Environmental Microbiome

RESEARCH

Open Access

The hidden genetic reservoir: structural variants as drivers of marine microbial and viral microdiversity

Jose M. Haro-Moreno[1], Juan J. Roda-Garcia[1], Carmen Molina-Pardines[1] and Mario López-Pérez[1*]

Abstract

Background Intraspecific genetic diversity is fundamental to understanding microbial adaptation, evolution, and contributions to ecosystem stability. However, traditional short-read metagenomics often underrepresents this diversity, particularly structural variants (SVs), due to assembly limitations in complex natural populations. To overcome these constraints, we employed third-generation (long-read) metagenomics to investigate the eco-evolutionary role of SVs in microbial and viral marine populations. Our analysis focused on the cellular metagenome fraction (0.22–5 µm size range) across distinct ecological niches within the photic zone of the marine water column.

Results Insertions and deletions emerged as the predominant SVs in the marine microbiome, occurring at similar frequencies across genomes. These SVs were not only found within the core genome but also in the flexible genome, serving as a source of genetic variability within genomic islands. Insertions were significantly larger, reaching more than 2 Kb, in streamlined microbes such as *Pelagibacter* (SAR11 clade) or the archaeon *Nitrosopumilus*. In contrast, SVs in viral populations were smaller and more uniform in size (~ 430 bp). Functionally, SVs were enriched in genes linked to nutrient uptake, amino acid metabolism, and regulatory networks due to the presence of non-coding RNAs. These SVs often encompassed entire genes or operons, acting as an important reservoir of niche-specific diversity that supports the emergence of ecological lineages better adapted to environmental gradients, such as rhodopsin-containing subpopulations in shallower waters. In viruses, SV-driven genetic plasticity facilitated host range adaptation and the evolution of mechanisms modulating host metabolism. We identified long-term genetically stable populations of cyanophages and pelagiphages, wherein SVs represented the primary source of genomic diversification. Notably, certain subpopulations of pelagimyophages carry SVs encoding a *pstS* gene, which enhances host phosphate uptake and increases viral replication efficiency—a beneficial adaptation in phosphate-depleted environments such as the oligotrophic Mediterranean Sea.

Conclusions By capturing SVs directly from natural populations, this study provides new insights into microbial evolution, phage-host interactions, and the broader implications of genomic plasticity for ecosystem resilience in marine environments. Furthermore, these results highlight the transformative potential of third-generation sequencing to unveil previously hidden layers of microbial and viral diversity.

*Correspondence:
Mario López-Pérez
mario.lopezp@umh.es

Full list of author information is available at the end of the article

Keywords Long-read sequencing, Structural variants, Bacteriophage, PacBio CCS long-reads, Metagenome-assembled genomes, Microdiversity, Third-generation metagenomics, Marine virome

Background

The photic zone of the water column represents one of the most ecologically significant ecosystems on Earth. Also known as the epipelagic zone, this region constitutes the sunlit upper layer of the ocean, extending from the surface to approximately 200 m deep. The microbial communities that inhabit this region play a key role in maintaining global ecological balance. Through photosynthesis, they sustain primary productivity, drive key biogeochemical cycles, and are essential for the preservation of marine biodiversity [1]. At the same time, bacteriophages (viruses that infect bacteria, hereinafter referred to as "phages"), which outnumber any other biological component in the ocean (10^7–10^{10} virions/ml), play a fundamental role as mediators of microbial diversity and dynamics [2]. Through lytic and lysogenic infection, phages not only regulate microbial population abundance but also drive processes such as the release of dissolved organic matter through the "viral shunt", recycling nutrients like nitrogen and phosphorus to sustain microbial productivity and influencing carbon sequestration via the biological pump [3].

However, the photic water column is not a homogeneous entity; it exhibits strong stratification due to steep gradients of physicochemical parameters such as temperature, light availability, salinity, and nutrient limitation [4, 5]. This stratification has profound implications for the distribution, composition, and metabolic potential of the marine microbiome, shaping its role in biogeochemical cycles [6]. In recent years, culture-independent approaches such as metagenomics and single-cell genomics have significantly advanced our understanding of the diversity, structure and seasonal dynamics of microbial and viral communities inhabiting the water column [4, 6–9]. Despite the advances brought by second-generation sequencing technologies, these methods have technical limitations, particularly in resolving microdiversity and reconstructing complete genomes due to their reliance on short reads. The presence of repetitive regions, high microbial diversity, and low sequencing coverage often results in fragmented genomes, limiting the resolution of metagenomic analyses [10, 11]. The advent of long-read sequencing has provided a powerful solution to these challenges. When applied to microbial and viral communities, long-read sequencing produces a greater number of contigs with larger average sizes compared to assemblies generated using short-read technologies such as Illumina [11, 12]. This advancement has led to substantial enhancements in the recovery of complete or near-complete genomes assembled from metagenomes, including the flexible genome, as well as improvements in the host assignment of viral sequences [11, 12].

Beyond achieving genome completeness, one of the key advantages of long-read sequencing is its ability to detect structural variations (SVs), including insertions, deletions, duplications, and rearrangements, within complex microbial communities [13], revealing genomic features that often remain unresolved in conventional short-read assemblies. While traditional genomic islands, typically spanning more than 10 Kb, have been extensively studied in marine microbiome research [14–17], fine-scale genomic variations (> 50 bp) and their role in microbial genome plasticity remain largely unexplored. A comparative genomics analysis of the marine copiotroph bacterium *Alteromonas* revealed that 50% of its flexible genome is concentrated in small regions containing one or a few genes, typically forming a single operon and distributed as small synteny breaks across the core genome [18]. Similarly, studies using long-read sequencing in SAR11 have highlighted that its flexible genome is predominantly concentrated in small regions containing single genes, underscoring the significance of small-scale genomic variations as a critical source of microbial genomic diversity [19]. Although the biological roles of these SVs have been extensively explored in eukaryotes, particularly in the context of human and crop diseases [20–22], their contributions to prokaryotic adaptation and evolution remain comparatively underexplored. Only a few studies have applied this approach using long-read sequencing, revealing an enrichment of deletions in genes associated with restriction-modification systems and transporters in freshwater pelagic bacterioplankton [23], as well as SVs linked to recombination, DNA methylation, and antibiotic resistance in the human gut phageome [24]. Similarly, in the human gut microbiome, it has been suggested that detected SVs may modulate the functionality of bacteria involved in host metabolism and thus could potentially be linked to human health [25].

Despite these findings, the eco-evolutionary impact of SVs on natural populations within marine microbiomes remains poorly understood. Addressing this gap is crucial for understanding how SVs contribute to the resilience, functional diversity, and evolutionary dynamics of microbial and viral communities in marine ecosystems. For this reason, in this study, we applied long-read metagenomics to investigate fine-scale genetic variations, i.e. SVs, across natural populations inhabiting the photic zone of the marine water column. To overcome the high DNA input requirements of long-read sequencing, we analyzed viral DNA from the cellular metagenome

fraction (0.22–5 μm), which includes naturally amplified viral genomes, lysogenic viruses, and cell-associated virions. Our findings underscore the significance of SVs in shaping the genomic plasticity of both prokaryotes and viruses, as well as their role in intra-species diversity, facilitating the adaptation of natural populations to their environment or host contexts. This study provides key insights into the importance of SVs in microbial and viral evolution, paving the way for future research on the genomic mechanisms underpinning microbiome functionality and resilience in marine ecosystems.

Results

Recovery of prokaryotic genomes

To capture the greatest possible bacterioplankton diversity in the water column, four samples were collected from a single offshore site in the Western Mediterranean Sea. Three samples were taken from different depths of the epipelagic and stratified (summer) water column, corresponding to the upper photic layer (UP; Med-OCT2021-15 m), the deep chlorophyll maximum (DCM; Med-SEP2022-60 m), and the lower photic layer (LP; Med-OCT2021-75 m) [26]. An additional sample was collected during winter, in the mixing period (MIX; MedWinter-JAN2019) [11] (Table S1).

Metagenomic hybrid assembly and binning (see Methods) yielded a total of 380 bacterial and 38 archaeal MAGs (≥50% completeness and ≤5% contamination) (Fig. 1 and Table S2). Of these, 54.2% were classified as novel species based on the Genome Taxonomy Database (GTDB) classification [27], and 53.7% contained at least one copy of the 16S rRNA gene (Table S2). A total of 58 MAGs met the criteria for high-quality draft genomes [10]. Notably, ten of these high-quality MAGs were assembled onto a single circular chromosome, including groups with no cultivated representatives such as two genera (MedAcidi-G3 and MedAcidi-G2A) within the order Acidimicrobiales (Actinobacteriota) [28], an archaeon from the family *Thalassarchaeaceae* as well as one genome from the SAR86 clade (family TMED112) (Fig. 1 and Table S2). In addition, a genome from the SAR11 clade (Pelagibacterales) was also recovered on a single circular chromosome, a group often underrepresented in MAGs due to its high microdiversity, despite its high abundance [14].

Structural variants as an important source of genomic variation

One significant advantage of long-read sequencing is its ability to accurately identify SVs, which are often missed or difficult to detect with short-read technologies. In this study, we identified population-level SVs by mapping long reads to MAGs and detecting variants using Sniffles2 [29]. These SVs reflect strain-level variation within microbial populations, as the input data consists of reads from complex metagenomic samples rather than clonal isolates. Thus, the detected SVs represent subpopulation-level genomic diversity relative to the assembled MAGs. A total of 46 MAGs were selected for SV detection based on their high completeness, low fragmentation (fewer contigs), and high read recruitment in at least one metagenomic sample (Table S2 – S3). Among the four types of SVs detected, insertions (INS; 52.1% of the total SVs) and deletions (DEL; 46.2%) were much more abundant than duplications and inversions (Fig. 2A). A strong and statistically significant correlation was observed between the number of INS and DEL per genome (Spearman $r = 0.89$; p-value < 0.05). The size distribution of INS and DEL showed a bimodal pattern, with a mean INS size of 491 bp, which was higher than that observed for DEL (208 bp) (Fig. 2B). SVs larger than 1 Kb comprised 39.7% of INS and 26.8% in the case of DEL. To explore potential links between SVs and other genomic features, we normalized SV counts by genome size (SV/Mb) and compared them with various genomic and microdiversity parameters derived from short-read mapping (Table S4). Among these, only the number of single nucleotide variants per genome size (SNV/Mb) showed a moderate correlation with SV density ($r = 0.66$; p-value < 0.05). To assess the effectiveness of long-read metagenomics in resolving intra-species diversity, we analyzed the presence of these SVs in short-read Illumina assemblies from the same environmental samples. The results revealed low recovery rates in short-read datasets. In the UP sample, which exhibited the highest recovery, only 37% of the SVs could be retrieved. In the DCM and LP samples, this percentage dropped below 30%, while in the winter sample, recovery was as low as 15%.

