Minimum Information about an Uncultivated Virus Genome (MIUViG): a community consensus on standards and best practices for describing genome sequences from uncultivated viruses


*Correspondence to: SR sroux@lbl.gov & EAE-F eaeloefadrosh@lbl.gov

Abstract

In light of the unprecedented diversity of viruses uncovered by culture-independent technologies, we here present a set of standards for describing sequence data from uncultivated virus genomes (UViGs). The proposed Minimum Information about an Uncultivated Virus Genome (MIUViG) standards have been developed within the framework of the Genomic Standards Consortium as an extension of the Minimum Information about any (x) Sequence (MIxS), and builds on the Minimum Information about a Single Amplified Genome (MISAG) and Metagenome-Assembled Genome (MIMAG) developed for uncultivated bacteria and archaea. These standards include features specific for UViG including the evaluation of virus origin and genome quality, and provide a framework for performing and reporting genome annotation, taxonomic classification, estimation of biogeographic distribution, and in silico host prediction. Community-wide adoption of the MIUViG standards will result in a greater inclusion of sequence data from uncultivated virus genomes in public databases, enhancing future comparative studies and enabling a more systematic and comprehensive exploration of the global virosphere.
Introduction

Viruses represent a ubiquitous component of life on Earth and, based on current estimates, virus particles significantly outnumber living cells in most habitats\(^{1,2}\). Only a small fraction of this vast virus diversity has been isolated and cultivated in the laboratory, yet great progress has been made in mapping the virus genomic sequence space based on genomes reconstructed from uncultivated viruses\(^{3,4}\). Virus genomes are now frequently sequenced and assembled \textit{de novo} directly from biotic and abiotic environments, and without laboratory isolation of the virus-host system. In the last two years alone, more than 750,000 uncultivated virus genomes (UViGs) have been identified from shotgun metagenome and metatranscriptome datasets\(^{5-10}\). These UViGs form a genome database that is five-fold larger than the one based on isolated viruses (Fig. 1), and represent \(\geq 95\%\) of the taxonomic diversity derived from publicly available virus sequences\(^{11,12}\). Although still skewed towards double-stranded DNA (dsDNA) genomes, these UViGs provide unprecedented opportunities for assessing global virus diversity, evaluating ecological structures and drivers of virus communities, improving our understanding of the evolutionary history of viruses, and investigating virus-host interactions.

Analysis and interpretation of genomes in the absence of a cultured isolate presents challenges, whether the genomes derive from microbial cells or viruses. In particular, these sequences are often not complete genomes, and phenotypic properties such as virion structure and host range (in the case of viruses) can only be predicted indirectly, usually by computational methods. To address some of these challenges, standards were recently proposed for reporting of uncultivated microbial genomes derived from single cell or shotgun metagenome approaches\(^{13}\). Although some aspects of the proposed Minimum Information about a Single Amplified Genome (MISAG) and Metagenome-Assembled Genome (MIMAG) standards are directly applicable to UViGs, a formalized set of standards specific to viruses is needed to provide alternative or additional criteria. Notably, the extraordinary diversity among viruses in genomic composition and content, replication strategy, and host specificity means that the completeness, quality, taxonomy, and ecological significance of UViGs must be evaluated by virus-specific metrics.

The Genomic Standards Consortium (\url{http://gensc.org}) maintains up-to-date metadata checklists for the Minimum Information about any (x) Sequence (MIxS), encompassing genome and metagenome sequences\(^{14}\), marker gene sequences\(^{15}\), and single amplified and metagenome-assembled bacterial and archaeal genomes\(^{13}\). Here, we provide a specific set of standards that extend the MIxS checklists to the identification, quality assessment, analysis, and public reporting of UViG sequences (Table 1 and Supplementary Table 1), along with recommendations on how to perform these analyses. The metadata checklist for the publication and database submission of UViG is designed to be flexible enough to accommodate technological changes and methodological advancements over time (Table 1). The information gathered through this checklist can be directly submitted alongside novel UViG sequences.
to member databases of the International Nucleotide Sequence Database Collaboration (i.e. DDBJ, EMBL-EBI, and NCBI), which host and display these metadata along with the UViG sequence. These MIUViG standards can also be used alongside existing guidelines for virus genome analysis, especially those issued by the International Committee on Taxonomy of Viruses (ICTV), which recently endorsed the incorporation of UViGs into the official virus classification scheme\(^{16-20}\) (https://talk.ictvonline.org). Finally, although these MIUViG standards and best practices were designed for genomes of viruses infecting microorganisms, they can also be applied to viruses infecting animals, fungi, and plants, and matched with comparable standards which already exist for epidemiological analysis of these viruses\(^{21}\) (Supplementary Table 2).

**Minimum Information about an Uncultivated Virus Genome (MIUViG)**

*Sources of UViGs.*

UViGs can be identified within a broad range of DNA and RNA sequence datasets (Fig. 2). First, some approaches aim at enriching virus particles from an environmental sample, such as viral metagenomics and single-virus genomics. Viral metagenomes are typically obtained through a combination of filtration steps and DNase/RNase treatments, DNA and/or RNA extraction depending on the targeted viruses, reverse-transcription for RNA viruses, and shotgun sequencing\(^{4,22-26}\). Targeted sequence capture approaches can also be used to recover members of specific virus groups (Fig. 2), which has already proven useful for cases in which viruses represent a minor part of the templates, e.g., clinical samples\(^{27,28}\). In contrast, single-virus genomics use flow cytometry to sort individual virus particles for genome amplification and sequencing, delivering viral single amplified genomes (vSAGs)\(^{10,29-31}\) (Fig. 2). Typically, both viral metagenomes and viral single-virus genomes are currently sequenced with short-read high-throughput methods (e.g. Illumina) and assembled using similar algorithms as for microbial genomes and metagenomes\(^{32}\). This includes assembly approaches for single samples or multiple samples combined. However, because of the relatively small size of virus genomes (92% of virus genomes currently represented in the National Center for Biotechnology Information (NCBI) Viral RefSeq database are < 100 kb\(^{11}\)), the short-read-based genome assembly step could soon be avoided by leveraging long read sequencing technologies\(^{33}\) (e.g., PacBio or Nanopore, Fig. 2). Sequencing virus genomes from a single template will notably enable the identification of individual genotypes within mixed populations. The main advantages of these virus-targeted datasets include an improved *de novo* assembly of both abundant and rare viruses, a greater confidence that the sequence is of virus origin, and the ability to sequence both active and “inactive” or “cryptic” viruses, i.e., viruses for which virions are present in the sample but without opportunities for infection. However, virus-targeted datasets have a number of limitations, including (i) an over-representation of virulent viruses with high burst size (i.e. high number of virus particles released from each infected cell), and (ii) an
under-representation of larger viruses with capsids $\geq 0.2 \, \mu m$, such as giant viruses, due to the selective filtration step often used to separate virus particles from cells\textsuperscript{34}. In addition, in silico approaches are often the only option available to determine the host range of these viruses (see below).

