ARTICLE

doi:10.1038/nature20167

Redefining the invertebrate RNA virosphere

Mang Shi[1,2]*, Xian-Dan Lin[3]*, Jun-Hua Tian[4]*, Liang-Jun Chen[1]*, Xiao Chen[5]*, Ci-Xiu Li[1]*, Xin-Cheng Qin[1], Jun Li[6], Jian-Ping Cao[7], John-Sebastian Eden[2], Jan Buchmann[2], Wen Wang[1], Jianguo Xu[1], Edward C. Holmes[1,2] & Yong-Zhen Zhang[1]

Current knowledge of RNA virus biodiversity is both biased and fragmentary, reflecting a focus on culturable or disease-causing agents. Here we profile the transcriptomes of over 220 invertebrate species sampled across nine animal phyla and report the discovery of 1,445 RNA viruses, including some that are sufficiently divergent to comprise new families. The identified viruses fill major gaps in the RNA virus phylogeny and reveal an evolutionary history that is characterized by both host switching and co-divergence. The invertebrate virome also reveals remarkable genomic flexibility that includes frequent recombination, lateral gene transfer among viruses and hosts, gene gain and loss, and complex genomic rearrangements. Together, these data present a view of the RNA virosphere that is more phylogenetically and genomically diverse than that depicted in current classification schemes and provide a more solid foundation for studies in virus ecology and evolution.

RNA viruses are likely to exist in every species of cellular life[1]. Despite this ubiquity, much of our knowledge of the biodiversity and evolution of RNA viruses, as well as their range of genomic structures, comes from those viruses that can be cultured and that act as agents of disease in humans or economically important animals and plants. However, these only represent a tiny fraction of eukaryotic diversity. This sparse sampling is apparent from studies of invertebrate viruses. Although invertebrates comprise the vast majority of the Metazoa (animals), little is known about the nature of the 'virosphere' of these organisms[2]. Metagenomic studies of invertebrate viruses have only recently been undertaken but often reveal far greater viral biodiversity than seen in vertebrates[3–8]. Arthropods, for example, commonly act as viral vectors and studies of arthropod RNA viruses have revealed that changes in genome size, structure and segmentation have occurred more frequently and on a larger scale than previously realized[3,9], with some arthropod viruses likely to be ancestors of those that infect vertebrates[3]. However, these studies are of limited scope and there are still substantial gaps in our knowledge of RNA virus biodiversity at both the phylogenetic and genomic scales for most invertebrates, a fact that may have important implications for our understanding of virus evolution, ecology and emergence[10]. We describe here a large-scale meta-transcriptomic survey of diverse invertebrate taxa aimed at revealing the hidden diversity of RNA viruses. The data obtained enable us to re-examine and re-define the invertebrate virosphere, providing a new perspective on the fundamental patterns and processes of viral evolution.

RNA viruses in invertebrates

We performed deep transcriptome sequencing on more than 220 invertebrate species, representing 9 metazoan phyla (Arthropoda, Annelida, Sipuncula, Mollusca, Nematoda, Platyhelminthes, Cnidaria, Echinodermata, and the Chordata subphylum Tunicata), most of which have not previously been screened for viruses (Supplementary Table 1). Accordingly, we extracted total RNA from these species and prepared 87 RNA sequencing (RNA-seq) libraries for Illumina HiSeq sequencing (Supplementary Table 1). In total, we generated 6 trillion bases of 90–100 bp paired-end reads that were assembled *de novo* for virus characterization.

These transcriptome data allowed us to identify at least 1,445 phylogenetically distinct virus genomes or genome segments that contained an RNA-dependent RNA polymerase (RdRp) domain (Supplementary Table 2). The majority of these virus genomes have greater than 20-fold coverage and are sequenced to their complete or near-complete length. Sequence alignments and structural comparisons revealed extensive sequence divergence within these newly discovered RdRp domains, with most sharing less than 40% amino acid identity with those RNA viruses described previously.

To assess the amount of viral RNA in each library, we removed all rRNA reads, including those from the host species, and determined the proportion of the remaining sequence data that mapped to viral RNA. This revealed that viral RNA comprised from 0.05% to 87% of the total RNA sequenced (rRNA excluded) within each library, although the very high levels in some cases may reflect degradation or inefficient extraction of host RNA (Fig. 1). Each library contains 1–20 virus species per host species and 1–6 virus species that represent more than 0.1% of the total RNA sequenced (rRNA excluded) (Fig. 1). Although some libraries contain far higher numbers of virus transcripts, most have low levels of RNA and may therefore be associated with other cellular organisms that are present within the host (see below).

These transcriptome data also contain a substantial proportion of transcripts that carry divergent reverse transcriptase enzymes, potentially derived from retrotransposons (Fig. 1). These can be distinguished from RNA viruses by their replicase components, the lack of consistently inherited structural proteins and the presence of DNA copies. Additionally, although RNAs produced by DNA viruses (including bacteriophages) and bacteria were present in the transcriptome data, these were generally at lower quantities than RNA viruses and will not be discussed further.

[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, Changping, 100206 Beijing, China. [2]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 2006, Australia. [3]Wenzhou Center for Disease Control and Prevention, Wenzhou, 325001 Zhejiang, China. [4]Wuhan Center for Disease Control and Prevention, Wuhan, 430015 Hubei, China. [5]Guangxi Mangrove Research Center, Beihai, 536000 Guangxi, China. [6]Systems Biology and Bioinformatics Group, School of Biological Sciences, Faculty of Sciences, University of Hong Kong, Hong Kong, China. [7]National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Shanghai, China.
*These authors contributed equally to this work.

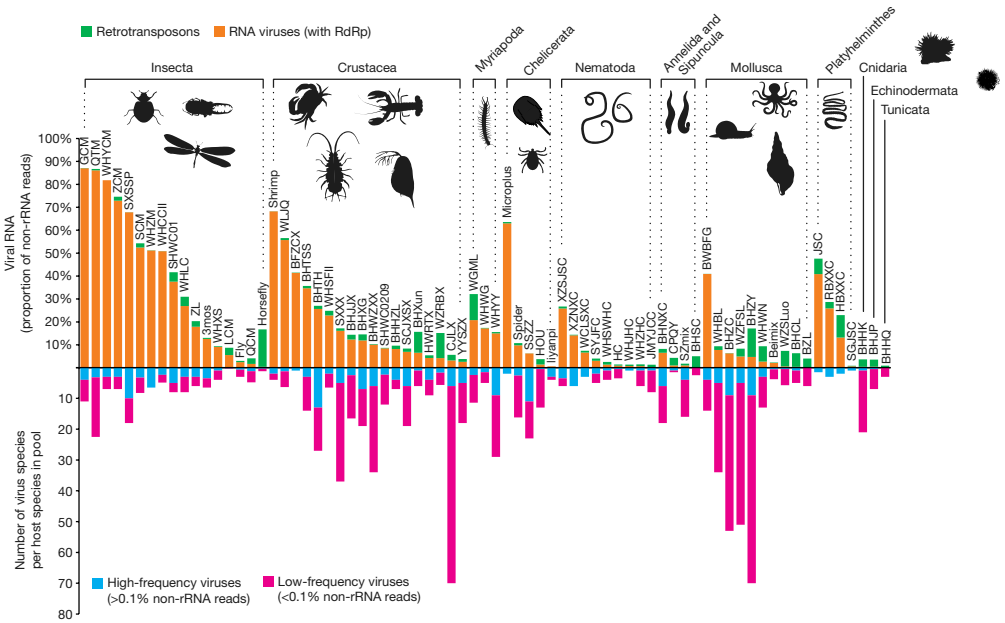


Figure 1 | The frequency and diversity of viral RNA transcripts in invertebrate transcriptomes. The top graph shows the percentage of non-rRNA reads mapped to viral RNA (RdRp; orange bars) and retrotransposons (green bars) in each library. The short name of each library is shown on top of each bar, while major host classifications are shown above the bar graph. The bottom graph shows a summary of the normalized number of virus species within each library. The total number is further subdivided to identify those RNA viruses at either high (blue) or low (magenta) frequency.

Evolution of invertebrate RNA viruses

To place these newly discovered viruses within the context of known viral biodiversity, we collated NCBI reference virus genomes from all established families and floating genera of RNA viruses, as well as non-reference virus genomes from unclassified taxa. The RdRp is the only conserved-sequence domain across all RNA viruses and was therefore used for phylogenetic inference. Phylogenetic analysis revealed that the genetic diversity of the newly discovered RNA viruses surpassed that described previously and could not always be easily incorporated into current virus classifications (Fig. 2; see Supplementary Data 1–21 for detailed phylogenies). In particular, many of the newly discovered viruses occupy topological positions that fall between families or genera and thus fill major phylogenetic gaps, so that RNA viruses now occupy a more continual spectrum of phylogenetic diversity.

