Predicting Reservoir Hosts and Arthropod Vectors from Evolutionary Signatures in RNA Virus Genomes*

Authors: Simon A. Babayan[1,2], Richard J. Orton[3] and Daniel G. Streicker[1,3†]

Affiliations:

[1] Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, G12 8QQ, Scotland, UK

[2] The Moredun Research Institute, Pentlands Science Park, EH26 0PZ, Scotland, UK

[3] MRC-University of Glasgow Centre for Virus Research, Glasgow, G61 1QH, Scotland, UK

† Correspondence to: daniel.streicker@glasgow.ac.uk

Abstract

Identifying the animal origins of RNA viruses requires years of field and laboratory studies that stall responses to emerging infectious diseases. Using large genomic and ecological datasets, we demonstrate that the animal reservoirs and the existence and identity of arthropod vectors can be predicted directly from viral genome sequences using machine learning. We illustrate the ability of these models to predict the epidemiology of diverse viruses across most human-infective families of single-stranded RNA viruses, including 69 viruses with previously elusive or never-investigated reservoirs or vectors. Models such as these, which capitalize on the proliferation of low-cost genomic sequencing, can narrow the time lag between virus discovery and targeted research, surveillance and management.

One Sentence Summary: The natural hosts of RNA viruses can be predicted directly from their genome sequences.

Main text:

Preventing emerging viral infections including Ebola, SARS, and Zika requires identifying which reservoir hosts and/or blood-feeding arthropod vectors perpetuate viruses in nature. Current practice requires combining evidence from field surveillance, phylogenetics, laboratory experiments, and real-world interventions, but is time consuming and often inconclusive (*1*). This creates prolonged periods of uncertainty that may amplify economic and health losses. We aimed to develop a general model to predict reservoir hosts and arthropod vectors across single-stranded RNA (ssRNA) viruses, the viral group most commonly implicated in zoonotic disease outbreaks (*2*), building on the modern expansion of low-cost viral sequence data (*3*).

We collected a single representative genome sequence per viral species or strain from twelve taxonomic groups (11 families and 1 order) of ssRNA viruses that can infect humans; 80% of all human-infective groups (Fig. 1A). For each virus, we used extensive literature searches to determine currently-accepted reservoir hosts (437 viruses; 11 reservoir groups), whether transmission involves an arthropod vector (527 viruses) and if so, the identity of arthropod vectors (98 viruses; 4 vector groups). To maximize predictive scope reservoir and vector groups included the most frequent sources of emerging human viruses as well as other common hosts in human-infective viral families (e.g., fish, plants and insects) (*2*, *4*).

Because related viruses often have closely-related hosts due to co-speciation and preferential host switching among related host species, we designed an algorithm to predict host associations from viral phylogenetic relatedness (*5*, *6*). This phylogenetic neighborhood (PN) model identified the reservoir hosts of 58.1 ± 0.07% (standard deviation) of viruses, whether or not viruses were transmitted by an arthropod vector (95% ± 0.24) and the vector identity of arthropod-borne viruses (67.2 ± 0.12%). Biases in viral genome composition can also inform host-virus associations. Specifically, viral codon pair and dinucleotide biases are

