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)

Summary

Genomic sequences of viral origin, especially those derived from retroviruses, constitute a large proportion 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, which we used to screen 37,254 eukaryotic genome assemblies. We identified 781,111 VRs in 7,103 (19%) genome assemblies, occupying up to 16% of individual genomes, including 12% for a human protozoan pathogen. Moreover, these VRs established 343 class-level associations between viral and eukaryotic taxa, which included 305 (89%) associations for which experimental confirmation 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 expands the known dsDNA virosphere. The resulting catalogue of viral genomic elements offers opportunities for wider exploration of the significance of virus–host coevolutionary processes.

Main

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 endogenous retroviral elements can occasionally cause diseases, provide essential host functions through 'co-option' or contribute antiviral functions[6–9]. While recent studies have started to reveal genomic signatures of dsDNA viruses in protist genomes[10,11], the global extent of endogenous elements derived from this type

31 of virus in diverse eukaryotes remains underexplored[2,12].

32 Given their large coding capacity compared with RNA viruses, dsDNA viruses could substantially affect the
33 evolution of their host genomes if endogenized[13]. Evidence from a study at the protein sequence level suggests that
34 horizontal transfer of functionally diverse genes has occurred from dsDNA viruses to eukaryotes[14]. However, the
35 extent to which these viruses contribute genes and genomic regions in different eukaryotic lineages remains unclear.

36 Eukaryotic dsDNA viruses are polyphyletic and highly diverse. They range from small viruses, such as
37 papillomavirids[15] (including papillomaviruses) and viruses from the polinton-like supergroup[16,17], to giant
38 representatives like megaviricetes[18]. Furthermore, the known eukaryotic dsDNA virosphere is rapidly expanding
39 thanks to culture-independent sequencing approaches (i.e. 'metagenomics')[19], which have led to the discovery of
40 numerous novel dsDNA viral groups, including polinton-like viruses[20] (PLVs, also classified as *Aquintoviricetes*[17]),
41 *Mirusviricota*[21], *Mriyaviricetes*[22] and *Egovirales*[23]. The sequence data of these viral lineages provide new
42 opportunities to assess the impact of their integration on the evolution of eukaryotic genomes[2,10,11,24]. Nonetheless,
43 most studies performed to date have focused on specific viral or eukaryotic lineages[10,11,25], while there has been no
44 comprehensive, up-to-date framework to systematically evaluate dsDNA viral elements across eukaryotic assemblies.

45

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

47 We conducted a systematic analysis to detect dsDNA VRs across eukaryotic genomic assemblies (**Fig. 1a**). In
48 total, 37,254 assemblies representing diverse eukaryotic lineages were analysed (**Fig. 1b, Supplementary Table 1**),
49 and open reading frames (ORFs) were predicted. To aid the identification of viral signatures within eukaryotic
50 contigs, scaffolds and chromosomes (collectively referred to as 'contigs' hereafter), we also curated and constructed
51 a comprehensive reference dataset of cellular protein sequences.

52 Our study incorporated 2,126 reference viral genomes from 11 eukaryotic dsDNA viral taxa (**Table 1,**
53 **Supplementary Table 2**), most of which correspond to classes, taxonomically. Hidden Markov models (HMMs)
54 were constructed for 7,755 orthologous groups (OGs) derived from these viral genomes and searched against cellular
55 reference sequences. Among them, 1,842 viral OGs (VOGs), defined as OGs without similarity to any cellular
56 reference sequences, and 1,344 taxonomic OGs (TOGs), characterized as VOGs strictly specific to a single viral
57 taxon, were selected to capture viral and viral lineage-specific signatures, respectively. On the basis of the presence
58 of ORFs with significant similarity to VOGs or TOGs within eukaryotic contigs, we identified 946,599 virus–contig
59 pairs as potential virus–host associations.

60

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

Viral taxon	Representative viruses	Average genome size
Adenoviridae	Canine adenovirus 1[26]	33.5 kb
Herpesvirales	Herpes simplex virus[27]; Salmonid herpesvirus 1[28]	159.1 kb
Lavidaviridae	Mimivirus-dependent virus Sputnik ("virophage")[29]	22.1 kb
Megaviricetes	Megavirus chilensis[30]; Invertebrate iridescent virus 31[31]	382.4 kb
Mirusviricota	Circular genomes uncovered from eukaryotes[11,24]	217.7 kb
Mriyaviricetes	Yaravirus brasiliensis[32]	37.4 kb
Naldaviricetes	Cryptophlebia leucotreta granulovirus[33]; White spot syndrome virus[34]	134.1 kb
Papovavircetes	Human papillomavirus[35]; Simian virus 40[36]	6.6 kb
PLVs	Gezel-14T ("virophage")[37]	18.3 kb
Pokkesviricetes	African swine fever virus[38]; Monkeypox virus[39]	182.0 kb
Orthopolintovirales	Terrapene box turtle adintovirus[40]	7.8 kb

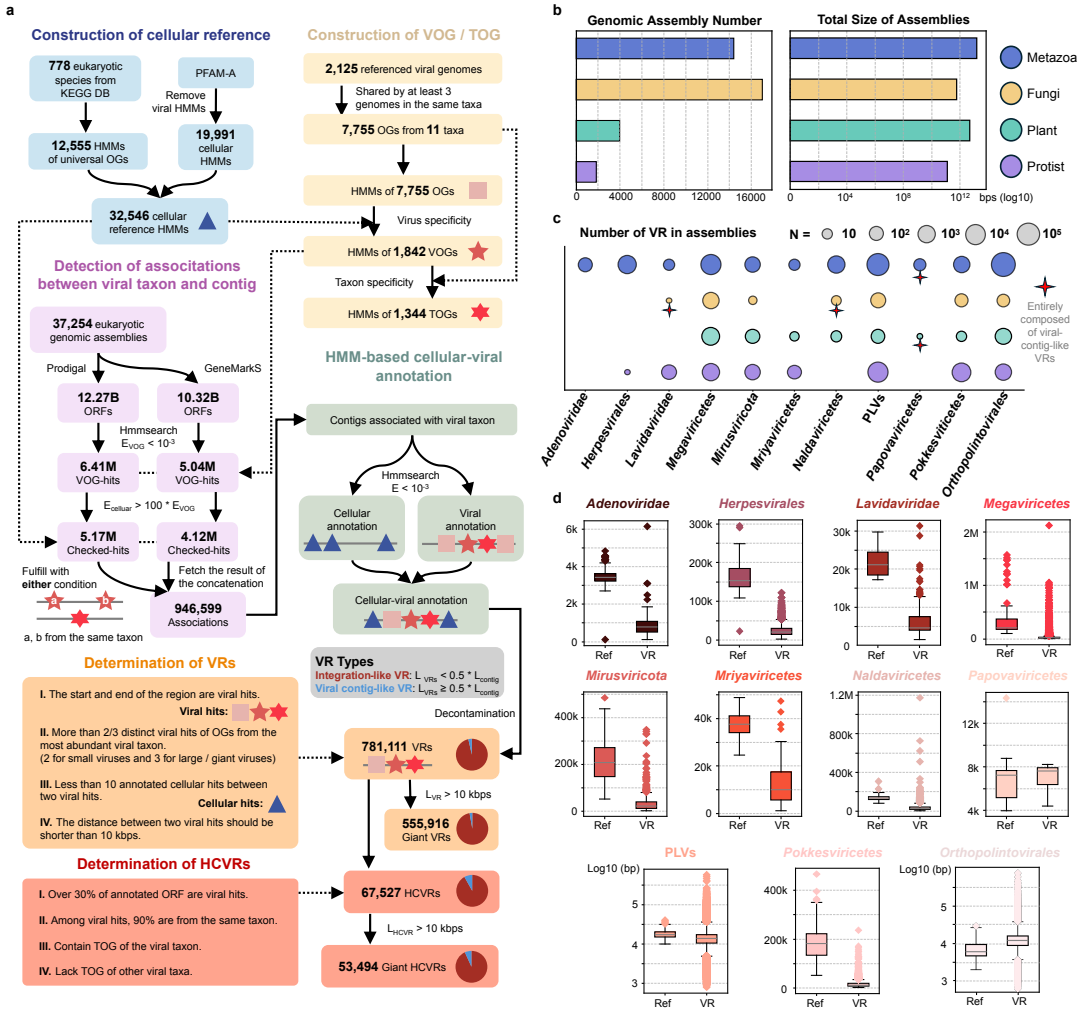


Fig. 1 | Viral regions identified in eukaryotic genome assemblies.

