EPIDEMIOLOGY

Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes

Simon A. Babayan[1,2], Richard J. Orton[3], Daniel G. Streicker[1,3]*

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 animal reservoirs and the existence and identity of arthropod vectors can be predicted directly from viral genome sequences via 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.

Preventing emerging viral infections—including Ebola, SARS, and Zika—requires identification of reservoir hosts and/or blood-feeding arthropod vectors that 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 12 taxonomic groups (11 families and one order) of ssRNA viruses that can infect humans; that is, 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, four 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 owing to cospeciation and preferential host switching among related host species, we first designed an algorithm to predict host associations from viral phylogenetic relatedness alone (*5*, *6*). This phylogenetic neighborhood (PN) model identified the reservoir hosts of 58.1 ± 0.07% (±SD) of viruses, whether or not 95 ± 0.24% of viruses were transmitted by an arthropod vector, as well as the vector identity of 67.2 ± 0.12% of arthropod-borne viruses. Biases in viral genome composition can also inform host-virus associations. Specifically, viral codon pair and dinucleotide biases are reported to mimic those of their hosts, representing either a genome-wide strategy for adaptation to specific host groups or genomic imprinting by the host cellular machinery that viruses co-opt for replication (*7*). In any case, genomic biases can coarsely discriminate viruses from different host groups within several well-studied viral families (*8*–*10*). However, whether genomic biases can predict hosts from smaller or less-studied groups of viruses remains unresolved (*11*). We quantified 4229 traits from the 536 viral genomes in our dataset, including all possible codon pair, dinucleotide, codon, and amino acid biases (*6*) (fig. S1). When all traits were weighted equally, dissimilarity-based clustering grouped viruses predominately by viral taxonomy; however, paraphyly of most viral groups implied selective forces on viral genomic biases that outweighed phylogenetic history (Fig. 1, B and C). Generalized linear mixed models further revealed that even after controlling for effects of viral taxonomy, some genomic biases of viruses were correlated with their reservoir and vector associations, suggesting host effects on viral genomes that transcend viral groups (figs. S2 to S7). We hypothesized that combining host-associated genomic biases with viral PNs could maximize prediction of reservoirs and vectors from viral sequence data.

We addressed this challenge by using supervised machine learning, a class of statistical models that can integrate multiple traits that carry a weak signal in isolation but build a strong signal when optimally weighted (*12*). Gradient boosting machines (GBMs) (*13*) outperformed seven alternative classifiers in predicting host associations from viral genomic biases and identified the most informative genomic traits for each aspect of viral ecology (figs. S8 to S12). GBMs combining selected genomic traits (SelGen) with viral PNs predicted reservoir hosts with up to 83.5% accuracy, distinguishing all 11 reservoir groups, including taxonomic divisions within the birds (i.e., Neoaves versus Galloanserae) and bats [i.e., Pteropodiformes ("Pterobat") versus Vespertilioniformes ("Vespbat")] (Fig. 2A). Reservoirs of arthropod-borne and non–arthropod-borne viruses were predicted equally well (χ^2 test, P = 0.5). Averaging predictions across observations of each virus in models trained on different data subsets (i.e., "bagging") improved prediction of most reservoir groups, such that the reservoirs of 71.9% of all viruses in the study were correctly assigned. GBMs lacking PN or SelGen misclassified the reservoirs of 33 and 22 more viruses, respectively (Fig. 2, B and C).

We trained two additional sets of models that focused on arthropod-borne transmission (*6*). The first nearly perfectly identified which viruses were transmitted by arthropod vectors. Combined GBMs were most accurate overall (bagged accuracy = 97.0%) (Fig. 2D and fig. S11). Only 5 out of 527 viruses were misclassified by all three GBMs (PN, SelGen, and combined), potentially reflecting uncertainty in some currently accepted transmission routes (supplementary text). The second set of models distinguished transmission by all four vector classes (bagged accuracy = 90.8%) (Fig. 2, E and F). Ranking traits according to their predictive power showed that midge and sandfly vectors were identified predominately from genomic biases, whereas mosquito and tick vectors were strongly correlated with viral phylogeny (fig. S12). Accuracy declined by 9.2 and 2.0 percentage points for GBMs lacking SelGen or PN (Fig. 2G). Thus, although phylogeny and genome-wide biases are partially correlated, algorithms successfully exploited independent information in each for all three prediction types.

