nature microbiology

Article

https://doi.org/10.1038/s41564-025-02167-5

Single-particle genomics uncovers abundant non-canonical marine viruses from nanolitre volumes

Received: 14 March 2025

Accepted: 29 September 2025

Published online: 5 November 2025

Check for updates

Alaina R. Weinheimer[1], Julia M. Brown[1], Brian Thompson[1], Greta Leonaviciene[2,3], Vaidotas Kiseliovas[3], Simonas Jocys[3], Jacob Munson-McGee[1], Gregory Gavelis[1], Corianna Mascena[1], Linas Mazutis[2], Nicole J. Poulton[1], Rapolas Zilionis[2,3] & Ramunas Stepanauskas[1]

Viruses and other extracellular genetic elements play essential roles in marine communities. However, methods to capture their full diversity remain limited by the constraints of bulk sequencing assemblers or pre-sorting throughput. Here we introduce environmental micro-compartment genomics (EMCG), which vastly improves the throughput and efficiency of single-particle genomic sequencing obtained from nanolitre volumes by compartmentalizing particles of a sample into picolitre-sized, semi-permeable capsules for in-capsule DNA amplification and barcoding. From 300 nanolitres of seawater, EMCG obtained genomic sequences of 2,037 particles. The microbiome composition agreed with other methods, and the virus-like assembly lengths indicated that most were near complete. Many viral assemblies belonged to the *Naomiviridae*, lacked metagenomic representation and aligned to outlier contigs of abundant, putative host lineages, suggesting their use of non-canonical DNA and overlooked ecological importance. This approach provides opportunities for high-throughput, quantitative and cost-effective genome analyses of individual cells and extracellular particles across complex microbiomes.

The ocean's biological diversity, biomass and biogeochemical processes are dominated by microscopic life, with most evading cultivation under laboratory conditions[1–3]. In the past two decades, cultivation-independent research tools, such as amplicon sequencing, metagenomics and single-cell genomics, have made major progress in deciphering the taxonomic composition and coding potential of the unicellular bacteria, archaea and eukaryotes inhabiting marine environments[4–7]. Less is known about the abundance, genome content and roles of various types of extracellular DNA, the concentration of which in seawater is comparable to that of DNA inside all planktonic cells[8–10].

Our current knowledge of environmental extracellular DNA suggests two general categories: (1) free DNA released from dead as well as metabolically active cells[11] and (2) protected genomic fragments within extracellular membrane vesicles or distinct genetic elements such as viruses and tycheposons, which exist in a continuum of selfishness and autonomy[12–15]. The metagenomic research of marine extracellular DNA has primarily focused on viruses, which outnumber microbial cells by roughly tenfold[16], revealing an extraordinary degree of genetic diversity and novelty[17–19]. However, due to the extensive microdiversity and lateral gene transfer among marine viruses, their metagenome-assembled genomes (MAGs) are often fragmented and incomplete, and potential analytical blind spots remain poorly understood[20–22]. Furthermore, the scarcity of viable alternatives hinders the validation and improvement of viral MAG quality. Genomic sequencing of individual viral particles and their de novo assembly into single amplified genomes (SAGs) may

[1]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA. [2]Life Sciences Center, Vilnius University, Vilnius, Lithuania. [3]Atrandi Biosciences, Vilnius, Lithuania. e-mail: rstepanauskas@bigelow.org

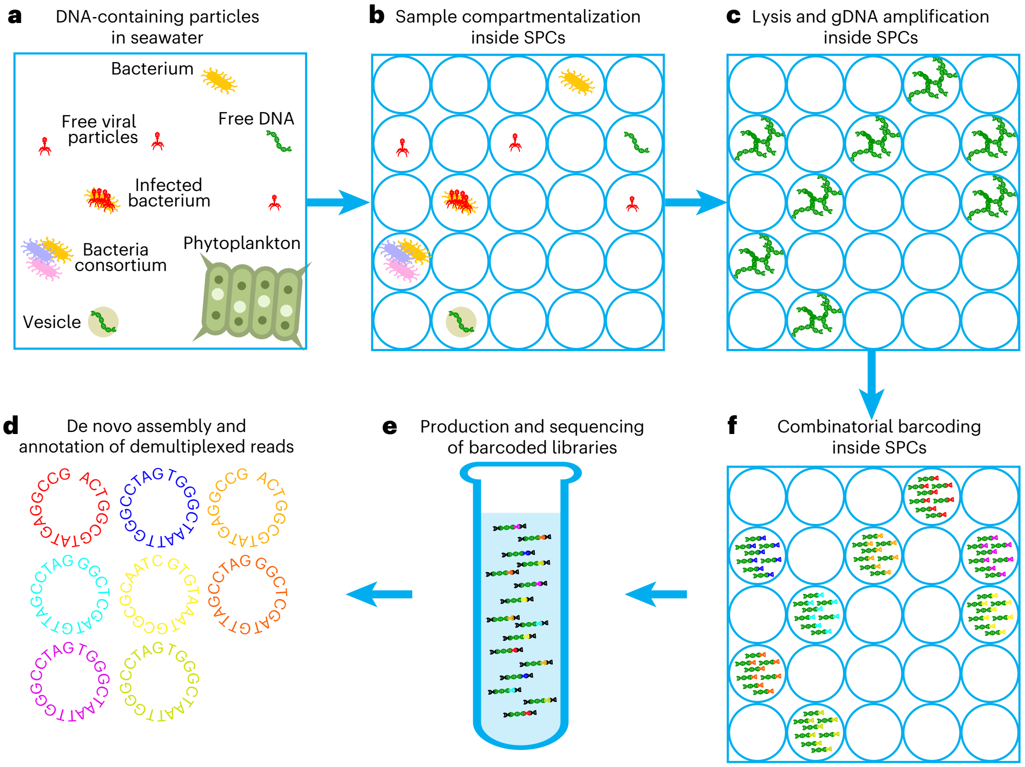
a DNA-containing particles in seawater
Bacterium
Free viral particles
Free DNA
Infected bacterium
Bacteria consortium
Phytoplankton
Vesicle
b Sample compartmentalization inside SPCs
c Lysis and gDNA amplification inside SPCs
d De novo assembly and annotation of demultiplexed reads
e Production and sequencing of barcoded libraries
f Combinatorial barcoding inside SPCs

Fig. 1 | Schematic overview of the EMCG workflow. Each panel displays a key step in the EMCG workflow. Items are not drawn to scale. **a**, An aquatic sample containing diverse types of DNA-containing particles is used in the analysis. **b**, Particles with diameters smaller than SPCs (~70 µm) are randomly compartmentalized inside SPCs. **c**, Genomic DNA is released from cells and extracellular particles with an alkaline treatment and amplified using a modified multiple displacement amplification, WGA-X. **d**, A series of barcodes are ligated to the amplified DNA within each SPC in a combinatorial manner to provide each particle with a unique tag for its DNA fragments. **e**, SPCs are dissolved and the barcoded DNA fragments are pooled to prepare a single sequencing library, which is sequenced using Illumina technology. **f**, Sequence reads are demultiplexed on the basis of their SPC-specific barcodes, assembled de novo separately for each SPC and annotated using standard bioinformatics tools.

circumvent these limitations[22–25], but so far, it has been hampered by relatively low throughput and the confines of optical sensitivity of fluorescence-activated virus sorting (FAVS) for particle separation. The global abundance and genome content of other extracellular genetic elements remain even less understood.

High-throughput droplet microfluidics-based studies provide a promising avenue to improve the limitations of FAVS-based single-particle genomics. Such methods have revolutionized biomedical sciences by providing an inexpensive means to measure gene expression of tens of thousands of individual animal cells[26,27]. Several recent reports have also demonstrated the feasibility of diverse microdroplet- and microgel-based approaches to analyse the genomic DNA of individual prokaryotic cells[28–32] and individual viral particles[33]. Here we introduce an approach that builds on these improvements in which samples are parcelled into picolitre-sized compartments using semi-permeable capsules (SPCs)[29], separating individual particles for DNA amplification, barcoding and sequencing, which we call Environmental Micro-Compartment Genomics (EMCG). We applied EMCG to a raw, coastal seawater sample of 300 nanolitres and obtained over 2,000 single amplified genomes. These corresponded to a broad range of cellular and extracellular entities, which largely overlapped with known members of marine communities. Surprisingly, however, a substantial fraction of SAGs lacked metagenomic representation and were similar to *Naomiviridae*, which use deoxyuridine (dU) instead of deoxythymidine (dT) in their genomes that prevents direct sequencing[34], suggesting that viruses with non-canonical DNA are abundant in seawater, yet remain invisible to conventional metagenomics. Collectively, this work demonstrates that EMCG enables a high-throughput, quantitative recovery of single-particle genome sequences with improved assembly completeness, thus providing a more comprehensive understanding of the extracellular genomic realm.

Results

Quantification and identification of particles captured by EMCG

We report the development and pilot application of EMCG, which is based on compartmentalization of aquatic samples inside SPCs (Fig. 1). The shell of SPCs is composed of polymers that allow the free passage of enzymes and other reagents while blocking the escape and entry of high-molecular-weight DNA. This enables multistep biochemical processing across thousands of encapsulated particles for lysis, DNA amplification and combinatorial barcoding. DNA is then pooled for shotgun sequencing, and sequence reads are then separated on the basis of individual barcodes for separate assemblies, resulting in thousands of individual genomic assemblies from a singular sequencing run. All steps are accomplished using manual pipetting without the need for complex and expensive instrumentation. The method represents a substantial increase in analytical throughput and reduction in equipment costs compared with fluorescence-activated cell sorting (FACS)-based single-cell and single-virus genomics.

