Egoviruses: distant relatives of poxviruses abundant in the gut microbiome of humans and animals worldwide

Morgan Gaïa[1,2], Hans-Joachim Ruscheweyh[3], A. Murat Eren[4,5,6,7,8], Eugene V. Koonin[9], Shinichi Sunagawa[3], Mart Krupovic[10], and Tom O. Delmont[1,2#]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France.
[2] Research Federation for the study of Global Ocean systems ecology and evolution, FR2022/Tara GOSEE, Paris, France.
[3] Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, 8093, Zürich, Switzerland.
[4] Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany.
[5] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany.
[6] Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany.
[7] Marine Biological Laboratory, Woods Hole, MA, USA.
[8] Max Planck Institute for Marine Microbiology, Bremen, Germany.
[9] National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA
[10] Institut Pasteur, Université Paris Cité, Archaeal Virology Unit, Paris, France.

#Corresponding author (contact: tomodelmont@gmail.com)

Abstract: Large and giant double-stranded DNA viruses within the phylum *Nucleocytoviricota* are diverse and prevalent in the environment where they substantially affect the ecology and evolution of eukaryotes. Until now, these viruses were only sporadically found in the digestive system of vertebrates. Here, we present the identification and genomic characterization of a proposed third order of viruses within the class *Pokkesviricetes* that currently consists of poxviruses and asfuviruses. Members of this newly identified order we provisionally named "*Egovirales*" are abundant in the digestive system of vertebrates worldwide and occur in high abundances in >10% of livestock animals, >2% of humans, and wild animals. Egoviruses have linear genomes up to 360 kbp in length that likely produce multilayered icosahedral capsids, similar to those of asfuviruses. The diversity of egoviruses already far exceeds that of all known poxviruses and animal-associated asfuviruses. Phylogenetic analyses and patterns of virus distribution across vertebrates suggest that egoviruses can be either specialists or generalists associated with a single or multiple vertebrate species, respectively. Notably, one egovirus clade is human-specific, evolutionarily constrained, and spread across continents, demonstrating a long-lasting association between *Egovirales* and the human population on the global scale. Egoviruses not only expand the ecological and evolutionary scope of *Pokkesviricetes*, but also appear to be the most diverse, widespread, and abundant group of double-stranded DNA viruses infecting eukaryotic cells in the digestive system of vertebrates.

Key words: Giant viruses; gut microbiome; *Poxviruses*; *Asfuvirales*; eukaryotic viruses; human virus; metagenomics.

Introduction:

Pokkesviricetes is a class of eukaryotic viruses with large, linear double-stranded DNA genomes within the phylum *Nucleocytoviricota*. The class is divided into two orders, *Chitovirales* (poxviruses) and *Asfuvirales*[1]. Poxviruses infect a broad diversity of animals including humans and insects[2,3]. They include the now eradicated smallpox virus, monkeypox viruses that continue to sporadically infect humans, molluscum contagiosum virus causing benign tumors in immunocompromised individuals, as well as vaccinia virus, a classic model of molecular virology[4–7]. Asfuviruses infect hosts ranging from unicellular eukaryotes to animals[8] and include the African swine fever virus, which causes deadly outbreaks in pigs and boars[9], as well as a wider range of viruses mostly characterized in co-culture with free-living amoeba[10,11] or directly from the environment (e.g., within plankton) using genome-resolved metagenomics[12–14]. Aside from the few viruses of global concern, as declared by the World Health Organization, the class *Pokkesviricetes* is of notable relevance to our understanding of the diversity, ecology, and evolution of eukaryotic DNA viruses[15,16].

Decades of investigations using cultivation, PCR surveys, serology, and metagenomics have indicated that viruses of the phylum *Nucleocytoviricota* are mainly transient and thus rarely abundant in the digestive system of humans and other animals (the gut microbiome)[13,17,18]. Until now, not a single *Nucleocytoviricota* clade had been found to be consistently associated with or prevalent in the gut microbiome, in a sharp contrast with the broad distribution and remarkable diversity and abundance of these viruses in the environment. The scarcity of eukaryotic double-stranded DNA viruses in the gut[13,17,18] could in part be explained by the fact that this environment is overwhelmingly dominated by bacteria and bacteriophages, with the eukaryotic component of the normal gut microbiome being far less diverse (e.g., mainly fungi and *Blastocystis* for humans)[19–22]. Overall, it has been widely assumed that *Nucleocytoviricota* viruses are not prevalent in the gut microbiome due to the limited number of eukaryotic cells they are able to infect in these ecosystems.

Here, we explored the global diversity and distribution of *Pokkesviricetes* by surveying publicly available metagenomic assemblies from a broad range of environmental samples and animal-associated microbiomes. We discovered a third *Pokkesviricetes* order (proposed "*Egovirales*") comprising numerous viruses that appear to be almost exclusively found in the gut microbiomes of humans, livestock, and wild animals. Notably, despite the dominance of bacteria in this environment, we found these viruses to be abundant in more than 2% of the human population. Egoviruses have large, complex genomes and represent some of the most abundant eukaryotic DNA viruses naturally occurring in vertebrates.

Results and discussion

"*Egovirales*": a third *Pokkesviricetes* order

DNA–dependent RNA polymerases are evolutionarily informative gene markers present in most viruses of the phylum *Nucleocytoviricota*, including all *Pokkesviricetes*[1,23]. We used a comprehensive hidden Markov model (HMM) for largest RNA polymerase subunit A (RNApolA)[23,24] to search 85,924 metagenomic assemblies covering a wide range of

environmental samples as well as thousands of samples from the digestive system of humans (e.g., [25–31]), livestock and wild animals (e.g., [32–40]) (Table S1). This search identified 1,969,342 RNApolA genes, which we linked to main cellular and viral taxonomic lineages using a custom reference database (see Methods). Among the *Pokkesviricetes*, we identified a single poxvirus most closely related to *Eptesipox* occurring in the gut microbiome of a marsupial carnivore from Australia[32] as well as 240 asfuviruses occurring only in the environment, predominantly in aquatic ecosystems (Table S2). These results are in line with our current knowledge on the global niche partitioning of *Pokkesviricetes* and contribute, alongside previous metagenomic surveys[13,14], to the rapid expansion of the known diversity of asfuviruses in the environment.

