A manuscript re-submitted to ISME Journal

Unexpected myriad of cooccurring viral strains and species in one of the most abundant and microdiverse viruses on Earth

Francisco Martinez-Hernandez[1], Awa Diop[2], Inmaculada Garcia-Heredia[1], Louis-Marie Bobay[2], Manuel Martinez-Garcia[1*]

[1]Department of Physiology, Genetics, and Microbiology, University of Alicante. Alicante, Spain.

[2]Department of Biology, University of North Carolina at Greensboro, USA.

***Corresponding author:** m.martinez@ua.es

Abstract

Viral genetic microdiversity drives adaptation, pathogenicity and speciation and has critical consequences for the viral-host arms race occurring at the strain and species levels, which ultimately impact microbial community structure and biogeochemical cycles. Despite the fact that most efforts have focused on viral macrodiversity, little is known about the microdiversity of ecologically important viruses on Earth. Recently, single-virus genomics discovered the putatively most abundant ocean virus in temperate and tropical waters: the uncultured dsDNA virus vSAG 37-F6 infecting *Pelagibacter*, the most abundant marine bacteria. In this study, we report the cooccurrence of up to ≈1,500 different viral strains (>95% nucleotide identity) and ≈30 related species (80-95% nucleotide identity) in a single oceanic sample. Viral microdiversity was maintained over space and time, and most alleles were the result of synonymous mutations without any apparent adaptive benefits to cope with host translation codon bias and efficiency. Gene flow analysis used to delimitate species according to the biological species concept (BSC) revealed the impact of recombination in shaping vSAG 37-F6 virus and *Pelagibacter* speciation. Data demonstrated that this large viral microdiversity somehow mirrors the host species diversity since ≈50% of the 926 analyzed *Pelagibacter* genomes were found to belong to independent BSC species that do not significantly engage in gene flow with one another. The host range of this evolutionarily successful virus revealed that a single viral species can infect multiple *Pelagibacter* BSC species, indicating that this virus crosses not only formal BSC barriers but also biomes since viral ancestors are found in freshwater.

Introduction

Recent advances in viral ecology, mainly based on viral metagenomics (hereinafter viromics), have allowed us to highly expand the diversity of the global virosphere [1–7]. To date, most viromic surveys have relied on short read assembly[1], which mostly recovers the genome of dominant viruses but frequently overlooks relevant information about the genetic microdiversity of cooccurring viruses[8–10]. Viral microdiversity (nucleotide differences within the same viral species) has important consequences on viral ecology, and understanding microdiversity patterns of ecologically relevant viruses in nature is important for increasing knowledge about speciation, pathogenicity, microbial community structure and host dynamics, which overall impact biogeochemical processes[11, 12].

The continuous arms race within the viral-host system is an important engine generating this viral microdiversity, which in some cases leads to amino acid changes (nonsynonymous mutations) in viral

proteins with a significant impact on viral fitness. In marine cyanophages, only a small number of genetic changes generated phenotypic diversification, affecting the successful infection of different *Synechococcus* spp. strains [13, 14]. Similarly, a single nonsynonymous mutation in the tail fiber of *Pseudomonas* virus LUZ7 drove host range expansion [15]. Paradoxically, synonymous mutations are thought to have a neutral evolutionary impact, although recent data suggest that they might provide an advantage for viruses to counteract host defense systems based on DNA recognition [16] or to adapt codon usage in accordance with the host's [17–21]. Many microdiversity studies have been conducted with reference viral isolates. In the marine ecosystem, for instance, using viral tagging methodology, "*discrete populations"* of cooccurring cyanophages were obtained from a single strain isolate of *Synechococcus* spp. [22]. In the human gut, the recently cultured ubiquitous crAssphage virus diverges intraindividually and generates a continuous replacement of different strains in the long term [23]. However, it is particularly challenging to address microdiversity for uncultured viruses [24–26]. Recently, the uncultured virus vSAG 37-F6 was discovered to be putatively the most abundant marine virus in temperate and tropical waters of the open ocean[8]. This virus obtained by single-virus genomics (SVGs) was shown to be widespread across the oceans and present from surface to deep waters. Nevertheless, despite its high abundance, this genome could not be assembled from metagenomic data [8–10]. The host of vSAG 37-F6, the dominant *Pelagibacter* spp., was later discovered by using single-cell genomic data mining, since related viral contigs were present in different single-amplified genomes (SAGs) of *Pelagibacter* spp. [27]. Thus, this virus is thought to be responsible for channelizing an enormous amount of carbon through the viral shunt [28], which has a major impact on a global scale. Here, we estimate the level of microdiversity of this relevant virus in nature that surprisingly reaches up to more than a thousand cooccurring strains in a single sample and explore the biological meaning of viral genetic microdiversity. In addition, we delved into the existence of true biological species within the vSAG 37-F6 virus and its host based on the biological species concept (BSC). Members of the same BSC are characterized by their capacity for gene exchange by homologous recombination. Although prokaryotes and viruses have an asexual mode of reproduction, it has been described that several microorganisms, including some types of viruses, such as cyanophages, engage in sufficient levels of homologous recombination to potentially distinguish biological species [29–31]. Furthermore, we also investigated whether viral infection respects the BSC barriers, i.e., whether one viral species can infect one or more different prokaryote species based on BSC. Altogether, our data helped us better understand the genetic patterns and viral species structure (i.e., number of cooccurring viral species and strains) and evolutionary forces (recombination vs mutation), probably shaping one of the most abundant and ecologically relevant viruses in nature.

Results

Estimating the number of cooccurring species/strains in one of the most abundant marine viruses, vSAG 37-F6

The putatively most abundant virus vSAG 37-F6 in temperate and tropical waters of the open ocean was originally discovered by SVG in the Mediterranean Sea and was overlooked for years by other standard viromic technologies despite huge metagenomic sequencing efforts[8]. Here, high-throughput amplicon sequencing targeting different genomic regions of vSAG 37-F6 and close relatives (Supplementary Fig. 1

and Supplementary Table 1) was performed for several Mediterranean viral samples (surface, DCM, 1000 m, and 2,000 m depth) to ascertain the level of cooccurring genetic microdiversity of this virus. For instance, one of those genomic regions partially encompassed the *gene 9* encoding a conserved capsid protein of virus vSAG 37-F6, which is one of the most abundant viral proteins in temperate and tropical waters of the open ocean, as previously demonstrated [8]. Sequencing data were used to unveil and estimate the number of putative strains and species by applying two different nucleotide thresholds for clustering dsDNA viruses as per recent recommendations [1, 3]: >95% nucleotide identity to estimate the number of potentially cooccurring strains (i.e., genetic microdiversity; Fig. 1 and Table 1) and a ≈80-95% cutoff to ascertain the number of viral species or "virus operational taxonomic units" (vOTUs) related to virus vSAG 37-F6 present in the same natural sample. Recently, a joint effort of viral and microbial ecologists suggested formalizing the use of species-rank virus groups and named these vOTUs to avoid confusion with other terms and proposed standard thresholds of 95% average nucleotide identity[1, 3] as a practical value for viral species-like delineation [1, 3, 8], as used here in our study.

Unexpectedly, microdiversity data showed that up to 1,422 different putative viral strains could cooccur in the same sample and location, such as the Blanes Bay Microbial Observatory (Fig. 1 and Table 1), where this virus was originally discovered. At the species level, an average of ≈10 cooccurring putative species related to vSAG 37-F6 were detected (Fig. 1 and Table 1). In offshore samples, vSAG 37-F6 species dominated either in the surface or deep samples (Fig. 1, Supplementary Fig. 2 and Table 1) since 97% of sequenced strains were assigned to this species. In coastal surface seawater samples, other related vSAG 37-F6 viral species (nucleotide identity ≈80% with virus vSAG 37-F6) dominated. Remarkably, many vSAG 37-F6 strains were shared across samples, although a significant fraction of strains was unique in each environment (Table 1). Thus, our empirical data unveiled a vast local coexisting (micro)diversity of this dominant virus that is maintained over space and time, since the analyzed samples were distantly located and collected years apart.

