

Genomic catalogue of giant viruses reveals expanded diversity and functional potential

Received: 7 October 2025

Accepted: 23 June 2026

Published online: 13 August 2026



Yumary M. Vasquez^{1,28}✉, Tiago Nardi^{1,2,28}, Gbocho Masato Terasaki^{1,3}, Petra Byl^{1,4}, Tomáš Brůna¹, Juan C. Villada¹, Miguel F. Romero-Gutiérrez^{1,5}, Thomas Mock^{1,6}, Timothy Y. James^{1,7}, GVMAGs data consortium*, Tanja Woyke^{1,8} & Frederik Schulz¹✉

Nucleocytoplasmic large DNA and *Mirusviricota* viruses exhibit taxonomic richness which continually expands due to metagenomic sequencing. Here we curate a database of giant virus metagenome-assembled genomes (GVMAGs V2), comprising 8,508 species-level representatives from 18,727 GVMAGs, a sixfold increase from the previous giant virus phylogenetic frameworks. Phylogenomics revealed 712 previously undescribed genera, 13 previously unknown viral families and an order we propose named *Mycodnavirales*. By accounting for alternative and custom genetic codes, we improved gene calling in over 1,300 GVMAGs, enabling more accurate identification of protein-coding genes. Database mining uncovered putative endogenous viral elements in hosts spanning algae, fungi and parasitic protists, highlighting that giant virus integration is widespread and evolutionarily persistent. Protein-level analysis revealed enriched genes for pollutant degradation in *Algavirales* and widespread biosynthetic gene clusters linked to antimicrobial-like and antibiotic resistance gene-like activity. This public resource will serve as a foundation for expanding giant virus diversity, uncovering virus–host interactions and exploring viral evolution.

Nucleocytoplasmic large DNA viruses (*Nucleocytooviricota*), commonly referred to as giant viruses, represent a diverse group of viruses that challenge traditional virological paradigms due to their large size, genomic complexity and unique biological features^{1–3}. Their particles rival small bacteria in size and their genomes can exceed 2.5 Mb, encoding hundreds to thousands of proteins, some drawn from cellular translation and metabolic pathways^{1,3,4}. Metagenomic surveys have discovered giant viruses across Earth's major biomes, suggesting they have an important yet still poorly understood ecological role³.

Despite these surveys, the taxonomic and functional landscape of *Nucleocytooviricota* remains fragmentary because of sampling biases towards marine and freshwater ecosystems and isolated algae and protists^{3,5–8}. Existing giant virus genome catalogues have laid groundwork for *Nucleocytooviricota* research, yet they encompass only a fraction of the group's diversity and were generated with gene-calling pipelines that assume the standard genetic code. Consequently, some open reading frames remain unannotated and occasional cellular sequences may fail quality filters. Expanding *Nucleocytooviricota* metagenome resources with new sampling ecosystems and improved

¹United States Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ²Department of Biology and Biotechnology, University of Pavia, Pavia, Italy. ³Department of Applied Mathematics, University of California, Merced, Merced, CA, USA. ⁴School of Ocean and Earth Science and Technology, University of Hawai'i Mānoa, Honolulu, HI, USA. ⁵Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City, Mexico. ⁶School of Environmental Sciences, University of East Anglia, Norwich, UK. ⁷Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA. ⁸Life and Environmental Sciences, University of California, Merced, Merced, CA, USA. ²⁸These authors contributed equally: Yumary M. Vasquez, Tiago Nardi. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: ymvasquez@lbl.gov; fschulz@lbl.gov

protein prediction will enable more comprehensive comparative analysis and clarify lineage-specific metabolic innovations, increasing our understanding of *Nucleocytoviricota* diversity and evolution.

Here, we have assembled GVMAGs V2, the largest curated collection of giant virus metagenome-assembled genomes (GVMAGs) so far, a fourfold increase from the original GVMAGs database⁶ and a sixfold increase from the previous 1,382 species-level representatives⁷. GVMAGs contain one or more contigs and reflect giant virus populations at varying levels of completeness. We screened public assemblies in IMG/M and integrated data from additional sources, recovering 18,727 GVMAGs^{7,9–15}. Dereplication at 95% average nucleotide identity (ANI) resulted in 8,508 species-level representatives and resolved 712 previously unrecognised genera. Our phylogenomic analysis supports the addition of 13 previously uncharacterized viral families but expands the existing taxonomic framework, which previously consisted of six orders, 32 families and 344 genera⁷. We also propose a new order (*Mycodnavirales*), basal to *Imitervirales*, and incorporate genomes from the recently described *Duplodnaviria* phylum, *Mirusviricota*¹⁶, into our comparative analyses, extending the evolutionary and functional scope of the dataset.

To enhance genome annotation accuracy, we accounted for alternative genetic codes during gene prediction. This increased complete profile hits and average scores in 1,342 GVMAGs, indicating improvement in annotation quality and recovering >1,300 previously missed genes. We mined PhycoCosm, MycoCosm, Phytozome and IMG/M databases to recover endogenous viral elements in eukaryotic genomes^{12,17–19}. Orthologue clustering of 2.5 million proteins resolved 135,998 gene families. Of the 75,781 orthologues shared by at least three GVMAGs, only 33% had functional assignments. Module-level pathway reconstruction revealed order- and family-specific enrichment for formaldehyde assimilation and aromatic degradation, offering insights into virus-driven metabolic reprogramming. We also identified widespread biosynthetic gene clusters (BGCs) and antibiotic resistance gene-like homologues within GVMAGs, highlighting potential roles in virus–host interactions. To enable exploration of these data, we provide an interactive agentic system in ref. 20. Thus, GVMAGs V2 provides a phylogenetic and functional resource for comparative analyses of giant virus evolution, ecology and metabolic potential.

Results

An expanded giant virus genome database

Expanding the genomic catalogue of giant viruses is essential for improving taxonomic resolution and uncovering previously uncharacterized lineages. We collated 18,727 GVMAGs from sources including IMG/M database, Lake Mendota co-assemblies and the giant virus taxonomy framework, among others^{7,9–15} (Fig. 1a). These GVMAGs span anthropogenic (for example, landfills), freshwater, marine and terrestrial environments (for example, soil) (Supplementary Fig. 1). Following dereplication using ANI, we identified 8,508 non-redundant species-level GVMAGs²¹ (hereafter known as species-level-clusters; Fig. 1b). The species-level clusters indicated here represent a fourfold increase from the original GVMAGs database⁶ and a sixfold increase from the previous 1,382 species-level representative framework⁷. Using a phylogeny based-pairwise distance matrix (PDM) and filtering for a minimum number of giant virus orthologous groups, we generated a final phylogenetic tree containing 2,037 GVMAGs, including five *Proculovirales* genomes and 194 *Egovirales* genomes¹⁶ (Fig. 1c,d).

Using the final tree, we quantified relative evolutionary distance (RED) scores from 0 to 1, where lower values represent deeper evolutionary lineages near the root and higher values indicate narrowly defined clades positioned near the terminal branches. RED scores for *Nucleocytoviricota* classes ranged from 0.18 to 0.2, orders from 0.26 to 0.32, families from 0.35 to 0.75 and genera from 0.76 to 0.98

(Extended Data Fig. 1 and Supplementary Table 1). We identified 13 previously uncharacterized family classifications within *Nucleocytoviricota*, none of which contain single cultivated representatives. *Imitervirales*, the largest *Nucleocytoviricota* order, displayed five previously uncharacterized families, increasing the 11 families identified in 2021⁷ (Fig. 2). Previously uncharacterized families were also identified in *Pimascovirales* ($n = 4$), *Pandoravirales* ($n = 3$) and *Asfuvirales* ($n = 1$). Most newly proposed families contain both clustered and singleton GVMAGs, except IM_24 and MY_01, which contain only singletons. At the genus level, we identified 712 previously uncharacterized genera, including 343 singletons and 369 genera containing 1,092 species-level clusters and 3,767 GVMAGs (taking into account all strain GVMAGs in the species-level clusters).

The final tree revealed a clade of *Imitervirales* GVMAGs that split from the main clade (Fig. 2), echoing previous findings²². This clade comprises two previously identified *Imitervirales* families: IM_18 (hereafter known as MY_03) and IM_02 (hereafter known as MY_02; *Mycodnaviridae*), together with a newly identified family (hereafter known as MY_01). Its members have similar GC percentage and assembly lengths to the larger *Imitervirales* group, but their monophyly outside *Imitervirales* is strongly supported, with a RED score consistent with order rank. The most robustly supported position is basal to *Imitervirales*, with high support values (99/79) at the common node. An alternative topology placing this clade sister to *Algavirales* had lower support and was observed in the GVOG4 and GVOG8 trees (Supplementary Figs. 2 and 3). Single-gene trees also showed variable placement (Supplementary Fig. 4), with Packaging ATPase and DNA polymerase family B positioning these GVMAGs within *Imitervirales*, whereas poxvirus late transcription factor VLT3 supported the sister-to-*Algavirales* topology. DNA topoisomerase II and transcription initiation factor IIB gene trees showed clustering with other *Imitervirales* GVMAGs as a sister clade to *Algavirales*, whereas DEAD/SNF2-like helicase grouped them with *Asfuvirales*. We designate this divergent clade as a new taxonomic order, tentatively named *Mycodnavirales*, derived from *Mycodnaviridae*, a proposed family associated with early-diverging fungi^{22,23}. To note, major capsid proteins of *Mycodnavirales* are significantly shorter than those of *Imitervirales* (median 400 versus 506 amino acids; Welch's t -test, $P = 7.0 \times 10^{-16}$), with families MY_02 and MY_03 having median lengths of 349 and 400 amino acids, respectively.

In our expanded dataset, we observed the division of known families into multiple non-monophyletic clades, denoted with the clade name and a new number identifier. For example, the *Imitervirales* family IM_09 separated into IM_09, IM_19 and IM_20, each comprising a mix of IMG/M GVMAGs and genomes from other sources. A similar pattern is observed in the *Asfuvirales* family AF_01, which separated into AF_01, AF_04, AF_05 and AF_06. Each newly defined family (except AF_05) contains GVMAGs from multiple independent sources, supporting their taxonomic validity. We also observed former singleton genera forming well-supported families through the addition of related GVMAGs, including four families that emerged from individual singletons (AF_03, MY_01, IM_21 and IM_24) and four that formed from pairs of singletons (PM_14, PV_07, PV_08 and IM_23). This transformation from singleton genera to robust families demonstrates how additional genomic sampling can reveal previously hidden viral diversity and establish more accurate phylogenetic relationships.

Expansion of the viral phylum *Mirusviricota*

Mirusviricota is a recently characterized phylum of giant viruses linked to herpesviruses¹⁶ and potentially associated with diverse eukaryotes, particularly protists and algae^{24,25}. We recovered 611 *Mirusviricota* metagenome-assembled genomes (MAGs) from aquatic systems (that is, marine, freshwater, non-marine saline/alkaline, deep subsurface and thermal springs), terrestrial systems (that is, soils and rock endoliths), plant-associated niches (that is, roots, peat moss and phyllosphere)

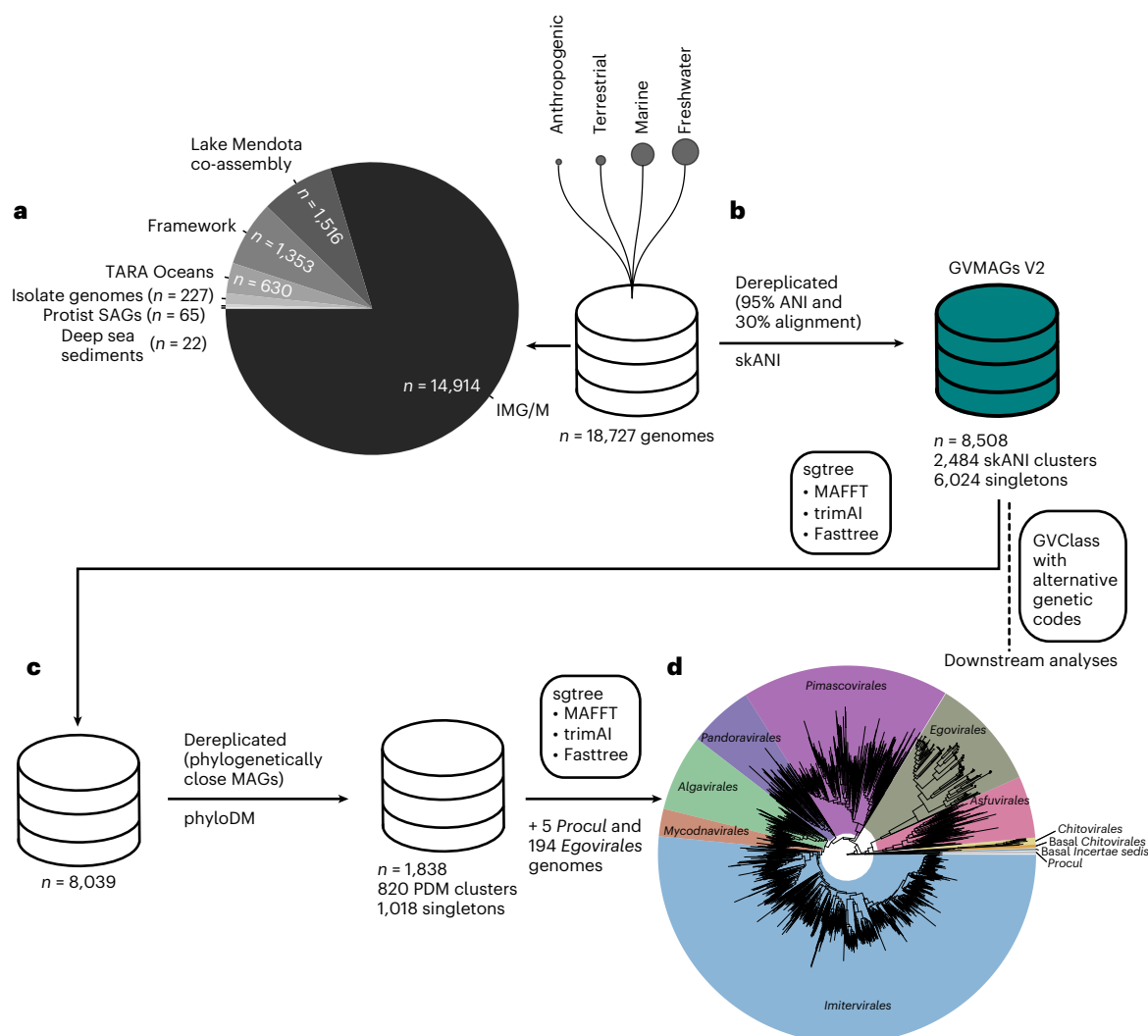


Fig. 1 | Workflow for genomic dereplication and phylogenetic tree construction.

