EnVhog: an extended view of the viral protein space on Earth

Pérez-Bucio Rubén[1,2], François Enault[1,*,†], Clovis Galiez[2,*,†]

[1] Université Clermont Auvergne, CNRS, LMGE, F-63000 Clermont-Ferrand, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[†] these authors equally contributed to the work

Abstract

Over the last twenty years, hundreds of metagenomic studies have generated millions of viral genomic sequences from a wide variety of ecosystems. Despite this, the overall genetic diversity of viruses remains elusive, both in terms of the number of protein families they encode and the diversity of these families. Indeed, even if it is recognized that the organization of the viral protein sequence space requires sensitive homology detection methods, such methods have never been applied at a large scale. To produce a more realistic and comprehensive view of the protein diversity in the viral world, we have (i) collected thousands of viromes and identified viral contigs and proteins within them, (ii) retrieved viral proteins available in different public databases, and (iii) applied sensitive similarity searches to cluster all these proteins into families. More than 46 million deduplicated proteins were clustered into less than 2.3 million protein families. An iterative procedure to detect and remove genomic sequences of cellular origin, specially developed here, showed that only a very small fraction of sequences were likely to be cellular contamination (~2 % of contigs, 7 K clusters). The remaining 2,203,457 clusters were coined enVhogs (for environmental Viral homologous groups). Their multiple sequence alignments have been transformed into HMMs to constitute the EnVhog database. Even if only a small proportion of enVhogs were annotated (15.9 %), they encompass almost half of the protein dataset (44.8 %). Applied to the annotation of four recently published viromes from diverse environments (sulfuric soil, grassland, surface seawater and human gut), enVhog HMMs doubled the number of viral sequences characterized, and increased by 54%-74% the number of proteins functionally annotated. EnVhog, the largest comprehensive compilation of viral protein information to date, will thus further help to determine the functions of proteins encoded in newly sequenced viral genomes, and help to improve the accuracy of viral sequence detection tools.

EnVhog database is available at http://envhog.u-ga.fr/envhog.

* to whom correspondence should be addressed, Email: francois.enault@uca.fr, clovis.galiez@univ-grenoble-alpes.fr

Introduction

With the remarkable estimated amount of at least 10^{31} viral particles on the planet, viruses are probably the most abundant and diverse biological entity on Earth (Mushegian 2020). Viruses can be found in almost all studied ecosystems, ranging from ocean samples to host-associated microenvironments such as the human gut (Cobián Güemes et al. 2016). Viruses are key players in all these ecosystems, from driving cellular mortality and evolution, to influencing global-scale biogeochemical cycles (Wieczynski et al. 2023). In addition to their large number, viruses are diverse, as evidenced by the existence of numerous groups with independent origins (Koonin et al. 2022). The many strategies implemented by these entities are reflected in the nature of their genetic support and also in the size of their genomes, which ranges from 1.7 Kb to 2.3 Mb. Thus, despite their small individual size, the number of different protein families encountered in viruses is high, especially as many cellular genes, even ribosomal ones, have been sporadically incorporated into viral genomes (Mizuno et al. 2019). Moreover, each viral protein family is diverse (Terzian et al. 2021), with high sequence heterogeneity, particularly for structural proteins.

Due to this diversity in terms of number of genome families, protein families and of the heterogeneity of most of these families, viral sequence space still remains far from well characterized. Virus genomes are accessed through isolation on a host for more than 40 years. Yet, as only a small fraction of microbes have culture representatives, most viruses on Earth remain uncultivated. To discover previously unseen genomes of viruses, metagenomic sequencing is the most commonly used methodology (Rosario and Breitbart, 2011). High throughput sequencing can be applied to study (i) microbiomes, by considering all microbes including viruses, or (ii) viromes by specifically targeting viruses mostly through prior filtration. Whether from microbial or viral metagenomes, the viral origin of the contigs must be verified as even viromes often contain a significant amount of cellular DNA (Roux et al. 2013). The identification of viral contigs from these datasets remains a difficult task as (i) some proteins are shared by the viral and cellular world and (ii) many proteins on contigs of viral origin are not similar to any known viral sequence. Considering this last point, no homologs are detected for many new viral proteins because these proteins are too distantly related from known ones or because they correspond to unseen proteins. Without identified homologs, no functional annotation is possible for such viral proteins.

To improve this crucial goal of identifying homologs for new viral proteins, two avenues can be explored. First, the most straightforward solution is to use the largest possible database of viral proteins. Several viral sequence databases have been developed in recent years, made of either cultivated viruses (RefSeqVirus) or non-cultivated viral genomes from metagenomes of marine systems with efam (Zayed et al. 2021), or from all ecosystems such as GL-UVAB (Coutinho et al. 2019) or IMG/VR (Camargo et al. 2023). The second solution to better identify similarities between sequences of interest and reference proteins is to use remote homology detection methods. The most sensitive method that can be used routinely is based on Hidden Markov Model (HMM) comparisons (Söding 2005). Many viral protein HMM databases exists, here again either based on viral isolates such as POGs/pVOGs (Kristensen et al 2013; Grazziotin et al 2017), VOGDB (http://vogdb.org/) and PHROG (Terzian et al. 2021) or (ii) based on viral reference sequences and curated viral contigs from metagenomes, such as VPF (Paez-Espino et al, 2016) or efam (Zayed et al. 2021). Although these databases provide

protein families in the form of HMMs, the creation of these families did not involve any remote homology detection step, with the exception of PHROG. To build these PHROGs, viral proteins were first clustered based on BLAST-like similarity searches as the other protein family databases. Then, HMMs of these protein clusters were generated and compared to each other using HMM-HMM comparisons and further used to build super-clusters (Terzian et al. 2021). This two step strategy proved to find more distant homologous relationships and result in generating more diverse protein families (Olo Ndela et al. 2021; Olo Ndela et al. 2022).
Here, we developed EnVhog, a database of viral protein families, whose interest lies both (i) in the incorporation of viral protein sequences from any possible sources, with an effort to collect all sequences from published viromes and (ii) in the clustering of these sequences using remote homology in the form of profile-sequence clustering.

Materials & Methods

Collecting viral contigs from in-house assembled public viromes

To create the Envhog database, all publicly available viral metagenomic studies published from 2006 to 2019 were first identified. A total of 3,257 viromes were collected from various sources, all as raw reads except for two that were already assembled. Each virome was downloaded, reads were trimmed and assembled, using softwares adapted to the sequencing technologies used (Table S1). In order to detect viral sequences among these viromes, contigs longer than 1.5 Kb were processed by VirSorter2 (Guo et al. 2021), VIBRANT (contigs containing at least four ORFs) (Kieft et al. 2020) and DeepVirFinder (score > 0.9 and p-value < 0.05) (Ren et al. 2020). Sequences detected as viral by at least two of these three softwares were considered as viral and were kept for further analysis. Proteins were predicted using prodigal (v2.6.3; Hyatt et al. 2010) and only full-length proteins with less than five unknown amino acids (Xs) were considered here.

Collecting viral genomic information and proteins from existing databases

In addition, viral genomic information was collected from multiple existing databases. Reference genomes of (pro-)viruses were downloaded from PHROG (Terzian *et al.* 2021) and RefSeqVirus (accessed at 17/10/2022). Viral genomic sequences, mostly from uncultivated viruses, were also downloaded from IMG/VR (v4.1, Camargo et al. 2023), GL-UVAB (Coutinho *et al.* 2019), and efam (Zayed et al. 2021). Proteins encoded by these genomes and contigs were retrieved and added to virome proteins. All contigs were affiliated to a taxonomy using MMseqs2 based on the taxonomic affiliation of RefSeqVirus sequences (Mirdita et al. 2021).

