Correcting the surveillance landscape reveals the macroecological associations of viral reservoir status in small mammals

# Abstract

Small mammals, particularly rodents and shrews, act as primary reservoirs for Arenaviruses and Hantaviruses, zoonotic pathogens causing substantial global morbidity. However, our understanding of reservoir ecology is obscured by biased surveillance efforts, where sampling preferentially targets synanthropic species and high-income regions. It remains unclear whether observed patterns of reservoir competence, such as the association with synanthropy, are biological realities or artefacts of surveillance bias. We conducted a systematic review and data synthesis of global surveillance efforts (1960–2025), creating a harmonised database of over 590,000 tested individuals contributing 716,000 assay results. We then integrated this with macroecological trait data and phylogenetics to model reservoir probability using Bayesian Phylogenetic Dyadic Generalised Linear Mixed Models (GLMMs). We identified substantial taxonomic and geographic biases; surveillance is heavily skewed towards the Palearctic and synanthropic species, while 46% of host genera remain entirely unsampled. Geographically, surveillance intensity correlates strongly with GDP and night-light intensity rather than host biodiversity. After statistically correcting for these biases, we demonstrate that reservoir status is a predictable biological trait. A fast pace of life (e.g., early maturity, large litters) significantly increases the probability of reservoir status (**result**), independent of sampling effort. Synanthropy remains a strong predictor, but its effect is largely explained by the underlying life-history traits of commensal species. [Results pending: Co-phylogenetic analysis and Conclusion].

# Introduction

Zoonotic spillover is an ecological process structured by the distribution of reservoir hosts, pathogen transmission within these populations (hazard), and opportunities for transmission to susceptible humans (risk) [ref]. Arenaviruses (Arenaviridae) and Hantaviruses (Orthohantaviridae) represent two of the most significant families of viral zoonotic pathogens carried by small mammals, responsible for a spectrum of human diseases ranging from haemorrhagic fevers (e.g., Lassa fever, Junin virus) to severe pulmonary and renal syndromes (e.g., Hantavirus Pulmonary Syndrome) [ref]. These pathogens disproportionately impact communities with limited healthcare infrastructure, where outbreaks further strain already scarce resources [ref]. Historically, the understanding of these pathogens focussed on a one-host-one-virus paradigm, implying a tight evolutionary co-divergence between virus and host [ref]. However, recent advances have revealed complex multi-host networks where spillover risk is determined not just by the presence of a specific host, but by the ecological dynamics of the entire reservoir community [ref]. Consequently, identifying which species are competent reservoirs and distinguishing them from incidental hosts, is critical for predicting areas of emergence risk.

Despite decades of research, the global map of reservoir status remains fundamentally distorted. Surveillance is rarely random; it is guided by convenience, funding, and accessibility a phenomenon often termed the streetlight effect [ref]. Researchers predominantly sample species that are easy to capture (synanthropes), proximate to research infrastructure (high-income nations), or implicated in known outbreaks, or as part of a public health response to an emerging outbreak (reactive rodent control) [ref]. This creates a surveillance landscape that reflects human activity as much, or more than, viral distribution. For example, the pervasive association between reservoir status and synanthropic rodents (e.g., *Rattus spp.*, *Mus spp.*) is well-documented, yet it remains unclear whether this reflects a genuine biological susceptibility of commensal animals or simply the fact that they are the most frequently sampled mammals [ref]. Here, we argue that disentangling the ecology of these viruses is impossible without first quantifying and statistically correcting for this anthropogenic filter.

Once surveillance bias is accounted for, macroecological and evolutionary theory offer competing hypotheses for what determines reservoir competence. Life-history theory posits a trade-off between reproduction and somatic maintenance; the Fast Life-History hypothesis suggests that r-selected species (those that mature early and produce large litters) invest less in costly adaptive immunity, potentially making them more tolerant of chronic viral infections [ref]. Alternatively, the Ecological Opportunity hypothesis suggests that generalist species with broad diets, large distributions, and wide habitat tolerances encounter a wider diversity of pathogens, increasing the likelihood of successful viral spillover and adaptation [ref]. At the community level, these individual traits scale up to landscape-level hazard: does high biodiversity amplify risk by supporting more host species, or dilute it by suppressing the relative abundance of high-competence generalists [ref]? Finally, beyond ecological traits, the evolutionary history between host and virus plays a critical role. While strict co-speciation implies restricted host ranges constrained by deep evolutionary time, phylogenetic incongruence between host and virus would suggest that host-switching, facilitated by ecological contact, is a more dominant driver of viral diversity than previously appreciated [ref].

