ISME

www.nature.com/ismecomms

ARTICLE OPEN

Check for updates

Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space

Dominik Lücking[1], Coraline Mercier[1], Tomas Alarcón-Schumacher[1] and Susanne Erdmann[1]✉

Environmental virus metagenomes, commonly referred to as "viromes", are typically generated by physically separating virus-like particles (VLPs) from the microbial fraction based on their size and mass. However, most methods used to purify VLPs, enrich extracellular vesicles (EVs) and gene transfer agents (GTAs) simultaneously. Consequently, the sequence space traditionally referred to as a "virome" contains host-associated sequences, transported via EVs or GTAs. We therefore propose to call the genetic material isolated from size-fractionated (0.22 µm) and DNase-treated samples *protected environmental DNA (peDNA)*. This sequence space contains viral genomes, DNA transduced by viruses and DNA transported in EVs and GTAs. Since there is no genetic signature for peDNA transported in EVs, GTAs and virus particles, we rely on the successful removal of contaminating remaining cellular and free DNA when analyzing peDNA. Using marine samples collected from the North Sea, we generated a thoroughly purified peDNA dataset and developed a bioinformatic pipeline to determine the potential origin of the purified DNA. This pipeline was applied to our dataset as well as existing global marine "viromes". Through this pipeline, we identified known GTA and EV producers, as well as organisms with actively transducing proviruses as the source of the peDNA, thus confirming the reliability of our approach. Additionally, we identified novel and widespread EV producers, and found quantitative evidence suggesting that EV-mediated gene transfer plays a significant role in driving horizontal gene transfer (HGT) in the world's oceans.

ISME Communications; https://doi.org/10.1038/s43705-023-00317-6

INTRODUCTION

The presence of extracellular entities strongly shapes microbial communities. Particles of various origins mediate the transport of genetic material from one cell to another, thus playing a crucial role in horizontal gene transfer (HGT) [1, 2]. Most prominently, viruses are highly abundant and diverse drivers of ecological and evolutionary interactions within a community [3–6]. However, due to the limited culturability of their hosts, viruses often escape traditional culture-based approaches [7], leading to the development of culture-independent techniques to study their fundamental impact on microbial communities, global biogeochemical cycles and their effect on climate change. Similar to metagenomic studies, researchers have sequenced and analyzed the genetic content of the viral fraction on a community level, leading to the advent of viral metagenomics or "viromics" [8]. This approach traditionally relies on the physical, pre-sequencing separation of virus-like particles (VLPs) from microbial cells. Methods like sequential size filtration, ultracentrifugation, tangential flow filtration, and flow cytometry exploit the distinct physical properties of VLPs when compared to microbes [9–12]. Additionally, bioinformatic methods were developed to identify virus-like sequences among microbial sequences [13–16]. This led to the discovery of many diverse viruses, fulfilling crucial functions in their respective microbial community [17]. Interestingly, even after the most thorough removal of microbial cells, non-viral genes were shown to be present in "viromes" generated from many diverse environments [7, 18]. The contamination of "viromes" with microbial sequences originating from remaining intact cells and free extracellular DNA has been reported in several studies. Consequently, tools have been developed to estimate the proportion of true viral DNA in a given virome, using the abundance of reads with homologs in available prokaryotic databases or a specific set of microbial marker genes (e.g., 16S rRNA gene) [18–20]. These tools fall short of assessing the true degree of contamination of a "virome", because the abundance of prokaryotic-, non-virus-like genes is not necessary due to microbial contaminations, but can be the result of horizontal gene transfer processes.

Long before the development of modern "viromics", studies showed that viruses carry and distribute random microbial genes [21, 22] or specific "auxiliary metabolic genes" (AMGs) in addition to bona fide viral genes (e.g., genes necessary for particle assembly or viral genome replication), in a well-described process termed "transduction". Here, either genetic material adjacent to the integrated viral genome (specialized transduction) or random snippets of the host genome (general transduction) are packaged into the viral particles [1]. AMGs have been shown to fundamentally alter the metabolism of microbes by providing genes otherwise unavailable to their host [8, 23], further demonstrating the need for a good understanding of non-viral DNA in viromes.

Viromes are traditionally generated by separating VLPs from cells. However, methods that enrich VLPs by removing larger and heavier microbial cells also enrich entities similar in size and mass to VLPs. Most prominently, gene transfer agents (GTAs) and

[1]Max-Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany. ✉email: serdmann@mpi-bremen.de

Received: 25 May 2023 Revised: 28 September 2023 Accepted: 5 October 2023
Published online: 17 October 2023

extracellular vesicles (EVs, also referred to as membrane vesicles, MVs, or outer membrane vesicles, OMVs) are particles with similar physical properties and both have been shown to be involved in HGT, thus contributing to the presence of non-viral DNA in "viromes".

GTAs are particles transporting host DNA from one cell to another. They likely derived from defective prophages, and retained functional genes for the head and tail components of a head-tailed virus particle, including the genes for DNA packaging. Therefore, mass and size (40–60 µm) of GTA particles are very similar to head-tailed viruses, making it hard to differentiate them from viruses solely based on morphology. Notably and in contrast to true viruses, GTAs do not specifically package the GTA-producing gene cluster into the particle, but transport short segments of the host genome. Up to this date, several distinct gene clusters have been identified that produce GTAs [24–28].

Prokaryotic EVs are small (10–300 nm) spherical structures derived from the cell membrane [29]. EVs represent compartments that protect their cargo from degradation and are used for the transport of a variety of different components across the extracellular space. This includes the transport of nutrients, toxins, antigens, lipids, proteins, RNA, and DNA [30–36]. Recent studies showed an abundance of EVs in marine environments of up to 10^6 vesicles per milliliter [33], produced across diverse taxa. EVs, produced by highly abundant marine heterotrophs and autotrophs, such as Pelagibacter, Marinobacter, and Prochlorococcus, have been shown to transport fragments of chromosomal and plasmid DNA [37–41], thus contributing to the fraction of non-viral DNA within viromes.

In this study, we aimed to explore the non-viral sequence space of viromics datasets. First, we generated our own dataset, carefully avoiding possible contaminations. Then we categorized the sequences from this dataset and publicly available viromics datasets as virus- or non-virus-derived. Subsequently, we explore the non-virus-derived sequence space to detect the extent of non-viral DNA potentially being horizontally transferred between cells. We then identify the means of transport (GTA-, EV-, or virus-driven) for the sequences by linking the datasets to existing microbial metagenomes and genomes. We identify potential novel EV- and GTA producers and metagenomics-assembled genomes (MAGs) with an actively transducing virus. We propose using the term "protected extracellular DNA" (peDNA) for DNA sequence data derived from appropriately purified environmental fractions <0.2 µm, so far referred to as viromics datasets.

RESULTS AND DISCUSSION

Viromics datasets represent the sequence space of protected extracellular DNA (peDNA)

The majority of samples prepared for "viromics" include GTAs and EVs, in addition to virus particles. All three entities are small protein- and/or lipid-containing particles that can enclose cellular DNA [3] or were found to bind cellular DNA on their surface [42], thus inflating the sequence space that traditionally has been described as a "virome". In contrast to free extracellular DNA, DNA that is enclosed in or tightly associated with particles, or DNA that is tightly enclosed in protein/DNA or DNA/RNA complexes, is protected against degradation by extracellular nucleases occurring in the environment or nucleases used to clean samples from free extracellular DNA. Hence, we propose the term "protected extracellular DNA" (peDNA) to describe the entirety of DNA transported by viruses, GTAs and EVs (Fig. 1). We will use this term throughout this work.

Purification of environmental samples for the generation of peDNA datasets is essential to explore of the entire dataset

Previously, the percentage of 16S/18S rRNA-mapping reads was used as a proxy for host contamination in virome datasets [19].

