Genome-resolved expansion of *Nucleocytoviricota* and *Mirusviricota* reveals new diversity, functional potential, and biotechnological applications

AUTHORS: Yumary M. Vasquez[†]*, Tiago Nardi[†]*, Gbocho Masato Terasaki, Petra Byl, Tomáš Brůna, Juan C. Villada, Miguel F. Romero, Thomas Mock, Timothy Y. James, GVMAGs data consortium, Tanja Woyke, Frederik Schulz*

† Shared first authorship

* Corresponding authors

GVMAGs data consortium

1. **Katherine (Trina) McMahon**
   trina.mcmahon@wisc.edu
   Departments of Civil and Environmental Engineering and Bacteriology, University of Wisconsin - Madison, Madison, WI, USA.
2. **Rick Cavicchioli**
   r.cavicchioli@unsw.edu.au
   School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney, Australia.
3. **Steven Hallam**
   shallam@mail.ubc.ca
   Department of Microbiology & Immunology, University of British Columbia, Vancouver, Canada.
4. **Barbara Campbell**
   bcampb7@clemson.edu
   Department of Biological Sciences College of Science, Clemson University, SC, USA
5. **Byron Crump**
   byron.crump@oregonstate.edu
   College of Earth, Ocean, and Atmospheric Sciences Oregon State University, Corvallis, OR
6. **Dana Hunt**
   dana.hunt@duke.edu
   Microbial Ecology, Duke University, NC, USA
7. **Karthik Anantharaman**
   karthik@bact.wisc.edu
   Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA.
8. **Holly Simon**
   simonholly128@gmail.com
   Division of Environmental and Biomolecular Systems, Institute of Environmental Health, Oregon Health & Science University, Portland, OR, USA.
9. **R. Michael McKay**
   robert.mckay@uwindsor.ca
   Great Lakes Institute for Environmental Research, University of Windsor, Ontario Canada
10. **David Walsh**
    david.walsh@concordia.ca
    Biology Department, Concordia, University, Montreal Canada
11. **Steven Hallam**
    shallam@interchange.ubc.ca
    Department of Microbiology & Immunology, University of British Columbia, Vancouver, Canada.
12. **Joan Slonczewski**
    slonczewski@kenyon.edu
    Department of Biology, Kenyon College, Gambier, OH, USA
13. **Konstantinos Konstantinidis**
    kostas@ce.gatech.edu
    School of Civil & Environmental Engineering, Georgia Tech, Atlanta, GA, USA
14. **David L. Valentine**
    valentine@geol.ucsb.edu
    Department of Earth Science and Marine Science Institute, UC Santa Barbara, Santa Barbara, CA, USA
15. **Neslihan Tas**
    ntas@lbl.gov
    Environmental Genomics and Systems Biology, Lawrence Berkeley National Lab, Berkeley CA, USA

16. **Pia Moisander**
pmoisander@umassd.edu
Department of Biology and Department of Estuarine and Ocean Sciences University of Massachusetts Dartmouth, MA, USA
17. **Roland Hatzenpichler**
roland.hatzenpichler@montana.edu
Department of Chemistry and Biochemistry, Thermal Biology Institute, Montana State University, Bozeman, MT, USA
18. **Matthew Sullivan**
sullivan.948@osu.edu
College of Arts and Sciences and College of Engineering Oregon State University Columbus, OH, USA
19. **Sean Crowe**
scrowe@eoas.ubc.ca
Departments of Microbiology and Immunology and Earth, Ocean, and Atmospheric Sciences. University of British Columbia, Vancouver, Canada
20. **Nathan Ahlgren**
nahlgren@clarku.edu
Clark University, Lasry Center for Bioscience, Worcester, MA, USA

ABSTRACT:
Nucleocytoplasmic large DNA viruses (NCLDV) of the phyla *Nucleocytoviricota* and viruses of the newly proposed *Duplodnaviria* phylum, *Mirusviricota*, exhibit taxonomic richness which continues to expand due to metagenomic sequencing of Earth's biomes. Giant viruses contain complex genomes encoding genes of both viral and cellular origin, representing a reservoir of unexplored biological functions with potential implications for ecology, evolution, and biotechnology. Here, we present the largest curated database of giant virus metagenome-assembled genomes (GVMAGs V2), comprising 8,508 species-level clusters inferred from 18,727 genomes, originating from marine, freshwater, anthropogenic and terrestrial environments, a six-fold increase from the previous giant virus phylogenetic frameworks. Phylogenomics and relative evolutionary distance analysis revealed 712 novel genera, 13 previously unknown viral families and a new proposed order, tentatively named *Mycodnavirales.* We improved gene calling of 12% of giant virus genomes by accounting for alternative and custom genetic codes, enabling more accurate identification of protein-coding genes. Database mining uncovered endogenous viral elements in a broad spectrum of eukaryotes, spanning algae, fungi, and parasitic protists highlighting that giant virus integration is both widespread and evolutionarily persistent. Orthologous clustering of 2.5 million proteins identified 135,998 orthogroups representing comprehensive metabolic capabilities, such as enrichment of genes involved in aromatic compound degradation (commonly associated with bioremediation) in *Algavirales* genomes. Furthermore, we detected widespread biosynthetic gene clusters underpinning antimicrobial activity and antibiotic resistance, suggesting roles of giant viruses in host defense and in the dissemination of antibiotic resistance genes. Conversely, 67% of orthogroups have unknown functions, underscoring a substantial unexplored potential. This comprehensive publicly available database provides a critical resource for the giant virus research community and a foundation for uncovering virus-host interactions, exploring viral evolution, and identifying reservoirs for novel enzymes with the potential to advance biotechnological applications.

MAIN:
Nucleocytoplasmic large DNA viruses (*Nucleocytoviricota*), commonly referred to as giant viruses (GVs), represent a diverse group of viruses that challenge traditional virological paradigms due to their large size, genomic complexity, and unique biological features (Raoult *et al.*, 2004; Abergel *et al.*, 2007; Yutin *et al.*, 2013; Schulz *et al.*, 2017). 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 (Raoult *et al.*, 2004; Philippe *et al.*, 2013; Schulz, Abergel and Woyke, 2022). Metagenomic surveys have discovered giant viruses across Earth's major biomes, suggesting they have a significant yet still poorly understood ecological role (Schulz, Abergel and Woyke, 2022).

Despite these metagenomic surveys, the taxonomic and functional landscape of *Nucleocytoviricota* remains fragmentary because of sampling biases towards marine and freshwater ecosystems and isolated algae and protists (Legendre *et al.*, 2018; Mihara *et al.*, 2018; Moniruzzaman, Weinheimer, *et al.*, 2020; Schulz *et al.*, 2020; Aylward *et al.*, 2021, 2023).

Existing giant virus genome catalogues have laid groundwork for *Nucleocytoviricota* research, yet they still encompass only a fraction of the group's diversity and were generated with gene-calling pipelines that assume the standard genetic code for all genomes. Consequently, some open reading frames remain unannotated and occasional cellular sequences may fail quality filters. Expanding *Nucleocytoviricota* metagenome resources with new sampling ecosystems and with better protein prediction will lead to a more comprehensive comparative analysis. Additionally, these resources can be used to clarify lineage-specific metabolic innovations and increase our understanding of *Nucleocytoviricota* diversity and evolution.

Here we have assembled GVMAGs V2, the largest curated collection of giant virus metagenome-assembled genomes (GVMAGs) to date, a four-fold increase from the original GVMAGs database (Schulz et al., 2020) and a six-fold increase from the previous 1,382 species-level representatives (Aylward et al., 2021). We screened public assemblies in IMG/M and integrated data from additional sources, recovering 18,727 giant virus genomes (Benson *et al.*, 2012; Tully *et al.*, 2017; Bäckström *et al.*, 2019; Moniruzzaman, Martinez-Gutierrez, *et al.*, 2020; Aylward *et al.*, 2021; Chen *et al.*, 2023; Gaïa *et al.*, 2023; Oliver *et al.*, 2024; Schulz *et al.*, 2024; Rohwer *et al.*, 2025). De-replication at 95% average nucleotide identity resulted in 8,508 species-level clusters, and resolved 712 previously unrecognised genera. Subsequent phylogenomic analysis supported the proposal of 13 new viral families, markedly expanding the current taxonomy. Additionally, we propose a new order (*Mycodnavirales*) which is basal to *Imitervirales,* and incorporate genomes from the recently described *Duplodnaviria* phylum, *Mirusviricota* (Gaïa *et al.*, 2023)*,* 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. We increased more complete profile hits and higher average scores in 1,066 genomes, indicating an improvement in annotation quality and recovering >1300 previously missed genes. We mined PhycoCosm, MycoCosm and IMG/M databases to recover endogenous viral elements in eukaryotic genomes (Grigoriev et al., 2014, 2021; Chen et al., 2019). Ortholog clustering of 2.5 million proteins resolved 135,998 gene families, only 33% of which have functional assignments, revealing a significant reservoir of unknown biology. Module-level pathway reconstruction revealed order- and family-specific enrichment for photosynthesis, formaldehyde assimilation and aromatic degradation offering new insights into virus-driven metabolic reprogramming. In addition, we identified widespread biosynthetic gene clusters and antibiotic resistance genes within giant virus genomes, highlighting their potential as reservoirs and vehicles for novel enzymes, bioactive compounds, and resistance determinants of biotechnological interest. To enable easy exploration of these data, we provide an interactive, agentic system (https://newlineages.com/pages/tools/gv-ai-agent). Thus, GVMAGs V2 provides a phylogenetic and functional resource for enabling comparative analyses of giant virus evolution, ecology, and biotechnological potential.

RESULTS & DISCUSSION

New families, genera, and order from an expanded giant virus genome database

Expanding the genomic catalog of giant viruses is essential for improving taxonomic resolution and uncovering novel lineages. We collated a total of 18,727 genomes from multiple data sources, including genomes recovered from the IMG/M database (Chen *et al.*, 2023; Villada *et al.*, 2025), a co-assembly and single assembly of MAGs from the freshwater Lake Mendota (Rohwer *et al.*, 2023; Oliver *et al.*, 2024), genomes from the giant virus taxonomy framework (Moniruzzaman, Martinez-Gutierrez, et al., 2020; Aylward et al., 2021), TARA Oceans sequencing efforts (Tully *et al.*, 2017; Gaïa *et al.*, 2023), isolate giant virus genomes (Benson *et al.*, 2012), a recent protists single cell genome study (Schulz et al., 2024), and from deep sea sediments (Bäckström et al., 2019) (Figure 1a). First, dereplication using Average Nucleotide Identity (ANI) resulted in 8,508 non-redundant species-level genomes (Roux *et al.*, 2019), separated into 2,484 clusters and 6,024 singletons (hereafter known as species-level-clusters; see Methods for more information) (Figure 1b, Supplemental Table 1). The species-level-clusters indicated here represent a four-fold increase from the original GVMAGs database (Schulz et al., 2020) and a six-fold increase from the previous 1,382 species-level representative framework (Aylward et al., 2021). Following phylogenetic tree building, 469 genomes were removed, including genomes that contained markers shorter than 50% of the marker median length. Using a phylogenetic based-pairwise distance matrix (PDM), genomes were further clustered into 851 clusters and 1,756 singletons (hereafter known as PDM-clusters; see Methods for more information) (Figure 1c). Lastly, filtering for a minimum number of giant virus orthologous groups (GVOGs), the final phylogenetic tree contained 2,037 genomes, including the addition of 4 *Proculovirales* genomes and 192 *Egovirales* genomes (Gaïa *et al.*, 2024) (Figure 1d). This phylogenetic tree was created using a concatenated alignment of seven GVOG markers identified in Aylward et al. (2021) (hereafter identified as GVOG7 PDM-clustered tree). This tree does not include *Mirusviricota* genomes as they are missing most of the *Nucleocytoviricota* hallmark genes.