Clustering all proteins

As some genomic sequences may be present in several databases, proteins were first deduplicated using the *linclust* module of MMseqs2 modules (Steinegger and Soeding 2018; Mirdita et al. 2019) with a minimum identity of 99% and a coverage of at least 99% of the shortest sequence. These deduplicated protein sequences were then clustered into homologous groups using a three-step iterative process, here again based on different MMseqs2 modules (Steinegger and Soeding 2017). The first clustering step, hereafter called shallow clustering,

aimed to gather together very similar proteins. Shallow clusters were built using the *linclust* module with a minimum identity of 90% and a coverage of at least 95% of the shortest sequence. Shallow clusters were subsequently submitted to a second clustering step, called standard clustering. To this end, representative proteins of each shallow cluster were compared and grouped, using the *cluster* module with an e-value of at least 10^{-3}, a coverage of 70% of the longest sequence in the alignment, and 30% of identity. Then, standard clusters were grouped into deep clusters using remote homology relationship detection. First, a profile (position-specific scoring matrix) and a consensus sequence were built for each standard cluster. Then, all profiles were compared to all consensus sequences (MMseqs2 with sensitivity -s 4, e-value threshold of 10^{-3}, a coverage of at least 60% of the mutual alignment) and the similarities identified were used to group corresponding standard clusters into deep clusters (set-cover clustering). The parameters used in the second and third clustering steps were defined based on results described in the PhaMMseqs pipeline (Gauthier et al. 2022). For more details about the parameters and clustering strategy, see Supplementary information 3.
Deep clusters containing at least two deduplicated proteins constitute the final protein groups, hereafter termed enVhogs for *environmental Viral homologous protein groups*. The number of effective sequences (NEFF) was used (Meier and Soeding 2015) to evaluate the sequence diversity inside each enVhog. The higher the NEFF value, the more diverse are proteins in the cluster, and a minimal value of one meaning the proteins of the cluster are identical.

Estimating the viralness of each enVhog

The occurrence of each enVhog was then quantified in the viral and cellular worlds (Kristensen et al. 2013). To this end, KEGG prokaryotic genomes (version Jan 2018) were first stripped of their prophages, identified using VirSorter (Roux et al. 2015a). Second, the remaining KEGG and RefSeqVirus proteins were separately clustered (MMseqs2, coverage>0.5) and compared to enVhogs (MMseqs2, bit-score>50). All proteins inside a cluster similar to a given enVhog were considered as similar to this enVhog. For each enVhog, the proportion of cells and viruses with at least a protein similar, named Cp and Vp respectively, were computed as follows. First, cellular species are not evenly represented in KEGG, 2612 of the 2068 species having only one genome whereas for example *Chlamydia trachomatis* account for as many as 73 of the 4755 genomes. For each enVhog and each cellular species, the ratio of strains with a match was computed and the Cp was defined as the sum of these ratios for all species, divided by the total number of species (2612). Second, as viruses have multiple independent origins and very variable gene content, two viruses belonging to different viral groups such as classes or order, will only share a limited number of proteins. For example, a protein of a microvirus strain will only be found in other microviruses and the quantitative presence of this protein has to be computed considering only this viral clade. Therefore, RefSeqVirus genomes were first clustered into clades using MCL (inflation of 1.2, weight=percent of shared proteins between virus pairs based on the RefSeqVirus protein clusters), the 18,846 viruses being gathered into 596 clusters and 497 singletons. For each enVhog, only viral clades with at least one protein similar to this enVhog were considered to compute Vp, the other clades being considered as evolutionarily unrelated. Yet, for virus clades only containing one virus (497 singleton viruses), all proteins will automatically be considered as conserved in all the viruses of this clade, and enVhog similar to these proteins will have a Vp of 1. To reduce this bias, a global pseudocount

of 24 over all clades was added to the true number of viruses in a clade to avoid giving a too strong weight to these small virus clades. The importance of this pseudocount has been scaled down by the number of clades the enVhog appears in. Formally, the Vp amounts to:

$$V_p(enVhog) = \frac{1}{\#similar\ clades} \sum_{C:similar\ clades} \frac{\#viruses\ with\ \geq 1\ similar\ protein\ in\ C}{\#viruses\ in\ C + \frac{24}{\#similar\ clades}}$$

EnVhog viral quotients were computed slightly differently as in (Kristensen et al. 2013):

$$V_q(enVhog) = \frac{V_p}{V_p + (C_p - 0.05)^+}$$

Here, we removed 0.05 to Cp as we realized that some distant (or non-complete) prophages were not detected by VirSorter in KEGG genomes and were still considered part of the cellular world, blurring the signal for proteins scarcely conserved in viral genomes.

Considering Vq and Cp, enVhogs were categorized as : (i) 'Viral' if conserved in the viral clades they appear in and rarely found in cellular genomes (Vq≥0.8), (ii) 'Cellular' if more often conserved in cells than in viruses (Vq<0.5 and Cp>0.05), (iii) 'Shared' if conserved by both worlds (0.5<Vq<0.8 and Cp>0.05) and (iv) 'Unknown' if not similar to any known viral sequence and not (significantly) conserved in cellular genomes (Vq==0 and Cp<0.05). In addition, enVhogs with a viral hallmark annotation were added to the 'Viral' category. See bottom part of Fig. 3 for a graphical view of the classification.

Annotating enVhogs

EnVhog clusters were functionally annotated using PHROG (Terzian et al. 2021) and Pfam (Mistry et al. 2021) databases. The moderate size of PHROG allowed us to run HHblits (one iteration, e-values <0.001, and default parameters) in all-against-all comparison between enVhogs and PHROG HMMs. To annotate the enVhogs with Pfam-A profile database (accessed at 22-Oct-2022), we used MMseqs2 *search* with Pfam-A profiles as queries and enVhog representative sequences as target (parameters: sensitivity 5.7, Coverage 0%, E-value 0.001, Coverage mode 0).

Annotation of proteins from environmental samples using enVhogs

Viral contigs assembled from viromes sampled in four distinct ecosystem types were downloaded, namely sulfuric soils (Bi et al. 2024), grassland soils (Santos-Medellín et al. 2023), surface marine water (Ni et al. 2024) and human gut (Garmaeva et al. 2024). Proteins were predicted using prodigal (parameters -c -m -p meta) and a random set of 1000 proteins was sampled in each of the four sets. These proteins were compared to RefSeqVirus proteins using MMseqs (e-value<1e-3). RefSeqVirus proteins were considered as unannotated if they contained the words "hypothetical protein", "unknown" or whose annotations consisted only of an ORF number. Virome proteins were also compared to PHROG and EnVhog HMMs using HHBlits (e-value<1e-3).

To evaluate the results obtained using more sensitive searches against EnVhog, the 1,159,978 complete proteins (prodigal; parameters -c -m -p meta) from human gut viral contigs (Zolfo et al. 2024) were grouped into clusters (mmseqs; >70% of bidirectional coverage). For each of the 65,409 clusters, a multiple sequence alignment was computed (MMSeqs2 result2msa) and the

corresponding HMM was compared to enVhog HMMs (HHblits; E-value<1e-3). Finally, to test the results obtained using EnVhog on a microbiome, the same strategy was used for proteins from a microbiome of IBDMDB (sample PSM6XBTR; Lloyd-Price et al. 2019), resulting in 41,505 clusters encompassing 75,278 proteins.

Results

Leveraging a large number of publicly accessible viral metagenomes

A set of 3,257 publicly available viromes were identified in 160 different publications. These viromes, sampled from all over the world, were classified as environmental or host-associated and further assigned to 31 environmental subcategories (Fig. 1; Table S1). In the 2,409 host-associated metagenomes, viral communities present in different tissues from a broad range of organisms were sampled, even if datasets from the human gut represent more than half of this category (1,372 viromes). Three softwares meant to detect viral sequences, VirSorter2 (Guo et al. 2021), VIBRANT (Kieft et al. 2020) and DeepVirFinder (Ren et al. 2020), were ran on contigs longer than 1.5 Kb and only sequences predicted as viral by two softwares were further considered as in Zayed et al. 2021. From the 12,533,593 contigs longer than 1.5 Kb, a set of 1,664,423 were detected as viral, from which 8,978,898 proteins were predicted.

