



Novel RNA viruses from the Atlantic Ocean: Ecogenomics, biogeography, and total viroplankton mass contribution from surface to the deep ocean

Marina Vila-Nistal^{1,2}  | Lucia Maestre-Carballea^{1,2} |
Francisco Martinez-Hernández¹ | Manuel Martinez-Garcia^{1,2} 

¹Department of Physiology, Genetics, and Microbiology, University of Alicante, Alicante, Spain

²Multidisciplinary Institute for Environmental Studies (IMEM), University of Alicante, Alicante, Spain

Correspondence

Manuel Martinez-Garcia, Department of Physiology, Genetics, and Microbiology and Multidisciplinary Institute for Environmental Studies (IMEM), University of Alicante, Carretera San Vicente del Raspeig, San Vicente del Raspeig, Alicante, 03690, Spain. Email: m.martinez@ua.es

Funding information

Generalitat Valenciana, Grant/Award Number: ACIF2020; Gordon and Betty Moore Foundation, Grant/Award Number: 5334; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2021-125175OB-I00; Ministerio de Economía y Competitividad, Grant/Award Number: RTI2018-094248-B-I00

Abstract

Marine viruses play a major role in the energy and nutrient cycle and affect the evolution of their hosts. Despite their importance, there is still little knowledge about RNA viruses. Here, we have explored the Atlantic Ocean, from surface to deep (4.296 m), and used viromics and quantitative methods to unveil the genomics, biogeography, and the mass contribution of RNA viruses to the total viroplankton. A total of 2481 putative RNA viral contigs (>500 bp) and 107 larger bona fide RNA viral genomes (>2.5 kb) were identified; 88 of them representing novel viruses belonging mostly to two clades: *Yangshan assemblage* (sister clade to the class *Alsuviricetes*) and *Nodaviridae*. These viruses were highly endemic and locally abundant, with little or no presence in other oceans since only $\approx 15\%$ of them were found in at least one of the *Tara* sampling metatranscriptomes. Quantitative data indicated that the abundance of RNA viruses in the surface and deep chlorophyll maximum zone was within $\approx 10^6$ VLP/mL representing a potential contribution of 5.2%–24.4% to the total viroplankton community (DNA and RNA viruses), with DNA viruses being the predominant members ($\approx 10^7$ VLP/mL). However, for the deep sample, the observed trend was the opposite, although as further discussed, several biases should be considered. Together these results contribute to our understanding of the diversity, abundance, and distribution of RNA viruses in the oceans and provide a basis for further investigation into their ecological roles and biogeography.

INTRODUCTION

Viruses are the most abundant organisms in the planet (Liang & Bushman, 2021) and can be found across all habitats. In the oceans, viruses play a major role in the energy and nutrient cycle. They take part in the transference of organic carbon from higher trophic levels to decomposers (Fuhrman, 1999; Suttle, 2007) cycling more than a quarter of the photosynthetically fixed carbon (Forterre, 2013). In addition, viruses drive the evolution of the plankton by serving as vehicles for gene transfer (Forterre, 2013) and increase the population

diversity by selectively killing their hosts, as the spreading of infections is density dependent, preventing the dominance of single species (Thingstad et al., 1993).

The focus of studies on marine viroplankton is primarily on DNA viruses. The reason behind this is that most of the majority of free viral particles found in the ocean are bacteriophages, which are predominantly dsDNA viruses (Culley & Steward, 2007; Maranger et al., 1994; Wommack et al., 1992). In contrast, RNA viruses tend to infect protists and other eukaryotic organisms, which makes their hosts fairly less abundant (Lang et al., 2009). Furthermore, the study of RNA

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Environmental Microbiology* published by Applied Microbiology International and John Wiley & Sons Ltd.

viruses presents both culture-dependent and -independent methodological limitations due to the difficulty of isolating viruses and hosts in addition to the relatively smaller genome size and higher taxonomic diversity that complicate the metagenomic surveys (Liao et al., 2022). These reasons explain why the study of RNA viruses has remained relatively unexplored by the scientific community until recent years (Callanan et al., 2020; Dominguez-Huerta et al., 2023; Zayed et al., 2022).

Over the course of the last decade, metagenomic studies have brought to light that RNA viruses are widespread and diverse, particularly the ssRNA viruses (Dominguez-Huerta et al., 2023). In certain marine environments, these viruses have been found to be as abundant as DNA viruses (Culley et al., 2006; Culley & Steward, 2007). Moreover, the significance of RNA viruses in biogeochemical cycles has been shown to be greater than previously believed, with specific RNA viral species having a direct impact on ocean carbon export (Dominguez-Huerta et al., 2022). But all these previous studies have been conducted in specific environments with limited areas of focus, primarily in coastal waters from temperate regions and the Antarctic ocean, or using metaviromes or metatranscriptomes available from the *Tara* dataset (Culley et al., 2006; Culley et al., 2014; Culley & Steward, 2007; Liao et al., 2022; Miranda et al., 2016).

In this study, we have explored the Atlantic Ocean, with sampling sites that span from coastal surface waters to 4296 m depth, and identified a total of 2.481 putative RNA viral scaffolds (>500 bp) and 107 bona fide RNA viruses with a genome size larger than 2.5 kb with 88 of them being previously unknown. We also analysed the community composition and estimated the relative mass contribution of DNA and RNA viruses to the viroplankton in each of the samples (Supplementary Table S3). The results indicate that most of the viruses were endemic, with only a few exceptions that were found in multiple oceanic regions, and that the population composition is uniform across depths and locations. We hope that our results will contribute to a better understanding of the viral ecology of the Atlantic Ocean.