All models misclassified some currently accepted hosts. We therefore analyzed whether attributes of predictions could help assess their veracity. Predictions with higher GBM probability [bagged prediction strength (BPS)] were correct more often than those diffused across multiple host groups (fig. S13, A to C). Furthermore, when models misclassified viruses, the true host was most often the second-ranked prediction, such that study-wide accuracy for reservoir and vector prediction rose to 81 and 95.9%, respectively, when considering the top two predictions as plausible (Fig. 2, C and G, and fig. S13, D and E). Consequently, BPS provides a confidence metric, such that weaker predictions imply that alternative hosts should be considered in order of their relative support.

We next used our trained models to predict the natural epidemiology of viruses with 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 and

[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, Penicuik EH26 0PZ, Scotland, UK. [3]MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, Scotland, UK.
*Corresponding author. Email: daniel.streicker@glasgow.ac.uk

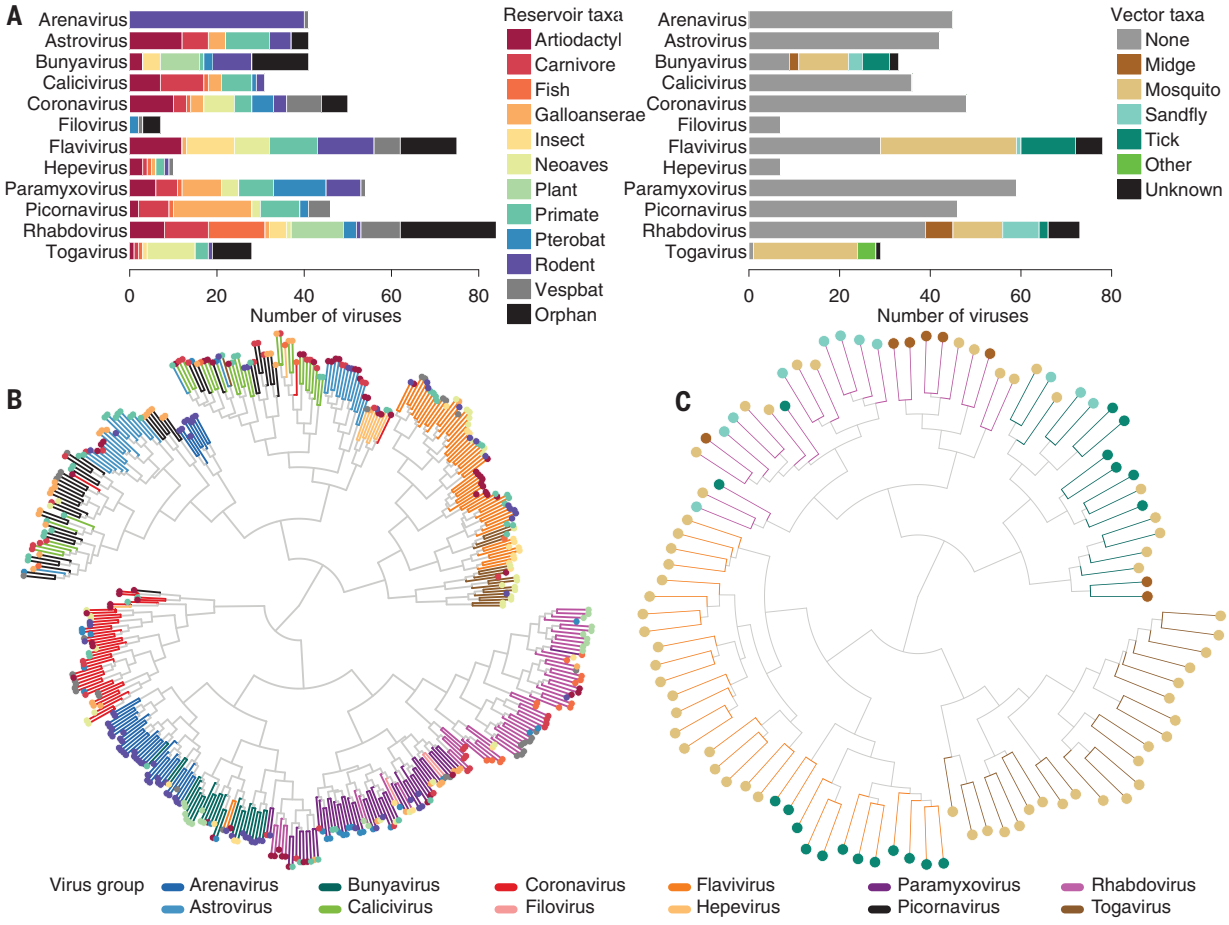

Fig. 1. Distribution and hierarchical clustering of reservoir host and arthropod vector associations across viral taxonomic groups. (**A**) Bar plots show the number of viruses in the dataset from each reservoir host and vector class and the number of orphan viruses in each viral group. The order Artiodactyla (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** and **C**) Dendrograms of 437 viruses with known reservoir hosts and 98 viruses with known arthropod vectors, estimated by hierarchically clustering 4229 biases calculated from viral genomes. Colors of tip symbols indicate reservoir or vector associations. Branch colors denote 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 [change in Akaike information criterion (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.

figs. S14 to 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 (suborder Vespertilioniformes) 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 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 the 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. 3, A and C). Such predictions may ultimately support earlier interventions targeting appropriate reservoirs or vectors that interrupt the critical early phases of outbreaks or limit future reemergence. 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 to include (fig. S18). As virus discoveries expand databases, evaluating predictive accuracy for additional host groups will be an important improvement.

In summary, we created a machine learning framework that leverages traits from individual viruses with network-derived information from

