ArHa manuscript 1: A global synthesis of rodent- and shrew-borne Arenaviruses and Hantaviruses

Table of contents

# Abstract

**Needs revisiting**

Rodents and shrews are key reservoirs for several zoonotic viruses in the families *Arenaviridae* and *Hantaviridae*, including pathogens causing viral haemorrhagic fevers with significant public health impact. Yet, our understanding of host-virus associations, discovery effort, and global sampling gaps remains fragmented. We conducted a systematic review of Arenavirus and Hantavirus detection in small mammals from 1977 to 2025, identifying 729 studies reporting on over 579,000 rodents (Order: Rodentia), 12,100 shrews (Order: Eulipotyphla; formerly Soricomorpha) and 2,500 other mammals.

From these, we extracted over 715,000 assays comprising 69,993 pathogen detection records, harmonised to the highest spatiotemporal and taxonomic resolution. We further linked 5,650 pathogen RNA sequences and 2,406 host DNA sequences from GenBank with associated assay and sampling metadata. The result is a global, open-access dataset capturing small mammal testing effort, pathogen detection and sequence availability.

We describe host-virus associations and use co-phylogenetic analysis to examine evolutionary congruence between pathogens and hosts. We assess the geographic and taxonomic sampling biases and identify undersampled regions and clades with high potential for viral discovery. Comparative analyses of virus with and without known zoonotic potential highlight traits associated with host breadth and spillover risk. This harmonised resource enables new insights into the ecology and evolution of rodent- and shrew-borne viruses and offers a foundational resource for prioritising surveillance and forecasting zoonotic emergence.

# Introduction

Broad Context: Arenaviruses and hantaviruses, hosted by small mammals (Rodentia and Eulipotyphla), are globally distributed and include significant zoonotic pathogens [Refs].

The Classic View vs. New Evidence: Historically, these viruses were considered to have tight, co-evolved relationships with single host species [Refs]. However, increasing surveillance efforts reveal a more complex picture, with a growing number of host species associated with each pathogen [Refs].

The Central Problem - The “Surveillance Landscape”: Our understanding of host range, viral sharing, and zoonotic risk is fundamentally shaped by a biased “surveillance landscape.” The true ecological picture is obscured by where, when, and for what we choose to look, often prioritizing areas and species near human populations (i.e., synanthropic species) or focusing efforts after major outbreak events [Refs]. An understanding of the host range of each viral species is driven by research effort, or more accurately, sampling effort across the small-mammal community in the endemic region of the pathogen.

The Knowledge Gap: The role of these multiple potential host species in sustaining transmission and their capacity for zoonotic spillover remains an open and critical question [Refs]. We lack a quantitative framework to map the boundaries of our collective knowledge and identify the key taxonomic, geographic, and temporal gaps in global surveillance. Furthermore, the translation of field sampling into publicly available genetic data, crucial for downstream analyses like phylogenetics and diagnostics development, may represent another significant bottleneck and potential bias in our understanding.

Study Aim: This study addresses this gap by synthesizing the global literature on Arenaviridae and Hantaviridae in wild small mammals. We construct a comprehensive, spatio-temporally explicit database to:

* Describe the current state of global surveillance data.
* Quantify the taxonomic, geographic, and temporal biases inherent in this data.
* Provide an evidence-based roadmap for future research and pandemic preparedness by identifying critical surveillance gaps and understanding the impact of historical events and data reporting lags.
* Use this synthesized data to test initial hypotheses about the ecological drivers of host-pathogen associations, employing network analysis and statistical models designed to explicitly account for the identified surveillance biases (particularly sampling effort and the influence of synanthropy).

# Methods

## Systematic review

To synthesise available data on the sampling of Arenviridae and Hantaviridae in wild-living animals we conducted a systematic review of the literature.

We searched NCBI PubMed and Clarivate Web of Science using the following terms.

* Rodent\* OR Shrew
* Arenvir\* OR Hantavir\*
* a AND b

Searches were conducted on YYYY-MM-DD

Inclusion and exclusion criteria were as follows:

* Inclusion
* Exclusion

A total of A citations were returned; after de-duplication, B citations were screened, supplemented with C manuscripts from citation chaining. Titles and abstracts were screened, with D records carried forward to full-text review. Of these, E manuscripts contained data suitable for extraction.

## Dataset construction and harmonisation

We developed a relational data extraction tool to produce four linked datasets: Descriptive, Host, Pathogen, and Sequence. Data was aligned with Darwin Core Terms to facilitate future deposition to the Global Biodiversity Information Facility (GBIF).

Key extracted variables included: publication metadata, sampling effort, host species and abundance, sampling location and time, pathogen assay results (tested, positive), and GenBank accession numbers.

A performed harmonisation of taxonomic and trait data. We used the Mammal Diversity Database (MDD) as the authoritative taxonomic backbone for hosts. All unique host scientific names from our literature review, as well as those from external trait datasets, were programmatically resolved against the GBIF taxonomy using the R package taxize. This generated a stable gbif\_id for each species, which was used as the primary key to join all host-level data. This process explicitly resolved taxonomic synonyms, ensuring data from different sources were correctly aggregated. External host trait data were sourced from EltonTraits 1.0 (for body mass, diet, and activity patterns) and a synanthropy database [ref ecke]. Geographic range polygons were sourced from the IUCN Red List [ref].

For pathogens, a similar process was used. Raw virus names were standardized against a manually curated dictionary, and these cleaned names were resolved against the NCBI Taxonomy database using taxize to retrieve a unique ncbi\_id and the complete taxonomic lineage for each virus.

## Analysis of surveillance biases

To quantify biases in the global surveillance landscape, we conducted a series of analyses across taxonomic, geographic, and temporal dimensions. All analyses were conducted in R (version 4.2.3). We quantified surveillance biases across taxonomic, geographic, and temporal dimensions.

### Host sampling biases

We quantified taxonomic sampling completeness by comparing the 745 host species in our database against the comprehensive Mammal Diversity Database checklist of 3,457 extant species in the orders Rodentia and Eulipotyphla. To formally test for factors associated with sampling intensity, we fitted a series of nested Generalized Additive Models (GAMs) with a Negative Binomial error distribution, using the total number of pathogen samples reported for each host species as the response variable. The primary analysis was conducted on all species with complete data for predictors including: biogeographic realms (dummy coded), log-transformed geographic range size, log-transformed adult body mass, and activity pattern. A focused sub-analysis was performed on the subset of 355 species with available data on synanthropy to test its influence. Model selection was guided by the Akaike Information Criterion (AIC).