EXPERIMENTAL PROCEDURES

Marine sample collection and processing

Five seawater samples from 10 to 30 L were collected from different depths and geographical points from the 20th to the 31st of October 2018 from aboard the research vessel *Ángeles Alvariño* of the Spanish Institute of Oceanography (Tel et al., 2016). The sampling sites were the following: sampling site 24: 29:09.8944 °N, 18:30.00 °W, sample depth (m): 4296.439 (24P_4296m),

sampling date: 21 October 2018; sampling site 103: 27 14.44 °N, 13:39.43 °W, samples depth (m): 5.080 (103_5m) and 26.875 (103_26m), sampling date: 26 October 2018; sampling site 602: 24:56.78 °N, 16:09.26 °W, sample at 75.451 (602_75m), sampling date: 28 October 2018; sampling site 501: 25:58.69 °N, 14:51.57 °W, sample depth (m): 3.383 (501_3m), sampling date: 29 October 2018. The chlorophyll concentration, depth, and other parameters from the sampling point were collected by an SBE 9 (Sea-Bird Scientific, Washington, USA) equipped with ECO-AFL/FL fluorescence sensor. All samples were filtered through a 0.2 µm membrane filter and then viruses present in the samples were concentrated to 20 mL using tangential flow filtration with a 30 kDa polyethersulfone Vivaflow 200 membrane (Sartorius). The concentrated samples and an eluate that will serve as a negative control in the following steps were filtered again through a 0.22 µm membrane to remove any cell remaining in the sample, which was later confirmed by SYBR Gold staining by epifluorescence microscopy. The sample 24P_4296m was filtrated twice through a 0.22 µm membrane, as it still had cells after the first filtration. These viral fractions were further concentrated to 250 µL using Amicon Ultra-0.5 mL centrifugal filter units (Merck Millipore, Massachusetts, USA).

To ensure the elimination of all external nucleic acids, each 200 µL concentrated sample was treated with 2 U of Turbo DNase I (Thermo Fisher Scientific, Massachusetts, USA) at 37°C for 1 h and with 4.22 µL of RNase for 30 min, followed by inactivation with 22.52 µL of inactivation buffer at room temperature for 5 min. Then, the viral fraction was recovered by centrifuging the samples at 9200g for 1.5 min.

For the nucleic acid extraction, the supernatant was collected in RNA-free tubes and treated with 1% proteinase K and 10% TE 10× at 65°C for 1 h with agitation and inactivated by a 5-min ice incubation. Total viral nucleic acids were extracted with MinElute Virus Spin Kit (Qiagen, Hilden, Germany) according to the manufacturer protocol with the subsequent modifications: the AL buffer was not mixed with carrier but with 21.25 µL of glycogen (1:20) and at the end, the columns were eluted three times, first with 50 µL of AVE buffer and then with 30 µL for the second and third elution. The quantity of extracted DNA and RNA was determined using a Qubit fluorometer (Thermo Fisher Scientific). In the case of the RNA, as the concentration was lower than the detection limit of the Qubit, we applied the spike-in method as detailed in Li et al. (2015) to measure it.

To eliminate the DNA of the nucleic acid extractions, samples were treated with 0.14 µL of Turbo DNase I (Thermo Fisher Scientific) 2 U/µL adding also 0.7 µL of 10% enzyme buffer and incubated at 37°C for 30 min. Then, 2 µL of inactivation buffer were added and the samples were incubated at room temperature for 5 min inverting the tube occasionally. Lastly, the

sample was centrifuged at 9200g for 1.5 min and the supernatant was transferred to a new collection tube.

Estimation of viral mass

To estimate the mass of the viral fraction that corresponds to RNA and DNA viruses (Figure S1), we inferred the number of RNA and DNA viruses from the ng obtained from each viral fraction applying the values of viral weight proposed by Miranda et al. (2016). We considered that most of the DNA and RNA extracted from the fraction that was filtrated by 0.2 μm to viruses although as discussed later on, some RNA and DNA could come from other sources, such as vesicles or recalcitrant free DNA and RNA to enzymatic digestion. The calculation of the length mean of marine RNA viruses was extracted from the IMG-VR 4 database and the mean length of marine DNA viruses was extracted from the GOV2 database (Gregory et al., 2019) to actualize the size of the virions from the data previously obtained (Miranda et al., 2016). The correlation between genome length and virion weight was performed using this new average length at it which applied the weight/length relation proposed by Miranda et al. (2016), which for RNA viruses is a length of 5.58 kb and a weight 3.12E^{-9} ng, and for a DNA virus 44.95 kb and 4.94E^{-8} ng.

To calculate the number of virions that are retained in the filtering steps, we performed a SYBR Gold staining of the DNA viruses fixating the samples at 0.1% of glutaraldehyde and calculated the number of virions before and after filtering through a 0.2 μm PES filter (ref. GSWP04700, Sigma-Aldrich, Missouri, USA) and a 0.02 μm filter membrane (ref. WHA68096002, Sigma-Aldrich). Then, we applied this corrective factor to the RNA and DNA viral density.

Sequencing and read treatment

RNA libraries and sequencing of samples were performed at the FISABIO genomics center (Valencia, Spain) using a Smart Seq Stranded Kit Ultra Low input RNA kit (Takara Bio Inc, Kusatsu, Japan) and a Next-Seq sequencer (2×250 bp, paired-end reads).

The reads were quality-filtered using Trimmomatic (v.0.36) (Bolger et al., 2014) with the following parameters: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:50. Then, each sample was decontaminated by a comparison with the negative control using BLAST, considering as a “contaminated read” any hit with an identity >95%, Qcov >70% and e-value < 1E^{-5} , which were discarded. Each metavirome was individually assembled using SPAdes (v. 3.13.0) (Nurk et al., 2013) with the options:—meta -k 21,33,55,77,99,127—only-assembler. All

assembled contigs with less than 500 bp were removed from the analyses and genomes were submitted to JGI-IMG for annotation (IMG Annotation Pipeline v.5.0.25) (Supplementary Table S2). Finally, the prediction of ORFs was done in Prodigal (v. 2.6.3) (Hyatt et al., 2010) using the option -p meta.

Bioinformatic analysis

To identify the scaffolds pertaining to RNA viruses, first several HMM for the RdRP were downloaded from Pfam and Callanan et al. (2020) and new models were created using hmmbuild (HMMER v.3.3.2) with RdRP sequences downloaded from NCBI and aligned to RdRP sequences obtained from JGI-IMG/VR. Two iterations were made to improve RdRP identification, comparing the proteins recruited with the first model to the NCBI nr/nt database of viral proteins using BLASTp (cut-off 1E^{-5}) to make sure that there were RdRP or viral hypothetical proteins. The same method was used for the CP and MP proteins (Callanan et al., 2020). To study the sample composition, the recovered viral scaffolds were compared through HMM to models with phylogenetic groups associated to them (Wolf et al., 2020).

