Charting the diversity of Uncultured Viruses of Archaea and Bacteria

Coutinho, FH¹; Rodríguez-Valera F¹

Author affiliations
1 – Evolutionary Genomics Group, Departamento de Produccion Vegetal y Microbiología, Universidad Miguel Hernández, Campus San Juan, San Juan, Alicante 03550, Spain
* Address correspondence to: fhernandes@umh.es

Abstract
Viruses of Archaea and Bacteria are among the most abundant and diverse biological entities on Earth. Unraveling their biodiversity has been challenging due to methodological limitations. Recent advances in culture-independent techniques, such as metagenomics, shed light on viral dark matter, revealing thousands of new viral genomes at an unprecedented scale. However, these novel genomes have not been properly classified and the evolutionary associations between them were not resolved. Here, we performed phylogenomic analysis of nearly 200,000 viral genomic sequences to establish GL-UVAB: Genomic Lineages of Uncultured Viruses of Archaea and Bacteria. GL-UVAB yielded a 44-fold increase in the amount of classified genomes. The pan-genome content of the identified lineages revealed their infection strategies, potential to modulate host physiology and mechanisms to escape resistance systems. Furthermore, using GL-UVAB for annotating metagenomes from multiple ecosystems revealed elusive habitat distribution patterns of viral communities. These findings expand the understanding of the diversity, evolution and ecology of viruses of prokaryotes.

Introduction
Grasping the biodiversity of viruses of Bacteria and Archaea has been a major challenge within the field of virology. Limitations for viral cultivation and purification associated with the absence of universal marker genes have been major drawbacks in the effort to chart and classify the biodiversity of these viruses[1,2]. The taxonomic classification system established for viruses of Bacteria and Archaea was originally based on morphological traits, but genetic studies demonstrated that the major taxa established through this approach are not monophyletic[3–5]. Thus, viral classification and taxonomy has come to rely heavily on comparative genomics. This shift has led the International Committee for the Taxonomy of Viruses (ICTV) to call for a scalable genome-based classification systems that can also be applied to uncultured viruses for which no phenotypic data is available[6].

Phylogenomic trees and genomic similarity networks incorporate full genomic data for comparison and clustering of viral genomes. Both phylogenomic and network based approaches have showed promising results for reconstructing phylogenies and classifying and identifying novel viral taxa[1,5,7]. These approaches circumvent the biases and limitations associated with morphological data or the use of phylogenetic markers, and are easily scalable to thousands of genomes[5,8]. Network methods rely on the identification of orthologous groups shared among genomes, which can be problematic for viruses due to the ratio in which their genes evolve. Additionally, the evolutionary associations among genome clusters identified by network approaches are not explicitly resolved by these methods[5,9]. Meanwhile, phylogenomic approaches provide trees in which the associations among genomes are easily interpretable under an evolutionary perspective. For these reasons, phylogenomic methods have been the standard approach for reconstructing phylogenies of prokaryotic viruses[1,4,7,8,10–14]. Previous studies have leveraged this method to investigate the genetic diversity of cultured viruses, but overlooked the uncultured diversity.

Thousands of novel viral genomes were recently discovered through culture-independent approaches, such as shotgun metagenomics, fosmid libraries, single-virus sequencing and prophage mining[4,11,15–18]. These new datasets unraveled an extensive biodiversity that had been overlooked by culture-based approaches. These genomes have the potential to fill many of the gaps in our understanding of the diversity of viruses of prokaryotes. Yet, achieving this goal requires that these genomes are properly organized in a robust evolutionary framework. Here, we applied a phylogenomic approach
to chart the diversity of uncultured viruses of Bacteria and Archaea aiming to gain insights on their genetic diversity, evolution and ecology.

**Results**

**Phylogenomic reconstruction**

A database of genomic sequences of prokaryotic viruses was compiled with isolated dsDNA viruses from RefSeq and uncultured viruses that were discovered across multiple ecosystems using approaches that bypassed culturing. This database amounted to 195,698 viral genomic sequences along with associated information of computational host predictions and ecosystem source (Table S1). Following pre-filtering and redundancy removal steps, a subset of 16,484 genomic sequences was selected for phylogenomic reconstruction (Methods and Figure S1). An all-versus-all comparison of the protein sequences encoded in this dataset was performed and used to calculate Dice distances between genomes. Essentially, the Dice distances between a pair of genomes decreases the more proteins that are shared between them and the higher their degree of identity. Finally, the obtained matrix of Dice distances was used to construct a phylogenomic tree through neighbor-joining (Figure 1). In addition, a benchmarking dataset containing a subset of 2,069 dsDNA prokaryotic viral genomes from NCBI RefSeq was analyzed in parallel for comparison and validation of results. The robustness of the tree topology was evaluated through a sub-sampling approach. Nodes displayed high confidence values (average 50% recovery), and 90% of all nodes were recovered at least once among the re-sampled trees. These figures were obtained when reducing the data used to calculate distances to approximately 64% of the amount used to establish the original tree, demonstrating that tree topology is robust even in the presence of incomplete or fragmented genomes, which might be the case for some of the uncultured viral genomes used.

