Viral and parasitic cell ratios per free-living microbe across ecosystems

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

[1] Ecologie Systématique Evolution, Centre National de la Recherche Scientifique - CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France
[2] 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 fluorescently-labelled viral-like particles (VLPs) counts. However, these exclude cell-infecting lytic and lysogenic viruses and potentially include many false positives, e.g. DNA-containing membrane vesicles, gene-transfer agents or inert particles. Here, we develop a metagenome-based VMR estimate (mVRM) accounting for all viral stages (virion-lytic-lysogenic) using normalised counts of viral DNA-polymerases and cellular universal single-copy genes as proxies for individuals. 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 6-14-32% salt solar-saltern ponds. Viruses outnumbered cells in non-blooming freshwater and marine plankton by around twofold. However, mVMR in 133 uniformly-analysed metagenomes from diverse ecosystems showed that free-living cells largely exceeded viruses and epibiont DPANN archaea and/or CPR bacteria in compact environments (sediments, soils, microbial mats, host-associated microbiomes). Along the Piggyback-the-winner model[5] lines, lysogenic genes significantly correlated with cell, but not necessarily biotope, density. While viruses likely are the most diverse biological entities on Earth[6,7], our results suggest that cells are the most abundant and that cellular parasites may exert population-size control in some high-density ecosystems.

Viruses are strict molecular parasites that play crucial roles in ecosystems[2,8]. Hugely diverse[6,7], they are thought to be at least one order of magnitude more abundant than microbial cells on Earth[1-4]. Abundance estimates are largely based on epifluorescence microscopy or flow cytometry (FCM) counts of viral-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 10:1[3,8,9]. However, epifluorescence-based VLP counts exclude lytic and lysogenic viruses infecting cells and, conversely, may include diverse false positives. These comprise cell-DNA-containing membrane vesicles[10-12] or capsids[4], gene-transfer agents[13] and, usually overlooked, non-specifically stained organic and/or mineral particles[14,15]. 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 are required prior to epifluorescence-based quantification[16]. VMRs are particularly disparate in sediments and soils[16-18] and vary depending on the protocol[19]. Nonetheless, several observations suggest that VLPs proportionally diminish with increasing cell density[5,20]. It has been hypothesized that lower VMR in cell-dense environments does not necessarily correlate with less viral load but with higher lysogenic prevalence (Piggyback-the-winner model)[5]. However, while VMR seems to decline with increasing microbial cell densities, evidence for enhanced lysogeny is unclear[21].

In principle, crowded cellular environments might hamper virus dispersal, limiting encounter rates with appropriate hosts and augmenting VLP decay. Nonetheless, substrate density and/or complexity are intrinsic dispersal-limitation factors *per se*. Here, we put forward the hypothesis that compact environments have much lower VMR than aquatic ecosystems and that obligate cellular epibiotic parasitism along with viral lysogeny are successful strategies under those conditions. The high diversity and relative abundance of DPANN archaea and CPR bacteria in sediments, microbial mats or subsurface ecosystems might support this idea[22-24]. Members of these taxa are widely diverse and, although mostly uncultured, their genome content indicates reduced metabolic capabilities incompatible with autonomy[23,25]. The few described DPANN and CPR members correspond to small obligate symbiotic cells physically attached to prokaryotic hosts. Some of them, notably CPR bacteria, are parasites/predators[26,27]; others can display mutualistic interactions (e.g. complementation of host's metabolism to degrade specific substrates[28]) although, upon environmental change, they likely behave as parasites, shifting along the mutualism-parasitism continuum[29]. Testing these hypotheses ultimately requires reliable quantification of virus-to-cell and episymbiont versus free-living cell ratios that can be comparatively applied across biomes. Here, we develop a metagenome-based strategy to estimate those rates that we first apply to *de novo* generated aquatic-system metagenomes integrating coexisting virus and cell fractions (50 kDa-200 µm) before extending it to a total of 133 metagenomes across a variety of ecosystems, notably including high-density habitats. Our results suggest that current measures based on fluorescently labelled VLP counts overestimate VMR and confirm that, in compact habitats, free-living cells massively outnumber viruses and, frequently, so do episymbiont cells.