A total of 1,236 DEL were identified among the selected genomes, of which 32.3% contained protein-coding sequences, resulting in the prediction of 738 proteins. Functional annotation of these proteins was performed using the KEGG Orthology database [30]. In the categories for which function could be determined, "transporters" represented the most prevalent category (27%), including multiple ABC transporter operons for branched-chain and polar amino acids (within SAR324, *Thioglobus*, Rhodobacteraceae, and *Synechococcus* MAGs), ferric iron (*Prochlorococcus*, *Synechococcus*, SAR324, Acidimicrobiales, SAR86, and HTCC2207), phosphonate (*Prochlorococcus*), putrescine/spermidine (SAR324), urea (*Pelagibacter*), ribose (*Thioglobus*), glycerol-3-phosphate (SAR324), polyol (*Thioglobus* and Rhodobacteraceae), and multiple sugars (*Pelagibacter* and SAR324). Additional predicted transporter systems included tripartite ATP-independent (TRAP) (*Pelagibacter* and SAR324) and TonB (SAR86 and Flavobacteriales) transporters (Fig. 2C). Other functional categories

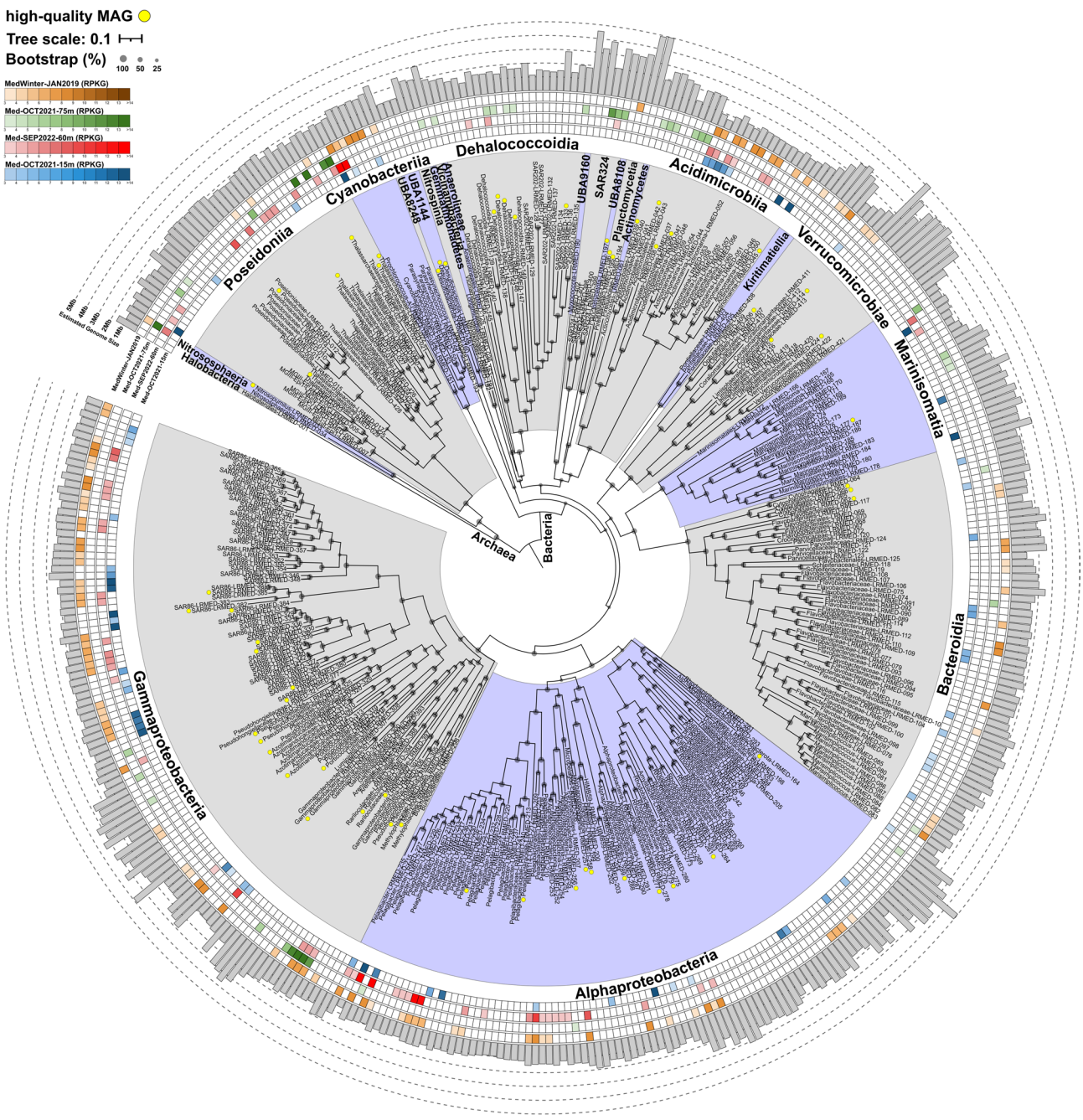

Fig. 1 Maximum likelihood phylogenomic tree of the 418 prokaryotic MAGs from this study. The sequences were grouped at the class level (grey or blue squares). Yellow dots show high-quality MAGs. Bootstrap values are indicated as grey circles on the nodes. The outer circles represent the abundance (measured in RPKG) in the metagenomic samples [Med-OCT2021-15 m (UP), blue; Med-SEP2022-60 m (DCM), red; Med-OCT2021-75 m (LP), green and MedWinter-JAN2019 (MIX), brown)]. The bars on the outside represent the estimated genome size of the MAGs, as calculated according to the percentage completeness obtained using CheckM2 (see Table S2). High-quality MAGs are defined as genomes with > 90% completeness, < 5% contamination, at least 18 out of the 20 tRNA genes encoding the standard amino acids, and the presence of at least one complete rRNA operon (16S, 23S, 5S) according to the MIMAG guidelines

represented were "genetic information processing" (9%), which encompasses fundamental processes such as DNA replication, transcription, and repair, "amino acid metabolism" (8%) and "lipopolysaccharide biosynthesis" (3%), among others (Fig. 2C). Remarkably, 16% of the SV-associated sequences could not be assigned to any known functional category, representing the second most abundant group. Among the DEL sequences, 13 were found to contain functional RNAs, including 15 non-coding RNAs (ncRNAs) and three transfer RNAs (tRNAs).

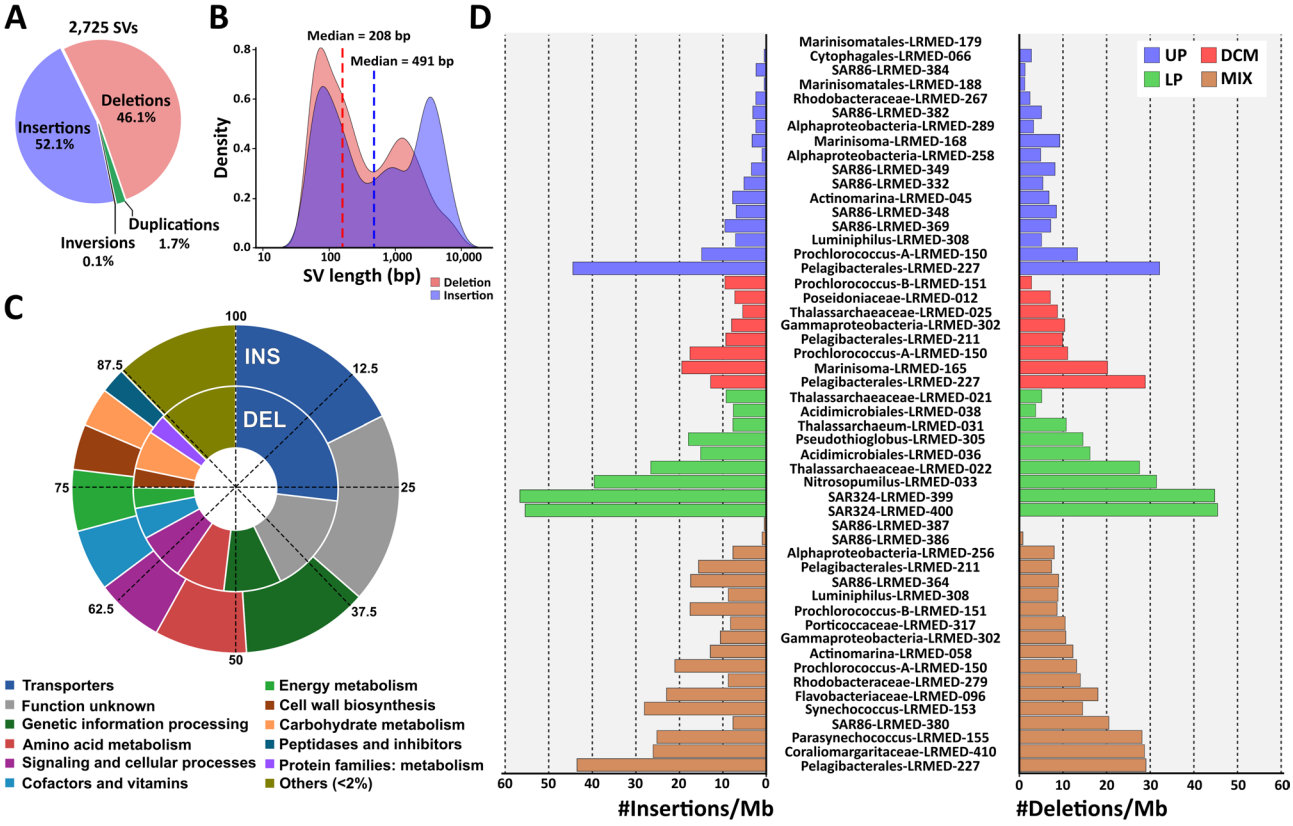

Fig. 2 Characterization of SVs in the marine microbiome. **A**. Proportion of each type of SV found in the MAGs. **B**. Length distribution of deletions (red) and insertions (blue) with the median length of each type of SV shown by the dotted line. **C**. Multilevel donut plot with functional classification for insertions (outer circle) and deletions (inner circle) found in all metagenomes. **D**. Number of insertions (right) and deletions (left) normalized per Mb of genome for MAGs distributed by metagenome. The colour of the bars indicates the metagenome to which they correspond [Med-OCT2021-15 m (UP), blue; Med-SEP2022-60 m (DCM), red; Med-OCT2021-75 m (LP), green and MedWinter-JAN2019 (MIX), brown)]. DEL: deletion, INS: insertion

Within the SAR86 clade, four DEL containing ncRNAs were detected in the UP and MIX samples. These ncRNAs were associated with nitrogen metabolism (glutamine riboswitch RF01739) and cobalamin biosynthesis (RF00174) (Table S5). Additionally, a glutamine-II riboswitch (RF01704), which regulates the synthesis of the glutamine synthetase regulatory protein [31], was identified within a DEL in Synechococcus-LRMED-153 (Table S5). Also related to nitrogen metabolism, a deletion of a ncRNA involved in regulating the urease gene cluster (RF02514) was identified in a DEL in the MIX sample within the MAG Pelagibacterales-LRMED-211. Moreover, identical ncRNAs were consistently detected within DEL in Pelagibacter-LRMED-227 and Prochlorococcus-A-LRMED-150 (high-light ecotype) across all three samples, localized to the same genomic region (Table S5). In Pelagibacter-LRMED-227, this ncRNA was linked to the regulation of glycine degradation (RF00504). In contrast, the redundant ncRNAs in Prochlorococcus-A-LRMED-150 exhibited no clearly defined function: one was associated with general cellular process regulation (RF02362), while the other was linked to stress response mechanisms (RF01701) (Table S5).

Compared to DEL, INS ($n = 1{,}395$) contained a higher proportion of protein-coding sequences (45.3%). While the overall functional annotation patterns were similar, INS showed a higher representation in the categories of "genetic information processing" (13%), "amino acid metabolism" (9%), "cofactor and vitamin metabolism" (6%) and "energy metabolism" (6%) (Fig. 2C). In contrast, INS showed a lower proportion in the "transporter" category (18%), with a functional classification similar to that observed in DEL. These included ABC, TRAP and TonB transporters involved in the acquisition of amino acids, sugars, metals (iron, nickel), osmolytes, and phosphonate. These transporters were predominantly found in *Pelagibacter*, SAR324, *Prochlorococcus* and *Synechococcus* MAGs. The percentage of sequences without functional annotation was 19%.