Global microdiversity of vSAG 37-F6 and other pelagiphages

A method based on metagenomic fragment recruitment using the Shannon index ($H = - \Sigma P_i \bullet \ln P_i$) [32] was used to analyze the global ocean genome microdiversity of vSAG 37-F6 and other pelagiphages, including lytic, lysogenic, isolated, and uncultured viruses. This *H* parameter (values from 0 to 1) calculates the genomic diversity at the single-nucleotide level (see methods). Briefly, higher values of *H* represent a more microdiverse genome (lower possibilities of finding the same nucleotide twice at a given genome position). Whole genome entropy was calculated using different cell metagenome and virome datasets [24, 33] (Supplementary Fig. 3 and Supplementary Table 2). Cell metagenomes inform about the microdiversity of those probably infectious viruses, while virome data (i.e., free viral particles in seawater) represent the total microdiversity pool of viruses. Overall, genome entropies values ranged from 0.012 to 0.17 (Fig. 2A). Higher values of microdiversity were always observed for each virus in the free viral fraction in seawater compared with cellular metagenomes.. Singularly, in the ocean panvirome and metagenome, the most microdiverse viral species was vSAG 37-F6, and its close viral relative pelagiphage MED40-C1 that was found in a single cell from the Mediterranean Sea[27]. vSAG 37-F6-like pelagiphages showed significant higher values of whole-genome entropy than those of other pelagiphages (p-value <0.05, Fig. 2 and

Supplementary Table 3). Remarkably, relevant differences were not observed in the maximum values of microdiversity for samples located several thousands of kilometers apart, collected at different seasons and depths, and even for samples with relevant variations in abundance (Supplementary Fig. 3). Most of the vast and conserved genetic microdiversity was generated by synonymous mutations (NSr mean = 40.29, Supplementary Table 4) that were stable over time and space (Fig 2B, Supplementary Figs 3-9 and Supplementary Table 4). Only a very low proportion of vSAG 37-F6 proteins (n = 3, unknown vSAG 37-F6 protein encoded by genes 8, 14, and 23), and viral relatives showed an unusually high ratio of nonsynonymous mutations (NSr mean = 61.47) suggestive of positive selection (e.g., unknown vSAG 37-F6 protein encoded by gene 8; Fig. 2B, Supplementary Figs. 3-9 and Supplementary Table 4). Data further suggest that this "hidden" vast genomic microdiversity of vSAG 37-F6 -mostly observed as synonymous mutations- never explored before in the oceans is strongly preserved and globally maintained in the long term since these results are not circumscribed to a specific location in a certain period of time but it is something general that is observed in samples spanning more than ten years from different oceans.

True biological species within vSAG 37-F6 and *Pelagibacter* host: do viruses respect biological species concept barriers?

Recently, it has been proposed that a universal biological species concept (BSC) definition can be used in all major lifeforms, including viruses, based on evidence of gene flow[34]. We then sought to investigate whether the vSAG 37-F6 virus, despite its high microdiversity, could be structured into true BSCs. Because members of the same biological species are characterized by their ability for gene exchange, we assessed the degree of recombination of vSAG 37-F6 with a set of most highly closely related viral genomes (n=32) sharing a high proportion of orthologous genes (i.e., core genome; Supplementary Fig. 10 and Supplementary Tables 5 and 6, see methods) to accurately determine whether polymorphic sites arose by mutation or recombination. Our analyses identified gene flow between homologous genes (i.e., homologous recombination) and estimated the ratio of homoplasic (*h=recombination*) to nonhomoplasic (*m=mutation*) polymorphisms along the core genome of each genus. Homoplasies are polymorphisms that are not compatible with vertical inheritance from a single ancestral mutation and likely result from the exchange of alleles through homologous recombination. High *h*/*m* ratios (≥1) are indicative of a substantial signal of gene flow, and low *h*/*m* ratios are indicative of clonal (<1) evolution. The data revealed that these viruses displayed a high *h/m* ratio (gray curve Fig 3A), suggesting that recombination might be an important force shaping the evolution of this virus. However, the *h/m* ratio was only slightly higher than that obtained from the dataset simulated in the absence of homologous recombination (pink curve, Fig 3A), which, as previously described [34], is used to assess the number of homoplasies introduced by convergent mutations. Therefore, these patterns indicate that the majority of homoplasies are introduced by mutations rather than recombination, suggesting that this analyzed virus is composed of a single clonal species or contains multiple biological species that do not recombine with one another. Following the same rationale for the host, we aimed to estimate gene flow (*h/m* ratio) and the number of true *Pelagibacter* BSC species within a dataset of 926 publicly available genomes by computing the pairwise core nucleotide identity (CNI) and by conducting a large-scale phylogenomic analysis (see methods; (Fig. 3B, Supplementary Fig. 11 and Supplementary Data 1). First, the 926 genomes were classified into 495 monophyletic clusters (i.e., putative

species) based on a >94% CNI threshold and the phylogenetic tree. These clusters were then tested for gene flow *within* clusters and *between* clusters by computing the *h/m* ratio using the core genomes of each of these clusters and for each pair of clusters (see Methods). Within-cluster analysis revealed that the number of homoplasies within most clusters was significantly higher (Supplementary Table 7) than the number of homoplasies expected from convergent mutations by generating sequences simulated under similar conditions but without recombination; this indicates that most of these clusters likely represent a single biological species. Clusters that did not show a clear signal of gene flow were found to contain genomes that did not engage in recombination with the rest of the viruses, and these genomes could be excluded from the cluster, thereby redefining all clusters into a biological species (Supplementary Table 7 and Supplementary data 1). Then, we tested for the signal of gene flow between pairs of clusters using the same approach (see Methods). Estimates of *h/m* were systematically compared to *h/m* ratios computed on the reference cluster while including one sequence simulated without recombination (see Methods). Using this approach, 54 clusters were found to engage in gene flow with another cluster, and cluster borders were redefined accordingly. Finally, this approach yielded a total of 441 clusters that can be considered true biological species, indicating a large diversity in this dataset, where approximately one out of two deposited *Pelagibacter* spp. genome represents a true biological species (Fig 3B, Sup Table 5). Most biological species were composed of highly related genomes (97% CNI on average); however, some contained more divergent genomes sharing as little as 80% CNI. This indicates that sequence thresholds do not accurately predict the borders of biological species and that highly divergent genomes are sometimes part of the same biological species. This large-scale genomic analysis further shows that recombination is a predominant force shaping *Pelagibacter* spp., which are composed of a highly diverse set of biological species that do not significantly engage in gene flow with one another. Recombination is possibly driving the evolution of vSAG 37-F6 as well, although convergent mutations cannot be ruled out, and additional genomes are needed to solve this question.

Further effort was then conducted to shed some light on the host range of vSAG 37-F6 and related pelagiphages in line with the described BSC conceptual approach (*n*=441 *Pelagibacter* BSC). Data showed that five different *Pelagibacter* single cells (*SAGs-MED 41,43, 45, 46 and 48* (Fig. 3C, Supplementary Fig. 12 and Supplementary Data 1) belonging to different BSCs (CNI values 74 – 88%) were infected by the same vSAG 37-F6-like pelagiphage species (strain sharing amino acid similarity > 98.5%, Supplementary Table 8 and 9). Our results indicate that this widespread and ubiquitous virus does not `respect´ true prokaryotic biological species boundaries, which represent a significant ecological example of general interest linking taxonomic and biological insights in probably one of the most abundant microbes in the biosphere.

Evolution and ancestors of vSAG 37-F6

Given the evolutionary success of vSAG 37-F6 and its host in the oceans and considering the transition and colonization of *Pelagibacter* spp. ancestors in freshwater (*Fonsibacter* spp., formerly described as LD-12 [35]), we sought to investigate whether vSAG 37-F6 viral relatives inhabit nonmarine environments. After mining the IMG/VR v.2.0 database [36] and other datasets [4] by searching orthologous genes of vSAG

37-F6 virus (amino acid similarity ≥ 50% and query coverage ≥ 95%), we identified several dozen viral genomes having hallmark genes of vSAG 37-F6 in low saline aquatic environments, such as inland lakes, lagoons, microbial mats and sediments (Supplementary Figs. 13 and 14, and Supplementary Table 10). More intriguingly, 101 vSAG 37-F6-related freshwater viruses were found in lakes located in North America, Canada (Lake Mendota and Simoncouche) and Europe that contained an ortholog of *gene 9* (Fig. 4, Supplementary Fig. 15, Supplementary Table 10) encoding the hallmark capsid protein of vSAG 37-F6 in addition to other ortholog genes. An in-silico search using a database of 5,500 freshwater metagenome-assembled genomes failed to find the host of these freshwater viruses. The phylogeny of *gene 9* (Fig. 4 and Supplementary Fig. 16*)* showed that freshwater and marine viruses, despite a long evolutionary history, preserved a large number of invariable amino acid site positions. Our results indicate that these freshwater viruses evolved from a vSAG 37-F6 viral ancestor and that after millions of years of evolution, they lost many vSAG 37-F6 viral genes, maintaining in all cases the capsid hallmark protein, which is one of the best examples of evolutionary success of viruses in nature since it is not only the most abundant viral protein in temperate and tropical waters of the open ocean [8, 37] but also remains functional in other biomes.