To develop a metagenome-based VMR estimate (mVMR) and compare it to classical epifluorescence-derived estimates (fVMR), we first sequenced global metagenomes (including genomes of cells and all viral stages) of selected aquatic ecosystems along a salinity gradient. These 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.1). To obtain more representative VMR, we sampled these ecosystems in late winter (see Methods), avoiding blooming periods when richness and evenness decrease[30]. An inherent problem for global plankton studies derives from the fact that organisms of different cell-size are separated by filtration and DNA purified from the biomass retained in different pore-size filters. Most available aquatic metagenomes comprise 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 VLPs ('metaviromes') are obtained after tangential filtration of <0.2 µm filtrates to retain particles larger than 50-100 kDa. Therefore, cellular metagenomes and metaviromes are usually generated from different water volumes/samples, sequenced from different libraries and analysed separately, such that mVMR cannot be directly and reliably estimated. 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 size (100 and 30 µm pore-size filters were used for the Etang des Vallées, EdV, to avoid the many suspended plant material-derived particles in this shallow lake). Tangential filtration was applied to the resulting <0.2 µm-filtrate to concentrate VLPs (>50 kDa). We retained aliquots of all sequential filtrates for FCM counts of fluorescently labelled cells and VLPs, and the final eluate (<50 kDa) as a blank (see Methods; Supplementary Table 1 and Extended Data Fig.2). FCM counts of cells in all fractions >0.2 µm were consistently similar for each sample type, suggesting that the vast majority of cells were prokaryotic. Cell abundances were higher in saltern ponds (5.2×10^6 to 1×10^7 cells.ml^{-1}) and freshwater samples (4.3 to 4.7×10^6 cells.ml^{-1}) compared to marine samples, one order of magnitude lower (4.5 to x10^5 cells.ml^{-1}) (Fig.1a; Supplementary Table 1). VLP counts were also consistent in large-size fractions of each sample and diminished in the tangential-filtration concentrate. This indicates that many VLPs were retained in larger fractions due to progressive filter saturation with biomass and highlights the importance of integrating 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). Estimated fVMR ranged from 3 to 16, with lower values for freshwater, 40 m-deep Mediterranean and 14%-salt pond (3-6) and higher for 20 m-deep Mediterranean (13-15) and 14-32% saltern ponds (9-16) (Fig.1a). We purified DNA from the biomass retained in all filters used during serial filtration as well as from tangential-filtration concentrates, keeping track of the corresponding water volumes. DNA amounts obtained from different fractions were then proportionally mixed for a given original water volume (Methods; Supplementary Table 2). We then sequenced metagenomes for 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 prokaryotes and pico-nano-microeukaryotes (50 kDa-200 µm; 'all' fraction). In this way, coexisting virus-cell DNA was sequenced from the same library, avoiding technical biases. As eukaryotes represented a minor proportion in these aquatic systems, the two metagenomes were expected to behave as quasi-replicates for mVMR. In addition, we sequenced two 'metavirome' fractions (50 kDa-0.2 µm; 'vir' fraction) for the marine samples as internal control (Extended Data Fig.1; Supplementary Table 3).

