ARTICLE

https://doi.org/10.1038/s41586-018-0012-7

The evolutionary history of vertebrate RNA viruses

Mang Shi[1,2,3,8], Xian-Dan Lin[4,8], Xiao Chen[5,8], Jun-Hua Tian[6,8], Liang-Jun Chen[1], Kun Li[1], Wen Wang[1], John-Sebastian Eden[3], Jin-Jin Shen[7], Li Liu[5], Edward C. Holmes[1,2,3] & Yong-Zhen Zhang[1,2]*

Our understanding of the diversity and evolution of vertebrate RNA viruses is largely limited to those found in mammalian and avian hosts and associated with overt disease. Here, using a large-scale meta-transcriptomic approach, we discover 214 vertebrate-associated viruses in reptiles, amphibians, lungfish, ray-finned fish, cartilaginous fish and jawless fish. The newly discovered viruses appear in every family or genus of RNA virus associated with vertebrate infection, including those containing human pathogens such as influenza virus, the *Arenaviridae* and *Filoviridae* families, and have branching orders that broadly reflected the phylogenetic history of their hosts. We establish a long evolutionary history for most groups of vertebrate RNA virus, and support this by evaluating evolutionary timescales using dated orthologous endogenous virus elements. We also identify new vertebrate-specific RNA viruses and genome architectures, and re-evaluate the evolution of vector-borne RNA viruses. In summary, this study reveals diverse virus–host associations across the entire evolutionary history of the vertebrates.

RNA viruses infect a wide range of hosts and contain enormous genetic and phenotypic diversity[1]. Because of their potential effect on public health and the agricultural industries, considerable attention has been directed towards describing the diversity and evolution of RNA viruses associated with vertebrates. Despite an increasingly widespread surveillance of invertebrate and vertebrate hosts, there are few direct links between invertebrate and vertebrate viruses, and vertebrate viruses tend to form monophyletic groups that are only distantly related to viruses found in invertebrates[1]. Within vertebrates, there has been a marked sampling bias towards mammals and birds[2], even though

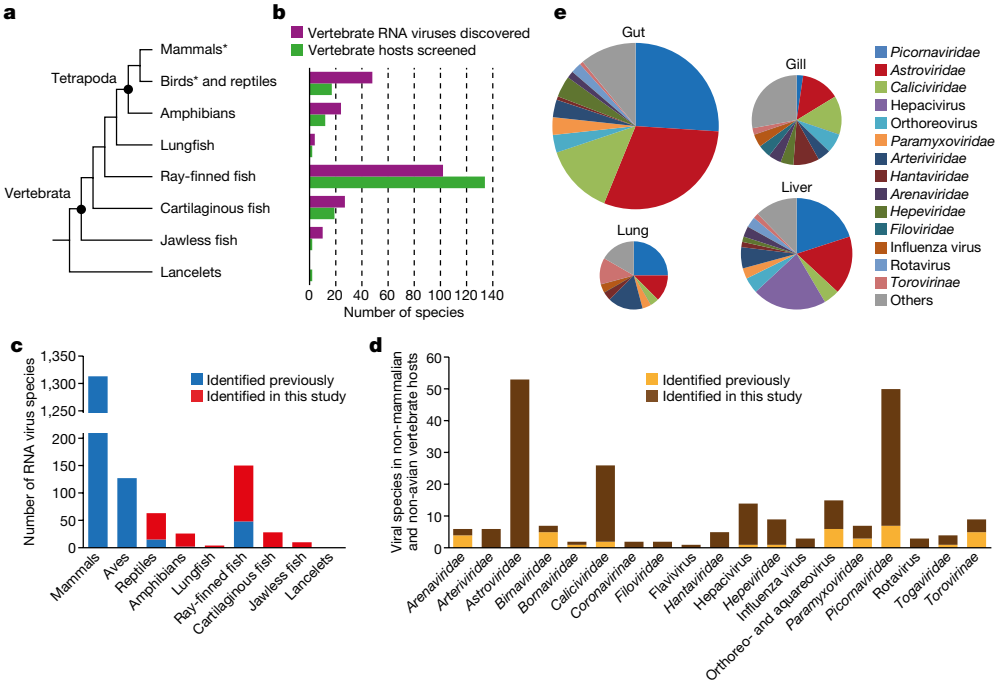


Fig. 1 | Identification of vertebrate-associated viruses in divergent vertebrate host groups. **a**, Phylogenetic relationships of the vertebrate host classes surveyed here. Asterisks denote hosts not surveyed in this study. **b**, Number of host species surveyed (purple) compared to the number of virus species discovered (green) in each host class. **c**, Number of viral species in each host class. Red and blue represent current and previously discovered RNA viruses, respectively. **d**, Number of non-avian and non-mammalian vertebrate virus species in each vertebrate-associated viral family or genus. Yellow and brown represent current and previously identified RNA viruses, respectively. **e**, Distribution of viruses identified in this study by tissue type.

[1]State Key Laboratory for Infectious Disease Prevention and Control, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China. [2]Shanghai Public Health Clinical Center & Institute of Biomedical Sciences, Fudan University, Shanghai, China. [3]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, New South Wales, Australia. [4]Wenzhou Center for Disease Control and Prevention, Wenzhou, China. [5]College of Marine Sciences, South China Agricultural University, Guangzhou, China. [6]Wuhan Center for Disease Control and Prevention, Wuhan, China. [7]Yancheng Center for Disease Control and Prevention, Yancheng, China. [8]These authors contributed equally: Mang Shi, Xian-Dan Lin, Xiao Chen, Jun-Hua Tian. *e-mail: zhangyongzhen@shphc.org.cn

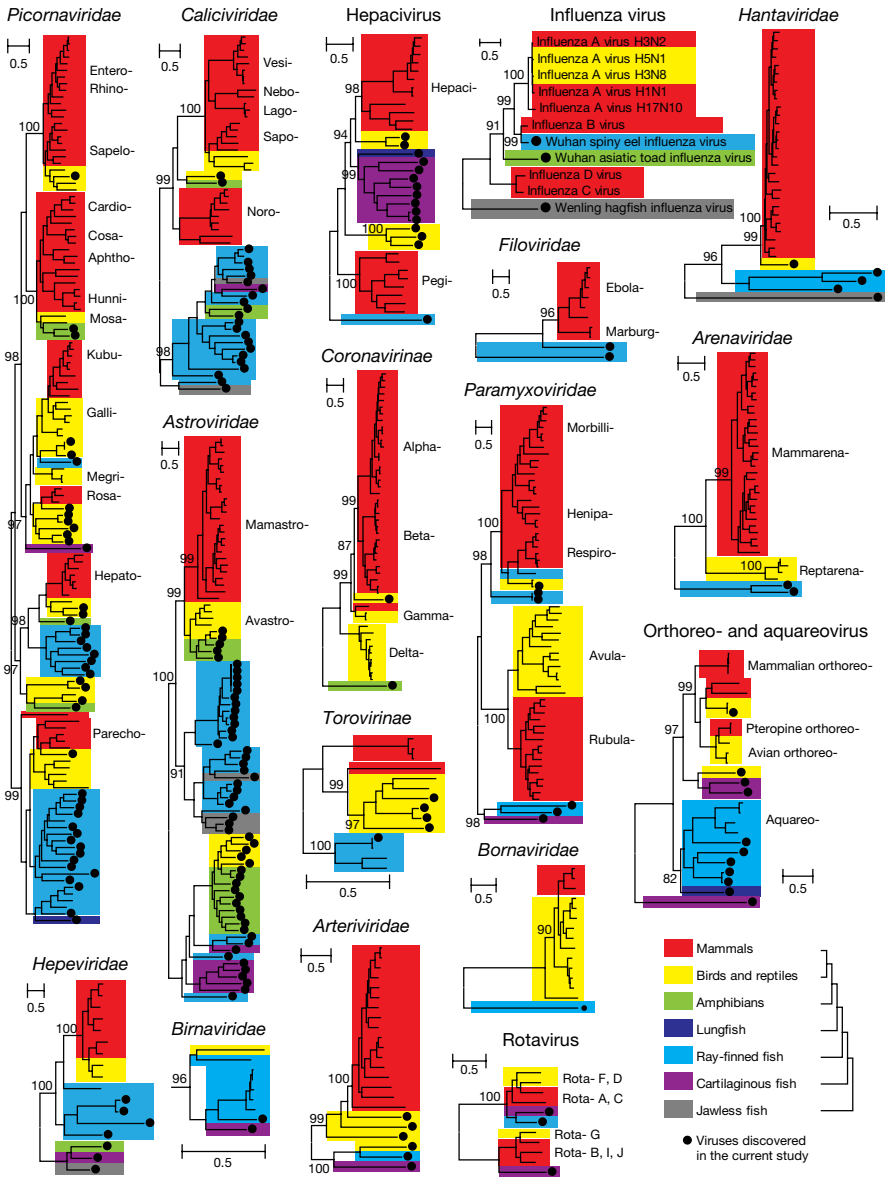


Fig. 2 | Evolutionary history of 17 major vertebrate-specific virus families or genera. Each phylogenetic tree was estimated using a maximum likelihood method, and is rooted using the corresponding broader scale tree that contains both vertebrate and invertebrate viruses (not shown). Within each phylogeny, the viruses newly identified here are marked with solid black circles. Host groups are indicated with different colours; mammals (red), birds and reptiles (yellow), amphibians (green), lungfish (deep blue), ray-finned fish (blue), cartilaginous fish (purple) and jawless fish (grey). The name of the virus family or genus is shown above each phylogeny, and the lower-order virus taxonomy is shown to the right when applicable.

they represent only a small proportion of total vertebrate diversity. Far less is known about those viruses infecting fish, amphibians and reptiles[3], despite their abundance, phenotypic diversity and central role in vertebrate evolution. Notably, the relatively few viruses from fish, amphibians and reptiles documented so far tend to form divergent lineages with respect to known vertebrate RNA viruses[4–7], which in part probably reflects the position of these hosts in the vertebrate phylogeny[8]. However, the extent of viral phylogenetic and genomic diversity in these taxa, their ancestry as well as the relative frequencies of virus–host co-divergence versus cross-species transmission in the evolution of vertebrate RNA viruses remains uncertain[9]. To better understand the origin and evolutionary history of vertebrate viruses, we screened for RNA viruses in a diverse set of species that covered much of the phylogenetic diversity of the vertebrates, including those basal vertebrate lineages in which viruses have only rarely been documented.

