Nucleic Acids Research, 2026, **54**, D902–D911
https://doi.org/10.1093/nar/gkaf1225
Advance access publication date: 29 November 2025
Database issue

OXFORD

VIRE: a metagenome-derived, planetary-scale virome resource with environmental context

Suguru Nishijima[1,2,*,†], **Anthony Fullam**[1,†], **Thomas S.B. Schmidt**[3,†], **Michael Kuhn**[1], **Peer Bork**[1,4,*]

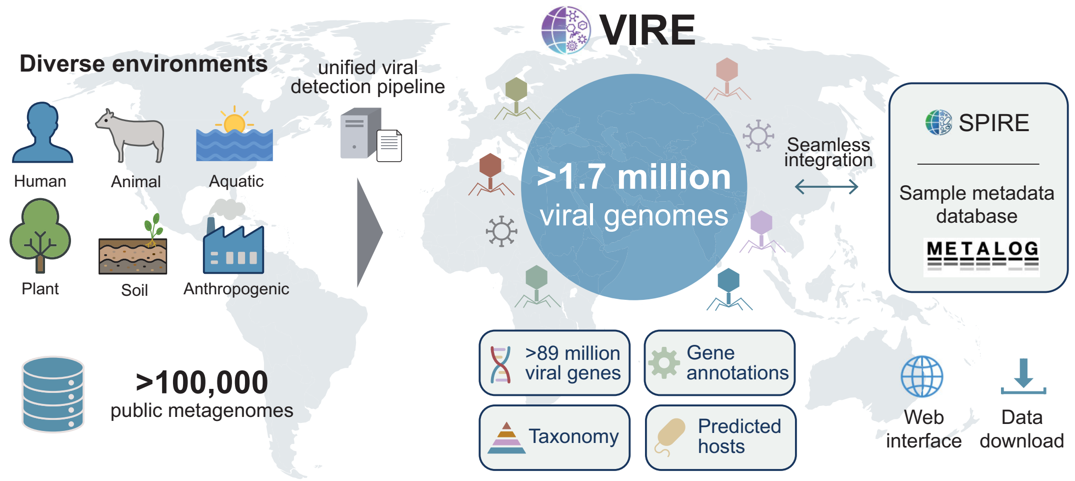
[1]Molecular Systems Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[2]Life Science Data Research Center, The University of Tokyo, 277-0882 Chiba, Japan
[3]APC Microbiome and School of Medicine, University College Cork, Cork, Ireland
[4]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany,

*To whom correspondence should be addressed. Email: peer.bork@embl.org
Correspondence may also be addressed to Suguru Nishijima. Email: nishijima.suguru@gmail.com
†The first three authors should be regarded as Joint First Authors.

Abstract

Viruses are the most abundant biological entities on Earth, yet their global diversity remains largely unexplored. Here, we present VIRE, a comprehensive resource comprising over 1.7 million high- and medium-quality viral genomes recovered from >100 000 publicly available metagenomes derived from samples that cover diverse ecosystems, including host-associated, aquatic, terrestrial, and anthropogenic environments. Using a unified and scalable pipeline, we systematically assembled viral genomes and provided detailed information on genome completeness, taxonomic classification, predicted lifestyle, and host assignment based on CRISPR spacer matches. VIRE contains >89 million predicted viral open reading frames, as well as detailed functional annotations derived from multiple databases. Importantly, VIRE is seamlessly integrated with related microbiome resources such as SPIRE (https://spire.embl.de) and Metalog (https://metalog.embl.de), enabling users to jointly explore viral genomes, metagenome-assembled genomes, and associated environmental or clinical metadata. Accessible at https://vire.embl.de, VIRE provides an open-access, scalable platform for investigating viral diversity, evolution, and ecology on a planetary scale.

Graphical abstract



Introduction

Viruses are estimated to number around 10^{31} particles on Earth, making them the most abundant biological entities on the planet [1, 2]. Among them, bacteriophages, viruses that infect bacteria, are now recognized as key players in microbial ecosystems. Phages shape microbial community structures [3, 4], facilitate horizontal gene transfer between bacteria [5, 6], and drive biogeochemical cycles on Earth [7–9]. Despite their ubiquity and ecological importance, our understanding of viral diversity has remained limited, largely due to the constraints of cultivation-based techniques. The advent of high-throughput sequencing technologies, particularly shotgun metagenomics, has revolutionized our ability to explore viral diversity directly from environmental samples [10–12]. Over the past decade, metagenomic analyses and improved bioinformatic pipelines have uncovered an enormous

Received: August 15, 2025. **Revised:** October 17, 2025. **Accepted:** October 17, 2025

diversity of previously unknown viral genomes from a wide range of environments, including the human gut [13–18], the ocean [19–24], and soil [25–28]. Yet these newly discovered genomes likely represent only the tip of the iceberg among a vast, largely unexplored viral "dark matter" across Earth's ecosystems. Understanding the genetic and ecological diversity of such environmental viruses is crucial for understanding viral function, evolution, and host-virus dynamics [29–31]. Moreover, characterizing viral reservoirs in natural environments contributes to pandemic preparedness by providing baseline data for identifying emerging zoonotic threats [32, 33]. To catalog the diversity of uncultivated viruses, previous studies have developed specialized viral genome databases from metagenomic datasets. However, most existing resources are environment-specific (e.g. human gut [13–15, 34], marine [21, 23, 35], or soil [25, 26, 36]) with only a few exceptions [37, 38].

Here, we present VIRE (Viral Integrated Resource across Ecosystems), a global-scale resource of viral genomes assembled from over 100 000 publicly available metagenomic samples spanning diverse environments. VIRE contains >1.7 million medium- to high-quality viral genomes reconstructed through a unified bioinformatics pipeline, making it the largest viral genome database to date. Each genome is accompanied by a rich set of metadata, including taxonomic classification, predicted host organisms, predicted lifestyle (lytic or temperate), and gene annotations derived from multiple functional databases. Importantly, VIRE is seamlessly linked with complementary resources such as SPIRE (https://spire.embl.de) [39] and Metalog (https://metalog.embl.de) [40], allowing users to access associated metagenome-assembled genomes (MAGs) and manually curated metadata of metagenomes, respectively. VIRE provides a comprehensive and scalable platform for exploring global viral diversity, serving as a valuable resource for virology, microbiome research, and microbial ecology.

Materials and methods

Identification of viral sequences from metagenomes

The core dataset of VIRE was constructed from a total of 101 623 metagenomic samples derived from 732 independent studies. The majority of these datasets were originally used in the SPIRE resource [39] and consist of publicly available shotgun metagenomes downloaded primarily from the European Nucleotide Archive (ENA) [41] or the Sequence Read Archive (SRA) [42], covering a wide range of environmental samples. The datasets were collected through a semi-automated process and manually curated to exclude certain data types, such as those from artificial experimental systems (e.g. *in vitro* mock communities, laboratory mice, or pathogen challenge studies), as well as amplicon-based and isolate-derived sequences. Moreover, additional virome samples specifically enriched for virus-like particles (VLPs, excluded from SPIRE) were incorporated into the dataset. Each metagenomic sample was annotated with a standardized environmental ontology called microntology, which assigns at least one of 92 terms describing the habitat of the associated microbial community [39].

