The Journal of Infectious Diseases

MAJOR ARTICLE

The Genetic Diversity of Nipah Virus Across Spatial Scales

Oscar Cortes-Azuero,[1,a] Noémie Lefrancq,[1,a] Birgit Nikolay,[2,a] Clifton McKee,[3,a] Julien Cappelle,[4,a] Vibol Hul,[5,a] Tey Putita Ou,[5,a] Thavry Hoem,[5,a] Philippe Lemey,[6,a] Mohammed Ziaur Rahman,[7,a] Ausraful Islam,[7,a] Emily S. Gurley,[3,a] Veasna Duong,[5,a,a] and Henrik Salje[1a,a]

[1]Department of Genetics, University of Cambridge, Cambridge, United Kingdom; [2]Department of Epidemiology and Training, Epicentre, Paris, France; [3]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland; [4]Joint Research Unit, Animal Santé Territoires Risques Ecosystèmes, Centre de coopération internationale en recherche agronomique pour le développement, Montpellier, France; [5]Virology Unit, Institut Pasteur du Cambodge, Pasteur Network, Phnom Penh, Cambodia; [6]Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium; and [7]Infectious Diseases Division, icddr,b, Dhaka, Bangladesh

Background. Nipah virus (NiV), a highly lethal virus in humans, circulates in *Pteropus* bats throughout South and Southeast Asia. Difficulty in obtaining viral genomes from bats means we have a poor understanding of NiV diversity.

Methods. We develop phylogenetic approaches applied to the most comprehensive collection of genomes to date (N = 257, 175 from bats, 73 from humans) from 6 countries over 22 years (1999–2020). We divide the 4 major NiV sublineages into 15 genetic clusters. Using Approximate Bayesian Computation fit to a spatial signature of viral diversity, we estimate the presence and the average size of genetic clusters per area.

Results. We find that, within any bat roost, there are an average of 2.4 co-circulating genetic clusters, rising to 5.5 clusters at areas of 1500–2000 km^2. We estimate that each genetic cluster occupies an average area of 1.3 million km^2 (95% confidence interval [CI], .6–2.3 million km^2), with 14 clusters in an area of 100 000 km^2 (95% CI, 6–24 km^2). In the few sites in Bangladesh and Cambodia where genomic surveillance has been concentrated, we estimate that most clusters have been identified, but only approximately 15% of overall NiV diversity has been uncovered.

Conclusions. Our findings are consistent with entrenched co-circulation of distinct lineages, even within roosts, coupled with slow migration over larger spatial scales.

Keywords. Nipah virus; phylogenetics; disease modeling; *Pteropus*; emerging pathogens.

Nipah virus (NiV) is a bat-borne virus and a World Health Organization priority pathogen [1]. Most infections in humans are fatal, and while most of them occur following zoonotic spillover, human-to-human transmission is responsible for around a third of known cases [2]. There are currently no approved treatments or vaccines. NiV was first identified in Malaysia in 1999 and has since recurred almost annually throughout South Asia [3–5]. *Pteropus* fruit bats are its reservoir hosts, and spillover pathways vary [6]. In the 1999 outbreak in Malaysia, human infection occurred through contact with pigs who had been infected from eating contaminated fruit [7, 8]. In Bangladesh, the primary cause of human infection is consumption of raw date palm sap from trees upon which infected bats have fed [4]. Infected horses have also been implicated [9], while the source of the outbreaks in Kerala, India, remain unknown.

Despite the substantial risk to human health, little is known about NiV's underlying genetic diversity. *Pteropus* bats are found throughout South and Southeast Asia and are commonly infected with NiV, with serostudies identifying NiV antibodies in 3%–83% of adult bats across the region [3, 10, 11]. However, infection patterns within bat populations remain unclear, including the number of discrete lineages circulating in roosts, the spatial spread of different lineages, or NiV's ability to transmit between different *Pteropus* species. We also do not understand whether specific lineages are linked to increased spillover risk, including specific routes of transmission. These critical knowledge gaps are problematic for assessments of spillover risk and vaccine development.

Characterizing NiV's genetic diversity is difficult as many isolates come from human cases, which remain rare. Even Bangladesh, the country with most spillover events, reports only an estimated approximately 15 spillovers annually [12]. The representativeness of viruses obtained from human infections is also unclear. This has motivated efforts to obtain NiV sequences from bat populations; however, this is difficult as bats are asymptomatic and viral shedding is infrequent [11].

Received 19 December 2023; accepted 25 April 2024; published online 29 April 2024

Presented in part: Epidemics9 Conference, Abstract number EPID2023_0652, Bologna, Italy, December 2023.

[a]V. D. and H. S. contributed equally to this work as joint senior authors.

Correspondence: Henrik Salje, PhD, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK (hs743@cam.ac.uk).

https://doi.org/10.1093/infdis/jiae221

NiV sequencing from bats has focused mainly on a few locations within Thailand, Bangladesh, and Cambodia [13–15]. While it is difficult to make inferences about NiV diversity using sequences from any 1 location, we can obtain a more complete picture by pooling information across different locations combined with appropriate analytical methods. Here we provide a comprehensive assessment of the diversity of NiV using all publicly available sequences coming from 6 countries, along with several previously unpublished sequences. We develop methods that are robust to strong biases in where and when sequences are obtained to track the diversity of NiV across spatial scales (within roost, district, country, and internationally) and quantify the extent to which diversity has been fully identified in locations that have implemented extensive surveillance efforts.

METHODS

Data Collection and Alignment

We collected all available NiV genomes in GenBank (N = 301, Supplementary Table 1) [16] and compiled their date, host species, and place of collection. We also included several previously unpublished sequences (N = 26), collected between 2013 and 2016 in 2 bat roosts in Cambodia. The sampling and screening approach for NiV for these sequences is explained elsewhere [15].

