Closest relatives of poxviruses are spread in the gut of humans and animals worldwide: the egoviruses

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, National Institutes of Health, 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[1–4]. Until now, these viruses were only sporadically found in the digestive system of vertebrates[5–7]. Here, we present the discovery of a diversified group of *Nucleocytoviricota* viruses dubbed egoviruses that almost exclusively occur in the digestive system of vertebrates worldwide. Egoviruses are most closely related to poxviruses and represent a third order within the class *Pokkesviricetes*[8]. They contain large linear genomes that include genes linked to multilayered icosahedral capsids only observed in asfuviruses. The widespread occurrence of egovirus genes in genomes of metamonads (with signal particularly enriched among *Trichomonas* and *Tritrichomonas*), and their co-existence in human fecal samples as demonstrated by our metagenomic survey point towards a preferential infection of unicellular eukaryotes known to prevail, often as symbionts or pathogens, in the gut of vertebrates worldwide[9]. Notably, the numerous *Egovirales* genes found throughout *Trichomonas vaginalis* genomes (>3% of their gene pools) designate this prominent sexually transmitted human pathogen as a likely vector to the spread of egoviruses. Notably, one egovirus clade is human-specific, evolutionarily constrained, and spread across continents, demonstrating a long-term association with the human population at a global scale. Egoviruses represent the only diverse, widespread, and abundant group of double-stranded DNA viruses infecting eukaryotes in the digestive system of vertebrates, with implications for human health, capsid evolution and the origin of poxviruses.

Key words: Giant viruses; gut microbiome; *Poxviruses*; *Asfuvirales*; eukaryotic viruses; human virus; *Metamonada*; 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*[8]. Poxviruses infect a broad diversity of animals including humans[10,11]. 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[12–15]. Asfuviruses infect hosts ranging from unicellular eukaryotes to animals[16] and include the African swine fever virus, which causes deadly outbreaks in pigs and boars[17], as well as a wider range of viruses mostly characterized in co-culture with free-living amoeba[18,19] or directly from the environment (e.g., within plankton) using genome-resolved metagenomics[7,20,21]. 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[22,23].

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)[5–7]. 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[5–7] 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[24–27]. 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 would be 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, these viruses are abundant in more than 2% of the human population worldwide and many of their genes occur in the genomes of co-existing parasitic metamonads (including the prominent *Trichomonas vaginalis*), pointing towards a lasting history of unicellular eukaryotic infections by egoviruses in the gut. 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*[8,28]. We used a comprehensive hidden Markov model (HMM) for the largest RNA polymerase

subunit A (RNApolA)[28,29] to search 85,123 metagenomic assemblies covering a wide range of environmental samples as well as thousands of samples from the digestive system of humans (e.g., [30–36]), livestock and wild animals (e.g., [37–45]) (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[37] 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[7,21], to the rapid expansion of the known diversity of asfuviruses in the environment.

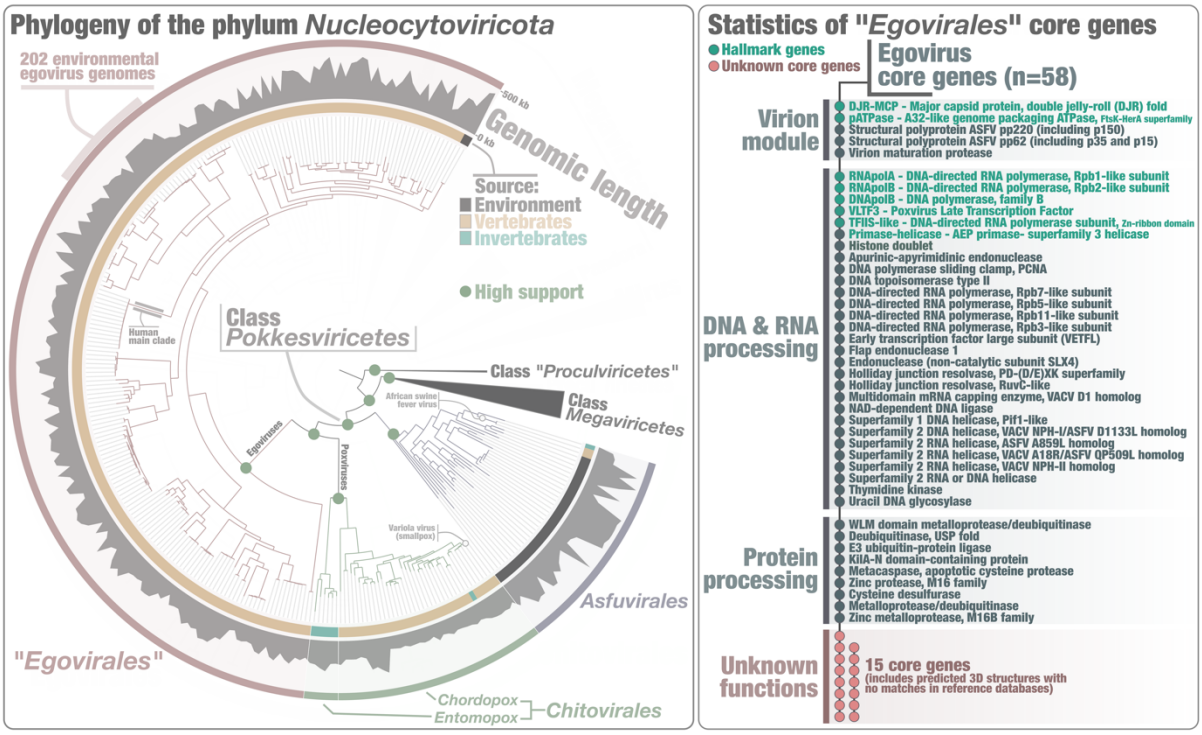


Figure 1: Evolution and predicted gene functions of egoviruses. Left panel displays a maximum-likelihood phylogenomic analysis of members of *Nucleocytoviricota*, along with 202 *Egovirales* genomes with a quality score above 50%. The tree covers 1,807 genomes and was built using concatenated multiple alignments of four hallmark proteins of *Nucleocyovoricota* (DNApolB, RNApolA, RNApolB, TFIIS; 4,110 sites) using IQTree with the LG+F+R10 model. The tree was rooted with the putative class *Proculviricetes* and visualized using anvi'o v8. Green dots at nodes in the tree highlight high supports (approximate likelihood ratio (aLRT) ≥ 80 and ultrafast bootstrap approximation (UFBoot) ≥ 95; see Methodds). Right panel summarizes the functional annotation of 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 and which we provisionally named egoviruses. We characterized 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*[46–48]. The egovirus genomes encompass all eight *Nucleocytoviricota* hallmark genes[8,28] (Table S3), with the DNA polymerase B coding region often split into two separate open reading frames. We built an HMM for the

Egovirales double jelly-roll major capsid protein (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, mainly from gut microbiomes of a wide diversity of vertebrates (Table S3). Using these two marker genes as guidance, we recovered a total of 224 metagenome-assembled genomes (MAGs) when including the 13 complete genomes. *Egovirales* MAGs display an average completion of 81% (minimal quality score of 50%; see Methods) and are up to 467 kbp in length (average of 229 kbp). Phylogenetic analysis of manually curated hallmark genes confidently identified the egovirus MAGs as a distinct major clade within the class *Pokkesviricetes* that is most closely related to poxviruses (Figures 1 and S1). Based on these findings, egoviruses should become a third order within *Pokkesviricetes* that we propose to name "*Egovirales*". *Egovirales* is the only clade within 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*[8], and explored clusters of homologous genes based on amino acid sequence similarity 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 S2 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, endonucleases, 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 in particular are represented in most members of the orders *Asfuvirales*, *Chitovirales* and the class *Megaviricetes*[7,21,29] (Table S4). Apart from the genes that could be functionally annotated, 18 core genes could not be assigned any function through sequence comparisons. However, structural modeling with ColabFold[49] followed by DALI searches against the protein structure database (PDB) allowed functional predictions for three additional core proteins. Notably, one of these was found to represent a histone doublet, likely implicated in genome condensation into nucleosome-like structures. Histones, including histone doublets, have been previously described in some members of the *Nucleocytoviricota,* in particular, marseilleviruses[50], but not in members of *Pokkesviricetes*. The remaining core genes are either exclusive to *Egovirales* or extremely 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 the major components of the virion morphogenetic

