AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Check for updates

| Bacteriophages | Research Article

A natural ANI gap that can define intra-species units of bacteriophages and other viruses

Borja Aldeguer-Riquelme,[1] Roth E. Conrad,[1] Josefa Antón,[2] Ramon Rossello-Mora,[3] Konstantinos T. Konstantinidis[1]

AUTHOR AFFILIATIONS See affiliation list on p. 14.

ABSTRACT Despite the importance of intra-species variants of viruses for causing disease and/or disrupting ecosystem functioning, there is no universally applicable standard to define these. A (natural) gap in whole-genome average nucleotide identity (ANI) values around 95% is commonly used to define species, especially for bacteriophages, but whether a similar gap exists within species that can be used to define intra-species units has not been evaluated yet. Whole-genome comparisons among members of 1,016 bacteriophage (*Caudoviricetes*) species revealed a region of low frequency of ANI values around 99.2%–99.8%, showing threefold or fewer pairs than expected for an even distribution. This second gap is prevalent in viruses infecting various cultured or uncultured hosts from a variety of environments, although a few exceptions to this pattern were also observed (3.7% of total species) and are likely attributed to cultivation biases or other factors. Similar results were observed for a limited set of eukaryotic viruses that are adequately sampled, including SARS-CoV-2, whose ANI-based clusters matched well with the WHO-defined variants of concern, indicating that our findings from bacteriophages might be more broadly applicable and the ANI-based clusters may represent functionally and/or ecologically distinct units. These units appear to be predominantly driven by (high) ecological cohesiveness coupled to either frequent recombination for bacteriophages or selection and clonal evolution for other viruses such as SARS-CoV-2, indicating that fundamentally different underlying mechanisms could lead to similar diversity patterns. Accordingly, we propose the ANI gap approach outlined above for defining viral intra-species units, for which we propose the term genomovars.

IMPORTANCE Viral species are composed of an ensemble of intra-species variants whose individual dynamics may have major implications for human and animal health and/or ecosystem functioning. However, the lack of universally accepted standards to define these intra-species variants has led researchers to use different approaches for this task, creating inconsistent intra-species units across different viral families and confusion in communication. By comparing hundreds of mostly bacteriophage genomes, we show that there is a widely distributed natural gap in whole-genome average nucleotide identity values in most, but not all, of these species that can be used to define intra-species units. Therefore, these results advance the molecular toolbox for tracking viral intra-species units and should facilitate future epidemiological and environmental studies.

KEYWORDS ANI gap, bacteriophages, genomovars, strains

Recognized viral species are often not homogeneous but consist of phenotypically and genotypically distinct variants (or intra-species populations), each of which could have distinct and major impacts on human and animal health, trophic webs, and ecosystem functioning. For example, bacteriophages may exert population control

Editor Vaughn S. Cooper, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

Address correspondence to Konstantinos T. Konstantinidis, kostas@ce.gatech.edu.

The authors declare no conflict of interest.

See the funding table on p. 14.

Received 11 June 2024
Accepted 24 June 2024
Published 22 July 2024

of relevant environmental bacteria, such as *Synechococcus*, whose temporal genetic diversity co-varies with cyanophages due to viral infections (1). Additional studies have documented the constant rise and fall of genome variants as the underlying mechanism for the overall stable abundance of bacteriophage species over time (2, 3). Another example was the emergence of a new coronavirus variant in 2019 that initiated the SARS-CoV-2 pandemic, which has killed an estimate of 6.95 million people across the world (World Health Organization, WHO). Beyond humans, the populations of European rabbits decreased by 60%–70% when a new, more lethal variant of the rabbit hemorrhagic disease virus appeared (4). As a consequence, species that feed on rabbits, such as the Iberian lynx, followed the decline in rabbit populations in a clear example of how viral variants can influence trophic webs. Therefore, viral variants are apparently an important unit of viral diversity. However, a widely accepted definition of what a viral variant is and how much diversity it should encompass remains elusive.

The International Committee on Taxonomy of Viruses (ICTV) oversees the development, regulation, and maintenance of a universal taxonomic classification of viruses. However, the ICTV does not regulate the classification and nomenclature of organisms below the species level nor does it provide a definition or criteria for intra-species units. The lack of a clear definition has led to confusion; most notably, the same (intra-species) terms, such as strain or variant, have been frequently used with different standards and meanings. For instance, van Regenmortel defined strain as a virus with unique phenotypic characteristics (5). Accordingly, viruses with the same phenotype but different genomic sequences are considered to be the same strain, and the author reserves the term "variant" for these cases. Others, however, employed "variant" to distinguish between viruses with different phenotypes (6, 7), contrasting with the definition given by van Regenmortel. The variety of definitions is also reflected in the diversity of criteria used to delineate intra-species units. Some authors used sequence identity values or clustering patterns of phylogenetic trees of single genes (8, 9), while others employed whole-genome similarities, with a range of identity thresholds (98%–100%) (10–12). It is important to note that the selection of the gene to use is typically an arbitrary decision, while different genes might produce different results (10). Furthermore, the whole-genome sequence identity thresholds were primarily employed for practical reasons and convenience, but their biological relevance in nature, if any, remains unknown.

Previous efforts have reported the existence of sequence-discrete viral units with intra-unit genome-wide ANI values being usually greater than 95%, a threshold that has been proposed as a reference standard to define viral species (or viral operational taxonomic units [vOTUs]), especially for bacteriophages (13–15). That is, the ANI values among genomes of the same species are higher than 95%, contrasting with <90% ANI to members of other species. Thus, there appears to be a natural gap in ANI values distribution between species, although the exact range of ANI values corresponding to the gap may differ in some species, with 90%–95% ANI being the most observed area by far. This is similar to the 95% ANI threshold commonly employed for microbial species definition (16, 17), and thus, sequence-discrete species seem to exist similarly for both microbes and their viruses. Recently, comparison of intra-species ANI values among genomes of the same bacterial species revealed the existence of a similar ANI gap between 99.2% and 99.8%, which has been proposed as a threshold to define intra-species units based on genomic data (18). Here, we aimed to test whether a similar intra-species gap exists for viruses, which can be used to define or refine existing intra-species units and assess the underlying molecular and/or ecological mechanisms for any such gap.

RESULTS AND DISCUSSION

An ANI gap within cultured viral species around 99.2%–99.8%

We tested the existence of a similar ANI gap to that observed previously within prokaryotic species (18) among viral genomes using a data set that included 75,012 bacterial viral (i.e., bacteriophage) genomes (viral isolates and prophages), which

represented 306 distinct species with a minimum of 20 genomes per species (Table S1). The ANI histogram based on all possible 51,522,103 intra-species pairwise genome comparisons revealed a substantial gap between 99.2% and 99.8% ANI (Fig. 1A). To ensure that this gap was not due to just a few highly sampled species, we subsampled