Next, viral scaffolds were recovered from metagenome assemblies using VirSorter2 (Guo, Bolduc, et al., 2021) with the following parameters:—include-groups RNA—min-length 500—min-score 0.5 -j 4 all. The quality of the matches was studied with CheckV (Nayfach et al., 2021) using the parameters:—include-groups RNA—min-length 500—min-score 0.5 -j 16 all, and then the result was analysed again with VirSorter2 following the pipeline described by Guo, Bolduc, et al. (2021) for virus identification. Finally, the results were annotated with DRAM (Shaffer et al., 2020) with the option: min_contig_size 500. We considered viral all those scaffolds that DRAM ranked as a D, meaning that they had a hit with Pfam, and a viral identity over 50%. The annotation was complemented by other methods, such as BLASTp against the NCBI nr/nt database of viral proteins (cut-off $1\text{e}-5$) and VIBRANT v.1.2.0 (Kieft et al., 2020).

To estimate the abundance of the identified marine viruses, we performed a viromic fragment recruitment using the metatranscriptomic datasets from *Tara* Oceans Expedition (Figure 4A,B) and the *Malaspina* Expedition, which are publicly available at the JGI, besides the dataset generated in this study. Fragment recruitment analyses were carried out with BLASTp (cut-off 1E^{-5}) and then filtered by identity percentage of 95 and query coverage of 70. To calculate the abundance of each virus in a way that allows the comparison between different samples, the number of nucleotides recruited by each virus were normalized by the viral genome length and the virome length using R software as follows: pb recruited/(Kb viral length \times Gb

virome size) (pbPKG). Only the viruses with pbPKG >500 were considered true hits and were taken into consideration for further analysis. Finally, the viral scaffolds over 2.5 kb were compared against the IMG-VR 4 database (Camargo et al., 2023) and the RNA viral scaffolds described by Zayed et al. (2022) to check for their novelty using a cut-off of 95% identity and 85% query coverage.

RESULTS AND DISCUSSION

Estimation of viral mass

In this study, we analysed a total of five samples from different depths and locations from the Atlantic Ocean pertaining to different oceanic layers: 501_3m, 103_5m, 103_26m, and 602_75m from surface, and 24P_4296m (4296 m) from the bathypelagic layer (Figure 1A, B). The sampling locations have different parameters in terms of chlorophyll concentration, from 0.987 mg/m³ at the deep chlorophyll maximum around 5 m to 0.048 in the surface or 0.008 in the deep ocean, and temperature, from 22°C in the most superficial sample to 2°C at 4296 m of depth.

To study the contribution of RNA viruses to the total viral community, we estimated the number of dsDNA and RNA viruses applying an updated experimental approach previously used by Miranda et al. (2016) and Steward et al. (2013) (see Table S1 in supplementary material). This method basically estimates the number of viral copy genomes of RNA and DNA viruses considering the total DNA and RNA mass extracted from viral fractions and the average genome size and weight of an RNA and DNA viral genome. It is worth noting, as previously demonstrated in seawater (Roux et al., 2016) that most of the extracted viral DNA corresponds to dsDNA viruses since ssDNA viruses as a whole represent only a minor fraction (<5%) of DNA virus communities.

The results showed that the abundance of DNA viruses ranged from 6×10^7 to 7×10^6 VLP/mL in surface and deep ocean (4296 m depth), respectively, in good agreement with other studies showing that viral abundances decrease with depth (Lara et al., 2017). Remarkably, in all samples collected from surface and DCM, abundance of DNA viruses was within $\approx 10^7$ VLP/mL and outnumbered RNA viruses that typically showed lower abundances within the order of $\approx 10^6$ VLP/mL. Thus, data showed that the contribution of RNA viruses to the total viroplankton community in surface, and DCM varies from 5.2% to 24.4%, similar to other values proposed by previous studies (Miranda et al., 2016). Unexpectedly, the only sample in which RNA viruses dominated the viroplankton community was that of the deep ocean, in which RNA viruses could represent up to 75% of the viroplankton community.

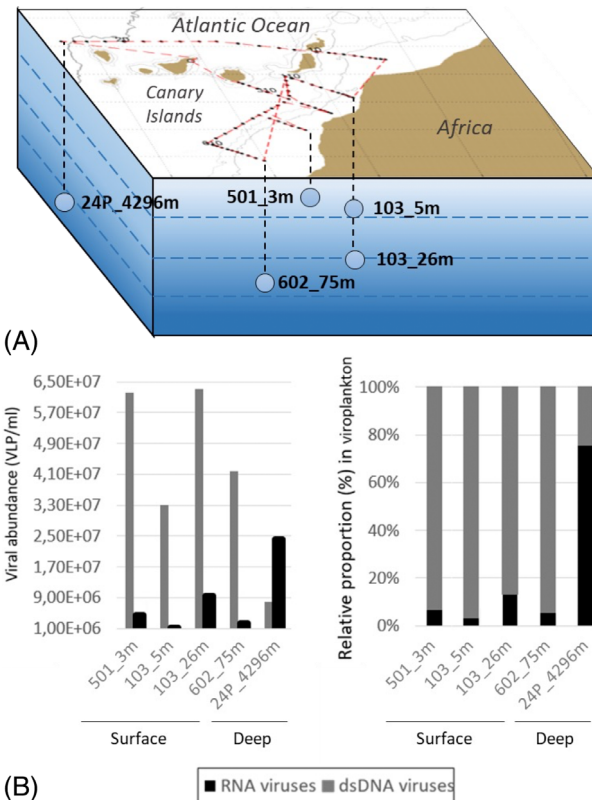


FIGURE 1 Estimation and inference of abundance of virus-like particles (VLP) in seawater. (A) Geographical location of the sampling. (B) Abundance of RNA and DNA viruses. VLP per mL of original seawater was calculated taking into consideration the mean length of DNA and RNA viruses according to the IMG-VR 4 database and the genome length/particle weight relation proposed by Miranda et al. (2016).

However, it is important to consider that the collection time of this sample due to its depth was unusual (several hours) compared to the rest of samples and this possible 'bottle effect' might impact and bias the microbial and viral community structure. In addition, this sample was subjected to an additional filtration step (see Experimental Procedures), while the rest of samples were only filtered once (0.2 µm filter; see Experimental Procedures for more details), so that could have affected mostly the retainment of DNA viruses that tend to have larger sizes and thus we cannot rule out that this might overall altered the original values and ratio of RNA:DNA viruses. In our study, for RNA viruses, we did not observe lower abundances in deeper samples as with DNA viruses.