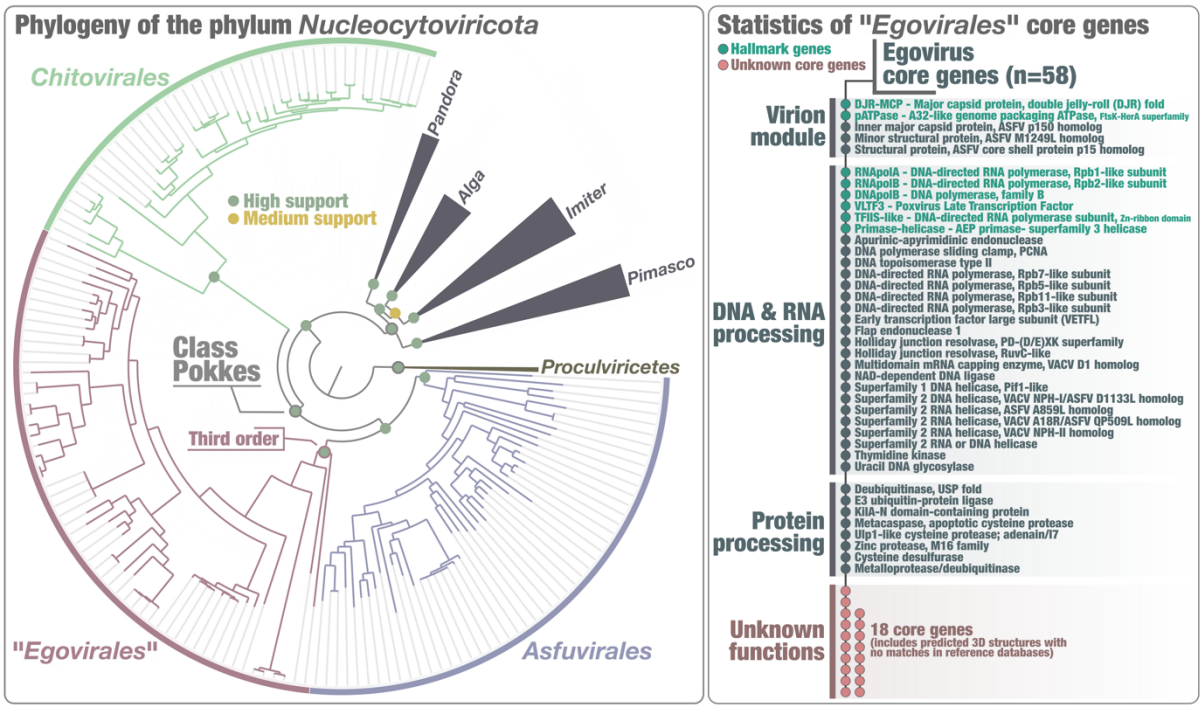


Figure 1: Evolution and gene functions of egoviruses. Left panel displays a maximum-likelihood phylogenomic analysis of members of *Nucleocytoviricota* from the GOEV database, along with the 57 *Egovirales* genomes >50 kbp in length. The tree was built using concatenated multiple alignments of four hallmark proteins of *Nucleocyovoricota* (DNApolB, RNApolA, RNApolB, TFIIS; 4,122 sites) using IQTree with the LG+F+R10 model. The tree was rooted with the putative class *Proculviricetes* and visualized using anvi'o v8. Highlighted supports at nodes (dots in the tree) were considered high (approximate likelihood ratio (aLRT) ≥ 80 and ultrafast bootstrap approximation (UFBoot) ≥ 95, in green) or medium (aLRT ≥ 80 or UFBoot ≥ 95, in yellow; see Methods). Right panel summarizes the functional annotation of the 58 *Egovirales* core genes.

In addition to poxviruses and asfuviruses, we identified a clade of previously undescribed DNA viruses related to these two viral orders that were prevalent in the gut microbiome, which we provisionally named egoviruses. We reconstructed 13 complete, non-redundant egovirus genomes (average nucleotide identity <98%) from the gut microbiomes of humans, buffalos, goats, chicken, deer, and pigs. These linear genomes are 160-360 kbp in length and contain 55 bp inverted terminal repeats, a feature characteristic of *Pokkesviricetes*. The egovirus genomes encompass all eight *Nucleocytoviricota* hallmark genes[1,23] (Table S3), with the DNA polymerase B coding region often split into two separate open reading frames. We also characterized 44 additional, partial, non-redundant egovirus genomes (up to 258 kbp in length with a

minimal cut-off of 50 kbp) mainly from gut microbiomes of a wider diversity of animal species (Table S3). Phylogenetic analysis of the hallmark genes confidently identified the 57 egovirus genomes as a distinct major clade within the class *Pokkesviricetes* (Figure 1 and Figure S1). Depending on the hallmark genes used, this clade displays affinity to either *Asfuvirales* (DNApolB and TFIIS) or *Chitovirales* (six remaining genes). Based on these findings, it appears that egoviruses should become a third order within *Pokkesviricetes* which we propose to name "*Egovirales*". *Egovirales* is the only clade from the phylum *Nucleocytoviricota* found to be consistently present in the gut microbiome.

A distinct functional layout of egovirus genomes

The 13 complete *Egovirales* genomes range from 21% to 42% in GC-content and encompass between 120 and 294 predicted genes (Table S3). We built a genomic database covering *Egovirales*, *Asfuvirales*, *Chitovirales* as well as representatives of the class *Megaviricetes*[1], and characterized gene clusters based on amino acid sequence similarities to characterize the gene pool of egoviruses (see Methods). We identified 58 core genes occurring in at least 10 of the 13 complete *Egovirales* genomes (>75% occurrence), with the majority being present in single copy (Figure 1 and Table S4). Aside from the *Nucleocytoviricota* hallmark genes, most of the functionally annotated core genes are implicated in genome replication, transcription and associated processes such as DNA repair, precursor synthesis and transcript maturation (various DNA and RNA helicases, endonuclease, DNA topoisomerase type II, thymidine kinase, RuvC-like Holliday junction resolvase, mRNA capping enzyme), transcription regulation, protein degradation (multiple proteases including a metacaspase), and post-translational modification of proteins (ubiquitin-protein ligase and deubiquitinylases). These core genes represent key functionalities of the viruses in the phylum *Nucleocytoviricota* that also occur in most members of the orders *Asfuvirales*, *Chitovirales* and *Megaviricetes*[13,14,24] (Table S4). Apart from the genes that could be functionally annotated, 18 core genes could not be assigned any function through sequence comparisons. These core genes are either exclusive to *Egovirales,* or are highly divergent from putative orthologs in other clades of the phylum *Nucleocytoviricota* (Table S4). Overall, the 58 *Egovirales* core genes recapitulate the core functionalities of *Nucleocytoviricota* (especially regarding DNA replication and transcription) and *Pokkesviricetes* in particular, while also including uncharacterized functions that could reflect a unique lifestyle adapted to the conditions of vertebrate digestive systems.

Analysis of the virion morphogenesis module sheds light on the key features of the egovirus virions. Egoviruses encode two structurally distinct major capsid proteins (MCPs), the A32-like genome packaging ATPase, a Ulp1-like capsid maturation protease and several minor capsid proteins (Figure 1). One of the two MCPs has the double jelly-roll (DJR) fold[41], with orthologs across *Nucleocytoviricota*, including poxviruses and asfurviruses (Figure 2). By contrast, the second MCP has an α-helical fold and is exclusive to asfuviruses in which it forms an internal icosahedral shell located inside of a larger, external icosahedral capsid formed from the DJR-MCP[42]. Furthermore, egoviruses encode a homolog of the asfuviral core shell protein p15 homolog which bridges the viral nucleoprotein complex and the internal capsid to the lipid membrane sandwiched between the inner and outer icosahedral shells[43]. The double DJR-MCP characteristic of *Nucleocytoviricota* was particularly challenging to identify in *Egovirales* due to its extensive protein sequence divergence compared to the closest reference proteins. Our

initial searches using a dedicated HMM, which successfully recovered DJR-MCPs from all previously described families of the phylum *Nucleocytoviricota*[23,24], systematically failed to recognize this protein. Eventually, HHpred[44] profile-profile comparisons yielded a partial match to the DJR-MCP of African swine fever virus, and structural modeling with ColabFold[45] confirmed that the predicted egovirus MCP has the DJR fold (Figure 2). The *Egovirales* DJR-MCPs are unusually large (average of 806 amino acids among the complete *Egovirales* genomes) and the region facing exterior of the capsid is highly diversified among egoviruses (Figure S2). These trends echo the previous observations on the large MCPs of faustoviruses from the order *Asfuvirales*[10]. Yet, *Egovirales* and *Chitovirales* are firmly established as sister clades in the DJR-MCP phylogeny (Figure S1), despite the substantial protein length differences between the two orders (Figure 2). So far, the extent of protein sequence diversification among egoviruses prevented the prediction of the high-confidence 3D structure for the second MCP. Collectively, the presence of the three structural proteins suggests that egoviruses form icosahedral multilayered capsids that are similar to those of *Asfuvirales*[42], but differ from the brick-shaped capsids of poxviruses[46].