**Clustering uncultured prokaryotic viruses into closely related lineages**

A three-step approach was applied to categorize diversity into hierarchical levels of increasing genomic relatedness: Level-1 (average Dice distances between genomes equal or below 0.925, and number of representatives equal or above 50), Level-2 (average Dice distances between genomes equal or below 0.825, and number of representatives equal or above 10) and Level-3 (average Dice distances between genomes equal or below 0.5, and number of representatives equal or above 3). These cutoffs were selected to establish
lineages at a degree of genomic similarity within the range observed for families, sub-
families and genera established by the ICTV (Table S4). At the first level, 10,357 genomic
sequences were assigned to 72 lineages (Figure 1). Of these, 68 of the Level-1 lineages
have at least one complete viral genome assigned to them. In addition, 50 of these Level-1
lineages had at least one reference viral genome assigned to them. At the second level,
10,356 sequences were assigned to 314 lineages, while at the third level, 6,971
sequences were assigned to 942 lineages. This three-level classification system was used
to establish the GL-UVAB. Apart from 6 cases, all of the Level-1 lineages are composed of
genomes assigned to a single taxonomic family as defined by the ICTV. Out of the 942
Level-3 lineages, 104 of them included genomes classified at the level of genus by the
ICTV. Of these, 74 lineages are consistent regarding the genus, meaning that all classified
genomes are assigned to the same genera. Meanwhile, only 6 genera were split among
more than one Level-3 lineage. Thus, the identified lineages of uncultured viruses are in
agreement with the ICTV established taxonomy.

Genome sequences that were not included in the phylogenomic reconstruction
were assigned to the lineage of their closest relatives as determined by the average amino
acid identity and percentage of shared genes. Following this lineage expansion step, a
total of 100,907 sequences were classified to at least one level (Table S1), which
represents a 44-fold improvement in the proportion of classified sequences compared to
the amount of RefSeq prokaryotic viral genomes classified by the NCBI taxonomy
database. The dataset that contributed most (~60%) of the classified sequences was from
a cross-ecosystem global analysis of metagenomes[16], followed by metagenomes from
the Tara and Malaspina expeditions [17] (~15%), global marine viromes[15] (~10%) and
prophages identified in bacterial genomes[18] (~10%) (Figure S3).

Targeted hosts and ecosystem sources of GL-UVAB lineages

GL-UVAB lineages differed regarding host prevalence (Figure 2A). Out of the 72
Level-1 lineages, 34 are predicted to infect a single host phylum, most often
Proteobacteria, Firmicutes or Actinobacteria, while 35 lineages are predicted to infect two
or more phyla. Level-3 lineages display the highest levels of host consistency. Among
Level-3 lineages with at least one annotated host, 99% of them are predicted to infect a
single phylum and 76% are predicted to infect a single genus. Lineages also differed
regarding the ecosystem sources from where their members were obtained (Figure 2B).
Nearly all lineages contained members obtained from multiple ecosystems but aquatic and
human-associated samples were consistently the main sources of genomes, highlighting diversity within these habitats and the potential for discoveries in other under-explored ecosystems. Trends of host and ecosystem prevalence observed for the expanded lineages (Figure S4) were consistent with those obtained from the analysis of the original dataset, corroborating the validity of these patterns. Applying the same lineage identification criteria to the benchmarking dataset of RefSeq viral genomes identified 12 Level-1 lineages, 49 Level-2 lineages and 162 Level-3 lineages (Figure S2). Among these Level-3 lineages, 145 (91%) are composed of genomes that infect within the same host order. This observation further corroborates the validity of the associations between GL-UVAB lineages and targeted hosts.

GL-UVAB lineages differ in habitat distribution and pan-genome content

The observed differences in host preference and ecosystem source among lineages led us to investigate the applicability of GL-UVAB as a reference database for deriving abundance profiles from metagenomes. We analyzed the abundances of 72 GL-UVAB Level-1 lineages across metagenomes from marine, freshwater, soil and human gut samples (Figure 3). Lineages 28, 21, and 36 were the most abundant in marine samples, in agreement with the high prevalence of *Cyanobacteria* and *Proteobacteria* as hosts of these lineages (Figure 2A). Meanwhile, the lineages 64 (which mostly infects *Proteobacteria* of classes Gamma and Delta), 23 (*Proteobacteria* and *Bacteroidetes*), and 21 where the dominant groups in freshwater habitats. In temperate soil samples, the most abundant lineages were 14 (*Actinobacteria*), 38 (*Beta- and Gamma- Proteobacteria*) and 62 (*Gammaproteobacteria*). Finally, human gut samples were dominated by lineages 64, 30 (*Firmicutes and Bacteroidetes*) and 10 (*Firmicutes* and *Actinobacteria*).