(a) VR discovery pipeline. Distinct background colours represent different parts of the pipeline. (b) Genomic assemblies used in this study. Colours indicate 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. Colours 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). These assemblies included 6,094 metazoan assemblies (42.25% of all metazoan assemblies) and 633 protist assemblies (34.39%) (**Fig. 1c**). Even though no dsDNA virus has been shown

experimentally to infect fungi and higher plants, 223 plant assemblies (5.59%) and 153 fungal assemblies (0.89%) were found to contain dsDNA VRs. 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%). Metazoan VRs were also taxonomically diverse and covered all viral taxa.

To distinguish likely integration events from potential viral genome contamination or unintegrated sequences, VRs were classified on the basis of 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 vast majority of VRs (96.43%) were classified into the former group (**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.

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

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

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

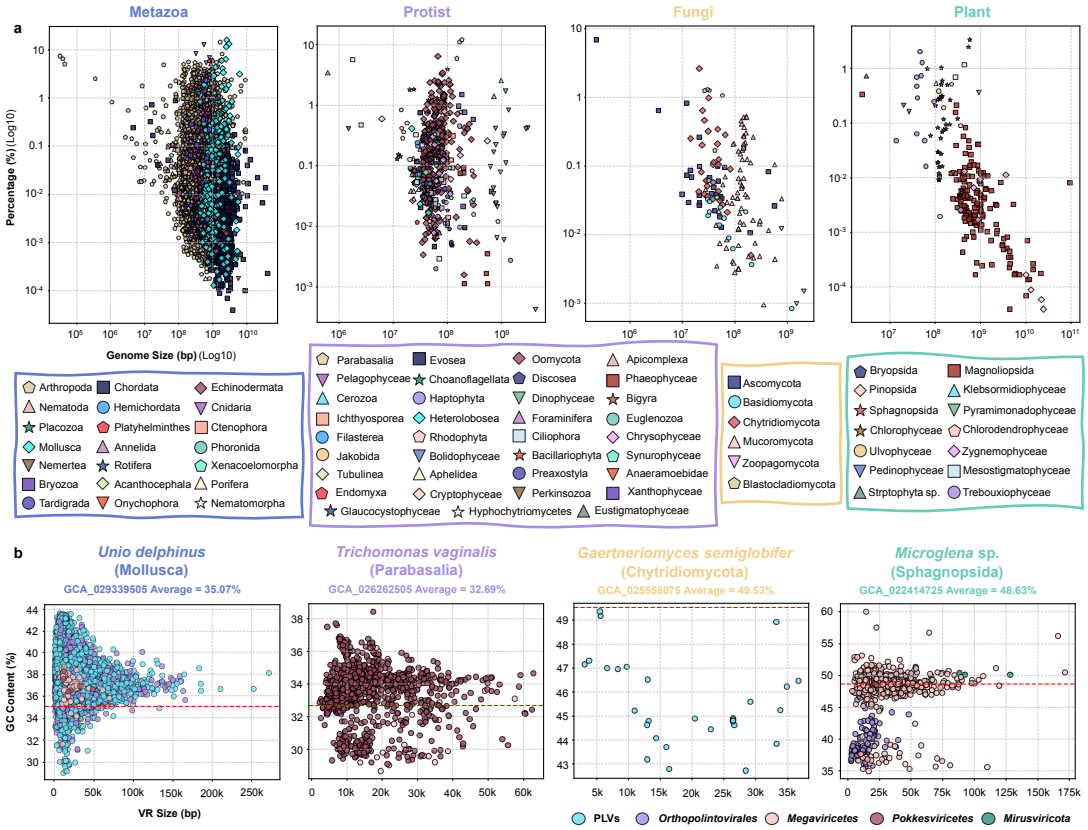


Fig. 2 | Extent of VRs in eukaryotic genomes.

(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 colour and shape corresponding to its taxon. The legend for each icon is provided in the box of the corresponding colour. 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 colours corresponding

122 to viral taxa. The horizontal axis represents VR size, while the vertical axis represents GC content. The red dashed
123 line indicates the average GC content of the respective assembly.

124 The GC content of eukaryotic assemblies negatively correlated with the abundance of VRs (**Supplementary**
125 **Fig. 1**). Among the 37,254 analysed assemblies, 46.61% (n = 17,364) exhibited GC content between 40% and 50%,
126 whereas only 34.79% (n = 2,471) of VR-containing assemblies fell within this GC range. By contrast, 32.98% (n =
127 12,288) of analysed assemblies showed GC content ranging from 30% to 40%, while 52.86% of VR-containing
128 assemblies belonged to this category. Furthermore, the majority of assemblies with a particularly high VR proportion
129 exhibited GC content in the 30%–40% range. Spearman's correlation analysis showed a significant negative
130 association between genomic GC content and VR proportion ($\rho = -0.135$, $p < 2.5 \times 10^{-30}$), suggesting a tendency
131 for dsDNA VRs to accumulate in AT-rich genomes. While the correlation is statistically significant, the small effect
132 size suggests that other factors are also related to the proportion of VRs.

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

141

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

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

152 On the basis of HCVRs, we systematically expanded the known associations between eukaryotes and dsDNA
153 viruses (**Fig. 3**). These associations may represent either present or past virus–host relationships. HCVRs were
154 identified in assemblies spanning over 30 eukaryotic phyla and 128 classes (or at the order level when the class-level
155 taxon was unavailable), resulting in 343 associations between viral and eukaryotic classes. Most (n = 277, 80.75%)
156 of these associations were supported by the presence of integration-like HCVRs. Among eukaryote–virus pairs
157 previously known from virus isolation studies, we successfully recovered 38 out of 47 associations through the
158 identification of HCVRs in their genomic assemblies, demonstrating that VR identification is an effective approach
159 for inferring virus–host relationships. Notably, many of the associations (n = 305, 88.92%) involved a viral taxon
160 from which no virus has yet been isolated for the paired eukaryotic taxon, underscoring the extent of the unexplored
161 dsDNA virosphere in eukaryotes. For example, no dsDNA viruses have been reported in fungi or streptophytes to
162 date, with their known virosphere consisting exclusively of RNA and single-stranded DNA viruses. However, our
163 data revealed their association (153 fungus and 169 streptophyte assemblies) with diverse dsDNA viruses,
164 particularly megaviricetes. Furthermore, even in well-studied lineages such as Arthropoda, Chordata and Mollusca,
165 we uncovered numerous previously unrecognized viral associations, highlighting the hidden and underestimated
166 associations between dsDNA viruses and eukaryotes.

