Metagenome-derived virus-microbe ratios across ecosystems

Purificación López-García[1*✉], Ana Gutiérrez-Preciado[1*], Mart Krupovic[2], Maria Ciobanu[1], Philippe Deschamps[1], Ludwig Jardillier[1], Mario López-Pérez[3], Francisco Rodríguez-Valera[3], David Moreira[1]

[1] Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France
[2] Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France
[3] Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Alicante, Spain

* These authors contributed equally to this work

✉ For correspondence, puri.lopez@universite-paris-saclay.fr

It is generally assumed that viruses outnumber cells on Earth by at least tenfold[1-4]. Virus-to-microbe ratios (VMR) are largely based on counts of fluorescently-labelled viral-like particles. However, these exclude intracellular viruses and potentially include false positives (DNA-containing vesicles, gene-transfer agents, unspecifically-stained inert particles). Here, we develop a metagenome-based VMR estimate (mVRM) that accounts for all DNA-virus stages (virion, intracellular lytic and lysogenic) by using normalised counts of major capsid proteins (MCPs) and cellular universal single-copy genes (USCGs) as proxies for virus and cell counts, respectively. After benchmarking this strategy using mock metagenomes with increasing VMR, we inferred mVMR across different biomes. To properly estimate mVMR in aquatic ecosystems, we generated metagenomes from co-occurring cellular and viral fractions (>50 kDa-200 µm size-range) in freshwater, seawater and solar-salterns. Viruses outnumbered cells in freshwater and marine plankton by, respectively, ~15-25 and ~2-7 fold. However, across 125 diverse non-aquatic metagenomes, viruses, on average, outnumber cells by two-fold. Putative provirus:virus ratios progressively increased in sediments/soils, microbial mats and animal-associated microbiomes as compared to plankton, consistent with the hypothesis that lysogeny increases with cell density[5,6,7]. Although viruses likely are the most diverse biological entities on Earth[8,9], their global numbers might be closer to those of cells than previously estimated.

Viruses are strict molecular parasites that play crucial roles in ecosystems[2,10]. Hugely diverse[8,9], viruses are thought to be at least one order of magnitude more abundant than microbial cells on Earth[1-4]. These abundance estimates are largely based on epifluorescence microscopy or flow cytometry (FCM) counts of virus-like particles (VLPs) and cells stained with nucleic acid dyes. These methods were originally applied to marine plankton, where measured virus-to-microbe ratios (VMR) frequently revolve, despite substantial variation, around 10 (Ref.[3,10,11]). However, epifluorescence-based VLP counts exclude lysogenic viruses and are blind to intracellular lytic viruses during the eclipse phase of infection. Conversely, these methods can include diverse false positives, including cell-DNA-containing membrane vesicles[12-15] or gene-transfer agents[4,16] and, usually overlooked, non-specifically stained organic and/or mineral particles[17,18]. The latter might be important confounding factors in high-density ecosystems such as sediments and soils, where cell and VLP detachment from the organic-mineral matrix is required prior to epifluorescence-based quantification[19]. VMR values are particularly disparate in sediments and soils[19-21] and broadly vary depending on the protocol used[22]. Nonetheless, several observations suggest that the VLP counts proportionally diminish with increasing cell density[7,23]. It has been hypothesized that lower VMR in cell-dense environments does not necessarily reflect low viral load but rather higher lysogenic prevalence (Piggyback-the-winner model)[23]. However, while VMR seems to decline with increasing microbial cell densities, evidence for enhanced lysogeny is ambiguous[6]. All these confounding factors challenge the accuracy of the existing VMR estimates and, hence, more reliable approaches for obtaining this ratio are needed. Here, we develop a metagenome-based strategy to estimate VMR for DNA viruses including all viral stages (viral particles –VP–, as well as intracellular lytic and lysogenic viruses). We first benchmark this method on mock metagenomes with increasing VMR. Then, we estimate mVMRs from *de novo* generated aquatic-system

metagenomes comprising coexisting VPs and cell fractions (50 kDa-200 µm) and compare the obtained estimates with classical epifluorescence-derived estimates (fVMR). Finally, we estimate mVMR from 125 additional metagenomes across diverse ecosystem types, notably including high-density habitats. Our results show that mVMR are adequate proxies that can more realistically account for DNA virus:cell ratios in natural ecosystems than fVMR.

Benchmarking mVMR in mock metagenomes

To develop a metagenome-based VMR estimate, we used universal single-copy genes (USCGs) and viral major capsid proteins (MCPs), usually present in single copy in viral genomes, as proxies for cells and viruses, respectively. We selected 15 ribosomal proteins that are well-represented across the three domains of life as USCGs (Supplementary Table 1). To limit differential representation of individual ribosomal protein genes, we averaged values for the 15 USCGs as a proxy for cell counts under the assumption that in most biomes cells contain one genome copy. Relatively rare cases of polyploidy have been described, e.g. in giant bacteria[24,25] and haloarchaea[26], and two chromosome copies usually occur in some archaea and actively transcribing bacteria[27]. However, in the wild, especially in energy-limited ecosystems, microbial cells are likely to exist in low average metabolic states and, consequently, limited ploidy. Should average ploidy be higher than one, this proxy might slightly overestimate cell numbers. Conversely, some markers do not capture well episymbiotic prokaryotes with reduced and divergent genomes, such as Candidate Phyla Radiation (CPR or Patescibacteria) and DPANN archaea[28,29], which might partly compensate for the ploidy excess in some cells. To count viruses, we used 132 Hidden Markov Models (HMMs) that recognize MCPs from virtually all known families of DNA viruses infecting bacteria, archaea and eukaryotes. In addition, to avoid missing MCPs from potential unknown viruses present in the studied natural metagenomes (see below), we designed specific HMMs for viral genomes assembled from them (Supplementary Table 2; see Methods). We then generated 10 mock metagenomes with increasing VMRs, ranging from ratios of 10 cells per viral genome (V001M10; the number following V and M indicates the ratio of viral to microbial cell genomes per mock metagenome) to 125 viral genomes per microbial cell (V125M01). We also included one metagenome with only cellular genomes (V000M01) and one with only viral genomes, i.e. a metavirome (V001M00). These mock metagenomes partly mimicked marine planktonic communities, and included genomes of archaea, bacteria, eukaryotes and viruses (Supplementary Table 3).

The assemblies in the mock metagenomes with the highest VMR (V042M01 and V125M01) were highly fragmented, exhibiting lower quality (Fig. 1a; Supplementary Table 4). To cope with potential gene detection problems in these assemblies, we applied two annotation strategies, one based on predicted gene annotation (Prokka[30]) and another based on the annotation of all possible open reading frames (OrfM[31]) allowing for the detection of individual protein domains even if genes were fragmented. Even if many more ORFs than genes were detected per mock metagenome (Extended Data Fig.1a), the amounts and relative proportions of USCGs per metagenome were remarkably similar between the two methods except for those with the highest VMR, where OrfM performed quantitatively better (Extended Data Fig.1b). The relative proportions of high-rank cellular taxa obtained with the two annotation strategies were closely similar (Extended Data Fig.1c-d) and in agreement with the expected taxonomic assignment (Supplementary Table 5). To estimate mVMR, we searched

for USCGs and MCPs in the assembled contigs from these 10 metagenomes using the respective protein family HMMs (see Methods) and, after mapping the original reads onto the identified contigs, obtained normalised USCG and MCP numbers by counting the reads per kilobase of gene sequence per million of mapped metagenome reads (RPKMs). The inferred mVMR values were also similar in for the two annotation strategies, and close to the expected VMR although, at very high VMR (V125M001), OrfM yielded slightly lower mVMR values than Prokka (Fig.1b). We could not detect USCGs in genes/ORFs predicted from the virus-only mock metavirome (V001M00). However, we did identify several MCPs in the metagenome containing only cellular genomes (V000M001). These were annotated as cellular viral-related or hypothetical proteins/domains by Prokka and OrfM and corresponded to proviruses present in the included cellular genomes (see Methods; Supplementary Table 3). Accordingly, it was possible to differentiate lysogenic from VPs and intracellular lytic viruses in these mock metagenomes (Fig.1c; Extended Data Fig.1c-d), which opens the possibility to do so in natural metagenomes. Although at very low VMR (V000M001-V001M10) OrfM slightly overestimated VP-lytic counts, the ratio of proviruses inferred by both annotation strategies was virtually identical at all mimicked VMRs (Fig.1c).

Estimation of mVMR in aquatic metagenomes

After benchmarking mVMR in mock metagenomes, we first sought to estimate mVMRs for aquatic ecosystems and compare the resulting values to classical epifluorescence-derived estimates (fVMR). However, published planktonic metagenomes are not suitable for the estimation of total mVMRs due to an inherent problem resulting from the fact that DNA is purified from the planktonic biomass retained on different pore-size filters upon filtration. Most of the available aquatic metagenomes contain the prokaryotic/picoeukaryotic fraction (0.2 to 3-5 µm cell size) and, more rarely, larger-size fractions where many protists are retained (>5 and up to 100-200 µm). Metagenomes from smaller plankton fractions containing VPs ('metaviromes') are obtained after tangential filtration of <0.2 µm filtrates to retain particles larger than 50-100 kDa. However, many VPs can be also retained in pre-filtration steps and end up in "cellular" fractions, for instance retained on membranes in-between pores and, especially, trapped among cells as filters saturate with biomass. A corollary is that cellular metagenomes and metaviromes are usually generated from different water volumes/samples, after imperfect particle fractionation steps, sequenced from different libraries and analysed separately, such that mVMR cannot be directly and reliably estimated. To overcome this problem, we generated integral microbial plankton metagenomes, including genomes of both cells and viruses (covering both extracellular and intracellular phases of the virus replication cycle), from the same water mass of selected aquatic ecosystems (Fig.2). They included a freshwater shallow lake in Northern France, the Western Mediterranean water column (sampled at 20 and 40 m depth) and three ponds of increasing salinity (6-14-32% salt) from a Mediterranean solar saltern (Extended Data Fig.2). To obtain more representative VMR, we sampled these ecosystems in late winter (see Methods), avoiding massive blooming periods when richness and evenness decrease[32].

To generate metagenomes of coexisting viral and microbial cell plankton, we carried out serial filtration steps from the same initial water mass through filters of 200 µm pore-diameter to exclude metazoans and then 20, 5 and 0.2 µm, retaining cells of different sizes. Tangential

filtration was applied to the resulting <0.2 µm-filtrate to concentrate viral particles (>50 kDa) (Fig.2). We retained aliquots of all sequential filtrates for fluorescent cytometry (FCM) counts of labelled cells and VLPs, and the final eluate (<50 kDa) as a blank (Supplementary Table 6 and Extended Data Fig.3). FCM counts of cells in all fractions >0.2 µm were consistently similar for each sample type, indicating that the vast majority of cells were prokaryotic. Cell abundances were higher in saltern ponds ($5.2x10^6$ to $1x10^7$ cells.ml^{-1}) and freshwater samples (4.3 to $4.7x10^6$ cells.ml^{-1}) compared to marine samples, for which the counts were about an order of magnitude lower (4.5 to $x10^5$ cells.ml^{-1}) (Fig.3a; Supplementary Table 6). VLP counts were also consistent in large-size fractions of each sample and diminished in the tangential-filtration concentrate. This observation indicates that many VLPs were retained in larger fractions due to progressive filter saturation with biomass and highlights the importance of combining different size fractions of the same filtration series for metagenome-based abundance estimates. VLP abundances appeared highest in saline-hypersaline environments ($5.6x10^7$, $4.1x10^7$ and $1.6x10^8$ VLP.ml^{-1} for 6%, 14% and 32% salt ponds). Although this could reflect the actual abundance of viruses, the very high VLP counts of the salt-saturated sample Sal32 might also include unspecifically fluorescently stained salt nanoprecipitates[18]. Estimated fVMR ranged from 3 to 16, with lower values for freshwater (~4), 40 m-deep Mediterranean (~5) and 14%-salt pond (~3) and higher for 20 m-deep Mediterranean (~14) and 6% (~10) and 32% (~13) saltern ponds (Fig.3a).

We purified DNA from the biomass retained on all filters used during serial filtration as well as from tangential-filtration concentrates, keeping track of the corresponding water volumes. DNA samples obtained from different fractions were then proportionally mixed for a given original water volume (Supplementary Table 7). We sequenced metagenomes from two plankton fractions including viral particles and cells up to i) 5 µm, i.e. prokaryotes and pico-nanoeukaryotes (50 kDa-5 µm; '02v' fraction), and ii) 200 µm (100 µm for EdV), including unicellular prokaryotes and pico-nano-microeukaryotes (50 kDa-200 µm; 'all' fraction) (Fig.2). In this way, co-occurring viral and cellular DNA was sequenced from the same library, avoiding technical biases. As eukaryotes represented a minor fraction in these aquatic systems, the two metagenomes were expected to behave as quasi-replicates for mVMR inference. In addition, we sequenced two 'metavirome' fractions (50 kDa-0.2 µm; 'vir' fraction) for the marine samples as internal control (Extended Data Fig.2; Supplementary Table 8). Normalised numbers of USCGs were generally congruent in our aquatic metagenomes, although with slight among-marker differences (Extended Data Fig.4). In most of our planktonic metagenomes, normalised viral counts (MCPs) were higher than USCG-based cell counts (Fig.3b). Most cell counts were bacterial (except for 32% salt ponds, dominated by haloarchaea); as expected, eukaryotic counts were comparatively low (Fig.3b).

Estimates of mVMR in our aquatic ecosystems varied between 1.5 and ~20, with marine samples exhibiting values below 8 (Fig.3c). mVMRs were similar to fVMR in the case of MedD (~4-6), although they were typically lower than fVMRs by a factor of 2-3 (MedS, ~3-5; Salt06, ~2-4; Salt14, ~1.5). By contrast, in the case of freshwater Etang des Vallées (EdV) samples, mVMR values were at least 3-4 times higher than fVMRs. This difference seems to suggest that, in addition to VPs, intracellular lytic viruses and proviruses were highly prevalent in the freshwater plankton of this highly diverse shallow lake[33]. The ratio of putative proviruses versus cells was indeed of ~3-7 for EdV samples (depending on the annotation strategy) but

very low (<0.3) for marine samples, which seems consistent with the prevalent genome streamlining of marine prokaryotes[34].

mVMR across ecosystems

We subsequently calculated mVMRs in additional metagenomes (for a total of 133) covering a wide biome diversity including marine and lake sediments, soils, microbial mats growing in various conditions (halophilic to thermophilic), microbialites and several types of metazoan-associated microbiomes (Extended Data Fig.2). Many of those metagenomes were retrieved from public databases (Supplementary Table 8) but 40 of them were generated in our laboratory, notably from poorly studied ecosystems. To avoid potential confounding factors derived from sequencing techniques and bioinformatic analysis, all chosen metagenomes were sequenced using the same technology and treated uniformly starting from raw reads. Cellular and viral counts varied substantially across the different metagenomes but both, USCG and MCP normalised counts (RPKM), were generally consistent for the two employed annotation strategies, Prokka and OrfM (Extended Data Fig.5). Also the broader taxonomic assignment into (mostly) free-living archaea and bacteria, symbiotic CPR and DPANN taxa as well as lytic virus (VPs and intracellular) and lysogenic (provirus) categories was in general good agreement between the two annotation strategies although a variable proportion of counts remained unassigned by OrfM (Supplementary Fig.2). CPR and DPANN counts outnumbered viral counts only in a small fraction of the studied metagenomes (Extended Data Fig.5; Supplementary Fig.2), suggesting that viral parasitism is generally more prevalent than prokaryotic parasitism.

At finer taxonomic scales, the diversity and relative proportions of cellular organisms and viruses were overall congruent with those reported for the different types of ecosystems and the specific studied metagenomes (Supplementary Table 8). Bacterial taxa were largely dominant across biomes, although archaea (haloarchaea, Euryarchaeota) dominated in 32% saltern ponds. Archaea were also abundant in marine sediments[35], several microbial mats[36] and the surface Mediterranean waters, the latter dominated by Marine Group II archaea (Extended Data Fig.4b). Diverse Alpha- and Gammaproteobacteria were highly abundant in marine plankton but also in microbialites and, to a lesser extent, microbial mats. Cyanobacteria were also present in considerable proportions in these ecosystems. Actinobacteria and Acidobacteria were dominant in soils. Some actinobacterial taxa were also abundant in the freshwater system[33]. Microbial eukaryotes consistently occurred in most ecosystems, albeit in minor proportions. They were more abundant in plankton samples and microbialites; diverse microalgae, notably diatoms and chlorophytes dominated the eukaryotic component (Extended Data Fig.4b). Diverse viral families were detected. As expected, dsDNA viruses largely dominated the studied microbial ecosystems, followed by ssDNA viruses (Extended Data Fig.6; Supplementary Table 9). In agreement with the pronounced dominance of bacteria as potential hosts in most biomes, the dominant identified VP-lytic viral families belonged to the class *Caudoviricetes* (tailed bacteriophages, with HK97-fold MCP; in particular members of the old families *Myoviridide*, *Siphoviridae* and *Podoviridae*), which collectively represent ~74% (n=8281) of all identified virus contigs across ecosystems. Members of *Caudoviricetes* also infect archaea[37]. The bacteriophage family *Microviridae*, with ssDNA genomes, also was highly represented in some ecosystems (Extended Data Fig.6b). Among eukaryotic viruses, the

most represented were members of the *Phycodnaviridae*, which typically parasitize algae, and *Mimiviridae*, which infect other protists (e.g. amoeba)[38].