### Pathogen sampling biases

We examined surveillance bias towards specific viruses by summarizing the total number of individuals tested and positive detections for each virus species. To quantify the breadth of surveillance, we calculated the number of unique virus species tested for within each host species, after parsing records that reported testing for multiple pathogens simultaneously. We also assessed the representativeness of genetic data by calculating the proportion of high-confidence (PCR or culture-based) positive detections for each virus that had at least one corresponding sequence deposited in GenBank.

### Geographic Sampling Biases

To quantify geographic biases in surveillance, we analyzed sampling effort at multiple spatial and taxonomic scales. First, we conducted a descriptive analysis to quantify sampling gaps within host ranges. We performed a spatial intersection between a global administrative level 2 (ADM2) polygon layer (GADM) and the complete IUCN range maps for all species in our study. This produced a comprehensive, lookup table identifying every ADM2 unit inhabited by each host species, which served as the foundation for subsequent analyses. Using this, we calculated the proportional area sampled for all host genera and the ten most-sampled species. This was defined as the sum of the areas of all inhabited ADM2 units containing at least one sampling record, divided by the total area of all inhabited ADM2 units across the taxon’s IUCN range. We also quantified sampling effort occurring outside of the IUCN-defined native ranges.

Next, to formally test hypotheses about factors associated with geographic sampling patterns, we used a hierarchical modeling approach. At a coarse, national scale, we tested the hypothesis that a country’s economic output is associated with surveillance effort by fitting two separate Bayesian regression models. To account for the confounding effect of country size, both models included the log-transformed total population as an offset term, effectively modeling the per-capita rate of sampling. Models were fitted using the brms package in R with weakly informative priors. The two models were; 1) a sampling intensity model, where the response variable was the log-transformed total number of hosts sampled and 2) a sampling breadth model, where the response variable was the number of unique host species sampled.

At a finer, subnational scale, we tested the hypothesis that sampling effort is concentrated in areas of high human influence and specific ecological contexts. For this, we fitted two separate Bayesian Generalized Linear Mixed Models (GLMMs) where the unit of analysis was the ADM2 polygon. Given the extreme data sparsity (98.3% of units had zero samples), we compared model fits and confirmed that a Zero-Inflated Negative Binomial (ZINB) distribution provided a significantly better fit than a standard Negative Binomial for both response variables. Similar to the national scale models, the two models were; 1) a sampling intensity model, where the response variable was the total number of hosts sampled and 2) a sampling breadth model, where the response variable was the number of unique host species sampled.

Both GLMMs included the log-transformed area of each polygon as an offset term (to account for the Modifiable Areal Unit Problem) and country as a random intercept (to account for non-independence). Fixed-effect predictors included mean human population density (WorldPop), mean nighttime light intensity (VIIRS), accessibility, and host species richness. All models were fitted using the brms package.

To identify surveillance hotspots and coldspots, we used the fitted subnational models to generate an expected baseline. For all ADM2 units with non-zero sampling, we calculated population-level predictions (the mean of the posterior expected value, posterior\_epred, with re\_formula=NA) for both models. We then calculated a log-scale residual (log10(Observed) - log10(Predicted)) for each unit. These residuals, representing the “surveillance gap,” were then mapped to visualize spatial patterns, and the top 10 hotspots and bottom 10 coldspots were extracted for analysis. Model fit was assessed using posterior predictive checks (PPCs) and randomized quantile residuals via the DHARMa package.

### Temporal Sampling Biases

We quantified temporal biases by aggregating surveillance effort and diversity by the year of sample collection. We used the start year of sampling reported in host records, coalescing with the publication year for records missing a sample date. We quantified two metrics: 1) Sampling Effort, by summing the the number of hosts sampled and counting the number of publications per year; and 2) Sampling Diversity, by counting the number of host species and number of pathogen species (stratified by family) per year. These trends were plotted to visualize patterns and provide context for subsequent modeling.

To formally analyze temporal biases, we modeled annual sampling intensity from 1966–2019 as a function of time. We fitted a Generalized Additive Model (GAM) with a Negative Binomial distribution and log-link. The model was structured to estimate a global smooth trend as well as continent-specific deviations from that trend, while also controlling for baseline differences in sampling effort between continents.

## Bipartite networks of host–pathogen associations

### Descriptive Network Analysis

To visualize and quantify the structure of observed host-pathogen interactions, we constructed bipartite networks separately for Arenaviridae and Hantaviridae using the igraph and bipartite packages. To account for diagnostic uncertainty, we stratified the analysis by detection method. First, we created edge lists based on unique host-species-pathogen-species pairs. We then filtered these lists to create two evidence strata; 1) the Acute Evidence Network, including only positive detections derived from direct methods (PCR, culture, sequencing) and 2) the All Evidence Network, including both direct and indirect detection methods (serology).

From these edge lists, we generated binary incidence matrices (host species × pathogen species) for each family and evidence stratum. We calculated key network-level structural metrics, including the number of connected components, modularity, and nestedness, using the bipartite package. We also calculated node-level degree centrality (number of links per species) to identify highly connected hosts and pathogens. Finally, to assess the potential confounding influence of surveillance bias on network structure, we calculated the total sampling effort (number of individuals tested) for each host species within each evidence stratum and tested its correlation (Spearman’s rank correlation) with host degree centrality. We also performed a modularity analysis on a combined network including both virus families to assess their ecological partitioning.

### Modeling Host-Virus Link Probability Using Dyadic GLMMs

To disentangle the effects of host sampling effort and the ecological associations of host-virus links, we employed a dyadic modeling approach using Bayesian GLMMs. We treated each potential host-pathogen pair (dyad) within a virus family (Arenaviridae or Hantaviridae) as the unit of analysis. The response variable was binary, coded as 1 if a positive interaction between that specific host and virus species was observed in our dataset (“known positive”), and 0 otherwise (representing both tested-negative and never-tested dyads).

We fitted separate models for each virus family using a Bernoulli distribution with a logit link. To explicitly control for surveillance bias, we included the log-transformed total sampling effort for each host species (scaled) as a fixed effect covariate. To test our ecological hypotheses, we included host log-transformed geographic range size (scaled) and host synanthropy status (categorical: Not Synanthropic (reference), Occasionally Synanthropic, Totally Synanthropic) as fixed effects. To account for the non-independence inherent in network data (i.e., multiple links involving the same host or pathogen), we included random intercepts for both host species and pathogen species Due to data availability, models including synanthropy were fitted on the relevant subset of host species. The primary objective was to determine if host ecological traits significantly predict the probability of a host-virus link after statistically controlling for sampling effort and accounting for baseline host and pathogen differences.