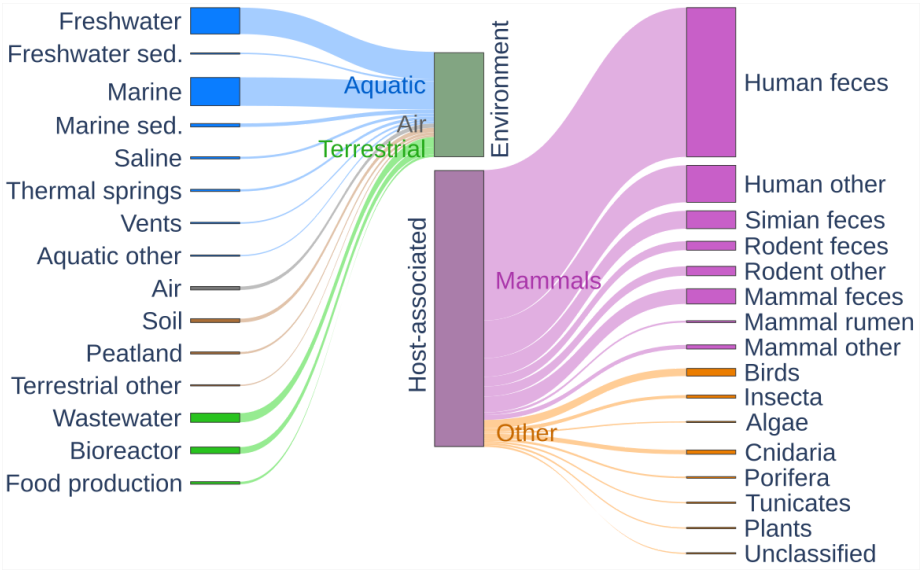

Figure 1 : Environmental categories of the viromes assembled for EnVhog*. Sankey plot of the ecosystem classification. The 3,257 collected viromes mostly fall into two categories: environmental viromes and host-associated viromes (at the center of the plot). Environmental viromes were further categorized into Aquatic, Air and Terrestrial, whereas host-associated datasets were separated into those associated to Mammals and to Other hosts.*

Creation of the final set of proteins from various sources

In order to create the largest possible set of viral proteins, sequences from different sources were collected and added to virome contigs described above. Reference genomes of (pro-)viruses were downloaded from PHROG (Terzian *et al.* 2021) and RefSeqVirus. Viral genomic sequences (*i.e.* complete genomes and genome fragments) obtained through culture-independent approaches, such as metagenomics, fosmid libraries, single-virus

sequencing and prophage mining were also downloaded from different databases. Among these, IMG/VR (v4.1, Camargo et al. 2023) is mostly made of viral sequences extracted from microbiomes, GL-UVAB (Coutinho *et al.* 2019) mostly from marine viromes and microbiomes and efam (Zayed *et al.* 2021) only from marine viromes. Together with the virome proteins described above, as many as 129.9 M proteins encoded by 7.7 M viral (fragments of) genomes were thus collected (Table S2). Databases compiled in this study are partially overlapping: reference phages are present both in PHROG and RefSeqVirus and identical metagenomes can be found in IMG/VR, GL-UVAB, efam and/or the 3,257 viromes. In addition, viruses with almost identical genomes can be found in RefSeq or in metagenomes from related samples. Protein sequences were thus deduplicated (>99 identity percent and coverage), the resulting 53.6 M distinct sequences being the dataset further described. It has to be noted that the 9 M proteins specifically collected here from public viromes account for 4.4 M deduplicated protein sequences not available in any other viral databases (Fig. 2, left).

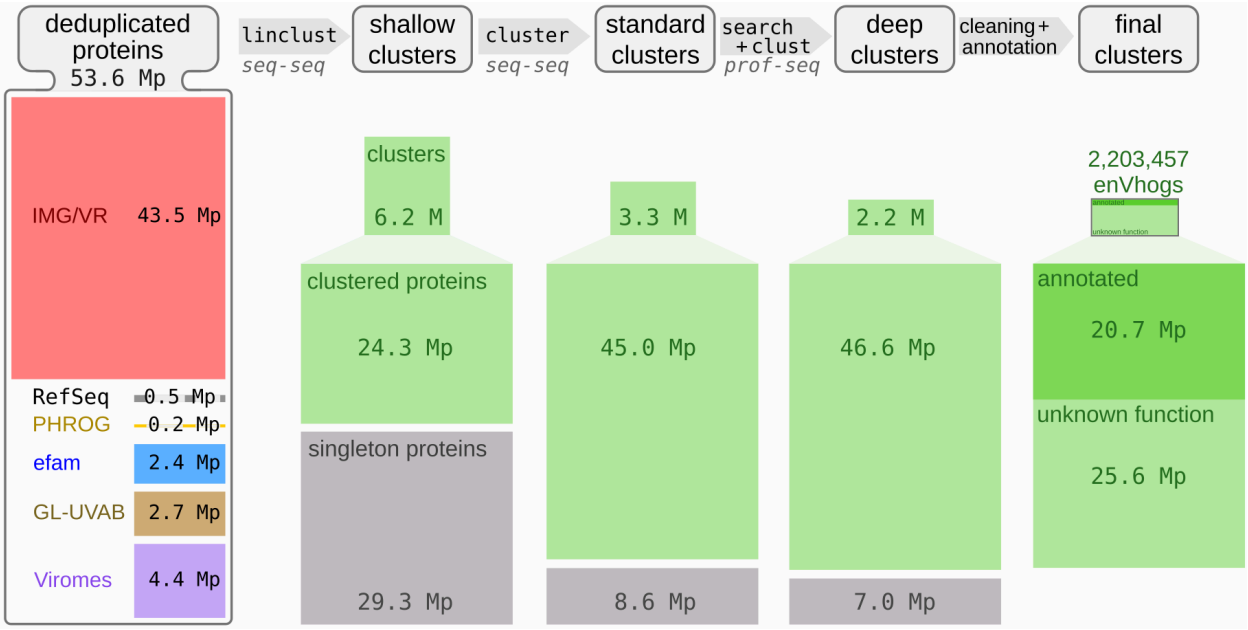

Figure 2: overview of the protein dataset and of the results of the different clustering steps. Left*: contribution of each of the data sources in terms of the number of the 53.6 M deduplicated proteins (Mp stands for Millions of proteins).* ***Right*** *: the three clustering steps, namely* shallow, standard *and* deep clustering *and their results in terms of number of clusters and singleton proteins. The number of clustered proteins is also indicated below the number of clusters. At the top, the MMseqs2 modules used, their philosophy and major parameters are indicated (seq-seq and prof-seq stand respectively for sequence-sequence and profile-sequence comparisons).*

Considering the taxonomic affiliation, most viral realms are significantly represented, with 0.04 M proteins from *Adnaviria* (archaeal filamentous viruses with A-form dsDNA genomes and alpha-helical MCP), 38.40 M from *Duplodnaviria* (dsDNA viruses with a HK97-fold MCP), 3.43 M from *Varidnaviria*, (dsDNA viruses with a vertical jelly roll MCP), 0.24 M from *Monodnaviria* (ssDNA viruses with a HUH superfamily endonuclease) and 0.48 M Riboviria (RNA viruses with a RNA-dependent RNA polymerase and viruses that encode reverse transcriptase), only *Ribozyviria* (hepatitis delta-like viruses with circular, negative-sense ssRNA genomes) being very rarely found in proteins obtained from metagenomes. The most prevalent viral class is *Caudoviricetes* (38.1 M proteins), the second most abundant class is *Megaviricetes* (2.6 M proteins) (Fig. S1). At the family level, 215 viral families are represented, 138 of which encompass more than a thousand proteins. In addition to these proteins encoded by affiliated contigs, as many as 10.39 M proteins are encoded by contigs that remained unclassified.