The mVMR estimates across biomes obtained with the two annotation strategies were generally consistent (Fig.4) although Prokka-based annotations tended to yield higher mVMRs in a few samples apparently exhibiting viral blooms. Although the inferred mVMR values were on average greater than 1 (global average 1.97, i.e. ~2 viruses per cell), cells outnumbered viruses (mVMR <1) in more than one third of the metagenomes (Fig.4). In only 2-5 (depending on the annotation method) metagenomes of the 133, viruses exceeded cells by more than one order of magnitude. The rest of the metagenomes displayed mVMR values between 1 and ~7. In terms of the grand ecosystem type, mVMR were the highest in aquatic ecosystems (average ~2.7) and the lowest in animal-associated microbiomes (~1.4; Fig.5a). Microbial mats and soils/sediments exhibited intermediate average values, the latter displaying a wider value distribution with a higher variance, in agreement with previous observations[19-21].

Applying similar criteria to those of above-mentioned mock and aquatic metagenomes, a fraction of the identified MCPs corresponded to putative proviruses which, on average, were less abundant than cells (ratios provirus:cell ranging from 0.27 to 0.43). Thus, cells in these microbial ecosystems appeared to contain on average less than one identifiable MCP-encoding provirus. Although the number of proviruses per cell is generally considered to be one or higher[39], the number of proviruses per cell in natural, including extreme, environments, is unknown. Also, the assignment of MCPs to proviruses based on the cellular annotation by Prokka is likely imperfect as it relies on current reference (provirus-containing) genomes in public databases. This should improve with the inclusion of more reference genomes, notably of uncultured microbes. As a consequence, the obtained counts of proviral MCPs likely underestimate the actual number of proviruses, which are counted as viruses (hence not impacting the total viral count). These caveats notwithstanding, our estimates might more closely reflect a real trend given that metagenomic contigs (>5 genes) encoding provirus-assigned MCPs have consistently and significantly ($p < 2.22e{-}16$) higher proportions of cellular genes than contigs with a (lytic) viral-assigned MCP (Extended Data Fig.7a). The proportion of lysogenic versus lytic (including dispersing VP) virus forms varied across metagenomes, with the highest ratios found in host-associated microbiomes and the lowest in marine plankton microbiomes (Extended Data Fig.7b). This trend was also visible in terms of grand ecosystem types. Host-associated microbiomes displayed significantly higher fractions of lysogens than, in that order, microbial mats, soils/sediments and plankton (Fig.5b). This observation seems to support the hypothesis that lysogeny increases as a function of cell density[6,23]. It might be thought that viruses from multicellular-host microbiomes are better identified, thus biasing these estimates, because animal microbiomes have been studied in greater depth. However, planktonic communities are comparatively better explored than soil/sediment and, especially, microbial mat communities, the latter often displaying up to 40% of poorly characterized archaeal and bacterial taxa in below-surface layers[36]. Nonetheless, plankton showed the lowest fraction of proviruses (Fig.5b). In addition, we used specific HMMs targeting MCPs of diverse viruses identified in our metagenomes. Taking all this evidence together, it appears likely that the observed lysogenic versus lytic viral ratios reflect a real trend.

To look for potential determinants explaining the distribution of viruses and cells across biomes, we carried out principal component analyses (PCA) of the studied metagenomes using diversity-related parameters, namely richness, alpha-diversity and evenness indices for cells and viruses, including putative proviruses (Methods; Supplementary Figs.3-4). The results obtained with OrfM and Prokka-based annotation strategies were very similar (Fig.6 and Extended Data Fig.8). As expected, cell alpha-diversity correlated with microbial mats and soils, being lower in host-related microbiomes, whereas viral diversity best correlated with marine and freshwater plankton. Cell evenness mostly correlated with some soil and microbialite samples, whereas viral evenness correlated with freshwater sediment samples (Fig.6a; Extended Data Fig.8a). Globally, alpha-diversity values estimated for cells, lytic and lysogenic viruses were significantly different among specific (Fig.6b; Extended Data Fig.8b) and grand ecosystem (Extended Data Fig.9a-b) types. Within grand ecosystem type, the significance of alpha-diversity differences between cells, lytic viruses and proviruses varied depending on the ecosystem, microbial mats showing higher statistical significance largely owing to lower variability of (lytic) viral diversity as compared with plankton and soils/sediments (Extended Data Fig.9c-d). The alpha-diversity of (VP, lytic) viral plankton was the highest followed, albeit at a distance, by sediments, both freshwater and marine (Fig.6b; Extended Data Fig.8b). This trend, along with higher viral abundances, is consistent with the notion that water masses constitute a reservoir of dispersing viral particles. By contrast, highly structured and heterogeneous environments are less prone to host a wide and abundant diversity of dispersing forms because, in the absence of their specific hosts, these particles can be easily used as food or otherwise degrade. Freshwater and marine sediments would appear as intermediates, receiving a constant input of sedimenting VPs, which can then enter a decay and degradation process associated to local prokaryotic communities.

Discussion

In this work, we developed and benchmarked on mock metagenome datasets a metagenome-based VMR metric based on virus and cell markers that is applicable throughout microbial ecosystems in a comparative manner. Our mVMR metric is a priori more accurate for the inference of actual viral-microbe ratios because, in comparison with fVMR values based on fluorescently labelled VLPs only, it accounts also for the actively infecting, intracellular lytic viruses as well as lysogenic forms. In addition to being more viral-inclusive, our approach is free from potential biases affecting fVMR[4,12-14,16-18]. In addition to identifying specifically stained vesicles and capsids containing non-viral DNA (transducing phages or gene transfer agents), fluorescence-based methods are affected by biases stemming from the differential treatment of contrasting sample types (water, soil, sediment, mucilaginous mats) and can include non-specifically stained organic and/or mineral particles from complex ecosystems. These artefacts may not only affect VLP counts, but possibly also cell counts, casting doubt on total cell estimates based on these methods in sediments, soils or the subsurface and, hence, global virus and cell counts[3,4]. Furthermore, because a sizable fraction of viral particles are defective[40-42], the real number of infective viruses derived from VLP counts remains uncertain.

Our approach is not fully devoid of potential biases that might result in slight relative over- or underestimates of cells and/or viruses, such as differences in cell ploidy or divergent viral and cellular individual gene markers. However, as metagenome sequencing and the exploration

of cellular and viral diversity across ecosystem types improve, detection problems linked to sequence divergence should be gradually solved. Obviously, being metagenome-based, our approach does not allow for the quantification of RNA viruses and, for reverse-transcribing viruses, only the replicative DNA intermediates can be identified. Unfortunately, a similar approach based on metatranscriptomic counts is not feasible because USCGs can be expressed at drastically different levels in different cells, preventing them to be used as markers for individuals. Like DNA viromes uncovered by metagenomic studies[9,43], the diversity of the global RNA virome is rapidly expanding thanks to metatranscriptomic analyses[44-46]. Nonetheless, dsDNA viruses dominate the prokaryotic virome[38] and, although several novel families of RNA viruses seem to infect bacteria[44,45], bacterial viruses jointly comprise only 3.4% of the RNA viral OTUs[44]. Therefore, given that most microbial ecosystems are overwhelmingly dominated by prokaryotes, mVMR estimates based on gene-marker proxies for individual viruses and cells, even if in need of further refinement (especially with progressively expanding databases), are likely to be the best current approximation of the real VMR, at least at the order-of-magnitude level.

This work shows that mVMRs are higher in planktonic systems approaching, albeit not necessarily reaching, the values of ~10:1 that are traditionally assumed to be the minimal values in current (fVMR-based) estimates. Water masses therefore seem to represent reservoirs of VPs that can more easily escape from cellular predation, whereby nucleic acids serve as an excellent source of P and N, even if the uptake of specific lytic viral nucleic acids comes at a cost[47]. From this perspective, aerosols likely are another huge reservoir of VPs. By contrast, in higher-density biomes, mVMR appear substantially lower, consistent with higher viral dispersal limitation and degradation rates, as well as with the idea that VMR declines as a function of microbial cell densities[7]. Even if mVMR are highly variable, especially in sediments and soils, mVMR estimates across non-aquatic ecosystems barely reach 2, being especially low in cell-dense multicellular host-associated microbiomes (<1). Estimates of putative proviruses versus dispersing and intracellularly infecting lytic viruses seem to follow the opposite trend (Fig.5), supporting the hypothesis that lysogeny increases with cell density[5,6,48]. Overall, our mVMR estimates across ecosystems question the current assumption that viruses outnumber cells at least by one order of magnitude. Even if ocean waters occupy an important fraction of Earth's volume, continental biomes together with the deep subsurface undoubtedly host most of the planet's biomass and cells[49,50]. Accordingly, although viruses likely are the most diverse biological entities on Earth[8,9], notably because most cellular organisms can be infected by different viruses, they might quantitatively be almost at equal footing with cells. In addition, deciphering the proportion of actively infecting viruses should provide more useful information about the maximum load of viral entities that microbial communities can support before ecosystems collapse. Future exploration of these poorly studied parameters should help understanding the dynamics of viral-cell interactions in the biosphere.

References

1 Bergh, O., Børsheim, K. Y., Bratbak, G. & Heldal, M. High abundance of viruses found in aquatic environments. *Nature* **340**, 467-468 (1989).

2 Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801-812 (2007).
3 Cobián Güemes, A. G. *et al.* Viruses as Winners in the Game of Life. *Annu Rev Virol* **3**, 197-214 (2016).
4 Mushegian, A. R. Are there 10^{31} virus particles on Earth, or more, or less? *J. Bacteriol.*, JB.00052-00020 (2020).
5 Knowles, B. *et al.* Variability and host density independence in inductions-based estimates of environmental lysogeny. *Nat Microbiol* **2**, 17064 (2017).
6 Weitz, J. S., Beckett, S. J., Brum, J. R., Cael, B. B. & Dushoff, J. Lysis, lysogeny and virus–microbe ratios. *Nature* **549**, E1-E3 (2017).
7 Wigington, C. H. *et al.* Re-examination of the relationship between marine virus and microbial cell abundances. *Nat Microbiol* **1**, 15024 (2016).
8 Paez-Espino, D. *et al.* Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* **7**, 157 (2019).
9 Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* **18**, 125-138 (2020).
10 Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548 (1999).
11 Suttle, C. A. Viruses in the sea. *Nature* **437**, 356-361 (2005).
12 Forterre, P., Soler, N., Krupovic, M., Marguet, E. & Ackermann, H. W. Fake virus particles generated by fluorescence microscopy. *Trends Microbiol.* **21**, 1-5 (2013).
13 Schatz, D. *et al.* Communication via extracellular vesicles enhances viral infection of a cosmopolitan alga. *Nat Microbiol* **2**, 1485-1492 (2017).
14 Toyofuku, M., Nomura, N. & Eberl, L. Types and origins of bacterial membrane vesicles. *Nature Reviews Microbiology* **17**, 13-24 (2019).
15 Liu, J. *et al.* Archaeal extracellular vesicles are produced in an ESCRT-dependent manner and promote gene transfer and nutrient cycling in extreme environments. *ISME J* **15**, 2892-2905 (2021).
16 Lang, A. S., Westbye, A. B. & Beatty, J. T. The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol* **4**, 87-104 (2017).
17 Klauth, P., Wilhelm, R., Klumpp, E., Poschen, L. & Groeneweg, J. Enumeration of soil bacteria with the green fluorescent nucleic acid dye Sytox green in the presence of soil particles. *J. Microbiol. Methods* **59**, 189-198 (2004).
18 Belilla, J. *et al.* Active Microbial Airborne Dispersal and Biomorphs as Confounding Factors for Life Detection in the Cell-Degrading Brines of the Polyextreme Dallol Geothermal Field. *mBio* **13**, e0030722 (2022).
19 Parikka, K. J., Le Romancer, M., Wauters, N. & Jacquet, S. Deciphering the virus-to-prokaryote ratio (VPR): insights into virus-host relationships in a variety of ecosystems. *Biol. Rev. Camb. Philos. Soc.* **92**, 1081-1100 (2017).
20 Danovaro, R. *et al.* Viriobenthos in freshwater and marine sediments: a review. *Freshw. Biol.* **53**, 1186-1213 (2008).
21 Williamson, K. E., Fuhrmann, J. J., Wommack, K. E. & Radosevich, M. Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annu Rev Virol* **4**, 201-219 (2017).
22 Wei, M. & Xu, K. New insights into the virus-to-prokaryote ratio (VPR) in marine sediments. *Front. Microbiol.* **11**, 1102 (2020).
23 Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 466-470 (2016).
24 Mendell, J. E., Clements, K. D., Choat, J. H. & Angert, E. R. Extreme polyploidy in a large bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 6730-6734 (2008).

25 Volland, J. M. *et al.* A centimeter-long bacterium with DNA contained in metabolically active, membrane-bound organelles. *Science* **376**, 1453-1458 (2022).
26 Ludt, K. & Soppa, J. Polyploidy in halophilic archaea: regulation, evolutionary advantages, and gene conversion. *Biochem. Soc. Trans.* **47**, 933-944 (2019).
27 Touchon, M. & Rocha, E. P. Coevolution of the Organization and Structure of Prokaryotic Genomes. *Cold Spring Harb. Perspect. Biol.* **8**, a018168 (2016).
28 Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629-645 (2018).
29 Dombrowski, N., Lee, J. H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366**, fnz008 (2019).
30 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
31 Woodcroft, B. J., Boyd, J. A. & Tyson, G. W. OrfM: a fast open reading frame predictor for metagenomic data. *Bioinformatics* **32**, 2702-2703 (2016).
32 Pommier, T. *et al.* Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquat. Microb. Ecol.* **61**, 221-233 (2010).
33 David, G. M. *et al.* Small freshwater ecosystems with dissimilar microbial communities exhibit similar temporal patterns. *Mol. Ecol.* **30**, 2162-2177 (2021).
34 Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11463-11468 (2013).
35 Dong, X. *et al.* Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nature Communications* **10**, 1816 (2019).
36 Gutierrez-Preciado, A. *et al.* Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nature ecology & evolution* **2**, 1700-1708 (2018).
37 Liu, Y. *et al.* Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. *PLoS Biol.* **19**, e3001442 (2021).
38 Koonin, E. V. *et al.* Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **84** (2020).
39 López-Leal, G. *et al.* Mining of Thousands of Prokaryotic Genomes Reveals High Abundance of Prophages with a Strictly Narrow Host Range. *mSystems* **7**, e00326-00322 (2022).
40 Rezelj, V. V., Levi, L. I. & Vignuzzi, M. The defective component of viral populations. *Curr. Opin. Virol.* **33**, 74-80 (2018).
41 Vignuzzi, M. & López, C. B. Defective viral genomes are key drivers of the virus–host interaction. *Nat Microbiol* **4**, 1075-1087 (2019).
42 Laurenceau, R., Raho, N., Forget, M., Arellano, A. A. & Chisholm, S. W. Frequency of mispackaging of Prochlorococcus DNA by cyanophage. *The ISME Journal* **15**, 129-140 (2021).
43 Zhang, Y. Z., Chen, Y. M., Wang, W., Qin, X. C. & Holmes, E. C. Expanding the RNA Virosphere by Unbiased Metagenomics. *Annu Rev Virol* **6**, 119-139 (2019).
44 Dominguez-Huerta, G. *et al.* Diversity and ecological footprint of Global Ocean RNA viruses. *Science* **376**, 1202-1208 (2022).
45 Neri, U. *et al.* Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037.e4018 (2022).
46 Edgar, R. C. *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nature* (2022).
47 Menge, D. N. & Weitz, J. S. Dangerous nutrients: evolution of phytoplankton resource uptake subject to virus attack. *J. Theor. Biol.* **257**, 104-115 (2009).

48 Weiss, R. A. Exchange of Genetic Sequences Between Viruses and Hosts. *Curr. Top. Microbiol. Immunol.* **407**, 1-29 (2017).
49 Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6578-6583 (1998).
50 Takamiya, H., Kouduka, M. & Suzuki, Y. The Deep Rocky Biosphere: New Geomicrobiological Insights and Prospects. *Front. Microbiol.* **12**, 785743 (2021).

FIGURE LEGENDS

Fig.1. Estimated metagenome virus-microbe ratios (mVMR) across mock metagenomes with increasing VMR. a, Statistics of the different mock metagenomes generated with an increasing number of virus genomes (V000 to V125) as compared to the number of microbial cells (M00 – M01 – M10). **b**, Expected versus observed viral ratios versus cells (cells normalised to 1) in the different mock metagenomes as a function of the annotation method (OrfM vs. Prokka) used. Cells were counted using universal single copy genes (USCGs) and viruses using major capsid proteins (MCPs). VP-lytic_MCPs include viral particles (VP) and cell-infecting lytic viruses; All_MCPs include also MCPs from lysogenic viruses (proviruses). **c**, Normalised counts (RPKM) of VP + lytic and proviruses per cell (normalised to 1) observed in mock metagenomes with our pipeline using OrfM and Prokka annotation strategies.

Fig.2. Strategies used to generate metagenomes in this study. Metagenomes from different planktonic fractions were obtained after sequential filtration of the same water mass, purification of DNA for each size fraction and proportional mixing per volume unit. Two metaviromes (particles <0.2 µm) were sequenced for control. Metagenomes of non-aquatic samples were generated after the in situ cell lysis and direct DNA purification prior to sequencing. * The metagenome WC-EdVall corresponds to the size fraction 50 kDa – 100 µm.