We identified 34 functional RNA types within INS, including 12 ncRNAs and 26 tRNAs. Among these, two INS, found in the MAGs Alphaproteobacteria-LRMED-289 (RF01849) and Flavobacteriaceae-LRMED-096 (RF03141), contained ncRNAs associated with ribosome function regulation (Table S5). The glycine riboswitch (RF00504), previously detected in a DEL

within Pelagibacter-LRMED-227, was also identified as an INS elsewhere in the genome. However, this INS occurred in a subpopulation associated with the winter metagenome (Table S5). The remaining INS containing ncRNAs were found in two *Prochlorococcus* MAGs. One ncRNA, Cyano-1 (RF01701), associated with stress response was present across multiple metagenomes in both *Prochlorococcus* genomes. In addition, two ncRNAs involved in ribosome function (RF01851) and the regulation of the photosynthetic protein isiA (RF01419) were exclusively detected in SVs within Prochlorococcus-A-LRMED-150 (Table S5). Although no prophages or mobile genetic elements were associated with either insertions or deletions, we identified 19 genes related to insertion sequence (IS) elements, primarily in populations linked to the LP and MIX samples. This suggests that transposition activity, potentially driven by IS elements, may contribute to SV formation in specific environmental contexts.

Genomic diversity of structural variants

Figure 2D displays the number of INS and DEL normalized per megabase (SV/Mb) across all genomes in each metagenome. The highest SV/Mb values were observed in two genomes of the SAR324 class (approximately 120 SV/Mb) within the lower photic zone metagenome. Despite both MAGs exceeding 3 Mb in size, no significant correlation was found between SV/Mb and estimated genome size within the dataset ($r = 0.15$; p-value > 0.05) (Table S4). Indeed, in the other metagenomes (UP, DCM, and MIX), the highest SV/Mb was detected in Pelagibacter-LRMED-227, followed by Nitrosopumilus-LRMED-033 (75 SV/Mb in LP). Both genomes have an estimated size of 1.3 Mb (Fig. 2D and Table S6). A notable disparity was observed in the values recorded across the diverse phyla and environmental contexts. For instance, within the SAR86 clade, the TMED112 family showed a range of one to ten SV/Mb, while the MAGs of the D2472 family in the same samples (UP and MIX) had a range of 15 to 31 SV/Mb (Fig. 2D and Table S6).

Following the analysis of the number of SVs, we evaluated the size distribution of both INS and DEL events. In Pelagibacter-LRMED-227, which was detected in three metagenomes, INS consistently exhibited a larger mean size than DEL. In the UP sample, INS showed a bimodal distribution with a mean size of 586 bp (Fig. 3A). In contrast, in both DCM and MIX samples, the INS size distribution showed a peak, with a mean size exceeding 2.5 Kb (Fig. 3B and C, respectively). The DEL size distribution across all three metagenomes followed a unimodal pattern, with a peak around 350 bp (Fig. 3A, B and C). In Prochlorococcus-A-LRMED-150, another abundant MAG in these metagenomes (UP, DCM and MIX), the DEL size distribution mirrored that of *Pelagibacter*, but with a smaller peak size of around 150 bp (Figure S1A, S1B and S1C). The INS sizes in *Prochlorococcus* were also smaller, averaging around 550 bp in the UP and MIX samples, and 723 bp in the DCM (Figure S1A, S1B and S1C).

In the remaining MAGs, we observed diverse size distribution patterns. For instance, in Coraliomargaritaceae-LRMED-410 and Parasynechococcus-LRMED-155, both from the winter sample, DEL had a larger mean size than INS, which is an atypical pattern not observed in other MAGs (Figure S2). In the lower photic zone sample, Nitrosopumilus-LRMED-033 exhibited a pattern like Pelagibacter-LRMED-227, with an INS peak near 3 Kb (Figure S2). In contrast, SAR324-LRMED-399, which had the highest number of SV/Mb, displayed both INS and DEL with mean sizes below 200 bp. In the DCM sample, Marinisoma-LRMED-165 showed a bimodal distribution for both SV categories, with mean sizes of 413 bp for DEL and 615 bp for INS (Figure S2). Meanwhile, in another MAG of the *Pelagibacter* genus, Pelagibacter-LRMED-211, which was exclusively found in the DCM, both DEL and INS exhibited a prominent peak around 1 Kb (Figure S2).

It is important to note that the classification of SVs as INS or DEL in metagenomic datasets depends on the nature of the reference assembly used. Since MAGs are used as reference frameworks, the presence or absence of a given sequence is relative: the same genomic region may be classified as either an INS or a DEL, depending on whether it is included in the reference assembly. This methodological limitation affects the interpretation of comparisons of the relative abundance between INS and DEL, although it does not invalidate their biological relevance as indicators of genomic microdiversity.

Impact of structural variants on the intraspecific metabolic potential in the marine microbiome

The widespread distribution and high abundance of two ecologically and evolutionarily distinct microbes, Pelagibacter-LRMED-227 and Prochlorococcus-A-LRMED-150, provided an opportunity to assess the role and impact of SVs in shaping subpopulation-level genetic diversity across distinct ecological niches represented by different metagenomic samples. A comparison of SV distribution across different ecological niches (UP, DCM, and MIX samples) revealed that most SVs were metagenome-specific, with only a small fraction shared between datasets, as these SVs were predicted in different metagenomes but at the same genomic position and coded for the same gene content. In Pelagibacter-LRMED-227, only 15 SVs (7.4% of the total SVs) were found among all three samples, while 24 SVs were shared exclusively between the DCM and MIX metagenomes (Figure S3). Conversely, 159 SVs (78.7%) were found exclusively in one of

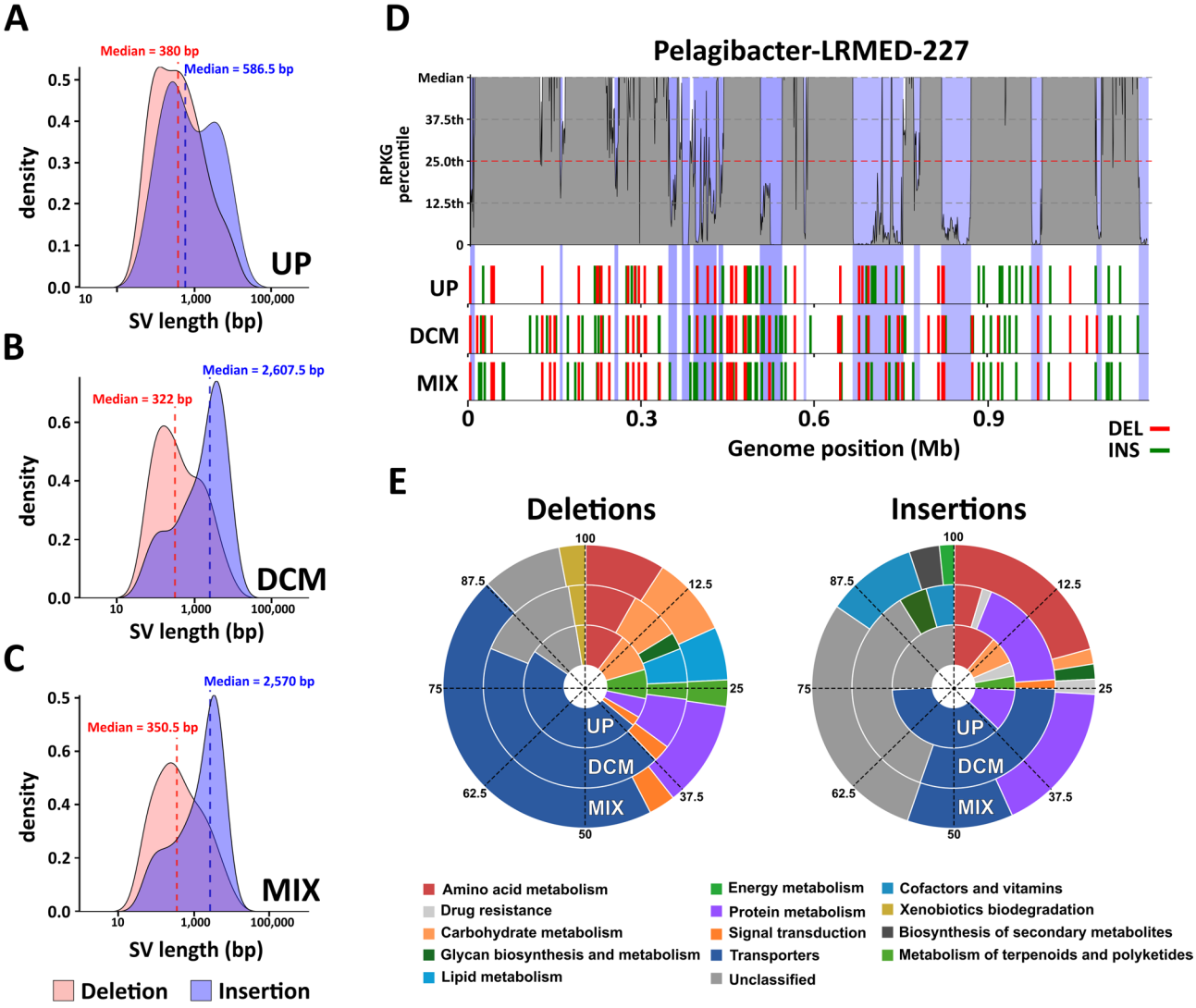

Fig. 3 Characterisation of SVs in *Pelagibacter*-LRMED-227. Density plot of the length distribution of deletions (red) and insertions (blue) with the median length of each SV type shown by the dotted line for **A**. Upper photic (UP) **B**. Deep chlorophyll maximum (DCM) and **C**. Lower photic (LP) metagenomes. **D**. The upper part shows the recruitment of short reads (Illumina) on the *Pelagibacter*-LRMED-227 genome in the Med-OCT2021-15 m sample (UP). The median RPKG of all genes in the genome was taken as a point to determine the core genome. The flexible genome (highlighted in blue) was determined as those values below the first quartile. The lower part indicates the position of insertions (green) and deletions (red) found in the UP (Med-OCT2021-15 m) DCM (Med-SEP2022-60 m) and MIX (MedWinter-JAN2019) metagenomes for Pelagibacter-LRMED-227. **E**. Multilevel donut plot with functional classification for deletions and insertions found in each of the metagenomes. DEL: deletion INS: insertion

the three metagenomes (Figure S3). A similar pattern was observed in Prochlorococcus-LRMED-150, where the majority of SVs (90%) were restricted to a single metagenome, and only one SV (~ 1%) was shared across all three (Figure S3). These findings highlight the role of SVs in niche-specific adaptation, emphasizing their potential contribution to microbial genome plasticity across distinct environmental conditions.

