Exploring the double-stranded DNA viral landscape in eukaryotic genomes

Hongda Zhao[1], Lingjie Meng[1], Ruixuan Zhang[1], Morgan Gaïa[2,3], Hiroyuki Ogata[1]*

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

[2] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, F-91057 Evry, France

[3] Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara GOsee, F-75016 Paris, France

* Corresponding author: Hiroyuki Ogata (ogata@kuicr.kyoto-u.ac.jp)

Abstract

Sequences of viral origin constitute a large proportion of the genomes of certain eukaryotes, yet the role of double-stranded (ds)DNA viruses in shaping the genomes of diverse eukaryotes remains underexplored. Here, we present a computational framework to identify dsDNA viral regions (VRs) in eukaryotic genomes, which was we used to screen 37,254 eukaryotic genome assemblies. In total, 781,111 VRs were identified across 7,103 (19%) assemblies. VRs were found to occupy 16% of a bivalve genome assembly and over 10% of diverse invertebrate and protist genomes. These VRs recovered 29 of the 43 class-level virus–eukaryote associations previously established by virus isolation, and additionally indicated 144 candidate associations for which virus isolation evidence is currently lacking. Some VRs form previously unrecognized deep viral clades, whereas others are phylogenetically related to known viruses. Our study provides a baseline for the extent to which dsDNA viral elements are embedded in diverse eukaryotic genomes and offers an overview of the endogenous dsDNA virosphere. The resulting catalog of viral genomic elements offers opportunities for wider exploration of the significance of virus–host coevolutionary processes.

Introduction

Endogenous viral elements are widespread across diverse eukaryotic genomes[1,2]. Endogenous retroviruses in particular account for a considerable proportion of certain eukaryotic genomes, including ~8% of the human genome[3–5]. These retroviral elements can occasionally cause diseases, provide essential host functions through 'co-option' or contribute antiviral functions[6–9].

In contrast to retroviruses, endogenous elements derived from dsDNA viruses remain underexplored[2,10]. A few

31 eukaryotic dsDNA viruses, such as human herpesvirus 6 (HHV-6) and the mavirus virophage, are known to integrate
32 their genomes into host chromosomes during their replication cycle[11,12]. Additional dsDNA viral lineages have
33 recently been revealed to persist within host cells as integrated or episomal forms, with the potential for reactivation
34 and virion production[12–15]. Domesticated dsDNA viral elements can enhance host survival, as exemplified in
35 parasitoid wasps, where *Bracoviriform* particles derived from endogenous viral elements are delivered into
36 parasitized lepidopteran larvae to suppress immunity and support wasp development[16,17]. Recent studies have begun
37 to uncover signatures of specific dsDNA virus lineages in protist genomes[18,19] and horizontal transfers of viral genes
38 to eukaryotes[20–23]. Given their large coding capacity compared with RNA viruses, endogenization of dsDNA viruses
39 may substantially affect the evolution of their host genomes[24].

40 Eukaryotic dsDNA viruses are polyphyletic and highly diverse. They range from small viruses, such as
41 papillomavirids[25] (including papillomaviruses) and viruses from the polinton-like supergroup[26,27], to giant
42 representatives like megaviricetes[28]. Of note, the known eukaryotic dsDNA virosphere is rapidly expanding thanks
43 to culture-independent metagenomics[29], which has led to discoveries of novel dsDNA viral groups, including
44 polinton-like viruses[30] (PLVs, also classified as *Aquintoviricetes*), *Mirusviricota*[31], *Mriyaviricetes*[32], and the
45 proposed *Egovirales* order[33]. Eukaryotic genome data are also accumulating at an impressive pace through
46 international initiatives such as the Earth BioGenome Project[34]. Therefore, it is timely to re-evaluate dsDNA viral
47 elements in diverse eukaryotes in light of these expanding datasets to better understand the endo-virosphere.

48

49 Results

50 Hundreds of thousands of dsDNA viral regions recovered from eukaryotic genomic assemblies

51 We conducted a systematic analysis to detect dsDNA VRs across eukaryotic genomic assemblies (**Fig. 1a**). In
52 this study, VRs are operationally defined by our detection framework; given the large number of identified VRs, we
53 do not aim to delineate the exact biological units or boundaries of individual VRs, which may be affected by complex
54 evolutionary processes such as rearrangements or recurrent insertions. In total, 37,254 assemblies encompassing all
55 sequenced eukaryotes (from GenBank, as of Feb. 2024) covering diverse lineages were analyzed (**Fig. 1b,**
56 **Supplementary Table 1**). Open reading frames (ORFs) were predicted for each assembly. To aid the identification
57 of viral signatures within eukaryotic contigs, scaffolds, and chromosomes (collectively referred to as 'contigs'
58 hereafter), we constructed a comprehensive reference dataset of cellular protein sequences derived from curated
59 databases.

60 Our study incorporated 2,126 reference viral genomes from 11 eukaryotic dsDNA viral taxa (**Table 1,**

61 **Supplementary Table 2**). These taxa, corresponding to one phylum, six classes, two orders, and two families, are
62 referred to as viral major taxa in this study. The dsDNA pararetroviruses (i.e., *Hepadnaviridae* and *Caulimoviridae*)
63 were not included in this dataset, as they are classified within the realm *Riboviria*, which is highly dominated by
64 RNA viruses. Hidden Markov models (HMMs) were constructed for 7,755 orthologous groups (OGs) derived from
65 these viral genomes. Among them, 1,842 viral OGs (VOGs), defined as OGs that exhibit only marginal or no
66 similarity to any cellular reference sequence, and 1,344 taxonomic OGs (TOGs), characterized as VOGs strictly
67 specific to a single viral major taxon, were selected to capture viral and viral lineage-specific signatures, respectively.
68 On the basis of the presence of ORFs with significant similarity to VOGs or TOGs within eukaryotic contigs (**Fig.**
69 **1a**), we identified 946,599 distinct combinations of viral major taxa and contigs (i.e., "virus–contig pairs").

70 **Table 1 | Eleven dsDNA viral major taxa included in this study and representative viruses.**

Major viral taxon	ICTV taxonomy	Representative viruses	Average genome size
Adenoviridae	*Varidnaviria › Bamfordvirae › Preplasmiviricota › Polisuviricotina › Pharingeaviricetes* (monotypic)	Canine adenovirus 1[35]	33.5 kb
Herpesvirales	*Duplodnaviria › Heunggongvirae › Peploviricota › Herviviricetes* (monotypic)	Herpes simplex virus[36]; Salmonid herpesvirus 1[37]	159.1 kb
Lavidaviridae *(Virophaviricetes)*	*Varidnaviria › Bamfordvirae › Preplasmiviricota › Polisuviricotina › Virophaviricetes*	Mimivirus-dependent virus Sputnik ("virophage")[38]	22.1 kb
Megaviricetes	*Varidnaviria › Bamfordvirae › Nucleocytoviricota › Megaviricetes*	Megavirus chilensis[39]; Invertebrate iridescent virus 31[40]	382.4 kb
Mirusviricota	Not yet assigned in ICTV. (*Duplodnaviria › Heunggongvirae*[31])	Circular genomes uncovered from eukaryotes[19,41]	217.7 kb
Mriyaviricetes	*Varidnaviria › Bamfordvirae › Nucleocytoviricota › Mriyaviricetes*	Yaravirus brasiliensis[42]	37.4 kb
Naldaviricetes	*Naldaviricetes* (Unclassified Class)	Cryptophlebia leucotreta granulovirus[43]; White spot syndrome virus[44]	134.1 kb
Papovavircetes	*Monodnaviria › Shotokuvirae › Cossaviricota › Papovaviricetes*	Human papillomavirus[45]; Simian virus 40[46]	6.6 kb
PLVs *(Aquitoviricetes)*	*Varidnaviria › Bamfordvirae › Preplasmiviricota › Polisuviricotina › Aquintoviricetes*	Gezel-14T ("virophage")[47]	18.3 kb
Pokkesviricetes	*Varidnaviria › Bamfordvirae › Nucleocytoviricota › Pokkesviricetes*	African swine fever virus[48]; Monkeypox virus[49]	182.0 kb
Orthopolintovirales (*Amphintovirales*)	*Varidnaviria › Bamfordvirae › Preplasmiviricota › Polisuviricotina › Polintoviricetes* (monotypic)	Terrapene box turtle adintovirus[50]	7.8 kb

71

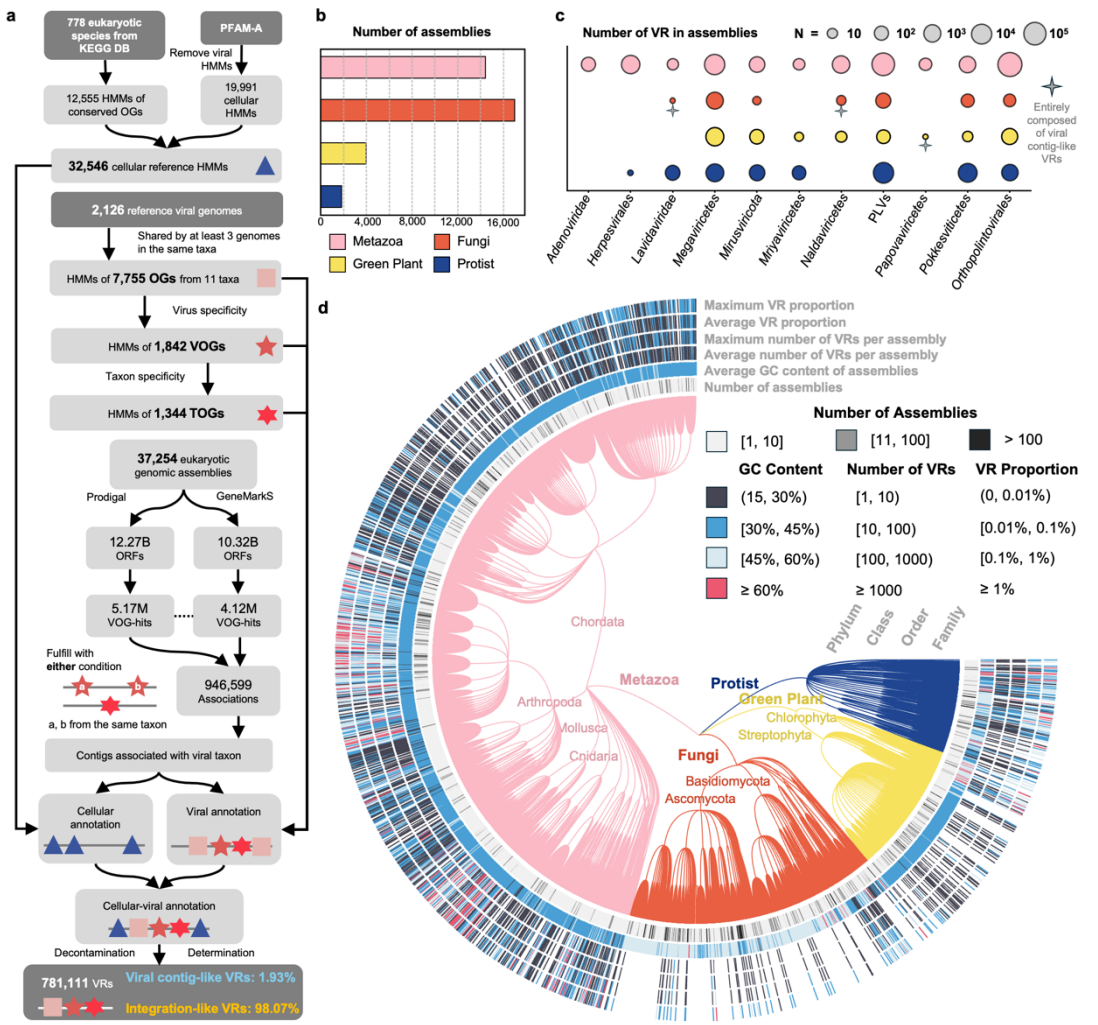


Fig. 1 | Viral regions identified in eukaryotic genome assemblies.

(a) VR discovery pipeline. Different symbols represent distinct types of hits corresponding to each ORF. Viral contig–like VRs occupied over half of the contig length and contained fewer than 20% of cellular ORFs; the remaining VRs were classified as integration-like. **(b) The number of eukaryotic genomic assemblies used in this study.** Colors indicate different eukaryotic groups. **(c) The number of identified VRs.** Colors match those in (b). The radius of each circle represents the logarithmically transformed count of VRs. **(d) VR content across the tree of eukaryotes.** The cladogram is based on taxonomy, and each leaf represents a eukaryotic family. The attributes shown in the outer ring are explained in the diagram in the upper-right corner.

For each virus–contig pair, we scanned the contig using HMMs for the OGs from the viral major taxon to delineate VRs as regions enriched in clustered viral ORFs. This analysis identified a total of 781,111 VRs in 7,103

83 genome assemblies (19% of all assemblies) (**Supplementary Table 3**). These assemblies included 6,094 metazoan
84 assemblies (42.25% of all metazoan assemblies) and 633 protist assemblies (34.39%). Even though no dsDNA
85 viruses, except for pararetroviruses, have been isolated from fungi or higher plants, 168 Streptophyta assemblies
86 (4.52%) and 153 fungal assemblies (0.89%) were found to contain dsDNA VRs (e.g., *Megaviricetes* and
87 *Orthopolintovirales*). These results are consistent with previous observations of imiterviral (*Megaviricetes*) signals
88 in a bryophyte, a lycophyte[51,52], and zoosporic fungi[53]. The majority of VRs were taxonomically associated with
89 *Orthopolintovirales* and PLVs (**Fig. 1c, Supplementary Table 4**).

90 The distribution of VRs across different host lineages was uneven, with metazoans accounting for the largest
91 proportion (757,529 VRs, 96.98%), followed by protists (20,186 VRs, 2.58%), green plants (2,442 VRs, 0.31%),
92 and fungi (954 VRs, 0.12%) (**Fig. 1c**). The mean of the family-level average numbers of VRs per assembly is 101.84,
93 19.23, 1.72, and 0.28 for metazoans, protists, green plants, and fungi, respectively. VRs were identified in 2,951
94 genera, but approximately 80% of VRs were derived from 300 genera (**Supplementary Fig. 1a**), and 4.3% originated
95 from a single genus, *Unio* (bivalve). We examined the relationship between the abundance of VRs and the
96 completeness of genome assemblies (**Supplementary Fig. 1b**). Neither the VR content (i.e., the proportion of VRs
97 in individual assemblies) nor the number of VRs was correlated with assembly completeness, indicating that
98 potential viral contamination in low-quality assemblies had no major influence on the overall detection of VRs.