Fig.3. Epifluorescence- and normalised metagenome-based counts and ratios of cells and viruses in selected aquatic ecosystems across a salinity gradient. a, Flow-cytometry counts of cells and viral-like particles (VLPs) and estimated virus-to-microbe ratios (histogram bars). Sample names indicate the plankton size fraction considered (Unf, all size fractions; <200, <100, <30 and <5, fractions below those numbers in µm; <0.2, fraction between 50 kDa and 0.2 µm; FE, final eluate <50 kDa; see Supplementary Table 1). **b**, Normalised counts (RPKM) of viruses (MCPs corresponding to VP + lytic and lysogenic –proviruses– forms) and cells from the three cellular domains (USCGs) in global (cells + viruses) metagenomes from the same original aquatic samples as in (a). Metagenomes correspond to size plankton fractions indicated by their suffixes: -all, fraction between 50 kDa and 200 µm (100 µm for EV); 02v, fraction between 50 kDa and 5 µm (see Extended Data Fig.2 and Supplementary Tables 4-5). Two metaviromes (fraction 50 kDa-0.2 µm) were sequenced for Mediterranean (MedS, MedD) plankton samples as internal control. **c**, Metagenome-based normalised ratios of lytic (VP, actively infecting) and lysogenic (proviruses) viruses per cell (bacteria + archaea + eukaryotes). RPKM, reads per kilo base per million mapped metagenome reads. Lytic and lysogenic mVMRs were estimated after OrfM and Prokka annotation strategies.

Fig.4. Metagenome-based normalised ratios of viral particles/lytic and lysogenic viruses per cell across ecosystem types. Viruses and cells were counted using, respectively, MCPs and USCGs as proxies and cell counts were normalised to 1. All the metagenomes were analysed using the same pipeline from raw reads. **a**, Normalised estimates of virus-cell ratios using the OrfM annotation strategy. A zoom of the zone marked by the green rectangle is shown below. **b**, Similar estimates based on Prokka annotation. RPKM, reads per kilo base per million mapped metagenome reads.

Fig. 5. mVMR and ratios of proviruses versus lytic viruses and viral particles per grand ecosystem type. a, Boxplot indicating mVMR. **b**, Boxplot showing the provirus:virus (VP-lytic) ratios. Values are based on Prokka annotation. Significant differences between ratios are indicated.

Fig. 6. Diversity and evenness of viruses, proviruses and cells across ecosystems. a, Principal component analysis of different metagenomes maximizing the variance explained by viruses (VPs and infecting lytic-viruses) and proviruses (MCPs counts in RPKMs) and microbial cells (USCG counts in RPKMs). **b**, Boxplot showing the alpha diversity of viruses (VP, lytic), proviruses and cells as a function of ecosystems (OrfM annotation strategy; a boxplot from data derived from Prokka annotation can be seen in Extended Data Fig. 7). Diversity differences between the three types were significant ($p<0.001$ or 0.0001). These differences were also seen across and within large ecosystem categories (Extended Data Fig. 8). VP, viral particle (virion).

Methods

Mock metagenomes

Ten mock metagenomes were built from a selection of 50 viral and 50 cellular genomes retrieved from the NCBI reference sequence database RefSeq (https://www.ncbi.nlm.nih.gov/) and GTDB (https://gtdb.ecogenomic.org/). Genomes were selected to partly mimic a simplified microbial community of marine plankton that included viruses, archaea, bacteria and eukaryotes (Supplementary Table 3). To simulate the output of Illumina HiSeq sequencing, all genomic sequences were fragmented into 10 million reads of 100 bp using cMessi[51]. Eight mock metagenomes were built by varying the relative abundance of the viral vs. cellular reads in a range from 10 cells per viral genome to 125 viral genomes per microbial cell (V001M10 to V125M01; names reflect the virus:microbe ratio). The two additional metagenomes contained either none of the selected viral genomes (V000M01) or only viral sequences (V001M00). The generated reads were assembled with Metaspades[52] v3.13.1 with default parameters and k-mer iteration cycles of "21,25,31,35,41,45,51,55". Assembly statistics are shown in Supplementary Table 4. Gene annotation was performed on assembled contigs with Prokka[30] v1.14.5 in metagenome mode (contigs >200 bp). In parallel, to more exhaustively identify gene/protein domains, all ORFs were predicted with OrfM[31].

Metagenome-based counting of cells and viruses

To count cells, we used a set of 15 universal single-copy genes (USCGs; Supplementary Table 1) highly conserved across the three cellular domains, including the reduced and often fast-evolving members of the Candidate Phyla Radiation (CRP, Patescibacteria) and the DPANN archaea. USCGs were identified via their PFAM motifs (Pfam-A database v3.1b2)[53] and searched in the two datasets (Prokka-predicted genes and OrfM-predicted ORFs) with HMMER hmmsearch[54] v3.2.1 with the trust cut-off threshold. To assign relevant hits to taxa, they were analysed using Diamond[55] v0.9.25.126 against the NCBI non-redundant database formatted by Kaiju[56] v1.7.1: kaiju_db_nr_euk.faa, which contains all NCBI reference sequences for archaea, bacteria, eukaryotes and viruses. Best hits with more than 35% identity over at least 70% of query lengths were included in a best-hit table in kaiju-like output format using an ad hoc Perl script (mimicKaijuOutput.pl). They were then assigned to taxa using the kaiju script addTaxonName. Most genomes used to build the mock metagenomes had only one copy of the set of 15 USCGs (Supplementary Table 3). Accordingly, the average of counts for the selected 15 USCGs in metagenomes represents an adequate proxy for individual cell counts. Five of the chosen ribosomal markers are known to have two copies in halophilic archaea (Extended Data Fig.4) and, accordingly, we divided by two haloarchaeal counts for them prior to averaging total counts. To estimate the number of viruses in metagenomes, we used as a proxy genes encoding major capsid proteins (MCPs). We used a total of 132 HMM profiles covering the MCPs of (almost) all families of DNA viruses or viruses with intermediate DNA stages (ssDNA, dsDNA, RT-DNA) known to infect bacteria, archaea and eukaryotes. For the analysis of the mock and planktonic metagenomes, we specifically included HMM profiles of environmentally abundant marine viruses discovered by metagenomics (e.g., viruses of MGII Euryarchaeota, MGI Thaumarchaeota, marine Thermoplasmata, etc) and, to capture viruses of the rest of metagenomes (see below), we built specific HMM profiles. For this,

metagenome contigs assigned to viruses whose MCPs were not captured by existing HMM were used to build new HMM profiles (a total of 132; Supplementary Table 2). For reference MCP sequences representing each family of DNA viruses, homologs were retrieved from the UniRef database, filtered to 30% sequence identity (3 iterations, inclusion value of E=1e-3) and aligned using hhblits[57], part of the hhsuite package 3.3.0 (Ref.[58]). The resulting alignments were converted to profile HMMs using hmmbuild included in the HMMER package v3.3.2 (Ref.[59]). Most MCP HMMs retrieved genes were functionally annotated as either capsid or hypothetical proteins and were assigned to viruses by our taxonomic assignment pipeline. However, several HMMs retrieved genes annotated as (diverse) encapsulins and simultaneously assigned to bacteria (Supplementary Fig. 1). The HMMs preferentially retrieving bacterial encapsulins were excluded from further analyses. MCPs assigned to cellular taxa that had "hypothetical protein" function (Supplementary Fig. 1a) or phage-related function were counted as proviruses. Manual inspection of randomly selected large contigs containing putative proviruses identified typical cellular genes nearby putative provirus MCPs. MCP HMM profiles that consistently retrieved genes assigned to cellular taxa and had clear cellular functions were discarded. A final set of 132 MCP HMMs was retained and used to search against all the Viral Ref Seq (NCBI) using HMMER hmmsearch[54] v3.2.1 with an e value of 1e-11 threshold. One gene can be attributed to different MCPs, so the gene IDs were hashed retaining the MCP assignment with the lowest e-value. Since multiple MCPs can occasionally be found in large viral genomes, we counted only one MCP per contig and considered that, under these constraints, 1 MCP represents 1 virus. To assess metagenome-inferred virus to microbe ratios (mVMR) in mock metagenomes, nucleotide sequences of identified cellular USCGs and viral MCPs were indexed with Bowtie[60] v2.3.5.1 and the original mock metagenome reads mapped back onto them. Mapped reads were retrieved with Samtools[61] v1.9 and, for each gene, the Reads Per Kilobase (of gene sequence) per Million of mapped metagenome reads (RPKM) were calculated using ad hoc Perl scripts, allowing normalization for gene length and sample sequencing depth. Subsequently, we used the USCG that recovered the most RPKMs per metagenome and the sum of MCP RPKMs to obtain normalised mVMRs. For an easier visualization of the proportion of viruses per microbial cell, we further normalised the viral RPKMs with respect to the number of cells (cell value = 1). Plots were generated with ad hoc R scripts using the ggplot2 package[62].

Water sampling, filtration, DNA purification and mixes for metagenome sequencing

Samples from the Mediterranean Sea were collected on 27/02/2019 at two depths (20 and 40 m; bottom depth 220 m) from a previously studied site off the Alicante coast (38°4'6.64"N, 0°13'55.18"W)[63]. The CTD (Seabird) data showed a typically winter-mixed water column, with homogeneous temperature (14.8°C), pH (8.4) and dissolved oxygen (8.5 ppm) throughout at least 60 m depth. Water was collected through a hose with a 200-µm Nitex mesh-capped end attached to the CTD and directly pumped onto a serial filtration system mounted with 20, 5 and 0.22 µm pore-size polycarbonate filters (Millipore). Biomass-containing filters were frozen on dry ice for transport to the laboratory prior to immediate DNA purification. The resulting <0.2 µm filtrate (200 l) was recovered in 50-l carboys and used for tangential filtration, which took place overnight on arrival at 4°C (cold room) using 6 Vivaflow200 filtration units (Sartorius) operating in two parallel series. The volume of concentrated viral fractions (>50 kDa) was

further reduced using Amicon® Ultra 15 ml Centrifugal Filters (Merck-Millipore) prior to phenol/chloroform extraction. Water samples (10 l) with increasing salt concentration (6, 14 and 32%) were collected in carboys on 28/02/2019 from the Bras del Port solar saltern ponds: Sal6 (38°11'52.30"N, 0°36'13.15"W), Sal14 (38°12'4.74"N, 0°35'43.14"W), Sal32 (38°11'48.00"N, 0°35'42.80"W). Brine samples were taken to the Alicante laboratory for immediate serial filtration (as above) and DNA purification. Freshwater (10 l) from a shallow lake in North-western France (Etang des Vallées; 48°41'23,0''N, 01°54'59.2''E)[64] was collected on 18/03/2019. Filtration was immediately carried out at the Orsay laboratory as above except we used filters of 100, 30, 5 and 0.22 µm pore-size, to avoid the many suspended plant material-derived particles in this shallow lake) In all cases, 4-ml aliquots of the different size-fraction filtrates and the final eluate (<50 kDa) were taken for flow cytometry and fixed overnight at 4°C with 0.5% glutaraldehyde then frozen in dry ice and stored at -80°C (Supplementary Table 6). DNA from cellular fractions was purified from filters. These were cut into small pieces with a scalpel, added to 5 ml CTAB lysis buffer (SERVA) with 1 mg/ml lysozyme and incubated at 37°C for 1 h. After addition of proteinase K (0.2 mg/ml) and vortexing, the mix was incubated at 55°C for one additional hour prior to phenol/chloroform/isoamyl alcohol (25:24:1, pH 8) and chloroform/isoamyl alcohol extraction. Nucleic acids were precipitated overnight at -20°C with 1/10 volume of 3 M sodium acetate, pH 5.2, and 2.5 volumes of absolute ethanol. After centrifugation at 4°C, 13000 x g for 30 min, the pellet was washed with 70% ethanol, centrifuged again and speed-vac dried before resuspension in 10 mM Tris-HCl, pH 8.0. To purify DNA from viral concentrates, we incubated them with 20 mM EDTA, 50 ng/µl proteinase K and 0.2% SDS (final concentrations) for 1h at 55°C and proceeded with the phenol/chloroform extraction as before. To recover the maximum of the aqueous phase, we used MaXtract™ High Density (Qiagen). DNA from the different cell fractions was fluorometrically quantified using Qubit® and mixed proportionally according to the same amount of original water sample as indicated in Supplementary Table 7. For all samples, we prepared a DNA mix 'all' which included DNA from the viral and all cellular fractions below 200 µm (100 µm for EdV) and a mix '0.2v' which contained DNA from viral particles and cells smaller than 5 µm (prokaryotes and small protists). All these mixes were used for metagenome sequencing. As an internal control for our fractionation process, we also sequenced DNA from the viral fraction (metaviromes) from the Mediterranean plankton (Extended Data Fig 2, Supplementary Table 8).

Flow cytometry

We quantified virus-like particles (VLP) and cells (up to ~10 µm) by flow cytometry (FCM) in the glutaraldehyde-fixed samples spanning different salinities and size fractions described above (Supplementary Table 6). We used a CytoFLEX S (Beckman Coulter) equipped with two active diode lasers (488 nm and 561 nm) following well-established protocols[65,66]. The flow cytometer settings and gains were, for VLP detection, forward scatter (FSC) 500, side scatter (SSC) 500 and green fluorescence (FITC) 400 with a threshold set to FITC 260; and, for cell detection, FSC 250, SSC 700 and FITC 70 with a threshold set to FITC 1000. The sheath fluid consisted of 0.1 µm-filtered Milli-Q water. Multifluorescent beads of 0.2 and 1µm (Polyscience) were used as size reference for VLP and cells. Samples were thawed on ice and diluted up to 1000 times for VLP counts and 10 times for cell counts with 0.02 µm-filtered

autoclaved Milli-Q water, TE buffer (Sigma) or seawater. VLP and cells were stained with SYBRGreen I (Sigma) for 10 min in the dark at 80°C and let cool down for 5 min at room temperature[66]. Background noise was systematically checked on blanks (Extended Data Fig.3). Samples were recorded with an event rate per second of 100-1,000 (VLP) and <300 (cells). For VLP, we defined three gates (V1, V2, V3) on a FITC-A vs SSC-A plot (Extended Data Fig.3). The VLP signal intensity was determined on marine plankton controls, with gates set to exclude background based on FITC signal. Cells were stained with SYBRGreen I for 10 min at room temperature in the dark. For cell counts, we established two gates, HDNA and LDNA, on a FITC-A *vs* SSC-A plot, as previously described[65]. Sample counts were analysed with CytExpert and Kaluza softwares (Beckman Coulter). VLP and cell counts were obtained by correcting the measured total counts for noise (typically having the lowest green fluorescence) using blanks. VLP and cell abundances were respectively expressed as VLP.ml^{-1}or Cells.ml^{-1} (Supplementary Table 6).

DNA purification of sediment and microbial mats, sequencing and selection of natural metagenomes

We also included in this analysis metagenomes of freshwater sediments and microbial mats from previous and ongoing studies in our laboratory (Extended Data Fig 2, Supplementary Table 8). DNA from Lake Baikal sediment samples collected in July 2017 was purified using the Power Soil™ DNA purification kit (Qiagen, Germany)[67]. DNA from microbial mat samples collected from the Salada de Chiprana (Spain, December 2013), lakes Bezymyannoe and Reid (Antarctica, January 2017) and several hot springs around Lake Baikal (Southern Siberia, July 2017), fixed *in situ* in ethanol as previously described[36], was purified using the Power Biofilm™ DNA purification kit (Qiagen, Germany). DNA was quantified using Qubit®. Sequencing was performed using paired-end (2x125 bp) Illumina HiSeq by Eurofins Genomics (Ebersberg, Germany). In total, we generated 14 metagenomes from planktonic fractions coming from increasingly salty waters (freshwater, marine and three solar saltern ponds at 6-14-32% salt), 8 metagenomes from freshwater sediment and 23 metagenomes of diverse microbial mats. To these, we added 13 microbial mat and 7 metagenomes from other microbial mats and microbialites (18 of which were previously generated in our laboratory), 11 metagenomes from freshwater and soda lake sediments, 7 marine sediment metagenomes, 18 soil metagenomes and 32 animal-associated microbiomes from diverse organs and animals (honeybees, frogs, humans), all of which were Illumina-sequenced. In total, we included 133 metagenomes in this study, 40 metagenomes generated in our laboratory and 93 metagenomes from other published studies[35,36,68-79] (Extended Data Fig.2; an extended description, GenBank accession numbers and statistics are provided in Supplementary Table 8).