The approach implemented in our study to estimate the number of RNA viruses, very similar to that used by Miranda et al. and Steward et al., is not exempt of biases. Here, using the viral databases from IMG-v4 (Camargo et al., 2023) and GOV 2.0 (Gregory et al., 2019), we have updated the mean genome size values for marine RNA and dsDNA viruses considering only high-quality genomes in those databases ($\geq 90\%$

completeness; see Table S1 in supplementary material). The mean genome sizes of RNA and dsDNA viruses therefore varied slightly from those of Steward et al. and Miranda et al. (see Table S1 in supplementary material). It is important to remark that according to data published by Roux et al. (2016), we assumed that most of the extracted viral DNA in our study is from dsDNA viruses since ssDNA viruses as a whole represent only a minor fraction (<5%) of DNA virus communities in seawater (Roux et al., 2016). As in Miranda et al. and Steward et al., we assumed that the bias of viral factor loss during experimental processing equally affects to both DNA and RNA viruses with no preference according to the type of virus. In our study, to polish the methodology published by Miranda et al. and Steward et al., we aimed to experimentally correct the quantitative data of RNA and DNA viruses by calculating the viral factor loss due to experimental processing for each one of the samples by SYBR Gold epifluorescence microscopy counting. For that, we performed direct counting of DNA viruses using the original seawater samples and the final concentrated and purified viral fractions that were later used for DNA extractions. For the DNA viruses, when comparing the number of viruses obtained with both methods, SYBR Gold counting and mass estimation from the extracted viral DNA (Figure S2), there is a strong correlation ($R^2 = 0.951$). Obviously, the best approach to achieve a robust and exact estimation of total RNA viruses in environmental samples would be a direct counting of RNA viruses with fluorescent RNA dyes. In previous experiments, we tested unsuccessfully different commercially available RNA dyes (SYBR Green II, SYTO RNA Select, RiboGreen, Pironyn Y and Styryl-TO) by flow cytometry and epifluorescence microscopy (data not shown). A combination of lack of sensitivity and specificity to only RNA precludes such experiments in natural viral communities. Finally, we cannot rule out that some extracted RNA (and DNA as well) might have other sources, such as extracellular vesicles that could co-purify with viruses. However, as demonstrated by Biller et al., although some vesicles contain sufficient nucleic acids to be visible and stained with SYBR fluorescent DNA dyes used to enumerate viruses, this represents only a small proportion (<0.01%–1%) of vesicles.

RNA viral scaffolds identification

In our dataset, using different detection methods (see Experimental Procedures for details), we identified a total of 2481 scaffolds over 500 pb. When comparing the number of unique scaffolds identified by each method, the one with the most unique scaffolds was the JGI annotation (2500 unique scaffolds), followed by HMM models (108) and Virsorter2 with 0 unique scaffolds. Even though the annotation by JGI was the

method that identified the larger fraction of unique scaffolds, it lacks the confidence of the HMM to say that the scaffolds really pertain to RNA viruses, as the HMM pertained to proteins that constitute RNA viral markers as the viral coat protein, the viral maturation protein, and the RNA-dependent RNA polymerase. The number of RNA viral scaffolds recovered in each sample ranged from 46 to 1088 (see supplementary material). When comparing the number of RNA viral scaffolds with several parameters, although it is worth noting that the data is strongly influenced by the sample 24P_4296m, we found an inverse correlation between the number of RNA viral scaffolds and the increase of depth ($R^2 = 0.319$). A direct correlation was observed between the number of RNA viral scaffolds and the concentration of chlorophyll ($R^2 = 0.647$). Similarly, a positive correlation was found between the number of RNA viral scaffolds and the amount of RNA at the sampling point ($R^2 = 0.984$) (see Table S1 in supplementary material). It is well known that marine viral productivity decreases with depth (Lara et al., 2017), which in turn is obviously related with primary production, mainly driven by phytoplankton. This would explain the lower number of RNA viral scaffolds found at the deep ocean sample 24P_4296m, mainly dominated by heterotrophic prokaryotes.

In general, the mean size of the scaffolds was between 577 and 1177 pb, with the sample 602_75m having both the larger scaffold and the larger mean (Figure 2A). The global annotation of the scaffolds from the different samples using DramV resulted in the identification of 31 ORFs that encoded six different genes, with the majority corresponding to the coat protein (Figure 2B). Next, we performed a deeper annotation of the viruses above 2.5 kb and we confirmed that they were RNA viruses due to the presence of RNA viral markers proteins as RdRP and CP (Figure 2C).

RdRP distribution

RdRP is likely one of the best hallmark genes for RNA viruses (Liao et al., 2022). Hence, to study the composition of the viral population of each sample (Figure 3A), we observed the distribution of different RdRP models on the samples (Figure 3C) complemented it with a principal component analysis (PCA) (Figure 3B). In our samples, we identified the presence of 11 different RdRP models, with only four of these models were present in almost all the samples (v2_KX18, v2_RdRP_1, v2_RdRP_3, and v2_MK01). The number of RdRP-positive scaffolds identified in each sample were 46 for 501_3m, 701 for 103_5m, 1088 for 103_26m, 791 for 602_75m, and 66 for 24P_4296m. Our data show no specific RdRP models associated with depth nor geographic location, except for the sample 24P_4296m that shows massive