99 To distinguish potential endogenizations from unintegrated viral sequences, we classified VRs into
100 'integration-like' or 'viral contig-like' categories based on the VR and contig lengths as well as the cellular OG
101 content (**Fig. 1a**). The majority (n = 766,069; 98%) of VRs were classified into the integration-like category, whereas
102 only a small subset of VRs (n = 15,042; 2%) was classified as viral contig-like and lacked signatures of integration.
103

104 **General features of viral regions**

105 We investigated several characteristics of VRs to assess their quality as *bona fide* virally derived regions.
106 Regions derived from viral genomes are expected to be rich in coding sequences compared with eukaryotic genomic
107 regions. The ORF density of a VR or a eukaryotic assembly was calculated as the total length of ORFs divided by
108 the length of the VR or the assembly, respectively. Considering the effect of the definition of a VR (i.e., it starts with
109 an ORF and ends with another ORF) on the computed ORF density, we only considered VRs longer than 50 kb. The
110 analyzed VRs exhibited higher coding potential than their host genomes (**Supplementary Fig. 1d**), with most
111 showing 20–40% higher ORF density, suggesting their distinct origin from the surrounding genomic regions.

112 We investigated tetranucleotide frequency (TNF) differences between VRs and their host genomic regions (see

113 Methods for details). The proportion of VRs showing TNF values distinct from the host genomes was highly
114 dependent on their assigned viral major taxa (**Supplementary Fig. 1f**). VRs assigned to *Herpesvirales*,
115 *Papovaviricetes*, and *Lavidaviridae* tend to exhibit TNF compositions distinct from their host assemblies, whereas
116 most VRs related to *Orthopolintovirales*, *Pokkesviricetes*, and *Mriyaviricetes* display compositions more similar to
117 those of their hosts.

118 We assessed the cellular homologs within the identified VRs to further qualify their viral likeness
119 (**Supplementary Fig. 1c**). Nearly three-quarters of VRs contained less than 20% of ORFs with best hits to cellular
120 references, indicating a general enrichment of viral genes in the VRs.

121 The VRs showed shorter ORFs than those in the reference viral genomes. The normal length range of a viral
122 ORF was defined based on the length distribution of ORFs in the same OG from reference viral genomes of the same
123 viral major taxon. For most VRs, nearly half of their ORFs fell outside this range (**Supplementary Fig. 1e**),
124 suggesting that pseudogenization through degradation is a common phenomenon in VRs.

125 We noticed that some VRs exhibited chimeric signatures from multiple viral taxa (**Supplementary Table 3**).
126 For example, the previously identified *Nucleocytoviricota*-like 1.5 Mb endogenous viral element in a fungal
127 chromosome (CP080821.1)[54] corresponded to four VRs. One of the VRs contained *Pokkesviricetes*-specific
128 homologs and *Megaviricetes*-specific homologs. These observations may reflect either the existence of chimeric
129 viruses from unknown clades or the accumulation of multiple independent viral insertions within the same locus.

130

131 **dsDNA viral elements make substantial contributions to the genome composition of diverse eukaryotes**

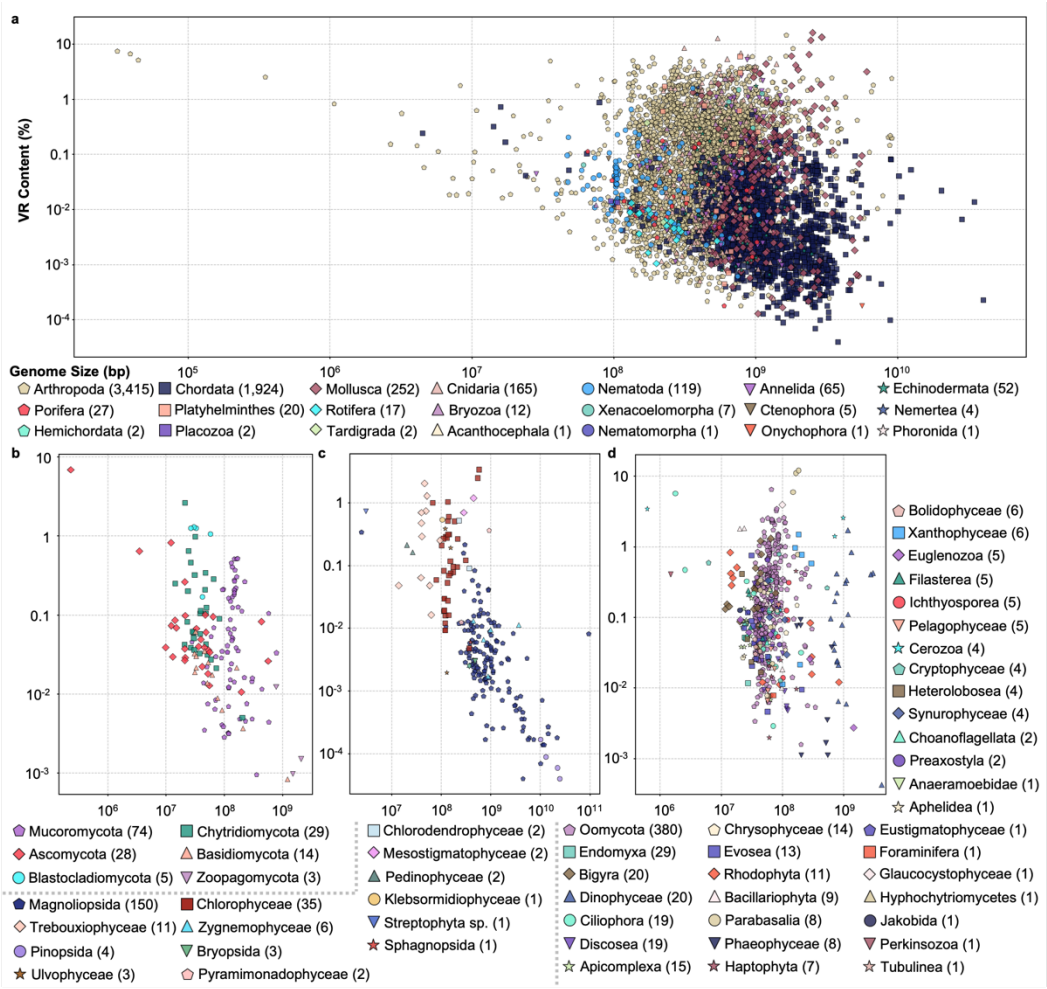
132 The VR content showed substantial variation across host lineages (**Fig. 1d**). The VR content of metazoans
133 ranged from trace levels up to over 10% (**Fig. 2**). High VR content was observed for animal species belonging to
134 different taxa, such as Bivalvia (*Unio delphinus*, 16.20%; *Sinohyriopsis cumingii*, 13.38%; *Unio pictorum*, 11.71%),
135 Insecta (*Neoaliturus tenellus*, 14.68%), and Anthozoa (*Ricordea florida*, 12.90%). The three bivalve species belong
136 to the same freshwater mussel family Unionidae[55]. VRs accounted for 0.004% (75 VRs from 59 assemblies) of the
137 human genome on average, and those in vertebrate genomes accounted for 0.014% on average, with the highest
138 proportion (0.725%) observed in the assembly of a fish, *Gasterosteus aculeatus*.

139 Protists showing VR content higher than 1% belonged to Parabasalia, Oomycota, Ciliophora, Cercozoa,
140 Bacillariophyta, Glaucocystophyceae, Dinophyceae, Xanthophyceae, and Bigyra. *Trichomonas vaginalis*, a
141 flagellated protozoan pathogen (phylum Parabasalia), showed the highest VR content (12.13%) among protists. The
142 genome was enriched in VRs assigned to *Pokkesviricetes* (1,514 VRs, 97%) or *Megaviricetes* (43 VRs, 3%). The

143 *Pokkesviricetes* VRs showed specific similarities to the genomes of recently discovered egovirals in this class[33].

144 In fungi and streptophytes, the VR content was generally low and seldom exceeded 1% (**Fig. 2**). The highest
145 VR content was observed in a lichen assembly of *Cladonia squamosa* (6.93%) among fungi and in the unicellular
146 biflagellate alga *Mesostigma viride* (1.19%) among streptophytes.

147 The high VR content appears to result, at least in part, from intragenomic proliferation of existing elements,
148 although the contribution of this process varies among assemblies. Among the 435 assemblies containing more than
149 1% VR content, 10 assemblies had a proportion of nearly identical VRs (≥90% identity and ≥90% coverage) above
150 30% (with the maximum of 50% for an insect genome, *Sphinx pinastri*) (**Supplementary Table 5**). These assemblies
151 were enriched in VRs annotated as *Orthopolintovirales* and PLVs. In contrast, the proportion of nearly identical VRs
152 was below 20% in 400 assemblies, including the cases of the three bivalve species with over 10% VR content. These
153 results suggest that some of the identified VRs may have proliferated within the genome like transposons, but the
154 majority of VRs did not show high intragenomic similarity, thus lacking signals of recent proliferation.



155

156 **Fig. 2 | The proportion of identified VRs within different eukaryotic groups.**

157 **(a–d) metazoans, fungi, green plants and protists.** Each dot represents a genomic assembly, with its color and

158 shape corresponding to its taxonomic group. The number shown in each label of eukaryotic group indicates the total
159 number of assemblies (dots) included in the figure. Groups with fewer assemblies are plotted in the upper layer of
160 the plot to enhance their visibility. The horizontal axis represents the assembly size, while the vertical axis represents
161 the percentage of the genome occupied by VRs, both of which are log10-scaled.
162

163 **Expansion of known virus–eukaryote associations**

164 The viral taxonomy of VRs may reveal previously unrecognized virus–eukaryote associations. However,
165 taxonomic assignment may not be unequivocal when a VR harbors features from multiple viral lineages. To identify
166 VRs with homogeneous viral taxonomic signals and thus ensure reliable relationships between viral and eukaryotic
167 lineages, we applied more stringent criteria including the proportion of OGs conserved in a specific viral taxon,
168 uniform taxonomic signals in TOGs, and the eukaryotic assembly quality (see Methods for details). As a result, a
169 subset of VRs was defined as highly taxonomically confidence viral regions (HCVRs). This filtering yielded 50,058
170 HCVRs in metazoans (6.61% of all metazoan VRs), 381 in fungi (39.94%), 366 in green plants (14.99%), and 269
171 in protists (1.33%). An independent taxonomic assessment of the HCVRs using an extended viral genome database
172 provided consistent results (**Supplementary Fig. 2**), supporting the accuracy of our viral taxonomic annotations. Of
173 these, 49,530 HCVRs (96.98%) were classified as integration-like, whereas 1,544 (3.02%) were viral contig-like.

174 The HCVRs greatly expanded current knowledge of eukaryote–dsDNA virus relationships by uncovering 173
175 putative associations between viral major taxa and eukaryotic classes (or orders when the class-level taxa were
176 unavailable) (**Fig. 3**). These spanned more than 6 eukaryotic supergroups[56], 27 phyla and 79 classes. Most (n = 159,
177 91.91%) of these associations were supported by the presence of integration-like HCVRs. Of 43 associations
178 between viral major taxa and eukaryotic classes currently known from isolation studies, the HCVRs recovered 29
179 (67%) associations. Notably, most of the associations (n = 144, 83.24%) involved a viral taxon from which no virus
180 has yet been isolated from the paired eukaryotic taxon.

181 The quality-controlled catalog of VRs and HCVRs provides multifaceted starting points for exploring the
182 largely uncharacterized endogenous dsDNA virosphere. In the following sections, we illustrate examples of such
183 investigations using detailed comparative genomics and phylogenetic approaches.