The genome-wide localization of SVs across different metagenomes revealed that 25% of these variants occurred at the same genomic position in both microbes (Figs. 3D and S1D), suggesting the presence of conserved SV hotspots analogous to genomic islands. To identify flexible regions, short reads from second-generation sequencing of the most abundant metagenome were mapped onto the genome (see Methods). This analysis allowed us to determine whether SVs were predominantly located within the core or flexible genome. In Pelagibacter-LRMED-211, approximately 65% of the SVs were associated with the core genome, while the remaining genomic variants were found within the flexible genome (Fig. 3D). For instance, in the winter metagenome, we identified a 5 Kb INS in the core genome that exhibited an allelic frequency of 1.0. This INS encompassed two transporters (Major Facilitator Superfamily and EamA-like transporter family) and an exoglucanase

(Figure S4A). An allelic frequency of 1.0 indicates that this variant is fixed within the population associated with this MAG in that particular environment, suggesting strong selection or ecological filtering. Moreover, a 2.5 Kb DEL was identified within a metagenomic island of Pelagibacter-LRMED-211, affecting the genes involved in glycerol metabolism (Figure S4B). As this region belongs to the flexible genome, allelic frequencies were lower and more variable across metagenomes, 0.34 in the DCM sample and 0.69 in the MIX sample, reflecting subpopulation heterogeneity (Figure S4B).

In Prochlorococcus-A-LRMED-150, INS and DEL exhibited distinct genomic distribution patterns. While the majority of DEL (62%) were located in coding regions, INS showed the opposite trend, with a higher proportion (62%) found in intergenic regions. Unlike Pelagibacter-LRMED-211, where most SVs were associated with the core genome, Prochlorococcus-A-LRMED-150 exhibited a predominant localization of SVs within the flexible genome. Specifically, 76% of DEL were found in the flexible genome, compared to 56% of INS (Figure S1D). For instance, a > 3 Kb deletion was identified in the core genome of the upper photic zone population, disrupting an ABC transporter involved in phosphonate uptake. This variant had an allelic frequency of 0.2, indicating its presence in a minor subpopulation relative to the reference genome (Figure S4C). In addition, a 2 Kb deletion located within a metagenomic island affected both a porin gene and a glutamine riboswitch—a non-coding RNA associated with nitrogen metabolism (Figure S4D). This deletion showed environment-dependent allelic frequencies, being fixed in the DCM population (allelic frequency = 1.0) but present at a lower frequency in the winter population (allelic frequency = 0.33) (Figure S4D).

Functional analysis was performed on the coding regions in which DEL and INS were detected in Pelagibacter-LRMED-227 across three metagenomes. DEL displayed a consistent profile across all environments, with a notable enrichment in transporter sequences, accounting for approximately half of the total (Fig. 3E). Around 10% of the sequences in each metagenome were associated with "amino acid" and "carbohydrate metabolism", as well as uncharacterized functions (Fig. 3E). Notable differences included an enrichment in "terpenoid and polyketide metabolism" in the upper photic zone (UP), while "lipid metabolism" was more prevalent in the deep chlorophyll maximum (DCM) and mixed layer (MIX) samples.

In contrast, INS showed higher variability between metagenomes. Among the annotated categories, "transporters" were most prevalent in the UP and DCM, followed by "protein metabolism", particularly peptidases and lipid biosynthesis. In the winter sample, "amino acid metabolism" was enriched (21%) with genes involved in glycine, serine, threonine, cysteine, and methionine metabolism, followed by "protein metabolism" (17%) and "transporters" (12%) categories (Fig. 3E). "Carbohydrate metabolism" was more prominent in the UP, while "energy metabolism" was enriched in the DCM and "cofactor and vitamins metabolism" (e.g., pantothenate and coenzyme A biosynthesis) in the MIX (Fig. 3E). Lastly, the proportion of unannotated sequences in INS was higher compared to DEL, ranging from 25% in the UP to 36% in the DCM. Among transporters, the TRAP transport system, DctPQM, was the most represented within SVs. Seven paralogs were identified in the Pelagibacter-LRMED-227 genome, with two showing partial deletions in a subset of the UP population. In the DCM and MIX samples, we observed that part of the population had lost the same paralog while acquiring another located in the same genomic position. As suggested in previous studies, members of this clade possess variable genomic regions linked to specific gene clusters, such as the TRAP. In this way, basic biological functions are maintained while expanding the range of substrates that can be used by the population [19]. Additionally, in the DCM, an INS event was associated with an ammonium transporter. Despite the presence of three paralogs in the genome, the acquisition of an environmentally variable copy with distinct substrate affinity could confer an adaptive advantage in this zone, characterized by increased ammonium availability [19].

Due to the limited number of functionally annotated DEL sequences in Prochlorococcus-A-LRMED-150, only INS were analyzed across the three environments. The proportion of unannotated sequences exhibited significant variation. In the UP and DCM samples, unannotated sequences were the most prevalent, accounting for 35% and 28% of the total, respectively (Figure S1E). In contrast, the MIX environment had the lowest proportion of unannotated sequences (19%) but showed the highest representation of sequences related to "genetic information processing" (38%) (Figure S1E). This category comprises genes associated with transcription factors, in addition to several integrases that may facilitate the insertion of other gene cassettes. The "genetic information processing" category was also abundant in the DCM, while the UP sample exhibited an enrichment of "transporters," primarily associated with phosphonate uptake. In the winter sample, however, no transporter-related genes were identified. Instead, a higher proportion of genes involved in the "metabolism of cofactors and vitamins," such as those related to thiamine biosynthesis, were observed (Figure S1E).

Several individual cases illustrate how SVs can enhance metabolic diversity within species, thereby increasing the adaptive capacity of specific subpopulations to varying physicochemical conditions. For instance, we observed

the loss of light-related genes in subpopulations associated with the metagenome of the lower photic zone (LP), compared to their counterparts in the upper layers. In Pseudothioglobus-LRMED-305, a DEL with an allelic frequency of 1.0 resulted in the complete loss of the flotillin-associated rhodopsin (FArhodopsin) gene cluster [32, 33] in LP-associated populations (Figure S5A). Similarly, in Acidimicrobiales-LRMED-036, a DEL of the bacteriorhodopsin gene was identified in the depth-associated population (Figure S5B and S6A). Both rhodopsins are typically enriched in microbes inhabiting the photic water column due to increased light exposure [32]. In the archaeon Nitrosopumilus-LRMED-033, an LP-specific DEL (allelic frequency 0.35) removed a genomic fragment containing a (6 – 4) DNA photolyase, an enzyme involved in UV damage repair (Figure S5C and S6B). In the same sample, INS was also detected in a subset of the Nitrosopumilus-LRMED-033 population, which contains a V-type ATPase. This enzyme has been associated with adaptations to high-pressure conditions characteristic of deep-water environments [34]. Population-level variations related to nutrient bioavailability were also observed. In SAR324-LRMED-400, a 5 Kb INS containing a two-component system (PdtaR/PdtaS) (allelic frequency 1.0), known for its role in nutrient starvation responses, was identified in the subset of the population associated with the LP sample [35] (Figure S5D). In the nutrient-depleted surface waters, an insertion (allelic frequency = 0.75) in SAR86-LRMED-332 encoded an alkaline phosphatase, enhancing phosphate acquisition under phosphorus limitation (Figure S5E). Furthermore, in the winter metagenome, a subpopulation of Flavobacteriaceae-LRMED-096 harbored a 7 Kb insertion (allelic frequency = 1.0) encompassing the complete histidine biosynthesis operon (Figure S5F). These examples underscore how SVs, even when involving small genomic regions, can become fixed within specific environmental populations (allelic frequency = 1.0), likely driven by natural selection. Such SVs modulate metabolic capabilities and contribute to the emergence of genomically distinct clonal lineages, ultimately enhancing population-level functional diversity in response to ecological pressures.

Recovery of metagenomic viral sequences and putative host prediction

We identified a total of 742 metagenomic viral sequences (hereafter referred to as viral genomes, or VGs; see Methods), ranging in size from 13 to 437 Kb (average sequence length of 56 Kb) and GC content spanning from 23 to 64% (Table S7). Among them, 94 sequences were identified as complete viral genomes. Host prediction was computationally feasible for 41% of the sequences, with Alphaproteobacteria emerging as the most represented host group, including 42 sequences linked to *Pelagibacter* (Table S7). Other predicted hosts included members of the classes Bacteroidia, Gammaproteobacteria, and Cyanobacteriia, with the latter including the genera *Prochlorococcus* and *Synechococcus* (Table S7). Taxonomic classification revealed that most viral sequences belonged to the class Caudoviricetes, with the family Autographiviridae (T7-like viruses) being the most prevalent (Table S7). Additionally, seven VGs were assigned to the class Megaviricetes, which includes giant viruses that infect marine eukaryotic algae.

Structural variants in the marine virome

To analyze the intrapopulation microdiversity introduced by SVs within viral populations, we first selected VGs exhibiting the highest relative abundance in at least one of the metagenomic datasets (Table S3). A sequence was considered present in a given metagenomic sample if it recruited at least ten reads per kilobase of sequence and per gigabase of metagenome (RPKG), with a minimum identity of 99% and ≥ 80% contig coverage. A total of 92 VGs met these criteria, five of which were present in more than one metagenome (Table S3). These VGs were subsequently used for long-read mapping and SV detection. Despite their high abundance, 40 of these VGs exhibited no detectable SVs. However, the remaining 52 VGs revealed a total of 380 SVs (Table S8). Consistent with previous findings for the prokaryotic fraction, INS (48.4% of total SVs) and DEL (46.6%) were found to be significantly more prevalent than duplications and inversions (Fig. 4A and Table S8). Furthermore, a robust and statistically significant correlation was identified between the number of INS and DEL per sequence ($r = 0.80$; $p < 0.05$). While the prokaryotic fraction exhibited a bimodal size distribution for both INS and DEL, the viral fraction displayed a unimodal distribution. The mean size of INS and DEL in the viral fraction was similar, averaging approximately 430 bp for both types of SVs (Fig. 4B). The recovery rate of these SVs from short-read assemblies of the same metagenomes was higher than that observed in the prokaryotic fraction. Except for the winter sample, where recovery was only 20%, all other samples exhibited recovery rates above 68%, reaching a maximum of 76% in the UP sample. This phenomenon can be attributed to the low genomic heterogeneity of marine phage populations coming from cellular metagenomes, which predominantly comprise natural amplified phages generated during the replication process within the lytic cycle.