To describe and accommodate the extraordinary diversity of viruses discovered here better, we merged previously defined virus families, orders and floating genera, to produce 16 clades of RNA viruses. For simplicity, we have abbreviated these clades as 'Astro', 'Birna', 'Hepe–Virga', 'Hypo', 'Luteo–Sobemo', 'Narna–Levi', 'Bunya–Arena', 'Mono–Chu', 'Orthomyxo', 'Nido', 'Partiti–Picobirna', 'Permutotetra', 'Picorna–Calici', 'Reo', 'Tombus–Noda' and 'Toti–Chryso', reflecting the presence of representative viral families or orders within each clade (Fig. 2 and Supplementary Data 1–20). Notably, these clades resemble, but do not necessarily correspond to, the 'supergroups' of RNA viruses proposed previously[11]. We also identified at least five clades of RNA viruses in which RdRp domains are so divergent that they might be considered as new virus families or orders, although phylogenetic analyses of such divergent taxa should be treated with caution (Supplementary Data 21). Reflecting the location of their sampling, we provisionally named these divergent lineages after ancient Chinese states from the Chunqiu period, specifically; 'Yuevirus', 'Qinvirus', 'Zhaovirus', 'Weivirus' and 'Yanvirus'.

Since our sample processing involves the entire individual invertebrate, it is possible that a substantial proportion of the viruses discovered here were associated with undigested food, gut microflora or parasites that exist within the organisms investigated. We therefore estimated the proportion of each viral transcript within the library and assumed that the more common the virus, the more likely that it was associated with that host (although this may not equate to active infection). Generally, those viruses that made up a higher proportion of total RNA levels (>0.1% total RNA) were not closely related to those known to infect vertebrates, plants or fungi, suggesting that viral RNA quantity may be a useful indicator of their true host. To assess the likely host species further, we screened for endogenous virus elements (EVEs) related to the exogenous viruses described here[12]. Although some viruses, such as the *Picornavirales*, rarely possess EVEs, and there is little genome data for species within the Annelida and Mollusca phyla, the EVE data helped confirm the host taxon by establishing their evolutionary ancestry in that host. In particular, the host taxa containing EVEs often closely matched those containing related exogenous viruses (Supplementary Data 3, 7–9, 11, 18, 20, 21). However, the EVE data also suggested alternative or additional hosts in a number of cases (Supplementary Data 5, 6, 21). For example, the highly divergent new RNA virus (Weivirus) identified in the mollusc transcriptome was related to EVEs identified in alveolates (protist) genomes (Supplementary Data 21). Finally, we also examined the presence of variant genetic codes as a guide to the likely host organisms. The most common variant genetic codes observed were the invertebrate mitochondrial code (Supplementary Data 6, 11) and the ciliate, dasycladacean and hexamita nuclear codes. The latter is found in the new RNA virus Zhaovirus (Supplementary Data 21), as well as in a cluster of viruses from the Tombus–Noda clade (Supplementary Data 19), indicating that these viruses are more likely to be associated with protists than with invertebrates.

Overall, the host spectrum for the RNA viruses described here is broad, including different phyla and sometimes different kingdoms (Supplementary Data 1–21). Much of our sampling was directed towards the Arthropoda, meaning that definite statements on host range cannot be made. Despite this bias, the diversity of arthropod viruses is notable as they appear in multiple lineages within each major clade (Extended Data Fig. 1 and Supplementary Data 1–21). Also of note were the phyla Mollusca, Annelida, and Sipuncula (collectively the superphylum Lophotrochozoa) that diverged early from Nematoda and Arthropoda in the metazoan phylogeny[13]. Notably, the viromes of these phyla either contained extremely divergent viruses (such as in the Bunya–Arena and Orthomyxo clades; Supplementary Data 7 and 9, respectively) or had substantial overlap with the arthropod virome (for example, several viruses in the aquatic picorna-like clade are commonly present in both Crustacea and Lophotrochozoa; Supplementary Data 14). Although only a limited number of species were available for the remaining phyla, it is notable that the Platyhelminthes and Cnidaria did not contain particularly divergent viruses, despite their basal position within Metazoa. Finally, although the phylum Echinodermata and the subphylum Tunicata of Chordata are more closely related to vertebrates than the other invertebrate taxa studied here, we did not identify any viruses that were clearly ancestors of vertebrate-specific virus families.

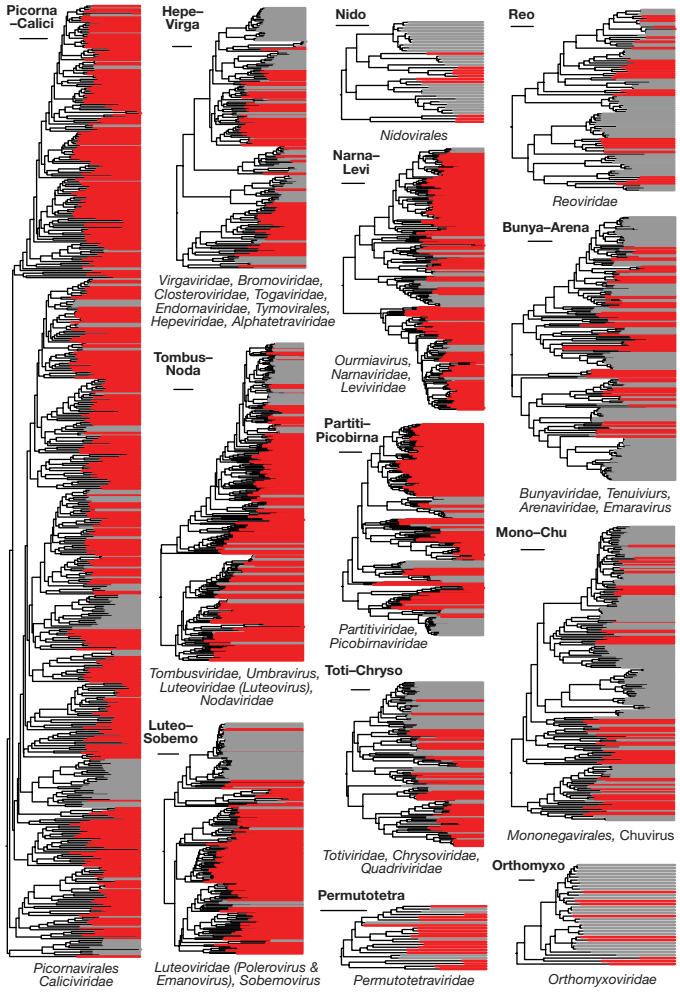


Figure 2 | Phylogenetic diversity of RNA viruses. Thirteen phylogenetic trees representing the major clades of RNA virus RdRp domains (see main text for definitions). Within each tree, the viruses discovered here are shaded red, while those described previously are shaded grey. The name of each clade is shown to the top left of each phylogeny and the names of the families or genera within clade are shown below the tree. Each scale bar indicates 0.5 amino acid substitutions per site. More detailed trees for each clade are shown in Supplementary Data 1–21, and their genome structures are shown in Supplementary Data 22–36.

Although viruses from divergent host taxa tend to form separate phylogenetic groups, suggesting that these virus–host associations have been established over long evolutionary timescales, there was generally little resemblance between the phylogenetic histories of viruses and their hosts, such that strict virus–host co-divergence cannot always be assumed. Indeed, there are clear examples of cross-species transmission of viruses among divergent host taxa. For example, several viruses that infect plants (such as Tenuivirus and Fijivirus; Supplementary Data 7 and 18, respectively) may be derived from arthropod viruses as they are nested within arthropod viruses on the phylogenies and the phylogenetic divergence between the plant and arthropod viruses was shallow.

Patterns of RNA virus genome evolution

Despite the presence of conserved RdRp sequences, the evolutionary histories of the structural and non-structural parts of the virus genomes characterized here often differed substantially (Fig. 3 and Supplementary Data 22–36). A single RdRp clade may contain coat proteins from diverse clades, and vice versa (Fig. 3 and Extended Data Table 1), indicative of widespread recombination among structural and non-structural genomic regions over long evolutionary timescales. Such incongruence is commonly observed within and between the major groups of positive-sense RNA viruses (that is, the Tombus–Noda, Luteo–Sobemo, Hepe–Virga, Permutotetra, Astro and Narna–Levi clades). Notably, the occurrence of recombination seems to be unaffected by genome organization. There were major differences between the tree topologies of the RdRp and coat protein domains for unsegmented viruses of the 'Tombus–Noda' clade (Extended Data Fig. 2), and the correlation coefficient between the two genetic distance matrices for these proteins was low (Extended Data Fig. 2b). By contrast, the 'Picorna–Calici' clade had a more stable genome structure (Supplementary Data 33) and lower rate of genetic exchange (Extended Data Fig. 2b). Envelope glycoproteins were also involved in inter-virus recombination, although such events were rarer. Specifically, we documented recombination events involving the glycoproteins of highly divergent virus groups, including within negative-sense RNA viruses, between negative- and positive-sense RNA viruses, and even between negative-sense RNA and DNA viruses (Extended Data Table 1 and Supplementary Data 27, 28).