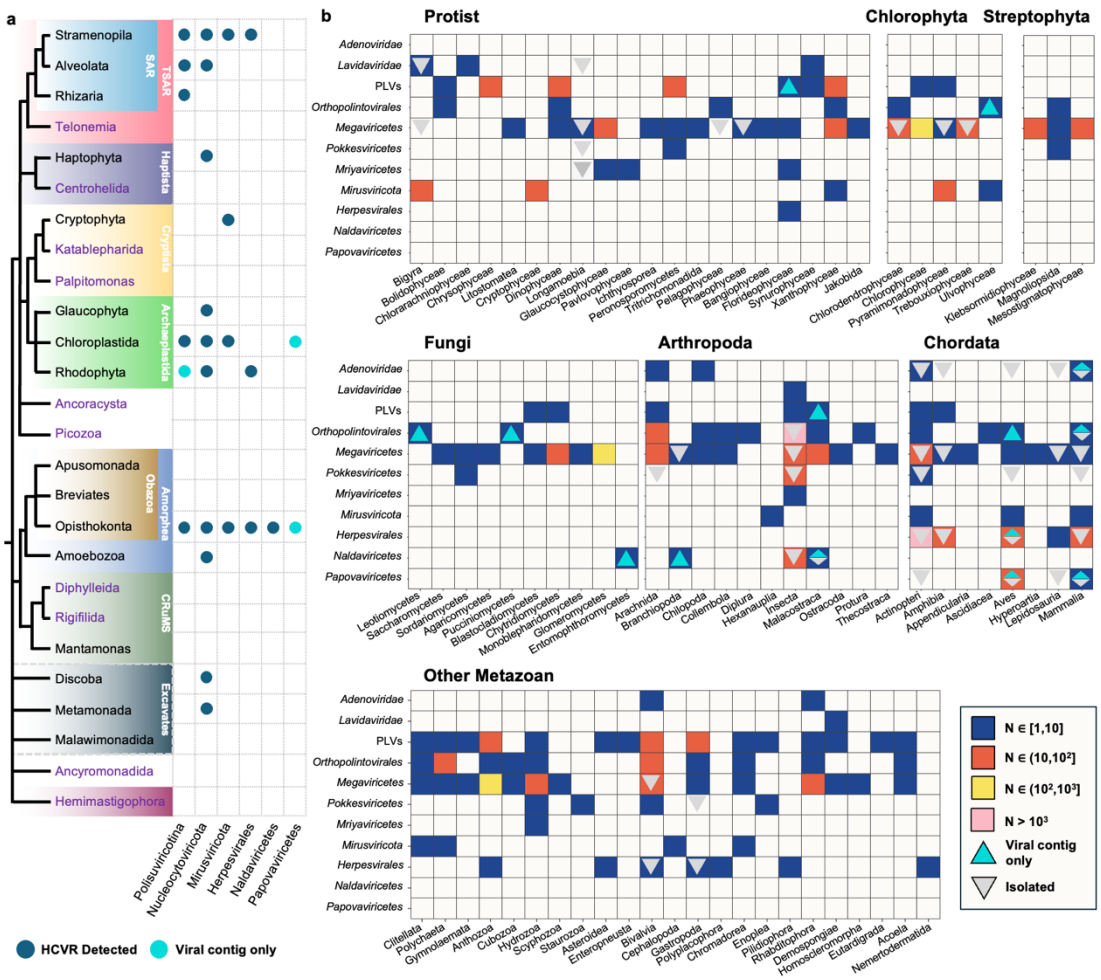


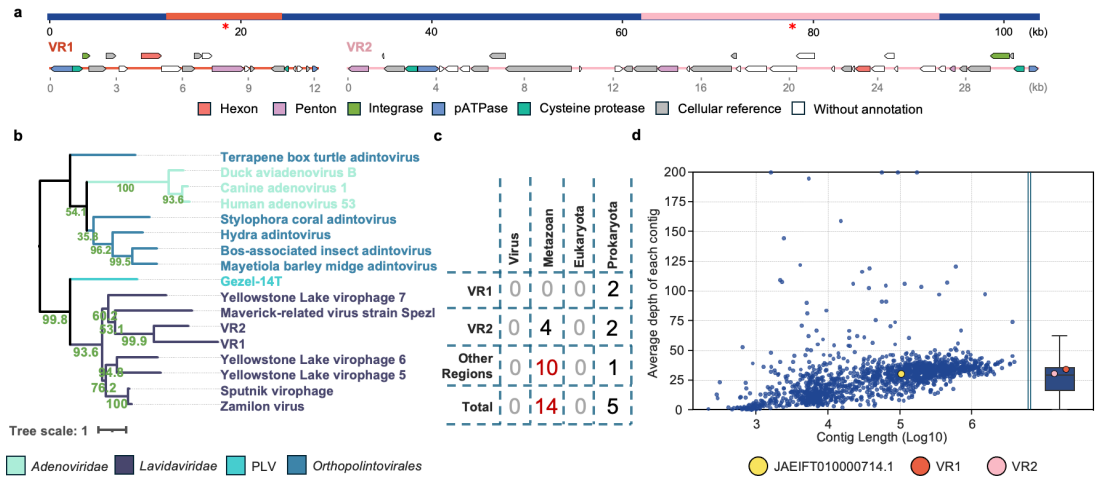
Fig. 3 | HCVR distribution in eukaryotes.

(a) The distribution of HCVRs across the eukaryotic tree of life[56]**.** *Polisuviricotina* includes *Adenoviridae*, *Lavidaviridae*, PLVs, and *Orthopolintovirales*. *Nucleocytoviricota* includes *Megaviricetes*, *Pokkesviricetes,* and *Mriyaviricetes*. Taxa shown in purple indicate lineages for which no genome assemblies are currently available in GenBank. **(b) The distribution of HCVRs in distinct eukaryotic groups.** In each heatmap, the horizontal axis represents the eukaryotic taxon, the vertical axis represents the viral taxon, and the color gradient indicates the number of HCVRs in each unit. The color of each eukaryotic taxon corresponds to its phylum, which is indicated alongside the heatmap. A light blue triangle marks units where all HCVRs are viral contig-like, while a gray triangle indicates the presence of viruses previously isolated from these eukaryotes.

Previously unrecognized virus–eukaryote associations

Even in well-studied metazoan lineages such as Arthropoda, Chordata, and Mollusca, we revealed numerous previously unrecognized putative associations (n = 88, 79.28% of all 111 detected virus–metazoan associations). One such association was the *Lavidaviridae*–insect relationship. Lavidavirids are satellite viruses that are only

199 known to infect protists together with their helper imitervirals (*Megaviricetes*), which are also only known to infect
200 protists. A recent study detected signals of PLVs in insect genomes[57], whereas no lavidavirid–metazoan relationship
201 has been described to date. In our study, two *Lavidaviridae* HCVRs were identified within a single contig from the
202 pomace fly *Lordiphosa mommai* (**Fig. 4a**). Both HCVRs contained core genes of *Lavidaviridae*, including those
203 encoding the hexon and penton capsid proteins, the packaging ATPase (pATPase), a cysteine protease, as well as an
204 integrase that may facilitate their integration into the eukaryotic genome. We reconstructed the phylogeny of the
205 hexon (**Fig. 4b**) and pATPase (**Supplementary Fig. 3**) proteins to infer their evolutionary relationships within
206 *Polisuviricotina* (*Lavidaviridae*, *Adenoviridae*, PLVs, and *Orthopolintovirales*). For both genes, the proteins from
207 the two HCVRs consistently clustered with *Lavidaviridae* members, confirming their taxonomic identity. The contig
208 harboring the HCVRs encoded 19 ORFs with similarities to sequences in the database. Of these, 14 showed best
209 matches to metazoan proteins and 5 to prokaryotic proteins (**Fig. 4c**), with most metazoan ORFs located outside the
210 VRs. We also examined the publicly available nanopore long reads to assess the depth of this assembly; both the
211 contig and the two HCVRs exhibited depth levels comparable to the genomic background (**Fig. 4d**). These results
212 confirm the eukaryotic origin of this contig and the integration of the HCVRs. We also found that a *Lavidaviridae*
213 HCVR integrated into a 14 Mb scaffold of *Ephydatia muelleri* (sponge), which also contains a VR classified in
214 *Megaviricetes* (**Supplementary Table 3**). These findings suggest the existence of diverse lavidavirids in metazoans,
215 possibly together with their helper viruses.



216

217 **Fig. 4 | The *Lavidaviridae* HCVRs from *L. mommai*.**

218 **(a) Two *Lavidaviridae* HCVRs within a contig.** The thick blue line represents the full contig; shaded regions
219 indicate HCVRs. Their magnified genome maps are displayed below the contig. ORFs in the magnified genome
220 maps are color-coded. **(b) Maximum-likelihood phylogenetic tree reconstructed using the hexon protein.** The
221 tree is midpoint-rooted, and ultrafast bootstrap support values are shown along the branches. The best-fit substitution

222 model was Q.pfam+F+I+R3. **(c) Taxonomic distribution of ORFs across different regions of the contig. (d)**
223 **Average depth of this assembly.** Each dot represents a contig; the vertical axis indicates the average depth, and the
224 horizontal axis indicates the contig length. The boxplot on the right shows the distribution of average depths across
225 all contigs.

226 Viruses of *Imitervirales* are known to infect protists. A previous study identified *Imitervirales* signals in a 400-
227 kb genomic scaffold of a cnidarian, *Hydra vulgaris*[52]. The entire length of the scaffold appeared to be part of a viral
228 genome, thus lacking evidence of integration. In the present study, we identified a cluster of *Imitervirales* VRs
229 localized on chromosome 9 (53 Mb, CM039220.1) of *H. vulgaris*. With the largest HCVR reaching 240 kb in length,
230 our results confirm the integration of imiterviral genomes in *Hydra* (**Supplementary Table 3**).

231 Another previously undescribed virus–host association was suggested for mirusviruses. The genome of a parrot
232 (*Agapornis swindernianus*) showed two HCVRs and one VR of *Mirusviricota* (**Supplementary Table 3**), a viral
233 lineage previously reported only from unicellular eukaryotes. The longest HCVR spans nearly 35 kilobases and
234 encodes two ORFs with best matches to metazoan proteins and one to a fungal protein (**Supplementary Fig. 4a**).
235 Structural prediction of the HK97 MCP revealed the highest TM-score (0.825) when compared with the mirusviral
236 MCP references reported previously[19] (**Supplementary Fig. 4b**), whereas the highest TM-scores for herpesviral and
237 caudoviral MCPs were 0.564 and 0.631, respectively. Phylogenetic analysis incorporating recently reported
238 sequences[58] further showed that this HK97 MCP does not cluster within any previously described putative orders
239 (**Supplementary Fig. 4c**). The average sequencing depth of the assembly was 60×, while the depth of the contigs
240 was 20×. Together with its short size, the contig harboring the mirusvirus signal may be part of episomal viral DNA.
241

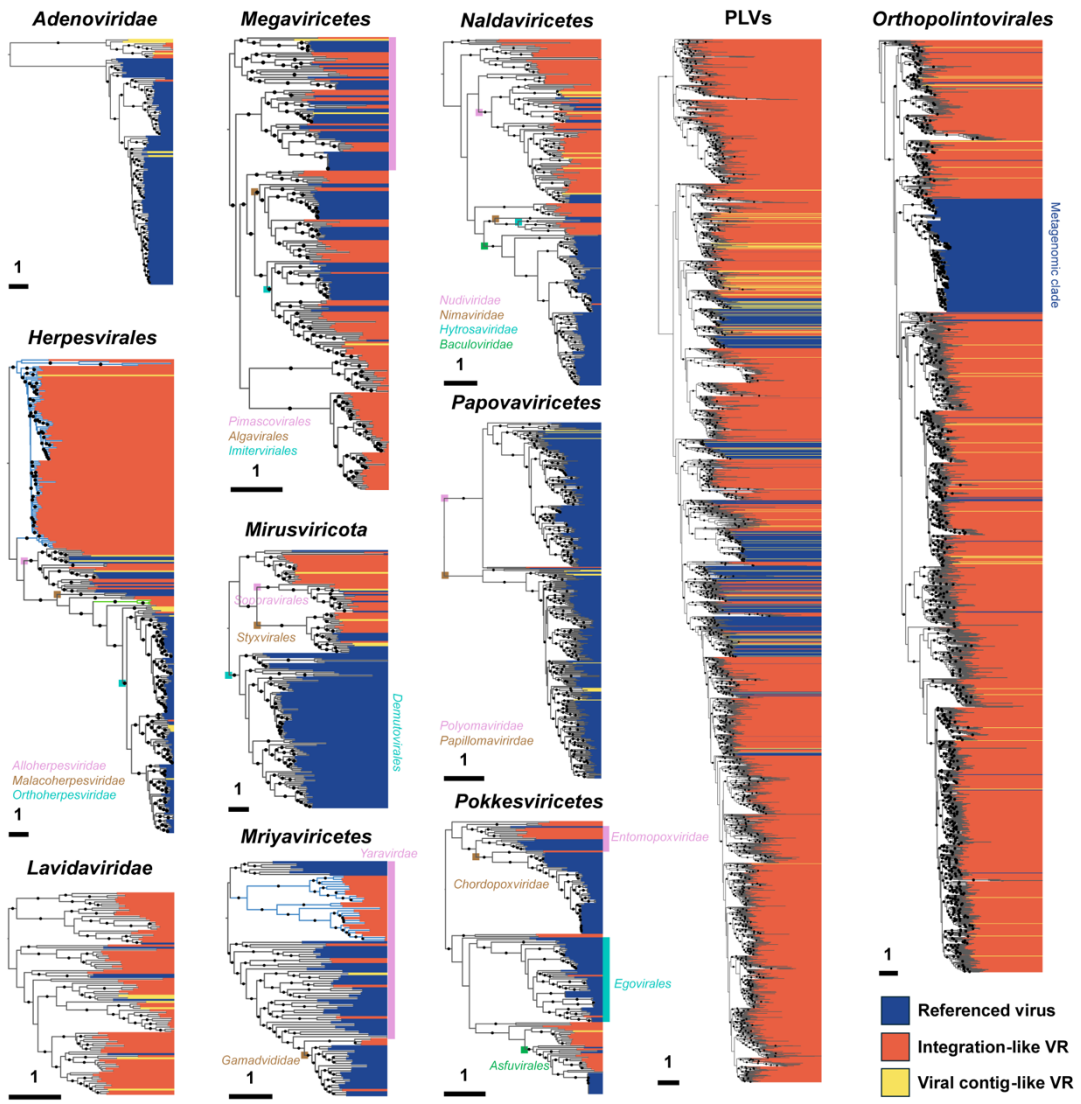
242 **Host interaction factors in high-confidence viral regions**

243 Genes in HCVRs were functionally annotated using COG, PFAM, and KEGG. The most prevalent COG
244 category was "Replication, recombination and repair", with 93.33% of HCVRs carrying at least one gene in this
245 category. Consistent results were obtained by PFAM analysis with the dominance of replication related functions,
246 such as *rve* (90.45%), endonuclease VII (76.69%), and *polB* (75.89%) (**Supplementary Fig. 5a**). These HCVRs
247 primarily originated from *Orthopolintovirales*, *Megaviricetes*, and *Herpesvirales*, all of which are known to encode
248 replication-associated genes essential for propagation of these viral lineages.

249 KEGG Orthologs (KOs) were assigned to 22,299 genes, which were distributed across 51 of the 56
250 KEGG/BRITE functional categories, indicating broad functional representation (**Supplementary Fig. 5b**). The three
251 most frequently detected KOs were GOLGB1 (n = 5,147; membrane traffic system), BIRC7_8 (n = 3,760), and

252 BIRC4 (n = 3,754; ubiquitin system) (**Supplementary Fig. 5c**). These genes were predominantly derived from
253 *Orthopolintovirales* HCVRs. GOLGB1 (i.e., Giantin) is a coiled-coil protein that regulates Golgi architecture and
254 function[59] and has been reported to interact with viral proteins and be involved in the formation of viral factories in
255 the case of RNA viruses[60,61]. BIRC7_8 and BIRC4 belong to the baculoviral IAP repeat-containing protein (BIRC)
256 family and interact with caspases to inhibit apoptosis in eukaryotic cells[62]. GOLGB1, BIRC7_8, and BIRC4 genes
257 identified in HCVRs were broadly distributed across multiple assemblies rather than being confined to a small
258 number of eukaryotic species. They were predominantly detected in insects, spanning 8, 18, and 17 orders,
259 respectively (**Supplementary Table 6**). Overall, HCVRs are enriched in core viral replication genes, and they also
260 contain host-interaction factors.
261
262 **Previously unrecognized viral lineages**
263 To investigate the evolutionary origins of the identified VRs, we reconstructed phylogenetic trees of VRs and
264 reference viral sequences using a representative marker gene conserved among reference genomes for each viral
265 major taxon (**Fig. 5, Supplementary Fig. 6, Supplementary Data**). For taxa with the DNA polymerase as the
266 phylogenetic marker, additional trees were constructed by including eukaryotic homologs (**Supplementary Fig. 7**)[63],
267 which supported the viral origin of these genes in the VRs.
268 The phylogenetic trees revealed that many of the VRs were closely related to reference viral genomes, although
269 several clades were enriched in VRs with few or no reference sequences. These clades may represent previously
270 unrecognized viral lineages. The *Alloherpesviridae* had previously been reported to integrate into the genomes of 15
271 fish species[64]. Our approach identified two VR-only clades (branches colored blue in **Fig. 5**), which were distantly
272 related to the *Alloherpesviridae* and comprised MCP marker genes from 72 fish species. An additional VR-only
273 clade (branches colored green in **Fig. 5**) was identified between the *Orthoherpesviridae* and the *Malacoherpesviridae*
274 clades. Orthoherpesvirids and malacoherpesvirids are known to infect vertebrates and molluscs, while the sequences
275 in the VR-only clade were derived from a ribbon worm genome (*Notospermus geniculatus*), thus expanding previous
276 observations[65]. Regarding *Mriyaviricetes*, a recently discovered viral clade known to infect amoebae and
277 haptophytes[32], we identified a novel VR-only clade (branches colored blue in **Fig. 5**). This clade comprises sequences
278 from 10 distinct eukaryotic species, including those from Cnidaria, Mollusca, Echinodermata, and
279 Glaucocystophyceae, suggesting a high host diversity for this group of viruses. The novelty of VR-only clades in
280 other viral taxa requires further investigation using more comprehensive datasets.
281 The phylogenetic distribution of identified integration-like VRs showed a bias across viral major taxa. There

282 were very few such signals in *Adenoviridae* and *Papovaviricetes*, whereas in other viral taxa, such as *Lavidaviridae*,
283 *Megaviricetes*, *Mriyaviricetes*, and PLVs, integration-like VRs were widespread. An imbalanced distribution of
284 integration-like VRs was also obvious within several viral taxa. For example, VRs were found concentrated in deep
285 branching lineages of *Herpesvirales*, including *Alloherpesviridae* and the above-mentioned related lineages. VRs
286 from *Mirusviricota* were not identified in the clade of viruses detected in marine ecosystems (the *Demutovirales*[58])
287 but mainly found in other orders. Regarding *Orthopolintovirales*, VRs were widely distributed across the tree, except
288 for one clade composed exclusively of reference sequences derived from aquatic environmental metagenomes.
289 Viruses with an integration life stage are known to appear sporadically during viral evolution (e.g., HHV-6 among
290 diverse herpesviruses[11]). The replication compartment (nucleus or cytoplasm) can also differ between the lineages
291 even within a virus group (e.g., mirusviruses[58]). The imbalanced phylogenetic distributions of VRs across viral
292 phylogenetic trees thus may be due to differences in replication strategy and host groups.