Using the GVOG7 PDM-clustered tree, we quantified relative evolutionary distance (RED) scores ranging from 0 to 1, where groups with scores closer to zero represent deeper evolutionary lineages near the root, while those with scores closer to one indicate narrowly defined clades positioned near the terminal branches. The RED scores of the *Nucleocytoviricota* classes range from 0.18 to 0.2, whereas the values for the orders range from 0.26 to 0.32, families range from 0.35 to 0.75 and genera range from 0.76 to 0.98 (Supplemental Figure S1, Supplemental Table 2). 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 labeled with “incertae sedis”. Using this method, we recovered a total of 61 families, including 33 “incertae sedis” cluster representatives, which likely represent additional families with currently insufficient sampling, and 1,030 genera. 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 nonredundant identifiers following previous nomenclature (Aylward *et al.*, 2021). We identified 13 new family classifications within the entire *Nucleocytoviricota,* none of which contain single cultivated representatives. *Imitervirales*, the largest giant virus order, displayed five new family classifications, a notable increase from the 11 families identified in 2021 (Aylward *et al.*, 2021)

(Figure 2). New family classifications were also identified in *Pimascovirales* (n = 4), *Pandoravirales* (n = 3) and one new family within *Asfuvirales*. Most newly proposed families are a mixture of clustered and singleton genomes, with the exception of genomes in the newly proposed families: IM_24 and MY_01. These families only contain singletons with no species-level- or PDM-clusters, possibly indicating new genera and species that have yet to be fully described and require further sampling to confirm. At the genus level, we identified 712 new genera. Of these, 343 are singletons represented by only one genome. The remaining 369 genera contained 1,092 species-level clusters and 3,767 genomes (taking into account all strain genomes in the species-level clusters).

The GVOG7 phylogenetic tree additionally revealed a clade of *Imitervirales* genomes that split from the main clade (Figure 2), echoing findings from previous studies (Myers *et al.*, 2024). This clade comprises two previously identified *Imitervirales* families: IM_18 (hereafter known as MY_03) and IM_02 (hereafter known as MY_02; *Mycodnaviridae*), and a newly identified family (hereafter known as MY_01). Their members contain similar GC percentage and assembly length as the larger *Imitervirales* group. However, its monophyly outside of *Imitervirales* is strongly supported, with a RED score consistent with the order ranking. The most robustly supported position for this clade is as a basal group to the *Imitervirales*, indicated by high support values (99/79) at the common node in the final tree. In contrast, there is lower support for an alternative topology that positions this clade as a sister group to *Algavirales*, a configuration observed in the GVOG4 and GVOG8 trees (Supplemental Figures S2 and S3). This discrepancy is further highlighted in single gene trees (Supplemental Figure S4), where the placement varies: Packaging ATPase (GVOG760) and DNA polymerase family B (GVOG054) support positioning within *Imitervirales*, while Poxvirus Late Transcription Factor VLTF3 (GVOG890) supports the sister-to-*Algavirales* topology. Interestingly, DNA topoisomerase II (GVOG461) and Transcription initiation factor IIB (GVOG172) show a relationship where these genomes form a clade with other *Imitervirales* genomes outside the main *Imitervirales* and appear as a sister clade to *Algavirales*. One marker DEAD/SNF2-like helicase (GVOG013) unexpectedly places these genomes with *Asfuvirales*. Additionally, two markers DNA-directed RNA polymerase beta subunit (GVOG022; see Methods for additional information on GVOG022) and DNA-directed RNA polymerase alpha subunit (GVOG023) involve only a limited number of genomes, exhibiting high uncertainty due to restricted phylogenetic signals from the large number of taxa and short sequence lengths. Consequently, we designate the divergent clade as a new taxonomic order tentatively named *Mycodnavirales*, with the name derived from *Mycodnaviridae,* a proposed family of giant viruses associated with early-diverging fungi (Myers *et al.*, 2024; Perini *et al.*, 2024).

Furthermore, in our expanded data set we observe the division of known families into multiple non-monophyletic clades, which we denote with the clade name and a new number identifier. This phenomenon is evident in the *Imitervirales* family IM_09, which separates into three distinct families (IM_09, IM_19, IM_20), each comprising a mix of IMG/M genomes and genomes from other sources. A similar pattern is observed in the *Asfuvirales* family AF_01, where the family splits into four families (AF_01, AF_04, AF_05, AF_06). Each of these newly defined families (except AF_05) contains genomes from multiple independent sources, supporting their

taxonomic validity. Second, former singleton genera now form well-supported families through the addition of related genomes. We identified four families that emerged from individual singletons (AF_03, MY_01, IM_21, IM_24) and four additional families that formed from pairs of singletons (PM_14, PV_07, PV_08, 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 the new viral phylum *Mirusviricota*

Mirusviricota is a new phylum of giant viruses linked to herpesviruses (Gaïa *et al.*, 2023), and seem to be associated with diverse eukaryotes, in particular protists and algae (Zhao *et al.*, 2024; Archibald *et al.*, 2025). We recovered 611 *Mirusviricota* MAGs from biomes spanning aquatic systems (*i.e.*, marine, freshwater, non-marine saline/alkaline, deep subsurface, and thermal springs), terrestrial systems (*i.e.,* soils and rock endoliths), plant-associated niches (*i.e.,* roots, peat moss, and phyllosphere) and engineered environments (*i.e.,* anaerobic bioreactors, soil microcosms, and plant-growth chambers). We combined our newly recovered *Mirusviricota* MAGs from the IMG/M dataset, with 88 *Mirusviricota* MAGs from TARA Oceans (Gaïa *et al.*, 2023).

Dereplication of *Mirusviricota* MAGs resulted in 111 species-level-clusters and 298 singletons. Only 145 clusters and singletons contained enough markers to be further clustered with phylogenetic distance clustering (PDM), resulting in 55 clusters and singletons, with 2 clusters solely consisting of novel genomes (*i.e.,* IMG/M genomes). The complete taxonomic classification for all 55 Mirusviricota representatives follows a sequential naming convention (orders: m1/m2, families: m1_1/m2_1, genera: m2_1g_1) (Figure 2). The majority of PDM-clusters (n = 49) are positioned with *Pandoravirales*, consistent with previous phylogenetic reconstructions (Fang *et al.*, 2025). Following a similar RED score-based approach as we did for *Nucleocytoviricota*, we were able to establish a taxonomic framework of *Mirusviricota* comprising two orders (Supplemental Tables 3 and 4). To better understand the evolutionary relationship of *Mirusviricota* and *Nucleocytovirictoa*, we built a combined species tree (Supplemental Figure S5). In this tree, the *Mirusviricota* m2 order contains 3 families subdivided into 13 genera (each containing multiple genomes), with one genus comprising only two genomes that was retained due to the specific characteristics of *Mirusviricota*. This genus occupies a position basal to the common node containing 47 other *Mirusviricota* sequences and a *Pandoravirales* genome (“sylv”), which likely represents long branch attraction as evidenced by low support values. The *Mirusviricota* m1 order, consisting of five genomes, appears basal to all *Megaviricetes*. While support for grouping these five genomes together is robust, their position outside the main *Mirusviricota* group remains uncertain. This clade could represent a genuine evolutionary separation through independent acquisition of *Nucleocytoviricota* genes or an artifactual placement given their relative proximity to the main group and the moderate support for internal nodes within *Megaviricetes*. This second order contains a single family with all genomes classified in independent genera. One outlier genome, clusters near the basal *Chitovirales* lineage. Nevertheless, 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 unexplored (Shackelton and Holmes, 2008). Our analysis of 8,508 dereplicated GVMAGs showed improvements in protein-coding gene prediction in *Nucleocytoviricota* when applying alternative genetic codes (Figure 3). Leveraging alternate genetic codes, we performed a comparison of giant virus-specific gene calling optimization with GVClass and the widely used Prodigal with the -p meta option using a two-metric (Δ complete best hits vs. Δ average best-hit score) quadrant analysis (Figure 3A; Supplemental Figure S6) (Hyatt *et al.*, 2010; Pitot, Brůna and Schulz, 2024). Of the 8,508 genomes analysed, 6,965 (~82%) fall directly on one of the axes *(i.e.,* the two programs perform identically for at least one of the two metrics). The remaining 1,543 genomes show a clear directional shift with 1,066 genomes (12% of the total) residing in Quadrant I, where the optimized gene calling in GVClass recovers both more complete profile hits and higher average scores, indicating an improvement in annotation quality. The other 476 genomes (5.6%) plot in Quadrant IV, gaining additional hits at the expense of a modest decrease in score. The increase in annotation quality observed with GVClass highlights the importance of considering genetic code variations in genomic analyses, as they can lead to more accurate predictions of gene content and functionality.

To determine which specific genetic codes provided the greatest prediction accuracy across our dataset, we conducted a comparative performance analysis of multiple translation tables (Figure 3B; Supplemental Figure S7). The standard genetic code (code 1) proved to be the best optimized overall code for giant virus protein predictions, followed by one of the pre-trained models for codes 1, 4, and 11 (-p meta flag in prodigal; codemeta). Genetic codes with a single stop codon reassignments were favored by giant virus genomes more 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, respectively. Different viral families showed optimization for specific genetic codes. Members of *Chitovirales* showed optimized gene calling at code 11, making it the only family that showed the best optimization for a code outside of code 1 or pretrained models (-p meta flag in prodigal; codemeta). Currently, all confirmed giant virus isolates have been isolated from hosts that use the standard genetic code 1 (*e.g.,* amoebae, algae and other protists). Host-virus genetic code alignment is believed to facilitate efficient translation of viral proteins during infection (Bahir *et al.*, 2009; Colson *et al.*, 2013). Within a shared genetic code, hosts and viruses may still differ in their codon usage preferences. As is the case for multiple amoeba species, codon usage between giant viruses and their hosts differ, suggesting that codon usage matching alone may be a poor predictor of host range (Willemsen, Manzano-Marín and Horn, 2025). Nevertheless, considering alternative genetic codes in host prediction can reveal plausible candidates for giant viruses infecting new hosts not yet identified.