a, A set of 18,727 genomes was compiled from public databases, with the majority ($n = 14,914$) from IMG/M. **b**, An initial dereplication using skANI (95% ANI, 30% alignment) produced the GVMAGs V2 database, containing 8,508 representative genomes (2,484 clusters, 6,024 singletons). **c**, After an intermediate filtering step (resulting in $n = 8,039$ genomes), a secondary

dereplication of phylogenetically close MAGs using a phylogenetic distance matrix (phyloDM) refined the set to 1,838 genomes (820 clusters, 1,018 singletons). **d**, Finally, 199 genomes (5 *Proculovirales* and 194 *Egovirales*) were added, and a phylogenetic tree was constructed from the final set of 2,037 genomes.

and engineered environments (that is, anaerobic bioreactors, soil microcosms and plant-growth chambers). We combined these with 88 *Mirusviricota* MAGs from TARA Oceans¹⁶.

Dereplication resulted in 111 species-level-clusters and 298 singletons. Of these, only 145 clusters and singletons had enough markers for phylogenetic clustering, resulting in 55 clusters and singletons, including two composed of IMG/M GVMAGs. Most clusters ($n = 49$) are positioned with *Pandoravirales*, consistent with previous reconstructions²⁶ (Extended Data Fig. 2). Using the same RED-based approach as for *Nucleocytoviricota*, we established a *Mirusviricota* framework comprising two orders (Fig. 2, m1 and m2; Supplementary Tables 2 and 3). The m1 order, consisting of five GVMAGs in one family with independent genera, appears basal to all *Megaviricetes*, although its position outside the main *Mirusviricota* group remains uncertain. The m2 order contains three families subdivided into 13 genera (each containing multiple GVMAGs). One m2 genus occupies a position basal to 47 other *Mirusviricota* sequences and a *Pandoravirales* genome ('sylv'), a placement that may reflect long-branch attraction. Our RED score analysis supports proposing *Mirusviricota* as its own phylum-level lineage.

Genetic code diversity improves annotation in giant viruses

Alternative genetic codes, including stop codon reassignments, can result in incomplete annotation of protein-coding genes if unaccounted for, yet their distribution and impact in giant viruses remain largely unexplored²⁷. Our analysis of 8,508 dereplicated GVMAGs showed improved protein-coding gene prediction in *Nucleocytoviricota* when applying alternative genetic codes (Fig. 3). We compared giant virus-specific gene-calling optimization with GVCClass against Prodigal using the meta mode by evaluating changes in the number of complete best hits and the average best-hit score^{28,29} (Fig. 3a). Of the 8,508 GVMAGs analysed, 6,448 (~76%) fall on one of the axes (that is, the two programs perform identically for at least one of the two metrics). The remaining GVMAGs show a directional shift with 1,342 GVMAGs (16% of the total) in quadrant 1, where GVCClass recovers more complete profile hits and higher average scores. Another 716 GVMAGs (8%) plot in quadrant 4, gaining additional hits at the expense of a decrease in score.

To determine which genetic codes provided the greatest prediction accuracy, we compared multiple translation tables (Fig. 3b and Extended Data Fig. 3). The standard genetic code (code 1) and

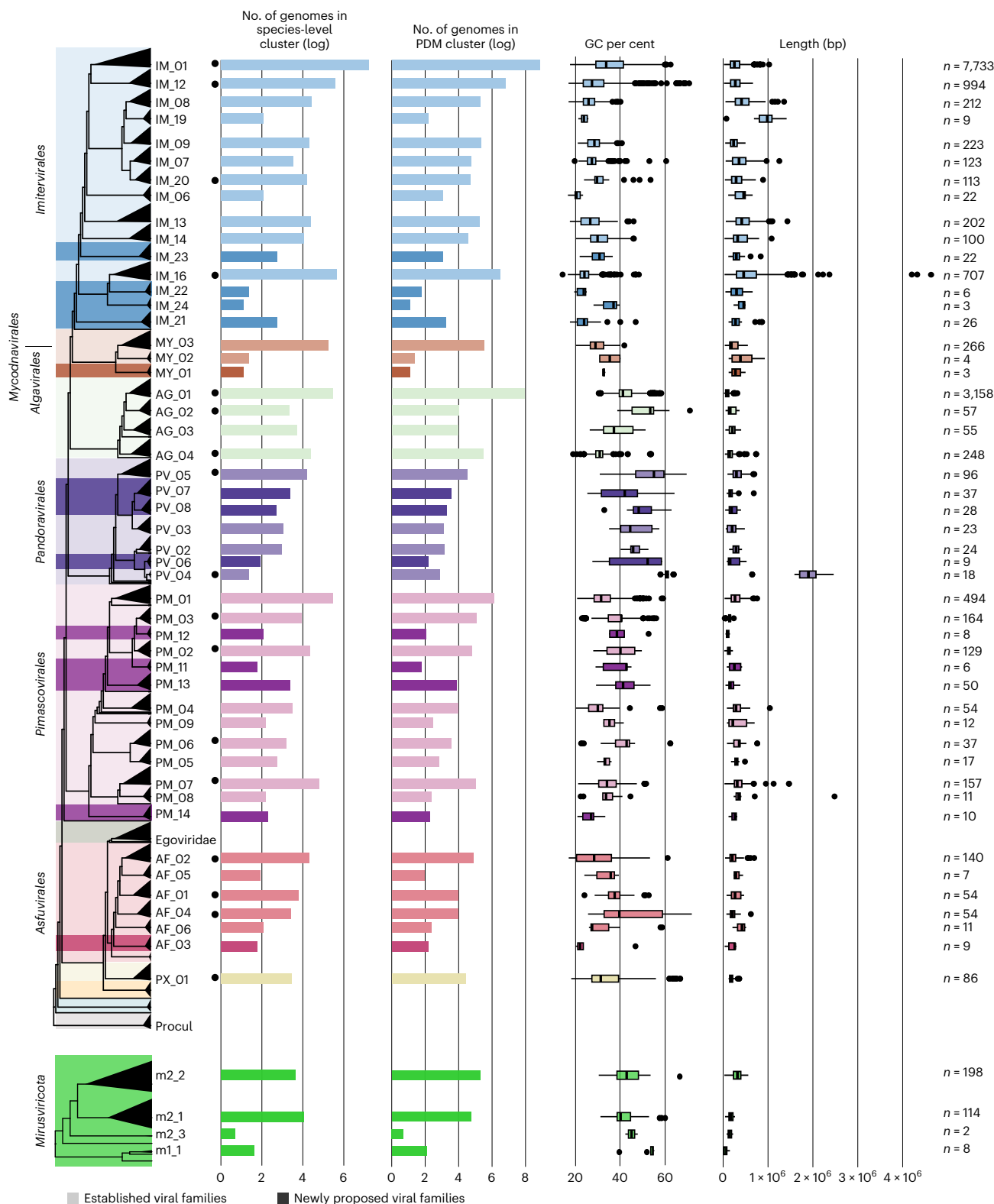


Fig. 2 | Phylogenetic and genomic features of *Nucleocytoviricota* and *Mirusviricota* families. Left: the phylogenetic tree displays viral families organized by order (indicated by coloured blocks) within the *Nucleocytoviricota* phylum, alongside families from the *Mirusviricota* phylum. Newly proposed families are indicated by dark coloured shading, distinguishing them from established families in light coloured shading. A solid black circle marks any family containing at least one cultured isolate. To the right of the tree, four data panels visualize key metrics for each family: (1) the natural log-transformed number of genomes in species-level clusters, (2) the natural log-transformed number of genomes in phylogeny based-pairwise distance matrix (PDM) clusters, (3) the distribution of genomic GC content

(%) and (4) the distribution of genome assembly length (bp). For box plots, the centre line represents the median, box limits indicate the 25th and 75th percentiles, whiskers extend to the most extreme values within 1.5× the interquartile range and the points beyond the whiskers represent outliers. No inferential statistical tests were performed for these panels. For each family, *n* represents the number of genomes assigned to that family and used to calculate the boxplot. The figure also highlights the refinement of polyphyletic families based on RED score analysis; for instance, the original *Imitervirales* family IM_09 is resolved into three distinct families (IM_09, IM_19 and IM_20) and the *Asfuvirales* family AF_01 is split into four (AF_01, AF_04, AF_05 and AF_06).

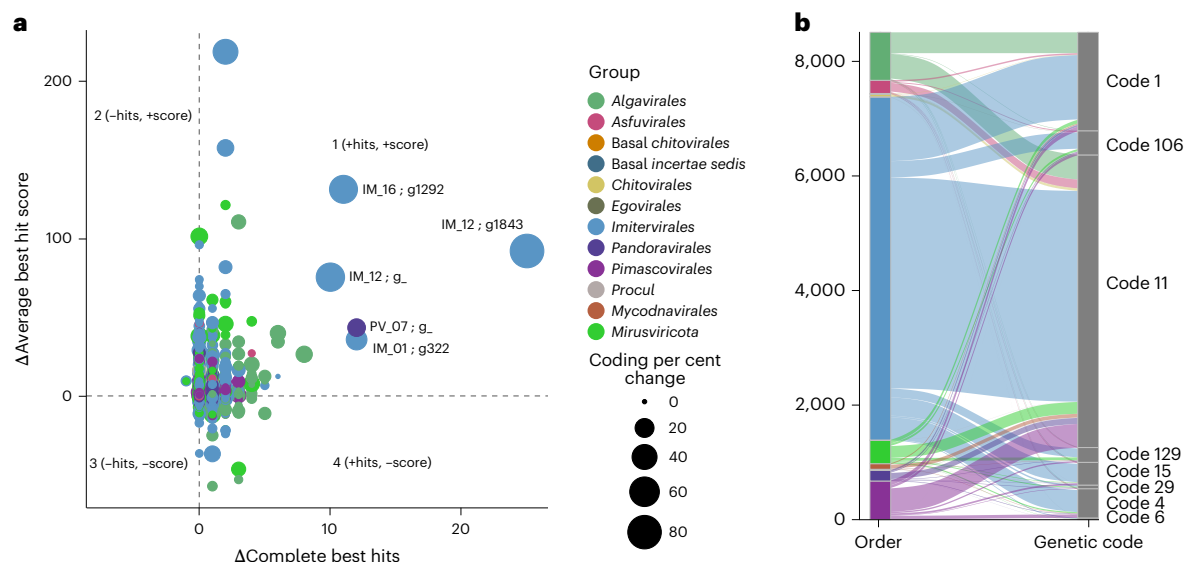


Fig. 3 | Improved gene prediction with GVClass and the distribution of alternative genetic codes. a, A scatter plot comparing the performance of a giant virus-specific annotation tool (GVClass) against a standard gene predictor (Prodigal-meta). Each point represents a single genome, coloured by taxonomic group. The x-axis shows the change (Δ) in the number of complete best hits, and the y-axis shows the change (Δ) in the average best-hit score. Points in quadrant 1 represent genomes where GVClass improved both metrics, indicating more effective protein identification. The size of each point corresponds to the percent

change in the total number of protein-coding genes. Specific genomes with substantial improvements are labelled with their family and genus taxonomy. **b**, An alluvial plot illustrating the distribution of predicted genetic codes across viral taxonomic groups. The left column represents the viral lineage, and the right column represents the genetic codes. The connecting flows reveal strong lineage-specific patterns, such as the predominant use of code 11 within the *Imitervirales* and the diversity of codes used by *Mirusviricota*.

the bacterial, archaeal and plant plastid code (code 11) were the best optimized overall codes. Genetic codes with a single stop codon reassignment were favoured more often than those with multiple reassignments. For example, code 106, reassigning TAA to glutamine, and code 129, reassigning TAG to tyrosine, were more optimal than their counterparts, code 6 and code 29.

