# 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

Genomic sequences of viral origin, especially those derived from retroviruses, constitute a large fraction of eukaryotic genomes, yet the role of double-stranded (ds) DNA viruses in shaping eukaryotic genomes remains underexplored. Here, we present a computational framework to identify dsDNA viral regions (VRs) in eukaryotic genomes. By systematically screening 37,254 eukaryotic genome assemblies, we identified 781,111 VRs in 7,103 genome assemblies. These VRs accounted for up to 16% of individual genomes and represented 343 associations between viral and eukaryotic taxa at the class level, which notably included 305 (89%) associations lacking evidence from isolation studies. Some VRs are phylogenetically nested within clades containing known viruses, while others form deep-branching clades composed solely of VR-derived sequences. Our study greatly expands the known dsDNA endogenous virosphere and the resulting VR catalog offers opportunities for wider exploration of the virus–host coevolutionary processes.

## Main

Endogenous viral elements are widespread across diverse eukaryotic genomes[1,2]. In particular, endogenous retroviral elements are known to account for a considerable proportion of certain eukaryotic genomes, including ~8% of the human genome[3–5]. Endogenous retroviral elements occasionally cause diseases, provide essential host functions through “co-option”, or contribute antiviral functions[6–9]. In contrast, endogenous elements derived from dsDNA viruses remain underexplored[2,10], although recent studies started to uncover their potential impact, particularly in protist genomes[11,12].

Given their large coding capacity compared with RNA viruses, dsDNA viruses may have a large potential to

affect the evolution of their host genomes if endogenized[13]. Evidence from a protein sequence–level study suggests horizontal transfers of functionally diverse genes from dsDNA viruses to eukaryotes[14]. However, the extent to which these viruses contribute genes and genomic regions of different eukaryotes remains unclear.

Eukaryotic dsDNA viruses are polyphyletic and highly diverse. They range from small viruses, such as papillomavirids[15] (including papillomaviruses) and viruses from the polinton-like supergroup[16,17], to giant representatives like megaviricetes[18]. Furthermore, the known eukaryotic dsDNA virosphere is rapidly expanding thanks to culture-independent sequencing approaches (i.e., "metagenomics")[19] which led to the discoveries of numerous novel dsDNA viral groups, including polinton-like viruses[20] (PLVs, also classified as *Aquintoviricetes*[16]), *Mirusviricota*[21], *Mriyaviricetes*[22], and *Egovirales*[23]. The sequence data of these viral lineages provide new opportunities to assess the impact of their integration on the evolution of eukaryotic genomes[2,12,24]. Nonetheless, most existing studies have focused on specific viral or eukaryotic lineages[11,12,25], and a comprehensive, up-to-date framework to systematically evaluate dsDNA viral elements across eukaryotic assemblies was lacking.

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

We conducted a systematic analysis to detect dsDNA VRs across eukaryotic genomic assemblies (**Fig. 1a**). In total, 37,254 assemblies representing diverse eukaryotic lineages were analyzed (**Fig. 1b, Supplementary Table 1**). Open reading frames (ORFs) were predicted using two distinct translation approaches commonly applied to prokaryotic and viral genomes. To aid in the identification of viral signatures within eukaryotic contigs, scaffolds, and chromosomes (collectively referred to as "contigs" hereafter), we also curated and constructed a comprehensive reference dataset of cellular protein sequences.

Our study incorporated 2,126 reference viral genomes from eleven eukaryotic dsDNA viral taxa (**Table 1, Supplementary Table 2**), most of which correspond to class-level taxa. Hidden Markov Models (HMMs) were constructed for 7,755 orthologous group (OGs) derived from these viral genomes and searched against cellular reference sequences. Among them, 1,842 viral OGs (VOGs), defined as OGs without similarity to any cellular reference sequences, and 1,344 taxonomic OGs (TOGs), characterized as VOGs strictly specific to a single viral taxon, were selected to capture viral and viral lineage-specific signatures, respectively. Based on the presence of ORFs with significant similarity to VOGs or TOGs within eukaryotic contigs, we identified 946,599 virus–contig pairs as potential virus–host associations.

**Table 1 | Eleven dsDNA viral taxa involved in this study and representative viruses.**

| Viral taxon | Representative viruses |
|---|---|
| *Adenoviridae* | Canine adenovirus 1[26] |
| *Herpesvirales* | Herpes simplex virus[27]; Salmonid herpesvirus 1[28] |
| *Lavidaviridae* | Mimivirus-dependent virus Sputnik (“virophage”)[29] |
| *Megaviricetes* | Megavirus chilensis[30]; Invertebrate iridescent virus 31[31] |
| *Mirusviricota* | Circular genomes uncovered from eukaryotes[12,24] |
| *Mriyaviricetes* | Yaravirus brasiliensis[32] |
| *Naldaviricetes* | Cryptophlebia leucotreta granulovirus; White spot syndrome virus[33,34] |
| PLVs | Gezel-14T (“virophage”)[35] |
| *Papovavircetes* | Human papillomavirus[36]; Simian virus 40[37] |
| *Pokkesviricetes* | African swine fever virus[38]; Monkeypox virus[39] |
| *Orthopolintovirales* | Terrapene box turtle adintovirus[40] |

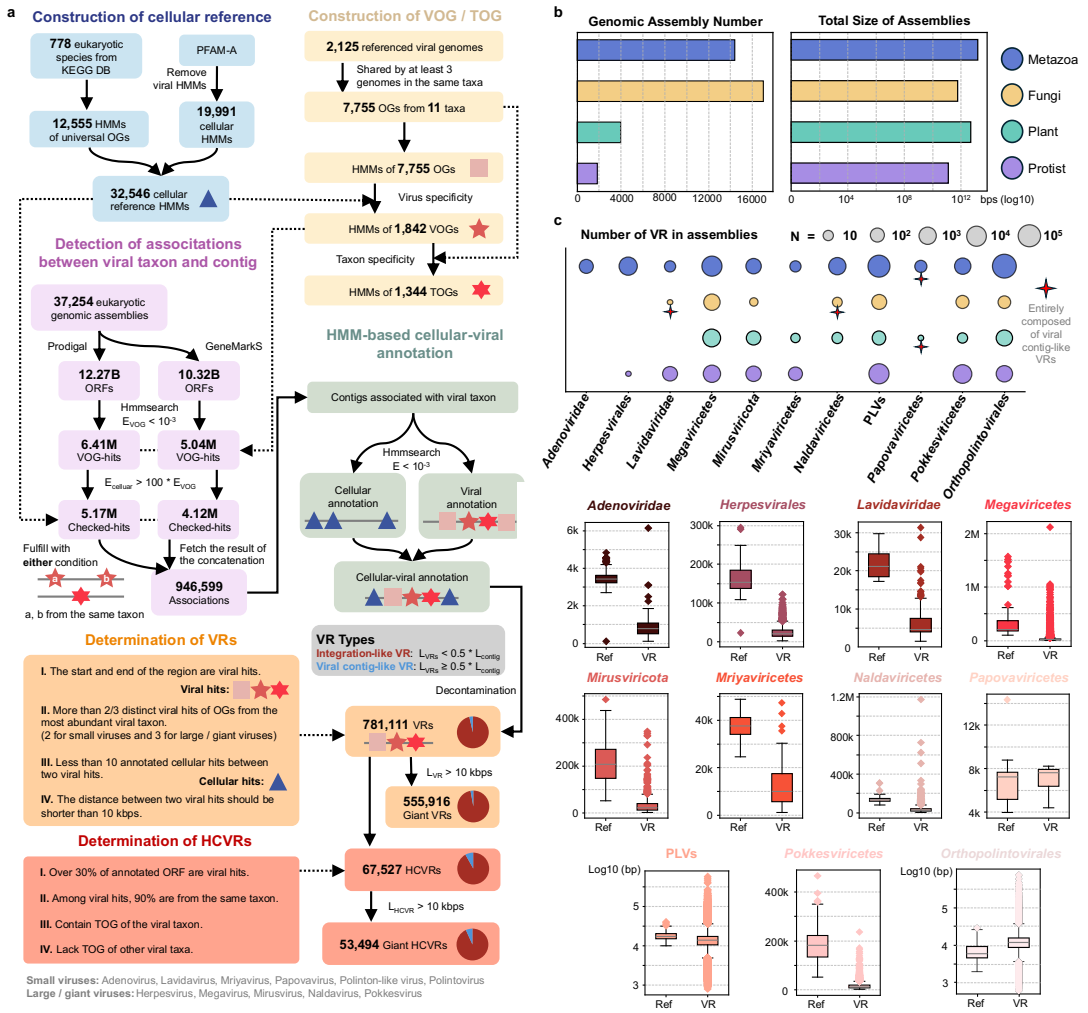


**Fig. 1 | Viral regions identified in eukaryotic genome assemblies.**

(a) VR discovery pipeline. Distinct background color represents different part of the pipeline. (b) Genomic assemblies used in this study. Colors indicate to different eukaryotic groups. The left panel shows the number of genomic assemblies in eukaryotic groups, while the right panel shows the total assembly lengths. (c) Number of identified VRs. Colors match those in (b). The radius of each circle represents the logarithmically transformed count of VRs. (d) Length distribution of reference virus genomes and identified VRs. Each boxplot displays the size range of complete reference viral genomes (left) and the corresponding VRs (right) for each viral taxon, measured in base pairs. The vertical axes for PLVs and *Orthopolintovirales* are log10-transformed.

