Diversity and evolution of the animal virome

Erin Harvey[1,2,3] *and Edward C. Holmes*[1,2,3]✉

Abstract | The COVID-19 pandemic has given the study of virus evolution and ecology new relevance. Although viruses were first identified more than a century ago, we likely know less about their diversity than that of any other biological entity. Most documented animal viruses have been sampled from just two phyla — the Chordata and the Arthropoda — with a strong bias towards viruses that infect humans or animals of economic and social importance, often in association with strong disease phenotypes. Fortunately, the recent development of unbiased metagenomic next-generation sequencing is providing a richer view of the animal virome and shedding new light on virus evolution. In this Review, we explore our changing understanding of the diversity, composition and evolution of the animal virome. We outline the factors that determine the phylogenetic diversity and genomic structure of animal viruses on evolutionary timescales and show how this impacts assessment of the risk of disease emergence in the short term. We also describe the ongoing challenges in metagenomic analysis and outline key themes for future research. A central question is how major events in the evolutionary history of animals, such as the origin of the vertebrates and periodic mass extinction events, have shaped the diversity and evolution of the viruses they carry.

Metagenomic next-generation sequencing (mNGS). The parallel high-throughput sequencing of the total genetic material (RNA or DNA) extracted from a sample. This method offers scalability and speed that cannot be achieved by earlier sequencing technologies.

[1]*Sydney Institute for Infectious Diseases, University of Sydney, Sydney, NSW, Australia.*

[2]*School of Life and Environmental Sciences, University of Sydney, Sydney, NSW, Australia.*

[3]*School of Medical Sciences, University of Sydney, Sydney, NSW, Australia.*

✉*e-mail: edward.holmes@sydney.edu.au*

https://doi.org/10.1038/s41579-021-00665-x

Viruses are the most diverse and abundant biological entity, infecting species from all of life's domains, regularly jumping to new hosts, and occasionally causing serious disease[1,2]. Although the diseases that we now know are caused by viruses have been documented for millennia, viruses were not formally identified until the late 1800s[3]. The first viruses were discovered in the context of strong disease phenotypes, and for much of its history virology was heavily biased towards research on viruses associated with overt disease, particularly from plants and animals of direct human relevance[4]. This has changed with advances in metagenomic next-generation sequencing (mNGS), which has enabled a broader characterization of virus diversity[5–9]. Yet despite these technological developments, our understanding of animal viruses remains strongly skewed towards those infecting a relatively small number of taxa (FIGS 1,2). In addition, as metagenomic datasets continue to grow in both size and complexity, so does the challenge of their analysis[10].

The development of increasingly large-scale and affordable mNGS technologies has ushered in a new age in our understanding of the diversity of the viral universe — the so-called virosphere — and the evolutionary and ecological processes that give rise to it. Paradoxically, however, the more animal viruses that are sequenced, the clearer it has become that most of this immense virosphere remains uncharacterized[7,11]. Few of the more than 1.5 million species within the kingdom Animalia have been surveyed for viruses, and most of those characterized come from a single phylum — the Chordata. Similarly, because mosquitoes and ticks are common disease vectors, most virological studies of invertebrates have focused on the Arthropoda, although this is just 1 of 21 invertebrate phyla[12–14] (FIG. 2). In addition, many metagenomic studies of animal viromes largely involve cataloguing the viral diversity present in the species in question. Although an important first step, by designing appropriate sampling schemes, metagenomic data can also address specific hypotheses on the evolutionary and ecological factors that shape the structure of viromes[15,16].

In this Review, we explore our current knowledge of the structure, diversity and evolution of the animal virome, particularly since the advent of mNGS. As most recent data have been generated by total RNA sequencing (also called 'metatranscriptomics'), we necessarily devote the greatest attention to the diversity and evolution of RNA viruses, although in many cases similar conclusions can be drawn for viruses with DNA genomes. A key message is that profound sampling biases have restricted our understanding not only of virus biodiversity but also of fundamental aspects of virus evolution. We argue that placing those viruses that cause zoonotic disease in humans in the context of a wider


a
Virus superclade or family
Zhaovirus
Yueviridae
Yanvirus
Weivirus
Toti–Chryso clade
Tombus–Noda clade
Solinviviridae
Sinhaliviridae
Retroviridae
Reoviridae
Qinviridae
Potyviridae
Pneumoviridae
Picorna–Calici clade
Permutotetraviridae
Partiti–Picobirna clade
Orthomyxoviridae
Olifoviridae
Nido clade
Narna–Levi clade
Nanhypoviridae
Nanghoshaviridae
Mono–Chu clade
Metaviridae
Matonaviridae
Luteo–Sobemo clade
Levi–Narna clade
Kitaviridae
Hepe–Virga clade
Hepadnaviridae
Hantaviridae
Gresnaviridae
Flaviviridae
Deltavirus
Carmotetraviridae
Bunya–Arena clade
Birnaviridae
Astroviridae
Aspiviridae
Amnoonviridae
Smacoviridae
Redondoviridae
Poxviridae
Polyomaviridae
Polydnaviridae
Parvoviridae
Papillomaviridae
Nudiviridae
Nimaviridae
Iridoviridae
Hytrosaviridae
Herpesviridae
Genomoviridae
Geminiviridae
Cruciviridae
Circoviridae
Bidnaviridae
Baculoviridae
Bacilladnaviridae
Asfarviridae
Ascoviridae
Anelloviridae
Alloherpesviridae
Adomaviridae
Adintoviridae
Adenoviridae
Chordata
Echinodermata
Nematoda
Arthropoda
Rotifera
Platyhelminthes
Mollusca
Annelida
Sipuncula
Cnidaria
Porifera
Ctenophora
Placozoa
b
Zhaovirus
Weivirus
Toti–Chryso clade
Tombus–Noda clade
Smacoviridae
Retroviridae
Reoviridae
Redondoviridae
Qinviridae
Poxviridae
Polyomaviridae
Pneumoviridae
Picorna–Calici clade
Phenuiviridae
Peribunyaviridae
Parvoviridae
Partiti–Picobirna clade
Papillomaviridae
Orthomyxoviridae
Olifoviridae
Nido clade
Narna–Levi clade
Nanhypoviridae
Nanghoshaviridae
Nairoviridae
Mono–Chu clade
Matonaviridae
Luteo–Sobemo clade
Levi–Narna clade
Iridoviridae
Herpesviridae
Hepe–Virga clade
Hepadnaviridae
Hantaviridae
Gresnaviridae
Genomoviridae
Geminiviridae
Flaviviridae
Deltavirus
Circoviridae
Bunya–Arena clade
Birnaviridae
Bacilladnaviridae
Astroviridae
Asfarviridae
Arenaviridae
Anelloviridae
Amnoonviridae
Alloherpesviridae
Adomaviridae
Adintoviridae
Adenoviridae
Mammalia
Aves
Reptilia
Amphibia
Sarcopterygii
Actinopterygii
Chondrichthyes
Cyclostomata
Ascidiacea
Leptocardii
Amnitoe
(amniotic egg)
Tetrapod (four legs)
Four limbs
Lungs
Gnathostomes (jaw)
Evolution of adaptive immunity
Vertebrata (vertebral column)
Craniata (head)
Chordates (notochord)

◀ Fig. 1 | **Phylogenetic diversity of animal viruses.** Schematic phylogenies showing each phylum within the kingdom Animalia (part **a**) and each animal class within the Chordata (part **b**), as well as the major events and traits acquired during chordate evolutionary history. In both part **a** and part **b**, the virus families and clades associated with each animal group are shown as identified from US National Center for Biotechnology Information (NCBI) GenBank nucleotide accession numbers. The animal phyla are those used for virus host taxonomy assignment within GenBank and the phylogeny is based on REFS[12,13]. The figure is reliant on the host species assigned to a given virus sequence in the NCBI GenBank sequence database, such that these associations may not have been experimentally verified.

sampling of animal viromes provides a more nuanced view of the frequency of host-jumping and emergence events, and hence assessments of zoonotic risk. We also give special emphasis to a central but rarely addressed question: whether major events in animal evolution — moments of evolutionary 'transition' such as the origin of the vertebrates or of adaptive immunity — also changed the phylogenetic diversity of the viruses that infect these species.

Diversity, composition and evolution of the animal virome

Metagenomics has widened the aperture through which we can view the diversity of the animal virome. Total RNA sequencing has enabled the rapid and comprehensive identification of viruses without the use of time-consuming and restrictive steps of cell culture or microscopy[5,17–20] (BOX 1). These studies have shown that animals are infected by viruses spanning the full range of genome types (that is, single-stranded RNA, in both positive-sense and negative-sense orientations, double-stranded RNA, retroviruses, single-stranded DNA and double-stranded DNA) as well as viruses with both segmented and unsegmented genomes. According to a recent (July 2021) classification by the International Committee on Taxonomy of Viruses, animal viruses can be placed into 5 (of 6) realms, 5 (of 10) kingdoms, 11 (of 17) phyla, 26 (of 39) classes, 36 (of 59) orders and 99 (of 189) families[21].

However, despite the broadening of species sampling through mNGS, our knowledge of the animal virome is still dominated by viruses associated with humans or human activities. As an illustration, ~75% of animal virus entries in the US National Center for Biotechnology Information nucleotide sequence database derive from humans, and most of the animal entries are from species of anthropogenic significance, either as disease hosts or vectors, or those of economic or social importance (FIG. 2). Major sampling biases mean that there are also marked differences in the extent and pattern of the diversity of viruses associated with different animal groups, such as different phyla or vertebrate classes (FIG. 1). The greatest diversity of known viruses resides within the vertebrates, closely followed by arthropods, with the phylum Mollusca a distant third. It is no coincidence that these phyla contain anthropogenically significant species, such as vectors of disease in the case of arthropods and farmed shellfish in the case of molluscs. Other phyla have evidently been sampled far less frequently. For example, as viruses are ubiquitous within the environment, it is unlikely that there is truly a lack of viruses infecting phyla such as the Placozoa (FIG. 1). Similarly, recent explorations of the fish virome have revealed a multitude of novel DNA and RNA viruses, with virus families previously only described in mammals or birds now also found in fish, indicative of their antiquity[22–29] (FIG. 3). Of the 37 families and clades of viruses found in mammals, 27 are also found in ray-finned fish (the Actinopterygii; FIG. 1). That these virus families and clades are seemingly absent from phylogenetic 'intermediate' taxa (such as Amphibia and Sarcopterygii) is again likely a signature of inadequate sampling (FIG. 1).