The three steps clustering procedure including profile-sequence comparisons

To provide a database of viral proteins organized into homologous groups or families, the next step was to detect homology relationships between sequences. Ecosystems not being evenly represented in public viromes (Fig. 1) and in the other databases, with an overrepresentation of the human gut and marine systems, we postulated that very similar proteins might be present in these similar samples, known to harbor viruses with very similar genomes. Thus, we decided to perform a first clustering step, called here shallow clustering, to eliminate protein sequence microdiversity and reduce the volume of data to be processed in the following clustering procedure. The 53.6 M proteins were thus clustered into groups of very similar proteins, ie shallow clusters (identity > 90 %, coverage > 95%). This threshold of amino-acid identity roughly corresponds to proteins encoded by genomes from the same genus, usually defined by a 70-80 % nucleotidic identity percent. A total of 35.5 M shallow clusters were obtained, made of 6.2 M clusters of at least two proteins, and 29.3 M singleton proteins (Fig. 2). Secondly, the 35.5 M representative proteins of shallow clusters were grouped into standard clusters. For this step, MMSeqs2 (BLASTp-like) sequence-sequence comparisons were performed and used as a basis for the clustering (e-value < 10^{-3}, identity>30%, coverage>70%). This methodology is classically used by most databases to generate their protein families, for example in efam (Zayed *et al.* 2021). A total of 11.9 M standard clusters, made of 3.3 M clusters (containing 45 M proteins) and 8.6 M singletons were obtained. A third clustering step was performed to further group homologous proteins. Indeed, most viral protein families were described to be diverse and BLASTp-like tools are not sensitive enough to detect all homology relationships (Terzian *et al.* 2021). Standard cluster profiles were thus compared to all cluster consensus sequences (e-value < 10^{-3}, coverage > 60%), and deep clusters were generated. Alongside 7 M proteins that remained as singletons, a total of 2.2 M deep clusters (2,210,249 precisely) were obtained, these clusters being hereafter termed enVhogs. These enVhogs encompass 46.6 M proteins, *i.e.* 87% of the initial set of 53.6 M proteins.

Cluster structure and diversity

The cluster structure at each clustering step highlights the significant increase in cluster size at each step (Fig. S2). Indeed, 50% of the initial 53.6 M deduplicated proteins were grouped in deep clusters containing more than 131 proteins, compared to 42 and 2 for standard and shallow clusters respectively. Deep clusters of at least ten proteins contain more than 75% of the proteins (75.7%), compared to 66.9% for standard clusters. A total of 165 deep clusters larger than 10,000 proteins were generated, and only 91 for standard clusters, two deep clusters exceeding 40,000 proteins (43,904 and 48,424).

To quantify sequence heterogeneity inside enVhogs (*i.e.* deep clusters), NEFF values were computed for each cluster. This sequence diversity measure is the average entropy of all columns constituting a multiple sequence alignment. NEFF values were also computed for PHROGs and Uniclust30 clusters (Mirdita et al. 2016). EnVhogs were more diverse than UniClust30 homologous groups at every cluster size (Fig. S3) and had a diversity comparable to PHROGs of similar size. For example, for clusters containing from 6 to 20 proteins, the average NEFF values were of 2.5, 2.4 and 2.1 for EnVhog, PHROG and Uniclust30 respectively, and of 5, 5.2 and 4 for clusters containing more than 100 proteins. It has to be noted that small clusters (containing between 1 and 5 proteins) were more diverse in EnVhog compared to the two other databases (1.44 versus 1.26 and 1.19 for PHROG and Uniclust30 respectively).

Estimation of the viralness of enVhogs

It has been shown that a fraction of protein families from cultivated viruses are shared by (some) cellular genomes (Kristensen *et al.* 2013). Thus, the viralness of an enVhog, *i.e.* its conservation in the viral world compared to that in the cellular world, is an interesting measure to have for future users. The viralness of each enVhog was estimated through the computation of a viral quotient (Vq) based on its conservation in viral clades it appears in and on its conservation in cellular species. Furthermore, these indices allowed us to categorize each enVhog as being Viral, Shared, Cellular or Unknown. First, as many as 498,505 enVhogs were categorized as Viral thanks to their high Vq, representing 22.5% of the enVhogs and 55.6% of the proteins in enVhogs (Fig. 3 and Table S3). In addition, 15,394 of the 77,759 enVhogs with a viral hallmark annotation that were not in this Viral category considering these criteria were now considered as Viral (0.9% of proteins). Conversely, only 35,097 enVhogs (6.2% of proteins) were categorized as Cellular, among which only 0.38% of total enVhog (0.27% proteins) had even no similarity with any known viral sequence ('No Viral Hit' column on Fig. 3). Only 0.26% of enVhogs (1.5% of proteins) were considered as Shared, the great majority of enVhogs being Unknown, either because not similar to any sequence (72% of enVhogs, 29.27% of proteins) or not similar to any virus and not significantly conserved in cellular genomes (3.5% of enVhogs, 2.4% of proteins).

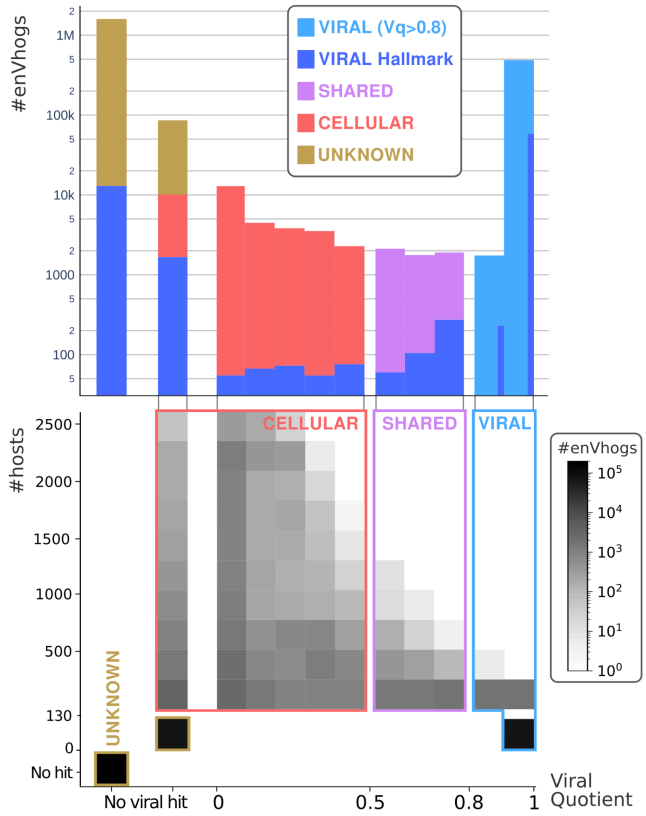

Figure 3: distribution of enVhogs in the four categories (Viral, Shared, Cellular and Unknown). *Top: Number of enVhogs in the four categories according to their Viral Quotient (Vq) in the x-axis. Bars on the left correspond to enVhogs with no hit in viral and cellular databases and no hit against the viral protein database. Y-axis, representing the number of enVhogs, is in log-scale. Bars are divided by the number of enVhogs belonging to each of the four categories whose definition is explained in the bottom half of the figure. EnVhogs tagged as Viral with Viral Quotient < 0.8 are proteins identified as viral hallmarks. Bottom : heatmap of the number of enVhogs (colors corresponding to log-scale) according to their conservation in cellular species (y-axis) and to their viral quotient (Vq on the x-axis). The four categories were defined based on thresholds on these hostness and Vq, and are here underlined.*