We aligned the sequences using MUSCLE on MEGA-X [17]. Among the sequences, 175 were sampled from 6 different bat host species: *Pteropus lylei* (n = 120), *Pteropus medius* (formerly *Pteropus giganteus*, n = 41), *Pteropus vampyrus* (n = 6), *Pteropus hypomelanus* (n = 6), *Hipposideros larvatus* (n = 1), and a *Taphozous* bat of undetermined species (n = 1). Other sources of sequences were humans (n = 73), pigs (n = 7), a dog (n = 1), and an uncertain host (n = 1). Sequence length varied from 153 bp to 18.2k bp (full genome). There were 64 full-length genomes, 185 genomes from the nucleocapsid (N) gene, and the remainder covering different parts of the NiV genome (Supplementary Table 2).

Phylogenetic Analyses

We evaluated temporal signal and performed model selection using IQ-Tree [18, 19] and Bayesian evaluation of temporal signal (BETS) [20, 21]. We then reconstructed a time-resolved phylogeny using BEAST (version 1.10.4) [22].

To assess whether some sublineages were more likely to result in spillovers than others, we considered only sequences from Bangladesh, as it is the only country with human cases regularly documented. We assessed whether human cases occurred with higher frequency in any 1 of the sublineages as compared to the distribution of sublineages in bats using a Fisher exact test.

We next analyzed the speed at which NiV has spread across South and Southeast Asia by computing the mean spatial pairwise distance as a function of the pairwise evolutionary distance for each pair of sequences.

Characterization of NiV Genetic Clusters

We used PhyCLIP, a phylogenetic clustering Python module, to cluster the sequences in the tree into different genetic clusters [23]. We analyzed sequence cluster distribution across countries and bat host species. We also analyzed the spatial distribution of bat species in our dataset. We implemented logistic regression to explore genetic clusters' spatial signature, their relationship with bat host species, and bat host species spatial signature. We conducted sensitivity analyses where we increased or reduced the number of clusters (see Supplementary Material).

Rarefaction Analysis

We explored how the observed number of genetic clusters evolved with sampling, and how it could evolve if more sequences were to be sampled. Focusing on locations with ≥10 available sequences and ≥2 observed genetic clusters, we implemented a rarefaction analysis using the iNEXT package [24, 25]. We then investigated the relationship between the estimated underlying number of discrete clusters in a location and different ecological variables (mean pairwise spatial distance between sequences, human population density, and percentage of tree coverage) from that location using Poisson regression.

Approximate Bayesian Computation

We implemented Approximate Bayesian Computation to characterize NiV genetic clusters' spatial footprint (Supplementary Figure 1). We estimated the number of genetic clusters and their average size in areas where *Pteropus* bats circulate. For each country with *Pteropus* circulation, we explored different estimates of the geographic range of *Pteropus* bats.

Additional details on the methods can be found in the Supplementary Material.

Ethics Statement

This project was conducted using publicly available sequence data with no identifiable information and therefore did not require ethical approval.

RESULTS

We analyzed 257 sequences from 6 countries (Figure 1*A*) covering a 22-year time period (1999–2020, Supplementary Figure 2). Seventy-three (28%) of the sequences came from human infections, 175 (68%) from bats (Figure 1*B*), and 5 (4%) from other sources. BETS [26] analysis supported a model including the sampling dates of sequences over a model in which they were considered contemporaneous (Bayes factor: 415). We built a time-resolved phylogeny and found that sequences could be broadly characterized into 2 major genotypes, I and II, previously reported [27], and 4 minor genotypes, IA, IB,