We applied EMCG to a 300-nanolitre aliquot of coastal seawater from the Gulf of Maine collected on 18 April 2022, which produced 2,037 EMCG single amplified genomes (emSAGs) (cumulative length 86 Mbp; Supplementary Table 1). Using SPC volume, rate of occupancy and emSAG taxonomic assignments as input variables, we estimated the abundance of prokaryoplankton cells and DNA virus-like particles in the analysed sample at 1.93 × 10^9 l^{-1} and 1.79 × 10^{10} l^{-1}, respectively. This agreed with flow-cytometric counts in the same sample (1.84 × 10^9 l^{-1} and 1.75 × 10^{10} l^{-1}) and with the global average of microscopic counts of

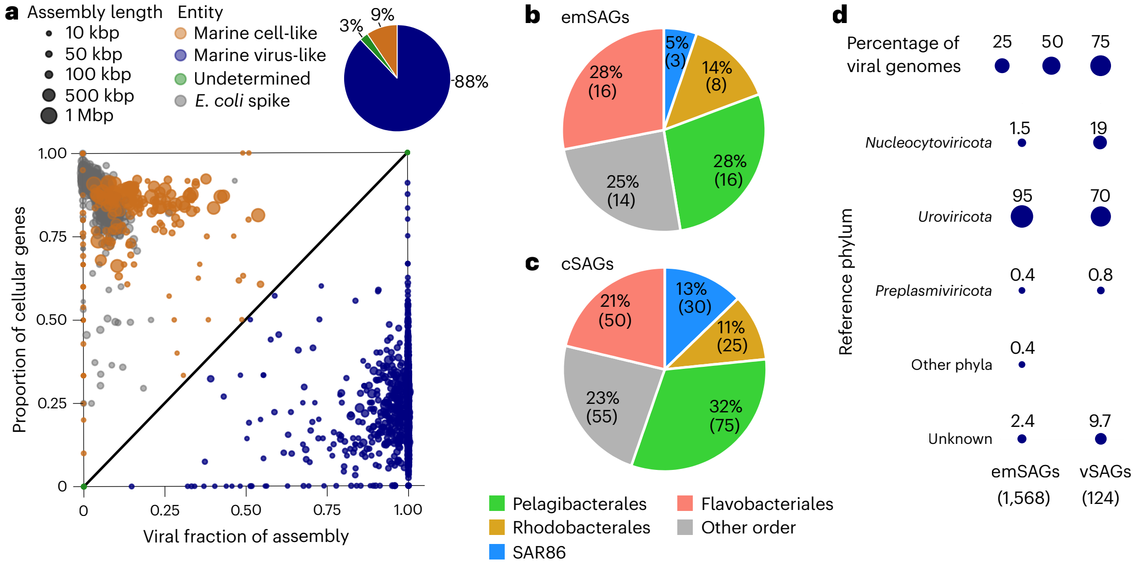

Fig. 2 | Compositional overview of de novo genome assemblies obtained from coastal seawater using EMCG and other methods. a, The fraction of prokarya-like genes relative to the fraction of virus-like sequence length in individual genome assemblies, which was used to tentatively distinguish cellular (orange) and virus-like (blue) SAGs. The SPCs containing *E. coli* cells were included in the EMCG workflow as internal standards (grey). The pie chart summarizes the fractions of cell-like and virus-like marine emSAG assemblies. **b**,**c**, Taxonomic composition of cellular SAGs recovered from EMCG (emSAGs, **b**) and prokarya-targeted FACS (cSAGs, **c**). **d**, Taxonomic phyla assignments of virus-like SAGs obtained using EMCG (emSAGs) and virus-targeting FAVS (vSAGs), arranged in order of decreasing average genome length in Viral RefSeq: *Nucleocytoviricota* (average genome length 676 kb), *Uroviricota* (56 kb), *Preplasmiviricota* (25 kb).

prokaryoplankton and viral particles in the coastal ocean[16], indicating that emSAG assemblies represented DNA-containing particles in the analysed sample in a randomly sampled, holistic and thus potentially quantitative manner.

For comparative purposes, we prepared a metagenome of virus-sized particles from the same seawater sample. The 127 million metagenomic paired-end reads assembled into 17,145 virus-like metagenome-assembled genomes (vMAGs) (cumulative length 44 Mbp; Supplementary Table 2). In addition, the same seawater sample was used to produce 260 cellular SAGs (cSAGs) and 190 virus-like SAGs (vSAGs) using a previously described FACS-based workflow[24] (Supplementary Table 3).

We performed a tentative separation of emSAGs into cell-like and virus-like genomes based on the proportion of prokaryote-like genes versus the fraction of sequence detected as viral, along with co-processing of *E. coli* emSAGs serving as internal standards to quantify the frequency of barcode co-use by multiple emSAGs (Fig. 2a, and Supplementary Tables 1, 4 and 5). This resulted in 1,791 virus-like marine emSAGs, 193 cell-like marine emSAGs, 53 marine emSAGs without assignments, 339 *E. coli* emSAGs, and 3 emSAGs that contained both *E. coli* and marine contigs. The 1.4% frequency at which the same barcode was incorporated into more than one emSAG (marine and *E. coli*) demonstrated fidelity of emSAG barcoding and computational demultiplexing. Notably, the virus-like emSAGs may include other protected genetic elements that could not be distinguished with current methods, as viruses exist along a continuum of diverse mobile genetic elements[13]. Likewise, the cell-like emSAGs may be derived from intact cells, free DNA, or membrane vesicles containing cellular DNA, but the distinction between these particle types cannot be made using current bioinformatics tools with genetic data alone. Nonetheless, the 9:1 ratio between virus-like and cell-like assemblies agreed well with the previously reported ~10:1 average in the surface ocean[16].

The taxonomic composition of cell-like emSAGs that could be classified (57 emSAGs) was similar to the composition of prokaryoplankton cSAGs that could be classified (235 cSAGs) from the same sample (Fig. 2b,c) and in agreement with previous studies of the same coastal location[35]. The absence of eukaryotic emSAGs (unicellular algae and heterotrophic protists) in this dataset can be explained by their abundance typically being ~100-fold lower than that of prokarya[36] and by some unicellular eukarya exceeding the diameter of our SPCs. Together, these findings provide further evidence that emSAGs quantitatively represent the composition of DNA-containing particles, ranging from prokaryotic to small virus sizes, in the analysed sample. In agreement with previous studies, they revealed viruses as the most abundant type of DNA-based extracellular genetic element in the analysed sample[15].

Of the virus-like SAGs that shared at least two genes with a classified reference virus[37,38], tailed double-stranded (ds)DNA phages of the phylum *Uroviricota* (all of the class *Caudoviricetes*) dominated both virus-like emSAGs (95.2%) and virus-like FACS-based vSAGs (70.2%), while the relative abundance of other viral groups differed substantially between the two methods (Fig. 2d and Supplementary Tables 1 and 3). Eukaryotic giant viruses of the phylum *Nucleocytoviricota* were over 10× more abundant in the vSAGs; in contrast, prokaryotic viruses of phyla with smaller genomes, including *Phixiviricota* and *Cressdnaviricota* (both single-stranded (ss)DNA viruses), were represented only by emSAGs. The absence of small viruses and the enrichment of giant viruses in vSAGs relative to emSAGs probably reflected optical detection limits of FAVS. EMCG appears to circumvent biases against small particles because it does not rely on particle optical detection.

An important feature of microbial single-cell genomics is the recovery of viral DNA from cells interacting with viruses or containing prophages[39–42]. Our analyses revealed the presence of DNA sequences of viral origin in 33 (17%) of cell-like emSAGs (Supplementary Table 1). None of these sequences were flanked by cellular DNA, suggesting lytic rather than lysogenic interactions, although suboptimal host infections may also cause some viral DNA occurrences[43]. Surprisingly, we found no matches between the sequences of viral origin in cell-like emSAGs and virus-like emSAGs in the analysed sample. This may be caused either by viral diversity exceeding the scale of our experiment or by an offset in time when a particular virus is present inside hosts versus the water column. Some evidence for the latter possibility is provided by the presence of sequences matching emSAGs of this study in cSAGs reported previously from the Gulf of Maine (see section on *Naomiviridae* below).

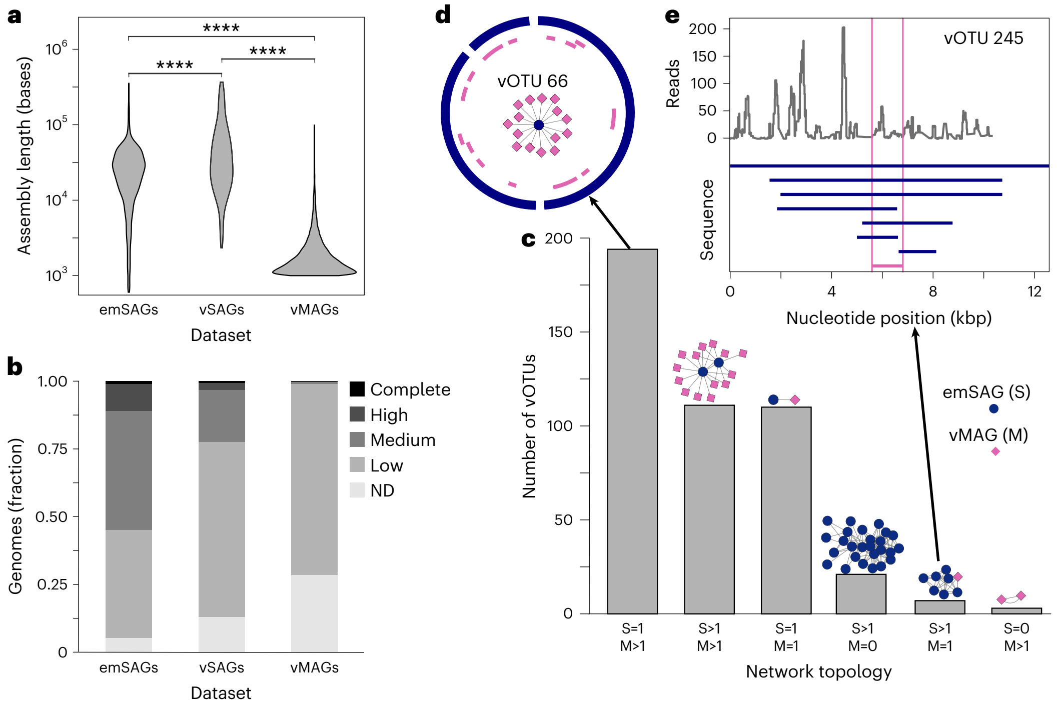

Fig. 3 | Quality of virus-like emSAGs, virus FACS-based vSAGs and viral vMAGs. **a**, Genome assembly length. Stars indicate significant differences (two-sided Wilcoxon test, ****$p < 0.001$). **b**, Quality assignments by CheckV (ND, not determined). **c**, Counts of vOTUs with various network topologies. Topology examples are provided in the insets, where nodes are emSAGs (pink) and vMAGs (blue) connected by shared vOTU edges. **d**, Illustration of alignments of multiple vMAGs against a single emSAG that form vOTU 66 (see SAG and MAGs listed in Supplementary Table 8). **e**, Illustration of a vOTU containing a single MAG and multiple SAGs from vOTU 245. Top: sequence depth of viral metagenome reads mapped on the longest emSAG in this vOTU. Bottom: alignments of the emSAGs and vMAG that form this vOTU (see SAGs and MAGs listed in Supplementary Table 7). The pink bar marks the region covered by the only vMAG of this vOTU.