Functional characterization of viral structural variants

A total of 456 proteins were predicted within the SVs obtained from the VGs. However, only 43% of these proteins ($n = 197$; Table S9) could be annotated. Among the annotated proteins, 94 were classified as hypothetical

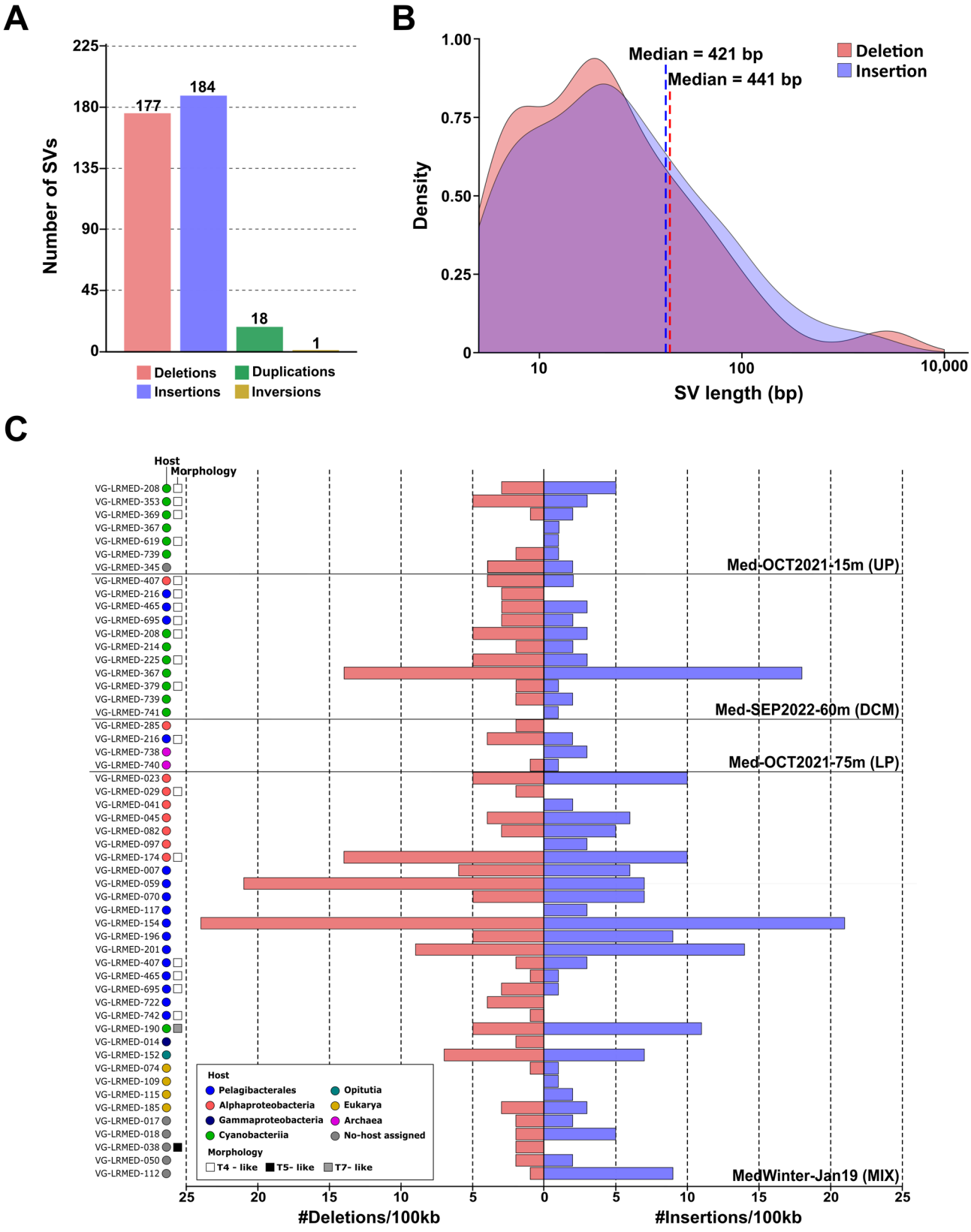

Fig. 4 Characterization of Structural Variants in the marine virome. **A.** Number of SVs detected in viral genomes. **B**. Length distribution of deletions (red) and insertions (blue) with the median length of each type of SV shown by the dotted line. **C**. Number of deletions (right) and insertions (left) normalized per 100 Kb of genome for viral genomes distributed by metagenome. The colour of the bars indicates the SV type (deletions and insertions are highlighted in red and blue, respectively). VG: Viral Genomes

proteins, even though we used three specialized viral databases: Prokaryotic Virus Orthologous Groups (pVOGs) [36], Database of Virus Orthologous Groups (VOGDB) [37], and vFam [38]. The second most represented category was auxiliary metabolic genes (AMGs), several of which are associated with cyanophages, including those related to photosystems and oxidative stress management, such as 2OG-Fe(II) oxygenase (Table S9). We also identified several genes annotated as glycosyltransferase, phosphoheptose isomerase, cytidyltransferase, and carboxylesterase, all of which are involved in the synthesis and modification of the host cell envelope and surface structures. Following AMGs, the most represented category was related to infection mechanisms, with a notable prevalence of tail fibers. These structures play a crucial role in host recognition, attachment, and the initiation of infection, and their persistence within the population may contribute to an expanded host range. In the categories related to packaging and immune evasion, we observed an enrichment of exo- and endonucleases, as well as methyltransferases, respectively.

In a myophage putatively infecting *Pelagibacter* (pelagimyophage, VG-LRMED-216), we detected an SV harbouring a phosphate-binding protein (PstS) (Table S9). In Cyanobacteria, the rate of phosphate uptake has been shown to increase in cells infected with cyanophages carrying this gene [39]. Therefore, it is plausible that a similar effect could occur in Pelagibacterales. To evaluate the putative origin and possible role of the novel PstS protein, we performed a maximum-likelihood phylogenetic tree including several sequences from a dereplicated set of nearly 1,500 Pelagibacterales genomes [14, 40, 41], as well as three PstS sequences recovered from viral *Pelagibacter* myophages [12, 42]. The phylogenetic tree of Fig. 5A indicated that the *Pelagibacter* viral PstS sequences clustered far from those of cyanobacterial viruses, forming a monophyletic branch within the *Pelagibacter* bacterial sequences, which were further split into two major branches, regardless of their genomic classification. The amino acid alignment (Figure S7) confirmed the dissimilarity of *Pelagibacter* and Cyanobacteria viral PstS, as they shared only 24% of their amino acid sequences. In contrast, amino acid identity between bacterial and viral *Pelagibacter* PstS sequences was 61%, indicating a significant level of divergence (Figure S7). We modelled the PstS protein structures for Pelagibacter HTCC7214 (bacterial) and MAVG-MED-5 (viral) using AlphaFold3 [43], obtaining high-quality structures (mean predicted local distance difference test [pLDDT] > 90) (Fig. 5B). The alignment of both structures using TM-align [44] resulted in a near-perfect fit (tm-score 0.98), suggesting that viral PstS can replace bacterial PstS during infection (Fig. 5B).

Structural variations in environmental persistent phages

As observed for their hosts, we identified several of these viruses in more than one sample (Table S3). The sequences classified as myophages associated with *Pelagibacter* and *Prochlorococcus* were determined to be complete genomes (circularized). Alignment with short-read metagenomic sequences confirmed full genome coverage (average sample coverage > 10x) with no gaps or indications of metagenomic islands. Interestingly, comparing these sequences with previous virome studies from the cellular fraction of the same geographic region [12, 42] revealed the presence of syntenic sequences and identical genetic content. These studies have already shown that myophage genomes recruit significantly more reads from cellular metagenomes than from viromes. Based on previous recruitment data [12], these myophages appear to be endemic to the western Mediterranean Sea. Moreover, the alignment of reads from samples collected up to a decade earlier revealed a similar pattern, with full genome coverage, indicating high persistence and low genetic variability in some viral genomes (Figure S8).

Despite the accumulation of mutations, this finding allowed us to investigate the role of SVs in the persistence of these natural myophage populations. For instance, VG-LRMED-465, a pelagimyophage, recruited reads across its entire genome not only in the DCM and MIX samples from our study but also in the Med-OCT2015-30 m sample (BioSample SAMN05992380), collected in the Mediterranean Sea in 2015 [6] (Figure S8A). However, SV analysis revealed eight SVs in the DCM and four in the winter sample. Three of these SVs were located in the same genomic region in both metagenomes, specifically within structural genes. Among the four SVs exclusive to the DCM, two were located in the region between the *vrlC* gene and the neck *gp14* gene, designated as the host recognition cluster (HRC) [45] (Figure S8A). Another pelagimyophage, VG-LRMED-695, was not only detected in two metagenomes (DCM and MIX) but also fully covered in a cellular metagenome from the TARA expedition (TARA_007; ERR315856), collected over a decade earlier (2009) [7]. SV analysis showed population variations associated with DCM and MIX, with eight and six SVs, respectively (Figure S8A). Only one SV was shared between the two metagenomes, while three SVs in the DCM were associated with the HRC. Unfortunately, due to poor annotation, we were unable to infer additional functional insights from the remaining SVs.

Similarly, we identified two cyanomyophages (VG-LRMED-739 and VG-LRMED-741) with complete genome coverage in the UP and MIX samples, as well as in the TARA_007 sample (ERR315856) [7]. VG-LRMED-739 exhibited six SVs in UP and seven in DCM (Figure S8B). Up to five SVs in both metagenomes were located within the same gene in the HRC region, which,

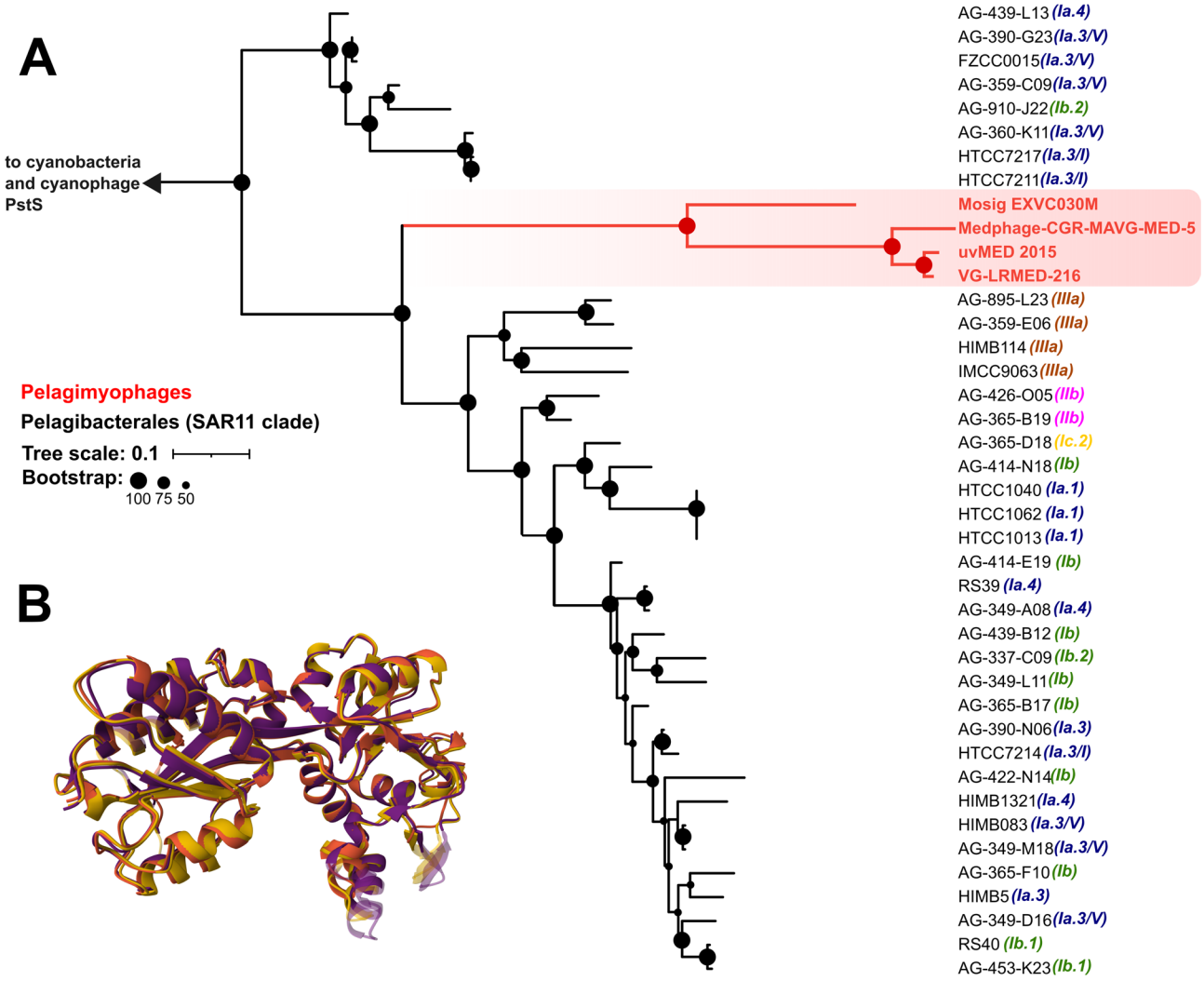

Fig. 5 Phylogeny of phosphate ABC transporter, substrate-binding protein (PstS) sequences from *Pelagibacter* and Pelagimyophages **A**. Maximum likelihood phylogenetic tree of PstS sequences. Red branches contain sequences of viral origin and black branches contain bacterial sequences. Numbers in brackets indicate the subclade to which each genome belongs within the SAR11 clade. Several cyanophage PstS sequences were used as outgroup. **B**. Superimposed structural models predicted with AlphaFold3 for the PstS proteins of bacterial *Pelagibacter* HTCC7214 and HTCC7217 (orange and red structures) and the pelagimyophage MAVG-MED-5 (purple structure)