We inspected the pan-genome of the identified lineages by clustering their protein encoding genes into orthologous groups (OGs). A total of 79,281 OGs containing at least three proteins were identified. These OGs displayed a sparse distribution, i.e., were only detected in a small fraction of genomes within lineages (Table S3). The most conserved OGs encoded functions associated with nucleic acid metabolism and viral particle assembly. Few OGs encoded putative auxiliary metabolic genes (AMGs), and those where never shared by all the members of a lineage. A total of 1,755 promiscuous OGs, present in the pan-genome of three or more Level-1 lineages were identified.

**Discussion**
Only a small fraction of prokaryotic viruses can be cultivated through currently available laboratory techniques. This limitation has left many gaps in our understanding of their biodiversity. The results presented here help to bridge these gaps by leveraging on a large dataset of viral genomic sequences obtained without cultivation from multiple ecosystems. Categorizing this new diversity into a robust evolutionary framework brings us closer to properly grasping the extension of viral biodiversity.

The average dice distance cutoffs used for defining lineages were chosen to classify as many genomes as possible while maintaining cohesiveness within lineages regarding similarity between genomes, targeted hosts and taxonomic classification as defined by the ICTV. These goals were achieved, as the GL-UVAB lineages are formed by groups of closely related genomes (Table S4), which was reflected in their targeted hosts (Figure 2A), pan-genome content (Table S2) and taxonomic classification (Table S5). Using different cutoffs for lineage identification would have resulted in the distinct lineages and consequently different patterns of host and source prevalence, pan-genome content, and habitat distribution. Nevertheless, GL-UVAB was conceived to be an evolving system. We encourage researchers to adapt the GL-UVAB approach to suit the needs of the specific questions under investigation. For example, performing species-level clustering genomes would require average Dice distance cutoffs even lower than those used to delineate Level-3 lineages.

The re-sampling approach showed that the tree topology was robust even in the presence of fragmented genomes. Still, incomplete genomes are more likely to be misclassified by our pipeline. Yet, in the absence of complete genomes, these fragments are the best available option for obtaining at least a provisional classification of uncultured viruses. As the quality and amount of uncultured viral genomes discovered increases, new data can be used to update and improve the GL-UVAB.

The consistency of the targeted hosts among lineages identified with our phylogenomic approach suggests that the assignment to GL-UVAB lineages provides a rough estimate of the hosts of uncultured viruses. This is of fundamental importance, considering the growing diversity of viral genomes discovered from metagenomic datasets for which no host information is initially available[19,20]. Host prevalence analysis indicated that approximately half of the Level-1 lineages are capable of infecting more than a single host phylum (Figure 2A). The ability to interact with the molecular machinery of the host is a major driver of the evolution of prokaryotic viruses. Thus, closely related genomes (that belong to the same lineages) likely have undergone similar evolutionary
pressures that ensure host infectivity, leading to the observed pattern of higher host consistency among the lowest level of hierarchical classification, i.e., Level-3 lineages. Meanwhile, the ability of some lineages to infect across multiple host phyla is likely an indication of the high level of genomic plasticity of viruses that allows them to evolve to infect new organisms that are not closely related to their original hosts.

Metagenomic studies of viral ecology strive to elucidate the habitat distribution patterns of taxa across ecosystems. The abundance patterns observed for the GL-UVAB lineages (Figure 3) are a reflection of their distinctive trends of host prevalence (Figure 2A). As expected, the GL-UVAB lineages that dominated at each ecosystem often targeted taxa that are the most abundant at these habitats[21,22], e.g., lineages that target *Proteobacteria* and *Cyanobacteria* at aquatic samples and lineages that target *Bacteroidetes* and *Firmicutes* in the human gut. Although this observation might seem obvious, it does not emerge when using cultured viral genomes for the taxonomic annotation of metagenomes. Instead, the same taxa are often observed with similar abundance patterns regardless of the ecosystem sampled. This occurs because established taxa have no discernible host or ecosystem preferences and because much of viral diversity is not encompassed by viral taxonomy[14,23,24]. Thus, the cohesiveness of GL-UVAB lineages regarding phylogeny, host preference and ecology allows for meaningful habitat-taxa associations to be observed.

A detailed investigation of the pan-genome content of the Level-1 lineage 21, revealed some of the strategies applied by these viruses during infection. This lineage was among the dominant group in both freshwater and marine samples and infects *Cyanobacteria* and *Proteobacteria*. The pan-genome of lineage 21 includes OGs encoding high-light inducible proteins, photosystem II D1 proteins and a transaldolase. These proteins are involved in photosynthesis and carbon fixation pathways[25]. Therefore, the success of this group across aquatic ecosystems might be linked to their capacity to use such proteins as AMGs to modulate the metabolism of their Cyanobacterial hosts during infection, redirecting it to the synthesis of building blocks to be used for the assembly of novel viral particles[25].