293

294 **Fig. 5 | Reconstructed phylogenetic trees of VRs and viruses.**

295 Each tree corresponds to a specific viral taxon, as indicated above. Scale bars are shown in the lower left corners,
296 and all trees are midpoint-rooted. Black circles indicate bootstrap support values below 90. Blue shading indicates
297 reference viral sequences, orange integration-like VRs, and yellow viral contig-like VRs. For *Adenoviridae*,
298 *Herpesvirales*, *Megaviricetes*, *Naldaviricetes*, and *Pokkesviricetes*, DNA polymerase sequences were used for tree
299 reconstruction. For *Lavidaviridae*, *Mirusviricota*, *Mriyaviricetes*, and *Orthopolintovirales*, the major capsid protein
300 sequences were used. For *Papovaviricetes* and PLVs, E1 helicase and pATPase were used, respectively. Trees for
301 *Orthopolintovirales* and *Megaviricetes* were manually curated.
302

303 **Discussion**

304 The computational framework presented in this work detected dsDNA VRs in 19% of the analyzed eukaryotic
305 genome assemblies, with substantial variations in VR content across organisms. A notable observation was the high
306 abundance of VRs in some invertebrates such as molluscs, arthropods, and cnidarians. VRs accounted for more than
307 0.1% up to 16% of the genome in 1,920 metazoan assemblies from 892 genera, suggesting a role of dsDNA viruses
308 in shaping these animal genomes (**Fig. 2, Supplementary Table 1**). Phylogenetic trees (**Fig. 5**) of viral and VR
309 sequences revealed an obvious separation between the currently observable "virosphere" and "endo-virosphere"
310 based on genomic data. These results suggest that some viruses have a greater propensity to leave traces in eukaryotic
311 genomes, whereas others do not. Our study revealed many previously unrecognized associations (n = 144) between
312 viral major taxa and eukaryotic classes, whereas only 43 associations are currently supported by evidence from virus
313 isolation (**Fig. 3**). We conducted several targeted bioinformatics analyses to investigate selected cases, such as the
314 lavidavirid–insect association. Similar approaches can be applied to further explore VRs and candidate associations.
315 It has become increasingly common to detect entirely novel viral signals in sequence data prior to virus isolation or
316 microscopic observation[14,15,53]. In this context, the putative associations and viral signals presented here may serve
317 as a gateway to currently hidden areas of the virosphere.

318 As eukaryotes diversified in lifestyle[66], cellular architecture, and antiviral defenses[67,68], their associated viral
319 lineages, as well as the abundance of endogenized viral elements, are expected to be diverse. In land plants and fungi,
320 it has been suggested that structural features such as cell walls and plasmodesmata act as barriers against large
321 dsDNA viruses[69], which may explain the observed scarcity of dsDNA viral signals. Protists, which encompass most
322 of the major eukaryotic supergroups[56], exhibit substantial viral diversity despite representing the group with the
323 fewest assemblies in our dataset (**Fig. 1b and 1c**). This finding underscores the vast, largely unexplored diversity of
324 the protist virosphere and highlights the need for further studies on protist–virus interactions and antiviral

325 mechanisms. In metazoans, vertebrates in particular, diverse and complex antiviral immune systems have evolved,
326 encompassing both innate and adaptive immunity[70,71]. These defenses may limit dsDNA viral infections and purge
327 endogenous viral elements[72], providing a plausible explanation for the relatively low density of dsDNA viral
328 sequences observed in vertebrate genomes. In contrast, invertebrates lack adaptive immunity and maintain extensive
329 interactions with symbionts and environmental microbes[73], factors that may collectively account for the high
330 proportions of dsDNA viral elements observed in their genomes (**Fig. 2**). Host ecological and immunological
331 characteristics may have shaped the evolutionary trajectories and diversity of endogenous dsDNA viruses.

332 The present study was limited to reference genomes from only 11 major viral taxa. As more viral groups are
333 discovered and characterized, additional genomic regions may be recognized as being of viral origin. Furthermore,
334 resolving virus–host taxonomic relationships at a finer taxonomic scale will require additional analytical efforts.
335 While these represent limitations of the present study, the catalog of VRs that we provide here should serve as a
336 valuable resource and foundation for future investigations of viral diversity and virus–host coevolutionary processes.
337

338 ## Methods

339 **Construction of cellular reference HMMs**

340 Cellular reference HMMs were constructed in two sequential steps. First, protein sequences were downloaded
341 from the KEGG database[74], covering 778 eukaryotic species from different genera (one species per genus).
342 Sequences shorter than 50 amino acids or longer than 4,000 amino acids were removed. OGs were identified using
343 OrthoFinder v2.5.5[75]. OGs present in at least 60% of genomes within at least one of the four major eukaryotic groups
344 (animals, green plants, fungi, or protists) were retained, yielding 12,555 OGs. For each OG, up to 200 sequences
345 were randomly selected and aligned using Muscle v5.1[76]. Alignments were trimmed with trimAl v1.4.1[77] using the
346 '-gt 0.05' option, and trimmed alignments were used to construct HMMs with HMMER v3.4[78].

347 The second step was based on HMMs from Pfam-A v37[79]. To minimize potential viral contamination, HMMs
348 containing virus-related keywords (i.e., 'virus', 'viral', 'phage', 'capsid', 'double jelly roll', 'terminase') were
349 excluded. The remaining HMMs were then evaluated using hmmscan (E-value < 10^{-3}) against both the collected
350 viral proteins (**Supplementary Table 2**) and eukaryotic proteins from KEGG. HMMs that produced more hits to
351 viral sequences than to eukaryotic sequences were removed, leaving 19,991 HMMs. In total, 32,546 HMMs from
352 both steps were retained as cellular reference models.
353

354 **Construction of VOGs and TOGs**

355 We collected reference genomes of the 11 eukaryotic dsDNA viral major taxa from various sources. For six
356 viral taxa (*Adenoviridae*, *Herpesvirales*, *Lavidaviridae*, *Megaviricetes*, *Naldaviricetes*, and *Papovaviricetes*), we
357 obtained protein sequences from RefSeq genomes in NCBI Virus (as of June 10, 2024). For *Mirusviricota*,
358 *Mriyaviricetes*, PLVs, *Pokkesviricetes*, and *Orthopolintovirales*, we additionally gathered genomic data from other
359 sources (see details in Supplementary Table 2). Gene prediction was performed for large/giant viruses using Prodigal
360 v2.6.3[80] with default options and for smaller viruses using GeneMarkS v4.30[81] with the metagenomic model and
361 other default options. For each viral taxon, we identified OGs using Proteinortho v6.3.2[82] with parameters following
362 a previously described approach[83] (-e = 1e-3 -identity = 15 -p = blastp+ -selfblast -cov = 40). For each taxon, we
363 constructed viral HMMs based on OGs shared by at least three genomes using the same methodology as described
364 in the previous section.
365 To identify OGs exhibiting viral specificity (i.e., VOGs), we performed hmmsearch using both cellular
366 reference HMMs and viral HMMs against proteins from all OGs identified in viral genomes. An OG was classified
367 as a VOG if at least one protein within the OG showed an E-value smaller than 10^{-5} against viral HMMs. For all
368 proteins in this OG, the minimum E-value for the cellular reference HMM had to exceed 10^{-5} and be at least 10^{3}
369 times greater than that of the corresponding viral HMM, ensuring viral specificity. The classification of TOGs was
370 based on VOGs. Proteins within a TOG were required to exhibit the best HMM match to viral HMMs from their
371 own taxon, while the minimum E-values for HMMs from other viral major taxa had to exceed 10^{-5} and be at least
372 10^{3} times higher than those of their own taxon, thereby confirming taxonomic specificity.
373
374 **Detection of relationships between viral major taxa and eukaryotic contigs**
375 We first retrieved eukaryotic taxonomic information using the Bio.Entrez module of Biopython[84], followed by
376 manual curation. For each eukaryotic assembly, Prodigal (with configurations adjusted for long input sequences)
377 and GeneMarkS were used to identify ORFs. The existence of introns was not considered here, as the goal of this
378 study is to identify viral-like regions. Thus, the ORF density of a eukaryotic assembly does not reflect the true coding
379 density. We then performed hmmsearch to identify ORFs with significant similarity to VOGs. To inclusively capture
380 candidate contigs in the initial screening step, an E-value threshold of 10^{-3} was applied. Because some ORFs may
381 match multiple VOGs, only the best hit was considered in each case.
382 We conducted a secondary hmmsearch using the -Z parameter to standardize E-value calculations across
383 searches, with database sizes specified for each eukaryotic group (Metazoa: 4,962,163; Fungi: 193,281;
384 Viridiplantae: 237,367; Protists: 317,655). If an ORF had a hit to a cellular model with an E-value less than 100

385 times that of its best VOG match, the hit was removed as a potential cellular sequence. The remaining ORFs with
386 VOG hits were considered viral-like ORFs. A virus–contig pair for a specific viral major taxon was recorded if the
387 contig contained either (i) at least one viral-like ORF with a TOG hit or (ii) two viral-like ORFs with hits to distinct
388 VOGs from the viral major taxon. If a virus–contig pair was detected based on the ORFs predicted by either Prodigal
389 or GeneMarkS, the corresponding ORF set was used for downstream analyses. When the same virus–contig pair was
390 detected based on ORFs predicted by both methods, only one ORF set was retained. The selected method was the
391 one showing better overall performance, defined as detecting a greater number of contigs specifically associated
392 with the corresponding viral major taxon and eukaryotic kingdom.
393
394 **Determination of viral regions**
395 We performed hmmsearch on all ORFs within each contig against OGs from the corresponding viral taxon and
396 cellular references (E-value < 10^{-3}) to obtain cellular and viral annotations. VRs were identified on the basis of these
397 annotations and the genomic locations of ORFs. A region was classified as a VR if it contained multiple distinct viral
398 hits originating from a single viral taxon. Specifically, at least two distinct viral hits were required for *Adenoviridae*,
399 *Lavidaviridae*, *Mriyaviricetes*, PLVs, *Papovaviricetes*, and *Orthopolintovirales*, whereas at least three distinct viral
400 hits were required for *Herpesvirales*, *Megaviricetes*, *Mirusviricota*, *Naldaviricetes*, and *Pokkesviricetes*. In both
401 cases, all viral hits had to originate from the dominant viral major taxon within the region. Additionally, no more
402 than 10 ORFs with cellular annotations were allowed between two viral ORFs, and the genomic distance between
403 them could not exceed 10 kb. The length of each VR was determined from the start of the first ORF with a viral best
404 hit to the end of the last ORF with a viral best hit.
405
406 **Annotation of viral regions**
407 Viral–cellular annotation tables were generated for each VR based on the minimum E-value comparison
408 between HMM hits from the cellular reference and OGs within the corresponding viral taxon. Functional annotation
409 of VRs was performed using eggNOG-mapper v2.1.12[85]. Annotation quality was assessed by manual inspection of
410 assigned COG categories. Genes were defined as having unknown function if they were assigned to COG categories
411 '–', 'R', or 'S', or if no hits were reported by eggNOG-mapper. The frequency of each functional label (i.e., COG
412 category, PFAM, and KEGG KO) was calculated at the region level by first scoring the presence or absence of the
413 label in each VR and then computing the average proportion of regions containing that label. For proteins associated
414 with multiple hits (e.g., multiple PFAM domains), all identified domains were counted. KEGG annotations were

415 retrieved from the KEGG Orthology database (https://www.genome.jp/kegg/ko.html) on October 22, 2025.

416

417 **Redundancy analysis of VRs in individual assemblies**

418 We extracted nucleotide sequences of VRs from each assembly and clustered them using MMseqs2 v16-
419 747c6[86] under different thresholds for minimum sequence identity and alignment coverage. Nearly identical VRs
420 were defined using sequence identity and coverage thresholds both set at 0.9.

421

422 **TNF composition analysis**

423 For each assembly containing identified VRs, the assembly was divided into viral and cellular components. If
424 the total length of VRs exceeded half the length of the contig on which they were located, the contig was classified
425 as viral. In other cases, the genomic regions adjacent to the VRs within the same contig were treated as cellular
426 components. These regions were combined with contigs lacking VRs to calculate the average cellular TNF vector
427 for each assembly. From the cellular regions, 1,000 random fragments of 3,000 bp were sampled, and their Pearson
428 correlation coefficients with the average cellular TNF vector were calculated to establish an assembly-specific
429 baseline. Pearson correlation coefficients were then calculated between the TNF vector of each VR and the average
430 cellular TNF vector to assess whether the nucleotide composition of VRs differed from that of the cellular
431 components, as described in the main text and Supplementary Fig. 1f.