An alternative approach for UViG detection is to identify virus sequences in non-targeted or cell-targeted datasets. Virus sequences will frequently appear in nominal cell fractions such as sorted cells, organismal tissues, or environmental samples collected on 0.2 $\mu m$ filters, for a variety of reasons\textsuperscript{6,35–37}. These sequences could originate from viruses actively replicating within the sample cells, from temperate viruses stably associated with the host genomes (i.e., provirus or prophage) either integrated or existing as an episomal element in the host cell, or through co-sampling of free virus particles along with the target cells. For the purpose of uncovering novel virus genomes, exploring these cellular datasets presents three main advantages: (i) lytic, temperate, and persistent infections ongoing in the microbial community will be broadly detected, (ii) sampling biases resulting from the selection of virus particles based on physical properties will be limited, and (iii) this approach can leverage the vast amount of metagenomic data generated for purposes other than virus discovery. However, these UViG datasets may be biased toward viruses infecting the dominant host cells in the sample, whereas rare viruses or viruses infecting rare hosts could be under-represented, if captured at all.

The broad range of datasets from which UViGs can be extracted (Fig. 2) reflects both the pervasiveness of viruses and their critical importance in multiple fields, such as evolutionary biology, microbial ecology, and infectious diseases. Some of these techniques are better suited towards addressing specific biological questions but from the virus discovery standpoint, these approaches are mostly complementary. To highlight the differences and complementarity between approaches, we compared the number of large UViGs (here virus contigs $\geq 10$kb) assembled from virus-targeted and microbial cell-targeted metagenomes from the same samples obtained through the Tara Oceans expedition\textsuperscript{38,39}, after we subsampled them to the same number of reads (Supplementary Fig. 1). Metagenomes targeting the nominal virus fraction yielded, on average, 20 times more UViGs than their microbe-targeted counterparts. However, at the current sequencing depth, UViGs derived from microbial metagenomes were not subsets of the UViGs identified in the viral metagenomes, with an average 74\% of the UViGs unique to the microbial fraction (range: 34–98\%). This comparison illustrates how integrating virus sequences from samples across different size fractions and/or processed with different techniques is highly valuable for exploring the virus genome sequence space\textsuperscript{40}.

Identification of virus sequences in genome and metagenome assemblies.

Regardless of what type of dataset is analyzed, the virus origin of sequences needs to be validated. Notably, even samples enriched for virus particles can contain a substantial amount of cellular DNA\textsuperscript{41}. What appears as contamination can result from difficulties in separating virus particles and cellular
fractions, e.g., due to the presence of ultra-small bacterial cells or the capture of dissolved extracellular DNA within the virus fraction. However, cellular sequences can also derive from genome fragments of cellular origin that were encased within virus capsids or comparable particles, e.g., through transduction events, DNA-containing membrane vesicles, or gene transfer agents.

A number of bioinformatic tools and protocols have been developed to identify sequences from bacteriophages and archaeal viruses, eukaryotic viruses, or combined bacteriophages, archaeal viruses, and large eukaryotic viruses (Supplementary Table 3). These approaches rely on a few fundamental characteristics: a sequence will be considered to be of virus origin if it is significantly similar to that of known viruses in terms of gene content or nucleotide usage pattern, or if it is mostly unrelated to any known virus and cellular genome but contains one or more viral hallmark genes. Any reported UViG should thus be accompanied by a list of virus detection tool(s) and protocol(s) used alongside the thresholds applied (Table 1).

Substantial challenges still need to be overcome to accurately identify integrated proviruses and define their precise boundaries in the host genome (Box #2). Notably, no high-throughput approach is currently available to accurately distinguish active proviruses still able to replicate and produce virions from decayed proviruses (inactive remnants of a past infection). Hence, although prediction methods are continuously improving, UViGs detected as proviruses should be clearly marked as such, as they come with their own specific caveats (Table 1).

Quality estimation of UViGs.

To standardize the description of UViG sequences in peer-reviewed publications and databases, we propose to formally define three categories: (i) genome fragment(s), (ii) high-quality draft genome, and (iii) finished genome (Table 2, Fig. 3). These categories mirror the classification system recently proposed for microbial SAGs and MAGs, and can be matched to categories previously proposed for complete-genome sequencing of small viruses for epidemiology and surveillance (Supplementary Table 2). Determining UViG quality is more challenging than for microbial MAGs or SAGs, largely because many virus taxa lack reliable sets of single-copy marker genes that can be used to estimate completeness of a draft genome, although notable exceptions exist, such as for large eukaryotic dsDNA viruses. Instead, the approaches adopted by the research community to estimate UViG sequence completeness have relied on (i) identifying circular contigs or contigs with inverted terminal repeats as putative complete genomes, and (ii) comparing linear contigs to known complete reference genome sequences. For the latter case, a taxonomic assignment of the UViG to a (candidate) (sub)family or genus is typically required, as genome length is relatively homogeneous at these ranks (±10%, Supplementary Fig. 2, Supplementary Table 4). This assignment can be based on the detection of specific marker genes, e.g. clade-specific Viral Orthologous Groups (Supplementary Table 5), or...
derived from genome-based classification tools (see below section “Taxonomic classification of UViGs.”). Estimating completeness is also more difficult for segmented genomes, which require either a closely related reference genome or in vitro experiments beyond the initial sequencing\textsuperscript{21}. A detailed example of how this quality tier classification can be performed on the Global Ocean Virome dataset\textsuperscript{7} is presented in Supplementary Text and Supplementary Table 6.

Contigs or genome bins representing < 90 % of the expected genome length, or for which no expected genome length can be determined, would be considered genome fragments. Pragmatically, this category would include some UViG fragments large enough to be assigned to known virus groups based on gene content and Average Nucleotide Identity (ANI), when applicable. However, high-quality draft or finished genomes would be required to establish new formal taxa (Fig. 3). Sequences from UViG fragments can be used in phylogenetic and diversity studies, either as references for virus OTUs (see section “Distribution and abundance of UViGs”), or through the analysis of virus marker genes encoded in these genome fragments, for example capsid proteins, terminases, ribonucleotide reductases, and DNA- or RNA-dependent RNA polymerases\textsuperscript{54–59}. Similarly, UViG fragments are useful for exploring the functional gene complement of unknown viruses and tentatively linking them to potential hosts. Importantly however, current methods for automatic virus sequence identification are challenged by short (< 10kb) sequences, which should be interpreted with utmost caution.

Contigs or genome bins (i.e., a collection of contigs) predicted as complete based on circularity or the presence of inverted terminal repeats, or representing ≥ 90% of the expected genome sequence, would be considered high-quality drafts, consistent with standards for microbial genomes\textsuperscript{13,60}. Of note, repeat regions can lead to erroneous assembly of partial genomes as circular contigs\textsuperscript{61}. Thus, the length of the assembled circular contig should be considered when assessing UViG completeness (Box #2). For UViGs not derived from a consensus assembly, i.e. single long reads, an average base calling quality > 99% (i.e. phred score > 20) is required to qualify as a high-quality draft genome. Among these high-quality drafts, genome sequences assembled in a single contig, or one per segment, with extensive manual review, editing, and annotation would be considered a finished genome. Annotation should include identification of putative gene functions, structural, replication, or lysogeny modules, and transcriptional units. This category is thus reserved for only the highest quality, manually curated UViGs, and required for the establishment of novel virus species (Fig. 3, Table 2).