The data also show that the evolution of structural genes involves the gain and loss of genes, which can occur in both segmented and unsegmented viruses. We found viruses with multiple copies of structural genes, such as coat protein genes in the Hepe–Virga clade (Extended Data Fig. 3a) and glycoprotein genes in the Mono–Chu clade (Extended Data Fig. 3b). In addition, their diverse positions on the phylogeny suggest that these additional gene copies were independently acquired through lateral gene transfer rather than being generated *de novo* by gene duplication (Extended Data Fig. 3a). In other taxa, a reduction in the number of genes encoding structural proteins has been observed. Structural genes are, for example, relatively more frequently lost in negative-sense RNA viruses (Extended Data Fig. 3c), and viruses with such 'reduced' genomes are found in Nematoda, Arthropoda, and Platyhelminthes (Supplementary Data 27–28). Across the dataset as a whole, gene loss most often involved the glycoprotein, although several viruses within the Bunya–Arena clade may lack both glycoprotein and nucleoprotein genes (Supplementary Data 27; see Methods). Viruses with no structural proteins are also present in some positive-sense and double-stranded RNA (dsRNA) viruses, such as the *Endornaviridae*, *Hypoviridae*, *Narnaviridae* and *Umbravirus* in the Hepe–Virga, Hypo, Narna–Levi, and Tombus–Noda clades, respectively[14], and we can tentatively identify clusters of viruses whose genome may only contain a replicase (Supplementary Data 23, 26, 27, 31).

We identified a number of protein domains in the non-structural part of the genome that are shared among divergent viruses, and even with cellular organisms (Extended Data Table 1). These include the RNA

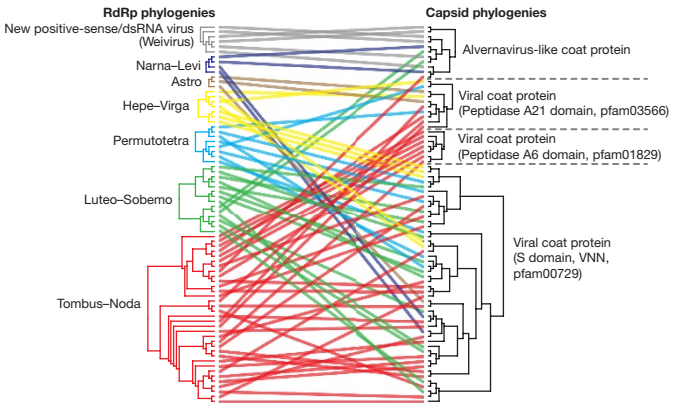


Figure 3 | Genetic exchange among RNA viruses. Comparison of the phylogenetic trees of 76 representative viral genomes with different types of structural protein (that is, four major types of capsid protein) and the equivalent phylogenies obtained for their RdRp amino acid sequences (eight clades as defined in the text, shown in different colours). Line colours correspond to those of the RdRp clade as shown to the left of the figure. Widespread recombination can be inferred when RdRp clades are associated with different types of structural protein, and vice versa.

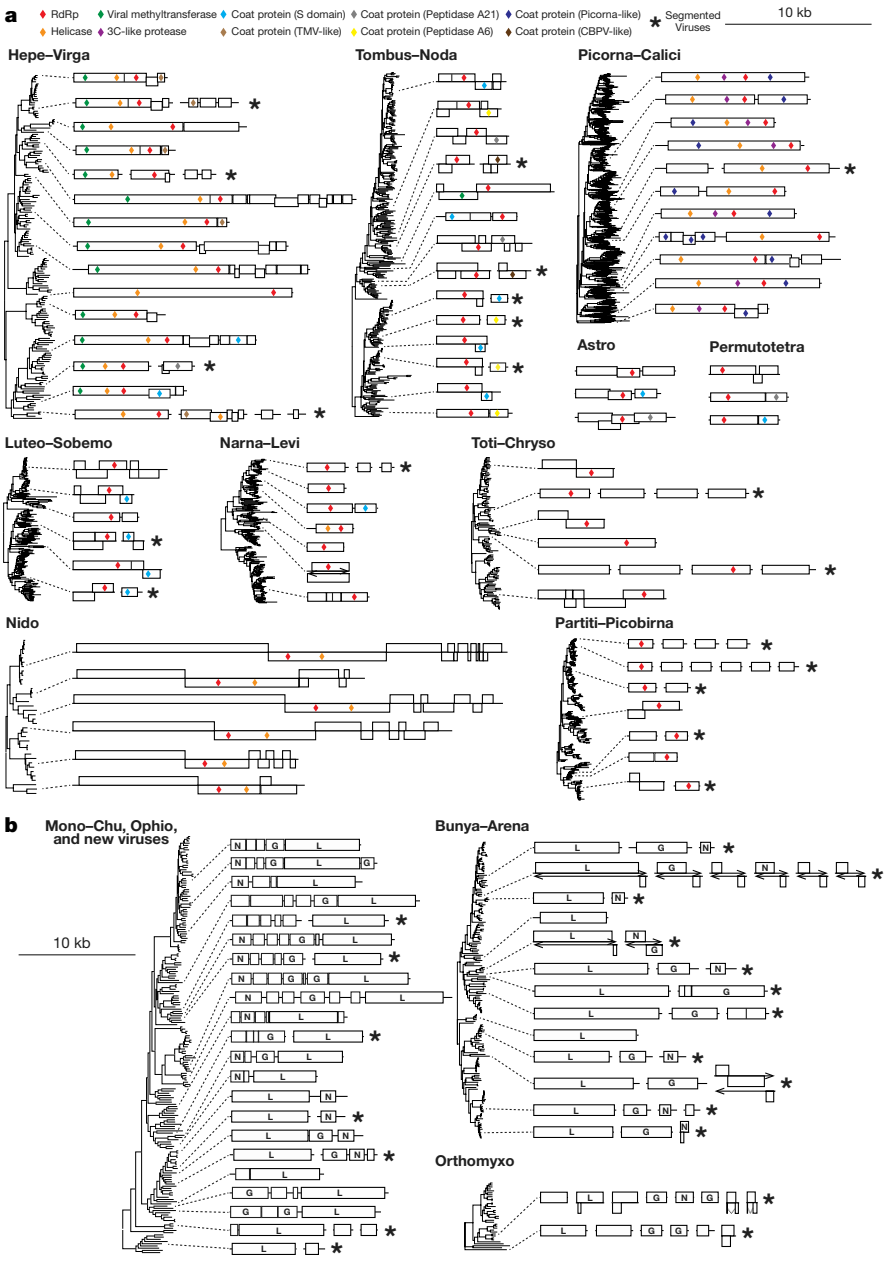


Figure 4 | Evolution of genome organization in RNA viruses. a, Genome evolution in ten representative clades of positive-sense and dsRNA viruses. Genome order follows the RdRp phylogeny. The genomes are drawn to a unified length scale shown at the top. The panel shows the pattern of segmentation (asterisks), the locations of major nonstructural and structural domains (coloured diamonds) and the arrangement of open reading frames. **b**, Genome evolution of negative-sense RNA viruses. N, G, and L indicate homologues from nucleoproteins, glycoproteins, and the polymerase, respectively. More detailed depictions of genome evolution are shown in Supplementary Data 22–36.

helicase, methyltransferase, exonuclease, protease, ADP–ribose binding protein (the macro domain), dsRNA binding protein, and even the *Escherichia coli* swarming motility protein (the NADAR domain). For example, we identified the parallel acquisition of eukaryotic-origin exonuclease domains (Extended Data Fig. 4a). Notably, the virus genomes that contain these exonuclease domains were highly divergent, although all were present in the same host (*Ligia exotica*, Extended Data Fig. 4a). In the case of the serine protease, virus diversity appeared both within and outside of the phylogenetic diversity of cellular proteins (Extended Data Fig. 4b), indicative of independent gene acquisitions. Also of note was the wide, but highly sporadic, phylogenetic distributions of some of these domains (such as Macro and NADAR; Extended Data Table 1) and the variable insertion locations in the genome (Extended Data Fig. 4c), again pointing to multiple, independent gene acquisitions.