Quality of virus-like genome assemblies

The quality of recovered virus-like genomic sequences differed across the three methodologies (EMCG (emSAGs), FAVS (vSAGs) and viral metagenomics (vMAGs)). The average length of virus-like emSAGs (27 kbp) was over 10× greater than that of vMAGs (2 kbp), although lower than that of vSAGs (56 kbp) (Fig. 3a), probably due to FAVS favouring larger viral particles (Fig. 2). The CheckV-estimated quality of the genomes[44] indicated that emSAGs had a higher proportion of high- and medium-quality genome assemblies (964 emSAGs, 54%) than both metagenomic contigs (156 vMAGs, 1%) and vSAGs (22%). These results collectively indicate that EMCG more successfully recovered high-quality genomes than metagenomics and FAVS-based single-viral sequencing (Fig. 3b).

To assess the diversity recovered, we used network analysis to cluster emSAGs and vMAGs into nominal species-like operational taxonomic units (vOTUs) using single-linkages clustering with a cut-off of ≥95% average nucleotide identity (ANI) across ≥85% of coding nucleotides of the smaller genome[45] (Supplementary Table 6). Most emSAGs and vMAGs did not cluster with other genome assemblies (44% and 89%, respectively), with 11% of vMAGs clustering with an emSAG and 44% of emSAGs clustering with a vMAG. Among the multimember clusters, the most abundant cluster configuration comprised multiple vMAGs clustering with a single emSAG by aligning to various regions of the emSAG (Fig. 3c,d), with 4.7% of vMAGs belonging to such clusters. This reflected the substantially greater completeness of emSAG assemblies compared with vMAGs (Fig. 2b) and demonstrated the important capacity of emSAGs to link together vMAGs that otherwise might have been identified as different viruses due to the lack of overlapping genomic regions.

When using metagenomics (bulk sequencing), populations of near-identical particles are expected to assemble into a single MAG representing a consensus genome. In contrast, each emSAG represents an individual viral particle, and therefore, EMCG can recover multiple closely related genomes. We thus expected to observe viral genomic clusters that contained multiple emSAGs and a single MAG. Yet, such clusters comprised only 1.6% of multimember vOTUs (7 total) (Fig. 3c). An alignment of genomes from one such vOTU revealed that the vMAG corresponded to a relatively short genome region that was shared by the greatest number of emSAGs and was one of several regions with an elevated metagenome read coverage (Fig. 3e and Supplementary Table 7). Overall, the emSAGs exhibited highly uneven coverage by viral metagenomic reads, indicating high variability in nucleotide conservation among viral particles that probably hindered metagenomic assembly, leading to fragmented, low-quality vMAGs.

Unexpected abundance of particles with putatively non-canonical DNA

Most large clusters of virus-like genomes (vOTUs) containing emSAGs shared similarities with phages isolated from known marine bacterial groups such as Flavobacteria, *Synechococcus* and *Pelagibacter* (Fig. 4a), highlighting the general congruence of EMCG with previous studies of marine viruses. Given the random, quantitative sampling of particles by EMCG (Figs. 1 and 2) and ocean mixing, we expected the number of emSAGs per vOTU to correlate with the fraction of viral metagenome reads mapped to a given vOTU. Surprisingly, this correlation was weak (Pearson's correlation $p < 0.01$; R value 0.27; Supplementary Table 1), and the vOTU with the highest number of emSAGs (26) did not recruit any viral metagenomic reads and lacked vMAG members (Fig. 4a). This indicated that differences in the composition of viral emSAGs and vMAGs (Fig. 3c) were not caused solely by metagenome assembly fragmentation but also by access to different DNA sources.

To further examine the taxonomic diversity of the virus-like emSAGs, we applied a gene-sharing network approach with reference viruses of INPHARED[37] and remaining non-prokaryotic viruses on Viral

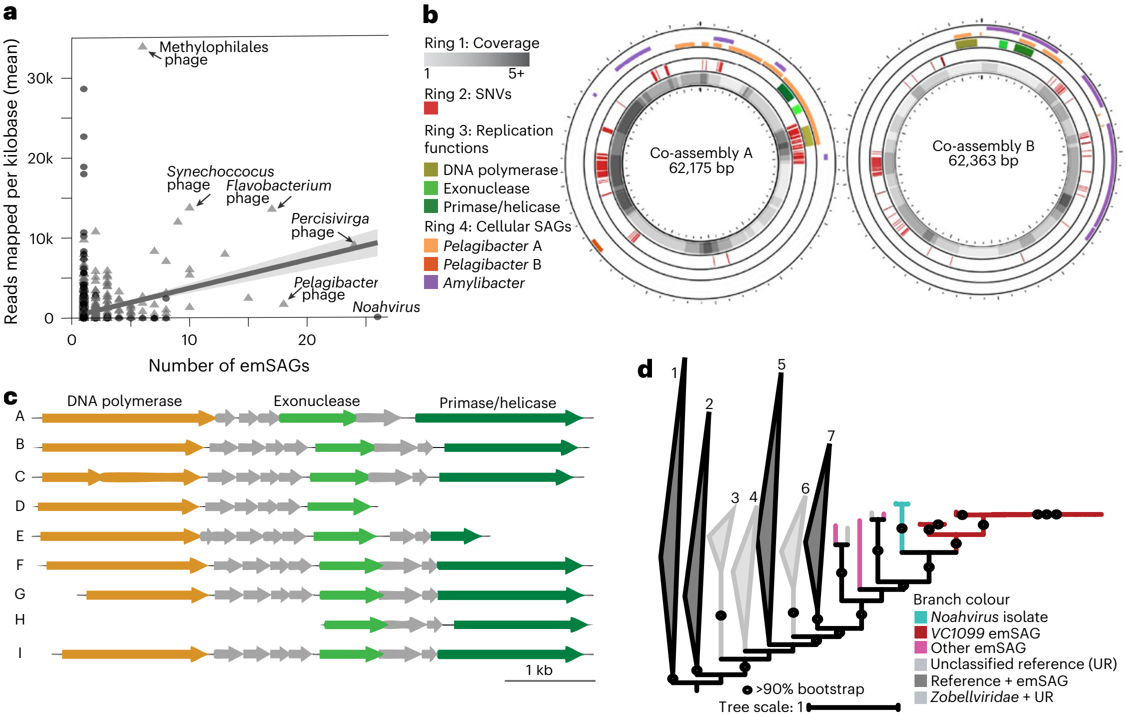

Fig. 4 | Characterization of emSAGs related to *Naomiviridae*. **a**, Relationship between the number of emSAGs in a vOTU and the average number of viral metagenomic reads recruited per kilobase of that vOTU. Triangle-shaped vOTUs contained a vMAG, circle vOTUs had no vMAGs. Line and shaded area indicate linear regression and its standard error, respectively. **b**, Genomic maps of co-assemblies of VC1099 emSAGs. Rings indicate co-assembly coverage depth by contigs of emSAGs (innermost ring 1), single nucleotide variants (SNVs; ring 2), known replication genes (ring 3) and alignments to cellular SAGs indicating putative hosts (ring 4). **c**, Alignments of genome regions containing replication genes characteristic to *Naomiviridae* retrieved from emSAGs of VC1099 (genomes A–I listed in Supplementary Table 13). **d**, Phylogeny of virus-like DNA polymerase found in emSAGs and reference viral genomes. Branches are coloured by dataset or genome type. Dark grey collapsed branches contain a mix of emSAGs and references. Numbers above clade correspond to these members (UR, unclassified references): (1) UR, emSAGs; (2) *Zobellviridae*, *Autographiviridae*, UR, emSAGs; (3) *Autographiviridae*, UR; (4) *Grimontviridae*, UR; (5) *Zobellviridae*, *Straboviridae*, *Chaseviridae*, emSAGs; (6) *Zobellviridae*, UR; (7) UR, emSAGs.

Refseq[38], which clusters genomes into genera or subfamilies. This revealed that the emSAGs of the abundant vOTU, which lacked representation in the metagenome, belonged to the cluster VC1099, which contained cultured viruses of the *Noahvirus* genus in the *Naomiviridae* family (Supplementary Table 9). Of the 1,791 virus-like emSAGs, 1.8% clustered into VC1099, which translates to their abundance of ~3×10^8 l^{-1} in the analysed sample. Collectively, these results indicate that EMCG uncovered an abundant lineage of viruses related to *Noahvirus* that were undetectable using viral metagenomics.