Across environments, code 11 and code 1 were the best translation codes, with one exception: anthropogenic environments, including effluent, landfills and activated sludge, showed greater improvement with code 106 and code 4 than with code 1, although code 11 remained dominant (Supplementary Fig. 5). Terrestrial environments, such as grass, peat moss, roots and soil, also exhibited a higher number of GVMAGs using codes 129, 106 and 4, although these remained less common than codes 1 and 11 (Supplementary Fig. 6).

Currently, all confirmed giant virus isolates have been isolated from hosts using the standard genetic code 1 (for example, amoebae, algae and other protists). One *Imitervirales* GVMAG from family IM_12 showed an increase in coding density from 13% to 93% and recovery of additional functional genes, with complete best hits increasing from 2 to 27, when code 6 was used. Code 6 translates both TAA and TAG to glutamine rather than stop and is found in some ciliates, algae and diplomonad protists^{30–32}. The only known association of a giant virus with a protist that uses code 6 involves the ciliate *Paramecium bursaria*, which harbours symbiotic *Chlorella* spp. infected by giant viruses³³, but no giant viruses have been reported to directly infect the nucleus or cytoplasm of *P. bursaria* or use its translation machinery.

GEVs are widespread across datasets

Using updated hidden Markov models (HMMs) derived from the GVMAGs, here, we searched for giant endogenous viral elements (GEVs) in publicly available eukaryote genomes. We identified 142, 550, 76 and 40 eukaryotic genomes with putative GEVs from IMG/M, MycoCosm, PhycoCosm and Phytozome, respectively, as identified by both ViralRecall and GVClass^{12,17–19,29,34} (Supplementary Table 7). After phylogenetic filtering, we recovered 22 genomes with high-confidence GEVs (Supplementary Tables 4 and 8).

Our analysis uncovered high-confidence GEVs in lineages previously known to contain giant virus integrations and in lineages not previously considered hosts. We detected GEVs in well-studied groups such as Symbiodiniaceae dinoflagellates, Phaeophyceae brown algae (Stramenopiles), green algae (Chlorophyta) and fungi (Chytridiomycota)^{5,22,35–37} (Extended Data Fig. 4). High-confidence GEVs were also found in putative hosts, including Glaucophyta and possibly vascular plants, such as *Arabidopsis thaliana* and *Sinapis alba*, although these two organisms lacked the giant virus major capsid protein. Marker searches using only HMMs identified GEVs in parasitic eukaryotes, including cordyceps fungi and the human pathogen *Trichomonas*. Endogenization of smaller, Maverick/Polinton-like viruses, which can parasitize giant viruses, is widespread in *Trichomonas*^{38,39}.

Shared genes and enriched functions across giant virus orders

The gene repertoires of giant viruses encompass conserved functions and diverse auxiliary genes, yet most coding potential remains uncharacterized⁴⁰. Comparative genomic analysis of 2,504,503 genes from dereplicated GVMAGs identified 135,998 orthologues, spanning 1,693,149 genes. Only 75,781 orthologues were shared by at least three GVMAGs (1,571,367 genes; Extended Data Fig. 5). Of these, 10,132 (14%) contained paralogs (defined as multiple genes from a single genome in an orthogroup), and 24,783 (33%) had functional annotations. Known functions include diverse biological processes, including protein synthesis and modification, and DNA damage response and repair.

To understand the distribution and connectivity of orthologues across *Nucleocytoviricota* orders, we conducted orthogroup and network analysis. *Imitervirales* exhibited the highest number of order-specific orthologues ($n = 47,541$, 63% of total orthologues, 42% of total genes in orthogroups, 34% of all genes in *Imitervirales*), as well as the highest median connectivity ratio (0.91) (Fig. 4a, Extended Data Fig. 6, Supplementary Table 5 and Supplementary Figs. 7–9). Only 68 orthogroups were shared across the five major orders (*Imitervirales*, *Pimascovirales*, *Algavirales*, *Pandoravirales* and *Asfuvirales*).

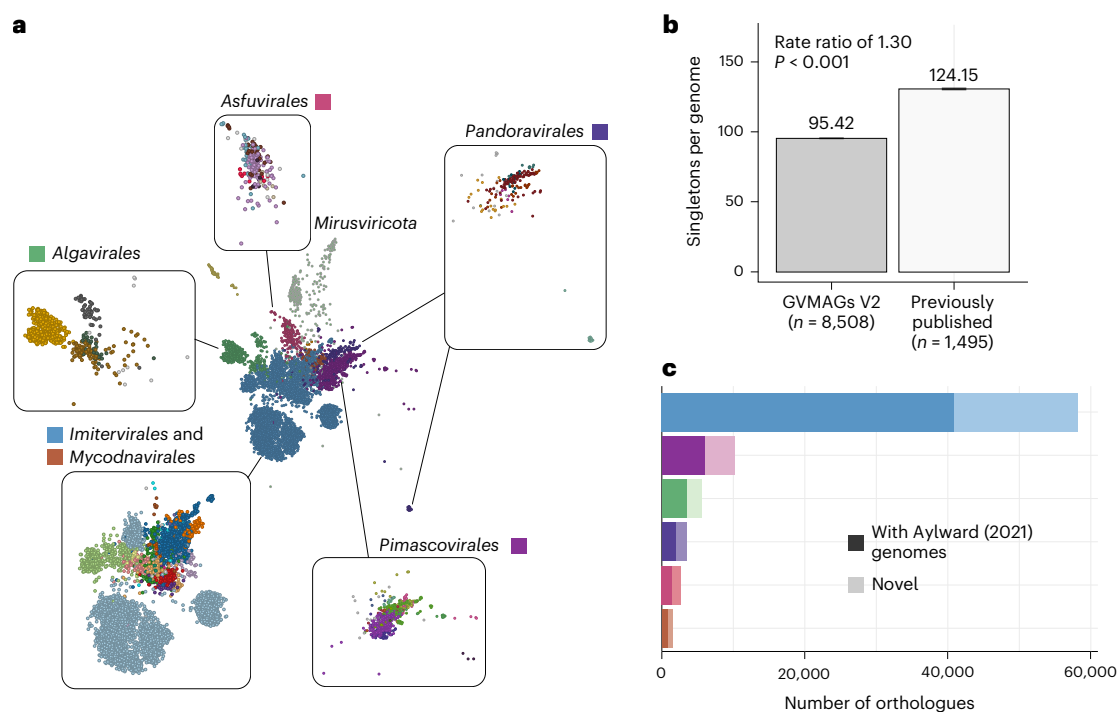


Fig. 4 | Gene-sharing networks and analysis of orthologous gene families across *Nucleocytoviricota* and *Mirusviricota*. **a**, A gene-sharing network where nodes represent individual genomes, clustered based on shared orthologous gene families. The main view shows genomes coloured by their respective viral order. The insets provide higher-resolution views of select orders, with genomes coloured by family to illustrate finer-scale relationships. For visual clarity, the edges representing shared orthologues have been omitted from this visualization (the full network is available in Extended Data Fig. 6). **b**, A bar chart comparing the average number of singleton genes (genes without a known orthologue) per MAG between the GVMAGs V2 dataset ($n = 8,508$ MAGs)

and a smaller dataset of previously published giant virus genomes ($n = 1,495$ MAGs). The previously published genomes exhibit a significantly higher rate of singletons (rate ratio of 1.30; exact $P = 1.09 \times 10^{-323}$). Statistical significance was assessed using a two-sided exact Poisson rate test comparing singleton frequencies. **c**, A stacked bar chart quantifying the number of orthologous gene families (orthologues) identified per viral order. The bars are segmented to show the contribution of this study: light portions indicate previously uncharacterized orthologues containing no members from the Aylward et al.⁷ reference dataset, while dark portions represent orthologues that were also present in that framework.

Imitervirales and the newly proposed *Mycodnavirales* shared 636 orthologues. *Mycodnavirales* possessed 433 order-specific orthologues (with an additional 9,364 singletons) (Extended Data Fig. 5). Enrichment analyses indicated that *Mycodnavirales* is enriched for Clusters of Orthologous Groups (COG) categories involved in cytoskeleton (Z), signal transduction mechanisms (T), inorganic ion transport and metabolism (P) and intracellular trafficking (U) (Supplementary Fig. 10). The newly proposed *Mycodnavirales* order is predicted to infect fungi, as evidenced by endogenous viral elements within fungal genomes and fungal homologues in giant metagenome-assembled viruses²². In our dataset, KEGG Orthology (KOs) groups unique to *Mycodnavirales* support a fungal host, including homologues annotated as pre-rRNA-processing protein Srd1p (K14858) and osmolarity two-component system, phosphorelay intermediate protein Ypd1p (K11232), genes involved in fungal virulence and stress response, respectively^{41,42} (Supplementary Fig. 11). We also found KO annotations for proteins involved in signal transduction, including mitogen-activated protein kinase-like protein (K04420, K06683) and histidine kinase-like protein (K07646, K07679), suggesting possible interference with host cell signalling pathways^{36,43,44}.

A total of 811,354 giant virus genes were identified as singletons, representing 32.4% of all genes analysed (Supplementary Figs. 12 and 13). To evaluate the impact of genome sampling on orthogroup formation, we compared singleton rates before and after adding 7,013 GVMAGs from IMG/M to 1,495 previously published genomes. This addition significantly reduced the singleton rate from 124.15 to 95.4 singletons per genome (rate ratio of 1.30, $P < 2.2 \times 10^{-16}$). The expanded dataset also led to a significant increase in identified orthologues, with 32,551 orthologues not reported in previous studies⁷ (Fig. 4c).

Having established patterns of gene sharing and taxonomic connectivity, we investigated differences in functional capacity. We identified 220 modules enriched across distinct families (Fig. 5a and Extended Data Fig. 7). For example, *Algavirales* showed enrichment for genes involved in the formaldehyde assimilation ribulose monophosphate pathway (M00345), which includes a sugar manipulation enzyme previously identified in prasinoviruses⁴⁵. By contrast, *Imitervirales* GVMAGs harbour genes for a different sugar manipulation process, the formaldehyde assimilation xylulose monophosphate pathway (M00344), as shown in⁴⁶. *Algavirales* GVMAGs from marine and freshwater environments, including shale gas reservoirs, estuarine waters, lakes, rivers, ponds and salt marshes, were also significantly enriched in three modules for aromatic degradation (M00545, M00551 and M00569).

Nucleocytoviricota and *Mirusviricota* as reservoirs for biotechnological potential

Antibiotic resistance genes (ARGs) and antimicrobial BGCs are important for biotechnological exploration. Recent findings indicate that giant viruses can act as reservoirs for 12 types of ARGs⁴⁷. However, because there is no evidence that they confer resistance in bacteria or relate to antibiotic resistance, this classification is debated^{48,49}. We therefore refer to them as viral homologues of drug-targeted enzymes and interpret their functions in the context of virus–host interactions. Our expanded dataset identified dihydrofolate reductase-like homologues associated with trimethoprim-resistance annotations (putative *dfr/dfrA*-like homologues; $n = 1,604$) as the most abundant ARG-like homologues, alongside glycopeptide resistance-associated *van*-like homologues ($n = 140$) and mupirocin resistance-associated homologues ($n = 114$) (Fig. 5b). We also found five *msbA*-like homologues