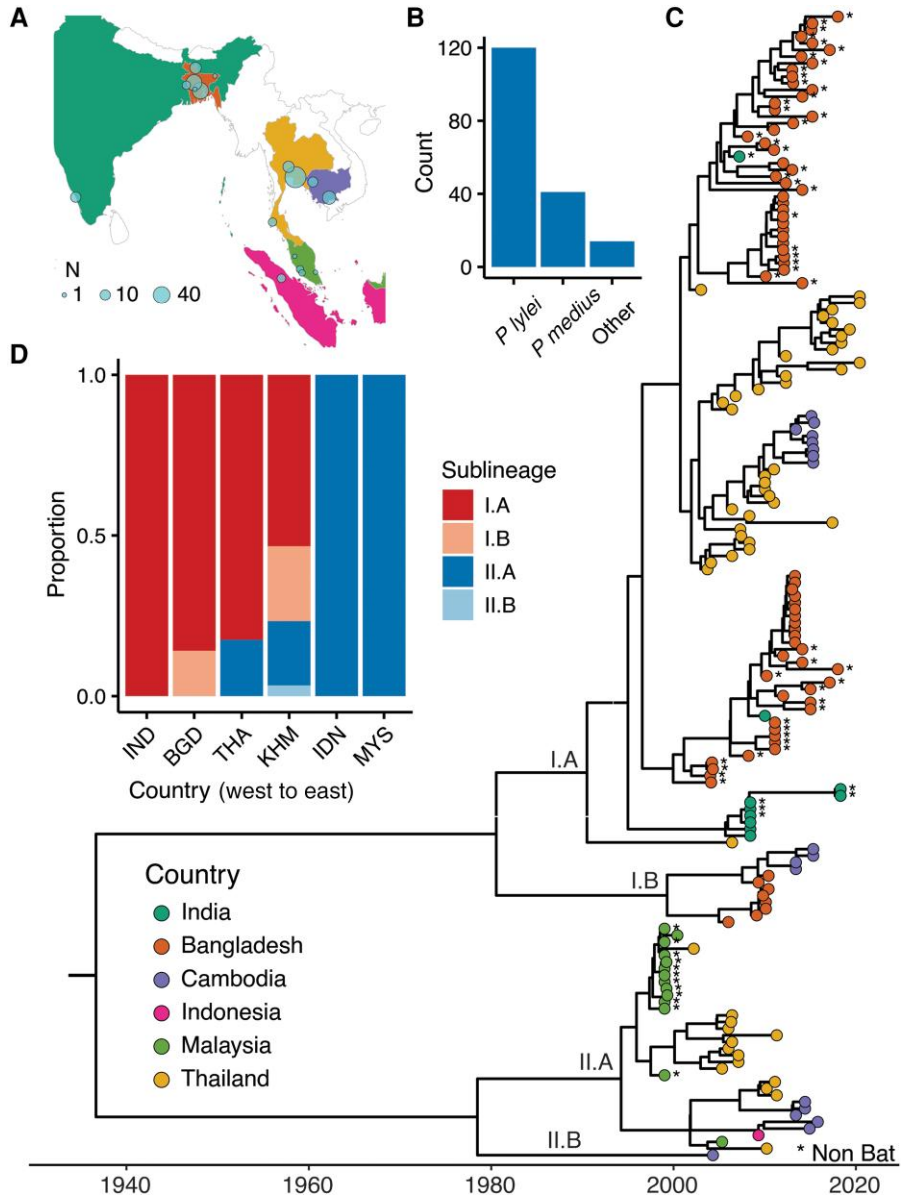

Figure 1. *A*, Country of origin sequences with location of bat roosts. The size of the circles is proportional to the number of samples from each location. *B*, Number of sequences obtained for the different bat species (total N = 175). *C*, Reconstructed time-resolved maximum clade credibility phylogeny; tips are colored with country of origin and non-bat sequences are marked with an asterisk. *D*, Proportion of sequences that come from each sublineage distribution for each country; the countries are ordered from West to East. Abbreviations: BGD, Bangladesh; IDN, Indonesia; IND, India; KHM, Cambodia; MYS, Malaysia; THA, Thailand.

IIA, and IIB (Figure 1*C*). We estimated a mean nucleotide substitution rate of 4.5×10^{-4} substitutions/site/year (95% confidence interval [CI], 2.9×10^{-4} to 6.0×10^{-4}), consistent with previous estimates [14]. Genotypes I and II diverged in 1937 (95% CI, 1838–1983), and the minor genotypes diverged in the 1970s. There was broad spatial structure in these genotypes, with countries on the eastern (Indonesia/Malaysia) and western edges (India) of the region having only 1 circulating sublineage. Cambodia, with a central position, had sequences from all 4 sublineages (Figure 1*D*).

Most sequences from human cases were from Bangladesh (n = 55/73), where there were 2 circulating sublineages (IA and IB). Lineage IA was found throughout 2004–2018, whereas IB was detected only through 2006–2009 (Supplementary Figure 3*A*). We found that all human sequences in Bangladesh came from genotype IA. In contrast, 64.9% (n = 24/37) of bat sequences came from IA and 35.1% (n = 13/37) from IB (Supplementary Figure 3*B*), suggesting that some genotypes may have higher probability of spilling over into human populations ($P = .007$ from Fisher exact test).

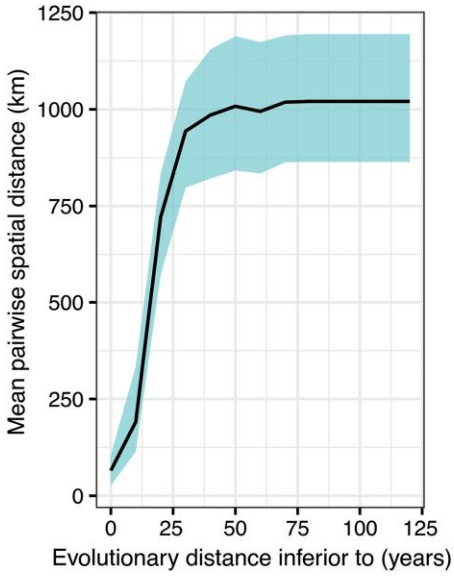

Figure 2. Mean pairwise spatial distance (in km) in function of pairwise evolutionary distance (in years). The shaded area represents 95% confidence intervals obtained from nonparametric bootstrapping of the sequences.

We used the time-resolved phylogeny to investigate spatial structure among sequences by comparing the evolutionary time and spatial separation of each pair of viruses (Figure 2). To mitigate sampling bias, we randomly sampled 1 sequence per bat roost per year, and we sampled 1 sequence per human case cluster. On average, each 10-year increase in evolutionary time was associated with a 186 km (95% CI, 161–212 km) increase in spatial separation, equivalent to 19 km annually.

To provide a finer scale characterization of genetic diversity, we adapted a genetic clustering method, PhyCLIP, to categorize sequences into different genetic clusters based on the distribution of pairwise evolutionary distances, resulting in 15 unique clusters (Figure 3*B*). We assessed our categorization's robustness by reimplementing the algorithm on 100 randomly selected posterior trees. We found highly consistent sequence grouping (median Adjusted Rand Index of 0.94, where 0 would indicate random assignment and 1 perfect consistency) [28]. Genetic clusters diverged between 1978 and 2002, with a mean evolutionary time from a cluster to the next-closest cluster of 15 years. Genetic clusters tended to aggregate within the same country (Figure 3*A*, left) and were also separated by bat species, with 11 of the clusters exclusively found within single bat species (Figure 3*A*, center).