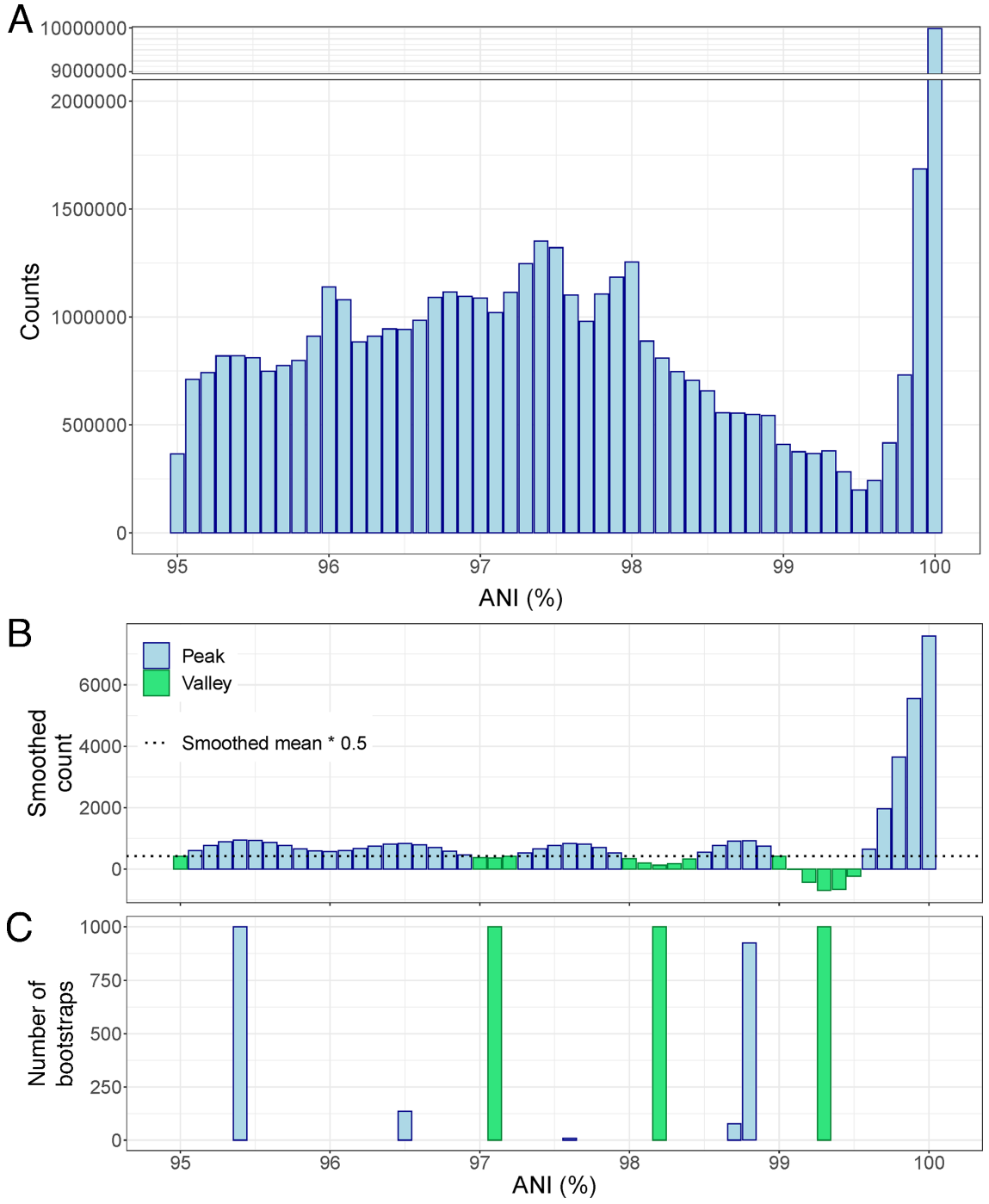

FIG 1 An intra-species ANI gap exists at around 99.2%–99.8% and is consistently detected after 1,000 subsampling rounds. (A) Histogram of ANI values without normalization to the same number of pairs per species. A total of 51,522,103 pairwise comparisons between 75,012 bacteriophage genome pairs showing >95% were used. Note the pronounced drop in pairs between 99% and 99.8% ANI (~300,000 per 0.1 units of ANI) compared to lower and higher ANI values (1,300,000 pairs at 97.4% ANI and 10,000,000 pairs at 100% ANI). (B) Histogram of smoothed counts (sm_method="gam") using data subsampled to get the same number of pairs per species (n = 150; see Materials and Methods). Green bars indicate valleys, while blue bars indicate intermediate values and peaks. (C) Peaks (blue) and valleys (green) detected after 1,000 bootstrap events of subsampling to 150 genome pairs per species and automatic peak and valley detection.

the data to the same number of pairs (150) per species. This subset corroborated the existence of the gap revealed based on the full data (Fig. S1). Specifically, we found 854 values to fall within the 95.5% ANI bin vs 316 for the 99.5% ANI bin for the subset. We estimated that if the ANI values were randomly and evenly distributed between 95% and 100% ANI, we would have expected 847 comparisons per every 0.1% bin of ANI (43,200 pairs in total, divided by 51 bins between 95% and 100%). In contrast, within the 99.4%–99.6% bins, there were only 374, 316, and 367 pairs, which are about 2.7 times fewer pairs than expected by chance. Indeed, bootstrapped (1,000) subsampling to the same number of pairs per species followed by automated peak and valley identification pointed to 99.3% as the deepest and most consistent valley in the ANI value distribution (Fig. 1B and C).

The species included in the data set belong to the Caudoviricetes class (realm *Duplodnaviria*) and infect 609 different bacterial host species classified in 15 phyla, associated with animals or humans as well as environmental sources, which highlights the diversity of the data set. Among the host species, *Escherichia coli* (19), *Mycobacterium smegmatis* (20), *Mycobacterium abscessus* (12), and *Serratia marcescens* (10) were the most predominant. Most of these viral species showed an intra-species ANI gap (see Fig. S2 for examples and https://github.com/baldeguer-riquelme/Viral-ANI-gap/ for all species evaluated). Indeed, 273 of the total 306 analyzed viral species (89.2%) showed an area of low frequency of pairs between 99.2% and 99.8% ANI when assessed individually, as opposed to collectively above (groups 1 and 2; see Fig. S3 for examples of each group). On the other hand, 26 species (8.5% of the total) did not show any peaks or valleys within the 99.2%–99.8% ANI, which do not provide evidence in favor or against the gap (undetermined distribution, group 3), and only seven species (2.5%) presented a contradictory distribution (group 4). That is, the latter species showed a peak rather than a valley of around 99.2%–99.8% (Table 1). Therefore, these results revealed a widely distributed, but not absolutely universal, natural genomic threshold for distinguishing intra-species units within the Caudoviricetes.

Support for the ANI gap by culture-independent, long-read metagenomic data

To assess whether or not culture-independent data are consistent with the existence of the ANI gap based on isolate genomes described above, we analyzed uncultured viral genomes recovered by fosmid sequencing and PacBio HiFi long-read metagenomes. Fosmids are cloning vectors of inserts up to 48 kbp long, thus allowing the recovery of complete or partial individual viral genomes. Similarly, PacBio HiFi provides high-quality consensus reads up to 25 kbp long. Therefore, both sequencing strategies could offer high resolution among co-occurring, uncultured viral genomes obtained directly from the environment with minimal sequencing error and bypassing isolation biases.

As shown in Fig. 2, the analysis of the uncultured viral genomes recovered by fosmid or long-read sequencing also supported the existence of the ANI gap. Indeed, 9 out of

TABLE 1 Classification of individual species based on the ANI distribution pattern[b]

	ANI gap				
	Group 1 (multiple clusters)	Group 2 (one cluster)	Group 3 (undetermined)	Group 4 (exceptions)	Total species
Prokaryotic	231 (75.5%)	42 (13.7%)	26 (8.5%)	7 (2.3%)	306
Fosmid	7 (53.8%)	2 (15.4%)	4 (30.8%)	0 (0%)	13
Long reads[a]	269 (38.6%)	305 (43.8%)	92 (13.2%)	31 (4.4%)	697

[a]For long reads, we found the gap to be shifted toward lower ANI values (98.8%–99.5%), so this range was employed to define the gap, and an average ANI ≥99.5% was used to define species in group 2.
[b]Group 1 refers to the species that show a valley between 99.2% and 99.8%; group 2 includes species with predominately high-identity genomes or highly clonal (average ANI >99.8%); group 3 represents species with no peak or valley between 99.2% and 99.8%, which is not consistent or contradictory with the gap (undetermined); and finally, group 4 includes species showing a peak rather than a valley between 99.2% and 99.8% (i.e., contradictory with the gap). The number of species within each group is shown, as well as the percentage they represent in parenthesis.