58 reported to mimic those of their hosts, representing either a genome-wide strategy for
59 adaptation to specific host groups or genomic imprinting by the host cellular machinery that
60 viruses co-opt for replication (*7*). Irrespective, genomic biases can coarsely discriminate
61 viruses from different host groups within several well-studied viral families (*8–10*). However,
62 whether genomic biases can predict hosts from smaller or less-studied groups of viruses
63 remains unresolved (*11*). We quantified 4229 traits from the 536 viral genomes in our
64 dataset, including all possible codon pair, dinucleotide, codon, and amino acid biases (*6*)(Fig.
65 S1). When all traits were weighted equally, dissimilarity-based clustering grouped viruses
66 predominately by viral taxonomy; however, paraphyly of most viral groups implied selective
67 forces on viral genomic biases that outweighed phylogenetic history (Fig. 1B,C). Generalized
68 linear mixed models further revealed that even after controlling for effects of viral taxonomy,
69 some genomic biases of viruses were correlated with their reservoir and vector associations,
70 suggesting host effects on viral genomes that transcend viral groups (Figs. S2–S7). We
71 hypothesized that combining host-associated genomic biases with viral PNs could maximize
72 prediction of reservoirs and vectors from viral sequence data.
73 We addressed this challenge using supervised machine learning, a class of statistical
74 models that can integrate multiple traits that carry weak signal in isolation, but build a strong
75 signal when optimally weighted (*12*). Gradient boosting machines (GBM, *13*) outperformed
76 seven alternative classifiers in predicting host associations from viral genomic biases and
77 identified the most informative genomic traits for each aspect of viral ecology (Figs. S8–
78 S12). GBMs combining selected genomic traits (SelGen) with viral PNs predicted reservoir
79 hosts with up to 83.5% accuracy, distinguishing all eleven reservoir groups, including
80 taxonomic divisions within the birds (i.e., Neoaves versus Galloanserae) and bats [i.e.,
81 Pteropodiformes ("Pterobat") versus Vespertilioniformes ("Vespbat")] (Fig. 2A). Reservoirs
82 of arthropod-borne and non-arthropod-borne viruses were predicted equally well (χ^2 test, p =
83 0.5). Averaging predictions across observations of each virus in models trained on different

84 data subsets (i.e., 'bagging') improved prediction of most reservoir groups, such that the
85 reservoirs of 71.9% of all viruses in the study were correctly assigned. GBMs lacking PN or
86 SelGen misclassified the reservoirs of 33 and 22 more viruses, respectively (Fig. 2B,C).
87 We trained two further sets of models that focused on arthropod-borne transmission (*6*).
88 The first nearly perfectly identified which viruses were transmitted by arthropod vectors.
89 Combined GBMs were most accurate overall (bagged accuracy = 97.0%, Fig. 2D, Fig. S11).
90 Only 5 out of 427 viruses were misclassified by all three GBMs (PN, SelGen and combined),
91 potentially reflecting uncertainty in some currently-accepted transmission routes
92 (Supplementary Text). The second set of models distinguished transmission by all four vector
93 classes (bagged accuracy = 90.8%; Fig. 2E,F). Ranking traits according to their predictive
94 power showed that midge and sandfly vectors were identified predominately from genomic
95 biases, while mosquito and tick vectors were strongly correlated with viral phylogeny (Fig.
96 S12). Accuracy declined by 9.2 and 2.0 percentage points for GBMs lacking SelGen or PN
97 (Fig. 2G). Thus, while phylogeny and genome-wide biases are partially correlated, algorithms
98 successfully exploited independent information in each for all three prediction types.
99 All models misclassified some currently-accepted hosts. We therefore analyzed
100 whether attributes of predictions could help assess their veracity. Predictions with higher
101 GBM probability ("bagged prediction strength", BPS) were correct more often than those
102 diffused across multiple host groups (Fig. S13A–C). Furthermore, when models misclassified
103 viruses, the true host was most often the second-ranked prediction, such that study-wide
104 accuracy for reservoir and vector prediction rose to 81% and 95.9% respectively when
105 considering the top two most plausible predictions (Fig. 2C,G, Fig. S13D,E). Consequently,
106 BPS provides a confidence metric, such that weaker predictions imply alternative hosts
107 should be considered in order of their relative support.
108 We next used our trained models to predict the natural epidemiology of viruses with
109 previously unknown hosts (hereafter "orphan" viruses). As expected from the accuracy of our