In contrast to bacterial and archaeal SAGs and MAGs\textsuperscript{13}, quality estimation of UViGs does not include a threshold on genome contamination, i.e. presence of sequence(s) originating from a different genome(s) alongside the genuine UViG. Most UViGs are represented by a single contig, and according to in silico simulations, chimeric contigs are relatively rare (< 2%)\textsuperscript{61}. Nevertheless, contamination should be evaluated whenever possible using (i) coverage by metagenome reads which should be even along the genome with no major deviance except for highly conserved genes\textsuperscript{62–64}, and (ii) single-copy
marker genes as for microbial MAGs (MIMAGs, Supplementary Table 5). In addition, UViG sequences often represent consensus genomes from a heterogeneous population. Although not included as a quality criterion, the structure of the underlying population can be estimated through read mapping and single nucleotide polymorphism (SNP) calling.

Functional annotation of UViGs.

Typically, functional annotation of UViGs consists of two parts: (i) predicting features on the genome sequence such as protein-coding genes, tRNAs, and integration sites, and (ii) assigning functions to the predicted features, or protein families for hypothetical proteins. Annotation pipelines have been proposed for different types of viruses, and major differences between virus genomes likely preclude the development of a single tool suitable to annotate every virus. Hence the computational approaches and softwares used to annotate UViGs must be reproducibly detailed (Table 1).

Of particular importance for viruses is the choice of methods and reference databases used to annotate predicted proteins. Notably, homologs of novel virus genes will often not be detected with standard methods for pairwise sequence similarity detection, such as BLAST, but instead require the use of more sensitive profile similarity approaches such as HMMER, PSI-BLAST, or HHPred, which can leverage databases of virus protein profiles (Supplementary Table 7, reviewed in ref. 76). Although sequence profiles for many protein families have been collected, they frequently remain unassociated with any specific function. Efforts to improve these functional annotations will be supported by information about the distribution, genome context, and diversity of these uncharacterized protein families. While these resources are being actively developed and improved, UViG analyses should always report (i) feature prediction method(s), (ii) sequence similarity search method(s), and (iii) database(s) searched (Table 1, Box #2).

Taxonomic classification of UViGs.

Taxonomic classification is an important step in the analysis of UViG as it provides information on its relationship to known viruses. Historically, the information and criteria used for virus classification have changed as knowledge on virus diversity and molecular biology approaches has improved, but classification has now broadly converged to genome-based analyses (Box #1). Because of stark differences in genome length, mutation rate, and evolution mode, however, the ICTV established specific demarcation criteria for each virus group (Supplementary Table 8). Meanwhile, since UViGs often represent new groups for which no formal demarcation criteria have been defined, establishing universal or near-universal cutoffs will enable the creation of primary groups approximating ICTV classification that could be scrutinized later by experts.
Recently, a consensus has emerged on using whole genome Average Nucleotide Identity (ANI) for classification at the species rank, which represents the primary data for many downstream ecological, evolutionary, and functional studies. This has been justified by population genetics studies and gene content analyses of NCBI RefSeq virus genomes (Supplementary Text and Supplementary Fig. 3). Here, we propose to formalize the use of these species-rank virus groups and, because these were alternatively termed “viral population,” “viral cluster,” or “contig cluster” in the literature, to uniquely designate these as virus Operational Taxonomic Units (vOTUs). We also suggest standard thresholds of 95% ANI over 85% alignment fraction (AF, relative to the shorter sequence), based on a comparison of sequences currently available in NCBI RefSeq and IMG/VR (Supplementary Text and Supplementary Figs. 3–4). Common thresholds will improve reproducibility and comparative analysis of distinct datasets, although partial genomes remain challenging to classify (Supplementary Figure 5). In addition, reporting the classification of new UViGs into vOTUs should include the clustering approach and cutoff used, the reference database used, if any, as well as the genome alignment approach since small differences have been observed between different methods (Table 1).

For higher taxonomic ranks, i.e. order, (sub)family, and genus, no consensus has yet been reached on which approach could be universally used, although several have been proposed. Regardless of the tool chosen, UViG reports including taxonomic classification must clearly indicate the methods and cutoffs applied, and any new taxon must be highlighted as preliminary, e.g. “genus-rank clusters,” “putative genus,” or “candidate genus,” but not simply “genus,” as the latter is reserved for ICTV-recognized groups (Table 1). For putative taxa to be officially accepted, authors should submit formal taxonomic proposals (“TaxoProps”) to the ICTV for consideration (https://talk.ictvonline.org/files/taxonomy-proposal-templates/).

Finally, information about the nature of the genome and the mode of expression, i.e. Baltimore classification, should be included in UViG description whenever possible. This information can usually be derived from the methods used to process the samples from which a UViG was assembled, which will often strongly select for or exclude specific types of genomes, and from the detection of specific marker genes (Supplementary Table 5). Similarly, the expected segmentation state of the genome, i.e., segmented or non-segmented, typically derived from taxonomic classification and comparison with the closest references, should be reported (Table 1).

Distribution and abundance of UViGs.

Abundance estimates of a vOTU across datasets provide valuable information on the distribution and potential ecological niche of the virus. The relative abundance and distribution of a virus can be estimated through short-read metagenome mapping. However, thresholds must be applied to (i) the nucleotide identity between the read and UViG sequence, and (ii) the percentage of the representative
UViG sequence covered by metagenome reads. Both parameters are critical to avoid false-positive detection\textsuperscript{61,62,97}. Alternatively, pseudo-alignment and abundance estimation through expectation-maximization as implemented e.g. in FastViromeExplorer\textsuperscript{98} can be used instead of coverage estimation through read mapping, with similar cutoffs applied on the coverage along the genome and total number of mapped reads.

The specific thresholds for nucleotide identity and coverage of the reference genome can be adjusted depending on the scientific objectives of a given study. For instance, increasing the coverage threshold from 10\% to 75\% led to a lower rate of incorrect detection (false discovery rate decreased from 8\% to 0\%) but at the cost of a lower sensitivity (decreased from 88\% to 82\%, based on simulated datasets from ref. \textsuperscript{61}). Thus, when reporting read mapping-based distributions and/or relative abundances, it is important to report the nucleotide identity and coverage thresholds, and provide an estimate of false-positive and false-negative rates for the combined thresholds, either computed \textit{de novo} or extracted from the literature, e.g. from refs \textsuperscript{61,62}. Finally, two important caveats should be considered when using read mapping to estimate virus distribution and relative abundance: (i) some amplification methods produce non-quantitative datasets, in which coverage can not be interpreted as relative abundance (Box #2), and (ii) there are currently no guidelines for integrating coverage data from different size fractions.