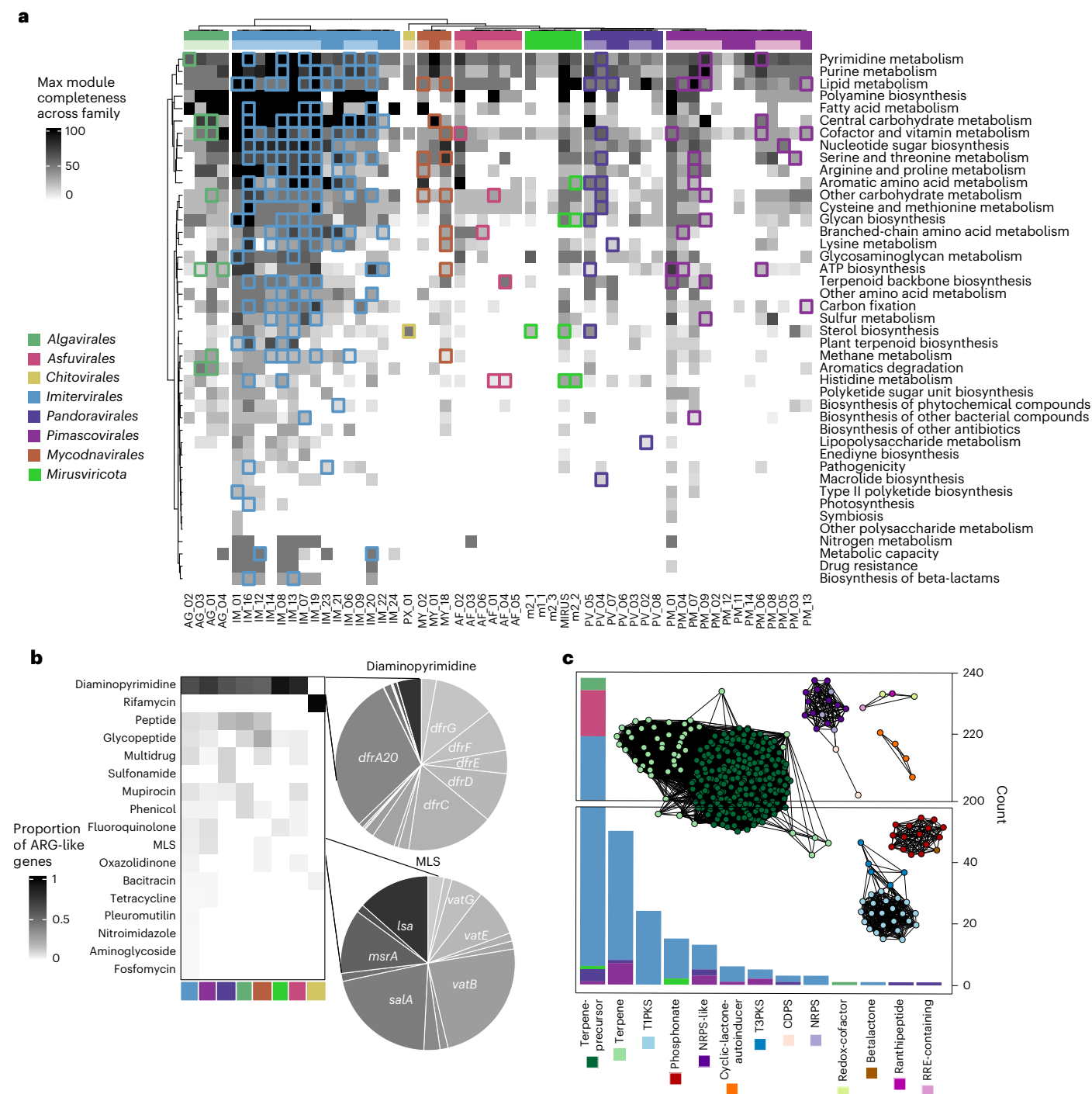


Fig. 5 | Functional profiling of metabolic capabilities, antibiotic resistance-like genes and biosynthetic-like gene clusters across *Nucleocytoviricota* and *Mirusviricota*. a, A heat map illustrating the maximum completeness of metabolic modules across viral families. Columns represent viral families (clustered within their respective orders), and rows represent metabolic pathways. The intensity of each cell corresponds to the maximum module completeness score observed within a family, from 0 (white) to 100 (black). The top colour bar indicates the viral order for each family, with darker shades denoting newly proposed families. Tiles with coloured borders indicate gene categories that contain modules enriched within a family. **b**, Analysis of antibiotic

resistance-like gene homologues (ARG-like). Left: the heat map shows the proportion of genomes within each viral order (bottom colour key) that encode genes for various ARG categories. The pie charts on the right provide a detailed breakdown of the specific genes found within two selected categories: diaminopyrimidine and macrolide–lincosamide–streptogramin (MLS). **c**, A quantification and network analysis of BGCs. The bar chart shows the total count of BGCs detected for each functional category. Above it, a similarity network visualizes the relationships between these BGCs. In the network, each node is a BGC, and they are coloured according to the same functional categories shown in the bar chart, revealing distinct clusters of related biosynthetic pathways.

annotated in the nitroimidazole resistance category in GVMAGs from freshwater lakes and estuaries, absent from previous surveys⁴⁷. Nitroimidazole resistance-like genes have been previously reported in giant viruses^{30,51}.

To further explore functional potential, we surveyed BGCs across our curated database, identifying BGC-like homologues in 326 GVMAGs within *Nucleocytoviricota* and *Mirusviricota* (Fig. 5c and Supplementary Table 6). Most predicted clusters were associated

with terpene-precursors ($n = 233$) and terpene ($n = 49$), compounds critical for membrane modification, signalling and defence⁵². Although terpene biosynthesis has only been examined in four viral genomes so far^{53,54}, our data show that putative terpene biosynthetic genes span all major *Nucleocytoviricota* orders and *Mirusviricota*. Terpenoid-related BGC homologues were found primarily in aquatic environments including thermal springs, deep subsurface habitats, and both freshwater and marine settings (Supplementary Fig. 14).

We also uncovered BGCs related to polyketides (TIPKS and T3PKS) and peptide-derived compounds (non-ribosomal peptide synthetases (NRPS)-like), commonly known for antimicrobial properties, primarily in *Imitervirales*, *Pandoravirales* and *Pimascovirales*. These compounds were found in freshwater, marine and mixed feedstock samples (Supplementary Fig. 14). Order-specific BGC-like homologues include a redox-cofactor cluster unique to *Algavirales* and RiPP recognition element clusters in *Pandoravirales*. Phosphonate gene clusters were identified in *Imitervirales* and *Mirusviricota*, along with quorum-sensing-like cyclic-autoinducer clusters in both *Imitervirales* and *Pimascovirales*.

Discussion

Our study expands the known diversity and functional potential of *Nucleocytoviricota* and *Mirusviricota* by combining large-scale genome recovery with phylogenomic classification, genetic code-aware annotation, endogenous viral element detection and comparative functional analyses. Building upon previous classifications⁷, we identified one previously uncharacterized order, 13 previously uncharacterized families and 712 previously undescribed genera, including singleton genomes that probably represent undescribed lineages. These findings show that giant virus diversity remains substantially underestimated and that genome-resolved metagenomics continues to reshape our understanding of giant virus evolution, ecology and metabolic potential.

The transition of several singleton genera into well-supported families demonstrates how additional genomes can reveal hidden viral diversity and establish more accurate phylogenetic relationships. Conversely, the division of known families into multiple non-monophyletic clades suggests that some current classifications remain sensitive to sampling depth. This pattern is exemplified by the proposed order *Mycodnavirales*. Although these genomes were previously associated with *Imitervirales*, their monophyly outside the main *Imitervirales* clade, order-level RED and distinct major capsid protein lengths support their recognition as a separate lineage. *Mycodnavirales*-specific genes, including homologues to fungal virulence and stress-response genes, further suggest a possible fungal host association and provide targets for future experimental work. Our analyses also expand *Mirusviricota* diversity and support its recognition as a phylum-level lineage, although the uncertain position of the m1 order highlights the difficulty of reconstructing deep viral evolutionary relationships and may reflect genuine evolutionary separation, independent acquisition of *Nucleocytoviricota* genes or limited phylogenetic resolution.

Expanded genome sampling also changes how we interpret giant virus gene content. Although giant viruses are known for large genomes and extensive gene repertoires, most genes in our dataset remain uncharacterized, reflecting both the novelty of giant virus biology and the limitations of current annotation databases. The unexpectedly low frequency of orthogroups with paralogs warrants further investigation, especially given that giant virus genomes often contain many gene duplications⁵⁵. Similarly, the high connectivity observed in *Imitervirales* may reflect the large number of *Imitervirales* GVMAGs and specific network edges, or it may indicate gene content shaped by consistent host associations and environmental pressures⁵⁶. The reduction in singleton rate after adding 7,013 IMG/M GVMAGs further shows that many apparent orphan genes may result from incomplete sampling. Thus, many genes currently interpreted as taxon-specific may have homologues in yet-to-be-sequenced viruses, emphasizing the need for continued genome recovery from diverse environments.

GVMAGs V2 also provides clues about host associations and the evolutionary constraints shaping genetic code variation. Although most GVMAGs were best annotated with the standard genetic code, several improved when alternative genetic codes were considered. These cases, which were checked for contamination, raise the possibility of previously unrecognized host–virus associations and suggest that giant virus–host coevolution may be broader than currently appreciated. Genetic codes involving single stop-codon reassignments were favoured more often than those involving multiple reassignments. According to the ‘ambush hypothesis’, multiple stop codons may reduce the cost of frameshift mutations by increasing the likelihood of premature translation termination⁵⁷. However, reassigning multiple stop codons may impose greater translational costs in giant viruses, whereas single reassignments may minimize translational disruption or facilitate adaptation to hosts with nonstandard termination biology.

Host–virus genetic code alignment is thought to facilitate efficient translation of viral proteins during infection^{58,59}, although codon usage matching alone may not reliably predict host range, as shown in some amoeba-associated giant viruses⁶⁰. Differences between viral and host genetic codes may also be influenced by immune evasion, replication site or requirements for efficient transcription and export⁶¹. The IM_12 genome that improved under code 6 is a plausible candidate for studying giant viruses that may infect hosts using genetic code 6 or use alternative genetic codes during infection. In parallel, the detection of GEVEs in algae, parasitic fungi and the human pathogen *Trichomonas* expands the range of possible giant virus associations and suggests that endogenization of giant virus DNA may be a recurring feature of eukaryotic genome evolution. These findings require further validation, particularly for candidates lacking clear major capsid protein evidence.

The functional analyses extend this picture from diversity and host association to potential ecological roles. Enrichment of metabolic modules across specific orders indicates that giant viruses may use lineage-specific strategies for host interaction and metabolic reprogramming. Aromatic degradation genes in *Algavirales* may enhance replication and survival in environments where microalgal hosts thrive; these hosts can also contribute to bioremediation through pollutant uptake and detoxification^{11,62}. The meta-cleavage modules enriched in *Algavirales* support degradation of natural and industrial pollutants in other microbes and could potentially be relevant to bioremediation⁶³. Likewise, antibiotic resistance gene-like homologues and BGCs-like regions highlight the functional complexity of giant virus genomes. Because there is no evidence that these viral homologues confer antibiotic resistance in bacteria, they should be interpreted cautiously as viral homologues of drug-targeted enzymes rather than confirmed resistance determinants. Their functions may instead relate to virus–host interactions, cellular metabolism or competition within host-associated microbial communities. Terpene-, polyketide- and peptide-related BGCs-like homologues further suggest that some giant viruses may affect microbial community dynamics or leverage these genes during virus–host interactions.

Several limitations should be considered. Incomplete MAGs may bias our view towards readily assembled viral genomes and obscure the full spectrum of diversity. Uneven environmental sampling leads to overrepresentation of certain ecosystems and host types. Limited host information constrains ecological interpretation and understanding of genetic code variation. Functional predictions also remain dependent on databases that are incomplete for giant viruses. Nevertheless, our large-scale genome mining uncovered increased giant virus diversity and functional potential relevant to biotechnology.

Together, these findings establish GVMAGs V2 as an expanded foundation for future research on giant virus evolution, diversity, host interactions and metabolic potential. By combining large-scale genome recovery, phylogenomic classification, genetic code-aware annotation, endogenous viral element detection and functional pathway

analysis, this resource enables comparative analyses and supports efforts to explore the ecological and biotechnological potential of giant virus-encoded genes and pathways.

Methods

Dereplication and phylogenetic inference of GVMAGs

A total of 18,727 genomes were downloaded from multiple sources, including 14,914 MAGs recovered from the IMG/M database⁶⁴, 1,516 genomes from a coassembly of MAGs from the freshwater Lake Mendota¹³, 1,353 genomes from the 2021 giant virus taxonomy framework⁷, 630 genomes for the TARA Oceans sequencing efforts^{9,16}, 227 isolate giant virus genomes, 65 from a sequencing effort for protists single-cell assembly genomes¹⁵ and 22 genomes from deep sea sediments¹⁰.

First, we did an initial dereplication using nucleotide identity. We calculated the pairwise ANI between MAGs using skani v 0.2.1 (settings: -slow -min-af 30 -m 200)⁶⁵. MAGs with at least 95% ANI and 30% alignment were considered in the same cluster (hereafter denoted as 'species-level cluster'). Each species-level cluster was chosen to have one MAG representative, preferring the MAG with the highest number of the nine giant virus orthologous groups (GVOGs) markers⁷. For reference, the nine GVOGs consist of 'GVOGm0003' (NCLDV major capsid protein), 'GVOGm0013' (DEAD/SNF2-like helicase), 'GVOGm0022' (DNA-directed RNA polymerase beta subunit), 'GVOGm0023' (DNA-directed RNA polymerase alpha subunit), 'GVOGm0054' (DNA polymerase family B), 'GVOGm0172' (transcription initiation factor IIB), 'GVOGm0461' (DNA topoisomerase II), 'GVOGm0760' (packaging ATPase), 'GVOGm0890' (poxvirus late transcription factor VLTf3).

Using the species-level cluster representatives and all genomes that do not fall into a species-level cluster (hereafter identified as singletons). To select the most likely correct copy of duplicated GVOGs we leveraged gene tree–species tree reconciliation based marker gene selection in sgtree (<https://github.com/NeLLi-team/sgtree>) to select the most likely correct copy of duplicated GVOG7 markers (includes DEAD/SNF2-like helicase, DNA-directed RNA polymerase alpha subunit, DNA polymerase family B, transcription initiation factor IIB, DNA topoisomerase II, packaging ATPase, poxvirus late transcription factor VLTf3) prior to alignment, supermatrix generation and tree inference⁷. This tree does not include *Mirusviricota* genomes as they are missing most of the *Nucleocytoviricota* hallmark genes. In brief, sgtree identifies the markers with HMMER v.3.4⁶⁶, aligns sequences using MAFFT v7.520, trims alignments with trimAl v1.4.rev1 (settings: -gt 0.1) and constructs the tree using Fasttree v2.1.11 (model: LG+G4)^{67–69}. To select between multiple copies of the same marker in a genome, sgtree infers a tree with each marker variant, compare the Robinson–Foulds distances of the trees to a tree generated with the best scoring markers according to HMMER and select the marker with the minimal Robinson–Foulds distance.