There have been major reconfigurations of viral genome organization through evolutionary history, including the number and arrangement of open reading frames, the order of structural and non-structural genes, and the occurrence and extent of segmentation (Fig. 4 and Supplementary Data 22–36). These features, which are often regarded as conservative traits, in reality show great flexibility at the deep evolutionary scale studied here. Examination of the newly discovered viruses reveals that the evolution of segmented genomes, or the loss of segmentation, has occurred frequently. Several viruses experience high frequencies of segmentation, including members of the Tombus–Noda and Mono–Chu clades, in which the break-up and reunification of structural and non-structural genes has occurred relatively frequently (Fig. 4 and Supplementary Data 28, 34). The dynamics of segmentation were also reflected by changes in the number of segments. For instance, the number of segments in both the Partiti–Picorbina and Bunya–Arena clades vary from 1 to 6 (Fig. 4). Notably, the change in segment numbers is not only associated with gene break-up and unification, but also with the gain and loss of genes. We also found a tri-segmented counterpart of the normally bi-segmented arenaviruses that exhibited no structural gene homology with other members of this group (Supplementary Data 7, 27). Despite this flexibility, it is also the case that a genomic plan comprising multiple segments can be conserved for extended time periods, as is seen in the reoviruses and orthomyxoviruses. In one instance in the latter, we identified a highly divergent virus from the earthworm that possessed a similar genomic plan to other orthomyxoviruses (with 6 segments), including those from vertebrates (Fig. 4 and Supplementary Data 29).

Discussion

We have used a simple, yet powerful, metagenomic approach to characterize the viromes of diverse invertebrates. This approach is relatively unbiased, as no attempt is made to enrich viral particles through filtering, centrifugation and nuclease treatment. Although some invertebrates seemingly harboured a high proportion of viral RNA, which evidently depends on the infection status in the host in question, we were unable to determine whether the viruses identified here have any impact on host biology, including as agents of disease. Despite this, it is clear that for many invertebrates infection by multiple RNA viruses is likely to be the norm rather than the exception[15–17].

The whole-transcriptome approach employed here allowed us to characterize the virome of a diverse array of invertebrates, providing a new perspective on viral biodiversity. A number of virus families that were previously only known to infect plants, fungi, and protists are now visible in invertebrates. Viruses infecting divergent phyla were dispersed throughout the phylogenetic trees and exhibited diverse patterns of clustering, reflecting a complex interplay between long-term virus-host associations, including co-divergence, as well as frequent host jumping[18].

Despite the relatively high frequency of potential cross-species virus transmission documented here, there were also probable cases of long-term virus–host co-divergence. The viruses found in several species of parasitic nematodes tend to form monophyletic clusters, within which the phylogeny of viruses mirrors that of their hosts. In addition, the genetic diversity of RNA viruses within the Narna–Levi clade can be placed into three groups: those that infect bacteria (leviviruses), those that infect mitochondria (that is, mitoviruses that utilize the mitochondrial genetic code), and those that infect other organisms. This separation could in theory have occurred when the α-proteobacteria became intracellular symbionts[19,20]. Finally, despite the massive expansion of virus diversity documented here, some vertebrate viruses (such as those from the *Picornaviridae*, *Paramyxoviridae* and *Hepeviridae* families) remain monophyletic, with viruses from mammals, birds, reptiles, and fish occupying similar phylogenetic relationships to those of their hosts groups, probably indicative of long-term co-divergence.

We have necessarily inferred phylogenetic trees using the relatively conserved gene encoding RdRp. However, the evolutionary history of the entire genome is evidently more complex and not necessarily consistent with that of the RdRp domain. Indeed, at deep evolutionary timescales it is easier to trace the evolutionary history of individual functional units, such as the RdRp, helicase, and capsid, rather than that of intact viral genomes. Such modular genome evolution is reflected in three aspects of genetic diversity. First, there is great flexibility in the organization of functional units, including changes in genome segmentation and gene order. Second, different functional 'units' within genomes can be acquired or removed independently, although such processes occur relatively infrequently, which is likely to reflect strong restrictions on virus genome size. Thus, the simplest virus genome may contain only the replication module, whereas the most complex can contain multiple structural units or accessory units. Third, there is clear evidence for the exchange of functional units among viruses, particularly for structural proteins that can seemingly move large phylogenetic distances. Hence, the macroevolution of RNA viruses parallels the modular evolution previously proposed for bacteriophages[21], albeit with differences in timescale and mechanism. However, in the face of such abundant diversity, it is notable that none of the RNA viruses described here has a genome that exceeds the previously defined maximum of approximately 32 kb, probably reflecting intrinsic size constraints owing to error-prone replication[10].

By sampling a diverse range of invertebrate taxa, we have revealed unprecedented levels of RNA virus genetic diversity that both re-shapes our understanding of the patterns and processes of their evolution and highlights the limitations of our knowledge on what are likely to be the most abundant organisms on earth[22]. A full understanding of virus evolution and ecology will require an extensive survey of diverse host organisms using the types of metagenomic approach outlined here.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 May; accepted 17 October 2016.
Published online 23 November 2016.

1. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of cells. *Biol. Direct* **1,** 29 (2006).
2. Junglen, S. & Drosten, C. Virus discovery and recent insights into virus diversity in arthropods. *Curr. Opin. Microbiol.* **16,** 507–513 (2013).
3. 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).
4. Bekal, S., Domier, L. L., Niblack, T. L. & Lambert, K. N. Discovery and initial analysis of novel viral genomes in the soybean cyst nematode. *J. Gen. Virol.* **92,** 1870–1879 (2011).
5. Ballinger, M. J., Bruenn, J. A., Hay, J., Czechowski, D. & Taylor, D. J. Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. *J. Virol.* **88,** 8783–8794 (2014).
6. 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).
7. 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).
8. Webster, C. L. *et al.* The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biol.* **13,** e1002210 (2015).
9. 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).
10. Holmes, E. C. The Evolution and Emergence of RNA Viruses. (Oxford Univ. Press, 2009).
11. Koonin, E. V. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J. Gen. Virol.* **72,** 2197–2206 (1991).
12. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13,** 283–296 (2012).
13. Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* **22,** 1246–1253 (2005).
14. 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 Press, 2012).
15. Gauthier, L. *et al.* Viral load estimation in asymptomatic honey bee colonies using the quantitative RT–PCR technique. *Apidologie (Celle)* **38,** 426–435 (2007).
16. Genersch, E. *et al.* The German bee monitoring project: a long term study to understand periodically high winter losses of honey bee colonies. *Apidologie (Celle)* **41,** 332–352 (2010).
17. Tentcheva, D. *et al.* Prevalence and seasonal variations of six bee viruses in *Apis mellifera* L. and Varroa destructor mite populations in France. *Appl. Environ. Microbiol.* **70,** 7185–7191 (2004).
18. Baranowski, E., Ruiz-Jarabo, C. M. & Domingo, E. Evolution of cell recognition by viruses. *Science* **292,** 1102–1105 (2001).
19. Andersson, S. G. & Kurland, C. G. Origins of mitochondria and hydrogenosomes. *Curr. Opin. Microbiol.* **2,** 535–541 (1999).
20. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283,** 1476–1481 (1999).
21. Botstein, D. A theory of modular evolution for bacteriophages. *Ann. NY Acad. Sci.* **354,** 484–490 (1980).
22. Suttle, C. A. Viruses in the sea. *Nature* **437,** 356–361 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements This study was supported by the National Natural Science Foundation of China (Grants 81290343, 81273014, 81672057), the Special National Project on Research and Development of Key Biosafety Technologies (Grants 2016YFC1201900, 2016YFC1200101), the 12th Five-Year Major National Science and Technology Projects of China (2014ZX10004001-005), and an NHMRC Australia Fellowship (GNT1037231).

Author Contributions Conceptualization: M.S. and Y.-Z.Z. Methodology: M.S., L.-J.C, C.-X.L., J.L., J.-S.E, J.B., E.C.H. and Y.-Z.Z. Investigation: M.S., X.-D.L., J.-H.T., L.-J.C, X.C., C.-X.L. and X.-C.Q. Writing (original draft): M.S., E.C.H. and Y.-Z.Z. Writing (review and editing): M.S., X.-D.L., J.-H.T., L.-J.C, X.C., C.-X.L, J.-S.E, J.X., E.C.H. and Y.-Z.Z. Funding Acquisition: J.X., E.C.H. and Y.-Z.Z. Resources (sampling): M.S., X.-D.L., J.-H.T., L.-J.C, X.C., C.-X.L., J.-P.C., W.W. and Y.-Z.Z. Resources (computational): M.S., J.L., J.B. and E.C.H. Supervision: E.C.H. and Y.-Z.Z.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.-Z.Z. (zhangyongzhen@icdc.cn).