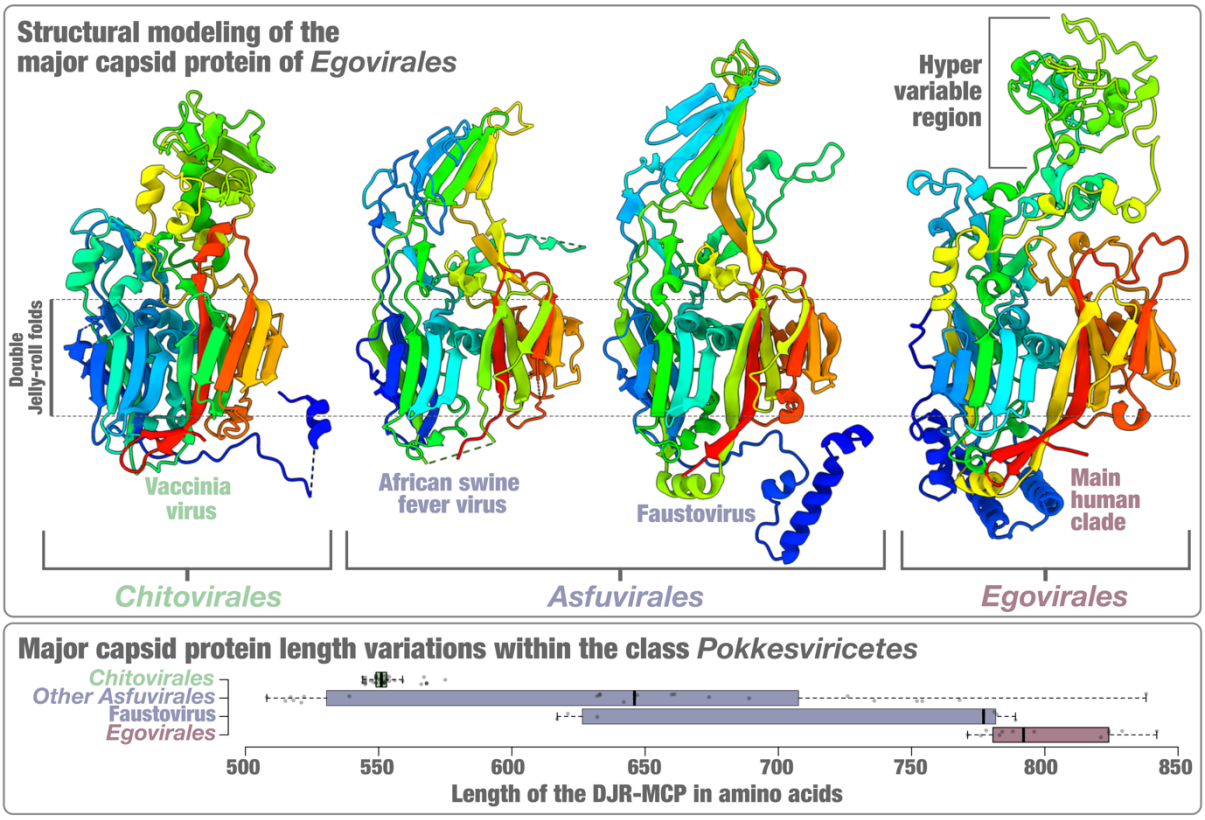


Figure 2: The *Egovirales* major capsid protein. Top panel displays structural modeling of an *Egovirales* DJR MCP (human main clade) and comparison to the structures of DJR MCPs (from left to right: 3sam, 6ku9, 5j7o) from three reference *Pokkesviricetes* MCPs. Bottom panel displays boxplots summarizing variation in protein length for the DJR-MCPs of egoviruses, faustoviruses, other clades of asfuviruses, and *Chitovirales*.

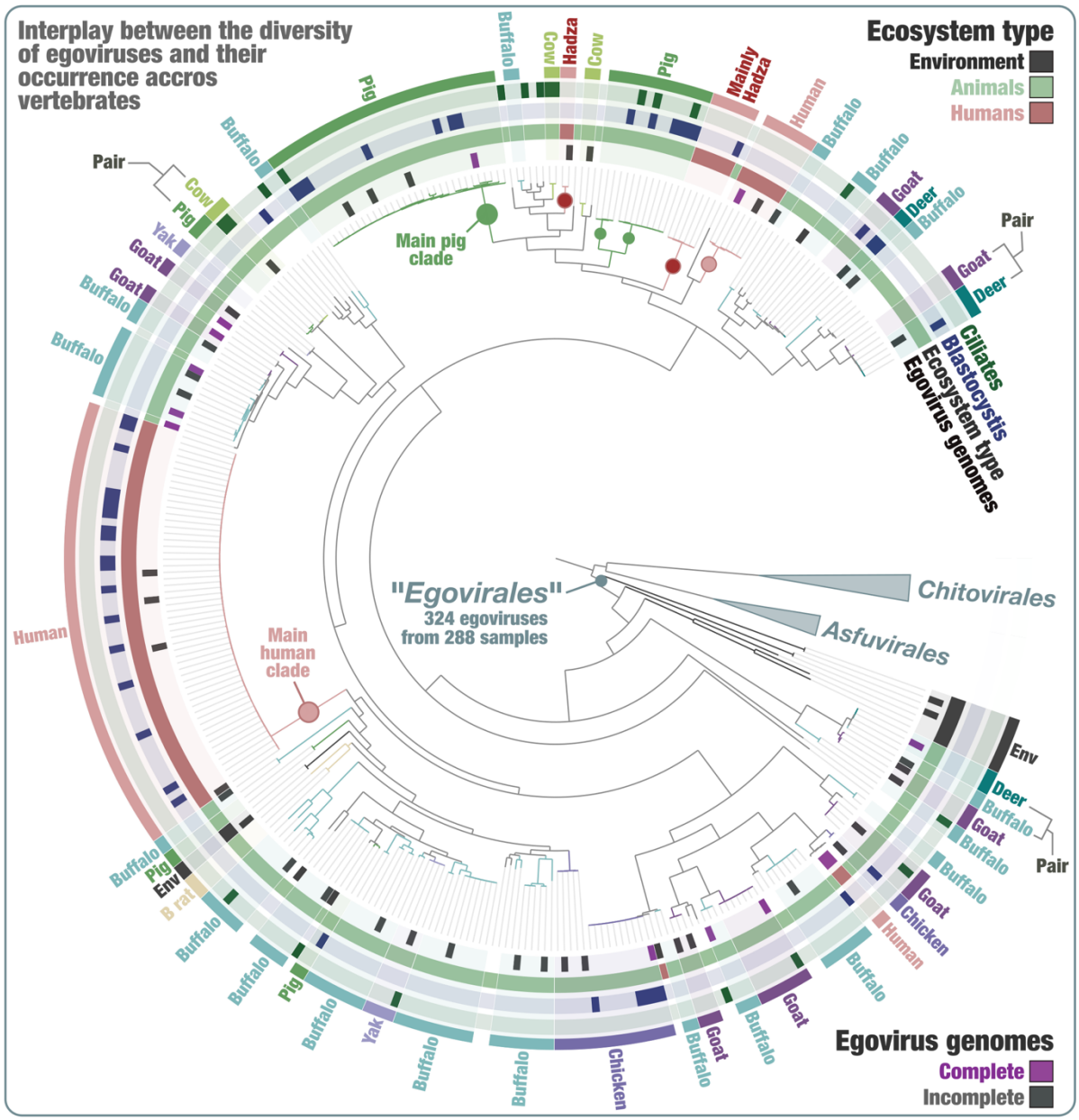
Egoviruses are prevalent in the gut microbiome of humans and other vertebrates

We built an HMM specifically for the *Egovirales* DJR-MCP to complement the RNApolA survey and expand the egovirus signal across metagenomes. Overall, we identified genes corresponding to the RNApolA and/or DJR-MCP of *Egovirales* in 864 metagenomic assemblies and used this data to perform a global survey of the occurrence and diversity

of egoviruses (Table S3). The RNApolA phylogeny showed that the *Egovirales* diversity extended far beyond the 57 complete and partial non-redundant genomes (Figure 3). Only 9 egovirus-like RNApolA genes were retrieved from environmental samples, including an aquifer[47], a cold sweep[48], and marine sediments[49]. These formed four relatively small clades, two of which (from marine sediments) were the most basal branches for *Egovirales*. In contrast, all the remaining RNApolA genes (>97% of the data) originated from gut microbiomes representing the bulk of *Egovirales* diversity. The DJR-MCP phylogeny recapitulated these trends, and in this case >99% of the data originated from gut microbiomes (Figure S3).

Among the 864 *Egovirales*-positive samples, egoviruses were detected in humans of various age and geographic location (n=217; positivity rate of 2.2%), with a notable increase of signal among hunter-gatherers from Tanzania (the Hadza; positivity rate of 5.1%). The extensive sequencing efforts in the study of Hadza populations[25] could in part explain this enhanced signal. Nevertheless, some of the human-associated clades were either only or mostly detected among the Hadza (Figure 3 and Figure S3), suggesting that lifestyle could influence the diversity of human-associated egoviruses. Egoviruses were also well detected in the digestive system of pigs (n=225; positivity rate of 11%), buffalos (n=193; positivity rate of 27.2%), goats (n=95; positivity rate of 15.4%), cows (n=44; positivity rate of 8.4%), chicken (n=38; positivity rate of 11.2%), and to a lesser extent, also in other livestock animals for which fewer samples were available (deer, yak, sheep, zebu, reindeer) (Tables S3 and S5). Finally, even fewer metagenomes were available for microbiomes of wild animals. Nevertheless, bamboo rats (rodents), a golden bamboo lemur, an olive baboon, and a white rhinoceros were found to be additional carriers of egoviruses, indicating that primates and other wild mammals also represent a considerable reservoir of egovirus diversity.