We filtered out gene markers shorter than 50% of the marker median length and then genomes with less than three GVOG4 ('GVOGm0461', 'GVOGm0022', 'GVOGm0023' and 'GVOGm0054') or two GVOG4 and three GVOG8 ('GVOGm0013', 'GVOGm0022', 'GVOGm0023', 'GVOGm0054', 'GVOGm0172', 'GVOGm0461', 'GVOGm0760' and 'GVOGm0890') markers were removed from further dereplication. To further dereplicate genomes, we used the sgtree phylogenetic tree and converted it into a PDM then further converted the distance into a similarity value (maximum 1) using PhyloDM (<https://github.com/aaronmussig/PhyloDM>). We generated a set of similarity tables, with different cut-off values for similarity, keeping the distance for similarity higher than the cut-off, from 0.98 to 0.05 with steps of 0.01. All similarity tables were clustered using MCL (inflation 2), obtaining clusters of phylogenetically close MAGs (hereafter identified as 'PDM-cluster'). Using the taxonomic assignments for the genomes from the giant virus taxonomy framework⁷, we counted how many genomes from different genera ended in the same MCL cluster for

each cut-off value. We selected the lowest cut-off number that did not collapse different genera in the same cluster. For each PDM-cluster we selected a representative genome, favouring the genomes with the highest number of GVOG markers. The final genome phylogenetic tree consists of the PDM-cluster representatives, all unclustered genomes (that is, singletons) and the addition of 5 Procul and 194 *Egrovirales* genomes, resulting in 2,037 genomes. Our final tree was inferred using IQtree v2.3.5 Q.pfam+R10+F with 1000 ultrafast bootstraps and 1000 SH-aLRT in sgtree and the GVOG7 HMM markers⁷⁰. The obtained tree was rooted on the Proculviricetes. We additionally tested our procedure with the genomes used in Aylward et al.⁷ to see if we were able to replicate the tree. The topology of the resulting trees was dependent more on the selection of the gene markers than to the parameters of the phylogenetic inference: our tree built without gene selection resulted in the splitting of a part of Imitervirales, with a common ancestor with *Algavirales* and *Pandoravirales*.

To build Supplementary trees to consider alternative topologies, we used the PDM-clustered representatives to build trees using concatenated GVOGs and single GVOG trees. We specifically used a concatenation of four markers (GVOG4) and a concatenation of eight markers (GVOG8). For GVOG4, we kept all representatives with at least three GVOG4 markers and re-ran sgtree, obtaining 1,838 genomes, plus the 5 Procul and 194 *Egrovirales*. The final tree was inferred using IQtree v2.3.5 Q.pfam+R10+F with 1,000 ultrafast bootstraps and 1,000 SH-aLRT. The GVOG4 tree had limited phylogenetic signal and reduced the number of genomes usable in the phylogeny with enough markers. The relations between the orders were not clearly solved as their internal nodes often had limited support. For example, the basal taxa (that is, basal *Chitovirales* and basal *Pokkesviricetes*) were nested inside Imitervirales and inside *Algavirales*, respectively.

For the GVOG8 tree, we added GVOGm0022 to the original GVOG7 markers. First, we identified GVOGm0022 sequences using HMMER v3.4 on the PDM-cluster representatives, aligned the protein sequences with MAFFT v7.520 and trimmed with trimAl v1.4.rev1 (settings: -gt 0.1). We built a single-gene tree on the GVOGm0022 sequences with IQtree v2.3.5 LG+F+R4. The tree was manually inspected to identify if paralogs were present, and we found two clades of Imitervirales ($n = 631$) showing presence of paralogs that form sister clades to Pimascovirales (Supplementary Fig. 15). Therefore, we subsetting the GVOGm0022 alignment, keeping the sequence with the best HMMER bit score for each genome. The sequences were realigned and trimmed, and added to the GVOG8 concatenated alignment, and the tree was inferred with IQtree v2.3.5 Q.pfam+R10+F with 1,000 ultrafast bootstraps and 1,000 SH-aLRT. The GVOG8 tree had low support for the internal nodes separating the orders, with *Algavirales* paraphyletic with some of its clades monophyletic with *Pandoravirales*, a topology inconsistent with other trees and literature.

For the 611 recovered *Mirusviricota* MAGs, we combined the genomes with 88 *Mirusviricota* MAGs from TARA Oceans¹⁶. Dereplication was done with skani v 0.2.1 (settings: -slow -min-af 30 -m 200) and MAGs with at least 95% ANI and 30% alignment were considered in the same cluster⁶⁵. However, only 145 clusters and singletons had enough markers for phylogenetic distance matrix clustering, resulting in 55 clusters and singletons.

Taxonomic assignments of GVMAGs

To assign the taxonomic rank of our final dataset, we used the R package Castor v1.8.2 to calculate the RED scores for the rooted phylogenetic tree where each node has a score calculating the relative position between the root and the tips^{71,72}. The calculated RED scores were used to assign the taxonomic classification without overlaps between different levels of taxonomic ranks, using the genomes from the Framework dataset as a reference. To prevent misclassifications based on isolated genome placements, we assigned family taxonomy if at least three genomes were present in the taxa, otherwise unclassified families

were labelled with ‘incertae sedis’. We further assigned genus taxonomy using the calculated RED scores. Using our expanded giant virus database, we identified new families and genera using RED score analysis on our species tree that was inferred from representative genomes obtained after PDM clustering of viral species. For newly proposed families and genera, we assigned non-redundant identifiers following previous nomenclature (Aylward et al.⁷). For *Mirusviricota* genomes, only clusters and singletons with enough markers for phylogenetic clustering (following steps as described above) were used for the RED score analysis. All 55 *Mirusviricota* representatives follow a sequential naming convention (orders: m1/m2, families: m1_1/m2_1, genera: m2_1g_1) (Fig. 2).

Coding potential and inferred functional capabilities of GVMAGs

To consider alternative genetic codes for giant virus genomes, we used GVClass v1.0²⁹. GVClass tests different genetic codes and evaluates results based on a *hmmsearch* using the general models. Currently, GVClass will select the optimal code based on the genetic code that yields the largest number of matches to general models (that is, specific gene orthogroups) with the highest average bitscore and the highest coding density. Genome completeness is calculated as the count of genes conserved in 50% of genomes of the respective *Nucleocytoviricota* order. When needed to confirm the lack of contamination, a BLASTP with diamond v2.1.9 search was performed^{73,74}.

We further identified orthologous genes using Proteinortho v6.3.3 (settings: -p=diamond -e=1e-5 -identity=25 -cov=50 -sim=0.80, --core) with all genomes included in the phylogenetic tree or were clustered through species-level or PDM⁷⁵. We performed a smaller orthologue search using the non-redundant GVMAGs. We used gephi to visualize the orthogroups shared by at least three genomes, using a bipartite graph where genomes and orthogroups are nodes and the presence of genes as weightless edges⁷⁶. We used a ForceAtlas2 layout with default settings and 100 iterations⁷⁷. ForceAtlas2 repels nodes from one another, while using edges to pull connected nodes together, thus creating a graph where highly connected nodes cluster together. We calculated a within-order and within-family connectivity ratio, defined by the number of edges shared with species within the same order divided by the total number of edges where a value closer to one indicates a stronger connection to its own order compared with its connections to other orders.

To evaluate the impact of genome sampling on orthogroup formation, we compared singleton rates before and after adding 7,013 genomes (those recovered from the IMG/M dataset not included in previous studies) to previously published 1,495 representatives (includes genomes from TARA Oceans, Aylward et al.⁷, isolate genomes and deep sea sediment genomes). We used a Poisson rate test for each set of genomes.

Finally, we assessed the completeness of metabolic modules based on the metabolic module definitions of KEGG⁷⁸. The KEGG KOs functional annotations of every genome were extracted from the results of EggNOG-mapper v2.1.10⁷⁹; when more than one KO was assigned to a protein, all the KOs were unnested and considered individual KOs for downstream analyses. The list of KOs per genome was then passed to the script *ko_mapper.py* from MicrobeAnnotator v2.0.5 to map KOs to their respective modules and compute metabolic module completeness per genome⁸⁰.

To identify family-specific enrichment of metabolic modules, we first converted module completeness scores to a simple binary variable (present/absent). Any value over 0 was considered present. For every module we then counted, within each focal family, how many genomes contained the module and how many did not and contrasted these numbers with the corresponding counts from all other families combined. The resulting 2 × 2 contingency tables were analysed with Fisher's exact test to obtain odds ratios and *P* values, which were subsequently

adjusted for multiple testing with the Benjamini–Hochberg procedure. Infinite odds ratios arising from zero counts were truncated to 5×10^3 for downstream display.

To identify antibiotic resistance gene homologues, we used DeepARG v1.0.2 which identifies a broader diversity of ARGs using its own database (--model LS --min-prob 0.8 --arg-alignment-identity 25 --arg-alignment-evalue 1e-10 --arg-num-alignments-per-entry 1000)⁸¹. For identification of BGCs, we used antiSMASH 8.0 and BLASTP (-evalue 1e-5) against the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database^{73,82,83}. Moreover, we used BLASTP against the NR database on all contigs containing putative BGCs to confirm they are nested within *Nucleocytoviricota* genes. We used BiG-SLICE v2⁸⁴ to cluster BGCs by similarity and cytoscape web 1.0⁸⁵ to visualize networks.

Identification of GEVEs

We screened 219 algal genomes and 2,487 eukaryotic bins from the PhycoCosm and MycoCosm databases for the presence of putative GEVEs^{18,19}, along with 570 eukaryotic MAGs in the IMG/M databases and 115 eukaryotic genomes from Phytozome. We first identified putative GEVEs using ViralRecall v2.0³⁴ and masked nucleotide genomes. We then used GVClass v1.0 on curated protein files generated by the respective database to identify contigs with hits to eukaryote Benchmarking Universal Single-Copy Orthologs (BUSCO)⁸⁶ and Giant Virus Orthologous Groups (GVOGs) or *Mirusvirus* Orthologous Groups (mOGs)²⁴ profile HMMs from the newly generated giant virus genomes from this analysis. For this analysis, an *e*-value of 1×10^{-5} was applied as a cut-off. Finally, we used the nearest neighbour phylogenetic tree result from GVClass to identify GEVEs that were phylogenetically placed within proximity to a *Nucleocytoviricota* gene. In GEVE analyses using GVClass, we consider GEVEs with at least one match to a viral major capsid protein or matches to at least three unique GVOG markers.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The complete dataset is available via Zenodo at <https://doi.org/10.5281/zenodo.20126600> (ref. 87) and can be explored interactively at <https://gvmagdb.newlineages.com/>. Nucleotide assembly files for MAGs are available via the European Nucleotide Archive under bioproject ID PRJEB113230. Source data are provided with this paper.

References

1. Raoult, D. et al. The 1.2-megabase genome sequence of *Mimivirus*. *Science* **306**, 1344–1350 (2004).
2. Schulz, F. et al. Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).
3. Schulz, F., Abergel, C. & Woyke, T. Giant virus biology and diversity in the era of genome-resolved metagenomics. *Nat. Rev. Microbiol.* **20**, 721–736 (2022).
4. Philippe, N. et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
5. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
6. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
7. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol.* **19**, e3001430 (2021).
8. Aylward, F. O. et al. Taxonomic update for giant viruses in the order Imitervirales (phylum *Nucleocytoviricota*). *Arch. Virol.* **168**, 283 (2023).

9. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5**, e3558 (2017).
10. Bäckström, D. et al. Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *mBio* **10**, e02497–18 (2019).
11. 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).
12. Chen, I.-M. A. et al. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
13. Oliver, T. et al. Coassembly and binning of a twenty-year metagenomic time-series from Lake Mendota. *Sci. Data* **11**, 966 (2024).
14. Rohwer, R. R. et al. Two decades of bacterial ecology and evolution in a freshwater lake. *Nat. Microbiol.* **10**, 246–257 (2025).
15. Schulz, F. et al. Single-cell genomics reveals complex microbial and viral associations in ciliates and testate amoebae. *Nat. Commun.* **16**, 10336 (2025).
16. Gaïa, M. et al. Egoviruses: distant relatives of poxviruses abundant in the gut microbiomes of humans and animals worldwide. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.03.23.586382> (2024).
17. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
18. Grigoriev, I. V. et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).
19. Grigoriev, I. V. et al. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* **49**, D1004–D1011 (2021).
20. Giant virus research agent. *NeLLi* <https://gvagent.newlineages.com/web> (2026).
21. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
22. Myers, J. M. et al. Discovery of giant viruses as past and present infections of zoospore fungi. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.01.04.574182> (2025).
23. Perini, L. et al. Giant viral signatures on the Greenland ice sheet. *Microbiome* **12**, 91 (2024).
24. Zhao, H., Meng, L., Hikida, H. & Ogata, H. Eukaryotic genomic data uncover an extensive host range of mirusviruses. *Curr. Biol.* **34**, 2633–2643 (2024).
25. Archibald, J. et al. Persistent mirusvirus infection in the marine protist *Aurantiochytrium*. *Nat. Commun.* **16**, 9922 (2025).
26. Fang, Y. et al. Genome-resolved year-round dynamics reveal a broad range of giant virus microdiversity. *mSystems* **10**, e0116824 (2025).
27. Shackelton, L. A. & Holmes, E. C. The role of alternative genetic codes in viral evolution and emergence. *J. Theor. Biol.* **254**, 128–134 (2008).
28. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
29. Pitot, T. M., Brûna, T. & Schulz, F. Conservative taxonomy and quality assessment of giant virus genomes with GVCClass. *npj Viruses* **2**, 60 (2024).
30. Schneider, S. U., Leible, M. B. & Yang, X.-P. Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. *Mol. Gen. Genet.* **218**, 445–452 (1989).
31. Hoffman, D. C., Anderson, R. C., DuBois, M. L. & Prescott, D. M. Macronuclear gene-sized molecules of hypotrichs. *Nucleic Acids Res.* **23**, 1279–1283 (1995).
32. Keeling, P. J. & Doolittle, W. F. A non-canonical genetic code in an early diverging eukaryotic lineage. *EMBO J.* **15**, 2285–2290 (1996).
33. Van Etten, J. L. Unusual life style of giant chlorella viruses. *Annu. Rev. Genet.* **37**, 153–195 (2003).
34. 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).
35. Benites, L. F., Stephens, T. G. & Bhattacharya, D. Multiple waves of viral invasions in Symbiodiniaceae algal genomes. *Virus Evol.* **8**, veac101 (2022).
36. Mckeown, D. et al. Multiple, diverse endogenous giant virus elements within the genome of a brown alga. *Virus Evol.* **11**, veaf009 (2025).
37. Erazo-Garcia, M. P. et al. Cryptic infection of a giant virus in a unicellular green alga. *Science* **388**, eads6303 (2025).
38. Pritham, E. J., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3–17 (2007).
39. Bellas, C. et al. Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proc. Natl Acad. Sci. USA* **120**, e2300465120 (2023).
40. Sun, T.-W. et al. Host range and coding potential of eukaryotic giant viruses. *Viruses* **12**, 1337 (2020).
41. Hess, S. M., Stanford, D. R. & Hopper, A. K. SRD1, a *S. cerevisiae* gene affecting pre-rRNA processing contains a C2/C2 zinc finger motif. *Nucleic Acids Res.* **22**, 1265–1271 (1994).
42. Posas, F. et al. Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 'two-component' osmosensor. *Cell* **86**, 865–875 (1996).
43. Delaroque, N. et al. The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. *Virology* **287**, 112–132 (2001).
44. Rolland, C. et al. Clandestinovirus: a giant virus with chromatin proteins and a potential to manipulate the cell cycle of its host *Vermamoeba vermiformis*. *Front. Microbiol.* **12**, 715608 (2021).
45. Weynberg, K., Allen, M. & Wilson, W. Marine prasinoviruses and their tiny plankton hosts: a review. *Viruses* **9**, 43 (2017).
46. Ha, A. D., Moniruzzaman, M. & Aylward, F. O. High transcriptional activity and diverse functional repertoires of hundreds of giant viruses in a coastal marine system. *mSystems* **6**, e0029321 (2021).
47. Yi, X. et al. Giant viruses as reservoirs of antibiotic resistance genes. *Nat. Commun.* **15**, 7536 (2024).
48. Barth, Z. K. Giant virus genomes are unlikely to be reservoirs of antibiotic resistance genes. *Nat. Commun.* **16**, 8539 (2025).
49. Yi, X., Liang, J.-L. & Li, J.-T. Reply to: giant virus genomes are unlikely to be reservoirs of antibiotic resistance genes. *Nat. Commun.* **16**, 8540 (2025).
50. Chatterjee, A. & Kondabagil, K. Giant viral genomic signatures in the previously reported gut metagenomes of pre-school children in rural India. *Arch. Virol.* **164**, 2819–2822 (2019).
51. Li, J.-T. et al. Discovering the potential role of giant viruses in antibiotic resistance spread. Preprint at *Research Square* <https://doi.org/10.21203/rs.3.rs-3747448/v1> (2024).
52. McGarvey, D. J. & Croteau, R. Terpenoid metabolism. *Plant Cell* **7**, 1015 (1995).
53. Jung, Y. et al. Function and structure of a terpene synthase encoded in a giant virus genome. *J. Am. Chem. Soc.* **145**, 25966–25970 (2023).
54. Park, C., Tsutsumi, S., Mitsunashi, T. & Fujita, M. Discovery of a new type of terpene synthase coded by an orphan gene in a giant virus. *Biochemistry* **64**, 3866–3871 (2025).
55. Machado, T. B. et al. Gene duplication as a major force driving the genome expansion in some giant viruses. *J. Virol.* **97**, e0130923 (2023).

56. Sun, T.-W. & Ku, C. Unraveling gene content variation across eukaryotic giant viruses based on network analyses and host associations. *Virus Evol.* **7**, veab081 (2021).
57. Seligmann, H. Localized context-dependent effects of the ‘ambush’ hypothesis: more off-frame stop codons downstream of shift codons. *DNA Cell Biol.* **38**, 786–795 (2019).
58. Bahir, I., Fromer, M., Prat, Y. & Linial, M. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* **5**, 311 (2009).
59. Colson, P., Fournous, G., Diene, S. M. & Raoult, D. Codon usage, amino acid usage, transfer RNA and amino-acyl-tRNA synthetases in *Mimiviruses*. *Intervirology* **56**, 364–375 (2013).
60. Willemsen, A., Manzano-Marín, A. & Horn, M. Novel High-quality amoeba genomes reveal widespread codon usage mismatch between giant viruses and their hosts. *Genome Biol. Evol.* **17**, evae271 (2025).
61. Mordstein, C. et al. Transcription, mRNA export, and immune evasion shape the codon usage of viruses. *Genome Biol. Evol.* **13**, evab106 (2021).
62. Dubey, S. et al. Advancement in algal bioremediation for organic, inorganic, and emerging pollutants. *Environ. Pollut.* **317**, 120840 (2023).
63. Ru, J., Xue, J., Sun, J., Cova, L. & Deng, L. Unveiling the hidden role of aquatic viruses in hydrocarbon pollution bioremediation. *J. Hazard. Mater.* **459**, 132299 (2023).
64. Villada, J. C. et al. A genomic catalog of Earth’s bacterial and archaeal symbionts. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.05.29.656868> (2025).
65. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
66. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
67. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
68. 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).
69. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
70. 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).
71. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).
72. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
73. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
74. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
75. 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).
76. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *Proc. Int. AAAI Conf. Web Soc. Media* **3**, 361–362 (2009).
77. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE* **9**, e98679 (2014).
78. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
79. 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. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
80. Ruiz-Perez, C. A., Conrad, R. E. & Konstantinidis, K. T. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC Bioinformatics* **22**, 11 (2021).
81. Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
82. Blin, K. et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res.* **53**, W32–W38 (2025).
83. Zdouc, M. M. et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res.* **53**, D678–D690 (2025).
84. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* **10**, giaa154 (2021).
85. Ono, K. et al. Cytoscape Web: bringing network biology to the browser. *Nucleic Acids Res.* **53**, W203–W212 (2025).
86. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
87. Vasquez, Y., Nardi, T. & Schulz, F. GVMAGs V2. *Zenodo* <https://doi.org/10.5281/zenodo.20126600> (2026).

Acknowledgements

We thank all JGI users whose data have been included in this work. We thank I. Shabalov, K. LaButti and S. Ahrendt for their help in accessing and understanding eukaryotic genomes. We thank all users who contributed data to this analysis.

Author contributions

T.N., Y.M.V., F.S. and T.W. designed the study. T.N. performed data collection, processed all GVMAGs and performed dereplication and phylogenetic analyses. Y.M.V. created figures, analysed genetic code diversity, functional enrichment and orthologue analyses, and ARG-like and BGC homologues. P.B. and Y.M.V. analysed giant endogenous viral elements across eukaryotic databases. G.M.T. and F.S. created and improved the giant virus AI agent. J.C.V., T.B. and M.F.R.-G. provided help in multiple analyses. T.M. and T.Y.J. provided restricted data for this analysis. GVMAGs data consortium provided public data for this analysis. Y.M.V., T.N. and F.S. wrote the paper. All authors provided input into the manuscript.

Funding

The work conducted by the US Department of Energy Joint Genome Institute (JGI) (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract no. DE-AC02-05CH11231. This project was funded by the European Community’s H2020 Programme H2020-MSCA-RISE 2019 under grant agreement no. 872767.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-026-02435-y>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41564-026-02435-y>.

Correspondence and requests for materials should be addressed to Yumary M. Vasquez or Frederik Schulz.

Peer review information *Nature Microbiology* thanks Laura Munoz Baena, Uri Sheyn and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

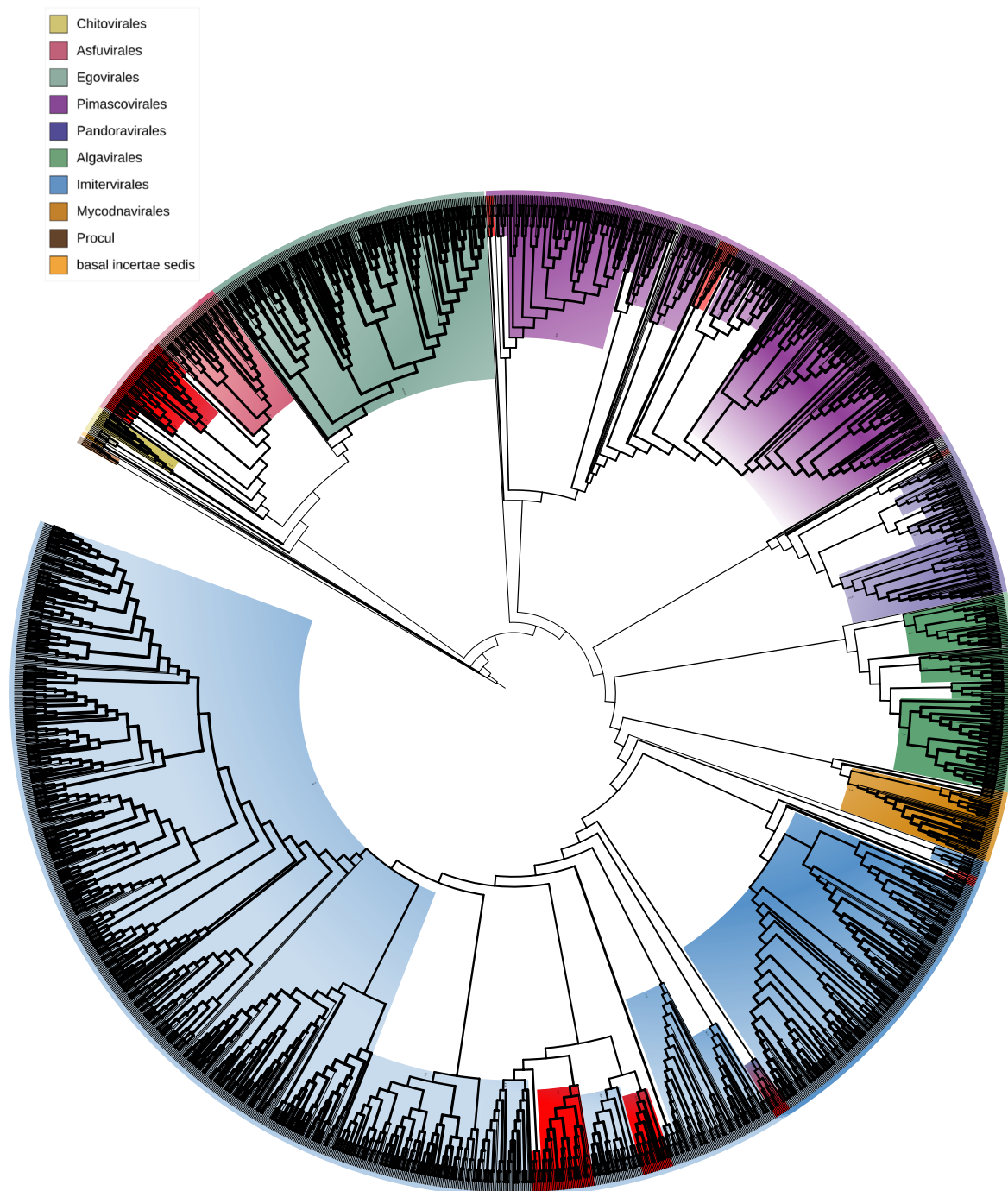
Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