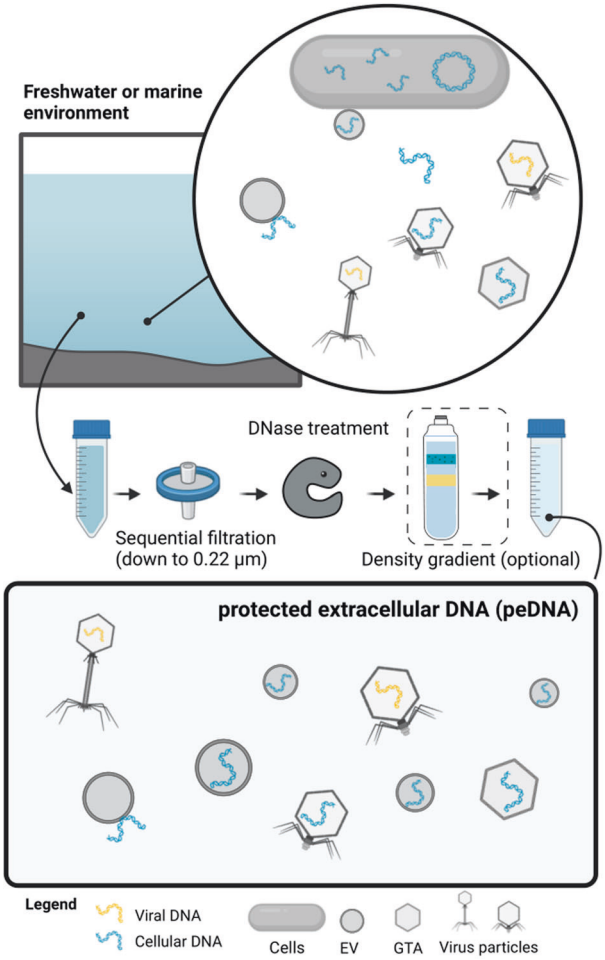

Fig. 1 Conceptual composition of protected extracellular DNA. The top panel depicts microbial entities present in a water body: microbial cells, viruses containing viral and microbial genetic material, gene transfer agents and extracellular vesicles containing host DNA. After size filtration (0.22 µm) and DNase treatment and, if applicable, purification via density gradients, microbial cells and free DNA are removed (middle panel). The remaining DNA makes up the sequence space of protected extracellular DNA, peDNA (bottom panel).

However, it has since been shown that GTAs and EVs enclose host DNA randomly, including 16S/18S rRNA genes [43]. We calculated SSU rRNA alignment rates for two highly purified samples: DNA extracted from virus isolates, purified by sequential plaque assays and 0.2 um size filtration [44] and DNA extracted from EVs purified from culture supernatants of *Prochlorococcus* [33] (Fig. 2). While the alignment rates were low for virus isolates (mean = 0.000437%), the percentage of 16S/18S-mapping reads was five orders of magnitude higher in DNA extracted from purified EVs (1.81%), even exceeding the mean alignment rate of publicly available microbial environmental metagenomes (mean = 0.078%, [19]). For these samples, Biller et al. confirmed the absence of microbial cells using electron microscopy. Thus, the presence of 16/18S rRNA-mapping reads neither proves or disproves contamination in peDNA samples. The only way to exclude contaminations with cellular DNA or extracellular free DNA is the rigorous purification of the sample before sequencing.

For this purpose, we generated a dataset from rigorously purified samples using several sequential filtration steps, DNase treatment and density gradient purification (Methods, Supplementary Fig. 1), resulting in cell-free samples containing virus-like particles, GTA particles and EVs. Subsequently, we compared the

SSU rRNA alignment rates of our dataset with one metagenomic peDNA dataset and two viromic datasets: Density gradient-purified EVs isolated from seawater samples ("Environmental EV enrichment") [33]; the "Tara Oceans virome" dataset [17], purified by size filtration and DNase treatment; and the "GDOCB virome" dataset [45]. GDOCB viromes are purified by flow cytometry. The process excludes free DNA and microbial cells with physical properties outside of the analyzed size spectrum, which makes it possible to assume that these viromes are free from contamination. Both datasets showed increased SSU rRNA alignment rates (Fig. 2), with some samples even exceeding the microbial metagenome alignment rate of 0.078%. Likewise, our dataset, while on average showing lower alignment rates (mean = 0.066%), contained samples exceeding that threshold. Lastly, Tara Oceans viromes mostly showed very low SSU rRNA alignment rates, with very few exceptions (mean = 0.031%). Overall, even thoroughly purified and confirmed contamination-free datasets show highly variable SSU rRNA alignment rates. We concluded that the majority of SSU rRNA hits in these datasets are enclosed in VLPs, GTA's or EVs rather than in contaminating microbial cells, and therefore included the datasets in the subsequent analysis.

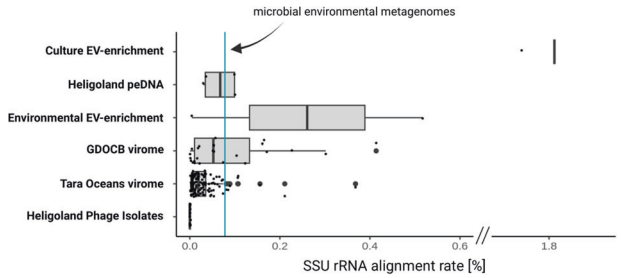

Fig. 2 Comparison of SSU rRNA alignment rates of diverse viromes and EV preparations. Each dot represents the percentage of reads aligning to either 16S or 18S rRNA genes. The cyan line indicates the average alignment rate for publicly available metagenomes from various environments [19]. "Culture EV enrichment": a cell-free preparation of EVs from *Prochlorococcus* cultures [33]. "Heligoland peDNA": dataset generated in this study from a highly purified (filtration, DNase treatment, gradient purification) <0.2 µm Heligoland water fraction. "Environmental EV enrichment": Density gradient-purified EVs isolated from seawater samples [33]. "GDOCB virome": <0.2 µm fraction enriched for VLPs by flow cytometry [42]. "Tara Oceans virome": <0.2 µm fraction purified by size filtrations and DNase treatment [17]. "Heligoland phage Isolates": DNA extracted from virus isolates, purified by 0.2 µm size filtration [41].

Separation of non-viral (nvpeDNA) from viral protected extracellular DNA (vpeDNA) indicates that EVs and GTAs could be very abundant entities in the ocean

The sequence space of protected extracellular DNA (peDNA) consists of virus genomes (viral protected extracellular DNA, vpeDNA) and non-viral, microbial DNA (non-viral extracellular DNA, nvpeDNA), deriving from transducing viruses, GTAs and EVs. nvpeDNA represents the sequence space that is potentially horizontally transferred between cells and therefore has major implications on the ecology and evolution of the organism in this environment and the environment itself.

In order to separate non-viral peDNA from viral peDNA, we developed a bioinformatic pipeline that, in brief, identifies virus sequences, separates those from non-viral sequences and calculates a non-viral to viral peDNA ratio in a given dataset (Fig. S1 and Methods). This pipeline was first applied to isolated viruses and purified EVs (Fig. 3A). As expected, vpeDNA made up >99% of the DNA of purified Heligoland phage isolates and nvpeDNA made up 98% of the DNA transported in purified EVs of a pure *Prochlorococcus* culture, verifying that the pipeline is reliably separating vpeDNA from nvpeDNA.