module conserved across *Nucleocytoviricota*, including poxviruses and asfuviruses (Figure 1). The core components of this module are the double jelly-roll (DJR) major capsid protein[51] (MCP; Figure 2), the A32-like genome packaging ATPase, and Ulp1-like capsid maturation protease. The 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*[28,29], systematically failed to recognize this protein in egoviruses. Eventually, HHpred[52] profile-profile comparisons yielded a partial match to the DJR-MCP of African swine fever virus, and structural modeling with ColabFold[49] 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 the exterior of the capsid is highly diversified among egoviruses (Figure S3). These trends echo the previous observations on the large MCPs of faustoviruses from the order *Asfuvirales*[18]. However, *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). In addition, egoviruses encode homologs of the major virion proteins exclusive to asfuviruses, namely, structural polyproteins pp220 and pp62. In African swine fever virus, both polyproteins are processed into multiple structural proteins, which play important roles in the formation of multilayered capsids. One of the major proteolytic products of pp220, p150, that is conserved in egoviruses, has an α-helical fold and forms an internal icosahedral shell located inside of a larger, external icosahedral capsid formed from the DJR-MCP[53]. So far, the extent of protein sequence diversification among egoviruses prevented the prediction of the high-confidence 3D structure for the p150 homolog. Structure-based searches allowed identification of the homologs of asfuviral core shell proteins p15 and p35, two of the three proteolytic products of pp62 polyprotein. The p15 and p35 bridge the viral nucleoprotein complex and the internal capsid to the lipid membrane sandwiched between the inner and outer icosahedral shells[54,55]. Collectively, the presence of these three structural proteins suggests that egoviruses form icosahedral multilayered capsids that are similar to those of *Asfuvirales*[53], but differ from the brick-shaped capsids of poxviruses[56].

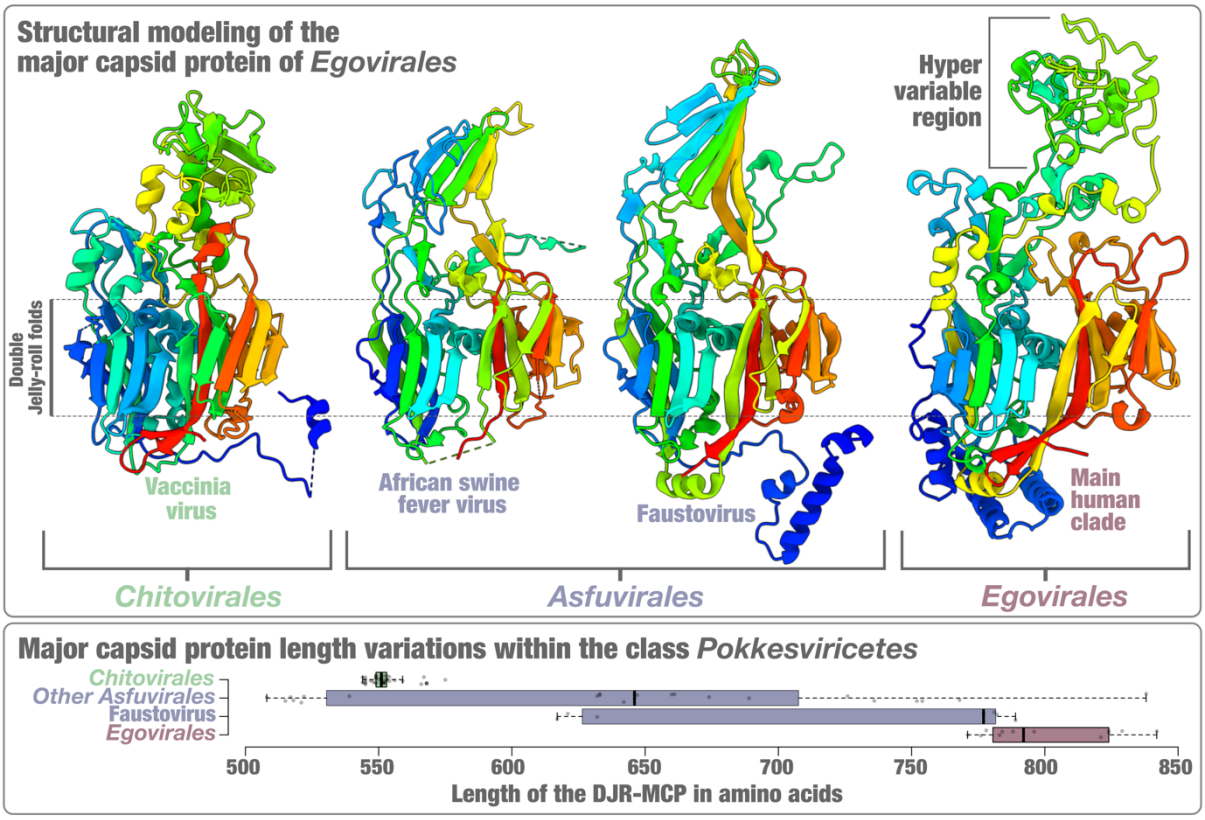


Figure 2: The *Egovirales* major capsid protein. Top panel displays structural modeling of an *Egovirales* DJR-MCP (human main clade) and comparison to the structures of DJR-MCPs from three reference *Pokkesviricetes* DJR-MCPs (from left to right: 3sam, 6ku9, 5j7o). Bottom panel displays boxplots summarizing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

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

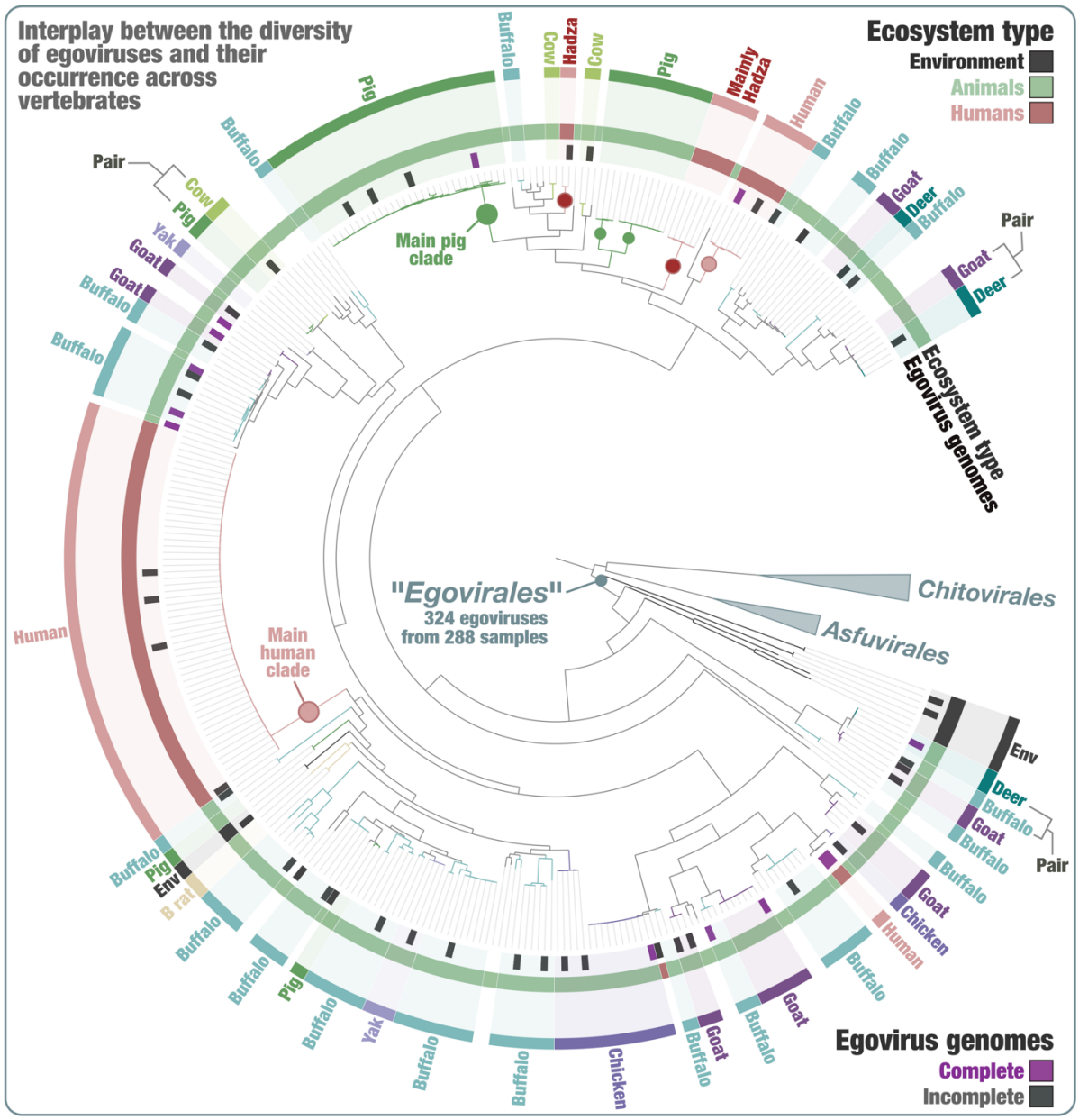
We used all the RNApolA and/or DJR-MCP genes of *Egovirales* found among the 864 metagenomic assemblies to perform a global survey of the occurrence and diversity of egoviruses (Table S3). With the RNApolA phylogeny, only 9 egovirus-like RNApolA genes were retrieved from environmental samples, including an aquifer[57], a cold sweep[58], and marine sediments[59] (Figure 3). These formed four small clades in the tree, two of which (from marine sediments) being 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 S4).