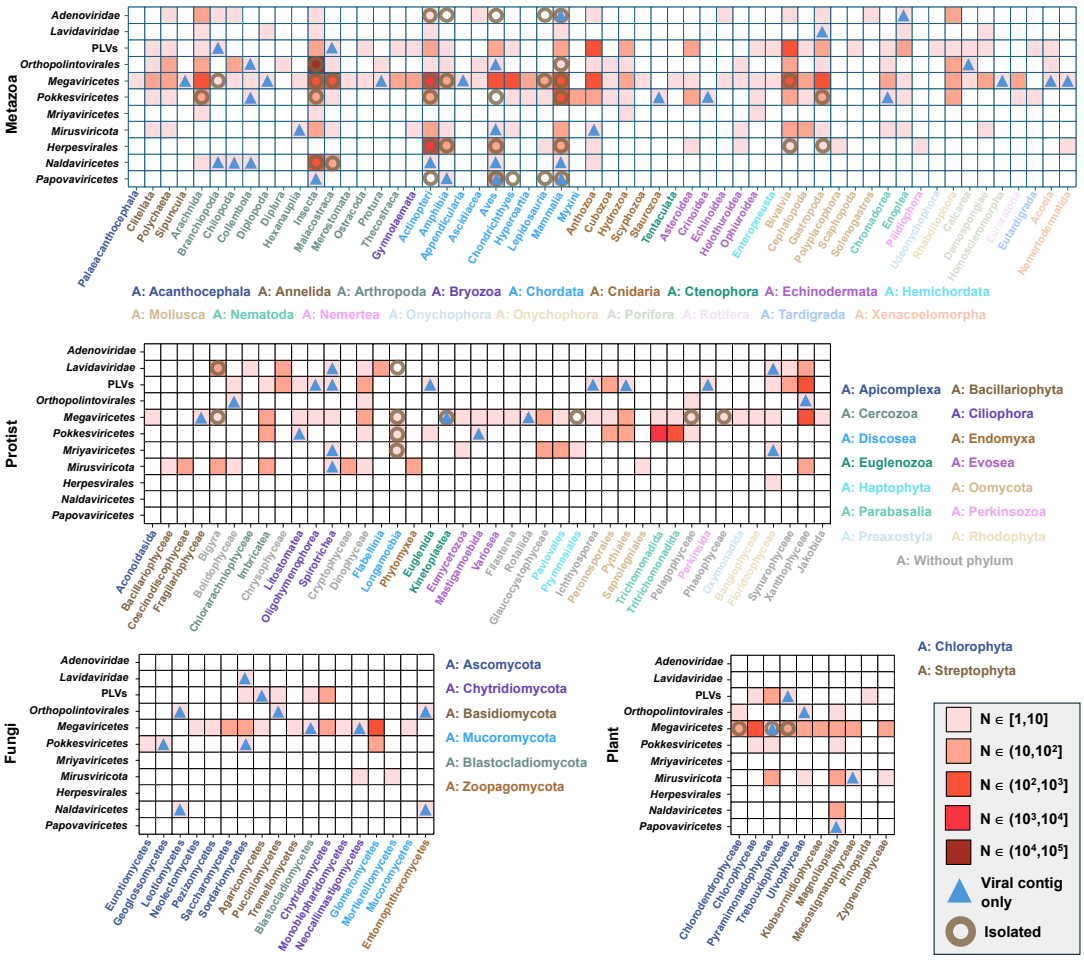


Fig. 3 | HCVR distribution in eukaryotes.

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 colour gradient indicates the number of HCVRs in each unit. The colour 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 coinfect protists together with nucleocytoviricots[16,17]. In our study, two *Lavidaviridae* HCVRs were identified within a single contig from the pomace fly, *Lordiphosa mommai* (**Fig. 4a**). Among 21 annotated ORFs on this contig, 16 showed best matches to metazoan proteins and 5 to viral proteins based on alignment annotation, supporting its eukaryotic origin. Both HCVRs contained core *Lavidaviridae* genes, including the hexon and penton capsid proteins, the packaging ATPase (pATPase), cysteine protease, but also an integrase, which could have been involved in the process of integration into the eukaryotic genome. We also observed abundant *Megaviricetes* HCVRs in Drosophilidae (the family of *L. mommai*), implying the potential coexistence of virophages

181 and their host viruses (**Supplementary Fig. 3**). This is consistent with previous studies reporting scenarios in which
182 virophages coexist and coevolve with nucleocytoviricots in insects[44]. Mammalian genomes, such as that of *Bubalus*
183 *bubalis* (water buffalo), were also found to harbour *Lavidaviridae* HCVRs and *Megaviricetes* HCVRs
184 (**Supplementary Table 3**). Bovidae showed high levels of VR signals for these viral taxa (**Supplementary Fig. 3**).
185 These findings suggest the existence of diverse virophages in metazoans including mammals.

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

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

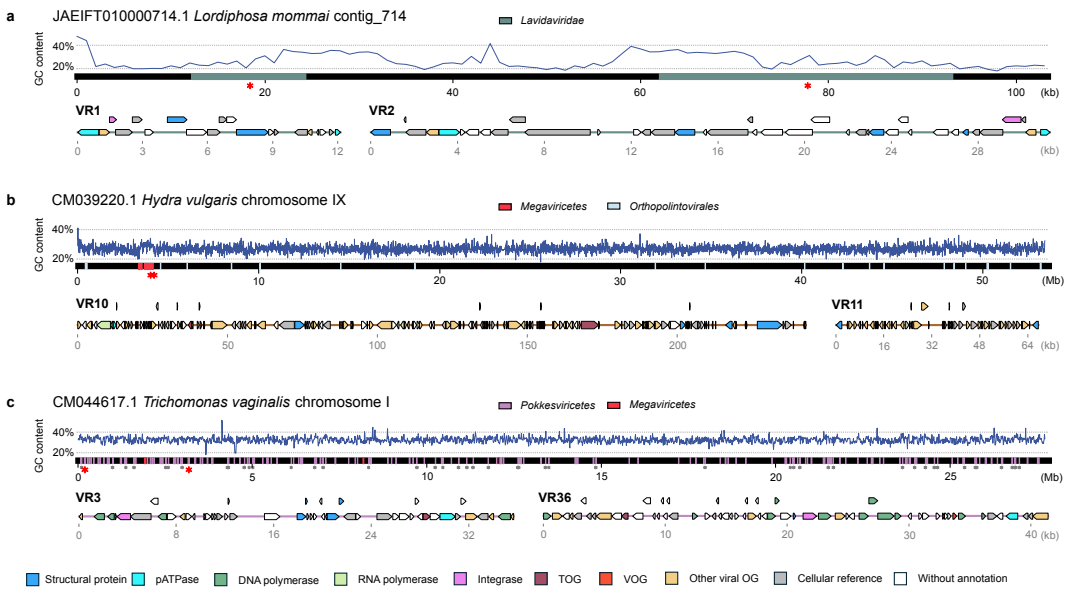


Fig. 4 | Representative HCVRs in eukaryotic genomes.

(a) Two *Lavidaviridae* HCVRs on a contig from *Lordiphosa mommai*. The thick black line represents the full contig; shaded regions indicate VRs, with colours 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 colour-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).

A palaeovirological 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 swinderianus*) showed two HCVRs and one VR of *Mirusviricota* (**Supplementary Table 3**), which have been associated only with unicellular eukaryotes[11,24]. The phylogenetic analysis confirmed the

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

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

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

236 Integrated VRs could be genetic fossils embedded in eukaryotic genomes and may reveal genomic features of
237 unknown (or extinct) viruses. We noticed that some VRs exhibit chimeric signatures from multiple viral taxa
238 (**Supplementary Table 3**). For example, the largest known continuous endogenous viral element corresponds to a
239 *Pokkesviricetes*-like insertion in a fungal chromosome (CP080821.1, from 1.4 to 2.9 Mb)[49]. The same endogenous
240 viral element also contains multiple *Megaviricetes*-specific homologs. This suggests genomic recombination of
241 distantly related viruses. Overall, our results demonstrate the power of a palaeovirological approach in uncovering
242 previously hidden viral lineages and providing novel insights into viral evolution.