Our limited knowledge of virus biodiversity has been put into sharp focus by the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, in late 2019 (REFS[30,31]). Ongoing metagenomic studies are beginning to identify a wealth of animal coronaviruses. Although these animals include rodents[32], the most notable hosts are arguably bats of the genus *Rhinolophus* (horseshoe bats), which are commonplace in China and parts of South-East Asia[33,34] as these sometimes carry viruses closely related to SARS-CoV-2 (FIG. 3). However, while it is probable that both bats and rodents harbour the greatest diversity of coronaviruses, this picture is very likely distorted by major sampling biases, as these two mammalian groups are also popular subjects of metagenomic studies due to their known role as reservoirs for a range of human infectious diseases. Indeed, as SARS-CoV-2 can infect and be transmitted among many animal species, resulting in large outbreaks in farmed mink[35] with transmission back to humans[36], and even reports of high virus prevalence in white-tailed deer in the USA[37], it is unlikely that the natural ecology of viruses closely related to SARS-CoV-2 involves only bats and pangolins[38,39].

Recent studies of other coronaviruses (that is, members of the family *Coronaviridae* of positive-sense RNA viruses) similarly provide informative examples of how metagenomic sequencing is leading to a new perspective on the diversity and antiquity of animal viruses. Historically, most attention has been directed towards those coronaviruses associated with mammals as these are most likely to emerge in humans[40]. However, a combination of mNGS and transcriptome database mining has led to the identification of divergent coronaviruses in a broader range of vertebrates, including amphibians and fish[28,41] (FIG. 3). Perhaps most surprising was the discovery of coronaviruses in a jawless vertebrate — the pouched lamprey (*Geotria australis*) from New Zealand[28]. Rather than falling basal to other vertebrate coronaviruses on a phylogenetic tree, as might be expected if they had co-diverged with their vertebrate hosts, the pouched lamprey viruses fell within the diversity of fish coronaviruses, highlighting the occurrence of host-jumping in aquatic environments[28] (FIG. 3). As appears to be true of many virus families, the evolutionary history of the coronaviruses reflects a combination of virus–host co-divergence that likely covers the entire evolutionary history of vertebrates over hundreds of millions of years and relatively frequent cross-species virus transmission among animals that inhabit the same environment and that can sometimes result in disease emergence.

Virosphere
The total assemblage of RNA viruses and DNA viruses on Earth, infecting hosts of any type.

Viromes
Total assemblages of viruses in individual organisms or species.

Metatranscriptomics
The study of the total expressed RNA — the transcriptome — within a sample. The RNA can be derived from expressed host genes as well as microbial species within the host, including both RNA virsuses and DNA viruses.

Zoonotic disease
An infectious disease that can be transmitted from animals to humans.

Emergence
Process by which novel infectious diseases (or pathogens) appear in species or previously known diseases rapidly increase in incidence or geographical range. Often associated with cross-species transmission.

Metagenomics
The simultaneous sequencing of all genetic material within a sample, including all the microorganisms present. It can involve the analysis of individual marker genes such as 16S or 18S ribosomal RNA or complete genomes.

Co-divergence
Evolutionary pattern in which the phylogenetic history of a virus or other pathogen matches that of the host organisms on long evolutionary timescales.

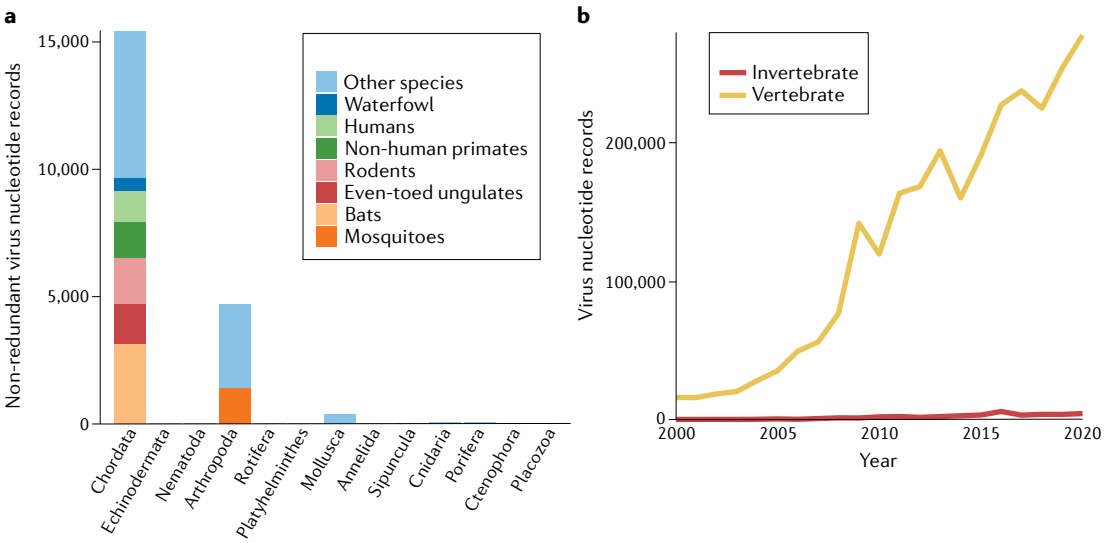


Fig. 2 | **Virome sequencing by animal phylum. a** | Graphical representation of the number of unique virus nucleotide entries in the US National Center for Biotechnology Information (NCBI) GenBank nucleotide sequence database sorted by virus species and host species showing that viruses associated with chordates far outnumber those from all other animal phyla. The proportions of these entries assigned to hosts of note are shown in different colours. Duplicate entries were excluded. **b** | Graphical representation of the rapid increase in vertebrate-associated virus entries in the NCBI GenBank sequence database over the past two decades and the comparatively low numbers of invertebrate-associated viruses identified over the same period.

An even more dramatic story can be told for hepatitis D virus (HDV). Until recently, HDV was described only in humans and in close association with human hepatitis B virus (HBV), performing an essential 'helper' role in its replication. The intimate relationship between HDV and HBV led to theories that HDV evolved in humans, perhaps as an escaped host gene[42]. However, recent metatranscriptomic studies have revealed that viruses closely related to HDV infect other vertebrates (mammals, birds, fish, snakes and amphibians) as well as a number of invertebrates[43–46] and in the absence of HBV-like viruses such that other viruses may act as helpers[46]. Similarly, it has traditionally been assumed that influenza viruses (family *Orthomyxoviridae*) are largely restricted to water birds of the orders Anseriformes and Charadriiformes, which act as reservoirs for their occasional emergence in mammals[47,48]. However, recent metagenomic studies have identified influenza virus-like viruses in fish, amphibians and even jawless vertebrates (that is, hagfish), and these viruses share common ancestry with a diverse set of invertebrate viruses[6,9]. Hence, as is true of many virus groups, the influenza viruses have a far older and more complex evolutionary history than previously envisaged[25] (FIG. 4). Indeed, the broader viral order *Articulavirales* of negative-sense viruses also contains divergent viruses sampled in fish as well as those from a variety of invertebrate species[5].

One fascinating insight from mNGS studies of animal viromes has been the recognition that invertebrates commonly carry a far greater diversity and abundance of viruses than vertebrates, in accord with their huge species numbers. In particular, large-scale metagenomic studies of invertebrates have uncovered novel virus families and genera, as well as viral lineages previously thought to be restricted to vertebrates[5,17,49–51]. These studies have similarly identified a wide diversity of novel genome structures in invertebrate viruses, in turn revealing that viral genome evolution is more fluid and dynamic than previously envisaged[5,17] (see later).

The first glimpse of the true breadth of the invertebrate virome came from a study of negative-sense RNA viruses in arthropods[17]. This was extended to cover other types of RNA virus in a broader range of invertebrate taxa[5], eventually leading to a myriad of metagenomic studies[52–56]. More recently, metagenomic studies have begun to focus on individual invertebrate species, such as flies of the genus *Drosophila*[57,58] and various species of mosquito[54,59–61]. Although these studies still reflect a limited sample of animals from the commonplace, easy to obtain and sometimes scientifically important arthropods, it is evident that viruses are copious in many invertebrate taxa. Indeed, some invertebrate RNA viruses reach abundance values as high as 87% of the non-ribosomal RNA reads in a single sequencing dataset[5]. That invertebrate species can possess such high virus abundance with no clear signs of disease (although these may be difficult to identify in such short-lived animals) further suggests that many of these viruses may be commensal and tolerated by their invertebrate hosts. Finally, not only are invertebrate viruses diverse but they often fall as basal lineages on phylogenetic trees of animal viruses, implying that they have ancient associations with animals[62,63]. Indeed, it is likely that many virus families will have an evolutionary ancestry that dates at least to the origin of vertebrates and perhaps even to the origin of animals.

Genome plasticity of animal viromes

The genome structures of animal viruses are characterized by a remarkable plasticity, reflected in major differences in genome length, genome organization

Multicomponent viruses
Also referred to 'multipartite viruses'. Viruses in which the genome segments are contained within separate virus particles. These are relatively commonplace in positive-sense RNA viruses of plants such as members of the *Bromoviridae*.