models on viruses with known hosts, model-projected reservoirs and vectors often matched those suspected from epidemiological investigations (Fig. 3, Figs. S14–S16). For example, we predicted an artiodactyl reservoir for human enteric coronavirus 4408, a suspected spillover infection from cows into humans; a primate reservoir of O'nyong-nyong virus, for which humans are the presumed reservoir; and that outbreaks of Tembusu virus in domestic ducks follow cross-species transmission from wild Neoaves (*14–16*). Other results pointed to unexpected reservoirs. For example, all four orphan ebolaviruses had greater support for the commonly-accepted Pterobat (suborder Pteropodiformes) than for Vespbat reservoirs, but surprisingly, Bundibugyo and Tai Forest ebolaviruses had equal or stronger support for primate reservoirs. This indicates that signals learned from primate viruses from divergent viral families occurred in these ebolavirus genomes. Neither of species of ebolavirus has been detected in bats (*17*) and the slow evolution of genomic biases in Filoviruses implied that the observed signal could not have evolved during short chains of transmission in primates (Fig. S17). The possibility of an undiscovered primate ebolavirus reservoir therefore deserves empirical validation. For viruses without conjectured reservoirs or vectors, we generate candidates for prioritized surveillance. For example, Bas-Congo virus caused an outbreak of hemorrhagic fever in the Democratic Republic of Congo and was detected in humans only (*18*). Our models predicted an Artiodactyl reservoir, a high probability of arthropod-borne transmission, and midges as the likely vector of this emerging disease (Fig. 3A,C). Such predictions may ultimately support earlier interventions targeting appropriate reservoirs or vectors that interrupt the critical early phases of outbreaks or limit future re-emergence. Likewise, our models can provide ecological insights for virus discovery programs (Fig. 3B).

By virtue of using slowly-evolving biases spread across viral genomes, our models predict taxa that maintain long-term viral circulation rather than "bridge hosts" that sustain insufficient chains of transmission to imprint evolutionary signals in viral genomes (e.g., pig hosts of bat-borne Nipah virus). Similarly, sustained transmission by divergent hosts may

create conflicting signals that obscure model predictions (Supplementary Text). Finally, models predict only the reservoir and vector groups used for training and will erroneously assign a host from these same categories if applied to viruses from host groups that were too rare include (Fig. S18). As virus discoveries expand databases, evaluating predictive accuracy for additional host groups will be an important improvement.

In conclusion, we created a machine learning framework that leverages traits from individual viruses with network-derived information from their relatives to predict: (i) the reservoir hosts of twelve key groups of RNA viruses, (ii) whether their transmission involves an arthropod vector and (iii) the identity of that vector. Our models make these predictions, supply quantitative measures of confidence, and provide relative support for alternatives from single genome sequences, with no requirement for experiments, longitudinal surveillance, or genomes of candidate reservoirs or vectors. As viral genomes are now produced within hours of detection (*19*), algorithms that rapidly generate field-testable hypotheses from sequence data narrow the gap between virus discovery and actionable understanding of virus ecology.

References and Notes

1. M. Viana *et al*., Assembling evidence for identifying reservoirs of infection. *Trends Ecol. Evol.* **29**, 270–279 (2014).
2. M. Woolhouse, E. Gaunt, Ecological origins of novel human pathogens. *Crit. Rev. Microbiol.* **33**, 231–242 (2007).
3. R. R. Kao, D. T. Haydon, S. J. Lycett, P. R. Murcia, Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol.* **22**, 282–291 (2014).
4. K. J. Olival *et al*., Host and viral traits predict zoonotic spillover from mammals. *Nature*. **546**, 646–650 (2017).
5. J. L. Geoghegan, S. Duchêne, E. C. Holmes, Comparative analysis estimates the

relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog*. **13**, e1006215 (2017).

6. Materials and methods are available as supplementary materials on Science Online.

7. M. A. Martínez *et al*., Synonymous virus genome recoding as a tool to impact viral fitness. *Trends Microbiol*. **24**, 134–147 (2016).

