Viral competence data improves rodent reservoir predictions for American orthohantaviruses

Short title: Competence data improves virus reservoir predictions

Nathaniel Mull^{1*}, Colin J. Carlson², Kristian M. Forbes¹, Daniel J. Becker³

¹Department of Biological Sciences, University of Arkansas, Fayetteville, AR, USA

²Center for Global Health Science and Security, Georgetown University Medical Center, Washington, D.C., USA

³Department of Biology, University of Oklahoma, Norman, OK, USA

*Corresponding author

Abstract

Identifying reservoir host species is crucial for understanding the risk of pathogen spillover from wildlife to people. Orthohantaviruses are zoonotic pathogens primarily carried by rodents that cause the diseases hemorrhagic fever with renal syndrome (HFRS) and hantavirus cardiopulmonary syndrome (HCPS) in humans. Given their diversity and abundance, many orthohantaviruses are expected to be undiscovered, and several host relationships remain unclear, particularly in the Americas. Despite the increasing use of predictive models for understanding zoonotic reservoirs, explicit comparisons between different evidence types for demonstrating host associations, and relevance to model performance in applied settings, have not been previously made. Using multiple machine learning methods, we identified phylogenetic patterns in and predicted unidentified reservoir hosts of New World orthohantaviruses based on evidence of infection (RT-PCR data) and competence (live virus isolation data). Infection data were driven by phylogeny, unlike competence data, and boosted regression tree (BRT) models using competence data displayed higher accuracy and a narrower list of predicted reservoirs than those using infection data. Eight species were identified by both BRT models as likely orthohantavirus hosts, with a total of 98 species identified by our infection models and 14 species identified by our competence models. Hosts predicted by competence models are concentrated in the northeastern United States (particularly *Myodes gapperi* and *Reithrodontomys megalotis*) and northern South America (several members of tribe Oryzomyini) and should be key targets for empirical monitoring. More broadly, these results demonstrate the value of infection competence data for predictive models of zoonotic pathogen hosts, which can be applied across a range of settings and host-pathogen systems.

Author Summary

Human diseases with wildlife origins constitute a significant risk for human health. Orthohantaviruses are viruses found primarily in rodents that cause disease with high rates of mortality and other complications in humans. An important step in disease prevention is to identify which rodent species carry and transmit orthohantaviruses. By incorporating species relatedness and evidence of different levels of host capacity to be infected and transmit virus, we used predictive modeling to determine unidentified rodent hosts of orthohantaviruses. Models using host competence data outperformed models using host infection data, highlighting the importance of stronger data in model optimization. Our results highlighted roughly a dozen key target species to be monitored that are concentrated in two geographic regions—northeastern United States and northern South America. More broadly, the approaches used in this study can be applied to a variety of other host-pathogen systems that threaten public health.

Introduction

Identifying reservoir host species (those that maintain and transmit a particular pathogen; Haydon et al. 2002) is crucial for understanding the risk of pathogen spillover from wildlife to people (zoonotic transmission; Viana et al. 2014; Plowright et al. 2017). By elucidating possible sources of zoonotic exposure, targeted strategies can be implemented to prevent or at least mitigate spillover risk. Large-scale surveillance of wildlife, often involving non-targeted sampling of a large diversity and abundance of animals, is commonly conducted shortly after disease outbreaks to search for the pathogen reservoirs (e.g., Leroy et al. 2005; Poon et al. 2005). Such studies are often expensive, time-consuming, and inefficient, particularly when there is

little information to direct sampling effort (e.g., Johara et al. 2001; Poon et al. 2005; Pourrut et al. 2009). Therefore, it is imperative to develop efficient methods for identifying reservoir hosts.

Recent advances in trait-based models have increased precision, and in turn predictive power, to facilitate identification of unknown reservoirs of viruses in nature (Becker et al. 2020a; Crowley et al. 2020). However, significant questions remain about how modelers should implement these approaches, particularly in regards to the type and level of evidence for virus infection and ability for onward transmission (Becker et al. 2020b; Worsley-Tonks et al. 2020). Most models are based on serology data (i.e., antibodies), which tend to be abundant due to its relative ease and cost-effectiveness to collect. However, such information often only provides evidence of virus exposure, not necessarily current infection (Gilbert et al. 2013). Polymerase chain reaction (PCR), on the other hand, provides stronger evidence of current infection, and is a better predictor of host competence than serology data (Tolsá et al. 2018). However, virus infection does not necessarily equate to onward transmission potential. Instead, the “gold standard” and least common evidence for competent reservoir hosts (i.e., those capable of transmitting virus) is live virus isolation (Corona et al. 2018). Current understanding of how these different types of evidence alter predictive capacity is limited, despite clear differences in host associations and relevance to model performance in applied settings (i.e., future efforts to search for reservoirs).

Orthohantaviruses (Bunyavirales: Hantaviridae) are an ideal virus group to examine using predictive models, due to their broad implications for human health as zoonotic pathogens, the predicted large number of unidentified viruses (Vaheri et al. 2008), and the varying types of virus infection and competence evidence currently available from wildlife surveys. There are currently 58 described orthohantaviruses, primarily found in rodents (Laenen et al. 2019), that

cause two main human diseases: hemorrhagic fever with renal syndrome (HFRS, which is common throughout the Old World) and hantavirus cardiopulmonary syndrome (HCPS or HPS, which is common throughout the New World). Because each human case is thought to be an independent spillover event from an infected rodent (Forbes et al. 2018; Avšič-Županc et al. 2019), identifying orthohantavirus reservoir host species is critical for efforts to mitigate human disease.