Metagenomic reads were assembled using MEGAHIT v1.2.9 [43], generating contigs ($n = 24\,883\,275\,724$). For all samples except newly added ones, we used assemblies that had already been generated during the construction of the SPIRE database [39], while newly added virome samples were assembled *de novo* in this study. Contigs longer than 5 kb from bulk metagenomes and those longer than 2 kb from virome metagenomes ($n = 346\,532\,161$) were subjected to viral detection using geNomad v1.5.2 [44] and CheckV v1.0.1 (database version 1.4) [45]. Contigs with a viral score of ≥0.7 by geNomad and classified as at least medium-quality by CheckV, defined as completeness ≥50% and contamination <10% [46], were considered putative viral genomes ($n = 1\,778\,826$). Viral sequences with a CheckV kmer_freq score ≥2 (indicative of possible concatemeric repeats) were excluded ($n = 635$). To further improve specificity, Barrnap (https://github.com/tseemann/barrnap) was used to screen for bacterial ribosomal RNA genes (5S, 16S, and 23S ribosomal RNAs), which are rarely found in viral genomes, and contigs encoding any of these genes were removed ($n = 5049$). All data processing steps were implemented in a Nextflow pipeline [47], ensuring reproducibility and scalability.

Collection of viral genomes from GenBank and RefSeq

To obtain a set of high-confidence viral genomes with reliable taxonomic classification, we downloaded viral genomes labeled as "complete genome" and taxonomically annotated in the International Committee on Taxonomy of Viruses (ICTV) [48] Release 40 from GenBank (accessed in July 2025; $n = 12\,395$) [49]. For segmented viruses, such as influenza viruses, individual genome segments were concatenated into a single sequence using a string of ten "N" nucleotides as separators. In addition, we retrieved viral genomes that were not included in the above but were registered as viral genomes in RefSeq (accessed in July 2025; $n = 8052$) [50]. These genomes were processed in the same manner as metagenome-derived viral genomes as described below.

Clustering viral sequences into species- and genus-level groups

All viral genomes were clustered into species- and genus-level groups using vclust v1.2.2-b687638 [51]. Clustering was performed at 95% and 70% average nucleotide identity (ANI) and 85% alignment fraction (AF) with the Leiden algorithm for species-level and genus-level clusters, respectively, following current guidelines proposed by the ICTV [48].

Rarefaction curves for each environment were generated by progressively subsampling increasing proportions of the full viral genome dataset (10%, 20%, …, 100%). For each sampling fraction, genomes were randomly sampled without replacement 10 times, and the resulting numbers of species-level (>95% ANI) and genus-level (>70% ANI) clusters were calculated. The mean values across the 10 iterations were then plotted to produce the curves.

The species discovery coefficient (α) was calculated for each environment following the approach described previously [52]. In brief, we first determined the number of newly discovered species from the rarefaction analysis for successive increments of sampling effort. We then fitted a log–log linear regression model relating the number of newly discovered species to the cumulative number of species observed, and calculated α as the regression slope plus one.

To evaluate the novelty of genomes in VIRE, we compared viral genomes in VIRE with those from IMG/VR v4 (2022-12-

19_7.1) [37]. Because IMG/VR contains low-quality genomes, only those annotated as medium-quality, high-quality, or complete ($n = 1\,059\,662$) were included for comparison. Clustering was performed using vclust with the Leiden algorithm under the thresholds of ANI > 95% and AF > 85%.

Host, gene, and lifestyle annotations

To infer bacteriophage host, we employed a CRISPR spacer–based method designed to minimize false positives [53]. We extracted CRISPR spacers from ~1.2 million MAGs from the SPIRE resource ($n = 9\,510\,889$) [39] and ~1.0 million isolate genomes from the proGenomes v3 database ($n = 18\,937\,140$) [54] using minced (https://github.com/ctSkennerton/minced). To reduce misbinning-derived contamination, additional filtering was applied to spacers derived from MAGs: for each contig containing a CRISPR locus, genes were predicted and aligned using DIAMOND v0.9.19.120 [55] against the reference gene set of the representative species in SPIRE. If fewer than 50% of genes matched any other MAG of the same genus (excluding self), the spacer was discarded. Spacers derived from contigs shorter than 10 kb were also excluded. The resulting filtered spacers from SPIRE ($n = 5\,702\,293$) and proGenomes ($n = 18\,937\,140$) were then aligned to the viral genome sequences using BLASTN v2.5.0 [56], allowing only perfect matches or alignments with a single mismatch or indel under a >95% AF. When a CRISPR spacer matched a viral genome under these criteria, the host taxonomy was assigned according to the GTDB-Tk classification v2.4.0 [57] based on release 220 of GTDB [58].

Protein-coding genes were predicted from the identified viral genomes using prodigal-gv v2.11.0 [44, 59], an algorithm optimized for viral gene calling. Functional annotations were then assigned using eggNOG-mapper v2.1.13 [60], MetaCerberus v1.4.0 [61], and RGI v5.2.1 [62]. These tools provided annotation across multiple databases and functional categories, including eggNOG orthology [63], KEGG Orthology [64], COG [65], PHROG [66], pVOG [67], Pfam [68], TIGRFAM [69], dbCAN [70], and antibiotic resistance genes [62]. To identify auxiliary metabolic genes (AMGs), we used the previously curated set of KEGG orthology terms [71] and calculated the proportion of genes assigned to AMGs relative to the total number of genes in each environment. These proportions were then summarized by functional category for metabolism according to the KEGG database.

The lifestyle (lytic or temperate) of each phage genome was predicted using BACPHLIP v0.9.3 [72], and those with a score of >0.8 were treated as temperate phages. Information on the genetic code of each viral genome was obtained from geNomad and used in downstream analyses.

Results

Overview of the viral genomes in VIRE

The VIRE database primarily consists of 1 784 510 medium- or high-quality viral genomes (defined as at least >50% completeness and <10% contamination) reconstructed from a total of 101 623 publicly available bulk and virome (VLP-enriched) metagenomic datasets (Supplementary Fig. S1). Quality assessments by CheckV [22] classified these as 384 035 complete, 417 105 high-quality, and 983 370 medium-quality genomes (Fig. 1A). In addition to these metagenome-derived sequences, VIRE includes 12 916 viral genomes downloaded from RefSeq/GenBank [49, 50] (Fig. 1A). Taxonomic classification with geNomad [44] revealed that the majority of sequences (87.2%) belong to *Duplodnaviria*, a realm that encompasses tailed double-stranded DNA bacteriophages (Fig. 1B). This is followed by *Monodnaviria* (9.2%), comprising single-stranded DNA (ssDNA) viruses; unclassified viruses (2.6%); and *Varidnaviria* (0.5%), which includes giant viruses. At the order level, *Petitvirales*, *Tubulavirales*, and *Sanitavirales* (*Monodnaviria*) and *Crassvirales* and *Autographivirales* (*Duplodnaviria*) were the most abundant (Supplementary Fig. S2). Environmental annotations using the microntology [39] indicated that the majority of metagenome-derived viral genomes (n = 1410837, 78.5%) originate from host-associated environments, most of which were human gut samples (n = 950399, 52.9%) (Fig. 1D). These are followed by viruses derived from aquatic (n = 336505, 18.7%), terrestrial (n = 115044, 6.4%), and anthropogenic environments (n = 56644, 3.2%). The average genome size of these viruses was 38.2 kb (Fig. 1C), and the largest metagenome-derived genome identified was an 836.2 kb phage genome from a bovine sample, classified within *Duplodnaviria*. This genome is among the largest phage genomes reported to date, comparable in size to previously described megaphages (e.g. 841 and 852 kb genomes) [73, 74].