The promiscuous distribution observed for multiple OGs could be the result of the positive selection of these genes following events of horizontal gene transfer. Indeed, promiscuous OGs often encoded proteins that might confer advantages during infection. Five of them encode thymidylate synthase, a protein involved in nucleotide synthesis. Meanwhile, three promiscuous OGs encode the *PhoH* protein, which mediates...
phosphorus acquisition in nutrient deprived conditions. These findings suggest a selective pressure favoring the acquisition of genes that allow viruses to modulate host metabolism towards the production of nucleic acids to be used for the synthesis of progeny DNA[25]. Nevertheless, the sparse distribution of these and other auxiliary metabolic genes across lineages suggest that such genes have been recently acquired by the members of each lineage through horizontal gene transfer (HGT), rather than being present in the common ancestor of the group. Multiple methylases were identified among promiscuous OGs. Viruses use these proteins to protect their DNA from host restriction modification systems[26]. Prokaryotes can acquire restriction modification systems through HGT[27], and our data suggest that viruses also benefit from HGT by acquiring novel methylases that allow them to escape these systems. Finally, lysins (e.g., peptidases and amidases) were a common function among promiscuous OGs. This finding is surprising because lysins are believed to be fine-tuned for the specific structure of host cell wall[28,29]. Acquisition of novel lysins might help viruses to expand their host spectra or as a mechanism to ensure infectivity following the emergence of resistance mutations that lead to alterations in the structure of the host cell wall.

In conclusion, by analyzing thousands of uncultured viral genomes we were able to categorize the diversity of these biological entities. This was achieved by identifying lineages of uncultured viruses through a robust and scalable phylogenomic approach. Analyzing host and source prevalence, pan-genome content and abundance in metagenomes painted a more accurate picture of viral biodiversity across ecosystems. As these uncultured viruses are isolated and their morphology and host spectra are elucidated, they will be properly integrated into the ICTV classification system. Here we provided an initial framework for their taxonomic classification as well as insights regarding their ecology and genetic diversity. Future studies will continue to shed light on the viral dark matter across our planet’s many ecosystems. Our work provides the initial steps for a genome-based classification of these yet undiscovered evolutionary lineages, providing a solid framework to investigate the biology of prokaryotic viruses.

**Methods**

**Viral genome database**

The NCBI RefSeq dataset was used as a starting set of reference viral genome sequences. Host information for these sequences was retrieved from GenBank files, and their taxonomic classification was obtained both from the NCBI Taxonomy database and
from the ICTV [30]. Additionally, genomic sequences were compiled from studies that used high throughput approaches to obtain viral genomes through culture-independent analysis, along with their associated host and ecosystem source information whenever available. These genomes were obtained from: environmental metagenomes and metaviromes[3,12,15–17,31,32], fosmid libraries of Mediterranean viruses[4,11], single virus genomes[33] and prophages integrated into prokaryotic genomes[18]. This dataset (henceforth referred to as Vir_DB_Nuc) contained a total of 195,698 viral genome sequences (Table S1 and Supplementary file 1). Protein encoding genes (PEGs) were predicted from Vir_DB_Nuc using the metagenomic mode of Prodigal[34], which identified 4,332,223 protein sequences (henceforth referred to as Vir_DB_Prot, Supplementary file 2). The Vir_DB_Prot dataset was queried against the NCBI-nr protein database using Diamond[35] for taxonomic and functional annotation.

Sequence pre-filtering

Identifying viral sequences within metagenomic and metaviromic datasets can be problematic. Because each study used different strategies to achieve that goal we pre-filtered genome sequences from Vir_DB_Nuc to ensure that only bona fide prokaryotic viral genomes were included in downstream analyses. First, the Vir_DB_Prot dataset was queried against the prokaryotic virus orthologous groups (pVOGs)[36] protein database using diamond[35] (more sensitive mode, BLOSUM45 matrix, identity ≥ 30%, bitscore ≥ 50, alignment length ≥ 30 amino acids and e-value ≤ 0.01). For each sequence, the percentage of proteins mapped to the pVOGs database and the added viral quotient (AVQ) were calculated (as the sum of the individual viral quotient of the best hit of each protein). Hits to pVOGs that presented homology with proteins from Eukaryotic viruses were not considered. Sequences with 20% or more of the proteins mapped to the pVOGs database and with an AVQ equal to or greater than 5 were classified as bona fide prokaryotic viral genomes. This initial round of recruitment yielded 26,610 genomic sequences (Vir_DB_Nuc_R1). Next, proteins from the Vir_DB_Nuc_R1 dataset were used as bait for a second recruitment round. The remaining protein sequences (which were not recruited in the first round) were queried against Vir_DB_Nuc_R1 through diamond as described above. Genomic sequences from which at least 20% of the derived proteins mapped to a single genome from Vir_DB_Nuc_R1, yielding a minimum of three hits, were recruited to Vir_DB_Nuc_R2 (78,295 genomic sequences). Finally, a last step of manual curation was performed, which recruited mostly long contigs with high AVQ that did not
match the percentage criteria of the automatic recruiting steps due to their high number of
coded proteins. This step recruited a total of 6,421 genomic sequences
(Vir_DB_Nuc_R3).