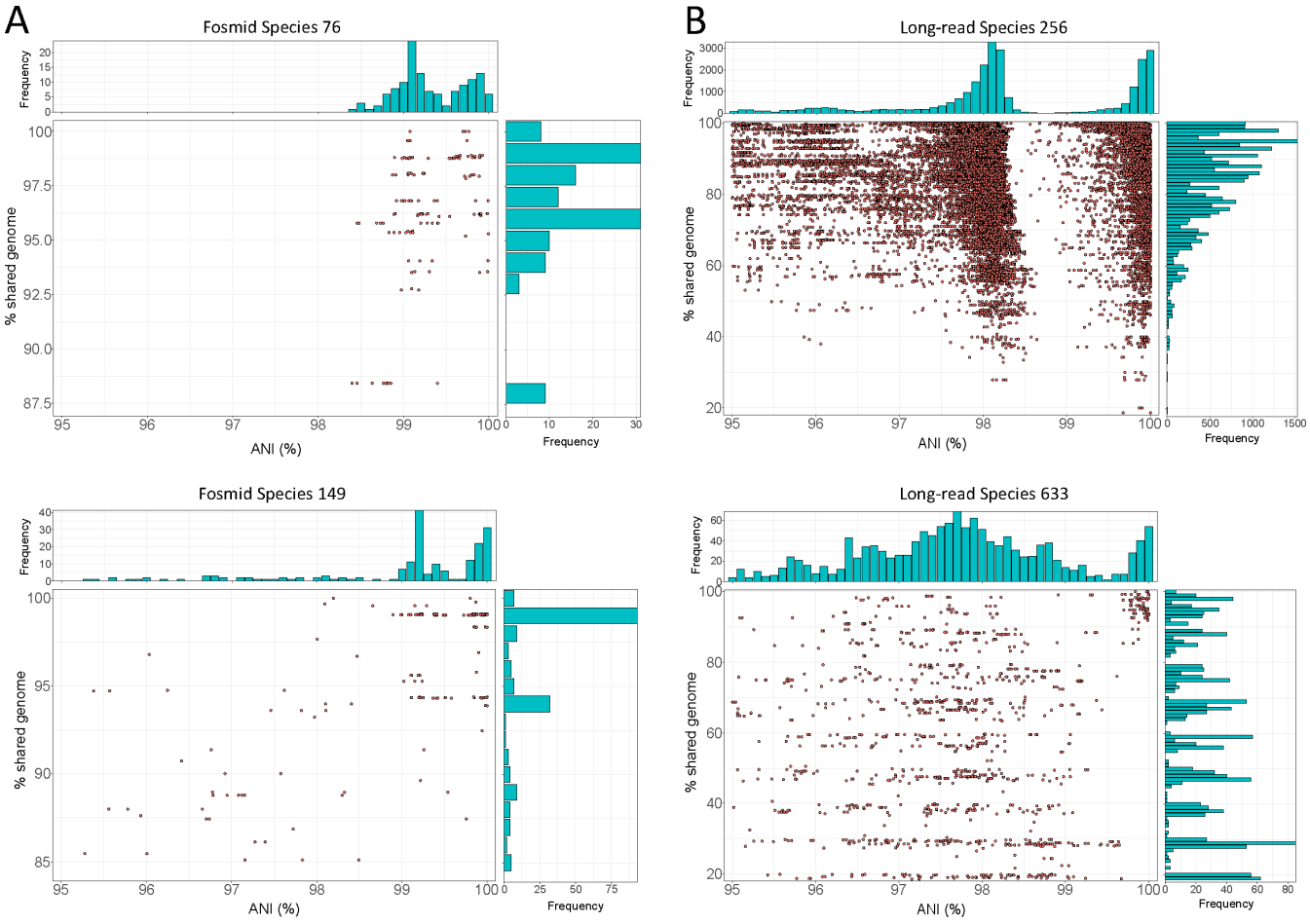

FIG 2 Fraction of shared gene content vs ANI for uncultured viral species recovered by fosmids (A) and long-read metagenomes (B). In the main panel of each figure, dots represent individual pairwise comparison of genomes of the same species, while histograms at the top and right side show the frequency of values for ANI and shared genome, respectively, similarly to Fig. 1. Note the low frequency of pairs with ANI values around 99.5%. Plots for each viral species with more than 10 genomes can be found in https://github.com/baldeguer-riquelme/Viral-ANI-gap/.

the total 13 species adequately covered by fosmid sequencing presented the ANI gap, 4 species displayed an undetermined distribution, and none were inconsistent with the gap, similar to the results reported above for isolate genomes (Table 1). For 9 of these 13 species, we detected overlapping ends, indicating complete and circular genomes that truly represent single, distinct species rather than different regions of the same genome. The latter cannot be completely ruled out for the remaining four species. Fosmid metadata did not provide helpful information related to the taxonomy and/or ecology of the corresponding species, and thus we can only conclude that these species represented marine taxa/samples, while their hosts remain unknown. Nevertheless, a previous study showed that these fosmids better represent the *in situ* abundant viruses than isolates (20) and are, most likely, a closer representation of the actual diversity in nature. Regarding long-read sequences, we did also observe a gap albeit slightly shifted toward lower ANI values (98.8%–99.5%). Remarkably, 574 out of a total of 697 detected species in the long-read data sets displayed a clear ANI gap (group 1: 269; group 2: 305), 92 did not present a peak or valley at the gap (undetermined species), and only 31 showed an incompatible distribution (Table 1). The species displaying the ANI gap represented three distinct environments (i.e., human gut, chicken gut, and seawater), supporting its widespread existence in nature. It should be mentioned that long-read data mostly represent fragmented genomes, resulting in a higher dispersion of ANI and percentage of shared genome values around the mean; nevertheless, the gap was

evident even for these partial genomes (Fig. 2B). Further analyses showed that the slight downward shift of the ANI gap observed for long-read data was due to (i) a relatively high-sequencing error rate (0.5% vs 0.01% in the isolate genomes described above) and (ii) the lack of an assembly step that corrects or minimizes the impact of such errors.

SARS-CoV-2 clusters defined by the ANI gap match WHO variants

The SARS-CoV-2 is probably the most sequenced virus to date, with more than 8 million genomes deposited in the NCBI database at the time of this writing. Furthermore, epidemiologic studies have provided detailed information on the phenotypic characteristics of the virus such as transmissibility or virulence of different viral variants. The WHO used these genotypic and phenotypic information to identify variants of concern (VOCs), which include the most dangerous variants and the main drivers of the SARS-CoV-2 pandemic (21, 22). To date, five SARS-CoV-2 variants, named Alpha, Beta, Gamma, Delta, and Omicron, have been declared as VOCs by the WHO (23), and an additional one, named Epsilon, has been also declared by the Centers for Disease Control and Prevention (CDC) of the United States (24).

To assess whether a similar ANI gap to the one revealed for bacteriophages above exists for human/animal viruses such as SARS-CoV-2 and minimize the potential bias introduced by different protocols, reagents, or assembly pipelines, we analyzed only the SARS-CoV-2 genomes deposited by the CDC (USA) and the Robert-Koch Institute (RKI, Germany), two of the main submitting institutions. Then, we randomly subsampled the data set to get the same number of genomes for each one of the six VOCs. The analysis of the CDC genomes showed a weak but evident signal of an ANI gap at around 99.8% (Fig. 3A). When the analysis was restricted to high-quality genomes only (without any *N*s), the gap was even more clearly observed (Fig. 3B), and the ANI values were less dispersed, especially at lower ANI values. This result highlights that undetermined positions (i.e., *N*s) increase the noise of the ANI calculation, and thus high-quality genomes should be used for accurate results, especially for relatively short genomes such as SARS-CoV-2. The genomes sequenced by the RKI yielded similar results (Fig. S4).

Remarkably, the distribution of genome pairs into the same or different variants overlapped, almost perfectly, with the highest and lowest peaks, respectively (Fig. 3C). Considering the ANI value with the lowest number of pairs (99.83%) as a threshold to define variants, we found that 99.4% of all pairs above this threshold represented genomes belonging to the same variant. Conversely, only 3.2% of all pairs below 99.83% ANI involved genomes of the same variant. This result indicates that classification and identification of variants based on the intra-species ANI gap produce almost the same result as the combination of genotypic and phenotypic data employed by the international institutions. While the latter data are certainly needed for other purposes such as virulence assessment, the ANI gap offers a simple and complementary approach to identify viral variants. Furthermore, this result demonstrates that the resulting ANI-based clusters are associated with significantly distinct phenotypic characteristics. While we are able to confirm the latter only for the SARS-CoV-2 based on the abovementioned results, we hypothesize that clusters defined by the ANI gap within other viral species may also present significant phenotypic and/or ecological differences.

In addition to SARS-CoV-2, the existence of the ANI gap was also examined for a small set of eukaryotic viruses that have been adequately sampled and belonged to the *Duplodnaviria*, *Riboviria,* and *Varidnaviria* realms. Similarly to the results reported above for bacteriophages, the gap was observed for a variety of viruses that infect humans, mosquitoes, birds, ruminants, or pigs (Fig. S5). All three realms included species that showed a clear ANI gap (Table S1). However, given the small size of the eukaryotic data set and the fact that different thresholds for intra-species units are frequently used for these viruses compared to bacteriophages (19, 25–27), the results reported here for the eukaryotic genome data set cannot be broadly extrapolated to all (or most) eukaryotic viruses. Nonetheless, the fact that at least some eukaryotic viruses display the intraspecies gap indicates that the gap might be a widespread feature of viruses,