All viral genomes in VIRE were clustered into species- and genus-level groups, operationally defined as genome clusters sharing 95% and 70% ANI, following current ICTV recommendations [48]. This resulted in 706 281 nonredundant species-level and 527 020 genus-level representative sequences. The largest species-level cluster corresponded to phiX174, a bacteriophage genome commonly used as a spike-in for Illumina sequencing quality control. This cluster was detected across diverse environments, including host-associated, aquatic, terrestrial, and anthropogenic samples, suggesting that this control DNA sequence is often incompletely removed from metagenomic datasets before deposition into public archives [75].

Rarefaction analysis revealed that the number of species- and genus-level clusters continued to increase with additional viral genomes across all environments (Fig. 1E and Supplementary Fig. S3), consistent with previous studies [37]. To further quantify this, we calculated species discovery coefficients from the exponent of power laws fitted to rarefaction curves, as described previously [52]. Species discovery coefficients typically range from 0 to 1, where values closer to 1 indicate that rarefaction curves are far from saturation and additional sampling will continue to reveal new species, while values near 0 indicate that diversity has already been largely captured and the discovery of novel lineages is slowing down. In our analyses, high coefficient values (>0.6) were observed for most environments (Fig. 1D). In particular, hydrothermal vent, tundra, and human airways samples all showed coefficients of ≥0.8, indicating that additional sampling effort is expected to discover new lineages at nearly unmitigated rates in these habitats. Other environments, such as agriculture, rhizosphere, rumen, plant host, hot spring, and air, also displayed high coefficients. The lowest coefficient was observed for the human gut and skin, likely reflecting the relatively high sampling effort and relatively low alpha diversity of their microbial community [76], respectively. Nevertheless, even in these environments, the coefficients remained above 0.5, suggesting that the rarefaction curves are still far from saturation. These results indicate that viral diversity across many environments

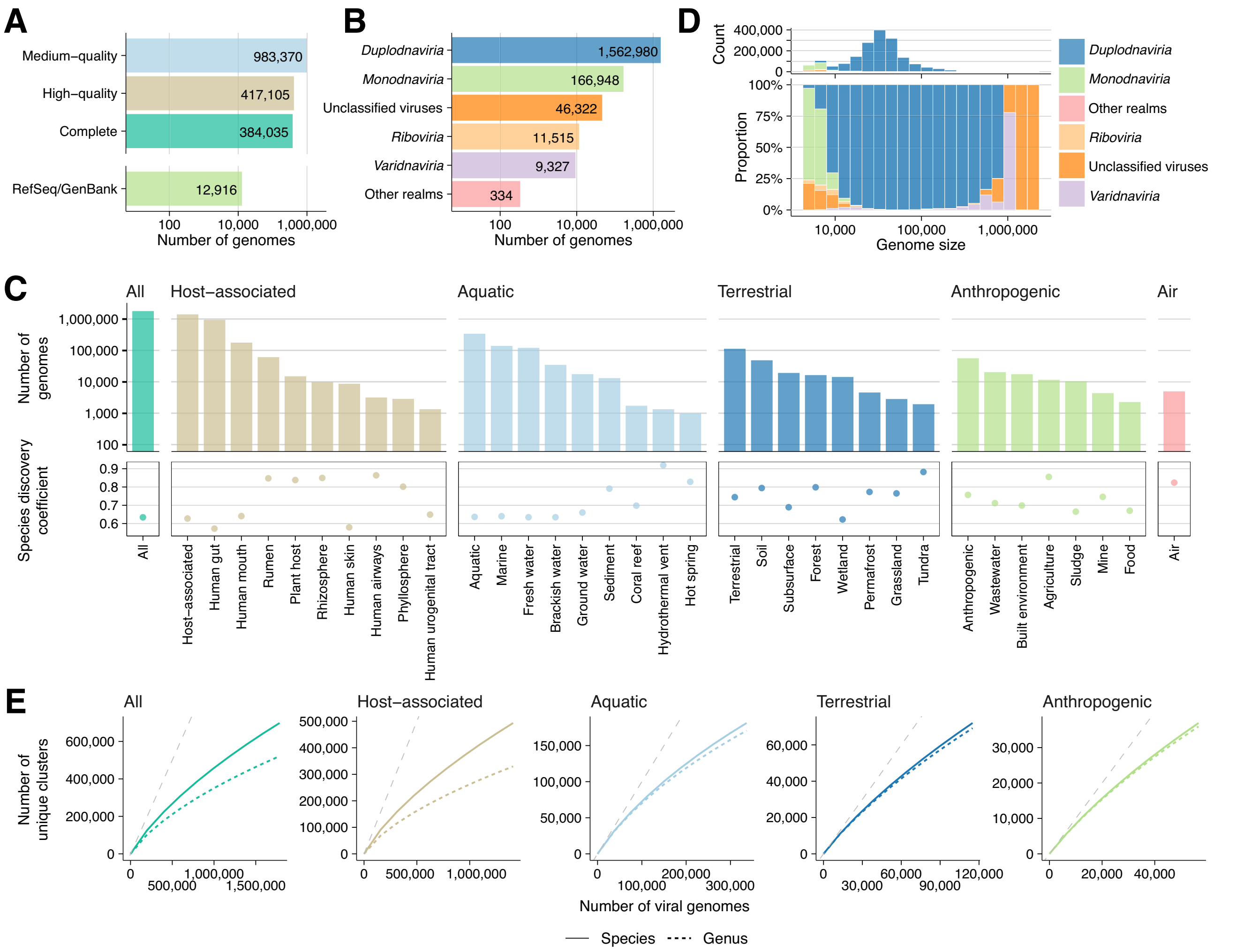


Figure 1. Number of viral genomes included in VIRE. (**A**) Bar plot showing the number of metagenome-derived viral genomes by quality category, as assessed by CheckV, together with viral genomes from RefSeq/GenBank. (**B**) Bar plot showing the number of viral genomes by predicted viral realm, based on classification by geNomad. (**C**) Bar plot showing the number of viral genomes by environment. Each metagenomic sample was annotated with microntology terms, and the number of viral genomes was aggregated based on the presence of each keyword. (**D**) Histogram (top) and stacked bar plot (bottom) showing the distribution of viral genome lengths, with colors indicating the predicted viral realms. For clarity, viral genomes shorter than 5 kb were excluded from the plot. (**E**) Rarefaction curves of viral genomes. All viral genomes were clustered at 95% ANI and 85% ANI to define species- and genus-level groups, respectively. For each environment, genomes were randomly subsampled 10 times, and the average number of recovered species/genus-level clusters was plotted. Dashed gray lines indicate the 1:1 line (diagonal) for reference.

remains substantially undersampled and highlight the need for continued expansion of metagenomic sampling efforts.