With few exceptions, including mole- and shrew-borne orthohantaviruses (Arai et al. 2007; Arai et al. 2008; Kang et al. 2009; Kang et al. 2011), most known orthohantaviruses infect rodents in the families Cricetidae and Muridae (superfamily Muroidea), including all orthohantaviruses that cause disease in humans (Forbes et al. 2018). Because spillover is constrained by phylogenetic distance (Streicker et al. 2010), undiscovered orthohantaviruses are also likely to be found among muroid rodents. Additionally, although the majority of described American orthohantaviruses cause disease in humans (13/22), knowledge of host relationships is weak for these viruses, and frequent discovery of novel orthohantaviruses indicates a high likelihood of unknown viruses in this part of the world (Mull et al. 2020). Efforts to predict novel orthohantavirus reservoirs can therefore be focused within New World muroids for maximum precision and impact.

In this study, we used machine learning approaches to predict reservoir hosts of unknown American orthohantaviruses. Predictions were generated by combining muroid phylogenetic and trait data with two levels of evidence for the propensity of a species to host orthohantaviruses: (1) RT-PCR (termed infection) and (2) live virus isolation (termed competence). Model performance was compared using these two evidence types to determine the power of our various methods for identifying undiscovered orthohantavirus hosts. Finally, host predictions

were incorporated with geospatial data on human exposure potential to predict geographic areas with greatest zoonotic spillover risk. Generated results will guide ongoing and future efforts to discover novel orthohantaviruses and determine virus-host relationships. More broadly, determining effective modeling approaches, specifically the role of infection versus competence data, is critical to optimizing tools for identifying and understanding potential zoonotic threats to human health and security.

Results

Phylogenetic patterns

Across our 601 New World muroid rodent species, 9.32% displayed evidence of orthohantavirus infection, whereas only 2% were found positive for virus isolation (Fig 1). We identified intermediate phylogenetic signal in infection ($D = 0.81$) but little phylogenetic signal in competence ($D = 0.90$). For the former, phylogenetic patterns in infection departed from both randomness ($p < 0.001$) and Brownian motion ($p < 0.001$), whereas competence departed from Brownian motion ($p < 0.001$) but not phylogenetic randomness ($p = 0.16$). Results from phylogenetic factorization were qualitatively similar. We identified two rodent clades with significantly greater propensities to have orthohantavirus infection. A subclade of the genus *Peromyscus* ($n = 24$) and the whole genus *Oligoryzomys* ($n = 20$) had 37.5% and 40% of species predicted to be capable of becoming infected, respectively, compared to 8% of the paraphyletic remainder. In contrast, our analyses identified no taxonomic patterns in competence.

Model performance

Both infection and competence BRT models distinguished orthohantavirus positive and negative rodent species with high accuracy ($AUC = 0.92 \pm 0.002$). However, BRTs trained on host competence performed significantly better ($AUC = 0.94 \pm 0.003$) than those trained on infection ($AUC = 0.91 \pm 0.003$; $t = 6.24$, $p < 0.001$; Fig 2A), resulting in a moderate effect size ($d = 0.62$; Cohen 1988). Despite this difference in model performance, both models identified similar species traits as predictive of positivity (infection and competence). Rankings of variable importance were strongly correlated ($\rho = 0.88$, $p < 0.001$), even after removing traits with zero relative importance ($n = 42$ remaining features; $\rho = 0.82$, $p < 0.001$). Consistently important features for both response variables included PubMed citations, litter size, and both mammal richness and mean precipitation within the species range. Consistently unimportant features included the genera *Neotoma*, *Rhipidomys*, *Nectomys*, and *Handleyomys*. Major discrepancies included the genus *Peromyscus* and activity cycle being important predictors of infection but not competence and the genus *Oryzomys* being an important predictor of competence but not infection (Fig 2B, S2 Table). Partial dependence plots suggested that the directions of effects were largely consistent across models, with positive species being well-studied, located in mammal-rich regions, and characterized by a faster life history (S1 Fig). However, our secondary BRTs showed that citations were not predictable by traits ($AUC = 0.49 \pm 0.001$), suggesting that the trait profile of positive rodents is not confounded by the traits of well-studied species.

Model prediction

Predicted probabilities of being an orthohantavirus host varied widely across the 601 rodent species but were only weakly positively correlated between infection and competence BRTs ($\rho = 0.14$, $p < 0.001$; Fig 3A). Many species with intermediate-to-high propensity scores

from models based on infection had a low corresponding probability of being competent. Whereas both predictions displayed moderate phylogenetic signal ($\lambda = 0.58$ and 0.54 , respectively), the taxonomic patterns identified by phylogenetic factorization largely differed between models (Fig 3B, S3 Table). For both infection and competence models, the genus *Oligoryzomys* ($n = 20$) had a greater mean probability of orthohantavirus hosting compared to the paraphyletic remainder. Predictions from infection models otherwise largely mirrored observed patterns in the data, with a subclade of the genus *Peromyscus* ($n = 25$) also having greater propensity scores ($\bar{x} = 0.51$), although a subclade of the genus *Oxymycterus* ($n = 6$) and the subfamily Arvicolinae (including voles, lemmings, and muskrats; $n = 43$) had greater and lower probabilities of infection ($\bar{x} = 0.58$ and $\bar{x} = 0.20$, respectively). Predictions from competence models instead were clustered in the genus *Oryzomys* ($n = 6$), a subclade of *Oecomys* ($n = 7$), a smaller subclade of *Peromyscus* ($n = 7$), and the genus *Sigmodon* ($n = 13$), all of which had greater probabilities of being reservoirs.