To investigate the spatial overlap of pteropid bat species in the region, we analyzed inter-roost distances (Supplementary Figure 4) and compared species linked to each roost. This assumed all pteropid bats in any roost were of the same species. As many countries in the region only have 1 pteropid species, this is likely a reasonable assumption. We found that bat species were spatially structured at a country level, with 4 (67%) countries having sequences from a single pteropid bat species (Figure 3*A*, right). We estimated that 96.8% (95% CI, 96.4%–97.2%) of bat roosts separated by <100 km were of the same species, dropping to 53.4% (95% CI, 47.9%–58.8%) for roosts 500–1000 km apart (Supplementary Figure 5).

We found that there were an average of 2.41 (95% CI, 1.92–2.94) different genetic clusters per bat roost. The probability that 2 sequences belonged to the same genetic cluster fell from 36.6% (95% CI, 30.6%–45.1%) when they were found within <100 km of each other to 5.7% (95% CI, 2.5%–9.7%) when they were 500–1000 km apart. Using logistic regression, we found that each additional 100 km in spatial distance separating roosts was associated with 0.75 (95% CI, .73–.77) times the odds of being part of the same genetic cluster (Figure 3*C*). As the probability of being from the same bat species is strongly linked to the distance between locations, we could not disentangle the role of spatial segregation between lineages from being solely a spatial effect or due to the different *Pteropus* species occupying different locations. However, on average, we found that sequence pairs coming from the same bat species had 19.3 (95% CI, 15.9–23.6) times the odds of belonging to the same genetic cluster as pairs from different bat species (Figure 3*D*). Importantly, our characterization of the changing probability of being from the same genetic cluster as a function of distance is robust to biased observation processes and can therefore be considered a spatial signature of NiV ecology. Our estimates of spatial dependence were robust to broad variations in the definition of a genetic cluster that results in greater or fewer clusters (Supplementary Figure 6).

We estimated the average number of genetic clusters circulating within an area using Approximate Bayesian Computation to fit our observed spatial distribution of genetic clusters (Figure 4*A*). We made a simplifying assumption that the spatial footprint covered by a genetic cluster is equivalent throughout the region, and that it can be captured by a multivariable normal distribution. We estimate that, on average, 95% of infections from a genetic cluster were found within an area of 1.3 million km^2 (95% CI, .6–2.3 million km^2) and that there were an average of 14 discrete genetic clusters per land area of 100 000 km^2 (95% CI, 6–24 km^2) (Figure 4*B*).

We explored different approximations of the circulation area of *Pteropus* bats to estimate NiV diversity. For each country in the region, we used estimates of *Pteropus* bats' geographic range from the International Union for Conservation of Nature (IUCN) [29], and the Global Forest Watch's area covered by tree cover as an alternative, indirect marker of geographic range. For this second approach, we separately considered over 10%, 30%, or 50% of tree cover as a proxy of the area covered by *Pteropus* bats (Table 1, Supplementary Figures 7 and 8) [30]. For most countries, we found little variation among estimates of diversity using these approaches. For example, in Thailand, we estimate 15 unique genetic clusters (95% CI, 6–26) using