Reviewer Information *Nature* thanks E. Ghedin, D. Obbard and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. These experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Sample collection and processing. This study was based on the analysis of 87 libraries of invertebrate samples obtained from various locations in China (Supplementary Table 1). The sampling included land and freshwater organisms from Anhui, Beijing, Hubei, Xinjiang and Zhejiang provinces. Marine and coastal samples were obtained from Zhejiang (East China Sea) and Guangxi (South China Sea) provinces.

A total of 54 of the libraries were from arthropods (phylum Arthropoda). Of these arthropod libraries, 19 have been described previously[3,6,9], with 35 additional libraries newly obtained here. We sampled across all four subphyla from the phylum Arthropoda: (i) for the subphylum Chelicerata we sampled from the class Merostomata (horseshoe crabs) and Arachnida (spiders and ticks); (ii) for the subphylum Crustacea we sampled from the classes Branchiopoda (water fleas), Maxillopoda (barnacles), and Malacostraca (crabs, shrimps, crayfish, woodlice, wharf roaches, and so on); (iii) for the subphylum Hexapoda we sampled from nine orders (Coleoptera, Dermaptera, Diptera, Hemiptera, Hymenoptera, Lepidoptera, Odonata, Orthoptera, and Siphonaptera), all within the class Insecta; and (iv) for the subphylum Myriapoda we sampled from the classes Chilopoda (centipedes) and Diplopoda (millipedes).

All of the taxa from the phyla Nematoda and Platyhelminthes sampled here were parasites. The majority of the nematodes were from the family Ascaridida (class Secernentea) and found in the stomachs of pigs (libraries HC and WHZHC) and birds (libraries HC, JMYJCC, SYJFC, and WHJHC). We also collected three divergent nematode species within the class Secernentea from the stomachs of mice (library WHSWHC), the thoracic cavity of birds (library XSNXC), and the stomachs of snakes (library XZSJSC). In addition, we included a nematode species that infects mosquitoes (library WCLSXC) tentatively classified in the genus *Romanomermis* (class Adenophorea). For the phylum Platyhelminthes, our sampling comprised one species of tapeworm (*Taenia* sp.) discovered in the liver of rodents, as well as two species of blood flukes (*Schistosoma mansoni* and *S. japonicum*) initially obtained from an intermediate host (*Oncomelania*) and then matured in rabbits.

We also sampled three phyla within the superphylum Lophotrochozoa, namely Mollusca, Sipuncula, and Annelida. Our samples from the phylum Mollusca comprised a number of marine and freshwater species, including the classes Bivalvia (such as clams, mussels, oysters), Gastropoda (Chinese land snails, various sea snails, etc.), and Cephalopoda (octopus). For the phylum Sipuncula (peanut worms), we sampled the most common species in the shallow waters of the South China Sea, namely *Phascolosoma esculenta* and *Sipunculus nudus*. For the phylum Annelida, we sampled three representative classes: Polychaeta (sandworms), Oligochaeta (earthworms), and Hirudinea (leeches).

All the samples described above were protostomes. For the deuterostomes our samples comprised the phyla Chordata and Echinodermata. Within the Chordata, we sampled two species from the subphylum Tunicata, the invertebrate group most closely related to the vertebrates. Within the Echinodermata, we sampled representatives from the classes Echinoidea (sea urchins) and Holothuroidea (sea cucumbers). Finally, we examined one species from the radially symmetric phylum Cnidaria (sea anemones) as a representative of basal lineages within the Metazoa.

All samples were captured alive and stored at −80 °C. A proportion of the animals were left in Petri dishes (including pillworms, earthworms, centipedes, fiddler crabs, hermit crabs, woodlice, sandworms, oysters, Paphia shells, razor shells, Murex snails, Turritella sea snails, and Chinese land snails) or in purified sea water (horseshoe crabs, penaeid shrimps and mantis shrimp) for up to 24 h to clean stomach contents before transferring to −80 °C. Sample processing often involved the entire animal. However, for those animals with large body sizes, hard shells, or tissues that were difficult to homogenize, dissection was performed to obtain the entire inner organs (visceral mass for molluscs) or parts of different inner organs (Supplementary Table 1). During the dissection, the content of gut was intentionally excluded to reduce contamination.

Host species identification was initially carried out by experienced field biologists. Further confirmation was based on analysing the cytochrome *c* oxidase subunit I (COI) gene. COI sequences were first obtained from the assembled contigs and then by Sanger sequencing. They were subsequently compared against the NCBI non-redundant nucleotide database and the BOLD database (http://www.boldsystems.org/) for host species identification and confirmation.

RNA library construction and sequencing. On the basis of the complexity of the component samples, our libraries were divided into two categories: (i) simple libraries that contained single or multiple individuals from one or two closely related species; (ii) mixed libraries that contained multiple species from a particular taxonomic group. For example, the library WLJQ is a mix of individuals from the order Decapoda sampled in the East China Sea. For some of the mixed libraries, we later sequenced individual species (that is, before pooling) to assist genome characterization.

To construct each library, the processed samples were first washed with a standard, sterile, RNA and DNA-free PBS solution (GIBCO). This washing was performed three times, and each time the solution was pipetted to agitate the solution and remove the surface organisms/material while keeping the organisms/tissue intact. The samples were then homogenized in 500–700 μl PBS solution using the Mixer mill MM400 (Restsch). Total RNA was extracted using TRIzol LS reagent (Invitrogen) and subsequently purified using EZNA Total RNA Kit (OMEGA). Aliquots of the resultant RNA solutions were then pooled in equal quantity and quality checked using an Agilent 2100 Bioanalyzer (Agilent Technologies) before library construction and sequencing. For most libraries we used the TruSeq total RNA Library Preparation protocol (Illumina). rRNA was removed using either the Ribo-Zero-Gold (Human–Mouse–Rat) Kit (Illumina) or the Ribo-Zero-Gold (Epidemiology) Kit (Illumina). For five libraries we the used TruSeq mRNA Library Preparation protocol (Illumina) that only targeted RNA with poly(A) tails, although these sequencing results were not used in the quantification. The information on library construction methods for each pool can be found in Supplementary Table 1. Paired-end (90 or 100 bp) sequencing of each RNA library was performed on the HiSeq 2000 platform (Illumina). All library preparation and sequencing was carried out by BGI Tech.

Sequence assembly and RNA virus discovery. For each library, sequencing reads were quality trimmed and assembled *de novo* using the Trinity program[23] with default parameter settings. No filtering of host/bacterial reads was performed before the assembly. The assembled contigs were first compared (using blastx) against the database of all reference RNA virus proteins downloaded from GenBank, which include those within the taxonomic classes ssRNA viruses (txid 439488), dsRNA viruses (txid 35325), and *Deltavirus* (txid 39759). We set the *e*-value to 1×10^{-5} to maintain high sensitivity and a low false-positive rate. To detect highly divergent viruses, we performed domain-based blast by comparing the assembled contigs against the Conserved Domain Database (CDD) version 3.14 with an expected value threshold of 1×10^{-2}. Sequences with positive hits to the domain RNA_dep_RNAP (cd01699) were retained. After the initial screening, potential false-positives were discovered by (i) comparing (blastx) putative viral contigs against the entire non-redundant protein database, and (ii) inspecting the sequence alignment for conserved domains. The quality-filtered viral sequences were incorporated into the reference protein database for a second round of blastx.

To identify potential retroviruses and retrotransposons, we examined the domain blast results for any hit to the superfamily reverse-transcriptase-like domain (RT_like, cl02808), excluding those related to RNA_dep_RNAP (cd01699). To avoid false positives, we used a higher *e*-value threshold (1×10^{-5}). All putative reverse transcriptases recovered were aligned to related proteins to determine the presence of key motifs.

Confirmation and extension of virus genomes. Viral contigs with unassembled overlaps or from the same scaffold were merged using the SeqMan program implemented in the Lasergene software package v7.1 (DNAstar). Gaps were filled by RT–PCR and Sanger sequencing. To confirm the assembly results, reads were mapped back to the full length genome with Bowtie2 (ref. 24) and inspected using the Integrated Genomics Viewer[25]. For genomes with novel structures or that contained sequences originating from lateral gene transfer events, we verified the complete or near complete viral genome by designing overlapping primers based on the assembled sequences (Supplementary Table 2). To check these viruses have no DNA stage, we used PCR and Sanger sequencing to examine the DNA extracted from the same set of viruses on which we performed RNA genome confirmation (Supplementary Table 2). Finally, genome termini were determined by RNA circularization or 5′/3′ RACE kits (TaKaRa) as described previously[3].