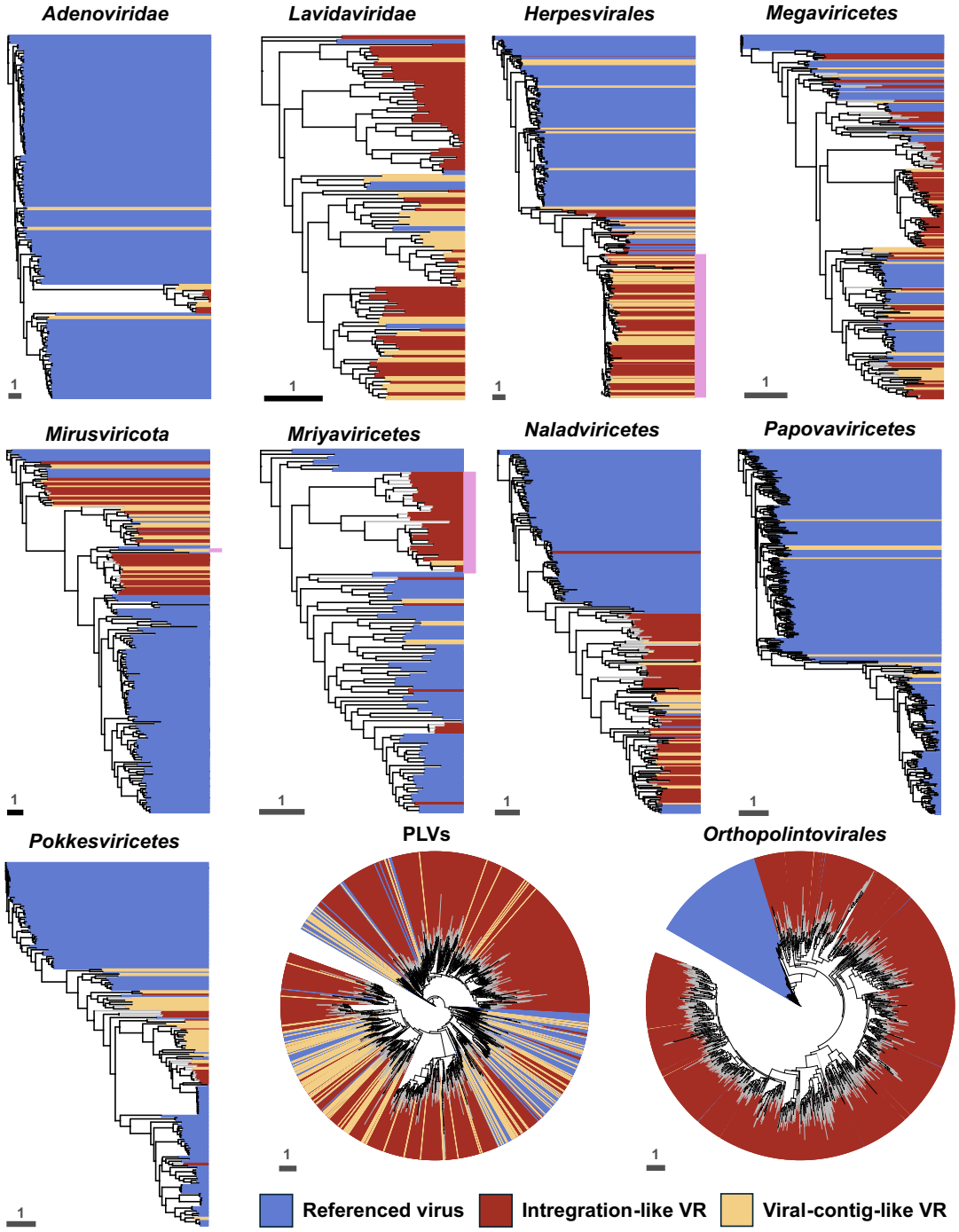


Fig. 5 | Reconstructed phylogenetic trees of viral regions and viruses.

Each tree corresponds to a specific viral taxon, as indicated above. Scale bars are shown in the lower left corners. Grey 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

250 capsid protein sequences were used. For *Papovaviricetes* and PLVs, E1 helicase and pATPase were used, respectively.
251 Trees for *Orthopolintovirales* and *Megaviricetes* were manually curated. The three examples mentioned in the main
252 text are marked by pink rectangles.
253

254 ## Discussion

255 The virosphere is thought to harbour the greatest biological diversity and plays a pivotal role in shaping global
256 ecosystems[50]. In addition to being abundant in the environment[51], viruses leave traces of their existence within host
257 genomes and occasionally influence host biology and evolution[1,52,53]. From retroviruses and Maverick/Polinton
258 elements to giant endogenous viral insertions derived from nucleocytoviricots, diverse parasitic genetic elements of
259 viral origin have now been identified in eukaryotic genomes[3,13,54,55]. However, the lack of systematic surveys has
260 limited our understanding of the phylogenetic breadth of this phenomenon. In this study, we demonstrated the
261 widespread occurrence of dsDNA VRs across nearly all major eukaryotic groups. In many eukaryotic lineages (i.e.
262 2,347 assemblies), these elements accounted for 0.1%–10% of the genome (**Fig. 2**). By resolving the strongly
263 supported taxonomic affiliations of these viral signatures, our results uncovered numerous previously
264 uncharacterized virus–host associations (**Fig. 3**). Phylogenetic analyses revealed that endogenized VR lineages are
265 unevenly distributed across viral taxonomic trees (**Fig. 5**). Some viral lineages contain numerous isolates yet show
266 no evidence of endogenization, whereas others with few or no known isolates exhibit widespread integration signals.
267 These patterns highlight the existence of previously unrecognized features of the virosphere.

268 Viruses have undergone long a period of coevolution with their hosts to gradually adapt to diverse hosts[19,56,57].
269 As distinct eukaryotes evolved different lifestyles and antiviral defences[58,59], their associated viruses also specialised
270 accordingly. Our results seem to reflect this trend (**Fig. 3**). As for land plants and fungi, it has been suggested that
271 structural features such as cell walls and plasmodesmata act as barriers against large dsDNA viruses[60], which may
272 explain the observed scarcity of dsDNA viral signals. Protists, which encompass most of the major eukaryotic
273 supergroups[61], exhibited substantial viral diversity despite representing the group with the fewest assemblies in our
274 dataset (**Fig. 1b and 1c**). This finding underscores the vast, largely unexplored diversity of the protist virosphere and
275 highlights the need for further studies on protist–virus interactions and antiviral mechanisms. Metazoans, vertebrates
276 in particular, have evolved diverse and complex antiviral immune systems[62,63]. These defences may limit dsDNA
277 viral infections and purge endogenous viral elements[64], explaining the relatively low density of dsDNA viral
278 sequences observed in vertebrate genomes. However, some invertebrates (e.g. molluscs, arthropods, cnidarians, and
279 flatworms) may lack such defences, which may account for higher proportions of dsDNA viral elements in their

genomes (**Fig. 2**). In addition, naldaviricetes are generally considered arthropod-specific viruses[65]. Consistent with this, most of the observed associations for *Naldaviricetes* were with arthropod genome assemblies. However, rare associations were also detected in vertebrates, anthozoans, fungi, and magnoliopsids (**Supplementary Table 3**). Although the possibility of contamination through incorporation of exogenous elements cannot be ruled out for some cases, these associations and their VRs provide valuable insights into the evolutionary history and host range expansions of this viral clade. Together, these findings highlight how host ecological and immunological characteristics have shaped the evolutionary trajectories and diversity of endogenous dsDNA viruses.

Although most VRs are considered genomic fossils or traces of ancient viral integration, some may represent currently active viruses or viral elements that influence host biology. One well-known example is *Bracoviriform*, which is distantly related to *Naldaviricetes*; it generates viral particles and contributes to the life cycle of its parasitoid wasp host[65,66]. Meanwhile, a brown algal virus of *Megaviricetes* infects cell wall-free zoospores or gametes and persists latently in all host somatic cells[67]. Moreover, endogenization of a *Lavidaviridae* virophage in *Cafeteria roenbergensis* has been shown to defend its host against lethal giant virus infections[68]. Recent studies have also identified an episomal virus from *Mirusviricota*[24,69] and an endogenized giant virus (i.e. punuivirus from *Megaviricetes*) that appear to produce virus-like particles in single-celled hosts[43]. These findings suggest that lysogenic lifestyles may be widespread among dsDNA eukaryotic viruses. In our study, we identified many VRs as large as reference viral genomes (**Fig. 1d**), further characterization of which may help uncover cryptic dsDNA viral lifestyles.

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

Methods

Construction of cellular reference HMMs

The cellular reference HMMs were constructed in two steps. First, we downloaded proteins from the KEGG database[70], covering 778 eukaryotic species from different genera. Sequences shorter than 50 amino acids or longer than 4,000 amino acids were removed. OGs were identified using OrthoFinder v2.5.5[71] and 12,555 OGs present in

310 at least 60% of animal, plant, fungal or protist genomes were retained. From each OG, we randomly selected 200
311 sequences for alignment using Muscle v5.1[72], trimmed them with trimAl v1.4.1[73] ('-gt 0.05') and used them to
312 construct HMMs with HMMER v3.4[74].

313 The second step was based on HMMs from Pfam-A v37[75]. To remove potential viral contamination, we filtered
314 out HMMs containing virus-related keywords (i.e. 'virus', 'viral', 'phage', 'capsid', 'double jelly roll', 'terminase').
315 The remaining models were then evaluated using hmmscan (E-value < 10^{-3}) against previously collected viral and
316 eukaryotic proteins. HMMs that produced more significant hits in eukaryotic sequences than viral sequences were
317 removed, leaving 19,991 HMMs. In total, 32,546 HMMs from both steps were retained for the cellular reference
318 models.