Lastly, we stratified our results into binary predictions using a 95% sensitivity threshold (S4 Table). This revealed a total 98 likely undiscovered hosts based on infection models versus only 14 undiscovered hosts based on competence models, of which 8 were also predicted by the former (Table 1). Mapping the geographic distribution of undetected hosts alongside known orthohantavirus-positive rodent species revealed that while predictions from infection models largely recapitulated the distributions of known RT-PCR-positive species, competence models suggested novel hotspots of overlapping reservoirs in the northeastern United States and northern South America, particularly along the Andes Mountains (Fig 4).

Table 1. Predicted undiscovered hosts of hantaviruses: a priority list for future sampling efforts.

173

<i>Genus</i>	<i>species</i>
Abrothrix	olivaceus*
Akodon	boliviensis, leucolimnaeus, mystax, paranaensis, pervalens
Baiomys	musculus, taylori
Brucepattersonius	griserufescens, soricinus
Calomys	callosus, cerqueirai, expulsus, musculus, tocantinsi, venustus
Cerradomys	subflavus, vivoi
Chelemys	macronyx
Eligmodontia	bolsonensis, typus
Holochilus	brasiliensis, lagigliai, sciureus
Melanomys	caliginosus
Microryzomys	minutus
Microtus	pinetorum
Mus	musculus
Myodes	gapperi
Neacomys	musseri, spinosus* , tenuipes
Necomys	lenguarum, punctulatus, urichi*
Nectomys	apicalis, magdalenae, squamipes
Neotoma	leucodon
Nyctomys	sumichrasti
Oecomys	bicolor* , catherinae, concolor* , roberti, sydandersoni* , trinitatis
Oligoryzomys	andinus, brendae, delticola, destructor* , elurus, flavescens , moojeni, rupestris, victus
Ondatra	zibethicus
Onychomys	torridus
Oryzomys	antillarum*
Oxymycterus	amazonicus, angularis, caparoe, dasytrichus, josei, quaestor, roberti
Peromyscus	carletoni, crinitus, difficilis, fraterculus, gratus, keeni, melanophrys, mexicanus, nasutus, pembertoni, polionotus, sagax, schmidlyi

Phyllotis	xanthopygus
Rattus	exulans
Reithrodon	auritus
Reithrodontomys	megalotis
Rhipidomys	cariri, couesi, emiliae, gardneri, ipukensis, itoan, leucodactylus, macrurus, mastacalis, nitela, ochrogaster, tribei
Sigmodon	fulviventer, hirsutus, planifrons, toltecus
Thomasomys	cinereiventer, onkiro, popayanus, ucucha, vulcani

Plain text species are predicted by the infection model only, bolded species are predicted by the competence model, and starred species are predicted by both infection and competence models.

We then mapped these total estimated sets of predicted unknown orthohantavirus hosts and reservoirs against anthropogenic impacts (Fig 5), as a proxy for cumulative and current excess spillover risk contributed by these environmental drivers. Hantavirus hosts coincide most with areas experiencing high anthropogenic impacts in central America and the Atlantic forests of Brazil and Uruguay. Reservoirs are distributed more evenly and extensively throughout the Americas, especially in both the Amazon and in high-latitude temperate ecosystems in North America. The greatest coincidence of those species with emergence risk factors may be in North American rural population centers and agricultural communities.

Discussion

In this study, we identified rodent species that are likely to host orthohantaviruses and demonstrated that the inclusion of competence data improves both model performance and generates distinct predictions from traditional models using RT-PCR data (i.e., infection). Determining the reservoir host of a particular orthohantavirus can be challenging, as orthohantaviruses are difficult to isolate (Strandin et al. 2020) and, like many infectious diseases,

infection prevalence varies drastically across space and time (Walsh et al. 2007; Vadell et al. 2011; Holsomback et al. 2013). However, predictive modeling enables the detection of novel hosts in the absence of field data and in turn facilitates targeted field surveillance that can ultimately be used to mitigate hazards posed by zoonotic viruses (Becker et al. 2019).

Orthohantaviruses have traditionally been considered to follow evolutionary cophylogenies with their hosts, with few cross-species infections denoting distinct lineages (Hjelle et al. 1995; Herbreteau et al. 2006; Song et al. 2007). However, the discovery of additional orthohantaviruses has since expanded the diversity of hosts and demonstrated host switches in their evolutionary history (Blasdell et al. 2011; Guo et al. 2013). Indeed, orthohantaviruses have been isolated from species among all four subfamilies of muroid rodents in the Americas. Within those subfamilies, orthohantaviruses have been isolated from seven genera, and the subset of hosts predicted by both models would expand this range by four additional genera (Table 1). Additionally, the contrasting phylogenetic patterns of infection and competence, alongside the differing importance of taxonomic predictors in our two models, suggests that many orthohantavirus host-switches among disparate species have occurred, and frequent viral sharing among related species likely helps to account for the clusters of closely related cophylogenies (Fig 1).