For each potential virus-contig pair, we scanned the contig using HMMs of the OGs from the viral taxon, alongside the cellular reference HMMs, to delineate VRs. This analysis identified a total of 781,111 VRs in 7,103 genome assemblies (19% of all assemblies). The distribution of VRs across different host lineages was uneven, with metazoans accounting for the largest proportion (757,529 VRs, 96.98%), followed by protists (20,186 VRs, 2.58%), plants (2,442 VRs, 0.31%), and fungi (954 VRs, 0.12%). The assemblies containing VRs included 6,094 metazoan assemblies (42.25% of all metazoan assemblies) and 633 protist assemblies (34.39%) (**Fig. 1c**). Metazoan genome data included VRs from every viral taxon. Despite the lack of confirmed dsDNA viruses infecting fungi and plants, 223 plant assemblies (5.59%) and 153 fungal assemblies (0.89%) were found to contain dsDNA VRs. To distinguish likely integration events from potential viral genome contamination or unintegrated sequences, VRs were classified based on their proportion of the total contig length: those comprising less than 50% of the contig length were defined as “integration-like”, whereas those occupying more than 50% were designated as “viral-contig-like”. The majority of VRs (96.43%) were classified as integration-like VRs (**Supplementary Table 3**).

In terms of viral taxonomy, the majority of VRs were associated with *Orthopolintovirales* (n=633,714) and PLVs (n=122,043), followed by *Megaviricetes* (n=16,525), *Pokkesviricetes* (n=4,182), *Herpesvirales* (n=2,442), *Naldaviricetes* (n=1,085), *Mirusviricota* (n=682), *Lavidaviridae* (n=188), *Mriyaviricetes* (n=110), *Adenoviridae* (n=104) and *Papovaviricetes* (n=36). Notably, all *Papovaviricetes* VRs were viral-contig-like, with no evidence of integration-like signatures.

Most VRs were shorter than the reference viral genomes of corresponding lineages (**Fig. 1d**), although some viral taxa, notably *Orthopolintovirales* and PLVs, exhibited a broad size distribution. In rare cases, VRs exceeded the length of their reference genomes, possibly due to the accumulation of repetitive sequences, multiple viral insertions, or the incorporation of host-derived sequences. Overall, the length of VRs exhibited greater variability than typically observed for reference viral genomes.

92

93 **dsDNA viral elements largely contribute to the genome composition of diverse eukaryotes**

94 To assess the extent of dsDNA viral elements in eukaryotic genomes, we quantified the proportion of identified
95 VRs across genomic assemblies. In metazoans and protists, VR proportion typically ranged from trace levels up to
96 10%, whereas in fungi and plants, it was generally low and seldom exceeded 1% (**Fig. 2a**). Notably, several
97 assemblies from specific metazoan groups (i.e., Mollusca, Cnidaria, and Arthropoda) and protists (i.e., Parabasalia)
98 exhibited VR proportions exceeding 10%, suggesting a substantial contribution of dsDNA viral sequences to their
99 genome composition. We additionally checked the VRs in the human genome (75 VRs from 59 assemblies), which
100 on average accounted for 0.004%, while those in vertebrate genomes averaged 0.014%, with the highest proportion
101 (0.725%) observed in the assembly of a fish, *Gasterosteus aculeatus*.

102 To characterize the genomic landscape of VRs, we selected representative assemblies from each major
103 eukaryotic group for in-depth analysis. These included *Unio delphinus*, a freshwater bivalve with the highest VR
104 proportion in metazoans (16.20%); *Trichomonas vaginalis*, a flagellated protozoan parasite with the highest VR
105 proportion in protists (12.13%); *Gaertneriomyces semiglobifer*, a zoospore-producing fungus with the second-
106 highest VR proportion in fungi (2.66%); and *Microglena* sp. YARC, a polar marine green alga with the highest VR
107 proportion in plants (3.40%) (**Fig. 2b**). All selected assemblies contained VRs from two or more viral taxa, except
108 for *G. semiglobifer*, which exclusively harbored VRs from PLVs. In addition, the GC content of VRs often differed
109 from that of the host genome, particularly in *Microglena*, indicating that viral insertions could influence host genome
110 composition[13].

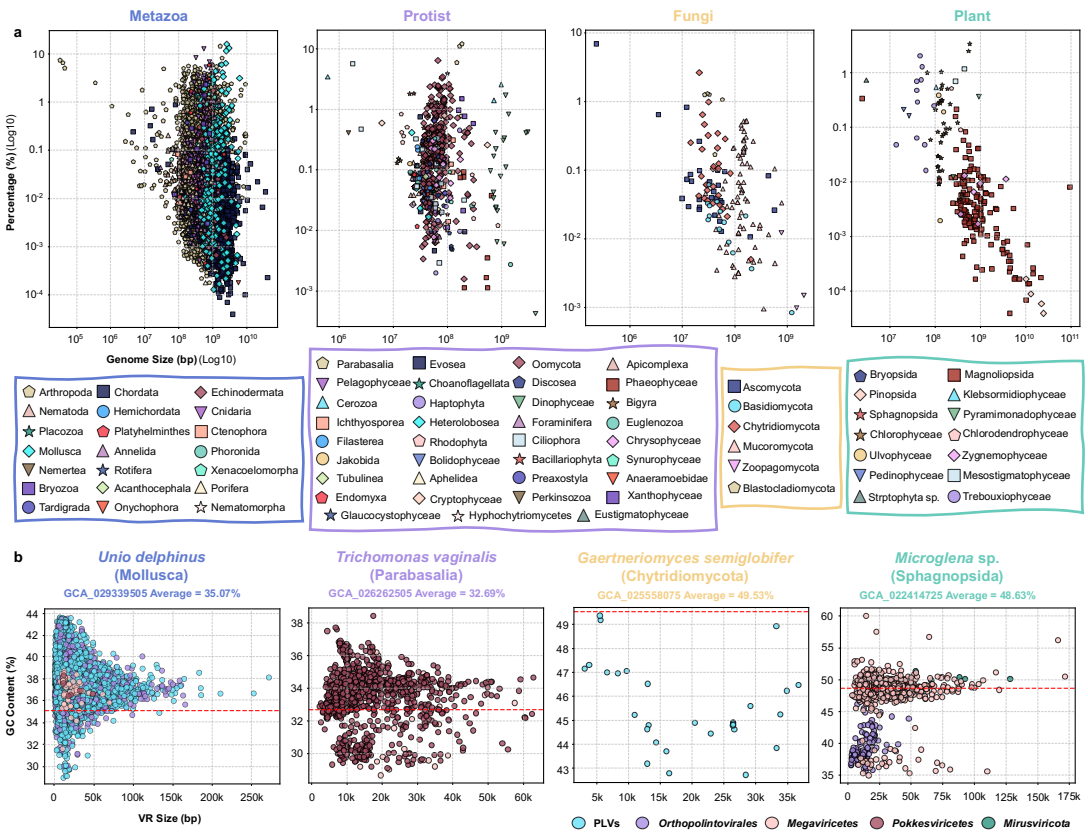



**Fig. 2 | Extent of VRs in eukaryotic genomes.**

**(a) The proportion of identified VRs within different eukaryotic groups.** From left to right: metazoans (blue), protists (purple), fungi (yellow), and plants (green). Each dot represents a genomic assembly, with its color and shape corresponding to its taxon. The legend for each icon is provided in the corresponding-colored box. The horizontal axis represents the assembly size, while the vertical axis represents the percentage of the genome occupied by VRs, both of which are log10-scaled. **(b) GC content and length distribution of VRs in the four selected assemblies.** The order of the assemblies follows that in (a). Each dot represents a VR, with colors corresponding to viral taxa. The horizontal axis represents VR size, while the vertical axis represents GC content. The red dashed line indicates the average GC content of the respective assembly.

The GC content of eukaryotic assemblies negatively correlated with the abundance of VRs (**Supplementary Fig. 1**). Among the 37,254 analyzed assemblies, 46.6% (n = 17,364) exhibited a GC content between 40% and 50%, whereas only 34.7% (n=2,471) of VR-containing assemblies fell within the same GC range. In contrast, 32.9% (n = 12,288) of analyzed assemblies showed a GC content ranging from 30% to 40%, while 52.8% of VR-containing assemblies belonged to this category. Furthermore, the majority of assemblies with particularly high VR proportion exhibited GC content in the 30%–40% range. A Spearman correlation analysis showed a significant negative association between genomic GC content and VR proportion ($\rho = -0.135$, $p < 2.5 \times 10^{-30}$), suggesting a tendency for

128 dsDNA VRs to accumulate in AT-rich genomes. While the correlation is statistically significant, the small effect size
129 indicates the existence of other factors related to the proportion of VRs.

130 The proportion of ORFs with known functions per VR was relatively low, averaging 19.63% based on COG
131 category annotations (**Supplementary Fig. 2a**). Among the annotated ORFs, those assigned to category "L"
132 (replication, recombination, and repair) were the most prevalent (**Supplementary Fig. 2b**). Overall, 47.22% of VRs
133 contained at least one replication-related gene, with the highest proportion observed in *Papovaviricetes* (98.57%)
134 and the lowest in *Lavidaviridae* (2.22%). We further classified genes within category "L", focusing on key
135 replication-associated components such as DNA ligases, polymerases, nucleases, helicases, and integrases (see
136 Methods for details) (**Supplementary Fig. 2c**). The prevalence and diversity of these replication-related genes
137 implied that part of VRs may retain the capacity to replicate and mobilize through various mechanisms[41–43].