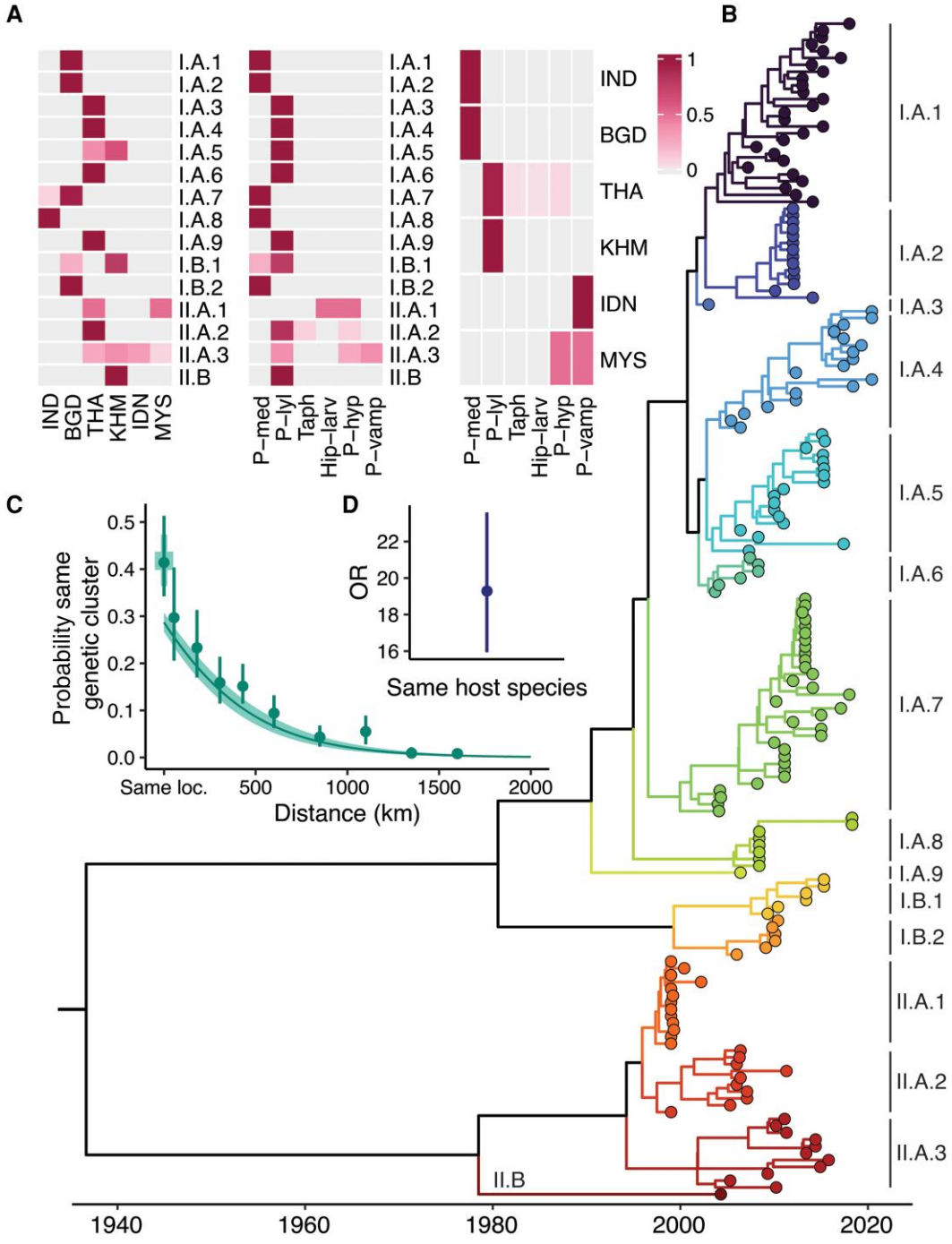

Figure 3. *A*, Proportion of all sequences of each genetic cluster by the country where it was detected (left) and by the bat species from which it was found (middle). The right plot shows the proportion of roosts that belong to each bat species for each country. *B*, Time-resolved phylogeny divided into 15 distinct clades using an adapted form of PhyCLIP. Each color represents a different genetic cluster. *C*, Proportion of sequence pairs belonging to the same cluster as a function of their spatial distance. Dots represent median values, and error bars represent 95% bootstrap confidence intervals (CIs). The line represents the fit of a logistic model. The green shaded region represents 95% CIs of the model fit. *D*, Odds ratio of belonging to the same genetic cluster if sequence pairs were sampled from the same bat host species or not. The error bar represents 95% bootstrap CIs. Abbreviations: BGD, Bangladesh; IDN, Indonesia; IND, India; KHM, Cambodia; MYS, Malaysia; OR, Odds Ratio; THA, Thailand.

the IUCN's geographic range, and 17 unique genetic clusters (95% CI, 7–29) using the area with >10% of tree cover. In Bangladesh, estimates ranged between 15 genetic clusters (95% CI, 7–29) and 11 genetic clusters (95% CI, 4–20), respectively. Taking South and Southeast Asia as a whole, we estimate between 66 and 118 separate NiV genetic clusters using the different approaches (range of confidence intervals: 28–211), compared to 15 currently detected, suggesting that

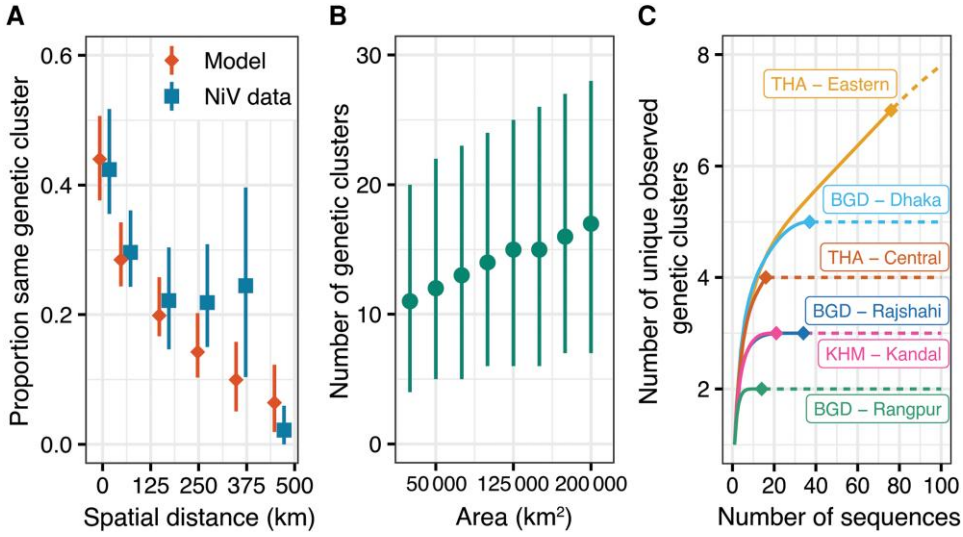

Figure 4. *A*, Model fit using approximate Bayesian computation on the proportion of sequence pairs that belong to the same genetic cluster in function of their spatial distance. The orange dots represent the median and the lines represent the 95% confidence intervals (CIs). Median proportions calculated on Nipah virus (NiV) data and used to calibrate the model are represented in blue. *B*, Predicted number of genetic clusters in function of area (in square kilometers). The points and bars represent point estimates of the number of distinct genetic clusters as a function of area with 95% CIs. *C*, Estimated number of clusters in function of number of sampled NiV sequences for 6 regions in South and Southeast Asia. The dots represent the current number of observed samples and clusters in each region, the solid lines represent interpolated values based on observed data, and the dashed lines represent predicted values if additional sampling was conducted. Abbreviations: BGD, Bangladesh; KHM, Cambodia; NiV, Nipah virus; THA, Thailand.