To estimate mVMRs, we used universal single-copy genes (USCGs) and viral DNA polymerases (DNApols) as proxies for individuals. We selected 19 USCGs well-represented across the three domains of life (see Methods). Most of them were ribosomal proteins but we also retained four genes involved in amino acid and nucleotide biosynthesis (Supplementary Table 4). Normalised numbers of these USCGs were globally congruent in our aquatic metagenomes, although with slight among-marker differences (Extended Data Fig.3). For instance, five ribosomal markers have two copies in halophilic archaea (which we took into account) and amino acid and nucleotide metabolism genes are underrepresented in DPANN and CPR groups (Supplementary Table 5), possibly as a consequence of their dependency on host metabolism[23,25]. To limit differential representation, we averaged values for the 19 USCGs as a proxy for cellular counts under the assumption that in most biomes cells have one genome copy. Rare cases of polyploidy have been described in giant bacteria[31] and two chromosome copies usually occur in some archaea and actively transcribing bacteria[32]. However, in the wild, especially in energy-limited ecosystems, microbial cells are likely to have low average metabolic states and, consequently, limited ploidy. Should average ploidy be higher than one, this proxy might slightly overestimate cell numbers. At the same time, some markers do not capture well episymbiotic prokaryotes with reduced and divergent genomes (Extended Data Fig.3), which might partly compensate for ploidy excess in some cells. From a lifestyle perspective and for the purpose of this study, we considered DPANN and CPR groups as episymbiont/parasites and the rest of cellular taxa as free-living. Non-DPANN archaea are not known to be parasites[33] and, although several non-CPR bacteria are parasites, they affect eukaryotes, which are in minor proportions in most microbial ecosystems. Likewise, parasitic protists exist, but their relative abundance is low as compared to prokaryotes in microbial ecosystems and most of them do not belong to parasitic clades, e.g.[34]. To count viruses, we used DNApols identified using Virus Orthologous Groups (VOGs)[35] as viral markers (Supplementary Table 6). DNApols were chosen as reasonable proxies because they are the most conserved, widespread and relatively easy-to-identify proteins in DNA viruses[36,37]. In principle, DNApols exclude RNA viruses. However, dsDNA viruses largely dominate the prokaryotic virome[37] and, since most microbial ecosystems are overwhelmingly dominated by prokaryotes, counting DNApols might be suitable to estimate the order of magnitude of viruses. To use DNApols as proxies for individual viruses, we needed to validate the assumption that, on average, one DNApol occurs per viral genome. To do so, we plotted all available complete viral genomes at the NCBI Viral RefSeq database as a function of their genome size, viral class and DNApol copies (Extended Data Fig.4; see Supplementary Fig.1 for their distribution as a function of host and viral family). Although the number of DNApols varied from 0 (Supplementary Table 7) to over 7 copies, proportionally to genome size[36], the average number of DNApols per dsDNA viral genome was 1.36. This might lead to a slight overestimation of dsDNA viruses that potentially might compensate for undetected RNA viruses, although some negative-strand ssRNA and reverse-transcribed viruses were also identified, along with some ssDNA viruses, by our selected VOGs (Extended Data Fig.4; Supplementary Table 6). Therefore, the 1 DNApol ≈ 1 virus appears a reasonable assumption for mVMR estimates in prokaryote-dominated biomes. We then summed counts of all the different VOG-identified DNApols as an approximation of the number of viruses in our aquatic metagenomes. VOGs allowed the identification of slightly more DNApols on average than Pfams but also many cellular polymerases, which were excluded from viral counts (Extended Data Fig.5; see Supplementary Fig.2 for specific VOG DNApol counts). The number of viral DNApols was considerably higher in marine 'metaviromes' than in other planktonic fractions, indicating that, as expected, VLPs were enriched after tangential filtration. However, although the number of total cellular genes considerably decreased in 'metaviromes', they were still present (Extended Data Fig.6). This explains the detection of polymerases that likely correspond to ultrasmall prokaryotes and free cellular DNA after filtration-induced cell lysis.