138

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

140 While VRs can indicate a potential viral origin for a genomic region, they have limitations in precisely
141 assigning viral lineage, particularly when features from multiple viral lineages are present. To address this limitation
142 in characterizing viral taxa of VRs, we applied more stringent criteria and defined high-confidence viral regions
143 (HCVRs) as VRs with at least 30% ORFs related to viruses, of which 90% are associated to a same viral taxon (**Fig.
144 1a & Methods for details**). This filtering yielded 63,562 HCVRs in metazoans (8.39% of all metazoan VRs), 537
145 in fungi (56.28%), 619 in plants (25.34%), and 2,809 in protists (13.91%). Compared to VRs, HCVRs provide a
146 more reliable representation of viral identity and exhibit more homogenous viral taxonomic signals. More
147 importantly, they enable a clearer one-to-one relationship between viral and eukaryotic lineages, facilitating more
148 confident inference of virus–host associations

149 Based on HCVRs, we systematically expanded the known associations between eukaryotes and dsDNA viruses
150 (**Fig. 3**). These associations may represent either present or past virus-host relationships. HCVRs were identified in
151 assemblies spanning over 30 eukaryotic phyla and 128 classes (or order-level when class-level taxon was
152 unavailable), resulting in 343 associations between viral and eukaryotic classes. Most (n = 277, 80.75%) of these
153 associations were supported by the presence of integration-like HCVRs. Among eukaryote-virus pairs previously
154 known from virus isolation studies, we successfully recovered 38 out of 47 associations through the identification
155 of HCVRs in their genomic assemblies, demonstrating that VR identification is an effective approach for inferring
156 virus-host relationships. Notably, many of the associations (n = 305, 88.92%) involved a viral taxon from which no
157 virus has yet been isolated for the paired eukaryotic taxon, underscoring the extent of unexplored dsDNA virosphere

in eukaryotes. For example, no dsDNA viruses have been reported in fungi or streptophytes to date, with their known virosphere consisting exclusively of RNA and single-stranded DNA viruses. However, our data revealed their association (153 fungi and 169 streptophytes assemblies) with diverse dsDNA viruses, particularly megaviricetes. Furthermore, even in well-studied lineages such as Arthropoda, Chordata, and Mollusca, we uncovered numerous previously unrecognized viral associations, highlighting the hidden and underestimated associations between dsDNA viruses and eukaryotes.

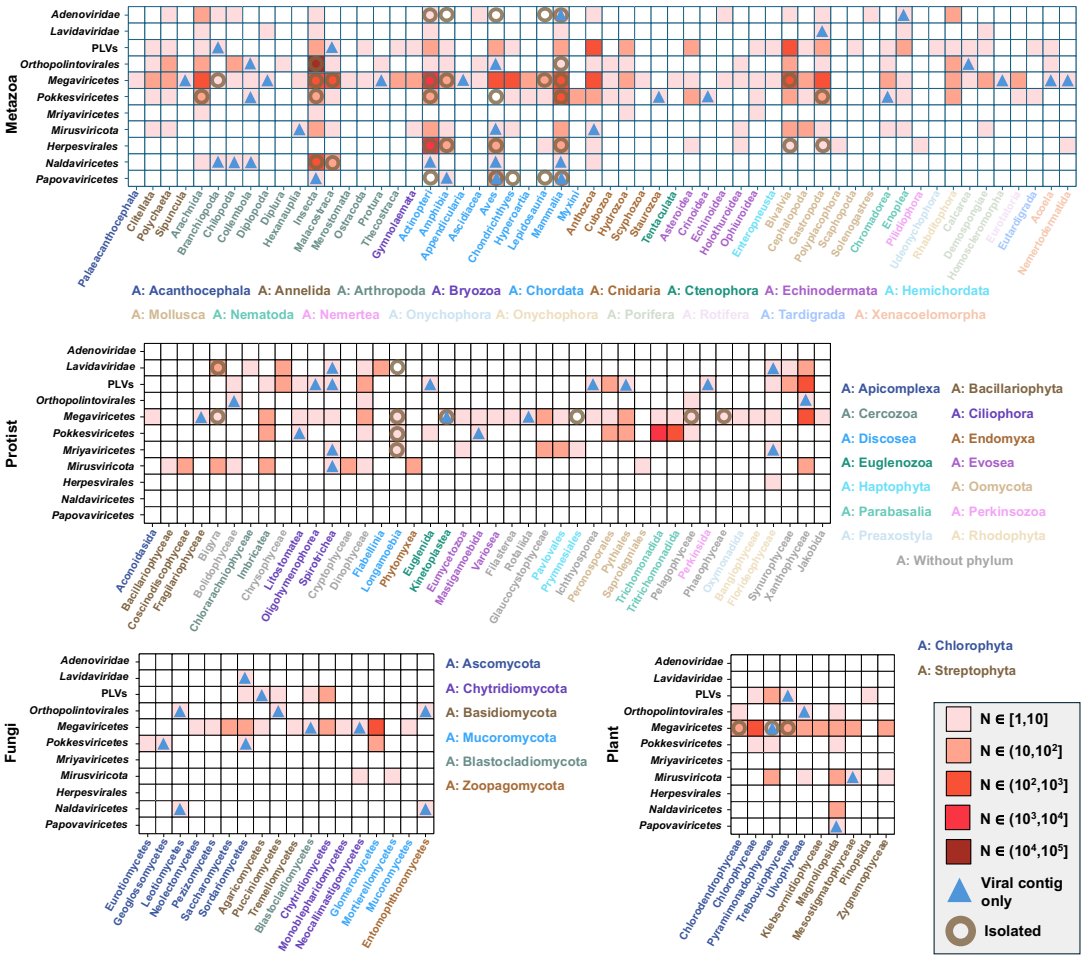



**Fig. 3 | HCVR distribution in eukaryotes.**

From top left to bottom right, the heatmaps represent metazoans, protists, fungi, and plants. 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 referenced alongside the heatmap. A blue triangle marks units where all HCVRs are viral contig-like, while a brown ring indicates the presence of viruses previously isolated from these eukaryotes.

For instance, virophages are known to coinfect protists together with nucleocytoviricots[16,17]. In our study, two

172 *Lavidaviridae* HCVRs were identified within a single contig from the pomace fly, *Lordiphosa mommai* (**Fig. 4a**).
173 Among 21 annotated ORFs on this contig, 16 showed best matches to metazoan proteins and 5 to viral proteins based
174 on alignment annotation, supporting its eukaryotic origin. Both HCVRs contained core *Lavidaviridae* genes,
175 including the hexon and penton capsid proteins, the packaging ATPase (pATPase), cysteine protease, but also an
176 integrase, which could have been involved in the integration process in the eukaryotic genome. We also observed
177 abundant *Megaviricetes* HCVRs in Drosophilidae (the family of *L. mommai*), implying potential coexistence of
178 virophages and their host viruses (**Supplementary Fig. 3**). This is consistent with previous studies reporting
179 scenarios in which virophages coexist and coevolve with nucleocytoviricots in insects[44]. Mammalian genomes, such
180 as *Bubalus bubalis* (water buffalo), were also found to harbor *Lavidaviridae* HCVRs and *Megaviricetes* HCVRs
181 (**Supplementary Table 3**). Bovidae showed high levels of VR signals for these viral taxa (**Supplementary Fig. 3**).
182 These findings suggest the existence of diverse virophages in metazoans including mammals.

183 Megaviricetes are known to possess giant genomes and can form large viral insertions[13,18]. Their host range
184 varies from protists to metazoans[19], depending on the viral lineages in this class. Order *Imitervirales* in this class
185 includes only viruses infecting protists. Previously, *Imitervirales* viral signals was detected in a scaffold of a *Hydra*
186 (hydrozoan, phylum Cnidaria) assembly, implying potential viral integration[45,46]. In the present study, we identified
187 a cluster of *Imitervirales* endogenous viral elements localized to chromosome IX of *Hydra vulgaris*, with the largest
188 HCVR reaching 240 kb in length (**Fig. 4b**). From a chromosome-scale perspective, the GC content of this HCVR-
189 enriched region was distinct from that of other regions of the chromosome. Furthermore, dsDNA viral elements were
190 found to be abundant in other cnidarian assemblies (**Fig. 2, Supplementary Fig. 3**). Taken together, our results
191 expand the known dsDNA viral associations in metazoan lineages (n=176, 86.69% of all detected metazoan
192 associations), including lineages such cnidarians whose virosphere remains poorly characterized.

193 *Egovirales* is the recently discovered order within the class *Pokkesviricetes*, which includes notorious
194 pathogens such as monkeypox virus (**Table 1**), through a metagenomic survey. Frequent gene exchange between
195 egovirals and *Trichomonas* has been previously reported[23]. In the present study, we identified a total of 174
196 *Pokkesviricetes* VRs located on chromosome I of *Trichomonas vaginalis*, a common pathogenic protozoan, among
197 which 49 were classified as HCVRs (**Fig. 4c**). These VRs reached up to 52 kb in length (41 kb for the HCVR subset)
198 and are distributed evenly along the chromosome. Egovirals may integrate their genomes into the host chromosomes
199 in their lifecycle, although further studies are needed to confirm this.

**Fig. 4 | Representative HCVRs in eukaryotic genomes.**

**(a) Two *Lavidaviridae* HCVRs on a contig from *Lordiphosa mommai*.** The thick black line represents the full contig; shaded regions indicate VRs, with colors corresponding to viral taxonomic affiliations (as indicated above). GC content across the chromosome is shown above the contig (1 kb window size). The red asterisk marks the two longest HCVRs, whose magnified genomic maps are displayed below the contig. ORFs in the magnified genome maps are color-coded. **(b) Clustered *Megaviricetes* HCVRs on chromosome IX of *Hydra vulgaris*.** Similar to (a), except GC content was calculated with a 10 kb window. **(c) Abundant *Pokkesviricetes* HCVRs on chromosome I of *Trichomonas vaginalis*.** Visualization follows the same scheme as in (b).