Table 1. Estimations of the Number of Genetic Clusters for the 6 Countries Represented in Our Dataset, According to Different Estimates of *Pteropus* Range

	IUCN *Pteropus* Range		Tree Cover (10% Threshold)		Tree Cover (30% Threshold)		Tree Cover (50% Threshold)	
Country	Area, km^2	No. of Clusters (95% CI)	Area, km^2	No. of Clusters (95% CI)	Area, km^2	No. of Clusters (95% CI)	Area, km^2	No. of Clusters (95% CI)
India	2 987 747	55 (24–89)	490 910	23 (10–37)	388 304	21 (9–34)	304 653	19 (8–31)
Bangladesh	127 346	15 (6–25)	26 607	11 (4–20)	19 390	11 (4–20)	15 229	10 (4–19)
Thailand	138 538	15 (6–26)	219 195	17 (7–29)	199 624	17 (7–28)	179 669	16 (7–27)
Cambodia	62 957	13 (5–22)	101 985	14 (6–24)	88 099	14 (6–24)	74 823	13 (5–23)
Malaysia	323 929	20 (8–32)	297 861	19 (8–31)	294 306	19 (8–31)	289 772	19 (8–31)
Indonesia	1 656 164	40 (17–63)	1 650 987	39 (17–63)	1 606 412	39 (17–62)	1 550 634	38 (17–61)

Table shows estimations of the number of circulating genetic clusters of NiV using the IUCN's estimations of *Pteropus* geographic range and the Global Forest Watch's estimations of tree cover per country above thresholds of 10%, 30%, and 50%, respectively.

Abbreviations: CI, confidence interval; IUCN, International Union for Conservation of Nature; NiV, Nipah virus.

approximately 80%–90% of circulating genetic clusters remain undetected.

Finally, we consider the extent to which genetic diversity has been fully uncovered in the 6 long-term established surveillance sites in the region (3 in Bangladesh, 2 in Thailand, and 1 in Cambodia), and whether there exist predictors of the estimated total number of genetic clusters in any one place. There have been between 14 and 76 sequences obtained in these sites, resulting in the detection of 2–7 different genetic clusters to date (Figure 4*C*). We estimated the number of new lineages that would be detected with additional sampling. This approach assumes equal probability of detection and similar levels of circulation of all clusters within an area. It also assumes stability in the clusters circulating in a location over time. We estimated that all of the circulating genetic clusters have been identified in 5 of the 6 locations (Figure 4*C*, Supplementary Figure 4). The total number of clusters circulating within a subnational division was not significantly associated with the size of the study area or with human population density (Supplementary Figure 9*A* and 9*B*). There was some evidence of an increase in genetic diversity with the percentage of forest cover ($P = .015$) (Supplementary Figure 9*C*).

DISCUSSION

We analyzed NiV sequences, alongside host and location information from multiple countries, to characterize the underlying

genetic diversity of a pathogen that poses a major risk to human health. We found that NiV is strongly spatially structured, with slow viral movement across the region and limited genetic similarity in viruses that circulate in different countries. These findings are consistent with a previous analysis using data from Bangladesh only [31]. The evolutionary time separating viruses sampled in the 2 extremes of the Nipah region (ie, India to Malaysia) was over 140 years, suggesting substantial entrenchment within communities, with greatest diversity observed in the central region. The extent to which the spatially structured nature of bat species, mixing patterns of bats across roosts, and preexisting immunity contribute to these observations remains unclear.

The transmission dynamics of NiV within bat populations, including long-term immunity after infection, remains poorly understood. We found substantial overlap in the spatial footprint of genetic clusters, to the extent that even individual bat roosts host >1 distinct genetic cluster. *Pteropus* bat roost size highly depends on species, typically hosting hundreds to thousands of bats at a time [15, 32–34]. Maintaining a sufficiently large susceptible population to sustain multiple independent transmission chains in populations of this size likely requires long durations of viral shedding, frequent reinfection, or coinfection, as suggested through modeling of bat immune profiles [35]. As movement between bat roosts is common, the wider population across multiple roosts may also facilitate the maintenance of multiple lineages. In support of a key role of roost population size in maintaining diversity, it is notable that the *P lylei* roosts in Cambodia, which had the greatest diversity with the co-circulation of 3 NiV sublineages within each roost, typically have thousands of bats per roost, many more than other locations [15].

The evolutionary separation between NiV lineages has previously been suggested as 1 possible explanation for the differences in the case fatality rate in Bangladesh (~70%) and Malaysia (~40%) [2, 36, 37]. However, it remains difficult to disentangle differences in the virus from human behavior or transmission route differences. Spillover from bats into pigs through the consumption of infected fruit drove the outbreak in Malaysia, whereas date palm sap consumption by humans appears key in Bangladesh [38]. Viral loads and inoculation routes and sites in these 2 transmission modes are likely to be very different, which could affect subsequent mortality. A primate model found increased fatality risk in strains of Bangladeshi origin as compared to Malaysian origin. However, studies on other animal models have provided less conclusive evidence of pathological or transmission differences between the 2 major clades, and the relevance of animal models to the human situation remains unclear [39–42]. Here we found evidence of differences in spillover or disease risk within a lineage. Genotype IB was found in 3 different years in bat roosts in an area of Bangladesh where human spillovers are frequently identified and where date palm sap consumption is common. However, no human cases were linked to this sublineage. Year-by-year variability in spillover risk, linked to temperature, may explain these findings, especially in the context when only a subset of human NiV cases are ever detected and have their viruses sequenced [10, 12, 43].

Using genetic cluster correlation over distance as a spatial signature of NiV, we estimated the spatial footprint of individual clusters and the number of circulating clusters. While crude, these ballpark figures are a useful marker of what we may be missing, especially in countries with little or no sampling. We estimated the presence of approximately 100 genetic clusters across the region, suggesting substantial undetected viral diversity (~80%–90%). Increased sampling across South and Southeast Asia will help uncover additional lineages, particularly in new locations. Long-term surveillance in locations with established sampling remains critical to characterize the evolutionary dynamics of NiV within roosts. Whole-genome sequencing should be prioritized, where possible, to maximize the signal that can be captured through genomic surveillance.