In most metagenomes of aquatic systems, normalised viral counts were higher than those for cells, essentially bacteria (except for 32% salt ponds, dominated by haloarchaea); eukaryotes, DPANN and CPR counts were comparatively low (Fig.1b). However, estimates of mVMRs in our aquatic ecosystems were considerably lower than fVMRs observed in the same samples (Fig.1c), despite the fact that they account for total viruses (virion, lytic and lysogenic forms). Viruses were more abundant than cells in freshwater and marine plankton but only by a factor of 2-2.5. Comparatively, fVMR estimates were 2-7 times higher (Fig.1a). In solar salterns, viruses were roughly as abundant as cells (1:1) in 6% and 14% salt plankton metagenomes and very low (~1:10) in the case of 32% salt brine, although archaeal cells were as abundant here as bacteria in lower-salt ponds (Fig.1b-c). This seems at odds with the high number of FCM-identified VLPs at the highest salinity (Fig.1a). However, previous studies have shown that, in extreme saturated brines devoid of cells, small halite crystals and silicate biomorphs can be non-specifically stained with DNA dyes such as SYBR Green and DRAQ5 and cell-sorted based on this fluorescence, as shown by subsequent scanning electron microscopy images of sorted particles[15]. This strongly suggests that, in addition to potential cell-DNA-containing vesicles, capsids and gene-transfer agents[4,10-13], tiny suspended mineral particles can be mistaken by viruses, resulting in inflated fVMRs. We subsequently calculated mVMRs in up to 133 total metagenomes 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.1; Supplementary Table 3). We generated 40 of those additional metagenomes, notably from poorly studied ecosystems. To avoid potential confounding factors, all chosen metagenomes were sequenced using the same technology and treated in the same way starting from raw reads (Methods). Although in general, dsDNA viruses dominated, ssDNA viruses had high abundances in some metagenomes and, occasionally, reverse-transcribed viruses captured (Extended Data Fig.7). Normalised metagenome counts showed that free-living bacteria and archaea, and sometimes parasitic episymbionts (DPANN-CPR) were more abundant than viruses (Extended Data Fig.8; Supplementary Fig.3). This translated in systematically low mVMRs for complex, dense ecosystems, with most virus:free-living cell ratios between 0-0.5:1 (Fig.2). Only in 6 out of the 119 non-planktonic metagenomes, viruses approached the 1:1 ratio. A single case exhibited a 2.5:1 virus-cell ratio but from extremely low number of viral and cellular normalised counts (Extended Data Fig.8), suggesting potential stochastic effects. Furthermore, although viruses appeared more abundant than parasitic cells in 38% of these metagenomes, DPANN-CPR groups exceeded viruses in 25% of them. Therefore, although viruses outnumbered cells in aquatic ecosystems, where DPANN-CPR groups were minor, albeit by a lower factor (~2:1) than generally credited based on epifluorescent VLP detection (~10:1), cells largely dominated viruses in high-density, complex ecosystems, with cellular episymbiosis often surpassing viral predation.

To better understand the distribution of viruses, episymbionts and free-living cells across biomes, we carried out principal component analyses (PCA) looking for potential determinants that best explained the variance in our metagenome-wide analysis. Virus counts clearly correlated with the planktonic realm, whereas episymbiont/parasite cells tightly correlated with free-living cells and some soil, marine sediment, microbial mat and animal-associated microbiomes (Fig.3a). We calculated richness, alpha-diversity and evenness indices for viruses, episymbiont and free-living cells using DNApols and two representative and universally distributed ribosomal markers to define operational taxonomic units (Methods; Supplementary Fig.4). Like viral abundance, viral diversity correlated with freshwater and marine plankton. However, despite our plankton samples were collected in winter, viral evenness correlated with more viral-deprived, complex ecosystems such as microbial mats, marine and most freshwater sediments and soils, being generally anticorrelated with animal microbiomes (Fig.3b). Likewise, parasite and free-living cells exhibited highly correlated diversity and evenness in this same type of habitats (Fig.3b) but did not significantly correlate with their respective abundance (Extended Data Fig.9a). To test whether viral lysogeny increase in cell-dense environments, as postulated by the Piggyback-the-winner model[5], we searched for a collection of lysogeny-related VOGs, including integrases, excisionases and prophage-related genes (Supplementary Table 8) in our metagenome collection and calculated their normalised counts. The three of them significantly correlated with total cell abundance, but not with viral abundance (Fig.3c). Coinciding with high cellular abundances (i.e. normalised USCG counts; Extended Data Fig.10), and therefore in agreement with the Piggyback-the-winner model, this trend was especially driven by some animal-related, notably gut, microbiomes, but also some (though not all) marine sediment and soil samples, as shown in a PCA (56.3% of the variance explained by axis 1; Fig. 3c) and regression analyses (Supplementary Fig.5). This highlights the heterogeneity of complex environments and potentially explains conflicting observations of lysogeny-related genes in other complex metagenomes[5,21]. Furthermore, lysogeny-related genes correlated with free-living but not episymbiotic cells (Extended Data Fig.9b). This is likely explained by the ongoing genome reduction process in DPANN and CPR groups. Pseudogenization and gene loss would preclude the long-term residence of prophages in these episymbiotic lineages.