Two *Imitervirales* GVMAGs showed an increase in coding density and recovery of additional functional genes when using code 6. The first one in the family IM_12 increased its coding density from 13% to 93%, complete best hits increased from 2 to 34, and estimated genome completeness from 3 to 42%. The second one in family IM_16 showed a 52% coding density

376 improvement, 13 additional complete best hits, and an increased genome completeness from
377 15% to 25%. Code 6 translates both TAA and TAG to glutamine rather than stop and is found in
378 some species belonging to ciliates, algae and diplomonad protists (Schneider, Leible and Yang,
379 1989; Hoffman *et al.*, 1995; Keeling and Doolittle, 1996). Currently, the only association of a
380 giant virus to a protist that uses genetic code 6 involves the ciliate *Paramecium bursaria* that
381 harbors a symbiotic algae (*Chlorella spp.*) in its cytoplasm, which is infected by giant viruses
382 (Van Etten, 2003). However, there are currently no reports on giant viruses that directly infect *P.*
383 *bursaria*'s nucleus or use its translation machinery, and the two genomes recovered in this study
384 represent the first plausible candidates for giant viruses that may infect hosts using genetic code
385 6.
386
387 <u>Giant endogenous viral elements (GEVEs) are widespread across datasets</u>
388 To identify putative hosts, we mined for giant endogenous viral elements (GEVEs) in publicly
389 available eukaryote genomes. We identified a total of 150, 807, and 84 genomes with putative
390 GEVEs from IMG/M, MycoCosm and Phycocosm, respectively, using a nearest neighbor
391 phylogeny-based search (Grigoriev *et al.*, 2014, 2021; Chen *et al.*, 2019) (Supplemental Figure
392 S8). Our analysis uncovered GEVEs in lineages where they were not previously known, further
393 expanding the recognized host range of giant viruses. These new putative hosts of giant viruses
394 include Glaucophyta and Rhodophyta. We detected additional GEVEs in well-studied host
395 groups such as Symbiodiniaceae dinoflagellates (Benites, Stephens and Bhattacharya, 2022)
396 and Phaeophyceae brown algae (Stramenopiles) (Mckeown *et al.*, 2025). In green algae
397 (Chlorophyta), where latent GEVEs have been experimentally studied (Moniruzzaman,
398 Weinheimer, *et al.*, 2020; Erazo-Garcia *et al.*, 2025), we again found viral fragments.
399 *Nucleocytoviricota* and *Mirusviricota* GEVEs were also found in parasitic eukaryotes, including
400 cordyceps fungi and the human pathogen *Trichomonas.* Endogenization of smaller,
401 Maverick/Polinton-like viruses, which can parasitize giant viruses, are also widespread in
402 *Trichomonas* (Pritham, Putliwala and Feschotte, 2007; Bellas *et al.*, 2023). Overall, our findings
403 demonstrate that integration of giant virus DNA is widespread across diverse eukaryotic
404 lineages.
405
406 <u>Shared genes and enriched functions across giant virus orders</u>
407 The gene repertoires of giant viruses encompass both conserved functions and diverse auxiliary
408 genes, yet the majority of their coding potential remains uncharacterized (Abergel, Legendre
409 and Claverie, 2015; Sun *et al.*, 2020). Ortholog clustering enables a systematic assessment of
410 this functional landscape across lineages. Comparative genomic analysis of 2,504,503 genes
411 from dereplicated GVMAGs identified 135,998 orthologs (ProteinOrtho), spanning 1,693,149
412 genes. Only 75,781 orthologues were shared by at least three genomes (1,571,367 genes). Of
413 the orthologs shared between at least three genomes, only 10,132 (14%) contained paralogs
414 (as defined as multiple genes from a single genome in an orthogroup), indicating that the vast
415 majority (86%) are composed of single-copy genes. As giant virus genomes are defined by a
416 high number of gene duplications (Machado *et al.*, 2023), this unexpected low frequency of
417 orthogroups with paralogs warrants further investigation to determine if this reflects a genuine
418 biological phenomenon or a methodological artifact related to the number of genomes analyzed.
419 Furthermore, only 24,783 (33%) of these orthogroups had functional annotations assigned by

EggNOG-mapper, highlighting a significant proportion of orthogroups of unknown function. Known functions encompass diverse biological processes, including protein synthesis and modification, and DNA damage response and repair.

Orthogroup analysis revealed distinct distributions of orthologs across the *Nucleocytoviricota* orders (Supplemental Table 6, Supplemental Figure S9). *Imitervirales* exhibited the highest number of order-specific orthologs (n = 47,541, 63% of total orthologs, 42% of total genes in orthogroups, 34% of all genes in *Imitervirales*), followed by *Pimascovirales* (n = 5,117, ~7% of total orthologs, ~2% of total genes in orthogroups). *Imitervirales* and *Pimascovirales* share 4,008 orthologs that are exclusive to both groups. When considering *Imitervirales* and the newly proposed *Mycodnavirales* order, 636 orthologs were shared. *Mycodnavirales* possessed 433 order-specific orthologs (with an additional 9,364 singletons). Only 68 orthogroups were shared across all five major orders (*Imitervirales*, *Pimascovirales*, *Algavirales*, *Pandoravirales*, and *Asfuvirales*).

Network analysis using a ForceAtlas2 layout (Figure 4A; Supplemental Figures S10 and S11) revealed a tendency for viruses within the same order and family to cluster together. Using the ortholog result that calculates the number of edges between two species, we calculated a within-order connectivity, 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 to its connections to other orders (Supplemental Figure S12). *Imitervirales* displayed the highest median connectivity ratio (0.91). This high connectivity is further influenced by the IM_01 family, which exhibits a median within-family connectivity of 0.79 (Supplemental Figure S13). Previous studies comparing giant viruses across diverse environments typically report lower inter-order gene conservation (Schulz *et al.*, 2020), a pattern we also find in all orders outside of *Imitervirales* (all mean within-order values < 0.30). It is thought that similar eukaryotic host communities and physicochemical conditions impose convergent functional constraints (Sun and Ku, 2021). The unusually high within-order connectivity in *Imitervirales* may indicate two possibilities: high connectivity reflects the large number of *Imitervirales* genomes and the high number of *Imitervirales*-specific edges, or that, unlike other orders, their gene content is more strongly shaped by consistent host associations and environmental pressures across diverse habitats.

A total of 811,354 giant virus genes represented singletons, genes not assigned to any orthogroup. This accounts for 32.4% of all total genes analyzed in this study. To assess how genome sampling affects orthogroup formation, we compared singleton rates (Figure 4B) 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.,* 2021, isolate genomes and deep sea sediment genomes). Using a Poisson rate test, we found that the singleton rate significantly decreased from 124.15 to 95.4 singletons per genome (rate ratio = 1.30, $p < 2.2\times10^{-16}$) after adding the new genomes. This 1.3-fold decrease in singleton genes demonstrates that expanding genomic sampling strongly reduces the number of orphan genes by revealing previously hidden orthologous relationships. These results suggest that the current number of apparent singletons in giant virus genomes is

likely inflated due to incomplete sampling of genomic diversity, and that many genes currently classified as taxon-specific may actually have undiscovered homologs in yet-to-be-sequenced organisms. In addition to reducing the singleton rate, our expanded dataset also led to a significant increase in the number of identified orthologs (Figure 4C). We identified 32,551 orthologs (out of the total 75,781 orthologues shared by at least three genomes) that do not contain genomes reported in the previous framework (Aylward *et al.*, 2021). This increase demonstrates the importance of comprehensive genomic sampling with newly sequenced genomes to understand the true extent of giant virus diversity and functional potential.

Having established patterns of gene sharing and taxonomic connectivity, we investigated differences in functional capacity using completeness of metabolic modules. Analysis of metabolic module completeness revealed variations across giant virus families (Figure 5A and Supplemental Figure S14). To identify family-specific enrichment of metabolic modules, we binarised module completeness scores (present/absent), compared each family's module frequency against that of all other families with 2 × 2 contingency tables, applied Fisher's exact tests, and adjusted the resulting P-values using the Benjamini-Hochberg procedure. We identified 220 distinct modules enriched across distinct families (Supplemental Figure S15). 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 prasinoviruses (Weynberg, Allen and Wilson, 2017). In contrast, *Imitervirales* genomes harbor genes for a different sugar manipulation process, the formaldehyde assimilation xylulose monophosphate pathway (M00344), as shown in Ha, Moniruzzaman and Aylward, 2021.

We additionally identified enriched modules that open new avenues for metabolic discovery. In particular, *Algavirales* is significantly enriched in three modules for aromatic degradation (M00545, M00551, M00569). These modules are found in *Algavirales* genomes recovered from marine and freshwater environments; including shale gas reservoirs, estuarine waters, lakes, rivers, ponds, and salt marshes. Giant viruses are already known to encode diverse metabolic genes that modulate host carbon metabolism during infection (Moniruzzaman, Martinez-Gutierrez, *et al.*, 2020); the presence of aromatic degradation genes in *Algavirales* likely reflects a similar strategy to infect host cells and reprogram host metabolism. The same meta-cleavage modules found in *Algavirales* support the degradation of natural and industrial pollutants in other microbial species, and could be potentially used in bioremediation (Ferrández, Garciá and Díaz, 1997; Elbehery and Deng, 2022; Ru *et al.*, 2023; Zhu *et al.*, 2024). Thus, giant viruses may represent an overlooked reservoir, and vehicle, of catabolic genes with potential for biotechnological exploitation.

Nucleocytoviricota and *Mirusviricota* as reservoirs for biotechnological potential

One important category often discussed for biotechnological exploration are antibiotic resistance genes (ARGs) and antimicrobial biosynthetic gene clusters. Recent work found that giant viruses can act as reservoirs of ARGs, with 12 ARG types being encoded by giant viruses (Yi *et al.*, 2024). Our genome dataset expanded this work (Figure 5B, Supplemental Table 7). Consistent with previous work, trimethoprim resistance genes (*dfr*) are the most abundant ARGs

in our dataset (n = 1,604 genes). Similarly, we found a large number of glycopeptides (*van;* n = 144 genes) and mupirocin resistance genes (n = 114). Our dataset also showed evidence of genes related to beta-lactam (two genes) and nitroimidazole (five genes) resistance that were absent from a recent larger scale survey of ARGs across giant virus genomes (Yi *et al.* 2024). This survey leveraged the previous giant virus taxonomy framework (Aylward et al., 2021), which included 1,382 species-level representatives compared to the 8,508 analyzed here. Beta-lactamase genes have been reported in giant viruses (Colson et al., 2020; Rigou et al., 2022), indicating that the earlier framework missed known functions likely due to limited genome representation at the time. Similarly, nitroimidazole resistance genes have been previously reported in giant viruses, but not identified in Yi *et al.*, 2024 (Chatterjee and Kondabagil, 2019; Li *et al.*, 2024). The beta-lactam and nitroimidazole resistance genes are found in three different ecological niches: freshwater lake, estuary water, and saline lake water. Beta-lactams are among the most widely prescribed antibiotics (Yimenu *et al.*, 2019; Centers for Disease Control and Prevention, 2022), and nitroimidazoles are often used to treat anaerobic bacterial and protozoal infections (Alexander, Gielen and Sartorelli, 1986). By recovering previously known but overlooked genes, absent from recent large scale surveys, our work provides an updated view of ARGs within giant viruses and their presence in diverse aquatic ecosystems highlights the need to consider giant viruses in antimicrobial resistance surveillance frameworks. Giant viruses may serve as reservoirs of biotechnologically relevant functions, as understanding their resistance mechanisms could enable the development of new genetic tools to address antibiotic resistance.