like pelagimyophages, was located between the structural protein (VrlC) and the neck protein (Figure S8B). Although the encoded protein was annotated as a structural protein, its predicted structure using AlphaFold3 exhibited similarity in the PDB100 database to a receptor-binding tail spike domain of the cyanophage Pam1 [46], as well as to an Mtd_N domain (Q775D6) in UniProt from a tail fiber receptor-binding protein. This suggests that structural variations may be involved in facilitating phage-host recognition and attachment. In contrast, cyanomyophage VG-LRMED-741 showed fewer structural variations, with no SVs detected in UP and only two INS in DCM (Figure S8B). One of these insertions was also located in the HRC, and, as observed in VG-LRMED-739, within a protein containing the same Mtd_N domain (Q775D6). These SVs identified in phage tail genes are relatively short (100–300 bp) compared to the full length of the genes (~6 Kb). Therefore, they are unlikely to represent complete deletions or replacements, as these genes are essential for phage function. Although AlphaFold-based structural predictions were performed on tail proteins affected by SVs, the models did not reveal clear changes associated with host interaction domains. The affected regions did not localize consistently to receptor-binding motifs, limiting our ability to infer functional consequences (data not shown). These results highlight the important role of SVs in the marine virome as a major contributor to intraspecies diversity in natural persistent phage populations.

Discussion

During the past decade, advancements in second-generation sequencing technologies have enabled the recovery of individual genomes, both of prokaryotic microbes (MAGs) and their associated viruses, directly from complex environmental samples, eliminating the need for cultivation. The subsequent exponential growth of these sequences in public databases has significantly enhanced our understanding of bacterial population evolution and structure, the adaptive mechanisms of bacteria to their environments, and their interactions with other biological entities [6, 7, 9, 45].

The formation of microbial species is driven by consortia of subpopulations that coexist within a shared habitat [47, 48]. Consequently, using these genetic sequences (MAGs and VGs), which provide static representations of individual organisms within complex environmental populations, restricts our ability to fully capture the genetic variability present in natural habitats, as well as the metabolic capacities of microorganisms and their ecological roles. This is particularly relevant given that flexible genes are often niche-specific. For instance, within the order Alphaproteobacteria, the genetic diversity of Pelagibacterales (SAR11) and HIMB59 enables populations to adapt to fluctuations in specific micronutrients in the environment, driving the emergence of ecologically distinct lineages within species [14, 19, 49].

In this sense, third-generation metagenomics emerges as a promising approach for analyzing intraspecific diversity and deciphering evolution at the level of natural populations. Our results show that a substantial fraction of the species-associated environmental pangenome is encoded within SVs. These SVs exhibited high heterogeneity among the different taxa, although the relative abundance of INS and DEL remained balanced. Although metagenomic detection of SVs in marine systems is unprecedented to date, a strong correlation between INS and DEL has also been observed in the only study conducted on aquatic microbiomes, specifically in Lake Biwa [23]. Since SVs can contain one or multiple genes, or even entire operons, these genomic variations have a greater impact on metabolism than single nucleotide variants. Consistent with the oligotrophic nature of the system, it is reasonable that the most represented categories among these SVs were transporters and core metabolic functions such as amino acid metabolism. The biological significance of this remarkable diversity of transporters suggests that a single population may be capable of processing thousands of different organic compounds, which aligns with the high chemical diversity of dissolved organic matter (DOM) found throughout the water column [50]. However, in the previous study conducted on a lacustrine system, the focus was solely on deletions, In that study, although these categories were well represented, they were proportionally less dominant than others, such as prokaryotic defense systems and motility-related proteins [23]. The enrichment of these latter categories was associated with large, flagellated genomes, whereas our system is dominated by streamlined microbes with an osmotrophic lifestyle.

Consistent with findings from studies of the human gut microbiome [25], the size distribution of SVs across our marine samples showed that approximately 60% were shorter than 500 bp. These small, non-coding SVs merit further investigation, as they may represent neutral variants maintained by genetic drift or exert regulatory functions, for instance by altering mRNA secondary structures and modulating the expression of adjacent genes. Supporting this possibility, we identified several ncRNAs within SV regions that are known to regulate gene expression and contribute to microbial physiological responses to environmental stimuli, including nutrient limitation and stress conditions. Similarly, in streamlined microbes like Pelagibacter, with an average intergenic spacing of just 2 bp, SVs often insert into coding regions. Even when the SVs lack coding sequences, their insertion may strongly affect the phenotype. SVs inserted within a gene are also likely to contain partial coding sequences. However, due to their fragmentary nature, these sequences may not be detected by protein-coding gene prediction tools, which could lead to an underestimation of the actual proportion of SVs containing coding sequences. The quantification of truncated genes could represent a fruitful avenue for future research. Rather than canonical genomic islands, which are fragments of more than 10 Kb, these SVs appear to be the primary source of intra-species richness in microbes with streamlined genomes such as *Pelagibacter*, *Prochlorococcus* or *Nitrosopumilus* that dominate the different layers of the water column. These findings align with recent studies that, through a combination of single-amplified genomes and long-read metagenomics, have demonstrated that the flexible SAR11 genome is organized into small regions, each typically containing a single gene [19]. These small genomic variations can have a profound impact on the population, allowing the creation of ecological lineages adapted to specific conditions. An example of this is the rhodopsin, as observed in the Pseudothioglobus-LRMED-305 and Acidimicrobiales-LRMED-036 genomes. This protein endows the population with the capacity to adapt to environments characterized by low nutrient availability and fluctuations in light within the water column. This phenomenon could give rise to a subpopulation with enhanced metabolic flexibility, increasing its competitiveness with other heterotrophic bacteria and facilitating the colonisation of new niches. Although rhodopsin was assembled in the MAGs from both cases, the subpopulation in the

LP sample lacks this gene, pointing to either an incipient speciation event or an adaptation of this subpopulation to the local environmental conditions. Remarkably, these SVs were identified not only within the core genome but also in the flexible genome, acting as a source of genetic variability within genomic islands. This variability facilitates the rapid acquisition of adaptive functions and contributes to their evolutionary dynamics. Moreover, SVs may even represent fundamental units in the biogenesis of genomic islands. The presence of these SVs in the same genomic position across different metagenomes suggests hotspots for the insertion of such genomic variations, akin to tRNA sites in genomic islands. The conservation of insertion sites in the genome may serve as an evolutionary mechanism favouring HGT exchange between clonal lineages within the population cohabiting in the same environment, as previously suggested [19]. Therefore, further research should investigate this phenomenon in depth to elucidate the dynamics of insertion and deletion events.

Similar to prokaryotes, we have found that SVs also contribute to the genomic heterogeneity of viral populations. To date, the only study using the same approach to determine these small-scale genetic variations focused on the human gut phageome [24]. Although DEL and INS were the most abundant SVs in both environments, their average size was larger in the marine virome (430 bp vs. 170 bp). At the functional level, human gut phage communities exhibited an enrichment in genes involved in recombination, DNA methylation, and antibiotic resistance. In contrast, our study revealed a greater abundance of genes associated with the evolutionary dynamics of phage-host interactions, including those related to recognition and attachment processes. Additionally, SVs can modulate host metabolism during infection via AMGs. While there is a bias toward cyanobacterial AMGs due to their extensive study, we identified an enrichment in AMGs involved in host cell wall synthesis and modification. These AMGs have been suggested to induce changes in the host glycotype as a mechanism to prevent superinfection [51]. In line with previous findings in cyanophages [39], we identified a high-affinity phosphate transport system (PstS) in pelagimyophages, which modulates host metabolism to enhance phosphate uptake, thereby maximizing viral replication success. This effect is particularly relevant in the Mediterranean Sea, where phosphate concentrations are limiting compared to the global ocean [52]. At the primary sequence level, the identity between the phage and host proteins was 60%. However, the AlphaFold-predicted tertiary structure revealed a 98% identity, suggesting that the phage-encoded PstS subunit is fully compatible with the host's PstAB transmembrane complex, although experimental validation is still required. Previous studies on cyanophages have demonstrated that the phage-encoded PstS variant does not increase affinity for host phosphate. Nevertheless, a higher concentration of the substrate-binding protein enhances the rate at which the transporter encounters phosphate, ultimately increasing substrate uptake efficiency [51].

Previous work on cyanophages proposed the existence of stable viral genomic clusters that persist for over a decade in specific locations [53]. In our study, metagenomic reads allowed us to recover nearly identical complete viral genomes, encompassing not only cyanophages but also pelagiphages, despite differences in origin, year, and season of sampling. These findings suggest high persistence and low genetic variability in some phage genomes, which remain stable for at least eight years, exhibiting an endemic distribution in the Mediterranean Sea. We refer to these as environmental persistent genomes. This high persistence may indicate strong specialization between bacteria and their viral predators, maintaining only subpopulations differentiated by small-scale SVs, primarily associated with host recognition. It has been hypothesised that viruses may adhere to a seed bank model and exhibit long-term stability [54]. However, it should be noted that these viruses were obtained from the prokaryotic fraction, which consists predominantly of cells infected with lytic cycle-active viruses. Therefore, at least in these myophages, the observed genetic stability is consistent with the Constant-Diversity model [55], which may better reflect evolutionary dynamics at the ecological scale. Phage populations could act as regulators, maintaining a balance among the subpopulations that comprise natural prokaryotic species. In this context, SVs in phages represent a key mechanism for generating genetic diversity within natural microbial communities. The persistence of nearly identical phage genomes over time mirrors patterns observed in long-term studies such as the HOT time series [56], as well as the central strain stability reported by Ignacio-Espinoza et al. (2019) [57]. These findings suggest that stable virus-host interactions may be a common feature in oligotrophic environments like the Mediterranean Sea, despite ongoing microdiversity. Nevertheless, other studies, particularly those investigating Prochlorococcus-infecting phages [58, 59], have uncovered complex diurnal dynamics and shifts in life strategies, which may not be fully captured by our approach. As our analyses are based on VGs derived from the cellular fraction, likely enriched in actively replicating viruses, we acknowledge a potential bias toward lytic cycles.

All these genomic findings on the role of SVs not only deepen our understanding of microbial diversity but also provide a more comprehensive perspective on the ecology and evolution of microbial communities. However, despite the challenges of cultivating most of these

microbes in pure culture, experimental validation is essential to fully comprehend these mechanisms. Therefore, these genomic data should serve as the foundation for future research, where genomic insights are supported by experimental evidence to unravel the ecological role of the high levels of genomic heterogeneity observed in marine prokaryotic populations.