Sequence analysis, annotation and taxonomic assignation of natural metagenomes

We treated the 133 metagenomes with exactly the same bioinformatics pipeline starting from raw Illumina sequences. Read quality was verified with FastQC[80] v0.11.8 and cleaned with Trimmomatic[81] v0.39, adjusting the parameters as needed (usually LEADING:3 TRAILING:3 MAXINFO:30:0.8 MINLEN:36) and eliminating the Illumina adapters if any. Clean reads were assembled with Metaspades[52] v3.13.1 with default parameters and k-mer iteration cycles of "21,25,31,35,41,45,51,55". A few complex metagenomes that failed to assemble in our

server were assembled with Megahit[82] v1.2.6 with default parameters except for a minimum contig length of 200 bp and starting kmer sizes of 23 to 93 with an increasing step of 10 (Supplementary Table 8). Gene annotation was performed with Prokka[30] v1.14.5 in metagenome mode (contigs >200 bp). We assigned coding sequences to PFAMs (Pfam-A database v3.1b2) with HMMER hmmsearch[54] v3.2.1 with the trust cut-off threshold. Taxonomy was assigned to each gene as described above for the mock metagenomes. Taxonomy was assigned to each gene as described above for the mock metagenomes. Taxon-assigned genes were then classified using an *ad hoc* Perl script (assignTaxa.pl) into the following major groups, further classified in three major lifestyle categories: i) CPR (Patescibacteria) bacteria and DPANN archaea (Supplementary Table 10), considered as cellular episymbionts/parasites strictly dependent on free-living organisms; ii) 'Other Archaea', 'Other Bacteria' and 'Eukaryotes', classed as free-living (FL), and iii) lytic and dispersing (VPs) viruses and lysogenic forms (essentially proviruses), as described above. Although some non-CPR bacteria are also parasites (e.g. of eukaryotes), these are likely to be minor in microbial ecosystems, including animal's healthy microbiomes and hence, would not affect counts at the order of magnitude scale. Likewise, some eukaryotes are parasites of other eukaryotes. However, eukaryotes were not very abundant as compared to prokaryotes and the vast majority of them belonged to non-parasitic lineages. An ad hoc Perl script (tax2PFAM.pl) linked each gene's taxonomy to its Pfam assignment. A subset of MCPs was classified as 'putative provirus' when the corresponding taxonomy was annotated as cellular. To further sustain provirus assignation, we determined the taxonomic affiliation of all the genes in MCP-containing contigs and their distribution visualized as boxplots with an R script using the ggplot2 package. Cellular genes were much more represented in contigs with putative provirus-assigned MCP. Metagenome-inferred proteins with >35% amino acid identity over >70% sequence length to viral genome-encoded proteins were assigned to the respective family, host and virus group. Viral taxonomy and Baltimore classification were determined with the VPF-Class tool[83] (Supplementary Table 9). MCP-containing contigs were used as an input for VPF-Class; we attributed a classification when values were over a membership ratio score of 0.75. The relative abundance of MCPs from these contigs in metagenomes were plotted using the R ggplot2 package visualizing the different viral families and Baltimore classification.

Diversity indices

To calculate diversity estimates, we selected the most abundant USCG per group of metagenome (Supplementary Table 11). Amino acid sequences for each ribosomal protein were retrieved with *ad hoc* Perl scripts for each metagenome group and clustered with cd-hit[84] at 99% identity (parameters -c 0.99, -n 5, -d 0, -T 16, -M 0). These clusters were considered operational taxonomic units (OTUs) and were assigned to CPR, DPANN, (non-DPANN) archaea, (non-CPR) bacteria and eukaryotes. For viruses, we used the whole set of viral MCPs under the premise that most viral genomes have only one MCP (see above). We obtained a total of 19,374 clusters for the cellular markers for OrfM and 8,610 for Prokka; and 67,874 viral MCP clusters for OrfM and 28,319 for Prokka. Abundance matrices of these clusters were built with an *ad hoc* Perl script and several diversity indices (alpha diversity, Shannon entropy, Pielou's evenness and Simpson's dominance) were calculated with an *ad hoc* R script using the Vegan package[85]. Indices for all metagenomes were plotted with ggplot2. To compare if

the difference between the viral and cellular diversity was significant, a Wilcoxon test was applied to the median of the alpha diversity per ecosystem, using an ad hoc R script with the ggpubr library. Visualization of the statistical test was represented as a box plot with the same R script and the ggplot2 library. Principal component analyses (PCA) of different diversity indices as a function of metagenomes were carried out with an ad hoc R script, and visualized using the ggplot2 package.

Data availability

Metagenome Illumina sequences generated in our laboratory and not yet published have been deposited in GenBank (National Center for Biotechnology Information) Short Read Archive with BioProject number PRJNA756245. Accession numbers for all used metagenomes are provided in Supplementary Table 8.

Code availability

All code and HMM profiles are available on GitLab (https://gitlab.com/anagtz/mvmrs). Individual frequency tables of USCG-based taxonomic diversity for all studied metagenomes can be retrieved at DOI 10.6084/m9.figshare.21341823.

51 Mende, D. R. *et al.* Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. *PLoS One* **7**, e31386 (2012).

52 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824-834 (2017).

53 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279-285 (2016).

54 Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200-w204 (2018).

55 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60 (2015).

56 Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**, 11257 (2016).

57 Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **9**, 173-175 (2011).

58 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

59 Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

60 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

61 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

62 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

63 Ghai, R. *et al.* Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J* **4**, 1154-1166 (2010).

64 Simon, M. *et al.* Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. *ISME J* **9**, 1941-1953 (2015).
65 Gasol, J. M. & del Giorgio, P. A. Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Scientia Marina* **64**, 197-224 (2000).
66 Brussaard, C., Payet, J., Winter, C. & Weinbauer, M. in *ASLO.* (eds S.W. Wilhelm, M.G. Weinbauer, & C.A. Suttle) 102-109.
67 Reboul, G. *et al.* Marine signature taxa and core microbial community stability along latitudinal and vertical gradients in sediments of the deepest freshwater lake. *The ISME Journal* **15**, 3412-3417 (2021).
68 Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nature Communications* **9**, 4999 (2018).
69 Zhang, X. *et al.* Metagenomics Reveals Microbial Diversity and Metabolic Potentials of Seawater and Surface Sediment From a Hadal Biosphere at the Yap Trench. *Front. Microbiol.* **9** (2018).
70 Ruuskanen, M. O. *et al.* Microbial genomes retrieved from High Arctic lake sediments encode for adaptation to cold and oligotrophic environments. *Limnology and Oceanography* **65**, S233-S247 (2020).
71 Vavourakis, C. D. *et al.* Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. *BMC Biol.* **17**, 69 (2019).
72 Ji, M. *et al.* Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* **552**, 400-403 (2017).
73 Xu, J. *et al.* The structure and function of the global citrus rhizosphere microbiome. *Nature Communications* **9**, 4894 (2018).
74 Wilbanks, E. G. *et al.* Microscale sulfur cycling in the phototrophic pink berry consortia of the Sippewissett Salt Marsh. *Environ. Microbiol.* **16**, 3398-3415 (2014).
75 Saghaï, A. *et al.* Comparative metagenomics unveils functions and genome features of microbialite-associated communities along a depth gradient. *Environ. Microbiol.* **18**, 4990-5004 (2016).
76 Rebollar, E. A. *et al.* The Skin Microbiome of the Neotropical Frog Craugastor fitzingeri: Inferring Potential Bacterial-Host-Pathogen Interactions From Metagenomic Data. *Front. Microbiol.* **9** (2018).
77 Regan, T. *et al.* Characterisation of the British honey bee metagenome. *Nature communications* **9**, 4995-4995 (2018).
78 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66 (2017).
79 Vangay, P. *et al.* US Immigration Westernizes the Human Gut Microbiome. *Cell* **175**, 962-972.e910 (2018).
80 Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **7**, 1338 (2018).
81 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
82 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3-11 (2016).
83 Pons, J. C. *et al.* VPF-Class: Taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics* **37**, 1805-1813 (2021).
84 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

85 Vegan: Community Ecology Package. R package version 1.17-9. (http://CRAN.R-project.org/package=vegan, 2011).

Acknowledgements

We thank José Manuel Haro Moreno for help during sampling in Alicante, Paola Bertolino for help during sampling in several other ecosystems, and Antonio Camacho y Antonio Picazo for CTD measurements during marine plankton sampling. We thank the Salinas Bras del Port for sampling permission and the UNICELL platform for flow cytometry analyses. We acknowledge support from the Moore-Simons Project on the Origin of the Eukaryotic Cell through Grant GBMF9739 (P.L.-G.) and the European Research Council Advanced Grants FP7 No. 322669 (P.L.-G.) and H2020 No. 787904 (D.M.).

Author contributions. P.L.-G. and D.M. conceived the study and designed experiments, organized and participated in sampling and successive filtration steps, did the tangential filtration and quantitatively purified DNA from the different cell and viral fractions. A.G.-P. participated in sampling and carried out all bioinformatic analyses, with informatics support by P.D. M.K. designed the HMM profiles for MCPs. M.C. carried out flow cytometry counts of viruses and cells with the assistance of L.J. M.L.-P. and F.R.-V. co-organized sampling in marine and solar saltern sites and helped with filtration on board. P.L.-G. wrote the manuscript with input from A.G.-P. and M.C. All authors read, commented and approved the manuscript.

Competing interests. The authors declare no competing interests.

Additional Information

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to P. López-García

Extended Data Figure Legends

Extended Data Fig.1. Predicted gene content of the generated mock metagenomes. a, prediction of genes and open reading frames (ORFs) based on, respectively Prokka and OrfM annotation of assembled contigs. **b**, Prediction of the relative proportion of different cellular taxa in the mock metagenomes based on the averaged number of selected universal single copy genes (USCGs). CPR, Candidate Phyla Radiation, host-dependent epibiotic bacteria; DPANN, clade of host-dependent epibiotic archaea. **c**, Relative proportion of cellular and viral entities estimated from Prokka-annotated mock metagenomes with USCGs and major capsid proteins (MCPs) as markers for individuals. **d**, the same estimation for OrfM annotation.

Extended Data Fig.2. Origin and main characteristics of analysed metagenomes. Extended information about metagenome origin and statistics are given in Supplementary Table 6. *metagenomes generated in our laboratory (this and previous or ongoing studies).

Extended Data Fig. 3. Cytograms of SYBR Green I-stained cells and virus-like particles (VLPs) in several aquatic environments spanning different salinities. Data are shown for plankton below 200 µm cell size (100 µm for EV). Three VLP subpopulations (V1-V2-V3) and two major cellular subpopulations containing high DNA (HDNA) and low DNA (LDNA) levels are defined. Blanks are also shown: MQF, Milli-Q water, TE, Tris EDTA buffer, SW, seawater. FE, final eluate (<50 kDa); b, fluorescent 0.2µm beads. More detailed sample descriptions are given in Extended Data Fig.2 and Supplementary Tables 4 and 6.

Extended Data Fig. 4. Normalised counts of selected universal single copy genes (USCGs) and relative abundance of major cellular taxa across metagenomes. a, Normalised counts of selected USCGs in aquatic metagenomes of coexisting viral and cell fractions. Suffixes in sample names indicate the plankton size fraction: _all, >50 kDa - 200 µm (100 µm in EdV); _02V, >50 kDa - 5 µm; _vir, >50 kDa - 0.2 µm ('metavirome'). Samples correspond, from left to right to plankton of the Etang des Vallées, France (EdV), Mediterranean water column, 20 m depth (MedS) and 40 m (MedD), and Bras del Port solar saltern ponds at 6-14-32% salt (Sal06, Sal14, Sal32). Counts are expressed in reads per kb per million mapped reads. **b**, Frequency of major cellular taxa identified in all studied metagenomes based on phylogenetically assigned USCG normalised counts.

Extended Data Fig. 5. Normalised metagenome counts of viruses and unicellular organisms across ecosystem types. a, Counts derived from the OrfM annotation strategy. **b**, Counts derived from the Prokka annotation. Viruses were counted using MCPs (sum of all) and cells for each large cellular group using the most abundant USCG for that group. MCP counts assigned to cellular contigs were considered proviruses, the rest were considered as viral particles (virions) or cell-infecting lytic forms. All the metagenomes were analysed using the same pipeline from raw reads. RPKM, reads per kilo base per million mapped metagenome reads. For each panel, a zoom on the area covered by the green rectangle is shown underneath. the Lines connecting points are shown to ease visualization.

Extended Data Fig. 6. Relative proportions of viral types and classes identified across metagenomes using major capsid proteins as markers. a, Relative abundance of different viral Baltimore classes identified by MCP counts in all analysed metagenomes. **b**, Histograms

showing the relative proportion of VP-lytic viruses that can be assigned to specific viral classes. **c**, Histograms showing the relative proportion of proviruses that can be assigned to specific viral classes.

Extended Data Fig. 7. Provirus-assigned contigs and lysogenic versus lytic viruses across metagenomes. a, Difference in the percentage of assigned genes to viruses and cells in metagenomic contigs (>5 genes) containing MCPs assigned to, respectively, proviruses and viruses across grand ecosystem types. Note that contigs containing putative provirus MCPs have much more negative values, indicating that those contigs have more cellular genes. **b,** ratios of proviruses versus lytic viruses and viral particles across more specific ecosystem types**.**

Extended Data Fig. 8. Number, diversity and evenness of viruses, proviruses and cells across ecosystems. a, Principal component analysis of different metagenomes maximizing the variance explained by viruses (VPs and infecting lytic-viruses) and proviruses (MCPs counts in RPKMs) and microbial cells (USCG counts in RPKMs). **b**, Boxplot showing the alpha diversity of viruses (VP, lytic), proviruses and cells as a function of ecosystems (Prokka annotation strategy). Diversity differences between the three types were significant ($p<0.001$ or 0.0001). VP, viral particle (virion).

Extended Data Fig. 9. Alpha diversity of cells, lytic/dispersing viruses and proviruses at the level of grand type of ecosystems. a, Box plot showing alpha diversity values and their correlation across grand ecosystem types. Values were calculated from counts derived from the OrfM annotation strategy. **b**, Similar bow plot with values inferred from the Prokka annotation strategy. **c**, Box plots showing alpha diversity values for cells, viruses (VP, lytic) and proviruses within ecosystem types (OrfM annotation). **d**, Box plots showing alpha diversity values for cells, viruses (VP, lytic) and proviruses within ecosystem types (Prokka annotation). VP, viral particle; ns, not significant.

LEGENDS FOR SUPPLEMENTARY FIGURES

Supplementary Fig. 1. Detection of (some) cellular genes/proteins using by some of the HMM models used to detect major viral capsid proteins.

Supplementary Fig. 2. Relative abundance of cells, lytic/dispersing and lysogenic viruses across the studied metagenomes. Cells were counted using the most represented USCG per metagenome. Viruses (VP, lytic) and proviruses were counted summing the different MCPs. Counts derived from two different annotation strategies based on OrfM and Prokka.

Supplementary Fig.3. Relative abundance and proportions of clusters of lysogenic (proviruses) versus lytic viruses in different sets of metagenomes. a, Relative counts. **b**, Frequency of counts. Colours indicate the assignment of clusters to either all lytic viruses, proviruses or to a mix of the two types.

Supplementary Fig. 4. Viral (all stages) and cellular diversity indices for the studied metagenomes. Data were derived from metagenomes annotated using either Prokka or OrfM.

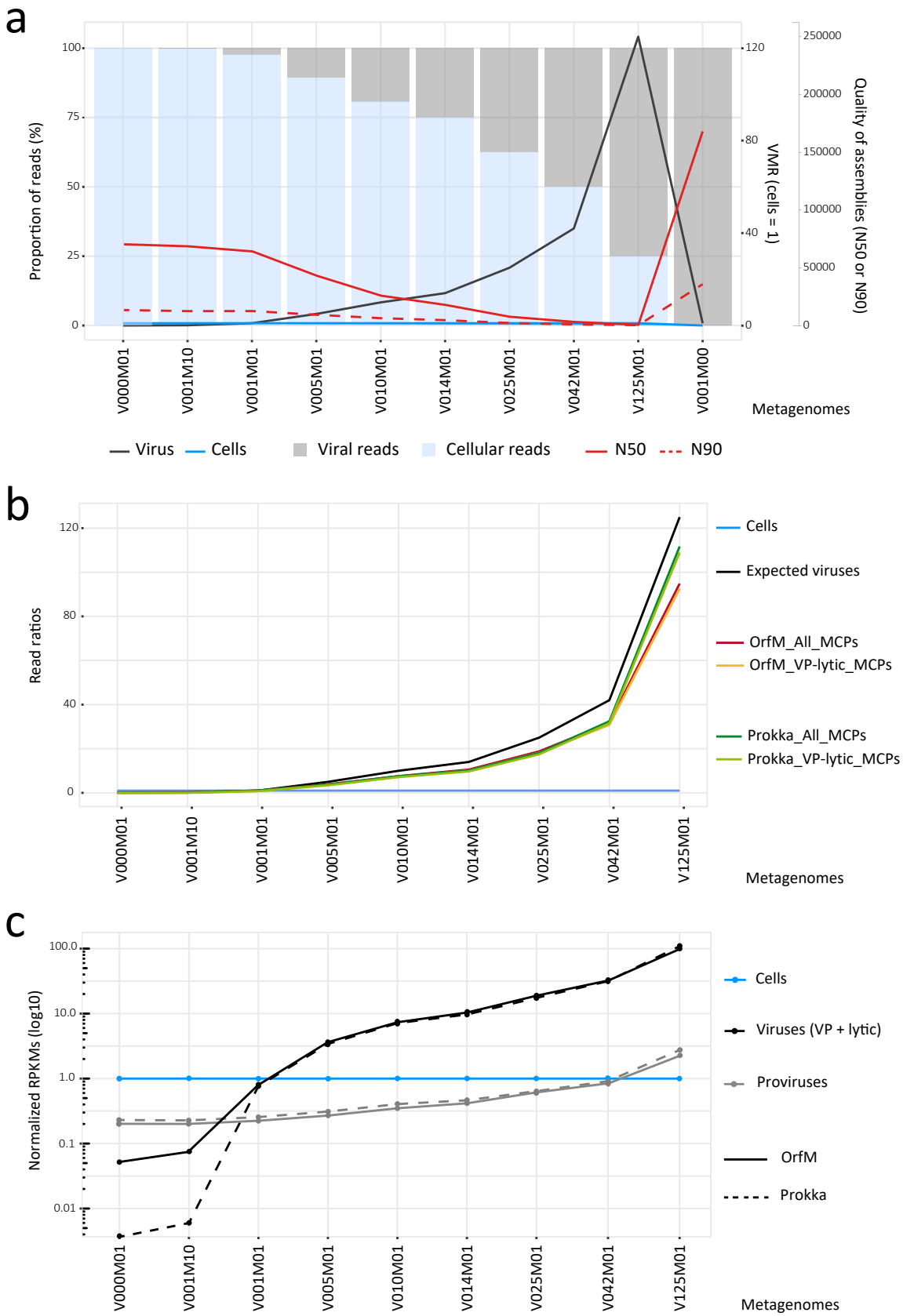


Fig.1. Estimated metagenome virus-microbe ratios (mVMR) across mock metagenomes with increasing VMR. a, Statistics of the different mock metagenomes generated with an increasing number of virus genomes (V000 to V125) as compared to the number of microbial cells (M00 – M01 – M10). **b**, Expected versus observed viral ratios versus cells (cells normalized to 1) in the different mock metagenomes as a function of the annotation method (OrfM vs. Prokka) used. Cells were counted using universal single copy genes (USCGs) and viruses using major capsid proteins (MCPs). VP-lytic_MCPs include VP and lytic viruses infecting cells; All_MCPs include also MCPs from lysogenic viruses (proviruses). **c**, Normalised counts (RPKM) of VP + lytic and proviruses per cell (normalized to 1) observed in mock metagenomes with our pipeline using OrfM and Prokka annotation strategies.