319

320 **Construction of VOGs and TOGs**

321 We collected reference viral genomes of the 11 eukaryotic dsDNA viral taxa from various sources. For six
322 viral taxa (*Adenoviridae*, *Herpesvirales*, *Lavidaviridae*, *Megaviricetes*, *Naldaviricetes* and *Papovaviricetes*), we
323 obtained protein sequences from RefSeq genomes in NCBI (National Center for Biotechnology Information)-Virus
324 (as of June 10, 2024). For *Mirusviricota*, *Mriyaviricetes*, PLVs, *Pokkesviricetes* and *Orthopolintovirales*, we
325 additionally gathered genomic data from other sources (see details in Supplementary Table 2). Gene prediction was
326 performed using Prodigal v2.6.3[76] (with default parameters for large/giant viruses) and GeneMarkS v4.30[77] (with
327 the metagenomic model for smaller viruses). For each viral taxon, we identified OGs using Proteinortho v6.3.2[78]
328 with the following parameters: -e = 1e-3 -identity = 15 -p = blastp+ -selfblast -cov = 40[79]. For each taxon, we
329 constructed viral HMMs based on OGs shared by at least three genomes using the same methodology as described
330 in the previous section.

331 To determine VOGs, we performed hmmsearch using both cellular reference HMMs and viral HMMs against
332 the proteins from identified viral OGs. An OG was classified as a VOG if its proteins showed a minimum E-value
333 smaller than 10^{-5} for viral HMMs. In addition, the minimum E-value for cellular reference HMMs had to exceed
334 10^{-5} and be at least 10^3 times greater than that of the viral HMMs, ensuring viral specificity. The classification of
335 TOGs was based on VOGs. Proteins within a TOG were required to exhibit the best HMM match to viral HMMs
336 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
337 times higher than those from other taxa, thereby confirming taxonomic specificity.

338

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

340 We first retrieved eukaryotic taxonomic information using the Bio.Entrez module of Biopython[80], followed by
341 manual curation. For each eukaryotic assembly, Prodigal (with configurations adjusted for long input sequences)
342 and GeneMarkS were used to predict viral-like ORFs. We then performed hmmsearch to identify ORFs with
343 significant similarity to VOGs (E-value < 10^{-3}). Because some ORFs may match multiple VOGs, only the best hit
344 was considered in each case.

345 To refine these results, we conducted a secondary hmmsearch using the -Z parameter to ensure consistency in
346 E-value calculation across different searches (Metazoa: 4,962,163; Fungus: 193,281; Plant: 237,367; Protist:
347 317,655). If an ORF had a hit to a cellular model with an E-value less than 100 times that of its best VOG match,
348 the hit was removed to avoid potential contamination from cellular sequences. For each contig, viral-like ORFs
349 predicted by both Prodigal and GeneMarkS were cross-validated. A virus–contig pair was recorded if the contig
350 either (i) contained at least one filtered TOG hit or (ii) contained two different filtered VOG hits. The amino acid
351 sequences of filtered contigs with a viral signal were used for downstream analysis, based on the ORF predictions
352 from the method that initially detected the signal. If both methods identified viral signals, we selected the prediction
353 method that demonstrated better overall performance for that viral taxon in the corresponding eukaryotic group.

354

355 **Determination of viral regions**

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

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

368

369 **Annotation of viral regions**

370 Viral–cellular annotation tables were generated for each VR based on the minimax E-value between HMM
371 hits from the cellular reference and OGs within the corresponding viral taxon. Genes shown in Fig. 4 were annotated
372 using the blastp function of DIAMOND v2.1.9[81] against the NCBI Non-Redundant database, followed by manual
373 curation. For the estimation of eukaryotic gene content within the *L. mommai* contig, annotations were retrieved
374 using the TrEMBL database[82].

375 Functional annotation of VRs was performed using eggNOG-mapper v2.1.12[74]. Annotation quality was
376 assessed by manually inspecting the assigned COG categories. Genes were defined as having unknown function if
377 they (i) were assigned to COG categories '–', 'R' or 'S', or (ii) lacked hits in the eggNOG-mapper output. The
378 average proportion of each COG category per viral region was then calculated. For PFAM domain annotation, only
379 PFAM categories found in more than 20% of VRs were retained for downstream analysis.

380

381 **Phylogenetic analysis**

382 For each viral taxon, two marker genes were initially selected as candidates for phylogenetic analysis. One
383 seed sequence was collected for each marker gene from public reference datasets. The selected seed sequences are
384 listed below, with different functional genes chosen depending on the lineage: *Adenoviridae*: YP_002213842.1
385 (DNA polymerase) and AP_000211.1 (hexon); *Herpesvirales*: YP_010084960.1 (DNA polymerase) and
386 YP_010084972.1 (capsid protein); *Lavidaviridae*: YP_002122370.1 (cysteine protease) and YP_002122381.1
387 (capsid protein); *Megaviricetes*: YP_003986825.1 (DNA polymerase) and YP_003986929.1 (capsid protein);
388 *Mirusviricota*: TARA_AON_NCLDV_00056_000000000009 (DNA polymerase) and
389 TARA_AON_NCLDV_00032_000000000015 (capsid protein); *Mriyaviricetes*: HUH_long_consensus (HUH
390 endonuclease) and Yaravirus_MCP_consensus (capsid protein); *Naldaviricetes*: YP_009116659.1 (DNA polymerase)
391 and YP_009116734.1 (capsid protein); *Papovaviricetes*: NP_041327.2 (E1 helicase) and NP_041332.2 (L1 capsid
392 protein); PLVs: SAF1_17 (pATPase) and SAF1_16 (capsid protein); *Pokkesviricetes*: YP_009162448.1 (DNA
393 polymerase) and YP_009162470.1 (capsid protein); *Orthopolintovirales*: YP_010796957.1 (DNA polymerase) and
394 Polinton-1_DY (capsid protein).

395 We then compared the coverage of seed sequence alignments against the reference genome dataset. Five
396 approaches were used: (i) BLASTP (BLAST+ v2.16.0[83]) with an E-value cut-off of 1e−5; (ii) PSI-BLAST[84], where
397 all reference genes within a given lineage were pooled into one database and searched against corresponding seed
398 sequences with five iterations using an E-value cut-off of 1e−5; (iii) HMMER, using HMM profiles built from initial
399 reference OGs to identify marker genes, applying hmmscan to the reference sequences; (iv) JackHMMER[85], where

all reference genes within a lineage were pooled into one database and searched against corresponding seed sequences with five iterations; and (v) MMseqs2 v16-747c6[86], where all seed and reference protein sequences were pooled and clustered using the cluster function with a minimum sequence identity of 0.2 and a coverage threshold of 0.5. The coverage of each marker gene on reference genomes is shown in Supplementary Fig. 4.

Owing to its superior sensitivity, JackHMMER was used to identify all marker genes in the viral regions using the same parameters and a database comprising VR protein sequences. Redundant sequences from viral regions were removed using Cd-hit v4.8.1[87]. Most lineages were filtered using a minimum sequence identity of 0.8 and an alignment coverage threshold of 0.8 for the shorter sequence. For *Orthopolintovirales*, identity and coverage thresholds were set to 0.6; for PLVs, more relaxed thresholds of 0.4 were applied due to the large number of detected sequences.

The dereplicated viral region sequences were then combined with reference sequences. Multiple sequence alignments were performed using Clustal Omega v1.2.4[87] and trimmed with trimAl using a gap threshold of 0.05. Aligned sequences with gaps as the predominant (90%) sites were removed. Phylogenetic trees were reconstructed using IQ-TREE v2.2.0[88] 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; and *Orthopolintovirales*: VT+F+I+G4. The phylogenetic trees were visualized with iTOL v7.1[89].

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.

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.