8. B. D. Greenbaum, A. J. Levine, G. Bhanot, R. Rabadan, Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog*. **4**, e1000079 1–9 (2008).

9. F. P. Lobo *et al*., Virus-host coevolution: Common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS One*. **4**, e6282 (2009).

10. A. Kapoor, P. Simmonds, W. I. Lipkin, S. Zaidi, E. Delwart, Use of nucleotide composition analysis to infer hosts for three novel Picorna-like viruses. *J. Virol*. **84**, 10322–10328 (2010).

11. F. Di Giallonardo, T. E. Schlub, M. Shi, E. C. Holmes, Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. *J. Virol*. **91**, e02381-16 (2017).

12. M. K. K. Leung, A. Delong, B. Alipanahi, B. J. Frey, Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE*. **104**, 176–197 (2016).

13. B. J. H. Friedman, Greedy function aproximation: A gradient boosting machine. *Ann. Stat*. **29**, 1189–1232 (2001).

14. M. G. Han, D.-S. Cheon, X. Zhang, L. J. Saif, Cross-protection against a human enteric coronavirus and a virulent bovine enteric Coronavirus in gnotobiotic calves. *J. Virol*. **80**, 12350–12356 (2006).

15. A. D. LaBeaud *et al*., High rates of O'nyong nyong and Chikungunya virus transmission in coastal Kenya. *PLoS Negl. Trop. Dis*. **9**, e0003436 (2015).

16. Y. Tang *et al*., Characterization of a Tembusu virus isolated from naturally infected House sparrows (*Passer domesticus*) in northern China. *Transbound. Emerg. Dis.* **60**, 152–158 (2013).

17. K. J. Olival, D. T. S. Hayman, Filoviruses in bats: current knowledge and future directions. *Viruses*. **6**, 1759–1788 (2014).

18. G. Grard *et al*., A novel Rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog*. **8**, e1002924 (2012).

19. J. Quick *et al*., Real-time, portable genome sequencing for Ebola surveillance. *Nature*. **530**, 228 (2016).

20. J. M. Hutcheon, J. A. W. Kirsch, A moveable face: deconstructing the Microchiroptera and a new classification of extant bats. *Acta Chiropterologica*. **8**, 1–10 (2006).

21. D. Kunec, N. Osterrieder, Codon pair bias is a direct consequence of dinucleotide bias. *Cell Rep*. **14**, 55–67 (2016).

22. J. R. Coleman *et al*., Virus attenuation by genome-scale changes in codon pair bias. *Science*. **320**, 1784–1787 (2008).

23. F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classif*. **31**, 274–295 (2014).

24. L. J. Harmon, J. T. Weir, C. D. Brock, R. E. Glor, W. Challenger, GEIGER: investigating evolutionary radiations. *Bioinformatics*. **24**, 129–131 (2007).

25. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*. **57**, 289–300 (1995).

26. S. F. Altschul *et al*., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. **25** (1997), pp. 3389–3402.

27. E. Paradis, *Analysis of Phylogenetics and Evolution with R* (Springer Science & Business Media, 2011).

28. D. Charif, J. R. Lobry, in *Structural approaches to sequence evolution* (Springer,

2007), pp. 207–232.

29. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794 (2016).

30. The H20.ai team, h2o: R interface for H2O (2017), (available at https://github.com/h2oai/h2o-3).

31. R Core Team, R: A language and environment for statistical computing (2017), (available at https://www.r-project.org/).

32. G. Dudas *et al*., Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*. **544**, 309 (2017).

33. A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, G. M. Smith, *Mixed effects models and extensions in ecology with R* (Springer Verlag, 2009).

34. F. Tulloch, N. J. Atkinson, D. J. Evans, M. D. Ryan, P. Simmonds, RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG / UpA dinucleotide frequencies. *Elife*. **3**, 1–15 (2014).

