Biodiversity databases as underutilized resources for pathogen discovery: a quantitative synthesis of bat and rodent tissue collections in natural history museums

Short title: Museums as resources for pathogen discovery: a quantitative synthesis

Colleen E. Cronin[1]*, Maya M. Juman[2]*^, Alexander J. Richardson[3], Christopher A. Whittier[1], Daniel J. Becker[4], Adam W. Ferguson[5]

[1]Cummings School of Veterinary Medicine, Tufts University, North Grafton, MA, USA
[2]Department of Veterinary Medicine, University of Cambridge, UK
[3]McCourt School of Public Policy, Georgetown University, Washington, DC, USA
[4]School of Biological Sciences, University of Oklahoma, Norman, OK, USA
[5]Gantz Family Collections Center, Field Museum of Natural History, Chicago, IL, USA

*These authors contributed equally to this work.
^Corresponding author (email: mmj38@cam.ac.uk)

Note: This study involves no new data collection or live participants and therefore does not require an Ethics Statement.

Abstract

Zoonotic spillovers are becoming increasingly frequent, and the devastating effects of the SARS-CoV-2 pandemic demonstrate our continued inability to combat their consequences effectively. Natural history museums can enhance the study of zoonoses by serving as valuable resources for understanding the ecology and evolutionary history of pathogens and their wildlife hosts. Despite the growing interest in the role museums can play in pathobiology research and zoonotic risk assessment, there remains a lack of centralized resources for locating tissue samples that may be leveraged for pathogen discovery. Using the world's most significant global aggregator of museum specimens, The Global Biodiversity Information Facility (GBIF), we examined how such tools could be adopted to identify specimens that might be sources of viral genetic material. Focusing on tissue samples from the mammalian orders Rodentia and Chiroptera, speciose taxa that host a high diversity of known zoonotic viruses, we examined temporal, spatial, and taxonomic gaps and patterns in the available tissue samples. Our analyses reveal a heavy bias toward tissue samples collected from the Americas (and consequently, taxonomic groups found in the Americas), with most collected samples housed in North American institutions. This limits the scope of future pathogen discovery efforts and presents a barrier to pandemic preparedness in the Global South. We also examine gaps in metadata quality (e.g., descriptions of preservation method and storage medium) and outline recommendations for GBIF to facilitate future biosurveillance projects and effectively incorporate natural history museums into One Health disease research.

Author Summary

The SARS-CoV-2 pandemic has driven a greater research focus on understanding wildlife reservoirs of zoonotic pathogens. As a complement to field-based sampling of wildlife,

natural history museum collections house millions of specimens that could be used by researchers to study pathogens quickly, safely, and cost-effectively. Digital databases of museum specimens and their associated tissue samples were originally created for biodiversity research, but these could be adapted to guide pathogen discovery research. Using the largest of these databases, The Global Biodiversity Information Facility, we explored tissue samples from rodents and bats, which have been shown to carry a disproportionate number of zoonotic pathogens. We searched for keywords that would indicate the presence of relevant tissue samples and then mapped the results visually. This revealed disproportionate distributions of tissue samples across time, space, and taxonomy. Our study is the first attempt to assess how a biodiversity database can be used for this novel purpose. We suggest changes that would improve these databases for zoonotic disease research.

Introduction

Most emerging infectious diseases are zoonotic in origin (1), but the ecology of these pathogens and the exact drivers of their emergence remain elusive (2). This gap in knowledge has resulted in solutions that are often reactive rather than preventative, which limits our ability to combat rapidly emerging zoonoses (3). The first step in predicting and eventually mitigating spillover is identifying existing and likely host species.

Wild animals harbor thousands of viruses that have yet to be discovered, many of which likely have zoonotic potential (2). Species from the orders Rodentia and Chiroptera have the highest number of reported zoonotic viruses; this is at least in part a function of the high species richness in both orders (5), although some studies suggest that even after controlling for sampling bias and species richness, bats have a significantly higher proportion of zoonotic viruses than all other mammalian orders (6, 7). As common hosts for zoonotic viruses, bats and

rodents are important groups for exploring limitations or gaps in resources for studying zoonotic pathogens. Within these orders, certain species are more likely to carry and transmit zoonotic viruses. Studies have linked zoonotic pathogen richness with larger ranges and more litters at a younger age in rodents (8) and range overlap, smaller litters, more litters per year, and longevity in bats (9, 10). This suggests that species traits can be used as a guide to predict zoonotic risk and focus research efforts on higher-risk mammalian species.

Host–virus associations are crucial data that allow researchers to better predict and mitigate spillover risk, such as the spatial and host range of a pathogen (11, 12, 13). Understanding the ecological and evolutionary relationship between a host and virus can also provide valuable insights into cellular interactions, changes in prevalence and distribution, and factors that influence the likelihood of spillover into humans (14, 15). However, the rarity of many viral infections and high spatiotemporal variation in prevalence necessitate the collection and screening of large sample sizes from hosts (16), which can involve prohibitively expensive and time-consuming field work.

In contrast to prospective field sampling, natural history museums already hold an estimated 1.1 billion specimens around the globe (17). Mammal collections in North America alone contain more than five million specimens (18), and new specimens are added each year (19). In the last decade, there has been growing interest in using these specimens for pathobiology research. Natural history museums are thereby uniquely positioned to circumvent many of the issues involved in field-based sampling, and they also enable researchers to cheaply and efficiently study pathogen occurrence and diversity across space and time (19, 20, 21, 22, 23, 24).

A classic example of collections-based viral host identification is the Sin Nombre hantavirus that emerged in 1993 in the American Southwest and resulted in 27 human fatalities. Researchers used specimens from the Museum of Southwestern Biology to not only identify the definitive reservoir as the common deer mouse (*Peromyscus maniculatus*) but also determine that the virus had been circulating in wild deer mouse populations for at least a decade prior to the first known spillover event (25, 26). Since then, over 40 hantaviruses have been discovered in various small mammals, many with the use of museum specimens, and archived tissue samples have provided new insights into the evolutionary origins, genetic diversity, and spatial and temporal distribution of these viruses (20, 27). In more recent studies, museum specimens have been used to study RNA viruses (28, 29), DNA viruses (30), bacteria (31), and metazoan parasites (32). Recent methodological advances have even facilitated the recovery of RNA from formalin-fixed specimens (33), an approach that could enable viral screening of much older material. However, despite mounting evidence that museum collections are critical resources for the identification, tracking, and mitigation of zoonotic pathogens, they remain largely overlooked by disease ecologists and virologists (19).

Incorporating natural history museums into One Health research also presents its own set of challenges. Most specimens were collected at times when the research questions being asked today were unimaginable (16). Traditional specimen preservation methods (e.g., formalin fixation; storage in ethanol, DMSO, or EDTA) are insufficient for preventing the degradation of viral genetic material, particularly RNA. In addition to long-term storage requirements, the utility of a tissue sample for viral research depends on the initial storage methods in the field as well as any changes during transport (34). Ideally, samples would be flash-frozen in liquid nitrogen as soon as possible (35), either without a buffer (e.g., if collecting tissue for host

genomics) or with a buffer that preserves genetic material but inactivates live virus (e.g., DNA/RNA Shield) (36). After initial storage, freeze–thaw cycles should be avoided (34). However, prior to 2014, there were no best-practice standards for the documentation and storage of genetic resources in natural history museums, and storage methods varied greatly across collections (34). Moreover, the Systematic Collections Committee of the American Society of Mammalogists only formally published their standards and guidelines on genetic resource maintenance in 2019 (35).

Despite this lack of guidance, museums have made great strides in recent decades to embrace the concepts of the "holistic specimen" (37), "extended specimen" (38), and "open specimen" (39). As a result, the preservation of tissue samples, metazoan parasites, and other associated microbiological samples is becoming increasingly common (24, 37). However, for these samples to be useful to the broader scientific community, the relevant metadata must also be archived in a publicly accessible digital format, and the push to digitize and share these data lags behind that of primary biodiversity metadata (40). For many decades, museums held their digital archives on various internal databases, each using different recording formats and software packages (37, 40). The standardization and open access of these data remain pressing challenges for natural history museums in the context of One Health (37, 41).

Publicly available biodiversity databases address many of these issues (42), and each database often has a specific focus related to spatial coverage and record type (40). Early efforts to digitize and mobilize museum data were also often taxon-specific, such as the Mammal Networked Information System (43). Today, the Global Biodiversity Information Facility (GBIF) is one of the largest biodiversity databases in the world, with over one billion records from institutions in over 50 countries (40, 44). By implementing standard terminology such as Darwin

Core (45) and new tools such as the Integrated Publishing Toolkit (46), GBIF has allowed museums to easily share their data online (37). Most importantly, GBIF works as a downstream data aggregator, collecting data from a network of biodiversity databases, organizations, and data papers such that information can be shared and searched in a centralized fashion (40). Many researchers have heralded the potential of data aggregators such as GBIF to facilitate use of museum collections in other research initiatives (19, 23, 37, 41, 44, 47). However, studies that quantitatively evaluate the spatial, temporal, and taxonomic coverage of GBIF for the purpose of pathobiology are lacking.