Buffalos had by far the highest *Egovirales* positivity rate among the surveyed vertebrate species, with viruses covering a wide range of distantly related clades (Figure 3). The considered cohorts cover seven Chinese provinces and include three buffalo breeds and different segments of their digestive system[33], allowing for a more refined survey of the occurrence of egoviruses within the gut microbiome of a herbivorous animal species (Table S6). First, we found up to four distinct egoviruses co-occurring in the same sample, contrasting with the human gut microbiome, where no more than one egovirus could be detected in the same individual. Second, the four ruminant stomach compartments displayed the highest prevalence of egoviruses: 79% of positive samples for the reticulum, 45% for the abomasum, 38% for the omasum, and 20% for the rumen. Egoviruses were also detected in the cecum (n=19%) and distal colon (n=10%). Finally, the positivity rate reached 25.8% among fecal samples. Based on the fecal samples, egoviruses were detected in all seven Chinese provinces with a positivity rate ranging from 6.7% to 52%, demonstrating the widespread occurrence of *Egovirales* in the buffalo gut microbiome.



Positivity rates across vertebrate species

	Humans	Hadza	Buffalos	Chicken	Goats	Pigs	Cows
Samples screened	8,841	388	670	339	447	1,406	368
Egovirus positivity rate	2.2%	5.1%	27.2%	11.2%	15.4%	11.0%	8.4%
Blastocystis positivity rate	4.9%	8.8%	1.3%	30.4%	2.7%	3.8%	0%
Egovirales-positive samples with signal for Blastocystis	31.8%	50%	0.55%	36.8%	2.9%	16.8%	0%
Ciliates positivity rate	0%	0%	13.1%	0%	2.7%	2.6%	52.5%
Egovirales-positive samples with signal for Ciliates	0%	0%	5.5%	0%	0%	7.7%	64.5%

Figure 3: Overlay of the evolutionary tree of egoviruses and their occurrence among vertebrate species. The top panel figure displays a maximum-likelihood phylogeny of the RNApolA of egoviruses (n=324), using representatives of *Chitovirales* and *Asfuvirales* for rooting (1,357 sites), and the LG+F+R7 model. Layers of information summarize the occurrence of the egovirus genomes >50 kbp in length (including the complete genomes), ecosystem type (environment vs. human vs. other animals), and co-occurrence of *Blastocystis* and ciliates. Finally, selections were made for egovirus clades occurring in the same animal species. Bottom panel summarizes the positivity rate for egoviruses, *Blastocystis* and ciliates in gut metagenomes from five vertebrate species.

The dispersal of *Egovirales* clearly indicates that clades of closely related egoviruses tend to occur in the same animal species. This was most apparent for the two main clades

associated with humans and pigs (Figures 3 and S3). We characterized the main human clade in 138 metagenomic assemblies from 18 studies that cover various countries from multiple continents, including Tanzania (the Hadza), Nepal, Fiji, USA, England, Sweden, Denmark, Germany, Italy, and Spain (Table S3). This clade alone appears to be abundant in more than 1% of the human population. We found that this clade could (1) be associated with various stages of the human lifespan (e.g., in a 3-year old child and a pregnant woman) and (2) remain abundant over time in longitudinal studies of the same individuals (up to nearly one year of sampling). Average amino acid identity was 99.4% among the RNApolA genes and 99.9% among the MCP genes, demonstrating a striking genomic stability of this virus clade among humans at a global scale. The main pig clade was found in 109 metagenomic assemblies from 4 studies that also cover countries from multiple continents. Average amino acid identity was 96.7% among the RNApolA genes and 95.4% among the MCP genes, indicating a more diverse clade compared to the one associated with humans. These results demonstrate that closely related egoviruses forming distinct clades can cover very large geographic regions while being strictly associated with a single vertebrate species. In addition, we found instances of closely related egoviruses occurring in multiple vertebrate species. Examples include goat-deer, pig-cow, and deer-buffalo pairs (Figures 3 and S3). Overall, the results demonstrate the worldwide spread and strong preference of egoviruses towards one or multiple vertebrate species. This congruence between the phylogenetic signal and distribution patterns of egoviruses appears to reflect coevolution with the vertebrate hosts (direct or indirect) and echoes trends observed with well-studied animal-infecting viruses, such as the poxviruses[50], papillomaviruses[51], and foamy viruses[52].

Unicellular eukaryotes in *Egovirales*-positive metagenomic assenblies

Finally, we assessed whether the detection of egoviruses in the gut samples coincides with the presence of any particular group of unicellular eukaryotes, which could serve as egovirus hosts. We identified eukaryotic RNApolA genes in 225 out of the 864 *Egovirales*-positive samples. Two clades explained nearly 95% of this signal: the genus *Blastocystis* (a Stramenopile) and a larger group of ciliates (Figure 3; Table S7). *Blastocystis* is highly prevalent in the digestive system of humans[53–55] and various animals[56,57], whereas ciliates are prevalent in the rumen microbiome of herbivorous mammals (e.g., cow, sheep, buffalo)[58–61] and can also occur in other vertebrates such as pigs[62,63]. The limited eukaryotic diversity co-detected with egoviruses across metagenomic assemblies suggests that *Blastocystis* and ciliates are the only two host candidates which could be considered among unicellular eukaryotes. *Blastocystis* was detected in 137 of the 864 *Egovirales*-positive samples, mainly in humans (n=65), pigs (n=52) and chicken (n=14). In addition, we only detected the ciliates in 87 out of 864 *Egovirales*-positive samples, mainly in pigs (n=40), cows (n=20) and buffalos (n=11). Thus, at the scale of all vertebrate species (Figure 3; Tables S5 and S7), no clear patterns could be found among the metagenomic assemblies that would support one unicellular eukaryotic clade as being the prime host for egoviruses.

Discussion:

Our global metagenomic survey of viruses of the class *Pokkesviricetes* (phylum *Nucleocytoviricota*) led to the discovery of a putative third order, "*Egovirales*", characterized by linear genomes that can reach 360 kbp in length, and occurring mostly

in the gut microbiomes of humans, livestock, and wild animals. Egoviruses have a highly divergent major capsid protein (DJR-MCP) and likely produce viral particles with an icosahedral multilayered structure resembling the virions of African swine fever virus. Egoviruses not only fill major gaps in our knowledge of the ecology and evolution of *Pokkesviricetes*, but also represent the only so far discovered group of viruses with large double-stranded DNA genomes that is consistently detected in human and animal gut microbiomes. Given its evolutionary trajectory and distribution patterns, this previously overlooked major group of viruses with complex genomes is of prime interest to human and veterinary virology.

Until now, the gut microbiome has been considered to contain a limited diversity and abundance of eukaryotic DNA viruses[13,17,18], with the most prevalent ones, such as anelloviruses[64], having small genomes of only a few kilobases. Egoviruses with their large genomes are changing this view, being projected from the overall metagenomic signal to be abundant in the digestive system of ~180 million humans and billions of livestock and wild animals worldwide. These viruses were discovered by metagenome analysis, an approach that only targets the most abundant viral genomes and were detected by this method alone in more than 2% of the analyzed human gut microbiomes and in even higher fractions of many animal microbiomes. In particular, a substantial proportion (and possibly a majority) of the ~200 million buffalos contain one or multiple egovirus species thriving in their stomach compartments. The diversity of egoviruses was found to be particularly high among buffalos and is likely far from being fully characterized in most vertebrate species, including humans. In addition, longitudinal studies indicate that egoviruses can remain abundant over long periods of time in the gut of the same human individual. Thus, egoviruses are diverse, widespread, and resilient viral members of the gut microbiome.

Apart from their likely ecological importance in vertebrate digestive systems, egoviruses are clarifying evolutionary events that shaped the diversity and functioning of the class *Pokkesviricetes*. First, the evolutionary association between poxviruses and asfuviruses has been debated in part due to long-branching attraction artifacts often associated with poxviruses in phylogenies[23,65], so the discovery of a third order (*Egovirales*) evolutionarily related to both *Asfuvirales* and *Chitovirales* is cementing the monophyly of the class *Pokkesviricetes*. Second, egoviruses provide insights into the nature of the differences in the particle structure between the two previously identified orders of *Pokkesviricetes*: multilayered icosahedral capsids (*Asfuvirales*) versus brick-shaped capsids (*Chitovirales*). On the one hand, we found the DJR-MCP of egoviruses to be most closely related to that of poxviruses despite substantial protein length differences and the fact that in poxviruses, the DJR protein is incorporated into virion assembly intermediates but not the mature virions. These findings seem to favor an evolutionary scenario in which *Egovirales* and *Chitovirales* evolved from a common ancestor, with *Asfuvirales* diverging earlier in evolution. On the other hand, asfuviruses and egoviruses are the only *Nucleocytoviricota* clades known (and predicted) to harbor an inner MCP. These observations suggest that the last common ancestor of *Pokkesviricetes* had a multilayered icosahedral capsid that was replaced by brick-shaped particles in the ancestors of poxviruses following their split from egoviruses. Finally, the global distribution of egoviruses in gut samples further emphasizes the strong evolutionary association between viruses of the class *Pokkesviricetes* and vertebrates. More broadly, an evolutionary and ecological divide is now clearly emerging within the phylum

Nucleocytoviricota between the class *Megaviricetes* that is mainly associated with protist hosts inhabiting diverse environments and the now substantially expanded class *Pokkesviricetes* enriched in viruses associated with animals.