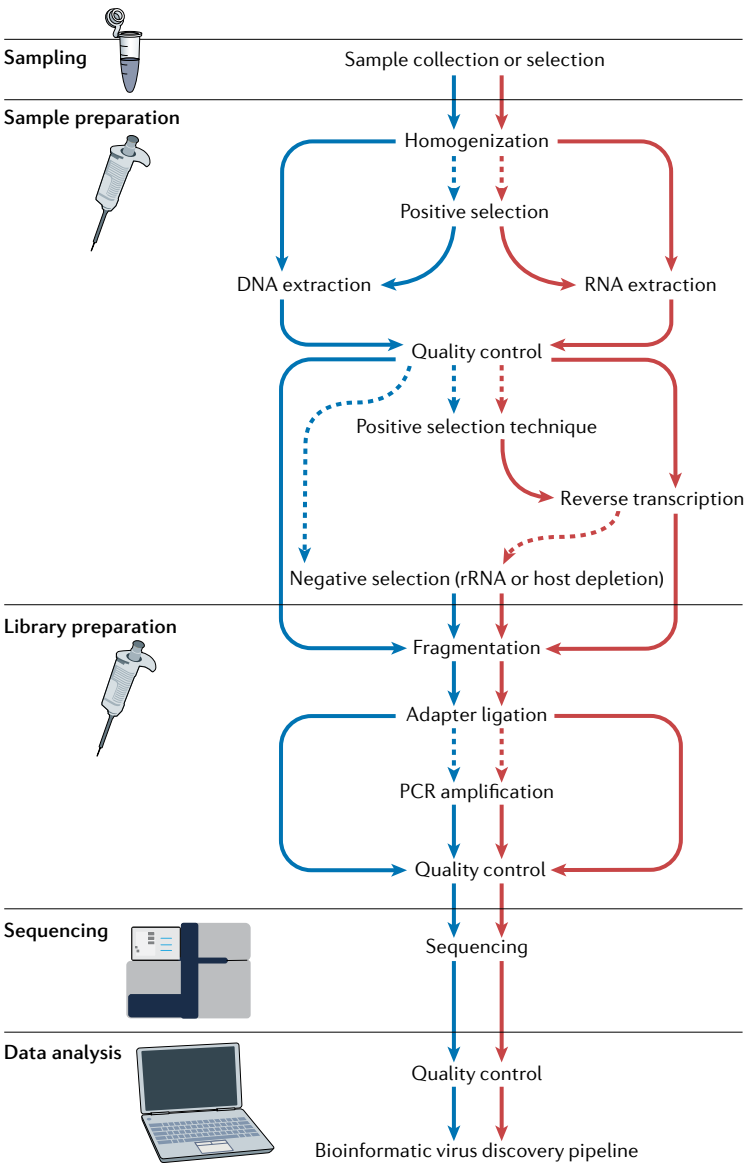
(for example, the number and orientation of genes) and the number of genome segments present in specific virus families (FIG. 5). Traditionally, individual families of RNA viruses were thought to possess characteristic patterns of segmentation, with those containing multiple segments (such as members of the *Orthomyxoviridae*) generally considered as constituting phylogenetic groups distinct from those characterized by a single segment. Metagenomic data have drastically changed this picture. It is now clear that genome segmentation has been gained and lost multiple times in evolutionary history, with the RNA virus orders *Nodamurales* and *Monjiviricetes* providing important examples[5,17] (FIG. 5). Similarly, the number of segments in the *Articulavirales* ranges from 4 to 10 (FIG. 4).

Of particular importance is that invertebrate viruses often have more complex genome structures than their vertebrate counterparts. A good example is presented by the *Flaviviridae*, a family of single-stranded, positive-sense RNA viruses that includes dengue virus, Zika virus and hepatitis C virus. All these familiar human pathogens are characterized by an unsegmented genome encoding a single polyprotein. Although this simple genome structure was once considered archetypal, the discovery of 'flavi-like' viruses with far more complex genome structures in a range of invertebrate taxa, such as Jingmenvirus from ticks, presents a very different picture[6,64] (FIG. 5). The jingmenviruses comprise four or five segments, two of which show sequence similarity to the non-structural proteins NS5 and NS2B–NS3 of the *Flaviviridae*[64]. The two remaining segments exhibit no sequence similarity to known virus genes but likely encode structural proteins. Remarkably, these different segments may sometimes be associated with different virus particles, such that these viruses can be considered multicomponent viruses — a pattern of genome organization commonly seen in positive-sense RNA viruses of plants[65]. More dramatically, the recently discovered *Chuviridae* family of

Box 1 | Metagenomic next-generation sequencing for virus discovery

'Metagenomics' describes the high-throughput sequencing of the total nucleic acids (DNA or RNA) extracted from a sample, including water, soil or plant and animal tissues[138]. Whereas metagenomics has traditionally been associated with DNA sequencing, metatranscriptomics — total RNA sequencing — is now commonly used in virological studies. Metatranscriptomics is particularly useful for characterizing the animal virosphere as it detects all the organisms that are transcribing RNA in the sample, including the RNA viruses, which are excluded from DNA sequencing. Although a metatranscriptome will include host RNA transcripts, it necessarily excludes the bulk of the host genome, providing additional power for pathogen detection[120,121].

The preparation of nucleic acid samples for next-generation sequencing is termed 'library preparation', and involves fragmentation of input material, ligation of sequencing adapters and PCR amplification (see the figure; DNA metagenomics in blue on the left and RNA metagenomics in red on the right). At this stage, positive (enrichment) or negative (depletion) selection steps can be taken to target the sequencing output towards a desired genetic material, although all currently available techniques have significant limitations. In metatranscriptomics, depletion or enrichment is necessary as 'host' sequences account for the bulk of transcripts within any sample and mask the presence of virus transcripts that are at lower abundance[139]. Filtration is performed before library preparation and is used to select 'virus-sized' particles (see the figure), although this technique also removes all large virus particles. Similarly, ultracentrifugation can be used to select virus particles on the basis of their density, although this technique has a number of limitations, including cost, contamination risk and sample size restrictions[4]. Library preparation enrichment steps rely on sequence-based selection or nuclease treatments, such as VirCapSeq, which uses biotinylated oligonucleotides to capture known (or closely related) virus sequences[140]. Virion enrichment involves the depletion of unencapsulated nucleic acids, utilizing the virus capsid. Importantly, comparative studies have shown that virus-specific selection steps reduce the diversity of viruses detected, such that ribosomal RNA (rRNA) depletion and bioinformatic filtering of virus sequences remains the most unbiased and hence comprehensive approach[141].



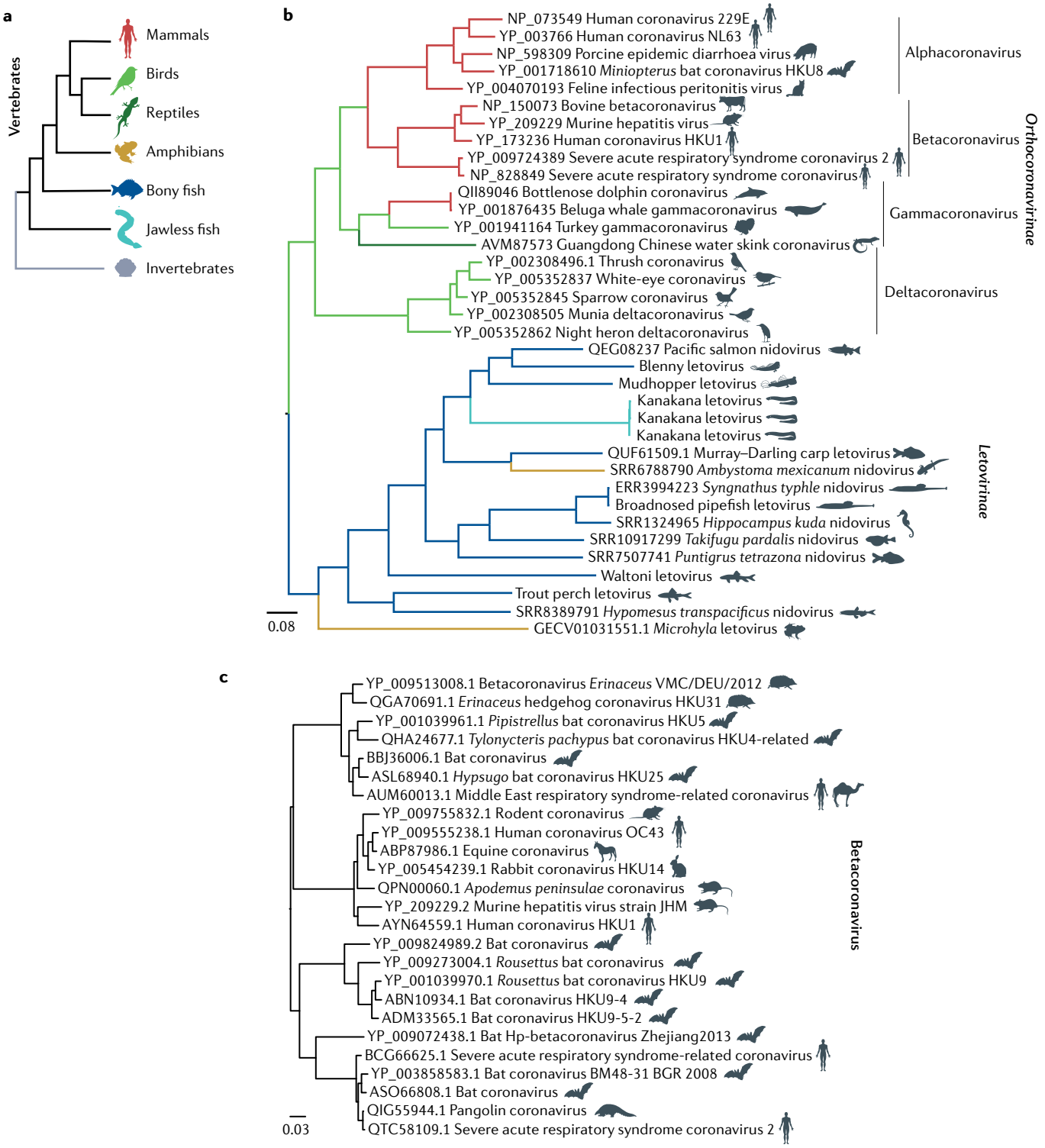


Fig. 3 | **Recent phylogenetic and genomic expansion of the coronaviruses.** The figure illustrates how a combination of virus–host co-divergence and sporadic host-jumping has shaped the evolutionary history of the family *Coronaviridae*. The phylogenetic history of the major host taxa (part **a**) is broadly reflected in the phylogeny of the subfamilies *Coronavirinae* and *Letovirinae* (part **b**), with the former largely associated with mammals and the latter with fish and other aquatic animals. The major host taxon is indicated in part **b** by the branch colour corresponding to the host group shown in part **a**, and the host species is indicated by the animal silhouette at the tree tip. An expanded maximum likelihood phylogeny of the genus Betacoronavirus containing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (part **c**) with animal silhouettes at the tree tips showing that most of these viruses are associated with bats, which are important reservoir hosts for these viruses. The phylogeny was estimated using ORF1ab protein using IQ-TREE[137] and was midpoint rooted for clarity. The scale bars depict the number of amino acid substitutions per site. Parts **a** and **b** adapted from REF.[28], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