In this work, we have developed a metagenome-based VMR metric that is applicable in a comparative manner throughout microbial ecosystems. Our approach is not fully devoid of potential biases that might affect its accuracy resulting in slight relative over- or underestimates of cells and/or viruses, such as differences in cell ploidy or DNApol copy number per individual and limitations to detect some RNA viruses or divergent viral DNApols but also cellular USCGs. However, given that most prokaryotic viruses are DNA viruses and that prokaryotes dominate the Earth microbiome, mVMR are likely much better estimates than fVMR at least at the order-of-magnitude level. Indeed, fluorescently-based approaches are incomplete (only count VLPs) and artefact-prone. Beyond identifying specifically stained vesicles and capsids containing non-viral DNA, fluorescence-based methods are affected by biases linked to the differential treatment of contrasting sample types (water, soil, sediment, mucilaginous mats) and may 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 numbers[3,4]. In addition, mVMR is inclusive, allowing to count lytic viruses infecting cells as well as lysogenic forms. Comparative estimates of mVMRs across uniformly-analysed metagenomes appear low to very low in complex, high-density ecosystems. Obviously, there is variability and VMRs should increase in blooming periods, albeit accompanied by a lower viral alpha-diversity. However, viruses appear only more abundant in plankton but just by a factor of 2:1, and not 10:1, as generally believed. Consequently, contrary to a pervasive belief, viruses are not more abundant than cells on Earth. On the contrary, given that most complex, cell-rich ecosystems seem to harbour few viruses, cells appear to largely outweigh viruses in the biosphere. Furthermore, since a portion of viral particles are defective[38-40], the real number of infective viruses is even smaller, such that attempts to infer infection rates in natural ecosystems based on VLP counts may be futile. Our study also reveals lysogeny and episymbiosis as important strategies in complex ecosystems where dispersal is limited. Furthermore, episymbiotic bacteria and archaea outnumber viruses in some dense biotopes, suggesting that prokaryotic parasitism and predation constitute an unforeseen mechanism of population control in yet poorly studied microbial ecosystems.

Methods

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)[41]. 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 (10l) from a shallow lake in North-western France (Etang des Vallées; 48°41'23,0''N, 01°54'59.2''E)[34] 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. 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 1). 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 2. 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 1, Supplementary Table 3).

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 1). We used a CytoFLEX S (Beckman Coulter) equipped with two active diode lasers (488 nm and 561 nm) following well-established protocols[42,43]. 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[43]. Background noise was systematically checked on blanks (Extended Data Fig.2). 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.2). 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[42]. 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 1)

DNA purification of sediment and microbial mats, sequencing and selection of 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 1, Supplementary Table 3). DNA from Lake Baikal sediment samples collected in July 2017 (Reboul et al., in prep) was purified using the Power Soil™ DNA purification kit (Qiagen, Germany). 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[44], 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[44-57] (Extended Data Fig.1; an extended description, GenBank accession numbers and statistics are provided in Supplementary Table 3).