One critical question is the nature of the host cells infected by this viral clade. Although metagenomics alone does not offer conclusive answers, there are insights that merit consideration. For instance, our data show that egovirus clades are spread across continents and display a strong preference towards one or multiple vertebrate species, resembling poxviruses in that respect, but not asfuviruses[50]. One clade of closely related egoviruses is strictly associated with humans who lived in different parts of the world, which suggests a long-lasting association between *Egovirales* and humans. Given the remarkable differences in lifestyles among global human populations that are well reflected in the gut microbiome[66,67], the presence of egoviruses both in humans from industrialized and non-industrialized countries makes an exogeneous or transitionary member of the gut biome an unlikely host for egoviruses. Thus, it appears likely that egoviruses infect either animal cells or unicellular eukaryotes that broadly reside in the animal digestive systems. Eukaryotic signal in *Egovirales*-positive metagenomic assemblies suggests two potential unicellular eukaryotic taxa as hosts, namely, the evolutionarily constrained *Blastocystis* genus, and a more diverse group of ciliates. We found that in the guts of humans and pigs, *Blastocystis* was much more likely to be detected in *Egovirales*-positive samples compared to *Egovirales*-negative samples. However, buffalos not only displayed the opposite trend, but also provided a critical perspective, having the highest *Egovirales* signal among the surveyed animals but a very limited detectable co-occurrence of these viruses with unicellular eukaryotes. Thus, it remains unclear whether egoviruses infect unicellular eukaryotes. An alternative and perhaps more likely possibility is that egoviruses infect animal epithelial cells of the gastrointestinal tract that are shed constantly and abundantly from the gut barrier[68,69]. Regardless of the actual hosts of egoviruses, their global distribution highlights a previously overlooked role of vertebrate digestive systems in sustaining eukaryotic viral diversity and spreading viruses of the phylum *Nucleocytoviricota* across countries and continents.

Conclusion:

Egoviruses are DNA viruses with large double-stranded genomes that form a third order-level clade in the class *Pokkesviricetes* (phylum *Nucleocytoviricota*), along with asfuviruses (order *Asfuvirales*) and poxvirus*es* (order *Chitovirales*). We propose to designate the egovirus clade the order "*Egovirales*". In a sharp contrast to both asfuviruses and poxviruses, we found that egoviruses are highly prevalent in the guts of vertebrates and are abundant in more than 2% of the human population. We anticipate that the known diversity of egoviruses will soon far exceed that of all poxviruses and animal-associated asfuviruses combined, transforming our assessment of animal-associated eukaryotic DNA viruses in general and the diversity of the prominent class *Pokkesviricetes* in particular.

Methods:

Metagenomic database: Metagenomic sequencing datasets were processed as previously described[70,71] to extract single-copy marker genes from metagenome

assembled genomes to build the database for the taxonomic profiling tool mOTUs[72]. Briefly, BBMap[73] (v.38.71) was used to quality control sequencing reads from all samples by removing adapters from the reads, removing reads that mapped to quality control sequences (PhiX genome) and discarding low-quality reads (trimq=14, maq=20, maxns=1, and minlength=45). Quality-controlled reads were merged using bbmerge.sh with a minimum overlap of 16 bases, resulting in merged, unmerged paired, and single reads. The reads from metagenomic samples were assembled into contigs using the SPAdes assembler[74] (v3.15.2) in metagenomic mode. Contigs were length-filtered (≥ 500 bp) and gene sequences were predicted using Prodigal[75] (v2.6.3) with the parameters -c -q -m -p meta.

A first metagenomic survey using the RNApolA: A broad-spectrum HMM profile for the RNApolA hallmark gene[23] was searched against the predicted gene sequences from the metagenomic database using hmmsearch (HMMER[76] 3.3.1) with an e-value cutoff of $1e^{-50}$. This search resulted in the identification of nearly two million RNApolA genes. Positive hits were then linked back to their source contig.

Taxonomic affiliation of RNApolA sequences. We collected amino acid sequences corresponding to a broad range of reference RNApolA genes for bacterial, archaeal, eukaryotic, and giant virus lineages characterized by means of cultivation and environmental geonomics[24,77,78]. We generated a custom diamond database[79] (v.2.1.8) from this resource and summarized the taxonomic affiliation of each sequence in a complementary metadata file. We then performed a diamond blast search of RNApolA genes identified from the metagenomic survey using this custom database and exploited the associated metadata file to assess both their taxonomic affiliation and novelty score (based on the percent identity).

Phylogenetic signal pointing towards egoviruses. We used cdhit[80] and a 70% identity cut-off at the protein level to remove redundancy among the nearly two million RNApolA genes identified from the metagenomic survey. This step provided a database of 17,293 unique RNApolA proteins longer than 250 amino acids that encapsulate a wide range of cellular and viral diversity recovered from the metagenomic survey. Then, we used the phylogenetic workflow and visualization capabilities of anvi'o[81,82] (v.8) using muscle[83] and fast-tree[84] to identify and remove RNApolA proteins corresponding to bacteria and plastids. Focusing on the 10,853 remaining unique RNApolA proteins, we iteratively used the same anvi'o workflow but this time to identify and remove long branching artifacts. A total of 13 iterations were performed, after which 8,672 unique RNApolA proteins remained. Using this curated database, we explored the phylogenetic signal of RNApolA proteins using taxonomic annotations and the novelty score for guidance. This led to the identification of a monophyletic clade corresponding to the egoviruses, which is most closely related to asfuviruses and poxviruses. We incorporated these sequences in our custom diamond database and performed a second diamond blast on the nearly two million RNApolA proteins, in order to identify *Egovirales* RNApolA proteins (percent identity >50%) and study their source contigs.

A database for 57 complete and partial egovirus genomes. We analyzed all contigs containing an *Egovirales* RNApolA gene and generated a database of 57 non-redundant *Egovirales* genomes >50 kbp (average nucleotide identity <98%), which was processed with anvi'o[81,82] (v.8.1) to predict gene sequences and functionally annotate hallmark

genes using specific HMMs. Among these genomes, 13 were found to be complete based on the occurrence of 55 bp inverted terminal repeats.

Functional annotation of *Egovirales* core genes. We collected reference culture genomes and metagenome-assembled genomes corresponding to *Asfuvirales* and *Chitovirales*, as well as few *Megaviricetes* representatives from the GVDB collection[1]. We added the 57 egoviruses genomes into this "*Nucleocytoviricota* genomic resource" and used Orthofinder[85] (V.2.5.5 in "diamond_ultra_sens" mode) to characterize gene clusters. Gene clusters occurring in at least 10 out of the 13 *Egovirales* complete genomes were labelled as "*Egovirales* core genes", and those occurring as a single copy in at least 10 out of the same 13 genomes were labelled as "*Egovirales* single copy core genes". Finally, HHpred[44] profile-profile comparisons and ColabFold[45] (3D structure predictions and comparisons) were used to functionally annotate the *Egovirales* core genes. In the case of the DJR-MCP of *Egovirales*, ColabFold[45] successfully predicted its 3D structure and confirmed the presence of the double-jelly roll (DJR) protein components characteristic of the phylum *Nucleocytoviricota*.

A second metagenomic survey using the DJR-MCP: We collected DJR-MCPs from the 57 non-redundant egovirus genomes to generate with HMMER[76] (v3.1b2) an HMM dedicated to the *Egovirales* DJR-MCP. This HMM was then searched against all predicted gene sequences from the metagenomic database, using hmmsearch (HMMER[76] v.3.3.1) and an e-value cutoff of $1e^{-10}$. Positive hits were then linked back to their source contig, expanding the scope of *Egovirales*-positive metagenomic samples.