Contig labeling to further dig in protein dark matter

Most enVhog display no similarity to any known sequence and were categorized as Unknown. To refine this, we devised an algorithm that categorizes the nature of contigs from the category of the envhog they encode, and then refine the category of enVhogs using the ones of the contigs they appear in, and so on iteratively. Briefly, the category (Viral, Shared, Cellular or Unknown) of each enVhog was first transferred to all the proteins constituting them. Then, each contig was tagged as either Viral, Cellular or Unknown if encoding a majority of proteins from the corresponding category, the remaining contigs that do not encode proteins from a particular category being tagged as Hybrid. The predicted categories of the contigs are transferred back to the proteins and then used to classify the previously unclassified enVhogs: for each yet unlabeled enVhog (i.e. enVhog in category U or S), given its number of occurrences in the different contig categories (V/C/H), a simple Bayesian decision rule (see SI3 and Fig. S5 for details) decides if there is enough evidence to label the enVhog as Viral or Cellular. When labelled as Cellular, if the Cp (conservation in cellular species, see Materials and Methods) is below 0.05, then the category of the enVhog falls back to Shared class. This strategy (from

enVhogs to contigs and back to enVhogs) was run iteratively three times. See supplementary information 6 for further details.
This procedure allowed us to classify as Viral most of the enVhogs, especially the ones originally categorized as Unknown because their proteins were not similar to any known sequences (Fig. 4). Indeed, Viral enVhogs now account for 89% of the whole set (81% of proteins) whereas Unknown enVhogs only account for less than 2% of proteins and enVhogs.

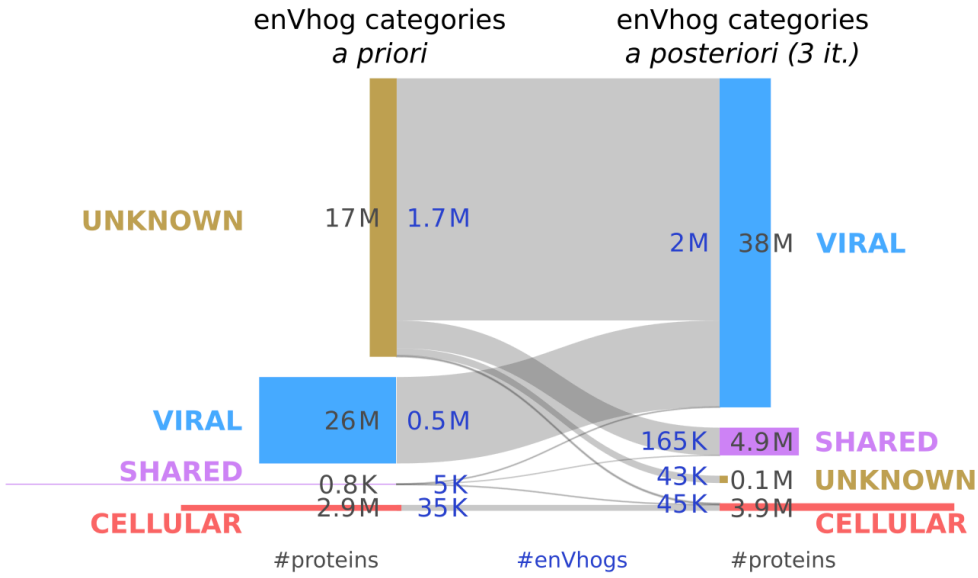

Figure 4: EnVhog classes. *Sankey plot of the evolution of categories of enVhogs using the iterative procedure. The initial number of enVhogs (in dark blue) and the number of associated deduplicated proteins (in dark grey) in the four categories (Unknown, Viral, Shared and Cellular) are represented on the left. The same numbers are displayed after three iterations of the procedure (on the right). Height of boxes is proportional to the corresponding number of enVhogs and their surface to the number of proteins.*

In addition to ascertaining the viral origin of most contigs, these results also highlight a limited set of 109,959 contigs that are predicted to be of cellular origin. Despite the strict criteria used to identify viral contigs and our use of well-established data sources, it is not surprising to find a small percentage (1.43%) of contigs predicted to be of cellular origin by a complementary approach such as the procedure developed here. Some contigs from all databases were predicted as Cellular (66,541, 36,359, 5,434, 1,516 and 109 in viromes, IMGVR, eFAM, GLUVAB and PHROG respectively). To improve the quality of our database, we decided to remove (i) these contigs, (ii) the proteins they encode and (iii) the 6,792 enVhogs having lost more than half of their deduplicated proteins and those which became too small (containing less than two proteins). Unsurprisingly, the filtered out enVhogs are enriched with hits against PFAM (35.5%) and cellular proteins (64%) compared to the rest of enVhogs (7.9% and 14.7% respectively). The remaining 2,203,457 enVhogs constitute the final set of enVhog database and is simply referred as enVhog hereafter.

Functional annotation of enVhogs

To provide a valuable viral protein resource, enVhogs were annotated using multiple databases and remote homology detection tools. A total of 292,089 enVhogs were similar (HHblits 1 iteration, eval<0.001) to at least one annotated PHROG (Terzian *et al.* 2021), leading to the annotation of 13.3% of enVhogs encompassing 39.9% of clustered proteins (18.5M). These results are very much alike those obtained for PHROGs, for which only 13.1% protein clusters are annotated yet representing a large proportion of proteins (50.6%). An additional 245,777 enVhogs were similar to unannotated PHROGs (23.3% of proteins). Among the ten large functional categories of PHROG, the most abundant functions were from the categories “Tail”, “DNA, RNA and nucleotide metabolism” and “Head and packaging” (see Table 2 for details). A total of 175,246 enVhogs were similar to Pfam-A profiles (13.7 M proteins). Out of these, 32,827 concerned enVhogs that had no hit in PHROGs. Leaving aside similarities to PHROGs or PFAMs of unknown function, 349,938 enVhogs were annotated (15.9% of enVhogs containing 44.8 % of proteins).

Annotation classification	#enVhogs	#proteins
unknown function	245,777	10,787,649
tail	78,210	3,284,209
DNA, RNA and nucleotide metabolism	65,269	4,621,857
head and packaging	51,848	4,563,053
other	26,931	2,005,222
transcription regulation	22,091	758,551
moron, auxiliary metabolic gene and host takeover	14,949	949,671
connector	12,148	1,120,256
lysis	12,128	785,524
integration and excision	8,514	376,582

Table 2. ***EnVhog annotation according to similarities to PHROG.*** *Total number of annotated enVhogs and proteins for the ten PhROG functional categories.*

EnVhog improves virome annotation

To quantify Envhog's contribution to annotating new viral proteins, we first collected sets of contigs assembled from recently published viromes from four different ecosystems. For a subsample of 1000 randomly chosen proteins in each of these datasets, we calculated the proportion of proteins similar to three databases: RefSeqVirus (sequence-sequence comparison), PHROG and EnVhog (sequence-HMM comparison). The use of EnVhog doubled both the proportion of sequences with identified homologs and the proportion of functionally annotated proteins, compared with the reference databases RefSeqVirus and PHROG. EnVhog indeed provided functional annotation for 40% to 56% of the proteins in the four new viromes, while RefSeqVirus and PHROG annotated only 19% to 36% and 23% to 35% of the same proteins respectively (Fig. 5A). While EnVhog's improvement over other BDs is the same in

every ecosystem, it is clear that viruses from some ecosystems remain better characterized than others. Unsurprisingly, more than 80% of proteins of viral communities from the human gut and from marine surface waters have homologs in EnVhog, half of the whole protein datasets being annotated (respectively 56 and 51%). In comparison, a particular, little-studied extreme ecosystem like sulfuric soils with a pH lower than three, has only 69% of its proteins characterized and 40% annotated.