In our study, postulated reservoir hosts are mostly concentrated in two regions, northeastern United States and northern South America, but southern Mexico and eastern Brazil are regions of likely spillover (Fig 4). Interestingly, all of these regions coincide with geographical gaps in known orthohantavirus distribution (Guzmán et al. 2017). In particular, not only would the discovery of an orthohantavirus hosted by *Myodes gapperi* bridge a geographic gap between Russia and North America, but it would also bridge a phylogenetic gap between

Eurasian and American viruses. Several *Microtus* species are the only arvicoline rodents currently known to host orthohantaviruses in the New World, despite a variety of arvicoline hosts in the Old World (Blasdell et al. 2011).

Competence models (virus isolation data) were both more accurate and precise in estimating reservoir hosts compared to infection models (RT-PCR data). Although model performance (AUC) for our infection models was high, the higher AUC of the competence models (despite fewer rodent species with live virus isolation records) indicates the importance of including stronger evidence for reservoir capacity in model performance (Fig 2; Becker et al. 2020b). Additionally, there was substantial overlap in predicted host species, but the competence model produced a more concise list of virus reservoir candidates (Table 1). For example, in the present study, *Peromyscus* demonstrate the tendency for certain taxa to be infected more often without being reservoirs (Figs 1 and 3). Considerations of such differences in model performance are crucial when focusing surveillance efforts on likely reservoirs, and adherence to predictions based on infection status alone could lead to wasteful sampling of misidentified species.

Although this study focused on New World orthohantaviruses to enable higher resolution results in this system, our modeling approach is transferable to many other systems. Old World orthohantaviruses represent the most obvious extension, particularly for regions with minimal surveillance, such as Africa, the eastern Mediterranean, and Southeast Asia (Herbreteau et al. 2006; Guo et al. 2013). However, other virus groups that pose a threat to human welfare would also benefit from predictive modeling. For example, the reservoir hosts, and likely virus diversity, of orthopoxviruses (e.g., cowpox virus, monkeypox virus) are still mostly unknown, despite common evidence of orthopoxvirus infection among a diverse assemblage of wildlife, particularly rodents (McInnes et al. 2006; Kinnunen et al. 2011) and carnivores (Emerson et al.

2009; Morgan et al. 2019). In cases like this, models incorporating multiple levels of infection evidence can help filter out sampling “noise” to empower host detection for the many known and future emerging infectious diseases (Jones et al. 2008).

Predicted species in this study represent priority targets for orthohantavirus surveillance, particularly *Myodes gapperi* in North America and several members of the tribe Oryzomyini in South America. However, field verification of these predictions will be necessary to ultimately determine the true diversity of novel reservoirs. More broadly, we demonstrate here that the inclusion of competence data strengthens trait-based predictive modeling, and tailoring models based on outcomes of field studies will further improve accuracy. These methods will increase efficiency in host surveillance not only for orthohantaviruses, but also for a range of other pathogens important in human and wildlife health.

Methods

Hantavirus data

A systematic literature search was conducted in Web of Science to identify empirical studies that reported orthohantavirus infections in New World muroid rodents via RT-PCR (specific to negative-sense RNA viruses) or virus isolation (S1-3 Appendix). We recorded the number of studies per rodent species with each of the following criteria: at least one individual RT-PCR-positive; all individuals RT-PCR-negative; or virus isolation from at least one individual. Because orthohantaviruses cause persistent and chronic infections in rodents (Forbes et al. 2018), serological tests are often used to demonstrate current or recent infection, and RT-PCR is performed only on samples from antibody-positive individuals for virus characterization (Vaheri et al. 2008). To preclude false positives in these studies, only rodents that had positive

RT-PCR results were considered RT-PCR-positive, and all other individuals were considered RT-PCR-negative, even if RT-PCR was not conducted. If a study used only serology without either RT-PCR or live virus isolation attempts, then the study was not included. When studies attempted virus isolation, additional RT-PCR results were recorded for specimen tissue analyses, but not infected cell cultures.

In studies that employed archived samples reported in a previous study (for the same level of evidence), those samples were omitted from our tallies to preclude pseudoreplication; instead, the original study was used. If a subsequent study examined a different level of evidence (e.g. virus isolation vs. RT-PCR), then we treated the two studies as a single report. In instances where the number or description of positive and negative results for each species was not clear in an article (including specimens reported at the genus level and outdated taxonomy that now represents multiple species), only definitive results were recorded. We manually matched select rodent species names between our orthohantavirus data and our phylogeny and trait data (see below). Species synonyms are provided in our online data repository. Since several *Rattus* and *Mus* are abundant in the Old and New World, only results derived in the Americas were included. Species without published evidence of orthohantavirus infection or competence were assigned pseudoabsences (Becker et al. 2020a).