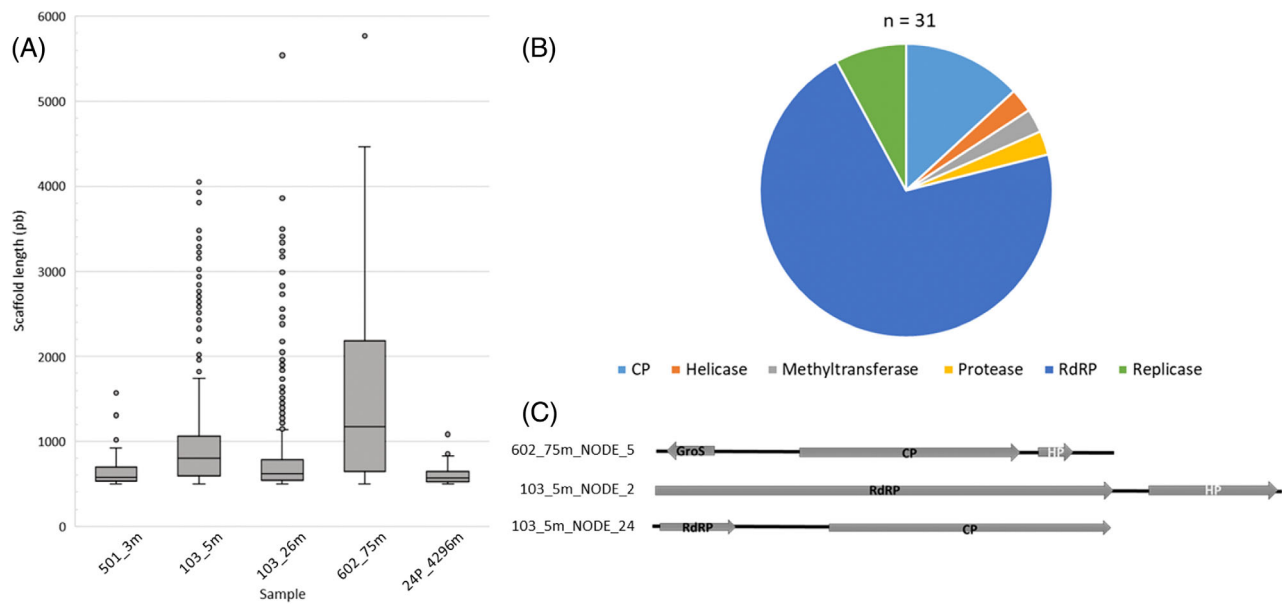


FIGURE 2 RNA viral scaffolds features and genome annotation. (A) Length distribution of the RNA genome fragments (over 500 pb) for each sample. (B) Global annotation of RNA viruses. Genome fragments were assigned to the D category of DramV and a viral identity score over 50% was applied according to (Guo, Vik, et al., 2021). (C) Example of three RNA viruses annotated in depth using different approaches: HMM for RdRP and CP, BLASTp against NCBI nr data base, Vibrant, as well as the VirSorter2-CheckV pipeline. CP, coat protein; HP, hypothetical protein; GroS, chaperonin; RdRP, RNA dependent RNA replicase.

differences with the rest of the samples as the lack of the most common model in the other samples (v2_KX18) or the great contribution of the model v2_RdRP_B to the total composition while in the rest of the samples this model was absent or minoritarian. For the other four samples, our results show the predominance of one model that contributes from 53% up to the 70% of the total composition (v2_KX18), followed by the model V2_RdRP_1 that goes from 10% to 22% and the model V2_RdRP_3, which has a lower contribution and is not present in the most superficial sample (Figure 3C). The richness of models is higher at the samples taken between 5 and 75 m of depth, and then we see a decrease in the number of different models in the deepest sample. This is consistent with the results of recent studies that support the idea that RNA viral diversity decreases with depth (Dominguez-Huerta et al., 2022). In conclusion, the sample with the most differences, both in the composition analysis and in the PCA, is the 24P_4296m followed by the 501_3m. The 24P_4296m is the sample taken at higher depth and most distant to the coastline, and the majority RdRP model matches one that is absent in all but one other sample. However, it must be taken into consideration that the two samples with the most different composition, 24P_4296m, and 501_3m, are also the two samples with the lower sample size.

Finally, we performed a new analysis using HMM from RdRP models with phylogenetic value, in which each model pertains to a specific viral group (Wolf et al., 2020). In this case, a total of 23 different models

were present in our dataset, with only 9 of them having 5 or more hits between all the samples (Figure 3E). The number of scaffolds RdRP positives identified in each sample were 4 for 501_3m sample, 61 for 103_5m, 153 for 103_26m, 56 for 602_75m and 2 for 24P_4296m. Consistent with the previous result, we see that the greater differences between the samples are in the sample 24P_4296m, which consists purely of *Fiersviridae* (formerly named *Leviviridae*), a viral family that is absent in the rest of samples except for 103_26m, in which is one of the most minoritarian. While there is a one-to-one ratio correlation between the two sets of models, it is worth noting that there are exceptions. Specifically, the HMM-RdRP model v2_RdRP_1 splits into two models from the other set, and the HMM-Wolf model *Yangshan assemblage*, sister clade to the class *Alsuviricetes*, which contains four different models from the previous set (as illustrated in Figure 3D). Overall, we can observe that the composition of the samples exhibits more differences compared to the first set of models. In this second set of models, no single model predominates clearly. However, most of the population in all the samples except the 24P_4296m is formed by the two models *Yangshan assemblage* and *Nodaviridae*. Despite these differences, we can also see a higher richness of models in the samples 103_26m and 602_75m, like what happened with the previous models. Most of these models infect protists, with the exceptions of *Fiersviridae*, that infects bacteria, and the *Ourmiaviruses*, which primarily infects herbaceous angiosperms. Furthermore, it is