Here, we test whether biodiversity data aggregators such as GBIF can be adopted for users to identify suitable tissues for pathobiological studies. We examine temporal, geographic, and taxonomic representation in digitized bat and rodent tissue samples that could be used for pathogen surveillance, including those from species predicted to carry the most undescribed zoonotic viruses (6). Finally, we outline recommendations for museums and database aggregators to improve data accuracy and facilitate the inclusion of natural history museums in One Health research.

Results

Of the 140 museum collections or institutions with tissue samples in GBIF, 26 showed evidence of possessing only bat tissue samples, 32 showed evidence of possessing only rodent tissue samples, and 89 showed evidence of possessing both bat and rodent tissue samples. In total, 588,425 (out of 4,586,821; 12.8%) listed specimens had associated tissue samples, of which 475,834 were from rodents and 112,591 were from bats (see Supplementary Data S1 for the full list of institutions and their associated tissue samples).

Among specimens with associated tissue samples, the search term “tiss” or “tejid” yielded the highest number of occurrences, appearing in over 59% of the dataset. Over 26% of tissue samples appeared to be in ethanol, while ~45% were frozen. Of reported tissue types, liver samples were the most common, followed by kidney, heart, spleen, lung, muscle, and colon (Table 1).

Table 1. Occurrence of preservation method and tissue type in total dataset and breakdown by taxonomic order.

Term	Total occurrences	Rodent occurrences	Bat occurrences
“blood” / “sangre” / “serum” / “suero” / “plasma”	66,382	64,995	1,387
“tiss” / “tejid”	348,165	275,050	73,115
“biopsy” / “biopsia”	1,169	7	1,162
“swab” / “torunda”	13,778	13,178	600
“ethanol” / “etanol” / “EtOH”	157,748	125,099	32,649
“froze” / “freez” / “congelad”	266,471	237,675	28,796
“DMSO” / “EDTA” / “VTM” / “RNAlater” / “shield” / “buffer” / “lysis”	2,217	1,531	686
“heart”	102,393	92,147	10,246
“kidney”	108,239	97,394	10,845
“liver”	150,432	136,200	14,232
“lung”	67,402	64,124	3,278
“muscle”	30,383	24,004	6,379
“spleen”	72,708	69,345	3,363
“colon”	1,010	718	292

Temporal distribution

The collection year of all bat and rodent samples in GBIF (not just those with associated tissue samples) ranged from 1591 to 2025, with a median year of 1972 that corresponded with an uptick in collecting (Fig. 1). Overall, 391,809 samples (8.5%) were missing collection date information. For samples with associated tissues, the years of collection ranged from 1806 to 2025, with a median collection year of 2005. Of these samples, 15,108 (2.6%) were missing collection date information.

We fitted a generalized additive model (GAM) to the tissue samples collected per year to identify periods of significant increase or decrease in the rates of tissue collection up until 2024 (48). From 1828 to 1852, tissue collection showed a significant increase, followed by a brief decline in the late 1860s. Tissue collection increased again from 1885 to 1914, followed by another brief decline from 1925 to 1932. After 1945, tissue collections grew rapidly until 2004, but since 2021 the rate of tissue collecting has been declining.

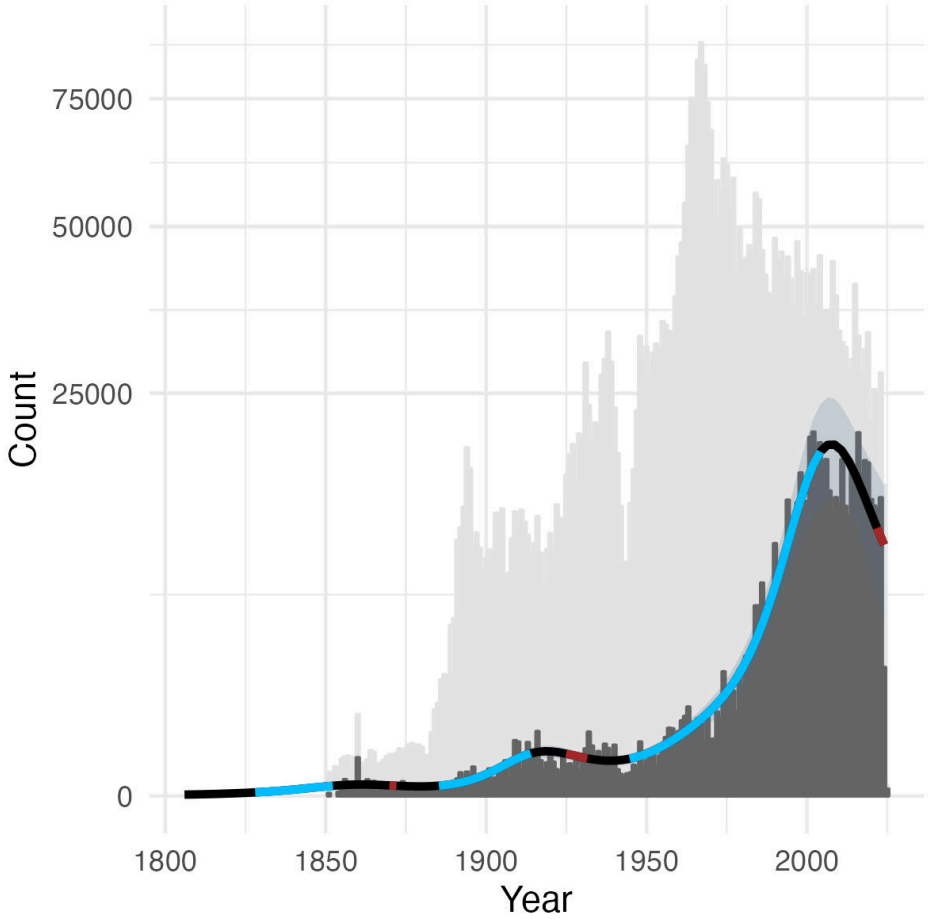


Fig. 1. Barplot indicating the collection year of all bat and rodent specimens in GBIF (light grey) and those with associated tissue samples (dark grey), with the number of samples per year shown on a square root scale. The black line indicates the fitted spline from a GAM of tissue samples

collected per year. Blue intervals indicate periods of significant increase in tissue collecting while red intervals indicate periods of significant decline.

Spatial biases

The dataset included tissue samples from 190 countries or territories with unique country codes. An additional 376 samples were from unknown countries (i.e., country code ZZ). However, sampling effort was highly heterogeneous across countries/territories, ranging from one sample (i.e., Netherlands Antilles, Bahrain, Bosnia and Herzegovina, Belarus, Cyprus, Georgia, Guam, Maldives, Oman, Saint Helena, Ascension and Tristan da Cunha, Slovenia, and Tajikistan) to 349,890 (i.e., the USA) (Fig. 2). The second most sampled country was Mexico, with 22,830 tissue samples. A generalized linear model (GLM) of binary sampling effort across all regions (i.e., Africa, Americas, Asia, Europe, Oceania) revealed disproportionate sampling of African, American, and Asian countries ($\chi^2 = 14.58$, $p = 0.0057$, $R^2 = 0.05$) (Supplementary Fig. 1). The number of samples per country also varied significantly across subregions ($\chi^2 = 2{,}214{,}834$, $p < 0.0001$, $R^2 = 0.62$), which was driven by the disproportionate number of samples from the USA, Mexico, and Canada (Supplementary Fig. 2).

Our analysis also revealed strong biases in the spatial distribution of reported tissue holdings (Fig. 2). The size of museum collections was heterogeneous across the 26 countries represented in the dataset, ranging from two samples (i.e., France, the Netherlands) to 533,322 samples in the USA. A GLM across all regions revealed disproportionately high fractions of countries with tissue holdings in Europe and the Americas ($\chi^2 = 20.10$, $p < 0.001$, $R^2 = 0.11$), and the number of samples per country also varied significantly across subregions ($\chi^2 = 2{,}478{,}857$, p

< 0.0001, $R^2 = 0.77$), driven by a disproportionate number of samples in North American museum collections (Supplementary Figs. 3,4).