Expanding diversity of vertebrate viruses

We performed a large-scale meta-transcriptomics survey of potential vertebrate-associated RNA viruses in more than 186 host species representing the extensive diversity within the phylum Chordata (Fig. 1a, b, Supplementary Table 1). This included animals from the classes Leptocardii (lancelets), Agnatha (jawless fish), Chondrichthyes (cartilaginous fish), Actinopterygii (ray-finned fish), Sarcopterygii (lungfish), Amphibia (frogs, salamanders and caecilians) and Reptilia (snakes, lizards and turtles). We extracted total RNA from the gut, liver, and lung or gill tissue of these animals, which was then organized into 126 libraries for high-throughput RNA sequencing (Supplementary Table 1). In total, we generated 806 billion bases of sequence reads that were assembled and screened for RNA viruses. Despite the very large number of viruses discovered, we focused on vertebrate-associated viruses, including vertebrate-specific viruses that exhibited relatively close evolutionary relationships to known virus families or genera thought to infect only vertebrate hosts, and 'vector-borne' viruses that are able to infect both vertebrate and invertebrate hosts (Supplementary Table 2). In the resultant phylogenies, the newly discovered viruses either grouped within these families/genera, or fell as immediate sister-groups (Extended Data Figs. 1 and 2). Because the host spectrum of the vertebrate-specific virus families or genera is relatively restricted[2] and generally does not contain viruses associated with other host types[1], we assume that vertebrates were their principle hosts, rather than any eukaryotic or prokaryotic microorganisms also present in the samples. Furthermore, at least 24% of the viruses were recovered from different tissues from the same individual and hence are likely to cause systemic infection (Supplementary Table 2). This gives further

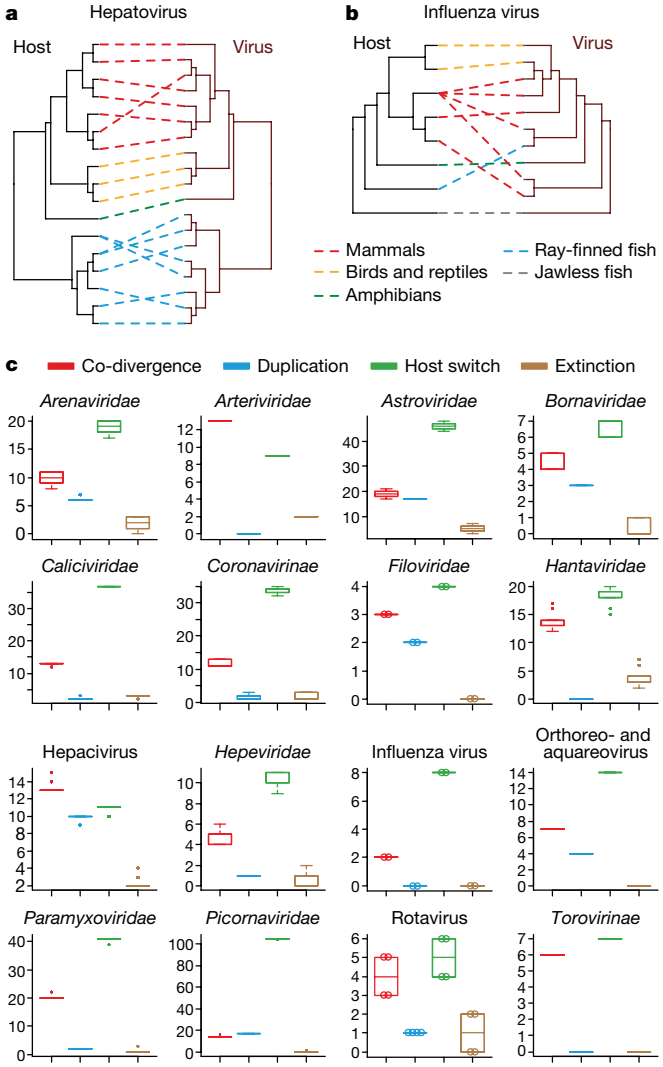


Fig. 3 | Long-term evolutionary relationships between vertebrate hosts and their associated viruses. a, **b**, Comparisons of hepatovirus (**a**) and influenza virus (**b**) phylogenies and their corresponding host phylogenies are presented as examples of virus–host co-divergence and host-switching, respectively. **c**, Estimation of co-phylogenetic events across the history of vertebrate-associated RNA viruses. Each boxplot illustrates the estimated median (centre line), upper and lower quartiles (box limits), 1.5 × interquartile range (whiskers), and outlier (points) of the co-divergence (red), duplication (blue), host-switching (green) and extinction (brown) events. Data from each estimation (hollow circles) are shown as overlays if there are less than 10 'solutions' provided.

support to a direct association within the vertebrate–host in which they were sampled.

In total, we identified 214 distinctive and previously undescribed putative viral species of vertebrates, of which 196 can be considered vertebrate-specific (Fig. 1b, Supplementary Table 2). Hence, these data reveal that RNA viruses are present in greater numbers and diversity in vertebrates other than birds and mammals than previously realized (Fig. 1c). In particular, it was notable that every vertebrate-specific viral family or genus known to infect mammals and birds is also present in amphibians, reptiles or fish (Fig. 1d). For most of the families or genera, the previously known hosts were either mammals (the *Arteriviridae*, *Filoviridae*, *Hantaviridae* and rubivirus) or mammals, birds and reptiles (*Arenaviridae*, *Astroviridae*, *Bornaviridae*, *Coronavirinae*, influenza virus and rotavirus). This is the first time, to our knowledge, that these viral groups have been identified in fish and/or amphibians (Fig. 1d). Particularly notable was the presence of divergent members of the *Arenaviridae*, *Filoviridae* and *Hantaviridae* families in ray-finned fish, suggesting that these previously mammal-dominated groups have relatives in aquatic vertebrates (Fig. 2). Similarly, for those virus groups previously known to contain fish viruses (*Caliciviridae*, *Hepeviridae*, *Paramyxoviridae* and *Picornaviridae*), we were able to greatly expand their genetic diversity, which now covers more phylogenetic space than in their mammalian counterparts (Fig. 2). Of particular note was influenza virus, for which we documented new viruses in jawless fish (hagfish), amphibians (Asiatic toad) and ray-finned fish (spiny eel), with the latter forming a sister-group to human influenza B virus (Fig. 2). Finally, it was notable that the viruses that were newly described in reptiles, amphibians and fish exhibited similar tissue tropisms as their mammalian counterparts[2], which again argues for their antiquity. For example, among the viruses discovered here, those of the Hepacivirus genus were mainly found in the liver, whereas members of the *Picornaviridae*, *Caliciviridae* and *Astroviridae* families dominate in the gut (Fig. 1e).

Long-term virus–host evolutionary relationship

On the basis of the distribution of host taxa on the virus tree, these data also revealed that virus phylogenetic history can mirror that of their hosts over long evolutionary timescales. Most notably, viruses from fish tend to fall basal to viruses in amphibians, reptiles, birds and mammals, reflecting their divergent phylogenetic position within vertebrates (Figs. 2 and 3a). This was supported by the observation that the virus phylogenies exhibited significant clustering by host taxonomy (that is, class), with $P < 0.001$ in the association index (AI)[10] for all family and genus comparisons, with the exception of influenza virus and rotavirus (Table 1). However, despite this overall host clustering, these data also revealed many examples of host-switching during virus evolutionary history. For example, the influenza virus identified in ray-finned fish was the closest relative of mammalian influenza B virus (76% amino acid identity), and the influenza viruses sampled from other tetrapods was more divergent (approximately 30–62% amino acid identity; Fig. 3b). Similarly, the viruses identified in lungfish (in *Picornaviridae*, hepacivirus and aquareovirus; Fig. 2) were more closely related to those from ray-finned fish than to those from tetrapods with whom they share a more recent common ancestor[11].

Table 1 | Phylogenetic test of virus–host association and co-divergence

	Test of host structure at the level of vertebrate class		Test of virus–host co-divergence		
Virus group	Association index ratio*	*P* value (AI)	Co-divergences	Number of costs	*P* value (no. of costs)
Arenaviridae	0.0000	<0.001	10–12	27	<0.01
Arteriviridae	0.4960	0.047	13	11	<0.01
Astroviridae	0.0878	<0.001	17–21	68	<0.01
Bornaviridae	0.0020	<0.001	4–5	10	0.05
Caliciviridae	0.2736	<0.001	12–13	42	<0.01
Coronavirinae	0.0834	<0.001	11–13	37	<0.01
Filoviridae	0.0017	<0.001	3	6	0.06
Hantavirus	0.0022	<0.001	12–17	22	<0.01
Hepacivirus	0.0002	<0.001	13–15	23	<0.01
Hepeviridae	0.2935	<0.001	4–6	12	0.13
Influenza virus	0.9173	0.65	2	8	0.8
Orthoreo- and aquareovirus	0.1015	<0.001	7	18	<0.01
Paramyxoviridae	0.0231	<0.001	20–22	44	<0.01
Picornaviridae	0.0294	<0.001	14–15	122	<0.01
Rotavirus	0.9275	0.34	3–5	7	0.52
Torovirinae	0.0072	<0.001	6	7	<0.01

The association index (AI) ratio is calculated as 'observed association index/null association index', in which the null association index is derived from 1,000 tree-tip randomizations. A ratio closer to 0 indicates a stronger host structure. The '*P* values (AI)' are outcomes from a Bayesian tip-association significance test (BaTS)[10], and derived from 1,000 tree tip randomizations without adjustment for multiple comparisons. The cost, that is, non-co-divergence, scheme included 'host-switching', 'host duplication', 'host loss' and 'failure to diverge' events, as specified in the model. The '*P* values (no. of costs)' are outcomes from a co-phylogeny test[12], and are derived from 100 tip-mapping randomizations without adjustment for multiple comparisons.