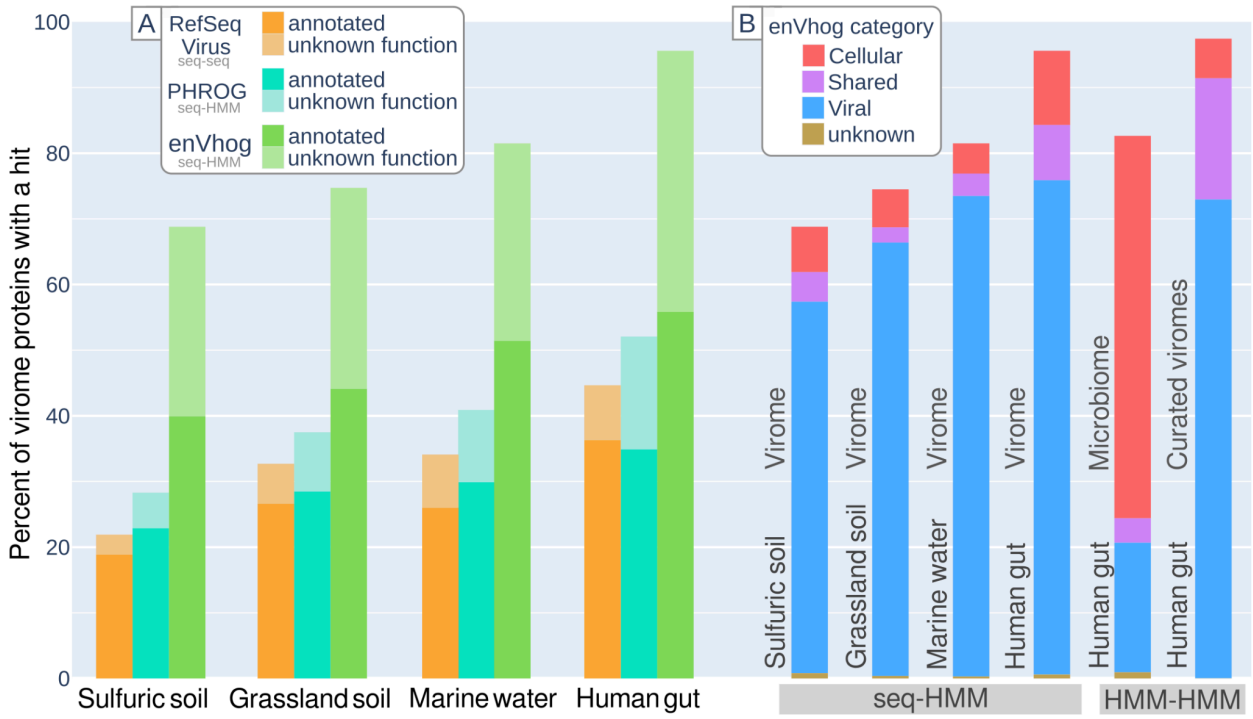

Figure 5: A: Functional annotation of virome. *Percent of characterized and annotated proteins from four sets of contigs from different ecosystems using different databases (RefSeq Virus, PHROG, EnVhog).* ***B: Categorical annotation of metagenomic proteins (V/SC/U).*** *Protein category (Viral/Shared/Cellular/Unknown) breakdown as identified by hits against enVhog on the four selected viromes, a microbiome and a curated set of human gut viruses extracted from metagenomes.*

Based on the same comparison results, the proportion of proteins similar to enVhogs of the four defined categories (Viral, Shared, Cellular, Unknown) was calculated for each of the four viromes (Fig. 5B). Results are comparable for the four viromes : a negligible proportion of Unknown enVhogs, a dominance of Viral enVhogs (>78%), the remainder being made up of Cellular and Shared enVhogs. While this pattern is expected for viromes, the same analysis was also applied to a human gut microbiome (arbitrary chosen sample PSM6XBTR of IBDMDB, Lloyd-Price et al. 2019). First, most proteins from the human gut microbiome are similar to an enVhog (82%) and in more than 70% of the cases, it concerned enVhogs of the Cellular category. This result suggests that most of the proteins encoded by microbes in the human digestive tract have at some point been carried by viral genomes. Second, the proportion of proteins detected as Viral by EnVhog amounts to 20% of the microbiome. As seen previously,

almost all human gut virome proteins are characterized and thus the uncharacterized proteins (18%) of the microbiome are probably mostly of cellular origin.
It should be noted that for the annotation of the latter dataset, its proteins were first clustered, with the resulting cluster HMMs being compared to the EnVhog HMMs. The same methodology was then applied to a set of human intestinal viral genomes (Zolfo et al. 2024). Thanks to this more sensitive strategy, a homolog in EnVhog could be identified for over 97% of the proteins in these genomes.

Discussion

More and more viral proteins have been predicted over the years, both from cultured virus genomes and, even more so, from contigs assembled from metagenomes. In order to analyze and use these millions of proteins efficiently, the first step is to cluster them into families of homologous proteins. Yet, the great sequence heterogeneity of most viral protein families (Terzian et al. 2021) makes homology detection difficult, and requires the use of sensitive similarity search tools. Here, we collected as many viral protein sequences as possible and clustered them with remote homology detection tools.

Moving closer to a comprehensive view of viral genomic sequence space

In addition to viral databases mostly built from sequences extracted from microbial genomes and microbiomes, an effort was made here to identify, collect, assemble and use more than 3,000 viral metagenomes from more than a hundred articles published from 2006 to 2019. As many as 4.4 million “new” proteins were added by these datasets, sampled from a wide range of ecosystems and coming from all over the world. This contribution of more than seven times the number of proteins of the reference viruses present in RefSeqVirus is therefore important for continuing to move closer to a comprehensive view of viral protein sequence space. Yet, the EnVhog database may still suffer from different biases. First, the collected samples, both for viromes or for microbiomes, are usually biased toward anthropocentric environments. Indeed, a larger proportion of protein were characterized in human gut or marine systems compared to soils or extreme environments. This results in a biased representation of the viral diversity on Earth. Second, the biochemistry processes are usually focused on DNA molecules (especially double stranded), RNA viruses being poorly represented in the available raw data. Third, contigs have been considered here as viral if they were predicted as such by at least two different softwares. Of course, this strategy increases accuracy, but at the cost of reduced sensitivity. Considering as viral all contigs predicted by at least one software would have doubled the number of contigs extracted from viromes. By focusing on the quality of the dataset, many contigs of genuine viral origin were undoubtedly excluded from the outset. EnVhog clusters, more representative of the whole viral diversity, could prove helpful to dig back into these datasets in order to identify *bona fide* viral contigs that were here not considered.

EnVhog : a large set of diverse families

Existing databases provide either individual proteins (IMGVR) or protein cluster HMMs (eFAM, PHROG, etc), the latter being generated by standard BLASTp-like comparisons except for the PHROG database. Our three-step clustering procedure, the last step of which uses profile-sequence comparisons, provides better clustering than standard methods. This result

illustrates once again that the diversity of viral families requires sensitive similarity search methods. The HMM-HMM comparisons used to cluster PHROG proteins are even more sensitive, and the detection of even more distantly related homology would make it possible to merge certain enVhogs. Yet, the computational greediness of such comparisons ruled us out of their application on this protein set and the number of viral protein families contained in this dataset is overestimated. Nonetheless, the internal diversity of enVhogs was greater than the one of UniClust30 clusters and was comparable to that of PHROGs. This is probably due to the better representativeness of metagenomic contigs, recovered randomly from a variety of ecosystems, compared with the reference viruses and prophages used in PHROG, isolated from a reduced number of hosts and ecosystems. Thus, enVhog HMMs are internally diverse, even small ones, and allow more efficient similarity searches than existing datasets.