This study aims to move beyond descriptive surveillance mapping to a mechanistic understanding of reservoir ecology. By synthesising over 50 years of global surveillance data, we first quantify the anthropogenic filter, mapping the taxonomic, geographic, and temporal biases in small mammal sampling. We then employ Bayesian phylogenetic models to test whether Pace of Life and Synanthropy predict host competence after accounting for the confounding effects of sampling effort. Concurrently, we assess the evolutionary congruence between hosts and viruses to determine the relative roles of host-switching versus co-divergence. Finally, we leverage these models to project the intrinsic reservoir potential of unsampled species, identifying potential hotspots of hazard that current surveillance has missed.

# Methods

## Systematic Review & Data Harmonisation

We conducted a systematic review following PRISMA guidelines (Supplementary Figure S1). We searched PubMed and Web of Science for primary research published between January 1, 1960, and ???, using search strings targeting Arenaviridae or Hantaviridae infection in small mammals (orders Rodentia, Eulipotyphla). Studies were included if they reported primary surveillance data (molecular or serological) with identifiable host taxonomy. Data were extracted into a relational database capturing sampling location, sampling date, sampling effort, assay methodology, and number of individuals tested/positive (Supplementary Methods 1).

Taxonomic harmonisation was performed using the taxize package (0.10) in R (4.2.3). All host names were resolved against the GBIF Backbone Taxonomy to handle synonyms and aligned with the Mammal Diversity Database (MDD v1.11) and IUCN Red List (v2024-2). Viral taxonomy was standardised to the International Committee on Taxonomy of Viruses (ICTV) 2024 release.

## Macroecological Trait Integration

Host life-history and ecological traits were aggregated from the COMBINE database. We prioritised traits hypothesised to influence viral competence including adult mass (g), litter size (n), litters per year (n), gestation length (days), weaning age (days), and sexual maturity age (days). To maximise sample size without introducing list-wise deletion bias, we employed phylogenetic imputation using the Rphylopars (0.3.10) package. This approach estimates missing trait values under a Brownian Motion model of evolution, leveraging the phylogenetic covariance between species (using the Upham *et al.* 2019 consensus mammal tree) to predict missing values.

To resolve multicolinearity among life-history traits we performed a Principal Component Analysis (PCA) on log-transformed reproductive and morphological variables. The first principal component (PC1) explained 48.7% of the variance and was extracted as a continuous metric of the Pace of Life, where low values represent “fast” strategies (early maturity, high reproductive output) and high values represent “slow” strategies (high body mass, prolonged gestation and longevity).

## Quantification of Surveillance Bias

We assessed geographic bias using two complementary approaches. First, to quantify taxonomic coverage, we intersected species’ IUCN range maps with GADM (Level 2) administrative boundaries. We aggregated these overlaps at the genus level to calculate the proportion of range sampled for each genus, comparing the total area of districts where sampling occurred against the genus’s total geographic range size. Second, to identify the sociodemographic drivers of surveillance intensity, we fitted a Zero-Inflated Negative Binomial (ZINB) model using the brms package (2.22). This model predicted the total count of tested individuals per district (aggregating all target species) as a function of night-time light intensity (VIIRS 2024), accessibility (travel time to major cities), and GDP per capita. The zero-inflation component modeled the probability of a district remaining entirely unsampled.

Temporal trends were analysed using Generalized Additive Models (GAMs) with the mgcv package (1.9-3). We modelled the annual sampling effort (1960–2025) for each continent using a negative binomial error structure with thin plate regression splines to characterise non-linear trajectories and detect reactionary surveillance pulses following major outbreaks of zoonotic viruses.

To quantify the disconnect between field surveillance and genomic data availability, we compared the number of PCR-positive detections reported in the literature against the number of unique sequences available in GenBank for each viral species. We calculated sequencing completeness ratios at the country level for both hosts and pathogens to identify regions where molecular detection is frequent but genomic data remains scarce (e.g., regions relying solely on diagnostic PCR without sequencing).

## Phylogenetic Dyadic Modelling

To identify the intrinsic drivers of reservoir status, we fitted a Bayesian Phylogenetic Dyadic Generalised Linear Mixed Model (GLMM). The unit of analysis was the host-virus dyad, defined as a potential association between a screened host species () and a viral species (). To ensure that absence of a reservoir association reflected genuine negative data rather than a lack of surveillance, the analytic dataset was restricted to host species with at least one recorded testing event (). Species present in the phylogeny but never sampled for Arenaviruses or Hantaviruses were excluded from the training set to prevent the absence of evidence from biasing trait estimates.