Discussion

At the oceanic global scale, five viral ecological zones have been observed, with maximum values of viral macro- and microdiversity detected in tropical surface waters and in the Arctic [3]. More recently, depth-dependent trends were observed in the frequency of polymorphic sites and nonsynonymous mutations in marine ecosystems among different viral genes, in line with the Red Queen dynamics [24]. Data also suggested seasonal variations of different uncultured viruses at the single nucleotide level and indicated that viral-host interaction is an important motor that drove viral diversification [12–14, 24]. In our study, we quantified the microdiversity structure at the strain and species levels of probably the most abundant ocean dsDNA virus in temperate and tropical waters, which has been overlooked in previous metagenomic studies [8–10]. Data indicate that thousand strains and different related species coexist in a single sample, forming a myriad of vSAG 37-F6 variants (nucleotide identity values ranging from 80 to 100%). Our contrasting microdiversity data from free viruses and cell metagenomes from the same site indicated that only a tiny fraction of all extant microdiversity was actively replicating (Fig. 2) since microdiversity values were more similar to those obtained from clonal expansions/replications of a single strain or a few strains in a sample (Supplementary Fig. 17). Furthermore, single-cell data suggest that the same viral strain infects distantly related *Pelagibacter BSCs,* and it has been described that viruses with broad host ranges commonly show low infection efficiencies [38]. Indeed, this has been previously observed in marine transcriptome datasets with an overall low transcriptional activity of vSAG 37-F6 per host cell regardless of abundance [39, 40]. A similar process has been observed in cyanophages of *Prochlorococcus*[41]. Thus, our data suggest that multiple, low efficiency, sequential infection cycles of different viral strains are maintained over space and time, generating a global large microdiversity, in line with the constant-diversity hypothesis [12], maintaining high overall abundances.

The high constant genetic microdiversity, mostly synonymous mutations, in all samples and oceans of vSAG 37-F6 seems to be evolutionarily preserved, which raises a fundamental question on whether preservation of these synonymous mutations provides a measurable fitness for vSAG 37-F6. Positive selection of genes/proteins (high nonsynonymous to synonymous substitution rates; i.e., *dN/dS*>1) that provide a fitness benefit is more obvious in biology. However, more intriguing is the interpretation of the large number of synonymous mutations observed in vSAG 37-F6 virus. Synonymous mutations can be related to viral codon usage optimization and adaptation to each host strain [17–21], or even they can generate new internal promoter sites that speed up viral transcription during infection, which ultimately is an advantage for viral replication [42, 43]. Here, we did not find any evidence supporting these lines of thought in vSAG 37-F6 (Supplementary Data 2). Furthermore, considering that *Pelagibacter* lacks CRISPR-Cas systems, the observed microdiversity does not seem apparently related to coping with the variability of host defense mechanisms of cooccurring host strains. Thus, the most parsimonious explanation is that this microdiversity might simply be evolutionarily neutral as a result of a large population size and high number of individuals/strains that fluctuate each after continuous, never ending infection cycles. These observations therefore imply that pelagiphages display truly gigantic effective population sizes, where standing microdiversity is ancient and maintained over long periods of time. This further suggests that microdiversity is not substantially affected by selective forces, such as selective sweeps, which are often thought to strongly impact viral evolution. One reason that could explain the high microdiversity of cooccurring viruses is likely related to the high diversity of their hosts based on BSC data. Indeed, most of the genomes of *Pelagibacter* spp. genomes analyzed in this study were found to constitute a single biological species, "sexually" isolated from other populations. It is therefore very likely that such a diverse population of isolated hosts contributes to maintaining high viral microdiversity. However, a positive selection of synonymous mutations cannot be ruled out since it has been described that synonymous mutations might provide certain benefits, such as improving the secondary structure of mRNA and therefore expression/translation [44], increasing transcriptional pausing favoring proper protein folding, reducing mRNA degradation [45] and/or 4) improving the binding sites of regulatory elements such as small RNA [46]. Furthermore, our data from putative active vSAG 37-F6-like viruses replicating in cells displaying high replication fidelity (Supplementary Fig. 17) along with previous culturomic studies on dsDNA pelagiphage isolates [47–50] do not point to error-prone polymerase, as with RNA viruses[39], as the cause of that observed high microdiversity.

Delineation of species is one of the most controversial paradigms addressed in microbiology [51, 52], especially in the era of metagenomics [53, 54]. Recently, the existence of true viral BSCs[34] driven by recombination has been proposed. Here, our gene flow analysis suggests signs of recombination in vSAG 37-F6 more in line with the BSC concept. However, at the same time, data pinpoint that mutation is a substantial contributor to the number of homoplasies detected in vSAG 37-F6, likely resulting from a large population size and individual abundance. Therefore, virus vSAG 37-F6 includes several viral variants that are likely clonal and/or composed of multiple biological species that do not engage in gene flow with one another. Most likely, both views are not mutually exclusive in the viral world, and different types of viruses in nature might behave more clonally or recombinantly. A global analysis of 627

mycobacteriophages [55] displayed rather continuous genetic diversity, such as vSAG 37-F6. On the other hand, as previously described, *"discrete populations"* of cyanophages have been isolated, detected [22, 56], and maintained in the long term by genetic recombination [56] and show an average nucleotide identity (ANI) between their homologous genes >98%. Nevertheless, the small genome size dataset along with the observed genomic microdiversity and the high genomic divergence preclude obtaining a robust conclusion on the BSC concept, if that truly exists, which in the case of the *Pelagibacter* host is more evident and driven mainly by recombination.

In viral ecology, in contrast to prokaryotes, we are far from unveiling fundamental questions such as the *in situ* abundances of cooccurring strains/species linked to viral community structure. In a previous study [57], we quantified the absolute abundances by digital PCR of free and infecting viral particles of a single viral strain out of the total pool of cooccurring strains belonging to vSAG 37-F6 species and reached up to several thousand viruses per mL (far from known values of total viruses in seawater; 10^6-10^7 per mL). This is somehow challenging since this virus is supposed to be putatively the most abundant virus in the temperate and tropical waters of the ocean[8], and *a priori* and intuitively, higher concentrations would be expected. However, our microdiversity data now help us to better conceive the structure of marine viral communities (Fig. 5). Each abundant cooccurring viral species in a sample, such as vSAG 37-F6, likely comprises up to thousands/hundreds of different strains, and each one of those in turn reaches several thousands of viral particles per mL. According to our data, the absolute *in situ* estimations of (micro)diverse viruses at the species/strain level seem to be a bottleneck in viral ecology.

Finally, high-quality sequencing data are critical to avoid spurious sequences in datasets (see our quality-trimming conditions in Methods). We estimated that in our sequencing data ($7.8x10^9$ nucleotides), potential errors represented only 0.03% of nucleotide positions. It is worth mentioning that our sequencing bias error was even lowered since only gene variants appearing at least ten times in each sample were considered. Thus, the effect of sequencing errors misleading our results is likely negligible. Additionally, unspecific PCR amplification bias of viruses actually not belonging to vSAG 37-F6 was ruled out in our study since the majority of sequences showed high nucleotide identity values (>95%) with the vSAG 37-F6 genome (Fig. 1, Supplementary Fig. 2). Here, we estimated the number of viral variants by ultradeep sequencing of hallmark genes from different genomic regions, which is a feasible conservative method. Undoubtedly, whole genome sequencing of all co-occurring viral variants in a sample would be the ideal method to capture the entire existing microdiversity of this virus. This could be addressed by an unprecedented large-scale sequencing project of hundred thousands of sorted single viruses recovered by SVG in combination with ultradeep metagenomic long-read sequencing [58–61], which could be further conducted under the umbrella of a large research consortium that could be applied to other ecologically relevant (uncultured) viruses in nature [62].

Methods

Marine Sample Collection and Processing

Mediterranean seawater samples were collected from three different locations (*Fig. 1*) i) Cape Huertas (Alicante coast, 38° 21′ 14.3″ N, 0° 25′ 36.6″ W on May 15, 2017), ii) Blanes Bay Microbial Observatory (BBMO) (41° 40′ 13.5″ N, 2° 48′ 0.6″ E; 2.7 miles offshore, on May 9, 2017) and iii) Mediterranean Sea REMEI Expedition. Samples from Cape Huertas and BBMO were collected from the surface (20 L each). From REMEI Expedition, a deep profile from the surface to 2,000 m depth samples was conducted, obtaining samples (100 L each) from the surface (5 m depth, 40° 49′ 16.2″ N, 3° 3′ 19.2″ E on September 27, 2017), deep chlorophyll maximum (DCM, 84 m depth, 40° 49′ 7.8″ N, 3° 3′ 58.8″ E on September 29, 2017), 1,000 m depth (40° 49′ 3.6″ N, 3° 3′ 55.8″ E on September 28, 2017), and 2,000 m depth (40° 49′ 21.6″ N, 3° 3′ 15″ E on September 27, 2017).

For all samples, seawater was filtered through a 0.22 µm membrane filter (Durapore membrane filters, Merck Millipore) to remove cell fraction. Then, the elute containing the viral fraction was concentrated by tangential flow filtration (TFF) using a Vivaflow 200 membrane (Sartorius) until a volume of 20 mL. Concentrated volume was filtered again through a 0.22 µm filter, to ensure the absence of cellular organisms. A final ultra-concentration was conducted using Amicon Ultra-15 centrifugal filters (100 KDa-cut off) until a 1 mL final volume was obtained.