Transcriptome annotation. For each library, we annotated the top 1,000 most common transcripts. The quantity of the transcripts were determined using the RSEM program[26] implemented in Trinity. The top 1000 common transcripts were then compared against four databases: (i) the non-redundant protein database (nr), (ii) the non-redundant nucleotide database (nt), (iii) the whole-genome shotgun database (wgs), and (iv) the Conserved Domain Database (CDD). The resultant information was used to identify the origin of sequences. Host and mitochondrial genes were identified using two criteria: (a) well-characterized domain/functional information from the domain-based blast results or, in the case of non-coding sequences, from the results of the blastn search against the nt database, and (b) the presence of identical sequences in the host genome or homologous genes in the genome of related host taxa. RNA virus genomes and retrotransposons were identified as described under the 'Sequence assembly and RNA virus discovery'

section. Bacterial contigs were identified if they exhibited high nucleotide similarity (>80%) to a particular bacterial genome. Finally, the transcripts of DNA viruses were identified if they shared protein sequence similarity with either viral polymerases or virus-specific genes (for example, capsid proteins) described previously. The remaining contigs were tentatively annotated as 'undetermined'.

Determination of additional virus genome segments. In those viruses with multiple segments we used various strategies to search for genome segments other than the RdRp. The majority of such segments were found based on their homology to the proteins of related reference viruses. For segments that encode proteins with no known homologues, we used *in silico* approaches that collectively utilize information on RNA quantity, protein structure, and/or conserved genome termini. To determine that these segments belonged to the same virus, we checked: (i) the sequencing depth of the segments; (ii) the presence of inverted complementary genome termini or conserved regulatory sequences in non-coding regions of the genome; (iii) whether the segments were found in the same samples; and (iv) the phylogenetic positions of related viral proteins. For example, Wuhan cricket virus 2 had six potential segments, of which only two exhibited homology to known viruses. The remaining segments were initially identified as 'undetermined' protein-coding contigs which, like the segments encoding the RdRp and capsid, were the most frequent contigs in the transcriptome. Further alignment of the six segments revealed conserved stretches at both the 5′ and 3′ ends (confirmed by RACE), implying they are derived from the same virus. Similarly, the five unknown segments of Changping earthworm virus 2, a divergent member of the 'Orthomyxo' clade, were identified by RNA quantity and the presence of the same inverted complementary genome termini. In addition, two of the segments were identified as potential glycoprotein genes because both encoded proteins had an N-terminal signal domain, a C-terminal or mid-point transmembrane domain and putative glycosylation sites.

Despite our best efforts to address the lack of sequence similarity, it remained difficult to fully characterize the genome segments in a number of divergent viruses. In some of these cases the viruses were likely to be unsegmented even though related viruses appeared to harbour multiple segments. For example, in the case of the nematode-associated lineage within the 'Bunya–Arena' clade (that is, that including Shayang *Ascaridia galli* virus 1), our annotation of contigs with similar abundance levels suggested they were either of host origin or derived from other viruses, rather than representing another protein-coding segment.

It is, however, critical to acknowledge that all segment identification is tentative at this stage, and requires confirmation from virus isolation.

Estimation of viral transcript frequency. To help determine the frequency of viral RNAs, we estimated the percentage of reads that mapped to viral RNA within the transcriptome of each host. To reduce any bias caused by the unequal efficiency of rRNA removal during library preparation, we first removed reads that mapped to rRNA contigs from each library. The remaining reads were then mapped to the entire collection of virus sequences within the library, from which we calculated the overall percentage of viral reads. The proportion of individual viral RNAs was then estimated based on the mapping results. To confirm those results in which viral RNA comprised a large percentage of transcripts within the host transcriptome, we re-extracted the total RNA from aliquots of the original homogenates and performed RNA-seq library preparation and sequencing as described above.

Inference of virus evolutionary history and virus nomenclature. To infer the phylogenetic relationships among RNA viruses we collected all replicase proteins translated from the virus sequence collections described above. For comparison, we downloaded from GenBank reference virus genomes from all established families and floating genera of RNA viruses (excluding retro-transcribing viruses) and non-reference virus genomes not included in the current classification scheme but that are relatively closely related to the viruses discovered here. The viral replicase sequences were then aligned using MAFFT version 7 employing the E-INS-i algorithm[27]. All alignments were trimmed so that they only contained the RdRp and its neighbouring conserved domains. All ambiguously aligned regions were then removed using the TrimAl program[28]. For each sequence alignment, the best-fit model of amino acid substitution was determined using ProtTest 3.4 (ref. 29). Phylogenetic trees were then inferred using the maximum likelihood approach (ML) implemented in PhyML version 3.0 (ref. 30), employing Subtree Pruning and Regrafting (SPR) branch-swapping. Branch support was accessed using an approximate likelihood ratio test (aLRT) with the Shimodaira–Hasegawa-like procedure as implemented in PhyML.

The (provisional) naming of viruses was based on the following approach: the name of a virus characterized by a high proportion of RNA transcripts (>0.1% of non-rRNA reads) contains information on the geographic location of sampling, the host common name (such as 'tick'), and a virus number; whereas the name of a virus characterized by a low proportion of RNA transcripts (<0.1% non-rRNA reads), for which host assignments are less certain, contains information on geographic region of origin, closest family/genus (for example, 'astro-like'), and virus number. The strain name of each virus (shown in each detailed tree, Supplementary Data 1–21) comprises the library abbreviation followed by its contig number.

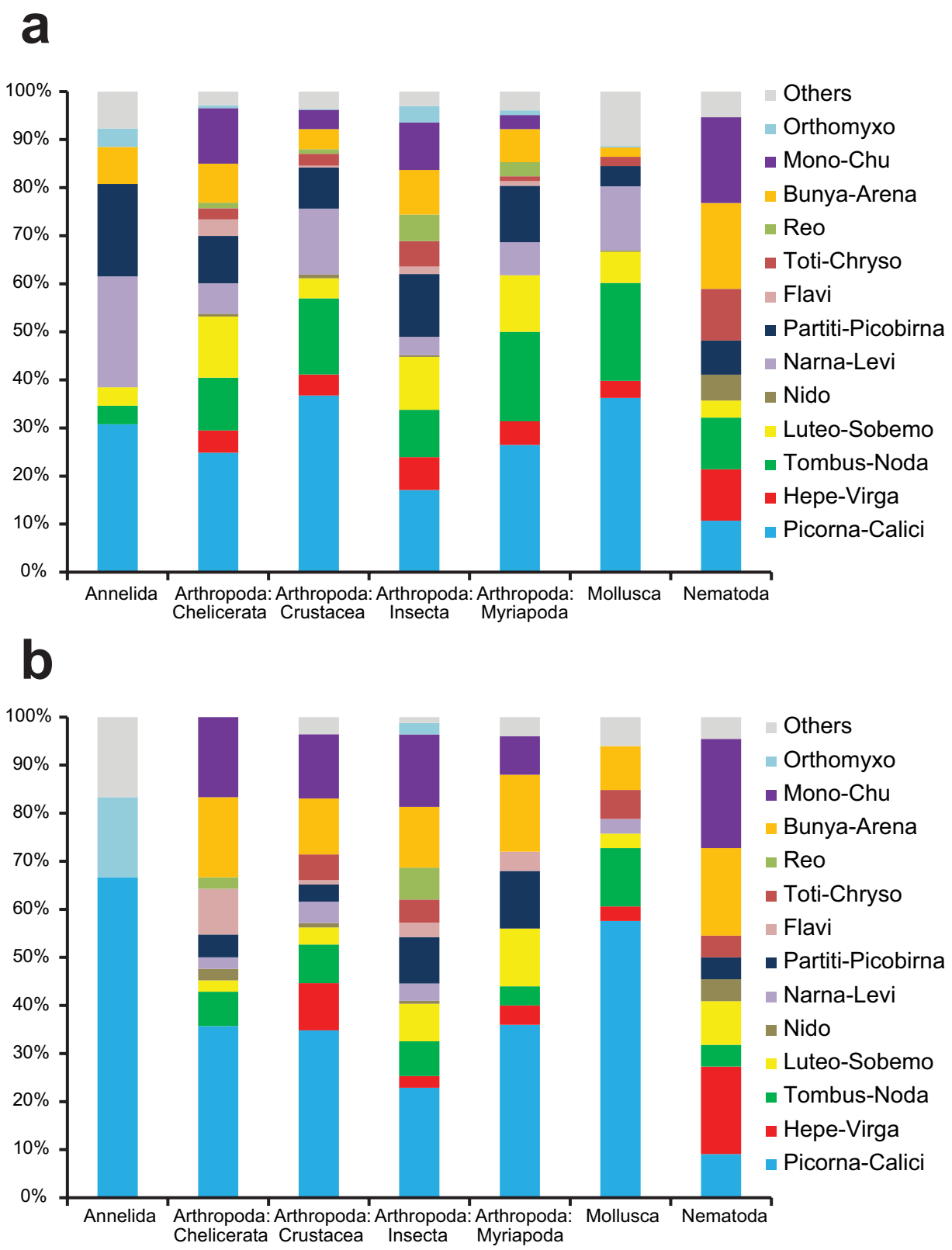
Virus genome annotation. Potential viral open reading frames (ORFs) were predicted based on two criteria: (i) the predicted amino acid sequences were longer than 100 amino acids in length, and (ii) if a short ORF (<200 amino acids) was completely nested within a larger one it was not regarded as a potential ORF unless it had a homologue in a closely related virus. The annotations of these ORFs were mainly based on comparisons to the Conserved Domain Database. Domains that were potentially subject to lateral gene transfer were further examined by sequence alignment and phylogenetic analyses. For the remaining ORFs, we predicated their potential functions by blast searches against the nr protein database with an *e*-value threshold of 1×10^{-5} and by primary protein structure predication using the programs SignalP, TMHMM, and NetNGlyc available through the website (http://www.cbs.dtu.dk/services/).