Consequently, we applied the pipeline to the entire Helgoland peDNA dataset and the other datasets that we verified earlier to be reasonably contamination-free peDNA datasets (Fig. 3A). Helgoland peDNA contained 39% non-viral reads. In the Tara Oceans Viromes, nvpeDNA made up, on average, 40% of all reads (105 samples). While viruses are considered the most abundant nucleic acid-containing biological entities in the ocean [3], these findings clearly indicate that EVs and GTAs, transferring cellular DNA, are likely very abundant entities as well. Surprisingly, in GDOCB viromes, the proportion of nvpeDNA to vpeDNA was even

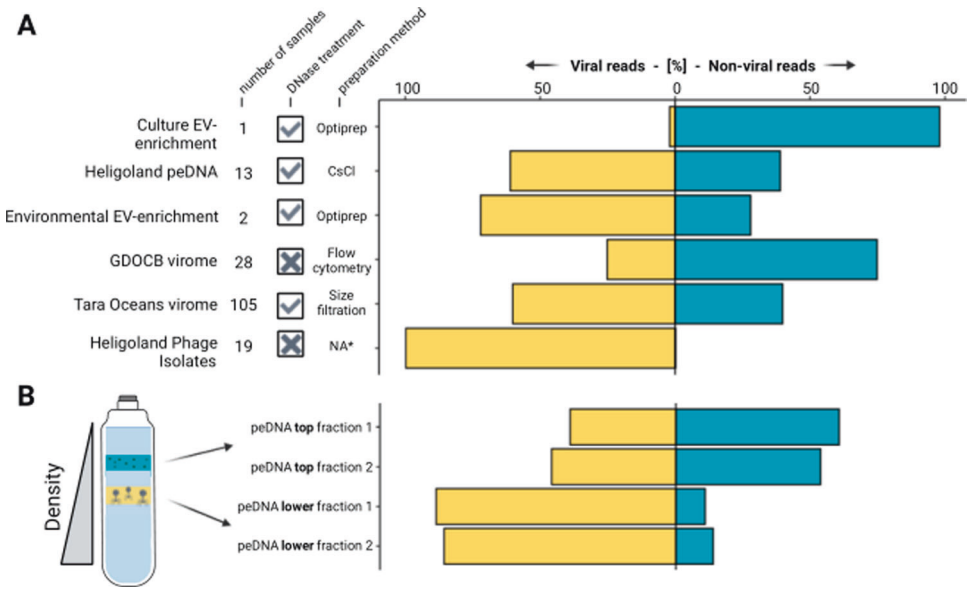

Fig. 3 Non-viral to viral peDNA ratios. A Non-viral to viral peDNA ratio across different studies. Each bar represents the percentage of read pairs mapping to contigs classified to be non-viral or viral for viromes, peDNA enrichments, EV enrichments and pure phage isolates. Sample size and purification methods are indicated for each sample. **B** Non-viral to viral peDNA ratio in different fractions of CsCl gradients. Left, schematic view of seawater samples running through CsCl gradients (adapted from [43]). Right, non-viral to viral peDNA ratio for top and bottom fractions in CsCl gradients for the Heligoland peDNA sample.

higher (75% nvpeDNA). However, this may be due to the comparably low sequencing depth, resulting in incomplete assembly and therefore hindering a reliable identification of virus contigs. Additionally, since the samples were not DNase treated, particles with the same size as the sorted VLPs might carry free DNA attached to the surface of the particles, therefore inflating nvpeDNA. In order to avoid any artificially introduced biases, this dataset was excluded from further downstream work.

We analyzed individual fractions of density gradients that were used to purify the Heligoland peDNA, because it was shown previously that density gradients can separate VLPs from EVs. VLPs, also including GTAs, were found to be more abundant in lower fractions of the gradients, while EV-like particles were more abundant in upper fractions [33, 46]. Indeed, in the upper gradient fractions of the Helgoland peDNA, nvpeDNA made up 60 and 54% of all reads in two biological replicates, while in the lower fraction nvpeDNA made up only 14 and 11% (Fig. 3B), confirming the previous observations. Thus the proportion of nvpeDNA is much higher in the upper fraction, that enriches EVs additionally to some VLPs, compared to the lower fraction, that enriches mainly virus particles and GTA's. This indicates that EVs are likely contributing significantly more to the nvpeDNA sequence space than GTA's and viruses.

Identifying the origin of non-viral protected extracellular DNA reveals that EVs could be the main driver of horizontal gene transfer in the oceans

For contamination-free peDNA samples, we consider three major possible origins of nvpeDNA: DNA transduced by viruses, DNA transported in GTA particles and DNA associated with EVs (Fig. 4A). While we cannot exclude that some of the nvpeDNA could originate from very stable protein/DNA or DNA/RNA complexes, we assume that the proportion of these complexes is rather small in comparison with DNA enclosed in particles.

It is inherently difficult to differentiate the origin of nvpeDNA based on sequence content. To this date, there are no reports on specific sequence signatures (e.g., marker genes) for DNA transported in EVs or GTAs, making DNA transported in EVs indistinguishable from DNA transported in GTA's or virus particles. Therefore, we developed a bioinformatic approach that tackles this differentiation from a different perspective. First, each read in the nvpeDNA fraction was linked to a given potential microbial host (Fig. S2). Then, the 20 most nvpeDNA recruiting MAGs per sample were selected. The main mechanism, which is most likely used to transport its DNA into the extracellular space was predicted, thus linking each read to either EV-, GTA- or transduction-associated transport. We confirmed that the abundance of these organisms (MAGs) in peDNA datasets does not correlate ($R^2 < 0.01$) with their abundance in the corresponding metagenomes (Fig. S3), and that none of the organisms identified are known to produce particularly small cells that could pass 0.2 µm filters. This additionally indicates that the high abundance of their genomic DNA in the nvpeDNA fraction is not due to cellular contamination with cells, but indeed the active transport of genomic DNA into the extracellular space via EVs, GTAs or virus particles. This approach was applied to nine Tara Oceans viromes which were linked to 2307 MAGs [44] and the entire Heligoland peDNA dataset which was linked to 457 MAGs sequenced from seawater coming from the same sampling station on Heligoland [47]. This yielded 200 MAGs (180 from Tara Oceans plus 20 from Heligoland), subsequently categorized as either GTA- or EV-producer or containing an actively transducing provirus. For details on the categorization approach, refer to the following sections, in brief: MAGs that contained an active (increased coverage in provirus region) provirus were labeled as "transducer". Similarly, the respective MAG was labeled as "GTA producer" if a complete or nearly complete GTA cluster could be identified. If neither an active provirus or a GTA cluster was identified, the MAG

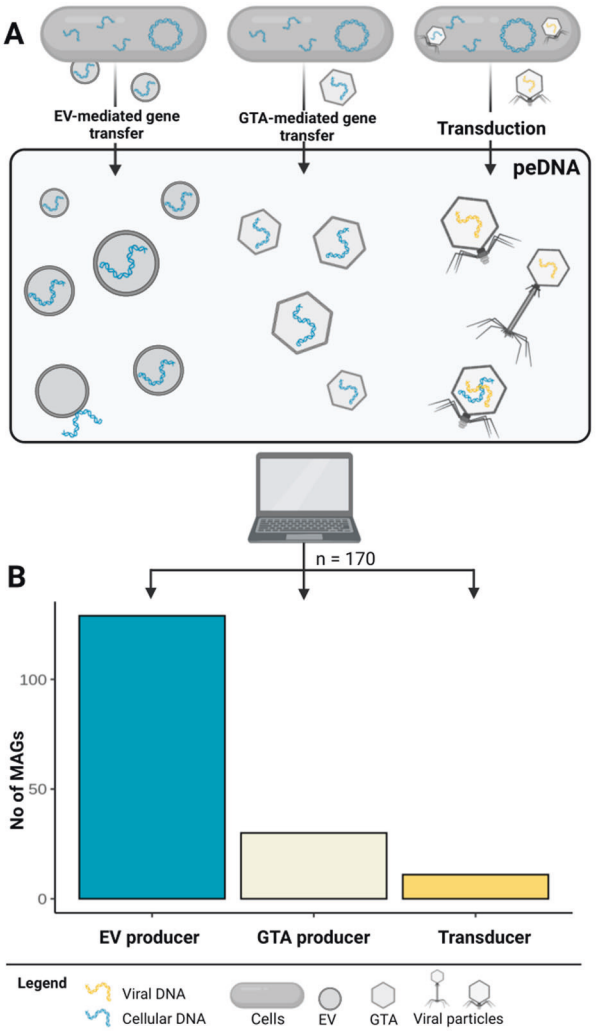

Fig. 4 Mechanisms of horizontal gene transfer that contribute to peDNA. A The origin of sequences comprising peDNA can be either EV-mediated or GTA-mediated gene transfer or via transduction. **B** Number of MAGs assigned to each mechanism. Of 200 analyzed metagenomic assembled MAGs, 170 could be assigned to predominantly use one of the three mechanisms in order to transport their genetic material into the extracellular space: 129 EV producers, 30 GTA producers and 11 genomes, with an actively transducing phage

was labeled as "EV producer" (Figs. S2 and S4). All labels were manually checked and verified by scrutinizing coverage plots (Fig. S5), prophage regions and GTA clusters.