\textit{In silico host prediction.}

Once a novel virus genome has been assembled, an important step toward understanding the ecological role of the associated virus is to predict its host(s). Most current experimental approaches to determine virus host range require the availability of a representative cultured virus, so \textit{in silico} approaches are often the only option for UViGs (reviewed in ref. \textsuperscript{99}; Supplementary Table 9). These bioinformatic approaches can be separated into four major types.

First, hosts can be predicted with relatively high precision based on sequence similarity between the UViG and a reference virus genome when a closely related virus is available\textsuperscript{100,101}. Second, hosts can be predicted based on sequence similarities between a UViG and a host genome. These sequence similarities can range from short exact matches (~ 20–100 bp), which include CRISPR spacers\textsuperscript{6,99,102}, to longer (>100 bp) nucleotide sequence matches, including proviruses integrated into a larger host contig\textsuperscript{99,103,104} (Supplementary Table 9). Host range predictions based on sequence similarity are the most reliable but require that a closely related host genome has been sequenced\textsuperscript{99}. Third, host taxonomy from domain down to genus rank can be predicted from nucleotide usage signatures reflecting coevolution between virus and host genomes in terms of GC content, kmer frequency, and codon usage\textsuperscript{36,105,106}. These approaches are usually less specific than sequence similarity-based ones, cannot reliably predict host range below the genus rank, but can provide a predicted host for a larger number of UViGs\textsuperscript{7} (Supplementary Table 9). Finally, host predictions can be computed from a comparison of abundance
profiles of host and virus sequences across spatial or temporal scales, either through abundance correlation \(^{34,107–109}\) or through more sophisticated model-based interaction predictors \(^{110,111}\). Although few datasets are currently available for robust evaluation of host prediction based on comparison of abundance profiles, we expect this approach to become more powerful and relevant as high-resolution time-series metagenomics becomes more common.

As all these bioinformatic approaches remain predictive, it is critical that robust false-discovery rate estimations are provided (Table 1). Moreover, computational tools do not predict quantitative infection characteristics (e.g. infection rate or burst size), which are important for understanding the impacts of viruses on host biology, and to date only apply to viruses infecting bacteria or archaea. Nevertheless, these predictions are important guides for subsequent \textit{in silico}, \textit{in vitro}, and \textit{in vivo} studies, including experimental validation to unequivocally demonstrate a viral infection of a given microbial host. Host predictions should thus be reported along with details regarding the specific tool(s) used and, importantly, their estimated accuracy as derived from either published benchmarks or from tests conducted in the study (Table 1). This information will allow virus-host databases\(^{100,112}\) to progressively incorporate UViGs while still controlling for the sensitivity and accuracy of the predictions provided to users.

\textbf{Public reporting of UViGs}

We recommend the following best practice for sharing and archiving UViGs and UViG-related data: data publication should center on the data resources of the International Nucleotide Sequence Database Collaboration (INSDC; \url{http://www.insdc.org/}), through one of the member databases at DDBJ (\url{https://www.ddbj.nig.ac.jp/index-e.html}), EMBL-EBI (European Nucleotide Archive, ENA; \url{https://www.ebi.ac.uk/ena}) or NCBI (GenBank and the Sequence Read Archive; \url{https://www.ncbi.nlm.nih.gov/nucleotide}). If needed, INSDC database curators can be contacted directly for large-scale batch dataset submissions. Where new data sets are generated as part of a UViG study, sequenced samples should be described according to the environment-relevant Minimum Information about any (x) Sequence (MIxS) checklists and raw read data should be reported in appropriate formats. High-quality and finished UViGs should be submitted as assemblies, the former reported as “draft,” accompanied by the required metadata (Table 1). Assemblies at other levels may be submitted, especially if these are central to the study, but they must be accompanied by the required metadata (Table 1). Where available, functional annotation and taxonomic classification should be provided to INSDC, while occurrence and abundance data can be reported as “Analysis” records to the ENA. For ICTV classification, only coding-complete genomes, i.e. complete high-quality and finished draft UViGs, are currently considered\(^{20}\). Finally, relevant INSDC accession numbers should be cited in peer-reviewed publications.
Conclusion

The MIUViG standards and best practices presented here provide the first virus-specific counterpart to the recently outlined MISAG and MIMAG. However, the field of virus genomics and metagenomics is rapidly changing. For instance, the recovery of high-quality UViGs will likely improve with new emerging sequencing strategies, which in the short term include the combination of short- and long-read sequencing and further developments in the direct sequencing of DNA and RNA with minimized library preparation steps. Meanwhile, a number of areas and resources are still under active development, such as approaches for genome-based classification of viruses, and the development of a unified, comprehensive, and annotated reference database of virus proteins. These standards are thus designed based on current knowledge of virus diversity and aim to provide a framework for the future exploration of virus genome sequence space while encouraging discussion about the analysis and reporting of UViGs. Community adoption of these standards, including through ongoing collaborations with other virus committees (e.g. ICTV) and data centers (e.g. DDBJ, EMBL-EBI, and NCBI), will enable the research community to better utilize and build on published uncultivated virus genomes.
**Glossary**

**UViGs:** Uncultivated Virus Genomes. Partial or complete genomes of viruses that are known exclusively from sequence data, as opposed to viruses that can be cultivated, cloned, characterized, and propagated on cell cultures or tissues.

**UpViGs:** Uncultivated proVirus Genomes. Partial or complete genomes of viruses that are known exclusively from sequence data and are integrated in a host genome fragment. These viruses are thus directly associated with a host but the boundaries of these proviruses can be difficult to accurately predict *in silico* and should thus be interpreted with utmost caution.

**vSAGs:** viral Single-Amplified Genomes. Partial or complete genomes of viruses assembled from sequencing of an individual virus particle, typically sorted using flow cytometry and amplified with Whole Genome Amplification techniques.

**MIUViG:** Minimum Information about an Uncultivated Virus Genome. Standards developed in the MIxS framework with the Genomic Standards Consortium for reporting uncultivated virus genome sequences (UVIGs).

**ICTV:** International Committee on Taxonomy of Viruses. The ICTV is a Committee of the Virology Division of the International Union of Microbiology Societies, whose primary mission is to develop, refine, and maintain a universal virus taxonomy that reflects their evolutionary relationships.

**MIxS:** Minimum Information about any (x) Sequence. Framework used as a single entry point to all minimum information checklists from the Genomic Standards Consortium.

**MIMAG/MISAG:** Minimum Information about a Metagenome-Assembled / Single Amplified Genome. Standards developed with the Genomic Standards Consortium in the MIxS framework for reporting bacterial and archaeal genome sequences.

**MAGs:** Metagenome-Assembled Genomes. Partial or complete genomes assembled from a (set of) metagenome(s). MAGs are usually genome bins, i.e. a collection of contigs predicted to belong to genomes from a single population.