# Results

## Description of the Dataset

Our systematic review yielded 729 unique studies suitable for extraction, with publication dates ranging from 1,974 to 2,025 The combined ddatasets represents a substantial global surveillance effort, documenting 595,962 individual small mammals sampled from 102 distinct countries across all major biogeographic realms. The host data is taxonomically broad, comprising 668 unique species from 214 genera and 23 familes within the order Rodentia and Elipotyphla.

From these host records, we synthesised 716,499 individual pathogen assays, which included 70,028 unique pathogen detection records. This effort targeted 87 distinct viral species within the Arenaviridae (35) and Hantaviridae (52) families. To create a link to genetic data, we associated these field records with 5,593 pathogen sequences and 2,375 host sequences from GenBank, resulting in a rich, multi-layered database that connects sampling, detection, and genetic characterisation at high resolution.

## The Surveillance Landscape: Quantified Biases

### Taxonomic and Trait-Based Biases in Host Surveillance

Our analysis reveals a profound taxonomic bias in global surveillance efforts. Of the 3,457 included extant species of rodents (2,819) and shrews (615), only 668 (19.30%) were found to have been sampled for either arenaviruses or hantaviruses in our systematic review. The sampling effort is highly concentrated in a few families, such as Cricetidae (260) and Muridae (187), while many other families remain substantially unsampled (Figure 1.). This bias is also geographic, with the proportion of small-mammal species in the Palearctic (179/731, 24.50%) and Nearctic (191/494, 38.70%) realms being significantly higher compared to other realms (Figure 2.). Furthermore, surveillance is not evenly distributed across conservation statuses; species listed as “Least Concern” by the IUCN (1,683) comprise the vast majority of all sampled records (559/1,683, 33.20) (Figure 3.)

|  |
| --- |
| Figure 1: Figure 1 Circular familial bias plot |

|  |
| --- |
| Figure 2: Figure 2 Circular realm bias plot |

|  |
| --- |
| Figure 3: Figure 3 IUCN status bias plot |

To disentangle the factors associated with this bias, we fitted a series of nested GAMs On our primary analytical dataset (n = 2,122 species). Model comparison showed that the inclusion of host traits significantly improved predictions of sampling intensity (ΔAIC > 500). In our final model for this dataset, which explained 47.50% of the deviance, both host geographic range size (GAM, edf = 3.62, *p* < 0.001) and adult body mass (GAM, edf = 5.80, *p* < 0.001) were significant, non-linear predictors of the number of pathogen samples collected. After accounting for these biological traits, the observed sampling bias towards the Neotropic and Afrotropic realms was no longer statistically significant, while persistent, strong positive effects remained for the Palearctic (Estimate = 1.4, 95% Confidence Interval (CI) = 0.62-2.18 *p* < 0.001) and Nearctic (Estimate = 1.92, 95% CI = 1.07-2.77, *p* < 0.001) realms. Host activity pattern was also a significant predictor, with nocturnal (Estimate = 0.71, 95% CI = 0.26-1.15 *p* = 0.002) and diurnal (Estimate = 0.84, 95% CI = 0.19-1.49, *p* = 0.011) species being associated with a greater sampling effort than cathemeral species.

### The Overwhelming Influence of Synanthropy

In a focused sub-analysis on the subset of 355 species with available synanthropy data, a species’ propensity to live near humans was the most powerful predictor of sampling intensity. The inclusion of synanthropy provided the best-fitting model (ΔAIC > 48), and species classified as Occasionally Synanthropic (Estimate = 2.11, 95% CI = 1.56-2.66, *p* < 0.001) or Totally Synanthropic (Estimate = 4.56, 95% CI = 2.46-6.66, *p* < 0.001) were associated with a substantial and significant increase in sampling effort. In this final model, which explained 45.10% of the deviance, the inclusion of synanthropy attenuated most other effects; host activity pattern and most biogeographic realms (including the Palearctic) became non-significant. Only host range size, body mass, and the Nearctic (Estimate = 1.87, 95% CI = 0.95-2.8, *p* < 0.001) and Indomalayan (Estimate = 1.22, 95% CI = 0.27-2.15, *p* = 0.011) realms retained a significant independent association with sampling effort.

### Taxonomic Biases in Pathogen Surveillance

Surveillance efforts for pathogens were similarly biased, concentrating heavily on a small number of well-known zoonotic agents (Figure 4.). Of the 87 virus species in the Arenaviridae and Hantaviridae families assayed for in our dataset, the top seven most-tested viruses (*Orthohantavirus sinnombreense*, *O. hantanense*, *O. puumalaense*, *O. andesense*, *O. seoulense*, *O. dobravaense* and *Mammarenavirus lassaense*), accounted for over 382,003 of the 716,499 total samples tested (53.30%), representing a significant proportion of the total surveillance effort.

|  |
| --- |
| Figure 4: Figure 4 Pathogen sampling plot |

We also found a strong bias in the breadth of viral screening applied to different hosts (Figure 5.). A few widespread, often synanthropic, host species serve as the primary targets for broad-spectrum virus screening; *M. musculus*, *R. norvegicus*, and *R. rattus* have been tested for 63, 46, and 40 distinct virus species, respectively. In contrast, the vast majority of host species have only been tested for a single virus, suggesting that most studies employ a targeted “case-finding” approach rather than broad, discovery-oriented surveillance.

|  |
| --- |
| Figure 5: Figure 5 Surveillance breadth plot |

### Could have something about co-detection - co-phylogeny here?

### Gaps in Genetic Data Representativeness

Our analysis reveals significant and uneven gaps between field sampling effort and the public availability of genetic data. Geographically, sequencing effort is highly variable (Figure 6.). While a few countries with smaller overall sampling effort show high completeness (e.g., Pakistan, 100% of 11 hosts have associated sequences), major surveillance gaps exist. For instance, in China, over 71,000 hosts were sampled, but these contributed to host sequences from records totaling fewer than 50 animals (<0.1% completeness). Similarly, in the USA, records representing over 185,000 sampled hosts yielded host sequences from records totaling only 40 animals (<0.03% completeness).