Analysis of recombination and lateral gene transfer. For each newly identified protein, we searched for any potential homologues against all RNA virus proteins (including those newly identified here), all DNA virus proteins, and those from the cellular organisms (using a subset of the nr database). We also identified homology if the proteins matched the same domain in the structure-based blast. On the basis of these results, we identified several well-established homologous protein clusters. We then mapped these protein clusters onto the RdRp phylogenies, which enabled us to identify topological inconsistencies that were likely to be the result of lateral gene transfer. To identify homologous recombination events, we compared the phylogenies of each homologous protein cluster to that of the RdRp. To measure the degree of phylogenetic incongruence, we transformed the two phylogenies into patristic genetic distance matrices and calculated the Pearson correlation coefficient.

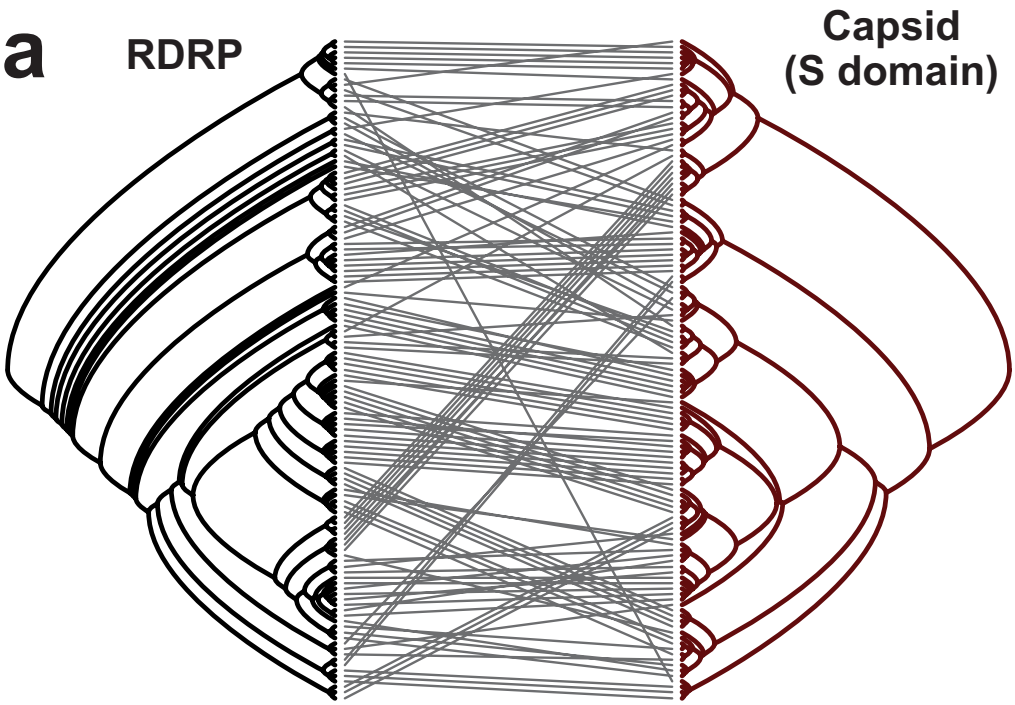
Analysis of endogenous virus elements. All genomes from cellular organisms available in GenBank were downloaded and incorporated into our local genome nucleotide database. Endogenous copies of the exogenous viruses (that is, EVEs) were detected using the tblastn algorithm against this database. The query involved amino acid sequences translated from both the virus genomes newly identified here as well as the reference virus genomes used in this study. The threshold for the search was set to 100 amino acids for length and 1×10^{-20} for *e*-value. For each potential endogenous virus, the query process was reversed to determine their corresponding phylogenetic group. The results were also checked manually to exclude those sequences involved in lateral gene transfer.

Data availability. All new sequence reads generated here are available at the NCBI Sequence Read Archive (SRA) database under the BioProject accession PRJNA318834 (Supplementary Table 1). All virus genome sequences generated in this study have been deposited in GenBank under the accession numbers KX882764–KX884872 (Supplementary Table 2). All viruses discovered in this study (fasta format), sequence alignments (fasta format), and phylogenetic trees (newick format) are available at https://figshare.com/articles/Redefining_the_invertebrate_RNA_virosphere/3792972.

23. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).
24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
25. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
26. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26,** 493–500 (2010).
27. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
28. 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).
29. 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).
30. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52,** 696–704 (2003).



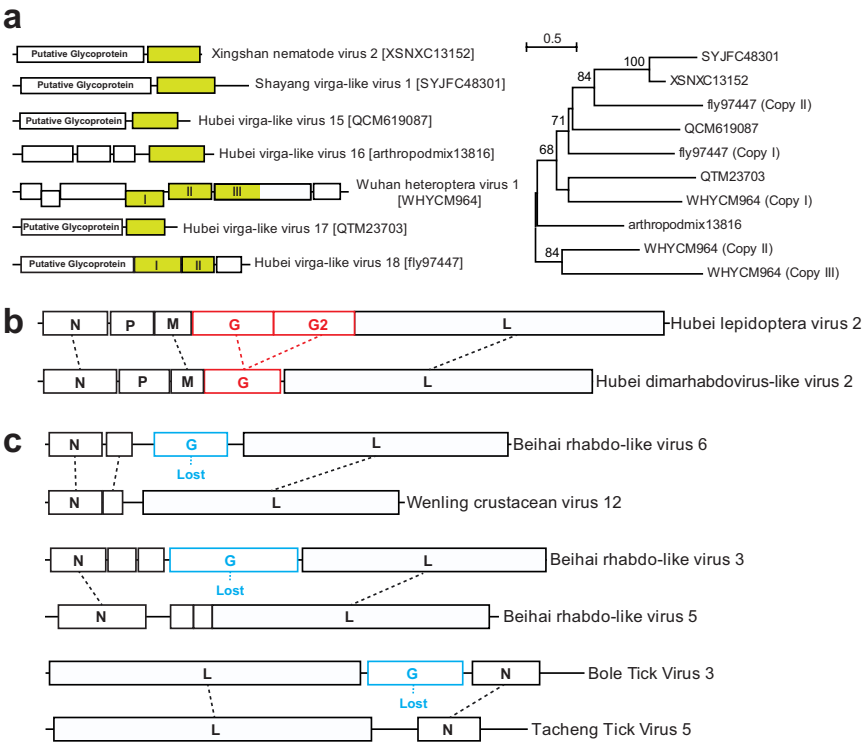
Extended Data Figure 1 | The contribution of major viral clades to the total virome of each host phylum/order. a, **b**, These analyses are based on viruses at all frequency levels (**a**), and viruses in which the frequency exceeds 0.1% of the total number of non-rRNA reads (**b**).