**SAGs:** Single Amplified Genomes. Partial or complete genomes assembled from sequencing of an individual cell, typically sorted using Fluorescence Activated Cell Sorting and amplified with Whole Genome Amplification techniques.
Box 1 – Virus taxonomy

Virus taxonomy has long been a subject of intense scrutiny and elaborate debates. Compared to the classification of cellular organisms, virus classification is associated with unique challenges. First, viruses are most likely polyphyletic, i.e., they arose multiple times independently. Thus, unlike ribosomal proteins or rRNAs for cellular organisms, no genes are systematically present among all virus genomes that could be used as universal taxonomic markers. Furthermore, viruses display a broad range of genomic characteristics, including ssRNA (or ssDNA) viruses encoding only a couple of proteins, dsRNA viruses with up to 12 segments, and large and complex dsDNA viruses with genome sizes that reach the realm of bacteria. Viruses exhibit high genetic diversity as they tend to evolve faster than cellular organisms, both in terms of their genetic sequence and in terms of their genome content. Due to this polyphyletic and diverse nature, viruses are not incorporated into the current universal tree of life and a “one-size-fits-all” virus taxonomy is difficult to attain, resulting in different classification rules for different groups of viruses.

A set of criteria to classify viruses was first formally proposed by the Virus Subcommittee of the International Nomenclature Committee at the 5th International Congress of Microbiology, held at Rio de Janeiro (Brazil), in August 1950. The virus classification criteria were purposefully based on stable properties of the virus itself, first among them being the virion morphology, genome type, and mode of replication, rather than more labile properties such as symptomatology after infection. A hierarchical categorization of viruses based on genome type and virion morphology was then proposed, and another operational classification scheme relying on nucleic acid type and method of genome expression was proposed by David Baltimore in 1971.

The need for a specific set of rules to name and classify viruses led to the establishment of the International Committee on Nomenclature of Viruses (ICNV), renamed as the International Committee on Taxonomy of Viruses (ICTV) in 1975. The ICTV is a committee of the Virology Division of the International Union of Microbiological Societies and is charged with the task of developing, refining and maintaining the official virus taxonomy, presented to the research community in “ICTV Reports” and interim update articles (“Virology Division news”) in Archives of Virology. Using some of the stable properties of viruses that were previously highlighted, experts within the ICTV progressively developed a universal virus taxonomy similar to the classical Linnaean hierarchical system, i.e. linking virus groups to familiar taxonomic ranks including Order, Family, Genus and Species.

In the post-genomic era, virus classification is now increasingly based on the comparison of genome and protein sequences, which provides a unique opportunity to evaluate phylogenetic and evolutionary relationships between viruses and reconcile virus taxonomy with their reconstructed evolutionary trajectory. The ICTV has undertaken the immense task of re-evaluating virus classification in light of...
this new sequence-based information\textsuperscript{16,20,116,117}. Importantly, with large sections of the virosphere still to be explored, virus taxonomy only represents our current best attempt at recapitulating virus evolutionary history based on available data. Thus, virus classification will necessarily remain dynamic, expanding and adjusting to new data as we discover novel viruses, and being refined with the progression of scientific understanding of virus evolution.

**Box 2 – Common pitfalls when analyzing sequence data for uncultivated virus genomes**

- *Mistaking a cellular genome fragment for a virus sequence*: Two situations are particularly prone to misidentification of a cellular sequence as viral. First, even viral metagenomes typically contain some level of cellular contamination\textsuperscript{41}. Any analysis should thus start with the identification of virus and cellular sequences, even for virus-targeted datasets — a process improved through the proper use of replicates, blanks, and other controls. Second, the boundaries of an integrated provirus can be challenging to identify even for dedicated softwares (e.g. PHAST, VirSorter). This can unfortunately lead to the erroneous inclusion of host gene(s) in the predicted virus genome, especially for genes on the edges of a predicted provirus or genome fragment. Thus, annotating these integrated virus genomes requires the greatest care and attention.

- *Partial genomes assembled as circular contigs*: Depending on the methods used, some partial genomes can be misassembled as circular contigs due to repeats\textsuperscript{61}. These erroneous circularized fragments could then be incorrectly identified as complete genomes. Hence, the size and gene content of circular contigs should always be validated to be consistent or at least plausible in comparison with known reference genomes.

- *Errors in gene prediction*: For novel viruses with little or no similarity to known references, gene prediction can be very challenging in the absence of concurrent transcriptomics or proteomics data. The result from automatic gene predictors applied to novel viruses should thus be checked for gene density (most viruses do not include large non-coding regions), as well as typical gene prediction errors such as internal stop codons causing artificially shortened genes.

- *Inaccurate functional annotation*: The annotation of open reading frames (ORFs) predicted from novel viruses often requires sensitive profile similarity approaches. While such sensitive searches are necessary to detect homology in the face of high rates of virus sequence evolution, the inferred function should be cautiously interpreted and remain general (e.g. “DNA polymerase”, “Membrane transporter”, or “PhoH-like protein”).

- *Clustering of partial genomes*: Incomplete genomes will often be difficult to classify using genome-based taxonomic classification methods. For example, the estimation of whole genome ANI from
partial genomes could vary by up to 50% from the complete genome value (Supplementary Fig. 5). Hence, the classification of genome fragments and their clustering into vOTUs should be interpreted only as an approximation of the true clustering values, and will likely change as more complete genomes become available.

- **Taxonomic classification of UViG:** Although virus classification primarily relies on genome sequences (see Box #1), no universal approach is currently available to classify viruses at different ranks. Classification of UViGs should thus be based on the best method available for the relevant type of virus and interpreted carefully.

- **Read mapping from non-quantitative datasets:** Amplified datasets, produced using e.g. Multiple Displacement Amplification or Sequence-Independent Single-Primer Amplification, are highly biased toward specific virus genome types and can selectively over-amplify specific genome regions. The coverage derived from read mapping based on these amplified datasets should thus not be interpreted as reflecting the relative abundance of the UViG in the initial sample.
Figure 1. Timeline of virus genome databases growth⁶,⁷,30,58,118–124. Genome sequences originate from isolates (blue and green) or from uncultivated viruses (UViGs, yellow). For genomes from isolates, both the total number of distinct genomes and the number of “reference” genomes, i.e. one genome per virus species, are indicated (in blue and green, respectively). These numbers are based on all virus sequences at NCBI and the NCBI Virus RefSeq database, respectively. UViGs can be obtained from metagenomes, proviruses identified within microbial genomes, or from single-virus genomes. A comprehensive database of UViGs is available at https://img.jgi.doe.gov/vr/. UpViG: Uncultivated provirus, i.e. virus genome integrated in its host genome (see glossary).
Figure 2. Identification and analysis of UViGs. Schematic of approaches used to obtain UViGs, which are largely similar to those used to obtain microbial Single Amplified Genomes (SAGs) and Metagenome-Assembled Genomes (MAGs). Additional steps or steps which have to be adapted for UViG are colored for sample preparation (orange) and for bioinformatics analysis (blue). Steps specifically required for virus targeting are highlighted in bold. *For viruses with short genomes, long-read technologies can provide complete genomes from shotgun sequencing in a single read, bypassing the assembly step. **Targeted sequence capture can be used to recover virus genomes from a known virus group. These genomes can be recovered from samples in which they represent a small fraction of the templates, e.g. clinical samples.
Figure 3. UViG classification and associated sequence analyses. The type(s) of analysis that can be performed for each quality category is indicated by the horizontal bar and labels on top. “Functional potential, host prediction” refers to typical functional annotation used in gene content analysis and the application of different \textit{in silico} host prediction tools. “Taxonomic classification” refers to classification of the contig to established groups using marker genes or gene content comparison. “Diversity and distribution” includes vOTU clustering and relative abundance estimation through metagenome read mapping, at the geographical scale or across anatomical sites for host-associated datasets. “Novel taxonomic groups” concerns the delineation of new proposed groups (e.g. families, genera) based exclusively on UViG sequences. “Novel reference species” refers to the proposal of a new entry in ICTV (\url{https://talk.ictvonline.org/files/taxonomy-proposal-templates/}). *Some of these approaches require a minimum contig size, e.g. contigs $\geq$ 10kb for taxonomic classification based on gene content\(^{84}\) or diversity estimation\(^{61}\), and will not be applicable to every genome fragment.