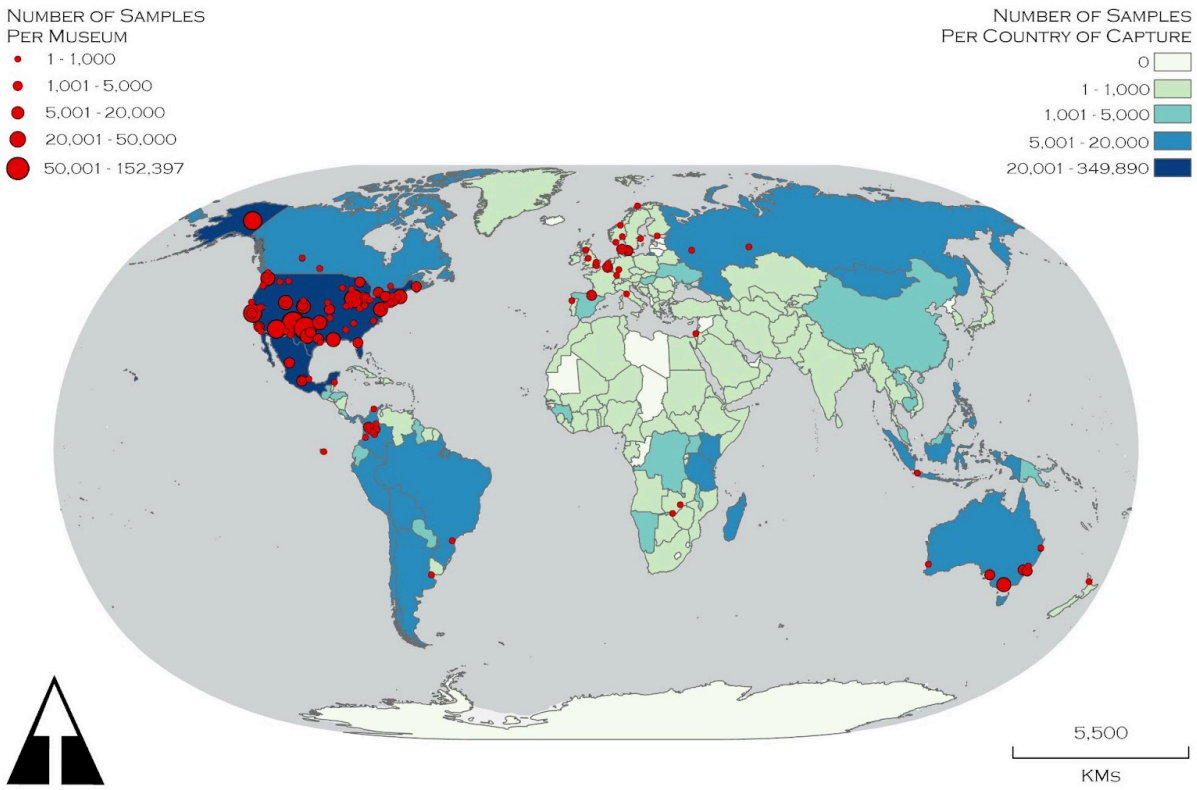


Fig. 2. Global map showing the spatial distribution of bat and rodent specimens with associated tissue samples in GBIF as of July 14, 2025. Countries of collection are shown with shading, while red dots indicate the locations of museum collections in the dataset. The map was constructed in ArcGIS Pro (49) with the World Natural Earth II projection (50) and the World Light Gray Base basemap (51). ISO two-digit country codes were used to match our dataset to the World Countries (Generalized) feature layer provided by Esri via ArcGIS Living Atlas of the World (52).

Taxonomic biases

The dataset initially listed 2,445 species and 619 genera of bats and rodents with associated tissue samples. Of these, 290 did not match to our taxonomic backbone (53). After manual taxonomic reconciliation using the Mammal Diversity Database, 47 of these 290 taxa were recently described or otherwise not in the mammalian phylogeny and could not be harmonized. This left us with 2,377 harmonized species with tissue samples, including 842 bat species and 1,416 rodent species. Based on our taxonomic backbone (53), these samples represent approximately 65% (842/1,287) and 59% (1,416/2,392) of recognized bat and rodent species, respectively.

We found intermediate phylogenetic signal in binary sampling effort across bat species ($D = 0.90$), departing from both phylogenetic randomness ($p < 0.001$) and Brownian motion models of evolution ($p < 0.001$). We then applied phylogenetic factorization, a graph-partitioning algorithm, which revealed that the superfamily Noctilionoidea (containing the families Myzopodidae, Mystacinidae, Thyropteridae, Noctilionidae, Furipteridae, Mormoopidae, and Phyllostomidae) was overrepresented, while the genera *Scoteanax*, *Harpiola*, *Murina*, and *Scotozous*—a subclade of Old World vesper bats—were underrepresented (Fig. 3a; Table 2a). Among sampled species, the number of tissue samples per bat species ranged from one to 4,726 and was distributed across the bat phylogeny with intermediate phylogenetic signal (Pagel's $\lambda = 0.28$), departing from both Brownian motion models of evolution ($p < 0.001$) and phylogenetic randomness ($p < 0.001$). Eleven *Myotis* species—all North American members of the subgenus *Pizonyx*—were overrepresented relative to other taxa; however, the genera *Artibeus* and *Dermanura* were underrepresented (Fig. 3b; Table 2b).

Similarly, we found intermediate phylogenetic signal in binary tissue sampling effort of rodents ($D = 0.82$), departing from both phylogenetic randomness ($p < 0.001$) and Brownian motion models of evolution ($p < 0.001$). Phylogenetic factorization revealed that some rodent clades (including the largely North American families Heteromyidae, Geomyidae, and Castoridae in the suborder Castorimorpha, as well as some genera in Cricetidae) have higher fractions of sampled species relative to other clades, while some genera have disproportionately lower fractions of sampled species (Fig. 3c; Table 2c). Among sampled species, the number of tissue samples per rodent species ranged from one to 54,687 and was distributed across the phylogeny with intermediate phylogenetic signal (Pagel's $\lambda = 0.49$), departing from both Brownian motion models of evolution ($p < 0.001$) and phylogenetic randomness ($p < 0.001$). For example, rodents in the families Heteromyidae, Geomyidae, and Castoridae (members of suborder Castorimorpha) were overrepresented in terms of sampling intensity relative to other taxa, while those in Calomyscidae, Nesomyidae, Cricetidae, and Muridae (clade Eumuroida) were underrepresented (Fig. 3d; Table 2d).

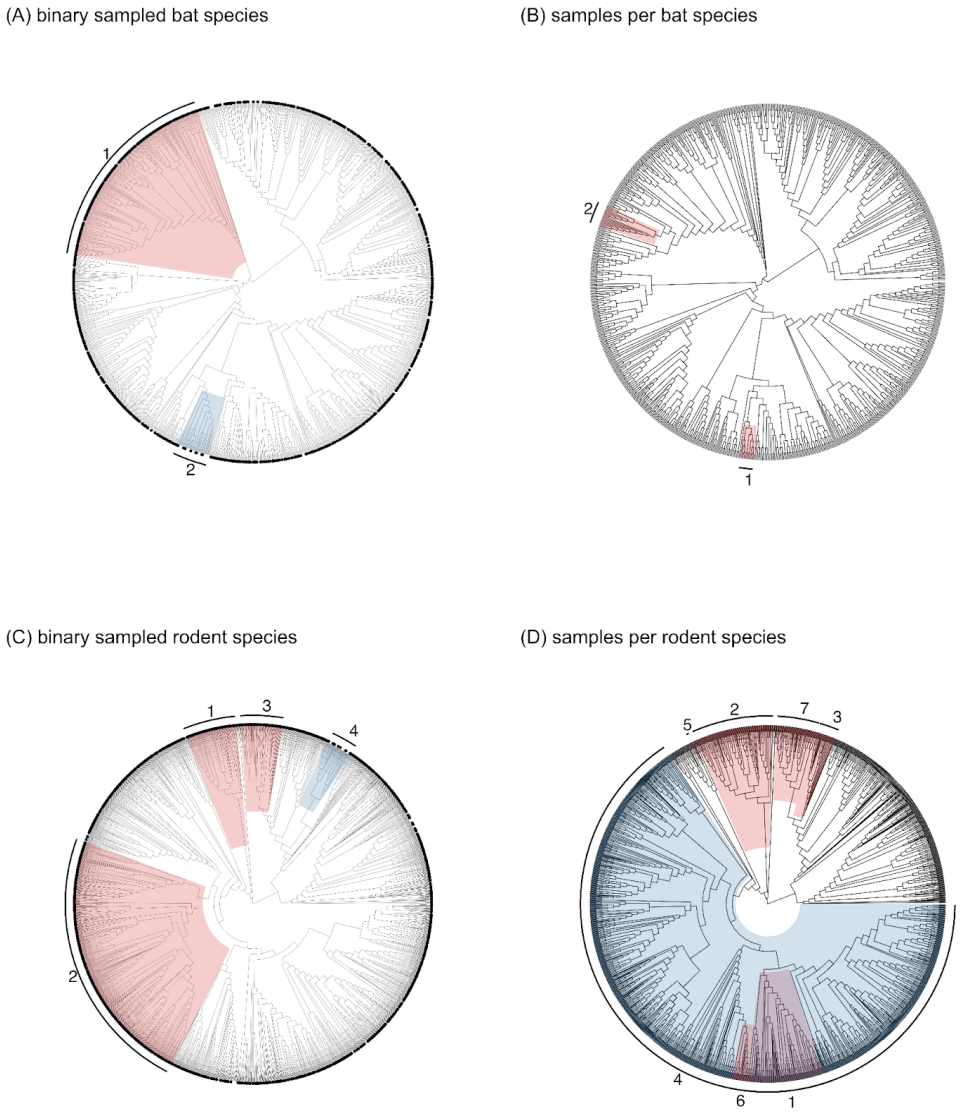
Table 2. Bat and rodent clades that differ significantly from the rest of their respective trees in terms of different measures of sampling effort, as identified by phylogenetic factorization.