GVMAGs data consortium

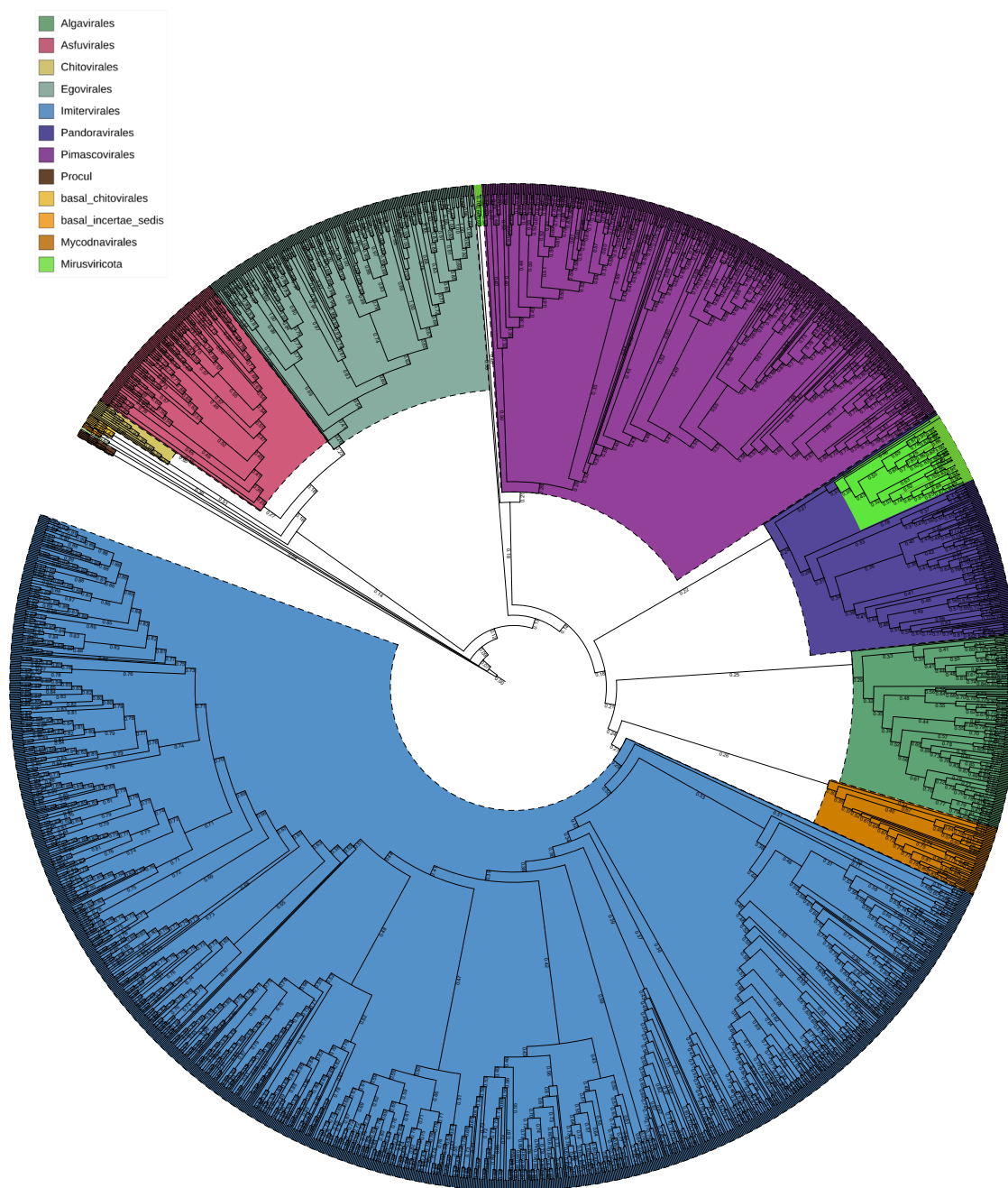
Katherine D. McMahon⁹, Rick Cavicchioli¹⁰, Steven Hallam¹¹, Barbara Campbell¹², Byron Crump¹³, Dana Hunt¹⁴, Karthik Anantharaman¹⁵, Holly Simon¹⁶, R. Michael McKay¹⁷, David Walsh¹⁸, Joan Slonczewski¹⁹, Konstantinos Konstantinidis²⁰, David L. Valentine²¹, Neslihan Tas²², Pia Moisander²³, Roland Hatzepichler²⁴, Matthew Sullivan²⁵, Sean Crowe^{11,26} & Nathan Ahlgren²⁷

⁹Departments of Civil and Environmental Engineering and Bacteriology, University of Wisconsin-Madison, Madison, WI, USA. ¹⁰School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney, New South Wales, Australia. ¹¹Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada. ¹²Department of Biological Sciences, College of Science, Clemson University, Clemson, SC, USA. ¹³College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA. ¹⁴Microbial Ecology, Duke University, Durham, NC, USA. ¹⁵Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA. ¹⁶Division of Environmental and Biomolecular Systems, Institute of Environmental Health, Oregon Health and Science University, Portland, OR, USA. ¹⁷Great Lakes Institute for Environmental Research, University of Windsor, Windsor, Ontario, Canada. ¹⁸Biology Department, Concordia University, Montreal, Quebec, Canada. ¹⁹Department of Biology, Kenyon College, Gambier, OH, USA. ²⁰School of Civil and Environmental Engineering, Georgia Tech, Atlanta, GA, USA. ²¹Department of Earth Science and Marine Science Institute, UC Santa Barbara, Santa Barbara, CA, USA. ²²Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ²³Department of Biology and Department of Estuarine and Ocean Sciences, University of Massachusetts Dartmouth, Dartmouth, MA, USA. ²⁴Department of Chemistry and Biochemistry, Thermal Biology Institute, Montana State University, Bozeman, MT, USA. ²⁵College of Arts and Sciences and College of Engineering, Ohio State University, Columbus, OH, USA. ²⁶Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, British Columbia, Canada. ²⁷Lasry Center for Bioscience, Clark University, Worcester, MA, USA.



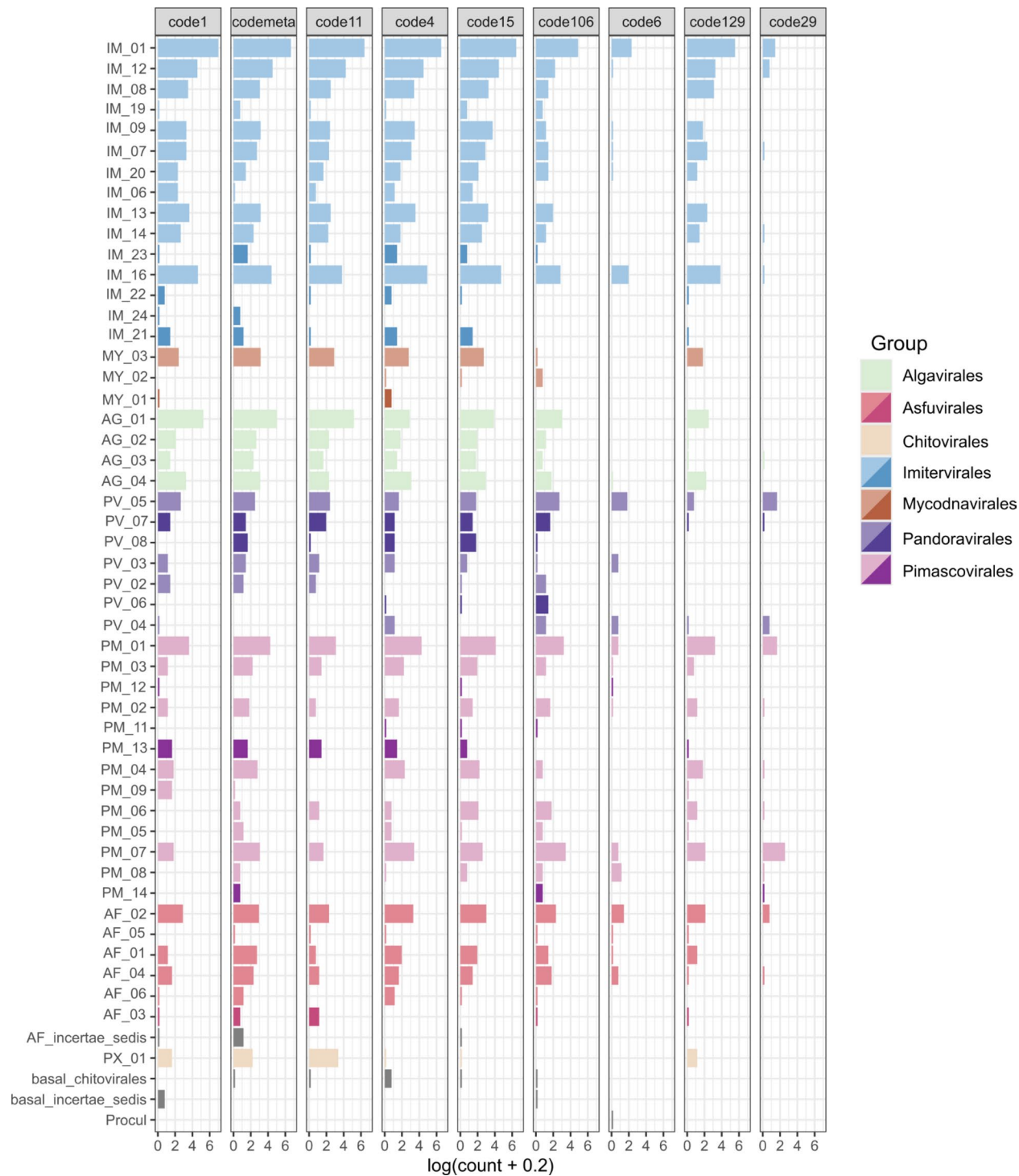
Extended Data Fig. 1 | Phylogenetic tree using GVOG7 markers with RED scores and taxonomic classification. GVOG7 markers consist of ('GVOGm0013' (DEAD/SNF2-like helicase), 'GVOGm0023' (DNA-directed RNA polymerase alpha subunit), 'GVOGm0054' (DNA polymerase family B), 'GVOGm0172' (Transcription initiation factor IIB), 'GVOGm0461' (DNA topoisomerase II), 'GVOGm0760' (Packaging ATPase), 'GVOGm0890' (Poxvirus Late Transcription

Factor VLTf3). The circular phylogenetic tree was constructed using seven GVOG marker genes. The tree displays distinct clades color-coded by taxonomic order. RED (Relative Evolutionary Divergence) scores are displayed on branches by width (*that is*, thicker widths indicate larger RED scores). Putative new families are highlighted in red throughout the tree. The interactive version of this tree is available at <https://itol.embl.de/shared/fmschulz>.

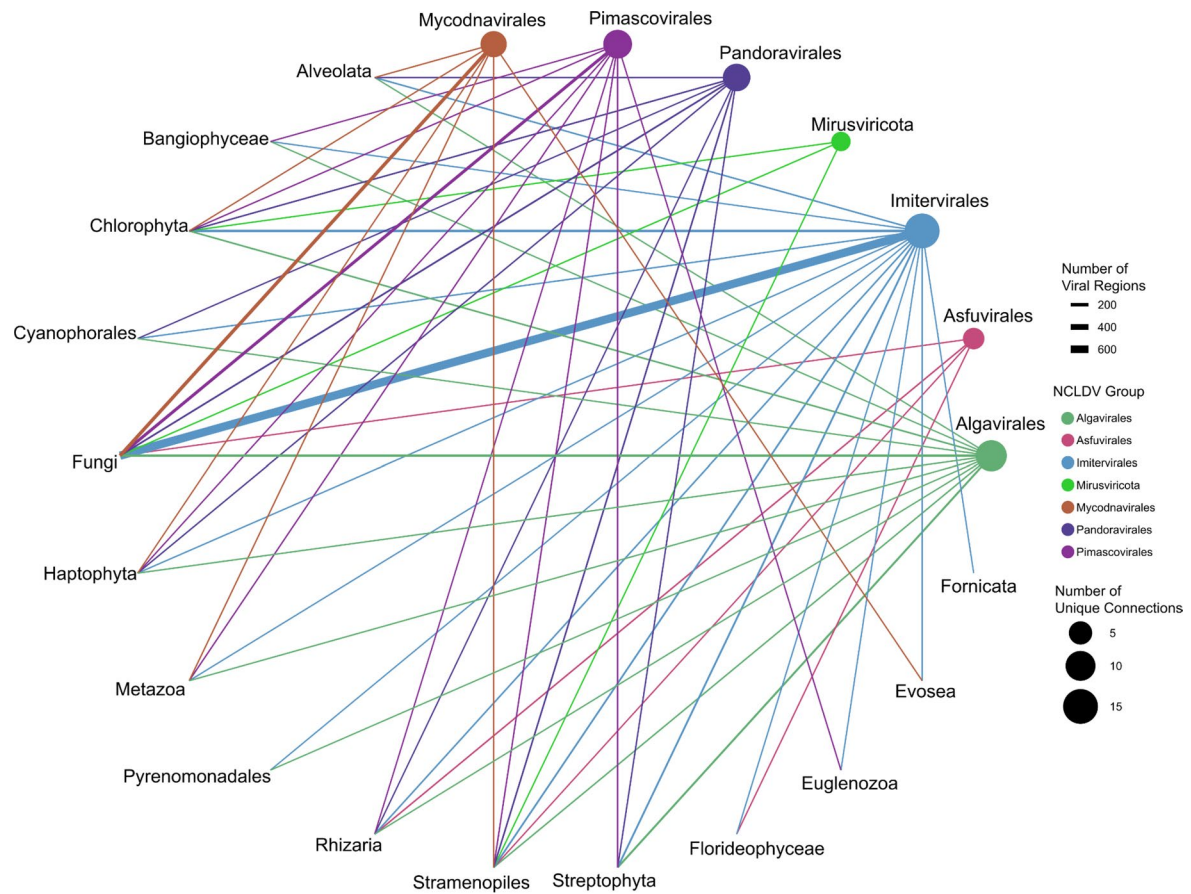


Extended Data Fig. 2 | Phylogenetic tree using GVOG7 markers with RED scores and taxonomic classification. GVOG7 markers consist of ('GVOGm0013' (DEAD/SNF2-like helicase), 'GVOGm0023' (DNA-directed RNA polymerase alpha subunit), 'GVOGm0054' (DNA polymerase family B), 'GVOGm0172' (Transcription initiation factor IIB), 'GVOGm0461' (DNA topoisomerase II), 'GVOGm0760'

(Packaging ATPase), 'GVOGm0890' (Poxvirus Late Transcription Factor VLTf3). The tree displays distinct clades color-coded by taxonomic order, 'Mirusviricota' represents the lime green clade nestled with Pandoravirales. RED (Relative Evolutionary Divergence) scores are displayed on branches. The interactive version of this tree is available at <https://itol.embl.de/shared/fmschulz>.