Table 1 (next page). \textbf{List of mandatory and optional metadata for UViGs.} Mandatory metadata are highlighted in blue. The status of metadata indicates if identical or similar information is asked for in the MIMAG / MISAG standards, with virus-specific metadata highlighted in orange, and metadata similar but adapted for UViGs in purple. If one of the mandatory metadata is missing, the value should be set as “Not applicable” for metadata that cannot be evaluated, or “Missing – Not collected” for the ones that could be assessed but for which the result is not currently available. MIMAG: metagenome-assembled genome; MISAG: minimum information about a single amplified genome. ANI: Average Nucleotide Identity. AF: Alignment Fraction.
<table>
<thead>
<tr>
<th>Metadata category</th>
<th>Metadata</th>
<th>Requirement</th>
<th>Description</th>
<th>Syntax</th>
<th>Example value</th>
<th>Specificity to UVGAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>General genome metadata</td>
<td>source of UVGs</td>
<td>Mandatory</td>
<td>Type of dataset from which the UVGs was obtained</td>
<td>[metagenome or viral target], [viral fraction metagenome (environment)], [sequence targeted metagenome (e.g. metagenome-viral) or viral fraction], [RNA metagenome (RNA-omics), sequence targeted viral RNA metagenome (RNA-omics)].</td>
<td>WGS metagenome, 2011.0, release 17.</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>assembly software</td>
<td>Mandatory</td>
<td>Tools used for assembly and/or binning, including version number and parameters</td>
<td>[software] (version) (parameters)</td>
<td>software</td>
<td>metapładys, 2011.0, release 17.0, default parameters otherwise.</td>
<td>Identical MIMAD / MISAQ</td>
</tr>
<tr>
<td>viral identification software</td>
<td>Mandatory</td>
<td>Tools used for the identification of UVGs as a viral genome, software or protocol name, including version number, parameters, and cutoffs used (see Table 1)</td>
<td>[software] (version) (parameters)</td>
<td>software</td>
<td>ViroFlux, 1.1.4, Virome database, category 2</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>predicted genome type</td>
<td>Mandatory</td>
<td>Type of genome predicted for the UVG</td>
<td>[DNA]</td>
<td>DNA</td>
<td>DNA-segmented, non-segmented, circularized</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>predicted genome structure</td>
<td>Mandatory</td>
<td>Expected structure of the viral genome</td>
<td>[RNA]</td>
<td>RNA</td>
<td>non-segmented, segmented</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>detection type</td>
<td>Mandatory</td>
<td>Type of UVG detection unit</td>
<td>[DNA/RNA]</td>
<td>DNA/RNA</td>
<td>segmented, non-segmented, circularized</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>Genome quality</td>
<td>assembly quality</td>
<td>Mandatory</td>
<td>The assembly quality categories, specific for viral genomes, are based on sets of criteria as follows:</td>
<td>[Estimated genome quality score]</td>
<td>[Estimated genome quality score]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>number of contigs</td>
<td>Mandatory</td>
<td>Total number of predicted contigs</td>
<td>[number]</td>
<td>number</td>
<td>[number]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>completeness score</td>
<td>Conditions</td>
<td>Estimated completeness of the UVG</td>
<td>[quality] (percentage)</td>
<td>quality</td>
<td>[percentage]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>completeness approach</td>
<td>Conditions</td>
<td>Approach used to estimate the UVG completeness, including reference genome in group used, and contig feature suggesting a complete genome</td>
<td>[method]</td>
<td>method</td>
<td>[method]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>Genome annotation</td>
<td>feature prediction</td>
<td>Conditions</td>
<td>Method used to predict UVGs features such as ORFs, protein, nucleotide sites, etc.</td>
<td>[software] (version) (parameters)</td>
<td>Prodigal, 2.6.3, default parameters</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>reference database</td>
<td>Conditions</td>
<td>List of databases used for ORF annotation, along with version number and reference to website or publication</td>
<td>[database], [version]</td>
<td>database</td>
<td>[version]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>similarity search method</td>
<td>Conditions</td>
<td>Tools used to compare DNA with database, along with version number and cutoffs used</td>
<td>[software] (version)</td>
<td>software</td>
<td>[version]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>taxonomic classification</td>
<td>Conditions</td>
<td>Method used for taxonomic classification, along with reference database used, classification scheme, and thresholds used to classify new genomes</td>
<td>[software] (version) (parameters)</td>
<td>software</td>
<td>[version]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>VGTU classification approach</td>
<td>Conditions</td>
<td>Tools and thresholds used to compare sequences when applying &quot;species-level&quot; VGTU</td>
<td>[software] (version) (parameters)</td>
<td>software</td>
<td>[version]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>VGTU database</td>
<td>Conditions</td>
<td>Reference database (e.g. sequences self-generated as part of the study) used to cluster new genomes in &quot;species-level&quot; VGTU</td>
<td>[database], [version]</td>
<td>database</td>
<td>[version]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>host prediction approach</td>
<td>Conditions</td>
<td>Tools or approach used for host prediction</td>
<td>[method]</td>
<td>method</td>
<td>[method]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>host prediction estimated accuracy</td>
<td>Conditions</td>
<td>For each tool or approach used for host prediction, estimated false-positive rate should be included, rather than the overall false-positive rate.</td>
<td>[accuracy]</td>
<td>accuracy</td>
<td>[accuracy]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>viral SAG metadata</td>
<td>sorting technology</td>
<td>Conditions</td>
<td>Method used to sort bacterial cells or particles of interest</td>
<td>[software] (version) (parameters)</td>
<td>SAMtools, 1.3, default parameters</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>single-cell or viral particle group</td>
<td>Conditions</td>
<td>Name of the kit used for library preparation used for single-cell or particle group</td>
<td>[kit]</td>
<td>kit</td>
<td>[kit]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>single-cell or viral particle type kit protocol</td>
<td>Conditions</td>
<td>Method used to amplify, generic DNA preparation for sequencing</td>
<td>[method]</td>
<td>method</td>
<td>[method]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>library amplification kit</td>
<td>Conditions</td>
<td>Kit used to amplify genomic DNA in preparation for sequencing</td>
<td>[kit]</td>
<td>kit</td>
<td>[kit]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>viral MGA metadata</td>
<td>size fraction selected</td>
<td>Conditions</td>
<td>Filtering pore size used in sample preparation</td>
<td>[filter] (pore size)</td>
<td>[filter] (pore size)</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>virus enrichment approach</td>
<td>Conditions</td>
<td>List of approaches used to enrich the sample for viruses, if any</td>
<td></td>
<td>list</td>
<td>[list]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>nucleic acid extraction</td>
<td>Conditions</td>
<td>Description of the approach used for whole genome amplification, if any</td>
<td>[software] (version)</td>
<td>software</td>
<td>[version]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>WGA amplification approach</td>
<td>Conditions</td>
<td>Method used to amplify genomic DNA in preparation for sequencing</td>
<td></td>
<td>method</td>
<td>[method]</td>
<td>Comparable to and adapted from MIMAD / MISAQ</td>
</tr>
<tr>
<td>binding parameters</td>
<td>Conditions</td>
<td>The parameters that have been applied during the extraction of genomic from metagenomic datasets</td>
<td>[parameters]</td>
<td>parameters</td>
<td>[parameters]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>binding software</td>
<td>Conditions</td>
<td>Tools used for the extraction of genomic from metagenomic datasets</td>
<td>[software]</td>
<td>software</td>
<td>[software]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>assembly post binding</td>
<td>Conditions</td>
<td>Tools used to determine the genome coverage if coverage is used as a binning parameter or the extraction of genome from metagenomic datasets</td>
<td>[software]</td>
<td>software</td>
<td>[software]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>MGA coverage software</td>
<td>Conditions</td>
<td>Tools used to determine the genome coverage if coverage is used as a binning parameter or the extraction of genome from metagenomic datasets</td>
<td>[software]</td>
<td>software</td>
<td>[software]</td>
<td>New and specific to UVGAs</td>
</tr>
<tr>
<td>Category</td>
<td>Genome fragment(s)</td>
<td>High-quality draft genome</td>
<td>Finished genome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>Single or multiple fragments</td>
<td>Single or multiple fragments where gaps span (mostly) repetitive regions.</td>
<td>Single contiguous sequence (per segment) without gaps or ambiguities.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td>&lt; 90% expected genome size or no expected genome size</td>
<td>Complete or ≥ 90% of expected genome size</td>
<td>Complete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required features</td>
<td>Minimal annotation</td>
<td>Minimal annotation</td>
<td>Comprehensive manual review and editing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Summary of required characteristics for each category. Complete genomes include sequences detected as circular, with terminal inverted repeats, or for which an integration site is identified.