(A) Whether or not a bat species has tissue samples in our dataset, across all Chiroptera (0: not sampled; 1: sampled); (B) number of tissue samples per sampled bat species; (C) whether or not a rodent species has tissue samples in our dataset, across all Rodentia (0: not sampled; 1: sampled); (D) number of tissue samples per sampled rodent species (Fig. 3). We report all significant clades detected through phylogenetic factorization (identified with Holm's

sequentially rejective test at a family-wise error rate of 5%), the taxa and number of species within those clades, and the mean sampling effort for the clade compared to the rest of the tree. Bat and rodent taxonomy follows the most recent mammal phylogeny (53).

Clade	Taxa	Number of species	Clade mean	Other mean
A) Binary sampling of bats				
A1	Myzopodidae, Mystacinidae, Thyropteridae, Noctilionidae, Furipteridae, Mormoopidae, Phyllostomidae	227	0.82	0.62
A2	*Scoteanax*, *Harpiola*, *Murina*, *Scotozous*	39	0.21	0.67
B) Logged number of samples of bats				
B1	*Myotis sodalis, Myotis volans, Myotis auriculus, Myotis septentrionalis, Myotis ciliolabrum, Myotis leibii, Myotis californicus, Myotis lucifugus, Myotis keenii, Myotis evotis, Myotis thysanodes*	11	2.55	1.29
B2	*Artibeus, Dermanura*	17	2.13	1.29
C) Binary sampling of rodents				
C1	Heteromyidae, Geomyidae, Castoridae	109	0.93	0.58
C2	*Neotoma, Xenomys, Nelsonia, Hodomys, Ochrotomys, Reithrodontomys, Isthmomys, Onychomys, Osgoodomys, Peromyscus, Megadontomys, Habromys, Podomys, Neotomodon, Scotinomys, Baiomys, Tylomys, Ototylomys, Otonyctomys, Nyctomys, Sigmodon, Rheomys, Ichthyomys, Wilfredomys, Anotomys, Reithrodon, Neusticomys, Chibchanomys, Phaenomys, Chinchillula, Abrawayaomys, Kunsia, Scapteromys, Akodon, Bibimys, Lenoxus, Blarinomys, Brucepattersonius, Thalpomys, Necromys, Podoxymys, Deltamys, Thaptomys, Oxymycterus, Juscelinomys, Rhipidomys, Thomasomys, Rhagomys, Chilomys, Aepeomys, Scolomys, Zygodontomys, Megaoryzomys, Hylaeamys, Mindomys, Nephelomys, Oecomys, Euryoryzomys, Transandinomys, Handleyomys, Neacomys, Microryzomys, Oreoryzomys, Oligoryzomys, Microakodontomys, Holochilus, Pseudoryzomys, Amphinectomys, Nectomys, Nesoryzomys, Aegialomys, Sigmodontomys, Melanomys, Oryzomys, Noronhomys, Pennatomys, Megalomys, Lundomys, Cerradomys, Sooretamys, Drymoreomys, Eremoryzomys, Neotomys, Euneomys, Irenomys, Andinomys, Punomys, Delomys, Calomys, Phyllotis, Paralomys, Galenomys, Auliscomys, Loxodontomys, Tapecomys, Salinomys, Andalgalomys, Graomys, Eligmodontia, Calassomys, Wiedomys, Juliomys, Chelemys, Pearsonomys, Geoxus, Notiomys, Abrothrix*	552	0.69	0.56
C3	*Tamias, Spermophilus, Urocitellus, Ictidomys, Poliocitellus, Xerospermophilus, Cynomys, Marmota, Otospermophilus, Callospermophilus, Notocitellus, Ammospermophilus*	90	0.86	0.58
C4	*Glaucomys, Eoglaucomys, Hylopetes, Petaurillus, Petinomys, Iomys, Pteromys, Biswamoyopterus, Pteromyscus, Aeretes, Trogopterus, Belomys, Aeromys, Eupetaurus, Petaurista*	55	0.20	0.60
D) Logged number of samples of rodents				
D1	Neotoma, Nelsonia, Hodomys, Ochrotomys, Reithrodontomys, Isthmomys, Onychomys, Osgoodomys, Peromyscus, Megadontomys, Habromys, Podomys, Neotomodon, Scotinomys, Baiomys	98	2.04	1.34
D2	Heteromyidae, Geomyidae, Castoridae	101	1.95	1.34
D3	*Tamias*	25	2.25	1.37
D4	Calomyscidae, Nesomyidae, Cricetidae, Muridae	833	1.37	1.41
D5	*Zapus, Napaeozapus*	4	3.20	1.38
D6	*Microtus gregalis, Microtus mexicanus, Microtus longicaudus, Microtus oregoni, Microtus townsendii, Microtus canicaudus, Microtus pennsylvanicus, Microtus montanus, Microtus ochrogaster, Microtus californicus, Microtus xanthognathus, Microtus richardsoni, Microtus umbrosus, Microtus chrotorrhinus, Microtus quasiater, Microtus pinetorum, Microtus guatemalensis, Microtus oaxacensis, Microtus abbreviatus, Microtus miurus*	20	2.30	1.37
D7	*Spermophilus, Urocitellus, Ictidomys, Poliocitellus, Xerospermophilus, Cynomys, Marmota, Otospermophilus, Callospermophilus, Notocitellus, Ammospermophilus*	52	1.55	1.38

Fig. 3. Host and viral taxonomic biases in pteropodid–paramyxovirus sampling effort. (A) Whether or not a bat species has tissue samples in our dataset, across all Chiroptera (black dots indicate species that have been sampled); (B) logged number of tissue samples per sampled bat species; (C) whether or not a rodent species has tissue samples in our dataset, across all Rodentia (black dots indicate viruses that have been detected); (D) logged number of tissue samples per sampled rodent species. Red and blue shading indicates clades with greater or lesser sampling effort, respectively, in comparison to all other taxa (as identified by phylogenetic factorization), with clade numbers corresponding to Table 2.



In 2017, Olival et al. used species-level trait data to predict the risk of spillover from other mammals into humans (6). Our dataset included tissue samples for the ten rodent and ten bat species with the highest predicted number of undiscovered zoonotic viruses (6), suggesting that GBIF can be used to locate samples for testing predictions (Table 3).

Table 3: Potentially viable tissue samples from the ten bat and ten rodent species with the highest number of predicted missing zoonoses as modeled by Olival et al., 2017 (6). A full list of museums housing these samples is presented in Supplementary Data S2.

Order	Species	Tissue samples (*n*)	Predicted missing zoonoses (*n*)
Rodentia	*Neacomys spinosus*	140	20.6
Rodentia	*Holochilus sciureus*	81	20.3
Rodentia	*Oligoryzomys fulvescens*	925	19.4
Rodentia	*Sigmodon alstoni*	13	19.2
Rodentia	*Coendou prehensilis*	16	18.8
Rodentia	*Oecomys bicolor*	161	17.8
Rodentia	*Dasymys incomtus*	40	17.2
Rodentia	*Zygodontomys brevicauda*	3,899	17.2
Rodentia	*Thryonomys swinderianus*	3	17.1
Rodentia	*Otomys tropicalis*	92	17.1
Chiroptera	*Artibeus planirostris*	764	31.5
Chiroptera	*Glossophaga longirostris*	124	30.0
Chiroptera	*Glossophaga commissarisi*	98	29.4
Chiroptera	*Lasiurus ega*	85	28.3
Chiroptera	*Natalus tumidirostris*	17	28.1
Chiroptera	*Vampyrodes caraccioli*	34	27.9

Chiroptera	*Chrotopterus auritus*	52	27.5
Chiroptera	*Eumops glaucinus*	71	26.8
Chiroptera	*Mimon crenulatum*	62	25.9
Chiroptera	*Sturnira tildae*	156	25.7

Discussion

Biodiversity databases have the potential to guide pathogen discovery and surveillance efforts, as they consolidate metadata on archived samples from wildlife globally. We used one of the most comprehensive of these databases, GBIF, to evaluate temporal, spatial, and taxonomic gaps in rodent and bat tissue holdings at natural history museum collections and other biorepositories reported in GBIF. Of the roughly 4.5 million rodent and bat records in GBIF, less than 13% (n = 588,425) included a term indicating a preserved tissue, serum, or biopsy sample that could potentially be used for pathogen diagnostics. Of that subset of records, only around half appeared to be frozen (Table 1).