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%) (Figure 4). The extensive sequencing efforts in the study of Hadza populations[30] could in part explain this enhanced signal. Nevertheless, some of the human-associated clades were either exclusively or mostly detected among the Hadza (Figure 3 and Figure S4), suggesting that lifestyle could influence the diversity of human-associated egoviruses. Egoviruses were also well represented in the gut 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 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[38], allowing for a more refined survey of the occurrence of egoviruses within the gut microbiome of an 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 was 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.

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 S4). 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 S4). 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[60], papillomaviruses[61], and foamy viruses[62].



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%

Figure 3: Overlay of the evolutionary tree of egoviruses and their occurrence among vertebrate species. The 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.

High rates of gene flux between egoviruses and co-existing metamonads

We first assessed whether the detection of egoviruses in the gut samples coincides with the presence of any group of unicellular eukaryotes, which could serve as egovirus hosts. We identified eukaryotic RNApolA genes in 225 out of the 864 *Egovirales*-positive samples. However, the two clades accounting for nearly 95% of this signal (*Blastocystis* and a larger group of ciliates) poorly correlated with the relative abundance of egoviruses

(Tables S5, S7and S8, Figure S5), and no Egovirales signal was observed in *Blastocystis* genomes. Considering the frequent transfers of genes between *Nucleocytoviricota* viruses and their hosts[7], we searched for possible gene fluxes between the viral order and known eukaryotic clades that would be indicative of past infection events. We compared our database of *Nucleocytoviricota* protein sequences to the TrEMBL protein database, which encompasses a wide range of eukaryotic lineages (see Methods). We found a distinct and strong signal connecting egoviruses and the high-ranking clade of *Metamonada*, which is entirely constituted of anaerobic amitochondriate unicellular eukaryotes and includes many vertebrate symbionts and parasites[63]. After integrating additional publicly available *Metamonada* genomes, this clade alone represented 74.3% of all best hits for *Egovirales* within the scope of eukaryotes, followed by *Amoebozoa* (11%) and *Opisthokonta* (9%) (Table S9). In sharp contrast, *Metamonada* only represented between 2% and 10% of eukaryotic best hits in the other *Nucleocytoviricota* orders.

In total, nearly 3% of egovirus genes had a best hit for *Metamonada*, far more than any other known *Nucleocytoviricota* orders (Figure 4; Table S9). Furthermore, the highest signal connecting egoviruses and *Metamonada* corresponded to two genera of the class *Parabasalia* and known to infect vertebrates at a global scale: *Trichomonas* and *Tritrichomonas* (Figure S6). Indeed, *Trichomonas vaginalis*, *Trichomonas tenax* and *Tritrichomonas musculus* had the highest representations of *Metamonada* hits, with the former being noticeably more detected in human-associated egoviruses (nearly 30% of all *Metamonada* hits) (Table S9). Finally, not only are many egovirus genes best matching to *Metamonada*, but the inverse is also true. For instance, *Trichomonas vaginalis* (strain G3) and *Tritrichomonas musculus* (strain EAF2021) contain 3.1% and 1.5% of their predicted gene sequences with a best match towards *Egovirales*, respectively (Table S10). Instead of representing distinct endogenous viral elements (as observed in various hosts of *Nucleocytoviricota*[4]), these *Egovirales*-related genes are spread within the *Metamonada* genomes with no apparent syntenic signal (Table S10). Despite a much lower co-existance signal compared to *Blastocystis* and ciliates, 12 *Egovirales*-positive samples (11 human samples and one from chicken) contain RNApolA genes affiliated with *Metamonada*, all of which are most closely related to *Parabasalia* (average amino acid sequence identify of 69% against either *Histomonas meleagridis* or *Tritrichomonas foetus*; Table S7). While comparative genomics exposed a high degree of gene flux between the newly described viral order and *Parabasalia* species, both our RNApolA metagenomic survey and the literature indicate that they co-exist inside vertebrates at a global scale.

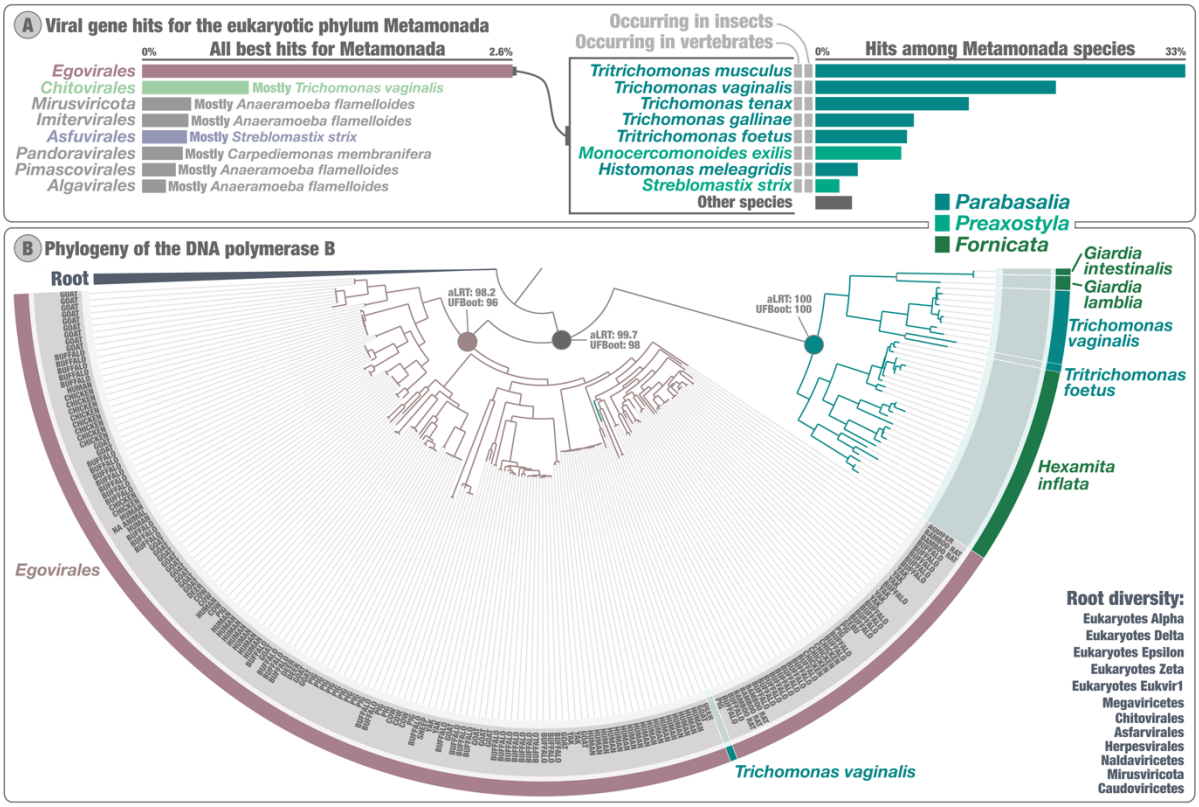


Figure 4: Gene flux between egoviruses and *Metamonada*. Panel A displays the percentage of viral genes with a best sequence similarity hit for *Metamonada* (see method). The score is summarized for each *Nucleocytoviricota* order and for the phylum *Mirusviricota*. The proportion of hits across *Metamonada* species is also described for the egovirus hits. Panel B is a maximum-likelihood phylogeny of the DNA polymerase B focusing on *Egovirales* and related sequences (1,950 taxa; 1,032 sites; LG+F+R10).