Supplementary Figures and Tables Legend

Supplementary Figure 1. Comparison of UViG recovery from microbial (“M”) and viral (“V”) metagenomes originating from the same Tara Oceans samples. Top panel represents the number of distinct virus contigs ≥ 10kb identified in each dataset, and the bottom panel depicts the ratio of “shared” (i.e. detected in both viral and microbial fraction of the sample) and “unique” (detected only in one fraction) contigs in each fraction. Datasets were originally analyzed in refs. 38,39.

Supplementary Figure 2. Genome length variation for different types of viruses and different taxonomic ranks. Genome length of virus genomes from NCBI RefSeq were compared at different taxonomic ranks and are presented separately for four main types of viruses (dsDNA, ssDNA, RNA and reverse-transcribing, viroids and satellites). Genome length variation was calculated as a coefficient of variation at the genus rank, i.e. standard deviation of genome length in the genus divided by average genome length in the genus (for genera with >1 genomes). Underlying data are available in Supplementary Table 5. Boxplots lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), while whisker extend from the nearest hinge to the smallest/largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). dsDNA: double-stranded DNA; ssDNA: single-stranded DNA.
Supplementary Figure 3. Pairwise Average Nucleotide Identity (ANI) and Alignment Fraction (AF) for NCBI Viral RefSeq genomes (A) and IMG/VR (B). Only genome pairs with ANI >60% and AF >20% were considered. ANI and AF were binned in 1% intervals, and are represented here as a heatmap (i.e. cell coloring represents the number of pairwise comparisons at the corresponding ANI and AF intervals). On the top right corner (i.e. AF and ANI close to 100%), three main groups of genome pairs are delineated with black dashed circles, and the proposed standard cutoff is highlighted in dark red. Note that for this clustering, the cutoff was applied as follows: pairs of genomes with ≥ 85% AF were first selected, and whole genome (wg) ANI was then calculated by multiplying the observed ANI by the observed AF. This wgANI was then compared to the corresponding whole genome ANI cutoff (i.e. 95% ANI * 85% AF = 80.75% wgANI). This allows for hits with ≤ 95% ANI but ≥ 85 % AF to be considered as well, i.e. a pair of genomes with 90% ANI on 100% AF would be considered as “passing” the cutoff. Examples of genome comparisons for each group are presented in Supplementary Figure 4 (from NCBI Viral RefSeq).

Supplementary Figure 4. Examples of pairwise genome comparisons from the three groups of genome pairs highlighted on Supplementary Figure 3. For each example, nucleotide similarity (blastn) and amino acid similarity (tblastx) are displayed, alongside the ANI, AF, and wgANI (i.e. ANI over the whole length of the shorter genome).

Supplementary Figure 5. Estimation of whole genome (wg) ANI from fragmented genomes. To evaluate the impact of genome fragmentation on wgANI estimation, pairs of genomes from NCBI RefSeq with wgANI ≥ 70% and ≥ 20kb were selected, random fragments were generated (from 1 to 45kb) from one of the two genomes, and then compared to the other complete genome. The resulting wgANI between the fragment and complete genome was then compared with the original values estimated from the two complete genomes (y-axis). Boxplots lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), while whisker extend from the nearest hinge to the smallest/largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles).

Supplementary Table 1. List of metadata from previous standards relevant for UViGs. The last 3 columns include information about whether an item is mandatory (M), conditional mandatory (C), optional (X), environment-dependent (E) or not applicable (-) in the MIMAG, MISAG, and MIUViG checklists. Items for which the MIUViG requirement differed from MIMAG and MISAG requirements are highlighted in yellow.
**Supplementary Table 2.** Comparison between UViGs categories and the quality categories proposed for small DNA/RNA virus whole-genome sequencing for epidemiology and surveillance by Ladner et al. 21.

**Supplementary Table 3.** List and characteristics of tools used to identify virus sequences in mixed datasets.

**Supplementary Table 4.** Variation in genome length for virus families and genera with 2 or more genomes, from NCBI RefSeq v83.

**Supplementary Table 5.** List of potential marker genes for virus orders, families, or genera, based on the VOGdb v83 ([http://vogdb.org/](http://vogdb.org/)).