432

433 **Classification of viral regions**

434 Viral contig-like VRs were defined based on two criteria (**Fig. 1a**). First, their total length was required to
435 exceed half the length of the corresponding contig. Second, the proportion of annotated ORFs assigned to cellular
436 OGs was required to be less than 20%. VRs that did not simultaneously satisfy both conditions were considered to
437 be potentially integrated into host genomes and were thus classified as integration-like VRs.

438 To classify HCVRs, we first identified conserved OGs for each viral taxon. Redundant reference viral genomes
439 were removed using vclust v1.3.1[87] (--metric ani --ani 0.95). An OG was considered conserved if it was present in
440 more than half of the filtered reference genomes of the corresponding taxon. We then calculated the acceptable range
441 for the proportion of conserved OGs based on the interquartile range (IQR; Q1 − 1.5×IQR to Q3 + 1.5×IQR). An
442 HCVR was required to contain at least one conserved OG, and its proportion of conserved OGs had to fall within
443 the calculated range.

444 We also evaluated the completeness of each assembly using BUSCO v5.1.2[88], applying metazoa_odb10,

445 fungi_odb10, viridiplantae_odb10, or eukaryota_odb10 datasets for metazoan, fungal, green plants, and protist
446 assemblies, respectively. VRs from assemblies with completeness scores below 30% were excluded from HCVR
447 classification. Additionally, we calculated the rate of unique hits for each VR using our annotation tables; VRs in
448 which fewer than 50% of hits were unique were excluded.

449 Finally, an HCVR was required to meet two additional conditions: (i) at least 20% of annotated ORFs within
450 the region had best hits to models constructed from viral sequences, with ≥ 90% of these hits assigned to the same
451 viral major taxon; and (ii) the region contained at least one TOG from its assigned viral major taxon and no TOGs
452 from other taxa.

453

454 **Independent taxonomic assessment of the HCVRs**

455 We performed DIAMOND blastp (v2.1.9, E-value < 10^{-3})[89] to compare all ORFs from the HCVRs against a
456 custom viral reference database and recorded the dominant taxon of the best hits. The custom database was built
457 from the reference viral genomes used in this study and complemented with high-confidence sequences from
458 corresponding taxa in the IMG/VR v7.1 database[90]. Notably, at the time when this version of the IMG/VR database
459 was compiled, several viral taxa (e.g., *Mirusviricota*, *Mriyaviricetes*, and PLVs) had not yet been formally
460 established. Therefore, TOGs from these taxa were used to identify potentially related contaminant sequences in the
461 database. Specifically, some sequences likely belonging to *Mirusviricota* or *Mriyaviricetes* were annotated as
462 *Nucleocytoviricota* in the database, whereas PLV-related sequences were annotated as *Polisuviricotina*. These
463 sequences were removed from the database prior to analysis.

464

465 **Distribution statistics of HCVRs in distinct eukaryotic groups**

466 The information on virus–host associations in specific eukaryotic lineages was obtained from the Virus–Host
467 database (https://www.genome.jp/virushostdb). All associations supported solely by viral contigs were manually
468 examined. As a result, the *Papovaviricetes*–Magnoliopsida association was removed, as the only supporting evidence
469 was a short viral contig (CBTL0109906317.1) showing clear similarity to human papillomavirus, whereas the
470 corresponding host assembly (GCA_900067735.1) was annotated as "contaminated" in GenBank.

471

472 **Detailed analysis of viral regions from *L. mommai***

473 We collected hexon gene sequences from the NCBI database, with the following accession numbers:
474 DAC81274.1, AP_000059.1, QKW89985.1, UYE94489.1, BBW89580.1, DAC81566.1, YP_010796969.1,

475 YP_004300296.1, YP_002122381.1, DAC81730.1, YP_010796952.1, YP_009177804.1, YP_009177819.1,
476 YP_009177692.1 and YP_008859635.1. Multiple sequence alignments were generated using Muscle, and the
477 alignments were trimmed with trimAl using a gap threshold of 0.1. Phylogenetic trees were reconstructed using IQ-
478 TREE v2.2.0[91] with the parameters "-m MFP -B 1000 -alrt 1000 -wbtl".

479 Alignment-based taxonomic annotation of ORFs was performed using DIAMOND (E-value < 0.1) against the
480 RefSeq database (downloaded on 2024-12-02). The taxonomic identity of the best hit for each ORF was recorded.

481 For depth analysis, we downloaded the dataset SRR13070751 from NCBI using the "prefetch" command of
482 sratoolkit v3.0.9 (https://github.com/ncbi/sra-tools) and converted it to FASTQ format with the "fastq-dump"
483 command. Reads were mapped to the assembly (GCA_018904225.1) using the "mem" command of BWA v0.7.17[92].
484 The average depth of each contig was then calculated with samtools v1.18[93].

485

486 **Detailed analysis of HK97 MCP from *A. swindernianus***

487 Alignment-based taxonomic annotation, phylogenetic reconstruction, and depth analysis were conducted using
488 the same approach as described for *L. mommai.* The dataset used for depth analysis was SRR25700359. The structure
489 of the MCP was predicted with AlphaFold v2.3.2[94] and visualized using ChimeraX v1.7[95]. Structural similarity
490 between HK97 MCPs was assessed using Foldseek v6-29e2557[97] with the easy-search module.

491

492 **Phylogenetic analysis of discovered VRs and reference viruses**

493 For each viral taxon, two marker genes were initially selected as candidates for phylogenetic analysis. One
494 seed sequence was collected for each marker gene from public reference datasets. The selected seed sequences are
495 listed as follows, with different functional genes chosen depending on the lineage. *Adenoviridae*: YP_002213842.1
496 (DNA polymerase) and AP_000211.1 (hexon); *Herpesvirales*: YP_010084960.1 (DNA polymerase) and
497 YP_010084972.1 (capsid protein); *Lavidaviridae*: YP_002122370.1 (cysteine protease) and YP_002122381.1
498 (capsid protein); *Megaviricetes*: YP_003986825.1 (DNA polymerase) and YP_003986929.1 (capsid protein);
499 *Mirusviricota*: TARA_AON_NCLDV_00056_000000000009 (DNA polymerase) and
500 TARA_AON_NCLDV_00032_000000000015 (capsid protein); *Mriyaviricetes*: HUH_long_consensus (HUH
501 endonuclease) and Yaravirus_MCP_consensus (capsid protein); *Naldaviricetes*: YP_009116659.1 (DNA
502 polymerase) and YP_009116734.1 (capsid protein); *Papovaviricetes*: NP_041327.2 (E1 helicase) and NP_041332.2
503 (L1 capsid protein); PLVs: SAF1_17 (pATPase) and SAF1_16 (capsid protein); *Pokkesviricetes*: YP_009162448.1
504 (DNA polymerase) and YP_009162470.1 (capsid protein); *Orthopolintovirales*: YP_010796957.1 (DNA

505 polymerase) and Polinton-1_DY (capsid protein).

506 We then compared the coverage of seed sequence alignments against the reference genome dataset. Five
507 approaches were used: (i) BLASTP (BLAST+ v2.16.0[96]) with an E-value cut-off of 1e-5; (ii) PSI-BLAST[97], where
508 all reference genes within a given lineage were pooled into one database and searched against corresponding seed
509 sequences with five iterations using an E-value cut-off of 1e-5; (iii) HMMER, using HMM profiles built from initial
510 reference OGs to identify marker genes, applying hmmscan to the reference sequences; (iv) JackHMMER[98], where
511 all reference genes within a lineage were pooled into one database and searched against corresponding seed
512 sequences with five iterations; and (v) MMseqs2, where all seed and reference protein sequences were pooled and
513 clustered using the cluster function with a minimum sequence identity of 0.2 and a coverage threshold of 0.5. The
514 coverage of each marker gene on reference genomes is shown in Supplementary Fig. 6.

515 Owing to its superior sensitivity, JackHMMER was used to identify all marker genes in the viral regions using
516 the same parameters and a database comprising VR protein sequences. The proportions of VRs encoding marker
517 genes were as follows: DNA polymerase for *Adenoviridae* (15%), *Herpesvirales* (21%), *Megaviricetes* (5%),
518 *Naldaviricetes* (28%), and *Pokkesviricetes* (2%); major capsid protein for *Lavidaviridae* (87%), *Mirusviricota* (10%),
519 and *Mriyaviricetes* (10%); E1 helicase for *Papovaviricetes* (100%); and pATPase for PLVs (86%). Redundant
520 sequences from viral regions were removed using CD-HIT v4.8.1[99]. Most lineages were filtered using a minimum
521 sequence identity of 0.8 and an alignment coverage threshold of 0.8 for the shorter sequence. For *Orthopolintovirales*,
522 identity and coverage thresholds were set to 0.6; for PLVs, more relaxed thresholds of 0.4 were applied due to the
523 large number of detected sequences. These lineage-specific thresholds ensured sufficient representation of
524 homologous sequences and supported reliable downstream phylogenetic analyses.

525 The dereplicated viral region sequences were then combined with reference sequences. Multiple sequence
526 alignments were performed using Clustal Omega v1.2.4[100] and trimmed with trimAl using a gap threshold of 0.05.
527 Aligned sequences with gaps as the predominant (90%) sites were removed. Phylogenetic trees were reconstructed
528 using IQ-TREE with the -m TEST option, allowing automatic model selection based on the Bayesian Information
529 Criterion. The selected models for each viral lineage were as follows: *Adenoviridae*: pfam+F+I+G4; *Herpesvirales*:
530 JTT+F+G4; *Lavidaviridae*: Blosum62+F+G4; *Megaviricetes*: VT+F+G4; *Mirusviricota*: Q.pfam+F+I+G4;
531 *Mriyaviricetes*: VT+F+G4; *Naldaviricetes*: VT+F+I+G4; *Papovaviricetes*: Q.yeast+F+I+G4; PLVs: VT+F+G4;
532 *Pokkesviricetes*: VT+F+I+G4; and *Orthopolintovirales*: VT+F+I+G4.

533 An additional phylogenetic analysis was independently performed for the HCVR dataset. JackHMMER was
534 used to identify all marker genes in HCVR sequences. Sequence alignment and trimming were performed using the

same procedures described above. Phylogenetic trees were reconstructed with IQ-TREE using the -m TEST option to enable automatic model selection. The selected substitution models for each viral lineage were as follows: *Adenoviridae*: Q.pfam+F+I+G4; *Herpesvirales*: Q.pfam+F+I+G4; *Lavidaviridae*: Q.pfam+F+G4; *Megaviricetes*: VT+F+G4; *Mirusviricota*: WAG+F+I+G4; *Mriyaviricetes*: Blosum62+F+G4; *Naldaviricetes*: Q.yeast+F+I+G4; *Papovaviricetes*: Q.yeast+F+I+G4; PLVs: Q.pfam+F+G4; *Pokkesviricetes*: VT+F+I+G4; and *Orthopolintovirales*: VT+F+G4. The phylogenetic trees were visualized with iTOL v7[101].

Acknowledgements

This work was supported by JSPS/KAKENHI (nos. 25KJ1513 and 22H00384), iJURC (2024-27) and HFSP (Ref. No.: RGP011/2024). Computational time was provided by the Supercomputer System, Institute for Chemical Research, Kyoto University (Kyoto, Japan). The authors thank Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript. The authors thank Dr. Kenji K. Kojima from Genetic Information Research Institute for his valuable suggestions.

Data Availability

HMM files of VOG/TOG and their annotations, amino acid sequences and annotation files of VRs, multiple sequence alignments for phylogenetic analysis, tree files and annotation files for visualization via iTOL have been deposited in Zenodo (https://doi.org/10.5281/zenodo.18919361).

Author Contributions

HZ designed the study, performed most of the analyses and wrote the first draft of the manuscript. LM contributed to the phylogenetic analyses. RZ contributed to the gene function analyses. MG contributed to the investigation of individual VRs. HO supervised the study and contributed to the design of the bioinformatics pipeline. All authors have read and approved the final version of the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

1. Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. *13*, 283–296. https://doi.org/10.1038/nrg3199.

2. Barreat, J.G.N., and Katzourakis, A. (2022). Paleovirology of the DNA viruses of eukaryotes. Trends Microbiol. *30*, 281–292. https://doi.org/10.1016/j.tim.2021.07.004.

3. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

4. Weiss, R.A. (2006). The discovery of endogenous retroviruses. Retrovirology *3*, 67. https://doi.org/10.1186/1742-4690-3-67.

5. Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., et al. (2018). The evolutionary history of vertebrate RNA viruses. Nature *556*, 197–202. https://doi.org/10.1038/s41586-018-0012-7.

6. Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., et al. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature *403*, 785–789. https://doi.org/10.1038/35001608.

7. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. https://www.science.org/doi/10.1126/science.aad5497

8. Ashley, J., Cordy, B., Lucia, D., Fradkin, L.G., Budnik, V., and Thomson, T. (2018). Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell *172*, 262-274.e11. https://doi.org/10.1016/j.cell.2017.12.022.

9. Ritsch, M., Brait, N., Harvey, E., Marz, M., and Lequime, S. (2024). Endogenous viral elements: insights into data availability and accessibility. Virus Evol. *10*, veae099. https://doi.org/10.1093/ve/veae099.

10. Moniruzzaman, M., and Aylward, F.O. (2023). Endogenous DNA viruses take center stage in eukaryotic genome evolution. Proc. Natl. Acad. Sci. USA. *120*, e2305212120. https://doi.org/10.1073/pnas.2305212120.

11. Morissette, G., and Flamand, L. (2010). Herpesviruses and Chromosomal Integration. J. Virol. *84*, 12100–12109. https://doi.org/10.1128/jvi.01169-10.

12. Fischer, M.G., and Hackl, T. (2016). Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature *540*, 288–291. https://doi.org/10.1038/nature20593.

13. Delaroque, N., Maier, I., Knippers, R., and Müller, D.G. (1999). Persistent virus integration into the genome of its algal host, Ectocarpus siliculosus (Phaeophyceae). J. Gen. Virol. *80*, 1367–1370. https://doi.org/10.1099/0022-1317-80-6-1367.

14. Chung, D., Brask, N., Matar, S., Gallot-Lavallée, L., Pringle, E.S., Duguay, B.A., Blais, C., Latimer, J., Haro, R., Slamovits, C.H., et al. (2025). Persistent mirusvirus infection in the marine protist Aurantiochytrium. Nat.