Many *Egovirales* genes with a best hit for *Metamonada* correspond to core genes conserved in a majority of egoviruses. This is notably the case for the DNApolB, mRNA capping enzyme, holliday junction resolvase, primase-Helicase, and thymidine kinase (Table S9). For these genes, phylogenetic trees including additional eukaryotic sequences display one or more sequences from a single *Metamonada* branching within the diversity of egoviruses with strong supports, more distant to other *Metamonada* sequences. For instance, the thymidine-kinase core gene occurs in three *Trichomonas* species as well as *Aduncisulcus paluster* and *Entamoeba histolyca* (Figure S7). In addition, primase helicase and mRNA capping core genes both occur in *Tritrichomonas musculus* (Figure S8), and one DNApolB from *Trichomonas vaginalis* branches within the clade of Egovirales (Figure 4). These positions suggest relatively recent transfers directed from virus to cell. Overall, the extent of signal between egoviruses and *Metamonada*, in the context of phylogenetic signal for *Egovirales* core genes and their co-existence in the gut, strongly indicates a prolonged viral-host relationship between the two clades.

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 467 kbp in length, and occur 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 encompass a limited diversity and abundance of eukaryotic DNA viruses[5–7], with the most prevalent ones, such as anelloviruses[64], having small, single-stranded DNA 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 (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, egovirus genome analysis helps to clarify evolutionary events that shaped the diversity and functionality of viruses in 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[28,65], so the discovery of a third order (*Egovirales*) evolutionarily related to both *Asfuvirales* and *Chitovirales* cements the monophyly of the class *Pokkesviricetes*. Second, egoviruses provide insights into the nature of the differences in the virion structure between the two previously identified orders of *Pokkesviricetes*, namely, 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 those of poxviruses despite substantial protein length differences and the fact that, in poxviruses, the DJR-MCP is incorporated into virion assembly intermediates but not into the mature virions[66–68]. Overall, phylogenetic analysis strongly favors an evolutionary scenario in which *Egovirales* and *Chitovirales* evolved from a common ancestor, with *Asfuvirales* diverging earlier in evolution. However, asfuviruses and egoviruses are the only *Nucleocytoviricota* clades known or predicted to harbor an inner MCP. These observations indicate 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 animal-associated viruses.

One critical question is the nature of the host cells infected by the egoviruses. 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[60]. One clade of closely related egoviruses is strictly associated with humans from different parts of the world, suggesting 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[69,70], the presence of egoviruses in humans from both 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. Our global comparative genomic analysis of *Nucleocytoviricota* and eukaryotes exposed a considerable rate of gene fluxes between egoviruses and metamonads, a phylum of unicellular eukaryotes that are prevalent in the gut, as shown in our metagenomic survey to co-exist with *Egovirales*. Most signal connected egoviruses and the parabasalids *Trichomonas* and *Tritrichomonas*. The extent of *Egovirales* genes spread in the genomes of *Trichomonas vaginalis* (more than 3%) was particularly notable, strongly indicate of a long-lasting history of infection events for this prominent sexually transmitted human pathogen.

Metamonads are amitochondriate unicellular eukaryotes that are primarily anaerobe, living in low-oxygen environments with free-living or symbiotic lifestyles. The parabasalids, in particular, include commensal species associated with the guts of insects (e.g., termites, which they aid for cellulose digestion), and parasitic species detected in association with the digestive and urogenital tracts of animals, including humans, and can cause pathologies[9,73–76]. Here, the discovery of egoviruses and extent of gene fluxes with co-existing metamonads point towards a long history of unicellular eukaryotic infections by *Nucleocytoviricota* in the gut of vertebrates, with *Trichomonas* and *Tritrichomonas* representing main vectors of the spread of the virus, including in the human population.

Conclusion

Egoviruses are DNA viruses with large genomes that form a third order-level clade in the class *Pokkesviricetes* (phylum *Nucleocytoviricota*), along with asfuviruses (order *Asfuvirales*) and poxviruses (order *Chitovirales*) to which egoviruses are most closely related. We propose to classify the egovirus clade as the order "*Egovirales*". In a sharp contrast to both asfuviruses and poxviruses, egoviruses are highly prevalent in the guts of vertebrates and are abundant in more than 2% of the human population, where they most likely infect unicellular eukaryotes affiliated with the metamonads. 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[77,78] to extract single-copy marker genes from metagenome assembled genomes to build the database for the taxonomic profiling tool mOTUs[79]. Briefly, BBMap[80] (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[81] (v3.15.2) in metagenomic mode. Contigs were length-filtered (≥ 500 bp) and gene sequences were predicted using Prodigal[82] (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[28] was searched against the predicted gene sequences from the metagenomic database using hmmsearch (HMMER[83] 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[29,84,85]. We generated a custom diamond database[86] (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[87] 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[88,89] (v.8) using muscle[90] and fast-tree[91] 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. This methodology recovered 13 complete *Egovirales* genomes.

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[8]. We added the 13 complete *Egovirales* genomes into this "*Nucleocytoviricota* genomic resource" and used Orthofinder[92] (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[52] profile-profile comparisons and ColabFold[49] (3D structure predictions and comparisons) were used to functionally annotate the *Egovirales* core genes. In the case of the DJR-MCP of *Egovirales*, ColabFold[49] successfully predicted its 3D structure and confirmed the presence of the double-jelly roll (DJR) protein components characteristic of the phylum *Nucleocytoviricota*.

Recovery of eight *Nucleocytoviricota* hallmark genes. We used HMMER[83] 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[8] 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 2024). We also added the Variola virus isolate VARV V563 genome from the GOEV database[29], and replaced the *Megaviricetes* representatives initially picked from the GVDB collection by others in the GOEV database[29] 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[93] v2.10.1) and phylogenetic reconstructions, as previously described[29]. 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. For the global phylogenies of the RNApolA and DJR-MCP of egoviruses (Figures 3 and S3), the database was restricted to egoviruses and representatives from the Chitovirales and Asfuvirales for rooting.

Alignments, trimming and individual gene phylogenetic analyses. From the genomic database created in the previous section (see "*Functional annotation of Egovirales core genes*"), protein sequences corresponding to the same hallmark gene were aligned with MAFFT[94] 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[95] v1.6.2, with the ModelFinder[96] option

to determine the best-fitting model. Supports were computed from 1,000 replicates for the Shimodaira-Hasegawa (SH)-like aLRT[97] and UFBoot[98]. 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. 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.

A second metagenomic survey using the DJR-MCP: We collected DJR-MCPs from *Egovirales* complete genomes to generate with HMMER[83] (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[83] 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.

A database of 224 egovirus MAGs. We analyzed all contigs containing an RNApolA gene or DJR-MCP gene affiliated to *Egovirales* in the context of their associated bins, which were generated using Metabat2[99] with differential coverage information, as previously described [77,78]. We removed contigs devoid of any *Egovirales* core gene found to be specific to this order (n=42), a stringent step to only work on high-quality viral MAGs. In addition, we removed MAGs with a quality score (completion minus redundancy) below 50% after creating a collection of 51 *Egovirales* single-copy-core genes and processing the database with anvi'o[88,89] (v.8.1) to predict gene sequences and functionally annotate hallmark genes using specific HMMs.