Among the 200 top peDNA-recruiting MAGs 170 could be assigned unambiguously to one of the three categories (Fig. 4B). Most importantly, the majority was identified as EV producers, confirming that EVs that have been shown to be very abundant in the marine environment [38], are not just abundant entities but also significantly contribute to the peDNA sequence space and thereby are likely one of the most important drivers of horizontal gene transfer.

Identification of four novel GTA producers with RcGTA-like clusters

Of the 170 unambiguously assigned MAGs, 30 were identified to contain a functional (>10 GTA-associated genes, core genes present) GTA cluster. For most identified GTA producers, the presence of a GTA cluster has been described elsewhere: *Roseobacter sp.* ($n = 16$), *Sulfitobacter sp.* ($n = 8$) and *Roseovarius sp.* ($n = 1$) are known GTA producers of the order *Rhodobacterales*

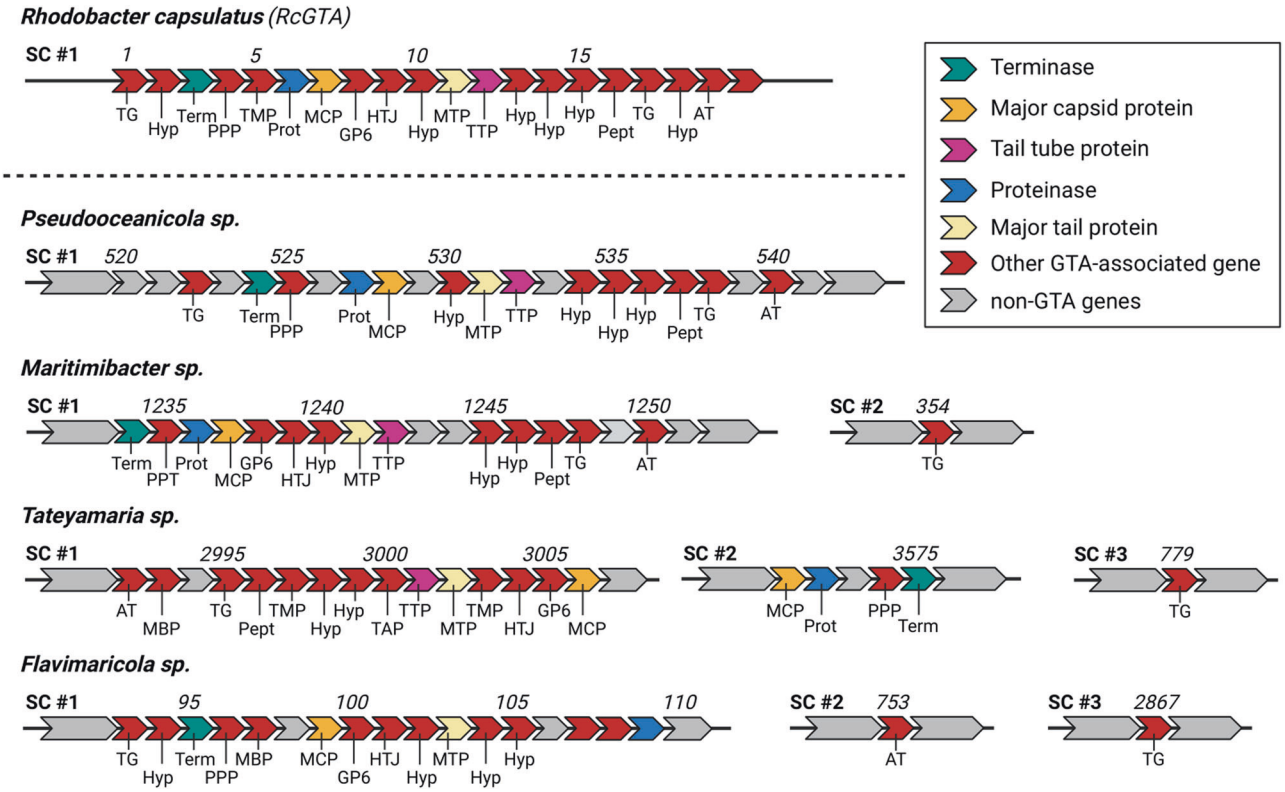

Fig. 5 Genome maps of four novel GTA producers. Organization of genes for the four identified, potentially novel GTAs, compared to the GTA cluster of *Rhodobacter capsulatus*. ORF number is given above the map, encoded protein function is indicated below. Non-GTA encoding genes are gray, GTA-associated genes are shown in red. In addition, core GTA encoding genes are colored accordingly. Encoding protein function is given below: AT acetyltransferase, MBP membrane bound protein, TG transglycosylase, Pept phage cell wall peptidase, TMP transmembrane protein, Hyp hypothetical, TAP tail assembly protein, TTP phage tail tube protein, MTP phage major tail protein, HTJ head-tail joining protein, GP6 gp6-like protein, MCP major capsid protein, Prot proteinase, PPP phage portal protein.

[48–50], indicating the efficiency of our approach. Additionally, we detected a functional GTA cluster in four more species: *Tateyamaria sp.*, *Pseudooceanicola sp.*, *Maritimibacter sp.* and *Flavimaricola sp.* all contained RcGTA homologs, including genes encoding the major capsid protein, a terminase, a proteinase and proteins associated with tail assembly (Fig. 5). The GTA cluster on the genome of the *Tateyamaria* species was distributed more widely over the genome in partial subclusters, as it has been described elsewhere [26]. As far as we know, these four species' clusters are not described elsewhere. We suggest that they are complete and functional because the core genes necessary to form GTA particles are present [51]; however, laboratory experiments are necessary to confirm their full functionality.

Only a few transducing proviruses could be identified with confidence in the peDNA sequence space

In order to label a MAG as containing an actively transducing provirus and therefore as a transducer, we relied on a combination of virus prediction tools, manual analysis of coverage plots and functional annotation of the proviral regions (see "Methods"—"Identification of potential transducers extracellular vesicle- and gene transfer agent producers"). The virus genome itself is present in all viral particles produced, in contrast to the transduced microbial DNA, which could be a randomly selected host DNA fragment (general transduction) or a specifically selected region (specialized transduction). In both cases, the coverage over the proviral region should be increased compared to the surrounding non-viral regions (Fig. 6C). Thus, only MAGs which showed the expected coverage profiles were labeled as transducers. Contrastingly, if a region recruited no reads from the peDNA fraction, the region was considered absent and the MAG was labeled as an EV producer (compare Fig. 6A *Haliea sp.* Station 158 SRF). In some (17 out of 200) cases, a clear assignment was not possible due to inconclusive coverage profiles or contradicting GTA and prophage predictions. These MAGs were labeled "unclear" and removed from further analysis. We identified 11 MAGs that carried an integrated and actively transducing provirus. Interestingly, a *Haliea sp.* with a proviral region was identified that recruited peDNA reads coming from one sampling station but none from the other. However, the non-viral part of the genome recruited high amounts of reads in both stations, albeit with lower coverage in the station where the provirus was absent (Fig. 6B). We hypothesize that there are two separate, distinct populations of the same *Haliea* species at the two stations: one, with the provirus integrated and actively transducing and one without the provirus. The fact that DNA from the population without the provirus is present in the peDNA fraction indicates that *Haliea sp.* transports its DNA into the extracellular space differently. In the absence of a GTA cluster, we hypothesize that *Haliea* sp. is capable of EV production and EV-mediated gene transfer. This demonstrates that transduction and EV-mediated gene transfer are not exclusive mechanisms of HGT but can overlap.