A majority of new viral proteins can be characterized by EnVhog

Due both to the large number of viral protein families represented and to their significant internal diversity, EnVhog is a database very efficient for identifying homologs to proteins encoded in newly discovered genomes. Indeed, more than two-thirds of the new viral proteins have homologs in EnVhog for the four ecosystems tested, a figure that rises to 96% for the highly-studied human digestive tract. Even if whole swathes of viral diversity may remain to be discovered, EnVhog contains a large proportion of the protein families encoded by viruses that are currently recognizable. In addition, the use of EnVhog increased by 54%-74% the proportion of annotated proteins in new genomes compared to existing databases. This improvement can be explained in two complementary ways. Firstly, the enVhogs themselves were annotated in depth using sensitive homology detection techniques (i.e. HMM-HMM comparison with PHROGs and PFAMs). Secondly, the resulting 350K diverse and annotated enVhogs, representing almost half the proteins in our database, proved useful for detecting distant homologies between new proteins and enVhogs (sequence-HMM). An even greater number of new proteins could be annotated when these proteins were themselves organized into homologous groups and compared with enVhogs using HMM-HMM tools. Functional annotation is crucial to the description and understanding of viruses through their genomes, and even if EnVhog represents a major step forward in identifying function encoded in viral proteins, a large fraction of the EnVhogs remains of unknown function. The functions of these enVhogs will have to be elucidated experimentally or using three-dimensional structure predictions and comparisons, a strategy complementary to our sequence-based approach.

Quantifying the viralness of these clusters, even of previously unseen ones

For each protein family, we quantified its conservation in the viral world compared to the cellular world, allowing us to assess the viral origin of about a quarter of the protein clusters. A viral quotient was computed for each enVhog allowing to classify them in Viral, Shared, Cellular and Unknown categories. A majority of enVhogs were originally classified as Unknown, *i.e.* with no similarity to known viruses, illustrating that reference viruses are far from representing an exhaustive view of the viral world. Then, these categories were refined by an iterative cluster-to-contig-to-cluster strategy. First, this strategy allowed us to identify and remove the small proportion of contigs and clusters whose viral origin was unclear and thus provide a database as clean as possible. Second, this strategy reduced the number of clusters initially

classified as Unknown from 75% to 2%, confirming our view that reference viral genomes are poorly representative of real diversity in the environment. In addition, 89% of the EnVhog database is constituted by proteins for which we are confident about their viral origin. EnVhogs classified as Cellular or Shared should not be removed as they represent families more or equally conserved in the cellular than in the viral world (Kristensen et al. 2013). These contain, among other things, auxiliary metabolic genes, and it would be interesting to analyze the extent of the functional diversity occasionally present in viruses on the basis of this dataset. Indeed, most microbiome proteins were similar to Cellular enVhogs, suggesting that most of the proteins encoded by microbes in the human digestive tract have at some point been carried by viral genomes. These categories also highlighted that 20% of the human gut microbiome proteins were of viral origin, a proportion likely to be realistic thanks to enVhog completeness. This proportion, markedly high, could be explained by (i) the significant proportion of prophages in microbial genomes (Roux et al. 2015a), a proportion still probably underestimated, (ii) the presence of free viral particles, both phages and NCLDV, captured and sequenced in microbiomes (Schulz et al. 2020), (iii) the presence of cells infected by viruses in their lytic cycle whose genome number of copies ease them to pass the detection threshold (López-García et al. 2023). These categories are thus useful for discriminating typical cellular and viral proteins, and will help quantify the presence of cellular DNA sequences in metagenomes and or identify of (pro)viruses in cellular (meta)genomes.

Conclusion

The EnVhog database is currently the most up-to-date ressource for viral protein families, covering a very broad spectrum of environments and geographic locations on Earth. Furthermore, the contamination of the database by sequences of cellular origin was further reduced compared to other databases thanks to an iterative cluster-to-contig procedure. Envhog significantly extends previous viral protein databases in terms of diversity and completeness, and provides highly sensitive HMMs for protein annotation. Through this work, we hope to provide a dataset that will help the microbiology and ecology community to identify new viruses and better understand the functioning of newly detected viral genomes.

Data availability

EnVhog clusters as MSA and HMMs, as well as profile database (MMseqs2 format), are available online for user's download at http://envhog.u-ga.fr/envhog/

In addition, information on contigs, proteins, intermediate clusters and enVhogs is available on the same website via various tables in tsv format :

- *Envhog_infos_final.tsv*: for each EnVhog cluster, a description of size, maximum protein length, diversity, functional annotations, viralness, hostess, viral quotient and category V/S/C/U as identified by our iterative algorithm.
- *Prot_dedup_shallow_standard_cluster_envhog_contig_final.tsv*: for each original protein, a list of its : (i) the original protein, (ii) the deduplicated protein group, shallow, standard and deep clusters and enVhog to which it belongs and (iv) its original contig.
- *enVhog_classes.tsv.gz*: for each enVhog, its categories V/S/C/U at each iteration of our iterative algorithm

- *Contig_infos_class_final.tsv*: for each contig, the contig identifier, its accession number in the original dataset, figures on the contig (length, number of proteins) and information on its taxonomy.
- *enVhog_contigs_classes.tsv.gz*: for each contig, its categories V/S/C/U at each iteration of our iterative algorithm

Acknowledgements

The computations presented in this paper were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr), which is supported by Grenoble research communities, and using the Froggy platform of the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr), supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche.

Funding

This work was developed in the framework of Grenoble Alpes Data Institute, supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02). This work was supported by the European Research Council's Horizon 2020 Framework Programme for Research and Innovation ("Virus-X", project no. 685778).

Supplementary materials

- *Supplementary tables:*

Table S1. Description of the 3,257 collected viromes.

Table S2. Dataset summary: *Neary 130 million proteins were selected from multiple resources. All proteins were cleaned to remove sequences with intermediate stop codons as well as sequences with 5 or more unknown aminoacids.*

Table S3. *Numbers and proportions of EnVhogs and proteins with different thresholds of Vq index.*

- Supplementary figures and supplementary information: see EnVhog_supp_mat external files.

References

Bi L, He JZ, Hu HW. Total metagenomes outperform viromes in recovering viral diversity from sulfuric soils. ISME Commun. 2024 Jan 27;4(1):ycae017. doi: 10.1093/ismeco/ycae017. PMID: 38481578; PMCID: PMC10936692.

Camargo AP, Nayfach S, Chen IA, Palaniappan K, Ratner A, Chu K, Ritter SJ, Reddy TBK, Mukherjee S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC, Roux S. 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(D1):D733-D743. doi: 10.1093/nar/gkac1037. PMID: 36399502; PMCID: PMC9825611.

Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F. Viruses as winners in the game of life. Annu Rev Virol. 2016;3:197–214. https://doi.org/10.1146/annurev-virology-100114-054952.

Coutinho FH, Edwards RA, Rodríguez-Valera F. Charting the diversity of uncultured viruses of Archaea and Bacteria. BMC Biol. 2019 Dec 29;17(1):109. doi: 10.1186/s12915-019-0723-8. PMID: 31884971; PMCID: PMC6936153.

Garmaeva S, Sinha T, Gulyaeva A, Kuzub N, Spreckels JE, Andreu-Sánchez S, Gacesa R, Vich Vila A, Brushett S, Kruk M; Lifelines NEXT cohort study; Dekens J, Sikkema J, Kuipers F, Shkoporov AN, Hill C, Scherjon S, Wijmenga C, Fu J, Kurilshikov A, Zhernakova A. Transmission and dynamics of mother-infant gut viruses during pregnancy and early life. Nat Commun. 2024 Mar 2;15(1):1945. doi: 10.1038/s41467-024-45257-4. PMID: 38431663; PMCID: PMC10908809.