|  |
| --- |
| Figure 6: Figure 6 Geographic univariate maps |

|  |
| --- |
| Figure 7: Figure 7 Geographic bivariate map |

This disparity is also evident at the taxonomic level for the most well-studied hosts (Figure 8). For most of the top-sampled species, such as *Peromyscus maniculatus* (113,804 individuals sampled) and *Myodes glareolus* (64,398 individuals), the proportion of individuals with either host or pathogen sequences is extremely low (<2%). Notably, some key hosts like *Mastomys natalensis* and *Suncus murinus* show a comparatively higher proportion of host sequencing (5.1% and 8.8%, respectively), indicating a targeted effort to generate host genetic data for these important reservoir species.

|  |
| --- |
| Figure 8: Figure 8 Host Representativeness |

A similar gap exists for pathogens (Figure 9). Several well-known viruses like PUUV are well-represented, with sequences available for a large proportion of their PCR-positive detections. However, other heavily studied pathogens are genetically under-represented. For example, despite being one of the most frequently detected arenaviruses, fewer than 15% of PCR-positive Mopeia mammarenavirus detections have associated sequences. More striking gaps exist for pathogens like Prospect Hill orthohantavirus, for which over 70 positive detections were reported in our dataset but no corresponding sequences were found.

|  |
| --- |
| Figure 9: Figure 9 Pathogen Representativeness |

## Geographic and Temporal Gaps in Surveillance

### Broad-Scale Sampling Patterns

Geographic sampling is highly concentrated. Our analysis of the top 10 most-sampled host species reveals that for most, surveillance has occurred in less than 15% of their inhabited administrative regions (Table X). A notable exception is *Oligoryzomys longicaudatus*, the primary reservoir for Andes virus, for which 74.6% of its inhabited area has been sampled. Furthermore, for highly invasive synanthropic species, a large proportion of all positive detections occurred outside their IUCN-defined native range (e.g., 49.3% for *Rattus norvegicus* and 32.6% for *Mus musculus*), highlighting a strong bias towards studying these species in their invaded, human-commensal contexts.

At the national level, we found a significant negative association between a country’s economic output and its per-capita surveillance rates. A Bayesian linear model, controlling for total population via an offset, showed that higher GDP per capita was associated with a lower rate of sampling intensity (log-transformed total hosts sampled; Estimate = -0.56, 95% Credible Interval [CrI] = -0.78 - -0.34). This pattern was even stronger for sampling breadth, where higher GDP per capita was also associated with a lower rate of unique host species sampled (Estimate = -1.72, 95% CrI = -1.95 - -1.49).

### Subnational Surveillance Model and Diagnostics

At the subnational level, the data was characterized by extreme sparsity, with 98.3% of the 38,897 ADM2 units reporting zero sampled hosts. This confirmed a two-part sampling process, and Zero-Inflated Negative Binomial (ZINB) models provided a significantly better fit than standard Negative Binomial models for both sampling intensity (ΔELPD = 846.5 ± 52.1) and breadth (ΔELPD = 619.4 ± 43.2).

Posterior predictive checks confirmed the ZINB structure was appropriate, as the proportion of zeros simulated by the models closely matched the observed data. Diagnostics using DHARMa confirmed the models successfully captured data dispersion (Dispersion test, *p* > 0.05) and showed no evidence of heteroskedasticity in the residual-vs-predicted plots. While the large dataset (n = 38,897) provided power to detect minor, statistically significant deviations (KS test, *p* < 0.05), the diagnostic checks support the use of the models for establishing a baseline for gap analysis.

### Factors Associated with Subnational Sampling Effort

The models’ zero-inflation component, which predicts the probability of an area being sampled at all, showed similar drivers for both intensity and breadth. A higher likelihood of any sampling (i.e., a lower probability of being an “excess zero”) was significantly associated with greater socioeconomic activity (higher nighttime lights; Intensity Est. = -5.11, 95% CrI = -9.62 - -1.28 and Breadth Est. = -4.41, 95% CrI = -10.53 - -0.33) and greater accessibility (lower travel time; Intensity Est. = -2.00, 95% CrI = -3.59 - -0.84). For the intensity model, higher host richness was also associated with a higher probability of being sampled (Est. = -0.68, 95% CrI = -1.38 - -0.12).

However, the models’ count component, predicting the magnitude of sampling once it occurs, revealed a critical divergence. For sampling intensity, higher nighttime lights (Est. = 0.50, 95% CrI = 0.14 - 0.87) and greater accessibility (Est. = -0.49, 95% CrI = -0.78 - -0.19) were associated with sampling more hosts, but host richness and population density were not. In sharp contrast, for sampling breadth, host richness was a strong positive predictor (Est. = 0.40, 95% CrI = 0.26 - 0.55), along with population density (Est. = 0.28, 95% CrI = 0.07 - 0.48), nighttime lights (Est. = 0.30, 95% CrI = 0.09 - 0.50), and accessibility (Est. = -0.51, 95% CrI = -0.69 - -0.29).

### Spatial Surveillance Gap Analysis

To identify surveillance gaps, we mapped the log-scale residuals (log10(Observed/Predicted)) from both models for all ADM2 units with non-zero sampling.

The sampling intensity (depth) model revealed a highly heterogeneous spatial pattern (Figure 10A). Hotspots and coldspots were often interspersed at a fine geographic scale. This is exemplified in insets (e.g., Figure 1C, showing parts of East Asia), which show adjacent districts with opposing surveillance gaps.

|  |
| --- |
| Figure 10: Figure 10 Sampling Intensity Hotspots and Coldspots |

In contrast, the sampling breadth (diversity) model showed significantly more spatial homogeneity (Figure 11A). Gaps and hotspots were clustered into larger, more geographically contiguous regions.