We benchmarked the accuracy of the automatic recruiting steps with two datasets.
First, a subset of Vir_DB_Nuc comprised of all the reference viral genomes was run
through the recruitment pipeline using the same criteria described above. None of the
7,036 eukaryotic viruses were recruited by the pipeline (i.e., 100% precision) and 2,136
out of 2,297 prokaryotic viruses were correctly recruited (i.e., 92.99% recall). We also
benchmarked the filtering pipeline with a dataset of 897 Gbp of genome sequence data
derived from the NCBI RefSeq prokaryote genomes spanning 880 genera from 35 phyla.
Sequences were split into fragments of 5, 10, 15, 20, 25, 50 and 100 Kbp to mimic
metagenomic scaffolds. Using the filtering criteria described above and the subsequent
length filtering for sequences longer than 30 Kbp would recruit only 109 sequences
(0.36%), all of which displayed homology to the prophage sequences described by Roux
et al. [18].

Phylogenomic reconstruction

Phylogenomic reconstruction was performed using a subset of genomes from
Vir_DB_Nuc that included all dsDNA RefSeq viral genomes annotated as complete for
which the host Domain was either Bacteria or Archaea and the uncultured bona fide
prokaryotic viruses from Vir_DB_Nuc_R3 with a length equal or greater than 30 Kbp or
annotated as complete and with a length equal or greater than 10 Kbp. These criteria were
established to ensure any incomplete genomes fragments encompassed a significant
fraction of the complete genome (considering the 47 Kbp median size of complete
genomes from RefSeq viruses of Bacteria and Archaea). Genome sequences were
clustered with CD-HIT[37] using a cut-off of 95% nucleotide identity and minimum 50%
coverage of the shorter sequence to remove redundant sequences. The non-redundant
dataset contained 16,484 viral genomic sequences that were used for phylogenomic
reconstruction (Vir_DB_Phy). Distances between genomes were calculated based on a
modified version of the Dice method[4]. First, an all-versus-all comparison of the PEGs
derived from the Vir_DB_Phy dataset was performed through diamond[35] (more sensitive
mode, identity ≥ 30%, bitscore ≥ 30, alignment length ≥ 30 amino acids and e-value ≤
0.01). Next, distances between genome sequences were calculated as follows: $D_{AB} = 1 -
\left(2*(AB) / (AA+BB)\right)$, where AB is the bitscore sum of all the valid hits of genome A against
genome B, while AA and BB are the bitscore sum of all the valid hits of genome A against itself and of all the valid hits of genome B against itself, respectively. The more homologous proteins are shared between genomes A and B, and the higher the degree of identity between these homologous proteins, the closer to zero the value of $D_{AB}$ will be. Non homologous proteins should produce no hits when comparing genomes A against B, but will hit with themselves when comparing A against A and B against B. Therefore, when estimating $D_{AB}$, non homologous proteins are penalized, increasing the value of $D_{AB}$. The obtained Dice distances matrix was used as input to build phylogenetic tree through Neighbor-Joining algorithm[38] implemented in the Phangorn package of R. The obtained tree was midpoint rooted.

**Tree topology validation by re-sampling**

A re-sampling approach was applied to test the consistency of the tree topology. First, 20% of the proteins encoded in the genomes used to build the tree were randomly selected. Then, distances between genomes were re-calculated after excluding any hits from the all-versus-all search in which either the query or subject sequences were selected for exclusion, which removes approximately 36% of all of the original hits. Finally, the obtained distance matrix was used to construct a new tree. This process was repeated over one hundred iterations. Next, we measured the frequency in which the nodes from the original tree were present in the re-sampled trees.

**Lineage identification**

Lineage identification was performed by parsing the phylogenetic tree to identify monophyletic clades that matched the established criteria for maximum average Dice distances between genomes, and for a minimum number of representatives. Lineages were identified in three steps, aimed at capturing diversity into levels of increasing genomic relatedness: Level-1 (average Dice distances between genomes equal or below 0.925, and number of representatives equal or above 50), Level-2 (average Dice distances between genomes equal or below 0.825, and number of representatives equal or above 10) and Level-3 (average Dice distances between genomes equal or below 0.5, and number of representatives equal or above 3). These cutoffs were selected based on the degree of genomic similarity observed for taxa of prokaryotic viruses described by the ICTV and NCBI Taxonomy databases (Table S4), with the goal of classifying as many sequences as possible while maximizing consistency within clades regarding pan-genome.
content and agreement with the currently accepted viral taxonomy. To trace the pan-genomes of the identified lineages the proteins derived from 16,484 genomes used for phylogenomic reconstruction were clustered into orthologous groups using the orthoMCL algorithm[39] implemented in the Get_Homologues pipeline[40]. The MCL inflation factor was set to 1 and all other parameters were set to default.