We next performed a more rigorous co-phylogenetic analysis[12,13] of the resemblance between the virus and host phylogenies at the species level. This revealed significantly more virus–host co-divergence than expected by chance alone (Table 1). However, these data also clearly show that host-switching has been commonplace during the evolutionary history of vertebrate RNA viruses, and is often more frequent than co-divergence across the phylogenies as a whole (Fig. 3c). Aside from phylogenetic position, host-switching is also suggested by the observation that single viruses are occasionally associated with multiple host species or even multiple host orders (such as Beihai fish astrovirus 1 and Wenling fish picornavirus 1; Supplementary Table 2). Collectively, these results suggest that there is a long-term association between the RNA viruses and their vertebrate hosts that stretches many millions of years, but that cross-species transmission has occurred frequently on this background of co-divergence.

To better determine the co-divergence history, we examined the temporal congruence between virus and host evolutionary histories[14,15]. As the large genetic distances between these viruses preclude molecular clock-based studies using heterochronous sequences[16,17], a more profitable approach involves the comparison of exogenous viruses and their endogenous relatives[18]. Previous studies have identified several dating calibration points in the *Filoviridae*[19] and *Bornaviridae*[18,20] families based on the presence of orthologous copies of endogenous virus elements (EVEs) in the genomes of related mammalian species with known times of divergence. Importantly, the viruses newly discovered here in ray-finned fish greatly expand the diversity in both the *Bornaviridae* (Fig. 4a) and *Filoviridae* (Fig. 4b). As a result, both the EVE clades and the calibration points (50 million years (Myr) ago and 30 Myr ago for the *Bornaviridae* and *Filoviridae*, respectively)[18,19] were now deeply nested within the diversity of exogenous viruses, with phylogenetic positions that were relatively distant from the root of the tree. This suggests that both viruses have ancient evolutionary histories that extend well beyond the calibration dates. Although no orthologous EVEs were found in the positive-sense and double-stranded RNA virus families studied here, that their (exogenous) protein sequence divergence was comparable to that of the *Bornaviridae* and *Filoviridae* is also compatible with long evolutionary histories.

Additional vertebrate-associated viruses

We discovered two potentially new groups of vertebrate-associated viruses: one distantly related to the *Astroviridae* and *Potyviridae* families, and another nested within the newly characterized Chuvirus group[21] (Extended Data Fig. 3). Several pieces of evidence support the association of these viruses with vertebrates: (i) they appear in several tissue types (gut, gill and liver), indicative of systemic infection (Extended Data Fig. 3); (ii) a search of the Transcriptome Shotgun Assembly (TSA) sequence database revealed that related viral sequences were found only in vertebrate transcriptomes, again involving several tissue types (Extended Data Fig. 3); and (iii) in the case of the vertebrate-associated chuviruses, EVEs were found in the genomes of several species of ray-finned fish (Extended Data Fig. 3).

In addition to the vertebrate-specific viruses, we discovered viruses in amphibians, fish and reptiles from genera that have previously been associated with vector-borne virus transmission, most notably alphaviruses, dimarhabdoviruses and flaviviruses. Among these, Wenzhou shark flavivirus is the first member of the Flavivirus genus identified in cartilaginous fish, and was found in all the tissue types analysed compatible with a systemic infection (Supplementary Table 2). In the phylogeny, Wenzhou shark flavivirus falls basal to the 'classic' vector-borne and insect-specific flaviviruses, and was more closely related to Tamana bat virus that has no known vector species (Extended Data Fig. 4). Similarly, in the case of the alphaviruses and dimarhabdoviruses, the fish viruses discovered here clustered with other fish viruses reported previously to form lineage(s) basal to those associated with vector-borne viruses (Extended Data Fig. 4). This complex mix of vectored and non-vectored viruses, with clear cases of the secondary loss of vector-borne transmission (Extended Data Fig. 4), raises the question

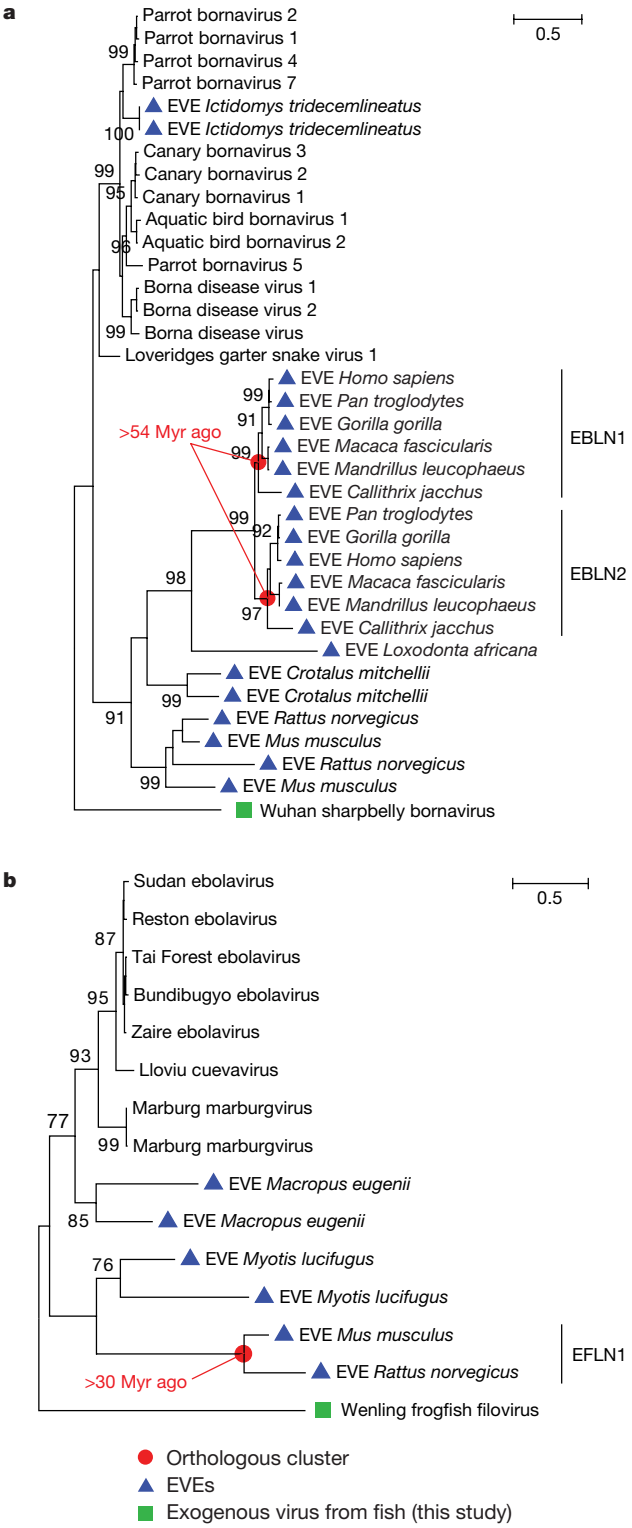


Fig. 4 | Evaluating the timescale of vertebrate virus evolution using EVEs. a, **b**, Phylogenies were based on the exogenous and endogenous nucleoproteins for the *Bornaviridae* (**a**) and *Filoviridae* (**b**) families. Within the trees, endogenous virus elements (EVEs) are highlighted with blue triangles, and the (divergent) exogenous viruses discovered in fish are highlighted with green squares. The nodes that represent orthologous clusters are highlighted with red circles, their associated divergence times are shown next to the nodes, and their corresponding names are given to the right of the phylogeny.