Extracellular DNA was removed by applying a DNase treatment using 5 U of Turbo DNase I (Ambion) for 1 h at 37ºC according to the manufacturer´s protocol. Then, the kit QIAamp Ultrasense Virus (Qiagen) was employed to perform the extraction of viral nucleic acids according to the manufacturer´s protocol.

Specific primer design and PCR conditions

Five specific vSAG 37-F6 primer sets (named as *37-F6 Seq 1, Seq 4, Seq 6, Seq 11, and Seq 14;* Supplementary Fig. 1 and Supplementary Table 1) were used to amplify and sequence conserved hypothetical proteins of the vSAG-37-F6 virus (genes 2, 7, 8 and 24) and four hallmark capsid protein genes (genes 5, 6, 9 and 10). Primer design was as described in our previous work [27]. Briefly, optimal PCR oligos were designed to specifically target different genomic regions of virus vSAG 37-F6, and examined for hairpins, self-dimers, and hetero-dimers using Integrated DNA Technologies (IDT) web-based PrimerQuest tool and IDT´s OligoAnalyzer 3.1. Then, to check primer specificity, they were compared with a custom, comprehensive viral database containing 331,723 viral genomes obtained from different methods [8, 33, 63–65] and the GenBank Nucleotide collection (nr/nt) using Primer-BLAST [66]. Primer sets targeted different specific 37-F6 genes and/or intergenic regions (Supplementary Fig. 1). For instance, the primer set named *37-F6 Seq 14* targeted *gene 9* encoding a structural capsid protein that resulted to be the most abundant viral protein intemperate and tropical waters of the ocean [8, 37]. All primer sets contained the Illumina specific adapters to allow the amplicon sequencing (Supplementary Table 1). All primers were successfully tested using the DNA template of the original vSAG 37-F6 genome. PCR conditions were as follows: 2.0 ng of environmental extracted DNA, 200 nM each of forward and reverse primers, 200 nM of dNTPs, 2mM $MgCl_2$, 6%, BSCA, 3% DMSO and 1X PCR buffer and 0.5 U of Taq DNA polymerase recombinant (Invitrogen), in a final volume of 25 µL. Thermal cycling conditions were: an initial denaturation of 94°C for 4 min, followed by 40 cycles of 20 sec at 94°C,

30 sec at 52°C and 1 min of 72°C, and a final extension of 30 min at 72ºC. PCR products were visualized on a 1% agarose gel to ensure the correct length of the amplicons and the absence of non-specific products.

PCR amplicon high-throughput sequencing and analysis

PCR products from environmental samples were cleaned using GeneJET PCR Purification Kit (Thermo Scientific) and sequenced with the MiSeq instrument (paired-end 2x300 bp; Illumina) in the Fisabio Foundation (Valencia, Spain). Overlapping forward and reverse sequences were trimmed using Trimmomatic (*trimmomatic-x.xx.jar SE -phred33 amplicons_seqX_zoneX.fastq.gz amplicons_seqX_sampleX._trimmed.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN: 250*) [67], obtaining, at least, Q30 in the 98.9% of the trimmed read length. Expected sequencing errors were estimated using the Geneious bioinformatic software for sequence data analysis v8.1.7 (https://www.geneious.com). After trimming, identical amplicon reads (100% nucleotide identity and length) were clustered**.** To minimize differences due to sequence errors, only amplicons that appeared at least 10 times were considered for further analysis. Second, reads were grouped with a nucleotide identity cut-off of 95% using cd-hit as a proxy for a tentative viral-species clustering [68] (Table 1). Then, to identify the cluster where the vSAG 37-F6 was assigned (vSAG 37-F6-species) the reference genome from each cluster was compared using Blastn against a custom viral database containing 434,772 viral sequences (≥ 10 Kb length) obtained from marine vSAGs including the vSAG 37-F6 virus [8], Mediterranean viral fosmids [64, 65], viral sequences from SAGs [63] assembled contigs from *Tara* expedition [33], archaeal virus contigs [69, 70], long read assembled contigs [58] and sequences from IMG/VR2 viral database [36]. The best bit-score hit was used to assign each cluster to its corresponding viral sequence. The vSAG 37-F6-species was the cluster assigned to this genome with ≥ 95% nucleotide identity. Different amplicons within each cluster were quantified as the indicative value of the microdiversity (different strains within a viral species). It is important to note that not all targeted sequenced genes provided the same pattern of representativeness and microdiversity level but it might differ since not all genomic regions evolve under same evolutionary pressure.

Global micro-diversity and biogeography analysis of vSAG 37-F6-like and other pelagiphages

The microdiversity of the nucleotide sequences of different pelagiphage genomes was analyzed using a highly stringent, sensitive metagenomic fragment recruitment (see below for employed parameters) to calculate the genomic Shannon index [32] $H = - \Sigma P_i \bullet \ln P_i$. The value of the genomic *H,* (compressed between 0 and 1) is related to P_i the probability to find a different nucleotide, in a given reference genome position, in different mapped reads. Higher values of *H* represent higher possibilities to find different nucleotides and therefore correspond to higher genome microdiversity (i.e. number of polymorphic nucleotide sites). Microdiversity was calculated at the whole-genome level (mean entropy for all the nucleotides of the viral genome, Fig. 2A) and at the protein level (mean entropy for all nucleotides encoding a protein, Supplementary Figs. 4-9). For this global genome microdiversity analysis, a total of nine pelagiphage genomes were employed in this study: the abundant and widespread

vSAG 37-F6 and two other highly related viruses found in two *Pelagibacter* single cells (MED40_C1 and SAG AG-422-I02) [27] and a collection of six reference pelagiphages obtained from different *P. ubique* strains (five lytic pelagiphages HTVC010P, HTVC008M, HTVC023P, HTVC111P, HTVC115P, and the lysogenic phage PNP1) [47, 48, 50]. A Tukey HSD test was performed to compare the global microdiversity of each pelagiphage. Previous studies reported that viral genetic microdiversity could correlate with viral abundance [24], our results showed that this correlation illustrate a situation when a virus is at low abundance in a metagenomic dataset. Our data show that, although abundant viruses can obtain higher microdiversity values, viruses are able to reach a maximum value of microdiversity, that remains constant although their abundance increase (Supplementary Figs. 3 and 4).

Reads from metagenomes (cell fraction) and viromes (virus fraction) obtained from *Tara Oceanic* expedition [33] and SPOT time series [24] (Supplementary Table 2) were mapped against the pelagiphages using the *very-sensitive* mode of Bowtie2 [25]. Every recruitment (each virus with each metagenome/virome) was performed by separate. Synonymous, non-synonymous mutations, and entropy were calculated for each protein separately (n=690) using DiversiTools (http://josephhughes.github.io/DiversiTools/ , Fig. 2B, Supplementary Figs 5-9 and Supplementary Table 4). The non-synonymous ratio was calculated as *NS ratio* = 100 * *NSm/(NSm + Sm*), considering only those proteins with an average amino acid coverage (AAcov) of at least 100x. This value is similar to the common dN/dS, allowing a parallel interpretation (NSr > 50, NSr <50 and NSr = 50 means positive, negative or neutral selection, respectively, as *dN/dS* >1, *dN/dS*<1 and *dN/dS* =1), and avoiding the erroneous value obtained by *dN/dS* when *dS* = 0.

Global diversity of the highly related vSAG 37-F6 viruses

To globally identify the most related vSAG 37-F6 phages, a blastp for each vSAG 37-F6 protein (n =25) was performed against the global viral protein database IMG/VR v.2.0 (n = 17,869,415 proteins) [36]. After blastp, only hits with at least 50% amino acid similarity and ≥ 95% query coverage were analyzed (i.e. homologous proteins). All viral genomes that contained at least 12 homologous proteins were selected, as highly similar vSAG 37-F6-like viruses (Supplementary Tables 5 and 6).

Construction of hidden Markov model profiles to analyze the diversity of the vSAG 37-F6

Other approximation to find viral relatives of virus vSAG 37-F6 was carried out using hidden Markov models (HMMs). Using the homologous proteins obtained from the vSAG 37-F6 similar viruses, only the proteins that appear in at least 22 similar viruses, with an amino acid similarity > 80% and query coverage > 95% were selected (Supplementary Fig. 10). Using these subsets of proteins (gene 9, 11, 22, and 24) an HMM was built for each one. Firstly, each group of proteins was aligned using *Clustal Omega aligner*, then *HMMER package v3.2.1* was employed to build the HMM profiles using the alignments and the *hmmbuild* tool. Finally, to find viruses containing proteins with these structural models, the *hmmsearch* tool was used. Viral contigs contained the four models in their genome were the most related viruses with the vSAG 37-F6 by this methodology (Supplementary Figs. 13 and 14).