**Paleovirology approach uncovers hidden viral lineages**

To investigate the evolutionary origins of the identified VRs, we reconstructed phylogenetic trees of VRs and reference viral sequences (**Fig. 5**). A representative marker gene that was broadly conserved among reference genomes was selected (i.e., typically involved in DNA replication or virion assembly) for each viral taxon (**Supplementary Fig. 4**). For taxa with the DNA polymerase as the phylogenetic marker, additional trees were constructed by including eukaryotic homologs (**Supplementary Fig. 5**)[47]. In the DNA polymerase-based phylogenetic trees, VR-derived sequences frequently formed monophyletic clades with reference viral sequences and exhibited clear distinctions from eukaryotic homologs, supporting the reliability of our VR identification framework. Another example confirming our VR identification is the case of the HK97 major capsid protein. The genome of a parrot (*Agapornis swindernianus*) showed two HCVRs and one VR of *Mirusviricota* (**Supplementary Table 3**), which have been associated only with unicellular eukaryotes[12,24]. The phylogenetic analysis confirmed the

221 *Mirusviricota* identity for the HK97 major capsid protein sequence derived from the parrot genome assembly.

222 Integration-like VRs were unevenly distributed in the trees. There were very few such signals in *Adenoviridae*
223 and *Papovaviricetes*. In contrast, in other viral taxa, such as *Lavidaviridae*, *Megaviricetes*, *Mriyaviricetes* and PLVs,
224 integration-like VRs were widespread and closely related to reference viral sequences, suggesting that the integration
225 of dsDNA viral sequences into eukaryotic genomes is a common phenomenon for these viral lineages.

226 We also identified clades of VRs that are distantly related to known viral taxa in the trees. For example, one
227 large VR-only clade was found to be distantly related to the *Alloherpesviridae* (**Fig. 5, Supplementary Fig. 5**),
228 which had previously been reported to integrate into the genomes of 15 fish species[48]. As comparison, our approach
229 identified homologous sequences from 82 assemblies spanning 73 fish species, substantially expanding the known
230 host range associated with this clade. For *Mriyaviricetes*, a recently discovered viral clade considered to be small
231 relatives of giant viruses[22], we identified a novel VR-only group comprising sequences from 10 distinct eukaryotic
232 species, including those from Cnidaria, Mollusca, Echinodermata, Haptophyta, and Glaucocystophyceae. Overall,
233 our results demonstrate the power of paleovirology approach in uncovering previously hidden viral lineages.

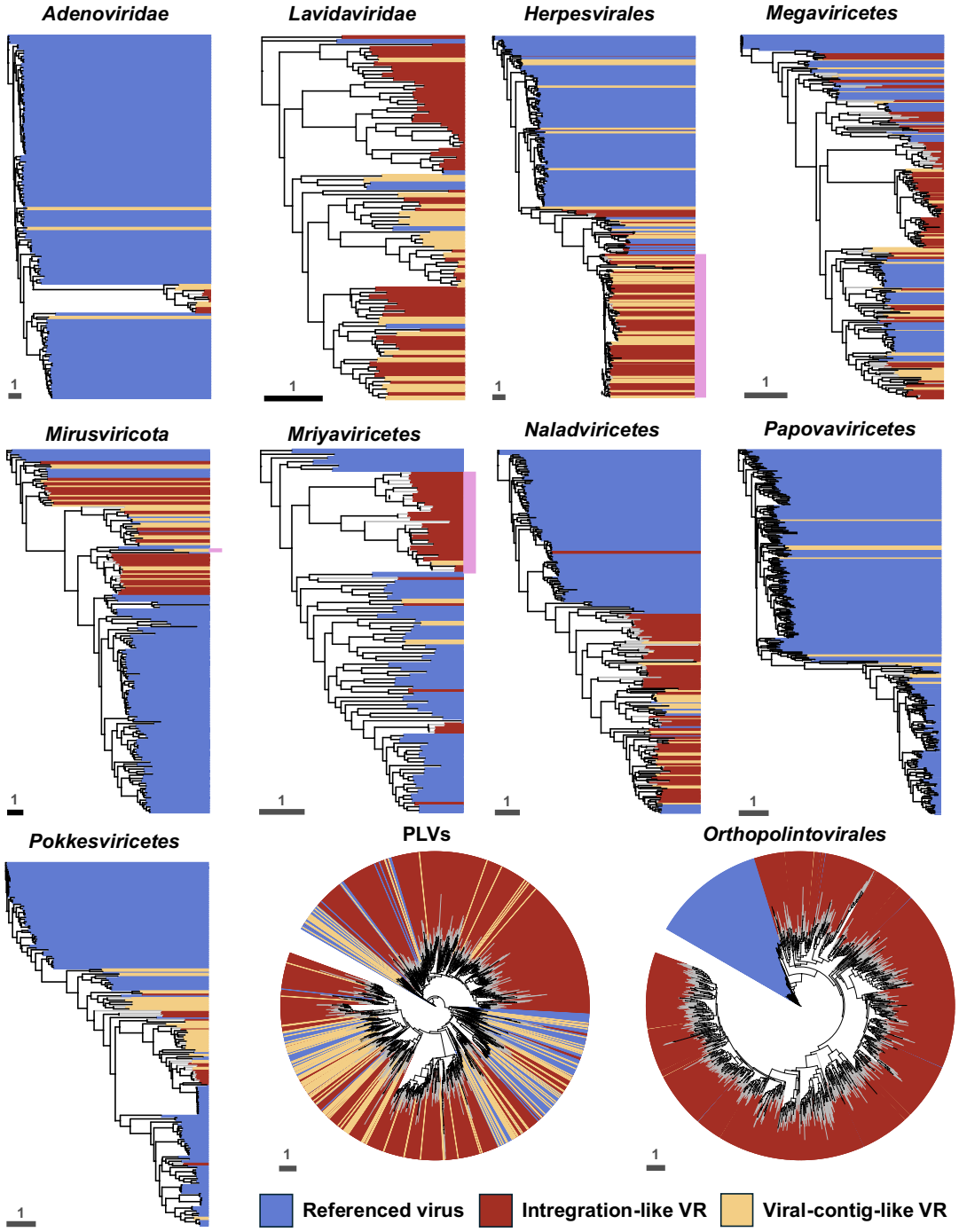



**Fig. 5 | Reconstructed phylogenetic trees of viral regions and viruses.**

Each tree corresponds to a specific viral taxon, as indicated above. Scale bars are shown in the lower left corners. Gray branches represent VRs not classified as HCVRs, and black branches represent HCVRs and reference viruses. Blue shading indicates reference viral sequences, red integration-like VRs, and yellow virus-contig-like VRs. For *Adenoviridae*, *Herpesvirales*, *Megaviricetes*, *Naldaviricetes*, and *Pokkesviricetes*, DNA polymerase sequences were used for tree reconstructions. For *Lavidaviridae*, *Mirusviricota*, *Mriyaviricetes*, and *Orthopolintovirales*, the major

241 capsid protein sequences were used. For *Papovaviricetes* and PLVs, E1 helicase and pATPase were used, respectively.
242 Trees for *Orthopolintovirales* and *Megaviricetes* were manually curated. The three examples mentioned in the main
243 text were marked by the pink rectangles.
244

245 ## Discussion

246 From retroviruses, Maverick/Polinton elements to giant endogenous viral insertions derived from
247 nucleocytoviricots, diverse parasitic genetic elements of viral origin have been identified in eukaryotic
248 genomes[3,13,49,50]. In this study, we demonstrated the widespread occurrence of dsDNA VRs across nearly all major
249 eukaryotic groups. In many eukaryotic lineages, these elements accounted for 0.1% up to 10% of the genome (**Fig.**
250 **2**). Furthermore, our results uncovered numerous previously uncharacterized virus–host associations (**Fig. 3**).

251 The virosphere is thought to harbor the greatest biological diversity and plays a pivotal role in shaping global
252 ecosystems[51]. Phylogenetic analyses revealed that endogenized VR lineages are unevenly distributed across viral
253 taxonomic trees (**Fig. 5**). Some viral lineages contain numerous isolates yet show no evidence of endogenization,
254 whereas others with few or no known isolates exhibit widespread integration signals, highlighting the existence of
255 previously unseen part of the virosphere.

256 Integrated VRs could be genetic fossils embedded in eukaryotic genomes and may reveal genomic features of
257 unknown (or past) viruses. We noticed that some VRs exhibit chimeric signatures from multiple viral taxa
258 (**Supplementary Table 3**). For example, the largest known continuous endogenous viral element is a
259 *Pokkesviricetes*-like insertion in a fungal chromosome (CP080821.1, from 1.4 to 2.9 Mb)[52]. This endogenous viral
260 element was reanalyzed in the present study and found to also contain multiple *Megaviricetes*-specific homologs.
261 Such features suggest genomic recombination of distantly related viruses, although the possibility of the existence
262 of mechanisms that make hot spots of viral integration is not excluded. Taken together, these findings suggest that
263 fossilized viral elements hold great potential to enrich our understanding of viral diversity, evolutionary history, and
264 virus–host interactions in eukaryotes[2].