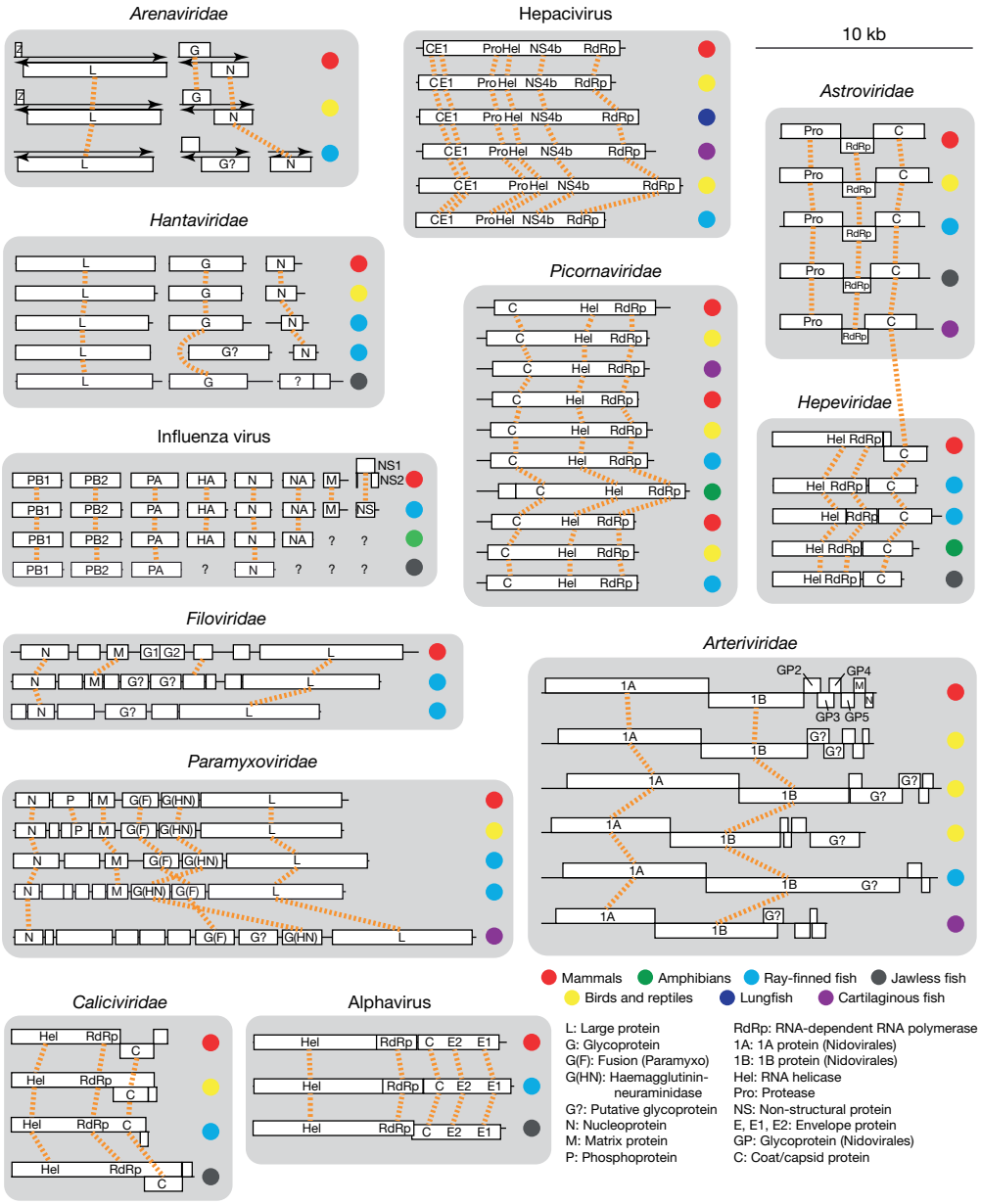


Fig. 5 | Evolution of vertebrate-associated virus genomes. Representative genomes from 12 vertebrate-associated virus families or genera are shown. The regions that encode major functional proteins or protein domains are labelled on each of the genomes. Homologous regions within or between viral families are connected by orange dotted lines. Host associations are labelled to the right of each genome using solid circles with different colours. The orientation of the positive-sense genomes are shown from 5′ to 3′, those of negative-sense genomes are from 3′ to 5′, and those of ambisense genomes (that is, arenaviruses) are indicated using arrows. More detailed depictions of genome evolution are presented in Extended Data Figs. 5 and 6.

of whether some of the vector-borne viruses were ultimately derived from vertebrate-specific or vector-specific viruses, or if the ability to infect both arthropods and vertebrates is the ancestral phenotype[22].

Genome evolution of vertebrate RNA viruses

The annotation of the virus genomes newly documented here showed a wider variety of genome architectures for vertebrate virus families or genera than previously observed[2], some of which may represent the ancestral types in the evolutionary history of these viruses (Fig. 5). Although the structures of these vertebrate virus genomes were more conserved than those of invertebrates[1,6,21,23], they still exhibited extensive variation, including genome length (hepacivirus), the organization of open reading frames (*Caliciviridae*), the complete re-configuration of the genome downstream of the non-structural genes (*Arteriviridae*), changes in the order and number of glycoproteins (*Paramyxoviridae*), inter-species re-assortment involving the M segment (hantavirus), inter-family recombination involving the capsid protein (*Astroviridae* and *Hepeviridae*)[24] and changes in segment numbers in the *Arenaviridae* family (Fig. 5, Extended Data Figs. 5 and 6). The latter is particularly interesting as the *Arenaviridae* were traditionally thought to be a family of bi-segmented negative-sense RNA viruses[2]. However, we discovered two arenavirus species in fish with genomes comprising three segments, similar to that of the divergent relative of the *Arenaviridae* family found in arthropods[21], and suggesting that there was a decrease in segment numbers from three to two (Extended Data Fig. 5). If so, this represents an important example of a reduction in segment number without a corresponding loss in gene content, hence compatible with segment merging.

Discussion

Despite a combination of rapid evolution and frequent host-switching, our large-scale analysis of virus diversity in previously undersampled hosts suggests that RNA viruses in vertebrates tend to broadly follow the evolutionary history of their hosts that began in the ocean and extends for hundreds of millions of years. These results, which apply to most of the vertebrate RNA virus families or genera, are in accord with recent analyses of viral evolution using palaeovirological data[18–20,25,26], and demonstrate the importance of conducting widespread taxonomic surveys of virus diversity when trying to reveal evolutionary history. These results also have broader implications for our understanding of virus evolution. In particular, it is clearly simplistic and perhaps erroneous to identify a specific host group as ancestral to another given that our sampling of RNA virus diversity is still so very limited. For example, on current data we suggest that it is premature to conclude that vertebrate RNA viruses necessarily originated in mosquitoes/ticks, since it is possible that the evolution of specific virus families may have followed that of the metazoans over an even longer period of co-divergence. In summary, our study reveals long-term virus–host relationships for each vertebrate-associated virus family that extend over geological timescales, further illustrating the ancient history of RNA viruses.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0012-7.

Received: 15 September 2017; Accepted: 23 February 2018;
Published online 4 April 2018.

1. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
2. King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. *Virus Taxonomy: 9th Report of the International Committee on Taxonomy of Viruses* (Elsevier Academic, Amsterdam, 2012).
3. Essbauer, S. & Ahne, W. Viruses of lower vertebrates. *J. Vet. Med. B Infect. Dis. Vet. Public Health* **48**, 403–475 (2001).
4. Batts, W., Yun, S., Hedrick, R. & Winton, J. A novel member of the family Hepeviridae from cutthroat trout (*Oncorhynchus clarkii*). *Virus Res.* **158**, 116–123 (2011).
5. Mikalsen, A. B. et al. Characterization of a novel calicivirus causing systemic infection in atlantic salmon *(Salmo salar* L.): proposal for a new genus of caliciviridae. *PLoS ONE* **9**, e107132 (2014).
6. Shi, M. et al. Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the *Flaviviridae* and related viruses. *J. Virol.* **90**, 659–669 (2015).
7. Stenglein, M. D. et al. Identification, characterization, and *in vitro* culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease. *MBio* **3**, e00180–12 (2012).
8. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol* **32**, 835–845 (2015).
9. Dill, J. A. et al. Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *J. Virol.* **90**, 7920–7933 (2016).
10. Wang, T. H., Donaldson, Y. K., Brettle, R. P., Bell, J. E. & Simmonds, P. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75**, 11686–11699 (2001).
11. Brinkmann, H., Venkatesh, B., Brenner, S. & Meyer, A. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proc. Natl Acad. Sci. USA* **101**, 4900–4905 (2004).
12. Conow, C., Fielder, D., Ovadia, Y. & Libeskind-Hadas, R. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol. Biol.* **5**, 16 (2010).
13. Geoghegan, J. L., Duchêne, S. & Holmes, E. C. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* **13**, e1006215 (2017).
14. Charleston, M. A. & Robertson, D. L. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol* **51**, 528–535 (2002).
15. de Vienne, D. M. et al. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytol.* **198**, 347–385 (2013).
16. Wertheim, J. O. & Kosakovsky Pond, S. L. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol* **28**, 3355–3365 (2011).
17. Zhang, Y. Z. & Holmes, E. C. What is the time-scale of hantavirus evolution? *Infect. Genet. Evol.* **25**, 144–145 (2014).
18. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6**, e1001191 (2010).
19. Taylor, D. J., Leach, R. W. & Bruenn, J. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol. Biol.* **10**, 193 (2010).
20. Horie, M. et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).
21. Li, C. X. et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* **4**, e05378 (2015).
22. Longdon, B. et al. The evolution, diversity, and host associations of rhabdoviruses. *Virus Evol.* **1**, vev014 (2015).
23. Qin, X. C. et al. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. *Proc. Natl Acad. Sci. USA* **111**, 6744–6749 (2014).
24. Kelly, A. G., Netzler, N. E. & White, P. A. Ancient recombination events and the origins of hepatitis E virus. *BMC Evol. Biol.* **16**, 210 (2016).
25. Han, G. Z. & Worobey, M. An endogenous foamy-like viral element in the coelacanth genome. *PLoS Pathog.* **8**, e1002790 (2012).
26. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat. Commun.* **8**, 13954 (2017).

Acknowledgements This study was supported by the Special National Project on Research and Development of Key Biosafety Technologies (2016YFC1201900, 2016YFC1200101) and the National Natural Science Foundation of China (Grants 81672057, 81611130073). E.C.H. and M.S. are funded by an ARC Australian Laureate Fellowship to E.C.H. (FL170100022). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank students at the Zoonosis branch of the China CDC, especially W.-C. Wu, J.-W. Shao, C.-X. Li, J.-J. Guo and K.-L. Song for assistance with virus and host sequence confirmation, and we thank B. Yu for help with the collection of animal samples. We acknowledge the University of Sydney high-performance computing (HPC) service at The University of Sydney for providing resources that have contributed to the research results reported within this paper

Reviewer information *Nature* thanks A. Rambaut and M. Worobey for their contribution to the peer review of this work.