|  |
| --- |
| Figure 11: Figure 11 Sampling Beadth Hotspots and Coldspots |

We extracted the top 10 and bottom 10 ADM2 districts to prioritize these gaps (Tables 1 & 2). We identified critical gaps (i.e., double coldspots), which are the districts appearing on both coldspot lists for intensity and breadth, an example being Zunyi, China. These represent the highest-priority surveillance gaps, as there is substantially lower sampling effort reported from this region. The coldspots identified in countries with higher proportions of sampled ADM2 districts (e.g., USA and China) reflects the within-country heterogeneity where surveillance effort is not evenly distributed. We further identified context-dependent hotspots (e.g., Al-Manshiyah, Egypt and Kepulauan Seribu, Indonesia) where sampling was substantially higher than would be expected based on sampling elsewhere in the country.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Hotspot and Coldspots Identified for Sampling Intensity   | Analysis Metric | Classification | Country | Country ADM2 Sampled | District (ADM2) | Observed | Predicted | Log-Residual | Country Context (Mean [Range]) | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Sampling intensity (number of individuals)** | **Hotspot** | Egypt | 4/343 (1.2%) | Al-Manshiyah | 861 | 18.82 | 3.99 | 2.56 [1.62 - 3.99] | | Tanzania | 46/186 (24.7%) | Morogoro Urban | 13,427 | 62.05 | 3.92 | 0.63 [-0.76 - 3.92] | | Namibia | 6/107 (5.6%) | Moses Garoeb | 780 | 18.03 | 3.39 | 0.19 [-1.34 - 3.39] | | Indonesia | 5/502 (1%) | Kepulauan Seribu | 229 | 10.59 | 3.32 | 1.14 [-0.27 - 3.32] | | Russia | 55/2445 (2.2%) | Yakovlevskiy rayon | 7,979 | 49.50 | 3.28 | 0.35 [-1.27 - 3.28] | | Hungary | 8/168 (4.8%) | Komlói | 5,471 | 42.02 | 3.26 | 1.23 [-0.79 - 3.26] | | Paraguay | 8/218 (3.7%) | Villa Ygatimí | 6,018 | 43.79 | 3.11 | 0.75 [-0.64 - 3.11] | | Brazil | 51/5572 (0.9%) | Jaborá | 1,545 | 24.26 | 3.01 | 0.8 [-1.57 - 3.01] | | Taiwan | 13/22 (59.1%) | Lienkiang | 138 | 8.50 | 3.00 | 1.2 [-0.08 - 3] | | United States | 221/3148 (7%) | Lake | 8,813 | 51.68 | 2.93 | 0.39 [-2.13 - 2.93] | | **Coldspot** | United States | 221/3148 (7%) | Weld | 1 | 1.00 | -1.82 | 0.39 [-2.13 - 2.93] | | China | 82/364 (22.5%) | Tongren | 1 | 1.00 | -1.83 | 0.09 [-2.06 - 2.77] | | United Kingdom | 23/183 (12.6%) | Derbyshire | 1 | 1.00 | -1.84 | -0.58 [-1.86 - 1.36] | | United Kingdom | 23/183 (12.6%) | Suffolk | 1 | 1.00 | -1.86 | -0.58 [-1.86 - 1.36] | | China | 82/364 (22.5%) | Taizhou | 2 | 1.35 | -1.94 | 0.09 [-2.06 - 2.77] | | China | 82/364 (22.5%) | Shaoxing | 2 | 1.35 | -1.95 | 0.09 [-2.06 - 2.77] | | China | 82/364 (22.5%) | Guiyang | 1 | 1.00 | -1.99 | 0.09 [-2.06 - 2.77] | | Finland | 17/21 (81%) | Pirkanmaa | 1 | 1.00 | -2.01 | -0.3 [-2.01 - 1.96] | | China | 82/364 (22.5%) | Zunyi | 1 | 1.00 | -2.06 | 0.09 [-2.06 - 2.77] | | United States | 221/3148 (7%) | Shelby | 1 | 1.00 | -2.13 | 0.39 [-2.13 - 2.93] | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Hotspot and Coldspots Identified for Sampling Breadth   | Analysis Metric | Classification | Country | Country ADM2 Sampled | District (ADM2) | Observed | Predicted | Log-Residual | Country Context (Mean [Range]) | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Sampling breadth (number of species)** | **Hotspot** | Egypt | 4/343 (1.2%) | Al-Manshiyah | 4 | 1.83 | 4.03 | 2.56 [1.62 - 3.99] | | Taiwan | 13/22 (59.1%) | Lienkiang | 3 | 1.61 | 3.42 | 1.2 [-0.08 - 3] | | Indonesia | 5/502 (1%) | Kepulauan Seribu | 2 | 1.35 | 3.41 | 1.14 [-0.27 - 3.32] | | Japan | 43/1811 (2.4%) | Yonaguni | 1 | 1.00 | 3.33 | 0.85 [-0.76 - 2.62] | | Estonia | 13/223 (5.8%) | Kuressaare | 2 | 1.35 | 3.30 | 1.24 [-0.44 - 2.65] | | Japan | 43/1811 (2.4%) | Tarama | 1 | 1.00 | 3.29 | 0.85 [-0.76 - 2.62] | | Russia | 55/2445 (2.2%) | Chistopol' | 1 | 1.00 | 3.20 | 0.35 [-1.27 - 3.28] | | Cambodia | 6/202 (3%) | Chamkar Mon | 18 | 3.51 | 3.14 | 1.56 [0.86 - 2.12] | | Egypt | 4/343 (1.2%) | Al-Laban | 1 | 1.00 | 3.02 | 2.56 [1.62 - 3.99] | | Namibia | 6/107 (5.6%) | Moses Garoeb | 8 | 2.47 | 3.00 | 0.19 [-1.34 - 3.39] | | **Coldspot** | China | 82/364 (22.5%) | Guiyang | 1 | 1.00 | -0.69 | 0.09 [-2.06 - 2.77] | | Greece | 3/14 (21.4%) | Central Macedonia | 1 | 1.00 | -0.70 | -0.79 [-1.71 - 0.37] | | China | 82/364 (22.5%) | Changji Hui | 1 | 1.00 | -0.71 | 0.09 [-2.06 - 2.77] | | China | 82/364 (22.5%) | Chongqing | 5 | 2.01 | -0.72 | 0.09 [-2.06 - 2.77] | | United States | 221/3148 (7%) | San Bernardino | 1 | 1.00 | -0.78 | 0.39 [-2.13 - 2.93] | | China | 82/364 (22.5%) | Qiandongnan Miao and Dong | 1 | 1.00 | -0.80 | 0.09 [-2.06 - 2.77] | | United States | 221/3148 (7%) | Maricopa | 1 | 1.00 | -0.86 | 0.39 [-2.13 - 2.93] | | China | 82/364 (22.5%) | Zunyi | 1 | 1.00 | -0.88 | 0.09 [-2.06 - 2.77] | | Brazil | 51/5572 (0.9%) | São Paulo | 1 | 1.00 | -0.92 | 0.8 [-1.57 - 3.01] | | United States | 221/3148 (7%) | Harris | 1 | 1.00 | -1.06 | 0.39 [-2.13 - 2.93] | |

Finally, conditional prediction plots visualizing the mu (count) and zi (zero-inflation) components for each predictor were generated to confirm these findings (Supplementary Figures X-Y).