To further explore the functional potential of giant viruses, we surveyed biosynthetic gene clusters (BGCs) across our curated database using antiSMASH (Figure 5C, Supplemental Table 8) (Blin *et al.*, 2025). We identified BGCs in 326 giant virus genomes, highlighting secondary metabolic capacities within *Nucleocytoviricota* and *Mirusviricota*. The majority of predicted clusters fell within terpene-precursor (n = 233) and terpene (n = 49) categories. While terpene biosynthesis has only been examined in four viral genomes to date (Jung *et al.*, 2023; Park *et al.*, 2025), our data reveal that putative terpene biosynthetic genes are far more widespread, spanning all major *Nucleocytoviricota* orders, including *Algavirales, Imitervirales, Asfuvirales, Pandoravirales,* and *Pimascovirales*, as well as the new phylum, *Mirusviricota*. This suggests that the capacity for terpenoid production, which are compounds involved in membrane modification, signaling, or defense (McGarvey and Croteau, 1995), may be a common and ecologically relevant feature of giant viruses, with potential implications for virus-host interactions. Beyond terpenes, our analysis uncovered BGCs associated with polyketides (T1PKS and T3PKS) and peptide-derived compounds (nonribosomal peptide synthetases (NRPS)-like), categories well-known for their antimicrobial activities (O'Hagan, 1991; Cane, Walsh and Khosla, 1998; Cane and Walsh, 1999). These gene clusters were primarily concentrated within *Imitervirales*, *Pandoravirales*, and *Pimascovirales*, raising the possibility that some giant viruses may influence microbial community dynamics within host niches. We also observed order-specific patterns, including redox-cofactor clusters uniquely in *Algavirales*, and BGCs such as RiPP recognition element (RRE)-containing and ranthipeptide clusters exclusively in *Pandoravirales*, hinting at unexplored diversity in viral small-molecule biosynthesis. Likewise, the detection of deazapurine, a nucleotide modification pathway

(Hutinet, Swarjo and de Crécy-Lagard, 2017), in *Pimascovirales*, phosphonate gene cluster in *Imitervirales* and *Mirusviricota*, and quorum-sensing-like cyclic-lactone autoinducer clusters in both *Imitervirales* and *Pimascovirales*, underscores the functional complexity of giant virus genomes.

CONCLUSION:

Our study significantly expands the known diversity and functional potential of *Nucleocytoviricota* and *Mirusviricota*. Through phylogenomic and comparative genome analyses, we identified new viral diversity, characterized both conserved and lineage-specific gene content, and revealed novel genetic code usage and metabolic potential across major viral clades. Building upon previous classifications (Aylward et al., 2021), we identified one new order, 13 new families and 712 novel genera, including numerous singleton genomes that likely represent yet-undescribed lineages. These findings highlight the current underestimation of giant virus diversity and the need for further genome-resolved metagenomics, especially in underexplored habitats such as polar regions and the deep sea.

While most genomes are best annotated with the standard genetic code, several showed notable improvements in gene prediction using alternative genetic code. These cases, which were checked for contamination, raise the possibility of previously unrecognized host-virus associations and suggest that the landscape of giant virus-host coevolution may be broader than currently appreciated. Our survey of publicly available eukaryotic genomes further expands the known host range of giant viruses through discovery of GEVEs into algae, several parasitic fungi, and the human pathogen *Trichomonas*. This indicates that endogenization of giant virus DNA is a recurring feature of eukaryotic genome evolution.

Despite our study expanding the coding potential of giant viruses, the vast majority of genes (~67%) remain uncharacterized, reflecting both the novelty of giant virus gene content and the limitations of current annotation databases. Several limitations should be considered. Incomplete metagenome-assembled genomes may bias our view toward more readily assembled viral genomes and obscure the full spectrum of viral diversity. Uneven environmental sampling leads to overrepresentation of certain ecosystems and host types. Limited host information constrains ecological interpretations and understanding of genetic code variation. Nevertheless, even within these constraints, our large-scale genome mining uncovered an expanded repertoire of biosynthetic gene clusters and antibiotic resistance genes, positioning giant viruses as an underexplored reservoir of novel enzymes, metabolic pathways, and bioactive compounds with potential applications in biotechnology, synthetic biology, and drug discovery. Thus, our findings provide a foundation for future experimental efforts aimed at exploring the biotechnological potential of giant virus encoded genes and pathways.

METHODS:

Dereplication & phylogenetic inference of GVMAGs

A total of 18,727 genomes were downloaded from multiple sources, including 14,914 MAGs recovered from the IMG/M database (Villada *et al.*, 2025), 1,516 genomes from a coassembly of

MAGs from the freshwater Lake Mendota (Oliver *et al.*, 2024), 1,353 genomes from the 2021 giant virus taxonomy framework (Aylward *et al.*, 2021), 630 genomes for the TARA Oceans sequencing efforts (Tully *et al.*, 2017; Gaïa *et al.*, 2023), 227 isolate giant virus genomes (Benson *et al.*, 2012), 65 from a sequencing effort for protists single-cell assembly genomes (Schulz *et al.*, 2024) and 22 genomes from deep sea sediments (Bäckström *et al.*, 2019).

First, we did an initial dereplication using nucleotide identity. We calculated the pairwise average nucleotide identity between MAGs using skani v 0.2.1 (settings: -slow -min-af 30 -m 200) (Shaw and Yu, 2023). 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) markers (Aylward *et al.*, 2021). Using the species-level cluster representatives and all genomes that do not fall into a species-level cluster (hereafter identified as singletons), we used sgtree (commit 2e7f93d https://github.com/NeLLi-team/sgtree) to perform phylogenetic tree construction using the GVOG7 markers (Aylward *et al.*, 2021). Briefly, sgtree 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+F+I+G4) (Katoh *et al.*, 2002; Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009; Price, Dehal and Arkin, 2010). We filtered out gene markers shorter than 50% of the marker median length, and then genomes with less than three GVOG4 ("GVOGm0461", "GVOGm0022", "GVOGm0023", "GVOGm0054") or two GVOG4 and three GVOG8 ("GVOGm0013", "GVOGm0022", "GVOGm0023", "GVOGm0054", "GVOGm0172", "GVOGm0461", "GVOGm0760", "GVOGm0890") markers were removed from further dereplication. To further dereplicate genomes, we used the sgtree phylogenetic tree and converted it into a pairwise distance matrix 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 cutoff values for similarity, keeping the distance for similarity higher than the cutoff, 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 framework (Aylward *et al.*, 2021), we counted how many genomes from different genera ended in the same MCL cluster for each cutoff value. We selected the lowest cutoff number that did not collapse different genera in the same dataset. For each PDM-cluster we selected a representative genome, favoring the genomes with the highest number of OGs markers. The final genome phylogenetic tree consists of the PDM-cluster representatives, all unclustered genomes (*i.e.*, singletons), and the addition of 4 Procul and 192 Egovirales 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 markers. The obtained tree was rooted on the Proculviricetes. We additionally tested our procedure with the genomes used in Aylward et. al (2021) 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 correction for the incomplete lineage sorting resulted in the splitting of a part of Imitervirales, with a common ancestor with Algavirales and Pandoravirales.

To build Supplemental trees in order 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 3 GVOG4 markers, and re-ran sgtree, obtaining 1810 genomes, plus the 4 Procul and 192 Egovirales. The final tree was inferred using IQtree v2.3.5 Q.pfam+R10+F with 1000 ultrafast bootstraps and 1000 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 (*i.e.,* basal Chitovirales and basal Pokkesviricetes) were nested inside Imitervirales and inside Algavirales, respectively.

For the GVOG8 tree, we added GVOG0022 to the original GVOG7 markers. First, we identified GVOG0022 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 GVOG0022 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 (Supplemental Figure S16). Therefore, we subsetted the GVOG0022 alignment, keeping the sequence with the best HHMER bitscore for each genome. The sequences were realigned and trimmed, and added to the GVOG8 concatenate, and the tree was inferred with IQtree v2.3.5 Q.pfam+R10+F with 1000 ultrafast bootstraps and 1000 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.

<u>Taxonomic assignments of GVMAGs</u>

To assign the taxonomic rank of our final dataset, we used the R package Castor v1.8.2 to calculate the Relative Evolutionary Distance (RED) scores for the rooted phylogenetic tree where each node has a score calculating the relative position between the root and the tips (Louca and Doebeli, 2018; Parks *et al.*, 2018). 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. We assigned family taxonomy if there were at least 3 genomes in the taxa, otherwise family taxonomy was classified as "*incertae sedis*". We further assigned genus taxonomy using the calculated RED scores.

<u>Coding potential and inferred functional capabilities of GVMAGs</u>

To consider alternative genetic codes for giant virus genomes, we used GVClass v1.0 (Pitot, Brůna and Schulz, 2024). 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 (*i.e.,* 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 (Camacho *et al.*, 2009; Buchfink, Reuter and Drost, 2021).

We further identified orthologous genes we used 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 PDM (Klemm, Stadler and Lechner, 2023). We performed a smaller ortholog search using the non-redundant GVMAGs. We used gephi to visualize the orthogroups shared by at least 3 genomes, using a bipartite graph where genomes and orthogroups are nodes and the presence of genes as weightless edges (Bastian, Heymann and Jacomy, 2009). We used a ForceAtlas2 layout with default settings and 100 iterations (Jacomy *et al.*, 2014). ForceAtlas2 repels nodes from one another, while using edges to pull connected nodes together, thus creating a graph where highly connected nodes cluster together.

Finally, we assessed the completeness of metabolic modules based on the metabolic module definitions of KEGG (Kanehisa and Goto, 2000). The KEGG KOs functional annotations of every genome were extracted from the results of EggNOG-mapper v2.1.10 (Cantalapiedra *et al.*, 2021); 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 genome (Ruiz-Perez, Conrad and Konstantinidis, 2021).

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 genes, we used DeepARG v1.0.2 which identifies a broader diversity of ARGs (--model LS --min-prob 0.8 --arg-alignment-identity 25 --arg-alignment-evalue 1e-10 --arg-num-alignments-per-entry 1000). For identification of biosynthetic gene clusters, we used antiSMASH 8.0 and BLASTP (-evalue 1e-5) against the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database (Camacho *et al.*, 2009; Blin *et al.*, 2025; Zdouc *et al.*, 2025). Additionally, we used BLASTP against the NR database on all contigs containing putative biosynthetic gene clusters to confirm they are nested within *Nucleocytoviricota* genes. We used BiG-SLiCE v2 (Kautsar *et al.*, 2021) to cluster or BGCs by similarity and cytoscape web 1.0 (Ono *et al.*, 2025) to visualize networks.