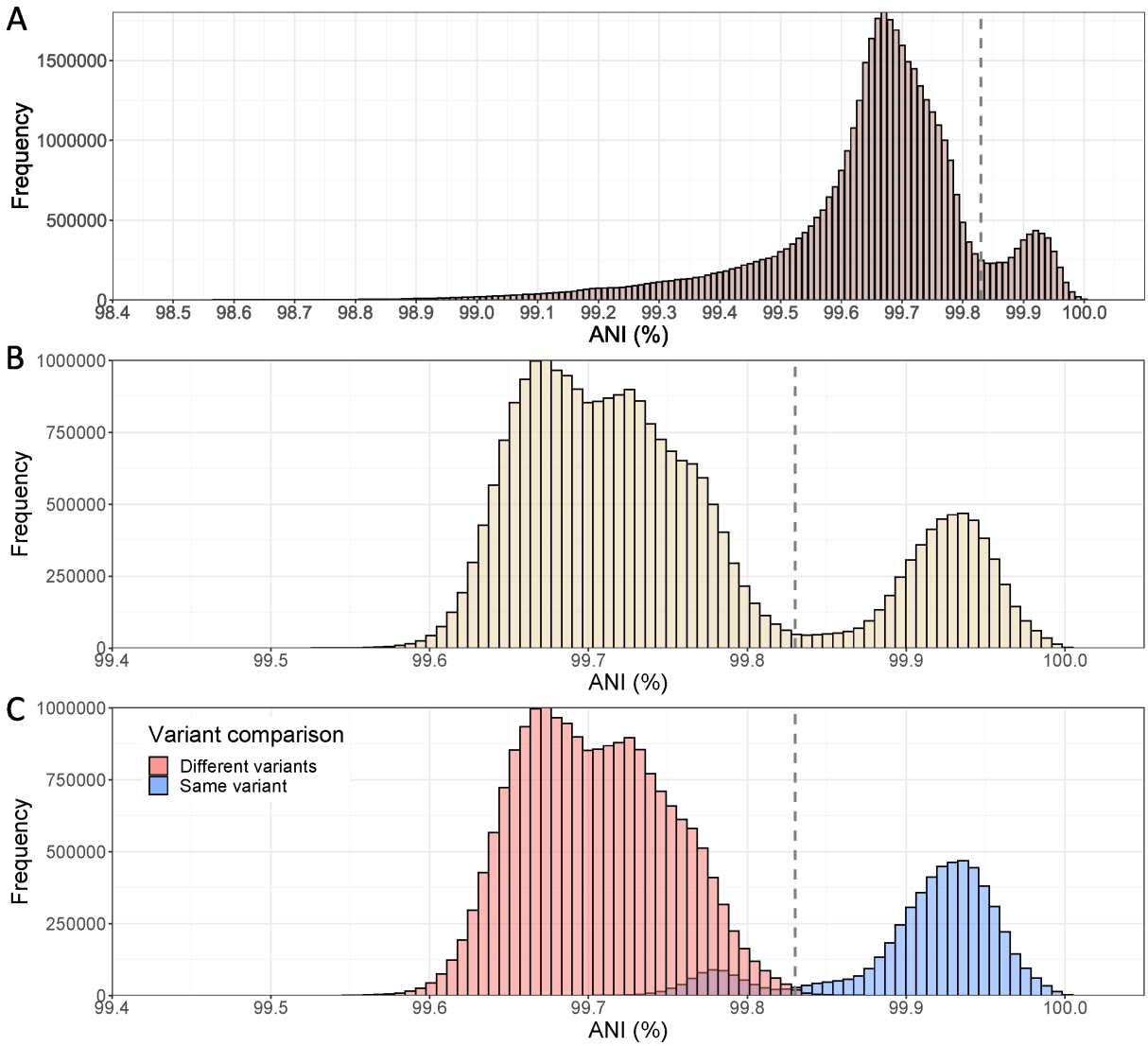

FIG 3 ANI histograms of the SARS-CoV-2 genomes sequenced by the US CDC. (A) Histogram with a random subset of 6,481 genomes (*N*s allowed) belonging to the Alpha (*n* = 1,250), Beta (*n* = 534), Delta (*n* = 1,250), Gamma (*n* = 1,250), Epsilon (*n* = 947), and Omicron (*n* = 1,250) VOCs was analyzed, that is, 41,996,880 pairs in total. A subtle ANI gap at around 99.8% can be observed (see vertical line). (B) Histogram with 5,041 high-quality genomes (i.e., without *N*s) belonging to the Alpha (*n* = 1,250), Beta (*n* = 99), Delta (*n* = 1,250), Gamma (*n* = 1,035), Epsilon (*n* = 157), and Omicron (*n* = 1,250) VOCs was analyzed, that is, 27,620,280 pairs in total. After removing genomes with *N*s, the data reveal a clearer bimodal distribution and a more pronounced ANI gap for the SARS-CoV-2 genomes. (C) Histogram shows the same data as in B, but the bars are colored based on whether the genomes compared are assigned to the same variants by WHO (in blue) or not (in pink). Note the limited overlap between the two groups.

not only bacteriophages. Interestingly, a threshold of 1% and 0.5% nucleotide differences between genomes was previously proposed to define human *Alphapapillomavirus* lineages and sublineages, respectively (28), and our 99.8% ANI threshold closely matches WHO's designation of VOCs for SARS-CoV-2. Future research should more rigorously assess the existence of an intra-species ANI gap in eukaryotic viruses by sequencing, for instance, a broader range of species.

Definition of intra-species clusters: a proposal for the term genomovar

Our results show that intra-species units of many bacteriophages may be distinguishable based on their ANI values. Given the multiple definitions of what constitutes a viral strain or a genetic variant, we propose the 99.5% ANI clusters, as the mean of the 99.2–

99.8 value, to be referred to as genomovars. This term was proposed decades ago to distinguish bacterial groups that typically belong to the same species and show distinct genotypic and phenotypic features, but these features did not represent enough of a diagnostic phenotype to qualify the groups as distinct species (29, 30). It has been recently proposed to use the term genomovar to refer to the intra-species genomic clusters of prokaryotic organisms that share more than 99.5% ANI because the term is well fit for this purpose (31). Therefore, our proposal herein is to use the same term for both viral and bacterial intra-species clusters in order to provide consistency and facilitate communication. We suggest the midpoint value (i.e., 99.5% ANI), rather than the upper value (i.e., 99.8% ANI), of the gap as a more conservative threshold, and in order to account for the variation in ANI gap values observed among different species. However, this ANI threshold should only be considered as a practical and convenient standard to define intra-species units and genomovars, and researchers are encouraged to adjust this threshold to better match the ANI value distribution revealed by the data of their phage of interest.

A remaining question is how to call the genomes whose pairwise identities fall within the area of the valley (gap). These genomes might represent an intermediate state of evolution toward becoming a new genomovar or members of a genomovar that has not been adequately sampled due, for instance, that it thrives in other conditions or sites than those preferably sampled to date. It is difficult at present to ascertain which of these scenarios is true and thus, what the biological meaning and classification of these intermediate genomes should be. Therefore, we recommend not to assign these intermediate genomes to an existing genomovar and instead call them as unassigned until further information from different sources is available (e.g., phenotypic properties) or designate them as new/novel genomovars. The latter option may be more practical, in our view.

What are the underlying mechanism(s) for the 99.5% ANI gap?

Viral infection can proceed through a lytic cycle, where a virus infects a host cell, replicates inside the cell, and finally, be released to the extracellular media as infectious virions, or through a lysogenic cycle, where the virus is first integrated into the host genome, often via a non-homologous recombination mechanism, until a signal activates the lytic cycle. These two infection strategies can generate different events and types of recombination. During lytic cycles, viral genome recombination may occur when two viruses co-infect the same cell, a phenomenon that has been observed in up to 50% of the total infected cells in the surface of the oceans and other environments (32). Once both genomes are inside the cell, recombination can take place coupled to replication (33), generating the widely described mosaicism of viral genomes (34). On the other hand, lysogenic viruses can incorporate some host genes into their genome during excision, generating a new recombinant virion. Thus, during lytic cycles, recombination usually—but not exclusively—happens between pairs of viruses, while in lysogenic cycles, recombination occurs mostly between a virus and its host. Since recombination is apparently an important mechanism that could drive viral genome evolution (33), we examined if it could be the underlying mechanism that maintains the intra-species ANI gap. Specifically, we tested the hypothesis that recent recombination is more frequent, and unbiased across the genome, within a cluster (e.g., a species or a genomovar) vs between clusters of genomes and thus can serve as the mechanism of cohesion for the cluster. For this, we first measured the fraction of genes showing >99.8% identity between genome pairs (F100) relative to the expected number of such high-identity genes based on their ANI value and assuming no recombination (F100 expected; see Materials and Methods for details), as a proxy for recent recombination events (Fig. 4A). For this analysis, we focused on the *Salinibacter ruber* phage species for which more genomes are available from a natural population (35). The data set included 177 high-quality viral genomes able to infect the same host (*Sal. ruber* strain M8) that were recovered from two ponds of the same saltern in Majorca Island, Spain, sampled (just)