Sequence analysis, annotation and taxonomic and lifestyle assignation

We treated the 133 metagenomes with exactly the same bioinformatic pipeline starting from raw Illumina sequences. Read quality was verified with FastQC[58] v0.11.8 and cleaned with Trimmomatic[59] 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[60] 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[61] 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 3). Gene annotation was performed with Prokka[62] v1.14.5 in metagenome mode (contigs >200 bp). We assigned coding sequences to PFAMs (Pfam-A database v3.1b2) with HMMER hmmsearch[63] v3.2.1. Simultaneously, we blasted annotated genes with Diamond[64] v0.9.25.126 against the non-redundant database used by Kaiju[65] v1.7.1: kaiju_db_nr_euk.faa, which contains all NCBI reference sequences for archaea, bacteria, eukaryotes and viruses. We then selected the best-hit having more than 35% identity over at least 70% coverage and created a best-hit table in kaiju output format-looking file using an *ad hoc* Perl script (mimicKaijuOutput.pl) and used the kaiju script addTaxonName to retrieve the taxonomic classification. 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) reduced CPR bacteria (genomes <1.2 Mbp) and DPANN archaea, considered as cellular episymbionts/parasites strictly dependent on free-living organisms (Supplementary Fig.6; Supplementary Table 5); ii) ‘Other Archaea’, ‘Other Bacteria’ and ‘Eukaryotes’, classed as free-living, and ‘Viruses’. 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. We carried out a Diamond blastp search of all genes assigned to viruses against the collection of complete viral genomes from NCBI (as of April 2020). 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-type (Baltimore classification). Taxonomy and host metadata were obtained from the NCBI and the Baltimore classification from ExPaSy ViralZone[66]. The relative abundance of genes from different viral families, hosts and viral group-type in metagenomes was plotted using the R ggplot2 package[67].

Metagenome-based counting of cells and viruses

To estimate the number of cells from metagenomes, we averaged counts of 19 Universal Single-Copy Genes (USCGs) as identified from their PFAMs. They included 15 ribosomal proteins and four metabolic genes present in the vast majority of organisms including most reduced CRP and DPANN members (Supplementary Table 5; Extended Data Fig.3). To estimate the total number of DNA viruses (viral particles and intracellular viruses −lytic or lysogenic−), we used as proxy the number of DNA polymerases (DNApol; Supplementary Table 6). In the case of multiple subunits, we selected one. We also excluded polymerase helper proteins. To identify viral DNA polymerases, we relied on the Virus Orthologous Groups (VOGs) database[35] release vog96, since VOGs allowed retrieving more viral DNA polymerases as compared to PFAMs in our aquatic ecosystems (Extended Data Fig.5). The rest of metagenomes were therefore also annotated against the VOG database. We then made the assumption that most viral genomes encode a single DNA polymerase, such that 1 DNApol ≈ 1 virus. To validate this assumption, we downloaded all complete genomes (6,569) from the 9,239 viral genomes available at NCBI (April, 2020) and ran a HMMER hmmsearch to count how many times a DNApol (VOG database) was present in each genome. Although many viruses have only one DNApol, some viruses have multiple polymerases (between 2 and 7, occasionally more), clearly correlating with genome size [36]. Violin plots indeed show a non-normal distribution of polymerases as a function of viral genome size to which a regression line generated with the non-parametric Loess method[68] can be fitted (Extended Data Fig.4). Accordingly, counting 1 DNApol as 1 virus would result in an overestimation of viruses. However, several viruses do not encode DNA polymerases, most of them RNA viruses (Supplementary Table 7). Nonetheless, given that dsDNA viruses largely dominate the prokaryotic virome[37], RNA viruses infecting prokaryotes being rare, it is likely that our counting approximation (1 DNApol ≈ 1 virus) compensates for DNA polymerase-lacking viruses in metagenomes. This might also compensate for excessively divergent polymerases missing from VOGs. For each metagenome, we created two databases of marker genes, one consisting of cellular USCGs (we excluded USCGs whose taxonomy was assigned to viruses) and another of exclusively viral DNApols (VOG-identified cellular polymerases were excluded). The nucleotide sequences of these two sets of marker genes were indexed with Bowtie2[69] v2.3.5.1 and metagenome reads mapped onto them. We retrieved the mapped reads with Samtools[70] v1.9 and calculated, for each gene, the Reads Per Kilobase (of gene sequence) per Million of mapped metagenome reads (RPKM) using *ad hoc* Perl scripts, allowing normalization for gene length and sample sequencing depth. For this, we first counted the total number of reads in each sample and divided it by 1,000,000 ("per million" scaling factor). Second, we divided the mapped gene read counts to the marker gene by the "per million" scaling factor to normalise for sequencing depth and obtain the number of marker reads per million of sequences (RPM). Finally, we divided RPM values by the corresponding marker gene length (kb) to which the read mapped to obtain RPKM values. Subsequently, we averaged the USCGs (RPKM) and summed viral DNApols (RPKM) to obtain normalised values of cells versus viruses per metagenome. USCGs (RPKM) estimates were calculated for broad phylogenetic clades (DPANN, Other Archaea, CPR, Other Bacteria, Eukaryotes) or for groups of particular lifestyle (BAE, including free-living bacteria, archaea and eukaryotes; SP, symbiont/parasites, including CPR and DPANN; a third, minor, category was implemented for not-assigned counts). For an easier visualization of the proportion of viruses and cellular parasites/symbionts per individual free-living cell, we further normalised the RPKMs of these three functional categories (BAE, SP and not assigned) with respect to the number of free-living cells (value = 1). Plots were generated with *ad hoc* R[71] scripts using the ggplot2 package.