265 Although most VRs are considered genomic fossils or traces of ancient viral integrations, some may represent
266 currently active viruses that influence host biology. One well-known example is *Bracoviriform*, which is distantly
267 related to *Naldaviricetes*; it generates viral particles and contributes to the life cycle of its parasitoid wasp host[53,54].
268 A brown algal virus of *Megaviricetes* is known to infect cell wall–free zoospores or gametes and persist latently in
269 all host somatic cells[55]. Endogenization of a *Lavidaviridae* virophage in *Cafeteria roenbergensis* has been shown to
270 defend its host against lethal giant virus infections[56]. Recent studies have also identified an episomal virus from

271 *Mirusviricota*[24,57] and an endogenized giant virus (i.e., punuivirus from *Megaviricetes*) that appears to produce virus-
272 like particles in single-celled hosts[43]. These findings suggest that lysogenic lifestyles may be widespread among
273 dsDNA eukaryotic viruses. In our study, we identified many VRs as large as reference viral genomes (**Fig. 1d**),
274 further characterization of which may help uncover cryptic dsDNA viral lifestyle.

275 The present study was limited to reference genomes from only 11 viral taxa at the class/phylum level. As more
276 viral groups are discovered and characterized, additional genomic regions may be recognized as of viral origins.
277 Furthermore, resolving virus–host taxonomic relationships at a finer taxonomic scale will require additional
278 methodological and/or analytical efforts. While these represent limitations of the present study, the catalog of VRs
279 we provide here will serve as a valuable resource and foundation for future investigations of viral diversity and
280 virus–host coevolutionary processes.

281

282

283 # Methods

284 **Construction of cellular reference HMMs**

285 The cellular reference HMMs were constructed in two parts. In the first part, we downloaded proteins from
286 the KEGG database, covering 778 eukaryotic species from different genera. Sequences shorter than 50 amino acids
287 or longer than 4,000 amino acids were removed. OGs were identified using OrthoFinder v2.5.5[58] and 12,555 OGs
288 present in at least 60% of animal, plant, fungal, or protist genomes were retained. From each OG, we randomly
289 selected 200 sequences for alignment using Muscle v5.1[59], trimmed with trimAl v1.4.1[60] ("-gt 0.05"), and used to
290 construct HMMs with HMMER v3.4[61].

291 The second part was based on HMMs from Pfam-A v37[62]. To remove potential viral contamination, we filtered
292 out HMMs containing virus-related keywords (i.e., "virus," "viral," "phage," "capsid," "double jelly roll,"
293 "terminase"). The remaining models were then evaluated using hmmscan (E-value < $10^{-3}$) against previously
294 collected viral and eukaryotic proteins. HMMs that produced more significant hits in eukaryotic sequences than viral
295 sequences were removed, leaving 19,991 HMMs. In total, 32,546 HMMs from both parts were retained for the
296 cellular reference models.

297

298 **Construction of VOGs and TOGs**

299 We collected reference viral genomes of the eleven eukaryotic dsDNA viral taxa from various sources. For six
300 viral taxa (*Adenoviridae*, *Herpesvirales*, *Lavidaviridae*, *Megaviricetes*, *Naldaviricetes*, and *Papovaviricetes*), we

301 obtained protein sequences from RefSeq genomes in NCBI (National Center for Biotechnology Information)-Virus
302 (as of June 10, 2024). For *Mirusviricota*, *Mriyaviricetes*, PLVs, *Pokkesviricetes* and *Orthopolintovirales*, we
303 additionally gathered genomic data from other sources (see details in Supplementary Table 2). Gene prediction was
304 performed using Prodigal v2.6.3[63] (with default parameters for large/giant viruses) and GeneMarkS v4.30[64] (with
305 the metagenomic model for smaller viruses). For each viral taxon, we identified OGs using Proteinortho v6.3.2[65]
306 with the parameters "-e = 1e-3 -identity = 15 -p = blastp+ -selfblast -cov = 40"[66]. For each taxon, we constructed
307 viral HMMs based on OGs shared by at least three genomes using the same methodology as described in the previous
308 section.

309 To determine VOGs, we performed hmmsearch using both cellular reference HMMs and viral HMMs against
310 the proteins from identified viral OGs. An OG was classified as a VOG if its proteins showed a minimum E-value
311 smaller than $10^{-5}$ for viral HMMs. In addition, the minimum E-value for cellular reference HMMs had to exceed
312 $10^{-5}$ and be at least $10^{3}$ times greater than that of the viral HMMs, ensuring viral specificity. The classification of
313 TOGs was based on VOGs. Proteins within a TOG were required to exhibit the best HMM match to viral HMMs
314 from their own taxon, while the minimum E-values for HMMs from other taxa had to exceed $10^{-5}$ and be at least $10^{3}$
315 times higher than those from other taxa, thereby confirming taxonomic specificity.

316

317 **Detection of relationships between viral taxon and eukaryotic contig**

318 We first retrieved eukaryotic taxonomic information using the Bio.Entrez module of Biopython[67], followed by
319 manual curation. For each eukaryotic assembly, Prodigal (with configurations adjusted for long input sequences)
320 and GeneMarkS were used to predict viral-like ORFs. We then performed hmmsearch to identify ORFs with
321 significant similarity to VOGs (E-value < $10^{-3}$). Since some ORFs may match multiple VOGs, only the best hit was
322 considered in each case.

323 To refine these results, we conducted a secondary hmmsearch using the -Z parameter to ensure consistency in
324 E-value calculation across different searches (Metazoa: 4,962,163; Fungi: 193,281; Plant: 237,367; Protist: 317,655).
325 If an ORF had a hit to a cellular model with an E-value less than 100 times that of its best VOG match, the hit was
326 removed to avoid potential contamination from cellular sequences. For each contig, viral-like ORFs predicted by
327 both Prodigal and GeneMarkS were cross-validated. A virus-contig pair was recorded if either (i) the contig
328 contained at least one filtered TOG hit or (ii) it contained two different filtered VOG hits. The amino acid sequence
329 of filtered contigs with viral signal were used for downstream analysis, based on the ORF predictions from the
330 method that initially detected the signal. If both methods identified viral signals, we selected the prediction method

that demonstrated better overall performance for that viral taxon in the corresponding eukaryotic group.

**Determination of viral regions**

We performed hmmsearch on all ORFs within each contig against OGs from the corresponding viral taxon and cellular references (E-value < $10^{-3}$) to obtain cellular-viral annotations. Viral regions (VRs) were identified based on these annotations and the genomic locations of ORFs. A region was classified as a VR if it contained at least two distinct viral hits for *Adenoviridae*, *Lavidaviridae*, *Mriyaviricetes*, PLVs, *Papovaviricetes*, and *Orthopolintovirales*; or (ii) at least three distinct viral hits for *Herpesvirales*, *Megaviricetes*, *Mirusviricota*, *Naldaviricetes*, and *Pokkesviricetes* In both cases, all viral hits had to originate from the most abundant viral taxon within the region. Additionally, no more than ten ORFs with cellular annotations were allowed between two viral ORFs, and the genomic distance between them could not exceed 10 kb. The length of each VR was determined from the start of the first ORF with a viral best-hit to the end of the last ORF with a viral best-hit.

A HCVR required that at least 30% of annotated ORFs within the region had viral OGs as their best hits, with at least 90% of these hits belonging to the same viral taxon. Additionally, a HCVR needed to contain at least one TOG from its assigned viral taxon while not containing any TOGs from other viral taxa.

**Annotation of viral regions**

Viral–cellular annotation tables were generated for each VR based on the minimax E-value between HMM hits from the cellular reference and OGs within the corresponding viral taxon. Genes shown in Fig. 4 were annotated using the blastp function of DIAMOND v2.1.9[68] against the NCBI Non-Redundant database, followed by manual curation. For the estimation of eukaryotic gene content within the *L. mommai* contig, annotations were retrieved using the TrEMBL database[69].

Functional annotation of VRs was performed using eggNOG-mapper v2.1.12[70]. Annotation quality was assessed by manually inspecting the assigned COG categories. Genes were defined as having unknown function if they (i) were assigned to COG categories "–", "R", or "S", or (ii) lacked hits in the eggNOG-mapper output. The average proportion of each COG category per viral region was then calculated. For PFAM domain annotation, only PFAM categories found in more than 20% of VRs were retained for downstream analysis.

**Phylogenetic analysis**

For each viral taxon, two marker genes were initially selected as candidates for phylogenetic analysis. One

361 seed sequence was collected for each marker gene from public reference datasets. The selected seed sequences are
362 listed below, with different functional genes chosen depending on the lineage. *Adenoviridae*: YP_002213842.1
363 (DNA polymerase) and AP_000211.1 (hexon); *Herpesvirales*: YP_010084960.1 (DNA polymerase) and
364 YP_010084972.1 (capsid protein); *Lavidaviridae*: YP_002122370.1 (cysteine protease) and YP_002122381.1
365 (capsid protein); *Megaviricetes*: YP_003986825.1 (DNA polymerase) and YP_003986929.1 (capsid protein);
366 *Mirusviricota*: TARA_AON_NCLDV_00056_000000000009 (DNA polymerase) and
367 TARA_AON_NCLDV_00032_000000000015 (capsid protein); *Mriyaviricetes*: HUH_long_consensus (HUH
368 endonuclease) and Yaravirus_MCP_consensus (capsid protein); *Naldaviricetes*: YP_009116659.1 (DNA polymerase)
369 and YP_009116734.1 (capsid protein); *Papovaviricetes*: NP_041327.2 (E1 helicase) and NP_041332.2 (L1 capsid
370 protein); PLVs: SAF1_17 (pATPase) and SAF1_16 (capsid protein); *Pokkesviricetes*: YP_009162448.1 (DNA
371 polymerase) and YP_009162470.1 (capsid protein); *Orthopolintovirales*: YP_010796957.1 (DNA polymerase) and
372 Polinton-1_DY (capsid protein).