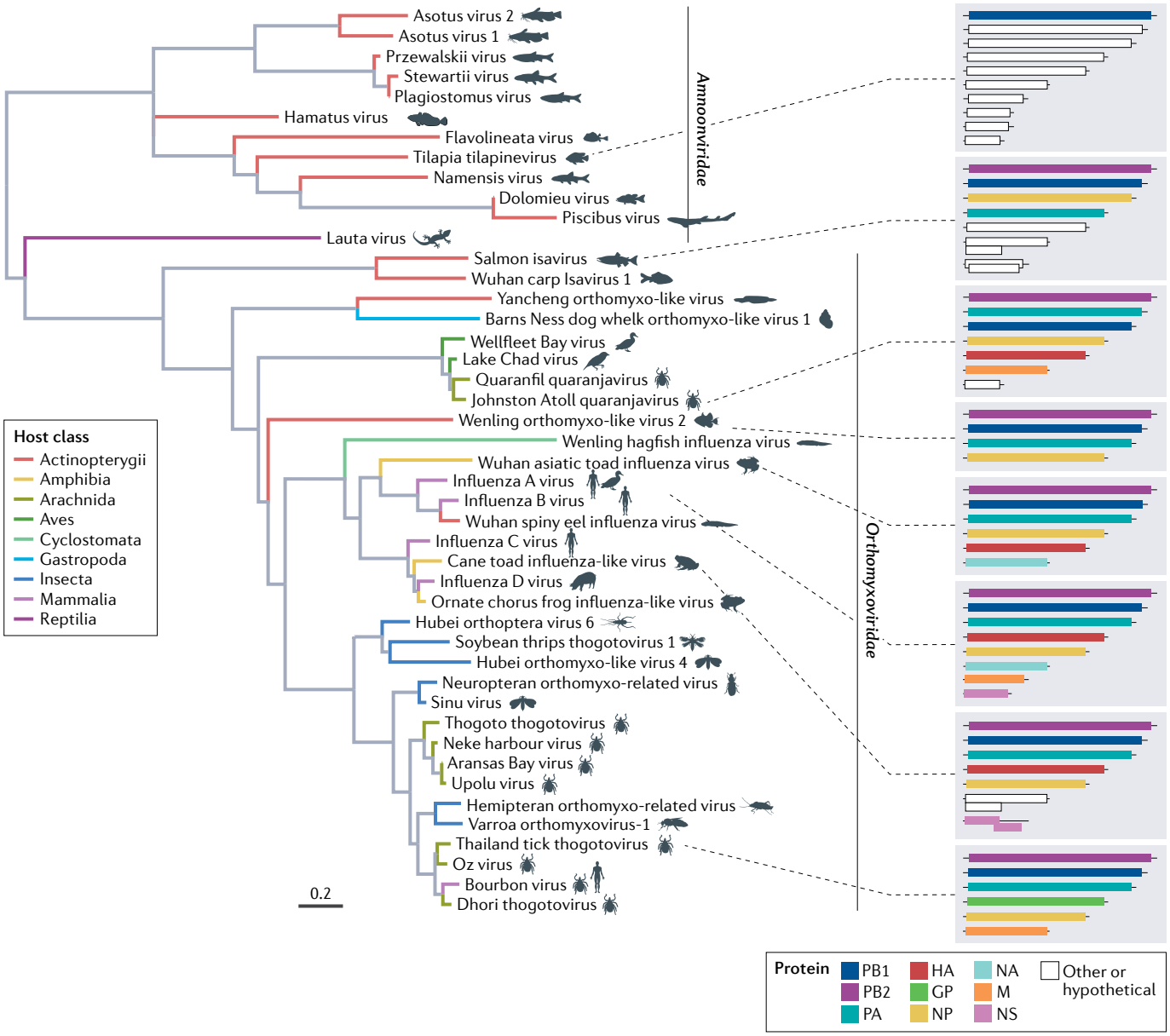


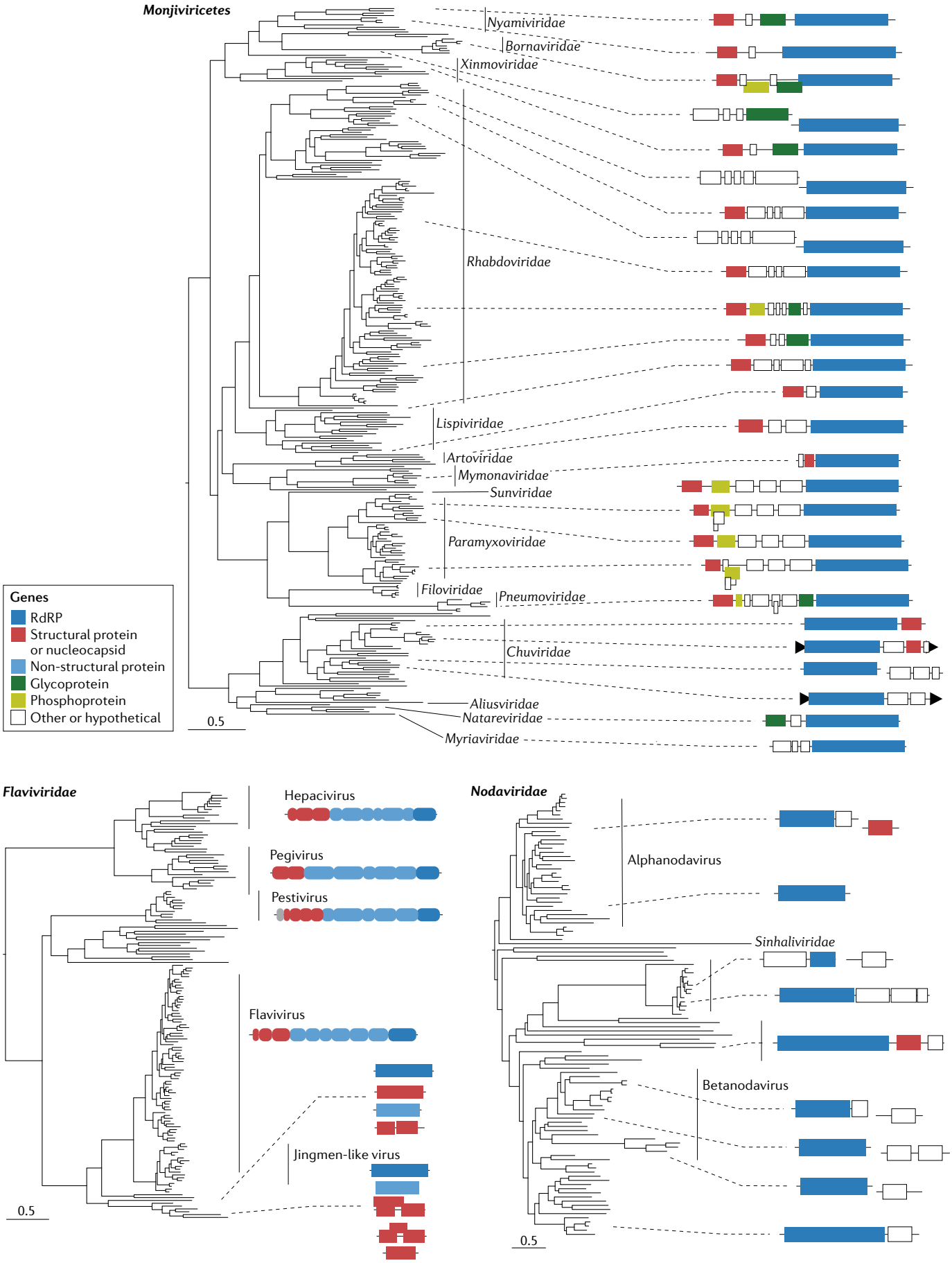
Fig. 4 | **The complexity of host associations in virus evolution.** Phylogeny of the order *Articulavirales* (negative-sense RNA viruses that include the influenza viruses from the family *Orthomyxoviridae*) showing the diverse set of animal hosts infected and the complex virus–host associations. As with many virus groups, this phylogenetic pattern is indicative of a history of cross-species transmission set on a background of ancient virus–host co-divergence. The animal host group is indicated by the colour of the terminal branch and the host is indicated by an animal silhouette. The maximum likelihood phylogeny (IQ-TREE[137]) was inferred using amino acid sequences of the protein PB1 (or equivalent) and was midpoint rooted for clarity. The scale bar depicts the number of amino acid substitutions per site. Virus genome structures, with segment lengths drawn to scale, are indicated where available, illustrating the variation in genome structure and segment number.

negative-sense RNA viruses contains viruses with unsegmented, bisegmented and even circular RNA genomes[22] (FIG. 5). To date, this fascinating group of viruses has been described in arthropods, nematodes and reptiles[5,17,66].

To evaluate whether any reduction in genome complexity is associated with the evolution of vertebrates will require a broader sampling of animals. One attractive, although untested, theory is that shorter genomes are selectively advantageous in vertebrates because fewer potential immune targets would be presented to hosts with more advanced adaptive immune responses. Testing this hypothesis will first require more detailed knowledge of the viromes of animal lineages that diverged close to the evolution of adaptive immunity.

Has host evolution shaped virus evolution?

As genome sequence data from animal viruses continue to accumulate, they can be used to address broader evolutionary questions. Viruses, by definition, have obligate associations with their hosts. Accordingly, changes in the


Monjiviricetes
Nyamiviridae
Bornaviridae
Xinmoviridae
Rhabdoviridae
Lispiviridae
Artoviridae
Mymonaviridae
Sunviridae
Paramyxoviridae
Filoviridae
Pneumoviridae
Chuviridae
Aliusviridae
Natareviridae
Myriaviridae
0.5
Genes
RdRP
Structural protein or nucleocapsid
Non-structural protein
Glycoprotein
Phosphoprotein
Other or hypothetical
Flaviviridae
Hepacivirus
Pegivirus
Pestivirus
Flavivirus
Jingmen-like virus
0.5
Nodaviridae
Alphanodavirus
Sinhaliviridae
Betanodavirus
0.5

◀ Fig. 5 | **The evolutionary flexibility of RNA virus genomes.** To illustrate the genome flexibility in RNA virus evolution in animals, phylogenies of the order *Monjiviricetes* and the families *Nodavidae* and *Flaviviridae* are labelled with representative genome structures. Genome structures differ in size, organization and number of segments. Key genes are indicated by different colours, and the relative length of the coding regions is indicated by size. Boxes positioned below the centreline of the genome indicate overlapping open reading frames and black triangles at the ends of a structure indicate circularization. In the case of genomes within the *Flaviviridae*, boxes with rounded corners indicate individual proteins within the single polyprotein that characterizes many members of this family. Notably, genome segmentation has been gained and lost multiple times during the evolution of the *Monjiviricetes* and *Nodaviridae*, and has evolved once within the family *Flaviviridae*, specifically in the jingmenviruses associated with invertebrates. In each case, maximum likelihood phylogenetic trees (IQ-TREE[137]) were estimated using the RNA-dependent RNA polymerase (RdRP; NS5 or NS5-like protein for the *Flaviviridae*). All trees were midpoint rooted for clarity only. The scale bars depict the number of amino acid substitutions per site.

number and diversity of host species through time are also expected to impact the number and diversity of the viruses they carry, albeit likely in a complex manner. A central issue is therefore whether and how the structure and phylogenetic diversity of animal viromes have been impacted by major events in the evolutionary history of their animal hosts. Although there has been some interest in documenting the generation, or 'birth', of lineages within individual virus species as this is central to the process of disease emergence[2,67], aside from a limited number of phylogenetic studies[68] and those examining local populations[69], far less is known about the rates and mechanisms of virus birth and death (that is, lineage extinction) on evolutionary timescales. We hypothesize that major events in the evolution of animals — key evolutionarily transitions — are likely to have had a major impact on the evolution of the viruses they harbour. To the best of our knowledge, no studies directly addressing this question have been undertaken to date, although similar work has been performed on other systems. For example, the diversification of pathogenic *Bartonella* bacteria has been proposed to reflect the expansion of the mammals[70].