At present, there are only two isolates of the *Noahvirus* genus, which were collected from the Grand Canal in Venice, Italy, and the coast of Puerto Morelos, Mexico, and found to infect the marine bacterium *Ruegeria pomeroyi*, a member of the abundant prokaryoplankton family Rhodobacteraceae (Alphaproteobacteria)[34]. This isolation study reported that *Noahvirus* distinctively uses a non-canonical nucleotide dU instead of dT, and their genomes could be sequenced only after producing canonical DNA copies with the help of the phi29 polymerase. The EMCG workflow involves the amplification of gDNA of individual particles with phi29 polymerase, which helps explain the recovery of *Naomiviridae* in emSAGs. In contrast, our viral metagenome was sequenced without pre-amplification, as pre-amplification can introduce severe biases in mixed community shotgun sequencing[46].

The genes *Noahvirus* use for dU synthesis have been hypothesized to include genes for a cytosine deaminase, a DNA polymerase and a dTTPase, although this has not been experimentally verified[34]. Of these genes, we only detected an orthologous DNA polymerase in several of the VC1099 emSAGs (Fig. 4c and Supplementary Table 10). The failure to find orthologues of *Noahvirus* cytosine deaminase and dTTPase suggests that these genes are either too divergent to detect at the sequence level or that different mechanisms are used for DNA modification in *Naomiviridae* than was previously hypothesized. The incompleteness of individual emSAGs may also lead to failed gene detection, although the likelihood of missing these genes altogether in all 33 VC1099 emSAGs due to assembly incompleteness is low, $<10^{-12}$. Furthermore, the successful co-assembly of contigs from these emSAGs into two circularized genomes suggests successful coverage of *Naomiviridae* core genome regions by the emSAGs (Fig. 4b). Orthologues of other *Noahvirus* replication genes were present in several VC1099 emSAGs, including an exonuclease and a DNA primase/helicase. Genomic synteny of these genes was consistent with cultured *Noahvirus* (Fig. 4c). A phylogenomic reconstruction of the DNA polymerase gene family in *Noahvirus*[34] with other reference viral genomes in INPHARED and RefSeq that encode this DNA polymerase family provided further evidence that VC1099 emSAGs and *Noahvirus* isolates are evolutionarily related (Fig. 4d).

To better understand the genome architecture of VC1099 viruses, we assembled all contigs from all marine emSAGs into constructs we refer to as co-assemblies. This produced two circularized co-assemblies, both of which contained exclusively *Naomiviridae* emSAG contigs (Fig. 4b). The two co-assemblies were ~62 kbp in length, similar to the length of the two previously published genomes of *Noahvirus* isolates (70 and 75 kbp)[34]. Mapping of individual emSAG contigs on these co-assemblies revealed a high frequency of nucleotide-level variability among individual *Naomiviridae* particles (Fig. 4b). Furthermore, these co-assemblies included only 39% of the contigs of the contributing emSAGs, although portions of all contigs could be mapped on the co-assemblies (Supplementary Tables 11). This indicated substantial genome content variation among individual *Naomiviridae* particles in the analysed sample that could not be fully represented in the two co-assemblies, and demonstrates the unique value of genome recovery from discrete, individual viral particles to circumvent the inherent loss of information that occurs during the assembly of consensus genomes from multiple entities, such as done in the construction of MAGs.

We next searched for contigs resembling the two VC1099 co-assemblies in cellular SAGs from this study and a previous, multiseason report of prokaryoplankton of FACS-based SAGs in the Gulf of Maine[35] to capture interactions with potential hosts. Small fragments of the phage co-assemblies aligned to contigs present within three cellular SAGs, one *Amylibacter* and two *Pelagibacter* SAGs (Supplementary Table 12 and Fig. 4b). These alignments varied in length from 1 to 17.5 kb and in their nucleotide identity from 73% to 98%. The cellular SAG contigs that aligned to *Naomiviridae* co-assemblies differed in tetramer composition from the rest of the cellular assembly, suggesting heterogeneous origins and lack of integration into hosts' chromosomes (Supplementary Figs. 1–3). The association of VC1099 with *Amylibacter*, of the Rhodobacteraceae family, is consistent with the taxonomy of the previously reported host *R. pomeroyi*[34]. Meanwhile, the detected association of *Naomiviridae* with *Pelagibacter* (Pelagibacteraceae family) suggests an important additional role of *Naomiviridae*, as *Pelagibacter* is the most abundant genus of marine planktonic bacteria and constitutes ~25% of all cells in the ocean[47]. Interestingly, all three putative host SAGs originate from a set of 276 cSAGs obtained from a seawater sample collected in April 2019, rather than from the other sampling months, October and July, which coincided with the sample collection for the EMCG study in April 2022. This indicates the potential seasonal recurrence of *Naomiviridae* infections in the Gulf of Maine. We also performed BLASTn searches for sequences similar to *Naomiviridae* co-assemblies in 12,715 GORG-Tropics cSAGs from the epipelagic depths of tropical and subtropical latitudes[5]. This produced no alignments greater than 100 bp, suggesting high specificity of *Naomiviridae* matches to cSAGs from the Gulf of Maine. Collectively, our findings expand the putative host range of *Naomiviridae* by including the most abundant lineage of marine bacterioplankton and suggest the dynamic nature of their ecology and evolution, highlighting the need for targeted studies on their roles in the ocean. EMCG opens a new window into the biology of viruses with non-canonical DNA, which may be instrumental not only in microbial ecology but also in the expanding use of xenobiotic nucleotides in synthetic biology[48].

Discussion

We report the development and pilot application of Environmental Micro-Compartment Genomics (EMCG), a novel method for high-throughput, quantitative genome sequencing of individual cells and extracellular genetic elements in environmental samples. Particles in a sample are randomly compartmentalized into semi-permeable capsules (SPCs), followed by multistep biochemical processing for lysis, DNA amplification and combinatorial barcoding via manual pipetting. This high-throughput separation of individuals and ability to pool barcoded DNA into a singular sequencing run enables thousands of particles to be sequenced from a sample, substantially increases the analytical throughput and reduces equipment costs compared with FACS-based single-cell and single-virus genomics.

The application of EMCG on a 300-nl coastal seawater sample recovered genome sequences of over 2,000 individual microbial cells and extracellular genetic elements, with the predominance of virus-like genome assemblies (Fig. 2). The comparison of EMCG outputs to FACS-based particle counting and FACS-based genomics of single cells and single viral particles demonstrated consistent genome recovery without obvious compositional biases, aside from demonstrating a bias of FAVS-based genomics for larger viral particles. The average completeness of virus-like genomes, which dominated the EMCG dataset, was substantially higher than the completeness of virus-like MAGs and FAVS-based SAGs obtained from the same sample (Fig. 3). Furthermore, EMCG recovered genomes from a broader taxonomic range of viruses compared with FAVS-based SAGs, and included viruses that might have been too small for optical detection by FAVS. Thus, the quality of virus-like genome assemblies obtained with EMCG has substantial advantages relative to other contemporary methods in environmental virology.

Unexpectedly, EMCG revealed the abundance of *Naomiviridae* in the analysed sample and indicated that they infect several of the most abundant marine bacterioplankton lineages (Fig. 4). The only previous report of this family of viruses found that they use dU instead of dT, which prevents their direct sequencing[34]. This and other non-canonical DNA modifications are hypothesized to help viruses avoid detection by hosts' immune systems[49]. Although several DNA modifications have been discovered in cultured prokaryotic viruses, and the initial *Naomiviridae* report indicated that relatives of these phages may be widespread in the ocean, the abundance of such viruses in the environment remains largely unknown[49,50]. Our results suggest that viruses with non-canonical DNA may be abundant in certain ocean regions and play crucial roles in marine microbial ecology, yet remain undetected by standard metagenomics methods.

One key limitation of EMCG, compared to the FACS-based genomics of single cells and viral particles, is the lack of information about analysed particles' optical properties. When available, such information can link single cells' genomes and phenomes and discriminate intact cells from subcellular particles[24,35]. Furthermore, the lack of direct control of sequence read distribution among SPCs can lead to both under- and over-sequencing of some SPCs, in particular when analysing particles with disparate genome sizes. The latter challenge could be mitigated by targeted particle or entity selection, for example, by filtration or FACS, upstream of EMCG. Compared with shotgun metagenomics, EMCG requires moderately higher upfront instrumentation costs, yet it offers an improved representation of diverse DNA sequences in the sample. Furthermore, due to the uneven DNA amplification from single molecule templates[51], EMCG may require deeper DNA sequencing compared with metagenomics to obtain the same cumulative assembly length.

Taken together, EMCG offers a novel opportunity for quantitative genome content analyses of viruses and other extracellular genetic elements in the ocean and other environments. The discovery of the prevalence of *Naomiviridae* in seawater reflects the importance of improving our understanding of the abundance, ecological roles and biotechnological potential of extracellular genetic elements with non-canonical DNA. The capacity of EMCG to decipher viral and microbial genome content at the resolution of individual particles from submicrolitre sample volumes may also find applications outside environmental microbiology, such as in infection diagnostics and forensics.

Methods

Field sample collection and storage

Gulf of Maine seawater sample was collected with a Niskin bottle from 1-m depth in Boothbay Harbor, Maine (43.84° N, 69.64° W) on 18 April 2022. Immediately after the collection, multiple 1-ml subsamples were transferred to cryovials, amended with 5% glycerol and 1× pH 8 TRIS-EDTA buffer (final concentrations) and stored at −80 °C. We refer to these subsamples as 'cryopreserved'.

Sample encapsulation and processing SPCs

SPCs containing cryopreserved seawater were generated using the FLUX SPC generator (CHP-SPC1, Atrandi Biosciences) and the corresponding SPC reagent kit (CKP-G34, Atrandi Biosciences). The Working Core (WCS) and Working Shell (WSS) Solutions were prepared following manufacturer instructions, except for dithiothreitol, which was excluded. Specifically, for each lane, 37.5 µl of diluted seawater was combined with 50 µl of Core Reagent and 12.5 µl of Photoinitiator. Before encapsulation, all samples were passed through a

40-µm-mesh-size cell strainer (Becton Dickinson) and diluted with sterile-filtered and UV-irradiated Sargasso Sea water to achieve 2–3% SPC occupancy, as determined through an iterative process by the microscopy of post-amplification SPCs (see below). This corresponds to 1–1.5% of the occupied SPCs harbouring >1 DNA-containing particle. Post encapsulation, the emulsion was transferred into a 2.0-ml microcentrifuge tube, the oil phase was aspirated from the bottom using a 1-ml pipette, and the SPCs were isolated by disrupting the emulsion with the Emulsion Breaker solution included in the kit. The recovered SPCs were then resuspended in molecular grade water containing 1× Wash Additive.