Phylogenetic analyses

We used a recently developed supertree of extant mammals to capture rodent phylogeny (Upham et al. 2019). The tree was simplified to our specified rodent species using the *ape* package in R (Paradis et al. 2004). Prior to predictive models, we conducted two assessments of phylogenetic signal (i.e., the propensity for related rodent species to be more similar in virus

positivity). For both response variables (infection and competence), we used the *caper* package to calculate D , where a value of 1 indicates a phylogenetically random trait distribution and a value of 0 indicates phylogenetic clustering under a Brownian motion model of evolution (Fritz and Purvis 2010). Significant departure from either model was quantified using a randomization test with 1,000 permutations. However, because traits may also arise under a punctuated equilibrium model of evolution, we next used a graph-partitioning algorithm, phylogenetic factorization, to flexibly identify clades with significantly different propensity to be infected or competent at various taxonomic depths (Washburne et al. 2019). We used the *phylofactor* package to partition both outcomes as Bernoulli-distributed response variables with generalized linear models. We determined the number of significant phylogenetic factors (clades) using a Holm's sequentially rejective 5% cutoff for the family-wise error rate.

Rodent traits

We used a published dataset of 55 traits describing the morphology, geography, taxonomy, and life history of rodent species. Trait data were primarily from PanTHERIA alongside derived covariates including postnatal growth rate, relative age to sexual maturity, relative age at first birth, production, and species density (Jones et al. 2009; Han et al. 2015). We also used the *picante* package to quantify evolutionary distinctiveness, a measure of how isolated a species is within our muroid phylogeny (Kembel et al. 2010; Redding and Mooers 2006). Finally, we included binary covariates for our muroid rodent genera to represent taxonomy. We excluded predictors with no variance or missing values for over 75% of species, resulting in a total set of 62 biological covariates (S1 Table). Lastly, we used the *easyPubMed* package

(accessed September 2020) to obtain the number of citations per species as a proxy for sampling effort (Olival et al. 2017; Fantini 2019).

Boosted regression trees

We used boosted regression trees (BRTs) to classify rodent species as orthohantavirus hosts based on our predictor matrix of traits. BRTs circumvent many statistical issues associated with traditional hypothesis testing (e.g., a large number of predictors, complex interactions, non-randomly missing covariates) and can uncover new and surprising patterns in data to help develop testable hypotheses or predictions (Hochachka et al. 2007). Using this machine learning approach, we modeled binomial virus positivity separately for infection and competence.

BRTs maximize classification accuracy by learning patterns of features that best distinguish positive and negative hosts (Elith et al. 2008). This generates recursive binary splits for randomly sampled predictor variables, and successive trees are built using residuals of the prior best-performing tree as the new response. Boosting generates an ensemble of linked trees, where each achieves increasingly more accurate classification. Prior to analysis, we randomly split data into training (70%) and test (30%) sets while preserving the proportion of positive labels using the *rsample* package. Models were then trained with the *gbm* package (Greenwell et al. 2020), with the maximum number of trees set to 10000, a learning rate of 0.01, and an interaction depth of three. BRTs used a Bernoulli error distribution and five-fold cross-validation, and we used the *ROCR* package to quantify accuracy as area under the receiver operator curve (AUC; Sing et al. 2005). As results can depend on random splits between training and test data, we used 100 partitions to generate an ensemble (Evans et al. 2017). To diagnose if

trait profiles of positive species are driven by study effort, we ran a secondary set of BRTs that modeled citation counts as a Poisson response (Plowright et al. 2019).

Model performance and prediction

To assess how BRT performance varied between infection and competence models (Becker et al. 2020b), we used a paired *t*-test to compare AUC. We also assessed similarity in variable importance between models by estimating the Spearman correlation coefficient between feature ranks. Next, we predicted the probability of a species being positive for either response. When predicting species status, we set citation counts per species to their mean across species as a *post hoc* method to correct for sampling effort and remove at least some bias (Becker et al. 2020a). Lastly, we also estimated the Spearman correlation coefficients for the mean predictions between infection and competence models.

We used these mean predictions to identify “false negative” orthohantavirus hosts (i.e., those without a prior recorded orthohantavirus infection or isolation). We identified taxonomic patterns in predictions using Pagel’s λ as an estimate of phylogenetic signal with the *caper* package (Orme et al. 2013) as well as a secondary phylogenetic factorization to identify clades with significantly different predicted probabilities. To identify potential unknown hosts or reservoirs, we estimated a 95% sensitivity threshold using the *presenceabsence* package (Freeman and Moisen 2008), which can stratify predictions at a 5% omission rate on known true positives. This threshold, while fairly inclusive, mostly selects species with comparable probabilities of being infected or competent to known hosts.

To visualize the spatial distribution of known and predicted rodent hosts, we used the IUCN Red List database of mammal geographic ranges and overlaid these shapefiles for

thresholded species based on infection and competence models. We finally mapped the distribution of known and predicted hosts and reservoirs against a proxy for cumulative anthropogenic impact on natural systems, given by the SEDAC Last of the Wild database's 2009 Human Footprint map (Venter et al. 2016; Venter et al. 2018). This qualitative descriptor encompasses several geospatial layers that describe anthropogenic impacts with relevance to human exposure to rodents and hantaviruses, particularly human occupation (i.e., built-up settlements and human population), agricultural intensification (i.e., crop lands and pasture lands), and ecosystem fragmentation (i.e., road and railway density).