Lineage expansion by closest relative identification

Sequences that did not pass the initial length and redundancy filters to be included in the phylogenomic tree were assigned to the lineages of their closest relatives. Closest relatives were defined as the genome with highest percentage of matched protein encoding genes (PEGs) as detected by Diamond searches. Potential ties were resolved by choosing the closest relative with the highest average amino acid identity (AAI) value. A minimum AAI of 50% and the percentage of matched PEGs of 50% was required for closest relative assignments.

Lineage benchmarking and validation with RefSeq prokaryotic viruses

We sought to validate our strategy for lineage identification. To that end a dataset of 2,069 genome sequences of dsDNA viruses of Archaea and Bacteria from the NCBI RefSeq database was used for benchmarking. The steps for distance calculation, tree construction and lineage identification were performed exactly as described for the full dataset. Next, we calculated statistics of distances between taxa established by ICTV and NCBI (Table S4).

Lineage abundance in metaviromes and metagenomes

The abundances of Vir_DB_Nuc sequences were estimated in viral metagenomes (viromes) from the following ecosystems: marine epipelagic samples[41], healthy human gut[42], freshwater lakes[43], and because no large scale viromes of mesophilic soils were available we used cellular metagenomes from this ecosystem[44,45]. Sequencing reads from these metagenomes and metaviromes were retrieved from the European Nucleotide Archive or NCBI Short Read Archive. Subsets of twenty million R1 reads from each sample were mapped to Vir_DB_Nuc using Bowtie2[46] using the sensitive-local alignment mode. Lineage abundances across samples were calculated by summing the relative abundances of individual genomes according to their assigned lineages.
Acknowledgements:

FHC and FRV were supported by the grants “VIREVO” CGL2016-76273-P [AEI/FEDER, EU], (co-funded with FEDER funds); Acciones de dinamización “REDES DE EXCELENCIA” CONSOLIDER- CGL2015-71523-REDC from the Spanish Ministério de Economía, Industria y Competitividad and PROMETEO II/2014/012 “AQUAMET” from Generalitat Valenciana. FHC was supported by APOSTD/2018/186 Post-doctoral fellowship from Generalitat Valenciana.

Author contributions:

FHC conceived, designed and performed experiments and analyzed the data. FHC and FRV wrote the manuscript.

Figure legends:

**Figure 1.** Phylogenomic reconstruction of 16,484 viral genomic sequences reveals major lineages of uncultured prokaryotic viruses. The tree was built through Neighbor-Joining based on Dice distances calculated between viral genomic sequences from both NCBI RefSeq and those reconstructed from metagenomes, fosmid libraries, single virus genomes and prophages integrated into prokaryote genomes. Tree was midpoint rooted. Branch lengths were omitted to better display tree topology. Each of the 72 Level-1 GL-UVAB lineages are highlighted by black colored branches and with their defining nodes in indicated by blue dots. Numeric identifiers for the lineages are displayed in the innermost ring within gray strips. The outermost ring depicts the ICTV family level classification assignments of RefSeq viral genomes that were included in the tree. Color codes for ICTV families are as follows: Myoviridae (Green), Siphoviridae (Red), Podoviridae (Yellow), Lipothrixviridae (Blue), Fuselloviridae (orange), Sphaerolipoviridae (Cyan), Rudiviridae (Purple) and Tectiviridae (Pink). Fore reference, representatives of isolated viral genomes are shown with their ICTV genus level classification shown in parentheses.

**Figure 2.** Prevalence of targeted host and ecosystem sources among Level-1 GL-UVAB lineages assigned through phylogenomic reconstruction. A) Frequency of infected host phyla across each of the 72 identified lineages. B) Frequency of ecosystem sources from which viral sequences were obtained across each of the 72 identified lineages. For clarity, only hosts and ecosystems with prevalence equal or above 1% are shown. Numbers in parentheses indicate the total number of genomes assigned to each lineage.
Figure 3. Abundance patterns of GL-UVAB Level-1 lineages across habitats. Bar plots display the average and standard errors of the relative abundances of GL-UVAB Level-1 lineages across metagenomes and metaviromes from marine, freshwater, human gut and soil ecosystems.

Supplementary Figure legends:

Figure S1. Flowchart summarizing the methodology used to establish GL-UVAB. The initial dataset of genomic sequences consisted of the NCBI RefSeq and viral genomic sequences obtained through culturing independent approaches adding up to 195,698 genomic sequences from which 4,332,223 protein encoding genes were identified. After the initial filtering, 16,484 sequences were selected for phylogenomic reconstruction. Dice distances were calculated between this set and the resulting distance matrix was used for phylogenomic reconstruction through Neighbor-Joining. The obtained tree was used to identify lineages at three levels, based on average Dice distances within nodes: Level-1 (average Dice distances between genomes equal or below 0.925, and number of representatives equal or above 50), Level-2 (average Dice distances between genomes equal or below 0.825, and number of representatives equal or above 10) and Level-3 (average Dice distances between genomes equal or below 0.5, and number of representatives equal or above 3). Lineage abundances were estimated in metagenomic datasets by read mapping. Lineage pan-genomes were determined by identifying clusters of orthologous genes. Finally, sequences that were not included in the original tree were assigned to the lineages by closest relative identification (CRI). Closest relatives were determined based on percentage of matched genes and average amino acid identity and setting a minimum value of 50% for both these variables.