Author contributions M.S. and Y.-Z.Z. conceived and designed the study. M.S., X.-D.L., X.C., J.-H.T., K.L., L.-J.C., J.-J.S., L.L. and Y.-Z.Z. organized field work, and collected samples. M.S., X.-D.L., X.C., J.-H.T., K.L., L.-J.C., W.W., J.-J.S., L.L. and Y.-Z.Z. performed experiments. M.S., J.-S.E., E.C.H. and Y.-Z.Z. analysed data. M.S., E.C.H. and Y.-Z.Z. wrote the paper with input from all authors. Y.-Z.Z. led the study.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0012-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0012-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.-Z.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample collection. The goal of this study was to survey animal species that were representative of biological diversity within the phylum Chordata, and that have only rarely been analysed for the presence of RNA viruses. Accordingly, we focused on amphibians, reptiles and fish rather than birds and mammals that have been studied in far greater detail (Fig. 1d). We also targeted species distributed at diverse locations across the vertebrate phylogeny (Fig. 1a), although those species associated with most basal vertebrate lineages are often rare. For each species we sampled 1–24 individuals to represent a population. No statistical methods were used to predetermine sample size. The procedures for sampling and sample processing were approved by the ethics committee of the National Institute for Communicable Disease Control and Prevention of the Chinese CDC.

In total, we sampled two species from the subphylum Cephalochordata (that is, lancelets), with the remainder from the subphylum Vertebrata (Supplementary Table 1, Fig. 1a). Within Vertebrata, we sampled two species each from the classes Agnatha (that is, jawless fish) and Sarcopterygii (that is, lungfish), as these are relatively rare. Most of our aquatic samples were from the classes Chondrichthyes (cartilaginous fish), from which we sampled 19 species, and Actinopterygii (bony fish), from which we sampled more than 130 species across 20 orders (Supplementary Table 1). With respect to land tetrapods, we sampled 12 species from the class Amphibia, including the orders Aura, Caudata and Gymnophiona, and 17 species from the class Reptilia, including the orders Testudines and Squamata (Supplementary Table 1).

With the exception of lungfish samples, which were obtained from Nigeria (*Protopterus annectens*) and Chile (*Lepidosiren* sp.), respectively, all other samples were collected in China (Supplementary Table 1). The marine species were sampled from the South China Sea, East China Sea and Yellow Sea, mostly from fishing vessels. The samples were kept at −20 °C on the boat before being transferred to −80 °C for storage. The remaining marine samples were either collected frozen from the returned ships at the dock, or purchased alive from local fisherman at nearby markets. The freshwater fish samples were captured alive using fishing rods or nets from rivers and lakes in Hubei, Heilongjiang and Guangdong provinces. The reptile and amphibian samples were caught by field biologists from a wide range of geographic locations, including Fujiang, Guangdong, Guangxi, Xinjiang and Zhejiang provinces.

For most of the animal samples, three types of internal organs were harvested, comprising the gut, liver and gill for jawless, cartilaginous, and ray-finned fish, and gut, liver and lung for amphibians and reptiles (Supplementary Table 1). For lungfish, all four types of tissue (that is, gut, liver, lung and gill) were obtained. For lancelets, the entire individual was used owing to their small body size. All specimens were stored at −80 °C for later RNA extraction.

Host species information was initially identified by experienced field biologists on capture based on morphological traits, and was later confirmed by sequencing and analysing the partial cytochrome *c* oxidase (COI) gene from each sample (approximately 600–700 nucleotides near 5′ of the gene).

RNA library construction and sequencing. RNA was extracted from individual animal specimens. For the initial screening of viruses, aliquots of RNA from several (that is, from 13 to 62) individuals of a particular taxonomic group or multiple taxonomic groups were pooled for library preparation and sequencing (Supplementary Table 1). After determining the presence of a specific virus, a subset of the initial pool or the individual un-pooled RNA extractions was subject to library construction and sequencing to obtain better genome coverage (Supplementary Table 1).

For each RNA extraction, we first transferred approximately 30 mg from the specimen to 500–700 µl standard, sterile, RNA and DNA-free 1× PBS solution (GIBCO). The tissue was then homogenized in the PBS solution using the Mixer mill MM400 (Restsch). Total RNA was extracted from the homogenates using TRIzol LS reagent (Invitrogen) and subsequently purified using RNeasy Plus Mini Kit (Qiagen). Aliquots of the resultant RNA solutions were then pooled in equal quantity. The quality of the pooled RNA was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies) before library construction and sequencing.

The TruSeq total RNA Library Preparation protocol (Illumina) was used for all library preparations. Ribosomal (r) RNA was removed using the Ribo-Zero Gold (Epidemiology) Kit (Illumina) for most of the libraries, with the exception of LXMC-PolyA and XYHYMC-PolyA for which poly(A) enrichment was used (Supplementary Table 1). The average fragmentation size for these libraries was either 200 bp or 300 bp. Accordingly, 100 bp and 150 bp paired-end sequencing of the RNA libraries were performed on the Hiseq 2500 and HiSeq 4000 platforms (Illumina), respectively. All library preparation and sequencing was carried out by BGI Tech (Shenzhen).

RNA virus discovery. For each library, sequencing reads were first adaptor- and quality-trimmed using the Trimmotic program[27] with the following options: SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, MINLEN:25. The remaining reads were assembled de novo using the Trinity program (version 2.1)[28] with default parameter settings. To identify viral contigs, the assembled contigs were compared (using blastx) against the database comprising reference RNA virus proteins downloaded from GenBank. The *E*-value cut-off for these comparisons was set at 1×10^{-5}. To eliminate false positives, these putative viral contigs were compared against the entire non-redundant nucleotide and protein database. The remaining contigs with unassembled overlaps were merged to form longer viral contigs using the SeqMan program implemented in the Lasergene software package (version 7.1, DNAstar).

Among all the virus contigs discovered, those likely to be associated with vertebrates (that is, vertebrate-specific viruses and vector-borne vertebrate viruses) were initially identified based on a closer relationship to established vertebrate-associated viruses than to other taxa in a Blast analysis (that is, known vertebrate-associated viral families/genera were the top blast hits). This relationship was later confirmed by more detailed phylogenetic analyses including viruses representative of both vertebrates and a wider variety of non-vertebrate organisms[1,6,21] (Extended Data Figs. 1 and 2).

For the vertebrate-associated viruses, we determined which samples contained the viruses and hence its potential host(s) using PCR with reverse transcription (RT–PCR) and sequencing. Accordingly, for each virus, we designed 2–3 pairs of primers based on the viral contigs and screened all the unpooled RNA extractions of the corresponding library. The target PCR products were then validated by Sanger sequencing.

Gaps in incomplete vertebrate-associated virus genomes were filled by either RT–PCR or by re-sequencing (using the meta-transcriptomics approach described above) on the individual RNA samples that contained the target virus. Genome termini were determined by RNA circularization as previously described[23], or by using the 5′/3′ RACE kits (TaKaRa). Confirmation of most of the viral genome sequences was performed by read mapping using Bowtie2[29], with the final majority consensus sequences determined from the final assembly of mapped reads using Geneious v.8[30]. For virus species with multiple variants in the same pool, we performed meta-transcriptomics or RT–PCR and Sanger sequencing of the entire genome from the individual positive sample. To exclude the possibility that these contigs belonged to expressed EVEs (see below), we used PCR and Sanger sequencing to examine the DNA extracted from the homogenates of the corresponding samples.

Searching existing databases for vertebrate viruses. To discover more vertebrate-associated viruses and hence enrich our dataset, we downloaded the entire Transcriptome Shotgun Assembly (TSA) sequence database which was then used as query to search against the virus protein database as previously described. Because not all transcriptome sequences have a TSA (assembled) entry, we also examined reads deposited in the Sequence Read Archive (SRA) database. We targeted basal vertebrate taxa with inadequate or limited sampling, including lancelets (NCBI taxonomy ID: 7736), jawless vertebrates (NCBI taxonomy ID: 1476529), cartilaginous fish (NCBI taxonomy ID: 7777), and lungfish (NCBI taxonomy ID: 7878). These reads were assembled using Trinity and compared against the virus protein database as described above. Unfortunately, no vertebrate-associated viruses were found in these read archives.

To reveal viruses that may have infected vertebrates in the evolutionary past, we searched within the vertebrate genomes for EVEs that were relatively closely related to the viruses discovered in this study, especially those that did not belong to any established vertebrate clade. Accordingly, we first downloaded all the assembled genome sequences within the taxonomic group Vertebrata (NCBI taxonomy ID: 7742) from the NCBI genome FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). We then compared the translated viral protein sequences discovered in this study against the all assembled vertebrate genomes using the tblastn program, with an *E*-value cut-off set at 1×10^{-20}. For each potential EVE, the query process was reversed to determine their phylogenetic positions. The alignment of EVEs and exogenous viruses was checked manually to exclude false-positives.

Virus genome characterization. For newly identified virus genomes, the prediction of the potential open reading frames (ORFs) was based on those from the related reference virus genomes. The annotation of ORFs was first based on comparisons against the Conserved Domain Database (CDD) and then against the non-redundant protein database. The remaining proteins were characterized by predicting their primary protein structure using the programs NetNGlyc, SignalP, and TMHMM (http://www.cbs.dtu.dk/services/). For example, the divergent glycoprotein genes of negative-sense RNA families were identified based on the presence of (i) an N-terminal signal peptide, (ii) a mid-point or C-terminal transmembrane domain, and (iii) putative *N*-linked glycosylation sites. Finally, the sequencing depth of each viral genome within the library was estimated based on the percentage of total reads that mapped to the target genome.

In the case of segmented viruses, most of the non-RdRp segments were recovered by homology comparisons. However, divergent members of the families *Hantaviridae* and *Arenaviridae* had glycoproteins that lacked clear homology with those of other family members. To look for these segments we first annotated

all contigs that were of similar sequencing depths by comparing them to the nr database. This removed most sequences of host origin. For the remaining contigs, we examined (i) the potential glycoprotein structure (that is, signal peptide, transmembrane domains and glycosylation sites), (ii) the presence of inverted complementary genome termini that are the same to those of other segments, (iii) whether all the segments were found in the same samples and (iv) whether its closest relative contained the related segment. Only when all four criteria were satisfied did we conclude that these segments most likely belonged to the same virus.