A binary indicator (response variable) () was set to 1 if the dyad showed evidence of Active Infection (PCR, sequencing, or viral isolation) or Specific Exposure (serology specific to the viral species level). All other dyads within the sampled subset were set to 0 (pseudo-absences).

The probability of reservoir status () was thus modelled as:

Where is the log-transformed total count of individuals of host tested for pathogens. This term acts as a control through defining an offset for detection probability. By using individuals tested rather than trap-nights, we explicitly control for downstream biases, species that are rare or difficult to trap yield fewer samples, resulting in lower detection power. The model accounts for this by widening the uncertainty intervals for species with low testing counts.

The model was applied to two analytic datasets:

* *The full dataset*: Comprising 49,270 dyadic pairs from 70 pathogens (including unclassified Arenaviridae and Hantaviridae) and 704 rodent and shrew species. In this model, we tested the evolutionary hypothesis using (PC1), a continuous variable where lower values correspond to a Fast life history strategy. This model was subsequently used for global spatial projection to maximise taxonomic coverage.
* *The synanthropy subset*: Comprising 19,180 dyadic pairs containing the same 70 pathogens but a reduced set of 274 host species for which detailed synanthropy data were available. In this subset, the ecological and evolutionary hypotheses were tested in tandem through the addition of the term (Categorical: Not Synanthropic, Occasionally Synanthropic, Totally Synanthropic). This model was used for mechanistic inference of ecological risk factors.

We incorporated three random effects to account for non-independence. is a structured random phylogenetic effect covarying according to the phylogenetic distance matrix () derived from the Upham *et al.* (2019) tree, controlling for shared evolutionary history. is an unstructured species-specific random intercept included to capture species-level variation (e.g., overdispersion) not explained by phylogeny (conceptually equivalent to Pagel’s ). Finally, viral identity is included as a random intercept controlling for uneven study effort across viral species (e.g., *Orthohantavirus sinnombreense* is studied more intensely than South American hantaviruses).

Models were fitted in Stan via the brms interface using 8 chains of 2,500 iterations (1,500 warmup). We utilised weakly informative priors ( for fixed effects; for intercepts) to regularise estimates. Convergence was assessed via and visual inspection of trace plots.

## Global Hazard Mapping

To map the global distribution of intrinsic zoonotic hazard, we projected the parameters of the Global Model (N = 49,270; trained on Life History + Sampling Effort) onto the full set of 2,766 mammal species for which life history data were available (imputed via Rphylopars as described above). We generated posterior predictions for reservoir probability () for all species, setting sampling effort to the maximum observed value to estimate intrinsic biological potential independent of surveillance history. Viral random effects were marginalised (set to zero) to predict generalised reservoir competence rather than specific viral associations.

Spatial projection was performed using the terra package (1.8-60). We rasterised IUCN Red List range maps for all 2,766 species onto a Mollweide equal-area projection at 20km resolution. For each grid cell, we calculated Community Competence (Mean Hazard) as the arithmetic mean of the predicted reservoir probabilities for all species present in that cell:

Where is the species richness in cell . To avoid artifacts from species-poor pixels, cells with <2 were masked.

To explicitly interrogate the relationship between biodiversity and hazard (Amplification vs. Dilution), we performed a bivariate classification using the biscale package (1.1). Grid cells were categorised into a 3×3 matrix based on tertiles (33rd and 66th percentiles) of species richness and community competence. We extracted mean values for all administrative districts (GADM Level 2) using the exactextractr package (0.10) to identify representative districts for each risk archetype.

## Evolutionary Analysis

Describe Co-phylogenetic methods to test for congruence between host and viral trees.

# Results

## The Anthropogenic Filter: Quantifying Surveillance Bias

Our systematic review identified approximately 570,000 tested individuals from 742 species, yet surveillance effort was highly unevenly distributed. We observed a strong dissociation between sampling intensity and host biodiversity. Despite 50 years of research, 46% of all Rodentia and Eulipotyphla genera remain entirely unsampled for *Arenaviridae* or *Hantaviridae*, with entire families such as *Gliridae* receiving negligible surveillance attention (Supplementary Figure X). In our Zero-Inflated Negative Binomial (ZINB) model of global surveillance, the strongest predictors of sampling intensity at the district level were night-light intensity (, 95% Credible Interval (CrI) [0.17, 0.85]) and accessibility (, 95% CrI [-0.77, 0.18]). Since accessibility was defined as travel time to major cities, this negative coefficient indicates that surveillance intensity significantly decreases as locations become more remote. Conversely, neither local species richness nor human population density were not important predictors of sampling effort (, 95% CrI [-0.27, 0.11] and , 95% CrI [-0.11, 0.43] respectively) ([Figure 1](#fig-geo)).