Identification of known and novel EV producers reveals that EV production is common amongst abundant marine bacteria

We identified 129 MAGs as EV producers. Most identified genera are known EV producers: *Marinobacter* ($n = 19$), *Alcanivorax* [18], *Flavobacteria* [9], *Thalassospira* [8], *Rheinheimera* [6], and *Polaribacter* [3], are known to produce high amounts of EVs [38, 52]. The fact that most organisms we labeled as EV producers are

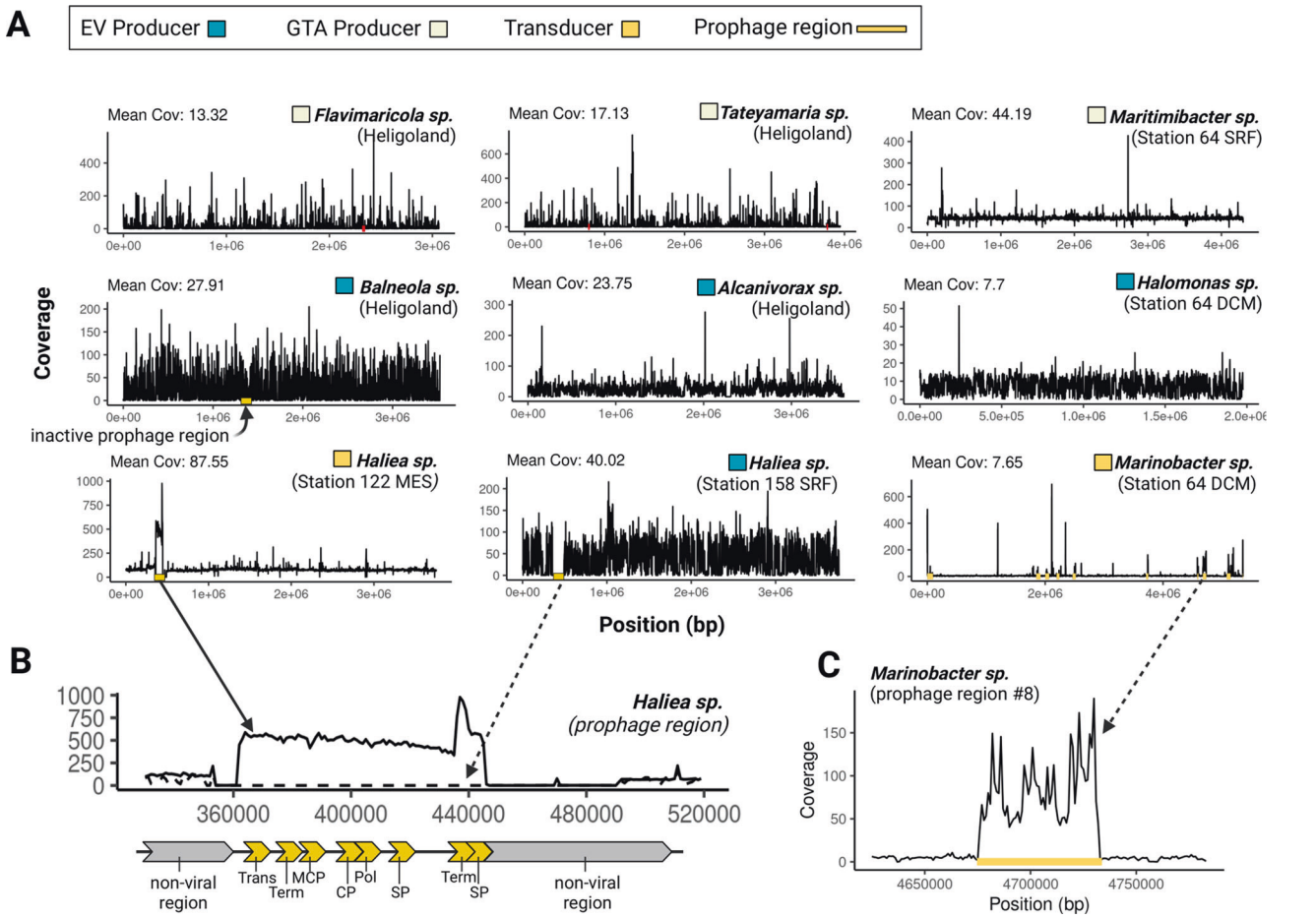

Fig. 6 Coverage plots of identified GTA- and EV-Producers and MAGs containing an actively transducing phage (transducer). A Coverage plots of 4 EV-producer, 2 genomes with an actively transducing phage and 3 GTA-producer (coverage blots of all analyzed MAGs see Fig. S12). Active prophage regions are indicated with yellow bars. **B** Genome map and detailed coverage plot of identified prophage region. On top, detailed coverage plot of the prophage region in two different samples. Coverage of reads from 122_MES (solid line) and Station 158_SRF (dotted) differs for this specific region. Below, schematic genome map for the genes identified in the prophage region and their approximate positions. Trans transposase, Term terminase, MCP major capsid protein, CP coat protein, Pol polymerase, SP shaft protein. **C** Coverage plot of an actively transducing phage. Close up of the coverage of prophage region #8 of *Marinobacter sp.* and surrounding non-viral regions.

already known EV producers again supports the efficiency of the approach. Interestingly, multiple MAGs of the genera *Haliea* ($n = 16$) and *Idiomarina* ($n = 8$) were identified to be EV producers. So far, EV production by either genus has not been described elsewhere and experimental confirmation is needed. However, this supports the observation by Biller et al., that many marine heterotrophs are actively producing EVs. While EVs could be used as a nutrient source [33, 38], or facilitate horizontal gene transfer (HGT) within microbial communities [37] potentially contributing to the evolution and adaptation of marine microbial populations, future research should aim to clarify the ecological and evolutionary role of EVs in the ocean.

The functional profile of peDNA links EV production to transposon induced gene mobilization

The functional profile of "viromes" or peDNA has been assessed previously [20, 53]. However, since these studies analyzed this sequence space from a virus perspective, mainly focusing on auxiliary metabolic genes (AMGs), they often excluded genes not directly associated with viral genomes. Here, we analyzed the functional profile of peDNA for each mode of transportation, EV- and GTA-mediated gene transfer, and transduction. In brief, each peDNA read that mapped to either an EV-, GTA producer or a transducer was classified into a cluster of orthologous groups (COG), and the resulting profile was then normalized with the

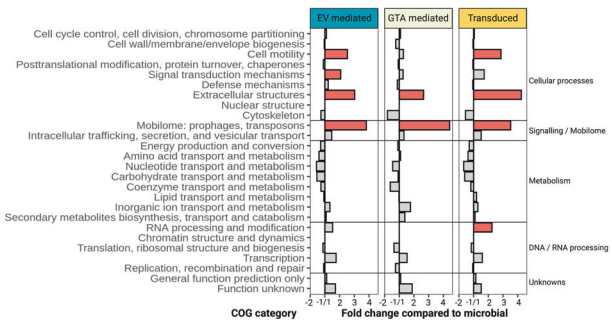

Fig. 7 Frequency of overrepresentation of clusters of orthologous groups (COG) categories for peDNA assigned to EV- and GTA producer and transducer. Bars represent the frequency of over-representation of genes assigned to each category, for each type of MAG (GTA producer, EV producer, transducer). Red bars indicate that a higher percentage of genes belonging to this category showed increased (two times standard deviation above mean) recruitment rates of peDNA reads.

profile of microbial reads (corresponding metagenome) mapping the respective sample (Fig. 7). The COG category "Mobilome (Prophages, transposons)" was overrepresented in all three groups. For transducers, this overrepresentation is mainly due to

an actively transducing provirus present in the peDNA. However, the overrepresentation of the COG category "Mobilome" in GTA and EV derived peDNA, is significant and we hypothesize, this is due to the presence of transposons. In fact, 76% of all reads assigned to the category "Mobilome" were assigned to a COG cluster containing the term "transposase" (Fig. S8). We suggest that transposon activity is also the reason for the overrepresentation of the other COG categories: "Extracellular structures", "Signal transduction mechanisms", and "Cell motility". All three categories are associated with the adaptation of the organism to a changing environment. Genes of these categories are often found on "genomic islands" (GIs), highly variable and mobile regions on the genome [54, 55]. At the same time, the occurrence of transposons on GIs is well-documented. Transposons have been shown to mobilize not only themselves but also adjacent "passenger genes", genes that are located in proximity to transposons and are therefore co-mobilized by transposons [56]. Evidence shows that environmental stressors increase the activity of transposons [57]. as well as the production of EVs [58], and the induction of proviruses [59] and GTAs [27]. Increased transposon activity increases the intracellular mobilization of genes surrounding transposons and therefore could lead to an increased uptake into EVs, GTAs or virus particles. Our data suggest that these two stress-induced mobilization mechanisms may be linked in a way that enhances the community's adaptability to the environment, by increasing genetic transfer between individual cells. However, whether the transposons and associated genes are indeed transferred in their complete active form or as fragments will require experimental evidence.