When we clustered the viral genomes from VIRE and those from IMG/VR v4 [37], the largest environmental viral genome database to date, at 95% ANI, we identified a total of 1 011 171 species-level clusters (Supplementary Fig. S4A). Of these, 56.1% were unique to VIRE. Compared with IMG/VR alone, VIRE effectively doubled the number of known species-level clusters. Furthermore, when comparing the proportion of viral genomes unique to VIRE across different environments, samples from rumen, coral reefs, built environments, and wastewater showed over 80% unique genomes not represented in IMG/VR (Supplementary Fig. S4B and C). In contrast, the human gut had the lowest proportion of unique genomes among host-associated environments. However, even in this well-studied environment, ∼42% of genomes were not present in IMG/VR. Other than the human gut environment, wetland, subsurface, and groundwater had lower proportions of unique genomes (Supplementary Fig. S4B and C). These findings demonstrate that VIRE contains novel viral genomes from a wide range of environments, including the extensively studied human gut.

Host annotation of phage genomes

Host annotation for phages in VIRE was performed systematically using CRISPR spacer–based predictions, a method recognized for its high specificity and low false-positive rate [53]. CRISPR spacers were extracted from ∼1.2 million bacterial/archaeal MAGs derived from the same set of metagenomic samples used for viral genome detection in VIRE, as included in the SPIRE resource [39], and an additional 1.0 million reference genomes from the proGenomes database [54], constituting the largest CRISPR spacer collection to date. These spacers were aligned to viral genomes using stringent matching criteria, resulting in host assignments for 46.8% of all viral genomes in the VIRE database. The predicted host organisms spanned 52 phyla, including both bacteria and archaea, and encompassed a total of 2367 genera, as defined by the GTDB taxonomy [58]. Among viruses with at least one predicted host, ∼40.7% were assigned to two or more host gen-

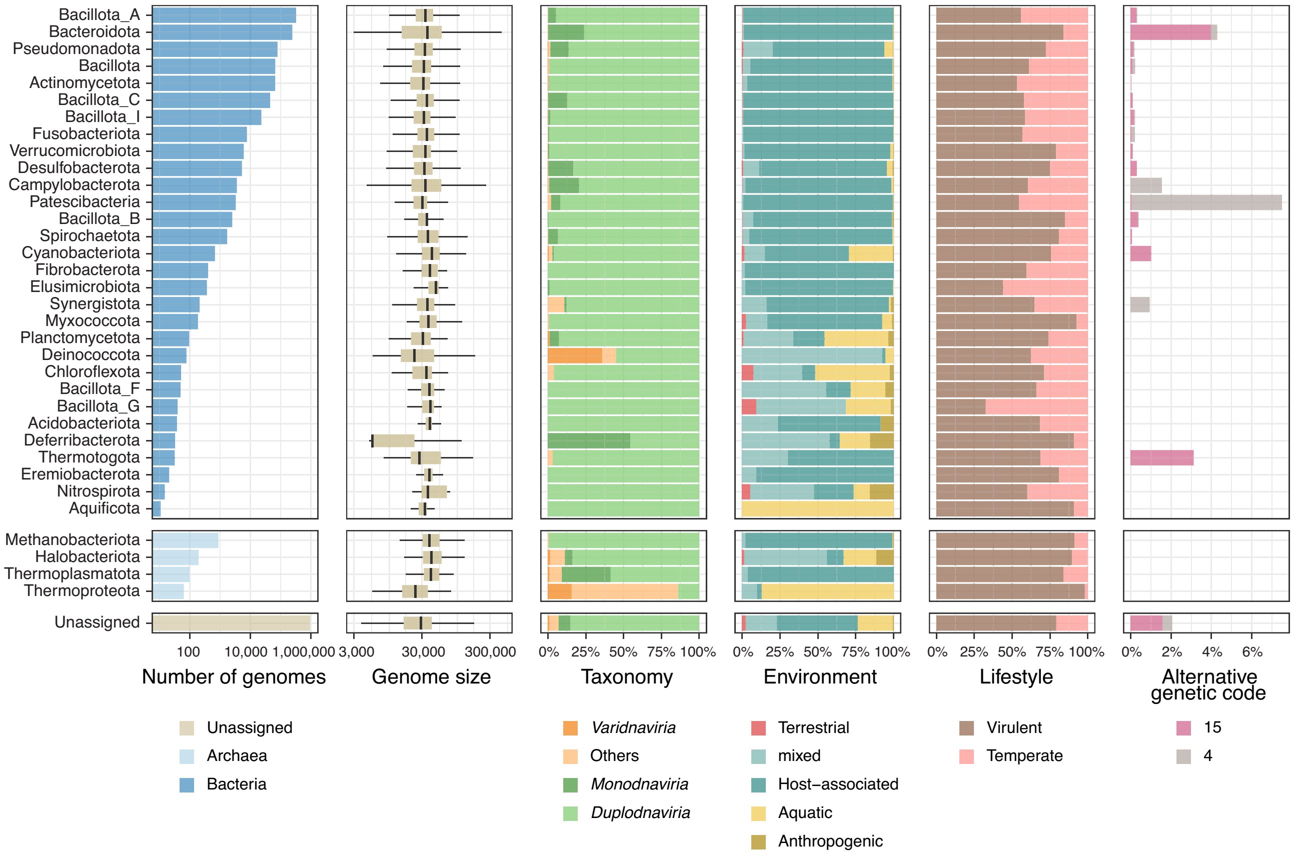


Figure 2. Prokaryotic host annotations for viruses in VIRE. Summary of viral features by predicted bacterial or archaeal host phylum. From left to right, the panels show: the number of viral genomes predicted to infect each phylum, genome size distribution, predicted viral taxonomy from geNomad, environmental source of the metagenomic samples, predicted viral lifestyle, and proportion of viruses predicted to use non-standard genetic codes, assessed by geNomad. Prokaryotic hosts were predicted by mapping CRISPR spacers derived from SPIRE MAGs and proGenomes reference genomes to the viral genomes. Taxonomic assignments for the MAGs and reference genomes were based on GTDB-Tk.

era, potentially representing broad-host-range phages as reported in recent studies [77, 78]. When stratified by environment, host-associated samples yielded the highest proportion of host-annotated viruses (57.4%), followed by those from anthropogenic (13.7%), terrestrial (8.5%), and aquatic samples (3.8%). The assigned host taxonomy was largely consistent with the bacterial taxonomies in the environment. For example, among host-associated viruses, the most frequently predicted hosts were *Faecalibacterium* spp., *Bacteroides* spp., and *Phocaeicola* spp., all of which are common and abundant members of the human gut microbiome. Moreover, at the phylum level, there was a strong positive correlation between the number of genomes/CRISPR spacers included in SPIRE/proGenomes and the number of viral genomes assigned to each phylum (Pearson's $r = 0.83$ and 0.86, respectively, Supplementary Fig. S5).