<u>Identification of giant endogenous viral elements</u>

Putative endogenous viral elements (EVEs) were identified from 194 algal genomes and 1,599 eukaryotic bins in the PhycoCosm and MycoCosm (Grigoriev *et al.*, 2014, 2021), along with 751 eukaryotic MAGs in the IMG/M databases with GVClass v1.0 (Pitot, Brůna and Schulz, 2024). We identified putative EVEs using a nearest neighbor search from the GVClass output where EVEs that were in proximity with an *Nucleocytoviricota* gene were used for further analysis. The putative EVEs were further assessed to check if the detected viral signature was integrated into a eukaryotic contig to avoid analyzing viruses that were not truly endogenized. To do this, we selected contigs with hits to eukaryote Benchmarking Universal Single-Copy Orthologs (BUSCO) (Manni *et al.*, 2021) and Giant Virus Orthologous Groups (GVOGs) or Mirusvirus Orthologous Groups (mOGs) (Zhao *et al.*, 2024) profile hidden Markov models HMMER v.3.4 (Eddy, 2011) for downstream analysis. The hosts of these high confidence EVEs were assigned taxonomy through a consensus of EukCC v2.1.3 (Saary, Mitchell and Finn, 2020) and BLASTp (Camacho *et al.*, 2009) of the BUSCO hits.

Data Availability

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/.

Acknowledgements

The work conducted by the U.S. 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 U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We thank all JGI users whose data has been included in this work. This project was funded by the European Community's H2020 Programme H2020-MSCA-RISE 2019 under grant agreement No. 872767.

Conflicts of Interest

F.S. serves as CEO of SampleX. This work was not funded by, nor does it benefit, SampleX. All other authors declare no competing interests.

Author Information

Yumary M. Vasquez, Tiago Nardi, Petra Byl, Gbocho Masato Terasaki, Miguel F. Romero, Tomáš Brůna, Juan C. Villada, Tanja Woyke, Frederik Schulz

United States Department of Energy Joint Genome Institute, Berkeley, CA, USA

Tiago Nardi

Department of Biology and Biotechnology, University of Pavia, Italy

Gbocho Masato Terasaki

Department of Applied Math, University of California, Merced, Merced, CA, USA

Petra Byl

School of Ocean and Earth Science and Technology, University of Hawai'i Mānoa, Honolulu, HI USA

Miguel F. Romero

Instituto de Biología, Universidad Nacional Autónoma de México, 3er. Circuito Exterior S/N, Ciudad Universitaria, CP 04510 CDMX, México

Thomas Mock
School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK
Timothy Y. James
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI USA

Figures:

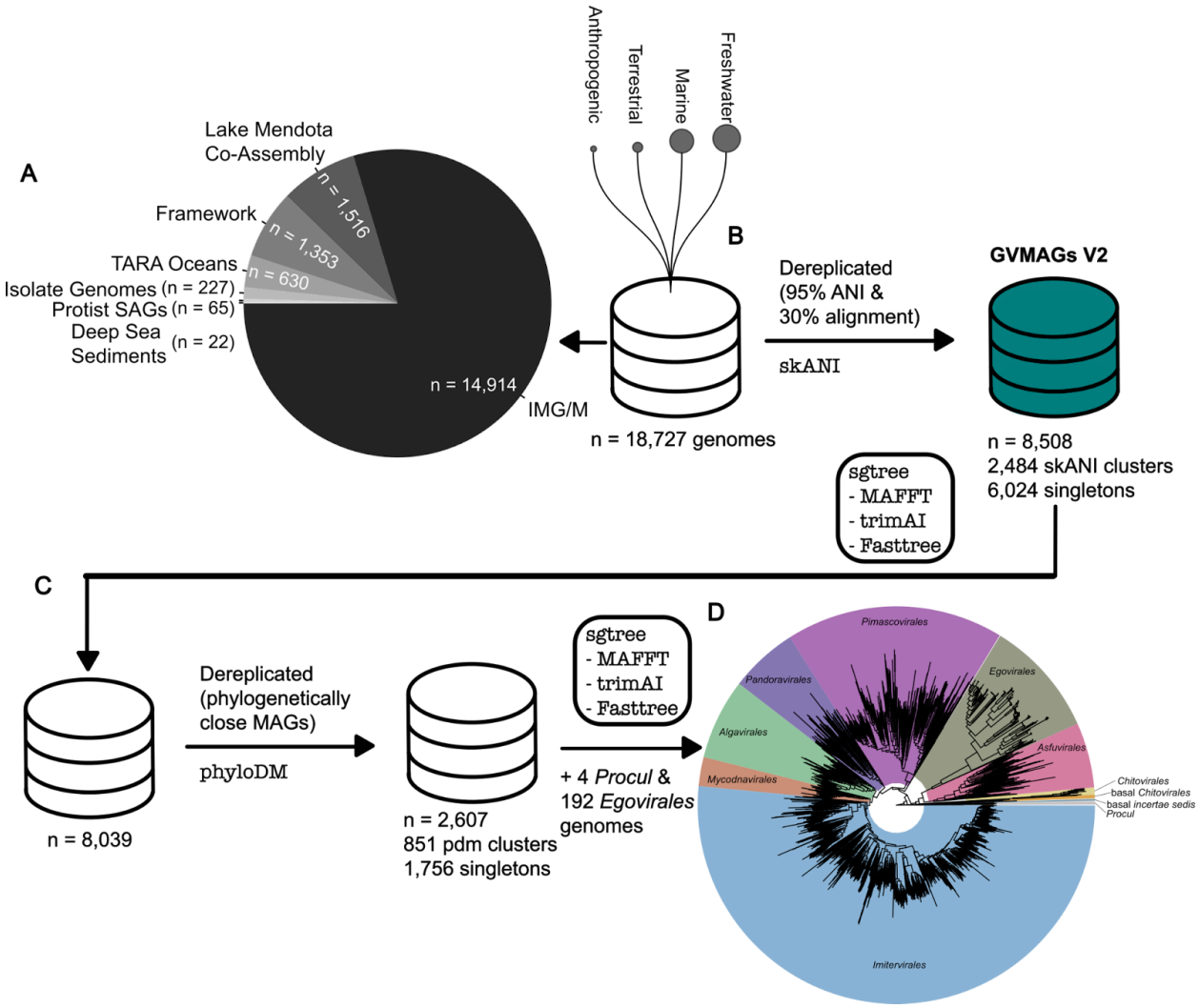

Figure 1: Workflow for genomic dereplication. A total of 18,727 genomes were obtained from public resources, predominantly from the IMG/M database (A). The workflow consisted of three main stages: B) Initial dereplication using skANI (sequence-based average nucleotide identity) with thresholds of 95% ANI and 30% minimum alignment coverage, generating preliminary clusters each represented by a single genome representing the GVMAGs V2 database; C) Construction of an initial phylogenetic tree, followed by secondary dereplication based on phylogenetic distance matrix (PDM) analysis of closely related MAGs (metagenome-assembled genomes), resulting in 2,607 representative genomes and; D) Incorporation of an additional 196 genomes in the final phylogenetic tree.

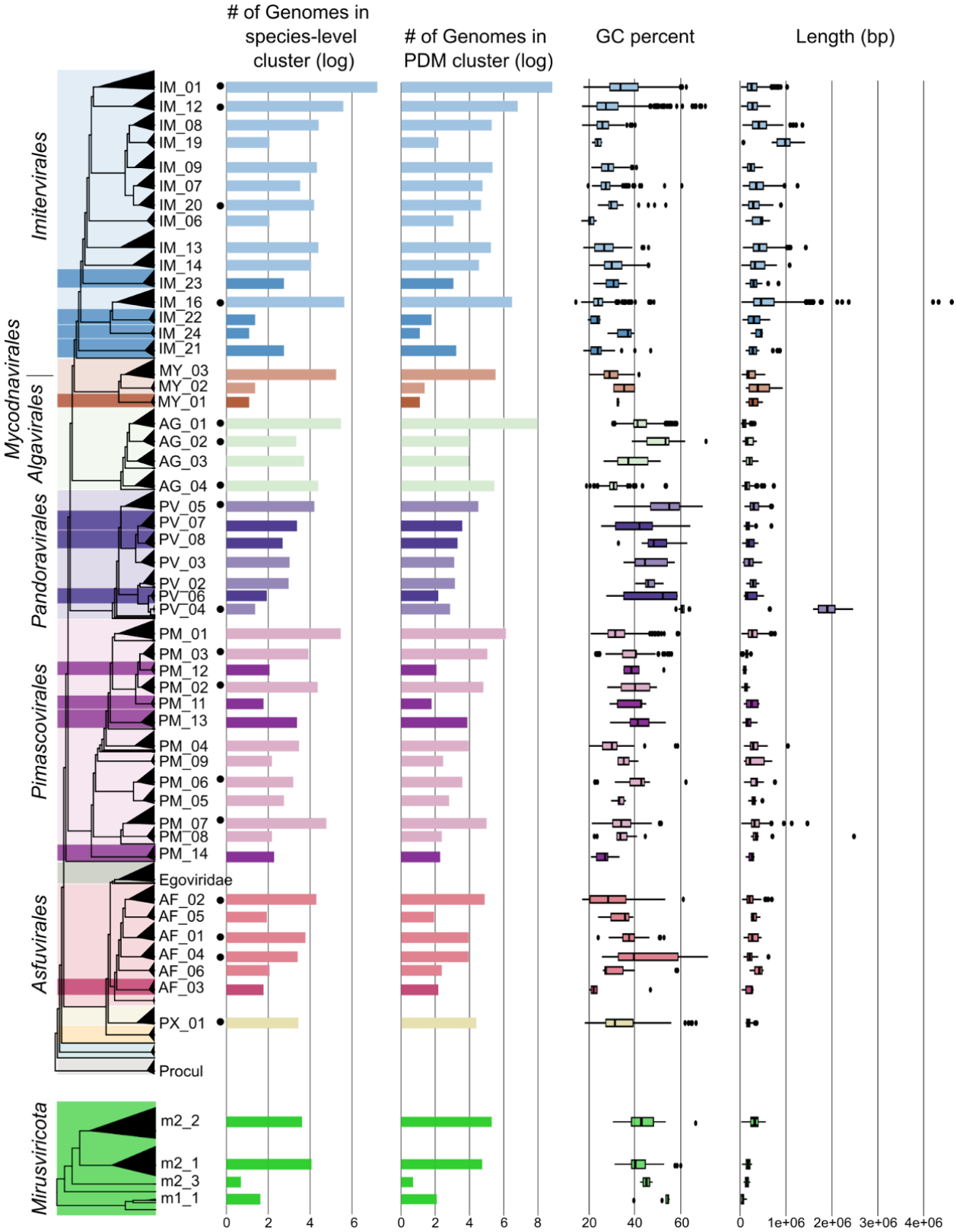

Figure 2: Expanded phylogenetic diversity of the *Nucleocytoviricota* and *Mirusviricota* phylums. The phylogenetic tree displays both established and newly proposed viral families

within *Nucleocytoviricota*, with newly proposed viral families highlighted in darker shades within their respective *Nucleocytoviricota* orders. *Mirusviricota,* the newly proposed *Duplodnaviria* phylum, is displayed with their proposed families based on similar RED score analysis as *Nucleocytoviricota.* Families containing at least one isolate (laboratory-cultured representative) are marked with a solid black circle. For some families, four metrics are visualized: (1) logarithmically-transformed counts of genomes in sequence-based (species-level) clusters; (2) logarithmically-transformed counts of genomes in phylogenetic distance matrix (PDM) clusters; (3) genomic GC content percentage and; (3) assembly length in base pairs. Imitervirales family IM_09 separates into three families (IM_09, IM_19, IM_20) and Asfuvirales family AF_01 splits into four families (AF_01, AF_04, AF_05, AF_06). These are not considered new families, but rather a polyphyletic family split into new families following RED score values.