Tissue samples were collected from 1806 onward, with a median collection date of 2005 (Fig. 1). This coincides with a shift toward the "extended specimen" model of collections, where associated tissues are preserved along with a voucher specimen, often in freezers or liquid nitrogen (54). The paucity of tissue samples from the 19th and first half of the 20th century limit our ability to use these collections as "time machines" for viral diagnostics further back than a century, although other pathogens like metazoan parasites are detectable (32). The continued collection and mindful preservation and storage of new biological materials represents an investment in future pathogen detection and disease prevention efforts. However, the growth of bat and rodent tissue collections has slowed in recent years (Fig. 1), threatening future pandemic preparedness and hindering our ability to study the effects of ongoing global change (20, 37, 55,

306 56, 57). Despite the obvious value of collections, the costs needed to collect and maintain
307 voucher specimens have made natural history museums vulnerable to changes in funding and
308 resource prioritization (37, 58, 59, 60, 61, 62, 63, 64, 65). Collections at academic institutions
309 have been especially impacted as a result of state budget cuts and shifting interest to other
310 scientific branches (66). Further, an increased focus on nonlethal samples or observations (e.g.,
311 camera-trap data, fur/skin samples, etc.) has contributed to the decline in collection rates, but
312 these approaches are less valuable for pathogen discovery efforts than whole specimen collection
313 (55, 57).

314 To effectively use GBIF and similar databases as resources for pathogen surveillance,
315 researchers must be able to locate appropriately preserved specimens for DNA and RNA
316 extraction (or other methods such as serology). The lack of standardized reporting of
317 preservation and storage details in GBIF is a serious barrier to selecting target specimens and
318 tissues for data collection. Of the reported tissue samples, we found that over one-quarter are
319 preserved in ethanol. If these fluid-preserved specimens were originally fixed in formalin, this
320 presents an issue for nucleic acid recovery (particularly RNA, which is problematic for the
321 detection of pandemic-causing RNA viruses such as coronaviruses and paramyxoviruses) (29).
322 New methodological advances are enabling us to extract nucleic acid material from older,
323 formalin-fixed specimens (33, 67, 68). However, this work is early in development and is more
324 expensive than traditional RNA extraction; therefore, the inclusion of more details in GBIF about
325 preservation methods would help facilitate these preliminary projects.

326 The tissue samples in our dataset were heterogeneously distributed across space, in terms
327 of both country of origin and institution country. Countries in the Americas were overrepresented
328 from multiple perspectives (Fig. 2). This bias has potential downstream impacts on our future

329 ability to screen tissues from a broad range of regions for potential emerging pathogens. Europe,
330 Asia, and Africa are all predicted hotspots for emerging zoonotic spillover events (1), but the
331 lack of samples with associated tissues from these continents may hinder the sample size of
332 future museum-based surveillance studies.

333 Such results are consistent with the current literature on where viral research related to
334 bats in particular is lacking. Western Asia is home to 96 bat species, yet a recent study found no
335 publications investigating bat-associated coronaviruses in most of the countries there (69).
336 Despite the high diversity of bat species in Central Africa and Southeast Asia, countries in these
337 regions are also severely underrepresented in the bat literature (69). Similarly, a review of global
338 coronavirus sampling in rodents and other small mammals found striking sampling gaps in
339 Western and South Asia, Eastern Europe, Australia, and much of the African continent (70). Our
340 results further support the observation that tissue samples of bat and rodent specimens are largely
341 deficient in these regions.

342 There is also a dramatic skew in the locations of museum collections with bat and rodent
343 tissue holdings, with most samples housed in North American institutions (Fig. 2). The
344 concentration of these resources in the Global North limits collection-based pathogen
345 surveillance and discovery efforts in underrepresented regions. Building biorepository capacity
346 in the Global South will be critical in fostering a more collaborative, fair, and comprehensive
347 approach toward pathogen spillover prediction and management (71).

348 Bat and rodent tissue samples in GBIF were also skewed in their taxonomic distributions.
349 Bats within the superfamily Noctilionoidea show higher proportions of sampled species relative
350 to other clades. Most of the families within this clade are found only in the Americas (e.g.,
351 including but not limited to Phyllostomidae and Mormoopidae), which is consistent with the

352 overrepresentation of North and South American countries in the dataset. However, many Asian,
353 African, and Oceanian bats are the demonstrated or hypothesized reservoirs of zoonotic viruses
354 like coronaviruses (72), paramyxoviruses (73), and filoviruses (74), suggesting that this sampling
355 bias might limit effective surveillance across all bat taxa for potential epidemic-causing viruses.
356 We identified a similar taxonomic and geographic skew in rodents, where species in the largely
357 New World families Heteromyidae, Geomyidae, and Castoridae are sampled more frequently
358 than those in other clades.

359 Predicting pathogen spillovers is a major research topic, and recent studies have
360 employed predictive models that leverage host–pathogen association data for this purpose (6, 9,
361 11, 75). The tissue samples in GBIF can be used to test such predictions. For example, there are
362 764 and 140 samples, respectively, from *Artibeus planirostris* and *Neacomys spinosus*, the bat
363 and rodent species predicted to carry the most undetected zoonotic viruses (6). By screening
364 tissue samples from these specimens, researchers could empirically test such predictions,
365 following a model-guided viral discovery framework (29). In turn, novel detections could be
366 iteratively integrated back into such models to improve forecasts and potentially mitigate the risk
367 of zoonotic spillover before it occurs (14, 75).

368 Given the availability of such tissue samples, the data from our study can be used as a
369 complementary approach to fieldwork, which would augment zoonotic pathogen research by
370 expediting the timeline and reducing the cost of studying emerging infectious disease risks. Our
371 results can also guide prospective collection efforts by illuminating which species and regions
372 have already been extensively sampled and which are lacking representation in museum
373 collections. Prioritizing subsequent collection efforts in areas that are species-rich but
374 underrepresented would greatly facilitate future pathogen discovery projects.

Our study is not without limitations. Although GBIF is an international database with information from institutions in over 50 countries, our results were limited to one database aggregator. As publishing data in GBIF is voluntary, many museums with viable tissue samples were potentially overlooked, especially smaller collections which may not be participating members of GBIF. Specimens collected by smaller museums or research teams are less likely to be properly recorded, and these "dark data" often remain hidden from the larger research community (76). Of the collections that have been digitized, many are not integrated into the network of databases or have inaccurate or outdated information (41, 76). Even records that have been incorporated into major databases shared with GBIF are often missing vital data, including but not limited to the date or location of collection (44). Furthermore, these databases and data aggregators were designed for the purpose of biodiversity research (47). As a result, standards regarding the inclusion and entry of data that are critical to searching for appropriate samples for pathobiology research (e.g., preservation method or buffer type) are commonly lacking.

Within GBIF, it is also possible that suitable samples could have been missed in our text mining search. Lack of data, including country name, tissue presence, and preservation details, would have led to omission of potentially relevant results. Inconsistent record keeping and lack of standardization also limited the feasibility of a fully exhaustive study. All researchers on our team were native English speakers and, despite our efforts to include the most common languages present in the dataset, this bias could have influenced the search terms we selected. Finally, we caution that it is not possible to extrapolate enough details about tissue preservation techniques to say for certain whether these samples are all suitable for pathobiology research (e.g., lack of information surrounding sample preparation after capture and long-term storage, such as temperature, number of freeze/thaw events, and types of buffers used, if any).

398 These limitations underscore an urgent need for changes in how tissue samples are
399 preserved and how this information is shared across the research community. Researchers have
400 long recognized the need for museums and their associated databases to be integrated with
401 pathogen discovery and research (19, 22, 77). Yet equally crucial is the documentation of these
402 resources to be housed in a centralized portal (78, 79). Large amounts of data, such as those
403 surrounding details of specimen collection and storage, are only useful if they are archived in a
404 standardized and easily accessible manner. While GBIF has made important strides in this
405 direction, it has largely prioritized conservation and biodiversity research, and this focus is
406 reflected in which data are collected and how they are stored. By including the needs of
407 pathogen-oriented researchers in its design, GBIF could further expand its scope to help prevent
408 and mitigate the spillover of zoonotic pathogens. Improved standardization and recording
409 practices, along with a universal portal, will allow researchers to more easily distinguish between
410 actual biases in specimen collections versus gaps in documentation. Proper data storage and
411 sharing will expose which specimens are lacking and which species need to be prioritized.
412 However, separate steps are needed to ensure that, once collected, specimens are used to their
413 greatest research potential.
414 While the benefits of preserving host voucher specimens and tissue collections have been
415 acknowledged by many within the infectious disease research community, large-scale support
416 has been slow (22). Archives of RNA and/or DNA sequences are now required by most top
417 journals in ecology, evolution, behavior, microbiology, and systematics, whereas specimen
418 deposition is required by only 13% of journals (39). Not only does specimen deposition facilitate
419 repeatability—a foundational tenet of scientific research—but it also allows the specimen to be
420 used in future unrelated studies, particularly if it is properly preserved and stored. Museums

serve a larger mission as data stewards and educators of the general public. Incorporating them into the One Health field may catalyze a shift in focus back to this common goal and foster a culture of collaboration as we work to prevent future pandemics.