Recovery of eight *Nucleocytoviricota* hallmark genes. We used HMMER[73] v3.1b2 to search and collect the eight hallmark genes of *Nucleocytoviricota* from the *Nucleocytoviricota* genomic resource (see "*Functional annotation of Egovirales core genes*" section). Given that genomes of faustoviruses from the GVDB collection[1] had partial DJR-MCP genes, we replaced those with 7 strains of faustoviruses from the NCBI GenBank collection (strains E9, LCD7, M6, S17, liban, vv10, and vv63; https://www.ncbi.nlm.nih.gov/genbank/ accessed February 2023). We also added the Variola virus isolate VARV V563 genome from the GOEV database[24], we replaced the *Megaviricetes* representatives by others in the GOEV database[24] based on their distribution as well as the occurrence of hallmark genes. In addition to the RNApolA and DJR-MCP protein sequences already recovered, we obtained sequences for the family-B DNA polymerase (DNApolB), the second largest DNA-dependent RNA polymerase subunits (RNApolB), the transcription factor IIS (TFIIS), the D5-like primase (Primase), the poxvirus late transcription factor (VLTF3) and the packaging ATPase (pATPase). The DNApolB gene, often present on two separate open reading frames in egoviruses, required manual curation through BLASTp alignments (BLAST[83] v2.10.1) and phylogenetic reconstructions, as previously described[24]. Briefly, when a single egovirus genome had two sequences for this gene, their positions were inspected in a maximum-likelihood single-protein phylogenetic tree, and their sequences were aligned with BLASTp to a close reference sequence and to each other. The two DNApolB sequences for a same egovirus systematically corresponded to a clear split: they were fused and accordingly labelled for further notice. Finally, we created a database of genomes containing at least seven out of the eight hallmark genes, for downstream phylogenetic analyses.

Alignments, trimming and individual gene phylogenetic analyses. From the genomic database created in the previous section, protein sequences corresponding to the same hallmark gene were aligned with MAFFT[84] v.7.464 and default parameters. The L-INS-i algorithm was used. Sites with more than 70% gaps were systematically trimmed using Goalign v0.3.5 (https://www.github.com/evolbioinfo/goalign). The maximum-likelihood phylogenetic reconstructions were performed using IQ-TREE[85] v1.6.2, with the ModelFinder[86] option to determine the best-fitting model. Supports were computed from 1,000 replicates for the Shimodaira-Hasegawa (SH)-like aLRT[87] and UFBoot[88]. Supports were deemed good when SH-like aLRT ≥80% and UFBoot ≥95%, as per IQ-TREE manual, and moderate when only one of the two conditions was met. The few cases of hallmark genes occurring more than once in each genome were investigated and resolved after removing recent paralogs and clear outliers. Aside from this analysis focused on a subset of *Nucleocytoviricota* genomes containing most hallmark genes, we also performed a global phylogeny of the RNApolA and DJR-MCP of egoviruses, using few Chitovirales and Asfuvirales representatives for rooting. All phylogenetic trees (Figure S1 for the eight individual tree phylogenies; Figure 3 and Figure S3 for the global phylogenies of the RNApolA and DJR-MCP of egoviruses) were visualized and rooted using anvi'o. Supports were extracted with iTOL (doi: 10.1093/nar/gkab301).

A phylogeny of *Nucleocytoviricota* using concatenated gene alignments. In addition to the single-protein phylogenies focused on the class *Pokkesviricetes*, we also performed a multi-protein phylogenetic analysis of the concatenated DNApolB, RNApolA, RNApolB and TFIIS protein sequences. We replaced the few *Megaviricetes* representatives used as outgroup by a much larger genomic database for the non-*Pokkesviricetes* nucleocytoviruses. We used gene markers for the DNApolB, RNApolA, RNApolB and TFIIS corresponding to the classes *Pokkesviricetes*, *Megaviricetes* (exclusion of *Medusavirus* to minimize long-branching artifacts) from the GOEV database[24], which were all already carefully manually curated. We added to this resource protein sequences for the DNApolB, RNApolA, RNApolB and TFIIS corresponding to all 57 egovirus genomes. This phylogenetic analysis followed the protocol described in the previous section, except that the FFT-NS-i algorithm of MAFFT v7.464 was used for the alignment of each individual protein dataset, before they were trimmed and concatenated. The phylogenetic tree (Figure 1) was visualized and rooted using anvi'o.

Data availability: Databases our study used or generated include (1) the RNApolA proteins identified from the metagenomic survey, (2) *Egovirales* RNApolA and DJR-MCPs proteins, (3) reference *Nucleocytoviricota* HMMs as well as two new HMMs targeting the DJR-MCP and VLTF3 of *Egovirales*, (4) the 57 Egovirales genomes >50 kbp in length (including 13 complete genomes), (5) and the 3D structure of the *Egovirales* DJR-MCP. These items along with the supplementary tables (Tables S01-S07) have been made publicly available at https://doi.org/10.6084/m9.figshare.25398895.v5.

Acknowledgments: Our survey was made possible thanks to the large amount of metagenomic data that has been produced and made publicly available in recent years. As a result, we are grateful to all the scientists involved in the metagenomic exploration of a wide range of environmental ecosystems as well as the human and animal gut microbiomes. We are also grateful to Lucas Paoli for helping facilitate the acquisition of results from the initial RNApolA survey as part parallel metagenomic investigations. This

study was supported by two grants from the French national research agency (ANR-23-CE02-0022 and ANR-23-CE02-0025).

Conflict of interest: Authors declare having no conflicts of interest.

Additional information: Correspondence should be addressed to Tom O. Delmont (tomodelmont@gmail.com).

Supplemental Tables

Table S1: Metadata associated with the metagenomic survey.
Table S2: Metagenomic signal for *Chitovirales* and *Asfuvirales* based on the RNApolA survey.
Table S3: Main statistics for egoviruses. The table includes metagenomic signal for *Egovirales* based on the RNApolA and MCP surveys, as well as main statistics regarding Egovirales-positive samples, and egovirus genomics.
Table S4: Gene clusters, core genes and functional annotations for the 13 complete *Egovirales* genomes.
Table S5: Occurrence and positivity rate of *Egovirales*, Ciliates, and *Blastocystis* among humans, buffalos, chicken, pigs, and cows.
Table S6: Metagenomic signal for *Egovirales* in buffalo samples. The survey covers seven Chinese provinces, three buffalo breeds and different segments of their digestive system.
Table S7: Metagenomic signal for eukaryotes among the *Egovirales*-positive samples based on the RNApolA survey.

Supplemental Figures

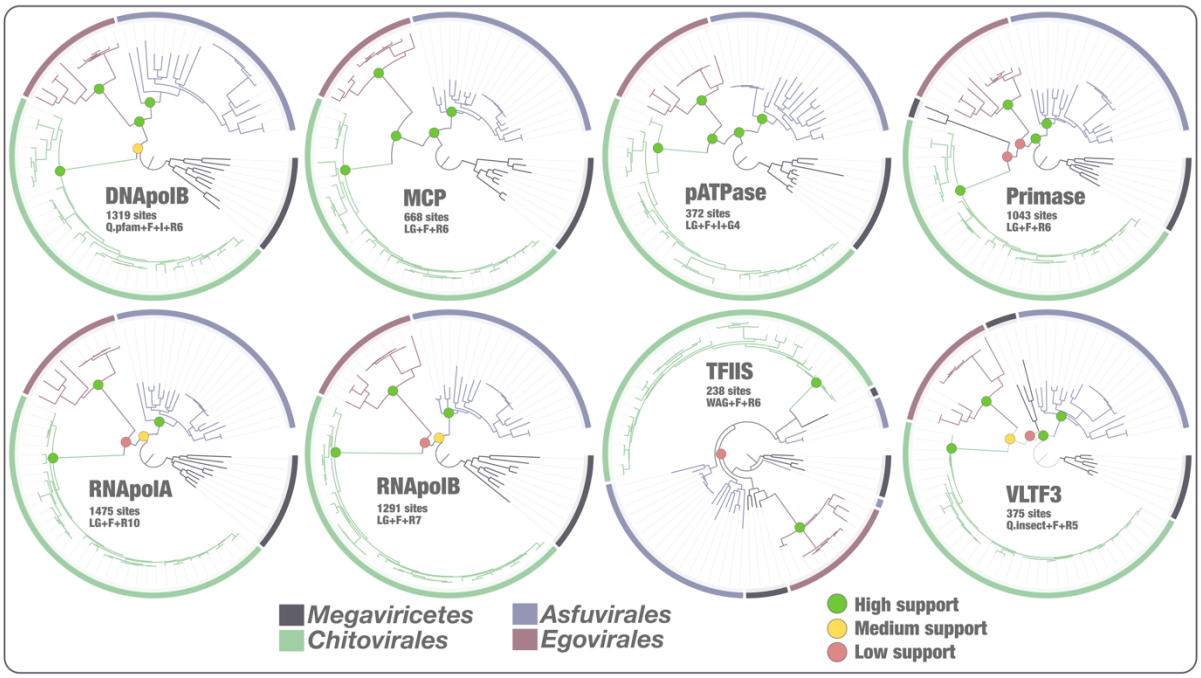


Figure S1: Phylogeny of *Pokkesviricetes* using seven individual hallmark genes. Each tree corresponds to a phylogeny of the same genomic resource (11 *Egovirales*, 36 *Chitovirales*, 21 *Asfuvirales*, and 9 *Megaviricetes*) using one of the seven hallmark genes shared by the tree orders of the class *Pokkesviricetes*. Each genome in this restrained genomic resource contains at least seven out of the eight hallmark genes.