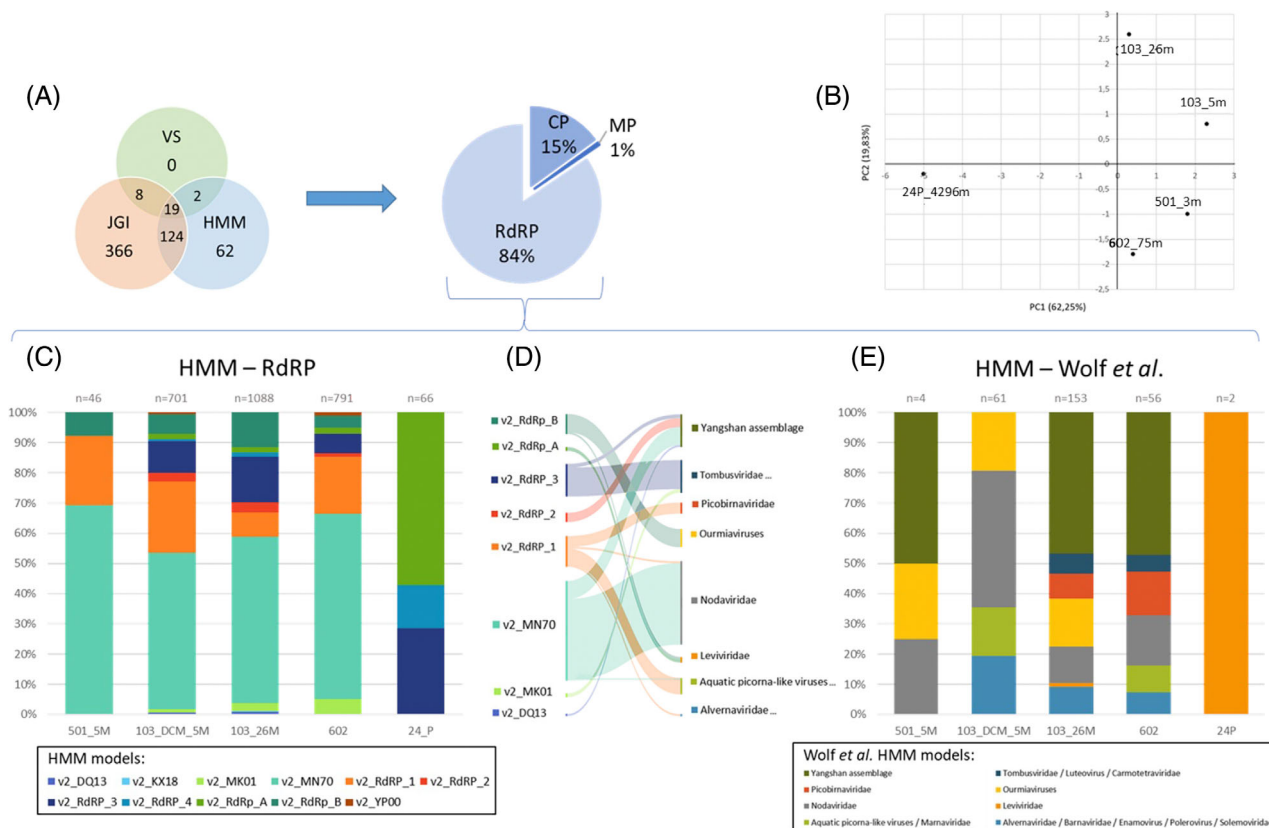


FIGURE 3 Composition of the RNA viral population for each sample. (A) Number of viral scaffolds identified by each method: VirSorter2-CheckV pipeline (VS), JGI annotation (JGI) and Hidden Markov Models (HMM), with being the RdRP model the one with most of the hits. (B) Principal component analysis of the HMM of the RdRP generated in this study. (C) Distribution of the different RdRP models both generated in this study and from the Wolf et al. (2020) representing for clarity only those models with more than five hits. (D) Sankey diagram showing the relation between the scaffolds that were assigned to both sets of RdRM models. CP, coat protein; HMM, Hidden Markov model; MP, maturation protein; RdRP, RNA-dependent RNA polymerase.

likely that some of the RNA viruses detected in our study could potentially infect phytoplankton, particularly in the surface samples. In our study, it was not possible to identify the hosts of these viruses by *in silico* approaches, as the genomic databases are mainly comprised of prokaryotes due to their predominant abundance, while the available genomic information of protists and other unicellular eukaryotes is very limited and insufficient for carrying out a proper *in silico* assignment of virus to host.

Biogeography and viral abundance

Of all the viral genomes identified, we selected those with a length over 2500 pb ($n = 107$) to study their abundance at different levels, but before that, we compared these viruses with the IMG-VR4 database and the viruses described by Zayed et al. (2022) and we got no significant hits for 88 viruses, which indicates that these viruses have not been described before (see Figure S3 in supplementary material). Of these viruses, from the 31 that had a positive hit with the HMM-Wolf

models, the majority were classified as pertaining to the *Yangshan assemblage* and the *Nodaviridae* family (see Figure S3 in supplementary material). First, we studied the endemicity of the viruses comparing the abundance of each virus in their respective sample with the abundance of that same virus in the other samples from this study. Here, we found that while some of the viruses have a lower abundance in their own sample, when we take into consideration their abundances in the different samples from our study dataset, the total abundance on these viruses is some of the highest. The opposite scenario can also be found, where some of the viruses that are highly abundant in the sample in which they were identified they are highly endemic and their abundance in other samples is practically irrelevant (see Figure S3 in supplementary material).

Next, to assess their global abundance, we analysed fragment recruitment-based abundances in metatranscriptomic datasets from the *Tara Ocean* (Salazar et al., 2019) and *Malaspina* (Duarte, 2015) expeditions.

To investigate the global distribution of viruses, we performed a fragment recruitment analysis using the *Tara Oceanic* metatranscriptome dataset and