Materials and Methods

To quantify the temporal, spatial, and taxonomic distribution of specimens viable for pathobiology research, we queried GBIF on July 14, 2025 for all records in the orders Chiroptera and Rodentia. We filtered data to those records in which occurrence status was listed as present and where the record was categorized as a material sample or preserved specimen (80, 81). We used R 4.2.1 (82) to clean the dataset, including 1) ensuring case uniformity; 2) assigning ISO country codes to avoid discrepancies in country names; and 3) matching codes to generalized country coordinates using a publicly available dataset (83). Several packages, especially those in or peripheral to the *tidyverse* collection (84), were used to import, manipulate, and visualize data, including *dplyr* (85), *tidyr* (86), *ggplot2* (87), *patchwork* (88), and *data.table* (89).

To use the maximum number of data points available for spatial analysis, we focused primarily on country-level capture locations. Rather than exclude records missing exact specimen coordinates, generalized country geolocations were added to records without coordinates. Although we acknowledge this limits visualizing exact within-country collection locations, this approach allows more specimens to be collectively mapped. Of the records with languages listed (2,207,368), English (1,873,092 or 84.9%) and Spanish (278,913 or 12.6%) were the two most common, amounting to over 97% of these records. Therefore, we used these two languages for our keyword searches. To ensure we captured any potential specimens of interest for pathobiology research, we included truncated word forms to account for

inconsistency and human error in data entries. We designed our search terms to capture tissue availability and thus included different organs, buffers, and preservation methods.

Using the *stringr* package (90), we searched for tissues in the fields “preparations,” “dynamicProperties,” “occurenceRemarks”, and “materialEntityRemarks” using the following search terms: “blood”, “sangre”, “tiss”, “tejid”, “serum”, “suero”, “plasma”, “biopsy”, “biopsia”, “swab”, “torunda”, “heart”, “kidney”, “liver”, “lung”, “muscle”, “spleen”, “colon”, “ethanol”, “etanol”, “EtOH”, “VTM”, “RNAlater”, “shield”, “lysis”, “DMSO”, “EDTA”, “buffer”, “froze”, “freez”, or “congelad”. We used the results of this search to generate a binary variable indicating the presence or absence of at least one of these search terms in at least one target field.

To map the museums that house these tissues, we used the Google Maps API. Any missing locations after the API query were obtained manually through the GBIF Registry of Scientific Collections, with preference given to the field “institution code.” Duplicates were assumed to be due to acquisitions and manual errors during data entry and were thus combined. We used ArcGIS Pro to visualize the spatial distribution of where bat and rodent specimens with associated tissue samples were collected (49).

To examine temporal patterns in tissue collections, we used the *mgcv* package and restricted maximum likelihood to fit a GAM to the number of tissue samples collected per year (up until 2024). We used penalized thin-plate regression splines to fit year as a nonlinear smoothing term, with number of tissue samples modeled as a negative binomial distribution (91). We then identified periods of significant increase or decline in tissue collection, defined as when the first derivative of the predicted fitted value was significantly nonzero (based on a 95% confidence interval) (48).

To further examine spatial patterns in tissue collection, we fitted two GLMs with collection region (i.e., Africa, Americas, Asia, Europe, Oceania) or subregion (i.e., Australia and New Zealand, Central Asia, Eastern Asia, Eastern Europe, Latin America and the Caribbean, Melanesia, Micronesia, Northern Africa, Northern America, Northern Europe, Polynesia, Southeastern Asia, Southern Asia, Southern Europe, Sub-Saharan Africa, Western Asia, Western Europe) as a categorical predictor. One GLM modeled binary sampling status (i.e., whether or not tissue samples have been collected from a country) as a binomial response, with all countries included and region as a predictor, whereas the other GLM modeled the number of tissue samples per country as a Poisson-distributed response, with only sampled countries included and subregion as a predictor. We also fitted two additional GLMs to assess biases in the spatial distribution of museum holdings. Similarly, one GLM modeled binary holding status (i.e., whether or not a country has any tissue samples) as a binomial response, with all countries included and region as a categorical predictor, whereas the other GLM modeled the number of tissue samples per country as a Poisson-distributed response, with all countries included and subregion as a categorical predictor. We assessed model output with the *car* package (92) and model fit with McFadden's R^2 using the *performance* package (93) and visualized differences across regions with the *visreg* package (94).

To assess taxonomic sampling biases in tissue collections, we aligned all sampled bat and rodent species to a mammalian phylogeny (53). This involved relabeling species names to fit this taxonomic backbone where necessary. We first harmonized species names with the *taxize* (95) and *rentrez* (96) packages, followed by subsequent manual relabeling of any remaining taxonomic discrepancies using synonyms listed on the Mammal Diversity Database v2.3 (97).

Recently described species missing from the mammal phylogeny were removed for the taxonomic analysis (53).

We then created a binary response variable for all bat and rodent species in the phylogeny, representing whether or not each species is present in our dataset with associated tissue samples. For those that were present, we calculated the total number of samples with tissues associated with each bat and rodent species. To assess phylogenetic signal in sampling effort within each of bats and rodents, we used the *cape*r package (98) to calculate *D* (0: phylogenetic clustering under Brownian motion; 0-1: intermediate phylogenetic signal; 1: phylogenetically random sampling distribution; >1: overdispersed) for binary sampling effort and Pagel's λ (0: phylogenetically random sampling distribution; 0-1: intermediate phylogenetic signal; 1: phylogenetic clustering under Brownian motion) for the $\log_{10}$-transformed number of samples. Finally, we applied a graph-partitioning algorithm, phylogenetic factorization, using the *phylofactor* package (99), to both the above sets of phylogenies. This approach partitioned binary sampling effort (binomial response) and the $\log_{10}$-transformed number of samples (Gaussian distribution) across the trees using iterative GLMs, producing a list of bat or rodent clades (containing at least three species) that differ significantly from the rest of the tree in sampling effort. We used Holm's sequentially rejective test (100) with a family-wise error rate of 5% to account for multiple comparisons and determine the number of statistically significantly different clades. Trees were visualized using the *ape*, *treeio,* and *ggtree* packages (101, 102, 103).

Acknowledgments

MMJ and DJB were supported by the National Science Foundation (DBI 2515340), and DJB was further supported by the Edward Mallinckrodt, Jr. Foundation. MMJ is also supported by a

Gates Cambridge Scholarship enabled by grant OPP1144 from the Bill & Melinda Gates Foundation.

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008 Feb;451(7181):990–3.

2. Carlson CJ, Brookson CB, Becker DJ, Cummings CA, Gibb R, Halliday FW, et al. Pathogens and planetary change. Nat Rev Biodivers. 2025 Jan;1(1):32–49.

3. Jones SD, Atshabar B, Schmid BV, Zuk M, Amramina A, Stenseth NC. Living with plague: lessons from the Soviet Union's antiplague system. Proc Natl Acad Sci USA. 2019 May;116(19):9155–63.

4. Carlson CJ, Zipfel CM, Garnier R, Bansal S. Global estimates of mammalian viral diversity accounting for host sharing. Nat Ecol Evol. 2019 July;3(7):1070–5.

5. Mollentze N, Streicker DG. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. Proceedings of the National Academy of Sciences. 2020 Apr 28;117(17):9423–30.

6. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. Nature. 2017 Jun;546(7660):646–50.

7. Irving AT, Ahn M, Goh G, Anderson DE, Wang LF. Lessons from the host defences of bats, a unique viral reservoir. Nature. 2021 Jan;589(7842):363–70.

8. Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. Proceedings of the National Academy of Sciences. 2015 Jun 2;112(22):7039-44.

9. Luis AD, Hayman DT, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JR, et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proc Biol Sci. 2013 Feb;280(1756):20122753.

10. Guy C, Ratcliffe JM, Mideo N. The influence of bat ecology on viral diversity and reservoir status. Ecol Evol. 2020 May 8;10(12):5748–58.

11. Grange ZL, Goldstein T, Johnson CK, Anthony S, Gilardi K, Daszak P, et al.; Expert Panel; PREDICT Consortium; University of Edinburgh Epigroup members those who wish to remain anonymous. Ranking the risk of animal-to-human spillover for newly discovered viruses. Proc Natl Acad Sci USA. 2021 Apr;118(15):e2002324118.

12. Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, et al. The phylogenetic range of bacterial and viral pathogens of vertebrates. Mol Ecol. 2020 Sept;29(17):3361–79.

13. Park AW. Phylogenetic aggregation increases zoonotic potential of mammalian viruses. Biology letters. 2019 Dec 24;15(12):20190668.

14. Albery GF, Becker DJ, Brierley L, Brook CE, Christofferson RC, Cohen LE, et al. The science of the host-virus network. Nat Microbiol. 2021 Dec;6(12):1483–92.

15. Gibb R, Albery GF, Becker DJ, Brierley L, Connor R, Dallas TA, et al. Data proliferation, reconciliation, and synthesis in viral ecology [Internet]. bioRxiv; 2021 [cited 2025 Nov 10]. p. 2021.01.14.426572. Available from: https://doi.org/10.1101/2021.01.14.426572

16. Plowright RK, Becker DJ, McCallum H, Manlove KR. Sampling to elucidate the dynamics of infections in reservoir hosts. Philos Trans R Soc Lond B Biol Sci. 2019 Sep;374(1782):20180336.