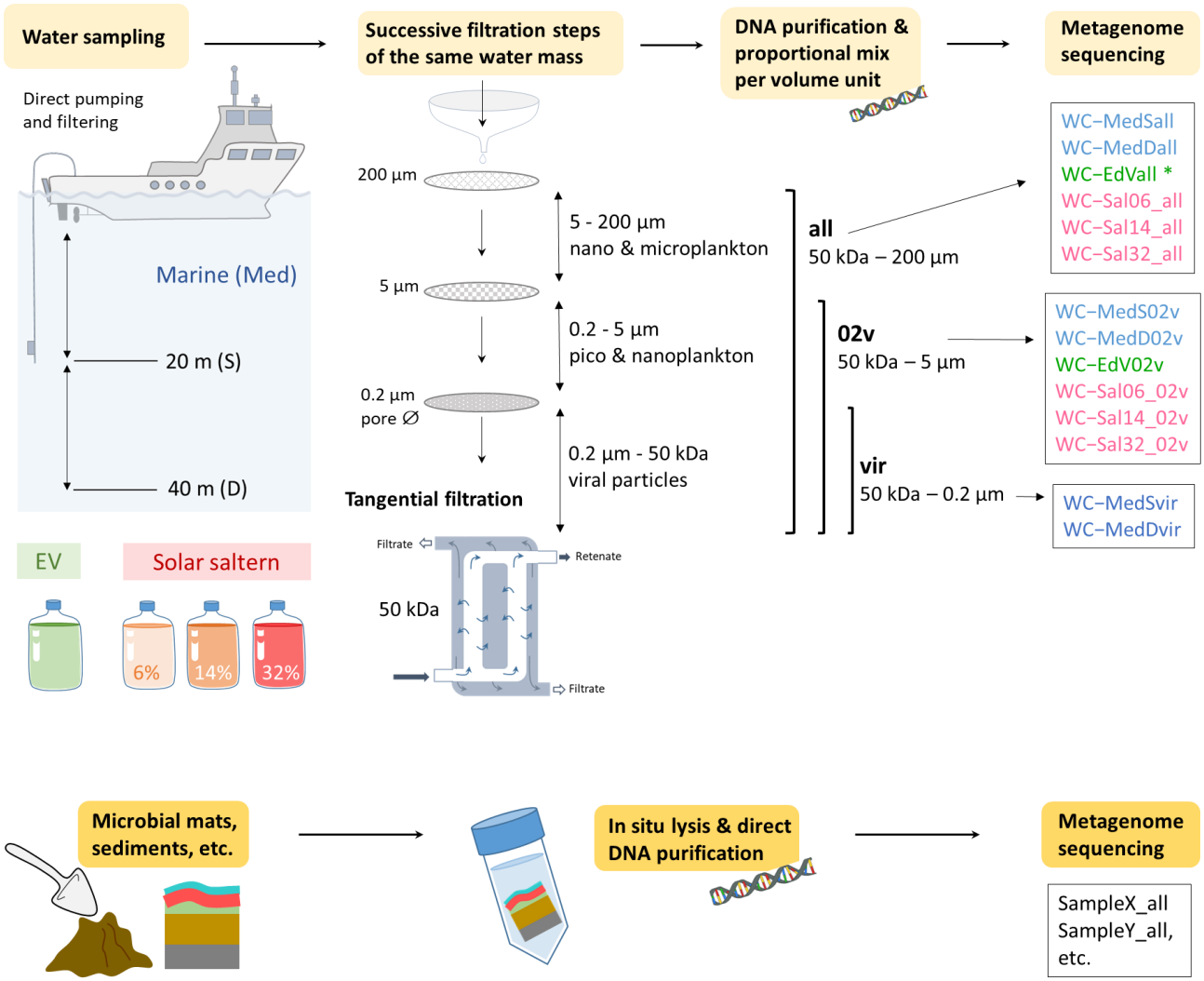


Fig.2. Strategies used to generate metagenomes in this study. Metagenomes from different planktonic fractions were obtained after sequential filtration of the same water mass, purification of DNA for each size fraction and proportional mixing per volume unit. Two metaviromes (particles <0.2 µm) were sequenced for control. Metagenomes of non-aquatic samples were generated after the in situ cell lysis and direct DNA purification prior to sequencing. * The metagenome WC-EdVall corresponds to the size fraction 50 kDa – 100 µm.

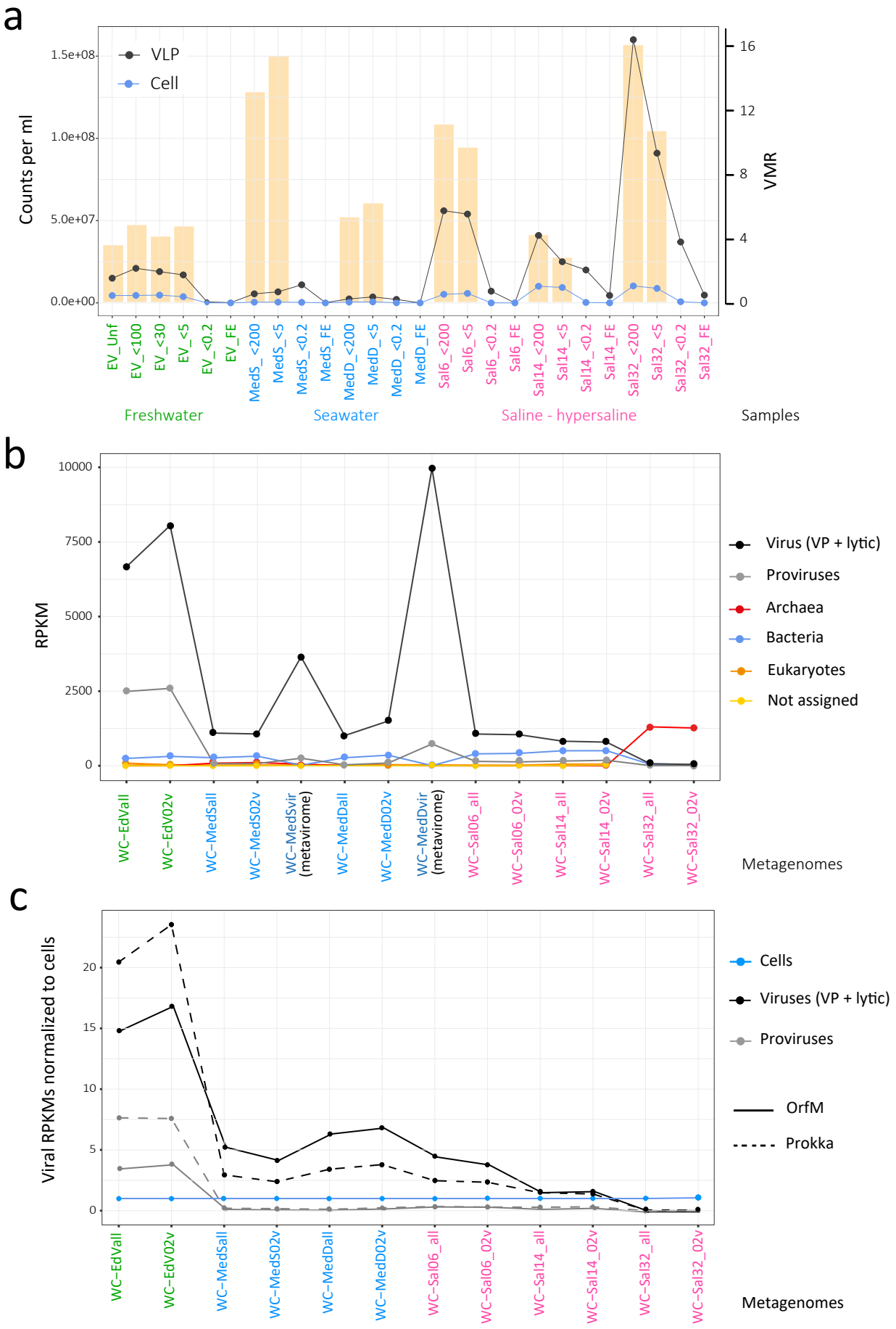


Fig.3. Epifluorescence- and normalised metagenome-based counts and ratios of cells and viruses in selected aquatic ecosystems across a salinity gradient. a, Flow-cytometry counts of cells and viral-like particles (VLPs) and estimated virus-to-microbe ratios (histogram bars). Sample names indicate the plankton size fraction considered (Unf, all size fractions; <200, <100, <30 and <5, fractions below those numbers in µm; <0.2, fraction between 50 kDa and 0.2 µm; FE, final eluate <50 kDa; see Supplementary Table 1). **b**, Normalised counts (RPKM) of viruses (MCPs corresponding to VP + lytic and lysogenic –proviruses– forms) and cells from the three cellular domains (USCGs) in global (cells + viruses) metagenomes from the same original aquatic samples as in (a). Metagenomes correspond to size plankton fractions indicated by their suffixes: -all, fraction between 50 kDa and 200 µm (100 µm for EV); 02v, fraction between 50 kDa and 5 µm (see Extended Data Fig.2 and Supplementary Tables 4-5). Two metaviromes (fraction 50 kDa-0.2 µm) were sequenced for Mediterranean (MedS, MedD) plankton samples as internal control. **c**, Metagenome-based normalised ratios of lytic (VP, actively infecting) and lysogenic (essentially proviruses) viruses per cell (bacteria + archaea + eukaryotes). RPKM, reads per kilo base per million mapped metagenome reads. Lytic and lysogenic mVMRs were estimated after OrfM and Prokka annotation strategies.

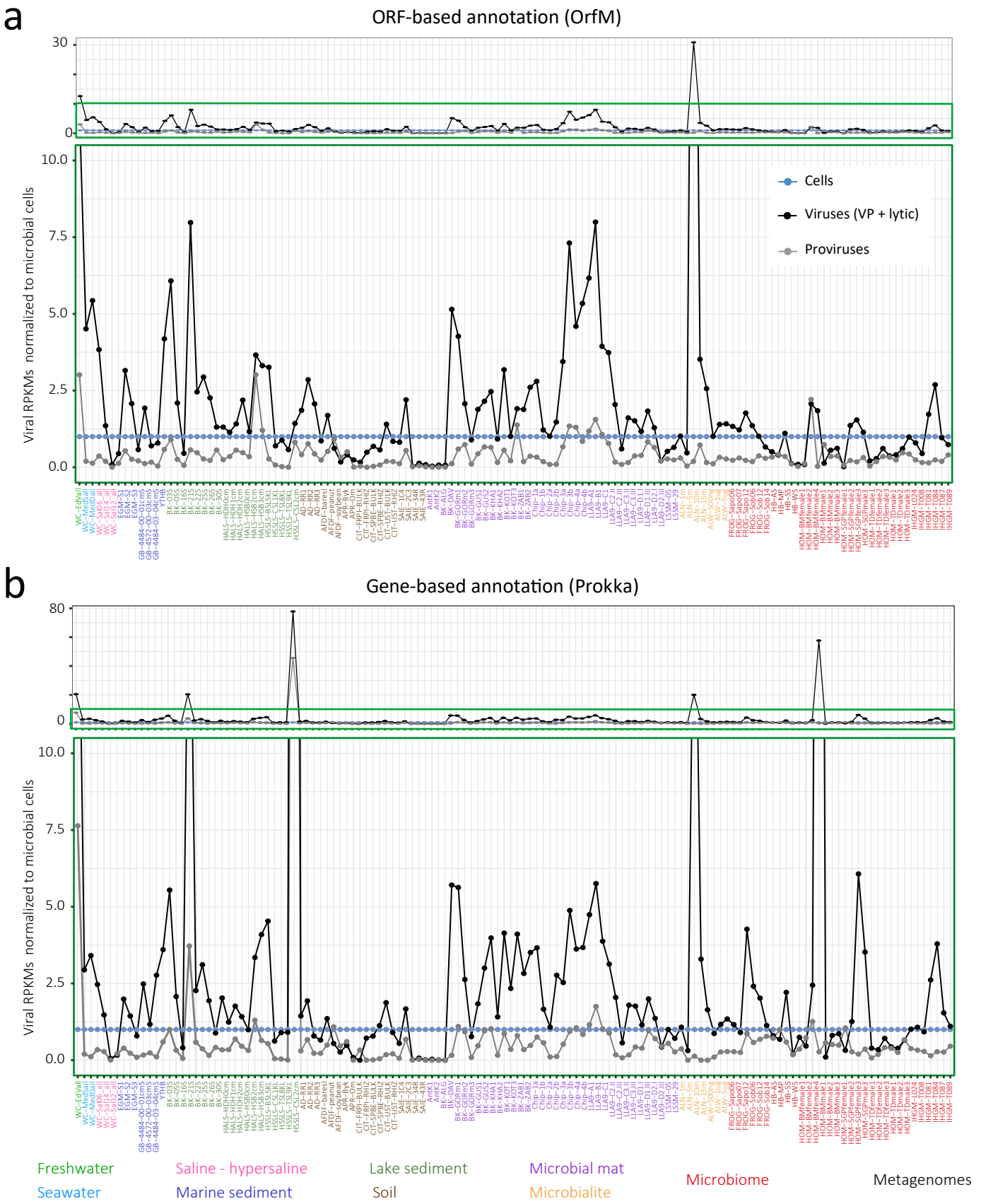


Fig.4. Metagenome-based normalised ratios of viral particles/lytic and lysogenic viruses per cell across ecosystem types. Viruses and cells were counted using, respectively, MCPs and USCGs as proxies and cell counts were normalised to 1. All the metagenomes were analysed using the same pipeline from raw reads. **a**, Normalised estimates of virus-cell ratios using the OrfM annotation strategy. A zoom of the zone marked by the green rectangle is shown below. **b**, Similar estimates based on Prokka annotation. RPKM, reads per kilo base per million mapped metagenome reads.

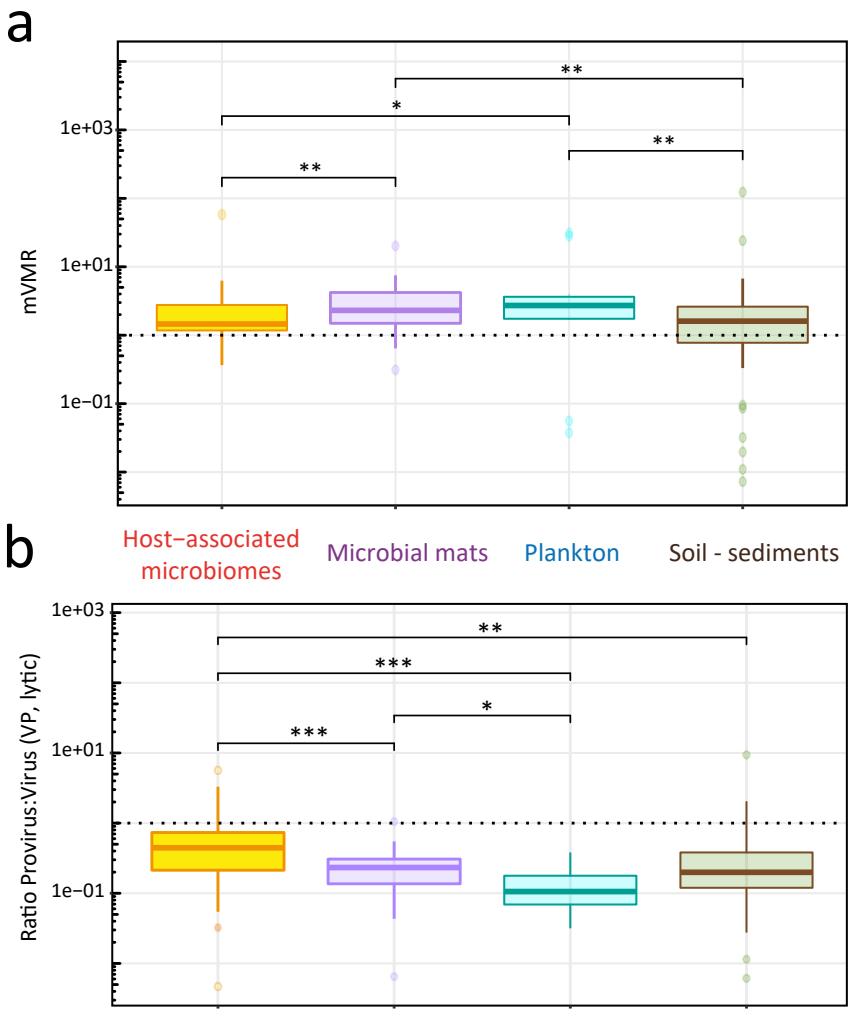


Fig. 5. mVMR and ratios of proviruses versus lytic viruses and viral particles per grand ecosystem type. a, Boxplot indicating mVMR. **b**, Boxplot showing the provirus:virus (VP-lytic) ratios. Values are based on Prokka annotation. Significant differences between ratios are indicated.