Our observations of the spatial structure of NiV need to be taken in the context of evolving bat ecology. Deforestation and land use changes across the region are leading to a more fragmented pattern of roosts and bringing bats closer to urban environments [7, 10, 44]. It has been shown for the related Hendra virus that stress in bats is linked to spillover risk [45]. It remains unclear how future climate changes and deforestation could lead to further shifts in bat population distribution and behavior, potentially increasing stress and spillover risk.

NiV remains a major threat to human health. Our findings have implications for monitoring and understanding this ongoing threat. We show that there is substantial undetected diversity in the region, with the potential for heterogeneity in risk of person-to-person transmission and disease severity across strains. Detection of changes in these characteristics will require investment in focused and stable surveillance, particularly outside Bangladesh. This is of particular importance since, outside Bangladesh, access to NiV testing, physician familiarity with the disease, and NiV disease surveillance are very limited [46]. Moreover, future changes in land use may heighten human exposure to NiV across the region.

This study is subject to some limitations. First, we rely on the few places where sequencing is conducted, which represent a small number of locations in 6 countries. While we used methods that minimize the impact of sampling biases, potential differences in NiV ecology in unsampled locations cannot be discounted. We also could not link sequences to specific bats. The most common sampling method requires the collection of bat urine under roosts using plastic or tarpaulin sheets that are polymerase chain reaction (PCR) tested, making it impossible to link urine to an individual bat and to ensure that multiple positive samples do not all come from the same bat. This

also means that most available NiV sequences from bats are short sequences from the PCR process. Despite their frequent short nature, we were still able to consistently place sequences in different clades. In particular, in Cambodia, where most sequences were very short (<400 nucleotides in length), we were able to identify multiple clades circulating within the same roosts. The classification of tips into genetic clusters necessarily relies on thresholds of evolutionary distance. However, in sensitivity analyses, changes in these thresholds resulted in minimal changes to the overall inferences on the genetic diversity of viruses circulating within any area. Finally, NiV sequences are ultimately reliant on the PCR primers used to detect the virus in the original sample. If the PCR primers are overly specific, they may systematically miss some viruses [47]. Future efforts may want to consider using broader primer sets.

This project has demonstrated that even sparsely sampled genetic data, including many short sequences, from large areas can provide meaningful characterization of underlying diversity in populations when considered together. We have shown that even individual roosts typically have multiple circulating transmission chains but with each genetic lineage covering a large spatial footprint, probably driven by bat mobility patterns. While most NiV diversity remains undetected, the surveillance sites that have been established appear to have uncovered a substantial proportion of the diversity in those locations.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Author contributions. Conceptualization: H. S. and O. C.-A. Methodology: H. S., O. C.-A., N. L., B. N., and C. M. Investigation: O. C.-A. and B. N. Visualization: O. C.-A. Supervision: H. S. and V. D. Writing—original draft and writing—review and editing: All authors.

Data availability. Data and code used in this analysis are available in a GitHub repository: https://github.com/ocortaz/nipah_genetic_diversity. GenBank accession numbers are provided in Supplementary Table 1, including new accession numbers for sequences not previously published. Novel sequences are not currently publicly accessible on GenBank, but they will be upon acceptance and/or request of the reviewers.

Disclaimer. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the US government.

Financial support. This work was supported by the European Research Council (grant number 804744), the Coalition for Epidemic Preparedness Innovations, the National Institutes of Health (grant numbers NIH R01 AI168287-01A1, R01AI160780, and U01AI168287), the European Commission Innovate program (ComAcross project, grant number DCI-ASIE/2013/315–047), and the Defense Advanced Research Projects Agency PREEMPT program (cooperative agreement D18AC00031).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. World Health Organization. Nipah research and development (R&D) roadmap. **2019**. https://cdn.who.int/media/docs/default-source/blue-print/nipah_rdblueprint_roadmap_advanceddraftoct2019.pdf?sfvrsn=4f0dc9ad_3&download=true. Accessed 10 July 2023.
2. Nikolay B, Salje H, Hossain MJ, et al. Transmission of Nipah virus—14 years of investigations in Bangladesh. N Engl J Med **2019**; 380:1804–14.
3. Plowright RK, Becker DJ, Crowley DE, et al. Prioritizing surveillance of Nipah virus in India. PLoS Negl Trop Dis **2019**; 13:e0007393.
4. Soman Pillai V, Krishna G, Valiya Veettil M. Nipah virus: past outbreaks and future containment. Viruses **2020**; 12: 465.
5. Lo MK, Lowe L, Hummel KB, et al. Characterization of Nipah virus from outbreaks in Bangladesh, 2008–2010. Emerg Infect Dis **2012**; 18:248–55.
6. Halpin K, Hyatt AD, Fogarty R, et al. Pteropid bats are confirmed as the reservoir hosts of henipaviruses: a comprehensive experimental study of virus transmission. Am J Trop Med Hyg **2011**; 85:946–51.
7. Chua KB, Chua BH, Wang CW. Anthropogenic deforestation, El Niño and the emergence of Nipah virus in Malaysia. Malays J Pathol **2002**; 24:15–21.
8. Pulliam JRC, Epstein JH, Dushoff J, et al. Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. J R Soc Interface **2012**; 9:89–101.
9. Ching PKG, de los Reyes VC, Sucaldito MN, et al. Outbreak of henipavirus infection, Philippines, 2014. Emerg Infect Dis **2015**; 21:328–31.
10. McKee CD, Islam A, Luby SP, et al. The ecology of Nipah virus in Bangladesh: a nexus of land-use change and opportunistic feeding behavior in bats. Viruses **2021**; 13:169.