When viral genomes were classified according to the predicted bacterial or archaeal host phyla, several distinctive patterns were observed (Fig. 2). Among viruses predicted to infect members of the Bacteroidota phylum, 4.0% were inferred to use genetic code 15 instead of the standard bacterial genetic code 11. Most of these viruses belonged to the *Crassvirales* order, a dominant viral group in the human gut that infects *Prevotella*, *Bacteroides*, and *Phocaeicola* (Supplementary Fig. S6). This observation is consistent with previous reports showing that some phages infecting these gut species have alternative genetic codes [79, 80]. Similarly, 7.5% of viruses predicted to infect the Patescibacteria phylum (formerly known as CPR) were inferred to use genetic code 4. This finding is in line with a prior study suggesting that certain Patescibacteria lineages, such as the Absconditabacterales order, utilize alternative genetic codes [81], indicating possible phage adaptation to host-specific translation systems. While the majority of host-assigned viruses were classified as either tailed bacteriophages (e.g. members of the *Caudoviricetes* order within *Duplodnaviria*) or ssDNA viruses (*Monodnaviria*), an exception was observed for viruses predicted to infect Deinococcota, 35.9% of which were assigned to *Varidnaviria*, a viral realm that also includes eukaryotic viruses. This group included members of the non-tailed *Sphaerolipoviridae* family, which are known to infect *Thermus* species in the Deinococcota phylum and inhabit hot springs [82]. In addition, viruses predicted to infect the Deferribacterota phylum had a small genome size (median size = 5612 bp), due to a relatively high proportion of *Monodnaviria* (54.5%), which are ssDNA viruses with relatively small genomes (∼6 kb). These viruses were predicted to infect *Mucispirillum* spp. inhabiting the guts of rodents and other animals.

Functional annotation of viral genes

From ∼1.7 million viral genomes, a total of 89 469 781 protein-coding genes were predicted. These genes were comprehensively annotated using multiple functional databases, including eggNOG [63], KEGG [64], COG [65], PHROG [66], pVOG [67], Pfam [68], TIGRFAM [69], dbCAN [70], and CARD [62]. Overall, 40.2% of these genes had at least one hit in any of these databases (Fig. 3A). Among them, eggNOG yielded the largest number of hits, with 51.4% of all viral genes having at least one eggNOG hit, including 27.0% assigned to functionally characterized groups. Consistent with this, eggNOG provided the highest number of unique annotations among the databases (Supplementary Fig. S7). The second-largest number of hits was obtained from PHROG, a database grouping distantly related viral gene families, in which hallmark genes of tailed bacteriophages, such as integrase, terminase large subunit, and portal protein, were frequently identified [66] (Fig. 3B). Additional functional insights were obtained from the broader KEGG annotations, where the most frequently assigned functions included ssDNA-binding proteins, DNA methyltransferases, and DNA polymerases (Fig. 3B). Furthermore, annotations based on the dbCAN database identified lysozymes, viral proteins known to degrade bacterial cell membranes (Fig. 3B). Given that such phage-derived endolysins have potential as alternatives to antibiotics in antimicrobial therapies [83, 84], such annotations may offer valuable information for the rational design of lysozyme-based therapeutics.

Using CARD [62], we identified 618 genes annotated as antibiotic resistance genes, the most abundant being the ACI-1 gene, which confers resistance to cephalosporins. Most viruses carrying this gene had no predicted host, but previous studies have reported that ACI-1 is encoded by prophages in *Negativicutes* inhabiting the human gut [85]. Other detected resistance genes included emrE (from *Escherichia coli*), lnuC, and tet(W/N/W). Of the viruses carrying the resistance genes, 97.2% originated from host-associated samples. Given that host-associated phages account for 78.5% of the VIRE dataset, this represents a statistically significant enrichment of resistance genes in viruses from host-associated environments (Fisher's exact test, $P < .01$). Nevertheless, the fact that only 618 out of 89 million genes were annotated as resistance genes is consistent with previous studies, which have shown that phages rarely encode antibiotic resistance genes [86].

Phages encode AMGs, which modulate the metabolic functions of their bacterial or archaeal hosts during infection [87, 88]. We examined the distribution of a previously curated set of AMGs [71] in VIRE and found substantial variation in both their abundance and types of AMGs across environments (Fig. 3C). Aquatic viruses, for example, showed a higher proportion of genes assigned as AMGs than those from other environments, spanning diverse functional categories, particularly cofactor and vitamin metabolism, carbohydrate metabolism, and glycan metabolism. Whether the enrichment of AMGs in the aquatic environment simply reflects the greater number of studies conducted in aquatic systems remains to be clarified. In contrast, viruses from host-associated samples, especially those from the human gut, oral cavity, and skin, were enriched in AMGs related to amino acid metabolism and energy metabolism, but had lower frequencies of AMGs associated with carbohydrate and glycan metabolism. Terrestrial viruses from subsurface, tundra, and wetland environments were comparatively enriched in AMGs involved in secondary metabolite biosynthesis and terpenoid/polyketide metabolism relative to the host-associated viruses. These findings suggest that virus–host interactions exhibit environment-specific metabolic signatures, reflecting ecological adaptation between viruses and microbial communities in distinct habitats.

Seamless integration with other microbiome resources

The metagenomic samples in VIRE use identifiers consistent with those used in our previously developed resources, SPIRE (https://spire.embl.de) [39] and Metalog (https://metalog.embl.de) [40], enabling seamless cross-referencing across the resources. SPIRE is a large-scale microbial genome resource consisting of ∼1 million MAGs, allowing users to compare viral genomes and microbial genomes derived from the same metagenomic samples. Metalog is a manually curated metadata repository for metagenomic studies, providing environmental descriptors (including geographic coordinates) extracted from original publications. For human gut samples, Metalog additionally provides host demographic information (e.g. age, sex, geographic origin) and detailed clinical metadata such as disease status and medication use. Additionally, taxonomic profiles of the samples based on mOTUs [89] and MetaPhlAn [90] are also available. By linking viral genomes, microbial MAGs, microbiome taxonomic profiles, and environmental or clinical metadata, VIRE enables large-scale, integrative analyses of microbial ecosystems with unprecedented depth and context.

Web interface and accessibility

VIRE is publicly accessible at vire.embl.de, where users can browse and download viral genome sequences along with their associated metadata. For each viral genome, the interface provides access to the genome sequence, key quality metrics (e.g. geNomad scores, CheckV completeness, and contamination), predicted genes, functional annotations, host predictions, species- and genus-level cluster assignments, and corresponding metagenome and study metadata. Data can be explored and downloaded by environmental category (host-associated, aquatic, terrestrial, engineered, or human gut) according to the microntology ontology, or by individual study, enabling flexible access tailored to diverse research needs. Community contributions, feature requests, and bug reports are welcome via https://vire.embl.de/contribute.

Future directions

As the volume of publicly available metagenomic data continues to expand, VIRE will be regularly updated to incorporate newly identified viral genomes. Planned developments include refining viral detection algorithms and gene annotation pipelines to improve the identification of novel viruses and functional elements. Future releases will also integrate long-read metagenomic and metatranscriptomic datasets, broadening the scope to encompass non-tailed phages and RNA viruses. Continued integration with companion microbiome resources such as SPIRE and Metalog will further facilitate comprehensive exploration of viral and microbial ecology across ecosystems and host-associated environments. Over time, the web platform will be enhanced to support more advanced query and analysis capabilities, making VIRE an increasingly powerful tool for the virome research community.