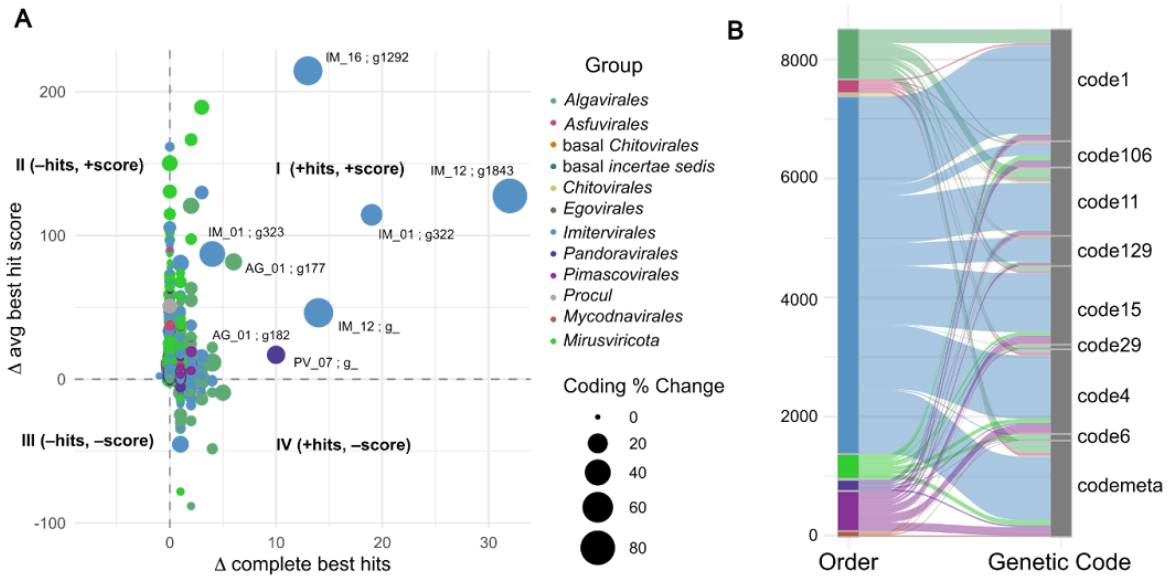

Figure 3. Discovery and distribution of novel genetic codes among *Nucleocytoviricota* orders and *Mirusviricota*. A) Change in complete best hits and the average best hit score between standard gene prediction (prodigal-meta) and a giant virus-specific annotation and classification tool (GVClass). Genomes with labeled family and genus names exhibit an increase of best hits of at least 3 and an increase of average best hit score of at least 15. Sizes denote the coding percent change per genome. Genomes within Quadrant 1 indicates an increase of best hits and average best hit score with GVClass, suggesting this tool is suitable for recovering previously unidentified proteins. B) Taxonomic distribution of predicted genetic codes across different orders, revealing lineage-specific patterns of genetic code usage.

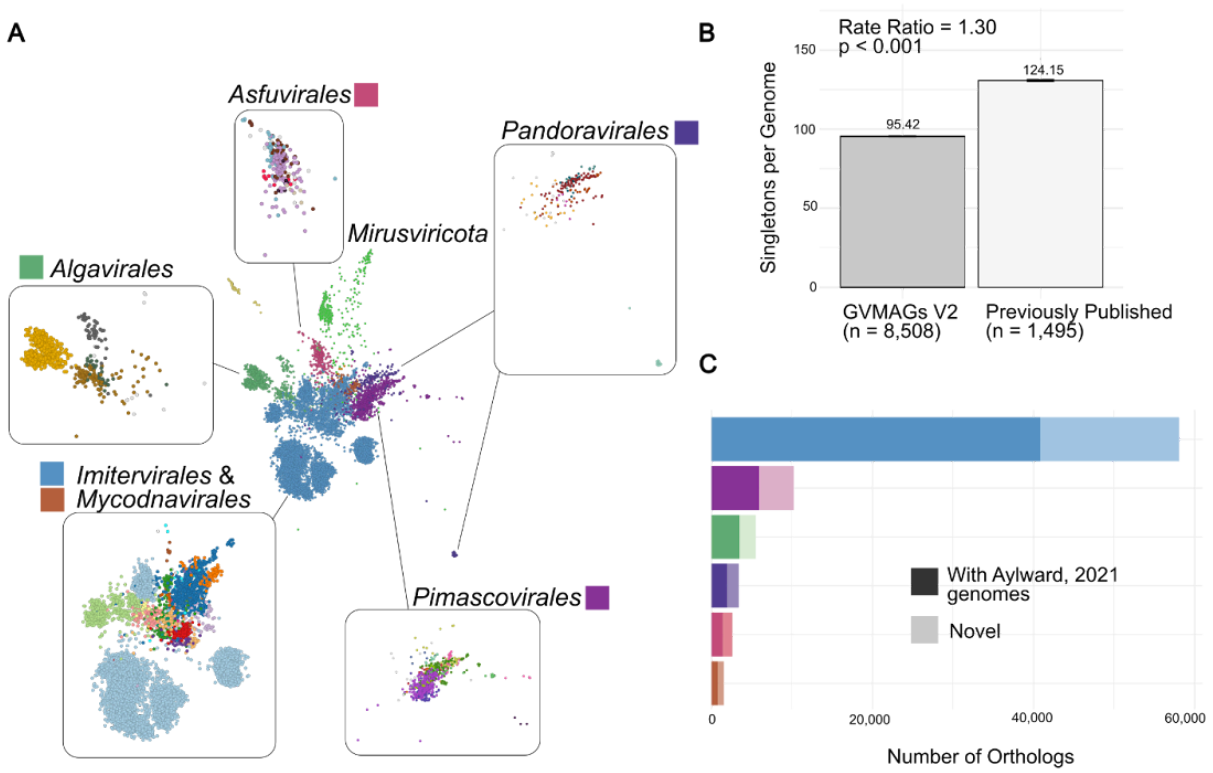

Figure 4: Orthologous gene families across *Nucleocytoviricota* and *Mirusviricota*. A) A bipartite network where genomes are connected through shared orthologous gene families (full network with all edges and ortholog nodes available in Supplemental Images). For clarity, edges and orthologs nodes have been removed. Network clusters colored by viral order. Inset panels: Order-specific subnetworks colored by family, revealing finer-scale patterns of gene sharing within taxonomic groups. B) Rate of singletons per dataset of our new GVMAGs V2 (n = 8,508) and a reduced dataset of only previously published giant virus genomes (n = 1,495). C) Quantification of orthologous gene families per order, with two-tone coloring indicating ortholog conservation, darker portions represent orthologs shared with genomes from the published giant virus framework (Aylward *et al*., 2021), while lighter portions show newly identified orthologs with no members from this framework.

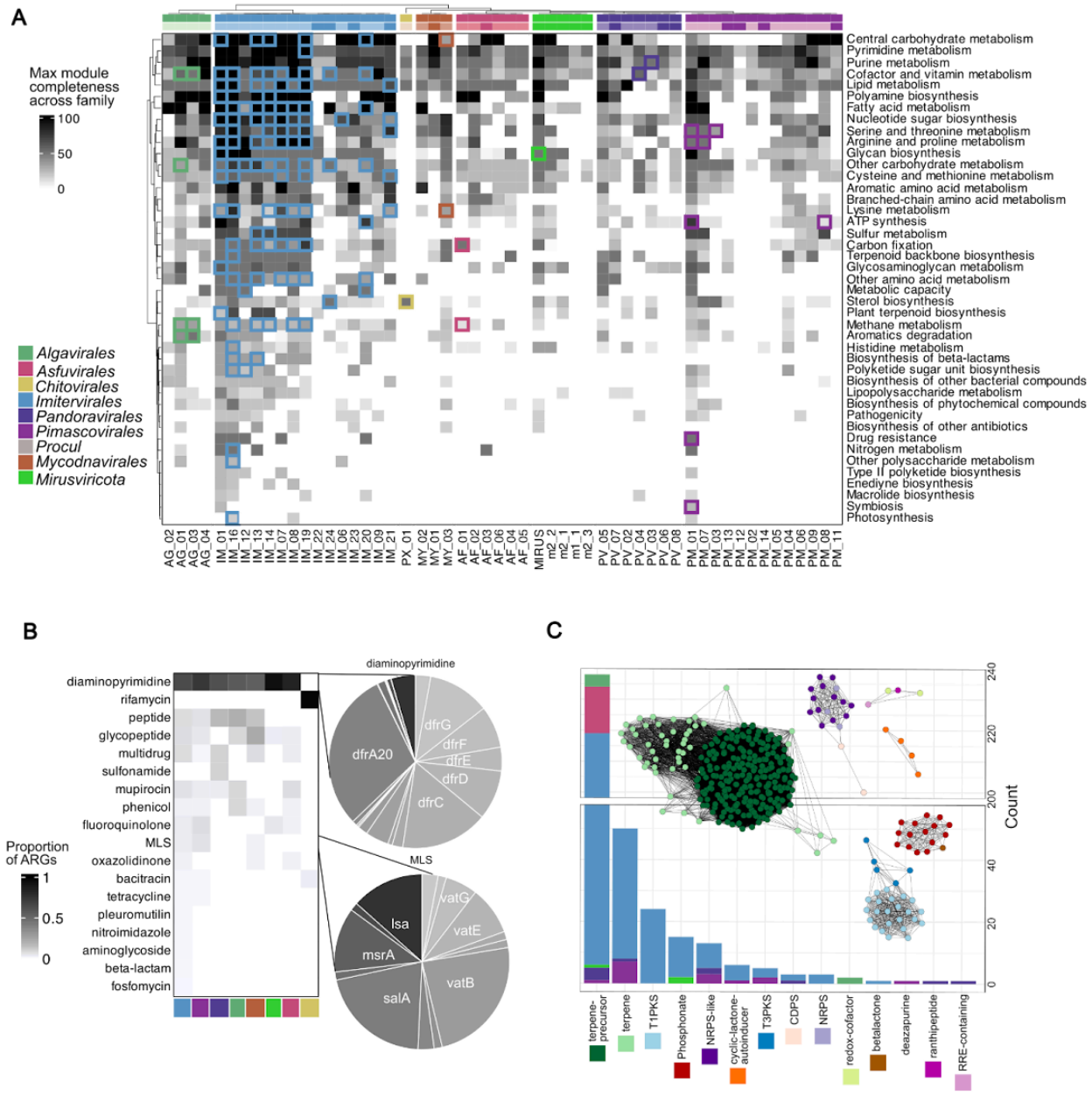

Figure 5. Functional capabilities across the *Nucleocytoviricota* and *Mirusviricota*. A) Overview of max completeness of various metabolic modules across different viral families. Hierarchical clustering was applied to both axes, however, clustering for families was done within order. Darker colors in the orders indicate newly identified families from this paper. Tiles with colored borders indicate gene categories that contain order-specific modules (extended figure can be found in Supplemental Figure S15). B) Antibiotic resistance genes (ARGs) within each *Nucleocytoviricota* order and *Mirusviricota*. Proportion of ARGs correspond to the proportion of each ARG category within each order. Two gene categories were expanded to show the proportion of genes present in the gene category. C) Total count of biosynthetic gene clusters (BGCs) in each BGC category. Bar chart colors correspond to the order colors shown in

(A). Network plots generated using the Cosmos layout in cytoscape are colored by the BGC category where the corresponding color is next to labels.