Diversity indices

To calculate diversity estimates, we retained as markers ribosomal proteins L14 and L16 (PF00238 and PF00252 respectively) from the previous set of USCGs, as they consistently retrieved members of the three domains of life across ecosystems (e.g. Extended Data Fig.3). Amino acid sequences for each ribosomal protein were retrieved with *ad hoc* Perl scripts for each ecosystem and clustered with cd-hit[72] 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, if the taxonomic assignment of the two genes were consistent (manual BLAST-based phylogenetic assignation was applied to conflicting clusters), or NA (not assigned to a particular phylogenetic clade) otherwise. For viruses, we used the whole set of viral DNApols under the premise that most viral genomes have only one DNApol (see above). We obtained a total of 21,205 clusters for the 2 cellular markers (10,159 for rpL14; 11,046 for rpL16) and 6,203 viral DNApol clusters. Diversity indices obtained for both ribosomal proteins were nearly identical (rpL16 results are displayed in Supplementary Fig.4). We also obtained diversity measurements for cellular lifestyle categories: free-living BAE (Bacteria, Archaea, Eucarya; 19,001 clusters) and parasitic/symbiotic (CPR and DPANN; 1,847 clusters). 334 protein ribosomal clusters remained unassigned. From the 6,203 viral clusters, 6 were phylogenetically classified as bacterial upon manual inspection and eliminated from the analysis. 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[73]. Indices for all metagenomes were plotted with ggplot2.

Identification of lysogeny-related genes

To test whether lysogenic viruses increase in high cell-density environments, we searched for integrase, excisionase and prophage-specific proteins as defined by their VOGs with hmmsearch (same parameters as for DNApol identification, see above) (Supplementary Table 8). Temperate, lysogeny-related genes for each metagenome were indexed and reads were mapped back to calculate their abundance in RPKM, as above.

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 numbers XXXXXX. Accession numbers for all used metagenomes are provided in Supplementary Table 3.

Code availability

All code is available on GitLab (https://gitlab.com/anagtz/mvmrs).

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. *Nature reviews. Microbiology* **5**, 801-812 (2007).

3 Cobián Güemes, A. G. *et al.* Viruses as Winners in the Game of Life. *Annual review of virology* **3**, 197-214 (2016).

4 Mushegian, A. R. Are there 10^{31} virus particles on Earth, or more, or less? *Journal of Bacteriology*, JB.00052-00020 (2020).

5 Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 466-470 (2016).

6 Paez-Espino, D. *et al.* Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* **7**, 157 (2019).

7 Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nature reviews. Microbiology* **18**, 125-138 (2020).

8 Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548 (1999).

9 Suttle, C. A. Viruses in the sea. *Nature* **437**, 356-361 (2005).

10 Forterre, P., Soler, N., Krupovic, M., Marguet, E. & Ackermann, H. W. Fake virus particles generated by fluorescence microscopy. *Trends in microbiology* **21**, 1-5 (2013).

11 Schatz, D. *et al.* Communication via extracellular vesicles enhances viral infection of a cosmopolitan alga. *Nature Microbiology* **2**, 1485-1492 (2017).