Identification of viral relatives of virus vSAG 37-F6 from non-marine environments

The highly abundant vSAG 37-F6 g9 was employed to mine the IMG/VR v.2.0 database [36] searching for similar vSAG 37-F6 viruses from non-marine habitats. This protein was previously found to be a suitable gene marker for this group[8, 27]. Homologous vSAG 37-F6 g9 proteins (amino acid similarity ≥ 50% and query coverage ≥ 95%) were found by blastp in the IMG/VR v.2.0 database [36]. Using the vConTACT2 [71] (default parameters), viral genomes containing the g9 homologous proteins were used to build a protein shared network, and group viral genomes in viral clusters (VCs). Finally, the vSAG 37-F6 related VCs were selected and analyzed to find viral contigs from different environments (Supplementary Figure 15).

Analysis of biological species concept within vSAG 37-F6-like group: estimation of the number of species and gene flow

Different viral species were defined based on the recently described viral biological species concept (BSC) within the virus vSAG 37-F6 and viral relatives[34]. Previously identified and related vSAG 37-F6 viral contigs found in SAGs MED40-C1, AG-422-I02, AG-470-G06, JGI BSCAE-1614-1.M18, AAA164I21, and AAA160P02, and the viral contig KT997850 (fosmid from the deep Mediterranean Sea) were also employed for this analysis [8, 27, 65], which resulted in a total of 32 viruses related to vSAG 37-F6 virus (Supplementary Tables 5 and 6). Estimation of the ratio of homoplasic (*h*) to non-homoplasic (*m*) polymorphisms along the genome was performed as previously described [34]. High *h*/*m* ratios are indicative of a substantial signal of gene flow, and low *h*/*m* ratios are indicative of clonal or nearly clonal evolution. In addition, a simulated analysis was performed to estimate the proportions of homoplasies expected to result from convergent mutations rather than recombination as in [34].

Estimation of gene flow analysis and the number of species in Pelagibacter spp.

Gene flow analysis was performed using the distance-based method of *ConSpeciFix* [72]. Groups of genomes were considered part of the same biological species when found to engage in gene flow, whereas genomes whose inclusion led to a substantial drop in gene flow (exclusion criterion) were classified as different biological species as previously described [73]. First, we collected a set of 926 genomes from different Pelagibacter strains obtained by single-cell genomics, metagenomics, and culturomics surveys available in public databases. Coding sequence (CDS) prediction was performed on all genomes using *Prodigal* v2.6.3. [74]. To build the core genome, orthologous clusters (OCs) were first identified by pairwise genome comparisons among the whole protein sequences from CDSs using *USEARCH* Global v8.0 [75] implemented in *CoreCruncher* [76]. OCs were defined as sharing at least 50% protein sequence identity and 50% sequence length. Each OC was considered part of the core genome if found in >85% of the genomes. Protein sequences from each core gene were then aligned using *MUSCLE* v3.8.31 [76] [77] with default parameters. The corresponding nucleotide sequence alignments were then generated by mapping each codon to the corresponding amino acid based on the protein sequence alignment using a python script, and the nucleotide alignments of each gene were concatenated into a single large alignment. Subsequently, core nucleotide identity (CNI) values were used to calculate genomic similarities from the core genome alignment of the 926 *Pelagibacter* genomes. Pairwise CNI was computed using the *distmat*

tool of EMBOSS version 6.6.0.0, [78] which calculates the pairwise nucleotide identities from the alignment as previously described [79]. Then, single linkage clustering was performed: all genome pairs with a CNI similarity threshold of 94% or higher were joined together and clustered into *de novo* species. A maximum likelihood phylogenomic tree based on the core nucleotide alignment among all *Pelagibacter* genomes was built using GTR+CAT model with the FastTree software version 2.0.0. [80]. Branch supports were evaluated by generating 100 bootstrap replicates using the same parameters.

From the clusters of genomes defined based on CNI, we selected each cluster with ≥15 genomes or more and used them as "reference clusters". Then, we tested each of this reference cluster against one genome of each other clusters (named "candidate clusters"). For each comparison of a candidate cluster against a reference cluster, the core genes shared by both clusters were then aligned and concatenated as described above. The resulting core genome was then used to compute a distance matrix using *RAxML* [81] version 8.2.12 with a GTR + Gamma model. From these distances and the core genome concatenate, the ratio of homoplasic to non-homoplasic alleles (*h/m*) was computed for each comparison (i.e. the set of genomes of the reference cluster + the candidate genome tested) and for the set of genomes of the reference cluster alone. From this step, graphs and statistics comparing *h/m* ratios between the genomes of each reference cluster with and without the candidate genome were inferred. In addition, a simulated genome was generated for each reference cluster to estimate the proportions of homoplasies expected to result from convergent mutations rather than recombination [73, 82]. The simulated sequence was first generated from the consensus sequence of the core genome concatenate of the reference cluster. Point mutations were then introduced *in silico* with a Jukes and Cantor model until the same sequence divergence was obtained as the one observed between the genomes of the reference cluster and the candidate genome. This analysis was conducted independently for each of the 3,458 comparisons of reference clusters against candidate clusters.

Analysis of vSAG 37-F6-like pelagiphage in Pelagibacter host cells

All proteins of the vSAG 37-F6 and the related viral contigs, found in the Pelagibacter MED40 and SAG AG-422-I02 were compared with every *Pelagibacter* spp proteins employing blastp (amino acid similarity ≥ 50% and query coverage ≥ 95%). Putative infected *Pelagibacter* spp genomes that contained at least 9 similar proteins were selected and their phylogenetic relationship was analyzed employing the BSC classification (Supplementary Fig. 12 and Supplementary Tables 8 and 9).

Analysis of viral-host codon usage and promoter sites detection

Codon usage was calculated, using the online tool https://www.kazusa.or.jp/codon/countcodon.html , for the vSAG 37-F6-like pelagiphages (MED40-C1 and the viral contig found in the SAG AG-422-I02), isolated pelagiphages (HTVC010P and HTVC023P), their hosts (SAGs MED40 and AG-422-I02, and the isolate Pelagibacter HTCC1062) and other marine phages for representative groups (Alteromonas phage AD45, Cellulophaga phage phi18, the Cyanophages P-SSP2 and S-TIM4, and the Flavobacterium phage 11).

To check the presence of promoter sites in the viral genome, the online tool *Bacterial Promoter Prediction* (BPP, http://www.bacpp.bioinfoucs.com/home) was employed, checking the sigma factors 24, 28, 32, 38, 54 and 70.

Data availability

vSAG 37-F6 Illumina amplicons sequenced in this study can be accessed at the SRA database in the BioSample accessions: SAMN18521786 – 18521791.

References

1. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIUVIG). Nat. Biotechnol. 2019; 37: 29–37.

2. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature 2016; 536: 425–430.

3. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell 2019; 177: 1109–1123.

4. Kavagutti VS, Andrei AŞ, Mehrshad M, Salcher MM, Ghai R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome 2019; 7: 1–15.

5. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, et al. Hidden diversity of soil giant viruses. Nat. Commun. 2018; 9: 1–9.

6. Trubl G, Jang H Bin, Roux S, Emerson JB, Solonenko N, Vik DR, et al. Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. mSystems 2018; 3.

7. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, et al. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host Microbe 2018; 24: 653-664.e6.

8. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Peña MJ, Martínez JM, et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat. Commun. 2017; 8: 1–13.

9. Aguirre de Cárcer D, Angly FE, Alcamí A. Evaluation of viral genome assembly and diversity estimation in deep metagenomes. BMC Genomics 2014; 15: 1–12.

10. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ 2017; 5: e3817.

11. Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. Genomic island variability facilitates *Prochlorococcus*-virus coexistence. Nature 2011; 474: 604–608.

12. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. 2009; 7: 828–836.

13. Marston MF, Pierciey FJ, Shepard A, Gearin G, Qi J, Yandava C, et al. Rapid diversification of coevolving marine *Synechococcus* and a virus. Proc. Natl. Acad. Sci. U.S.A. 2012; 109: 4544–4549.

14. Enav H, Kirzner S, Lindell D, Mandel-Gutfreund Y, Béjà O. Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. Nat. Comm. 2018; 9: 1–11.

15. Boon M, Holtappels D, Lood C, van Noort V, Lavigne R. Host Range Expansion of Pseudomonas Virus LUZ7 Is Driven by a Conserved Tail Fiber Mutation . PHAGE 2020; 1: 87–90.

16. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. Nat. Rev. Microbiol. 2020; 18: 113–119

17. Sørensen MA, Kurland CG, Pedersen S. Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. 1989; 207: 365–377.

18. Varenne S, Buc J, Lloubes R, Lazdunski C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide

chains. J. Mol. Biol. 1984; 180: 549–576.

19. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol. Cell 2015; 59: 744–754.