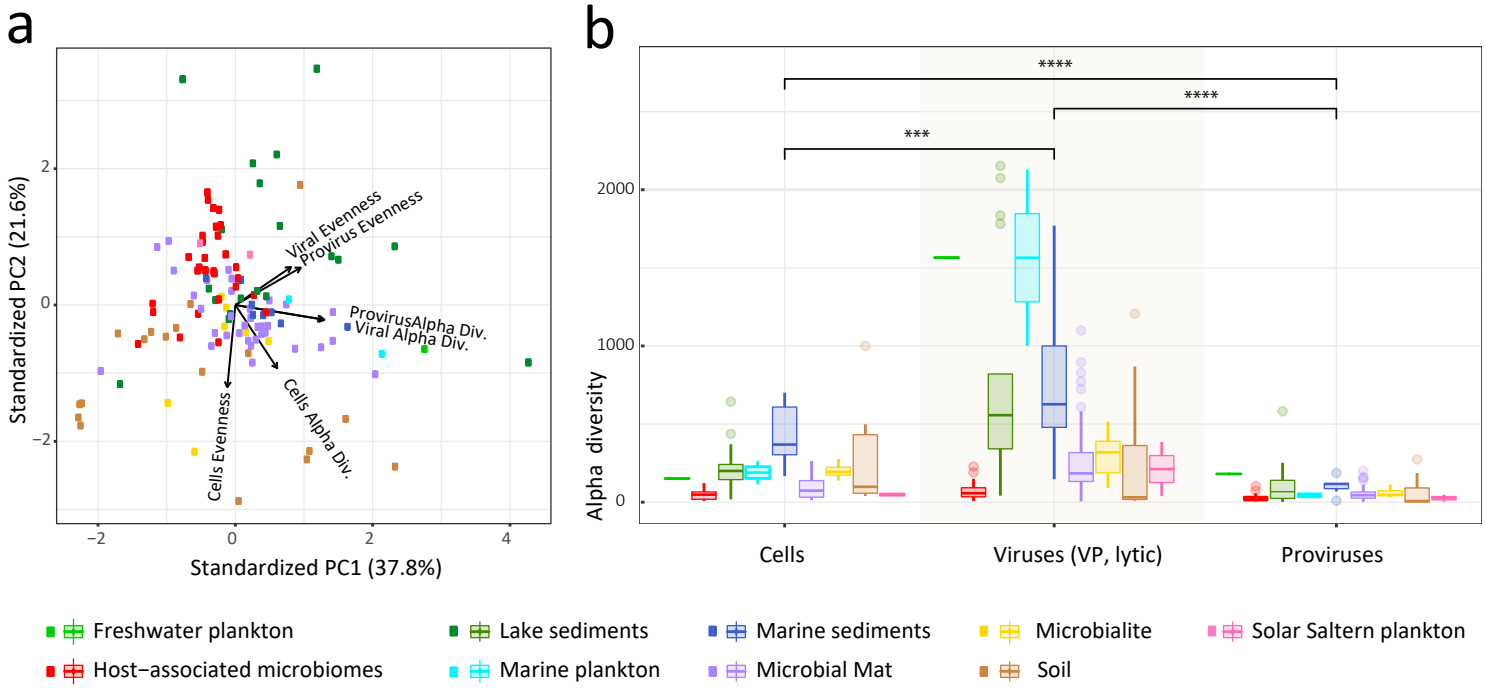
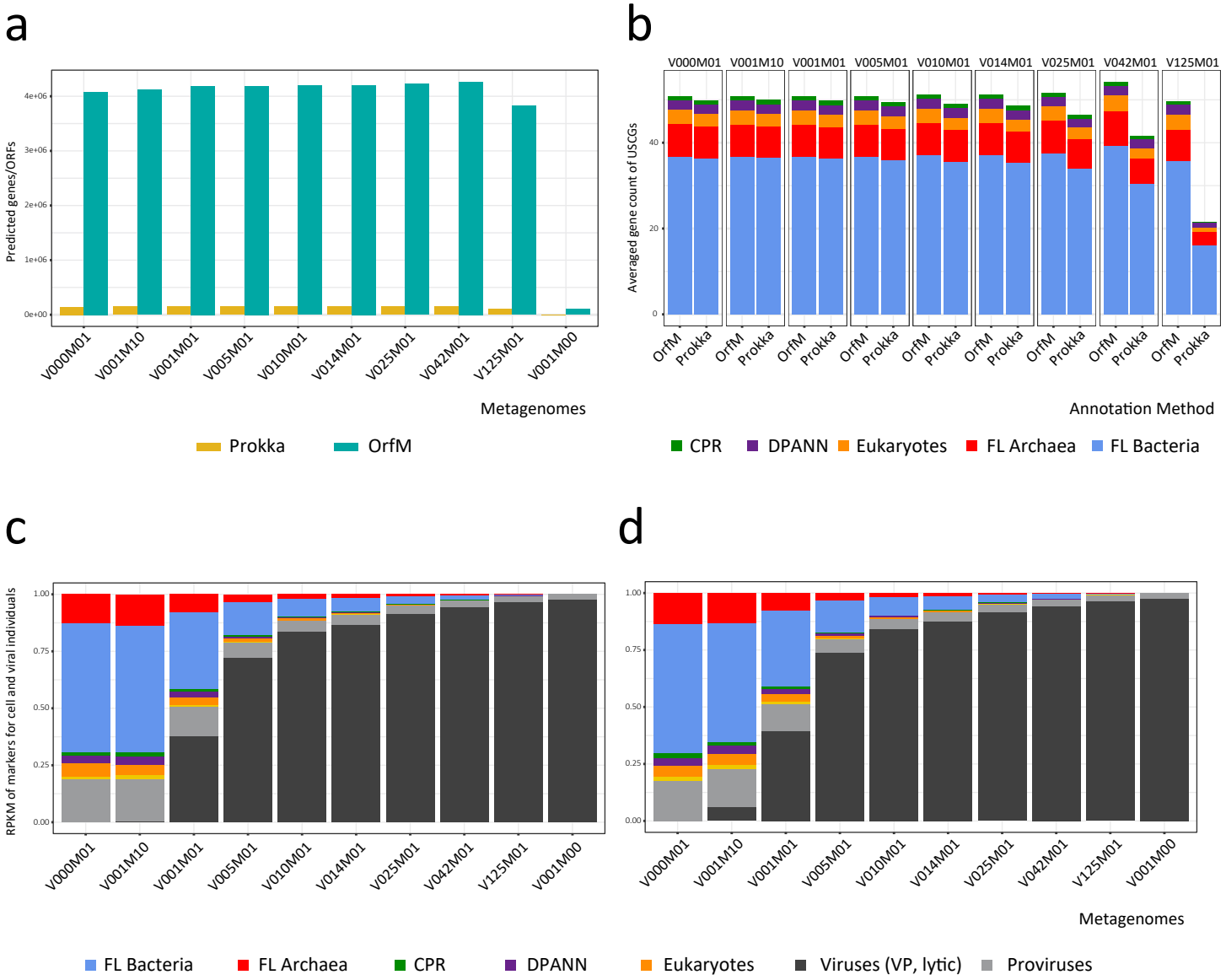


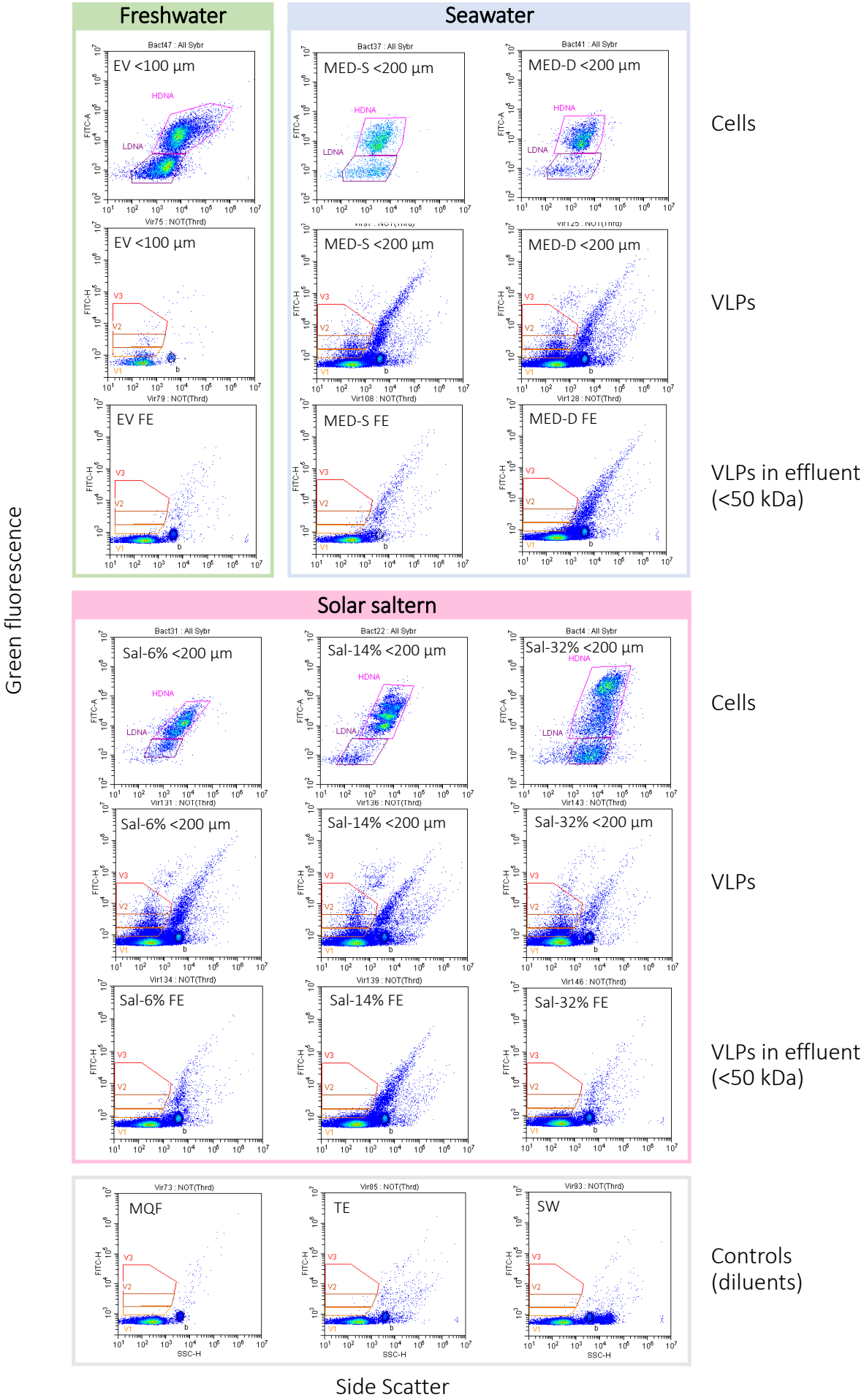
Fig.6. Diversity and evenness of viruses, proviruses and cells across ecosystems. a, Principal component analysis of different metagenomes maximizing the variance explained by viruses (VPs and infecting lytic-viruses) and proviruses (MCPs counts in RPKMs) and microbial cells (USCG counts in RPKMs). **b**, Boxplot showing the alpha diversity of viruses (VP, lytic), proviruses and cells as a function of ecosystems (OrfM annotation strategy; a boxplot from data derived from Prokka annotation can be seen in Extended Data Fig. 7). Diversity differences between the three types were significant ($p<0.001$ or 0.0001). These differences were also seen across and within large ecosystem categories (Extended Data Fig. 8). VP, viral particle (virion).



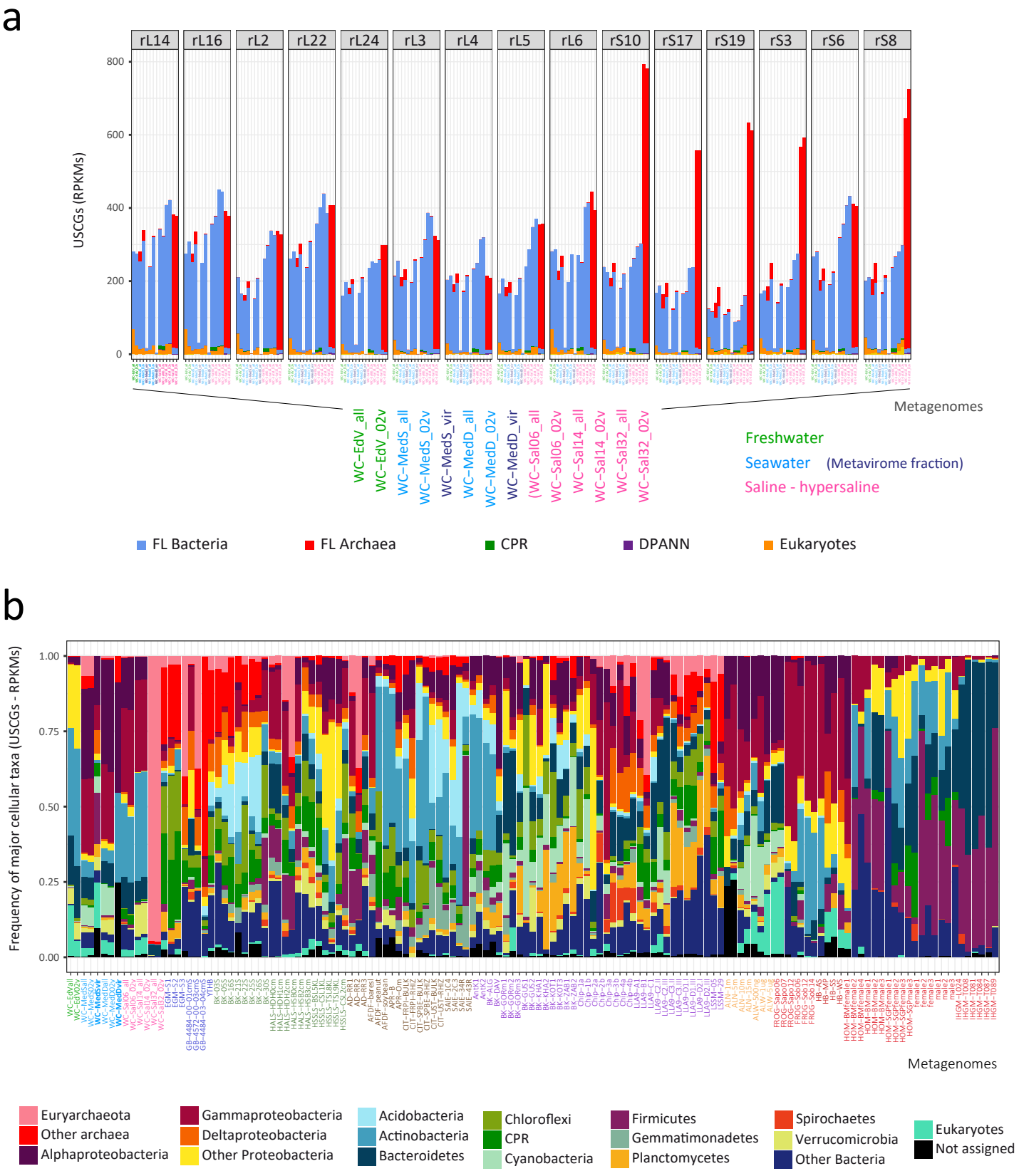
Extended Data Fig.1. Predicted gene content of the generated mock metagenomes. a, prediction of genes and open reading frames (ORFs) based on, respectively Prokka and OrfM annotation of assembled contigs. **b**, Prediction of the relative proportion of different cellular taxa in the mock metagenomes based on the averaged number of selected universal single copy genes (USCGs). CPR, Candidate Phyla Radiation, host-dependent epibiotic bacteria; DPANN, clade of host-dependent epibiotic archaea. **c**, Relative proportion of cellular and viral entities estimated from Prokka-annotated mock metagenomes with USCGs and major capsid proteins (MCPs) as markers for individuals. **d**, the same estimation for OrfM annotation.