|  |
| --- |
| **Figure 1 | Global surveillance inequality.** (a) Global map of surveillance residuals from a Zero-Inflated Negative Binomial (ZINB) model. Colours indicate the log-ratio of observed versus predicted surveillance effort (log10 [Observed/Predicted]). “Hotspots” (Orange, >0) represent districts sampled more intensely than predicted by their socioeconomic characteristics; “Coldspots” (Blue, <0) represent undersampled regions. Insets highlight key regional anomalies: the US Midwest (oversampled), West Africa (targeted Lassa surveillance), and Central Asia (undersampled). (b–d) Marginal effects of the top predictors on sampling intensity. Surveillance effort increases with (b) Night-time Light Intensity and decreases with (c) Remoteness, but is unrelated to (d) Host Species Richness. |

These geographic constraints generate profound taxonomic distortions. Surveillance is heavily skewed towards the Palearctic and Nearctic realms, which collectively account for 71% of all detected small mammals, leaving a substantial number of the most biodiverse regions in the tropics largely uncharacterised. At the genus level, sampling coverage is driven primarily by geographic ubiquity rather than evolutionary distinctiveness ([Figure 2](#fig-bias)). We observed a strong positive relationship between total geographic range size and the area sampled; genera with massive distributions (e.g., *Rattus*, *Mus*, *Apodemus*) are sampled across broad spatial extents, whereas range-restricted genera are frequently ignored. However, even for widespread genera, the proportional coverage remains low; despite *Mus* being sampled in 444 administrative districts, this represents less than 3.3% of its global range.

|  |
| --- |
| **Figure 2 | The disconnect between geographic distribution and surveillance coverage.** The relationship between the total geographic range size of a small mammal genus (y-axis) and the total area of administrative districts in which it has been sampled for Arenaviridae or Hantaviridae (x-axis). Both axes are log-transformed. Points represent individual genera (). The background heatmap and contours visualise the proportional coverage (Sampled Area / Total Range), interpolated from the observed data, with lighter colours indicating higher coverage. Labels highlight genera that represent extremes in total sample size (green), proportional coverage (red), or geographic range size (blue). Despite a positive correlation between range size and sampled area, the vast majority of genera occupy the dark purple region, indicating that even widespread taxa are sampled across only a small fraction of their distribution. |

Generalised Additive Models (GAMs) confirmed that host body mass and geographic range size were the dominant predictors of species-level sampling effort (*p* < 0.001), with larger and more widespread species receiving disproportionate attention. While totally synanthropic species exhibited higher mean sampling estimates than non-synanthropic species ( = 2.23), this effect was not statistically significant (*p* = 0.20) likely due to the rarity of this trait in the dataset. Instead, geographic range size remained the single strongest driver of surveillance effort ( = 422.7).

Temporally, surveillance has followed a reactive rather than proactive trajectory. Global sampling effort exhibits distinct pulses following major zoonotic discovery events, most notably the description of Lassa virus (1969) and the Sin Nombre virus outbreak (1993). GAMs reveal that while surveillance in the Americas grew exponentially since the mid-1990s (*p* < 0.001) before declining post 2005, effort in Africa and Asia has remained comparatively static and highly episodic, driven by short-term outbreak responses rather than sustained monitoring.

## The Genetic Data Gap

Beyond the biases in host sampling, we identified a substantial disconnect between viral detection and genomic characterisation. Of the [N] individual hosts that tested positive via molecular assays (PCR), valid sequence data was available for only [XX]% ([N] individuals). Sequencing completeness varied significantly by viral family: Hantaviridae detections had lower genomic coverage ([XX]%) compared to Arenaviridae ([XX]%). Spatially, the proportion of unsequenced positive results was highest in [Region] (**?@fig-gap**). As a result, [XX]% of the host-virus associations in the global dataset are currently supported only by serological or fragment-based evidence without lineage-specific sequence confirmation.

## Macroecological Associations with Reservoir Status

To disentangle intrinsic biological suitability from the surveillance biases identified above, we fitted Bayesian phylogenetic generalized linear mixed models (GLMMs) to the dyadic host-virus data (N = 49,280 pairs). As expected, sampling effort was the dominant predictor of reservoir discovery; the number of individuals tested had a strong positive effect on the probability of detection ( = 0.59, 95% CrI [0.53, 0.65]), confirming that widely detected reservoirs are often the most intensely scrutinised species ([Figure 3](#fig-macro) a). After controlling for sampling effort and phylogeny, we found support for the fast life history hypothesis. The first principal component of life history (PC1), representing the slow-fast continuum, showed a negative association with reservoir status ( = −0.13, 95% CrI [-0.30, 0.04]). Since lower PC1 scores correspond to species with smaller body mass, earlier maturity, and larger litter sizes, this indicates that fast-lived species are more likely to be competent reservoirs. While the 95% CrI overlaps zero, the probability of direction is high (>93% of posterior draws were negative), suggesting a consistent biological signal.