598 Commun. *16*, 9922. https://doi.org/10.1038/s41467-025-65172-6.

599 15. Erazo-Garcia, M.P., Sheyn, U., Barth, Z.K., Craig, R.J., Wessman, P., Jivaji, A.M., Ray, W.K., Svensson-Coelho,
600 M., Cornwallis, C.K., Rengefors, K., et al. (2025). Cryptic infection of a giant virus in a unicellular green alga.
601 Science *388*, eads6303. https://doi.org/10.1126/science.ads6303.

602 16. van Oers, M.M., Herniou, E.A., Jehle, J.A., Krell, P.J., Abd-Alla, A.M.M., Ribeiro, B.M., Theilmann, D.A., Hu,
603 Z., and Harrison, R.L. (2023). Developments in the classification and nomenclature of arthropod-infecting large
604 DNA viruses that contain pif genes. Arch. Virol. *168*, 182. https://doi.org/10.1007/s00705-023-05793-8.

605 17. Bézier, A., Annaheim, M., Herbinière, J., Wetterwald, C., Gyapay, G., Bernard-Samain, S., Wincker, P., Roditi,
606 I., Heller, M., Belghazi, M., et al. (2009). Polydnaviruses of Braconid Wasps Derive from an Ancestral
607 Nudivirus. Science *323*, 926–930. https://doi.org/10.1126/science.1166788.

608 18. Bellas, C., Hackl, T., Plakolb, M.-S., Koslová, A., Fischer, M.G., and Sommaruga, R. (2023). Large-scale
609 invasion of unicellular eukaryotic genomes by integrating DNA viruses. Proc. Natl. Acad. Sci. USA. *120*,
610 e2300465120. https://doi.org/10.1073/pnas.2300465120.

611 19. Zhao, H., Meng, L., Hikida, H., and Ogata, H. (2024). Eukaryotic genomic data uncover an extensive host
612 range of mirusviruses. Curr. Biol. *34*, 2633-2643.e3. https://doi.org/10.1016/j.cub.2024.04.085.

613 20. Irwin, N.A.T., Pittis, A.A., Richards, T.A., and Keeling, P.J. (2022). Systematic evaluation of horizontal gene
614 transfer between eukaryotes and viruses. Nat. Microbiol. *7*, 327–336. https://doi.org/10.1038/s41564-021-
615 01026-3.

616 21. Gallot-Lavallée, L., and Blanc, G. (2017). A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity
617 through the Eukaryotic Genomics Window. Viruses *9*, 17. https://doi.org/10.3390/v9010017.

618 22. Haapa-Paananen, S., Wahlberg, N., and Savilahti, H. (2014). Phylogenetic analysis of Maverick/Polinton giant
619 transposons across organisms. Mol. Phylogenet. Evol. *78*, 271–274.
620 https://doi.org/10.1016/j.ympev.2014.05.024.

621 23. Krupovic, M., and Koonin, E.V. (2015). Polintons: a hotbed of eukaryotic virus, transposon and plasmid
622 evolution. Nat. Rev. Microbiol. *13*, 105–115. https://doi.org/10.1038/nrmicro3389.

623 24. Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., and Aylward, F.O. (2020). Widespread
624 endogenization of giant viruses shapes genomes of green algae. Nature *588*, 141–145.
625 https://doi.org/10.1038/s41586-020-2924-2.

626 25. de Villiers, E.-M., Fauquet, C., Broker, T.R., Bernard, H.-U., and zur Hausen, H. (2004). Classification of
627 papillomaviruses. Virology *324*, 17–27. https://doi.org/10.1016/j.virol.2004.03.033.

628 26. Fischer, M.G. (2021). The Virophage Family Lavidaviridae. Curr. Issues Mol. Biol. *40*, 1–24.
629 https://doi.org/10.21775/cimb.040.001.

630 27. Koonin, E.V., Fischer, M.G., Kuhn, J.H., and Krupovic, M. (2024). The polinton-like supergroup of viruses:
631 evolution, molecular biology, and taxonomy. Microbiol. Mol. Biol. Rev. *88*, e00086-23.

https://doi.org/10.1128/mmbr.00086-23.

28. Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J.-M. (2004). The 1.2-Megabase Genome Sequence of Mimivirus. Science *306*, 1344–1350. https://doi.org/10.1126/science.1101485.

29. Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., et al. (2020). Giant virus diversity and host interactions through global metagenomics. Nature *578*, 432–436. https://doi.org/10.1038/s41586-020-1957-x.

30. Bellas, C.M., and Sommaruga, R. (2021). Polinton-like viruses are abundant in aquatic ecosystems. Microbiome *9*, 13. https://doi.org/10.1186/s40168-020-00956-0.

31. Gaïa, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A., Jaillon, O., Wincker, P., Ogata, H., Krupovic, M., et al. (2023). Mirusviruses link herpesviruses to giant viruses. Nature *616*, 783–789. https://doi.org/10.1038/s41586-023-05962-4.

32. Yutin, N., Mutz, P., Krupovic, M., and Koonin, E.V. (2024). Mriyaviruses: small relatives of giant viruses. mBio *15*, e01035-24. https://doi.org/10.1128/mbio.01035-24.

33. Gaia, M., Ruscheweyh, H.-J., Eren, A.M., Koonin, E.V., Sunagawa, S., Krupovic, M., and Delmont, T.O. (2025). Closest relatives of poxviruses are spread in the gut of humans and animals worldwide: the egoviruses. Preprint at bioRxiv, https://doi.org/10.1101/2024.03.23.586382 https://doi.org/10.1101/2024.03.23.586382.

34. Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G., et al. (2022). The Earth BioGenome Project 2020: Starting the clock. Proc. Natl. Acad. Sci. USA. *119*, e2115635118. https://doi.org/10.1073/pnas.2115635118.

35. Dowgier, G., Lahoreau, J., Lanave, G., Losurdo, M., Varello, K., Lucente, M.S., Ventriglia, G., Bozzetta, E., Martella, V., Buonavoglia, C., et al. (2018). Sequential circulation of canine adenoviruses 1 and 2 in captive wild carnivores, France. Vet. Microbiol. *221*, 67–73. https://doi.org/10.1016/j.vetmic.2018.05.025.

36. Whitley, R.J., and Roizman, B. (2001). Herpes simplex virus infections. Lancet *357*, 1513–1518. https://doi.org/10.1016/S0140-6736(00)04638-9.

37. Davison, A.J. (1998). The Genome of Salmonid Herpesvirus 1. J. Virol. *72*, 1974–1982. https://doi.org/10.1128/JVI.72.3.1974-1982.1998.

38. Zhang, X., Sun, S., Xiang, Y., Wong, J., Klose, T., Raoult, D., and Rossmann, M.G. (2012). Structure of Sputnik, a virophage, at 3.5-Å resolution. Proc. Natl. Acad. Sci. USA. *109*, 18431–18436. https://doi.org/10.1073/pnas.1211702109.

39. Arslan, D., Legendre, M., Seltzer, V., Abergel, C., and Claverie, J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc. Natl. Acad. Sci. USA. *108*, 17486–17491. https://doi.org/10.1073/pnas.1110889108.

40. Piégu, B., Guizard, S., Yeping, T., Cruaud, C., Asgari, S., Bideshi, D.K., Federici, B.A., and Bigot, Y. (2014).

666 Genome sequence of a crustacean iridovirus, IIV31, isolated from the pill bug, Armadillidium vulgare. J. Gen.
667 Virol. *95*, 1585–1590. https://doi.org/10.1099/vir.0.066076-0.

668 41. Collier, J.L., Rest, J.S., Gallot-Lavallée, L., Lavington, E., Kuo, A., Jenkins, J., Plott, C., Pangilinan, J., Daum,
669 C., Grigoriev, I.V., et al. (2023). The protist Aurantiochytrium has universal subtelomeric rDNAs and is a host
670 for mirusviruses. Curr. Biol. *33*, 5199-5207.e4. https://doi.org/10.1016/j.cub.2023.10.009.

671 42. Boratto, P.V.M., Oliveira, G.P., Machado, T.B., Andrade, A.C.S.P., Baudoin, J.-P., Klose, T., Schulz, F., Azza,
672 S., Decloquement, P., Chabrière, E., et al. (2020). Yaravirus: A novel 80-nm virus infecting Acanthamoeba
673 castellanii. Proc. Natl. Acad. Sci. USA. *117*, 16579–16586. https://doi.org/10.1073/pnas.2001637117.

674 43. Lange, M., and Jehle, J.A. (2003). The genome of the *Cryptophlebia leucotreta* granulovirus. Virology *317*,
675 220–236. https://doi.org/10.1016/S0042-6822(03)00515-4.

676 44. Lightner, D.V., Redman, R.M., Pantoja, C.R., Tang, K.F.J., Noble, B.L., Schofield, P., Mohney, L.L., Nunan,
677 L.M., and Navarro, S.A. (2012). Historic emergence, impact and current status of shrimp pathogens in the
678 Americas. J. Invertebr. Pathol. *110*, 174–183. https://doi.org/10.1016/j.jip.2012.03.006.

679 45. Doorbar, J., and Gallimore, P.H. (1987). Identification of proteins encoded by the L1 and L2 open reading
680 frames of human papillomavirus 1a. J. Virol. *61*, 2793–2799. https://doi.org/10.1128/jvi.61.9.2793-2799.1987.

681 46. Gazdar, A.F., Butel, J.S., and Carbone, M. (2002). SV40 and human tumours: myth, association or causality?
682 Nat. Rev. Cancer *2*, 957–964. https://doi.org/10.1038/nrc947.

683 47. Roitman, S., Rozenberg, A., Lavy, T., Brussaard, C.P.D., Kleifeld, O., and Béjà, O. (2023). Isolation and
684 infection cycle of a polinton-like virus virophage in an abundant marine alga. Nat. Microbiol. *8*, 332–346.
685 https://doi.org/10.1038/s41564-022-01305-7.

686 48. Karki, S., Moniruzzaman, M., and Aylward, F.O. (2021). Comparative Genomics and Environmental
687 Distribution of Large dsDNA Viruses in the Family Asfarviridae. Front. Microbiol. *12*, 657471.
688 https://doi.org/10.3389/fmicb.2021.657471.

689 49. Mitjà, O., Ogoina, D., Titanji, B.K., Galvan, C., Muyembe, J.-J., Marks, M., and Orkin, C.M. (2023).
690 Monkeypox. Lancet *401*, 60–74. https://doi.org/10.1016/S0140-6736(22)02075-X.

691 50. Starrett, G.J., Tisza, M.J., Welch, N.L., Belford, A.K., Peretti, A., Pastrana, D.V., and Buck, C.B. (2021).
692 Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. Virus
693 Evol. *7*, veaa055. https://doi.org/10.1093/ve/veaa055.

694 51. Maumus, F., Epert, A., Nogué, F., and Blanc, G. (2014). Plant genomes enclose footprints of past infections by
695 giant virus relatives. Nat. Commun. *5*, 4268. https://doi.org/10.1038/ncomms5268.

696 52. Filée, J. (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible
697 part of the iceberg? Virology *466–467*, 53–59. https://doi.org/10.1016/j.virol.2014.06.004.

698 53. Myers, J.M., Schulz, F., Rahimlou, S., Yadav, V., Amses, K.R., Simmons, D.R., Sun, S., Orozco-Quime, M.,
699 Heitman, J., Stajich, J.E., et al. (2025). Discovery of giant viruses as past and present infections of zoosporic

fungi. Preprint at bioRxiv, https://doi.org/10.1101/2024.01.04.574182.

54. Zhao, H., Zhang, R., Wu, J., Meng, L., Okazaki, Y., Hikida, H., and Ogata, H. (2023). A 1.5-Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus Rhizophagus irregularis. Virus Evol. *9*, vead064. https://doi.org/10.1093/ve/vead064.

55. Lopes-Lima, M., Froufe, E., Do, V.T., Ghamizi, M., Mock, K.E., Kebapçı, Ü., Klishko, O., Kovitvadhi, S., Kovitvadhi, U., Paulo, O.S., et al. (2017). Phylogeny of the most species-rich freshwater bivalve family (Bivalvia: Unionida: Unionidae): Defining modern subfamilies and tribes. Mol. Phylogenet. Evol. *106*, 174–191. https://doi.org/10.1016/j.ympev.2016.08.021.

56. Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The New Tree of Eukaryotes. Trends Ecol. Evol. *35*, 43–55. https://doi.org/10.1016/j.tree.2019.08.008.

57. Barth, Z.K., Hicklin, I., Thézé, J., Takatsuka, J., Nakai, M., Herniou, E.A., Brown, A.M., and Aylward, F.O. (2024). Genomic analysis of hyperparasitic viruses associated with entomopoxviruses. Virus Evol. *10*, veae051. https://doi.org/10.1093/ve/veae051.

58. Medvedeva, S., Guyet, U., Pelletier, E., Ruscheweyh, H.-J., Sunagawa, S., Ogata, H., Aylward, F.O., Gaïa, M., Yutin, N., Koonin, E.V., et al. (2026). Widespread and intron-rich mirusviruses are predicted to reproduce in nuclei of unicellular eukaryotes. Nat. Microbiol. *11*, 228–239. https://doi.org/10.1038/s41564-025-02190-6.

59. Barr, F.A., and Short, B. (2003). Golgins in the structure and dynamics of the Golgi apparatus. Curr. Opin. Cell Biol. *15*, 405–413. https://doi.org/10.1016/s0955-0674(03)00054-1.

60. Barbosa, N.S., Mendonça, L.R., Dias, M.V.S., Pontelli, M.C., da Silva, E.Z.M., Criado, M.F., da Silva-Januário, M.E., Schindler, M., Jamur, M.C., Oliver, C., et al. (2018). ESCRT machinery components are required for Orthobunyavirus particle production in Golgi compartments. PLoS Pathog. *14*, e1007047. https://doi.org/10.1371/journal.ppat.1007047.

61. Zhu, P., Liang, L., Shao, X., Luo, W., Jiang, S., Zhao, Q., Sun, N., Zhao, Y., Li, J., Wang, J., et al. (2016). Host Cellular Protein TRAPPC6AΔ Interacts with Influenza A Virus M2 Protein and Regulates Viral Propagation by Modulating M2 Trafficking. J. Virol. *91*, 10.1128/jvi.01757-16. https://doi.org/10.1128/jvi.01757-16.

62. Hay, B.A., Wassarman, D.A., and Rubin, G.M. (1995). Drosophila homologs of baculovirus inhibitor of apoptosis proteins function to block cell death. Cell *83*, 1253–1262. https://doi.org/10.1016/0092-8674(95)90150-7.

63. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P., and Venclovas, Č. (2020). Diversity and evolution of B-family DNA polymerases. Nucleic Acids Res. *48*, 10142–10156. https://doi.org/10.1093/nar/gkaa760.