Acknowledgements

This work was supported by the Viral Emergence Research Initiative (VERENA) consortium

References

- Arai S, Song J-W, Sumibcay L, Bennett SN, Nerurkar VR, Parmenter C, et al. Hantavirus in northern short-tailed shrew, United States. *Emerg Infect Dis.* 2007;13(9):1420–3.
- Arai S, Bennett SN, Sumibcay L, Cook JA, Song J-W, Hope A, et al. Short report: Phylogenetically distinct hantaviruses in the masked shrew (*Sorex cinereus*) and dusky shrew (*Sorex monticolus*) in the United States. *Am J Trop Med Hyg.* 2008;78(2):348–51.
- Avšič-Županc T, Saksia A, Korva M. Hantavirus infections. *Clin Microbiol Infect.* 2019;21:e6–16.
- Becker DJ, Albery GF, Sjodin AR, Poisot T, Dallas TA, Eskew EA, et al. Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization. *BioRxiv* 111344

[Preprint]. 2020a [cited 2020 Dec 23]. Available from:
<https://www.biorxiv.org/content/10.1101/2020.05.22.111344v3>

Becker DJ, Seifert SN, Carlson CJ. Beyond infection: Integrating competence into reservoir host prediction. *Trends Ecol Evol.* 2020b;35(12):1062–5.

Becker DJ, Washburne AD, Faust CL, Mordecai EA, Plowright RK. The problem of scale in the prediction and management of pathogen spillover. *Philos Trans R Soc B* 2019;374(1782):20190224.

Blasdell K, Henttonen H, Buchy P. Hantavirus genetic diversity. In: Morand S, Beaudeau F, Cabaret J, editors. *New frontiers of molecular epidemiology of infectious diseases*. Dordrecht: Springer; 2011. p. 179–216.

Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. New York: Routledge; 1988.

Corona TF, Böger B, da Rocha TC, Svoboda WK, Gomes EC. Comparative analysis of mouse inoculation test and virus isolation in cell culture for rabies diagnosis in animals of Parana, Brazil. *Rev Soc Bras Med Trop.* 2018;51(1):39–43.

Crowley D, Becker D, Washburne A, Plowright R. Identifying suspect bat reservoirs of emerging infections. *Vaccine* 2020;8(2):228.

Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol.* 2008;77(4):802–13.

Emerson GL, Li Y, Frace MA, Olsen-Rasmussen MA, Khristova ML, Govil D, et al. The phylogenetics and ecology of the orthopoxviruses endemic to North America. *PLoS One* 2009;4(10):e7666.

394 Evans MV, Dallas TA, Han BA, Murdock CC, Drake JM. Data-driven identification of potential
395 Zika virus vectors. *Elife* 2017;6:22053.

396 Fantini D. easyPubMed: Search and retrieve scientific publication records from PubMed. R
397 package version 2.13. 2019.

398 Freeman EA, Moisen G. PresenceAbsence: An R package for presence absence analysis. *J Stat*
399 *Softw.* 2008;23(11):31 P.

400 Forbes KM, Sironen T, Plyusnin A. Hantavirus maintenance and transmission in reservoir host
401 populations. *Curr Opin Virol.* 2018;28:1–6.

402 Fritz SA, Purvis A. Selectivity in mammalian extinction risk and threat types: A new measure of
403 phylogenetic signal strength in binary traits. *Conserv Biol.* 2010;24(4):1042–51.

404 Gilbert AT, Fooks AR, Hayman DTS, Horton DL, Müller T, Plowright R, et al. Deciphering
405 serology to understand the ecology of infectious diseases in wildlife. *Ecohealth*
406 2013;10(3):298–313.

407 Greenwell B, Boehmke B, Cunningham J, Ridgeway G. gbm: Generalized boosted regression
408 models. R package version 2.1.8. 2020

409 Guo W-P, Lin X-D, Wang W, Tian J-H, Cong M-L, Zhang H-L, et al. Phylogeny and origins of
410 hantaviruses harbored by bats, insectivores, and rodents. *PLoS Pathog.*
411 2013;9(2):e1003159.

412 Guzmán C, Calderón A, González M, Mattar S. Hantavirus infections. *Rev MVZ Córdoba*
413 2017;22:6101–17.

414 Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. *Proc*
415 *Natl Acad Sci U S A* 2015;112(22):7039–44.

416 Haydon DT, Cleveland S, Taylor LH, Laurenson MK. Identifying reservoirs of infection: A
417 conceptual and practical challenge. *Emerg Infect Dis.* 2002;8(12):1468–73.

418 Herbreteau V, Gonzalez J-P, Hugot, J-P. Implication of phylogenetic systematics of rodent-borne
419 hantaviruses allows understanding of their distribution. *Ann N Y Acad Sci.*
420 2006;1081:39–56.

421 Hjelle B, Lee S-W, Song W, Torrez-Martinez N, Song J-W, Yanagihara R, et al. Molecular
422 linkage of hantavirus pulmonary syndrome to the white-footed mouse, *Peromyscus*
423 *leucopus*: Genetic characterization of the M genome of New York virus. *J Virol.*
424 1995;69(12):8137–41.