Conclusions

The continuous advancement of sequencing technologies is profoundly transforming our understanding of genomic diversity in natural environments. Eliminating the need for the assembly step, long-read sequencing enhances the resolution of intragroup genomic diversity, thereby highlighting SVs as key drivers of gene content variation and uncovering the full metabolic potential at the intraspecific level in environmental samples. This remarkable genetic diversity underscores the importance of population-level analysis. This is particularly true in streamlined microbes, which maintain broad genetic repertoires across multiple subpopulations, enabling metabolic plasticity in response to environmental fluctuations. Consequently, these findings call for a re-evaluation of our understanding of genomic diversity in environmental samples, as current approaches have only provided access to a limited fraction of the total genomic diversity. Future research should integrate this genomic information with functional data to elucidate the role of SVs in microbial ecosystem stability, resilience, and virus-host interactions.

Methods

Sample collection, DNA extraction and sequencing

Four marine samples were collected from the same sampling site in the epipelagic Mediterranean Sea at 20 nautical miles off the coast of Alicante (Spain) (37.35361°N, 0.286194°W). Med-OCT2021-15 m, Med-SEP2022-60 m, and Med-OCT2021-75 m were collected in summer, during a strong stratification period [26]. We added to the comparison a winter sample, MedWinter-JAN2019 [11], collected in January 2019 at the same sampling location, when the water column was fully mixed (Table S1). For each depth, 200 L were collected and filtered on board as described [6]. Briefly, seawater samples were sequentially filtered through 20-, 5-, and 0.22-µm pore filter polycarbonate filters (Millipore). Water was directly pumped onto the series of filters to minimize the bottle effect. Filters were immediately frozen on dry ice and stored at − 80 °C until processing. DNA extraction was performed from the 0.22-µm filter (free-living bacteria) following the MagAttract Purification Kit protocol (QIAGEN). DNA was extracted from each sample and used for Illumina NextSeq 150 bp paired-end reads (60 Gb/sample) and PacBio Sequel II sequencing (one 8 M SMRT Cell Run, 30-h movie) (Novogene, South Korea). The accession numbers for the Illumina and PacBio metagenomes are provided in Table S1.

Metagenomic raw reads filtering and assembly

Illumina raw reads were trimmed using trimmomatic v0.39 [60] with the following options: LEADING:3, TRAILING:3, SLIDINGWINDOWN:4:15, MINLEN:50. PacBio raw reads were error corrected by generating Highly Accurate Single-Molecule Consensus Reads (CCS reads) using the CCS v6 program of the SMRTlink package. The minimum number of full-length subreads required to generate a CCS read was set to 15 (> 99.99% base call accuracy). Metagenomic assembly was performed individually, using a hybrid approach with metaspades v3.15 [61], on which the CCS15 reads > 1Kb were provided as single (-s) reads, Illumina reads as forward (-1) and reverse (-2), and with the following options: --meta --threads 64 --memory 950 -k 49,59,69,79,89,99. The choice of a hybrid assembly over a single PacBio de novo assembly was based on a previous result, which showed that this approach resulted in a better assembly yield and larger contig recovery [11].

Binning, genome reconstruction and phylogenomic characterization

Assembled contigs longer than or equal to 5 Kb were selected and subjected to metagenomic binning using SemiBin2 v2.1.0 [62], with the multi_easy_bin approach and the sequencing-type option set to long-read. Contig coverage was calculated using bowtie2 [63] against the Illumina metagenomes collected in this work, together with a set of metagenomic samples collected in late summer 2015 (15, 30, 45, 60, 75, 90, 1000 and 2000 m deep) and winter 2015 (20 and 80 m) that spanned the epipelagic water column [6]. The resulting MAGs were screened for completeness and contamination, as well as genomic features such as coding density, average gene length, and GC content, using CheckM2 [64]. MAGs were dereplicated at 99% nucleotide identity (ANI) using dRep [65]. Lastly, only MAGs with > 50% completeness and < 5% contamination were retained for further analysis. MAGs were classified according to the MIMAG (Minimum Information about a Metagenome-Assembled Genome) guidelines [10] defining high-quality MAGs as those with > 90% completeness, < 5% contamination, the presence of at least 18 out of the 20 tRNA genes encoding the standard amino acids, and the presence of at least one complete rRNA operon (16S, 23S, 5S). Taxonomy for the reconstructed MAGs was assigned using the Genome Taxonomy Database (GTDB) v2.4.0 [66] (release 09-RS220; April 24th, 2024). Open-reading frames (ORFs) were predicted with Prodigal v2.6.3 [67] with the metagenomic option. tRNA and rRNA genes were

predicted using tRNAscan-SE v2.0.5 [68] and barrnap v0.9 [69], respectively. Phylogenomic classification was conducted using PhyloPhlAn 3.0 with the parameters: -d phylophlan -t a --diversity high --accurate -f supermatrix_aa.cfg [70]. A phylogenetic tree was then inferred using the maximum-likelihood method with 1,000 ultrafast bootstrap replicates and the LG + G4 substitution model, implemented in IQ-TREE [71]. The average nucleotide identity based on metagenomic reads (ANIr) was calculated using the enveomics R package [72].

Viral contig recovery, taxonomic affiliation, and host prediction

Viral sequences were recovered from contigs larger than 10 Kb across the four metagenomes (Table S7) using VIBRANT v1.2.0 [73] with default parameters. Redundant nucleotide sequences were dereplicated at 95% identity using CD-HIT v4.8.1 [74]. Sequence quality was assessed with CheckV v1.0.1 [75], and only contigs classified as complete, high-quality, or medium-quality were retained. Potential host prediction was performed using the IPHOP program v.1.3.2 [76], which combines multiple approaches for virus-host prediction. Phabox v2.0 [77] and geNomad v1.8.0 [78] were used for the taxonomic assignment of viral sequences and to determine the phage lifestyle.

Metagenomic fragment recruitment

Genomes reconstructed in this study (MAGs and VGs) were used to recruit reads from our metagenomic datasets via BLASTn v2.10.1 [79]. We applied a stringent cut-off of 99% nucleotide identity over a minimum alignment length of 50 nucleotides, requiring ≥ 80% genome coverage by recruited reads. These recruited reads were then used to calculate RPKG values (reads recruited per kilobase of genome per gigabase of metagenome), providing a normalized metric for comparison across metagenomic samples. Illumina short reads were competitively mapped to a single fasta file containing all genomes using Bowtie2 [63], and analyzed using inStrain (v1.8.0) [80] in profile mode on predicted genes to assess microdiversity parameters. To delineate the core and flexible genomes and thereby determine the localization of SVs within Pelagibacter-LRMED-227 and Prochlorococcus-A-LRMED-150, we employed a metagenomic recruitment-based approach. Each gene was individually assessed based on its recruitment level in the metagenome where the reference microbe exhibited the highest overall recruitment, quantified as RPKG. From the full dataset, we calculated the median and percentile distribution of gene recruitment values. Genes with recruitment at or above the median were classified as part of the core genome, while metagenomic islands were defined as genomic regions in which gene recruitment fell below the first quartile (25th percentile).

Detection of Structural Variants

To ensure robustness in SV analysis, only MAGs and VGs meeting stringent quality criteria were selected. Specifically, MAGs with RPKG values ≥ 5 in at least one metagenome, ≥ 80% genome integrity and < 50 contigs (except Marinisomatales-LRMED-188, which has 69 contigs, thus allowing some genomes of this taxon to be included). Regarding VGs, only sequences with at least 80% coverage and an abundance value exceeding 10 RPKGs were selected for SV analysis. Then, high-fidelity PacBio CCS reads were subsequently aligned to the selected genomes using the NGMLR aligner (v0.2.7) [81]. The resulting BAM file was sorted using samtools [82] sort option and used as input for SV detection via Sniffles2 (v2.4) [29]. From the resulting VCF file containing the five types of SVs identified by Sniffles2, DEL, INS, duplications, and inversions were selected for further analysis. However, due to their potential to span multiple contigs when working with MAGs, translocations were excluded from further analysis. Variants classified as imprecise, larger than 50 Kb in length or with an allele frequency of less than 0.1 were excluded from further analysis. To functionally classify INS or DEL from all detected SVs in VGs and MAGs, respectively, ORFs were predicted using Prodigal v2.6.3 with the -meta option [67]. ORFs in deletion sequences were predicted directly from the excluded (i.e. deleted) region in the MAG. In the case of insertions, we predicted the ORFs from the SV sequence provided in the Sniffles2 VCF file. The resulting proteins were then annotated against the KEGG Orthology (KO) database (downloaded in January 2024 [30]) using hmmscan [83], applying the predefined trusted cut-off values specific to each KO HMM model. In addition, ncRNAs were predicted using cmscan from the INFERNAL v1.1.5 package [84], with the predefined covariance models using the Rfam database v15 [85] with the following options: --rfam --cut_ga --nohmmonly. In addition, viral ORFs were also annotated using HMM profiles from Prokaryotic Virus Orthologous Groups (pVOGs) [36], Virus Orthologous Groups (VOGDB) release 227 [37], and vFam release 227 [38], using hmmscan.

Detection of SVs in Illumina assemblies

Because Illumina assemblies are often fragmented, especially in regions of high variability such as SVs, we undertook a comprehensive evaluation of SV recovery from Illumina contigs. This was achieved by aligning previously predicted SVs (as detailed in the "Detection of Structural Variants" section) against these assemblies. More precisely, nucleotide sequences corresponding to INS and DEL SVs were aligned using BLASTn [79]. A

stringent threshold of 98% nucleotide identity and 98% of alignment length was applied to ensure high-confidence.

Analyses PstS sequences

A PstS protein predicted from a viral genome (VG-LRMED-216) putatively infecting *Pelagibacter* was compared against a subset of 37 cellular *Pelagibacter* PstS proteins, selected randomly from a dereplicated collection of nearly 1,500 genomes, using CD-HIT at 98% amino acid identity. *Pelagibacter* PstS sequences were taxonomically classified following the nomenclature in [14]. We included in the comparison three more viral sequences from pelagimyophages [42, 45, 86], as well as the cyanobacterial PstS sequences from *Prochlorococcus marinus* MIT9314 (KGG02356.1), MIT9301 (ABI23452.1), *Synechococcus* sp. MED-G71 (RCL53563.1), and their phages P-RSM4 (YP_004323306.1), P-SSM2 (YP_214479.1), and S-SSM7 (YP_004324340.1). Amino acid alignment was performed using muscle v5 [87], curated using trimal [88] with the -automated1 option, and a maximum-likelihood phylogenetic tree was performed using IQ-TREE2 with 1000 ultrafast bootstraps and WAG-G4 as the substitution model. Protein modelling of P-SSM2, *Pelagibacter* HTCC7214 and VG-LRMED-216 was performed using AlphaFold3 [43]. Structural alignments were made using tm-align [44] and visualized using pymol (https://github.com/schrodinger/pymol-open-source).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40793-025-00773-8.

Supplementary Tables
Supplementary Figure 1
Supplementary Figure 2
Supplementary Figure 3
Supplementary Figure 4
Supplementary Figure 5
Supplementary Figure 6
Supplementary Figure 7
Supplementary Figure 8
Supplementary Figure 9

Acknowledgements
Not applicable.

