1998).
13. Duplantier, J.-M., Orth, A., Catalan, J. & Bonhomme, F. Evidence for a mitochondrial lineage originating from the Arabian peninsula in the Madagascar house mouse (*Mus musculus*). *Heredity* **89**, 154–158 (2002).
14. Hardouin, E. A. *et al.* Eurasian house mouse (*Mus musculus* L.) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. *BMC Evol. Biol.* **15**, 26 (2015).
15. Adhikari, P. *et al.* First molecular evidence of *Mus musculus bactrianus* in Nepal inferred from the mitochondrial DNA cytochrome B gene sequences. *Mitochondrial DNA Part A* **29**, 561–566 (2018).
16. Zimmermann, K. Die hausmaus von helgoland. *Mus musculus sspec Z. Seaugetierkunde* **17**, 163–166 (1953).
17. Schwarz, E. & Schwarz, H. K. The wild and commensal stocks of the house mouse, *Mus musculus* Linnaeus. *J. Mammal.* **24**, 59–72 (1943).
18. Yonekawa, H. *et al.* Evolutionary relationships among five subspecies of *Mus musculus* based on restriction enzyme cleavage patterns of mitochondrial DNA. *Genetics* **98**, 801–816 (1981).
19. Yonekawa, H. *et al.* Hybrid origin of Japanese mice "*Mus musculus molossinus*": Evidence from restriction analysis of mitochondrial DNA. *Mol. Biol. Evol.* **5**, 63–78 (1988).
20. Schoch, C. L. *et al.* NCBI taxonomy: A comprehensive update on curation, resources and tools. *Database* **2020** (2020).
21. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289 (2011).
22. White, M. A., Ané, C., Dewey, C. N., Larget, B. R. & Payseur, B. A. Fine-scale phylogenetic discordance across the house mouse genome. *PLoS Genet.* **5**, e1000729 (2009).

23. Phifer-Rixey, M., Harr, B. & Hey, J. Further resolution of the house mouse (*Mus musculus*) phylogeny by integration over isolation-with-migration histories. *BMC Evol. Biol.* **20**, 1–9 (2020).
24. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat. Genet.* **43**, 648 (2011).
25. Fujiwara, K. *et al.* Insights into *Mus musculus* population structure across Eurasia revealed by whole-genome analysis. *Genome Biol. Evol.* **14**, evac068 (2022).
26. Ballard, J. W. O. & Whitlock, M. C. The incomplete natural history of mitochondria. *Mol. Ecol.* **13**, 729–744 (2004).
27. Willerslev, E. & Cooper, A. Ancient dna. *Proc. R. Soc. B Biol. Sci.* **272**, 3–16 (2005).
28. Burrell, A. S., Disotell, T. R. & Bergey, C. M. The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* **79**, 35–44 (2015).
29. Moraes-Barros, N. D. & Morgante, J. S. A simple protocol for the extraction and sequence analysis of DNA from study skin of museum collections. *Genet. Mol. Biol.* **30**, 1181–1185 (2007).
30. Andrews, K. R., De Barba, M., Russello, M. A. & Waits, L. P. Advances in using non-invasive, archival, and environmental samples for population genomic studies. In *Population Genomics Wildlife* 63–99 (Springer, 2018).
31. Bi, K. *et al.* Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* **22**, 6018–6032 (2013).
32. Rowe, K. C. *et al.* Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* **11**, 1082–1092 (2011).
33. Harr, B. *et al.* Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. *Sci. Data* **3**, 160075 (2016).
34. Phifer-Rixey, M. *et al.* The genomic basis of environmental adaptation in house mice. *PLoS Genet.* **14**, e1007672 (2018).
35. Davies, R. W. *Factors Influencing Genetic Variation in Wild Mice* (University of Oxford, 2015).
36. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7**, e34131 (2012).
37. Kirch, M., Romundset, A., Gilbert, M. T. P., Jones, F. C. & Foote, A. D. Ancient and modern stickleback genomes reveal the demographic constraints on adaptation. *Curr. Biol.* **31**, 2027–2036 (2021).
38. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucl. Acids Res.* **38**, e87–e87 (2010).
39. Teeter, K. C. *et al.* Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* **18**, 67–76 (2008).
40. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967. https://doi.org/10.1371/journal.pgen.1002967 (2012).
41. GBIF.org GBIF occurrence download (2022). https://doi.org/10.15468/dl.xuksm3.
42. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2. 0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
43. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint* http://arxiv.org/abs/1303.3997 (2013).
46. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. https://doi.org/10.1093/bioinformatics/btp698 (2010).
47. Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11–33. https://doi.org/10.1002/0471250953.bi1110s43 (2013).
48. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 (2011).
49. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7. https://doi.org/10.1186/s13742-015-0047-8 (2015).
50. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
51. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257. https://doi.org/10.1093/molbev/msu269 (2015).
52. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
53. Gao, X. & Starmer, J. Human population structure detection via multilocus genotype clustering. *BMC Genet.* **8**, 1–11 (2007).
54. Kolde, R. Pheatmap: Pretty heatmaps. *R Package Version* **1**, 747 (2012).
55. Martin, S. H. & Van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **205** (2017).
56. Paradis, E., Schliep, K. & Schwartz, R. APE 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **1**, 3. https://doi.org/10.1093/bioinformatics/bty633 (2018).

Author contributions

R.A.L. and B.L.D. conceptualized the project. R.A.L. performed all investigations and wrote the original draft with a major review by B.L.D. B.L.D. was responsible for project supervision. Museum specimens were provided by V.L.M. A.G. generated the geographic map for all samples. M.E.B. and J.R.C. extracted the DNA and prepared genome sequencing libraries. All authors read, reviewed, and approved the final manuscript.

Funding

RAL was supported by The Jackson Laboratory (JAX) Postdoctoral Scholar Award. Genome sequencing was completed with funds from a JAX Pyewacket Award to RAL and BLD. AG is supported by the National Science Foundation Graduate Research Fellowship Program under the Grant No. 1842474. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-25420-x.

Correspondence and requests for materials should be addressed to R.A.L. or B.L.D.