425 Hochachka WM, Caruana R, Fink D, Monsun ART, Riedewald M, Sorokina, D, et al. Data-
426 mining discovery of pattern and process in ecological systems. *J Wild Manage*
427 2007;71(7):2427–37.

428 Holsomback TA, Van Nice CJ, Clark RN, McIntyre NE, Abuzeineh AA, Salazar-Bravo J. Soio-
429 ecology of the marsh rice rat (*Oryzomys palustris*) and the spatio-temporal distribution of
430 Bayou virus in coastal Texas. *Geospat Health* 2013;7(2):289–98.

431 Johara MY, Field H, Rashdi AM, Morrissy C, van der Heide B, Rota P, et al. Nipah virus
432 infection in bats (order Chiroptera) in Peninsular Malaysia. *Emerg Infect Dis.*
433 2001;7(3):439–41.

434 Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme DL, et al. PanTHERIA: A species-
435 level database of life history, ecology, and geography of extant and recently extinct
436 mammals. *Ecology* 2009;90(9):2648

437 Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. Global trends in
438 emerging infectious diseases. *Nature* 2008;451(7181):990–3.

Kang HJ, Bennett SN, Dizney L, Sumibcay L, Arai S, Ruedas LA, et al. Host switch during evolution of a genetically distinct hantavirus in the American shrew mole (*Neurotrichus gibbsii*). *Virology* 2009;388(1):8–14.

Kang HJ, Bennett SN, Hope AG, Cook JA, Yanagihara R. Shared ancestry between a newfound mole-borne hantavirus and hantaviruses harbored by cricetid rodents. *J Virol.* 2011;85(15):7496–503.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly, DD, et al. Picante: R tools for integrating phylogenies and Ecology. *Bioinformatics* 2010;26(11):1463–4.

Kinnunen PM, Henttonen H, Hoffmann B, Kallio ER, Korthase C, Laakkonen J, et al. Orthopox virus infections in Eurasian wild rodents. *Vector Borne Zoonotic Dis.* 2011;11(8):1133–40.

Laenen L, Vergote V, Calisher CH, Klempa B, Klingström J, Kuhn JH, et al. *Hantaviridae*: Current classification and future perspectives. *Viruses* 2019;11(9):788.

Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, et al. Fruit bats as reservoirs of Ebola virus. *Nature* 2005;438(7068):575–6.

McInnes CJ, Wood AR, Thomas K, Sainsbury AW, Gurnell J, Dein FJ, et al. Genomic characterization of a novel poxvirus contributing to the decline of the red squirrel (*Sciurus vulgaris*) in the UK. *J Gen Virol.* 2006;87(Pt 8):2115–25.

Morgan CN, López-Perez AM, Martínez-Duque P, Jackson FR, Suzán G, Gallardo-Romero NF. Prevalence of antibodies to orthopoxvirus in wild carnivores of northwestern Chihuahua, Mexico. *J Wildl Dis.* 2019;55(3):637–44.

Mull N, Jackson R, Sironen T, Forbes KM. Ecology of neglected rodent-borne American orthohantaviruses. *Pathogens* 2020;9(5):325.

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

464 Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, et al. The Caper package:
465 Comparative analysis of phylogenetics and evolution in R. R package version 1.0.1.
466 2018.

467 Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language.
468 *Bioinformatics* 2004;20(2):289–90.

469 Plowright RK, Becker DJ, Crowley DE, Washburne AD, Huang T, Nameer PO, et al.
470 Prioritizing surveillance of Nipah virus in India. *PLoS Negl Trop Dis*.
471 2019;13(6):e0007393.

472 Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, et al. Pathways to
473 zoonotic spillover. *Nat Rev Microbiol*. 2017;15(8):502–10.

474 Poon LLM, Chu DKW, Chan KH, Wong OK, Ellis TM, Leung YHC, et al. Identification of a
475 novel coronavirus in bats. *J Virol*. 2005;79(4):2001–9.

476 Pourrut X, Souris M, Towner JS, Rollin PE, Nichol ST, Gonzalez J-P, et al. Large serological
477 survey showing circulation of Ebola and Marburg viruses in Gabonese bat populations,
478 and a high seroprevalence of both viruses in *Rousettus aegyptiacus*. *BMC Infect Dis*.
479 2009;9:159.

480 Redding DW, Mooers AØ. Incorporating evolutionary measures into conservation prioritization.
481 *Conserv Biol*. 2006;20(6):1670–8.

482 Strandin T, Smura T, Ahola P, Aaltonen K, Sironen T, Hepojoki J, et al. Orthohantavirus isolated
483 in reservoir host cells displays minimal genetic changes and retains wild-type infection
484 properties. *Viruses* 2020;12(4):457.

485 Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R.
486 Bioinformatics 2005;21(20):3940–1.

487 Song J-W, Baek LJ, Schmaljohn CS, Yanagihara R. Thottapalayam virus, a prototype
488 shrewborne hantavirus. Emerg Infect Dis. 2007;13(7):980–5.

489 Streicker DG, Trumelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE. Host
490 phylogeny constrains cross-species emergence and establishment of rabies virus in bats.
491 Science 2010;329(5992):676–9.

492 Tolsá MJ, García-Peña GE, Rico-Chávez O, Roche B, Suzán G. Macroecology of birds
493 potentially susceptible to West Nile virus. Proc Biol Sci. 2018;285(1893):20182178.