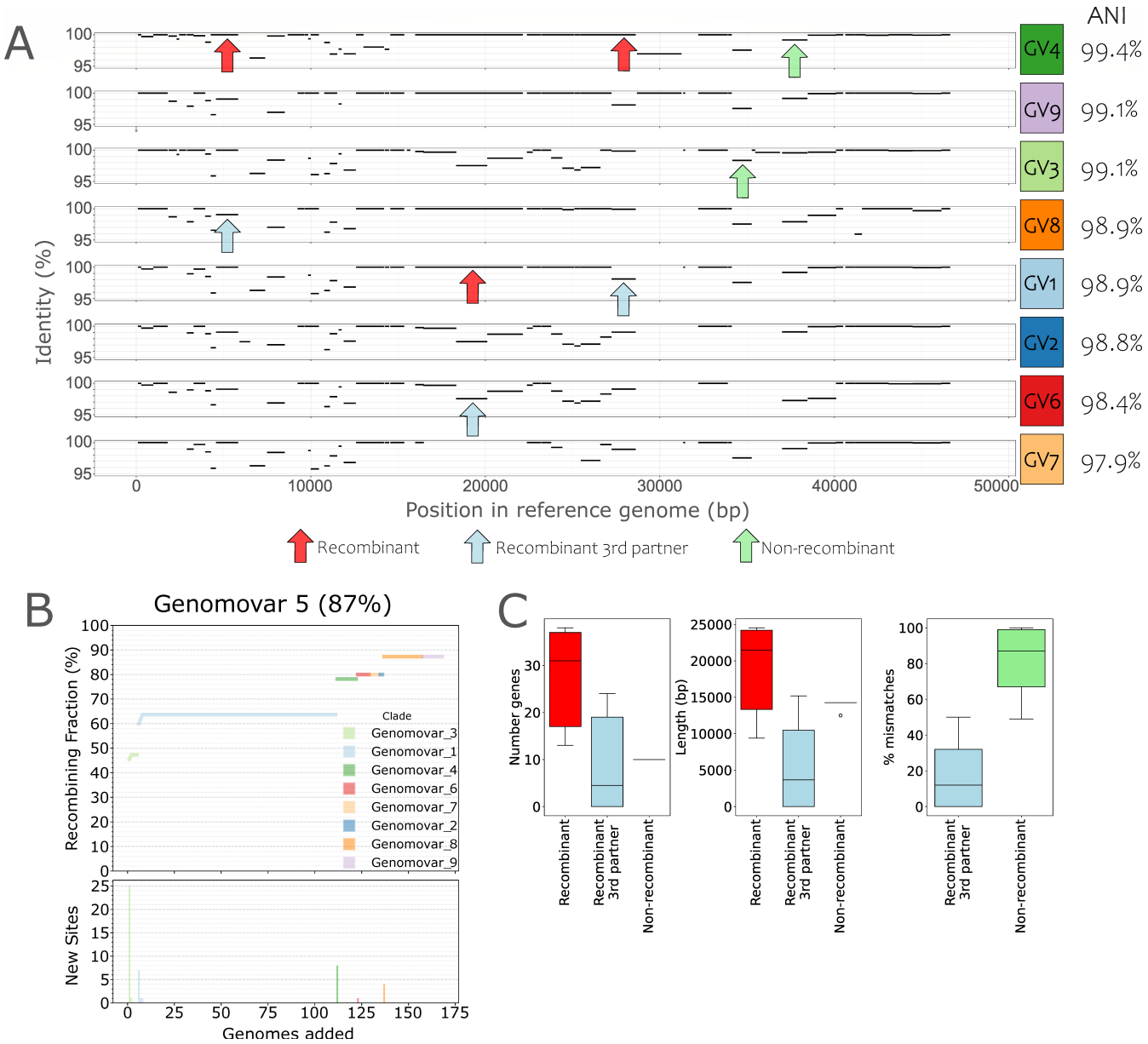

FIG 4 Quantifying the role of recombination as a force of cohesion of the intra-species units. (A) Gene identity plot of a reference genome (genome 13v8 of genomovar 5) against query genomes of different genomovars. Plots are sorted by decreasing ANI against the reference genome (rightmost value). Each black line in the plot represents the position of a gene of the reference genome (x-axis) shared by the query genome and the percent identity of the shared homolog (y-axis); no lines represent genes that are specific to the reference genome (no match in the query genome). For each pair of genomes, shared genes with identities above 99.8% were considered as (recently) recombinant (red arrows, recombination as a force of cohesion); genes with identities below 99.8% within the pair of genomes in question but above 99.8% with another genome were classified as recombinant with a third partner (blue arrows, recombination as a force of diversification); genes with identities below 99.8% for all analyzed pairs were considered non-recombinant (green arrows, representing recent point mutations). (B) Cumulative curve of the number of genes of the reference genome found to have recently recombined with a genome of another genomovar when adding the latter genomes sequentially in the analysis (x-axis). Top plot shows the number of recombined genes as a fraction of the total genes in the reference genome, while the bottom plot shows the number of recombined genes newly detected by each genome added to the analysis. (C) Quantification of the strength of recombination as a force of cohesion (red) and diversification (blue) following the gene classification described in panel A. Note that the impact of cohesive recombination, measured by the number of recombinant genes (left) or their total length (middle), is greater than diversifying recombination. GV, genomovar.

2 weeks apart in 2014. ANI distribution values highlighted a gap at 99.6%, consistent with the data reported above for all viral species (Fig. 1), which translated to nine distinct

99.6% ANI-based genomovars. Genomes sharing <99.6% ANI shared more high-identity genes than expected based on the model with no recombination (Fig. S6). Furthermore, we calculated the fraction of recombinant genes, defined as showing >99.8% nucleotide identity between a reference genome and all genomes of different genomovars (so, pairwise genome ANI <99.5%) and observed that recombination occurred in 83%–100% of the genes in the reference genome, depending on the genome used as reference (Fig. 4B; Fig. S7). Therefore, recombination is not only frequent enough but also occurs throughout the entire genome.

A recombination event could increase similarity between the two partner genomes, but it could also increase dissimilarity between the two genomes when recombination occurs with a third genome that is more divergent; recombination can be both a force of cohesion and diversification. To assess the relative importance of these two forces, we classified genes into three groups for each pairwise genome comparison: recombinant genes (>99.8% identity within the genome pair considered, a proxy for cohesion force), recombinant genes with a third partner (<99.8% identity within the genome pair considered but >99.8% identity with a third genome, a proxy for diversification force), and non-recombinant genes (<99.8% identity within the genome pair considered, proxy for point mutation force) (Fig. 4A). We randomly selected one reference genome from each genomovar and calculated the total number, length, and mismatches of genes classified in the three categories described above (Fig. 4C; Fig. S8). We observed that mismatches between a pair of genomes mainly involved "recombinant genes with a third partner" rather than "non-recombinant genes" because most of the low-identity alleles between the pair of genomes evaluated had a high identity (>99.8% identity) match with another genome, of a different genomovar, in our collection. These results suggested that recent point mutations have a relatively small impact on the evolution of these genomes relative to recombination. Furthermore, the number and length of "recombinant genes" were higher than that of "recombinant genes with a third partner" for five out of nine genomovars, while only three genomovars displayed higher numbers for "recombinant genes with a third partner," and one genomovar showed similar values for both categories. While our empirical approach, most likely, underestimates the frequency of recombinant genes with a third partner because our genome collections do not cover the total diversity of the natural population, cohesive recombination had similar contribution to diversifying recombination for at least a couple of the genome pairs evaluated when we added the genes found in the mutation group to the third partner group. These results do indicate that recombination as a cohesion force might be, overall, frequent enough.

It is important to note that it is not possible to perform this type of analysis for members of the same genomovar due to the high identity across the whole genome (i.e., there is no signal over the background level of sequence identity to detect recombination). Consistent with this assumption, F100 values among members of the same genomovar (>99.6% ANI) fall inside the confidence interval of the model that assumes no recombination (see purple dots in Fig. S6). However, since recombination between genomovars is frequent, recombination within the same genomovar is certainly expected and is likely even higher in frequency, given also that recombination generally increases with increasing sequence identity of the recombining partners (36). Therefore, recombination might also be the main force of cohesion and evolution of genomovars of DNA viruses, in addition to the force of cohesion at the species level.