20. Plotkin JB, Kudla G. Synonymous but not the same: The causes and consequences of codon bias. Nat. Rev. Genet. 2011; 12: 32–42

21. Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer 2019; 19: 359.

22. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, et al. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. Nature 2014; 513: 242–245.

23. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat. Microbiol. 2019; 4: 1727–1736.

24. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat. Microbiol. 2019; 1–7.

25. Coutinho FH, Rosselli R, Rodríguez-Valera F. Trends of Microdiversity Reveal Depth-Dependent Evolutionary Strategies of Viruses in the Mediterranean. mSystems 2019; 4: 1–17.

26. Needham DM, Sachdeva R, Fuhrman JA. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. ISME J 2017; 11: 1614–1629.

27. Martinez-Hernandez F, Fornas Ò, Lluesma Gomez M, Garcia-Heredia I, Maestre-Carballa L, López-Pérez M, et al. Single-cell genomics uncover *Pelagibacter* as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. ISME J 2019; 13: 232–236.

28. McMullen A, Martinez-Hernandez F, Martinez-Garcia M. Absolute

603 quantification of infecting viral particles by chip-based digital polymerase chain
604 reaction. Environ. Microbiol. Rep. 2019; 11: 855–860.

605 29. Marston MF, Amrich CG. Recombination and microdiversity in coastal marine
606 cyanophages. Environ. Microbiol. 2009; 11: 2893–2903.

607 30. Marston MF, Martiny JBH. Genomic diversification of marine cyanophages into
608 stable ecotypes. Environ. Microbiol. 2016; 18: 4240–4253.

609 31. Cordero OX. Endemic cyanophages and the puzzle of phage-bacteria
610 coevolution. Environ. Microbiol. 2017; 19: 420–422.

611 32. Shannon CE. The mathematical theory of communication. 1963. MD computing
612 1997; 14: 306–317.

613 33. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al.
614 Ecogenomics and potential biogeochemical impacts of globally abundant ocean
615 viruses. Nature 2016; 537: 689–693.

616 34. Bobay L-M, Ochman H. Biological species in the viral world. Proc. Natl. Acad.
617 Sci. U.S.A. 2018; 115: 6040–6045.

618 35. Henson MW, Lanclos VC, Faircloth BC, Thrash JC. Cultivation and genomics of
619 the first freshwater SAR11 (LD12) isolate. ISME J 2018; 12: 1846–1860.

620 36. Paez-Espino D, Roux S, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al.
621 IMG/VR v.2.0: an integrated data management and analysis system for cultivated
622 and environmental viral genomes. Nucleic Acids Res. 2019; 47: D678–D686.

623 37. Brum JR, Ignacio-Espinoza JC, Kim E-H, Trubl G, Jones RM, Roux S, et al.
624 Illuminating structural proteins in viral ‘dark matter’ with metaproteomics. Proc.
625 Natl. Acad. Sci. U.S.A. 2016; 113: 2436–2441.

626 38. Sakowski EG, Arora-Williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB,
627 et al. Interaction dynamics and virus–host range for estuarine actinophages
628 captured by epicPCR. Nat. Microbiol. 2021; 1–13.

629 39. Alonso-Sáez L, Morán XAG, Clokie MR. Low activity of lytic pelagiphages in

630 coastal marine waters. ISME J 2018; 12: 2100–2102.

631 40. Martinez-Hernandez F, Luo E, Tominaga K, Ogata H, Yoshida T, DeLong EF, et
632 al. Diel cycling of the cosmopolitan abundant Pelagibacter virus 37-F6: one of
633 the most abundant viruses in Earth. Environ. Microbiol. Rep. 2020; 1758-
634 2229.12825.

635 41. Mruwat N, Carlson MCG, Goldin S, Ribalet F, Kirzner S, Hulata Y, et al. A
636 single-cell polony method reveals low levels of infected *Prochlorococcus* in
637 oligotrophic waters despite high cyanophage abundances. ISME J 2021; 15: 41–
638 54.

639 42. de Avila e Silva S, Echeverrigaray S, Gerhardt GJL. BacPP: Bacterial promoter
640 prediction-A tool for accurate sigma-factor specific assignment in enterobacteria.
641 J. Theor. Biol. 2011; 287: 92–99.

642 43. Sampaio M, Rocha M, Oliveira H, Dias O. Predicting promoters in phage
643 genomes using PhagePromoter. Bioinformatics 2019; 35: 5301–5302.

644 44. Allert M, Cox JC, Hellinga HW. Multifactorial Determinants of Protein
645 Expression in Prokaryotic Open Reading Frames. J. Mol. Biol 2010; 402: 905–
646 918.

647 45. Dressaire C, Picard F, Redon E, Loubière P, Queinnec I, Girbal L, et al. Role of
648 mRNA Stability during Bacterial Adaptation. PLoS ONE 2013; 8.

649 46. Deana A, Belasco JG. Lost in translation: The influence of ribosomes on
650 bacterial mRNA decay. Genes Dev. 2005; 19: 2526–2533

651 47. Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, et al.
652 Abundant SAR11 viruses in the ocean. Nature 2013; 494: 357–360.

653 48. Zhang Z, Qin F, Chen F, Chu X, Luo H, Zhang R, et al. Culturing novel and
654 abundant pelagiphages in the ocean. Environ. Microbiol. 2020; 1462-
655 2920.15272.

656 49. Zhao Y, Qin F, Zhang R, Giovannoni SJ, Zhang Z, Sun J, et al. Pelagiphages in
657 the Podoviridae family integrate into host genomes. Environ. Microbiol. 2018.

50. Morris RM, Cain KR, Hvorecny KL, Kollman JM. Lysogenic host–virus interactions in SAR11 marine bacteria. Nat. Microbiol. 2020; 5: 1011–1015.

51. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 2006; 361: 1929–1940.

52. Rosselló-Mora R. Updating prokaryotic taxonomy. J. Bacteriol. 2005; 187: 6255–6257

53. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2017; 2: 1533–1542.

54. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. 2009; 106: 19126–19131.

55. Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, et al. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. eLife 2015; 4.

56. Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, et al. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC genomics 2016; 17: 930.

57. Martinez-Hernandez F, Garcia-Heredia I, Lluesma Gomez M, Maestre-Carballa L, Martínez Martínez J, Martinez-Garcia M. Droplet Digital PCR for Estimating Absolute Abundances of Widespread Pelagibacter Viruses. Front. Microbiol. 2019; 10: 1226.

58. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ 2019; 7: e6800.

59. Beaulaurier J, Luo E, Eppley JM, Uyl P Den, Dai X, Burger A, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res. 2020; 30: 437–446.

60. Murigneux V, Rai SK, Furtado A, Bruxner TJC, Tian W, Harliwong I, et al.

Comparison of long-read methods for sequencing and assembly of a plant genome. GigaScience 2020; 9.

61. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 2019; 37: 1155–1162.

62. Martínez Martínez J, Martinez-Hernandez F, Martinez-Garcia M. Single-virus genomics and beyond. Nat. Rev. Microbiol. . 2020; 18: 705–716

63. Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, et al. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. ISME J 2015; 9: 2386–2399.

64. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the marine virosphere using metagenomics. PLoS Genet. 2013; 9: e1003987.

65. Mizuno CM, Ghai R, Saghaï A, López-García P, Rodriguez-Valera F. Genomes of abundant and widespread viruses from the deep ocean. mBio 2016; 7: e00805-16.

66. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics 2012; 13: 134.

67. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114–2120.

68. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22: 1658–1659.

69. Philosof A, Yutin N, Flores-Uribe J, Sharon I, Koonin E V., Béjà O. Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota. Curr. Biol 2017; 27: 1362–1368.

70. Vik DR, Roux S, Brum JR, Bolduc B, Emerson JB, Padilla CC, et al. Putative archaeal viruses from the mesopelagic ocean. PeerJ 2017; 5: e3428.

71. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. 2019; 37: 632–639.

72. Bobay L-M, Ellis BS-H, Ochman H. ConSpeciFix: classifying prokaryotic species based on gene flow. Bioinformatics 2018; 34: 3738–3740.

73. Bobay L-M, Ochman H. Biological Species Are Universal across Life's Domains. Genome Biol. Evol. 2017; 9: 491–501.

74. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 2010; 11: 119.

75. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010; 26: 2460–2461.

76. Harris CD, Torrance EL, Raymann K, Bobay L-M. CoreCruncher : Fast and Robust Construction of Core Genomes in Large Prokaryotic Data Sets . Mol. Biol. Evol. 2020.

77. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792–1797.

78. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000. Elsevier Ltd. , 16: 276–277

79. Džunková M, Low SJ, Daly JN, Deng L, Rinke C, Hugenholtz P. Defining the human gut host–phage network through single-cell viral tagging. Nat. Microbiol. 2019; 4: 2192–2203.

80. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 2010; 5: e9490.

81. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30: 1312–1313.