A phylogeny of *Nucleocytoviricota* using concatenated gene alignments. In addition to the single-protein phylogenies, 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 *Megaviricetes* (exclusion of *Medusavirus* to minimize long-branching artifacts) from the GOEV database[29], which were all already carefully manually curated. We added to this resource protein sequences for the DNApolB, RNApolA, RNApolB and TFIIS corresponding to all 224 *Egovirales* MAGs. For the DNApolB, RNApolA, RNApolB, and TFIIS, only sequences longer than 100, 200, 200, and 25 amino-acids respectively were kept. 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.

Metagenomic coverage values for egoviruses and co-occurring eukaryotes: We generated a non-redundant database (average nucleotide identity <95%) of egovirus MCP genes, egovirus RNApolA genes, and eukaryotic RNApolA genes characterized from the *Egovirales*-positive samples. We performed read recruitment of all the *Egovirales*-positive metagenomic samples against the database. For that, quality controlled reads were mapped (Bowtie2[100], v2.4.116) against the database and the 'anvi-profile-blitz'[88,89] routine used alignments to generate mean coverage values. Taxonomic information was

included using the diamond blast for RNApolA genes (see "Taxonomic affiliation of RNApolA sequences" section) and the phylogenetic tree for MCP genes (see "Alignments, trimming and individual gene phylogenetic analyses" section).

All-vs-all protein searches: Egovirales proteins, *Pokkesviricetes* from the GVDB collection, and *Megaviricetes* from the GOEV collection, were used as queries for similarity searches against UniProt-TrEMBL (UniProt Knowledgebase Release 2024-03[101]) using MMseqs2[102] with default settings. Selected terms ("d_Viruses"; "metagenome"; "unclassified"; "uncultured", "DNA molecule") were used as filters to remove hits not clearly assigned to cellular organisms. Only best hits with >= 30% identity and >= 100 bit scores were conserved for each query. Considering the initially observed enrichment in *Metamonada* hits for *Egovirales* proteins, the same queries were searched against additional metamonad genomes downloaded from NCBI Genome (list in Supp Table XX) with identical settings. The bit score, less sensitive to variations in target dataset sizes than e-values, was used to compare the hits between the two runs and keep the best only. MMseqs2 was also used to cluster *Egovirales* proteins at 30%, 50%, and 80% identity. Separately, protein sequences from the set of metamonad genomes plus *Trichomonas vaginalis* genome from UniProt-TrEMBL (which comprised a significant number of hits from *Egovirales* proteins), were used as queries for similarity searches against the entire set of *Egovirales* proteins on one hand, and against UniProt-TrEMBL on the other hand, with MMseqs2 and default settings. The same filtering criteria as previously described were applied, and hits to *Metamonada* were removed from the UniProt-TrEMBL search in order to remove self-match; only the bet hits were conserved. Finally, the metamonads queries with hits to egoviruses were also searched against the *Pokkesviricetes* and *Megaviricetes* to check their specificity. The CDS information for metamonads with the highest number of *Egovirales* hits were extracted when available, and the positions of those hits were mapped onto it. MMseqs2 was used to cluster metamonad proteins at 30% and 50% identity with default settings.

Gene transfers with metamonads: Five genes were particularly prominent amongst *Metamonada* hits for *Egovirales* queries, corresponding to the DNA-dependent DNA polymerase, the mRNA capping enzyme, the Primase Helicase, and a Holliday-Junction Resolvase, according to our functional annotation of egoviruses' core genes. For all but the former, HMM profiles of the corresponding aligned egoviruses' sequences were generated, and matching sequences were collected from *Pokkesviricetes*, *Megaviricetes*, the set of additional *Metamonada* (see previous section), and from the TCS database (derived from EukProt v3 and representing the diversity of eukaryotes[103]. Redundant sequences (>=95% identity) were removed with cdhit, as well as sequences significantly shorter than average. For the DNApolB gene, since the metamonad hit was from UniProt-TrEMBL, we directly incorporated this sequence along with *Egovirales* sequences in a large set of DNApolB sequences[29]. Gene trees were reconstructed using the same approach as previously described, and long branches and/or ambiguous taxa were iteratively removed, as well as clades particularly distant from *Egovirales* sequences.

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) HMMs for the corre genes and single copy core genes of Egovirales, (5) the 224 *Egovirales* MAGs (including complete genomes), (6) the

3D structure of the *Egovirales* DJR-MCP, (7) and the non-redundant database of MCP and RNApolA genes covering the diversity of egoviruses and eukaryotes found in the *Egovirales*-positive samples. These items along with the supplementary tables (Tables S01-S08) have been made publicly available at https://doi.org/10.6084/m9.figshare.25398895.v7.

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.
Table S8: Metagenomic coverage values for egoviruses and eukaryotes among the *Egovirales*-positive samples based on the MCP and RNApolA surveys.

Supplemental Figures

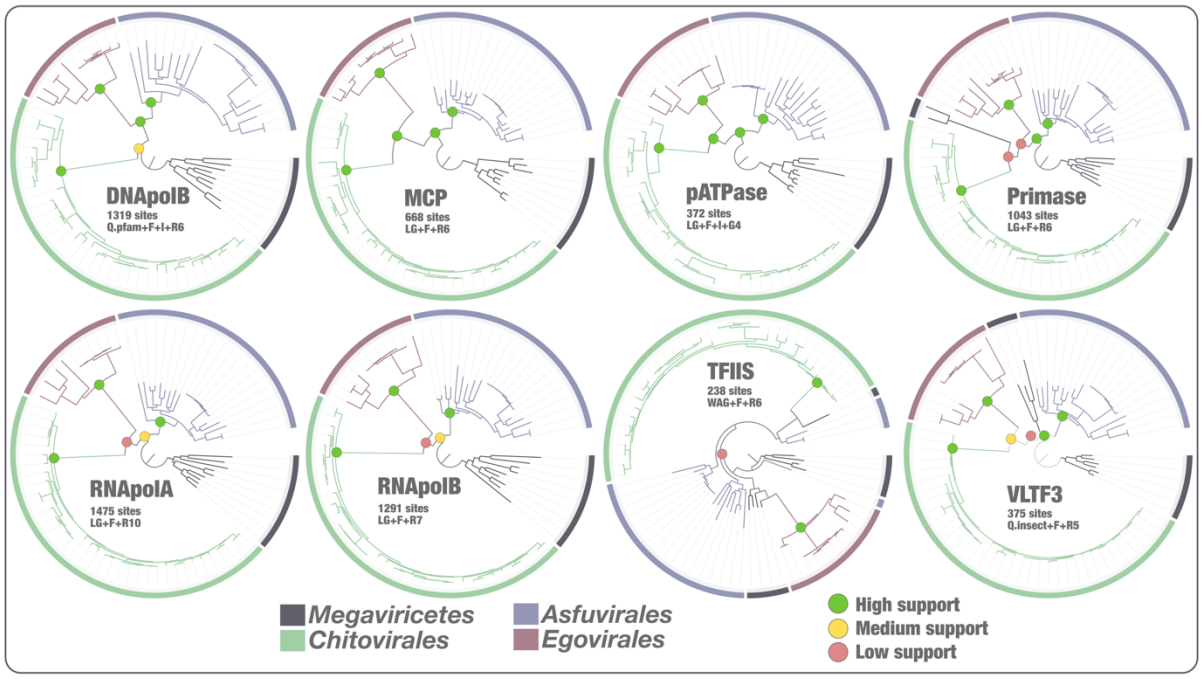


Figure S1: Phylogeny of *Pokkesviricetes* using seven individual hallmark genes. Each tree corresponds to a phylogeny of the same genomic resource (11 *Egovirales*, 36 *Chitovirales*, 21 *Asfuvirales*, and 9 *Megaviricetes*) using one of the eight hallmark genes shared by the tree orders of the class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