Abergel, C. *et al.* (2007) “Virus-Encoded Aminoacyl-tRNA Synthetases: Structural and Functional Characterization of Mimivirus TyrRS and MetRS,” *Journal of Virology*, 81(22), pp. 12406–12417. Available at: https://doi.org/10.1128/JVI.01107-07.

Abergel, C., Legendre, M. and Claverie, J.-M. (2015) “The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus,” *FEMS Microbiology Reviews*. Edited by A. Filloux, 39(6), pp. 779–796. Available at: https://doi.org/10.1093/femsre/fuv037.

Alexander, P., Gielen, J. and Sartorelli, A.C. (eds.) (1986) *Bioreduction in the activation of drugs: proceedings of the Second Biochemical Pharmacology Symposium, Oxford, UK, 25-26 July 1985*. First edition. *Biochemical Pharmacology Symposium*, Oxford: Pergamon Press.

Archibald, J. *et al.* (2025) “Persistent mirusvirus infection in the marine protist Aurantiochytrium.” Springer Science and Business Media LLC. Available at: https://doi.org/10.21203/rs.3.rs-5686297/v1.

Aylward, F.O. *et al.* (2021) “A phylogenomic framework for charting the diversity and evolution of giant viruses,” *PLOS Biology*, 19(10), p. e3001430. Available at: https://doi.org/10.1371/journal.pbio.3001430.

Aylward, F.O. *et al.* (2023) “Taxonomic update for giant viruses in the order Imitervirales (phylum Nucleocytoviricota),” *Archives of Virology*, 168(11), p. 283. Available at: https://doi.org/10.1007/s00705-023-05906-3.

Bäckström, D. *et al.* (2019) “Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism,” *mBio*. Edited by R.P. Novick, 10(2), pp. e02497-18. Available at: https://doi.org/10.1128/mBio.02497-18.

Bahir, I. *et al.* (2009) “Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences,” *Molecular Systems Biology*, 5(1), p. 311. Available at: https://doi.org/10.1038/msb.2009.71.

Bastian, M., Heymann, S. and Jacomy, M. (2009) “Gephi: An Open Source Software for Exploring and Manipulating Networks,” *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), pp. 361–362. Available at: https://doi.org/10.1609/icwsm.v3i1.13937.

Bellas, C. *et al.* (2023) “Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses,” *Proceedings of the National Academy of Sciences*, 120(16). Available at: https://doi.org/10.1073/pnas.2300465120.

Benites, L.F., Stephens, T.G. and Bhattacharya, D. (2022) “Multiple waves of viral invasions in Symbiodiniaceae algal genomes,” *Virus Evolution*, 8(2). Available at: https://doi.org/10.1093/ve/veac101.

Benson, D.A. *et al.* (2012) “GenBank,” *Nucleic Acids Research*, 41(D1), pp. D36–D42. Available at: https://doi.org/10.1093/nar/gks1195.

Blin, K. *et al.* (2025) “antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation,” *Nucleic Acids Research*, p. gkaf334. Available at: https://doi.org/10.1093/nar/gkaf334.

Buchfink, B., Reuter, K. and Drost, H.-G. (2021) “Sensitive protein alignments at tree-of-life scale using DIAMOND,” *Nature Methods*, 18(4), pp. 366–368. Available at: https://doi.org/10.1038/s41592-021-01101-x.

Camacho, C. *et al.* (2009) “BLAST+: architecture and applications,” *BMC Bioinformatics*, 10(1), p. 421. Available at: https://doi.org/10.1186/1471-2105-10-421.

Cane, D.E. and Walsh, C.T. (1999) “The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases,” *Chemistry & biology*, 6(12), pp. R319–R325.

Cane, D.E., Walsh, C.T. and Khosla, C. (1998) “Harnessing the biosynthetic code: combinations, permutations, and mutations,” *Science*, 282(5386), pp. 63–68.

Cantalapiedra, C.P. *et al.* (2021) “eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale,” *Molecular Biology and Evolution*. Edited by K. Tamura, 38(12), pp. 5825–5829. Available at: https://doi.org/10.1093/molbev/msab293.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) “trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses,” *Bioinformatics*, 25(15), pp. 1972–1973. Available at: https://doi.org/10.1093/bioinformatics/btp348.

Centers for Disease Control and Prevention (2022) *Outpatient Antibiotic Prescriptions — United States, 2022*.

Chatterjee, A. and Kondabagil, K. (2019) “Giant viral genomic signatures in the previously reported gut metagenomes of pre-school children in rural India,” *Archives of Virology*, 164(11), pp. 2819–2822. Available at: https://doi.org/10.1007/s00705-019-04387-7.

Chen, I.-M.A. *et al.* (2019) “IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes,” *Nucleic Acids Research*, 47(D1), pp. D666–D677. Available at: https://doi.org/10.1093/nar/gky901.

Chen, I.-M.A. *et al.* (2023) “The IMG/M data management and analysis system v.7: content updates and new features,” *Nucleic Acids Research*, 51(D1), pp. D723–D732. Available at: https://doi.org/10.1093/nar/gkac976.

Colson, P. *et al.* (2013) “Codon Usage, Amino Acid Usage, Transfer RNA and Amino-Acyl-tRNA Synthetases in Mimiviruses,” *Intervirology*, 56(6), pp. 364–375. Available at: https://doi.org/10.1159/000354557.

Colson, P. *et al.* (2020) “A protein of the metallo-hydrolase/oxidoreductase superfamily with both beta-lactamase and ribonuclease activity is linked with translation in giant viruses,” *Scientific Reports*, 10(1), p. 21685. Available at: https://doi.org/10.1038/s41598-020-78658-8.

Eddy, S.R. (2011) “Accelerated Profile HMM Searches,” *PLoS Computational Biology*. Edited by W.R. Pearson, 7(10), p. e1002195. Available at: https://doi.org/10.1371/journal.pcbi.1002195.

Elbehery, A.H.A. and Deng, L. (2022) “Insights into the global freshwater virome,” *Frontiers in Microbiology*, 13, p. 953500. Available at: https://doi.org/10.3389/fmicb.2022.953500.

Erazo-Garcia, M.P. *et al.* (2025) “Cryptic infection of a giant virus in a unicellular green alga,” *Science*, 388(6748). Available at: https://doi.org/10.1126/science.ads6303.

Fang, Y. *et al.* (2025) “Genome-resolved year-round dynamics reveal a broad range of giant virus microdiversity,” *mSystems*. Edited by X. Dong, 10(1), pp. e01168-24. Available at: https://doi.org/10.1128/msystems.01168-24.

Ferrández, A., Garciá, J.L. and Díaz, E. (1997) “Genetic characterization and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway of Escherichia coli K-12,” *Journal of Bacteriology*, 179(8), pp. 2573–2581. Available at: https://doi.org/10.1128/jb.179.8.2573-2581.1997.

Gaïa, M. *et al.* (2023) “Mirusviruses link herpesviruses to giant viruses,” *Nature*, 616(7958), pp. 783–789. Available at: https://doi.org/10.1038/s41586-023-05962-4.

Gaïa, M. *et al.* (2024) “Egoviruses: distant relatives of poxviruses abundant in the gut microbiomes of humans and animals worldwide.” Available at: https://doi.org/10.1101/2024.03.23.586382.

Grigoriev, I.V. *et al.* (2014) “MycoCosm portal: gearing up for 1000 fungal genomes,” *Nucleic Acids Research*, 42(D1), pp. D699–D704. Available at: https://doi.org/10.1093/nar/gkt1183.

Grigoriev, I.V. *et al.* (2021) “PhycoCosm, a comparative algal genomics resource,” *Nucleic Acids Research*, 49(D1), pp. D1004–D1011. Available at: https://doi.org/10.1093/nar/gkaa898.

Ha, A.D., Moniruzzaman, M. and Aylward, F.O. (2021) “High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System,” *mSystems*. Edited by S. Bordenstein, 6(4). Available at: https://doi.org/10.1128/msystems.00293-21.

Hoffman, D.C. *et al.* (1995) “Macronuclear gene-sized molecules of hypotrichs,” *Nucleic Acids Research*, 23(8), pp. 1279–1283. Available at: https://doi.org/10.1093/nar/23.8.1279.

Hutinet, G., Swarjo, M.A. and de Crécy-Lagard, V. (2017) “Deazaguanine derivatives, examples of crosstalk between RNA and DNA modification pathways,” *RNA biology*, 14(9), pp. 1175–1184.

Hyatt, D. *et al.* (2010) “Prodigal: prokaryotic gene recognition and translation initiation site identification,” *BMC Bioinformatics*, 11(1), p. 119. Available at: https://doi.org/10.1186/1471-2105-11-119.

Jacomy, M. *et al.* (2014) “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software,” *PLoS ONE*. Edited by M.R. Muldoon, 9(6), p. e98679. Available at: https://doi.org/10.1371/journal.pone.0098679.

Jung, Y. *et al.* (2023) “Function and Structure of a Terpene Synthase Encoded in a Giant Virus Genome,” *Journal of the American Chemical Society*, 145(48), pp. 25966–25970. Available at: https://doi.org/10.1021/jacs.3c10603.

Kanehisa, M. and Goto, S. (2000) “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, 28(1), pp. 27–30.

Katoh, K. *et al.* (2002) “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform,” *Nucleic Acids Research*, 30(14), pp. 3059–3066.

Kautsar, S.A. *et al.* (2021) “BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters,” *GigaScience*, 10(1). Available at: https://doi.org/10.1093/gigascience/giaa154.

Keeling, P.J. and Doolittle, W.F. (1996) “A non-canonical genetic code in an early diverging eukaryotic lineage.,” *The EMBO Journal*, 15(9), pp. 2285–2290. Available at: https://doi.org/10.1002/j.1460-2075.1996.tb00581.x.