|  |
| --- |
| **Figure 3 | Associations with reservoir status.** (a) Posterior log-odds distributions for the predictors of reservoir status in a Bayesian phylogenetic GLMM ( = 19,180). Percentages indicate the Probability of Direction (pd), representing the certainty that the effect is non-zero in the indicated direction. (b–d) Marginal effects of the predictors on the probability of being a reservoir host. Shaded areas represent 95% Credible Intervals. (b) Sampling effort is the strongest driver of discovery; the curve visualizes probability across the observed range of sampling effort. (c–d) Intrinsic biological drivers visualized conditional on maximum observed sampling effort to isolate biological potential from surveillance limitations. (c) Fast life history (low PC1 scores) is associated with a higher probability of reservoir competence. (d) Totally synanthropic species (obligate commensals) exhibit higher reservoir probability than non-synanthropic species, highlighting the role of ecological opportunity. |

In a subset analysis accounting for human affinity (N = 19,180), synanthropy emerged as a distinct risk factor independent of life history ([Figure 3](#fig-macro) a). Totally synanthropic species (obligate commensals) had a markedly higher probability of being reservoirs compared to non-synanthropic species ( = 0.62, 95% CrI [-0.09, 1.31]) ([Figure 3](#fig-macro) d), representing an approximate 1.8-fold increase in odds. Occasionally synanthropic species showed a much weaker effect ( = 0.17, 95% CrI [-0.15, 0.48]), suggesting that the risk is concentrated in species with high-intensity anthropogenic contact. The effect of life history remained stable in this model ( = −0.12), confirming that the reservoir competence of fast species is not merely a byproduct of them being commensal pests ([Figure 3](#fig-macro) c).

Variance partitioning revealed that reservoir status is a highly labile trait. The random effect variance attributed to specific host identity ( = 0.43) was more than four times larger than the variance attributed to phylogenetic history ( = 0.10). This indicates that while reservoir competence clusters within certain lineages (e.g., specific genera), it does not evolve slowly and predictably across the mammalian tree but rather emerges in specific species. Finally, we observed substantial variation in detection probability among viral species ( = 1.46), reflecting major differences in viral prevalence and the sensitivity of diagnostic assays used for different pathogens. Post-hoc interrogation of the viral random effects revealed a systematic detection bias: Arenaviridae species exhibited significantly higher random intercepts than Hantaviridae (Supplementary Figure 6), indicating that, all else being equal, arenaviruses had a higher baseline probability of detection in reservoir hosts.

Finally, we projected these trait-based estimates spatially to map the global landscape of zoonotic hazard ([Figure 4](#fig-hazard)). To maximise taxonomic coverage, we utilised the global model (N = 49,280) to predict reservoir probability for all small mammal species for which life history data were available (N = 2,766). Regional validation confirmed that intrinsic hazard is not randomly distributed. North American assemblages exhibited significantly lower mean reservoir potential compared to tropical regions (Pairwise t-test, p < 0.01; Supplementary Figure X). Conversely, no significant difference was observed between the Atlantic Forest, West Africa, and East Asia (p > 0.05).

|  |
| --- |
| **Figure 4 | The global landscape of intrinsic zoonotic potential.** Bivariate map illustrating the intersection of Species Richness (x-axis, blue) and Community Competence (mean intrinsic hazard, y-axis, pink). Community competence represents the average predicted probability of reservoir status for local small mammal assemblages (20km resolution), derived from the global Bayesian GLMM based on intrinsic pace-of-life traits (N = 2,766 species). Purple regions (High Richness + High Hazard) indicate true hotspots where high biodiversity converges with high intrinsic reservoir potential (e.g., West Africa, Atlantic Forest). Pink regions (Low Richness + High Hazard) indicate species-poor assemblages dominated by high-risk, fast-lived generalists (e.g., arid zones, agricultural frontiers). Blue regions (High Richness + Low Hazard) indicate biodiverse assemblages dominated by slow-lived specialists, potentially buffering spillover hazard via the dilution effect (e.g., Amazonia). Note that while synanthropy was identified as a risk factor in subset analyses (Figure 4), it is not included in this global projection due to data limitations for non-target species. |

Bivariate classification revealed that the highest intrinsic hazard was not found in the most biodiverse regions, but in species-poor assemblages (Class 1-3; Mean Hazard = 0.129, Mean Richness = 11.5 spp), supporting the hypothesis that ecological filtering favors hyper-competent generalists (e.g., pest monocultures). Conversely, the most species-rich assemblages (Class 3-1; Mean Richness = 29.3 spp) exhibited significantly lower community competence (Mean Hazard = 0.115), consistent with the dilution effect. However, we also identified a critical subset of hotspots (Class 3-3), where high biodiversity (Mean Richness = 26.3 spp) coincided with high intrinsic hazard (Mean Hazard = 0.127), suggesting regions where reservoir control may be complicated by complex community dynamics.