### Sampling Over Time

Temporal trends in global surveillance effort have been highly variable, increasing non-linearly since the 1960s (Figure 12). Sampling effort, measured by the number of hosts sampled, shows several distinct peaks, notably in 1994 (n = 60,041) and 1995 (n = 38,552), corresponding to the period immediately following the 1993 Sin Nombre virus outbreak. A second, broader period of high effort occurred between 2000 and 2013, where annual sampling often exceeded 20,000 individuals (e.g., 28,038 in 2002). The number of unique studies published per year followed a similar trajectory, peaking in 2004 (n = 61).

|  |
| --- |
| Figure 12: Figure 12 Temporal Trends in Sampling Effort. A) Annual sampling effort, number of individual hosts sampled, B) Annual sampling effort, number of published studies. |

The diversity of sampled taxa also increased over time. The number of unique host species sampled annually grew from 33 in 1964 to a maximum of 202 in 2015. This was mirrored in pathogen diversity, which also peaked in the 2010s (e.g., 28 hantavirus species and 14 arenavirus species sampled in 2016). For all metrics, sampling effort and diversity show a sharp decline after 2018, with the 490 hosts sampled in 2023 representing the lowest effort since the 1980s.

|  |
| --- |
| Figure 13: Figure 13 Annual Sampled Diversity of Hosts and Pathogens |

### Temporal Biases in Sampling

The temporal GAM (Deviance explained = 34.4%) identified a significant non-linear global trend in sampling intensity (edf = 3.90, *p* = 0.0315). The model summary revealed that the temporal trend for the Americas was significantly different from the global average (edf = 5.97, *p* < 0.001), while the trends for Africa, Asia, and Europe did not significantly deviate from the global pattern (*p* > 0.05).

Based on the model’s predictions, sampling intensity peaked at different times and magnitudes across continents (Figure 13). The Americas showed the earliest and highest predicted peak of 16,358 hosts in 1998, which coincided with high observed values (e.g., 52,079 hosts sampled in 1994). Europe, Africa, and Asia showed later, closely-timed peaks in 2006 (predicted: 4,674 hosts), 2008 (predicted: 3,430 hosts), and 2008 (predicted: 4,911 hosts), respectively. For all continents, observed sampling effort in the post-2018 period consistently fell below the modeled mean, with observed samples in 2022 (e.g., 133 for the Americas) being substantially lower than the predicted trend (e.g., 612 for the Americas).

|  |
| --- |
| Figure 14: Figure 14 Regional Deviations from Global Sampling Trend (pre-2020) |

## Host-Pathogen Network Structures

To visualise and quantify the structure of observed host-pathogen interactions, we constructed bipartite networks for Arenaviridae and Hantaviridae, stratifying our analysis by the strength of diagnostic evidence.

### Acute Infection Networks

We first analyzed high-confidence “Acute Evidence” networks, built using only direct detection methods (PCR, culture, or sequencing).

These acute networks for both families were highly fragmented and specialized. The Arenaviridae network consisted of 14 distinct components, involving 48 host species and 25 virus species. This network was characterized by high modularity (Q = 0.82) and low nestedness (NODF = 3.82), indicating a system of highly specific, co-evolved host-virus modules with minimal overlap.

The Hantaviridae acute network showed a similar, though slightly more generalized, structure. It was also highly fragmented (components = 14, Q = 0.71) but was more nested (NODF = 7.38) than its arenavirus counterpart.

|  |
| --- |
| Figure 15: Figure 15 Arenavidae Host Pathogen Network (Acute Evidence) |

|  |
| --- |
| Figure 16: Figure 16 Hantaviridae Host Pathogen Network (Acute Evidence) |

### The Impact of Serological Data

We next incorporated serological data to build “All Evidence” networks, which dramatically reshaped the perceived structure of host-pathogen interactions, primarily by nearly doubling the number of observed host species for both Arenaviridae (from 48 to 85 species) and Hantaviridae (from 97 to 174 species).

However, the effect of these new serological links on network topology differed starkly between the two virus families. For Arenaviridae, the addition of serological data collapsed the fragmented structure, reducing the number of distinct components by 50% (from 14 to 7). This was accompanied by a decrease in modularity (Q = 0.82 0.71) and a sharp increase in nestedness (NODF = 3.82 10.79), suggesting that serological data (potentially from cross-reactive assays) bridges previously isolated modules and creates a more generalized, nested structure.

In contrast, the Hantaviridae network remained highly fragmented. The number of components barely changed (from 14 to 13), indicating that serological links, while adding many new hosts, did not bridge the major, distinct modules (e.g., New World vs. Old World hantaviruses). This suggests that the primary hantavirus modules are either too antigenically distant for serological cross-reactivity or are sampled in such geographically and taxonomically distinct host communities that no co-occurrence is recorded.

|  |
| --- |
| Figure 17: Figure 17 Arenaviridae Host Pathogen Network (All Evidence) |

|  |
| --- |
| Figure 18: Figure 18 Hantaviridae Host Pathogen Network (All Evidence) |

### Keystone Hosts and Sampling Effort

Node-level analysis identified several “hub” hosts with a high degree of connectivity across the networks (Table 3). For Arenaviridae, Mastomys natalensis was the most connected host with 5 viral links, followed by Mastomys natalensis (5 links). For Hantaviridae, the most prominent hubs were Mus musculus (5 links), Rattus rattus and Apodemus agrarius (both with 5 links).

| Species | Arenaviridae, acute evidence | Hantaviridae, acute evidence | Arenaviridae, all evidence | Hantaviridae, all evidence |
| --- | --- | --- | --- | --- |
| Mus musculus | 2 | 5 | 5 | 7 |
| Sigmodon hispidus | 0 | 4 | 2 | 7 |
| Rattus rattus | 2 | 5 | 3 | 6 |
| Apodemus agrarius | 1 | 5 | 2 | 6 |
| Mastomys natalensis | 5 | 0 | 5 | 0 |
| Apodemus sylvaticus | 1 | 4 | 1 | 5 |
| Rattus norvegicus | 1 | 2 | 3 | 5 |
| Apodemus flavicollis | 0 | 4 | 1 | 5 |
| Oligoryzomys nigripes | 0 | 4 | 0 | 5 |
| Sorex araneus | 0 | 3 | 0 | 5 |
| Peromyscus leucopus | 0 | 2 | 0 | 5 |
| Reithrodontomys megalotis | 0 | 2 | 0 | 5 |
| Peromyscus maniculatus | 0 | 1 | 1 | 5 |
| Mus minutoides | 3 | 0 | 4 | 0 |
| Microtus arvalis | 1 | 2 | 1 | 4 |
| Peromyscus boylii | 1 | 2 | 2 | 4 |
| Rattus tanezumi | 1 | 3 | 1 | 4 |
| Akodon montensis | 0 | 4 | 0 | 4 |
| Microtus fortis | 0 | 4 | 0 | 4 |
| Microtus agrestis | 0 | 3 | 1 | 4 |