373 We then compared the coverage of seed sequence alignments against the reference genome dataset. Five
374 approaches were used: (i) BLASTP (BLAST+ v2.16.0)[71] with an E-value cutoff of 1e-5; (ii) PSI-BLAST[72], where
375 all reference genes within a given lineage were pooled into one database and searched against corresponding seed
376 sequences with 5 iterations using an E-value cutoff of 1e-5; (iii) HMMER, using HMM profiles built from initial
377 reference OGs to identify marker genes, applying hmmscan to the reference sequences; (iv) JackHMMER[73], where
378 all reference genes within a lineage were pooled into one database and searched against corresponding seed
379 sequences with 5 iterations; and (v) MMseqs2 v16-747c6[74], where all seed and reference protein sequences were
380 pooled and clustered using the cluster function with a minimum sequence identity of 0.2 and a coverage threshold
381 of 0.5. The coverage of each marker gene on reference genomes was shown in Supplementary Fig. 4.

382 Due to its superior sensitivity, JackHMMER was used to identify all marker genes in the viral regions using
383 the same parameters and a database comprising VR protein sequences. Redundant sequences from viral regions were
384 removed using Cd-hit v4.8.1[75]. Most lineages were filtered using a minimum sequence identity of 0.8 and an
385 alignment coverage threshold of 0.8 for the shorter sequence. For *Orthopolintovirales*, identity and coverage
386 thresholds were set to 0.6; for PLVs, more relaxed thresholds of 0.4 were applied due to the large number of detected
387 sequences.

388 The dereplicated viral region sequences were then combined with reference sequences. Multiple sequence
389 alignments were performed using Clustal Omega v1.2.4[76] and trimmed with trimAl using a gap threshold of 0.05.
390 Aligned sequences with gaps as the predominant (90%) sites were removed. Phylogenetic trees were reconstructed

using using IQ-TREE v2.2.0[77] with the -m TEST option, allowing automatic model selection based on the Bayesian Information Criterion. The selected models for each viral lineage were as follows: *Adenoviridae*: pfam+F+I+G4; *Herpesvirales*: JTT+F+G4; *Lavidaviridae*: Blosum62+F+G4; *Megaviricetes*: VT+F+G4; *Mirusviricota*: Q.pfam+F+I+G4; *Mriyaviricetes*: VT+F+G4; *Naldaviricetes*: VT+F+I+G4; *Papovaviricetes*: Q.yeast+F+I+G4; PLVs: VT+F+G4; *Pokkesviricetes*: VT+F+I+G4; *Orthopolintovirales*: VT+F+I+G4. The phylogenetic trees were visualized with iTOL v7.1 .

## Acknowledgments

This work was supported by JSPS/KAKENHI (nos. 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).

## 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, 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.

413 ## References


414 1. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat*
415 *Rev Genet* **13**, 283–296 (2012).
416 2. Barreat, J. G. N. & Katzourakis, A. Paleovirology of the DNA viruses of eukaryotes. *Trends in Microbiology*
417 **30**, 281–292 (2022).
418 3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
419 4. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
420 5. Weiss, R. A. The discovery of endogenous retroviruses. *Retrovirology* **3**, 67 (2006).
421 6. Ashley, J. *et al.* Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172**,
422 262-274.e11 (2018).
423 7. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis.
424 *Nature* **403**, 785–789 (2000).
425 8. Ritsch, M., Brait, N., Harvey, E., Marz, M. & Lequime, S. Endogenous viral elements: insights into data
426 availability and accessibility. *Virus Evol* **10**, veae099 (2024).
427 9. Blomberg, J., Ushameckis, D. & Jern, P. Evolutionary Aspects of Human Endogenous Retroviral Sequences
428 (HERVs) and Disease. in *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, 2013).
429 10. Moniruzzaman, M. & Aylward, F. O. Endogenous DNA viruses take center stage in eukaryotic genome
430 evolution. *Proceedings of the National Academy of Sciences* **120**, e2305212120 (2023).
431 11. Bellas, C. *et al.* Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses.
432 *Proceedings of the National Academy of Sciences* **120**, e2300465120 (2023).
433 12. Zhao, H., Meng, L., Hikida, H. & Ogata, H. Eukaryotic genomic data uncover an extensive host range of
434 mirusviruses. *Current Biology* **34**, 2633-2643.e3 (2024).
435 13. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread
436 endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
437 14. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer
438 between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336 (2022).
439 15. de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. & zur Hausen, H. Classification of
440 papillomaviruses. *Virology* **324**, 17–27 (2004).
441 16. Koonin, E. V., Fischer, M. G., Kuhn, J. H. & Krupovic, M. The polinton-like supergroup of viruses: evolution,
442 molecular biology, and taxonomy. *Microbiology and Molecular Biology Reviews* **0**, e00086-23 (2024).
443 17. Fischer, M. G. The Virophage Family Lavidaviridae. *Current Issues in Molecular Biology* **40**, 1–24 (2021).
444 18. Raoult, D. *et al.* The 1.2-Megabase Genome Sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
445 19. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–
446 436 (2020).
447 20. Bellas, C. M. & Sommaruga, R. Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* **9**, 13
448 (2021).
449 21. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature* **616**, 783–789 (2023).
450 22. Yutin, N., Mutz, P., Krupovic, M. & Koonin, E. V. Mriyaviruses: small relatives of giant viruses. *mBio* **15**,
451 e01035-24 (2024).
452 23. Gaia, M. *et al.* Closest relatives of poxviruses are spread in the gut of humans and animals worldwide: the
453 egoviruses. 2024.03.23.586382 Preprint at https://doi.org/10.1101/2024.03.23.586382 (2025).
454 24. Collier, J. L. *et al.* The protist Aurantiochytrium has universal subtelomeric rDNAs and is a host for
455 mirusviruses. *Current Biology* **33**, 5199-5207.e4 (2023).
456 25. Moniruzzaman, M., Erazo-Garcia, M. P. & Aylward, F. O. Endogenous giant viruses contribute to intraspecies

genomic variability in the model green alga Chlamydomonas reinhardtii. *Virus Evolution* **8**, veac102 (2022).

26. Dowgier, G. *et al.* Sequential circulation of canine adenoviruses 1 and 2 in captive wild carnivores, France. *Vet Microbiol* **221**, 67–73 (2018).

27. Whitley, R. J. & Roizman, B. Herpes simplex virus infections. *The Lancet* **357**, 1513–1518 (2001).

28. Davison, A. J. The genome of salmonid herpesvirus 1. *J Virol* **72**, 1974–1982 (1998).

29. Zhang, X. *et al.* Structure of Sputnik, a virophage, at 3.5-Å resolution. *Proceedings of the National Academy of Sciences* **109**, 18431–18436 (2012).

30. Arslan, D., Legendre, M., Seltzer, V., Abergel, C. & Claverie, J.-M. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences* **108**, 17486–17491 (2011).

31. Piégu, B. *et al.* Genome sequence of a crustacean iridovirus, IIV31, isolated from the pill bug, Armadillidium vulgare. *Journal of General Virology* **95**, 1585–1590 (2014).

32. Boratto, P. V. M. *et al.* Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. *Proceedings of the National Academy of Sciences* **117**, 16579–16586 (2020).

33. Lightner, D. V. *et al.* Historic emergence, impact and current status of shrimp pathogens in the Americas. *Journal of Invertebrate Pathology* **110**, 174–183 (2012).

34. Lange, M. & Jehle, J. A. The genome of the *Cryptophlebia leucotreta* granulovirus. *Virology* **317**, 220–236 (2003).

35. Roitman, S. *et al.* Isolation and infection cycle of a polinton-like virus virophage in an abundant marine alga. *Nat Microbiol* **8**, 332–346 (2023).

36. Doorbar, J. & Gallimore, P. H. Identification of proteins encoded by the L1 and L2 open reading frames of human papillomavirus 1a. *Journal of Virology* **61**, 2793–2799 (1987).

37. Gazdar, A. F., Butel, J. S. & Carbone, M. SV40 and human tumours: myth, association or causality? *Nat Rev Cancer* **2**, 957–964 (2002).

38. Karki, S., Moniruzzaman, M. & Aylward, F. O. Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae. *Front Microbiol* **12**, 657471 (2021).

39. Mitjà, O. *et al.* Monkeypox. *The Lancet* **401**, 60–74 (2023).

40. Starrett, G. J. *et al.* Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. *Virus Evolution* **7**, veaa055 (2021).

41. Krupovic, M., Kuhn, J. H., Fischer, M. G. & Koonin, E. V. Natural history of eukaryotic DNA viruses with double jelly-roll major capsid proteins. *Proceedings of the National Academy of Sciences* **121**, e2405771121 (2024).

42. Kawato, S., Nozaki, R., Kondo, H. & Hirono, I. Integrase-associated niche differentiation of endogenous large DNA viruses in crustaceans. *Microbiology Spectrum* **12**, e00559-23 (2023).

43. Erazo-Garcia, M. P. *et al.* Cryptic infection of a giant virus in a unicellular green alga. *Science* **0**, eads6303 (2025).

44. Barth, Z. K. *et al.* Genomic analysis of hyperparasitic viruses associated with entomopoxviruses. *Virus Evol* **10**, veae051 (2024).

45. Aylward, F. O. & Moniruzzaman, M. ViralRecall—A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in ‘Omic Data. *Viruses* **13**, 150 (2021).

46. Filée, J. Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? *Virology* **466–467**, 53–59 (2014).

47. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P. & Venclovas, Č. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res* **48**, 10142–10156 (2020).