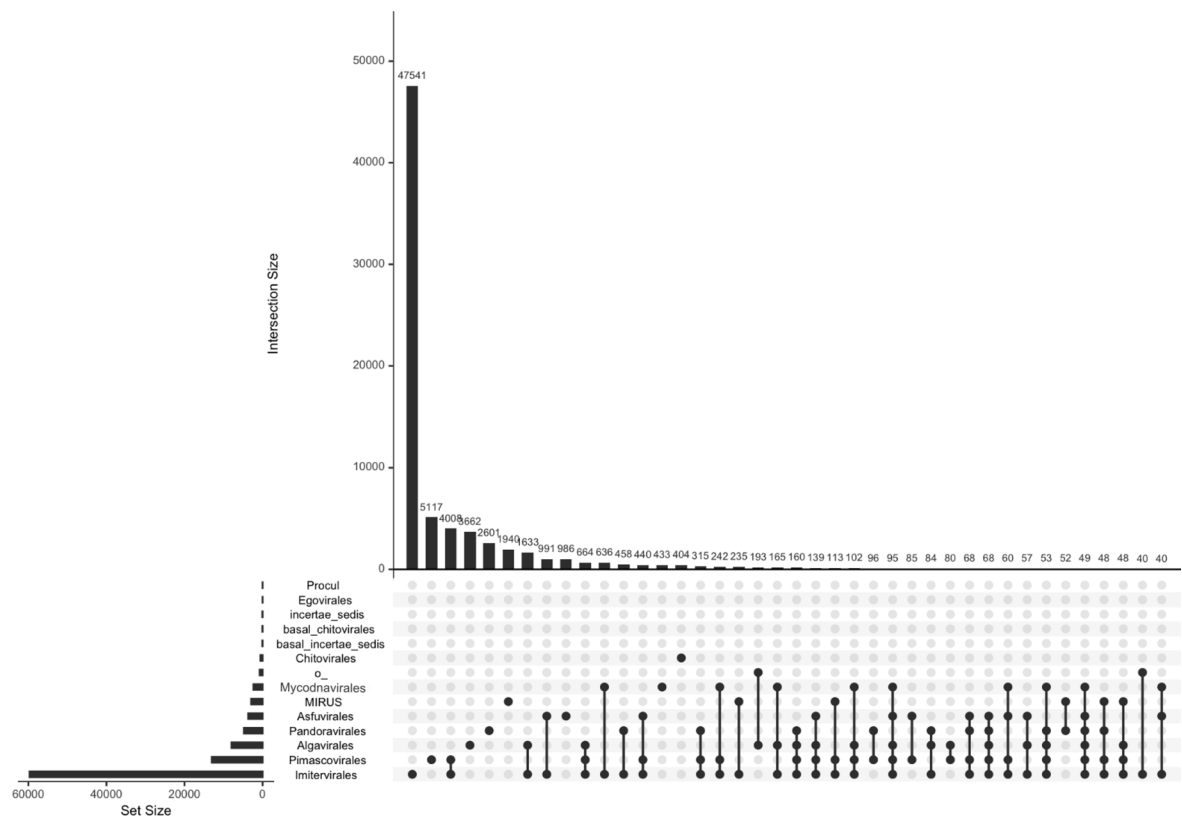


Extended Data Fig. 3 | Distribution of genetic codes across families. Frequency distribution of different genetic codes (code1, codemeta, code11, code4, code15, code106, code6, code129, and code29) across various families, the x-axis showing log-transformed (natural logarithm) genome counts, and families color-coded by taxonomic order. Dark colors within families indicate newly identified families.

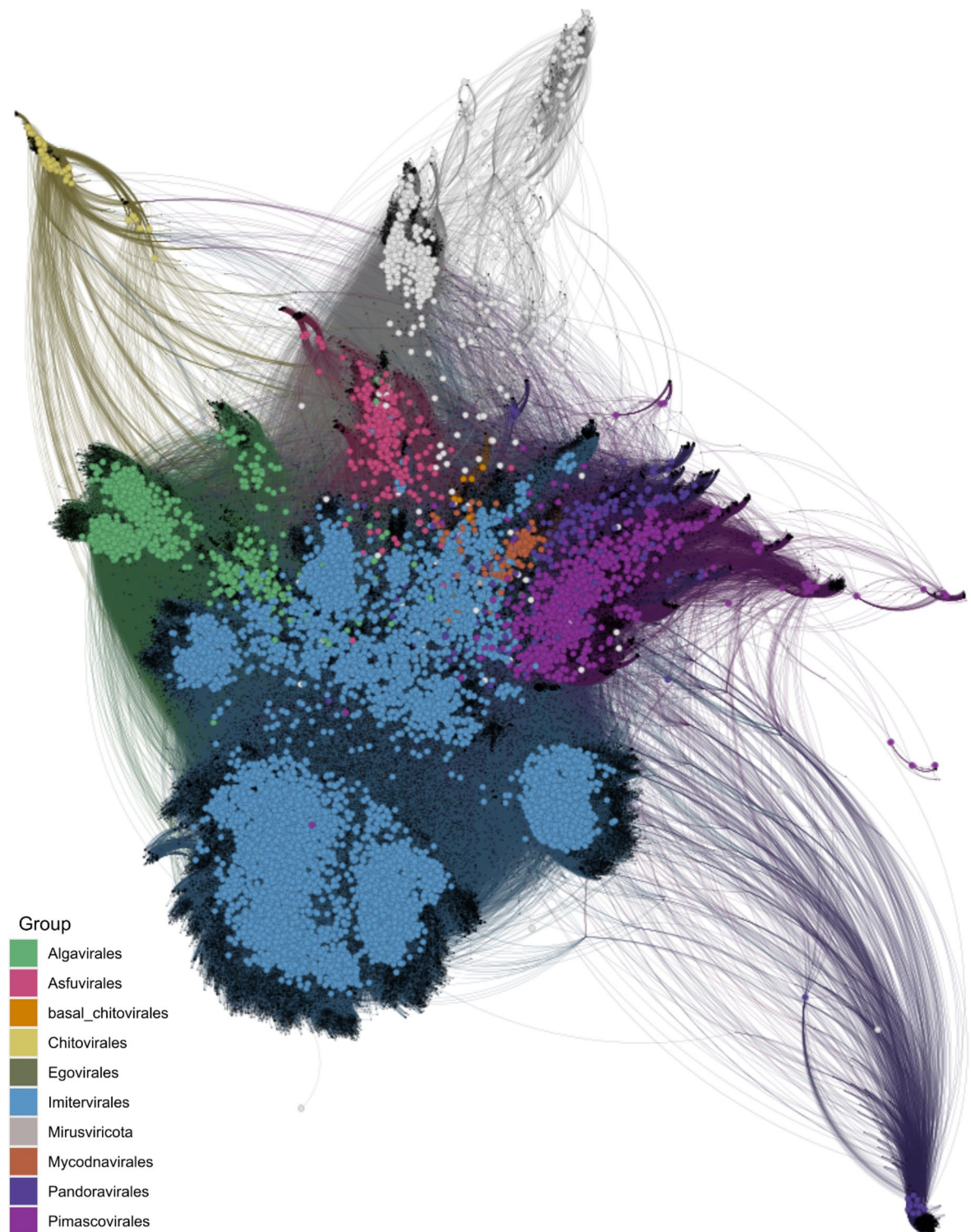


Extended Data Fig. 4 | Giant endogenized viral elements in IMG/M, PhycoCosm, Phytozome and MycoCosm. Putative *Nucleocytoviricota* origins are identified by color. Circle sizes are representative of the unique connections to a eukaryotic group. The width of the edges represent the number of viral regions found

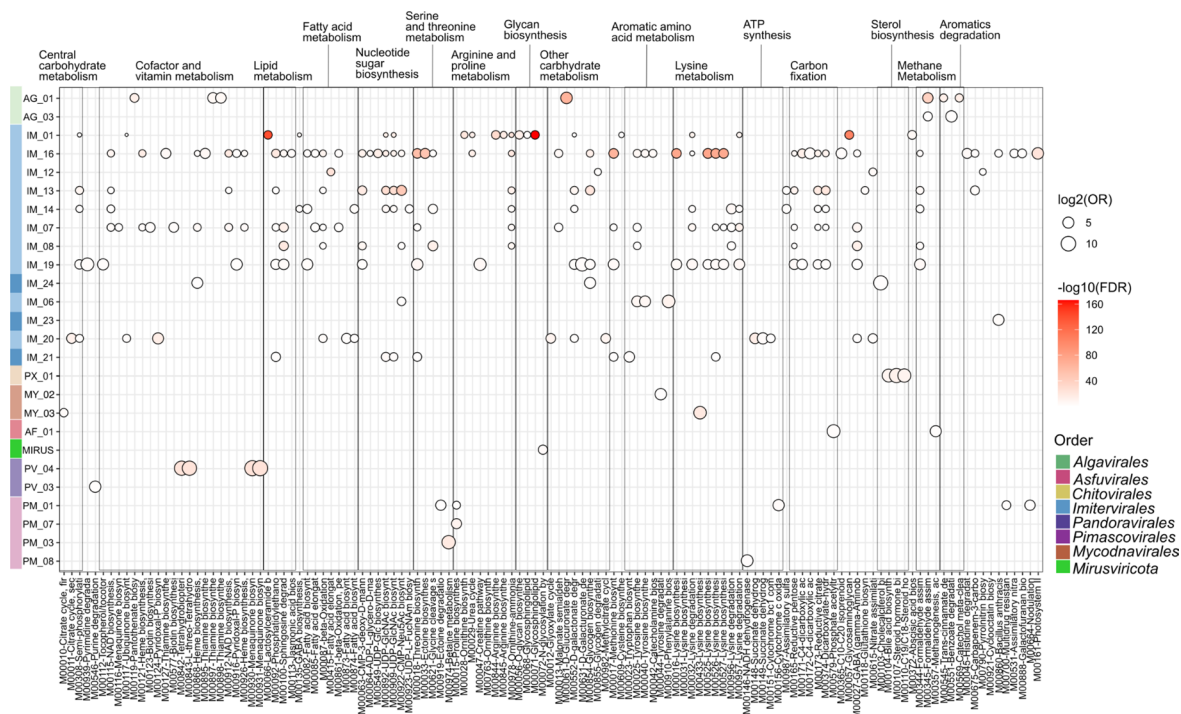
between groups using ViralRecall. Viral regions that did not have a taxonomic classification identified in GVClass were not included in these results. Complete ViralRecall results can be found in the Supplemental Tables.



Extended Data Fig. 5 | Upset plot of orthogroups across orders. Plot shows the distribution and overlap of orthogroups (clusters of orthologous genes) across different orders, with the top bar chart showing intersection sizes.



Extended Data Fig. 6 | Network analysis of orthogroup relationships across orders. Network visualization displays the relationships between orders based on shared orthogroups, with each colored cluster representing a different taxonomic order and edges connecting orders that share orthologous gene groups. Mirusviricota is shown in light grey.



Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| ☐ | ☒ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	no software was used to collect data
Data analysis	skani - v0.2.1, sgtree - commit 2e7f93d, MAFFT - v7.520, trimAl - v1.4.rev1, FastTree - v2.1.11, IQtree - v2.3.5, PhyloDM, MCL, Castor - v1.8.2, GVCClass - v1.0, diamond - v2.1.9, Proteinortho - v6.3.3, Gephi, EggNOG-mapper - v2.1.10, MicrobeAnnotator - v2.0.5, DeepARG - v1.0.2, antiSMASH - v8.0, BiG-SLiCE - v2, Cytoscape web - v1.0, HMMER - v3.4, EukCC - v2.1.3, ViralRecall - v1.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The complete dataset is freely available at <https://portal.nersc.gov/cfs/nelli/gvmagsV2/>.
Additional data is available at <https://yvasquez-lbl.github.io/GVMAGsV2/>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is based on metagenomic data generated at the Joint Genome Institute. We investigated genomes of nucleocytoplasmic large DNA viruses to generate a new phylogenetic framework, as well as understand the biotechnology potential of their genomes.
Research sample	Genomes were downloaded from IMG/M from the Joint Genome Institute. All 18,000 genomes are included in the information found in https://yvasquez-lbl.github.io/GVMAGsV2/
Sampling strategy	N/A
Data collection	Water samples were filtered and filters were placed in a -80C freezer until extraction and sequencing in 2018-2020.
Timing and spatial scale	N/A
Data exclusions	No data was excluded from this study.
Reproducibility	All data is available and detailed methods are provided to reproduce results.
Randomization	The study does not contain any experimental treatments so randomization is not needed.
Blinding	Blinding was not relevant to this study as these methods used are computational and blinding would not enhance the validity of the research outcomes.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involvement in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Not applicable
Novel plant genotypes	Not applicable
Authentication	Not applicable