64. Aswad, A., and Katzourakis, A. (2017). A novel viral lineage distantly related to herpesviruses discovered within fish genome sequence data. Virus Evol. *3*, vex016. https://doi.org/10.1093/ve/vex016.

65. Rosani, U., Gaia, M., Delmont, T.O., and Krupovic, M. (2023). Tracing the invertebrate herpesviruses in the global sequence datasets. Front. Mar. Sci. *10*. https://doi.org/10.3389/fmars.2023.1159754.

734 66. Guinet, B., Lepetit, D., Charlat, S., Buhl, P.N., Notton, D.G., Cruaud, A., Rasplus, J.-Y., Stigenberg, J., de
735 Vienne, D.M., Boussau, B., et al. (2023). Endoparasitoid lifestyle promotes endogenization and domestication
736 of dsDNA viruses. eLife *12*, e85993. https://doi.org/10.7554/eLife.85993.

737 67. Koonin, E.V. (2017). Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and
738 eukaryotes: common ancestry vs convergence. Biol. Direct *12*, 5. https://doi.org/10.1186/s13062-017-0177-2.

739 68. Marques, J.T., Meignin, C., and Imler, J.-L. (2024). An evolutionary perspective to innate antiviral immunity
740 in animals. Cell Rep. *43*. https://doi.org/10.1016/j.celrep.2024.114678.

741 69. Dolja, V.V., Krupovic, M., and Koonin, E.V. (2020). Deep Roots and Splendid Boughs of the Global Plant
742 Virome. Annu. Rev. Phytopathol. *58*, 23–53. https://doi.org/10.1146/annurev-phyto-030320-041346.

743 70. Paludan, S.R., Pradeu, T., Masters, S.L., and Mogensen, T.H. (2021). Constitutive immune mechanisms:
744 mediators of host defence and immune regulation. Nat. Rev. Immunol. *21*, 137–150.
745 https://doi.org/10.1038/s41577-020-0391-5.

746 71. Kassiotis, G. (2023). The Immunological Conundrum of Endogenous Retroelements. Annu. Rev. Immunol. *41*,
747 99–125. https://doi.org/10.1146/annurev-immunol-101721-033341.

748 72. Sarre, L.A., Kim, I.V., Ovchinnikov, V., Olivetta, M., Suga, H., Dudin, O., Sebé-Pedrós, A., and de Mendoza,
749 A. (2024). DNA methylation enables recurrent endogenization of giant viruses in an animal relative. Sci. Adv.
750 *10*, eado6406. https://doi.org/10.1126/sciadv.ado6406.

751 73. Rathour, R., Ma, Y., Xiong, J., Wang, X.-W., Petersen, J., and Zhang, X. (2025). Hemolymph microbiota and
752 host immunity of crustaceans and mollusks. ISME J. *19*, wraf133. https://doi.org/10.1093/ismejo/wraf133.

753 74. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y., and Ishiguro-Watanabe, M. (2025). KEGG: biological
754 systems database as a model of the real world. Nucleic Acids Res. *53*, D672–D677.
755 https://doi.org/10.1093/nar/gkae909.

756 75. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics.
757 Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

758 76. Edgar, R.C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence
759 homology and phylogeny. Nat. Commun. *13*, 6968. https://doi.org/10.1038/s41467-022-34630-w.

760 77. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment
761 trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.
762 https://doi.org/10.1093/bioinformatics/btp348.

763 78. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.
764 https://doi.org/10.1371/journal.pcbi.1002195.

765 79. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K.,
766 Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222.
767 https://doi.org/10.1093/nar/gkt1223.

80. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

81. Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. *29*, 2607–2618. https://doi.org/10.1093/nar/29.12.2607.

82. Klemm, P., Stadler, P.F., and Lechner, M. (2023). Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. Front. Bioinform. *3*. https://doi.org/10.3389/fbinf.2023.1322477.

83. Moniruzzaman, M., Martinez-Gutierrez, C.A., Weinheimer, A.R., and Aylward, F.O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat. Commun. *11*, 1710. https://doi.org/10.1038/s41467-020-15507-2.

84. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

85. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

86. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

87. Zielezinski, A., Gudyś, A., Barylski, J., Siminski, K., Rozwalak, P., Dutilh, B.E., and Deorowicz, S. (2025). Ultrafast and accurate sequence alignment and clustering of viral genomes. Nat. Methods *22*, 1191–1194. https://doi.org/10.1038/s41592-025-02701-7.

88. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

89. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

90. Camargo, A.P., Nayfach, S., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T.B.K., Mukherjee, S., Schulz, F., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. *51*, D733–D743. https://doi.org/10.1093/nar/gkac1037.

91. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

92. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform.

Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

93. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

94. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

95. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Sci. *32*, e4792. https://doi.org/10.1002/pro.4792.

96. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

97. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

98. Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431. https://doi.org/10.1186/1471-2105-11-431.

99. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

100. Sievers, F., and Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. *27*, 135–145. https://doi.org/10.1002/pro.3290.

101. Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. *52*, W78–W82. https://doi.org/10.1093/nar/gkae268.

828 **Supplemental dataset legends**

829 **Supplementary Table 1: Eukaryotic assemblies used in this study**

830 **Supplementary Table 2: Reference viral genomes used in this study**

831 **Supplementary Table 3: Information on identified VRs**

832 **Supplementary Table 4: Summary statistics of VRs across major viral taxa**

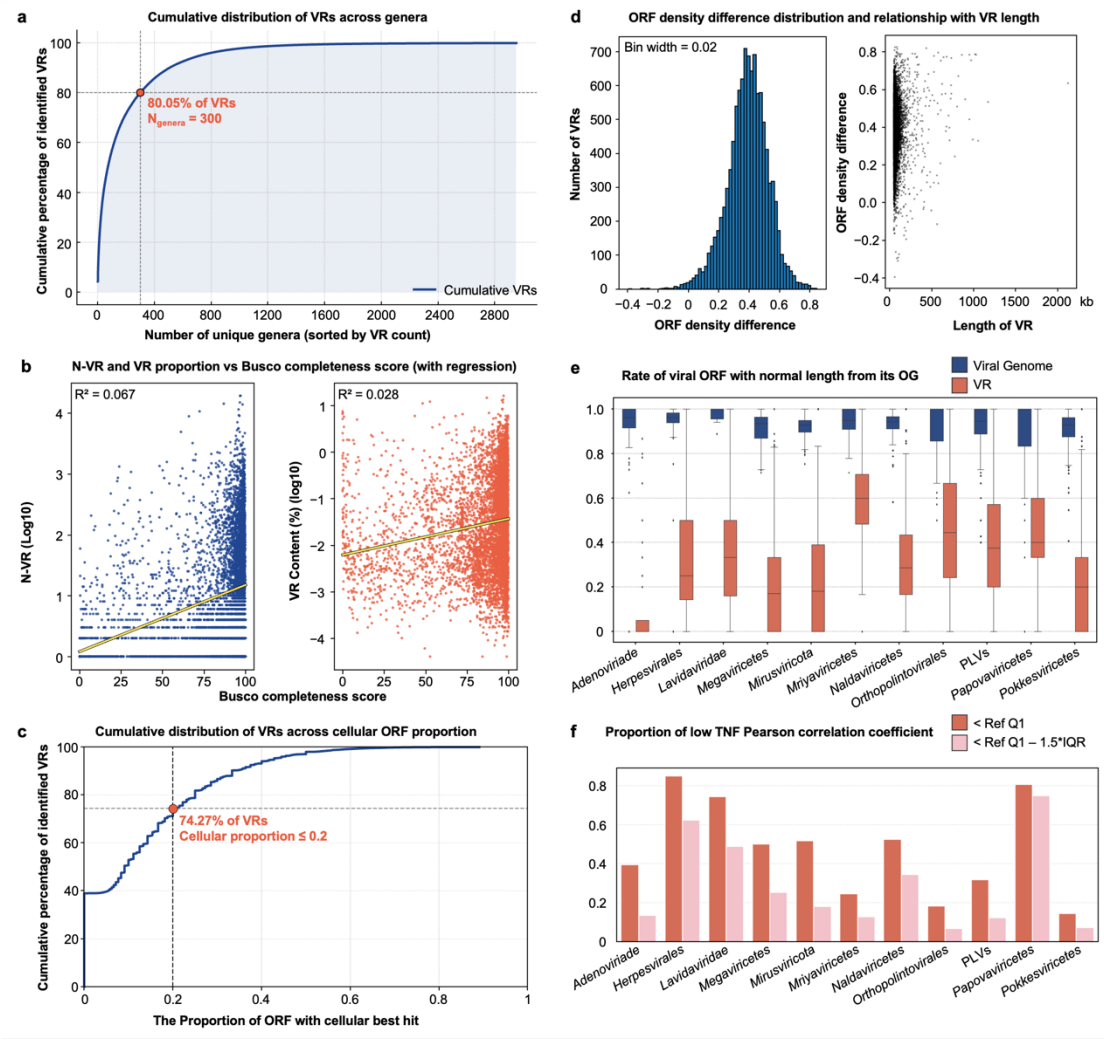
833 **Supplementary Table 5: Proportion of redundant VRs across individual assemblies**

834 **Supplementary Table 6: Host distribution of the three most abundant KOs**

835

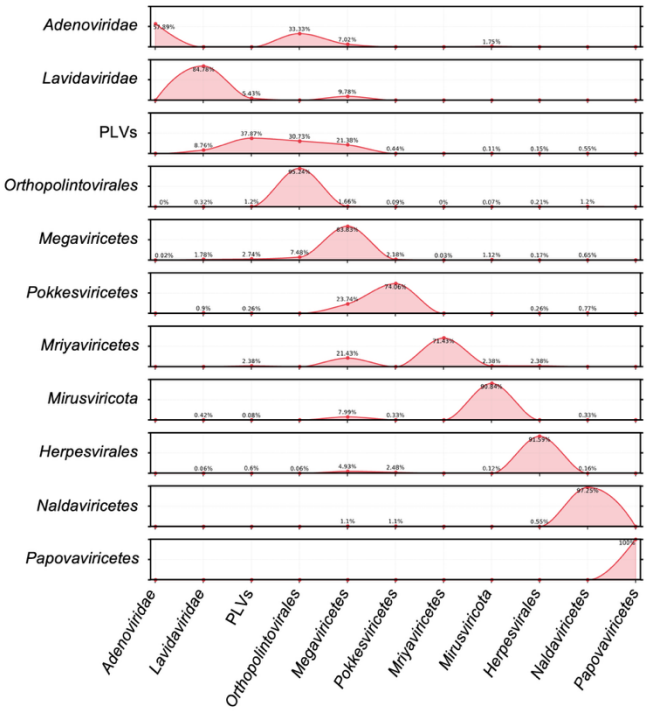
836

Supplementary Figures



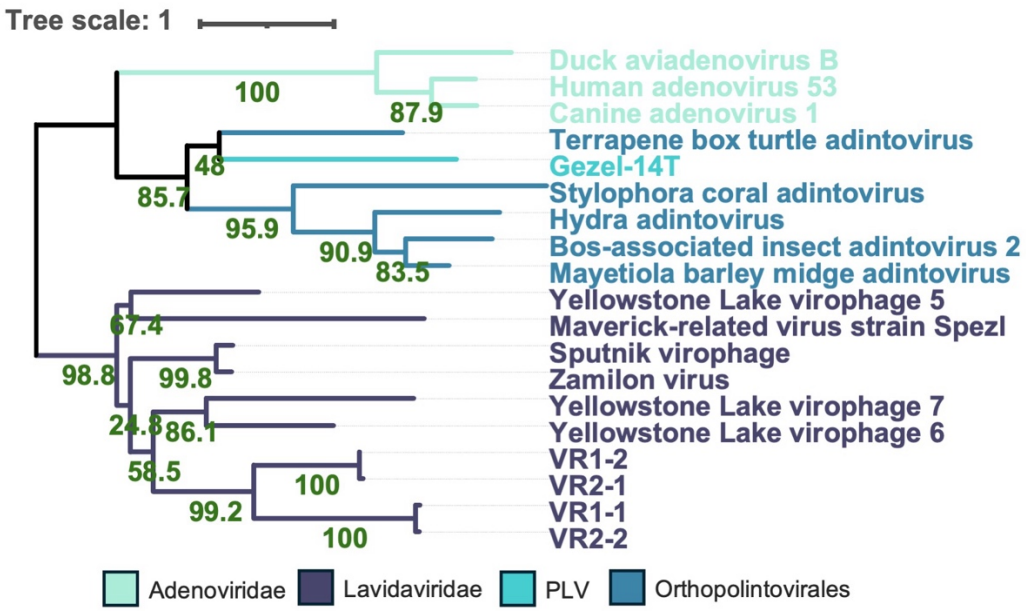
Supplementary Fig. 1 | Properties of VRs

(a–f) Subfigure titles are shown at the top of each panel. (b) The left plot represents the number of VRs, and the right plot represents the proportion of VRs in each assembly. Each dot corresponds to one assembly, and the regression line is shown in yellow. **(d)** The difference represents the ORF density of VRs minus that of the corresponding. **(e)** The y-axis represents, for each VR or viral genome, the proportion of ORFs within the normal length range of their corresponding OG. **(f)** The y-axis represents, for all VRs of a given taxon, the proportion of VRs whose TNF composition differs from that of the host assembly.