48. Aswad, A. & Katzourakis, A. A novel viral lineage distantly related to herpesviruses discovered within fish

genome sequence data. *Virus Evol* **3**, vex016 (2017).
49. Haapa-Paananen, S., Wahlberg, N. & Savilahti, H. Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. *Mol Phylogenet Evol* **78**, 271–274 (2014).
50. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol* **13**, 105–115 (2015).
51. Suttle, C. The viriosphere: the greatest biological diversity on Earth and driver of global processes. *Environmental Microbiology* **7**, 481–482 (2005).
52. Zhao, H. *et al.* A 1.5-Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus Rhizophagus irregularis. *Virus Evolution* **9**, vead064 (2023).
53. van Oers, M. M. *et al.* Developments in the classification and nomenclature of arthropod-infecting large DNA viruses that contain pif genes. *Arch Virol* **168**, 182 (2023).
54. Bézier, A. *et al.* Polydnaviruses of Braconid Wasps Derive from an Ancestral Nudivirus. *Science* **323**, 926–930 (2009).
55. Delaroque, N., Maier, I., Knippers, R. & Müller, D. G. Persistent virus integration into the genome of its algal host, Ectocarpus siliculosus (Phaeophyceae). *Journal of General Virology* **80**, 1367–1370 (1999).
56. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* **540**, 288–291 (2016).
57. Archibald, J. *et al.* Persistent mirusvirus infection in the marine protist Aurantiochytrium. Preprint at https://doi.org/10.21203/rs.3.rs-5686297/v1 (2025).
58. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
59. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022).
60. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
61. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011).
62. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Research* **42**, D222 (2014).
63. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
64. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research* **29**, 2607–2618 (2001).
65. Klemm, P., Stadler, P. F. & Lechner, M. Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. *Front. Bioinform.* **3**, (2023).
66. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun* **11**, 1710 (2020).
67. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
68. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
69. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28**, 45–48 (2000).
70. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular*

*Biology and Evolution* **38**, 5825–5829 (2021).

71. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

72. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

73. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

74. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

75. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

76. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**, 135–145 (2018).

77. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

78. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* **52**, W78–W82 (2024).

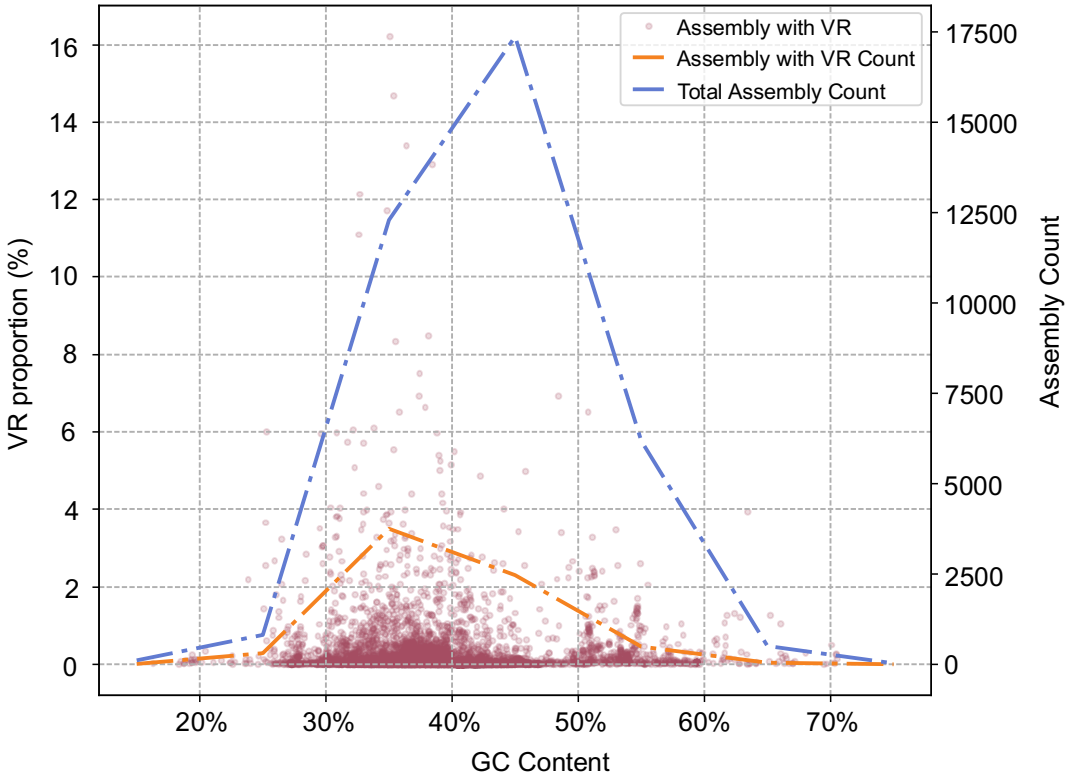
## Supplemental dataset legends

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

**Supplementary Table 2: Reference viral genomes in this study.**

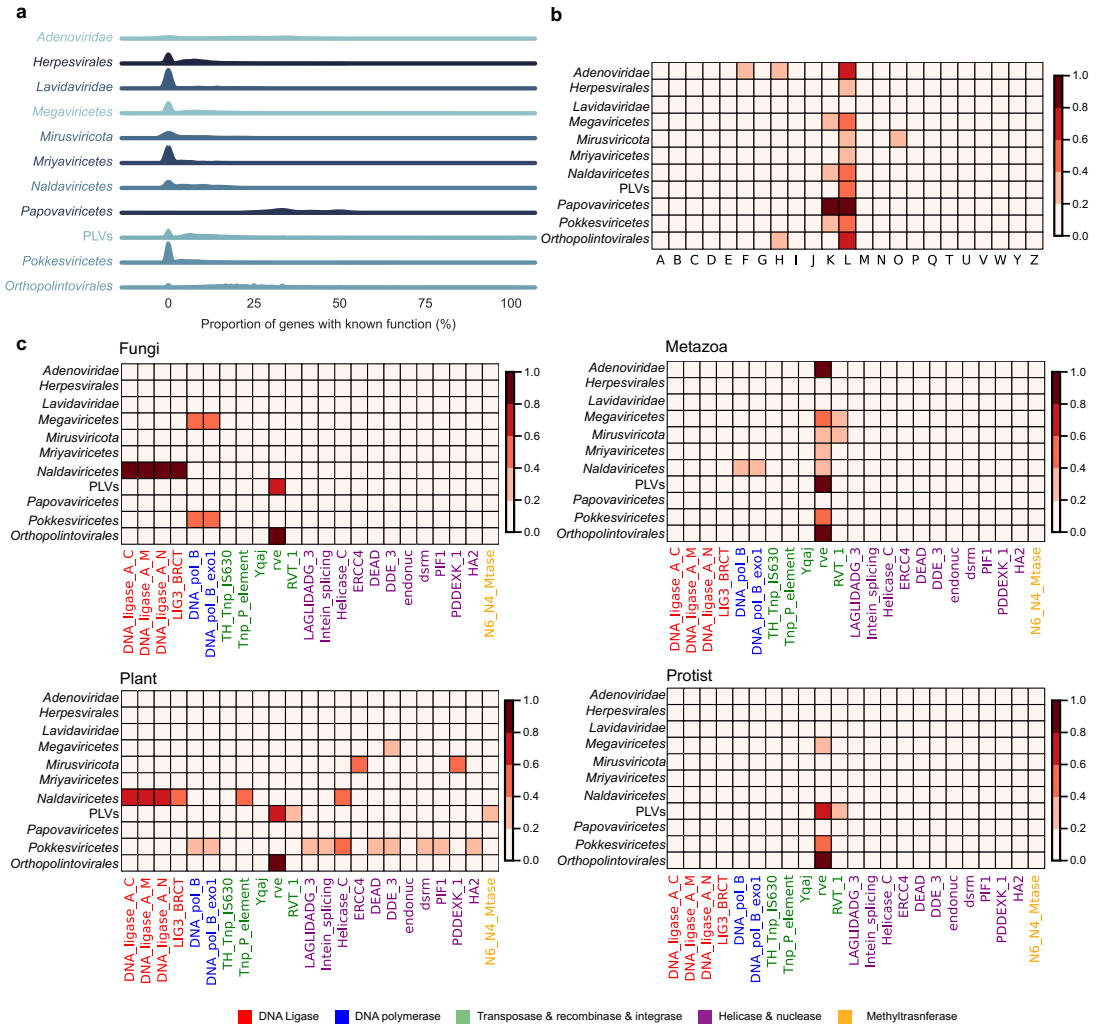
**Supplementary Table 3: Information about identified VRs.**