Colour code	Ecosystem type	Sample name	Short description / Location
	Freshwater plankton*	WC-EdVall	Etang des Vallées, freshwater pond, 50 kDa - 100 µm cell size fraction (viral particles and cells <100 µm)
	Freshwater plankton*	WC-EdV02v	Etang des Vallées, freshwater pond, 50 kDa - 5 µm cell size fraction (viral particles and cells <5µm)
	Marine plankton*	WC-MedSall	Mediterranean water column 20 m depth, 50 kDa - 200 µm cell size fraction (viral particles and cells <200 µm
	Marine plankton*	WC-MedS02v	Mediterranean water column 20 m depth, 50 kDa - 5 µm cell size fraction (viral particles and cells <5 µm)
	Marine plankton*	WC-MedSvir	Mediterranean water column 20 m depth, 50 kDa - 0.2 µm **(metavirome)**
	Marine plankton*	WC-MedDall	Mediterranean water column 40 m depth, 50 kDa - 200 µm cell size fraction (viral particles and cells <200 µm
	Marine plankton*	WC-MedD02v	Mediterranean water column 40 m depth, 50 kDa - 5 µm cell size fraction (viral particles and cells <5 µm)
	Marine plankton*	WC-MedDvir	Mediterranean water column 40 m depth, 50 kDa - 0.2 µm **(metavirome)**
	Plankton solar saltern*	WC-Sal6_all	Solar saltern Bras del Port, 6% salt, 50 kDa - 200 µm cell size fraction (viral particles and cells <200 µm)
	Plankton solar saltern*	WC-Sal6_02v	Solar saltern Bras del Port, 6% salt, 50 kDa - 5 µm cell size fraction (viral particles and cells <5 µm)
	Plankton solar saltern*	WC-Sal14_all	Solar saltern Bras del Port, 14% salt, 50 kDa - 200 µm cell size fraction (viral particles and cells <200 µm)
	Plankton solar saltern*	WC-Sal14_02v	Solar saltern Bras del Port, 14% salt, 50 kDa - 5 µm cell size fraction (viral particles and cells <5 µm)
	Plankton solar saltern*	WC-Sal32_all	Solar saltern Bras del Port, 32% salt, 50 kDa - 200 µm cell size fraction (viral particles and cells <200 µm)
	Plankton solar saltern*	WC-Sal32_02v	Solar saltern Bras del Port, 32% salt, 50 kDa - 5 µm cell size fraction (viral particles and cells <5 µm)
	Marine Sediments	EGM-S1	Eastern Gulf of Mexico
	Marine Sediments	EGM-S2	Eastern Gulf of Mexico
	Marine Sediments	EGM-S3	Eastern Gulf of Mexico
	Marine Sediments	GB-4484-00-01cmS	Guaymas sediment
	Marine Sediments	GB-4572-00-03cmS	Guaymas sediment
	Marine Sediments	GB-4484-03-04cmS	Guaymas sediment
	Marine Sediments	YTHB	Yap Trench Hadal Biosphere
	Lake Sediments*	BK-03S	Lake Baikal sediment, Central basin, 1081 m depth
	Lake Sediments*	BK-05S	Lake Baikal sediment, Central basin, 1450 m depth
	Lake Sediments*	BK-16S	Lake Baikal sediment, North basin, 846 m depth
	Lake Sediments*	BK-21S	Lake Baikal sediment, North basin, 373 m depth
	Lake Sediments*	BK-22S	Lake Baikal sediment, Central basin, 592 m depth
	Lake Sediments*	BK-25S	Lake Baikal sediment, Central basin, 323 m depth
	Lake Sediments*	BK-26S	Lake Baikal sediment, South basin, 1412 m depth
	Lake Sediments*	BK-30S	Lake Baikal sediment, South basin, 1381 m depth
	Lake Sediments	HALS-HDH0cm	Lake Hazen deep hole sediment, High Arctic
	Lake Sediments	HALS-HDH1cm	Lake Hazen deep hole sediment, High Arctic
	Lake Sediments	HALS-HDH2cm	Lake Hazen deep hole sediment, High Arctic
	Lake Sediments	HALS-HSB0cm	Lake Hazen Snowgoose Bay sediment, High Arctic
	Lake Sediments	HALS-HSB2cm	Lake Hazen Snowgoose Bay sediment, High Arctic
	Lake Sediments	HALS-HSB3cm	Lake Hazen Snowgoose Bay sediment, High Arctic
	Lake Sediments	HSSLS-CSL1KL	Hypersaline Soda Lake Sediments
	Lake Sediments	HSSLS-BSL5KL	Hypersaline Soda Lake Sediments
	Lake Sediments	HSSLS-TSL8KL	Hypersaline Soda Lake Sediments
	Lake Sediments	HSSLS-TSL9KL	Hypersaline Soda Lake Sediments
	Lake Sediments	HSSLS-CSL2cm	Hypersaline Soda Lake Sediments
	Soil	AD-RR1	Antarctic desert soil
	Soil	AD-RR2	Antarctic desert soil
	Soil	AD-RR3	Antarctic desert soil
	Soil	AFDF-baresl	Agricultural soil during fallow
	Soil	AFDF-peanut	Agricultural soil during fallow
	Soil	AFDF-soybean	Agricultural soil during fallow
	Soil	APR-Byk	Arctic permafrost, Russia
	Soil	APR-Om	Arctic permafrost, Russia
	Soil	CIT-FRPI-BULK	Citrus-associated soil
	Soil	CIT-FRPI-RHIZ	Citrus-associated soil, rhizosphere
	Soil	CIT-SPBE-BULK	Citrus-associated soil
	Soil	CIT-SPBE-RHIZ	Citrus-associated soil, rhizosphere
	Soil	CIT-UST-BULK	Citrus-associated soil
	Soil	CIT-UST-RHIZ	Citrus-associated soil, rhizosphere
	Soil	SAIE-1C4	South Africa soil
	Soil	SAIE-2C3	South Africa soil
	Soil	SAIE-34R	South Africa soil
	Soil	SAIE-43R	South Africa soil
	Microbial Mat*	AntK1	Antarctic microbial mat
	Microbial Mat*	AntK2	Antarctic microbial mat
	Microbial Mat*	BK-ALG	Thin microbial mat, brackish lake, Southern Siberia
	Microbial Mat*	BK-DAV	Thermophilic mat, Southern Siberia
	Microbial Mat*	BK-GORm1	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	BK-GORm2	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	BK-GORm3	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	BK-GUS1	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	BK-GUS2	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	BK-KHA1	Thermophilic mat, Southern Siberia
	Microbial Mat*	BK-KHA2	Thermophilic mat, Southern Siberia
	Microbial Mat*	BK-KOT1	Microbial mat, Southern Siberia
	Microbial Mat*	BK-KOT3	Microbial mat, Southern Siberia
	Microbial Mat*	BK-ZAB1	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	BK-ZAB2	Thermophilic microbial mat, Southern Siberia
	Microbial Mat*	Chip-1a	Halophilic mat, Spain
	Microbial Mat*	Chip-1b	Halophilic mat, Spain
	Microbial Mat*	Chip-2a	Halophilic mat, Spain
	Microbial Mat*	Chip-2b	Halophilic mat, Spain
	Microbial Mat*	Chip-3a	Halophilic mat, Spain
	Microbial Mat*	Chip-3b	Halophilic mat, Spain
	Microbial Mat*	Chip-4a	Halophilic mat, Spain
	Microbial Mat*	Chip-4b	Halophilic mat, Spain
	Microbial Mat*	LLA9-A1	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-B1	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-C1	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-C2.II	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-C2.III	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-C3.II	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-C3.III	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-D1.I	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-D1.III	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-D2.I	Salar de Llamara, microbial mat, Chile
	Microbial Mat*	LLA9-D2.III	Salar de Llamara, microbial mat, Chile
	Microbial Mat	LSSM-05	Little Sippewissett salt marsh Microbial Mats
	Microbial Mat	LSSM-29	Little Sippewissett salt marsh Microbial Mats
	Microbialite*	ALN-1m	Lake Alchichica microbialite fragment, Mexico
	Microbialite*	ALN-5m	Lake Alchichica microbialite fragment, Mexico
	Microbialite*	ALN-10m	Lake Alchichica microbialite fragment, Mexico
	Microbialite*	ALN-15m	Lake Alchichica microbialite fragment, Mexico
	Microbialite*	ALW-500ng	Lake Alchichica microbialite fragment, Mexico
	Microbialite*	ALW-1ug	Lake Alchichica microbialite fragment, Mexico
	Microbialite*	ALW-5ug	Lake Alchichica microbialite fragment, Mexico
	Host-associated microbiomes	FROG-Sapo06	Frog skin microbiome
	Host-associated microbiomes	FROG-Sapo07	Frog skin microbiome
	Host-associated microbiomes	FROG-Sapo12	Frog skin microbiome
	Host-associated microbiomes	FROG-Sob06	Frog skin microbiome
	Host-associated microbiomes	FROG-Sob12	Frog skin microbiome
	Host-associated microbiomes	FROG-Sob14	Frog skin microbiome
	Host-associated microbiomes	HB-AS	Ayrshire honeybee microbiome
	Host-associated microbiomes	HB-MP	Mid Perthshire honeybee microbiome
	Host-associated microbiomes	HB-SS	Shropshire honeybee microbiome
	Host-associated microbiomes	HB-WS	Wigtownshire honeybee microbiome
	Host-associated microbiomes	HOM-BMfemale1	Human buccal mucosa microbiome
	Host-associated microbiomes	HOM-BMfemale2	Human buccal mucosa microbiome
	Host-associated microbiomes	HOM-BMfemale4	Human buccal mucosa microbiome
	Host-associated microbiomes	HOM-BMmale1	Human buccal mucosa microbiome
	Host-associated microbiomes	HOM-BMmale2	Human buccal mucosa microbiome
	Host-associated microbiomes	HOM-BMmale3	Human buccal mucosa microbiome
	Host-associated microbiomes	HOM-SGPfemale1	Human subgingival microbiome
	Host-associated microbiomes	HOM-SGPfemale2	Human subgingival microbiome
	Host-associated microbiomes	HOM-SGPfemale3	Human subgingival microbiome
	Host-associated microbiomes	HOM-SGPmale3	Human subgingival microbiome
	Host-associated microbiomes	HOM-TDfemale1	Tongue dorsum microbiome
	Host-associated microbiomes	HOM-TDfemale2	Tongue dorsum microbiome
	Host-associated microbiomes	HOM-TDfemale3	Tongue dorsum microbiome
	Host-associated microbiomes	HOM-TDmale1	Tongue dorsum microbiome
	Host-associated microbiomes	HOM-TDmale2	Tongue dorsum microbiome
	Host-associated microbiomes	HOM-TDmale3	Tongue dorsum microbiome
	Host-associated microbiomes	IHGM-L024	Inmmigrant Human Gut Microbiome
	Host-associated microbiomes	IHGM-T008	Inmmigrant Human Gut Microbiome
	Host-associated microbiomes	IHGM-T081	Inmmigrant Human Gut Microbiome
	Host-associated microbiomes	IHGM-T084	Inmmigrant Human Gut Microbiome
	Host-associated microbiomes	IHGM-T087	Inmmigrant Human Gut Microbiome
	Host-associated microbiomes	IHGM-T089	Inmmigrant Human Gut Microbiome

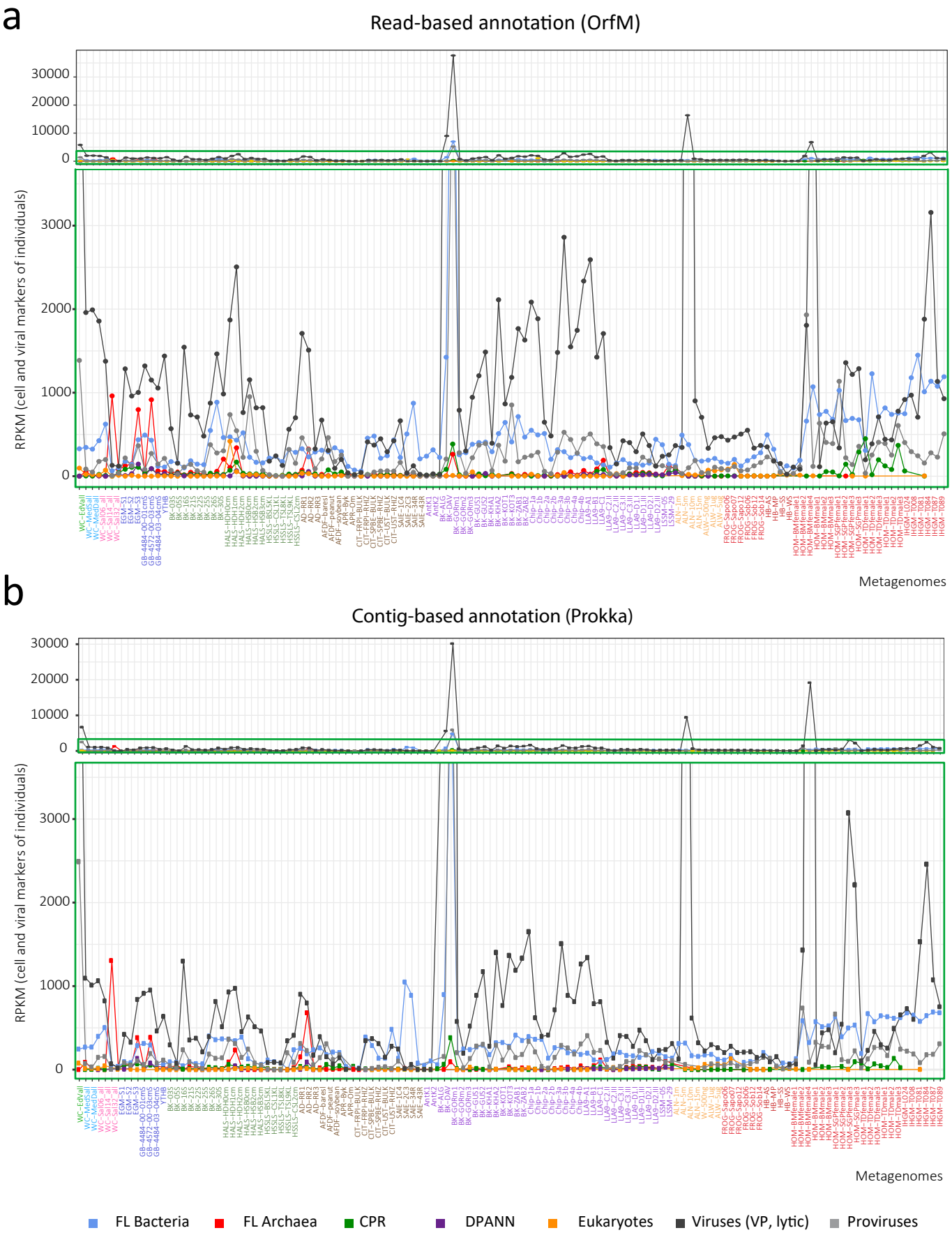
Extended Data Fig.2. Origin and main characteristics of analysed metagenomes. Extended information about metagenome origin and statistics are given in Supplementary Table 6. *metagenomes generated in our laboratory (this and previous or ongoing studies).