Alkaline lysis and DNA amplification in SPCs were performed following the WGA-X protocol[24] instead of using lysis and amplification reagents provided in the kit. Briefly, SPCs were treated with 0.4 M KOH, 10 mM EDTA and 100 mM dithiothreitol for a 10-min incubation at 20 °C, followed by neutralization and buffer replacement using Wash Buffer (10 mM Tris pH 7.5, 0.1% Triton X-100). The WGA-X reactions contained 0.2 U µl^{-1} Equiphi29 polymerase (Thermo Fisher), 1× Equiphi29 reaction buffer (Thermo Fisher), 0.4 mM each dNTP (New England BioLabs), 10 mM dithiothreitol (Thermo Fisher) and 40 µM random heptamers with two 3′-terminal phosphorothioated nucleotide bonds (Integrated DNA Technologies) (all final concentrations). These reactions were performed at 45 °C for 1 h and then inactivated by a 15-min incubation at 75 °C. Following the WGA-X, SPCs were washed with 1 ml wash buffer three times. To confirm DNA amplification, the SPCs were stained with 5 µM SYTO-9 (Thermo Fisher) in wash buffer. SPCs containing green fluorescence were quantified using epifluorescence microscopy. We refer to amplified DNA in individual SPCs as environmental micro-compartment single amplified genomes (emSAGs).

An axenic culture of *Escherichia coli* K12 DH1 (ATCC 33849) was obtained from the American Type Culture Collection and cultivated in an autoclaved LB medium overnight. The 1 ml subsamples of the culture were transferred to cryovials, amended with 5% glycerol and 1× pH 8 TRIS-EDTA buffer (final concentrations) and stored at −80 °C. Subsequently, the same dilution, SPC generation and DNA amplification procedures as described above were applied to these cryopreserved *E. coli* samples. Before DNA barcoding, the SPCs containing amplified *E. coli* DNA were spiked into SPCs containing amplified seawater DNA at the ratio of ~1:10 to assist in workflow performance evaluation.

The same techniques and reagents described above were used to encapsulate, lyse, amplify and microscopically examine a sample of UV-irradiated deionized water. This negative control produced no SPCs with detectable DNA amplification products, indicating no DNA amplification introduced by the analytical workflow.

Combinatorial barcoding of WGA products within SPCs was conducted using the Single-Microbe DNA Barcoding kit (CKP-BARK1, Atrandi Biosciences). Approximately 150 µl of post-WGA-X SPCs, containing ~6,000 SPCs with WGA products, were passed through DNA debranching and end preparation protocols to yield A-tailed dsDNA fragments. A 4-step, ligation-based combinatorial split-and-pool barcoding process was then employed, generating a barcode diversity of 24^4 (331,776) unique variants.

The barcoded emSAGs were released from SPCs using the Release Reagent (Atrandi Biosciences) and purified using 0.8× Ampure XP (Beckman Coulter) paramagnetic beads. Subsequently, the next-generation sequencing libraries were prepared following the protocol of the NEB Next Ultra II FS DNA Library Prep Kit for Illumina (E7805S, NEB) using a protocol-specific ligation adapter (Atrandi Biosciences) and Unique Dual Indexing (UDI) PCR primers as directed by the Single-Microbe DNA Barcoding kit. Libraries were quantified using the Agilent High Sensitivity DNA kit (5067-4626, Agilent) and the Qubit Fluorometer with a dsDNA HS Assay kit (Q32851, Thermo Fisher). The prepared SAG libraries were loaded on the Illumina NextSeq 2000 platform using a High Output P3 kit (300 cycles) with sequencing read lengths of R1 = 151 bp, R2 = 151 bp, i7 = 8 bp and i5 = 8 bp.

Read demultiplexing proceeded in two stages. In the first stage, we used bcl2fastq2 v.2.2.0 software (Illumina) with the parameter ‘–barcode-mismatches=3’ to convert basecall files to pairs of fastq files and separate reads on the basis of the Illumina library indexes. The second stage used the paired fastq files as input and the Pheniqs (v.2.1.0)[52] software to separate reads on the basis of the combinatorial Atrandi barcodes. To ensure demultiplexing stringency, we used a frequency threshold of ≥4 out of 1 million reads to exclude spurious barcode combinations and ran Pheniqs using the minimal distance decoding algorithm and distance tolerance set to allow ≤1 mismatch in each Atrandi barcode combination. It should be noted that an earlier version of the manual mistakenly indicated compatibility with the NEB Next Multiplex Oligos for Illumina (Dual Index Primers Set 2), NEB (P/N: E7780S) indexing primer kit, which compromised barcode-bearing fragment enrichment, resulting in only ~30% of reads displaying a recognizable cell barcode. This issue has been rectified in subsequent manual editions, ensuring reference to compatible indexing primers and achieving greater than 90% readability of cell barcodes in reads.

Following demultiplexing, we performed two-stage read trimming. First, Trim Galore! (v.0.6.7)[53] command ‘trim_galore–clip_R2 45–paired -o output ${input_forward_readfile} ${input_reverse_readfile}’ was used to remove Atrandi barcodes. Subsequently, Trimmomatic (v.0.32)[54] command ‘trimmomatic PE -phred33 ${ input_forward_readfile } ${input_reverse_readfile } LEADING:0 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36’ was used to remove Illumina adapters and perform general quality-based trimming. Only reads containing their pairs and longer than 36 bp were retained as previously benchmarked[24].

Quantification of DNA-containing particles

Flow-cytometric quantification of prokarya within the seawater sample (1.84×10^6 ml^{-1}) and *E. coli* culture (3.50×10^8 ml^{-1}) was performed with a BD Influx Mariner by counting events gated as prokarya (Supplementary Fig. 4) in a known sample volume[55]. The same quantification procedure was used on a FACS gate for virus-like particles, indicating 1.75×10^7 virus-like particles per ml. The volume analysed was determined by weight displacement before and after analysis. Details are found in the section ‘Fluorescence-activated sorting and downstream processing of individual cells and virus-like particles’.

The EMCG-based quantification of DNA-containing particles was based on measuring sample volumes encapsulated in SPCs, the rate of SPC occupancy by particles producing WGA-X products, and the taxonomic annotation of sequenced emSAGs. First, we used the following formula to estimate the volume of raw seawater captured in an average SPC (V_{sea}-1SPC):

$$V_{\mathrm{sea}} - 1\mathrm{SPC}(\mathrm{pl}) = 160 \times 0.2 \times 0.033 = 1.06$$

where ‘160’ is the microscopically determined average droplet volume in pl (assuming a spherical shape), ‘0.2’ is the volumetric fraction of the diluted sample in a droplet, and ‘0.033’ is the sample dilution factor before encapsulation.

Next, we estimated the abundance of DNA-containing particles in the analysed seawater (A_{DNA}) as follows:

$$A_{\mathrm{DNA}}\,(\mathrm{ml}^{-1}) = 0.022 \times 10^{9}/V_{\mathrm{sea}} - 1\mathrm{SPC} = 2.04 \times 10^{7}$$

where ‘0.022’ is the microscopically determined fraction of SPCs containing WGA-X amplicons and ‘10^9’ is the conversion factor from pl to ml.

Subsequently, we estimated the abundance of prokarya cells (A_{prok}) and virus-like particles (A_{vir}) as follows:

$$A_{\mathrm{prok}}(\mathrm{ml}^{-1}) = A_{\mathrm{DNA}} \times 193/2,037 = 1.93 \times 10^{6}$$

$$A_{\mathrm{vir}}(\mathrm{ml}^{-1}) = A_{\mathrm{DNA}} \times 1,791/2,037 = 1.79 \times 10^{7}$$

where '193', '1,791' and '2,037' are the respective counts of prokarya-like, virus-like and all emSAG genome assemblies of marine origin.

The total volume of seawater compartmentalized in SPCs that were used in the EMCG experiment (V_{sea}-sequenced-SPCs) was estimated as follows:

$$V_{\text{sea}} - \text{sequenced} - \text{SPCs(nl)} = V_{\text{sea}} - 1\text{SPC} \times 6,000/0.022/1,000 = 289$$

where '6,000' is the number of SPCs used in barcoding and sequencing, '0.022' is the microscopically determined fraction of SPCs containing WGA-X amplicons, and '1,000' is the conversion factor from pl to nl.

Fluorescence-activated sorting and downstream processing of individual cells and virus-like particles

The generation of FACS-based single amplified genomes of prokaryote cells (cSAGs) and individual viral particles (vSAGs) from the cryopreserved seawater sample was performed as previously described[24]. For prokaryote cell sorting, samples were incubated with the SYTO 9 nucleic acid stain (5 µM final concentration; Thermo Fisher) on ice for 10–60 min. For virus-like particle sorting, samples were incubated with the SYBRGreen nucleic acid stain (1:500 dilution, final concentration; Thermo Fisher) on ice for 60–120 min. Flow-cytometric analysis and sorting were performed using a BD InFlux Mariner flow cytometer equipped with a 488-nm laser for excitation and a 70-µm nozzle orifice (Becton Dickinson, formerly Cytopeia). The cytometer was triggered on green fluorescence (531/40 BP emission filter). The 'single-1 drop' mode was used for maximal sort purity. Sort gates for prokarya and viral-like particles were defined on the basis of particle green fluorescence and forward scatter (Supplementary Fig. 4). Individual cells and extracellular particles were deposited into 384-well microplates containing 600 nl of 1× TE buffer per well and stored at −80 °C until further processing. Of the 384 wells, 317 wells were dedicated for single cells, 64 wells were used as negative controls (no droplet deposition), and 3 wells received 10 cells each to serve as positive controls. The accuracy of droplet deposition into microplate wells was confirmed several times during each sort day by sorting 3.46-µm-diameter SPHERO Rainbow Fluorescent Particles (Spherotech) and microscopically examining their presence at the bottom of each well. In these examinations, <2% of wells did not contain beads and <0.4% of wells contained more than one bead. Index sort data were collected using the BD Sortware software.