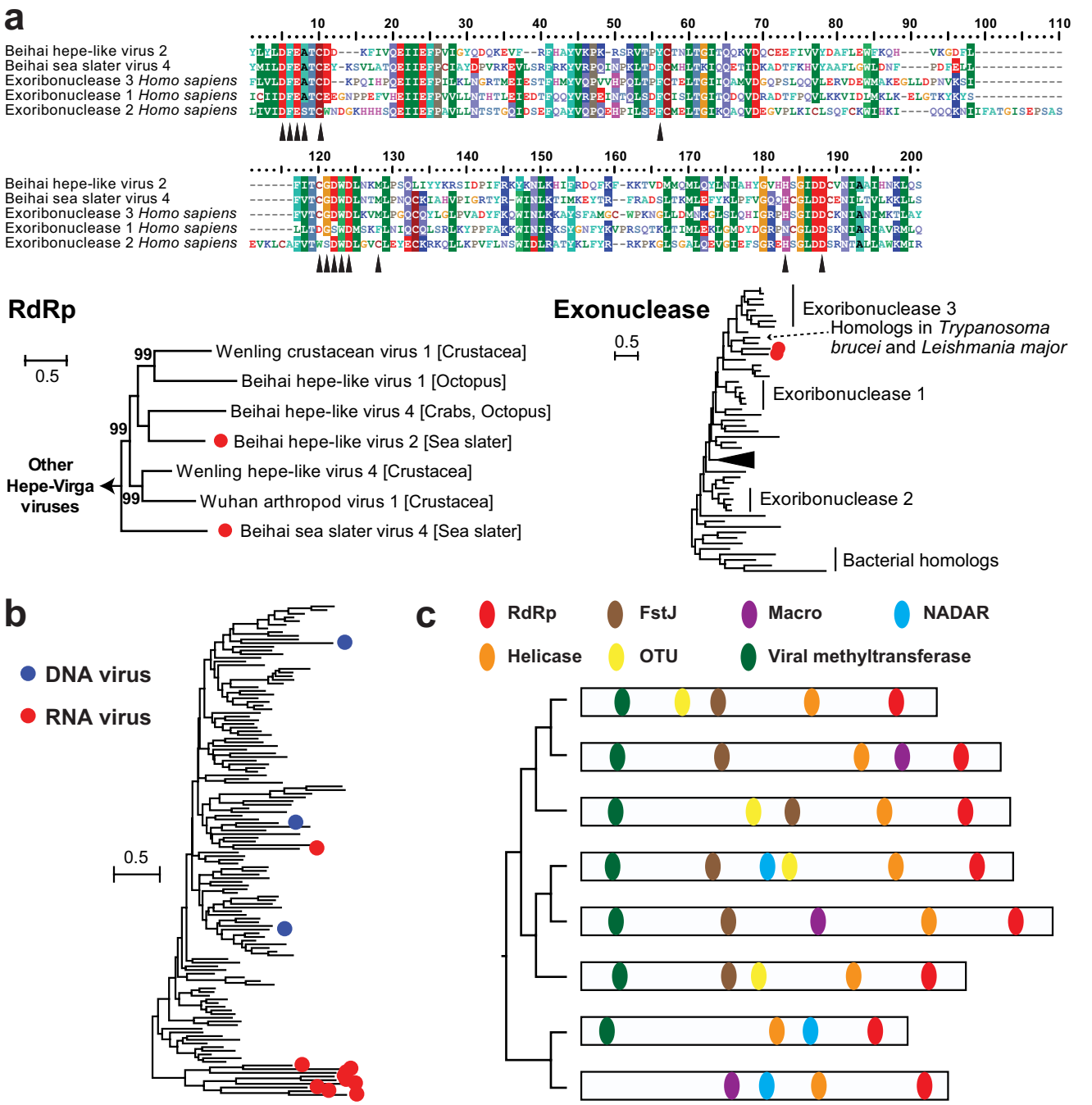
b

RdRp	Structural Protein	Correlation
Tombus-noda viruses (only unsegmented)	S domain	0.286
Aquatic Picorna-like Cluster	Picorna-like capsid	0.712
Dicistroviridae-related viruses	Picorna-like capsid	0.695
Dimarhabdovirus group	Nucleoprotein	0.725
Chuviruses and relatives (only unsegmented)	Chuvirus-like glycoprotein	0.559
New -ve RNA virus (Qinvirus)	Putative nucleoprotein	0.231
New +ve/ds RNA virus (Weivirus)	*Alvenoviridae*-like capsid	0.010

Extended Data Figure 2 | Phylogenetic incongruence between the RdRp and structural proteins. a, Match between the phylogenies of the RdRp and coat proteins (S-domain like) for non-segmented members of the Tombus–Noda clade. The relationship between the two phylogenies is displayed to maximize topological congruence. **b**, The degree of phylogenetic incongruence for different pairs of structural and non-structural phylogenies. The comparisons were based on patristic distances matrices derived from the phylogenies.



Extended Data Figure 3 | The gain and loss of RNA virus structural proteins. a, The parallel acquisition of multiple copies of structural proteins by viruses within the Hepe–Virga clade. Left panel shows an outline of the structural part of their genomes, with homologous structural genes marked in yellow and multiple copies of these proteins within the same genome labelled as 'I', 'II', and 'III'. Right panel shows a maximum-likelihood phylogeny depicting the evolutionary history of the corresponding structural proteins of these viruses. **b**, Acquisition of a glycoprotein in the genome of Hubei Lepidoptera virus 2 from the Mono–Chu Clade. Its genome is compared against that of a closely related virus (Hubei dimarhabdovirus-like virus 2). Homologous proteins are connected with dotted lines, and the target glycoprotein is shown in red. **c**, Three examples of glycoprotein loss in the Mono–Chu Clade. Homologous proteins are connected with dotted lines, and the target glycoproteins are shown in blue.



Extended Data Figure 4 | Lateral gene transfer between RNA viruses and cellular organisms. a, Evolutionary origin of two exoribonucleases (cd06133) in two sea-slater-associated viruses (Beihai hepe-like virus 2 and Beihai sea slater virus 4). Top, alignment of viral and (human) cellular exoribonucleases. The solid triangles indicate the key catalytic sites. Lower left panel shows the phylogenetic positions of the two viruses (marked with solid red circles) whose genomes contain these exoribonucleases. The host information for each virus is shown in parentheses. Lower right panel shows the phylogenetic position of the virus exoribonucleases (solid red circle) in the context of cellular exoribonucleases. **b**, Evolutionary origin of viral serine proteases (cd00190). The phylogeny contains serine proteases from RNA viruses (solid red circles), DNA viruses (solid blue circles) and cellular organisms. Serine proteases from RNA viruses are either highly divergent or group within the diversity of cellular proteins. **c**, Relative positions of different protein domains in the replicase of selected Hepe–Virga viruses. The domains are shown as ovals and marked with different colours, and comprise: RdRp (cd01699), Helicase (pfam01443), FstJ (pfam01728), OTU (OTU-like cysteine protease, pfam02338), Macro (cl00019), NADAR (cd15457), and viral methyltransferase (pfam01660). More detailed depictions of lateral gene transfer can be found in Supplementary Data 22–36.

Extended Data Table 1 | Distribution of homologous protein clusters across divergent taxonomic groups (RNA viruses, DNA viruses and cellular organisms)

Category	Protein (Domain)	CDD accession	Astro	Flavi	Hepe-Virga	Luteo-Sobemo	Narna-Levi	Nido	Permutotetra	Picorna-Calici	Reo	Tombus-Noda	Toti-Chryso	New +ve RNA Viruses	-ve RNA Viruses	DNA Virus	Cellular Organisms
Structural	Capsid (S domain)	pfam00729	X		X	X	X		X			X		X		X	
	Capsid (peptidase A21)	pfam03566	X		X				X			X					
	Capsid (Alvernavirus core like)	N/A				X	X					X		X			
	Glycoprotein (Okavirus like)	N/A						X							X		
	Glycoprotein (Ferak virus like)	N/A			X										X		
	Glycoprotein (Hemagglutinin-neuraminidase)	pfam00423						X							X		
	Glycoprotein (*Choristoneura rosaceana* alphabaculovirus GP64)	N/A													X	X	
	Glycoprotein (Ostreid herpesvirus 1 ORF68)	N/A													X	X	
Non-structural	RNA Helicase (picorna-like)	pfam00910					X			X							
	Viral methyltransferase	pfam01660			X							X					
	FtsJ-like methyltransferase	pfam01728		X	X										X		X
	Macro (ADP-ribose binding)	cl00019		X	X			X		X					X		X
	Ribonuclease III (dsRNA binding)	cl00054							X	X	X	X	X				X
	3'-5' exonucleases	cd06133			X												X
	E. coli swarming motility protein (NADAR)	cd15457			X			X		X							X
	OTU-like cysteine protease	pfam02338		X	X										X		X
	2OG-Fe(II) oxygenase	pfam13532			X					X		X				X	X
	Trypsin-like serine protease	cl21584	X	X	X	X		X		X						X	X
	RNA 2'-phosphotransferase	pfam01885								X							X