The evolution of the Metazoa more than 600 million years ago resulted in a huge increase in phenotypic diversity, eventually leading to the myriad of animal phyla that we see today. Similarly, there was a massive increase in the phenotypic diversity of animals concurrent with the origin of the Chordata more than 500 million years ago[71], while the evolution of jawed vertebrates (Gnathostomata) approximately 450 million years ago was associated with multiple rounds of full genome duplications and the evolution of adaptive immunity[72] (FIG. 1). It seems inevitable that these major events in host evolution will have had a profound impact on the extent, diversity and composition of the viruses the hosts carry. Major questions in this context include whether the evolution of new types of host cell led to a rise in virus diversity, and whether the evolution of adaptive immunity led to the extinction of many viral lineages and hence a marked reduction in diversity. It is tempting to speculate that the apparent reduction in virus abundance levels in vertebrates compared with invertebrates[7] (see earlier) in part reflects the evolution of adaptive immunity (FIG. 1). Similarly, the earlier evolutionary transition to multicellularity would have greatly increased the number and diversity of hosts cells, and their receptors, for viruses to infect.

Other events in host evolution may also have led to major reductions in virus diversity. Probable examples include mass extinction events[73], such as those that occurred at the Permian–Triassic boundary approximately 250 million years ago resulting in the loss of more than 80% of all marine species and ~70% of terrestrial vertebrate species[74], and the Cretaceous–Paleogene extinction event approximately 66 million years ago, which massively reduced the number of tetrapods and resulted in the extinction of non-avian dinosaurs[75]. Similarly, an overall decline in host population size and density coincident with the evolution of the vertebrates would have increased the impact of stochastic effects on virus populations subject to weaker natural selection[76]: with fewer potential hosts to infect, viral lineages would be expected to be lost more frequently leading to stronger genetic drift.

When sufficient data become available, a detailed phylogenetic analysis of animal viruses will provide meaningful insights into how host evolutionary transitions might have influenced the long-term macroevolution of viruses. The drastic reduction in the number of animal species associated with mass extinction events should be visible in the species distribution of viral lineages on phylogenetic trees. The first insights may come from comparisons of vertebrate and invertebrate viruses, particularly whether some viruses are restricted to either host type, or whether there is a marked phylogenetic gap between vertebrate and invertebrate viruses on phylogenetic trees of individual virus families that signifies a major transition in virus diversity. A provisional analysis of the limited and highly biased data currently available reveals that 16 of the 66 family or multifamily 'superclades' of viruses[9,17] are associated with vertebrates alone, whereas 17 are found in invertebrates with no vertebrate counterpart (FIG. 1). Broader investigations of this type should be a research priority.

Linking virus emergence to virus evolution

The phylogenetic analysis of virus orders, families and genera sits at the heart of studies of the diversity of viromes and their evolution[77]. On the one hand, there is often a broad congruence between the phylogenies of viruses and their animal hosts, with, for example, viruses sampled from fish and jawless vertebrates tending to fall in more basal phylogenetic positions than those sampled from mammals and birds (FIGS 3,4). Hence, these phylogenetic trees generally depict evolutionary events, particularly virus–host co-divergence, that have taken place on timescales of millions of years. Conversely, these phylogenetic analyses also reveal that cross-species virus transmission to new hosts has been commonplace throughout animal evolution[78]. In the short term, this same process of host-jumping is responsible for the emergence of novel pathogens such as SARS-CoV-2 (REFS[79–81]), with the vast majority of human viruses appearing in this way[2]. Indeed, disease emergence events occur over observable human history, and on timescales that are far shorter than depicted in most phylogenetic studies[82]. Hence, there is necessarily a marked temporal

Genetic drift
The change in frequency of a mutation in a population due to the chance effect of random sampling. Although genetic drift occurs in all populations of finite size, its effect is strongest in small populations.

Cross-species transmission
Also referred to as 'host-jumping' or 'host-switching'. The transmission of a virus from one host species to another.

Ectoparasites
Parasitic organisms that live on the skin of the host (rather than within a host), from which they derive their energy.

disconnect between evolutionary studies of animal viromes, such as those described in the preceding sections, and the timescale of disease emergence[11]. This in part explains why we still know little about the frequency with which host-jumping occurs in nature, or the rate at which cross-species transmission events are successful compared with those that die out[83].

Understanding the drivers of disease emergence on short timescales provides a means to link virus microevolution, as happens within populations, with virus macroevolution as reflected in broad-scale phylogenetic analyses. The historical domestication of animals and the development of animal husbandry provided many opportunities for viruses to jump to humans, with the emergence of measles virus from relatives (that is, rinderpest virus-like viruses) in cattle a likely case in point[84]. More recently, increased interactions with wildlife, following such factors as climate change, alterations in land use, the flourishing of live animal markets and the farming and trafficking of wild animals, have exposed the human population to novel pathogens, with urbanization, population growth and globalization allowing these emerging viruses to spread rapidly and far. Human immunodeficiency virus 1 (HIV-1) spread across Africa from its zoonotic origin in the Congo River basin region[67], and then to other continents, in part reflecting changes in colonial administration. By moving humans, animals and cargo great distances, air travel aided the spread of diseases and disease vectors into new environments. This includes the translocation of the *Aedes aegypti* mosquito from Africa to Asia and South America, enabling chikungunya virus, yellow fever virus, Zika virus and West Nile virus to establish animal transmission cycles in immunologically naive localities[85–87], and fuelling increasingly widespread outbreaks of Ebola virus infection in mammalian hosts[88]. Similarly, environmental changes such as increasing urbanization and climate change are leading to an increased prevalence of existing human pathogens such as yellow fever virus and dengue virus[85,86,89].

Deforestation forces wildlife into smaller, overlapping habitats, leading to new and greater interactions between and within species, fuelling disease spread[90,91]. Urbanization alters the way in which animals behave, changing their diets and interspecies and intraspecies interactions. Intensive farming creates opportunities for virus interspecies transmission and provides an environment in which a virus can spread rapidly through a population[92,93], with viruses moving from wildlife to domestic species as well between domestic animals. This is of special concern in poultry production, in which farmed birds regularly interact with wild birds, with virus transmission between them an occupational hazard. A powerful example is provided by the emergence of H5N1 avian influenza A virus in poultry and its subsequent zoonotic transmission to humans[94]. Backyard poultry populations within urban environments are of increasing concern as poultry-associated viruses such as Marek disease virus, infectious bursal disease virus and Newcastle disease virus (Avian orthoavulavirus 1) are being introduced into wild bird populations[91,95], and they also harbour multiple picornaviruses[96]. The reverse process is also possible, with viruses jumping from domestic animals to wildlife. The migration of humans and wildlife has similarly acted as a driver of disease emergence[97–100], with metagenomic studies revealing that very closely related animal viruses can be found in very diverse geographical regions[101]. A telling example is viruses associated with seabird ticks (*Ixodes uriae*) sampled as far apart as northern Sweden and the Antarctic peninsula, demonstrating that migratory birds and their ectoparasites can facilitate a global movement of viruses without human assistance[102].

It has often been proposed that RNA viruses have a higher rate of cross-species transmission and hence experience less frequent virus–host co-divergence than their DNA counterparts[2]. Although this is supported by large-scale comparative analyses, it is also the case that both DNA viruses and RNA viruses jump species boundaries more readily over evolutionary time, as reflected in phylogenetic comparisons, than might have been assumed[78]. Although most cross-species transmission events likely occur between animals that are relatively close in taxonomic space, such as among different species of mammals[77,82,103], some jumps may cover wide phylogenetic distances, including the possible transmission of hepadnaviruses from fish to mammals[22,104]. Again, sampling biases and data limitations make it difficult to draw precise conclusions on the frequency of cross-species transmission events in nature, although the more sampling that is done, the more examples are inevitably documented.

Metagenomics and zoonotic risk assessment

Determining the rate at which cross-species transmission events occur on epidemiological timescales of decades is of central importance in understanding disease emergence[103]. These data impact how we quantify zoonotic risk; that is, identifying those viruses with the potential ability to infect humans[105,106]. Before the metagenomic revolution, virus discovery studies in animals were focused on outbreaks with visible death and/or morbidity. As disease outbreaks in wildlife with low levels of death would generally not have been identified, a relatively high proportion of viruses appeared to be pathogenic[107]. However, the rebalancing of virome studies towards the sampling of seemingly healthy animals has shown that potentially pathogenic viruses may be more the exception rather than the rule, with studies of birds and bats important exemplars[107]. The broadening of animal sampling away from overt disease also changes the proportion of viruses that appear as potentially zoonotic, altering the denominator of emergence risk. Metagenomic studies have revealed that bats harbour a large and complex virome[18,20,33,108–111], with considerable discussion of the reasons why this might be so, particularly whether these animals possess immune systems that can tolerate a heavy burden of viral infection[73,112,113]. Although bats are implicated in the ultimate evolutionary origins of some important human viruses, only a tiny proportion of the huge number of bat viruses have ever successfully spread in humans, often entering our species via 'intermediate hosts', as appears to be true of some coronaviruses[40] (FIG. 3). The more bat viruses that

Spillover
The initial and sometimes transient appearance of a pathogen in a new species following a host jump. Can sometimes lead to a full-blown epidemic or pandemic.

are identified through metagenomic sequencing, so the relative frequency of those that are pathogenic and/or zoonotic declines.