Inferring virus evolutionary history. We examined the phylogenetic relationship among these viruses at two levels: (i) an overall evolutionary history that placed the vertebrate-associated viruses in the context of viruses sampled from other hosts, and (ii) family/genus specific phylogenies that provide a more detailed depiction of the evolutionary relationships within each of the vertebrate-associated virus families/genera. At the family/genus level, we included as background all reference virus replicase sequences (that is, RNA-dependent RNA polymerase; RdRp) as well as replicases from non-reference viruses that occupied a unique phylogenetic position and which had an established host association. At the overall level, we included viral replicases representative of a broader phylogenetic diversity[1,6,21] in addition to those used in the family/genus level phylogenies.

For each dataset, the virus replicases were aligned using the E-INS-i algorithm implemented in the program MAFFT (version 7)[31], with all ambiguously aligned regions were subsequently removed using TrimAl (version 1.2)[32]. The best-fit model of amino acid substitution in each dataset was determined using ProtTest (version 3.4)[33]. Phylogenetic trees were then inferred using the maximum likelihood method implemented in PhyML (version 3.0)[34], using the best-fit substitution model and Subtree Pruning and Regrafting (SPR) branch-swapping. Support for specific nodes on the trees was assessed using an approximate likelihood ratio test (aLRT) with the Shimodaira–Hasegawa-like procedure. In addition, phylogenetic trees were inferred using the Bayesian method implemented in the program MrBayes v.3.2[35], using the same amino acid substitution models. Because the tree topologies generated by the two programs were largely identical, only maximum likelihood phylogenies are shown here.

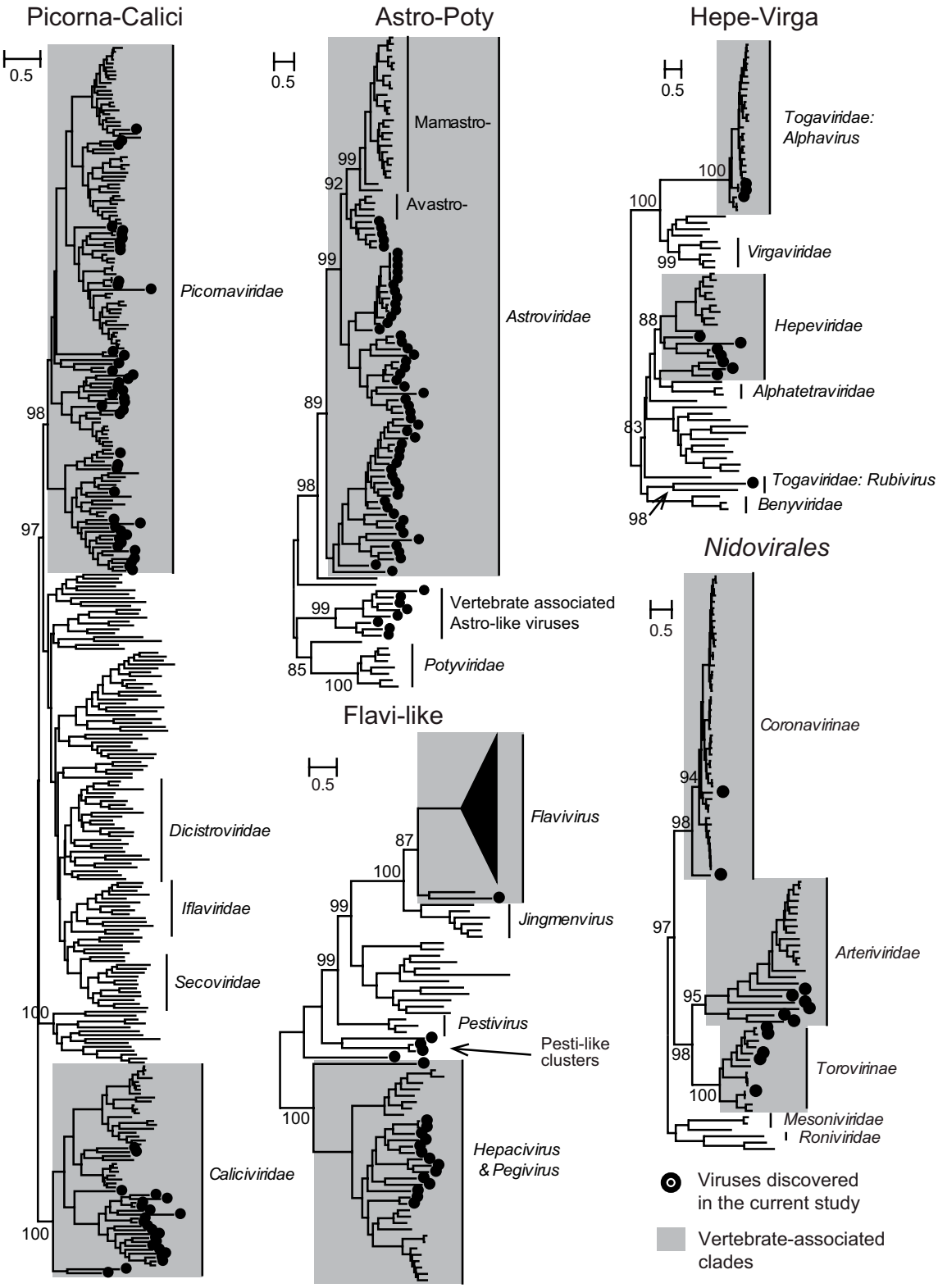
Examining virus–host evolutionary relationships. We used the BaTS (Bayesian tip-association significance testing) program[36] to test whether viruses cluster more strongly with particular host taxonomic groups than expected by chance alone. This analysis considered host phylogenetic structure at the level of vertebrate class: that is, mammals, reptiles and birds, amphibians, lungfish, bony fish, cartilaginous fish, and jawless fish. Accordingly, we estimated the association index[10] within BaTS to determine the strength of the association between virus phylogeny and host class. This was then compared to a null distribution generated using 1,000 replicates of state randomization across a credible set of trees generated by MrBayes as described above.

To examine the extent of virus-host co-divergence in each vertebrate-specific virus family/genus, we performed event-based co-phylogenetic reconstructions using the Jane program (version 4)[12]. The virus phylogenies were based on the family/genus level phylogenies estimated here, from which we removed those with no host information. All 'generalist' viruses (that is, those that infect more than three species of hosts) were included in the analyses as unresolved parallel lineages. The corresponding host topologies were obtained from both the TIMETREE website (http://www.timetree.org/) and a previous phylogeny of bony fish[37]. The 'cost' scheme for analyses in Jane was set as follows: co-divergence = 0, duplication = 1, host switch = 1, loss = 1, failure to diverge = 1. The number of generations and the population size were both set to 100. The significance of co-divergence was derived by comparing the estimated costs to null distributions calculated from 100 randomizations of host tip mapping.

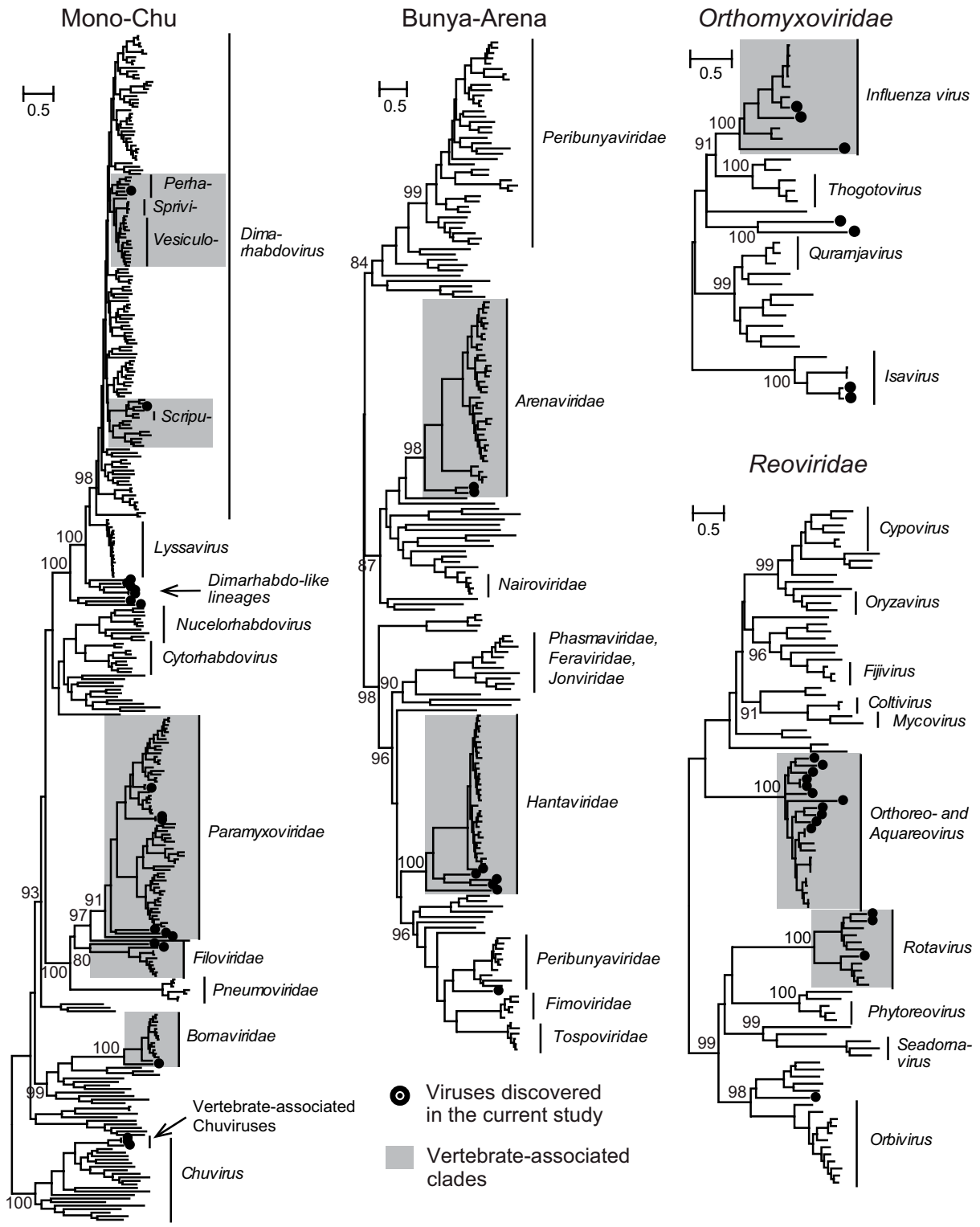
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All sequence reads generated in this study are available at the NCBI Sequence Read Archive (SRA) database under the BioProject accession PRJNA418053 (Supplementary Table 1). All viral sequences generated in this study have been deposited in GenBank under the accession numbers MG599863–MG600130 (Supplementary Table 2). All virus nucleotide sequences (fasta format), the unaligned and the aligned data set used in the phylogenetic analyses (fasta format), as well as the phylogenetic trees (newick and MEGA5 mts format), are available at the Figshare website at: https://Figshare.com/articles/The_evolutionary_history_of_vertebrate_RNA_viruses/5405620.