In contrast to the *Sal. ruber* bacteriophages mentioned above, we observed low frequency of recombination for SARS-CoV-2 genomes (e.g., F100 values fall inside the confidence interval of the model that assumes no recombination for most SARS-CoV-2 genomes; Fig. S9), consistent with recent literature (37–39), even though SARS-CoV-2 genomes also show a clear ANI gap similar to *Sal. ruber* bacteriophages (Fig. 3). Thus, it is likely that mutation rather than recombination is the main genetic mechanism driving SARS-CoV-2 genome diversification. This finding suggests that different mechanisms (i.e., cohesive recombination vs diversifying point mutation) could have similar results on the

intra-species diversity patterns of DNA and RNA viruses, that is, the existence of an ANI gap. Furthermore, the study of the SARS-CoV-2 dynamics has shown the emergence of new and more infectious SARS-CoV-2 genomovars that replace the existing genomovars, in a clear example of ecological competition (40, 41). These results indicate that competition between genomovars (i.e., selection) may also be an important underlying mechanism for the ANI gap in at least some viruses such as SARS-CoV-2. Accordingly, intermediate genomes (i.e., their ANI values fall within the gap) might present less competitive phenotypes on their way to extinction by natural selection or genomovars that thrive under other conditions and/or hosts and, hence, possibly not adequately sampled.

Conclusions

While the data set analyzed here certainly under-sampled total viral species diversity, it includes genomes for 1,016 species, primarily from the *Caudoviricetes* class (realm *Duplodnaviria*), as well as a smaller set of eukaryotic viruses belonging to the realms *Duplodnaviria*, *Riboviria,* and *Varidnaviria*. Thus, we believe that the data set is large enough to allow some initial views of the patterns of intra-species diversity at least for bacteriophages. Our findings highlight the accuracy and robustness of the ANI gap to distinguish viral variants and suggest that genomovars defined by this gap can carry distinct phenotypic (e.g., different virulence and/or infectivity) and/or ecological properties. Therefore, we consider the data presented here as strong evidence for the widespread—but not necessarily universal—existence of functionally distinct intra-species units for bacteriophages. Finding similar patterns within a few eukaryotic viral species evaluated (e.g., SARS-CoV-2) indicates that the results reported here might apply more broadly to viruses, but this assumption needs to be more rigorously tested in the future with more eukaryotic viral species. Our data support that the 99.5% ANI threshold can be useful for most bacteriophage species, but we also recognize that viral genome diversity is vast, and therefore, the threshold may need to be adjusted for specific viral species. For the latter, we suggest obtaining the ANI value distribution for the species in question and assessing whether, and at what range of ANI values, genomovar-discriminating valleys appear. Several viral species in our data set did not show this major ANI pattern presumably due to sampling bias or their true diversity, and they should be studied in the future to better understand the mechanisms driving intra-species diversity. Furthermore, it should be mentioned that the ANI gap reported here using fosmid and long-read metagenome sequences might not be observed in short-read metagenomic studies due to the assembly step merging highly similar sequences into a consensus (e.g., sequences sharing >97%–98% nucleotide identity) (42). Finally, our results indicated that the ANI gap may be the result of either recombination (*Sal. ruber* bacteriophages) or selection-driven diversifying mutation (SARS-CoV-2 genomes), confirming earlier hypotheses that gene exchange (recombinogenic speciation) and ecology (ecological speciation) can both explain the appearance and maintenance of species and intra-species units. It would be interesting to study additional species in the future to advance our understanding of the relative importance of these two processes and their interplay. We expect that the proposed ANI methodology and threshold advance the set of genomic tools to define and track the units of intra-species diversity, thus facilitating future epidemiological and environmental studies.

MATERIALS AND METHODS

The bacteriophage genomes analyzed here were retrieved from JGI IMG/VR database, selecting only those recovered from prokaryotic isolate genomes to avoid any effects (masking) from the assembly of metagenomic reads. Thus, the IMG/VR genomes analyzed here mainly represent proviruses or pseudolysogens. This data set was complemented with viral isolates from "The Actinobacteriophage Database"

(https://phagesdb.org/) (43) and NCBI using the following search string: "Viruses[Organism] NOT cellular organisms[ORGN] NOT wgs[PROP] NOT gbdiv syn[prop] AND (srcdb_refseq[PROP] OR nuccore genome samespecies[Filter])" (Table S1). Since there are no tools to confidently estimate viral genome completeness for all types of viruses (i.e., available tools work well for certain groups of viruses, such as *Caudoviricetes*), only genomes longer than 20 kbp were analyzed with the aim of reducing the potential noise introduced by very incomplete genomes. In addition, 177 recently published *Salinibacter ruber* bacteriophage genomes (35) were also included in the data set. As a result, the final data set comprised a total of 75,012 genomes longer than 20 kbp (Table S1) that correspond to 72,915 lysogens/pseudolysogens and 2,097 viral isolates.

Uncultured genome sequences of fosmids and long-read metagenomes were downloaded from the NCBI genome and SRA databases, respectively (Table S1). All fosmid sequences are of marine origin and have been described previously (44, 45). Presumably, due to the large amount of DNA required to perform long-read sequencing, we did not find any viral long-read metagenomes in public databases at the time of this writing. Instead, we used 19 PacBio HiFi cellular metagenomes from human and chicken guts as well as seawater samples described previously (46–50). Sequences were quality filtered using filtlong v0.2.1 (https://github.com/rrwick/Filtlong) with a minimum read length of 10 kbp and a minimum window quality of 99. Samples with at least 5,000 surviving reads were first analyzed by VirSorter2 v2.2.4 (51) to identify viral sequences and then with checkv v1.0.1 (52) to further refine these sequences as viral or host derived. Only those reads with at least one identified viral gene and more viral genes than host genes were retained for further analysis.

The 8.15 million SARS-CoV-2 genomes were downloaded from NCBI (accessed on 2 August 2023), and then, to retain only high-quality genomes, sequences with any undetermined position (*N*s) identified by the FastA.filterN.pl (content = 0) script of the enveomics collection (53) were discarded. For the CDC sequences, we subsampled them to 1,250 genomes per VOC (except Epsilon and Beta VOCs that had only 157 and 99 available genomes, respectively), that is, 5,041 genomes were used in total. The RKI data set was subsampled to 300 genomes per VOC (except Gamma and Epsilon VOCs with 112 and 2 genomes, respectively), which provided 1,314 genomes in total. In addition to SARS-CoV-2, eukaryotic viral genomes retrieved from NCBI using the search explained above were also analyzed (Table S1).

ANI values between viral genomes were calculated using FastANI v1.33 (16) "Many to Many" mode with a fragment length of 1 kbp to account for the shorter viral genomes (relative to the default 3 kbp for microbial genomes). Viral genomes were assigned to the same species when sharing more than 95% ANI, a previously proposed threshold (14) that we also corroborated within our data set (Fig. S10). Finally, self-matches were removed, and plots were drawn in R using ggplot2 v3.4.2 (54).

To challenge the robustness of the gap, we performed a subsampling analysis to get the same number of pairwise comparisons per species (150). The data were then smoothed using the smooth_data function (sm_method="gam") from the gcplyr R package v1.9.0 (55), and peaks and valleys were automatically identified using findpeaks from the pracma R package v2.4.4 (https://cran.r-project.org/package=pracma). This process was repeated 1,000 times, and the results were pooled and finally plotted using ggplot2 (we used the Bootstrap_analysis.R script available in https://github.com/baldeguer-riquelme/Viral-ANI-gap/).

To classify the observed ANI patterns, we defined four groups based on the detection of valleys and peaks using the approach outlined above, the average ANI value of the collection of genomes analyzed (of the specific species of interest), and the average smoothed counts. Briefly, the data (ANI values) were first smoothed, and peaks and valleys were automatically detected, as explained above. Then, areas of low frequency of pairs were defined as those ANI bins with a number of smoothed counts below the smoothed mean × 0.5. To validate a valley, it had to be detected by the findpeaks function and be in an area of low frequency of pairs. This approach ensures that

validated valleys are at the bottom of the ANI distribution. Group 1 included species displaying a validated valley between 99.2% and 99.8% ANI, and an average ANI below 99.8%. Species with an average ANI above 99.8% were classified into group 2 and represented highly clonal species. Both group 1 and group 2 species provide support to the existence of the gap since the area between 99.2% and 99.8% displays a low number of pairs. Species on group 2 and group 1 are then composed by one or several clusters, respectively. Group 3 included species that did not show a peak or a valley between 99.2% and 99.8% ANI and thus, does not provide strong evidence against—or in favor of—the existence of the gap (undetermined distributions). Finally, there was a group of species that showed a peak rather than a valley between 99.2% and 99.8% ANI, which is inconsistent with the existence of the gap. These species were classified in group 4. Selected examples of these groups are shown in Fig. S3; and all plots are available on https://github.com/baldeguer-riquelme/Viral-ANI-gap/. We manually reviewed all 1,016 species plots and moved 56 species (9 prokaryotic, 6 fosmids, and 41 long read) to a different group than the species was automatically assigned to using the methodology described above.