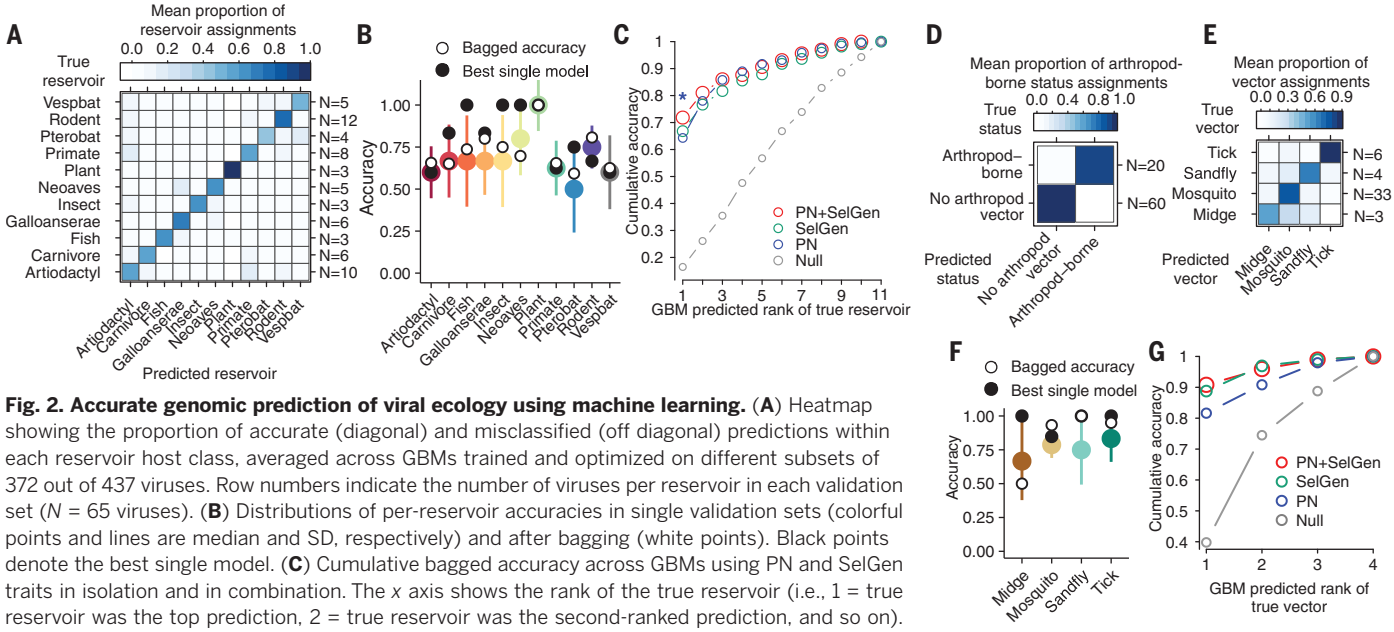

Fig. 2. Accurate genomic prediction of viral ecology using machine learning. (**A**) Heatmap showing the proportion of accurate (diagonal) and misclassified (off diagonal) predictions within each reservoir host class, averaged across GBMs trained and optimized on different subsets of 372 out of 437 viruses. Row numbers indicate the number of viruses per reservoir in each validation set (N = 65 viruses). (**B**) Distributions of per-reservoir accuracies in single validation sets (colorful points and lines are median and SD, respectively) and after bagging (white points). Black points denote the best single model. (**C**) Cumulative bagged accuracy across GBMs using PN and SelGen traits in isolation and in combination. The x axis shows the rank of the true reservoir (i.e., 1 = true reservoir was the top prediction, 2 = true reservoir was the