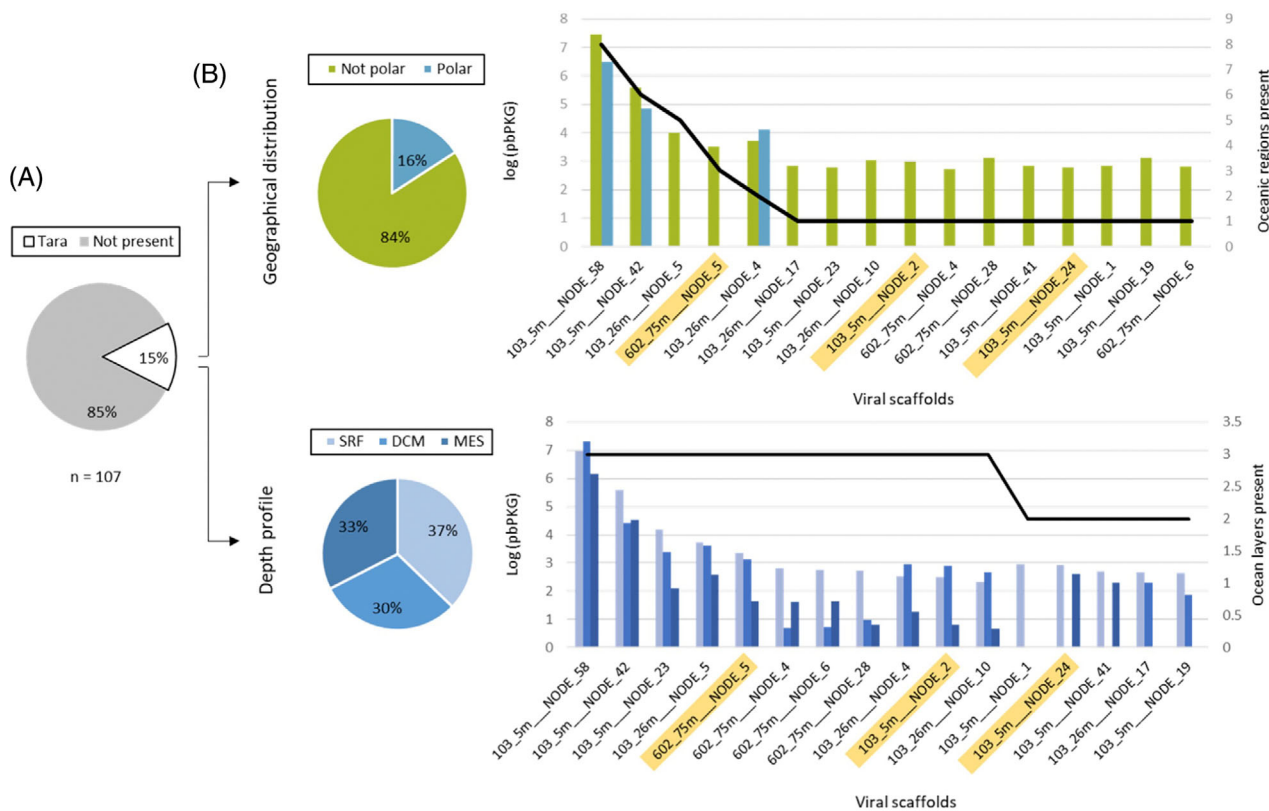


FIGURE 4 Relative abundance and distribution of marine RNA viruses in the global ocean. The abundance was calculated by fragment recruitment against the *Tara* Oceans metatranscriptome (expressed in pbPKG, with a cutoff of 500). (A) Geographical distribution of the viruses over 2.5 kb present in the *Tara* dataset indicating the number of *Tara* Oceanic regions in which they were present. (B) Depth profile indicating the number of ocean layers in which the viruses were present. Yellow are indicated those viruses with genome annotation shown in Figure 2.

determined the relative abundance of each virus in pbPKG (Supplementary Table S4). Our results indicate that most detected viruses in our study (85%) were endemic and restricted to our sampling region, with only 15% being present in other *Tara* samples. Among the viruses that were detected in the *Tara* datasets, they were mostly found in non-polar regions, with most viruses being restricted to a single oceanic region. Only three viruses were present in more than half of the oceanic regions. When we focused on the polar regions, we found no clear patterns of viral abundance, as two of the three viruses were widely distributed while the third was only present in a non-polar oceanic region.

We also examined the depth profile of the viruses in the oceans using the same *Tara* Oceanic dataset. Our analysis revealed that most of the viruses were present in all three major oceanic layers (surface [SRF], deep chlorophyll maximum [DCM], and mesopelagic zone [MES]). Notably, all the viruses identified were present in the SRF layer, but while most of the virus are more abundant at the SRF level, we observed some variations in the distribution patterns, as some viruses being more abundant in the DCM layer. There is also worth notice that some of the viruses are present in the DCM

layer and not present and the MES and vice versa. It is important to remark that so far, the size of publicly available database of marine RNA viruses is certainly more limited in comparison to that of DNA viruses, which have been traditionally more studied. Thus, we cannot rule out that some of the RNA viruses discovered here in the Atlantic Ocean could be also present in other oceanic regions yet to be more explored.

Lastly, we studied the abundance of the newly found viruses in the metatranscriptome from the Malaspina Expedition. While some of the viruses present in Malaspina can also be found in *Tara* sites, there are some viruses that despite being highly globally distributed in the Malaspina sites cannot be found in Malaspina and vice versa. The samples from the Malaspina Expedition all pertain to the deep-sea level, which corresponds to more than 4000 m of depth, which could explain these differences. It is worth noting that most of the viruses present in Malaspina dataset pertain to the sample 103_5m collected at 4.75 m, with the 91.7% of the viruses over 2.5 kb present in Malaspina pertaining to this sample compared to the 54.2% that represents for the total of viruses over 2.5 kb (see Table S5 in supplementary material).

CONCLUSIONS

This study accomplished the identification of a total of 2481 RNA viral scaffolds combining different approaches, such as RdRp viral hallmark gene. In total, 107 larger bona fide RNA viral genomes (>2.5 kb) were identified; 88 of them representing novel viruses not described before. The composition analysis of the populations of the samples revealed no specific RdRP model associated with depth or geographic location. The study also found that the sample with the most differences in composition was the 24P_4296m followed by the 501_3m. Biogeography and genomic data comparison showed that most of these 88 new viruses were endemic to the sampling region with only 15% being present in the *Tara* metatranscriptome datasets in only one oceanic region, and mostly in not polar regions, which highlight the high diversity of RNA viruses yet to be discovered. Furthermore, there were different patterns of distributions, as we found some viruses that were present in several ocean regions described at the *Tara* database. When we studied the abundance of the newly found viruses in the metatranscriptome from the Malaspina Expedition, we found that most of the viruses present in this dataset from deep-sea level were recovered from the sample collected at 5 m depth, and were not present in the deeper samples. Lastly, data abundance of RNA viruses in surface and DCM samples were within the order of 10^6 VLP/ml, with being DNA viruses the dominant viral members of the community. In contrast to DNA viruses, an inverse correlation of RNA viral abundance with depth was not clearly observed.

AUTHOR CONTRIBUTIONS

Marina Vila-Nistal: Data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (lead); validation (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). **Lucia Maestre-Carballea:** Data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); software (supporting); validation (supporting); visualization (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Francisco Martinez-Hernandez:** Data curation (supporting); formal analysis (supporting); investigation (supporting); software (supporting). **Manuel Martinez-Garcia:** Conceptualization (lead); data curation (supporting); formal analysis (supporting); funding acquisition (lead); investigation (supporting); methodology (supporting); project administration (lead); resources (lead); software (supporting); supervision (lead); validation (equal); visualization (supporting); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGEMENTS

We thank to the Spanish Institute of Oceanography for the organization of the Raprocan expedition and to Jesús Arrieta for the opportunity to be part of it. This work was supported by the *Generalitat Valenciana* ACIF2020 grant and by the research grants funded by Spanish Ministry of Science and Innovation (refs. RTI2018-094248-B-I00 and PID2021-125175OB-I00), and by the Gordon and Moore Foundation (ref. 5334).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated and analysed during the current study are available in the CyVerse Discovery Environment repository (https://de.cyverse.org/data/ds/iplant/home/mvn111/supplementary_material?type=folder&resourceId=fefe573a-ed9d-11ed-a7eb-90e2ba675364) and the JGI Genome portal with the GOLD Study ID Gs0154411.