The frequency of recombination was calculated based on the fraction of shared identical reciprocal best-match genes between genome pairs (F100) relative to those expected based on the ANI value of the genome and assuming no recombination, which can be considered a proxy for recent recombination events. For this, genes were first predicted using Prodigal v2.6.3, and then reciprocal best-match genes for each pairwise comparison were identified using BLASTn (v2.14.0). We assumed that genes sharing more than 99.8% identity represent recently recombined genes and labeled them accordingly. Finally, we defined the F100 value as the fraction of recombinant genes from the total number of reciprocal best-match genes for each pairwise comparison. To build the simulated model that only considers random mutations and no recombination, we employed the script Simulate_population_genomes.py, available at https://github.com/rotheconrad/Population-Genome-Simulator. To resemble the available *Salinibacter ruber* phage genomes as much as possible, the simulated population was created with the following parameters: -n 100 g 70 c 1 -cr 90 -mu 690 -sd 150. Cumulative recombinant gene curves were built using the 03 g_Recombinant_group_analysis.py script of the F100_Prok_Recombination pipeline. For each pairwise comparison, genes were classified into three groups: recombinant genes (>99.8% identity, proxy for cohesion force), recombinant genes with a third partner (<99.8% identity but >100% identity with a third genome, proxy for diversification force), and non-recombinant genes (<99.8% identity, proxy for mutation force), as described in the main text. Note that the genes are labeled separately for each pairwise comparison, and thus, the same gene might be classified into different categories depending on the genome pair analyzed. For example, a gene of genome A can be classified as "recombinant" with genome B and as "recombinant with a third partner" when compared to genome C. In addition, note that non-recombinant genes might actually be recombinant with a partner not included in the data set. Classification and plots were carried out using the 03 g_Recombinant_group_extra_code.py script available at https://github.com/baldeguer-riquelme/Viral-ANI-gap/. We used the F100 approach to detect evidence of recent recombination of SARS-CoV-2 genomes. Specifically, we compared the SARS-CoV-2 F100 values against a model with no recombination (Simulate_population_genomes.py script, parameters: -n 10 g 12 c 1 -cr 90 -mu 1180 -sd 2200).

ACKNOWLEDGMENTS

This work has been supported, in part, by the US National Science Foundation (Award No. 1759831 and 2129823) to K.T.K. and the METACIRCLE projects (PID2021-126114NB-C41 and PID2021-126114NB-C42) funded by the Spanish Ministry of Science and Innovation to J.A. and R.R.-M.

J.A. is a member of FAGOMA RED2022-134837-T.

AUTHOR AFFILIATIONS

[1]School of Civil & Environmental Engineering and School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA
[2]Department of Physiology, Genetics and Microbiology, University of Alicante, San Vicente del Raspeig, Spain
[3]Marine Microbiology Group, Department of Animal and Microbial Biodiversity, Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB), Esporles, Spain

AUTHOR ORCIDs

Borja Aldeguer-Riquelme http://orcid.org/0000-0003-4266-0712
Roth E. Conrad http://orcid.org/0000-0001-8155-8441
Josefa Antón http://orcid.org/0000-0002-5823-493X
Konstantinos T. Konstantinidis http://orcid.org/0000-0002-0954-4755

FUNDING

Funder	Grant(s)	Author(s)
National Science Foundation (NSF)	1759831	Konstantinos T. Konstantinidis
National Science Foundation (NSF)	2129823	Konstantinos T. Konstantinidis
Spanish Ministry of Science and Innovation	PID2021-126114NB-C41	Josefa Antón Ramon Rossello-Mora
Spanish Ministry of Science and Innovation	PID2021-126114NB-C42	Josefa Antón Ramon Rossello-Mora

AUTHOR CONTRIBUTIONS

Borja Aldeguer-Riquelme, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft | Roth E. Conrad, Data curation, Formal analysis | Josefa Antón, Investigation, Validation, Writing – review and editing | Ramon Rossello-Mora, Investigation, Validation, Writing – review and editing | Konstantinos T. Konstantinidis, Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing – review and editing

DIRECT CONTRIBUTION

This article is a direct contribution from Konstantinos T. Konstantinidis, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Simon Roux, JGI, and Francisco Murilo Zerbini, Universidade Federal de Viçosa.

DATA AVAILABILITY

Prokaryotic, eukaryotic, and fosmid viral genomes have been retrieved from IMG/VR, NCBI, and "The Actinobacteriophage Database" (https://phagesdb.org/) (see Table S1 for accession numbers). Long-read metagenomes were downloaded from SRA (see Table S1 for accession numbers). A more detailed description of the procedure and R scripts used in this study, as well as plots for each individual species and the ANI-pattern-group to which they were assigned, can be found on https://github.com/baldeguer-riquelme/Viral-ANI-gap/ . Further technical details and scripts for the recombination analyses are available at https://github.com/rotheconrad/F100_Prok_Recombination/.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental figures (mBio01536-24-s0001.pdf). Figures S1 to S10.
Table S1 (mBio01536-24-s0002.xlsx). Accession numbers of the sequences used in this study and associated metadata.

REFERENCES

1. Mühling M, Fuller NJ, Millard A, Somerfield PJ, Marie D, Wilson WH, Scanlan DJ, Post AF, Joint I, Mann NH. 2005. Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. Environ Microbiol 7:499–508. https://doi.org/10.1111/j.1462-2920.2005.00713.x
2. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, et al. 2010. Viral and microbial community dynamics in four aquatic environments. ISME J 4:739–751. https://doi.org/10.1038/ismej.2010.1
3. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. 2020. Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat Microbiol 5:265–271. https://doi.org/10.1038/s41564-019-0628-x
4. Monterroso P, Garrote G, Serronha A, Santos E, Delibes-Mateos M, Abrantes J, Perez de Ayala R, Silvestre F, Carvalho J, Vasco I, Lopes AM, Maio E, Magalhães MJ, Mills LS, Esteves PJ, Simón MÁ, Alves PC. 2016. Disease-mediated bottom-up regulation: an emergent virus affects a keystone prey, and alters the dynamics of trophic webs. Sci Rep 6:36072. https://doi.org/10.1038/srep36072
5. Van Regenmortel MHV. 2007. Virus species and virus identification: past and current controversies. Infect Genet Evol 7:133–144. https://doi.org/10.1016/j.meegid.2006.04.002
6. Callaway E. 2021. Beyond Omicron: what's next for COVID's viral evolution. Nature 600:204–207. https://doi.org/10.1038/d41586-021-03619-8
7. Tortuel D, Tahrioui A, David A, Cambronel M, Nilly F, Clamens T, Maillot O, Barreau M, Feuilloley MGJ, Lesouhaitier O, Filloux A, Bouffartigues E, Cornelis P, Chevalier S. 2022. Pf4 phage variant infection reduces virulence-associated traits in *Pseudomonas aeruginosa*. Microbiol Spectr 10:e0154822. https://doi.org/10.1128/spectrum.01548-22
8. Parra GI, Squires RB, Karangwa CK, Johnson JA, Lepore CJ, Sosnovtsev SV, Green KY. 2017. Static and evolving norovirus genotypes: implications for epidemiology and immunity. PLoS Pathog 13:e1006136. https://doi.org/10.1371/journal.ppat.1006136
9. Valastro V, Holmes EC, Britton P, Fusaro A, Jackwood MW, Cattoli G, Monne I. 2016. S1 gene-based phylogeny of infectious bronchitis virus: an attempt to harmonize virus classification. Infect Genet Evol 39:349–364. https://doi.org/10.1016/j.meegid.2016.02.015
10. Martinez-Hernandez F, Diop A, Garcia-Heredia I, Bobay LM, Martinez-Garcia M. 2022. Unexpected myriad of co-occurring viral strains and species in one of the most abundant and microdiverse viruses on Earth. ISME J 16:1025–1035. https://doi.org/10.1038/s41396-021-01150-2
11. Olm MR, Bhattacharya N, Crits-Christoph A, Firek BA, Baker R, Song YS, Morowitz MJ, Banfield JF. 2019. Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria. Sci Adv 5:eaax5727. https://doi.org/10.1126/sciadv.aax5727
12. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. 2021. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol 39:727–736. https://doi.org/10.1038/s41587-020-00797-0
13. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB. 2014. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. Nature 513:242–245. https://doi.org/10.1038/nature13459
14. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, et al. 2019. Minimum information about an uncultivated virus genome (MIUViG). Nat Biotechnol 37:29–37. https://doi.org/10.1038/nbt.4306
15. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, et al. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177:1109–1123. https://doi.org/10.1016/j.cell.2019.03.040
16. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9
17. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. https://doi.org/10.1099/ijs.0.64483-0
18. Rodriguez-R LM, Conrad RE, Viver T, Feistel DJ, Lindner BG, Venter SN, Orellana LH, Amann R, Rossello-Mora R, Konstantinidis KT. 2024. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio 15:e0269623. https://doi.org/10.1128/mbio.02696-23
19. Varsani A, Martin DP, Navas-Castillo J, Moriones E, Hernández-Zepeda C, Idris A, Murilo Zerbini F, Brown JK. 2014. Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch Virol 159:1873–1882. https://doi.org/10.1007/s00705-014-1982-x
20. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Peña MJ, Martínez JM, Anton J, Gasol JM, Rosselli R, Rodriguez-Valera F, Sullivan MB, Acinas SG, Martinez-Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat Commun 8:15892. https://doi.org/10.1038/ncomms15892
21. World Health Organization (WHO). 2023. Updated working definitions and primary actions for SARS-CoV-2 variants, 15 March 2023
22. World Health Organization (WHO). 2023. Tracking SARS-CoV-2 variants. Available from: https://www.who.int/activities/tracking-SARS-CoV-2-variants. Retrieved 02 Aug 2023.
23. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, Peacock SJ, Barclay WS, de Silva TI, Towers GJ, Robertson DL, COVID-19 Genomics UK Consortium. 2023. SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nat Rev Microbiol 21:162–177. https://doi.org/10.1038/s41579-022-00841-7
24. Center for Disease Control and Prevention. 2023. SARS-CoV-2 variant classifications and definitions. Available from: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html. Retrieved 02 Aug 2023.
25. Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R, Silva JCF, Fiallo-Olivé E, Briddon RW, Hernández-Zepeda C, Idris A, Malathi VG, Martin DP, Rivera-Bustamante R, Ueda S, Varsani A. 2015. Revision of *Begomovirus* taxonomy based on pairwise sequence comparisons. Arch Virol 160:1593–1619. https://doi.org/10.1007/s00705-015-2398-y
26. Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, Zerbini FM, Rivera-Bustamante R, Malathi VG, Briddon RW, Varsani A. 2013. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus *Mastrevirus* (family *Geminiviridae*). Arch Virol 158:1411–1424. https://doi.org/10.1007/s00705-012-1601-7
27. Zerbini FM, Herrera da Silva João Paulo. 2024. Taxonomic classification of *Geminiviruses* based on pairwise sequence comparisons. In Fontes ElizabethPB, Mäkinen K (ed), Plant-virus interactions. Springer US, New York, NY.
28. Burk RD, Harari A, Chen Z. 2013. Human papillomavirus genome variants. Virology 445:232–243. https://doi.org/10.1016/j.virol.2013.07.018
29. Ursing JB, Rossello-MoraRA, Garcia-ValdesE, Lalucat J. 1995. Taxonomic note: a pragmatic approach to the nomenclature of phenotypically similar genomic groups. Int J Syst Bacteriol45:604–604. https://doi.org/10.1099/00207713-45-3-604
30. Rossello R, Garcia-Valdes E, Lalucat J, Ursing J. 1991. Genotypic and phenotypic diversity of *Pseudomonas stutzeri*. Syst Appl Microbiol14:150–157. https://doi.org/10.1016/S0723-2020(11)80294-8