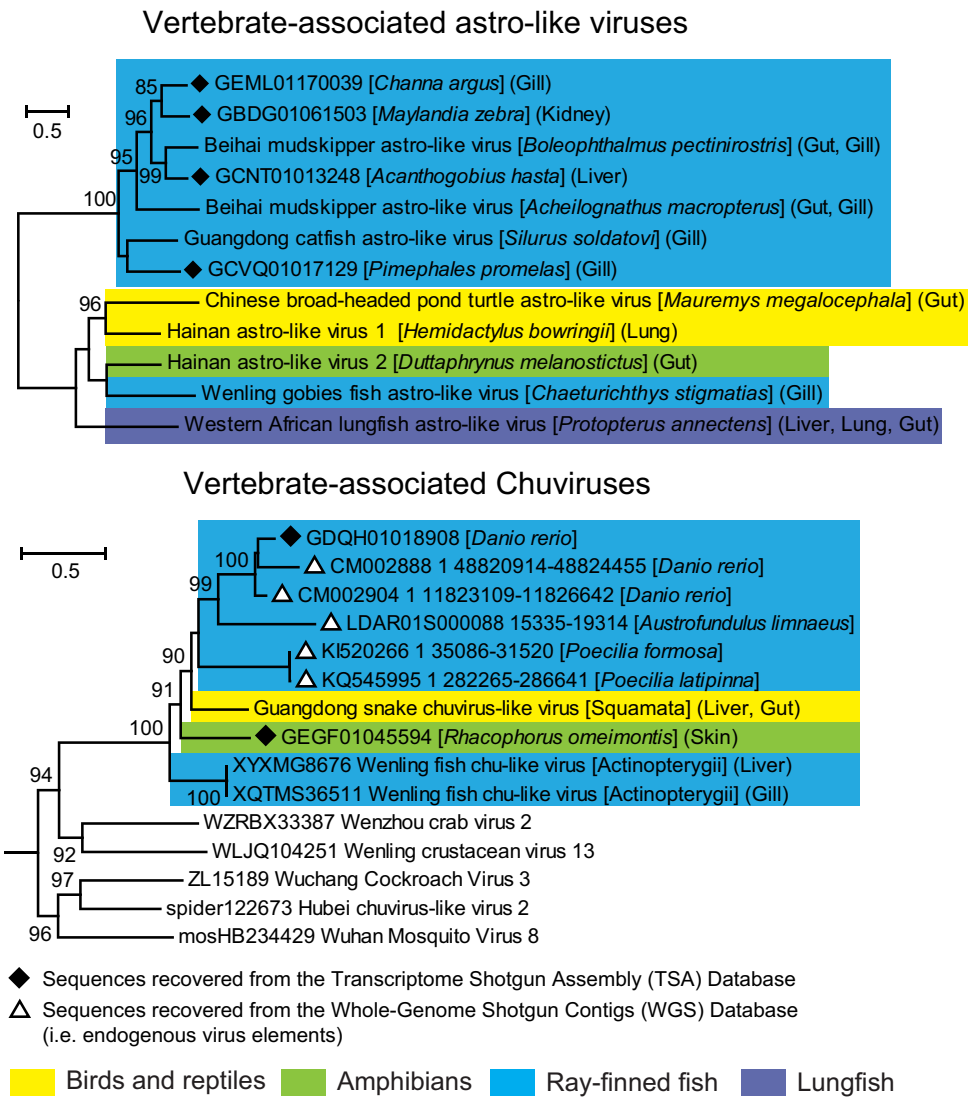
27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
28. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
31. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
32. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
33. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
34. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol* **52**, 696–704 (2003).
35. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
36. Parker, J., Rambaut, A. & Pybus, O. G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).
37. Betancur-R, R. et al. The tree of life and a new classification of bony fishes. *PLoS Curr.* **5**, https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288 (2013).



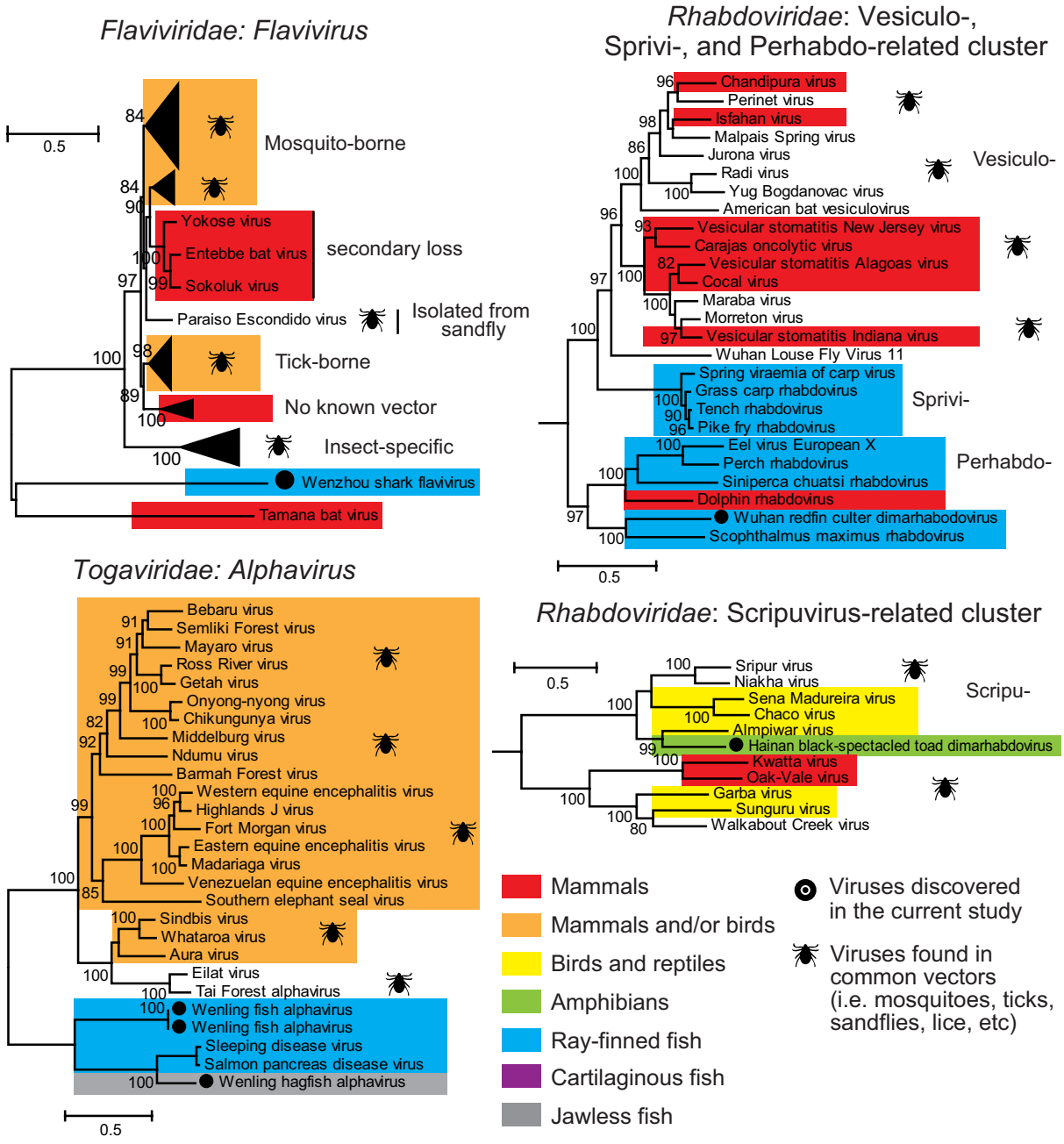
Extended Data Fig. 1 | Phylogenetic positions of vertebrate-associated positive-sense and double-stranded RNA viruses within the broader diversity of RNA viruses. Phylogenies were estimated using a maximum likelihood method and midpoint-rooted for clarity only. Viruses discovered here are labelled with solid black circles. The name of the major clade (phylogeny) is shown at the top of each tree, and taxonomic names are shown to the right. The vertebrate associated virus diversity is shaded in grey. All horizontal branch lengths are scaled to the number of amino acid substitutions per site.