The vast number of animal viruses described by metagenomics also complicates attempts to assess which of these will eventually emerge in humans[107,114]. There is no simple way to translate the long-term rates of virus evolution depicted in phylogenetic trees into short-term zoonotic risk assessments or pandemic predictions. Although revealing the diversity of the animal virome places newly emerged viruses into their true evolutionary context, it is arguably of less value for predicting whether some viruses have pandemic potential. There are many thousands of uncharacterized animal viruses that will differ in their natural propensity to infect humans. Large-scale metagenomic studies necessarily document virome composition in host species in a specific place at a particular point in time, often with little background ecological context. They should not be interpreted as exact descriptions of complete virome compositions in a species, particularly for hosts that occupy large geographical ranges, and do not necessarily inform on which viruses are able to emerge in humans. The snapshot of virus genetic diversity provided by metagenomics is also a static one in the face of the very rapid evolution of RNA viruses, which experience rates of nucleotide substitution approximately six orders of magnitude greater than those in their animal hosts[115]. The large-scale metagenomic sequencing of wildlife species will usually not identify the full spectrum of intrahost virus genetic variation, potentially missing low-frequency mutations that may facilitate host adaptation.

Most animal viruses sampled will lack some of the mutations they need to successfully replicate in and be transmitted among humans, with evolutionary optimization a necessity in the new host[116]. Hence, the vast majority of the viruses identified by metagenomic screening alone will have little chance of successfully spreading through human populations. As a topical case in point, although bat viruses that are closely related to SARS-CoV-2 have been identified, at the time of writing all those characterized lack an intact polybasic (furin) cleavage site at the S1–S2 junction in the virus spike protein that enhances human infectivity[117,118]. Similarly, although broad-scale screens have suggested that one of the closest relatives of SARS-CoV-2, virus RaTG13 sampled from *Rhinolophus affinis* bats in Yunnan province, China, had 'high zoonotic potential'[106], detailed virological studies revealed that this virus was unable to bind to the human ACE2 receptor[119]. Hence, although a potentially informative provisional screen, computational risk assessments of this kind may lack the precision necessary for actionable risk assessments. In addition, the identification of a virus sequence through metagenomics does not provide prima facie evidence that the virus can replicate in human cells, and evaluation of this key trait will require detailed experimental data, hugely increasing the associated costs and person hours.

Despite these limitations, the capacity of mNGS to detect the full range of microorganisms within a sample in a single run signifies a new age in clinical diagnostics[120,121]. In the same way, if not an exact prediction tool, mNGS will surely become a key component of future efforts for the surveillance for zoonotic pathogens at the human–animal interface. For example, to fully understand the emergence of SARS-CoV-2 and help prevent future epidemics, mNGS can be used to document the full host range of pathogens such as coronaviruses that seem best able to jump host species, and simultaneously reveal the barriers to cross-species virus transmission. As a case in point, a single study of a 1,100-hectare tropical botanical garden in Yunnan province, China, identified 24 novel bat coronaviruses, including close relatives of SARS-CoV-2 and of the animal pathogen porcine epidemic diarrhoea virus[39]. What other mammalian species within this single botanical garden carry coronaviruses are unknown, but a broader sampling of all the species in such an ecosystem will do much to reveal the patterns, rates and determinants of cross-species virus transmission at local scales.

The factors currently limiting the use of mNGS in studies of zoonotic risk assessment and disease emergence are that the technology detects only actively replicating viruses, is relatively expensive and generates a huge amount of data that require considerable computing power for detailed analysis. The deployment of metagenomics in resource-poor settings may therefore be challenging, even though these are the locations where humans likely interact most with wildlife species (as well as biting arthropods) and hence where the risk of virus spillover is perhaps greatest, and where approaches to reduce the exposure of humans to wildlife would likely have the greatest impact. In these instances, pathogen surveillance approaches based on immunological techniques, such as VirScan, which can be designed to detect past and present infection by hundreds of potential zoonotic pathogens with a single assay, represent a more practical solution[122]. Rather than recognizing only already known pathogens, approaches such as VirScan can in theory be extended to recognize peptides from those groups of viruses that are most likely to emerge in humans[107]. Given their past behaviour, the coronaviruses fall into this 'high-risk' category, as do the influenza viruses and the paramxyoviruses (within which the henipaviruses are an important example of an emerging threat[123]) and could be incorporated into broad-scale screening assays. Although such an approach will not capture all zoonotic viruses, it does provide some ability to detect potential threats.

Challenges and new research avenues

Although mNGS is transforming our understanding of animal viromes and their evolution, additional work is required on several fronts. We suggest that the priority for future sampling and sequencing should be those animal taxa that have been only poorly studied to date, particularly those that occupy key positions on the animal phylogeny, including those that mark evolutionary transitions. It will also be important to sample animals across their full range of habitats to determine whether virome structures differ substantially within individual

host species. Similarly, given the rapidity of RNA virus evolution, a priority should be to determine how virome structures within individual animal species change over time, for instance by annually sampling the same species at the same locations. More broadly, it is essential that future metagenomic studies of virus populations test explicit ecological and/or evolutionary hypotheses, such as exploring the impact of changing land use on virome structures, rather than simply presenting descriptive lists of the viruses present.

Host associations cannot always be relied upon in metagenomic studies, as viruses infecting symbionts, components of host diet, and contaminant microorganisms and laboratory reagents are also sequenced as part of the metagenome. For example, RNA virus families associated with plants, such as the *Tombusviridae* and *Luteoviridae*, are often detected in animal metagenomes as they are probably a dietary component, while the *Leviviridae*, a family of RNA bacteriophages, are likely associated with the microbial communities within animal hosts[124,125]. Clearly, erroneous host assignments may lead to erroneous conclusions on virus ecology and evolution. As a consequence, new bioinformatic tools are required that can accurately assign virus sequences to the true hosts, perhaps using statistical approaches that jointly consider levels of virus abundance and phylogenetic relationships. Although the analysis of dinucleotide frequencies provides a potential way to distinguish viruses infecting different host phyla, it is unable to provide a fine-scale host discrimination[126].

Future virome analyses will similarly be enabled by the development of methods that can identify highly divergent viral sequences, as it is clear that a large proportion of the virosphere comprises sequences that are so divergent from the sequences of known viruses that they are currently 'invisible' to discovery strategies based on sequence similarity alone[7]. Although this problem is particularly acute for host taxa that are the most divergent from the usual animal species usually considered in virus metagenomics studies, such as archaea, bacteria and basal eukaryotes, many animal taxa likely carry RNA viruses that are hidden within the 'dark matter' of uncharacterized sequences[127]. Arguably the simplest way to shed light on this hidden and likely diverse virosphere is through the detection and characterization of conserved protein structures as these retain the signal of homology and hence evolutionary relatedness for longer than primary sequences[128,129]. An informative example is provided by enveloped viruses, which require a protein capable of inducing the fusion of viral and cellular membranes for entry. Structural studies of multiple virus families have revealed that they fold into only three structural classes[130]. The amino acid sequences of these virus proteins show no detectable conservation among classes, and their relatedness is made apparent only through structural studies[131]. Fortunately, the 'resolution revolution' that has accompanied the development of cryo-electron microscopy has enabled the determination of more protein structures that are difficult to crystallize[132]. Hence, an important area for future research will be to use these structures to guide the identification of highly diverse viruses in metagenomic data, perhaps by determining the 'profiles' of physicochemical and structural features that distinguish virus proteins[133]. Detecting highly divergent viruses may also provide answers to some of the most profound questions in virus evolution, such as whether the absence of RNA viruses in archaea and their low frequency in bacteria is simply because they are too divergent in sequence to be detected[134].

Although the analysis of protein structure provides a potential means to reveal more of the diversity of the virosphere, it also presents a fundamental problem: that any novel viruses identified are so divergent in sequence that they cannot be incorporated into phylogenetic or other evolutionary analyses. This is even true in the case of the canonical RNA-dependent RNA polymerase, which is routinely used to infer multifamily phylogenies of RNA viruses (a variety of genes are used as phylogenetic markers in the DNA viruses). Even with currently available data, attempts to infer the evolutionary relationships among all extant RNA viruses are unconvincing, with pairwise identities in amino acid sequence alignments that are often less than expected by chance[135]. This raises the vexing question of how viable it is to infer a 'global' phylogeny of RNA viruses using sequence data alone. The most profitable approach may again involve methods that are able to accurately infer the distant evolutionary relationships on the basis of shared features of protein structure. Although these are not unsurmountable challenges, and the foundations of this approach have been laid[136], little productive work has been done in this area.

Conclusions

Metagenomic sequencing has radically changed our understanding of the diversity, structure and evolution of the animal virome, particularly in the case of RNA viruses. Yet it has also made the gaps in our knowledge more apparent than ever. As stressed throughout this Review, relatively little is known about the factors that shape virome structure outside anthropocentrically important species. Large-scale studies of a wider range of animal taxa are needed to provide a better understanding of the biological and phylogenetic diversity of viruses and the evolutionary and ecological processes that have given rise to it. Not only do we need to explain the large-scale patterns of virus diversity on evolutionary timescales, but to understand disease emergence and zoonotic risk it is essential to determine the factors that shape the ecology and evolution of viruses on shorter and more relevant timescales of years or decades, rather than millennia. Human activity is already leading to shifts in the diversity of the animal virome, although we usually see these effects only after they lead to a novel zoonotic event. Although metagenomics is shedding new light on the diversity of the virosphere, greater emphasis should be given to revealing the processes that determine cross-species transmission events among animals and hence that underpin disease outbreaks.