31. Rodriguez-R LM, Conrad RE, Viver T, Feistel DJ, Lindner BG, Venter SN, Orellana LH, Amann R, Rossello-Mora R, Konstantinidis KT. 2024. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio 15:e02696–23. https://doi.org/10.1128/mbio.02696-23
32. Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. Elife 4:e08490. https://doi.org/10.7554/eLife.08490
33. Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F. 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. Infect Genet Evol 30:296–307. https://doi.org/10.1016/j.meegid.2014.12.022
34. Mavrich TN, Hatfull GF. 2017. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol 2:17112. https://doi.org/10.1038/nmicrobiol.2017.112
35. Ramos-Barbero MD, Aldeguer-Riquelme B, Viver T, Villamor J, Carrillo-Bautista M, López-Pascual C, Konstantinidis K, Martinez-Garcia M, Santos F, Rossello-Mora R, Anton J. 2023. Experimental evolution at ecological scale allows linking viral genotypes to specific host strains. Res sq. Res sq,. https://doi.org/10.21203/rs.3.rs-3621737/v1
36. Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480. https://doi.org/10.1126/science.1127573
37. VanInsberghe D, Neish AS, Lowen AC, Koelle K. 2021. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. Virus Evol 7:veab059. https://doi.org/10.1093/ve/veab059
38. Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, Smith K, De Maio N, Haussler D, Lanfear R, Corbett-Detig R. 2022. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature 609:994–997. https://doi.org/10.1038/s41586-022-05189-9
39. Preska Steinberg A, Silander OK, Kussell E. 2023. Correlated substitutions reveal SARS-like coronaviruses recombine frequently with a diverse set of structured gene pools. Proc Natl Acad Sci U S A 120:e2206945119. https://doi.org/10.1073/pnas.2206945119
40. Chen J, Gu C, Ruan Z, Tang M. 2023. Competition of SARS-CoV-2 variants on the pandemic transmission dynamics. Chaos Solitons Fractals 169:113193. https://doi.org/10.1016/j.chaos.2023.113193
41. Beesley LJ, Moran KR, Wagh K, Castro LA, Theiler J, Yoon H, Fischer W, Hengartner NW, Korber B, Del Valle SY. 2023. SARS-CoV-2 variant transition dynamics are associated with vaccination rates, number of co-circulating variants, and convalescent immunity. EBioMedicine 91:104534. https://doi.org/10.1016/j.ebiom.2023.104534
42. Meyer F, Fritz A, Deng Z-L, Koslicki D, Lesker TR, Gurevich A, Robertson G, Alser M, Antipov D, Beghini F, et al. 2022. Critical assessment of metagenome interpretation: the second round of challenges. Nat Methods 19:429–440. https://doi.org/10.1038/s41592-022-01431-4
43. Russell DA, Hatfull GF. 2017. PhagesDB: the actinobacteriophage database. Bioinformatics 33:784–786. https://doi.org/10.1093/bioinformatics/btw711
44. Ghai R, Martin-Cuadrado AB, Molto AG, Heredia IG, Cabrera R, Martin J, Verdú M, Deschamps P, Moreira D, López-García P, Mira A, Rodriguez-Valera F. 2010. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J 4:1154–1166. https://doi.org/10.1038/ismej.2010.44
45. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. PLoS Genet 9:e1003987. https://doi.org/10.1371/journal.pgen.1003987
46. Zhang Y, Jiang F, Yang B, Wang S, Wang H, Wang A, Xu D, Fan W. 2022. Improved microbial genomes and gene catalog of the chicken gut from metagenomic sequencing of high-fidelity long reads. Gigascience 11:giac116. https://doi.org/10.1093/gigascience/giac116
47. Nhu NTK, Phan M-D, Hancock SJ, Peters KM, Alvarez-Fraga L, Forde BM, Andersen SB, Miliya T, Harris PNA, Beatson SA, Schlebusch S, Bergh H, Turner P, Brauner A, Westerlund-Wikström B, Irwin AD, Schembri MA. 2023. High-risk *Escherichia coli* clones that cause neonatal meningitis and association with recrudescent infection. medRxiv. https://doi.org/10.1101/2023.10.05.23296362
48. Plaza Oñate F, Roume H, Almeida M. 2022. Recovery of metagenome-assembled genomes from a human fecal sample with Pacific Biosciences high-fidelity sequencing. Microbiol Resour Announc 11:e0025022. https://doi.org/10.1128/mra.00250-22
49. Kim CY, Ma J, Lee I. 2022. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. Nat Commun 13:6367. https://doi.org/10.1038/s41467-022-34149-0
50. Patin NV, Goodwin KD. 2022. Long-read sequencing improves recovery of picoeukaryotic genomes and zooplankton marker genes from marine metagenomes. mSystems 7:e0059522. https://doi.org/10.1128/msystems.00595-22
51. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB, Roux S. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9:37. https://doi.org/10.1186/s40168-020-00990-y
52. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–585. https://doi.org/10.1038/s41587-020-00774-7
53. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection : a toolbox for specialized analyses of microbial genomes and metagenomes. Peer J. https://doi.org/10.7287/peerj.preprints.1900v1
54. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York. Available from: https://ggplot2.tidyverse.org
55. Blazanin M. 2024. gcplyr: an R package for microbial growth curve data analysis. bioRxiv. https://doi.org/10.1101/2023.04.30.538883