CONCLUSION AND OUTLOOK

In this study, we propose the term "protected extracellular DNA" (peDNA) to refer to genetic material obtained from size-filtered and DNase-treated samples, thereby accommodating non-viral, EV- or GTA-transported DNA in that sequence space. So far, there is no known sequence marker to distinguish horizontally transferred DNA from cellular DNA, therefore the removal of contaminating cells and free DNA is crucial when analyzing peDNA samples. The level of contamination however, should not be assessed using the presence of ribosomal subunit-mapping reads, since EVs have been shown to transport 16/18S rRNA genes.

In our study, we analyzed a carefully purified marine sample of peDNA and existing global marine datasets. We were able to link peDNA sequences to potential hosts and identify their primary mode of DNA transfer. Among the identified GTA and EV producers, most were shown to produce the respective particles in previous studies, confirming the validity of our approach, however, new potential GTA and EV producers were also identified. Overall, EV-mediated gene transfer was the most common mechanism and we hypothesize that EVs are a main driver of HGT in the ocean. Lastly, our findings suggest that EV-mediated gene transfer and transposon induced gene mobilization potentially work together and enhance the ability of microbial communities to adapt to a changing environment. Given the considerable ecological stressors imposed by climate change, comprehensive investigations into the role of EVs, GTAs and viruses for HGT is essential to understand genetic adaptability in marine microbes. This study highlights the need for further research into HGT mechanisms, and peDNA in general, since the community composition and function of marine microbes, and therefore the global oceans, is strongly shaped by the abundance of protected viral and non-viral extracellular DNA.

METHODS

Sampling and filtration

A visual overview of the sampling and filtration methods is given in Fig. S6.

Three seawater samples (G, H, I) of 100 liters were collected off the shore of Helgoland at the sampling station "Kabeltonne" (54°11'02.4'N 7°53'49.2'E). Each sample was sequentially filtered through 10, 3, 0.8, 0.45 and 0.22 µm (polyethersulfone filters, Merck Millipore, Burlington, MA, US). Filters were immediately stored at −20 °C for later DNA extraction. Flow through of the 0.22 µm filters was subsequently concentrated using tangential flow filtration with a 100 kDa cassette (Sartorius Stedim). The concentrated samples were stored at 4 °C, until further concentration down to 0.5 ml, using Amicon filter centrifugation (1 MDa AmiCon tube filters, 2500 × *g*). Finally, the concentrated sample was diluted with purified seawater (flow-through from the tangential flow filtration) to 2 ml. Two aliquots were created for each sample á 0.5 ml, one treated with DNase before gradient purification, one treated afterwards (see "Purification of peDNA samples").

Purification of peDNA samples

In order to remove free DNA, half of the samples were incubated with 100 U/ml DNase I (Thermo Scientific), supplied with the buffer provided with the enzyme, at 37 °C for 10 min, while the other half were DNase-treated after density gradient purification. EDTA was added to a final concentration of 5 mM and the enzyme was deactivated at 75 °C for 10 min. CsCl density gradients were prepared as following: Five CsCl solutions were prepared with 25, 30, 35, 40 and 60% CsCl solved in artificial sea water (480 mM NaCl, 27 mM $MgCl_2$, 2.8 mM $MgSO_4$, 9 mM KCl, 6 mM $NaHCO_3$, 10 mM $CaCl_2$) [60]. For each gradient, 1 ml of each solution was carefully layered on top of each other and stored at 4 °C overnight in order to establish the gradient. 0.5 ml of sample was carefully placed on top of the samples, before ultracentrifugation (20 h, 38,000 × *g*, 4 °C). For each gradient, individual 0.5 ml fractions were carefully extracted and incubated with 40% PEG 6000 (final concentration 10%) overnight at 4 °C. Particles were precipitated by centrifugation (13,000 × *g*, 45 min, 4 °C) and particle pellets dissolved in 1 ml artificial sea water. The other half of the samples were DNase-treated at this time point. A detailed overview of the samples is given in Table S1 and a visual overview in Fig. S7.

DNA extraction

Frozen polycarbonate filters (3, 0.8, 0.45 and 0.22 µm) were placed in a 50 ml tube together with 13.5 ml of extraction buffer (100 mM Tris-HCl pH 8.0, 100 mM EDTA pH 8.0, 100 mM Na-Phosphate buffer pH 8.0, 1.5 M NaCl, 1% CTAB). peDNA samples were processed directly. DNA was extracted as described elsewhere [61]. In brief, samples were treated with 10 mg/ml Proteinase K and incubated at 37 °C for 30 min on a shaker. Then, 1/10 vol of 20% SDS was added, before incubating again at 65 °C for 2 h on a shaker. After centrifugation (53,000 × *g*, 10 min, RT), the samples were transferred into a new tube and 1 vol of chloroform/isoamylalcohol was added and samples were thoroughly mixed, before centrifuging at 4000 × *g* for 20 min at RT. The aqueous phase upper phase was collected and transferred into a new tube. This step was repeated until no protein/polysaccharide layer was visible. DNA was then precipitated by adding 0.6 vol isopropanol and incubation for 1 h at room temperature. DNA was pelleted at 53,000 × *g* for 10 min at RT washed with 1 ml cold (4 °C) 80% ethanol and resuspended in 60 µl 1x TE buffer overnight. DNA concentration was assessed using a spectrophotometer (DS-11 FX+ by DeNovix®, Wilmington, DE, US), see Table S1.

Sequencing

DNA samples were pooled according to Table S1, assuring enough DNA content per sample for successful sequencing. Library preparation (FS DNA Library, NEBNext® Ultra™, Ipswitch, MA, US) and sequencing (Illumina HiSeq2500 by Illumina, San Diego, CA, US, 2 × 250 bp for peDNA samples and Illumina HiSeq3000, 2 × 150 bp for Filter DNA) was performed at the Max Planck Genome Centre Cologne (MP-GC).

Read trimming and assembly

Paired-end reads from Heligoland EV enrichments were trimmed using Trimmomatic [Bolger 2014] in paired-end mode, with the parameters LEADING:8 TRAILING:8 SLIDINGWINDOW:5:24 MINLEN:50. Paired-end reads from EV Enrichments [33] and paired-end reads from GDOCB [45] were trimmed using bbduk.sh, part of the BBTools suite [62] with the following parameters: bbduk.sh qtrim = rl trimq = 20 maq = 20 minlen = 30 ordered t = 8 ref = adapters.fa, where adapters.fa were fasta files containing adapters identified to be present in the reads using FastQC [63]. Reads from Heligoland EV enrichments were assembled using metaSPAdes [64] with default parameters.

Handling of external data

External datasets were downloaded from public servers. An overview of external datasets used in this study, with SRR, ERS and DRR accessions, is given in Table S1. Reads from Tara Ocean viromes [17] were already trimmed. Reads from EV enrichments [33] and GDOCB viromes [45] were assembled using metaSPAdes [64] with default parameters. For Tara Oceans viromes, assembled contigs were downloaded from https://www.ebi.ac.uk/. Tara Oceans MAGs were published elsewhere by Tully et al. [65], accessions are listed in Table S1.

Calculation of SSU alignment rates

SSU alignment rates were calculated using ViromeQC [19], which maps input reads against 16S and 16S rRNA subunits. This was done for all Tara Ocean viromes, Heligoland peDNA, EV enrichments (see Table S1 for an overview of all samples used).