494 Upham NS, Esselstyn JA, Jetz W. Inferring the mammal tree: Species-level sets of phylogenies
495 for questions in ecology, evolution, and conservation. PLoS Biol. 2019;17(12):e3000494.

496 Vadell MV, Bellomo C, Martín AS, Padula P, Villafañe IG. Hantavirus ecology in rodent
497 populations in three protected areas of Argentina. Trop Med Int Health
498 2011;16(10):1342–52.

499 Vaheri A, Vapalahti O, Plyusnin A. How to diagnose hantavirus infections and detect them in
500 rodents and insectivores. Rev Med Virol. 2008;18(4):277–88.

501 Venter O, Sanderson EW, Magrath A, Allan JR, Beher J, Jones KR, et al. Last of the wild
502 project, version 3 (LWP-3): 2009 human footprint, 2018 release. Published online 2018.

503 Venter O, Sanderson EW, Magrath A, Allan JR, Beher J, Jones KR, et al. Global Terrestrial
504 Human Footprint Maps for 1993 and 2009. Scientific Data. Published online
505 2016:160067.

506 Viana M, Mancy R, Biek R, Cleveland S, Cross PC, Lloyd-Smith JO, et al. Assembling evidence
507 for identifying reservoirs of infection. Trends Ecol Evol. 2014;29(5):270–9.

Walsh AS, Louis TA, Glass GE. Detecting multiple levels of effect during survey sampling using a Bayesian approach: Point prevalence estimates of a hantavirus in hispid cotton rats (*Sigmodon hispidus*). *Ecol Modell.* 2007;205(1-2):29–38.

Washburne AD, Silverman JD, Morton JT, Becker DJ, Crowley D, Mukherjee S, et al. Phylofactorization: A graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecol Monogr.* 2019;89(2):e01353.

Worsley-Tonks KEL, Escobar LE, Biek R, Castaneda-Guzman M, Craft ME, Streicker DG, et al. Using host traits to predict reservoir host species of rabies virus. *PLoS Negl Trop Dis.* 2020;14(12):e0008940.

Supporting information

S1 Appendix: Web of Science search terms for empirical study inclusion. The focused search includes New World orthohantavirus names and abbreviations along with several terms for PCR and virus isolation. A separate non-focused search was also conducted that did not include the PCR and virus isolation terms.

S2 Appendix: PRISMA diagram for empirical study inclusion.

S3 Appendix: Reference list for empirical studies used in analyses.

S1 Table. Feature coverage across the 601 muroid rodent species included in the BRT models. Variables are presented as given in their original sources (Jones et al. 2009; Han et al. 2015).

531

532 **S2 Table. Rodent trait importance and ranks for BRTs trained on infection and**

533 **competence.**

534

535 **S3 Table. Phylogenetic factorization of mean predicted probabilities for orthohantavirus**

536 **positivity for (i) infection and (ii) competence models.** The number of retained clades after a

537 5% family-wise error rate, taxa corresponding to those clades, number of species per clade, and

538 mean predicted probabilities for the clade compared to the paraphyletic remainder are shown.

539

540 **S4 Table. Sensitivity of estimated number of undiscovered hosts to thresholding method.**

541

542 **S1 Figure. Partial dependence plots for top traits for infection (left) and competence**

543 **(right).**

544

545

546 **Figures**

547 **Fig 1.**

548

549

Fig 2

553 **Fig 3.**

554

555

556 **Fig 4.**

557

558

559 **Fig 5.**

560

561

Figure captions

Fig 1. Phylogenetic distribution of orthohantavirus-positive muroid rodents in the New World. Species

with evidence of infection (A, RT-PCR) or competence (B, live virus isolation) are displayed in black.

Visualized in red are any clades identified through phylogenetic factorization for having greater virus positivity when compared to the paraphyletic remainder.

Fig 2. Performance of rodent orthohantavirus BRT models trained on infection versus competence

data as the response. (A) Area under the receiver operating characteristic curve (AUC) across 100

random splits of training (70%) and test (30%) data. Boxplots show the median and interquartile range

alongside paired AUC values. (B) Correlation between ranks of mean feature importance between

models. Mean relative importance is given in S2 Table.

Fig 3. Predicted probabilities of rodent orthohantavirus positivity based on infection and competence.

(A) Distribution of propensity scores stratified by known positive, currently negative, and unsampled

species. The scatterplot between predictions includes a smoothed curve and confidence intervals from a

generalized additive model. (B–C) Taxonomic patterns in predictions as identified through phylogenetic

factorization. Segments are scaled by probabilities and colored as in A. Clades identified with

significantly different mean predictions are shown in grey, and additional information (e.g., included

taxa, species richness) is included in S3 Table.

Fig 4. Distribution of orthohantavirus hosts. The distribution of known (A, B) and predicted

undiscovered (C, D) hosts of orthohantaviruses based on infection (A,C) and competence (B, D).

585 **Fig 5. Geographic emergence risk of novel orthohantaviruses.** Risk is presented as a function of total
586 richness of predicted unknown rodent hosts (inferred from infection data; A) and reservoirs (inferred
587 from competence; B) against the anthropogenic footprint on natural ecosystems.