17. Johnson KR, Owens IFP, the Global Collection Group. A global approach for natural history museum collections. Science. 2023 Mar 24;379(6638):1192–4.

18. Dunnum JL, McLean BS, Dowler RC, Bradley JE, Bradley RD, Carraway LN, et al.; Systematic Collections Committee of the American Society of Mammalogists. Mammal collections of the Western Hemisphere: a survey and directory of collections. J Mammal. 2018 Dec;99(6):1307–22.

19. Cook JA, Arai S, Armién B, Bates J, Bonilla CA, Cortez MB, et al. Integrating Biodiversity Infrastructure into Pathogen Discovery and Mitigation of Emerging Infectious Diseases. Bioscience. 2020 Jul;70(6):531–4.

20. McLean BS, Bell KC, Dunnum JL, Abrahamson B, Colella JP, Deardorff ER, et al. Natural history collections-based research: progress, promise, and best practices. J Mammal. 2016 Jan;97(1):287–97.

21. DiEuliis D, Johnson KR, Morse SS, Schindel DE. Opinion: specimen collections should have a much bigger role in infectious disease research and response. Proc Natl Acad Sci USA. 2016 Jan;113(1):4–7.

22. Dunnum JL, Yanagihara R, Johnson KM, Armien B, Batsaikhan N, Morgan L, et al. Biospecimen Repositories and Integrated Databases as Critical Infrastructure for Pathogen Discovery and Pathobiology Research. PLoS Negl Trop Dis. 2017 Jan;11(1):e0005133.

23. Schindel DE, Cook JA. The next generation of natural history collections. PLoS Biol. 2018 Jul;16(7):e2006125.

24. Thompson CW, Phelps KL, Allard MW, Cook JA, Dunnum JL, Ferguson AW, et al. Preserve a Voucher Specimen! The Critical Need for Integrating Natural History Collections in Infectious Disease Studies. MBio. 2021 Jan;12(1):e02698–20.

25. Yates TL, Mills JN, Parmenter CA, Ksiazek TG, Parmenter RR, Vande Castle JR, et al. The ecology and evolutionary history of an emergent disease: hantavirus pulmonary syndrome. Bioscience. 2002;52(11):989–98.

26. Schmitt CJ, Cook JA, Zamudio KR, Edwards SV. Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. Philos Trans R Soc Lond B Biol Sci. 2018 Nov 19;374(1763):20170387.

27. Yanagihara R, Gu SH, Arai S, Kang HJ, Song JW. Hantaviruses: rediscovery and new beginnings. Virus Res. 2014 Jul;187:6–14.

28. Gieraths U, Beheim-Schwarzbach J, Pickin MJ, Beyer A, Begeman L, Hoffmann B, et al. A century-old museum sample reveals a bandavirus with modern day presence in northern European bats [Internet]. bioRxiv; 2025 [cited 2025 Jun 19]. p. 2025.02.28.640763. Available from: https://doi.org/10.1101/2025.02.28.640763

29. Juman MM, McDonough MM, Ferguson AW, Han BA, Andemwana FB, Cisirika BM, et al. Museum collections and machine learning guide discovery of novel coronaviruses and paramyxoviruses [Internet]. bioRxiv; 2025 [cited 2025 Oct 15]. p. 2025.09.11.675601. Available from: https://doi.org/10.1101/2025.09.11.675601

30. Lund MC, Larsen BB, Rowsey DM, Otto HW, Gryseels S, Kraberger S, et al. Using archived and biocollection samples towards deciphering the DNA virus diversity associated with rodent species in the families Cricetidae and Heteromyidae. Virology. 2023 Aug 1;585:42–60.

31. Leber M, Moncrief ND, Gatens LJ, Michel M, Brinkerhoff RJ. Use of mammalian museum specimens to test hypotheses about the geographic expansion of Lyme disease in the southeastern United States. Ticks and Tick-borne Diseases. 2022 Nov 1;13(6):102018.

32. Wood CL, Welicky RL, Preisser WC, Leslie KL, Mastick N, Greene C, et al. A reconstruction of parasite burden reveals one century of climate-associated parasite decline. Proceedings of the National Academy of Sciences. 2023 Jan 17;120(3):e2211903120.

33. Speer KA, Hawkins MTR, Flores MFC, McGowen MR, Fleischer RC, Maldonado JE, et al. A comparative study of RNA yields from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens. Front Ecol Evol [Internet]. 2022 Aug 2 [cited 2025 Jun 19];10.

34. Zimkus BM, Ford LS. Best practices for genetic resources associated with natural history collections: recommendations for practical implementation. In Collection Forum 2014 (Vol. 28, No. 1-2, pp. 77-112). Department of Psychology, Indiana University, Bloomington, Indiana 47405: The Society for the Experimental Analysis of Behavior, Inc.

35. Phillips CD, Dunnum JL, Dowler RC, Bradley LC, Garner HJ, MacDonald KA, et al. Curatorial guidelines and standards of the American Society of Mammalogists for collections of genetic resources. J Mammal. 2019;100(5):1690–4.

36. Wilde THC, Shukla RK, Madden C, Vodovotz Y, Sharma A, McGraw WS, et al. Simian immunodeficiency virus and storage buffer: Field-friendly preservation methods for RNA viral detection in primate feces. mSphere. 2023 Dec 20;8(6):e0048423.

37. Cook JA, Light JE. The emerging role of mammal collections in 21st century mammalogy. J Mammal. 2019;100(3):733–50.

38. Webster MS. The extended specimen: emerging frontiers in collections-based ornithological research. Boca Rotan (FL): CRC Press; 2017.

39. Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ, Mclean BS. The Open-Specimen Movement. Bioscience. 2021;71(4):405–14.

40. Feng X, Enquist BJ, Park DS, Boyle B, Breshears DD, Gallagher RV, et al. A review of the heterogeneous landscape of biodiversity databases: opportunities and challenges for a synthesized biodiversity knowledge base. Glob Ecol Biogeogr. 2022;31(7):1242–60.

41. Cook JA, Edwards SV, Lacey EA, Guralnick RP, Soltis PS, Soltis DE, et al. Natural History Collections as Emerging Resources for Innovative Education. BioScience. 2014 Aug 1;64(8):725–34.

42. Poisot T, Becker DJ, Catchen MD, Gibb R, Shimabukuro PHF, Carlson CJ. Biodiversity science and biosurveillance are fellow travelers. BioScience. 2025 June 24;biaf091.

43. Stein BR, Wieczorek JR. Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics [Internet]. 2004 Nov 11 [cited 2025 Nov 10];1. Available from: https://journals.ku.edu/jbi/article/view/7

44. Hobern D, Baptiste B, Copas K, Guralnick R, Hahn A, van Huis E, et al. Connecting data and expertise: a new alliance for biodiversity knowledge. Biodivers Data J. 2019 Mar;7:e33679.

45. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE [Internet]. 2012 [cited 2023 Feb 8];7(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3253084/

46. Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, et al. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PLoS One. 2014 Aug;9(8):e102623.

47. Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, et al. The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. Bioscience. 2020 Jan;70(1):23–30.

48. Gibb R, Albery GF, Mollentze N, Eskew EA, Brierley L, Ryan SJ, Seifert SN, Carlson CJ. Mammal virus diversity estimates are unstable due to accelerating discovery effort. Biol Lett. 2022 Jan;18(1):20210427. doi: 10.1098/rsbl.2021.0427. Epub 2022 Jan 5. PMID: 34982955; PMCID: PMC8727147.

49. Esri Inc. ArcGIS Pro Version 3.5.4 [Internet]. Redlands, CA. 2025. Available from: https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

50. Šavrič B., Patterson, T., Jenny, B. 2015. The Natural Earth II world map projection. International Journal of Cartography, 1(2), 123-133. DOI: 10.1080/23729333.2015.1093312

51. Esri. 2025. World Light Gray Base [basemap]. Scale: 1:140,000,000

52. Esri. "World Countries (Generalized)" [data set]. ArcGIS Living Atlas of the World. Redlands, CA. 2022 [accessed 2023 Mar 16]. Available from: https://services.arcgis.com/P3ePLMYs2RVChkJx/arcgis/rest/services/World_Countries_(Generalized)/FeatureServer

53. Upham NS, Esselstyn JA, Jetz W. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biol. 2019;17(12):3000494.

54. Timm RM, Colella JP, McLaren SB, Genoways HH. Innovations that changed mammalogy: frozen tissue collections. J Mammal. 2025 Aug 1;106(4):1045–52.

55. Rohwer VG, Rohwer Y, Dillman CB. Declining growth of natural history collections fails future generations. PLOS Biology. 2022 Apr 19;20(4):e3001613.

56. Hope AG, Sandercock BK, Malaney JL. Benefits for biodiversity sciences and limited impacts on communities of small mammals. Bioscience. 2018 Jan;68(1):35–42.

57. Malaney JL, Cook JA. A perfect storm for mammalogy: declining sample availability in a period of rapid environmental degradation. J Mammal. 2018;99(4):773–88.