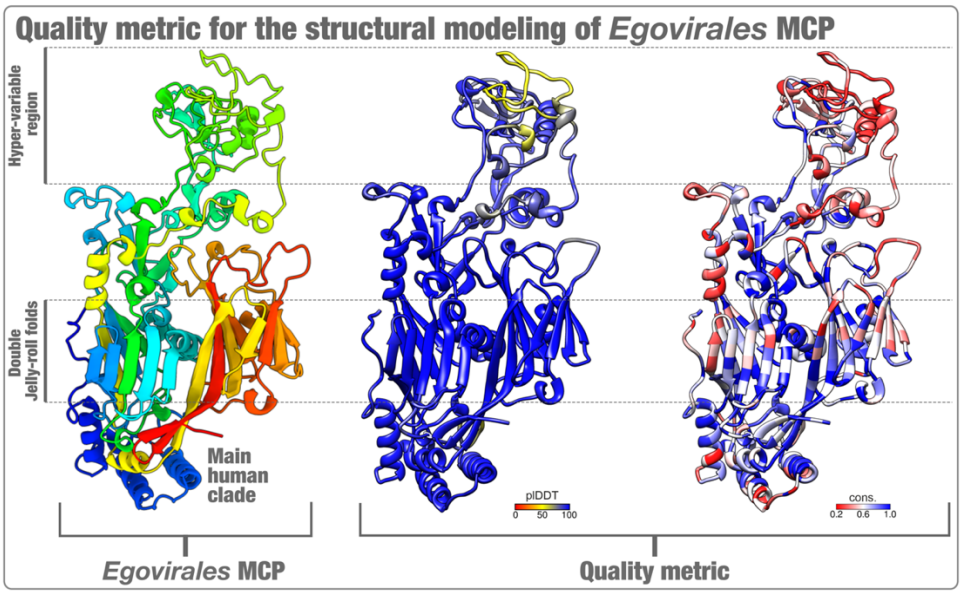


Figure S2: Quality metric of the structural modeling of the Egovirales major capsid protein. Left panel displays the structural modeling of an *Egovirales* DJR MCP (human main clade), as predicted from Colabfold[45]. Right panels display the pIDDT and conservation levels (quality metric).

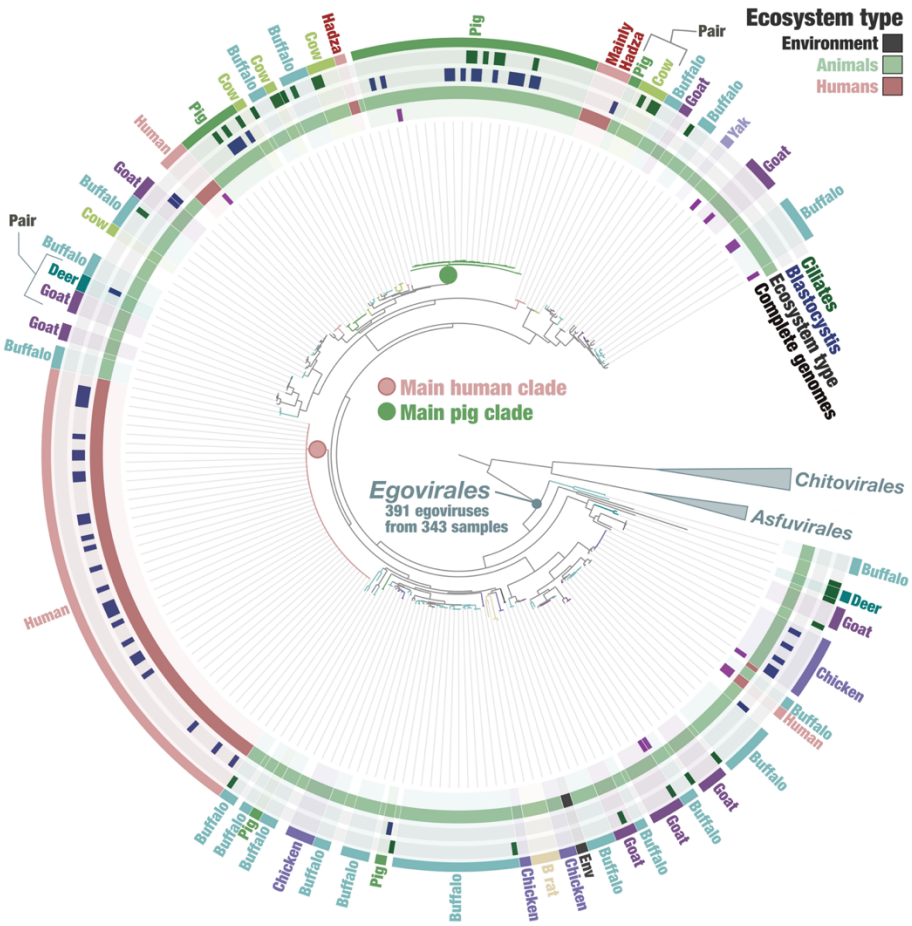


Figure S3: Overlay of the evolutionary tree of egoviruses and their occurrence among vertebrate species. The figure displays a maximum-likelihood phylogeny of the MCP of egoviruses (>600 amino acids; n=391), using representatives of *Chitovirales* and *Asfuvirales* for rooting (784 sites), and the LG+F+R10 model. Layers of information summarize the occurrence of the 13 complete egovirus genomes, ecosystem type (environment vs. human vs. other animals), and co-occurrence of *Blastocystis* and ciliates. Finally, selections were made for egovirus clades occurring in the same animal species.

References:

1. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* **19**, e3001430 (2021).
2. McFadden, G. Poxvirus tropism. *Nature Reviews Microbiology 2005 3:3* **3**, 201–213 (2005).
3. Smith, S. A. & Kotwal, G. J. Immune response to poxvirus infections in various animals. *Crit Rev Microbiol* **28**, 149–185 (2002).
4. Lu, B. *et al.* Outbreak of Vaccinia Virus Infection from Occupational Exposure, China, 2017. *Emerg Infect Dis* **25**, 1192 (2019).
5. Wilson, M. E., Hughes, J. M., McCollum, A. M. & Damon, I. K. Human Monkeypox. *Clinical Infectious Diseases* **58**, 260–267 (2014).
6. Chen, X., Anstey, A. V. & Bugert, J. J. Molluscum contagiosum virus infection. *Lancet Infect Dis* **13**, 877–888 (2013).
7. Mühlemann, B. *et al.* Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science (1979)* **369**, (2020).
8. Hannat, S., La Scola, B., Andreani, J. & Aherfi, S. Asfarviruses and Closely Related Giant Viruses. *Viruses 2023, Vol. 15, Page 1015* **15**, 1015 (2023).
9. Dixon, L. K., Sun, H. & Roberts, H. African swine fever. *Antiviral Res* **165**, 34–41 (2019).
10. Reteno, D. G. *et al.* Faustovirus, an Asfarvirus-Related New Lineage of Giant Viruses Infecting Amoebae. *J Virol* **89**, 6585 (2015).
11. Hannat, S., La Scola, B., Andreani, J. & Aherfi, S. Asfarviruses and Closely Related Giant Viruses. *Viruses 2023, Vol. 15, Page 1015* **15**, 1015 (2023).
12. Karki, S., Moniruzzaman, M. & Aylward, F. O. Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae. *Front Microbiol* **12**, (2021).
13. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* (2020) doi:10.1038/s41586-020-1957-x.
14. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications 2020 11:1* **11**, 1–11 (2020).
15. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, (2020).
16. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).
17. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology 2021 19:8* **19**, 514–527 (2021).
18. Colson, P., Aherfi, S., La Scola, B. & Raoult, D. The role of giant viruses of amoebas in humans. *Curr Opin Microbiol* **31**, 199–208 (2016).
19. Laforest-Lapointe, I. & Arrieta, M.-C. Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems* **3**, (2018).
20. Scanlan, P. D. & Marchesi, J. R. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME J* **2**, 1183–1193 (2008).
21. Parfrey, L. W., Walters, W. A. & Knight, R. Microbial Eukaryotes in the Human Microbiome: Ecology, Evolution, and Future Directions. *Front Microbiol* **2**, (2011).