35. M. A. Takata *et al*., CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*. **550**, 124–127 (2017).

36. Z. A. Memish *et al*., Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg. Infect. Dis*. **19**, 1819 (2013).

37. Y. Yang *et al*., Receptor usage and cell entry of bat coronavirus HKU4 provide insight into bat-to-human transmission of MERS coronavirus. *Proc. Natl. Acad. Sci*. **111**, 12516–12521 (2014).

38. Q. Wang *et al*., Bat origins of MERS-CoV supported by bat coronavirus HKU4 usage of human receptor CD26. *Cell Host Microbe*. **16**, 328–337 (2014).

39. M. A. Müller *et al*., MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983–1997. *Emerg. Infect. Dis*. **20**, 2093 (2014).

40. E. I. Azhar *et al*., Evidence for camel-to-human transmission of MERS Coronavirus. *N. Engl. J. Med.* **370**, 2499–2505 (2014).

41. H. A. Mohd, J. A. Al-Tawfiq, Z. A. Memish, Middle East Respiratory Syndrome Coronavirus (MERS-CoV) origin and animal reservoir. *Virol. J.* **13**, 87 (2016).

42. G. Dudas, A. Rambaut, MERS-CoV recombination: implications about the reservoir and potential for adaptation. *Virus Evol.* **2** (2016).

Acknowledgements: We thank R. Biek, B. Mable, M. Viana, D. Haydon, P. Johnson, S. Altizer, B. Brennan, A. Szemiel, M. Palmarini and three anonymous reviewers for helpful feedback. **Funding:** S.B. was supported by a Glasgow University Research Fellowship and the BBSRC (BB/M012956/1). D.S. was supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and Royal Society (102507/Z/13/Z). Additional funding was provided from the Medical Research Council (MC_UU_12014/12). **Author contributions:** D.S. conceived the research; D.S. and R.O. collected the data; D.S., R.O., and S.B. analyzed the data; and D.S. and S.B. wrote and revised the manuscript. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** Data and code reported in this paper are available at https://github.com/DanielStreicker/ViralHostPredictor