However, a host’s apparent importance in the network was strongly correlated with sampling effort (Figure 12). We found a significant positive association between a host’s network degree (number of viral links) and the total number of individuals sampled for that species across all studies (Spearman’s ρ = 0.62, *p* < 0.001 for the Hantaviridae “All Evidence” network; ρ = 0.41, *p* < 0.001 for the Arenaviridae “All Evidence” network). This finding underscores the potential for sampling bias to inflate the perceived importance of well-studied hosts and motivates the use of formal models to disentangle this effect in the subsequent analysis.

|  |
| --- |
| Figure 19: Figure 19 Degree vs. Effort |

### Ecological Partitioning of Virus Families

To formally test the degree of separation between the two viral families, we constructed a single comprehensive network combining all “All Evidence” interactions for both Arenaviridae and Hantaviridae and performed a modularity analysis.

The analysis revealed a partitioned system (Modularity Q = 0.56), identifying 77 distinct modules within the combined network. Critically, these modules were almost perfectly segregated by virus family. Of the 77 modules, 50 (64.9%) consisted exclusively of hantaviruses and their associated hosts, and 27 (35.1%) consisted exclusively of arenaviruses. We found zero mixed modules containing viruses from both families.

### Synanthropy Associated with Increased Probability of Host-Virus Links After Controlling for Sampling Effort

To formally test the association between host traits and the probability of a host-virus link while controlling for sampling effort, we used Dyadic Bayesian GLMMs.

For Arenaviridae, after accounting for sampling effort and host range size, total synanthropy was strongly associated with an increased probability of a host-virus link. Totally Synanthropic hosts had 6.73 times the odds (95% CrI = 2.15 - 20.2) of being a reported host compared to Non-Synanthropic species. Occasionally Synanthropic hosts showed a non-significant trend towards higher odds (OR = 1.38, 95% CrI = 0.82 - 2.32).

For Hantaviridae, the pattern differed. Occasionally Synanthropic hosts had significantly higher odds (OR = 1.67, 95% CrI = 1.10 - 2.57) of being linked to a hantavirus compared to Non-Synanthropic species, controlling for effort and range size. Totally Synanthropic hosts also showed elevated odds, but the association was weaker (OR = 1.85, 95% CrI = 0.73 - 4.52). The full posterior distributions for these Odds Ratios are visualized in Figure 16.

Across all models for both virus families, host sampling effort showed a strong and consistent positive association with link probability (e.g., Arenaviridae Full Model: Est. = 1.48, 95% CrI = 1.21 - 1.78; Hantaviridae Full Model: Est. = 1.41, 95% CrI = 1.25 = 1.60), confirming its critical role in explaining observed interactions. In contrast, the association with host geographic range size was less consistent; a significant negative association was found for Arenaviridae in the synanthropy subset (Est. = -0.38, 95% CrI = -0.67 - -0.12), but credible intervals overlapped zero in the other models.

|  |
| --- |
| Figure 20: Figure 20 Odds Ratios of Host-Virus Links by Synanthropy Status |

# Discussion

Our synthesis reveals a global surveillance landscape for rodent-borne viruses that is highly fragmented and biased across taxonomic, geographic, and temporal dimensions. The data demonstrates that our collective knowledge of these important pathogens has been shaped as much by human factors—such as research funding, convenience, and a focus on known threats—as by the underlying host ecology. This “street-light effect,” where we search most intensely in the most familiar places, has profound implications for our ability to anticipate and mitigate future zoonotic threats.

A key finding of our study is the successful disentangling of biological drivers from anthropogenic research biases. While predictable ecological traits like host geographic range size and body mass are significant predictors of sampling intensity, our models reveal that these factors do not fully account for the observed geographic disparities. The initial, strong sampling signals in the Afrotropic and Neotropic realms were largely explained away once host traits were included, suggesting that surveillance in these regions, while less frequent overall, may be more closely tied to specific ecological hypotheses. In stark contrast, the Palearctic and Nearctic realms retained a strong, significant association with higher sampling effort, even after controlling for host biology. This provides compelling quantitative evidence for a persistent “researcher bias,” where surveillance is disproportionately concentrated in the Global North, likely reflecting the geographic location of major research institutions and funding bodies.

This interpretation is powerfully reinforced by our sub-analysis on synanthropy. A species’ propensity to live alongside humans was the most dominant predictor of sampling effort, attenuating most other effects, including the strong Palearctic bias. This suggests that a substantial portion of global surveillance is focused on a handful of commensal or “peri-domestic” species (e.g., *Mus musculus*, *Rattus* spp.). While this is a logical strategy for detecting immediate spillover threats from known reservoirs, it creates a profound knowledge gap regarding the vast majority of viral diversity circulating in wild, non-synanthropic hosts, leaving us vulnerable to novel pathogens emerging from less-studied wildlife reservoirs.

Further nuance emerges from our subnational gap analysis using Zero-Inflated Negative Binomial (ZINB) models. By comparing observed sampling against model predictions, we identified specific geographic hotspots (oversampling) and coldspots (undersampling).

* Critically, the spatial patterns differed between sampling intensity (number of hosts) and breadth (number of species). Intensity gaps were highly heterogeneous, with hotspots and coldspots often adjacent, suggesting localized drivers. Breadth gaps were more spatially homogeneous, implying regional factors might shape species diversity sampling.
* The analysis also highlighted “high-expectation” coldspots, particularly in well-resourced countries like the USA and China, revealing significant within-country heterogeneity where overall effort doesn’t translate to uniform coverage.
* Conversely, some hotspots (e.g., in Egypt, Indonesia) represented truly exceptional efforts exceeding expectations on both intensity and breadth, while others represented “surprise” sampling in areas predicted to have very low effort.

Temporal biases also compound these patterns. While global sampling effort increased non-linearly over time, our GAM analysis revealed significant regional deviations.