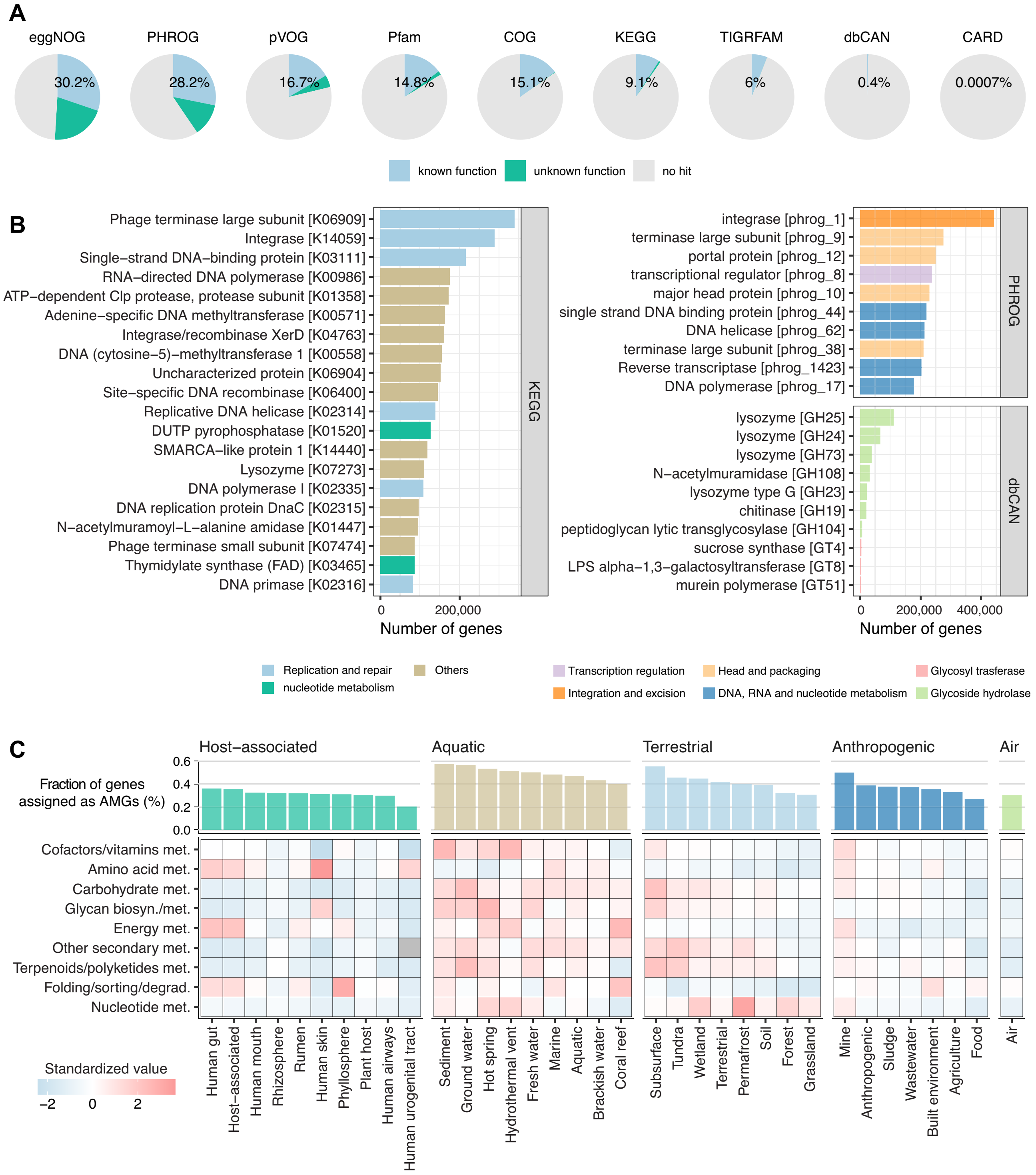


Figure 3. Viral genes and functional annotations. (**A**) Pie chart showing the proportion of viral genes annotated by each functional database. Blue color represents the proportion of genes matching known functions, while green indicates hits to hypothetical or uncharacterized proteins. The numerical values indicate the proportion of genes assigned to known functions. (**B**) Bar plot representing the number of functional annotations derived from KEGG, PHROG, and dbCAN databases. The top 20 functions from KEGG and the top 10 functions from PHROG and dbCAN are displayed. (**C**) Heatmap illustrating the distribution of AMGs across environments. Curated KEGG orthology terms [71]corresponding to AMGs were detected in viral genomes, and the percentage of AMGs in each category was calculated. Colors represent the standardized value (z-score) of the proportion of each AMG relative to the total number of genes in that environment. The bar plot above shows the percentage of genes assigned as AMGs for each environment.

Discussion

VIRE is a large-scale viral genome resource constructed from over 100 000 publicly available metagenomes using a consistent and standardized pipeline. The underlying metagenomes include a wide range of environments, offering a comprehensive resource for investigating viral diversity on a planetary scale. Each viral genome in VIRE is accompanied by comprehensive annotations, including genome quality metrics, taxonomic classification, predicted host, and gene-level functional annotations, all generated using state-of-the-art tools and databases. These features make VIRE a powerful platform for comparative viromics, host–virus interaction, and the exploration of viral functions across diverse ecosystems. A key strength of VIRE is its seamless integration with other metagenome-based resources, such as SPIRE and Metalog. This interoperability allows users to link viral genomes with MAGs and curated environmental or clinical metadata from the same samples, enabling multi-dimensional analyses of microbial communities in their ecological and host contexts. We anticipate that VIRE will serve as a foundational resource for advancing our understanding of viral diversity, evolution, and ecological roles and will remain a critical resource for the broader microbiome research community.

Acknowledgements

We thank members of the Bork group at EMBL for their support and constructive discussions. S.N. was partially supported by the Overseas Postdoctoral Fellowships of the Uehara Memorial Foundation and Takeda Science Foundation. T.S.B.S. was supported by Research Ireland under Grant Number 12/RC/2273-P2.

Author contributions: Suguru Nishijima (Conceptualization [lead], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Visualization [lead], Writing – original draft [lead], Writing – review & editing [lead]), Anthony Fullam (Data curation [equal], Resources [equal], Software [lead], Writing – original draft [equal], Writing – review & editing [equal]), Thomas S.B. Schmidt (Data curation [lead], Methodology [equal], Resources [lead], Writing – original draft [equal], Writing – review & editing [equal]), Michael Kuhn (Project administration [equal], Resources [equal], Supervision [equal], Writing – review & editing [equal]), and Peer Bork (Conceptualization [equal], Funding acquisition [equal], Project administration [lead], Supervision [lead], Writing – review & editing [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

Takeda Science Foundation, Uehara Memorial Foundation, and Research Ireland [12/RC/2273-P2]. Funding to pay the Open Access publication charges for this article was provided by Internal EMBL funding for publication.

Data availability

All data in the VIRE resource are freely accessible and downloadable via https://vire.embl.de. All metagenomic datasets used for the construction of this database are publicly available through the European Nucleotide Archive (ENA) [41] or the Sequence Read Archive (SRA) [42]. No in-house and unpublished metagenomes were included. VIRE is released under the Creative Commons Attribution-ShareAlike 4.0 International License.