Figure S2. Phylogenomic reconstruction of 16,484 viral genomic sequences. The tree was built through Neighbor-Joining based on Dice distances calculated between viral genomic sequences from both NCBI RefSeq and those reconstructed from metagenomes, fosmid libraries and prophages integrated into prokaryote genomes. The tree was midpoint rooted. Nodes were collapsed if all the leaves belonged to the same Level-1 lineage to better display higher-order associations between lineages.
**Figure S3.** Phylogenomic reconstruction of 2,069 genomes of dsDNA viruses of Archaea and Bacteria from RefSeq. The tree was built through Neighbor-Joining based on Dice distances calculated between complete prokaryotic viral genomes from NCBI RefSeq. The tree was midpoint rooted. Branches are colored according to their family level taxonomic classification, the inner ring displays classification of genomes into subfamilies and the outer ring displays classifications into lineages identified for the benchmarking dataset.

**Figure S4.** Dataset sources of GL-UVAB genomes. Bar plots depict the abundance of sequences from each original publication that made up the full dataset (Total), those that were recruited during the pre-filtering step (Recruited) and those that were classified into GL-UVAB lineages (Classified).

**Figure S5.** Prevalence of targeted host and ecosystem sources among Level-1 GL-UVAB lineages assigned through phylogenomic reconstruction and lineage expansion by closest relative identification. A) Frequency of infected host phyla across each of the 72 identified lineages. B) Frequency of ecosystem sources from which viral sequences were obtained across each of the 72 identified lineages. For clarity, only hosts and ecosystems with prevalence within a lineage equal or above 1% are shown. Numbers in parentheses indicate the total number of genomes assigned to each lineage.

**Supplementary files:**

- **File S1.** Multifasta file containing the 195,698 viral genome sequences from Vir_DB_Nuc analyzed in this study.
- **File S2.** Multifasta file containing the 4,332,223 protein encoding gene sequences predicted from Vir_DB_Nuc.
- **File S3.** Newick format phylogenomic tree used to define GL-UVAB lineages. The tree was constructed with the Neighbor-Joining algorithm based on Dice distances calculated between 16,484 genomic sequences.
- **Table S1.** Table containing detailed information of all the genomic sequences from Vir_DB_Nuc analyzed in this study, including sequence identifier, NCBI access number,
original dataset, sequence length, number of identified PEGs, taxonomic classification (NCBI and ICTV), ecosystem source and host taxonomic classification and affiliation to the identified lineages at three hierarchical levels.

**Table S2.** Table describing the prevalence of OGs across GL-UVAB Level-1 lineages with associated taxonomic and functional annotation. Only OGs detected in at least 3 members of a lineage are shown.

**Table S3.** Statistics of the Dice distances observed between genomes of the taxa established by NCBI/ICTV and the lineages established by GL-UVAB.

**Table S4.** Prevalence of targeted hosts, ecosystem and dataset sources, taxonomic classification and completeness of genomes among the three levels of GL-UVAB lineages.
References

1. Rohwer F, Edwards R. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. J Bacteriol. 2002;184: 4529–4535. doi:10.1128/JB.184.16.4529

2. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. Cell. 2003;113: 171–182. doi:10.1016/S0092-8674(03)00233-2

3. Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, et al. Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. mSphere. 2017;2: 2:e00359-16. Available: https://doi.org/10.1128/mSphere.00359-16

4. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the marine virosphere using metagenomics. PLoS Genet. 2013;9: e1003987. doi:10.1371/journal.pgen.1003987

5. Bolduc B, Jang H Bin, Doulcier G, You Z-Q, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ. 2017;5: e3243. doi:10.7717/peerj.3243

6. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017; doi:10.1038/nrmicro.2016.177

7. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. Bioinformatics. 2017;33: 1–2. doi:10.1093/bioinformatics/btx157

8. Meier-Kolthoff JP, Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics. 2017;33: 3396–3404. doi:10.1093/bioinformatics/btx440

9. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol. 2008;25: 762–777. doi:10.1093/molbev/msn023

10. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun. 2017;8. doi:10.1038/ncomms15955

11. Mizuno CM, Ghai R, Saghaï A, López-García P, Rodriguez-Valera F. Genomes of Abundant and Widespread Viruses from the Deep Ocean. MBio. 2016;7: e00805-16. doi:10.1128/mBio.00805-16

12. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. Genome diversity of marine phages recovered from Mediterranean
metagenomes: Size matters. PLoS Genet. Public Library of Science; 2017;13:
e1007018.