**Supplementary Table 6.** List of UViGs from the GOV dataset7 considered as high-quality drafts or finished genomes. Example of UViGs classified as genome fragments with varying size and completeness estimations are also included at the bottom of the table. For genome fragments for which no complete genome is available, the expected genome size is displayed as greater than the size of the largest contig in the cluster (e.g. “> 20,000bp”), and no estimated completeness can be provided for these contigs.

**Supplementary Table 7.** List of databases providing collections of HMM profiles for virus protein families. This topic has been recently reviewed in Reyes et al. 76.

**Supplementary Table 8.** Current species demarcation criteria from ICTV 9th and 10th reports.

**Supplementary Table 9.** Approaches available for *in silico* host prediction.
Acknowledgments and funding sources

This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231 for SR; the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004 for BED; the Intramural Research Program of the National Library of Medicine, National Institutes of Health for EVK, IKM, JRB and NY; the Virus-X project (EU Horizon 2020, no. 685778) for FE and MK; Battelle Memorial Institute's prime contract with the US National Institute of Allergy and Infectious Diseases (NIAID) under Contract No. HHSN272200700016I for JHK; the GOA grant “Bacteriophage Biosystems” from KU Leuven for RL; the European Molecular Biology Laboratory for CA and GRC; National Science Foundation award 1456778, National Institutes of Health awards R01 AI132581 and R21 HD086833, and The Vanderbilt Microbiome Initiative award for SRB; National Science Foundation awards DEB-1239976 for MB and KR and DEB-1555854 for MB; the NSF Early Career award DEB-1555854, and NSF Dimensions of Biodiversity #1342701 for KCW and RAD; the Agence Nationale de la Recherche JCJC grant #ANR-13-JSV6-0004 and Investissements d'Avenir Méditerranée Infection #10-IAHU-03 for CD; the Gordon and Betty Moore Foundation Marine Microbiology Initiative #3779 and the Simons Foundation for JAF; the French government ‘Investissements d’Avenir’ programme OCEANOMICS ANR-11-BTBR-0008 and European FEDER Fund 1166-39417 for PH; the National Science Foundation award #1801367 and C-DEBI Research Grant for JML; the Gordon and Betty Moore Foundation grant #5334 and Ministry of Economy and Competitivity refs. CGL2013-40564-R and SAF2013-49267-EXP for MM; the Grant-in-Aid for Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports, and Technology (MEXT) of Japan No. 16H06429, 16K21723, and 16H06437 for HO and TY; the Ministry of Economy and Competitivity ref CGL2016-76273-P (cofunded with FEDER funds) for FRV; the Gordon and Betty Moore Foundation awards #3305 and #3790 and NSF Biological Oceanography OCE#1536989 for MBS; the ETH Zurich and Helmut Horten Foundation, and the Novartis Foundation for medical-biological Research (#17B077) for SS; NSF Biological Oceanography Grant #1635913 for RVT; the Australian Research Council Future Fellowship FT120100480 for NSW; a Gilead Sciences Cystic Fibrosis Research Scholarship for KLB; Gordon and Better Moore Foundation Grant 4971 for SWW; the NSF EPSCoR grant 1736030 for KEW; the National Science Foundation award DEB-4W4596 and National Institutes of Health award R01 GM117361 for MJY; the Gordon and Betty Moore Foundation #7000, and the National Oceanic and Atmospheric Administration (NOAA) under award NA15OAR4320071 for LZA. DDBJ is supported by ROIS and MEXT. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the
official policies, either expressed or implied, of the US Department of Health and Human Services or of the institutions and companies affiliated with the authors. BED, AK, MK, JK, AV, are members of the ICTV Executive Committee, but the views and opinions expressed are those of the authors and not those of the ICTV.

**Competing interests**
The authors declare no competing interests.
Affiliations
1 U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA
2 Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
3 Theoretical Biology and Bioinformatics Utrecht University, Utrecht, The Netherlands
4 Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands
5 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
6 Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, Ontario N1G 2W1, Canada
7 Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Paris, 75015, France
8 Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD 21702, USA.
9 KU Leuven, Laboratory of Gene Technology, 3001 Heverlee, Belgium
10 The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life sciences, Arizona State University, Tempe, AZ 85287-5001, USA
11 Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Observatory, Cape Town, 7925, South Africa
12 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge. CB10 1SD, United Kingdom
13 Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, 11562 Cairo, Egypt
14 Departments of Biological Sciences and Pathology, Microbiology, and Immunology, Vanderbilt Institute for Infection, Immunology and Inflammation, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA
15 European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany
16 College of Marine Science, University of South Florida, Saint Petersburg, FL 33701, USA
17 Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA
18 Aix-Marseille Université, CNRS, MEPHI, IHU Méditerranée Infection, Marseille, 13005, France
19 Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA
20 University of California, Davis, Department of Plant Pathology, Davis, CA 95616, USA
21 Université Clermont Auvergne, CNRS, LMGE, Clermont-Ferrand, F-63000, France
22 University of Southern California, Los Angeles, CA 90089, USA
23 Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO UM 110, Marseille,13288, France
24 Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, Queensland 4072, Australia
25 Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ 85719, USA
26 BIO5 Research Institute, University of Arizona, Tucson, AZ 85719, USA
27 Department of Marine Biology, Texas A&M University at Galveston, Galveston, TX 77553, USA
28 DDBJ Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
29 Department of Physiology, Genetics and Microbiology, University of Alicante, Alicante, Spain
30 Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan
31 Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, 78352, France
32 Department of Biology, Loyola University Chicago, Chicago, IL 60660, USA
33 Bioinformatics Program, Loyola University Chicago, Chicago, Illinois, USA
34 Department of Computer Science, Loyola University Chicago, Chicago, Illinois, USA
35 Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, Research Network “Chemistry Meets Biology”, University of Vienna, 1090 Vienna, Austria
36 Max Planck tandem group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia
37 Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Alicante, Spain
38 University of Maryland School of Medicine, Baltimore, MD 21201, USA
39 Center for Microbial Oceanography: Research and Education, Department of Oceanography, University of Hawai‘i at Mānoa, Honolulu, Hawai‘i 96822, United States
40 Department of Microbiology, Ohio State University, Columbus, OH 43210, United States of America
41 Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH 43210, United States of America
42 ETH Zurich, Department of Biology, 8093 Zurich, Switzerland
43 Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
44 Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
45 Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
46 Institute of Oceans and Fisheries, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
47 School of Biosciences, University of Exeter, Exeter EX4, UK
48 Department of Microbiology, Oregon State University, OR 97331, USA
49 Australian Institute of Marine Science, Townsville, Queensland 4811, Australia
50 Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA
51 Department of Microbiology, The University of Tennessee, Knoxville, TN 37919, USA
52 University of Delaware, Delaware Biotechnology Institute, Newark, DE 19711, USA
53 Microbial Physiology Group, Max Planck Institute for Marine Microbiology Bremen, Germany
54 Graduate School of Agriculture, Kyoto University, Kitashirakawa-Oiwake, Sakyō-ku, Kyoto, 606-8502, Japan
55 Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT 59715, USA
56 J Craig Venter Institute, La Jolla, CA 92037, USA
57 Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093, USA
References