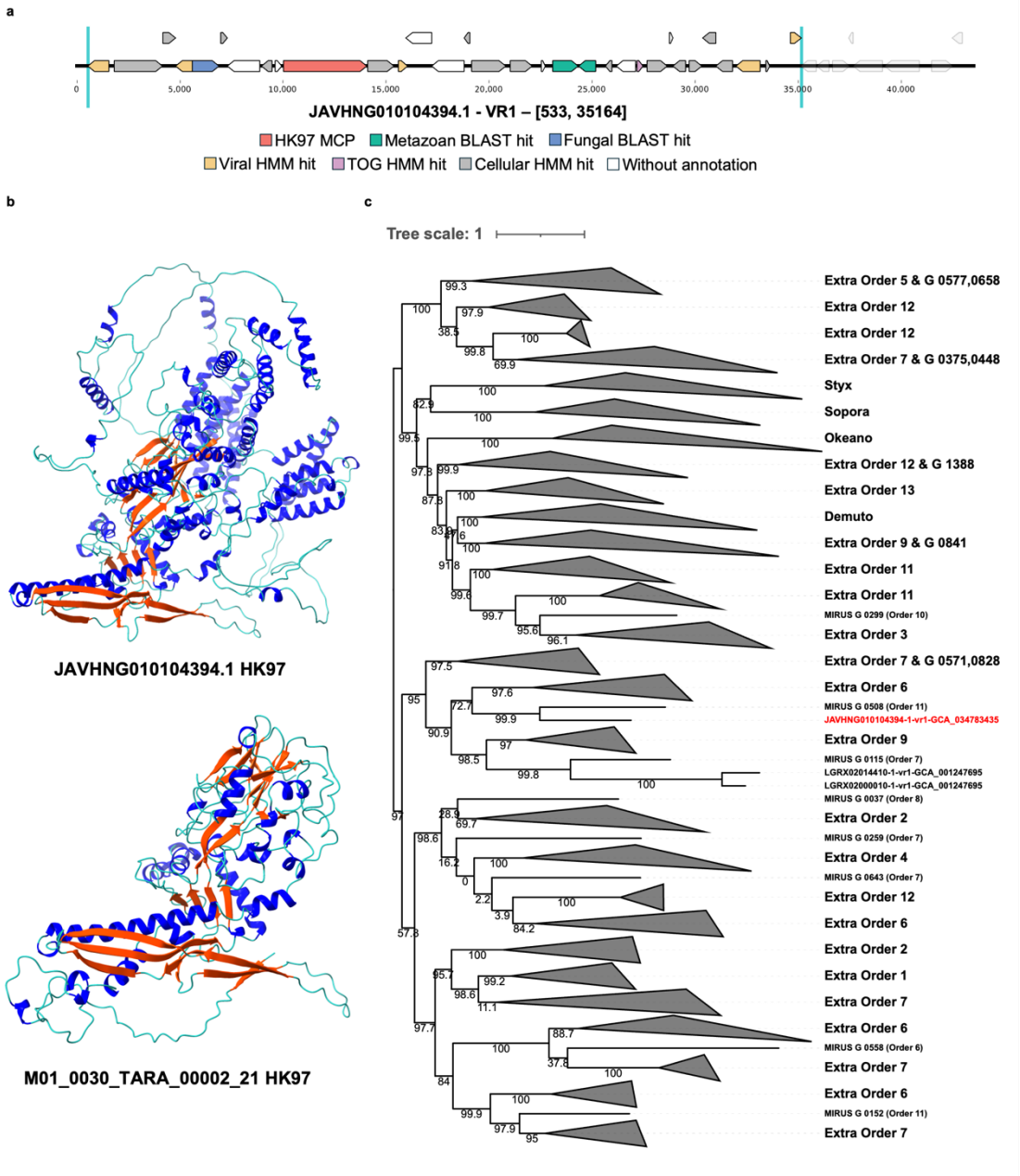
Supplementary Fig. 2 | Independent taxonomic assessment results of the HCVRs

Each subplot corresponds to one viral taxon and shows the taxonomic distribution of the best hits for all HCVRs belonging to that taxon in the alignment-based analysis. The vertical axis indicates the proportion, and the horizontal axis represents different viral taxa in the reference database.



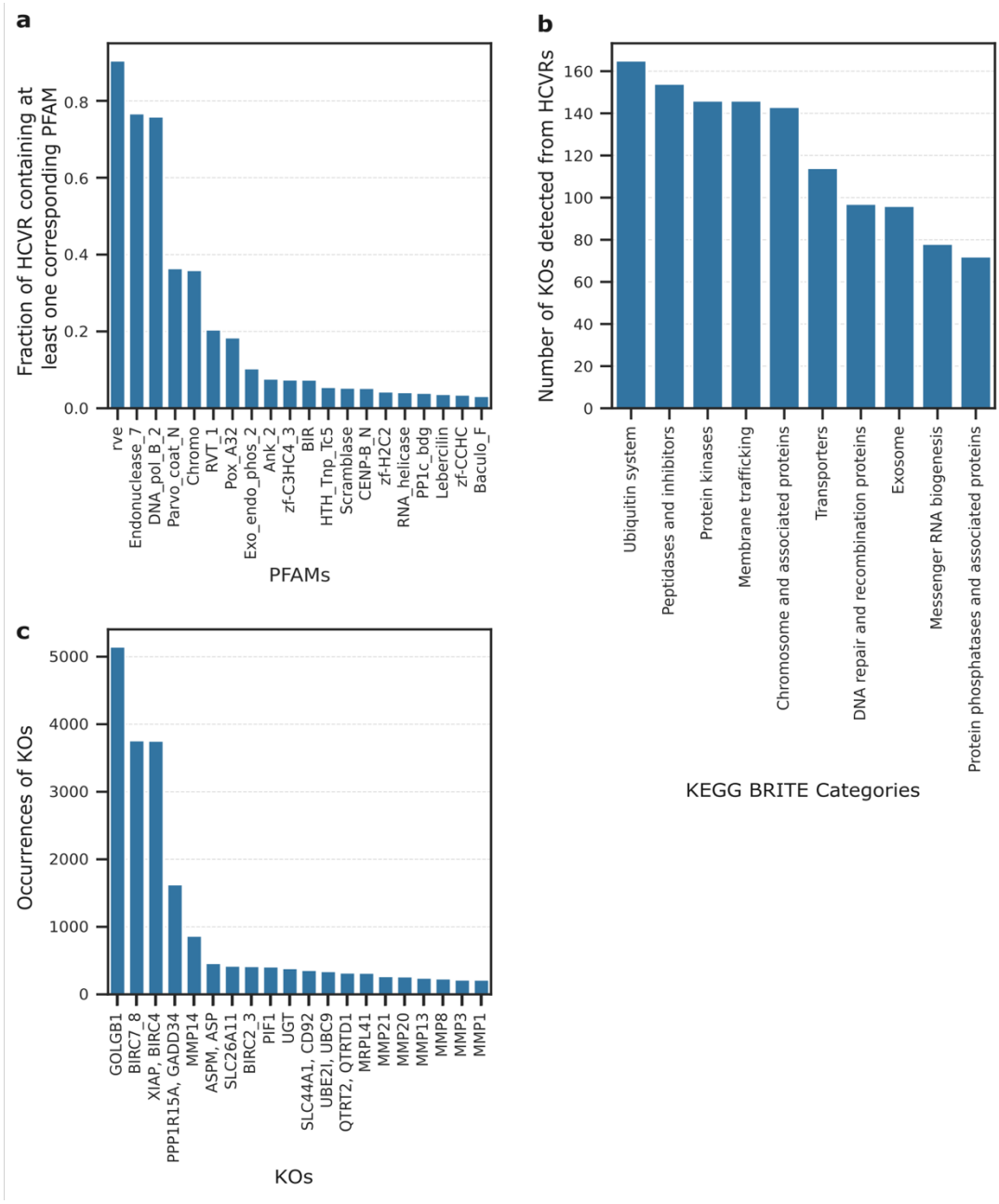
Supplementary Fig. 3 | Maximum-likelihood phylogenetic tree reconstructed using the pATPase sequences

The reference viral sequences included YP_010796965.1, AVL93344.1, YP_009177784.1, UYE94484.1, BBW89550.1, YP_004300293.1, DAC81269.1, YP_002122364.1, YP_009177673.1, YP_010796956.1, XAM99555.1, DAC81558.1, DAC81736.1, YP_009177816.1 and Q65945.1. The tree was midpoint-rooted, and ultrafast bootstrap support values were shown along the branches. The best-fit substitution model was Q.pfam+F+I+R3.



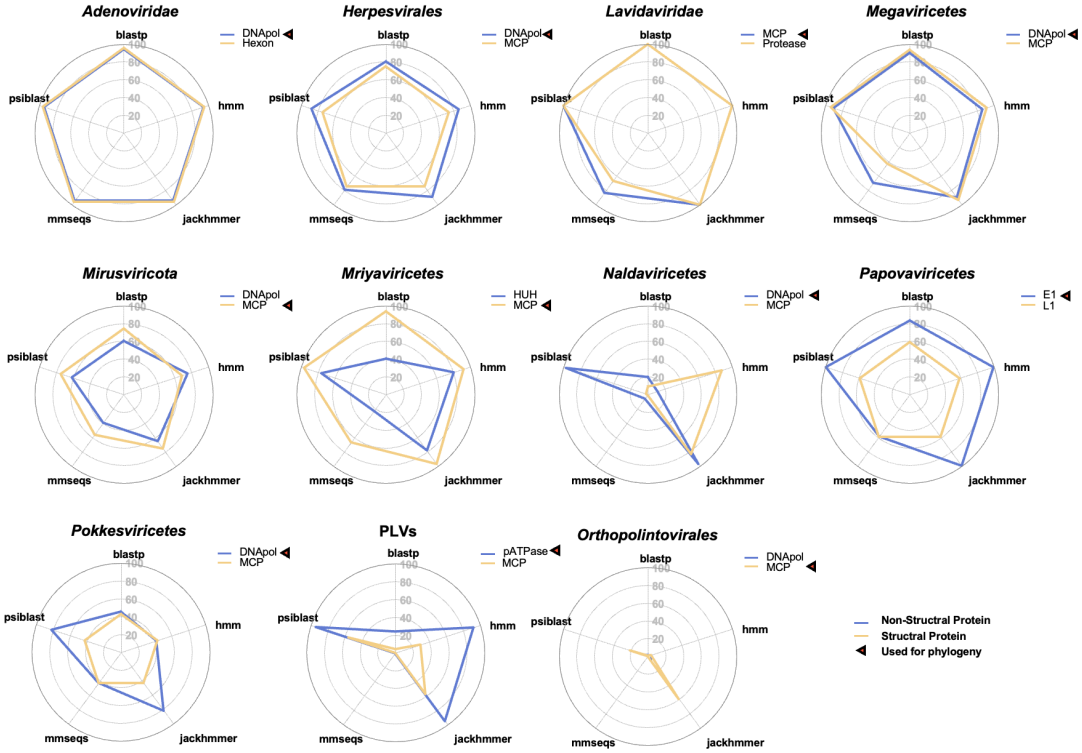
Supplementary Fig. 4 | *Mirusviricota* MCP from *A. swindernianus*

(a) Genomic context of the HCVR containing an HK97 MCP. The region delimited by two light blue vertical lines indicates the HCVR. ORFs in the magnified genome maps are color-coded. **(b) Predicted structure of the HK97 MCP encoded by this contig and its closest mirusviral reference sequence.** Distinct secondary structure elements are indicated in different colors. **(c) Maximum-likelihood phylogenetic tree reconstructed from the HK97 MCP.** The tree was midpoint-rooted, and ultrafast bootstrap support values were shown along the branches. The best-fit substitution model was VT+F+I+G4.



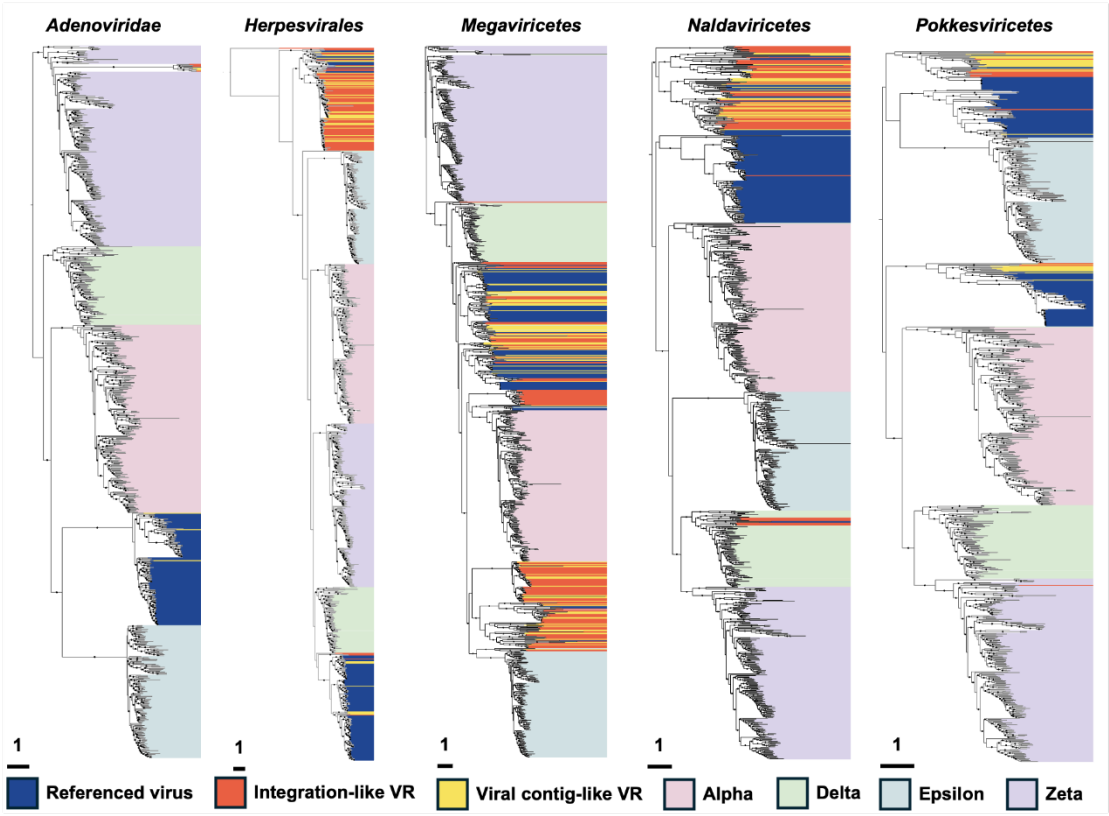
Supplementary Fig. 5 | Annotation statistics for proteins in HCVRs

(a) Top 20 PFAM domains by prevalence. A fraction of 0.8 represents 80% of HCVRs have at least one corresponding PFAM domain-containing gene. **(b)** Top 10 KEGG BRITE categories ranked by number of distinct KOs detected in HCVRs. **(c)** Occurrence of the top 20 KOs in HCVRs.



Supplementary Fig. 6 | Distribution of marker genes in reference viral genomes.

Each radar plot corresponds to a viral taxon, as labeled above. Concentric circles represent the proportion of genomes in which homologs of each gene were detected using the indicated method. Colored lines denote different marker genes, as shown in the legend (top right). Triangles indicate the genes selected for phylogenetic tree reconstruction.



Supplementary Fig. 7 | Phylogenetic trees of DNA polymerase including eukaryotic homologs.

Each tree corresponds to a specific viral taxon, as indicated in the panel title. Scale bars are shown in the lower left corners, and all trees are midpoint-rooted. Black circles indicate ultrafast bootstrap support values greater than 90. The color legend is shown at the bottom. The best-fit substitution model used for *Adenoviridae*, *Herpesvirales*, and *Naldaviricetes* was Q.pfam+I+G4, while Q.pfam+F+I+G4 was applied for *Megaviricetes* and *Pokkesviricetes*.