Published online 4 January 2022

1. Wasik, B. R. & Turner, P. E. On the biological success of viruses. *Annu. Rev. Microbiol.* **67**, 519–541 (2013).
2. Holmes, E. C. *The Evolution and Emergence of RNA Viruses* (Oxford University Press, 2009).
3. Loeffler, F. A. J. & Frosch, P. Berichte der Kommission zur Erforschung der Maul- und Klauenseuche bei dem Institut für Infektionskrankheiten in Berlin (G. Fischer, 1898).
4. Kumar, A., Murthy, S. & Kapoor, A. Evolution of selective-sequencing approaches for virus discovery and virome analysis. *Virus Res.* **239**, 172–179 (2017).
5. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
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. Zhang, Y. Z., Shi, M. & Holmes, E. C. Using metagenomics to characterize an expanding virosphere. *Cell* **172**, 1168–1172 (2018).
8. Ambrose, H. E. & Clewley, J. P. Virus discovery by sequence-independent genome amplification. *Rev. Med. Virol.* **16**, 365–383 (2006).
9. Shi, M. et al. The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
Major study of the phylogenetic diversity of RNA viruses carried by diverse vertebrates, showing that many of the virus families associated with mammals have a deep ancestry with evolutionary roots in fish.
10. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using diamond. *Nat. Methods* **18**, 366–368 (2021).
11. Geoghegan, J. L. & Holmes, E. C. Predicting virus emergence amid evolutionary noise. *Open Biol.* **7**, 170189 (2017).
12. Fernández, R. & Gabaldón, T. Gene gain and loss across the metazoan tree of life. *Nat. Ecol. Evol.* **4**, 524–533 (2020).
13. Laumer, C. E. et al. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. Biol. Sci.* **286**, 20190831 (2019).
14. Paraskevopoulou, S. et al. Viromics of extant insect orders unveil the evolution of the flavi-like superfamily. *Virus Evol.* **7**, veab030 (2021).
15. Murphy, F. A. Historical perspective: what constitutes discovery (of a new virus)? *Adv. Virus Res.* **95**, 197–220 (2016).
16. Greninger, A. L. A decade of RNA virus metagenomics is (not) enough. *Virus Res.* **244**, 218–229 (2018).
17. 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).
First article to show that invertebrates — in this case arthropods — harbour an enormous diversity of RNA viruses, often at high abundance. Provides the first description of the chuviruses, which are characterized by diverse genome structures.
18. Donaldson, E. F. et al. Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *J. Virol.* **84**, 13004–13018 (2010).
19. Li, L. et al. The fecal viral flora of California sea lions. *J. Virol.* **85**, 9909–9917 (2011).
20. Ge, X. et al. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. *J. Virol.* **86**, 4620–4630 (2012).
21. Walker, P. J. et al. Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch. Virol.* https://doi.org/10.1007/s00705-021-05156-1 (2021).
22. Lauber, C. et al. Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. *Cell Host Microbe* **22**, 387–399.e386 (2017).
Major study of the phylogenetic diversity of HBV-like viruses in fish, including the discovery of a group of related viruses — the nackednaviruses — that lack the envelope protein.
23. Geoghegan, J. L. et al. Hidden diversity and evolution of viruses in market fish. *Virus Evol.* **4**, vey031 (2018).
24. Zeigler Allen, L. et al. The Baltic Sea virome: diversity and transcriptional activity of DNA and RNA viruses. *mSystems* **2**, e00125–16 (2017).
25. Parry, R., Wille, M., Turnbull, O. M. H., Geoghegan, J. L. & Holmes, E. C. Divergent influenza-like viruses of amphibians and fish support an ancient evolutionary association. *Viruses* **12**, 1042 (2020).
26. Geoghegan, J. L. et al. Virome composition in marine fish revealed by meta-transcriptomics. *Virus Evol.* **7**, veab005 (2021).
27. Costa, V. A. et al. Metagenomic sequencing reveals a lack of virus exchange between native and invasive freshwater fish across the Murray–Darling Basin, Australia. *Virus Evol.* **7**, veab034 (2021).
28. Miller, A. K. et al. Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the *Coronaviridae*. *Virus Evol.* **7**, veab050 (2021).
29. López-Bueno, A. et al. Concurrence of iridovirus, polyomavirus, and a unique member of a new group of fish papillomaviruses in lymphocystis disease-affected gilthead sea bream. *J. Virol.* **90**, 8768–8779 (2016).
30. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
31. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
32. Wang, W. et al. Extensive genetic diversity and host range of rodent-borne coronaviruses. *Virus Evol.* **6**, veaa078 (2021).
33. Latinne, A. et al. Origin and cross-species transmission of bat coronaviruses in China. *Nat. Commun.* **11**, 4235 (2020).
34. Temmam, S. et al. Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-871965/v1 (2021).
35. van Aart, A. E. et al. SARS-CoV-2 infection in cats and dogs in infected mink farms. *Transbound. Emerg. Dis.* https://doi.org/10.1111/tbed.14173 (2021).
36. Oude Munnink, B. B. et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172–177 (2021).
Demonstration of the broad host range of SARS-CoV-2, reflected in a major outbreak in farmed mink. That the virus was able to spread back to humans shows that some animal species may become SARS-CoV-2 reservoirs.
37. Chandler, J. C. et al. SARS-CoV-2 exposure in wild white-tailed deer (*Odocoileus virginianus*). Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.29.454326 (2021).
38. Lam, T. T.-Y. et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285 (2020).
39. Zhou, H. et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell* **184**, 4380–4391 (2021).
40. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and sources of endemic human coronaviruses. *Adv. Virus Res.* **100**, 163–188 (2018).
Important review of the human coronaviruses highlighting their diversity, evolutionary history and zoonotic origins.
41. Edgar, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.07.241729 (2020).
42. Salehi-Ashtiani, K., Luptāk, A., Litovchick, A. & Szostak, J. W. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* **313**, 1788–1792 (2006).
43. Chang, W.-S. et al. Novel hepatitis D-like agents in vertebrates and invertebrates. *Virus Evol.* **5**, vez021 (2019).
44. Hetzel, U. et al. Identification of a novel deltavirus in boa constrictors. *mBio* **10**, e00014-19 (2019).
45. Iwamoto, M. et al. Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution. *Virus Evol.* **7**, veab003 (2021).
Overview of the evolution of deltaviruses (that is, HDV-like viruses) in birds and mammals. Reveals the ancient history and diversity of these viruses and shows that they are not exclusively associated with humans or HBV.
46. Paraskevopoulou, S. et al. Mammalian deltavirus without hepadnavirus coinfection in the neotropical rodent *Proechimys semispinosus*. *Proc. Natl Acad. Sci. USA* **117**, 17977–17983 (2020).
47. Taubenberger, J. K. & Kash, J. C. Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe* **7**, 440–451 (2010).
48. Joseph, U., Su, Y. C., Vijaykrishna, D. & Smith, G. J. The ecology and adaptive evolution of influenza A interspecies transmission. *Influenza Other Respir. Viruses* **11**, 74–84 (2017).
49. Wu, H. et al. Abundant and diverse RNA viruses in insects revealed by RNA-Seq analysis: ecological and evolutionary implications. *mSystems* **5**, e00039-20 (2020).
50. Van Eynde, B. et al. Exploration of the virome of the European brown shrimp (*Crangon crangon*). *J. Gen. Virol.* **101**, 651–666 (2020).
51. Laffy, P. W. et al. Reef invertebrate viromics: diversity, host specificity and functional capacity. *Environ. Microbiol.* **20**, 2125–2141 (2018).
52. Tokarz, R. et al. Virome analysis of *Amblyomma americanum*, *Dermacentor variabilis*, and *Ixodes scapularis* ticks reveals novel highly divergent vertebrate and invertebrate viruses. *J. Virol.* **88**, 11480–11492 (2014).
53. Käfer, S. et al. Re-assessing the diversity of negative strand RNA viruses in insects. *PLoS Pathog.* **15**, e1008224 (2019).
54. Ramírez, A. L. et al. Metagenomic analysis of the virome of mosquito excreta. *mSphere* **5**, e00587-20 (2020).
55. Brinkmann, A. et al. A metagenomic survey identifies Tamdy orthonairovirus as well as divergent phlebo-, rhabdo-, chu- and flavi-like viruses in Anatolia, Turkey. *Ticks Tick. Borne Dis.* **9**, 1173–1183 (2018).
56. Gudenkauf, B. M. & Hewson, I. Comparative metagenomics of viral assemblages inhabiting four phyla of marine invertebrates. *Front. Mar. Sci.* **3**, 23 (2016).
57. Medd, N. C. et al. The virome of *Drosophila suzukii*, an invasive pest of soft fruit. *Virus Evol.* **4**, vey009 (2018).
58. Webster, C. L. et al. The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biol.* **13**, e1002210 (2015).
59. Hameed, M. et al. A metagenomic analysis of mosquito virome collected from different animal farms at Yunnan–Myanmar border of China. *Front. Microbiol.* **11**, 591478 (2021).
60. Sadeghi, M. et al. Virome of >12 thousand Culex mosquitoes from throughout California. *Virology* **523**, 74–88 (2018).
Major metagenomic study of virome diversity in mosquitoes showing the power of this technology for high-throughput virus screening at a single location.
61. He, X. et al. Metagenomic sequencing reveals viral abundance and diversity in mosquitoes from the Shaanxi-Gansu-Ningxia region, China. *PLoS Negl. Trop. Dis.* **15**, e0009381 (2021).
62. Marklewitz, M., Zirkel, F., Kurth, A., Drosten, C. & Junglen, S. Evolutionary and phenotypic analysis of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. *Proc. Natl Acad. Sci. USA* **112**, 7536–7541 (2015).
63. Schmidlin, K. et al. A novel lineage of polyomaviruses identified in bark scorpions. *Virology* **563**, 58–63 (2021).
64. 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).
65. Ladner, J. T. et al. A multicomponent animal virus isolated from mosquitoes. *Cell Host Microbe* **20**, 357–367 (2016).
First description of a multicomponent virus in an animal. Highlights the complexity of genome evolution in RNA viruses, in this case in the flavi-like viruses.
66. Argenta, F. F. et al. Identification of reptarenaviruses, hartmaniviruses, and a novel chuvirus in captive native Brazilian boa constrictors with boid inclusion body disease. *J. Virol.* **94**, e00001-20 (2020).
67. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS pandemic. *Cold Spring Harb. Perspect. Med.* **1**, a006841 (2011).
68. Pybus, O. G., Rambaut, A., Holmes, E. C. & Harvey, P. H. New inferences from tree shape: numbers of missing taxa and population growth rates. *Syst. Biol.* **51**, 881–888 (2002).
69. Kapusinszky, B. et al. Local virus extinctions following a host population bottleneck. *J. Virol.* **89**, 8152–8161 (2015).
70. McKee, C. D., Bai, Y., Webb, C. T. & Kosoy, M. Y. Bats are key hosts in the radiation of mammal-associated *Bartonella* bacteria. *Infect. Genet. Evol.* **89**, 104719 (2021).
71. dos Reis, M. et al. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Curr. Biol.* **25**, 2939–2950 (2015).
72. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47–59 (2010).
73. Wang, L.-F., Walker, P. J. & Poon, L. L. M. Mass extinctions, biodiversity and mitochondrial function:

are bats 'special' as reservoirs for emerging viruses? *Curr. Opin. Virol.* **1**, 649–657 (2011).
One of the first articles to propose that bats are uniquely important hosts for emerging viruses and that host mass extinction events might play a key role in shaping the phylogenetic diversity of viruses.
74. Stanley, S. M. Estimates of the magnitudes of major marine mass extinctions in earth history. *Proc. Natl Acad. Sci. USA* **113**, E6325–E6334 (2016).
75. Raup, D. M. & Sepkoski, J. J. Mass extinctions in the marine fossil record. *Science* **215**, 1501–1503 (1982).
76. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
77. Carlson, C. J., Zipfel, C. M., Garnier, R. & Bansal, S. Global estimates of mammalian viral diversity accounting for host sharing. *Nat. Ecol. Evol.* **3**, 1070–1075 (2019).
78. 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).
79. Racaniello, V. Moving beyond metagenomics to find the next pandemic virus. *Proc. Natl Acad. Sci. USA* **113**, 2812–2814 (2016).
80. Morse, S. S. et al. Prediction and prevention of the next pandemic zoonosis. *Lancet* **380**, 1956–1965 (2012).
81. Smith, I. & Wang, L. F. Bats and their virome: an important source of emerging viruses capable of infecting humans. *Curr. Opin. Virol.* **3**, 84–91 (2013).
82. Olival, K. J. et al. Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
83. Wasik, B. R. et al. Onward transmission of viruses: how do viruses emerge to cause epidemics after spillover? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190017 (2019).
84. Dük, A. et al. Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science* **368**, 1367–1370 (2020).
85. Tabachnick, W. J. Climate change and the arboviruses: lessons from the evolution of the dengue and yellow fever viruses. *Annu. Rev. Virol.* **3**, 125–145 (2016).
86. Fritzell, C. et al. Current challenges and implications for dengue, chikungunya and Zika seroprevalence studies worldwide: a scoping review. *PLoS Negl. Trop. Dis.* **12**, e0006533 (2018).
87. Campbell-Lendrum, D., Manga, L., Bagayoko, M. & Sommerfeld, J. Climate change and vector-borne diseases: what are the implications for public health research and policy? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20130552 (2015).
88. Jacob, S. T. et al. Ebola virus disease. *Nat. Rev. Dis. Prim.* **6**, 13 (2020).
89. Gould, E. A. & Higgs, S. Impact of climate change and other factors on emerging arbovirus diseases. *Trans. R. Soc. Trop. Med. Hyg.* **103**, 109–121 (2009).
90. Marcondes, M. & Day, M. J. Current status and management of canine leishmaniasis in Latin America. *Res. Vet. Sci.* **123**, 261–272 (2019).
91. Brock, P. M. et al. Predictive analysis across spatial scales links zoonotic malaria to deforestation. *Proc. Biol. Sci.* **286**, 20182351 (2019).
92. Ayala, A. J., Yabsley, M. J. & Hernandez, S. M. A review of pathogen transmission at the backyard chicken–wild bird interface. *Front. Vet. Sci.* **7**, 662 (2020).
93. Munoz, O. et al. Genetic adaptation of influenza A viruses in domestic animals and their potential role in interspecies transmission: a literature review. *Ecohealth* **13**, 171–198 (2016).
94. Peiris, J. S., de Jong, M. D. & Guan, Y. Avian influenza virus (H5N1): a threat to human health. *Clin. Microbiol. Rev.* **20**, 243–267 (2007).
Review of the ecology and evolution of H5N1 avian influenza virus, particularly how it emerges in humans from its avian reservoir populations and its associated pandemic risk.
95. Schelling, E., Thur, B., Griot, C. & Audige, L. Epidemiological study of Newcastle disease in backyard poultry and wild bird populations in Switzerland. *Avian Pathol.* **28**, 263–272 (1999).
96. Boros, Á. et al. A diarrheic chicken simultaneously co-infected with multiple picornaviruses: complete genome analysis of avian picornaviruses representing up to six genera. *Virology* **489**, 63–74 (2016).
97. Lang, A. S. et al. Assessing the role of seabirds in the ecology of influenza A viruses. *Avian Dis.* **60**, 378–386 (2016).
98. Lickfett, T. M., Clark, E., Gehring, T. M. & Alm, E. W. Detection of influenza A viruses at migratory bird stopover sites in Michigan, USA. *Infect. Ecol. Epidemiol.* **8**, 1474709 (2018).
99. Rezza, G. Dengue and chikungunya: long-distance spread and outbreaks in naïve areas. *Pathog. Glob. Health* **108**, 349–355 (2014).
100. Fritzsche McKay, A. & Hoye, B. J. Are migratory animals superspreaders of infection? *Integr. Comp. Biol.* **56**, 260–267 (2016).
101. Jeong, S. et al. Introduction of avian influenza A(H6N5) virus into Asia from North America by wild birds. *Emerg. Infect. Dis.* **25**, 2138–2140 (2019).
102. Pettersson, J. H. O. et al. Circumpolar diversification of the *Ixodes uriae* tick virome. *PLoS Pathog.* **16**, e1008759 (2020).
103. Albery, G. F., Eskew, E. A., Ross, N. & Olival, K. J. Predicting the global mammalian viral sharing network using phylogeography. *Nat. Commun.* **11**, 2260 (2020).
104. 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).
105. Mollentze, N. & Streicker, D. G. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proc. Natl Acad. Sci. USA* **117**, 9423–9430 (2020).
106. Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS Biol.* **19**, e3001390 (2021).
107. Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS Biol.* **19**, e3001135 (2021).
108. Kohl, C. et al. The virome of German bats: comparing virus discovery approaches. *Sci. Rep.* **11**, 7430 (2021).
109. Li, L. et al. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J. Virol.* **84**, 6955–6965 (2010).
110. Kemenesi, G. et al. Molecular survey of RNA viruses in Hungarian bats: discovering novel astroviruses, coronaviruses, and caliciviruses. *Vector Borne Zoonotic Dis.* **14**, 846–855 (2014).
111. Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K. & Munster, V. J. Bat-borne virus diversity, spillover and emergence. *Nat. Rev. Microbiol.* **18**, 461–471 (2020).
Extensive review of the relevant biology of bats and the viruses they carry, particularly in the context of SARS-CoV-2.
112. Irving, A. T., Ahn, M., Goh, G., Anderson, D. E. & Wang, L.-F. Lessons from the host defences of bats, a unique viral reservoir. *Nature* **589**, 363–370 (2021).
Timely review outlining the reasons why bats might be uniquely important virus reservoirs and what this might mean for understanding future emergence events.
113. Banerjee, A. et al. Novel insights into immune systems of bats. *Front. Immunol.* **11**, 26 (2020).
114. Holmes, E. C., Rambaut, A. & Andersen, K. G. Pandemics: spend on surveillance, not prediction. *Nature* **558**, 180–182 (2018).
115. Sanjuán, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell Mol. Life Sci.* **73**, 4433–4448 (2016).
116. Plowright, R. K. et al. Pathways to zoonotic spillover. *Nat. Rev. Microbiol.* **15**, 502–510 (2017).
Benchmark review of the ecological processes by which viruses can spill over and emerge in new hosts, identifying this as a key process in virus evolution.
117. Holmes, E. C. et al. The origins of SARS-CoV-2: a critical review. *Cell* **184**, 4848–4856 (2021).
118. Johnson, B. A. et al. Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293–299 (2021).
119. Wrobel, A. G. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763–767 (2020).
Detailed structural virology study that demonstrates that even closely related human and animal coronaviruses can differ profoundly in receptor-binding ability.
120. Wilson, M. R. et al. Chronic meningitis investigated via metagenomic next-generation sequencing. *JAMA Neurol.* **75**, 947–955 (2018).
121. Wilson, M. R. et al. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. *N. Engl. J. Med.* **380**, 2327–2340 (2019).
Key article showing the importance of mNGS in a clinical diagnostic setting, in this case for the identification of the microbial pathogens associated with meningitis and encephalitis.
122. Xu, G. J. et al. Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).
Presents VirScan — a method for the high-throughput screening of viruses by identifying antiviral antibodies in human sera. Although originally designed to screen the human virome, the method could be adapted to detect zoonotic viruses.
123. Field, H. E., Mackenzie, J. S. & Daszak, P. Henipaviruses: emerging paramyxoviruses associated with fruit bats. *Curr. Top. Microbiol. Immunol.* **315**, 133–159 (2007).
124. Harvey, E. et al. Extensive diversity of RNA viruses in Australian ticks. *J. Virol.* **93**, e01358-18 (2019).
125. Wille, M. et al. Sustained RNA virome diversity in Antarctic penguins and their ticks. *ISME J.* **14**, 1768–1782 (2020).
126. Di Giallonardo, F., Schlub, T. E., Shi, M. & Holmes, E. C. Dinucleotide composition in RNA viruses is shaped more by virus family than host species. *J. Virol.* **91**, e02381-16 (2017).
127. Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res.* **239**, 136–142 (2017).
128. Bamford, D. H., Grimes, J. M. & Stuart, D. I. What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol.* **15**, 655–663 (2005).
Demonstrates how patterns of evolutionary relatedness are preserved in the structure of viral capsid proteins. Lays the foundation for how protein structural information can be used to infer phylogenetic relationships.
129. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence — a study of structural response in protein cores. *Proteins* **77**, 499–508 (2009).
130. Harrison, S. C. Viral membrane fusion. *Nat. Struct. Mol. Biol.* **15**, 690–698 (2008).
131. Fédry, J. et al. The ancient gamete fusogen HAP2 is a eukaryotic class II fusion protein. *Cell* **168**, 904–915 (2017).
Demonstration of how protein structure can reveal ancient evolutionary homologies, in this case between an algal gamete fusogen and a class II viral membrane fusion protein.
132. Henderson, R. Overview and future of single particle electron cryomicroscopy. *Arch. Biochem. Biophys.* **581**, 19–24 (2015).
133. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comp. Biol.* **7**, e1002195 (2011).
134. Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. V. Viruses of Archaea: structural, functional, environmental and evolutionary genomics. *Virus Res.* **244**, 181–193 (2018).
135. Holmes, E. C. & Duchêne, S. Can sequence phylogenies safely infer the origin of the global virome? *mBio* **10**, e00289-19 (2019).
136. Chang, G. S. et al. Phylogenetic profiles reveal evolutionary relationships within the 'twilight zone' of sequence similarity. *Proc. Natl Acad. Sci. USA* **105**, 13474–13479 (2008).
137. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
138. Wooley, J. C., Godzik, A. & Friedberg, I. A primer on metagenomics. *PLoS Comp. Biol.* **6**, e10006677 (2010).
139. O'Neil, D., Glowatz, H. & Schlumpberger, M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr. Protoc. Mol. Biol.* https://doi.org/10.1002/0471142727.mb0419s103 (2013).
140. Briese, T. et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio* **6**, e01491-15 (2015).
141. Chong, R. et al. Fecal viral diversity of captive and wild Tasmanian devils characterized using virion-enriched metagenomics and metatranscriptomics. *J. Virol.* **93**, e00205-19 (2019).

Author contributions

E.H. researched data for the article. Both authors contributed substantially to discussion of the content, wrote the article, and edited and reviewed the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Microbiology thanks Kevin Olival, Arvind Varsani and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.