While the absolute differences in mean community competence were modest, the ecological signal was highly consistent. Species-poor assemblages (Class 1-3) exhibited a 12% higher relative intrinsic hazard compared to the most biodiverse assemblages (Class 3-1; 0.129 vs. 0.115). Given the low variance within these classes (SD < 0.005), this indicates a robust macroecological pattern: ecological filtering reliably shifts community composition toward higher-competence phenotypes.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Representative districts for the three bivariate risk classes.   | Country | Province | District | Mean Hazard (Prob) | Mean Richness (Spp) | | --- | --- | --- | --- | --- | | **1-3: High Hazard / Low Richness** | | | | | | Denmark | Syddanmark | Fanø | 0.154 | 2.0 | | Estonia | Lääne-Viru | Kunda | 0.154 | 2.0 | | Netherlands | Noord-Holland | Waterland | 0.154 | 2.0 | | Norway | Finnmark | Hammerfest | 0.146 | 2.0 | | United States | Michigan | Keweenaw | 0.146 | 2.0 | | **3-1: Low Hazard / High Richness** | | | | | | Brazil | Pará | Juruti | 0.101 | 23.3 | | Thailand | Prachuap Khiri Khan | Muang Prachuap Khiri Khan | 0.102 | 25.3 | | Ecuador | Morona Santiago | Taisha | 0.103 | 27.1 | | Indonesia | Bengkulu | Bengkulu | 0.103 | 41.0 | | Malaysia | Sarawak | Miri | 0.104 | 47.0 | | **3-3: High Hazard / High Richness** | | | | | | Bolivia | Cochabamba | Germán Jordán | 0.135 | 27.7 | | México | Oaxaca | Santo Tomás Ocotepec | 0.135 | 22.0 | | Peru | Puno | San Antonio de Putina | 0.134 | 23.6 | | Kenya | Turkana | Turkana Central | 0.132 | 22.1 | | South Africa | KwaZulu-Natal | iLembe | 0.131 | 23.5 | |

To translate these macroecological patterns into actionable surveillance targets, we extracted representative administrative districts for each risk archetype (Table 1). The analysis highlights a stark global divide. Class 1-3, characterized by species-poor communities dominated by hyper-competent generalists, is typified by districts in the Global North, including Denmark (Fanø), The Netherlands (Waterland), and the USA (Michigan). In contrast, Class 3-1, representing the safest assemblages globally, is strictly pan-tropical. The districts with the lowest intrinsic hazard were consistently found in high-biodiversity rainforests across three continents, including Pará, Brazil (Amazon), Sarawak, Malaysia (Borneo), and Bengkulu, Indonesia. Finally, Class 3-3, was typified by high biodiversity coincides with high intrinsic hazard. This class includes Cochabamba, Bolivia (endemic for Machupo virus), validating the model’s ability to pinpoint complex, high-risk interfaces.

## Evolutionary Congruence

* **[Action]:** Report results of the co-phylogeny.
* *Expected:* “While broad congruence exists at the Family level, frequent host-switching is evident at the genus/species level…” (or whatever the data shows).

# Discussion

TBC

# Discussion notes

1. Surveillance Bias

The dissociation between sampling effort and biodiversity confirms a pervasive streetlight effect in global surveillance, where research intensity is dictated by logistical convenience and socioeconomic factors (GDP, light, accessibility) rather than ecological relevance. The strong prediction of sampling by night-lights and accessibility suggests that our current understanding of reservoir diversity is heavily biased towards human-modified, accessible landscapes, potentially blinding us to sylvatic cycles in remote, biodiverse regions. The strong phylogenetic signal in sampling effort (Pagel’s λ) indicates that researchers historically cluster their efforts within specific “familiar” lineages (e.g., Muridae, Cricetidae), leading to a self-reinforcing cycle where well-studied families get more attention while others (e.g., Gliridae, Tenrecidae) remain perpetually neglected.