* Notably, the Americas showed a unique trajectory, peaking sharply in the late 1990s, likely driven by research following the Sin Nombre virus outbreak, while Africa, Asia, and Europe followed a more congruent, later-peaking global trend.
* Furthermore, a marked decline in observed sampling effort across all continents post-2018 falls below the pre-COVID predicted trend, suggesting a combination of pandemic-related disruptions and data reporting lags.

Finally, our analysis highlights a critical bottleneck in the global surveillance pipeline. We found a stark incongruity between the immense effort of field sampling and the public availability of genetic data. In some of the most heavily sampled countries, such as China and the USA, the proportion of sampled animals that result in a public host or pathogen sequence is less than 1%. This gap is also present for key taxa; for example, despite thousands of positive detections, many common viruses like *Mopeia mammarenavirus* are genetically under-represented, while others like *Prospect Hill orthohantavirus* have limited sampling associated public sequences despite being frequently detected. This lack of genetic data severely hinders our ability to track viral evolution, conduct phylogeographic analyses, and develop broadly effective diagnostics.

### An Evidence-Based Roadmap for Future Surveillance

Our findings provide a data-driven framework for optimizing future surveillance. The priority should not simply be to sample more, but to sample more strategically to fill the most critical gaps. First, our models identify species that are predicted to be heavily sampled based on their traits (e.g., large range size) but are currently under-represented in the data; these taxa are prime candidates for targeted surveillance. Second, the subnational gap analysis pinpoints specific ADM2 units (“double coldspots”) that fall short on both intensity and breadth metrics, representing high-priority geographic areas. Third, while the Palearctic and Nearctic are well-sampled overall, our synanthropy analysis suggests a deep bias towards commensal species within these realms. Future efforts should therefore focus on the wild, non-commensal rodent and shrew fauna in these regions to better characterize the full scope of viral diversity. Finally, we call for a shift in policy and funding to address the genetic data sink, incentivizing and mandating the public deposition of sequence data from both hosts and pathogens as a standard outcome of surveillance projects.

### Ecological Insights from Network Analysis

Our descriptive network analysis revealed distinct structural properties for Arenaviridae and Hantaviridae interactions.

* Networks based on high-confidence “Acute Evidence” were highly fragmented and modular for both families, suggesting specialized host-virus relationships.
* The inclusion of serological data (“All Evidence”) dramatically altered perceived network structure, particularly for Arenaviridae, where it reduced fragmentation and increased nestedness, possibly due to cross-reactivity bridging modules. In contrast, the Hantaviridae network remained fragmented, implying stronger separation between its core viral groups.
* While node-level analysis identified apparent “keystone” hosts, their centrality (degree) was strongly correlated with sampling effort, highlighting the need for formal modeling to control for this bias.
* Modularity analysis of a combined network confirmed near-perfect ecological partitioning, with essentially zero overlap in host-virus modules between the two families.

To formally test hypotheses about ecological drivers while accounting for sampling bias, we used Dyadic GLMMs. These models provided several key insights:

* They statistically confirmed the overwhelming positive association between host sampling effort and the probability of observing a host-virus link for both families.
* After controlling for effort, synanthropy emerged as a significant correlate. Totally Synanthropic hosts showed substantially higher odds of being linked to Arenaviridae (OR ≈ 6.7) compared to non-synanthropic hosts. For Hantaviridae, the association was weaker overall, but Occasionally Synanthropic hosts showed significantly increased odds (OR ≈ 1.7).
* The association with host geographic range size was less consistent, showing a significant negative relationship for Arenaviridae in the synanthropy subset but not definitively across all models.

Limitations:

* Acknowledge limitations of the systematic review process (e.g., publication bias, language limitations, inability to capture all “grey literature,” potential inconsistencies in reported sampling effort or diagnostics across studies).
* Limitations of trait data (e.g., incomplete synanthropy data requiring subset analysis, potential inaccuracies in IUCN range maps vs. fine-scale occurrence).
* Limitations of statistical models (e.g., assumptions of ZINB/GAM/GLMMs, potential unmeasured confounding variables, ecological fallacy when interpreting aggregated subnational predictors).
* Challenges in modeling sparse networks (as noted with ERGM exploration, highlighting why Dyadic GLMMs were chosen).

Conclusion: This work provides both a valuable, novel data resource and a quantitative framework for understanding the landscape of zoonotic virus surveillance. By identifying specific taxonomic, geographic, and temporal gaps, quantifying the influence of sampling biases versus ecological traits, and highlighting the critical disconnect between field effort and genetic data availability, we offer an evidence base to guide more effective and equitable future surveillance strategies.

# Results

## Description of the Dataset

* Summary of the total number of unique studies, sampling sites, and years represented.
* Overview of the number of host species sampled and their distribution across families and genera.
* Summary of the number and diversity of viruses tested, detected, and sequenced.

## The Surveillance Landscape: Quantified Biases

* Taxonomic Gaps: Presentation of results showing over- and under-sampling of specific host and pathogen taxa. Identification of key host genera and virus species that are under-represented relative to their known diversity or host associations.
* Geographic Gaps: Presentation of maps and statistical results highlighting major geographic regions where known hosts exist but sampling is absent. Results of the correlation between sampling effort and socio-economic factors.
* Temporal Trends: Presentation of plots showing trends in sampling effort and detected diversity over time, highlighting periods of intensified research.

## Host-Pathogen Network Structures

* Visualization and description of the Arenaviridae and Hantaviridae networks.
* Results of the quantitative comparison of the two network structures.
* Identification of keystone host and pathogen taxa.

# Discussion

Our synthesis reveals a global surveillance landscape for rodent-borne viruses that is highly fragmented and biased across taxonomic, geographic, and temporal dimensions. Surveillance is concentrated in specific taxa, regions, and time periods, leaving vast “known unknowns.”

An Evidence-Based Roadmap for Future Surveillance:

* Synthesize the taxonomic, geographic, and temporal gap analyses to provide a clear, actionable “roadmap.”
* Explicitly list the top host genera, geographic regions, and under-studied time periods that represent the most critical priorities for future surveillance efforts.

Ecological Insights from Network Analysis: Discuss the implications of the observed network structures. For example, do the differences between the Arena- and Hantavirus networks suggest different evolutionary strategies or transmission dynamics?

Limitations: Acknowledge limitations of the systematic review process (e.g., publication bias, language limitations, inability to capture all “grey literature”).

Conclusion: This work provides both a valuable, novel data resource and a quantitative framework for understanding the landscape of zoonotic virus surveillance. By identifying where, in which species, and when we are not looking, we provide a crucial tool for guiding more strategic and efficient global health research and pandemic preparedness.