58. Anderson S. It costs more to store a whale than a mouse: libraries, collections, and the cost of knowledge. Curator (N Y N Y). 1973;16(1):30–44.

59. Yates TL. The role of voucher specimens in mammal collections: characterization and funding responsibilities. Acta Zool Fenn. 1985;170:81–2.

60. Allmon WD. The value of natural history collections. Curator (N Y N Y). 1994;37(2):83–9.

61. Blackmore S, Donlon N, Watson E. Calculating the financial value of systematic biology collections: proceedings of the international conference, Manchester. In: Nudds JR, Pettitt CW, editors. The value and valuation of natural science collections. London, United Kingdom: The Geological Society; 1997. pp. 17–21.

62. Bradley RD, Bradley LC, Garner HJ, Baker RJ. Cost of collecting and preparing mammal voucher specimens for natural history collections. Occasional Papers, Museum of Texas Tech University; 2012 Dec;313:1-14.

63. Baker RJ, Bradley LC, Garner HJ, Bradley RD. “Door to Drawer” costs of curation, installation, documentation, databasing, and long-term care of mammal voucher specimens in natural history collections. Occasional Papers, Museum of Texas Tech University; 2014 Apr;323:1–15.

64. Dalton R. Natural history collections in crisis as funding is slashed. Nature. 2003 Jun;423(6940):575.

65. Kemp C. Museums: the endangered dead. Nature. 2015 Feb;518(7539):292–4.

66. Gropp RE. Are university natural science collections going extinct? Bioscience. 2003;53(6):550.

686 67. Hahn EE, Alexander M, Stiller J, Grewe PM, Holleley CE. Hot alkaline lysis gDNA
687 extraction from formalin-fixed archival tissues. PLOS ONE. 2024 Jan 2;19(1):e0296491.
688 68. Hämmerle M, Guellil M, Trgovec-Greif L, Cheronet O, Sawyer S, Ruiz-Gartzia I, et al.
689 Screening great ape museum specimens for DNA viruses. Sci Rep. 2024 Nov 30;14(1):29806.
690 69. Phelps KL, Hamel L, Alhmoud N, Ali S, Bilgin R, Sidamonidze K, et al. Bat Research
691 Networks and Viral Surveillance: Gaps and Opportunities in Western Asia. Viruses. 2019
692 Mar;11(3):240.
693 70. Jeeves SP, Kotwa JD, Pearl DL, Pickering BS, Bowman J, Mubareka S, et al. Coronaviruses
694 in wild rodent and eulipotyphlan small mammals: a review of diversity, ecological implications
695 and surveillance considerations. Journal of General Virology. 2025;106(7):002130.
696 71. Colella JP, Agwanda BR, Anwarali Khan FA, Bates J, Carrión Bonilla CA, de la Sancha NU,
697 et al. Build international biorepository capacity. Science. 2020 Nov 13;370(6518):773–4.
698 72. Cohen LE, Fagre AC, Chen B, Carlson CJ, Becker DJ. Coronavirus sampling and
699 surveillance in bats from 1996–2019: a systematic review and meta-analysis. Nat Microbiol.
700 2023 June;8(6):1176–86.
701 73. Juman MM, Restif O, Becker DJ. Paramyxoviruses in Old World fruit bats (Pteropodidae):
702 an open database and synthesis of sampling effort, viral positivity, and coevolution. PLoS Negl
703 Trop Dis. 2025 Nov;19(11):e0013698.
704 74. Olival KJ, Hayman DTS. Filoviruses in Bats: Current Knowledge and Future Directions.
705 Viruses. 2014 Apr;6(4):1759–88.
706 75. Becker DJ, Albery GF, Sjodin AR, Poisot T, Bergner LM, Dallas TA, et al. Optimizing
707 predictive models to prioritize viral discovery in zoonotic reservoirs. bioRxiv. 2021 Jan
708 1;2020.05.22.111344.

76. Heidorn PB. Shedding light on the dark data in the long tail of science. Libr Trends. 2008;57(2):280–99.

77. Colella JP, Bates J, Burneo SF, Camacho MA, Carrion Bonilla C, Constable I, et al. Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. PLoS Pathog. 2021 Jun;17(6):e1009583.

78. Wen J, Ickert-Bond SM, Appelhans MS, Dorr LJ, Funk VA. Collections-based systematics: opportunities and outlook for 2050. J Syst Evol. 2015;53(6):477–88.

79. Funk VA. Collections-based science in the 21st Century. J Syst Evol. 2018;56(3):175–93.

80. GBIF.org (14 July 2025) GBIF Occurrence Download https://doi.org/10.15468/dl.s8tqad

81. GBIF.org (13 July 2025) GBIF Occurrence Download https://doi.org/10.15468/dl.7j3bhm

82. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2021. Available from: https://www.R-project.org/.

83. Wang A. avenews [Internet]. 2022 [cited 2023 Feb 8]. Available from: https://github.com/albertyw/avenews/blob/6f2d89a5f68807348400869fb560e0a2bee94809/old/data/average-latitude-longitude-countries.csv

84. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

85. Wickham H, François R, Henry L, Müller K, Vaughan D. _dplyr: A Grammar of Data Manipulation_. R package version 1.1.4 [Internet]. 2023. Available from: <https://CRAN.R-project.org/package=dplyr>.

86. Wickham H, Vaughan D, Girlich M. tidyr: Tidy Messy Data. R package version 1.3.1 [Internet]. 2024. Available from: <https://CRAN.R-project.org/package=tidyr>.

87. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. R package version 4.0.0 [Internet]. 2016. Available from: https://ggplot2.tidyverse.org.

88. Pedersen T. patchwork: The Composer of Plots. R package version 1.3.0 [Internet]. 2024. Available from: <https://CRAN.R-project.org/package=patchwork>.

89. Dowle M, Srinivasan A. data.table: Extension of `data.frame`. R package version 1.14.4 [Internet]. 2022. Available from: <https://CRAN.R-project.org/package=data.table>.

90. Wickham H.. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.1 [Internet]. 2023. Available from: <https://CRAN.R-project.org/package=stringr>.

91. Wood SN. Generalized Additive Models: An Introduction with R, Second Edition. 2nd ed. New York: Chapman and Hall/CRC; 2017. 496 p.

92. Fox J, Weisberg S. An {R} Companion to Applied Regression, Third Edition. Thousand Oaks, CA, 2019. Sage.

93. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. J Open Source Softw. 2021 Apr 21;6(60):3139.

94. Breheny P, Burchett W (2017). Visualization of Regression Models Using visreg. The R Journal, 9: 56-71.

95. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R [Internet]. F1000Research; 2013 [cited 2025 Nov 1]. Available from: https://f1000research.com/articles/2-191

96. Winter DJ. rentrez: an R package for the NCBI eUtils API. 2017. The R Journal 9(2):520-526

97. Burgin CJ, Zijlstra JS, Becker MA, Handika H, Alston JM, Widness J, et al. How many mammal species are there now? Updates and trends in taxonomic, nomenclatural, and geographic knowledge [Internet]. bioRxiv; 2025 [cited 2025 Oct 5]. p. 2025.02.27.640393. Available from: https://www.biorxiv.org/content/10.1101/2025.02.27.640393v1

98. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, et al. caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 1.0.3 [Internet]. 2023. Available from: https://CRAN.R-project.org/package=caper

99. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, et al. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 2017, Feb 9;5:e2969.

100. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat. 1979;6(2):65–70.

101. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019, 35, 526-528. doi:10.1093/bioinformatics/bty633 <https://doi.org/10.1093/bioinformatics/bty633>.

102. Wang, LG, Lam TTY, Xu S, Dai Z, Zhou L, Feng T, et al. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Molecular Biology and Evolution 2020, 37(2):599-603.

103. Yu G, Smith D, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017, 8(1):28-36.

Supporting information captions

Supplementary Fig. 1. Predicted values and 95% confidence intervals from a generalized linear model of binary sampling of a country for bat and rodent tissues (0: no sampling; 1: at least one bat or rodent tissue sample collected from country) across five global regions, generated with the *visreg* package in R.

Supplementary Fig. 2. Predicted values and 95% confidence intervals from a generalized linear model of number of bat and rodent tissue samples collected from countries across all global subregions, generated with the *visreg* package in R.

Supplementary Fig. 3. Predicted values and 95% confidence intervals from a generalized linear model of the binary tissue holding status of a country (0: no bat or rodent tissue samples; 1: at least one bat or rodent tissue sample) across five global regions, generated with the *visreg* package in R.

Supplementary Fig. 4. Predicted values and 95% confidence intervals from a generalized linear model of number of bat and rodent tissue samples housed in countries across all global subregions, generated with the *visreg* package in R.

Supplementary Data S1. Complete list of institutions included in our filtered dataset, including the number of bat, rodent, and total samples as well as the country code for each institution.

Supplementary Data S2. Complete list of institutions holding tissue samples from the ten rodent and ten bat species predicted to harbor the greatest number of undiscovered zoonoses.