Figures

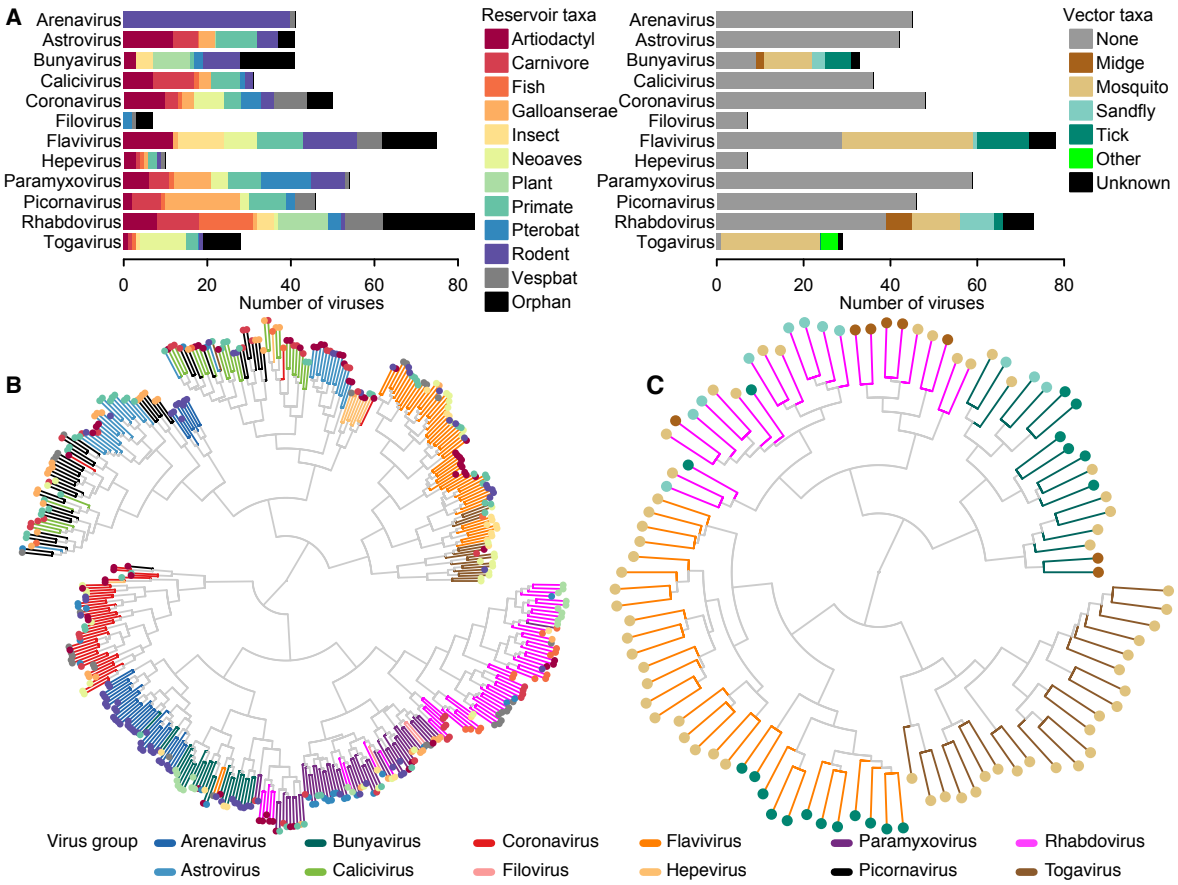
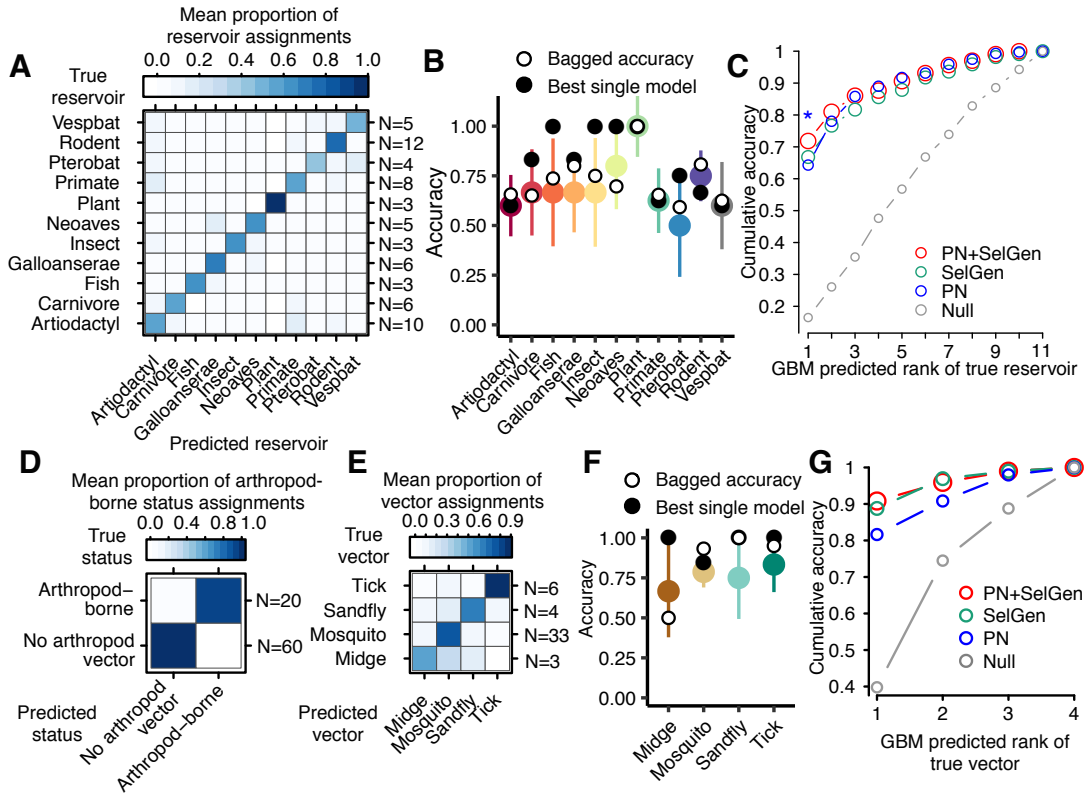