Author contributions
MLP conceived the study. JHM, CMP, JRG, and MLP analyzed the data and participated in the collection of environmental samples. JHM and MLP contributed to write the manuscript. All authors revised the manuscript and approved the final version.

Funding
This research was funded by the grant "MICRO3GEN" (PID2023-150293NB-I00), awarded by the Spanish Ministry of Economy, Industry and Competitiveness, and co-financed with FEDER funds, to MLP. CMP received support through a Ph.D. fellowship from the Spanish Ministry of Science and Innovation (PRE2021-098122).

Data availability
Illumina raw reads and PacBio CCS reads have been submitted to NCBI SRA and are available under BioProject accession numbers PRJNA1088973 and PRJNA674982. In addition, MAGs have been deposited under BioProject accession number PRJNA1088973. Viral sequences are publicly available in Zenodo (10.5281/zenodo.14650439).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Microbial Genomics and Evolution Group, División de Microbiología, Universidad Miguel Hernández, Apartado 18, San Juan, Alicante 03550, Spain

Received: 20 April 2025 / Accepted: 18 August 2025
Published online: 25 August 2025

References

1. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. Volume 281. New York, NY: Science; 1998. pp. 237–40.
2. Weinbauer MG. Ecology of prokaryotic viruses. FEMS Microbiol Rev. 2004;28:127–81.
3. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018;3:754–66.
4. Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N, et al. Community genomics among microbial assemblages in the ocean's interior. Science. 2006;311:496–503.
5. Konstantinidis KT, Braff J, Karl DM, DeLong EF. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical Gyre. Appl Environ Microbiol. 2009;75:5345–55.
6. Haro-Moreno JM, López-Pérez M, de la Torre JR, Picazo A, Camacho A, Rodriguez-Valera F. Fine metagenomic profile of the mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome. 2018;6:128.
7. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean Microbiome. Volume 348. New York, NY: Science; 2015. p. 1261359.
8. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2017;2:1367–73.
9. 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–e112314.
10. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725–31.
11. Haro-Moreno JM, López-Pérez M, Rodriguez-Valera F. Enhanced recovery of microbial genes and genomes from a marine water column using Long-Read metagenomics. Front Microbiol. 2021;12:708782.
12. Zaragoza-Solas A, Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. Long-Read metagenomics improves the recovery of viral diversity from complex natural marine samples. mSystems. 2022;7:e0019222.

13. Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. Nat Methods. 2023;20:1143–58.
14. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-Garcia JJ, Gomez ML, et al. Ecogenomics of the SAR11 clade. Environ Microbiol. 2020;22:1748–63.
15. Roda-Garcia JJ, Haro-Moreno JM, Huschet LA, Rodriguez-Valera F, López-Pérez M. Phylogenomics of SAR116 clade reveals two subclades with different evolutionary trajectories and an important role in the ocean sulfur cycle. mSystems. 2021;6:e0094421.
16. Roda-Garcia JJ, Haro-Moreno JM, Rodriguez-Valera F, Almagro-Moreno S, López-Pérez M. Single-amplified genomes reveal most streamlined free-living marine bacteria. Environ Microbiol. 2023;25:1136–54.
17. López-Pérez M, Haro-Moreno JM, Iranzo J, Rodriguez-Valera F. Genomes of the Candidatus Actinomarinales Order: Highly Streamlined Marine Epipelagic Actinobacteria. mSystems. 2020;5.
18. López-Pérez M, Rodriguez-Valera F. Pangenome evolution in themarine bacterium alteromonas. Genome Biol Evol. 2016;8:1556–70.
19. Molina-Pardines C, Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. Extensive paralogism in the environmental pangenome: a key factor in the ecological success of natural SAR11 populations. Microbiome. 2025;13:41.
20. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021;53:779–86.
21. Wu Z, Jiang Z, Li T, Xie C, Zhao L, Yang J, et al. Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. Nat Commun. 2021;12:6501.
22. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, et al. De Novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021;373:655–62.
23. Okazaki Y, Nakano S, Toyoda A, Tamaki H. Long-Read-Resolved, Ecosystem-Wide exploration of nucleotide and structural microdiversity of lake bacterioplankton genomes. mSystems. 2022;7:e00433–22.
24. Lai S, Wang H, Bork P, Chen W-H, Zhao X-M. Long-read sequencing reveals extensive gut phageome structural variations driven by genetic exchange with bacterial hosts. Sci Adv. 2024;10:eadn3316.
25. Chen L, Zhao N, Cao J, Liu X, Xu J, Ma Y, et al. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. Nat Commun. 2022;13:3175.
26. Haro-Moreno JM, López-Pérez M, Molina-Pardines C, Rodriguez-Valera F. Large diversity in the O-chain biosynthetic cluster within populations of pelagibacterales. mBio. 2025;16:e03455–24.
27. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996.
28. Roda-Garcia JJ, Haro-Moreno JM, López-Pérez M. Evolutionary pathways for deep-sea adaptation in marine planktonic actinobacteriota. Front Microbiol. 2023;14:1159270.
29. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80.
30. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
31. Klähn S, Bolay P, Wright PR, Atilho RM, Brewer KI, Hagemann M, et al. A glutamine riboswitch is a key element for the regulation of glutamine synthetase in cyanobacteria. Nucleic Acids Res. 2018;46:10082–94.
32. Haro-Moreno JM, López-Pérez M, Alekseev A, Podoliak E, Kovalev K, Gordeliy V, et al. Flotillin-associated rhodopsin (FArhodopsin), a widespread paralog of proteorhodopsin in aquatic bacteria with streamlined genomes. mSystems. 2023;8:e0000823.
33. Kovalev K, Stetsenko A, Trunk F, Marin E, Haro-Moreno JM, Lamm GHU et al. Structural basis for no retinal binding in flotillin-associated rhodopsins. Structure. 2025 [cited 2025 Jul 30];0. Available from: https://www.cell.com/structure/abstract/S0969-2126(25)00222-9
34. Wang B, Qin W, Ren Y, Zhou X, Jung M-Y, Han P, et al. Expansion of thaumarchaeota habitat range is correlated with horizontal transfer of ATPase operons. ISME J. 2019;13:3067–79.
35. Hariharan VN, Yadav R, Thakur C, Singh A, Gopinathan R, Singh DP, et al. Cyclic di-GMP sensing histidine kinase PdtaS controls mycobacterial adaptation to carbon sources. FASEB J. 2021;35:e21475.
36. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45:D491–8.
37. Trgovec-Greif L, Hellinger H-J, Mainguy J, Pfundner A, Frishman D, Kiening M, et al. VOGDB—Database of virus orthologous groups. Viruses. 2024;16:1191.
38. Bzhalava Z, Hultin E, Dillner J. Extension of the viral ecology in humans using viral profile hidden Markov models. PLoS ONE. 2018;13:e0190938.
39. Zhao F, Lin X, Cai K, Jiang Y, Ni T, Chen Y, et al. Biochemical and structural characterization of the cyanophage-encoded phosphate-binding protein: implications for enhanced phosphate uptake of infected cyanobacteria. Environ Microbiol. 2022;24:3037–50.
40. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, et al. Charting the complexity of the marine Microbiome through Single-Cell genomics. Cell. 2019;179:1623–e163511.
41. Thompson LR, Haroon MF, Shibl AA, Cahill MJ, Ngugi DK, Williams GJ et al. Red Sea SAR11 and Prochlorococcus single-cell genomes reflect globally distributed pangenomes. Applied and Environmental Microbiology. 2019 [cited 2019 Dec 16];85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31028022
42. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. Genome diversity of marine phages recovered from mediterranean metagenomes: size matters. PLoS Genet. 2017;13:e1007018.
43. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature. 2024;630:493–500.
44. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302–9.
45. Zaragoza-Solas A, Rodriguez-Valera F, López-Pérez M. Metagenome Mining Reveals Hidden Genomic Diversity of Pelagimyophages in Aquatic Environments. Petersen J, editor. mSystems. 2020;5:e00905-19.
46. Zhang J-T, Yang F, Du K, Li W-F, Chen Y, Jiang Y-L, et al. Structure and assembly pattern of a freshwater short-tailed cyanophage Pam1. Structure. 2022;30:240–e2514.
47. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-Cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science. 2014;344:416–20.
48. Gonzaga A, Martin-Cuadrado AB, López-Pérez M, Mizuno CM, García-Heredia I, Kimes NE, et al. Polyclonality of concurrent natural populations of alteromonas Macleodii. Genome Biol Evol. 2012;4:1360–74.
49. Molina-Pardines C, Haro-Moreno JM, López-Pérez M. Phosphate-related genomic Islands as drivers of environmental adaptation in the streamlined marine alphaproteobacterial HIMB59. mSystems. 2023;8:e00898–23.
50. Riedel T, Dittmar T. A method detection limit for the analysis of natural organic matter via fourier transform ion cyclotron resonance mass spectrometry. Anal Chem. 2014;86:8376–82.
51. Iyer LM, Koonin EV, Aravind L. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Genome Biol. 2002;3:research00121.
52. Tanhua T, Hainbucher D, Schroeder K, Cardin V, Álvarez M, Civitarese G. The mediterranean sea system: A review and an introduction to the special issue. Ocean Sci. 2013;9:789–803.
53. Marston MF, Martiny JBH. Genomic diversification of marine cyanophages into stable ecotypes. Environ Microbiol. 2016;18:4240–53.
54. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 2005;13:278–84.
55. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
56. Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics and their ecological interpretation. Nat Reviews: Microbiol. 2015;13:133–46.
57. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. Long-term stability and red Queen-like strain dynamics in marine viruses. Nat Microbiol. 2020;5:265–71.
58. Hevroni G, Flores-Uribe J, Béjà O, Philosof A. Seasonal and diel patterns of abundance and activity of viruses in the red sea. Proc Natl Acad Sci. 2020;117:29738–47.
59. Liu R, Liu Y, Chen Y, Zhan Y, Zeng Q. Cyanobacterial viruses exhibit diurnal rhythms during infection. Proc Natl Acad Sci. 2019;116:14077–82.
60. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30:2114–20.
61. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. HybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016;32:1009–15.

62. Pan S, Zhao X-M, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. Bioinformatics. 2023;39:i21–9.
63. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
64. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023;20:1203–12.
65. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
66. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2019;36:1925–7.
67. 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.
68. Lowe TM, Eddy SR. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1996;25:955–64.
69. Seemann T, Prokka. Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
70. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun. 2020;11:2500.
71. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
72. Rodriguez-r LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. Peer J Preprints. 2016 [cited 2017 Jan 31]; Available from: https://github.com/lmrodriguezr/enveomics
73. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8:90.
74. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: A web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
75. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39:578–85.
76. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLoS Biol. 2023;21:e3002083.
77. Shang J, Peng C, Liao H, Tang X, Sun Y. PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data. Bioinf Adv. 2023;3:vbad101.
78. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with GeNomad. Nat Biotechnol. 2024;42:1303–12.
79. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
80. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. InStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021;39:727–36.
81. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
82. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. Bioinformatics. 2009;25:2078–9.
83. Eddy SR. Accelerated profile HMM searches. Pearson WR, editor. PLoS Comput Biol. 2011;7:e1002195.
84. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25:1335–7.
85. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-Coding RNA analysis using the Rfam database. Curr Protocols Bioinf. 2018;62:e51.
86. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the Marine Virosphere Using Metagenomics. Rocha EPC, editor. PLoS Genetics. 2013;9:e1003987.
87. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun. 2022;13:6968.
88. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

Publisher's note