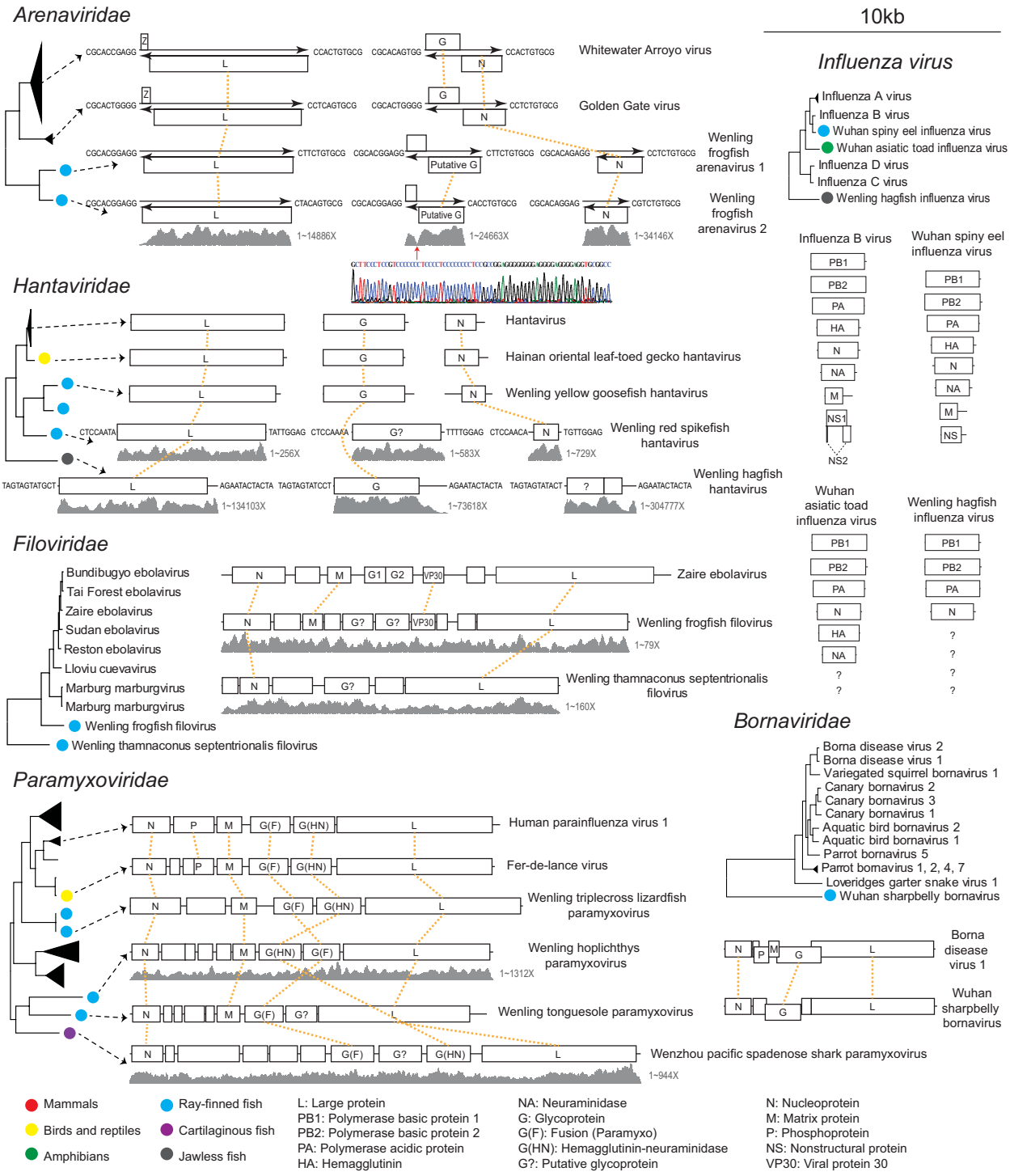
Extended Data Fig. 2 | Phylogenetic positions of vertebrate-associated negative-sense RNA viruses within the broader diversity of RNA viruses. Phylogenies were estimated using a maximum likelihood method and midpoint-rooted for clarity only. Viruses discovered here are labelled with solid black circles. The name of the major clade (phylogeny) is shown at the top of each tree, and taxonomic names are shown to the right. The vertebrate associated virus diversity is shaded in grey. All horizontal branch lengths are scaled to the number of amino acid substitutions per site.



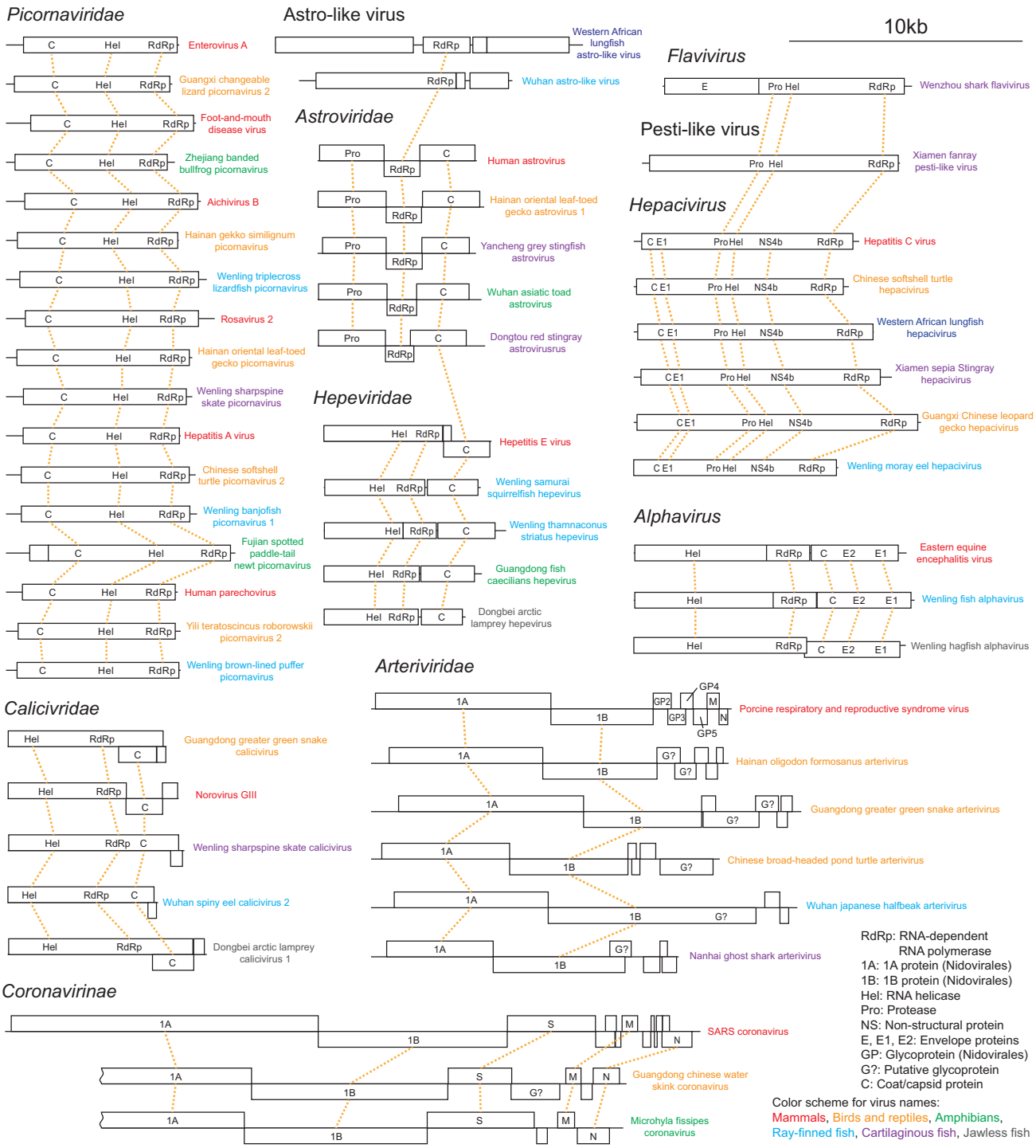
Extended Data Fig. 3 | The phylogenies of potentially new families of vertebrate-associated viruses. Viruses identified from vertebrate hosts are shaded with different colours. Sequences recovered from the Transcriptome Shotgun Assembly (TSA) database are marked with solid black diamonds, while those recovered from the Whole-Genome Shotgun (WGS) contigs database (that is, endogenous virus elements) are marked with open triangles. For vertebrate viruses, the relevant taxonomic and tissue information is provided in the sequence names.



Extended Data Fig. 4 | Evolutionary history of four groups of vector-borne RNA virus. Each phylogenetic tree was estimated using a maximum likelihood method. Within each phylogeny, the viruses newly identified here are marked with solid black circles, the vertebrate host groups are indicated by different colours, and the vector symbol is shown next to viruses known to be transmitted by vectors. The name of the virus family or genus is shown at the top of each phylogeny, and the lower level virus taxonomic names are shown to the right.



Extended Data Fig. 5 | Evolution of vertebrate-associated negative-sense RNA virus genomes. Representative genomes from negative-sense RNA virus families/genera are shown. The regions that encode major functional proteins or protein domains are labelled on each of the genomes. Homologous regions within each family are connected with orange dotted lines. Schematic phylogenetic relationships are shown next to the genomes diagrams. Coverage plots are shown underneath novel genome structures. Reverse-complementary sequences are shown for negative-sense RNA viruses with complete termini. A Sanger sequencing chromatogram is shown at a GC-rich hairpin-forming region of the Wenling frogfish arenavirus 2 genome, in which the coverage drops substantially. Host associations are labelled to the right of tree using solid circles with different colours. Host associations and abbreviation of functional domains are described at the bottom of the figure.



Extended Data Fig. 6 | Evolution of vertebrate-associated positive-sense RNA virus genomes. Representative genomes from positive-sense RNA virus families or genera are shown. The regions that encode major functional proteins or protein domains are labelled on each of the genomes. Homologous regions within or between viral families are connected by orange dotted lines. Host associations are reflected in the colour of the virus names. Host association colour schemes and the abbreviations of functional domains are described at the bottom of the figure.