11. Epstein JH, Anthony SJ, Islam A, et al. Nipah virus dynamics in bats and implications for spillover to humans. Proc Natl Acad Sci U S A **2020**; 117:29190–201.
12. Hegde ST, Salje H, Sazzad HMS, et al. Using healthcare-seeking behaviour to estimate the number of Nipah outbreaks missed by hospital-based surveillance in Bangladesh. Int J Epidemiol **2019**; 48:1219–27.
13. Reynes J-M, Counor D, Ong S, et al. Nipah virus in Lyle's flying foxes, Cambodia. Emerg Infect Dis **2005**; 11:1042–7.
14. Rahman MZ, Islam MM, Hossain ME, et al. Genetic diversity of Nipah virus in Bangladesh. Int J Infect Dis **2021**; 102:144–51.
15. Cappelle J, Hoem T, Hul V, et al. Nipah virus circulation at human-bat interfaces, Cambodia. Bull World Health Organ **2020**; 98:539–47.
16. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res **2016**; 44(D1):D67–72.
17. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA x: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol **2018**; 35:1547–9.
18. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **2015**; 32:268–74.
19. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods **2017**; 14:587–9.
20. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol **2012**; 29: 2157–67.
21. Baele G, Li WL, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol **2013**; 30:239–43.
22. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol **2018**; 4:vey016.
23. Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. Phylogenetic clustering by linear integer programming (PhyCLIP). Mol Biol Evol **2019**; 36:1580–95.
24. Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods Ecol Evol **2016**; 7:1451–6.
25. Chao A, Gotelli NJ, Hsieh TC, et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr **2014**; 84:45–67.
26. Duchene S, Lemey P, Stadler T, et al. Bayesian evaluation of temporal signal in measurably evolving populations. Mol Biol Evol **2020**; 37:3363–79.
27. Lo Presti A, Cella E, Giovanetti M, et al. Origin and evolution of Nipah virus. J Med Virol **2016**; 88:380–8.
28. Hubert L, Arabie P. Comparing partitions. J Classification **1985**; 2:193–218.
29. International Union for Conservation of Nature. The IUCN red list of threatened species. **2022**. https://www.iucnredlist.org. Accessed 10 July 2023.
30. Hansen MC, Potapov PV, Moore R, et al. High-resolution global maps of 21st-century forest cover change. Science **2013**; 342:850–3.
31. Whitmer SLM, Lo MK, Sazzad HMS, et al. Inference of Nipah virus evolution, 1999–2015. Virus Evol **2021**; 7:veaa062.
32. Hahn MB, Epstein JH, Gurley ES, et al. Roosting behaviour and habitat selection of *Pteropus giganteus* reveals potential links to Nipah virus epidemiology. J Appl Ecol **2014**; 51: 376–87.
33. Chaiyes A, Duengkae P, Wacharapluesadee S, Pongpattananurak N, Olival KJ, Hemachudha T. Assessing the distribution, roosting site characteristics, and population of *Pteropus lylei* in Thailand. Raffles Bull Zool **2017**; 65: 670–80.
34. Ravon S, Furey NM, Vibol HUL, Cappelle J. A rapid assessment of flying fox (*Pteropus* spp.) colonies in Cambodia. http://www.seabcru.org/wp-content/uploads/2014/09/Ravon-et-al.-2014.-Cambodian-Pteropus.pdf. Accessed 15 May 2023.
35. Glennon EE, Becker DJ, Peel AJ, et al. What is stirring in the reservoir? Modelling mechanisms of henipavirus circulation in fruit bat hosts. Philos Trans R Soc Lond B Biol Sci **2019**; 374:20190021.
36. Kasloff SB, Leung A, Pickering BS, et al. Pathogenicity of Nipah henipavirus Bangladesh in a swine host. Sci Rep **2019**; 9:5230.
37. Kenmoe S, Demanou M, Bigna JJ, et al. Case fatality rate and risk factors for Nipah virus encephalitis: a systematic review and meta-analysis. J Clin Virol **2019**; 117:19–26.
38. Gurley ES, Hegde ST, Hossain K, et al. Convergence of humans, bats, trees, and culture in Nipah virus transmission, Bangladesh. Emerg Infect Dis **2017**; 23:1446–53.
39. Gaudino M, Aurine N, Dumont C, et al. High pathogenicity of Nipah virus from *Pteropus lylei* fruit bats, Cambodia. Emerg Infect Dis **2020**; 26:104–13.
40. de Wit E, Munster VJ. Animal models of disease shed light on Nipah virus pathogenesis and transmission. J Pathol **2015**; 235:196–205.
41. Hegde ST, Lee KH, Styczynski A, et al. Potential for person-to-person transmission of henipaviruses: a systematic review of the literature. J Infect Dis **2024**; 229:733–42.
42. Mire CE, Satterfield BA, Geisbert JB, et al. Pathogenic differences between Nipah virus Bangladesh and Malaysia strains in primates: implications for antibody therapy. Sci Rep **2016**; 6:30916.
43. Cortes MC, Cauchemez S, Lefrancq N, et al. Characterization of the spatial and temporal distribution

of Nipah virus spillover events in Bangladesh, 2007–2013. J Infect Dis **2018**; 217:1390–4.
44. Chaiyes A, Duengkae P, Suksavate W, et al. Mapping risk of Nipah virus transmission from bats to humans in Thailand. Ecohealth **2022**; 19:175–89.
45. Eby P, Peel AJ, Hoegh A, et al. Pathogen spillover driven by rapid changes in bat ecology. Nature **2023**; 613:340–4.
46. Satter SM, Aquib WR, Sultana S, et al. Tackling a global epidemic threat: Nipah surveillance in Bangladesh, 2006–2021. PLoS Negl Trop Dis **2023**; 17: e0011617.
47. Annand EJ, Horsburgh BA, Xu K, et al. Novel Hendra virus variant detected by sentinel surveillance of horses in Australia. Emerg Infect Dis **2022**; 28:693–704.