Before genomic DNA amplification, sorted cells and viral particles were lysed and their DNA was denatured by two freeze–thaw cycles and the addition of 700 nl of a lysis buffer consisting of 0.4 M KOH, 10 mM EDTA and 100 mM dithiothreitol, and a subsequent 10-min incubation at 20 °C. The lysis was terminated by the addition of 700 nl of 1 M Tris-HCl, pH 4. Whole genome amplification was performed using WGA-X[24].

Cell and particle sorting, lysis and WGA-X setup were performed in a HEPA-filtered environment conforming to Class 1000 cleanroom specifications. Before cell sorting, the instrument, the reagents and the workspace were decontaminated for DNA using UV irradiation and sodium hypochlorite solution. To further reduce the risk of DNA contamination, and to improve accuracy and throughput, Bravo (Agilent Technologies) and Freedom Evo (Tecan) robotic liquid handlers were used for all liquid handling in 384-well plates at Bigelow Laboratory's Single Cell Genomics Center (SCGC).

Libraries for sequencing of the amplified DNA were created with Nextera XT (Illumina) reagents following manufacturer instructions except for purification steps, which were done with column cleanup kits (QIAGEN QIAQuick PCR Purification and MinElute PCR Purification kits), and library size selection, which was done using BluePippin (Sage Science) with a target size of 500 ± 50 bp. DNA concentration measurements were performed with Agilent HS D1000 TapeStation reagents following manufacturer instructions. Libraries were sequenced with NextSeq 2000 (Illumina) and P3 reagents in 2 × 100 bp mode.

SAG de novo assembly and quality control

The demultiplexed sequencing reads of emSAGs, cSAGs and vSAGs were quality trimmed with Trimmomatic (v.0.32)[54] using the following settings: -phred33 LEADING:0 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. Reads matching the *H. sapiens* reference assembly GRCh38 and a local database of WGA-X reagent contaminants[24] (≥95% identity of ≥100 bp alignments) as well as low-complexity reads (containing <5% of any nucleotide) were removed. The remaining reads were digitally normalized with kmernorm 1.05 (http://sourceforge.net/projects/kmernorm) using the settings '-k 21 -t 30 -c 3' and then assembled with SPAdes (v.3.0.0)[56] using the following settings: −careful−sc−phred-offset 33. Each end of the obtained contigs was trimmed by 100 bp and only contigs longer than 2,000 bp were retained. Contigs matching the *H. sapiens* reference assembly GRCh38 and a local database of WGA-X reagent contaminants (≥95% identity of ≥100 bp alignments) were removed. All SAGs were run through CheckM (v.1.0.7)[57] to determine cellular genome completeness and a tetramer frequency analysis[24]. The application of this workflow on cSAGs was evaluated for assembly errors using three bacterial benchmark cultures with diverse genome complexity and %GC, indicating no non-target and undefined bases in the assemblies and the following average frequencies of misassemblies, indels and mismatches per 100 kbp: 1.5, 3.0 and 5.0 (ref. 24). All SAGs were examined for bacterial and archaeal taxonomy via GTDB-Tk (v.1.4.1)[58]. The 16S rRNA gene regions longer than 500 bp were identified using local alignments provided by BLAST against CREST's[59] curated SILVA reference database SILVAMod (v.128)[60] and classified using a re-implementation of CREST's last common ancestor algorithm.

Co-use of barcode assessment with *E. coli* spike

The theoretically expected fraction of emSAG assemblies originating from more than one SPC due to random barcode co-use is 0.36% ($2,379 \times 0.5/24^4 = 0.0036$, where 2,379 is the number of obtained genome assemblies and 24^4 is the total number of barcode combinations). To experimentally verify the low frequency of barcode co-use in emSAGs, and due to the technical challenges in robustly assessing the purity of genome assemblies of novel extracellular genetic elements, we introduced a spike of SPCs containing *E. coli* in the barcoding procedure. Subsequently, contigs of the *E. coli* spike were detected in emSAG genome assemblies using QUAST[61]. Other contigs were considered of marine origin. Of the total of 2,379 emSAG assemblies with at least one >1-kbp contig, 2,037 assemblies had marine contigs only, 339 assemblies had *E. coli* contigs only, and 3 assemblies had a mix of both (Supplementary Table 4). The presence of marine contigs in 0.9% of emSAG assemblies containing *E. coli* sequences translates to the observed fraction of emSAG assemblies originating from more than one SPC as 1.0% ($3 \times 2,379/2,037/(339 + 3)$), thus confirming the high fidelity of EMCG barcoding and computational demultiplexing procedures.

SAG annotation and entity assignment

Genes for all SAGs were called using Prodigal[62] (flags -d, -a, -g) and predicted with all genetic codes considered initially (1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25). If code 11 yielded a coding density <80%, then the code yielding the highest coding density was deemed optimal[63] (Supplementary Tables 1, 3 and 5). The amino acid sequences of genes identified with the optimal genetic code were then aligned to Hidden Markov Model profiles of gene families of all eukaryotic, bacterial, archaeal and viral groups in the EggNOG database (v.5.0)[64] using HMMsearches (hmmsearch -E 0.001; HMMER 3.3.2)[65]. We used the highest bitscore as the criterion to determine a gene's best hit. The proportion of cellular genes in a SAG was calculated by dividing the number of best hits to a cellular EggNOG family by the total gene count. To identify viral contigs within the SAGs, untrimmed contigs were analysed with Virsorter2 (v.2.1)[66] (default settings) and

DeepVirFinder (v.1.0)[67]. Contigs with a Virsorter2 maxscore ≥0.5 or a DeepVirFinder score ≥0.5 and $p < 0.05$ were deemed 'viral'. All contigs of a SAG containing at least one virus-like contig were then analysed using CheckV[44] to trim off host regions. Next, the proportion of viral-like sequence (in basepairs) in a SAG was calculated on the basis of the total untrimmed length of that SAG. SAGs that had a higher fraction of cellular genes than the fraction of viral sequence were deemed 'cellular'. Meanwhile, SAGs that had a higher fraction of viral sequence than the fraction of cellular genes were deemed 'virus-like'. If these proportions were equal, the SAG assignment was considered 'undetermined' (Supplementary Tables 1, 3 and 5). Cell-like emSAG genome assemblies were deemed to contain a viral infection if they had at least one contig over 10 kbp in length that met the criteria used to detect viral sequences of SAGs initially described above using VirSorter2 and DeepVirFinder (Supplementary Table 2). The 10-kbp length cut-off used here is higher than the 1-kbp length used to categorize a SAG as cell-like or virus-like out of caution since viral detection altogether is more accurate at higher contig lengths, specifically 10 kbp for both tools used here (VirSorter2 and DeepVirFinder)[66]. However, this limits the detection of infections by viruses with shorter genomes such as some members of the *Inoviridae* and *Microviridae* families. Viral sequences of prophages were detected by CheckV as 'proviruses' and only reported if their contigs were flanked on both sides by host DNA. This did not apply to any of the sequences.

CheckV[44] was used to assess the quality of virus-like SAG assemblies (Supplementary Tables 1 and 3). Because CheckV only runs on individual contigs, we joined all viral contigs originating from the same SAG with 30 Ns before processing with CheckV[63]. Because no universally accepted method currently exists for the taxonomic classification of virus-like genome assemblies, we examined which reference viruses the viruses shared the most similarity with by analysing their protein content. Amino acid sequences of each virus-like assembly were aligned against all proteins in the INPHARED database (downloaded 2 August 2024)[37] and remaining non-prokaryotic viruses in Viral RefSeq (VRS) (downloaded 9 March 2024)[38] using BLASTp (*e*-value 0.00001; (v.2.14.1)[68]). Tentative classifications were reported for SAGs that had at least two genes with the highest bitscore values from BLAST hits to the same reference genome (Supplementary Tables 1 and 3). We note that this approach reveals the most similar reference phage to a virus-like SAG but not necessarily the specific taxonomic group this virus belongs to. To determine the genus or subfamily level of classification for the virus-like emSAGs, we combined the virus-like emSAG sequences with INPHARED and VRS and ran them in vContact2 (ref. 69) (default settings; Supplementary Table 9).

Viral metagenome preparation, assembly and annotation

A total of 40 l of seawater collected from the same sample that was for SAG analyses was filtered through a 0.2-μm-mesh-size filter, concentrated to 250 ml using a REXEED 30k MWCO tangential flow filter system (Asahi Kasei Medical)[70,71], and then stored at 4 °C overnight. The following morning, this primary concentrate was further concentrated to ~1.5 ml final volume (~30,000× concentration) using Amicon 100k MWCO ultracentrifugation filters (Millipore Sigma) in a refrigerated centrifuge. The collected material was resuspended off the filter in DNA-free, deionized water and stored at −80 °C. The DNA was extracted from a 30-μl aliquot of this secondary concentrate using previously described protocols[72]. Extracted DNA was subjected to an additional clean and concentration step using a Genomic DNA Clean and Concentrator kit (Zymo Research), eluting the purified DNA with elution buffer heated to 65 °C.