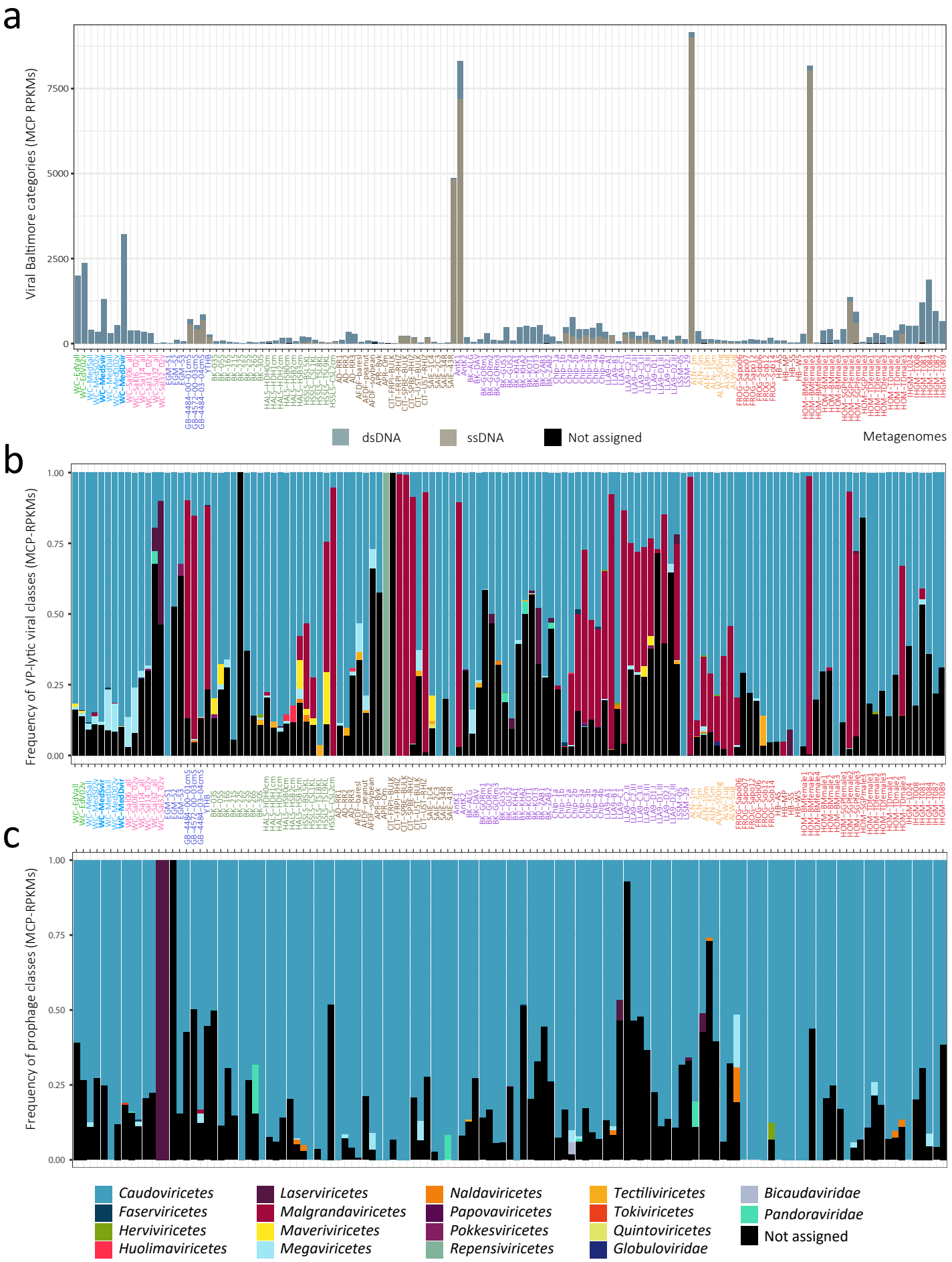
Extended Data Fig.3. Cytograms of SYBR Green I-stained cells and virus-like particles (VLPs) in several aquatic environments spanning different salinities. Data are shown for plankton below 200 µm cell size (100 µm for EV). Three VLP subpopulations (V1-V2-V3) and two major cellular subpopulations containing high DNA (HDNA) and low DNA (LDNA) levels are defined. Blanks are also shown: MQF, Milli-Q water, TE, Tris EDTA buffer, SW, seawater. FE, final eluate (<50 kDa); b, fluorescent 0.2µm beads. More detailed sample descriptions are given in Extended Data Fig.2 and Supplementary Tables 4 and 6.



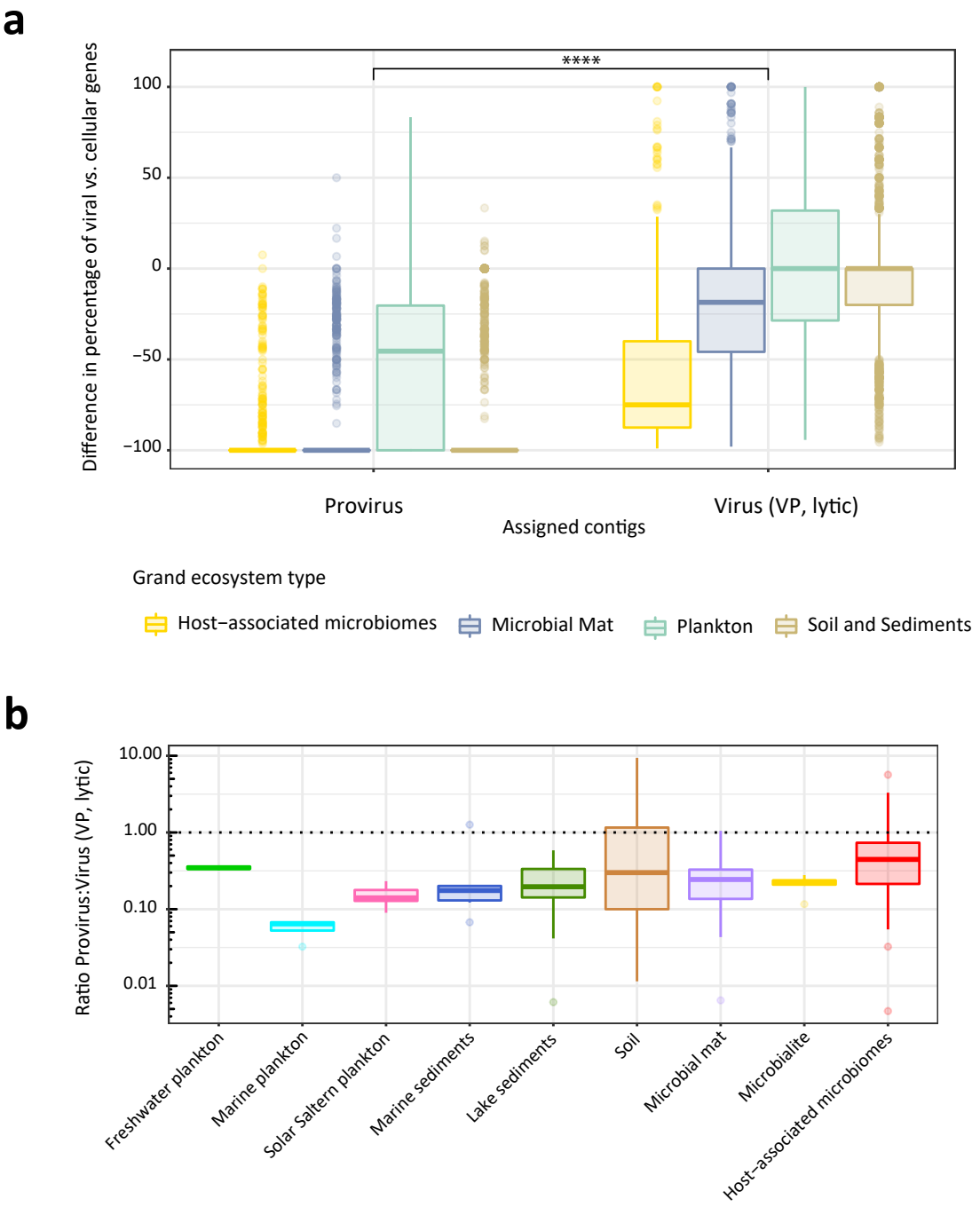
Extended Data Fig. 4. Normalised counts of selected universal single copy genes (USCGs) and relative abundance of major cellular taxa across metagenomes. a, Normalised counts of selected USCGs in aquatic metagenomes of coexisting viral and cell fractions. Suffixes in sample names indicate the plankton size fraction: _all, >50 kDa - 200 µm (100 µm in EdV); _02V, >50 kDa - 5 µm; _vir, >50 kDa - 0.2 µm ('metavirome'). Samples correspond, from left to right to plankton of the Etang des Vallées, France (EdV), Mediterranean water column, 20 m depth (MedS) and 40 m (MedD), and Bras del Port solar saltern ponds at 6-14-32% salt (Sal06, Sal14, Sal32). Counts are expressed in reads per kb per million mapped reads. **b**, Frequency of major cellular taxa identified in all studied metagenomes based on phylogenetically assigned USCG normalized counts.



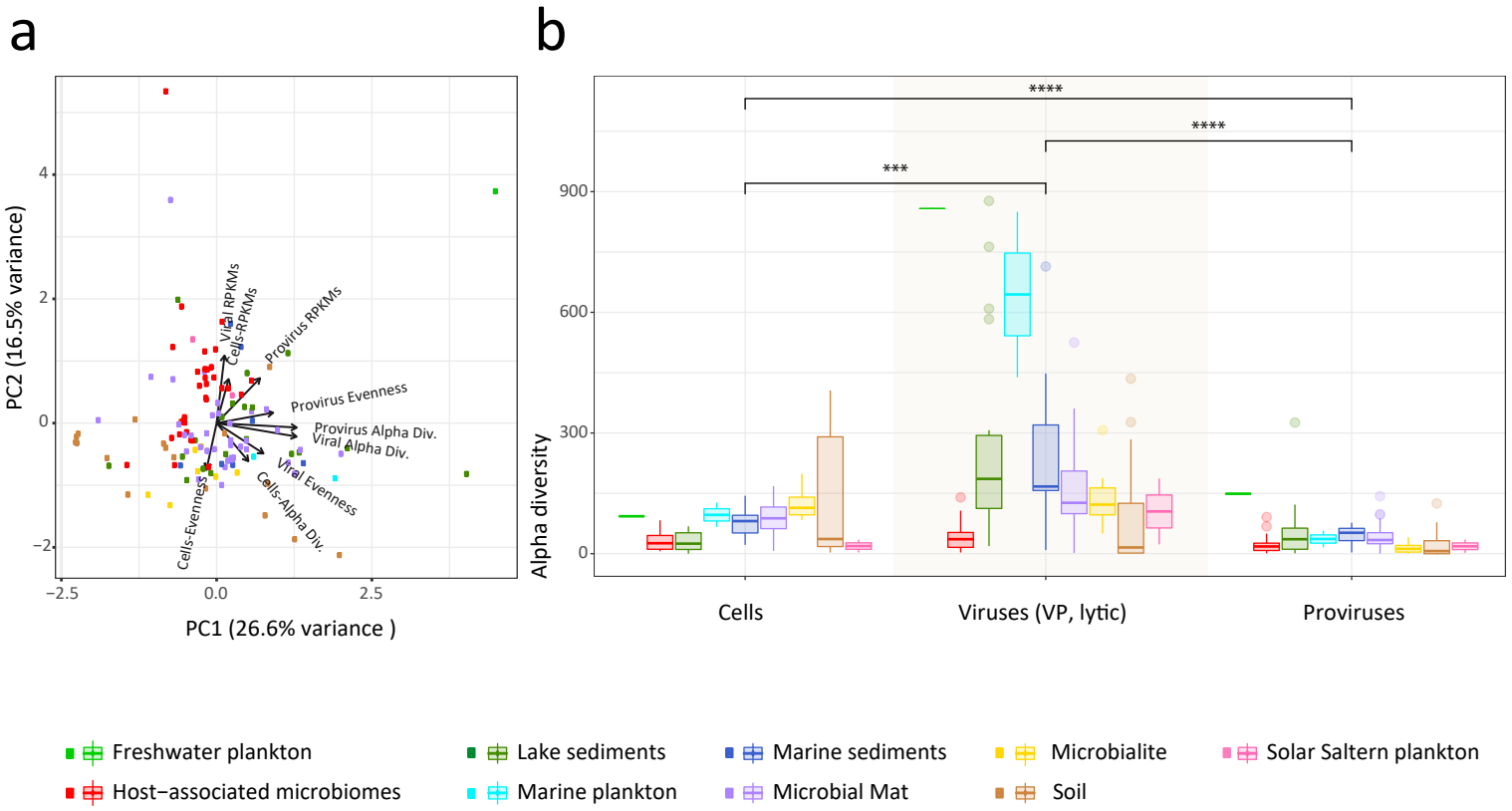
Extended Data Fig. 5. Normalised metagenome counts of viruses and unicellular organisms across ecosystem types. a, Counts derived from the OrfM annotation strategy. **b**, Counts derived from the Prokka annotation. Viruses were counted using MCPs (sum of all) and cells for each large cellular group using the most abundant USCG for that group. MCP counts assigned to cellular contigs were considered proviruses, the rest were considered as viral particles (virions) or cell-infecting lytic forms. All the metagenomes were analysed using the same pipeline from raw reads. RPKM, reads per kilo base per million mapped metagenome reads. For each panel, a zoom on the area covered by the green rectangle is shown underneath. the Lines connecting points are shown to ease visualization.



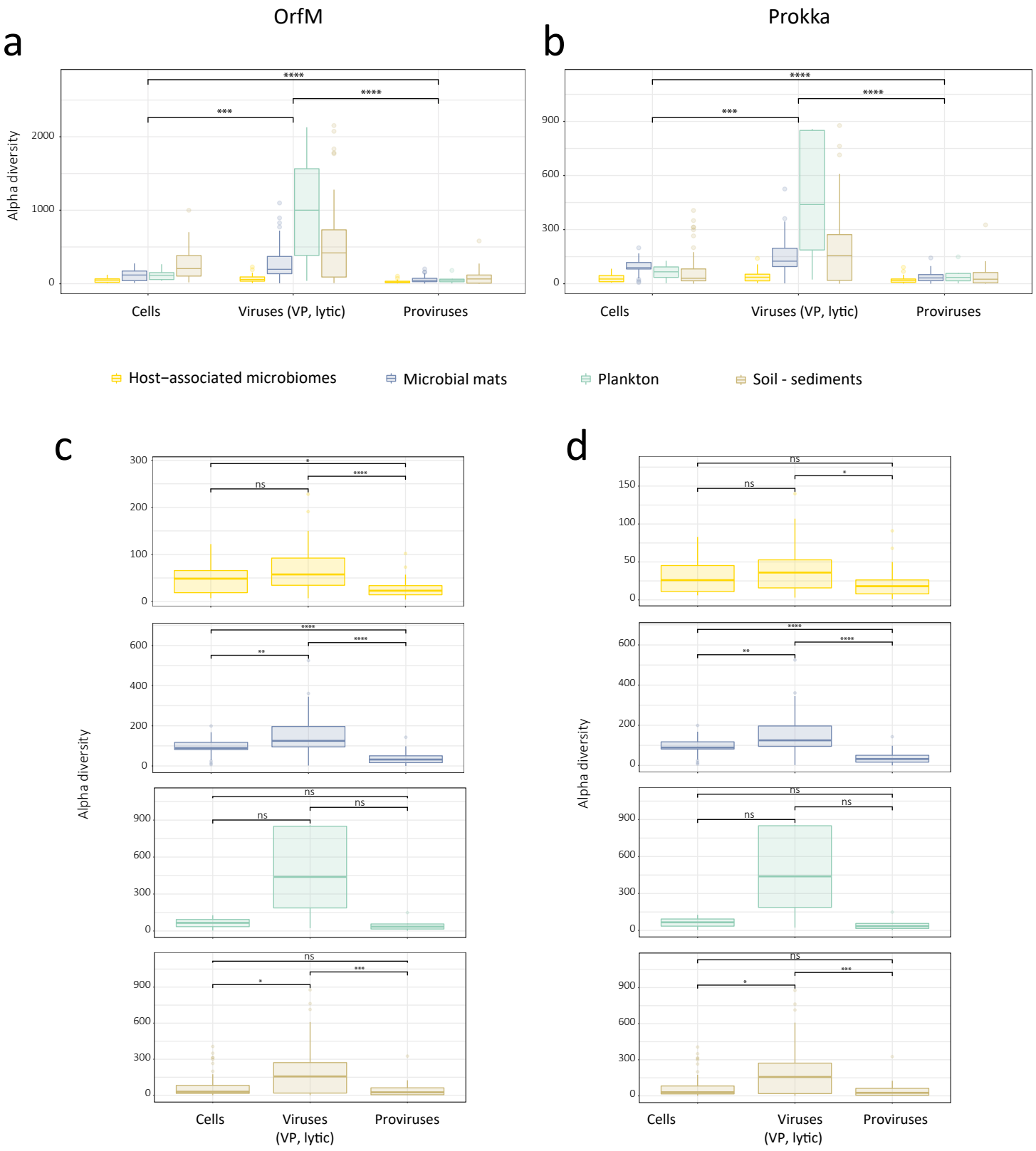
Extended Data Fig. 6. Relative proportions of viral types and classes identified across metagenomes using major capsid proteins as markers. a, Relative abundance of different viral Baltimore classes identified by MCP counts in all analysed metagenomes. **b**, Histograms showing the relative proportion of VP-lytic viruses that can be assigned to specific viral classes. **c**, Histograms showing the relative proportion of proviruses that can be assigned to specific viral classes.



Extended Data Fig. 7. Provirus-assigned contigs and lysogenic versus lytic viruses across metagenomes. a, Difference in the percentage of assigned genes to viruses and cells in metagenomic contigs (>5 genes) containing MCPs assigned to, respectively, proviruses and viruses across grand ecosystem types. Note that contigs containing putative proviral MCPs have much more negative values, indicating that those contigs have more cellular genes. **b**, ratios of proviruses versus lytic viruses and viral particles across more specific ecosystem types.



Extended Data Fig. 8. Number, diversity and evenness of viruses, proviruses and cells across ecosystems. a, Principal component analysis of different metagenomes maximizing the variance explained by viruses (VPs and infecting lytic-viruses) and proviruses (MCPs counts in RPKMs) and microbial cells (USCG counts in RPKMs). **b**, Boxplot showing the alpha diversity of viruses (VP, lytic), proviruses and cells as a function of ecosystems (Prokka annotation strategy). Diversity differences between the three types were significant ($p<0.001$ or 0.0001). VP, viral particle (virion).



Extended Data Fig. 9. Alpha diversity of cells, lytic/dispersing viruses and proviruses at the level of grand type of ecosystems. **a**, Box plot showing alpha diversity values and their correlation across grand ecosystem types. Values were calculated from counts derived from the OrfM annotation strategy. **b**, Similar bow plot with values inferred from the Prokka annotation strategy. **c**, Box plots showing alpha diversity values for cells, viruses (VP, lytic) and proviruses within ecosystem types (OrfM annotation). **d**, Box plots showing alpha diversity values for cells, viruses (VP, lytic) and proviruses within ecosystem types (Prokka annotation). VP, viral particle; ns, not significant.