Explicitly link the “Urban/Wealth Bias” to the specific finding that Night-lights and Accessibility were the top predictors in the ZINB model. Connect this to Colonial Science. The skew towards the Palearctic/Nearctic (71% of data) isn’t just a “bias”; it’s a structural failure that leaves the most biodiverse regions (and thus the “Dilution Zones”) unmonitored.

The pulse dynamics of sampling effort—particularly the surges following the description of Lassa virus (1969) and Sin Nombre (1993)—indicate that global surveillance is reactive rather than proactive. Funding and effort are mobilised in response to public health crises rather than sustained as a baseline monitoring strategy.

By mapping the ‘Hazard’ (reservoir distribution) rather than ‘Risk’ (spillover), we isolate the ecological potential for emergence from the anthropogenic factors (human density, behavior) that precipitate outbreaks.

1. The Genetic Data Gap

The significantly lower sequencing completeness for Hantaviridae likely reflects the widespread historical use of generic pan-hantavirus PCR primers. These assays effectively detect viral presence but often fail to generate amplifiable products for lineage-specific sequencing. This genetic gap severely limits our ability to resolve viral taxonomy or infer transmission networks. Without sequence data, we cannot distinguish between spillover of a known pathogen and the discovery of a novel, closely related variant. Consequently, a substantial proportion of known reservoir associations currently rely on serological or fragment-based evidence, which cannot definitively link the host to a specific viral species or exclude cross-reactivity. Because sequences are missing, many reservoirs in the literature might actually be spillover hosts (incidental infections). Without phylogeny, we can’t prove the virus is evolving in that host.

1. Mechanisms of Competence (GLMMs)

The strong link between synanthropy and reservoir status—even after controlling for life history—suggests that ecological opportunity plays a critical role. Commensal species may not necessarily be more competent physiologically, but their high population densities and frequent contact with humans/vectors in urban landscapes amplify transmission cycles. The robustness of the fast life history signal supports the pace-of-life hypothesis, suggesting that r-selected species invest fewer resources in costly immune defences (viral clearance), thereby facilitating tolerance and persistent infection. “Ecological filtering reliably shifts community composition toward higher-competence phenotypes.” The finding that species-specific variation far exceeds structured phylogenetic variation implies that reservoir competence is a labile trait. It does not evolve slowly down the tree but emerges erratically, likely driven by specific receptor compatibility or local ecological factors rather than deep evolutionary constraints.

While synanthropy is a risk factor (Figure 4), the Global Map (Figure 5) shows that even without explicitly modelling synanthropy, species-poor areas (often anthropogenic) are still high hazard. This implies that Fast Life History and Synanthropy functionally overlap—human-disturbed environments select for fast animals.