Gauthier CH, Cresawn SG, Hatfull GF. PhaMMseqs: a new pipeline for constructing phage gene phamilies using MMseqs2. G3 (Bethesda). 2022 Nov 4;12(11):jkac233. doi: 10.1093/g3journal/jkac233. PMID: 36161315; PMCID: PMC9635663.

Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017 Jan 4;45(D1):D491-D498. doi: 10.1093/nar/gkw975. Epub 2016 Oct 26. PMID: 27789703; PMCID: PMC5210652.

Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB, Roux S. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021 Feb 1;9(1):37. doi: 10.1186/s40168-020-00990-y. PMID: 33522966; PMCID: PMC7852108.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar 8;11:119. doi: 10.1186/1471-2105-11-119. PMID: 20211023; PMCID: PMC2848648.

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 Jun 10;8(1):90. doi: 10.1186/s40168-020-00867-0. PMID: 32522236; PMCID: PMC7288430.

Koonin EV, Krupovic M, Dolja VV. The global virome: How much diversity and how many independent origins? Environ Microbiol. 2023 Jan;25(1):40-44. doi: 10.1111/1462-2920.16207. Epub 2022 Sep 29. PMID: 36097140.

Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV. Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. J Bacteriol. 2013. 195(5):941-50. doi: 10.1128/JB.01801-12. PMID: 23222723; PMCID: PMC3571318.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019). https://doi.org/10.1038/s41586-019-1237-9

López-García P, Gutiérrez-Preciado A, Krupovic M, Ciobanu M, Deschamps P, Jardillier L, López-Pérez M, Rodríguez-Valera F, Moreira D. Metagenome-derived virus-microbe ratios across ecosystems. ISME J. 2023 Oct;17(10):1552-1563. doi: 10.1038/s41396-023-01431-y. Epub 2023 May 11. PMID: 37169871; PMCID: PMC10504350.

Meier A, Söding J. Context similarity scoring improves protein sequence alignments in the midnight zone. Bioinformatics. 2015 Mar 1;31(5):674-81. doi: 10.1093/bioinformatics/btu697. PMID: 25338715.

Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics. 2019 Aug 15;35(16):2856-2858. doi: 10.1093/bioinformatics/bty1057. PMID: 30615063; PMCID: PMC6691333.

Mirdita M., von den Driesch L., Galiez C., Martin M. J., Söding J., and Steinegger M., Uniclust databases of clustered and deeply annotated protein sequences and alignments, *Nucleic Acids Res.* 2016.

Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021 Sep 29;37(18):3029-3031. doi: 10.1093/bioinformatics/btab184. PMID: 33734313; PMCID: PMC8479651.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D412-D419. doi: 10.1093/nar/gkaa913. PMID: 33125078; PMCID: PMC7779014.

Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, Gillet R, Forterre P, Krupovic M. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun. 2019 Feb 14;10(1):752. doi: 10.1038/s41467-019-08672-6. PMID: 30765709; PMCID: PMC6375957.

Mushegian AR. Are There 10^{31} Virus Particles on Earth, or More, or Fewer? J Bacteriol. 2020 Apr 9;202(9):e00052-20. doi: 10.1128/JB.00052-20. PMID: 32071093; PMCID: PMC7148134.

Ni Y, Chu T, Yan S, Wang Y. Forty-nine metagenomic-assembled genomes from an aquatic virome expand *Caudoviricetes* by 45 potential new families and the newly uncovered Gossevirus of *Bamfordvirae*. J Gen Virol. 2024 Mar;105(3). doi: 10.1099/jgv.0.001967. PMID: 38446011.

Olo Ndela E, Enault F, Toussaint A. Transposable Prophages in Leptospira: An Ancient, Now Diverse, Group Predominant in Causative Agents of Weil's Disease. Int J Mol Sci. 2021 Dec 14;22(24):13434. doi: 10.3390/ijms222413434. PMID: 34948244; PMCID: PMC8705779.

Olo Ndela E, Roux S, Henke C, Sczyrba A, Sime Ngando T, Varsani A, Enault F. Reekeekee- and roodoodooviruses, two different Microviridae clades constituted by the smallest DNA phages. Virus Evol. 2022. 9(1):veac123. doi: 10.1093/ve/veac123. PMID: 36694818; PMCID: PMC9865509.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. Uncovering Earth's virome. Nature. 2016 Aug 25;536(7617):425-30. doi: 10.1038/nature19094. Epub 2016 Aug 17. PMID: 27533034.

Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020 Mar;8(1):64-77. doi: 10.1007/s40484-019-0187-4. PMID: 34084563; PMCID: PMC8172088.

Rosario K, Breitbart M. Exploring the viral world through metagenomics. Curr Opin Virol. 2011. 1(4):289-97. doi: 10.1016/j.coviro.2011.06.004. PMID: 22440785.

Roux S, Krupovic M, Debroas D, Forterre P, Enault F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. 2013 Dec 11;3(12):130160. doi: 10.1098/rsob.130160. PMID: 24335607; PMCID: PMC3877843.

Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015a May 28;3:e985. doi: 10.7717/peerj.985. PMID: 26038737; PMCID: PMC4451026.

Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife. 2015b Jul 22;4:e08490. doi: 10.7554/eLife.08490. PMID: 26200428; PMCID: PMC4533152.

Santos-Medellín C, Blazewicz SJ, Pett-Ridge J, Firestone MK, Emerson JB. Viral but not bacterial community successional patterns reflect extreme turnover shortly after rewetting dry soils. Nat Ecol Evol. 2023 Nov;7(11):1809-1822. doi: 10.1038/s41559-023-02207-5. Epub 2023 Sep 28. PMID: 37770548.

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. Giant virus diversity and host interactions through global metagenomics. Nature. 2020 Feb;578(7795):432-436. doi: 10.1038/s41586-020-1957-x. Epub 2020 Jan 22. PMID: 31968354; PMCID: PMC7162819.

Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005. 21(7):951-60. doi: 10.1093/bioinformatics/bti125.

Steinegger M and Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, doi: 10.1038/nbt.3988 (2017).

Steinegger M and Söding J. Clustering huge protein sequence sets in linear time. Nature Communications, doi: 10.1038/s41467-018-04964-5 (2018).

Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit MA, Enault F. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021 Aug 5;3(3):lqab067. doi: 10.1093/nargab/lqab067. PMID: 34377978; PMCID: PMC8341000.

Wieczynski DJ, Yoshimura KM, Denison ER, Geisen S, DeBruyn JM, Shaw AJ, Weston DJ, Pelletier DA, Wilhelm SW, Gibert JP. Viral infections likely mediate microbial controls on ecosystem responses to global warming. FEMS Microbiol Ecol. 2023 Feb 28;99(3):fiad016. doi: 10.1093/femsec/fiad016. PMID: 36828391.

Zayed AA, Lücking D, Mohssen M, Cronin D, Bolduc B, Gregory AC, Hargreaves KR, Piehowski PD, White RA, Huang EL, Adkins JN, Roux S, Moraru C, Sullivan MB. efam: an expanded, metaproteome-supported HMM profile database of viral protein families. Bioinformatics. 2021 Jun 16;37(22):4202–8. doi: 10.1093/bioinformatics/btab451. Epub ahead of print. PMID: 34132786; PMCID: PMC9502166.

Zolfo M, Silverj A, Blanco-Míguez A, Manghi P, Rota-Stabelli O, Heidrich V, Jensen J, Maharjan S, Franzosa E, Menni C, Visconti A, Pinto F, Ciciani M, Huttenhower C, Cereseto A, Asnicar F, Kitano H, Yamada T, Segata N. Discovering and exploring the hidden diversity of human gut viruses using highly enriched virome samples. bioRxiv. 2024 19:2024.02.19.580813. doi: 10.1101/2024.02.19.580813. PMID: 38464031; PMCID: PMC10925137.