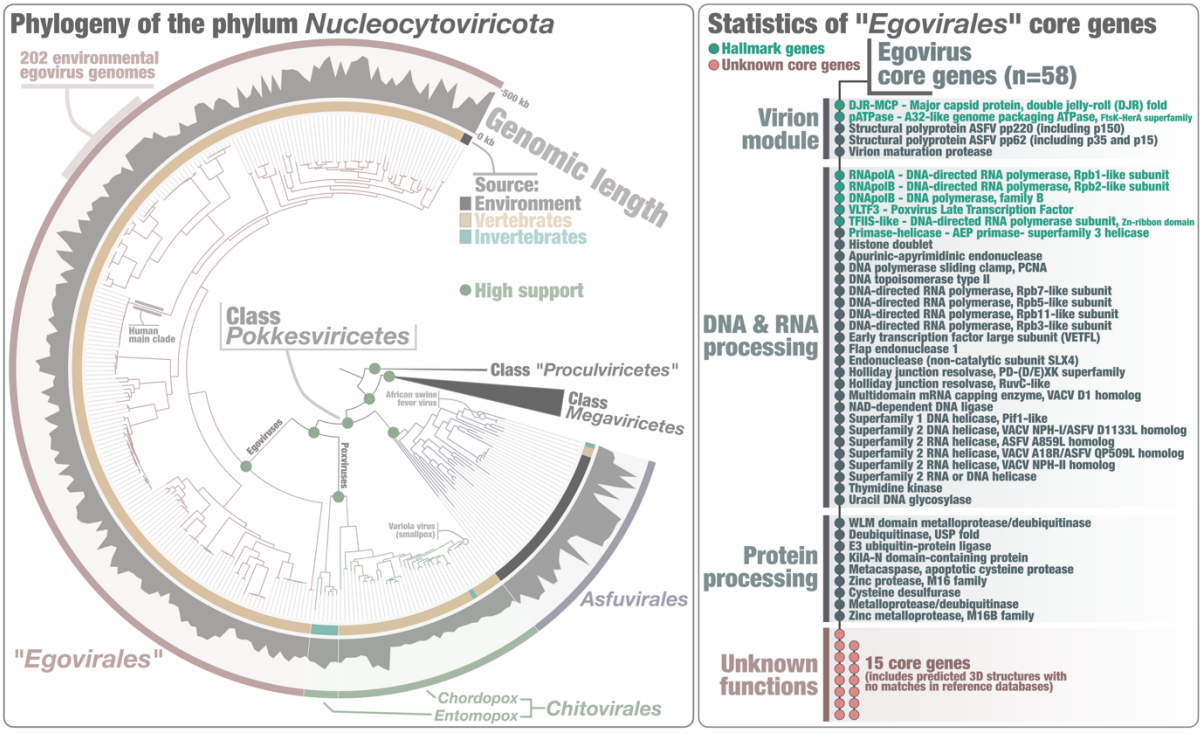


Figure S2: Evolution and predicted gene functions of egoviruses. Left panel displays a maximum-likelihood phylogenomic analysis of members of *Nucleocytoviricota*, along with 202 *Egovirales* genomes with a quality score above 50%. The tree covers 1,807 genomes and was built using concatenated multiple alignments of four hallmark proteins of *Nucleocyovoricota* (DNApolB, RNApolA, RNApolB, TFIIS; 4,110 sites) using IQTree with the LG+F+R10 model. The tree was rooted with the putative class *Proculviricetes*

and visualized using anvi'o v8. Green dots at nodes in the tree highlight high supports (approximate likelihood ratio (aLRT) ≥ 80 and ultrafast bootstrap approximation (UFBoot) ≥ 95; see Methodds). Right panel summarizes the functional annotation of 58 *Egovirales* core genes.

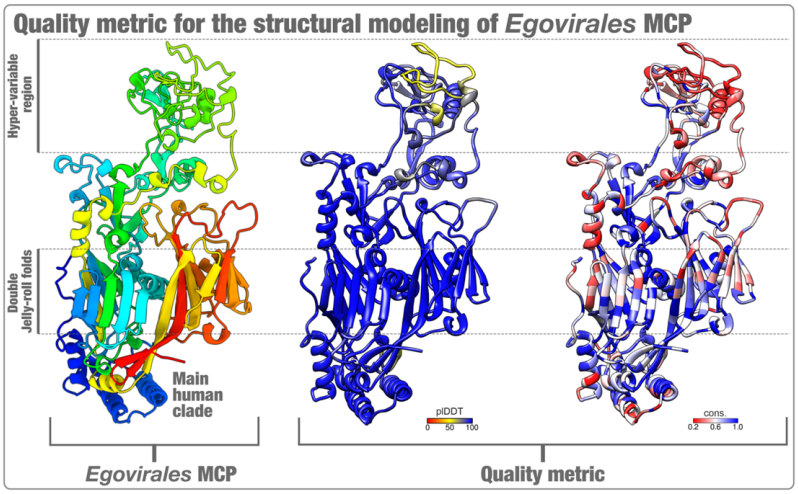


Figure S3: Quality metric of the structural modeling of the Egovirales major capsid protein. Left panel displays the structural modeling of an *Egovirales* DJR MCP (human main clade), as predicted from Colabfold[49]. Right panels display the pIDDT and conservation levels (quality metric).

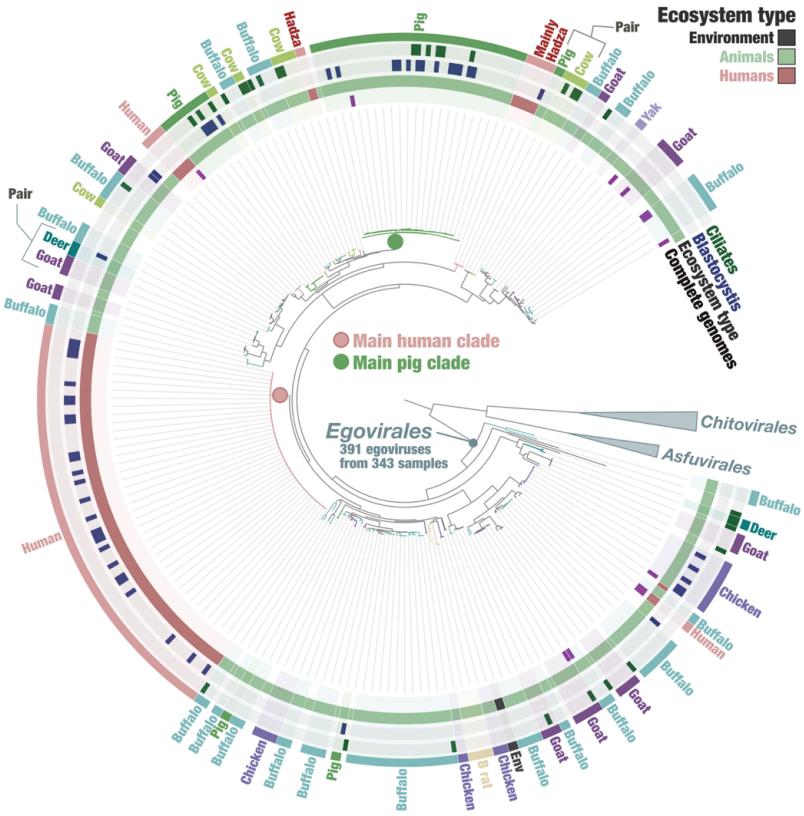


Figure S4: Overlay of the evolutionary tree of egoviruses and their occurrence among vertebrate species. The figure displays a maximum-likelihood phylogeny of the MCP of egoviruses (>600 amino acids; n=391), using representatives of *Chitovirales* and *Asfuvirales* for rooting (784 sites), and the LG+F+R10 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.

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%

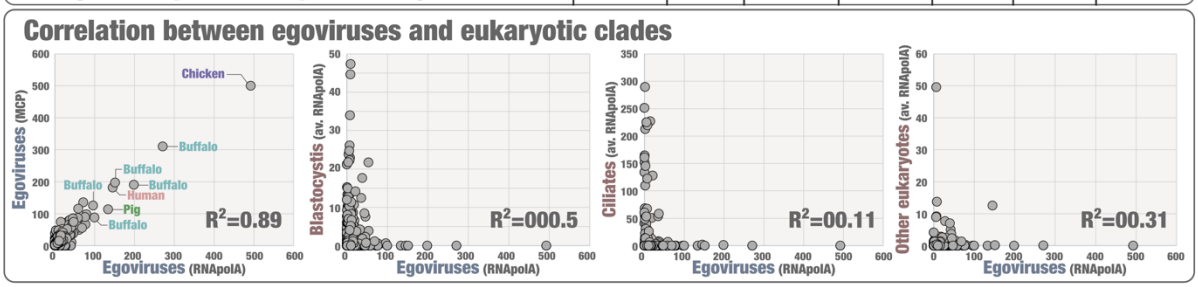


Figure S5: Prevalence of egoviruses and co-occurring unicellular eukaryotes. Top panel summarizes the positivity rate for egoviruses, *Blastocystis* and ciliates in gut metagenomes from five vertebrate species. Bottom panel displays the mean coverage correlation between egoviruses and clades of unicellular eukaryotes across metagenomes corresponding to the 864 *Egovirales*-positive samples.