second-ranked prediction, and so on). The y axis shows accuracy when considering increasing numbers of predictions as plausible. The asterisk indicates significantly higher accuracy in the combined model (χ^2 test: $P < 0.05$). Cumulative null model accuracy was estimated by training GBMs on 50 randomly generated traits that were simulated from normal distributions ranging from 0 to 2 and randomly assigned to viruses. (**D** and **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).

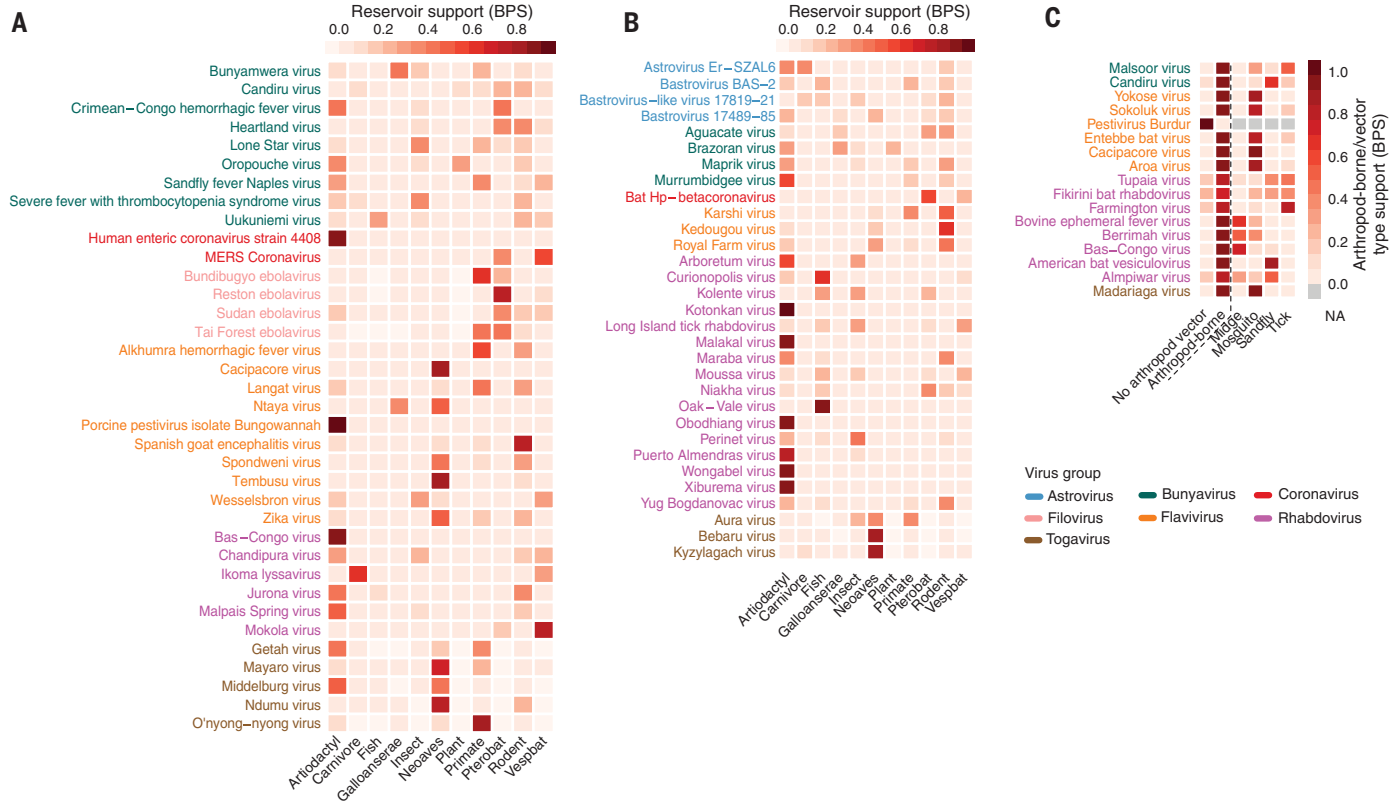

Fig. 3. Reservoir hosts and arthropod vectors of orphan viruses predicted from their genome sequences. (**A**) Predicted reservoirs for 36 viruses that emerged from unknown sources. (**B**) Thirty-one viruses discovered by active surveillance of wildlife or blood-feeding arthropods. (**C**) Predictions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