References

1. Hendrix RW, Smith MC, Burns RN *et al.* Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* 1999;96:2192–7. https://doi.org/10.1073/pnas.96.5.2192
2. Suttle CA. Viruses in the sea. *Nature* 2005;437:356–61. https://doi.org/10.1038/nature04160
3. Rohwer F, Thurber RV. Viruses manipulate the marine environment. *Nature* 2009;459:207–12. https://doi.org/10.1038/nature08060
4. Shkoporov AN, Hill C. Bacteriophages of the human gut: the 'known unknown' of the microbiome. *Cell Host Microbe* 2019;25:195–209. https://doi.org/10.1016/j.chom.2019.01.017
5. Brito IL. Examining horizontal gene transfer in microbial communities. *Nat Rev Microbiol* 2021;19:442–53. https://doi.org/10.1038/s41579-021-00534-7
6. Borodovich T, Shkoporov AN, Ross RP *et al.* Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *Gastroenterol Rep (Oxf.)* 2022;10:goac012. https://doi.org/10.1093/gastro/goac012
7. Zimmerman AE, Howard-Varona C, Needham DM *et al.* Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol* 2020;18:21–34. https://doi.org/10.1038/s41579-019-0270-x
8. Kuzyakov Y, Mason-Jones K. Viruses in soil: nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biol Biochem* 2018;127:305–17. https://doi.org/10.1016/j.soilbio.2018.09.032
9. Suttle CA. Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol* 2007;5:801–12. https://doi.org/10.1038/nrmicro1750
10. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2012;2:63–77. https://doi.org/10.1016/j.coviro.2011.12.004
11. Reyes A, Semenkovich NP, Whiteson K *et al.* Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* 2012;10:607–17. https://doi.org/10.1038/nrmicro2853
12. Brum JR, Sullivan MB. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol* 2015;13:147–59. https://doi.org/10.1038/nrmicro3404
13. Gregory AC, Zablocki O, Zayed AA *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* 2020;28:724–40.e8. https://doi.org/10.1016/j.chom.2020.08.003
14. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G *et al.* Massive expansion of human gut bacteriophage diversity. *Cell* 2021;184:1098–109.e9. https://doi.org/10.1016/j.cell.2021.01.029
15. Nayfach S, Páez-Espino D, Call L *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* 2021;6:960–70. https://doi.org/10.1038/s41564-021-00928-6
16. Benler S, Yutin N, Antipov D *et al.* Thousands of previously unknown phages discovered in whole-community human gut

metagenomes. *Microbiome* 2021;9:78. https://doi.org/10.1186/s40168-021-01017-w

17. Reyes A, Haynes M, Hanson N *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 2010;466:334–8. https://doi.org/10.1038/nature09199
18. Breitbart M, Hewson I, Felts B *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 2003;185:6220–3. https://doi.org/10.1128/JB.185.20.6220-6223.2003
19. Wolf YI, Silas S, Wang Y *et al.* Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* 2020;5:1262–70. https://doi.org/10.1038/s41564-020-0755-4
20. Mizuno CM, Rodriguez-Valera F, Kimes NE *et al.* Expanding the marine virosphere using metagenomics. *PLoS Genet* 2013;9:e1003987. https://doi.org/10.1371/journal.pgen.1003987
21. Gregory AC, Zayed AA, Conceição-Neto N *et al.* Marine DNA viral macro- and microdiversity from pole to pole. *Cell* 2019;177:1109–23.e14. https://doi.org/10.1016/j.cell.2019.03.040
22. Brum JR, Ignacio-Espinoza JC, Roux S *et al.* Patterns and ecological drivers of ocean viral communities. *Science* 2015;348:1261498. https://doi.org/10.1126/science.1261498
23. Roux S, Brum JR, Dutilh BE *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 2016;537:689–93. https://doi.org/10.1038/nature19366
24. Angly FE, Felts B, Breitbart M *et al.* The marine viromes of four oceanic regions. *PLoS Biol* 2006;4:e368. https://doi.org/10.1371/journal.pbio.0040368
25. Ma B, Wang Y, Zhao K *et al.* Biogeographic patterns and drivers of soil viromes. *Nat Ecol Evol* 2024;8:717–28. https://doi.org/10.1038/s41559-024-02347-2
26. Graham EB, Camargo AP, Wu R *et al.* A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nat Microbiol* 2024;9:1873–83. https://doi.org/10.1038/s41564-024-01686-x
27. Emerson JB, Roux S, Brum JR *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* 2018;3:870–80. https://doi.org/10.1038/s41564-018-0190-y
28. Chen Y-M, Sadiq S, Tian J-H *et al.* RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat Microbiol* 2022;7:1312–23. https://doi.org/10.1038/s41564-022-01180-2
29. Koonin EV, Dolja VV, Krupovic M *et al.* Global organization and proposed megataxonomy of the. *Microbiol Mol Biol Rev.* 2020;84:e00061–19.
30. Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 2015;479-480:2–25. https://doi.org/10.1016/j.virol.2015.02.039
31. Rosenwasser S, Ziv C, van Creveld SG *et al.* Virocell metabolism: metabolic innovations during host-virus interactions in the ocean. *Trends in Microbiology* 2016;24:821–32. https://doi.org/10.1016/j.tim.2016.06.006
32. Olival KJ, Hosseini PR, Zambrana-Torrelio C *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* 2017;546:646–50. https://doi.org/10.1038/nature22975
33. Carroll D, Daszak P, Wolfe ND *et al.* The global virome project. *Science* 2018;359:872–4. https://doi.org/10.1126/science.aap7463
34. Shah SA, Deng L, Thorsen J *et al.* Expanding known viral diversity in the healthy infant gut. *Nat Microbiol* 2023;8:986–98. https://doi.org/10.1038/s41564-023-01345-7
35. Jian H, Yi Y, Wang J *et al.* Diversity and distribution of viruses inhabiting the deepest ocean on Earth. *ISME J* 2021;15:3094–110. https://doi.org/10.1038/s41396-021-00994-y
36. Ter Horst AM, Santos-Medellín C, Sorensen JW *et al.* Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* 2021;9:233. https://doi.org/10.1186/s40168-021-01156-0
37. Camargo AP, Nayfach S, Chen I-MA *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res* 2023;51:D733–43. https://doi.org/10.1093/nar/gkac1037
38. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA *et al.* Uncovering Earth's virome. *Nature* 2016;536:425–30. https://doi.org/10.1038/nature19094
39. Schmidt TSB, Fullam A, Ferretti P *et al.* SPIRE: a searchable, planetary-scale microbiome resource. *Nucleic Acids Res* 2023;52:D777–83. https://doi.org/10.1093/nar/gkad943
40. Kuhn M, Schmidt TSB, Ferretti P *et al.* Metalog: curated and harmonised contextual data for global metagenomics samples. *Nucleic Acids Res* 2025;gkaf1118. https://doi.org/10.1093/nar/gkaf1118
41. O'Cathail C, Ahamed A, Burgin J *et al.* The European Nucleotide Archive in 2024. *Nucleic Acids Res* 2025;53:D49–55. https://doi.org/10.1093/nar/gkae975
42. Katz K, Shutov O, Lapoint R *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res* 2022;50:D387–90. https://doi.org/10.1093/nar/gkab1053
43. Li D, Liu C-M, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–6. https://doi.org/10.1093/bioinformatics/btv033
44. Camargo AP, Roux S, Schulz F *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024;42:1303–12. https://doi.org/10.1038/s41587-023-01953-y
45. Nayfach S, Camargo AP, Schulz F *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2020;39:578–85. https://doi.org/10.1038/s41587-020-00774-7
46. Roux S, Adriaenssens EM, Dutilh BE *et al.* Minimum information about an uncultivated virus genome (MIUViG). *Nat Biotechnol* 2019;37:29–37. https://doi.org/10.1038/nbt.4306
47. Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9. https://doi.org/10.1038/nbt.3820
48. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage taxonomy. *Viruses* 2021;13:506. https://doi.org/10.3390/v13030506 .
49. Sayers EW, Cavanaugh M, Frisse L *et al.* GenBank 2025 update. *Nucleic Acids Res* 2025;53:D56–61. https://doi.org/10.1093/nar/gkae1114
50. Goldfarb T, Kodali VK, Pujar S *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 2024;53:D243–57. https://doi.org/10.1093/nar/gkae1038
51. Zielezinski A, Gudyś A, Barylski J *et al.* Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nat Methods* 2025;22:1191–4. https://doi.org/10.1038/s41592-025-02701-7
52. Prasoodanan PK,V, Maistrenko OM, Fullam A *et al.* A census of hidden and discoverable microbial diversity beyond genome-centric approaches. bioRxiv, https://doi.org/10.1101/2025.06.26.661807, 26 June 2025, preprint: not peer reviewed.
53. Edwards RA, McNair K, Faust K *et al.* Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* 2016;40:258–72. https://doi.org/10.1093/femsre/fuv048
54. Fullam A, Letunic I, Schmidt TSB *et al.* proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res* 2023;51:D760–6. https://doi.org/10.1093/nar/gkac1078
55. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60. https://doi.org/10.1038/nmeth.3176
56. Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

57. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2019;36:1925–7. https://doi.org/10.1093/bioinformatics/btz848

58. Parks DH, Chuvochina M, Rinke C *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785–94. https://doi.org/10.1093/nar/gkab776

59. Hyatt D, Chen G-L, Locascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119. https://doi.org/10.1186/1471-2105-11-119

60. Huerta-Cepas J, Forslund K, Coelho LP *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 2017;34:2115–22. https://doi.org/10.1093/molbev/msx148

61. Figueroa JL, Iii, Dhungel E, Bellanger M *et al.* MetaCerberus: distributed highly parallelized HMM-based processing for robust functional annotation across the tree of life. *Bioinformatics* 2024;40:btae119. https://doi.org/10.1093/bioinformatics/btae119

62. Alcock BP, Raphenya AR, Lau TTY *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2019. 48 D517–525https://doi.org/10.1093/nar/gkz935

63. Hernández-Plaza A, Szklarczyk D, Botas J *et al.* eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res* 2023;51:D389–94. https://doi.org/10.1093/nar/gkac1022

64. Kanehisa M, Furumichi M, Sato Y *et al.* KEGG: biological systems database as a model of the real world. *Nucleic Acids Res* 2025;53:D672–7. https://doi.org/10.1093/nar/gkae909

65. Galperin MY, Wolf YI, Makarova KS *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021;49:D274–81. https://doi.org/10.1093/nar/gkaa1018

66. Terzian P, Olo Ndela E, Galiez C *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* 2021;3:lqab067.https://doi.org/10.1093/nargab/lqab067

67. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* 2017;45:D491–8. https://doi.org/10.1093/nar/gkw975

68. Paysan-Lafosse T, Andreeva A, Blum M *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res* 2025;53:D523–34. https://doi.org/10.1093/nar/gkae997

69. Li W, O'Neill KR, Haft DH *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;49:D1020–8. https://doi.org/10.1093/nar/gkaa1105

70. Zheng J, Ge Q, Yan Y *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* 2023;51:W115–21. https://doi.org/10.1093/nar/gkad328

71. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;8:90. https://doi.org/10.1186/s40168-020-00867-0

72. Hockenberry AJ, Wilke CO. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *Peer J* 2021;9:e11396. https://doi.org/10.7717/peerj.11396

73. Zhao L, Shi Y, Lau HC-H *et al.* Uncovering 1058 novel human Enteric DNA viruses through deep long-read third-generation sequencing and their clinical impact. *Gastroenterology* 2022;163:699–711. https://doi.org/10.1053/j.gastro.2022.05.048

74. Chen L. Discovery and analysis of an 841 kbp phage genome: the largest known to date. bioRxiv, https://doi.org/10.1101/2025.01.14.633092, 16 January 2025, preprint: not peer reviewed.

75. Mukherjee S, Huntemann M, Ivanova N *et al.* Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci* 2015;10:18. https://doi.org/10.1186/1944-3277-10-18

76. Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–14. https://doi.org/10.1038/nature11234

77. de Jonge PA, Nobrega FL, Brouns SJJ *et al.* Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol* 2019;27:51–63. https://doi.org/10.1016/j.tim.2018.08.006

78. Bignaud A, Conti DE, Thierry A *et al.* Phages with a broad host range are common across ecosystems. *Nat Microbiol* 2025;10:2537–49. https://doi.org/10.1038/s41564-025-02108-2 CrossRef]

79. Borges AL, Lou YC, Sachdeva R *et al.* Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nat Microbiol* 2022;7:918–27. https://doi.org/10.1038/s41564-022-01128-6

80. Peters SL, Borges AL, Giannone RJ *et al.* Experimental validation that human microbiome phages use alternative genetic coding. *Nat Commun* 2022;13:5710. https://doi.org/10.1038/s41467-022-32979-6

81. Shulgina Y, Eddy SR. A computational screen for alternative genetic codes in over 250,000 genomes. *eLife* 2021;10:e71402. https://doi.org/10.7554/eLife.71402

82. Demina TA, Pietilä MK, Svirskaitė J *et al.* HCIV-1 and other tailless icosahedral internal membrane-containing viruses of the family Sphaerolipoviridae. *Viruses* 2017;9:32. https://doi.org/10.3390/v9020032 .

83. Gontijo MTP, Jorge GP, Brocchi M. Current status of endolysin-based treatments against Gram-negative bacteria. *Antibiotics* 2021;10:1143. https://doi.org/10.3390/antibiotics10101143

84. Rahman MU, Wang W, Sun Q *et al.* Endolysin, a promising solution against antimicrobial resistance. *Antibiotics* 2021;10:1277. https://doi.org/10.3390/antibiotics10111277

85. Rands CM, Starikova EV, Brüssow H *et al.* ACI-1 beta-lactamase is widespread across human gut microbiomes in Negativicutes due to transposons harboured by tailed prophages. *Environ Microbiol* 2018;20:2288–300. https://doi.org/10.1111/1462-2920.14276

86. Enault F, Briet A, Bouteille L *et al.* Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* 2017;11:237–47. https://doi.org/10.1038/ismej.2016.90

87. Thompson LR, Zeng Q, Kelly L *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci USA* 2011;108:E757–64. https://doi.org/10.1073/pnas.1102164108

88. Tian F, Wainaina JM, Howard-Varona C *et al.* Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. *Microbiome* 2024;12:159.

89. Dmitrijeva M, Ruscheweyh H-J, Feer L *et al.* The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. *Nucleic Acids Res* 2025;53:D797–805. https://doi.org/10.1093/nar/gkae1004

90. Blanco-Míguez A, Beghini F, Cumbo F *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 2023;41:1633–44. https://doi.org/10.1038/s41587-023-01688-w

Received: August 15, 2025. **Revised:** October 17, 2025. **Accepted:** October 17, 2025