Calculation of the percentage of non-viral associated reads

The percentage of non-viral peDNA in viromic samples was calculated with a pipeline of bioinformatic tools. An overview is given in Fig. S1. Paired-end, trimmed input reads were assembled. Contigs shorter than 2000 bp were removed from downstream analysis. Then, contigs were subject to two viral-prediction steps: viral sequences were predicted (1) using a combination of VirSorter2 and CheckV as described previously [66] and (2) DeepVirFinder [16]. The results of both steps were summarized using a custom script (https://github.com/dluecking/peDNA_custom_scripts/) and each contig was labeled as either "viral" or "non-viral". Then, the initial input reads were mapped against labeled contigs, using bbmap.sh, part of the BBTools suite [62] with default parameters. Then the number of non-viral-contig mapping reads was divided by the number of total reads mapping against viral or non-viral contigs. This ratio of non-viral/viral reads is referred to as "percentage of non-viral to viral associated reads" or "nvpeDNA/peDNA read ratio" in this study.

Identification of potential transducers, extracellular vesicle- and gene transfer agent producers

In order to identify potential EV producers, GTA producers and MAGs with an actively transducing virus, a second bioinformatic pipeline was developed (see Fig. S5). First, MAGs were filtered by removal of MAGs shorter than 100,000 bp. Then, reads from the corresponding viromes/peDNA samples were mapped against the MAGs, using bbmap.sh with default parameters. For each sample (in total 9 samples from Tara Oceans, 1 combined Heligoland sample) the 20 most recruiting MAGs were selected for further downstream work. VirSorter2 (default parameters) was used in order to predict potential integrated proviruses. GTA clusters were predicted by searching for homologs of proteins of known GTA clusters using *diamond blastp* with default parameters (evalue $\leq 10^{-5}$, pident > 50%). Finally, a customscript (https://github.com/dluecking/peDNA_custom_scripts/) summarized the results and an automated label was given. Additionally, each label was manually curated and each MAG was labeled as either EV producer, GTA producer or an organism with an actively transducing virus (see Figs. S5 and S6).

Annotation of GTA producers and viral regions

Open reading frames were predicted using prodigal with the metagenome flag (*prodigal -i < fasta-file > -d < genes-out > -a <protein-out > -p meta*). Each ORF was then annotated using the InterProScan API (https://github.com/ebi-wp/webservice-clients-generator) with default parameters [67] and additionally checked manually. For the prophage region, shown in Fig. 6, the DNA sequence was submitted and annotated in PHASTER [68, 69]. Genome maps and presence-absence plots were generated using ggplot [70] and BioRender.com.

Identification of cluster of orthologous groups

In order to assess the functional profile of peDNA reads, each read was mapped to the 170 MAGs for which the primary transport mechanism was identified using bbmap, part of the BBTools suite [62] with minid = 95 and otherwise default parameters. This resulted in three sets of reads: EV-mediated, GTA-mediated, VLP-mediated. For each read partial ORFs were predicted using FragGeneScan [71] with the parameters -complete = 0, -train = illumina_5 and otherwise default parameters. The partial ORFs were then blasted against the COG database [72] using the diamond tool set [73] with the following parameters: *-f 6 –max-target-seqs 1 –query-cover 80 –subject-cover 10*. Each read was then assigned a COG cluster and consequently a COG category. For each category, the relative abundance was calculated using:

$$freq_{cat} = \frac{n_{hc}}{n_{tot}}$$

where *n_hc* is the number of reads in category *cat* that show high coverage and *n_tot* is the total number of reads assigned by this label. The same procedure was done for metagenome reads of the corresponding metagenome samples (see Supplementary Table S1—Sample overview). The fold change between categories was calculated pairwise with the formula:

$$fold\ change \frac{e_{cat-label} = freq_{cat-label}}{fre} q_{cat-microbial}$$

where *cat* refers to a specific COG category, label to either EV, GTA or virally transduced and microbial to the microbial counterpart of that sample. For visualization reasons, fold changes smaller 1 were calculated with the reversed formula:

$$fold\ change_{cat-label} = fre \frac{q_{cat-microbial}}{fre} q_{cat-label}$$

Fold changes between −1 and 1 are therefore not possible and this area is excluded from the plot.

In order to get a detailed resolution of EV-mediated reads belonging to COG category × -Mobilome a subset of 10 million reads for each sample (9 Tara Ocean stations and 1 sample from heligoland) were selected at random. From these, ~11 M protein fragments were predicted and blasted against nr with an *e*-value threshold of 10^{-5}, query coverage >80%, subject coverage >10%, resulting in a total of 34,826 assigned reads. The results were visualized in R.

Identification of transposable elements on EV producing genomes

Putative EV producing MAGs were annotated using DRAM (Distilled and Refined Annotation of Metabolism, https://github.com/WrightonLabCSU/DRAM), [74]. Transposases were then detected using the regex term "IS\\d\\+*|Tn\\d\\+*|attTn\\d\\+*|transposase|Transposase" among all annotations found.

Coverage plots, genome maps and schematic figures

Coverage plots of potential transducers, EV- and GTA producers were created using the R package ggplot2 [70]. Genome maps of potential GTA producers were created using the R package gggenes (https://github.com/wilkox/gggenes). Schematic genome maps and additional elements in figures were created with BioRender.com.

DATA AVAILABILITY

Heligoland metagenome and peDNA reads are available at the European Nucleotide Archive (ENA) under BioProject PRJEB60526. Custom scripts are available at https://github.com/dluecking/peDNA_custom_scripts/.

REFERENCES

1. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015;16:472–82.
2. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022;20:206–18.
3. Fuhrman JA. Marine viruses and their biogeochemical and ecological effects. Nature. 1999;399:541–8.
4. Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev. 2000;64:69–114.
5. Suttle CA. Viruses in the sea. Nature. 2005;437:356–61.
6. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. Exploring the vast diversity of marine viruses. Oceanography. 2007;20:135–9.
7. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3:504–10.
8. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, et al. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci. 2002;99:14250–5.