Klemm, P., Stadler, P.F. and Lechner, M. (2023) “Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs,” *Frontiers in Bioinformatics*, 3, p. 1322477. Available at: https://doi.org/10.3389/fbinf.2023.1322477.

Legendre, M. *et al.* (2018) “Diversity and evolution of the emerging Pandoraviridae family,” *Nature Communications*, 9(1), p. 2285. Available at: https://doi.org/10.1038/s41467-018-04698-4.

Li, J. *et al.* (2024) “Discovering the potential role of giant viruses in antibiotic resistance spread.” In Review. Available at: https://doi.org/10.21203/rs.3.rs-3747448/v1.

Louca, S. and Doebeli, M. (2018) “Efficient comparative phylogenetics on large trees,” *Bioinformatics*. Edited by A. Valencia, 34(6), pp. 1053–1055. Available at: https://doi.org/10.1093/bioinformatics/btx701.

Machado, T.B. *et al.* (2023) “Gene duplication as a major force driving the genome expansion in some giant viruses,” *Journal of Virology*, 97(12), pp. e01309-23. Available at: https://doi.org/10.1128/jvi.01309-23.

Manni, M. *et al.* (2021) “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes,” *Molecular Biology and Evolution*. Edited by J. Kelley, 38(10), pp. 4647–4654. Available at: https://doi.org/10.1093/molbev/msab199.

McGarvey, D.J. and Croteau, R. (1995) “Terpenoid metabolism,” *The plant cell*, 7(7), p. 1015.

Mckeown, D. *et al.* (2025) “Multiple, diverse endogenous giant virus elements within the genome of a brown alga,” *Virus Evolution*, 11(1). Available at: https://doi.org/10.1093/ve/veaf009.

Mihara, T. *et al.* (2018) “Taxon Richness of ‘Megaviridae’ Exceeds those of Bacteria and Archaea in the Ocean,” *Microbes and Environments*, 33(2), pp. 162–171. Available at: https://doi.org/10.1264/jsme2.ME17203.

Moniruzzaman, M., Martinez-Gutierrez, C.A., *et al.* (2020) “Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses,” *Nature Communications*, 11(1), p. 1710. Available at: https://doi.org/10.1038/s41467-020-15507-2.

Moniruzzaman, M., Weinheimer, A.R., *et al.* (2020) “Widespread endogenization of giant viruses shapes genomes of green algae,” *Nature*, 588(7836), pp. 141–145. Available at:

https://doi.org/10.1038/s41586-020-2924-2.

Myers, J.M. *et al.* (2024) “Large DNA viruses in early diverging fungal genomes are relics of past and present infections.” Available at: https://doi.org/10.1101/2024.01.04.574182.

O’Hagan, D. (1991) “The polyketide metabolites,” *(No Title)* [Preprint].

Oliver, T. *et al.* (2024) “Coassembly and binning of a twenty-year metagenomic time-series from Lake Mendota,” *Scientific Data*, 11(1), p. 966. Available at: https://doi.org/10.1038/s41597-024-03826-8.

Ono, K. *et al.* (2025) “Cytoscape Web: bringing network biology to the browser,” *Nucleic Acids Research*, 53(W1), pp. W203–W212. Available at: https://doi.org/10.1093/nar/gkaf365.

Park, C. *et al.* (2025) “Discovery of a New Type of Terpene Synthase Coded by an Orphan Gene in a Giant Virus,” *Biochemistry*, p. acs.biochem.5c00354. Available at: https://doi.org/10.1021/acs.biochem.5c00354.

Parks, D.H. *et al.* (2018) “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life,” *Nature Biotechnology*, 36(10), pp. 996–1004. Available at: https://doi.org/10.1038/nbt.4229.

Perini, L. *et al.* (2024) “Giant viral signatures on the Greenland ice sheet,” *Microbiome*, 12(1), p. 91. Available at: https://doi.org/10.1186/s40168-024-01796-y.

Philippe, N. *et al.* (2013) “Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes,” *Science*, 341(6143), pp. 281–286. Available at: https://doi.org/10.1126/science.1239181.

Pitot, T.M., Brůna, T. and Schulz, F. (2024) “Conservative taxonomy and quality assessment of giant virus genomes with GVClass,” *npj Viruses*, 2(1), p. 60. Available at: https://doi.org/10.1038/s44298-024-00069-7.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments,” *PLoS ONE*. Edited by A.F.Y. Poon, 5(3), p. e9490. Available at: https://doi.org/10.1371/journal.pone.0009490.

Pritham, E.J., Putliwala, T. and Feschotte, C. (2007) “Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses,” *Gene*, 390(1–2), pp. 3–17. Available at: https://doi.org/10.1016/j.gene.2006.08.008.

Raoult, D. *et al.* (2004) “The 1.2-Megabase Genome Sequence of Mimivirus,” *Science*, 306(5700), pp. 1344–1350. Available at: https://doi.org/10.1126/science.1101485.

Rigou, S. *et al.* (2022) “Past and present giant viruses diversity explored through permafrost metagenomics,” *Nature Communications*, 13(1), p. 5853. Available at: https://doi.org/10.1038/s41467-022-33633-x.

Rohwer, R.R. *et al.* (2023) “Species invasions shift microbial phenology in a two-decade freshwater time series,” *Proceedings of the National Academy of Sciences*, 120(11), p. e2211796120. Available at: https://doi.org/10.1073/pnas.2211796120.

Rohwer, R.R. *et al.* (2025) “Two decades of bacterial ecology and evolution in a freshwater lake,” *Nature Microbiology* [Preprint]. Available at: https://doi.org/10.1038/s41564-024-01888-3.

Roux, S. *et al.* (2019) “Minimum Information about an Uncultivated Virus Genome (MIUViG),” *Nature Biotechnology*, 37(1), pp. 29–37. Available at: https://doi.org/10.1038/nbt.4306.

Ru, J. *et al.* (2023) “Unveiling the hidden role of aquatic viruses in hydrocarbon pollution bioremediation,” *Journal of Hazardous Materials*, 459, p. 132299. Available at: https://doi.org/10.1016/j.jhazmat.2023.132299.

Ruiz-Perez, C.A., Conrad, R.E. and Konstantinidis, K.T. (2021) “MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes,” *BMC Bioinformatics*, 22(1), p. 11. Available at: https://doi.org/10.1186/s12859-020-03940-5.

Saary, P., Mitchell, A.L. and Finn, R.D. (2020) “Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC,” *Genome Biology*, 21(1). Available at: https://doi.org/10.1186/s13059-020-02155-4.

Schneider, S.U., Leible, M.B. and Yang, X.-P. (1989) “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,” *Molecular and General Genetics MGG*, 218(3), pp. 445–452. Available at: https://doi.org/10.1007/BF00332408.

Schulz, F. *et al.* (2017) “Towards a balanced view of the bacterial tree of life,” *Microbiome*, 5(1), p. 140. Available at: https://doi.org/10.1186/s40168-017-0360-9.

Schulz, F. *et al.* (2020) “Giant virus diversity and host interactions through global metagenomics,” *Nature*, 578(7795), pp. 432–436. Available at: https://doi.org/10.1038/s41586-020-1957-x.

Schulz, F. *et al.* (2024) “Protists as mediators of complex microbial and viral associations.” Available at: https://doi.org/10.1101/2024.12.29.630703.

Schulz, F., Abergel, C. and Woyke, T. (2022) “Giant virus biology and diversity in the era of genome-resolved metagenomics,” *Nature Reviews Microbiology*, 20(12), pp. 721–736. Available at: https://doi.org/10.1038/s41579-022-00754-5.

Shackelton, L.A. and Holmes, E.C. (2008) “The role of alternative genetic codes in viral evolution and emergence,” *Journal of Theoretical Biology*, 254(1), pp. 128–134. Available at: https://doi.org/10.1016/j.jtbi.2008.05.024.

Shaw, J. and Yu, Y.W. (2023) “Fast and robust metagenomic sequence comparison through sparse chaining with skani,” *Nature Methods*, 20(11), pp. 1661–1665. Available at: https://doi.org/10.1038/s41592-023-02018-3.

Sun, T.-W. *et al.* (2020) “Host Range and Coding Potential of Eukaryotic Giant Viruses,” *Viruses*, 12(11), p. 1337. Available at: https://doi.org/10.3390/v12111337.

Sun, T.-W. and Ku, C. (2021) “Unraveling gene content variation across eukaryotic giant viruses based on network analyses and host associations,” *Virus Evolution*, 7(2), p. veab081. Available at: https://doi.org/10.1093/ve/veab081.

Tully, B.J. *et al.* (2017) “290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology,” *PeerJ*, 5, p. e3558. Available at: https://doi.org/10.7717/peerj.3558.

Van Etten, J.L. (2003) “Unusual Life Style of Giant Chlorella Viruses,” *Annual Review of Genetics*, 37(1), pp. 153–195. Available at: https://doi.org/10.1146/annurev.genet.37.110801.143915.

Villada, J.C. *et al.* (2025) “A genomic catalog of Earth’s bacterial and archaeal symbionts.” Genomics. Available at: https://doi.org/10.1101/2025.05.29.656868.

Weynberg, K., Allen, M. and Wilson, W. (2017) “Marine Prasinoviruses and Their Tiny Plankton Hosts: A Review,” *Viruses*, 9(3), p. 43. Available at: https://doi.org/10.3390/v9030043.

Willemsen, A., Manzano-Marín, A. and Horn, M. (2025) “Novel High-Quality Amoeba Genomes Reveal Widespread Codon Usage Mismatch Between Giant Viruses and Their Hosts,” *Genome Biology and Evolution*. Edited by L. Santoferrara, 17(1), p. evae271. Available at: https://doi.org/10.1093/gbe/evae271.

Yi, X. *et al.* (2024) “Giant viruses as reservoirs of antibiotic resistance genes,” *Nature Communications*, 15(1), p. 7536. Available at: https://doi.org/10.1038/s41467-024-51936-z.

Yimenu, D.K. *et al.* (2019) “Assessment of Antibiotic Prescribing Patterns at Outpatient Pharmacy Using World Health Organization Prescribing Indicators,” *Journal of Primary Care & Community Health*, 10, p. 2150132719886942. Available at: https://doi.org/10.1177/2150132719886942.

Yutin, N. *et al.* (2013) “Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family,” *Virology Journal*, 10(1), p. 106. Available at: https://doi.org/10.1186/1743-422X-10-106.

Zdouc, M.M. *et al.* (2025) “MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration,” *Nucleic Acids Research*, 53(D1), pp. D678–D690. Available at: https://doi.org/10.1093/nar/gkae1115.

Zhao, H. *et al.* (2024) “Eukaryotic genomic data uncover an extensive host range of mirusviruses.” Cold Spring Harbor Laboratory. Available at: https://doi.org/10.1101/2024.01.18.576163.

Zhu, S. *et al.* (2024) “Symbiotic virus-bacteria interactions in biological treatment of coking wastewater manipulating bacterial physiological activities,” *Water Research*, 257, p. 121741. Available at: https://doi.org/10.1016/j.watres.2024.121741.