## Supplementary Figures



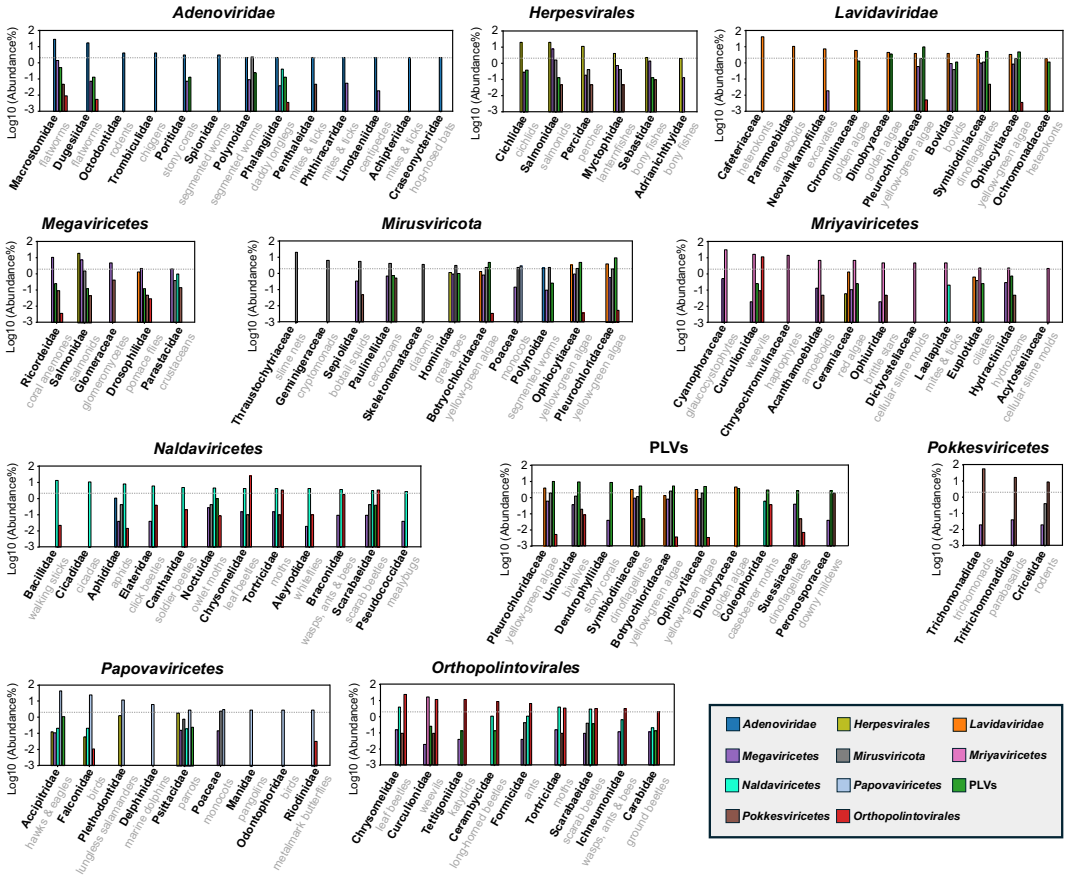

**Supplementary Fig. 1 | Relationship between genomic GC content and VR proportion.**

Each light red dot represents a single assembly containing VR. The horizontal axis indicates GC content, while the left vertical axis shows the proportion of VRs in each assembly. The blue dotted line represents the total number of assemblies within each GC content bin (ranging from 10% to 90% in 10% intervals), and the orange dotted line indicates the number of assemblies with detected VRs in each bin. Values for both dotted lines correspond to the right vertical axis.




**Supplementary Fig. 2 | Functions taken by VRs.**

(a) Proportion of functional annotated ORFs per viral region. (b) Average proportion of COG categories per VR. T The x-axis represents COG categories, the y-axis represents viral taxa, and the color of each cell indicates the average proportion. (c) Average proportion of PFAM across different host lineages. The x-axis represents PFAM models, the y-axis represents viral taxa, and the color scale indicates the average proportion. PFAM colors correspond to functional categories shown at the bottom.




585

586 **Supplementary Fig. 3 | Eukaryotic families with the highest numbers of HCVRs from each viral taxon.**

587 The vertical axis indicates the number of HCVRs in a given family as a percentage of the total HCVRs detected for

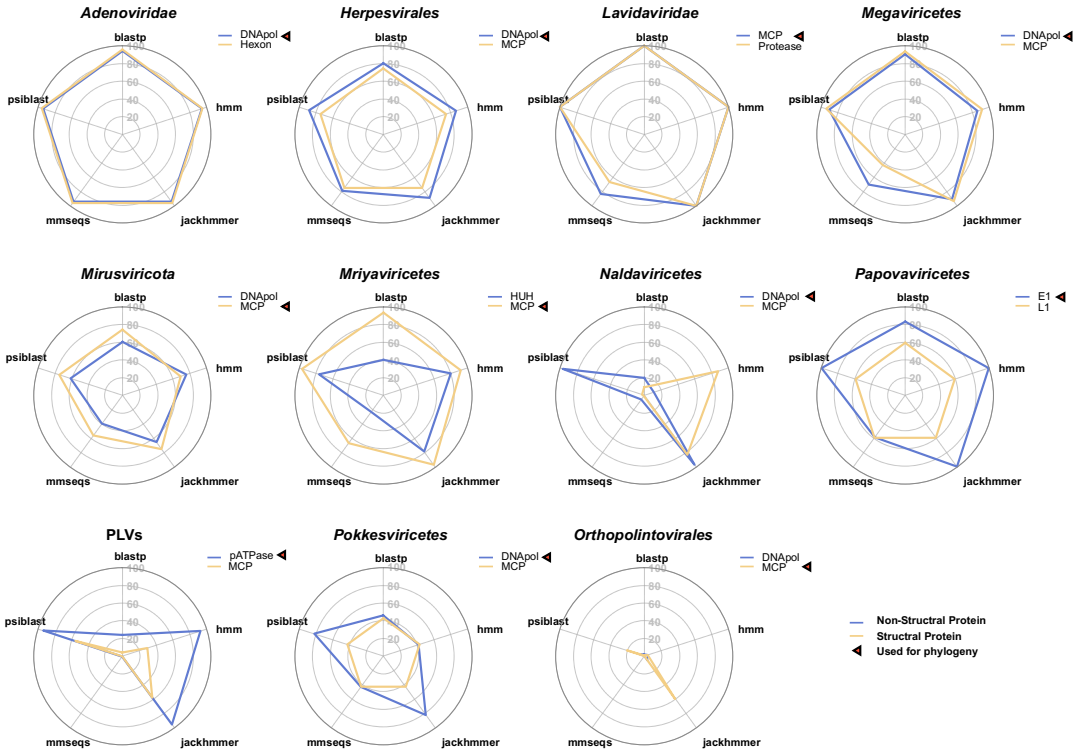
588 each viral taxon. Different colors represent distinct viral groups, as indicated in the legend at the bottom right. Each

589 bar corresponds to a eukaryotic family that contributes more than 2% of the total HCVRs for the respective viral

590 taxon, shown in descending order from left to right. Taxonomic names are shown in bold black font, and common

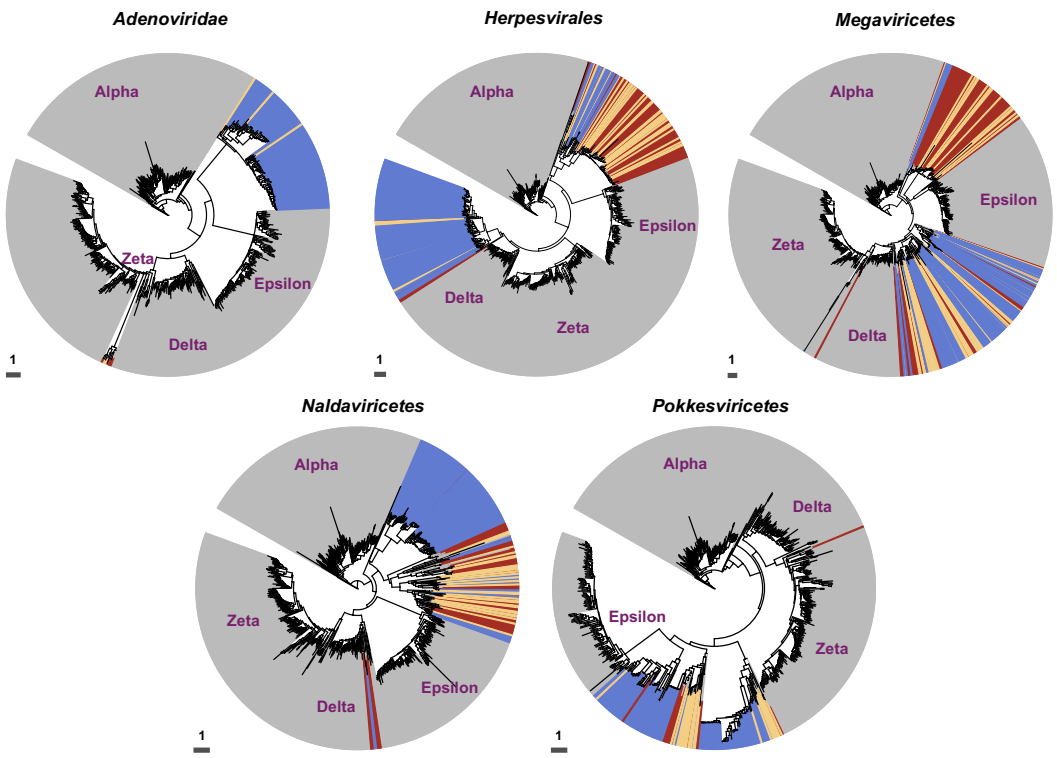
591 names are shown in gray.

592




**Supplementary Fig. 4 | Distribution of marker genes across 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). Red triangles indicate the genes selected for phylogenetic tree reconstruction.




**Supplementary Fig. 5 | Phylogenetic trees of DNA polymerase including eukaryotic references.** Each tree corresponds to a specific viral taxon, as indicated above. Scale bars are shown in the lower left corners. Gray branches represent VRs not classified as HCVRs. Blue shading indicates reference viral sequences; gray branches represent eukaryotic DNA polymerase references; yellow denotes virus-contig-like VRs; and red represents integration-like VRs. Eukaryotic DNA polymerase clades are labeled in purple text. 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*.