Libraries for DNA sequencing were created with Nextera XT (Illumina) reagents following manufacturer instructions, except for purification steps, which were done with column cleanup kits (QIAGEN QIAQuick PCR Purification and MinElute PCR Purification kits), and library size selection, which was done using BluePippin (Sage Science) with a target size of 370 ± 43 bp. DNA concentration measurements were performed with Agilent HS D1000 reagents for TapeStation following manufacturer instructions. Libraries were sequenced with NextSeq 2000 (Illumina) and P1 reagents in 2 × 150 bp mode. The obtained 127,098,956 read pairs were quality trimmed with Trimmomatic (v.0.3260)[54] using the following settings: -phred33 LEADING:0 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. Reads matching the *H. sapiens* reference assembly GRCh38 and a local database of reagent contaminants (≥95% identity of ≥100 bp alignments) as well as low-complexity reads (containing <5% of any nucleotide) were removed[24]. The remaining 123,812,680 read pairs were assembled into contigs using metaSPades[73] with default settings and filtered for scaffolds at least 1,000 bp in length. Virus-like contigs were detected and their quality was assessed using VirSorter2 (ref. 66), DeepVirFinder[67] and CheckV[44] using the same criteria as in SAG analyses, with a minimum contig length of 1,000 bp (see above; Supplementary Table 2).

vOTU clustering, genome alignments and read mapping

To assess the diversity of virus-like emSAGs and virus-like vMAGs, we clustered these together into either nominal virus species or vOTUs, which are most commonly defined by ≥95% ANI across ≥85% of the length of constituent genomes[45]. Given that virus-like emSAGs can contain multiple contigs with unknown orientations, we calculated their pairwise ANI values between all virus-like emSAGs and virus-like vMAGs as averages across their coding sequence lengths. Coding sequences of each genome were aligned using BLASTn[68] (*e*-value 0.0001). The alignment between two genes with the best bitscore was used. vOTUs were determined on the basis of single-linkage clustering using the threshold of ≥95% ANI across ≥85% of the length of the smallest constituent genome (Supplementary Table 6).

For the genome alignments within population_66 (Fig. 3d), we used the MUMmer4 (v.4.0.0rc1)[74] commands nucmer with default settings and show-cords -T with the largest genome assembly as the reference (Fig. 3d and Supplementary Table 8). For population_245, we visualized the region of the largest emSAG assembly to which the vMAG BBHvirome2_NODE_14273 aligned (Fig. 3e). For the mapping of reads from the virus-enriched metagenome, the reads were initially mapped to all genome assemblies of each dataset using bwa mem (2.2.1)[75] and samtools (1.15.1)[76]. An index was created using bwa index. Then the following commands and flags were used: bwa mem -t 4 | samtools view -ShuF4q2−threads 4 | samtools sort -m 8 G−threads 4. The samtools index was applied to a bam file output from above. Next, reads that mapped at 95% ID, as performed in ref. 77, were filtered. The metagenome reads were also mapped on individual genome assemblies using bwa mem and the 95% ID threshold (Supplementary Table 1). To determine the relative abundance of each vOTU in the metagenome, the average number of reads mapped to each emSAG within the vOTU was divided by the average assembly length of emSAGs within the vOTU (Fig. 4a). The command 'samtools depth' was used for positional information of mapping to an individual sequence (Fig. 4e).

Naomiviridae analyses

Gene annotations of *Noahvirus arc* and *Noahvirus PM1* were downloaded from GenBank and clustered with genes from emSAGs of VC1099 using OrthoFinder[78] (-S blast; v.2.5.5) (Supplementary Table 10). Clusters were then examined for those containing cytosine deaminase, DNA polymerase, dTTPase, exonuclease and DNA primase/helicase based on the annotation of the cultured genomes' genes. For emSAGs that contained more than one of these genes, corresponding genome neighbourhoods were plotted with genoPlotR[79] in R (v.4.2.2)[80] using RStudio[81] (v.2023.06.2 + 561) (Supplementary Table 13). To estimate the approximate likelihood of missing a gene in the collective VC1099 emSAGs, we used genome assembly completeness (C_A) estimates as a proxy. We calculated each emSAG's completeness by dividing its assembly length by the average genome length of the two reference

Naomiviridae (72,565.5 bp). We used $1{-}C_A$ as a heuristic estimate of the likelihood that a particular gene would be absent in a genome assembly due to assembly incompleteness. The probability of a specific gene being missed in all VC1099 emSAGs due to assembly incompleteness was estimated as $(1{-}C_{A1}) \times (1{-}C_{A2}) \ldots \times (1{-}C_{A33})$, where C_{A1} through C_{A33} are completeness estimates for the 33 emSAGs of VC1099.

To examine DNA polymerase diversity, amino acid sequences encoded by virus-like emSAGs, VRS and the INPHARED database were searched against HMM profiles provided in ref. 34 using hmmsearch in HMMERv3 (e-value < 10^{-5}). The best hit for a genome was selected on the basis of the highest bitscore. Due to visualization constraints, DNA polymerases found in VRS and INPHARED, excluding the *Noahvirus* copies, were clustered at 90% ID with CD-Hit[82] (v.4.7; -c 0.99; Supplementary Table 14). The resulting amino acids were then aligned with Muscle5 (v.5.2)[83] (-super5). The alignment was trimmed to remove regions that contained >10% gaps (trimAl[84]; -gt 0.1). A maximum likelihood phylogeny was reconstructed with this alignment using IQ-TREE 2 (ref. 85) with 1,000 ultrafast bootstraps and 8 independent runs, using the ModelFinder algorithm[86] to identify the optimal amino acid substitution model (-nt 8 -m TEST -wbt -bb 1000 –runs 5 -safe). The tree was visualized in iTOL (v.6)[87].

All contigs of marine emSAGs were used as input sequences in a single de novo assembly with Geneious Assembler (Dotmatics) using medium sensitivity and allowing circularization of contigs with ≥2 sequences. Of the obtained co-assemblies, only two were circularized and examined further. The two circularized co-assemblies, which we designated 'Co-assembly A' and 'Co-assembly B', consisted entirely of contigs derived from emSAGs of VC1099, suggesting that they represent genetic material of *Naomiviridae*. Contigs of these emSAGs were also aligned to the co-assembly using BLASTn (e-value < 10^{-5}) to identifiy the fraction of assemblies that matched the co-assembly (Supplementary Table 11). Functional annotation of these co-assemblies was performed in the same way as for emSAGs. Their amino acid sequences were then searched for against the *Noahviruses*' genes via BLASTp searches (e-value < 10^{-5}) to identify genes involved in replication. Putative hosts of these co-assemblies were predicted by BLASTn alignments (bitscores ≥ 1,000) of the contigs of cellular SAGs from the Gulf of Maine of this study and two previous studies in the Gulf of Maine and global tropics[5,35] to the two co-assemblies (Supplementary Table 12). The coverage, single nucleotide polymorphisms, replication genes and host alignments were mapped onto the co-assembly genomes in the web-based software Proksee on the web-based interface[88].

Data visualization and statistics

Data frames were managed with the pandas package[89] in Python 3.8.16 or the R packages dplyr[90], tidyverse[91] and reshape2 (ref. 92). Unless specified otherwise, plots were created in R using RStudio with the following packages: ggplot2 (ref. 93), ggpubr[94], cowplot[95]. Correlation tests were completed with the cor.test in base R (method = 'kendall'). Pairwise t-tests (Fig. 3a) were completed with stat_compare_means (label = 'p.signif'; ggpubr function).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Genomic data are available on figshare at https://doi.org/10.6084/m9.figshare.29087720.v1 (ref. 96) and on GenBank under BioProject accession PRJNA1227513.

Code availability

All bioinformatic tools, visualizations and statistics were completed with open-source software and programmes as described in the Methods. More detailed code is provided on GitHub as 'EMCG_code' at https://github.com/scubalaina/EMCG_code/tree/main (ref. 97).

References

1. Azam, F. et al. The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).
2. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl Acad. Sci. USA* **115**, 6506–6511 (2018).
3. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
4. Sunagawa, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
5. Pachiadaki, M. G. et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* **179**, 1623–1635.e11 (2019).
6. Paoli, L. et al. Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
7. Chen, J. et al. Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371–379 (2024).
8. Holm-Hansen, O., Sutcliffe, W. H. & Sharp, J. Measurement of deoxyribonucleic acid in the ocean and its ecological significance. *Limnol. Oceanogr.* **13**, 507–514 (1968).
9. Brum, J. R. Concentration, production and turnover of viruses and dissolved DNA pools at Stn ALOHA, North Pacific Subtropical Gyre. *Aquat. Microb. Ecol.* **41**, 103–113 (2005).
10. Linney, M. D., Schvarcz, C. R., Steward, G. F., DeLong, E. F. & Karl, D. M. A method for characterizing dissolved DNA and its application to the North Pacific Subtropical Gyre. *Limnol. Oceanogr. Methods* **19**, 210–221 (2021).
11. Nagler, M., Insam, H., Pietramellara, G. & Ascher-Jenull, J. Extracellular DNA in natural environments: features, relevance and applications. *Appl. Microbiol. Biotechnol.* **102**, 6343–6356 (2018).
12. Hackl, T. et al. Novel integrative elements and genomic plasticity in ocean ecosystems. *Cell* **186**, 47–62.e16 (2023).
13. Koonin, E. V. Viruses and mobile elements as drivers of evolutionary transitions. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150442 (2016).
14. Penadés, J. R. & Christie, G. E. The phage-inducible chromosomal islands: a family of highly evolved molecular parasites. *Annu. Rev. Virol.* **2**, 181–201 (2015).
15. Linney, M. D. et al. Microbial sources of exocellular DNA in the ocean. *Appl. Environ. Microbiol.* https://doi.org/10.1128/aem.02093-21 (2022).
16. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
17. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123.e14 (2019).
18. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
19. Luo, E., Eppley, J. M., Romano, A. E., Mende, D. R. & DeLong, E. F. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J.* **14**, 1304–1315 (2020).
20. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
21. García-López, R., Vázquez-Castellanos, J. F. & Moya, A. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Front. Bioeng. Biotechnol.* **3**, 141 (2015).
22. Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).