Fig. 1. Distribution and hierarchical clustering of reservoir host and arthropod vector associations across viral taxonomic groups. (A) Barplots show the number of viruses in the dataset from each reservoir host and vector class and the number of orphan viruses in each viral group. The order Artiodactyla (even-toed ungulates) includes the Bovidae, Camelidae, Suidae, Antilocapridae, and Giraffidae families. Galloanserae (ducks, fowl) and Neoaves (most other modern birds) are superorders within the class Aves (birds). (B,C) Dendrograms of 437 viruses with known reservoir hosts and 98 viruses with known arthropod vectors, estimated by hierarchically clustering 4229 genomic biases calculated from viral genomes. Colors of tip symbols indicate reservoir or vectors associations. Branch colors show viral taxonomic groups. Branch lengths are log(n+1) transformed for visualization. (B) Trait models with true viral taxonomic group associations were favored over those with randomly shuffled viral groups (ΔAIC = -1690.6) but also clustered significantly by reservoir (ΔAIC =

-540.7). (C) Arboviruses clustered by both viral taxonomy (ΔAIC = -238.1) and vector group (ΔAIC = -61.5). ΔAIC values are from models comparing true associations to the mean AIC from 500 tip trait randomizations.



295

296 **Fig. 2.** Accurate genomic prediction of viral ecology using machine learning. (A) Heatmap
297 showing the proportion of accurate (diagonal) and misclassified (off diagonal) predictions
298 within each reservoir host class, averaged across GBMs trained and optimized on different
299 subsets of 372 viruses. Row numbers indicate the number of viruses per reservoir in each
300 validation set (N = 65 viruses). (B) The distributions of per reservoir accuracies in single
301 validation sets (colorful points and lines are median and SD) and after bagging (white points).
302 Black points show the best single model. (C) Cumulative bagged accuracy across GBMs
303 using PN and SelGen traits in isolation and in combination. The x-axis shows the rank of the
304 true reservoir (i.e., 1 = true reservoir was the top prediction; 2 = true reservoir was the
305 second-ranked prediction and so on). The y-axis shows accuracy when considering increasing
306 numbers of predictions as plausible. The asterisk indicates significantly higher accuracy in
307 the combined model (χ^2 test: $p < 0.05$). Cumulative null model accuracy was estimated by
308 training GBMs on 50 randomly generated traits that were simulated from normal
309 distributions ranging from 0 to 2 and randomly assigned to viruses. (D,E) Heatmaps showing

the average proportion of accurate predictions of arthropod-borne status and vector identity (N = 80 and 46 viruses per validation set, respectively). (F) Distributions of per vector accuracies as in B. (G) Cumulative bagged accuracy in vector prediction across models as in C.