9. Li WKW, Dickie PM. Growth of bacteria in seawater filtered through 0.2 µm Nuclepore membranes: implications for dilution experiments. Mar Ecol Prog Ser. 1985;26:245–52.
10. Van der Pol E, Hoekstra AG, Sturk A, Otto C, Van leeuwen TG, Nieuwland R. Optical and non-optical methods for detection and characterization of microparticles and exosomes. J Thromb Haemost. 2010;8:2596–607.
11. McNamara RP, Dittmer DP. Modern techniques for the isolation of extracellular vesicles and viruses. J Neuroimmune Pharmacol. 2020;15:459–72.
12. Hillebrandt N, Vormittag P, Bluthardt N, Dietrich A, Hubbuch J. Integrated process for capture and purification of virus-like particles: enhancing process performance by cross-flow filtration. Front Bioeng Biotechnol. 2020;8. https://doi.org/10.3389/fbioe.2020.00489.
13. Duhaime MB, Sullivan MB. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology. 2012;434:181–6.
14. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife. 2015;4:e08490.
15. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.
16. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8:64–77.
17. 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;177:1109–23.e14.
18. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, et al. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS ONE. 2012;7:e33641.
19. Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD, et al. Detecting contamination in viromes using ViromeQC. Nat Biotechnol. 2019;37:1408–12.
20. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. 2013;3:130160.
21. Jiang S, Paul J, et al. Gene transfer by transduction in the marine environment. App Environ Microbiol. 1998;64:2780–7.
22. Zinder ND, Lederberg J. Genetic exchange in salmonella. J Bacteriol. 1952;64:679–99.
23. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci USA. 2004;101:11013–8.
24. Matson EG, Thompson MG, Humphrey SB, Zuerner RL, Stanton TB. Identification of genes of VSH-1, a prophage-like gene transfer agent of Brachyspira hyodysenteriae. J Bacteriol. 2005;187:5885–92.
25. Krupovic M, Forterre P, Bamford DH. Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol. 2010;397:144–60.
26. Lang AS, Zhaxybayeva O, Beatty JT. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol. 2012;10:472–82.
27. Québatte M, Christen M, Harms A, Körner J, Christen B, Dehio C. Gene transfer agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. Cell Syst. 2017;4:611–21.e6.
28. Tomasch J, Wang H, Hall ATK, Patzelt D, Preusse M, Petersen J, et al. Packaging of Dinoroseobacter shibae DNA into gene transfer agent particles is not random. Genome Biol Evol. 2018;10:359–69.
29. Deatherage BL, Cookson BT. Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life. Infect Immun. 2012;80:1948–57.
30. Dorward DW, Garon CF, Judd RC. Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae. J Bacteriol. 1989;171:2499–505.
31. Kadurugamuwa JL, Beveridge TJ. Bacteriolytic effect of membrane vesicles from Pseudomonas aeruginosa on other bacteria including pathogens: conceptually new antibiotics. J Bacteriol. 1996;178:2767–74.
32. Yaron S, Kolling GL, Simon L, Matthews KR. Vesicle-mediated transfer of virulence genes from Escherichia coli O157:H7 to other enteric bacteria. Appl Environ Microbiol. 2000;66:4414–20.
33. Biller SJ, Schubotz F, Roggensack SE, Thompson AW, Summons RE, Chisholm SW. Bacterial vesicles in marine ecosystems. Science. 2014;343:183–6.
34. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol. 2015;13:13–27.
35. Schwechheimer C, Kuehn MJ. Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions. Nat Rev Microbiol. 2015;13:605–19.
36. Schatz D, Rosenwasser S, Malitsky S, Wolf SG, Feldmesser E, Vardi A. Communication via extracellular vesicles enhances viral infection of a cosmopolitan alga. Nat Microbiol. 2017;2:1485–92.
37. Domingues S, Nielsen KM. Membrane vesicles and horizontal gene transfer in prokaryotes. Curr Opin Microbiol. 2017;38:16–21.
38. Biller SJ, Coe A, Arellano AA, Dooley K, Gong JS, Yeager EA, et al. Environmental and taxonomic drivers of bacterial extracellular vesicle production in marine ecosystems. bioRxiv. 2022. https://doi.org/10.1101/2022.01.18.476865v2.
39. Biller SJ, McDaniel LD, Breitbart M, Rogers E, Paul JH, Chisholm SW. Membrane vesicles in sea water: heterogeneous DNA content and implications for viral abundance estimates. ISME J. 2017;11:394–404.
40. Hackl T, Laurenceau R, Ankenbrand MJ, Bliem C, Cariani Z, Thomas E, et al. Novel integrative elements and genomic plasticity in ocean ecosystems. Cell. 2023;186:47–62.e16.
41. Linney MD, Eppley JM, Romano AE, Luo E, DeLong EF, Karl DM. Microbial sources of exocellular DNA in the ocean. Appl Environ Microbiol. 2022;88:e02093–21.
42. Bitto NJ, Chapman R, Pidot S, Costin A, Lo C, Choi J, et al. Bacterial membrane vesicles transport their DNA cargo into host cells. Sci Rep. 2017;7:7072.
43. Ricci V, Carcione D, Messina S, Colombo GI, D'Alessandra Y. Circulating 16S RNA in biofluids: extracellular vesicles as mirrors of human microbiome? Int J Mol Sci. 2020;21:8959.
44. Bartlau N, Wichels A, Krohne G, Adriaenssens EM, Heins A, Fuchs BM, et al. Highly diverse flavobacterial phages isolated from North Sea spring blooms. ISME J. 2022;16:555–68.
45. De Corte D, Martínez JM, Cretoiu MS, Takaki Y, Nunoura T, Sintes E, et al. Viral communities in the global deep ocean conveyor belt assessed by targeted viromics. Front Microbiol. 2019;10:1801.
46. Linney MD, Schvarcz CR, Steward GF, DeLong EF, Karl DM. A method for characterizing dissolved DNA and its application to the North Pacific Subtropical Gyre. Limnol Oceanogr Methods. 2021;19:210–21.
47. Orellana LH, Ben Francis T, Krüger K, Teeling H, Müller MC, Fuchs BM, et al. Niche differentiation among annually recurrent coastal Marine Group II Euryarchaeota. ISME J. 2019;13:3024–36.
48. Zhao Y, Wang K, Budinoff C, Buchan A, Lang A, Jiao N, et al. Gene transfer agent (GTA) genes reveal diverse and dynamic Roseobacter and Rhodobacter populations in the Chesapeake Bay. ISME J. 2009;3:364–73.
49. Ankrah NYD, Lane T, Budinoff CR, Hadden MK, Buchan A. Draft genome sequence of Sulfitobacter sp. CB2047, a member of the roseobacter clade of marine bacteria, isolated from an emiliania huxleyi bloom. Genome Announc. 2014;2:e01125–14.
50. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. High frequency of horizontal gene transfer in the oceans. Science. 2010;330:50.
51. Sherlock D, Leong JX, Fogg PCM. Identification of the first gene transfer agent (GTA) small terminase in rhodobacter capsulatus and its role in GTA production and packaging of DNA. J Virol. 2019;93:e01328–19.
52. Schuster AK. Production of extracellular DNA (eDNA) of the γ-proteobacterium Rheinheimera sp. F8 in biofilms. Germany: Technische Universitaet Berlin; 2017.
53. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, et al. Functional metagenomic profiling of nine biomes. Nature. 2008;452:629–32.
54. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, et al. Genomic islands and the ecology and evolution of Prochlorococcus. Science. 2006;311:1768–70.
55. Thompson AW, Huang K, Saito MA, Chisholm SW. Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability. ISME J. 2011;5:1580–94.
56. Pál C, Papp B. From passengers to drivers. Mob Genet Elem. 2013;3:e23617.
57. Capy P, Gasperi G, Biémont C, Bazin C. Stress and transposable elements: co-evolution or useful parasites? Heredity. 2000;85:101–6.
58. MacDonald IA, Kuehn MJ. Stress-induced outer membrane vesicle production by Pseudomonas aeruginosa. J Bacteriol. 2013;195:2971–81.
59. Choi J, Kotay SM, Goel R. Various physico-chemical stress factors cause prophage induction in Nitrosospira multiformis 25196–an ammonia oxidizing bacteria. Water Res. 2010;44:4550–8.
60. Carini P, Steindler L, Beszteri S, Giovannoni SJ. Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. ISME J. 2013;7:592–602.
61. Zhou J, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. Appl Environ Microbiol. 1996;62:316–22.
62. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Report No.: LBNL-7065E. Berkeley, CA (United States): Lawrence Berkeley National Lab. (LBNL); 2014. https://www.osti.gov/biblio/1241166.
63. Andrews S, et al. FastQC: a quality control tool for high throughput sequence data. Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute; 2010.
64. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
65. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5:170203.

66. Guo J, Vik D, Pratama AA, Roux S, Sullivan M. Viral sequence identification SOP with VirSorter2. 2021. https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-bwm5pc86.
67. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
68. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44:W16–21.
69. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39:W347–52.
70. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016. https://ggplot2.tidyverse.org.
71. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38:e191.
72. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49:D274–81.
73. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
74. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48:8883–900.

ACKNOWLEDGEMENTS

We thank the sampling crew of the Heligoland sampling station "Kabeltonne" for their help obtaining the sea water samples and the AWI research station on heligoland for providing laboratory workspace. We are grateful to Luis "Coto" Orellana and Isabella Maria Wilkie for providing access to their data. Finally, we thank A. Probst, A. Zayed and B. Fuchs for insights, discussions and feedback. We thank Daniela Thies and Ingrid Kunze (MPI for Marine Microbiology, Bremen, Germany) for assistance with some of the experiments. Finally, we want to thank the Max-Planck-Institute for Marine Microbiology and the Max-Planck-Society for continuous support.

AUTHORS CONTRIBUTIONS

DL performed the majority of the experimental laboratory and bioinformatic work. CM and TAS supported general laboratory work, sampling and purification of sea water samples. SE conceived and led the study. DL and SE performed the primary writing of the manuscript. All authors participated in the analysis and interpretation of the data and contributed to the writing of the manuscript.

FUNDING

This study was funded by the Max Planck Society (Munich, Germany) as part of the Max Planck Research Group Archaeal Virology to SE. Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-023-00317-6.

Correspondence and requests for materials should be addressed to Susanne Erdmann.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.