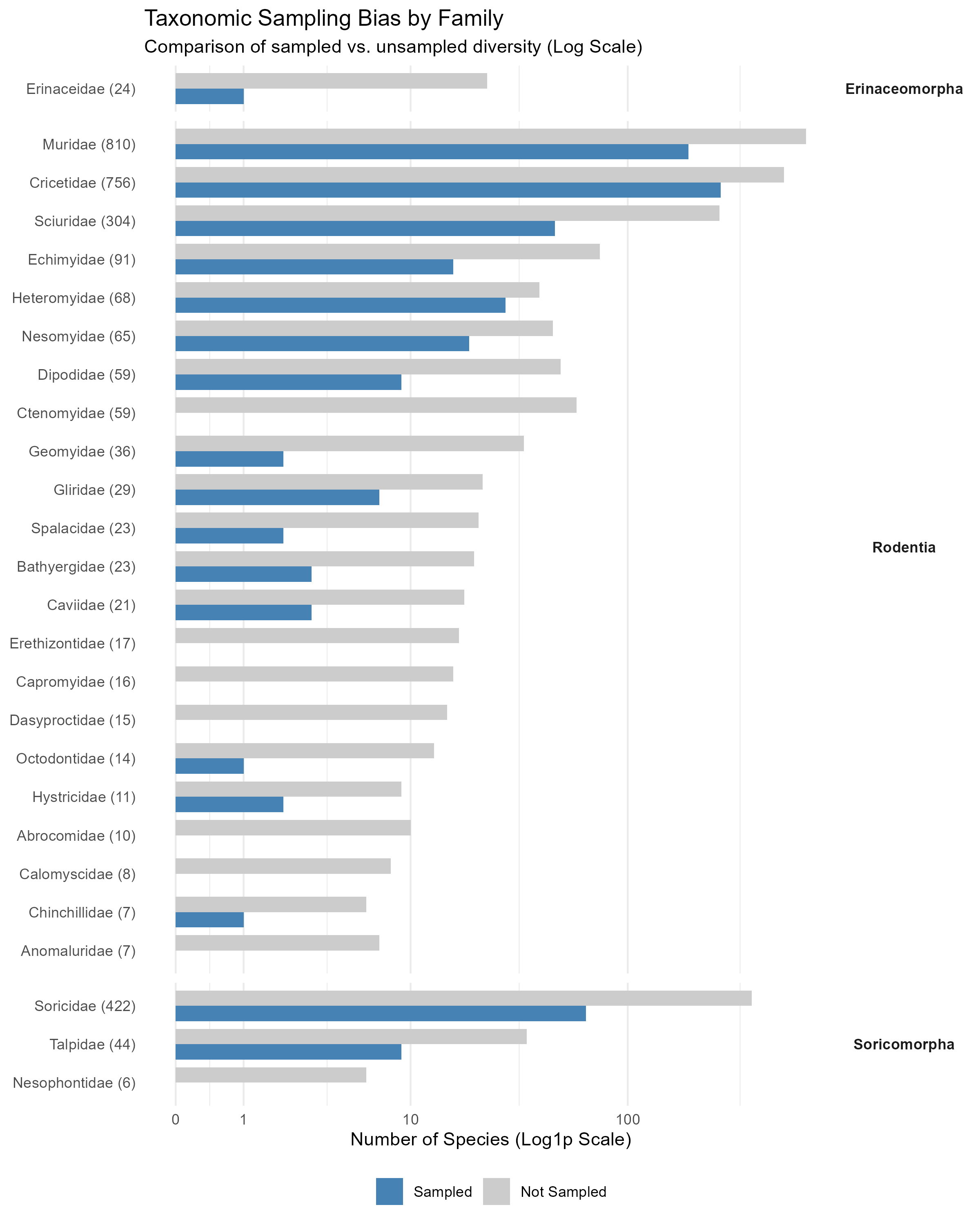
1. The Dilution Effect

High biodiversity is not synonymous with high hazard. In fact, the most species-rich districts (Class 3-1) had the lowest mean hazard (0.115). This supports the dilution effect. In complex rainforests (Amazon, Borneo), the community is dominated by specialised, slow-lived rodents (e.g., Sciurids, Echimyids) that are poor reservoirs. These safe species dilute the density of the risky generalists. “Contrary to the assumption that the ‘tropics are a ticking time bomb,’ our model reveals that intact, high-biodiversity ecosystems are intrinsically safer than the species-poor ‘pest monocultures’ of the Global North.”

We find Denmark, Netherlands, and USA (Michigan) as “High Hazard / Low Richness” (Class 1-3). Anthropogenic disturbance may as an evolutionary filter. By clearing forests and building cities/farms, humans filter out the safe specialists, leaving behind only the fast, hyper-competent generalists (mice, voles, rats). The intrinsic probability of a rodent being a reservoir is higher in a Dutch field or an American suburb than in the Amazon.

Focus on Class 3-3 (Purple). These are the Transition Zones (e.g., Turkana, Cochabamba, Savannah-Forest mosaics). These areas have the “worst of both worlds”: enough diversity to support many viral lineages, but enough disturbance/ecological opportunity to favour fast reservoirs (Mastomys, Calomys). Note that while Cochabamba is known, many other purple zones (e.g., parts of West Africa or India) are “Coldspots” in Figure 1 map. This is where the Proactive Surveillance should be targeted.

# Supplementary Materials



**Supplementary Figure 1 | Taxonomic gaps in surveillance effort.** The number of species sampled (blue) versus not sampled (grey) for Arenaviridae or Hantaviridae across all families within the orders Rodentia, Eulipotyphla, Soricomorpha, and Erinaceomorpha (N=2,766 species). Families are faceted by order and ordered by total species richness. While large families like Muridae and Cricetidae contain the highest number of sampled species, they also harbor the largest absolute number of neglected species. Entire families (e.g., Gliridae, Tenrecidae) remain largely or completely uncharacterized despite their potential relevance to zoonotic cycles.



**Supplementary Figure 2 | The global co-surveillance landscape.** A heatmap illustrating the intensity of surveillance for specific host-pathogen associations. The x-axis represents the 62 most sampled viral species (faceted by Family: Arenaviridae and Hantaviridae), and the y-axis represents the sampled host species. Cell colour indicates the log-transformed number of individuals tested for that specific pair. The matrix reveals a sparse, structure, indicating that surveillance is highly targeted: most host species are screened for only a narrow range of expected viruses (e.g., *Mastomys natalensis* for Lassa virus; *Myodes glareolus* for Puumala virus), leaving the majority of potential host-virus combinations untested.