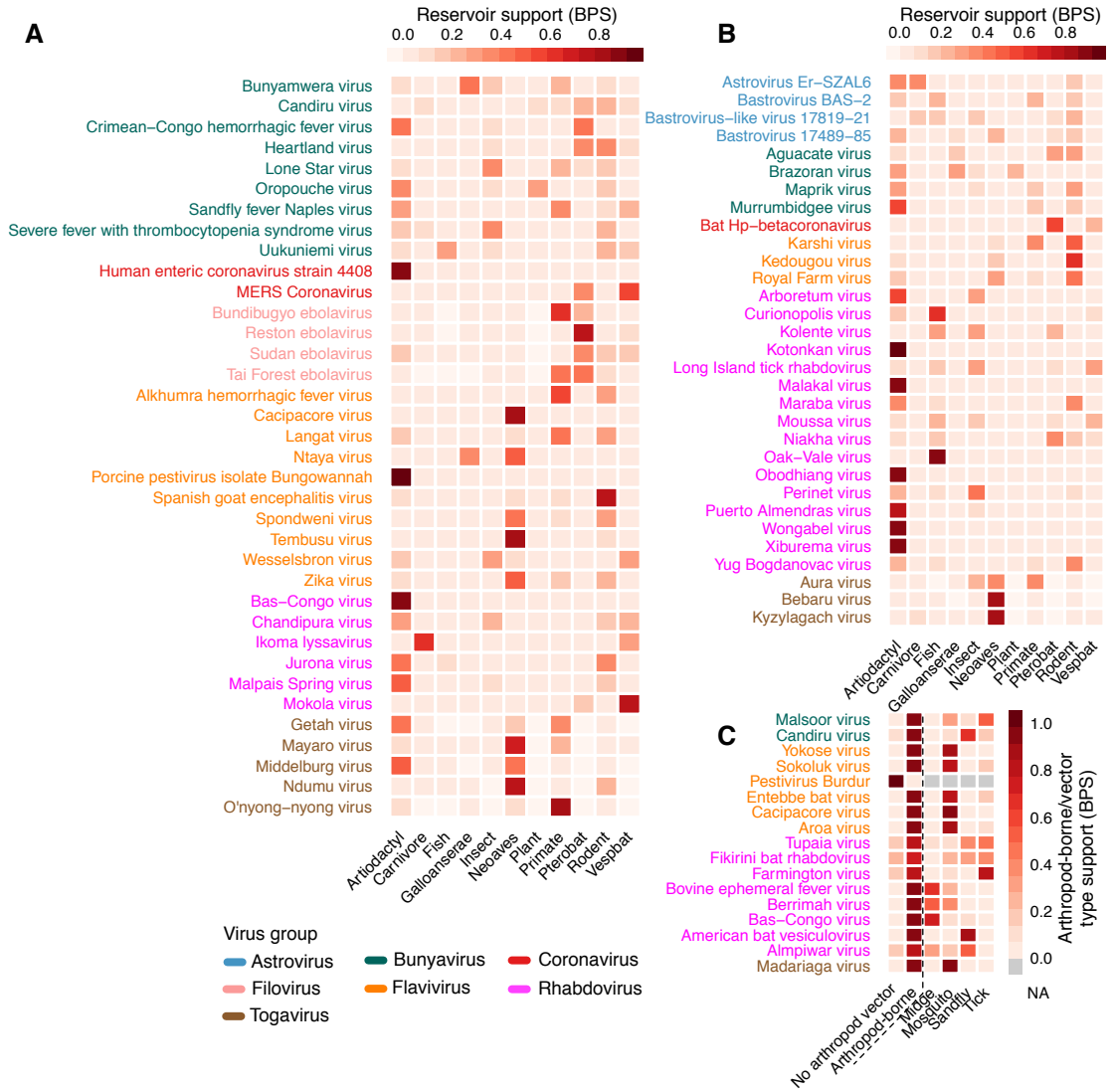


Fig. 3. Reservoir hosts and arthropod vectors of orphan viruses predicted from their genome sequences. (A) Predicted reservoirs for 36 viruses that emerged from unknown sources. (B) 31 viruses discovered by active surveillance of wildlife or blood-feeding arthropods. (C) Predictions of arthropod-borne status for 17 viruses (left of dashed line) and vector identities (last 4 columns, when applicable). Color gradients show the BPS for each class from the top 25% models from each set of GBMs. Figs. S14–S16 show the full probability distributions of predictions.

Supplementary Materials:

Materials and Methods

Supporting Text

Figs. S1–S18

References (20–43)

Appendix S1

Data S1