82. Swan BK, Ehrhardt CJ, Reifel KM, Moreno LI, Valentine DL. Archaeal and

bacterial communities respond differently to environmental gradients in anoxic sediments of a california hypersaline lake, the Salton Sea. Appl. Environ. Microbiol. 2010; 76: 757–768.

83. Baran N, Goldin S, Maidanik I, Lindell D. Quantification of diverse virus populations in the environment using the polony method. Nat. Microbiol. 2018; 3: 62–72.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation (RTI2018-094248-B-I00), Gordon and Betty Moore Foundation (grant 5334) and Generalitat Valenciana (ACIF/2015/332 and APOSTD/2020/237). We thank Dr. Josep Gasol for giving us access to collecting samples from REMEI Expedition.

Author information

Affiliations

Department of Physiology, Genetics, and Microbiology, University of Alicante, Carretera San Vicente del Raspeig, San Vicente del Raspeig, Alicante, Spain
Francisco Martinez-Hernandez, Inmaculada Garcia-Heredia & Manuel Martinez-Garcia

Department of Biology, University of North Carolina at Greensboro, USA.
Awa Diop & Louis-Marie Bobay

Contributions

M.M-G conceived and led the study. F.M-H. led the analyses and interpretation of data. A.D and L-M.B led the biological specie analysis and interpretation of data. M.M-G and F.M-H wrote the paper.

Corresponding author

Correspondence to Manuel Martinez-Garcia.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary Data

Supplementary Data 1. *Pelagibacter* spp. classification. Each *Pelagibacter* sp genome has been assigned to a cluster. Classification has been based on the pairwaise core genome comparison (Core Nucleotide Identity, CNI), or the gene flow (*h/m* ratio, Biological Species Concept, BSC).

Supplementary Data 2. Codon usage table of some viruses and their hosts. Percentage of codons for each virus and its host is showed. "*Total proportion*" columns represent the usage of each codon regarding all codons present in the genome (for all amino-acids), "*proportion*" columns show the usage percentage of each codon encoding for an specific amino-acid.

Main Figures and Tables

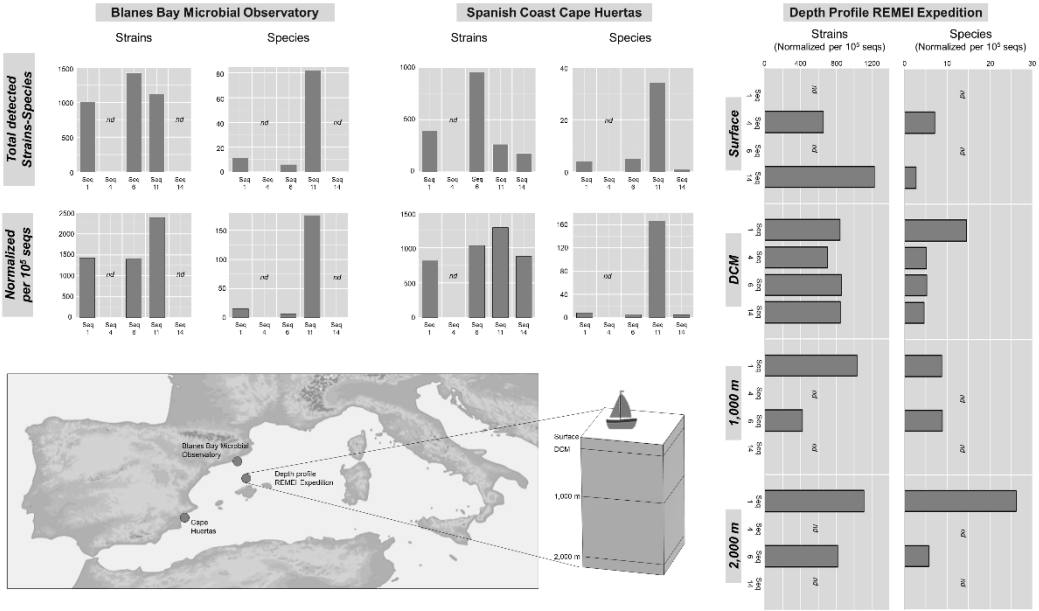

Fig 1. Local micro- and macrodiversity of virus vSAG 37-F6 at the strain and species levels. Genetic diversity of virus vSAG 37-F6 and viral relatives evaluated using Illumina amplicon sequencing at different locations from the Mediterranean Sea. Total detected vSAG 37-F6-like strains and species from coastal samples are showed in graph bar. To allow comparison between samples (from the coast and the offshore depth profile) absolute number of strains/species were normalized per each 10^5 sequenced amplicons.

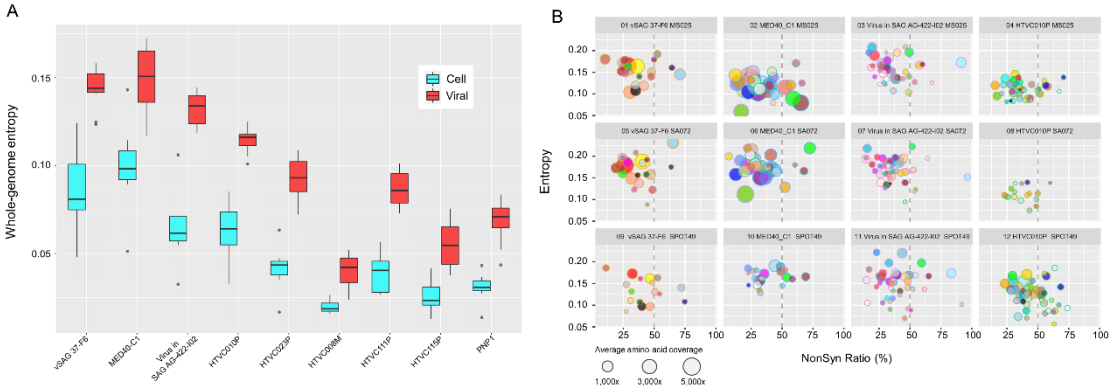

Fig 2. Global microdiversity of pelagiphages. Microdiversity at a global ocean scale of different pelagiphages. A) Whole genome entropy values (i.e. genomic microdiversity) obtained for all pelagiphages analyzed in the cell fraction (blue boxplots) and the viral fraction (red boxplots). Vertical lines indicate the standard deviation of the whole genome values calculated for each virome or metagenome. Significant differences were found between vSAG 37-F6-like pelagiphages and isolated genomes (not depicted for convenience in the figure but available in Supplementary Table 3). B) Non-synonymous and synonymous rates of pelagiphage proteomes. Each protein is represented by a circle. The area is proportional to their average amino acid coverage (abundance). Circles located to the right of the dashed line depicts proteins in which non-synonymous mutations prevail (i.e. dn/ds>1; positive selection), while circles located to the left of the dashed line depicts proteins in which synonymous mutations prevail (i.e. dn/ds<1; negative selection).

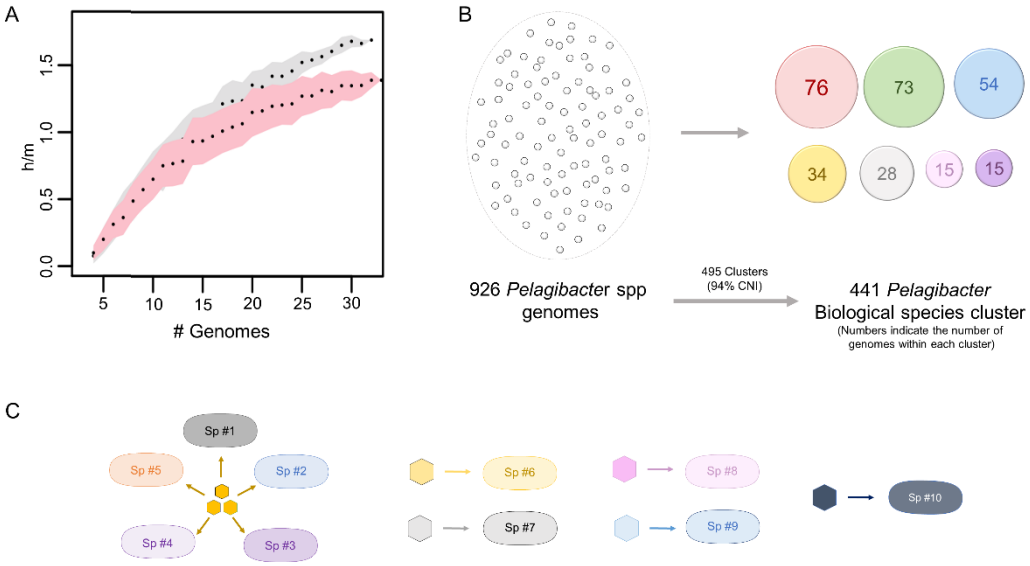

Fig 3. The biological species concept within vSAG 37-F6-like pelagiphages – Pelagibacter* spp.** ***A) Gene flow analysis based on homoplasies-mutation (h/m) rate of closed vSAG 37-F6 related viruses to determine existence of true biological species (BSC). Grey curve represents the h/m rate of the analyzed viruses, while the pink curve shows the value of a simulated dataset. ***B)*** Gene flow analysis based on homoplasies-mutation (h/m) rate of >900 *Pelagibacter* genomes to determine the number of true biological species in databases. ***C)*** Novel vSAG 37-F6-like pelagiphages were found infecting different true biological species of *Pelagibacter*. One viral species was able to infect up to five different *Pelagibacter* BSC.