22. Vargas-Albores, F., Garibay-Valdez, E., Medina-Félix, D. & Martínez-Porchas, M. The micro-eukaryotic community: An underrated component of the mammalian gut microbiota? *Front Microbiol* **14**, (2023).
23. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A* **116**, 19585–19592 (2019).
24. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature 2023 616:7958* **616**, 783–789 (2023).
25. Carter, M. M. *et al.* Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* **186**, 3111-3124.e13 (2023).
26. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature 2023 614:7946* **614**, 125–135 (2023).
27. Sayols-Baixeras, S. *et al.* Streptococcus Species Abundance in the Gut Is Linked to Subclinical Coronary Atherosclerosis in 8973 Participants From the SCAPIS Cohort. *Circulation* **148**, 459–472 (2023).
28. Hildebrand, F. *et al.* Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host Microbe* **29**, 1167-1176.e9 (2021).
29. Fukuyama, J. *et al.* Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput Biol* **13**, e1005706 (2017).
30. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature 2016 535:7612* **535**, 435–439 (2016).
31. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nature Medicine 2021 27:2* **27**, 321–332 (2021).
32. Levin, D. *et al.* Diversity and functional landscapes in the microbiota of animals in the wild. *Science (1979)* **372**, (2021).
33. Tong, F. *et al.* The microbiome of the buffalo digestive tract. *Nature Communications 2022 13:1* **13**, 1–16 (2022).
34. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology 2019 37:8* **37**, 953–961 (2019).
35. Guo, W. *et al.* Metagenomic study suggests that the gut microbiota of the giant panda (Ailuropoda melanoleuca) may not be specialized for fiber fermentation. *Front Microbiol* **9**, 324779 (2018).
36. Xie, F. *et al.* An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* **9**, 1–20 (2021).
37. Cao, Y. *et al.* The multi-kingdom microbiome of the goat gastrointestinal tract. *Microbiome* **11**, 1–20 (2023).
38. Gaio, D. *et al.* A large-scale metagenomic survey dataset of the post-weaning piglet gut lumen. *Gigascience* **10**, 1–8 (2021).
39. Andersen, V. D. *et al.* Predicting effects of changed antimicrobial usage on the abundance of antimicrobial resistance genes in finisher' gut microbiomes. *Prev Vet Med* **174**, 104853 (2020).
40. Xiao, L. *et al.* A reference gene catalogue of the pig gut microbiome. *Nature Microbiology 2016 1:12* **1**, 1–6 (2016).
41. Krupovič, M. & Bamford, D. H. Virus evolution: how far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol* **6**, 941–948 (2008).
42. Li, H., Liu, Q., Shao, L. & Xiang, Y. Structural Insights into the Assembly of the African Swine Fever Virus Inner Capsid. *J Virol* **97**, (2023).

43. Fu, D. *et al.* Structure of African swine fever virus p15 reveals its dual role for membrane-association and DNA binding. *Protein Cell* **11**, 606–612 (2020).
44. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **430**, 2237–2243 (2018).
45. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
46. Condit, R. C., Moussatche, N. & Traktman, P. In A Nutshell: Structure and Assembly of the Vaccinia Virion. *Adv Virus Res* **66**, 31–124 (2006).
47. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219 (2016).
48. Zhang, H. *et al.* Metagenome sequencing and 768 microbial genomes from cold seep in South China Sea. *Scientific Data 2022 9:1* **9**, 1–7 (2022).
49. Mu, D., Zhao, J., Wang, Z., Chen, G. & Du, Z. Draft Genome Sequence of Algoriphagus sp. Strain NH1, a Multidrug-Resistant Bacterium Isolated from Coastal Sediments of the Northern Yellow Sea in China. *Genome Announc* **4**, 1555–1570 (2016).
50. Haller, S. L., Peng, C., McFadden, G. & Rothenburg, S. Poxviruses and the Evolution of Host Range and Virulence. *Infect Genet Evol* **0**, 15 (2014).
51. Chen, Z. *et al.* Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog* **14**, e1007352 (2018).
52. Katzourakis, A. *et al.* Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. *Retrovirology* **11**, (2014).
53. Wawrzyniak, I. *et al.* Blastocystis, an unrecognized parasite: an overview of pathogenesis and diagnosis. *Ther Adv Infect Dis* **1**, 167 (2013).
54. Scanlan, P. D. *et al.* The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol Ecol* **90**, 326–330 (2014).
55. Alfellani, M. A. *et al.* Variable geographic distribution of Blastocystis subtypes and its potential implications. *Acta Trop* **126**, 11–18 (2013).
56. Deng, L. *et al.* First identification and molecular subtyping of Blastocystis sp. in zoo animals in southwestern China. *Parasit Vectors* **14**, 1–11 (2021).
57. Boreham, P. F. L. & Stenzel, D. J. Blastocystis in Humans and Animals: Morphology, Biology, and Epizootiology. *Adv Parasitol* **32**, 1–70 (1993).
58. Lin, C., Raskin, L. & Stahl, D. A. Microbial community structure in gastrointestinal tracts of domestic animals: comparative analyses using rRNA-targeted oligonucleotide probes. *FEMS Microbiol Ecol* **22**, 281–294 (1997).
59. Andersen, T. O. *et al.* Metabolic influence of core ciliates within the rumen microbiome. *ISME J* **17**, 1128 (2023).
60. Cedrola, F. *et al.* Trichostomatid Ciliates (Alveolata, Ciliophora, Trichostomatia) Systematics and Diversity: Past, Present, and Future. *Front Microbiol* **10**, 471263 (2020).
61. Selim, H. M. *et al.* Comparative study of rumen ciliates in buffalo, cattle and sheep in Egypt. *J Vet Med Sci* **58**, 799–801 (1996).
62. Pomajbíková, K. *et al.* Novel Insights into the Genetic Diversity of Balantidium and Balantidium-like Cyst-forming Ciliates. *PLoS Negl Trop Dis* **7**, (2013).
63. Nakauchi, K. The prevalence of Balantidium coli infection in fifty-six mammalian species. *J Vet Med Sci* **61**, 63–65 (1999).
64. Butkovic, A. *et al.* Evolution of anelloviruses from a circovirus-like ancestor through gradual augmentation of the jelly-roll capsid protein. *Virus Evol* **9**, (2023).

65. Woo, A. C., Gaia, M., Guglielmini, J., Da Cunha, V. & Forterre, P. Phylogeny of the Varidnaviria Morphogenesis Module: Congruence and Incongruence With the Tree of Life and Viral Taxonomy. *Front Microbiol* **12**, 1708 (2021).
66. Segata, N. Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. *Current Biology* **25**, R611–R613 (2015).
67. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nature Communications 2014 5:1* **5**, 1–12 (2014).
68. Sender, R. & Milo, R. The distribution of cellular turnover in the human body. *Nature Medicine 2021 27:1* **27**, 45–48 (2021).
69. Williams, J. M. *et al.* Epithelial Cell Shedding and Barrier Function: A Matter of Life and Death at the Small Intestinal Villus Tip. *Vet Pathol* **52**, 445 (2015).
70. Ruscheweyh, H. J. *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* **10**, 1–12 (2022).
71. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature 2022* 1–8 (2022) doi:10.1038/s41586-022-04862-3.
72. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**, 1196–1199 (2013).
73. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
74. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res* **27**, 824–834 (2017).
75. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
76. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–W37 (2011).
77. Delmont, T. O. *et al.* Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *The ISME Journal 2021* 1–10 (2021) doi:10.1038/s41396-021-01135-1.
78. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* 100123 (2022) doi:10.1016/J.XGEN.2022.100123.
79. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
80. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
81. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
82. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology 2020 6:1* **6**, 3–6 (2020).
83. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
85. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, (2015).
86. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* vol. 25 3389–3402 Preprint at https://doi.org/10.1093/nar/25.17.3389 (1997).

87. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
88. 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).
89. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods 2017 14:6* **14**, 587–589 (2017).
90. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321 (2010).
91. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).