ORCID

Marina Vila-Nistal  <https://orcid.org/0000-0002-5635-5607>

Manuel Martinez-Garcia  <https://orcid.org/0000-0001-5056-1525>

REFERENCES

- Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.
- Callanan, J., Stockdale, S.R., Shkoporov, A., Draper, L.A., Ross, R.P. & Hill, C. (2020) Expansion of known ssRNA phage genomes: from tens to over a thousand. *Science Advances*, 6, eaay5981.
- Camargo, A.P., Nayfach, S., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K. et al. (2023) IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research*, 51, D733–D743.
- Culley, A.I., Lang, A.S. & Suttle, C.A. (2006) Metagenomic analysis of coastal RNA virus communities. *Science*, 312, 1795–1798.
- Culley, A.I., Mueller, J.A., Belcaid, M., Wood-Charlson, E.M., Poisson, G. & Steward, G.F. (2014) The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *MBio*, 5, e01210–e01214.
- Culley, A.I. & Steward, G.F. (2007) New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Applied and Environmental Microbiology*, 73, 5937–5944.
- Dominguez-Huerta, G., Wainaina, J.M., Zayed, A.A., Culley, A.I., Kuhn, J.H. & Sullivan, M.B. (2023) The RNA virosphere: how big and diverse is it? *Environmental Microbiology*, 25, 209–215.
- Dominguez-Huerta, G., Zayed, A.A., Wainaina, J.M., Guo, J., Tian, F., Pratama, A.A. et al. (2022) Diversity and ecological footprint of Global Ocean RNA viruses. *Science*, 376, 1202–1208.
- Duarte, C.M. (2015) Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnology and Oceanography Bulletin*, 24, 11–14.

- Forterre, P. (2013) The virocell concept and environmental microbiology. *The ISME Journal*, 7, 233–236.
- Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature*, 399, 541–548.
- Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A. et al. (2019) Marine DNA viral macro- and microdiversity from pole to pole. *Cell*, 177, 1109.e14–1123.e14.
- Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O. et al. (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, 9, 37.
- Guo, J., Vik, D., Pratama, A.A., Roux, S. & Sullivan, M. (2021) Viral sequence identification SOP with VirSorter2 v3. <https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-5qpvoqyqebg4o/v3>
- Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 119.
- Kieft, K., Zhou, Z. & Anantharaman, K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8, 90.
- Lang, A.S., Rise, M.L., Culley, A.I. & Steward, G.F. (2009) RNA viruses in the sea. *FEMS Microbiology Reviews*, 33, 295–323.
- Lara, E., Vaqué, D., Sà, E.L., Boras, J.A., Gomes, A., Borrull, E. et al. (2017) Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Science Advances*, 3, e1602565.
- Li, X., Ben-Dov, I.Z., Mauro, M. & Williams, Z. (2015) Lowering the quantification limit of the QubitTM RNA HS assay using RNA spike-in. *BMC Molecular Biology*, 16, 9.
- Liang, G. & Bushman, F.D. (2021) The human virome: assembly, composition and host interactions. *Nature Reviews. Microbiology*, 19, 514–527.
- Liao, M., Xie, Y., Shi, M. & Cui, J. (2022) Over two decades of research on the marine RNA virosphere. *iMeta*, 1, e59.
- Maranger, R., Bird, D.F. & Juniper, S.K. (1994) Viral and bacterial dynamics in Arctic Sea ice during the spring algal bloom near resolute, N.W.T., Canada. *Mar Ecol Prog Ser*, 111, 121–127.
- Miranda, J.A., Culley, A.I., Schvarcz, C.R. & Steward, G.F. (2016) RNA viruses as major contributors to Antarctic virioplankton: RNA viruses in the Antarctic. *Environmental Microbiology*, 18, 3714–3727.
- Nayfach, S., Camargo, A.P., Schulz, F., Eloë-Fadrosh, E., Roux, S. & Kyrpides, N.C. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, 39, 578–585.
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A. et al. (2013) Assembling genomes and mini-metagenomes from highly chimeric reads. In: Deng, M., Jiang, R., Sun, F. & Zhang, X. (Eds.) *Research in computational molecular biology*. Berlin, Heidelberg: Springer Berlin Heidelberg: Lecture Notes in Computer Science, pp. 158–170.
- Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B. et al. (2016) Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ*, 4, e2777.
- Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M. et al. (2019) Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179, 1068.e21–1083.e21.
- Shaffer, M., Borton, M.A., McGivern, B.B., Zayed, A.A., La Rosa, S.L., Solden, L.M. et al. (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, 48, 8883–8900.
- Steward, G.F., Culley, A.I., Mueller, J.A., Wood-Charlson, E.M., Belcaid, M. & Poisson, G. (2013) Are we missing half of the viruses in the ocean? *The ISME Journal*, 7, 672–679.
- Suttle, C.A. (2007) Marine viruses—major players in the global ecosystem. *Nature Reviews. Microbiology*, 5, 801–812.
- Tel, E., Balbin, R., Cabanas, J.-M., Garcia, M.-J., Garcia-Martinez, M.C., Gonzalez-Pola, C. et al. (2016) IEOOS: the Spanish Institute of Oceanography Observing System. *Ocean Science*, 12, 345–353.
- Thingstad, T.F., Heldal, M., Bratbak, G. & Dundas, I. (1993) Are viruses important partners in pelagic food webs? *Trends in Ecology & Evolution*, 8, 209–213.
- Wolf, Y.I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D. et al. (2020) Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology*, 5, 1262–1270.
- Wommack, K.E., Hill, R.T., Kessel, M., Russek-Cohen, E. & Colwell, R.R. (1992) Distribution of viruses in the Chesapeake Bay. *Applied and Environmental Microbiology*, 58, 2965–2970.
- Zayed, A.A., Wainaina, J.M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M. et al. (2022) Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science*, 376, 156–162.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Vila-Nistal, M., Maestre-Carballe, L., Martínez-Hernández, F. & Martínez-García, M. (2023) Novel RNA viruses from the Atlantic Ocean: Ecogenomics, biogeography, and total virioplankton mass contribution from surface to the deep ocean. *Environmental Microbiology*, 25(12), 3151–3160. Available from: <https://doi.org/10.1111/1462-2920.16502>