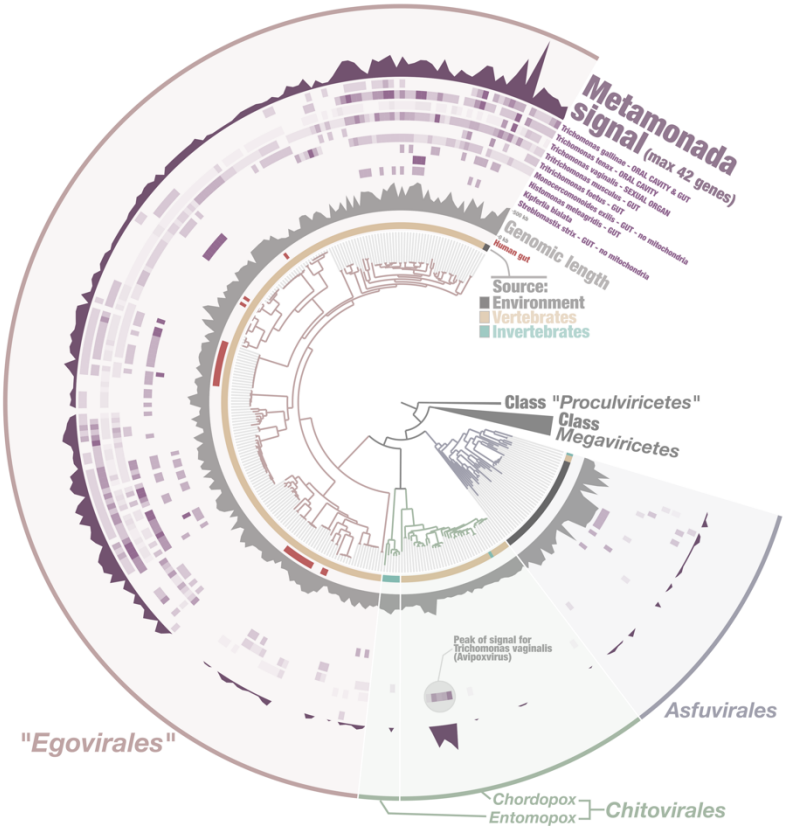


Figure S6: Gene flux between egoviruses and *Metamonada*. Inner tree corresponds to the maximum-likelihood phylogenomic analysis of Figure 1. The tree covers 1,807 genomes and was built using concatenated multiple alignments of four hallmark proteins of *Nucleocyovoricota* (DNApolB, RNApolA, RNApolB, TFIIS; 4,110 sites) using IQTree with the LG+F+R10 model. The tree was rooted with the putative class *Proculviricetes* and visualized using anvi'o v8. The number of egovirus genes best matching to that of a *Metamonada* genome (only the 9 genomes with highest signal for Egovirales are displayed) are summarized as additional layers of information.

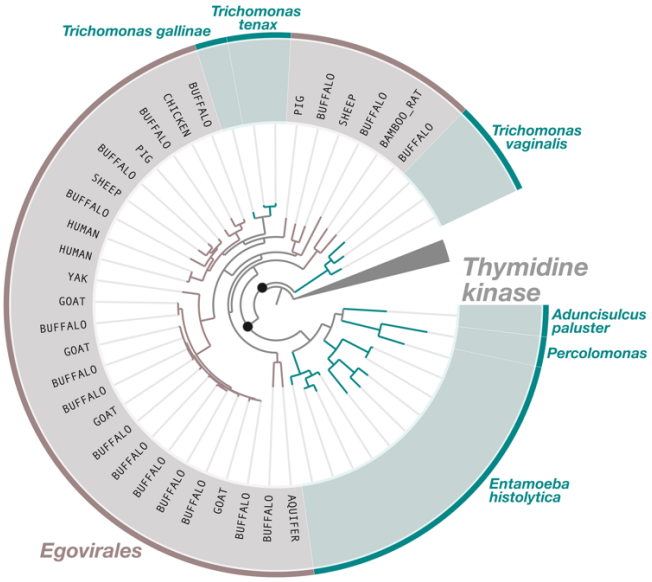


Figure S7: Phylogeny of the Thymidine kinase. The tree corresponds to the maximum-likelihood phylogeny of the thymidine kinase of *Nucleocytoviricota* and eukaryotes (minimum protein length: 50 aa; 613 taxa; 232 sites; model LG+F+R9), rooted at the level of a *Metamonada* clade closely related to *Egovirales*. The figure shows *Metamonada* genes closely related to that of *Egovirales*.

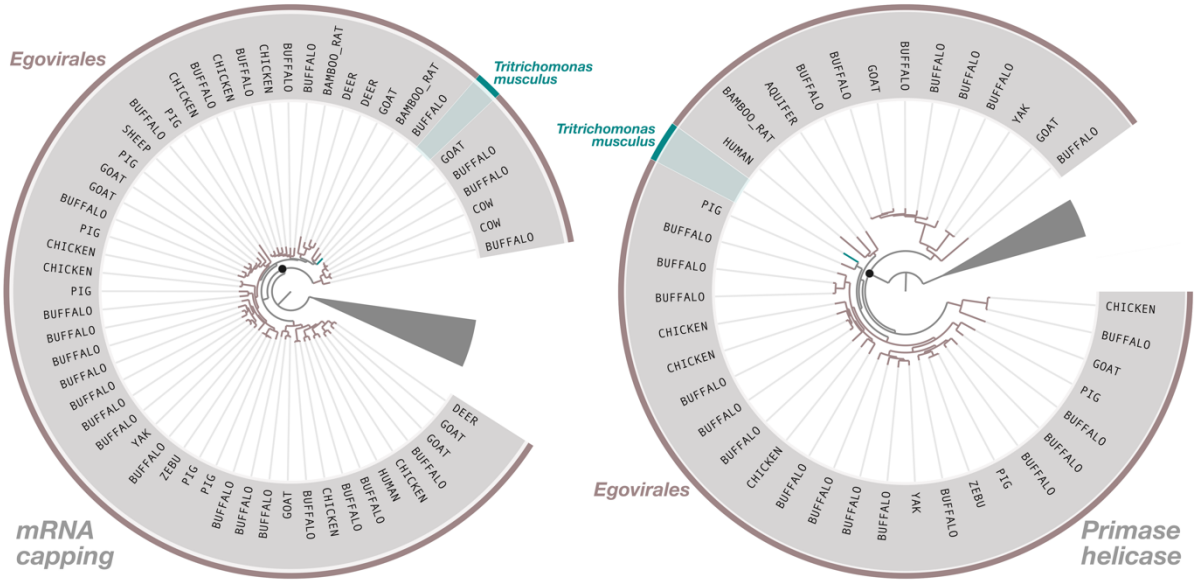


Figure S8: Phylogenies of the mRNA capping and primase helicase. Left tree corresponds to the maximum-likelihood phylogeny of the mRNA capping of *Nucleocytoviricota* and eukaryotes (minimum protein length: 100 aa; 1,514 taxa; 495 sites; model LG+F+R10), rooted at the level of *Egovirales*. Right tree corresponds to the maximum-likelihood phylogeny of the primase helicase of *Nucleocytoviricota* and eukaryotes (minimum protein length: 150 aa; 828 taxa; 904 sites; LG+F+R10), rooted at the level of *Egovirales*.

References:

1. Vincent, F., Sheyn, U., Porat, Z., Schatz, D. & Vardi, A. Visualizing active viral infection reveals diverse cell fates in synchronized algal bloom demise. *Proc Natl Acad Sci U S A* **118**, e2021586118 (2021).
2. Suttle, C. a. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* (2007) doi:10.1038/nrmicro1750.

3. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336 (2022).
4. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* (2020) doi:10.1038/s41586-020-2924-2.
5. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology 2021 19:8* **19**, 514–527 (2021).
6. 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).
7. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* (2020) doi:10.1038/s41586-020-1957-x.
8. 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).
9. Gerrick, E. R. & Howitt, M. R. The Lost Kingdom: commensal protists in the gut microbiota. *Trends Microbiol* (2025) doi:10.1016/J.TIM.2025.01.008.
10. McFadden, G. Poxvirus tropism. *Nature Reviews Microbiology 2005 3:3* **3**, 201–213 (2005).
11. Smith, S. A. & Kotwal, G. J. Immune response to poxvirus infections in various animals. *Crit Rev Microbiol* **28**, 149–185 (2002).
12. Lu, B. *et al.* Outbreak of Vaccinia Virus Infection from Occupational Exposure, China, 2017. *Emerg Infect Dis* **25**, 1192 (2019).
13. Wilson, M. E., Hughes, J. M., McCollum, A. M. & Damon, I. K. Human Monkeypox. *Clinical Infectious Diseases* **58**, 260–267 (2014).
14. Chen, X., Anstey, A. V. & Bugert, J. J. Molluscum contagiosum virus infection. *Lancet Infect Dis* **13**, 877–888 (2013).
15. Mühlemann, B. *et al.* Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science (1979)* **369**, (2020).
16. Hannat, S., La Scola, B., Andreani, J. & Aherfi, S. Asfarviruses and Closely Related Giant Viruses. *Viruses 2023, Vol. 15, Page 1015* **15**, 1015 (2023).
17. Dixon, L. K., Sun, H. & Roberts, H. African swine fever. *Antiviral Res* **165**, 34–41 (2019).
18. Reteno, D. G. *et al.* Faustovirus, an Asfarvirus-Related New Lineage of Giant Viruses Infecting Amoebae. *J Virol* **89**, 6585 (2015).
19. Hannat, S., La Scola, B., Andreani, J. & Aherfi, S. Asfarviruses and Closely Related Giant Viruses. *Viruses 2023, Vol. 15, Page 1015* **15**, 1015 (2023).
20. Karki, S., Moniruzzaman, M. & Aylward, F. O. Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae. *Front Microbiol* **12**, (2021).
21. 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).
22. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, (2020).
23. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).
24. Laforest-Lapointe, I. & Arrieta, M.-C. Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems* **3**, (2018).

25. 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).
26. Parfrey, L. W., Walters, W. A. & Knight, R. Microbial Eukaryotes in the Human Microbiome: Ecology, Evolution, and Future Directions. *Front Microbiol* **2**, (2011).
27. 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).
28. 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).
29. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature 2023 616:7958* **616**, 783–789 (2023).
30. Carter, M. M. *et al.* Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* **186**, 3111-3124.e13 (2023).
31. 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).
32. 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).
33. Hildebrand, F. *et al.* Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host Microbe* **29**, 1167-1176.e9 (2021).
34. Fukuyama, J. *et al.* Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput Biol* **13**, e1005706 (2017).
35. 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).
36. 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).
37. Levin, D. *et al.* Diversity and functional landscapes in the microbiota of animals in the wild. *Science (1979)* **372**, (2021).
38. Tong, F. *et al.* The microbiome of the buffalo digestive tract. *Nature Communications 2022 13:1* **13**, 1–16 (2022).
39. 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).
40. 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).
41. 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).
42. Cao, Y. *et al.* The multi-kingdom microbiome of the goat gastrointestinal tract. *Microbiome* **11**, 1–20 (2023).
43. Gaio, D. *et al.* A large-scale metagenomic survey dataset of the post-weaning piglet gut lumen. *Gigascience* **10**, 1–8 (2021).
44. 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).
45. Xiao, L. *et al.* A reference gene catalogue of the pig gut microbiome. *Nature Microbiology 2016 1:12* **1**, 1–6 (2016).

46. Wittek, R. *et al.* Inverted terminal repeats in rabbit poxvirus and vaccinia virus DNA. *J Virol* **28**, 171–181 (1978).
47. Alonso, C. *et al.* ICTV Virus Taxonomy Profile: Asfarviridae. *J Gen Virol* **99**, 613–614 (2018).
48. McInnes, C. J. *et al.* ICTV Virus Taxonomy Profile: Poxviridae 2023. *J Gen Virol* **104**, (2023).
49. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
50. Liu, Y. & Krupovic, M. Genome chromatinization in giant double-stranded DNA viruses. *Trends Biochem Sci* **47**, 3–5 (2022).
51. Krupovič, M. & Bamford, D. H. Virus evolution: how far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol* **6**, 941–948 (2008).
52. 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).
53. 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).
54. 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).
55. Li, G. *et al.* Crystal structure of the African swine fever virus structural protein p35 reveals its role for core shell assembly. *Protein Cell* **11**, 600–605 (2020).
56. Condit, R. C., Moussatche, N. & Traktman, P. In A Nutshell: Structure and Assembly of the Vaccinia Virion. *Adv Virus Res* **66**, 31–124 (2006).
57. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219 (2016).
58. 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).
59. 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).
60. Haller, S. L., Peng, C., McFadden, G. & Rothenburg, S. Poxviruses and the Evolution of Host Range and Virulence. *Infect Genet Evol* **0**, 15 (2014).
61. Chen, Z. *et al.* Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog* **14**, e1007352 (2018).
62. 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).
63. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. *Trends Ecol Evol* **35**, 43–55 (2020).
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. Hyun, J., Matsunami, H., Kim, T. G. & Wolf, M. Assembly mechanism of the pleomorphic immature poxvirus scaffold. *Nature Communications 2022 13:1* **13**, 1–10 (2022).
67. Hyun, J. K. *et al.* Membrane remodeling by the double-barrel scaffolding protein of poxvirus. *PLoS Pathog* **7**, (2011).

68. Bisht, H., Weisberg, A. S., Szajner, P. & Moss, B. Assembly and disassembly of the capsid-like external scaffold of immature virions during vaccinia virus morphogenesis. *J Virol* **83**, 9140–9150 (2009).
69. Segata, N. Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. *Current Biology* **25**, R611–R613 (2015).
70. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nature Communications 2014 5:1* **5**, 1–12 (2014).
71. Sender, R. & Milo, R. The distribution of cellular turnover in the human body. *Nature Medicine 2021 27:1* **27**, 45–48 (2021).
72. 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).
73. Gerrick, E. R. *et al.* Metabolic diversity in commensal protists regulates intestinal immunity and trans-kingdom competition. *Cell* **187**, 62-78.e20 (2024).
74. Zhang, N. *et al.* High prevalence of Pentatrichomonas hominis infection in gastrointestinal cancer patients. *Parasit Vectors* **12**, 1–9 (2019).
75. Popovic, A. *et al.* Commensal protist Tritrichomonas musculus exhibits a dynamic life cycle that induces extensive remodeling of the gut microbiota. *ISME J* **18**, 23 (2024).
76. Conrad, M. D., Bradic, M., Warring, S. D., Gorman, A. W. & Carlton, J. M. Getting trichy: tools and approaches to interrogating Trichomonas vaginalis in a post-genome world. *Trends Parasitol* **29**, 17–25 (2013).
77. Ruscheweyh, H. J. *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* **10**, 1–12 (2022).
78. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature 2022* 1–8 (2022) doi:10.1038/s41586-022-04862-3.
79. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**, 1196–1199 (2013).
80. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
81. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res* **27**, 824–834 (2017).
82. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
83. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–W37 (2011).
84. 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.
85. 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.
86. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
87. 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).
88. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
89. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology 2020 6:1* **6**, 3–6 (2020).

90. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
91. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
92. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, (2015).
93. 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).
94. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
95. 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).
96. 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).
97. 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).
98. 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).
99. Kang, D. D. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019**, (2019).
100. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4* **9**, 357–359 (2012).
101. Consortium, T. U. *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
102. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology 2017 35:11* **35**, 1026–1028 (2017).
103. Richter, D. J. *et al.* EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal* **2**, (2022).