436 ## References

437 1. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat.*
438 *Rev. Genet.* **13**, 283–296 (2012).
439 2. Barreat, J. G. N. & Katzourakis, A. Paleovirology of the DNA viruses of eukaryotes. *Trends Microbiol.* **30**,
440 281–292 (2022).
441 3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
442 4. Weiss, R. A. The discovery of endogenous retroviruses. *Retrovirology* **3**, 67 (2006).
443 5. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
444 6. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis.
445 *Nature* **403**, 785–789 (2000).
446 7. Blomberg, J., Ushameckis, D. & Jern, P. Evolutionary Aspects of Human Endogenous Retroviral Sequences
447 (HERVs) and Disease. *Retroviruses and primate genome evolution*, 204-238 (2005).
448 8. Ashley, J. *et al.* Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172**,
449 262-274.e11 (2018).
450 9. Ritsch, M., Brait, N., Harvey, E., Marz, M. & Lequime, S. Endogenous viral elements: insights into data
451 availability and accessibility. *Virus Evol.* **10**, veae099 (2024).
452 10. Bellas, C. *et al.* Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proc.*
453 *Natl. Acad. Sci. USA* **120**, e2300465120 (2023).
454 11. Zhao, H., Meng, L., Hikida, H. & Ogata, H. Eukaryotic genomic data uncover an extensive host range of
455 mirusviruses. *Curr. Biol.* **34**, 2633-2643.e3 (2024).
456 12. Moniruzzaman, M. & Aylward, F. O. Endogenous DNA viruses take center stage in eukaryotic genome
457 evolution. *Proc. Natl. Acad. Sci. USA* **120**, e2305212120 (2023).
458 13. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread
459 endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
460 14. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer
461 between eukaryotes and viruses. *Nat. Microbiol.* **7**, 327–336 (2022).
462 15. de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. & zur Hausen, H. Classification of
463 papillomaviruses. *Virology* **324**, 17–27 (2004).
464 16. Fischer, M. G. The Virophage Family Lavidaviridae. *Curr. Issues Mol. Biol.* **40**, 1–24 (2021).
465 17. Koonin, E. V., Fischer, M. G., Kuhn, J. H. & Krupovic, M. The polinton-like supergroup of viruses: evolution,
466 molecular biology, and taxonomy. *Microbiol. Mol. Biol. Rev.* **88**, e00086-23 (2024).
467 18. Raoult, D. *et al.* The 1.2-Megabase Genome Sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
468 19. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–
469 436 (2020).
470 20. Bellas, C. M. & Sommaruga, R. Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* **9**, 13
471 (2021).
472 21. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature* **616**, 783–789 (2023).
473 22. Yutin, N., Mutz, P., Krupovic, M. & Koonin, E. V. Mriyaviruses: small relatives of giant viruses. *mBio* **15**,
474 e01035-24 (2024).
475 23. Gaia, M. *et al.* Closest relatives of poxviruses are spread in the gut of humans and animals worldwide: the
476 egoviruses. 2024.03.23.586382 Preprint at https://doi.org/10.1101/2024.03.23.586382 (2025).
477 24. Collier, J. L. *et al.* The protist Aurantiochytrium has universal subtelomeric rDNAs and is a host for
478 mirusviruses. *Curr. Biol.* **33**, 5199-5207.e4 (2023).
479 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 Evol.* **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. *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. *Proc. Natl. Acad. Sci. USA* **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. *Proc. Natl. Acad. Sci. USA* **108**, 17486–17491 (2011).
31. Piégu, B. *et al.* Genome sequence of a crustacean iridovirus, IIV31, isolated from the pill bug, Armadillidium vulgare. *J. Gen. Virol.* **95**, 1585–1590 (2014).
32. Boratto, P. V. M. *et al.* Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. *Proc. Natl. Acad. Sci. USA* **117**, 16579–16586 (2020).
33. Lange, M. & Jehle, J. A. The genome of the *Cryptophlebia leucotreta* granulovirus. *Virology* **317**, 220–236 (2003).
34. Lightner, D. V. *et al.* Historic emergence, impact and current status of shrimp pathogens in the Americas. *J. Invertebr. Pathol.* **110**, 174–183 (2012).
35. Doorbar, J. & Gallimore, P. H. Identification of proteins encoded by the L1 and L2 open reading frames of human papillomavirus 1a. *J. Virol.* **61**, 2793–2799 (1987).
36. Gazdar, A. F., Butel, J. S. & Carbone, M. SV40 and human tumours: myth, association or causality? *Nat. Rev. Cancer* **2**, 957–964 (2002).
37. 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).
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. *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 Evol.* **7**, veaa055 (2021).
41. Kawato, S., Nozaki, R., Kondo, H. & Hirono, I. Integrase-associated niche differentiation of endogenous large DNA viruses in crustaceans. *Microbiol. Spectr.* **12**, e00559-23 (2023).
42. Krupovic, M., Kuhn, J. H., Fischer, M. G. & Koonin, E. V. Natural history of eukaryotic DNA viruses with double jelly-roll major capsid proteins. *Proc. Natl. Acad. Sci. USA* **121**, e2405771121 (2024).
43. Erazo-Garcia, M. P. *et al.* Cryptic infection of a giant virus in a unicellular green alga. *Science*, eads6303 (2025).
44. Barth, Z. K. *et al.* Genomic analysis of hyperparasitic viruses associated with entomopoxviruses. *Virus Evol.* **10**, veae051 (2024).
45. 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).
46. 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).
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).

524 49. Zhao, H. *et al.* A 1.5-Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus
525 Rhizophagus irregularis. *Virus Evol.* **9**, vead064 (2023).
526 50. Suttle, C. The viriosphere: the greatest biological diversity on Earth and driver of global processes. *Environ.*
527 *Microbiol.* **7**, 481–482 (2005).
528 51. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).
529 52. Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev.*
530 *Microbiol.* **17**, 355–370 (2019).
531 53. Dopkins, N. & Nixon, D. F. Activation of human endogenous retroviruses and its physiological consequences.
532 *Nat. Rev. Mol. Cell Biol.* **25**, 212–222 (2024).
533 54. Haapa-Paananen, S., Wahlberg, N. & Savilahti, H. Phylogenetic analysis of Maverick/Polinton giant
534 transposons across organisms. *Mol. Phylogenet. Evol.* **78**, 271–274 (2014).
535 55. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat.*
536 *Rev. Microbiol.* **13**, 105–115 (2015).
537 56. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war — host adaptation and its constraints on
538 virus evolution. *Nat. Rev. Microbiol.* **17**, 321–328 (2019).
539 57. Koonin, E. V., Dolja, V. V. & Krupovic, M. The logic of virus evolution. *Cell Host Microbe* **30**, 917–929
540 (2022).
541 58. Koonin, E. V. Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes:
542 common ancestry vs convergence. *Biol. Direct* **12**, 5 (2017).
543 59. Marques, J. T., Meignin, C. & Imler, J.-L. An evolutionary perspective to innate antiviral immunity in animals.
544 *Cell Rep.* **43**, (2024).
545 60. Dolja, V. V., Krupovic, M. & Koonin, E. V. Deep Roots and Splendid Boughs of the Global Plant Virome.
546 *Annu. Rev. Phytopathol.* **58**, 23–53 (2020).
547 61. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. *Trends Ecol. Evol.*
548 **35**, 43–55 (2020).
549 62. Paludan, S. R., Pradeu, T., Masters, S. L. & Mogensen, T. H. Constitutive immune mechanisms: mediators of
550 host defence and immune regulation. *Nat. Rev. Immunol.* **21**, 137–150 (2021).
551 63. Kassiotis, G. The Immunological Conundrum of Endogenous Retroelements. *Annu. Rev. Immunol.* **41**, 99–125
552 (2023).
553 64. Sarre, L. A. *et al.* DNA methylation enables recurrent endogenization of giant viruses in an animal relative.
554 *Sci. Adv.* **10**, eado6406 (2024).
555 65. van Oers, M. M. *et al.* Developments in the classification and nomenclature of arthropod-infecting large DNA
556 viruses that contain pif genes. *Arch. Virol.* **168**, 182 (2023).
557 66. Bézier, A. *et al.* Polydnaviruses of Braconid Wasps Derive from an Ancestral Nudivirus. *Science* **323**, 926–
558 930 (2009).
559 67. Delaroque, N., Maier, I., Knippers, R. & Müller, D. G. Persistent virus integration into the genome of its algal
560 host, Ectocarpus siliculosus (Phaeophyceae). *J. Gen. Virol.* **80**, 1367–1370 (1999).
561 68. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage
562 mavirus. *Nature* **540**, 288–291 (2016).
563 69. Archibald, J. *et al.* Persistent mirusvirus infection in the marine protist Aurantiochytrium. Preprint at
564 https://doi.org/10.21203/rs.3.rs-5686297/v1 (2025).
565 70. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems
566 database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
567 71. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome*

Biol. **20**, 238 (2019).

72. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).

73. 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).

74. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).

75. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222 (2014).

76. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).

77. 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 Res.* **29**, 2607–2618 (2001).

78. Klemm, P., Stadler, P. F. & Lechner, M. Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. *Front. Bioinform.* **3**, 1322477 (2023).

79. 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).

80. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

81. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

82. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).

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

84. 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).