23. Wilson, W. H. et al. Genomic exploration of individual giant ocean viruses. *ISME J.* **11**, 1736–1745 (2017).
24. Stepanauskas, R. et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.* **8**, 84 (2017).
25. Martinez-Hernandez, F., Fornas, O. & Martinez-Garcia, M. Into the dark: exploring the deep ocean with single-virus genomics. *Viruses* **14**, 1589 (2022).
26. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
27. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
28. Lan, F., Demaree, B., Ahmed, N. & Abate, A. R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646 (2017).
29. Leonaviciene, G., Leonavicius, K., Meskys, R. & Mazutis, L. Multi-step processing of single cells using semi-permeable capsules. *Lab Chip* **20**, 4052–4062 (2020).
30. Chijiiwa, R. et al. Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome* **8**, 5 (2020).
31. Zheng, W. et al. High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* **376**, eabm1483 (2022).
32. Ma, P. et al. Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell* **186**, 877–891.e14 (2023).
33. Nishikawa, Y. et al. Large-scale single-virus genomics uncovers hidden diversity of river water viruses and diversified gene profiles. *ISME J.* **18**, wrae124 (2024).
34. Rihtman, B. et al. A new family of globally distributed lytic roseophages with unusual deoxythymidine to deoxyuridine substitution. *Curr. Biol.* **31**, 3199–3206.e4 (2021).
35. Munson-McGee, J. H. et al. Decoupling of respiration rates and abundance in marine prokaryoplankton. *Nature* **612**, 764–770 (2022).
36. Gin, K. Y. H., Chisholm, S. W. & Olson, R. J. Seasonal and depth variation in microbial size spectra at the Bermuda Atlantic time series station. *Deep Sea Res. 1 Oceanogr. Res. Pap.* **46**, 1221–1245 (1999).
37. Cook, R. et al. INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. *PHAGE* **2**, 214–223 (2021).
38. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
39. Labonté, J. M. et al. Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* **9**, 2386–2399 (2015).
40. Labonté, J. M. et al. Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front. Microbiol.* **6**, 349 (2015).
41. Jarett, J. K. et al. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J.* **14**, 2527–2541 (2020).
42. Munson-McGee, J. H. et al. A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *ISME J.* **12**, 1706–1714 (2018).
43. Kauffman, K. M. et al. Resolving the structure of phage–bacteria interactions in the context of natural diversity. *Nat. Commun.* **13**, 372 (2022).
44. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
45. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
46. Yilmaz, S., Allgaier, M. & Hugenholtz, P. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat. Methods* **7**, 943–944 (2010).
47. Giovannoni, S. J. SAR11 bacteria: the most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.* **9**, 231–255 (2017).
48. Eremeeva, E. & Herdewijn, P. Non canonical genetic material. *Curr. Opin. Biotechnol.* **57**, 25–33 (2019).
49. Hutinet, G., Lee, Y.-J., de Crécy-Lagard, V. & Weigele, P. R. Hypermodified DNA in viruses of *E. coli* and *Salmonella*. *EcoSal Plus* **9**, eESP00282019 (2021).
50. Zhou, Y. et al. A widespread pathway for substitution of adenine by diaminopurine in phage genomes. *Science* **372**, 512–516 (2021).
51. Woyke, T. et al. Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4**, e5299 (2009).
52. Galanti, L., Shasha, D. & Gunsalus, K. C. Pheniqs 2.0: accurate, high-performance Bayesian decoding and confidence estimation for combinatorial barcode indexing. *BMC Bioinformatics* **22**, 359 (2021).
53. Krueger, F. Trim Galore! *Babraham Institute* https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
55. Marie, D., Simon, N. & Vaulot, D. in *Algal Culturing Techniques* (ed. Andersen, R. A.) 253–267 (Elsevier, 2005).
56. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
58. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
59. Lanzén, A. et al. CREST–classification resources for environmental sequence tags. *PLoS ONE* **7**, e49334 (2012).
60. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
61. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
62. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
63. Weinheimer, A. R. & Aylward, F. O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *ISME J.* **16**, 1657–1667 (2022).
64. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
65. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
66. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
67. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
68. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

69. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
70. Hill, V. R. et al. Development of a rapid method for simultaneous recovery of diverse microbes in drinking water by ultrafiltration with sodium polyphosphate and surfactants. *Appl. Environ. Microbiol.* **71**, 6878–6884 (2005).
71. Langenfeld, K., Chin, K., Roy, A., Wigginton, K. & Duhaime, M. B. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ* **9**, e11111 (2021).
72. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
73. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
74. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
75. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* https://doi.org/10.1109/ipdps.2019.00041 (IEEE, 2019).
76. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
77. Chang, T., Gavelis, G. S., Brown, J. M. & Stepanauskas, R. Genomic representativeness and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton. *Microbiome* **12**, 126 (2024).
78. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
79. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
80. Gardener, M. *Beginning R: The Statistical Programming Language* (John Wiley & Sons, 2012).
81. Verzani, J. *Getting Started with RStudio: An Integrated Development Environment for R* (O'Reilly Media, 2011).
82. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
83. Edgar, R. C. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
84. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
85. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
86. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
87. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
88. Grant, J. R. et al. Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Res.* **51**, W484–W492 (2023).
89. McKinney, W. Data Structures for Statistical Computing in Python. In *Proc. Python in Science Conference* 56–61 (SciPy, 2010).
90. Wickham, H. & Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (O'Reilly Media, 2016).
91. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
92. Wickham, H. Reshaping data with the reshape Package. *J. Stat. Softw*. **21**, 1–20 (2007).
93. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
94. Kassambara, A. *R Graphics Essentials for Great Data Visualization: 200 Practical Examples You Want to Know for Data Science* (STHDA, 2017).
95. Wimberly, M. C. *Geographic Data Science with R: Visualizing and Analyzing Environmental Change* (CRC Press, 2023).
96. Weinheimer, A. et al. Abundant non-canonical DNA viruses in seawater revealed by high-throughput single particle sequencing (reference data). Dataset. *figshare* https://doi.org/10.6084/m9.figshare.29087720.v1 (2025).
97. Weinheimer, A. R. EMCG_code. *GitHub* http://github.com/scubalaina/EMCG_code/tree/main (2025).

Acknowledgements

This work was supported by US National Science Foundation grants OCE-2116253 and OCE-2231327 to R.S., J.M.B. and N.J.P., and Simons Foundation grant 991222 to A.R.W. R.Z. received funding from the Research Council of Lithuania (LMTLT), Agreement No. P-MIP-24-597.

Author contributions

A.R.W. and R.S. led the data analysis and manuscript preparation, with input provided by all authors. G.G. developed high-throughput computational workflows. B.T., C.M., J.M.B., J.M.-M., N.J.P. and R.S. performed field and laboratory-based experimental work. G.L., V.K., S.J., L.M. and R.Z. provided microencapsulation technology for beta-testing and advised on its use. R.S., J.M.B. and N.J.P. conceived the study and secured funding.

Competing interests

This research involved the use of commercial products from Atrandi Biosciences, as described in the Methods section. L.M. and R.Z. are shareholders of Atrandi Biosciences. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-025-02167-5.

Correspondence and requests for materials should be addressed to Ramunas Stepanauskas.

Peer review information *Nature Microbiology* thanks Luis Bolaños and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

nature portfolio

Corresponding author(s): Ramunas Stepanauskas

Last updated by author(s): Sep 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	All bioinformatic tools, visualizations, and statistics were completed with open-source software and programs as described in the Methods. More detailed code is provided on the GitHub repository “EMCG_code” (link: https://github.com/scubalaina/EMCG_code/tree/main).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Genomic data is available on the FigShare repository: Abundant non-canonical DNA viruses in seawater revealed by high-throughput single particle sequencing (Reference data) (link https://figshare.com/s/63d43c0f989e5f3b4f06) and on GenBank in the BioProject accession PRJNA1227513.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Here we applied a novel method termed Environmental Compartment Genomics to seawater and compared the resulting prokaryotic and viral community to other sequencing methods of virus-enriched metagenomics, fluorescence activated cell-sorted genomics, and fluorescence activated virus-sorted genomics.
Research sample	Seawater was collected from the at the Bigelow Laboratory for Ocean Sciences dock in East Boothbay, Maine, USA in April 2022. The same sample was used to generate the metagenome, flow-cytometry derived genomes, and environmental compartment genomics-based genomes.
Sampling strategy	A singular large sample was collected and subsamples were taken as input for the sequencing methods compared.
Data collection	Sequence data was collected by Illumina shotgun sequencing.
Timing and spatial scale	N/A. A single sample was collected.
Data exclusions	No data was excluded.
Reproducibility	Several methods were applied to compare the resulting community from the new method, Environmental Micro-compartment Genomics (EMCG) to previous methods such as flow-cytometry based counts, FACS and FAVS-based sequencing, and metagenomics to ensure resulting community was consistent. Additionally, code and methods have been made explicit for reproducibility.
Randomization	Inherently EMCG , FACS/FAVS, and metagenomics randomly capture DNA-entities in the environment.
Blinding	N/A

Did the study involve field work? ☒ Yes □ No

Field work, collection and transport

Field conditions	April 18th, 2022. No precipitation. Maximum temperature was 13C and low was 2C.
Location	Boothbay Harbor, Maine (43.84ºN, 69.64°W)
Access & import/export	Walking from main laboratory building

Disturbance	No disturbance.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
□	☒ Flow cytometry
☒	□ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Stated in the manuscript: For prokaryote cell sorting, samples were incubated with the SYTO™ 9 nucleic acid stain (5 µM final concentration; Thermo Fisher Scientific) on ice for 10-60 min. For virus-like particle sorting, samples were incubated with the SYBRGreen nucleic acid stain (1:500 dilution, final concentration; Thermo Fisher Scientific) on ice for 60-120 min. Flow cytometric analysis and sorting were performed using a BD InFlux Mariner flow cytometer equipped with a 488 nm laser for excitation and a 70 µm nozzle orifice (Becton Dickinson, San Jose, CA - formerly Cytopeia). The cytometer was triggered on green fluorescence (531/40 BP emission filter). The “single-1 drop” mode was used for maximal sort purity. Sort gates for prokarya and viral-like particles were defined based on particle green fluorescence and forward scatter (Supplementary Fig. S4).
Instrument	BD InFlux Mariner
Software	BD Sortware
Cell population abundance	1.84 x 106 mL-1

Gating strategy	particle green fluorescence and forward scatter

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.