their relatives to predict: (i) the reservoir hosts of 12 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*., *Trends Ecol. Evol.* **29**, 270–279 (2014).
2. M. Woolhouse, E. Gaunt, *Crit. Rev. Microbiol.* **33**, 231–242 (2007).
3. R. R. Kao, D. T. Haydon, S. J. Lycett, P. R. Murcia, *Trends Microbiol.* **22**, 282–291 (2014).
4. K. J. Olival *et al*., *Nature* **546**, 646–650 (2017).
5. J. L. Geoghegan, S. Duchêne, E. C. Holmes, *PLOS Pathog.* **13**, e1006215 (2017).
6. Materials and methods are available as supplementary materials.
7. M. A. Martínez, A. Jordan-Paiz, S. Franco, M. Nevot, *Trends Microbiol.* **24**, 134–147 (2016).
8. B. D. Greenbaum, A. J. Levine, G. Bhanot, R. Rabadan, *PLOS Pathog.* **4**, e1000079 (2008).
9. F. P. Lobo *et al*., *PLOS ONE* **4**, e6282 (2009).
10. A. Kapoor, P. Simmonds, W. I. Lipkin, S. Zaidi, E. Delwart, *J. Virol.* **84**, 10322–10328 (2010).
11. F. Di Giallonardo, T. E. Schlub, M. Shi, E. C. Holmes, *J. Virol.* **91**, e02381-16 (2017).
12. M. K. K. Leung, A. Delong, B. Alipanahi, B. J. Frey, *Proc. IEEE* **104**, 176–197 (2016).
13. B. J. H. Friedman, *Ann. Stat.* **29**, 1189–1232 (2001).
14. M. G. Han, D.-S. Cheon, X. Zhang, L. J. Saif, *J. Virol.* **80**, 12350–12356 (2006).
15. A. D. LaBeaud *et al*., *PLOS Negl. Trop. Dis.* **9**, e0003436 (2015).
16. Y. Tang *et al*., *Transbound. Emerg. Dis.* **60**, 152–158 (2013).
17. K. J. Olival, D. T. S. Hayman, *Viruses* **6**, 1759–1788 (2014).
18. G. Grard *et al*., *PLOS Pathog.* **8**, e1002924 (2012).
19. J. Quick *et al*., *Nature* **530**, 228–232 (2016).

ACKNOWLEDGMENTS

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.A.B. was supported by a Glasgow University Research Fellowship and the BBSRC (BB/M012956/1). D.G.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.G.S. conceived of the research; D.G.S. and R.J.O. collected the data; D.G.S., R.J.O., and S.A.B. analyzed the data; and D.G.S. and S.A.B. wrote and revised the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code reported in this paper are available at https://github.com/DanielStreicker/ViralHostPredictor.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/362/6414/577/suppl/DC1
Materials and Methods
Supporting Text
Figs. S1 to S18
References (*20*–*43*)
Data S1
Appendix S1

14 September 2017; resubmitted 21 May 2018
Accepted 12 September 2018
10.1126/science.aap9072

Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes

Simon A. Babayan, Richard J. Orton and Daniel G. Streicker

Science **362** (6414), 577-580.
DOI: 10.1126/science.aap9072

Predicting hosts and vectors

During outbreaks of mysterious infections, events can rapidly become dangerous and confusing. A combination of increasing experience with outbreaks and genome-sequencing technology now means the pathogen can often be identified within days. But for some of the most frightening viral pathogens, the originating hosts and possible vectors often remain obscure. Babayan *et al.* took sequence data from more than 500 single-stranded RNA viruses (see the Perspective by Woolhouse) and used machine-learning algorithms to extract evolutionary signals imprinted in the virus sequence that offer information about its original hosts and if an arthropod vector, and what type, plays a part in the virus's natural ecology.

Science, this issue p. 577; see also p. 524