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

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

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

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

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

605 ## Supplemental dataset legends

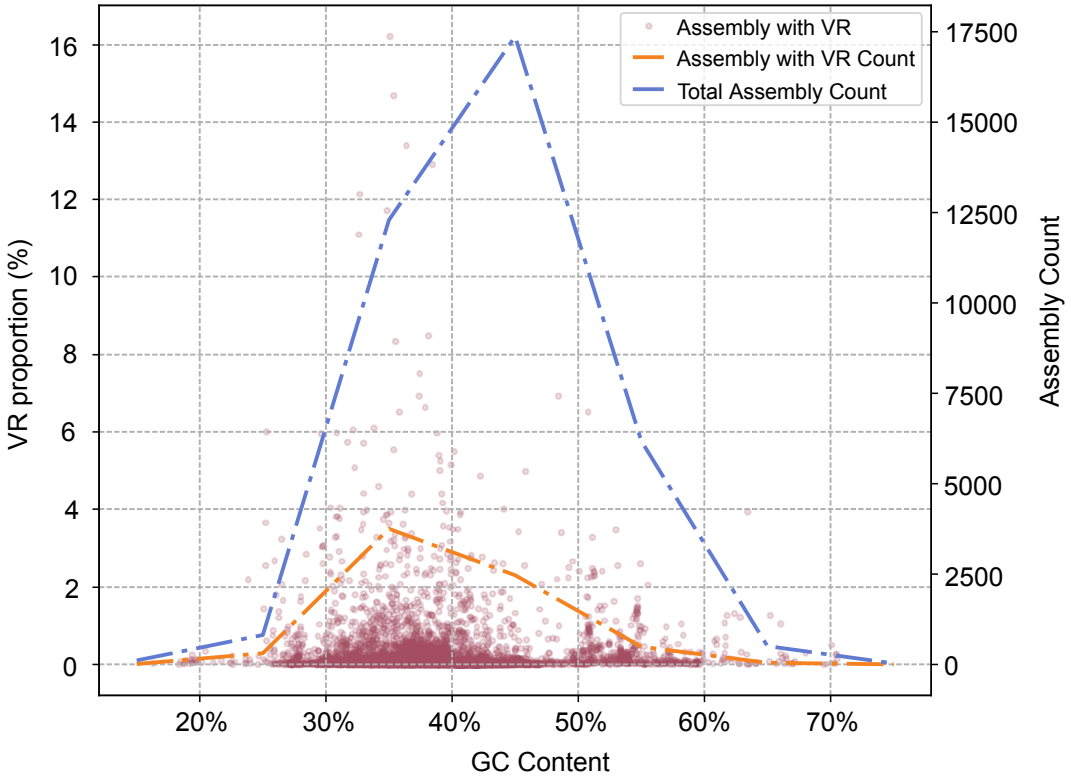
606 **Supplementary Table 1: Eukaryotic assemblies used in this study**

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

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

609

610 ## Supplementary Figures

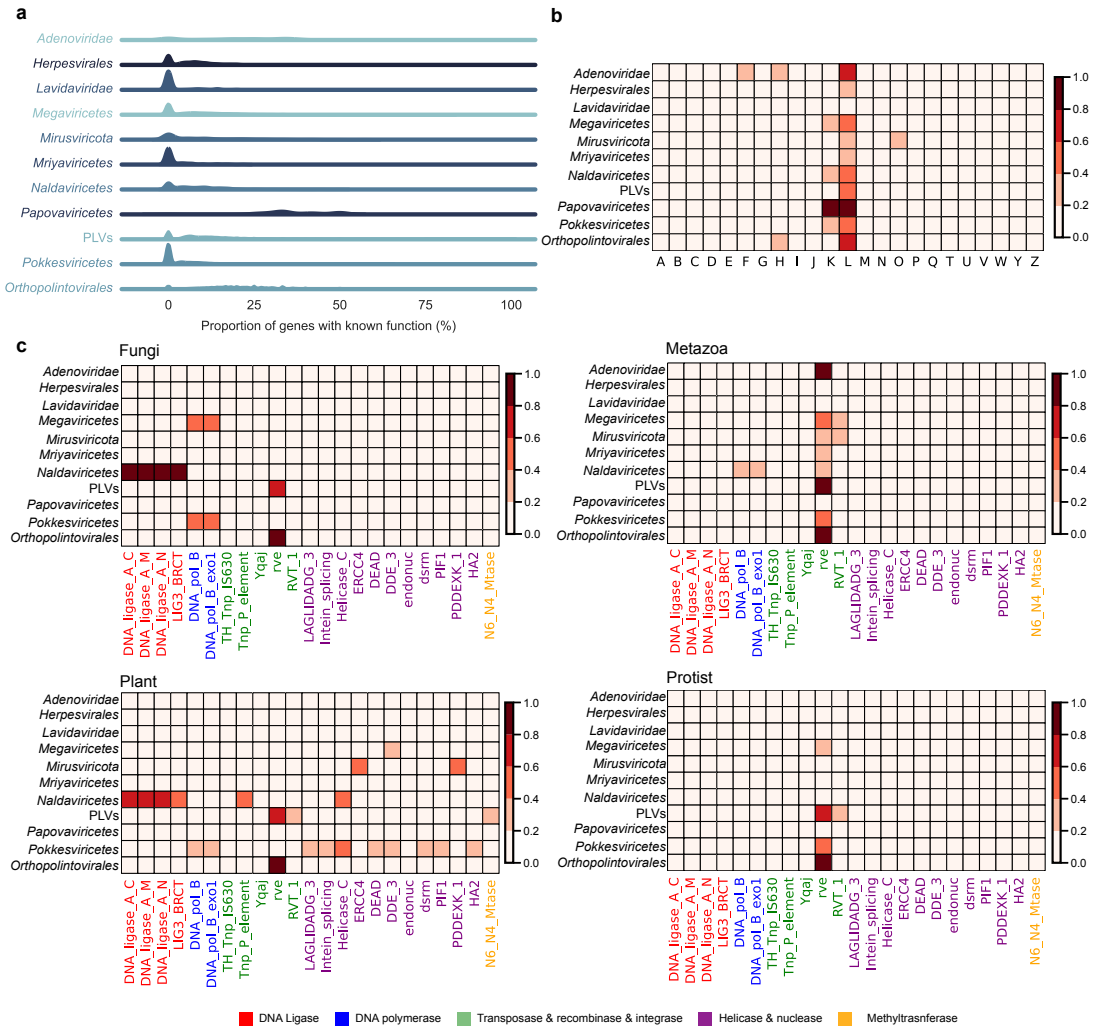


611

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

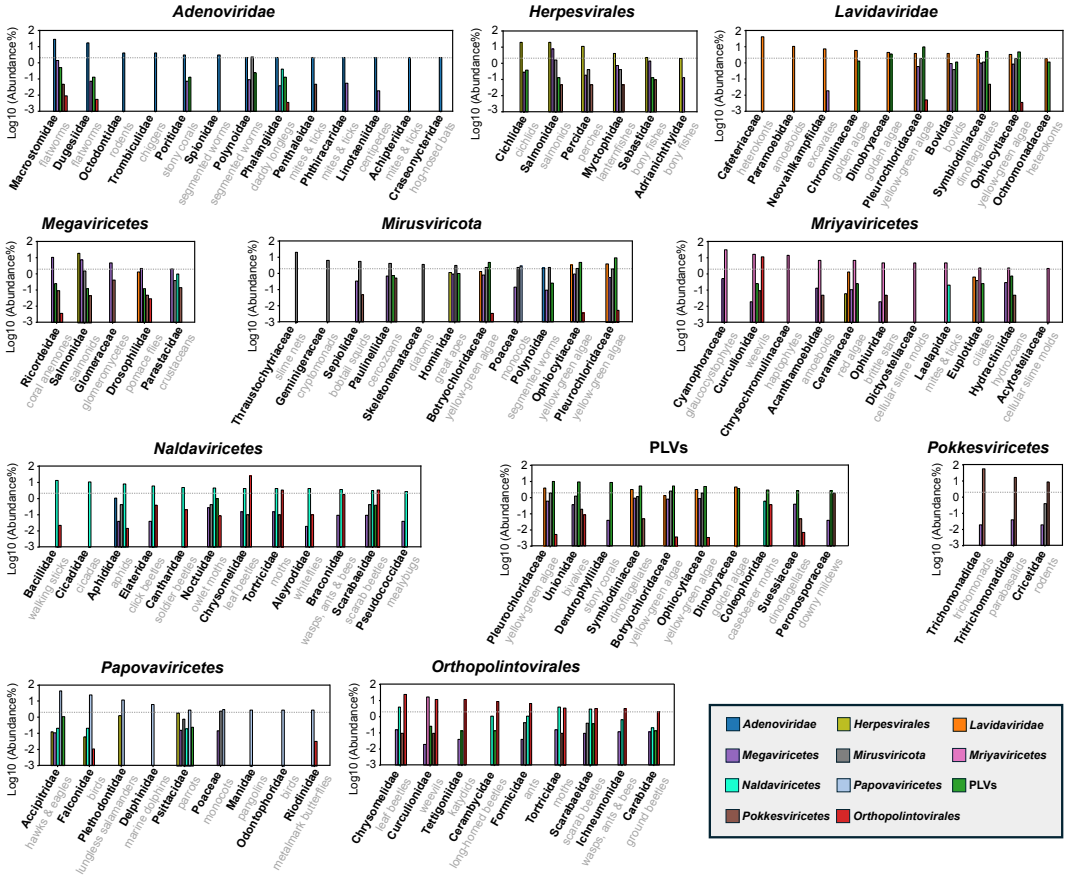
613 Each light red dot represents a single assembly containing VR. The horizontal axis indicates GC content, while the
614 left vertical axis shows the proportion of VRs in each assembly. The blue dotted line represents the total number of
615 assemblies within each GC content bin (ranging from 10% to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