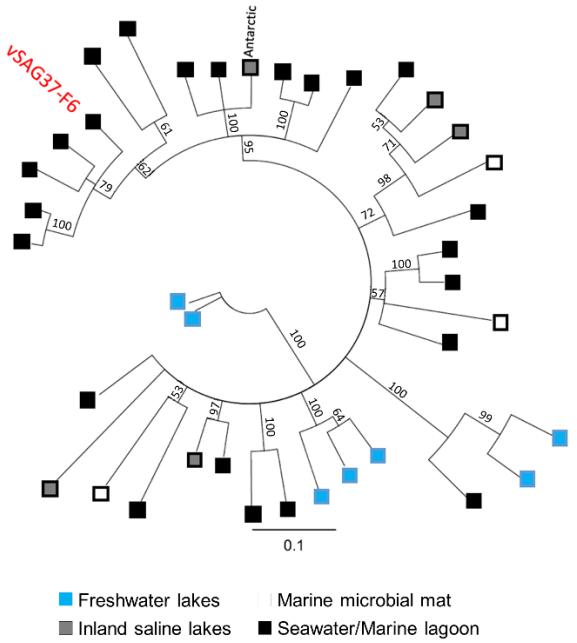

Fig 4. Global phylogeography and evolution of vSAG 37-F6-like viruses. Protein alignment and phylogeny of vSAG 37-F6 capsid protein found in non-marine environments.

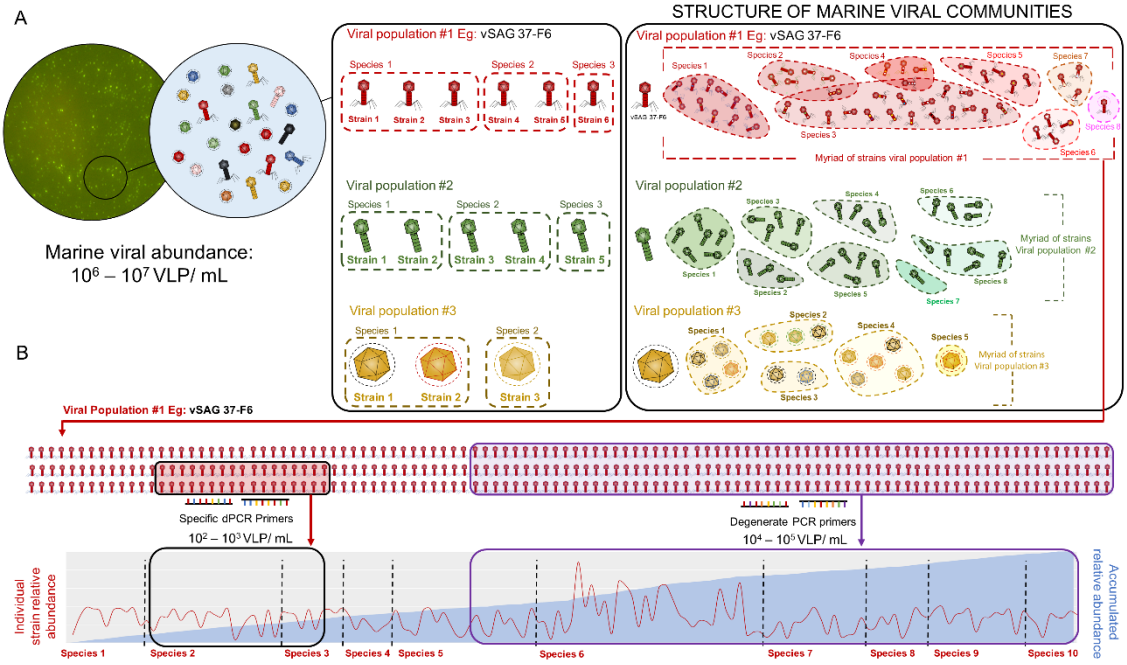

Fig. 5. ***Microstructure of viral communities in marine ecosystems***. A) Unprecedent values of co-occurring vSAG 37-F6 viral strains suggest a more complex structure of marine viral communities. Thousands of different strains within the same dsDNA viral species can co-exist in a sample generating a complex myriad of viral strains/variants. B) High microdiversity values hamper absolute in situ quantification of viruses at the species and strain levels in nature (e.g. digital PCR[57] or polony PCR[41, 83] targeting one strain or different viral species, respectively). Graph (bottom) depicts a conceptual model of the relative abundance of a microdiverse virus in nature. Red line represents the abundance of each strain, and blue area indicates the accumulated abundance.

Main Table

Table 1. Sequencing of hallmark genes of vSAG 37-F6 virus

Zone	Genome	Total Abundance[1]	# Strains[2]	Normalized # strains (per 100,000 seqs)[3]	# Species[4]	Normalized # Species (per 100,000 seqs)[5]	% ID vSAG 37-F6 sp[6]	Strains within vSAG 37-F6 sp[7] (%)	Relative abundance vSAG 37-F6 sp[8] (%)
BBMO	**Seq 1**	70,613.0	1,004.0	1,421.8	11.0	15.6	100	46.8	37.9
	Seq 4	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 6	101,840.0	1,422.0	1,396.3	6.0	5.9	95.2	82.7	53.6
	Seq 11	46,781.0	1,116.0	2,385.6	82.0	175.3	98.9	0.6	0.5
	Seq 14	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
Cape	**Seq 1**	48,374.0	392.0	810.4	4.0	8.3	98.8	60.2	47.1
Huertas	**Seq 4**	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 6	92,770.0	957.0	1,031.6	5.0	5.4	99.4	61.7	17.1
	Seq 11	20,475.0	264.0	1,289.4	34.0	166.1	98.9	8.0	10.5
	Seq 14	20,387.0	179.0	878.0	1.0	4.9	97.9	100	100
Surface	**Seq 1**	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 4	70,344.0	461.0	655.4	5.0	7.1	100	97.4	99.6
	Seq 6	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 11	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 14	36,737.0	452.0	1,230.4	1.0	2.7	97.9	100	100
DCM	**Seq 1**	27,455.0	232.0	845.0	4.0	14.6	99.8	98.7	99.8
	Seq 4	96,958.0	684.0	705.5	5.0	5.2	100	96.1	97.0
	Seq 6	95,132.0	815.0	856.7	5.0	5.3	96.6	96.4	85.3
	Seq 11	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 14	21,614.0	183.0	846.7	1.0	4.6	97.9	100	100
1,000 m	**Seq 1**	34,469.0	356.0	1,032.8	3.0	8.7	100	78.7	69.3
	Seq 4	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 6	34,662.0	146.0	421.2	3.0	8.7	100	43.2	34.6

	Seq 11	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 14	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
2,000 m	**Seq 1**	7,646.0	85.0	1,111.7	2.0	26.2	100	98.8	99.4
	Seq 4	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 6	87,596.0	717.0	818.5	5.0	5.7	100	70.0	52.4
	Seq 11	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
	Seq 14	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*	*nd*
All	**Seq 1**	188,577.0	1,279.0	678.2	13.0	6.9	99.7	56.7	57.5
	Seq 4	167,302.0	730.0	436.3	5.0	3.0	100	96.2	98.1
	Seq 6	412,000.0	3,003.0	728.9	7.0	1.7	97.4	36.4	51.0
	Seq 11	67,260.0	1,279.0	1,901.6	97.0	144.2	98.9	2.0	3.5
	Seq 14	78,738.0	597.0	758.2	1.0	1.3	97.9	100	100

[1] Only joined trimmed amplicons that appear at least 10 times in each zone were considered.

[2] Number of different sequences within the total amplicons, representing the number of different vSAG 37-F6 strains.

[3] Number of different vSAG 37-F6 strains normalized per each 100,000 sequenced amplicons.

[4] Number of 95% nucleotide identity clusters (C95) representing the number of vSAG 37-F6-like species.

[5] Number of different vSAG 37-F6-like species normalized per each 100,000 sequenced amplicons.

[6] Percentage of nucleotide identity between the vSAG 37-F6 genome and the reference genome of the assigned cluster (C95)

[7] Number of different sequences (strains) within the vSAG 37-F6 assigned C95 (vSAG 37-F6 species), representing the vSAG 37-F6 sp microdiversity.

[8] Percentage of total amplicons within the vSAG 37-F6 assigned C95