13. Adriaenssens EM, Edwards R, Nash JHE, Mahadevan P, Seto D, Ackermann HW, et al. Integration of genomic and proteomic analyses in the classification of the Siphoviridae family. Virology. Elsevier; 2015;477: 144–154. doi:10.1016/j.virol.2014.10.016

14. Bellas CM, Anesio AM, Barker G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. Front Microbiol. 2015;6: 1–14. doi:10.3389/fmicb.2015.00656

15. Coutinho FH, Silveira CB, Gregoracci GB, Edwards RA, Brussaard CPD, Dutilh BE, et al. Marine viruses discovered through metagenomics shed light on viral strategies throughout the oceans. Nat Commun. Nature Publishing Group; 2017;8: 1–12. doi:10.1038/ncomms15955

16. Paez-Espino D, Elof-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. Nature Publishing Group; 2016;536: 425–430. doi:10.1038/nature19094

17. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and biogeochemical impacts of uncultivated globally abundant ocean viruses. Nature. Nature Publishing Group; 2016;537: 589–693. doi:10.1101/053090

18. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife. eLife Sciences Publications Limited; 2015;4: e08490. doi:10.7554/eLife.08490

19. Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, et al. Diel cycling and long-term persistence of viruses in the ocean’s euphotic zone. Proc Natl Acad Sci. 2017; 201714821. doi:10.1073/pnas.1714821114

20. Luo E, Aylward FO, Mende DR, Delong EF. Bacteriophage Distributions and Temporal Variability in the Ocean’s Interior. MBio. 2017;8: 1–13.

21. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science (80- ). 2015;348: 1–10. doi:10.1126/science.1261359

22. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473: 174–180. doi:Doi 10.1038/Nature09944

23. Aguirre de Cárcer D, López-Bueno A, Pearce DA, Alcamí A. Biodiversity and distribution of polar freshwater DNA viruses. Sci Adv. 2015;1: e1400127. doi:10.1126/sciadv.1400127
24. Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, et al. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl Environ Microbiol. 2014;80:6888–6897. doi:10.1128/AEM.01525-14

25. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, et al. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci. 2011;108: E757–E764. doi:10.1073/pnas.1102164108

26. Samson JE, Magadân AH, Sabri M, Moineau S. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol. Nature Publishing Group; 2013;11: 675–687. doi:10.1038/nrmicro3096

27. Oliveira PH, Touchon M, Rocha EPC. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. 2014;42: 10618–10631. doi:10.1093/nar/gku734

28. Oliveira H, Melo LDR, Santos SB, Nobrega FL, Ferreira EC, Cerca N, et al. Molecular Aspects and Comparative Genomics of Bacteriophage Endolysins. J Virol. 2013;87: 4558–4570. doi:10.1128/JVI.03277-12

29. Fernández-Ruiz I, Coutinho FH, Rodriguez-Valera F. Thousands of Novel Endolysins Discovered in Uncultured Phage Genomes. Front Microbiol. 2018;9: 1033. doi:10.3389/fmicb.2018.01033

30. Krupovic M, Dutilh BE, Adriaenssens EM, Wittmann J, Vogensen FK, Sullivan MB, et al. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. Arch Virol. Springer Vienna; 2016;161: 1095–1099. doi:10.1007/s00705-015-2728-0

31. Philosof A, Yutin N, Sharon I, Koonin E V, Shmona K. Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota Identified by Genome-Centric Metagenomics. 2017;

32. Vik DR, Roux S, Brum JR, Bolduc B, Joanne B, Padilla CC, et al. Putative archaeal viruses from the mesopelagic ocean. PeerJ. 2017; 1–26. doi:10.7717/peerj.3428

33. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Peña MJ, Martínez JM, et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat Commun. 2017;8: 15892. doi:10.1038/ncomms15892

34. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

35. Buchfink B, Xie C, Huson DH. Fast and Sensitive Protein Alignment using DIAMOND. Nat Methods. 2015;12: 59–60. doi:10.1038/nmeth.3176
36. Grazziotin AL, Koonin E V., Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45: D491–D498. doi:10.1093/nar/gkw975

37. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565

38. Saitou N NM. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4: 406–425.

39. Li L, Stoeckert Jr. CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13: 2178–2189. doi:10.1101/gr.1224503

40. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79: 7696–7701. doi:10.1128/AEM.02411-13

41. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Patterns and ecological drivers of ocean viral communities. Science. 2015;348: 1261498. doi:10.1126/science.1261498

42. Minot S, Bryson A. Rapid evolution of the human gut virome. Proc …. 2013;110: 12450–12455. doi:10.1073/pnas.1300833110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1300833110

43. Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. Nat Commun. Springer US; 2017;8. doi:10.1038/s41467-017-01086-2

44. Rascovan N, Carbonetto B, Revale S, Reinert MD, Alvarez R, Goddeas AM, et al. The PAMPA datasets: a metagenomic survey of microbial communities in Argentinean pampean soils. Microbiome. 2013;1: 21. doi:10.1186/2049-2618-1-21

45. Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K, et al. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. Appl Environ Microbiol. 2014;80: 1777–1786. doi:10.1128/AEM.03712-13

46. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;9: 357–9. doi:10.1038/nmeth.1923