618

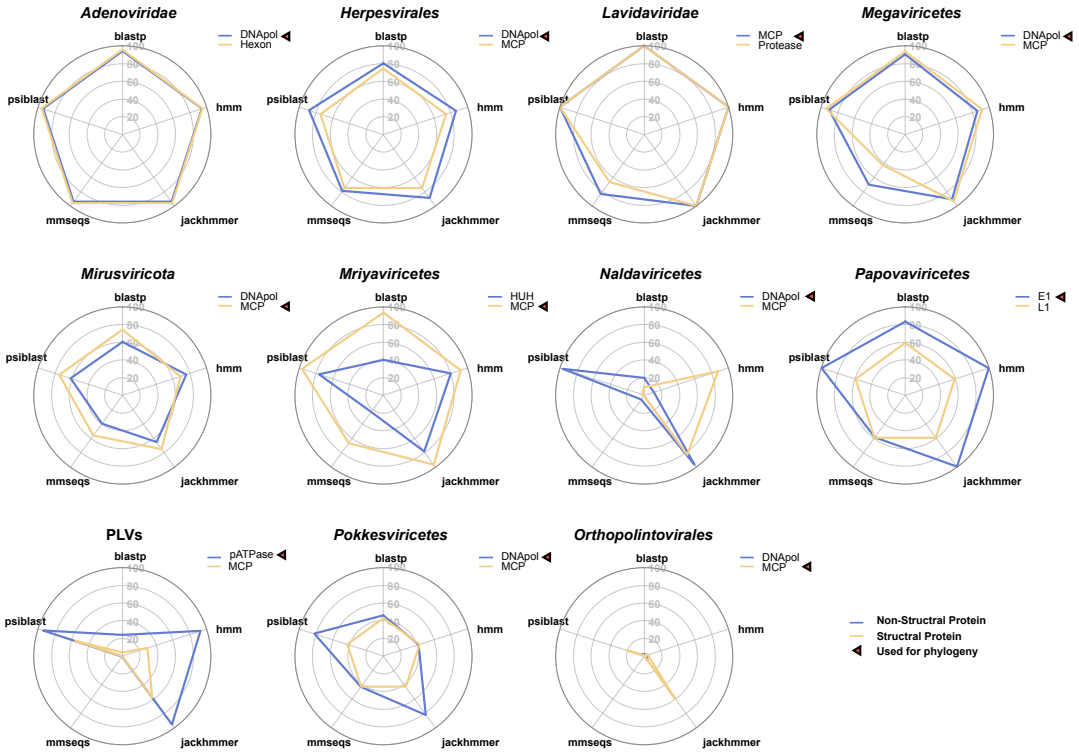


Supplementary Fig. 2 | Functions exhibited by VRs.

(a) Proportion of functional annotated ORFs per viral region. (b) Average proportion of COG categories per VR. The x-axis represents COG categories, the y-axis represents viral taxa and the colour 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 colour scale indicates the average proportion. PFAM colours correspond to functional categories shown at the bottom.

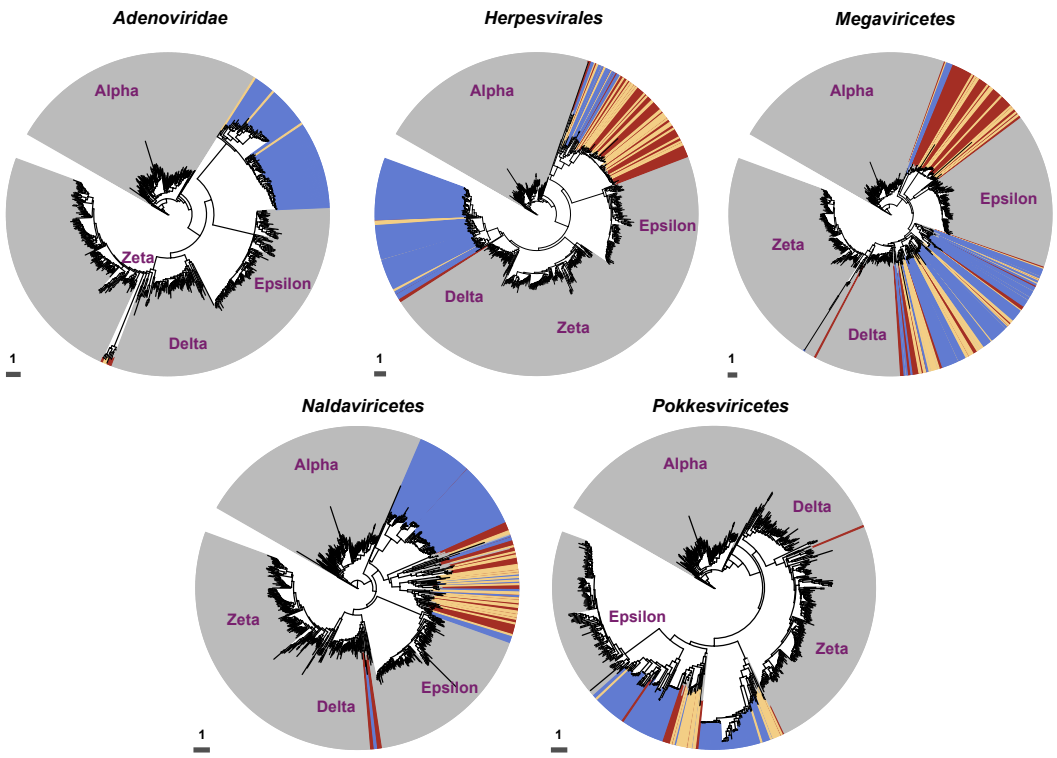


Supplementary Fig. 3 | Eukaryotic families with the highest numbers of HCVRs from each viral taxon. The vertical axis indicates the number of HCVRs in a given family as a percentage of the total HCVRs detected for each viral taxon. Different colours represent distinct viral groups, as indicated in the legend at the bottom right. Each bar corresponds to a eukaryotic family that contributes more than 2% of the total HCVRs for the respective viral taxon, shown in descending order from left to right. Taxonomic names are shown in bold black font, and common names are shown in grey.



Supplementary Fig. 4 | Distribution of marker genes across reference viral genomes.

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



642

643 **Supplementary Fig. 5 | Phylogenetic trees of DNA polymerase including eukaryotic references.**

644 Each tree corresponds to a specific viral taxon, as indicated above. Scale bars are shown in the lower left corners.

645 Grey branches represent VRs not classified as HCVRs. Blue shading indicates reference viral sequences; grey

646 branches represent eukaryotic DNA polymerase references; yellow denotes virus-contig-like VRs; and red represents

647 integration-like VRs. Eukaryotic DNA polymerase clades are labelled in purple text. The best-fit substitution model

648 used for *Adenoviridae*, *Herpesvirales* and *Naldaviricetes* was Q.pfam+I+G4, while Q.pfam+F+I+G4 was applied

649 for *Megaviricetes* and *Pokkesviricetes*.

650