12 Toyofuku, M., Nomura, N. & Eberl, L. Types and origins of bacterial membrane vesicles. *Nature Reviews Microbiology* **17**, 13-24 (2019).

13 Lang, A. S., Westbye, A. B. & Beatty, J. T. The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annual review of virology* **4**, 87-104 (2017).

14 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. *Journal of microbiological methods* **59**, 189-198 (2004).

15 Belilla, J. *et al.* Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. *Nature Ecol Evol* **3**, 1552–1561 (2019).

16 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. *Biological reviews of the Cambridge Philosophical Society* **92**, 1081-1100 (2017).

17 Danovaro, R. *et al.* Viriobenthos in freshwater and marine sediments: a review. *Freshwater Biology* **53**, 1186-1213 (2008).

18 Williamson, K. E., Fuhrmann, J. J., Wommack, K. E. & Radosevich, M. Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annual review of virology* **4**, 201-219 (2017).

19 Wei, M. & Xu, K. New insights into the virus-to-prokaryote ratio (VPR) in marine sediments. *Frontiers in microbiology* **11**, 1102 (2020).

20 Wigington, C. H. *et al.* Re-examination of the relationship between marine virus and microbial cell abundances. *Nature Microbiology* **1**, 15024 (2016).

21 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).

22 Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208-211 (2015).

23 Dombrowski, N., Lee, J. H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS microbiology letters* **366**, fnz008 (2019).

24 He, C. *et al.* Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat Microbiol* (2021).

25 Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature reviews. Microbiology* **16**, 629-645 (2018).

26 Bor, B. *et al.* Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 12277-12282 (2018).

27 Moreira, D., Zivanovic, Y., Lopez-Archilla, A. I., Iniesto, M. & Lopez-Garcia, P. Reductive evolution and unique infection and feeding mode in the CPR predatory bacterium *Vampirococcus lugosii*. *bioRxiv*, 2020.2011.2010.374967 (2020).

28 La Cono, V. *et al.* Symbiosis between nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides. *Proceedings of the National Academy of Sciences of the United States of America* **online early**, 202007232 (2020).

29 López-García, P. & Moreira, D. Physical connections: prokaryotes parasitizing their kin. *Environ Microbiol Rep*, [epub ahead of print] doi.10.1111/1758-2229.12910 (2020).

30 Pommier, T. *et al.* Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquatic Microbial Ecology* **61**, 221-233 (2010).

31 Mendell, J. E., Clements, K. D., Choat, J. H. & Angert, E. R. Extreme polyploidy in a large bacterium. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6730-6734 (2008).

32 Touchon, M. & Rocha, E. P. Coevolution of the Organization and Structure of Prokaryotic Genomes. *Cold Spring Harbor perspectives in biology* **8**, a018168 (2016).

33 Moissl-Eichinger, C. & Huber, H. Archaeal symbionts and parasites. *Curr Opin Microbiol* **14**, 364-370 (2011).

34 Simon, M. *et al.* Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. *The ISME journal* **9**, 1941-1953 (2015).

35 Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* **45**, D491-d498 (2017).

36 Kazlauskas, D., Krupovic, M. & Venclovas, Č. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res* **44**, 4551-4564 (2016).

37 Koonin, E. V. *et al.* Global organization and proposed megataxonomy of the virus world. *Microbiol Mol Biol Rev* **84** (2020).

38 Rezelj, V. V., Levi, L. I. & Vignuzzi, M. The defective component of viral populations. *Current opinion in virology* **33**, 74-80 (2018).

39 Vignuzzi, M. & López, C. B. Defective viral genomes are key drivers of the virus–host interaction. *Nature Microbiology* **4**, 1075-1087 (2019).

40 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).

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

42 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).

43 Brussaard, C., Payet, J., Winter, C. & Weinbauer, M. in *ASLO.* (eds S.W. Wilhelm, M.G. Weinbauer, & C.A. Suttle) 102-109.

44 Gutierrez-Preciado, A. *et al.* Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nature ecology & evolution* **2**, 1700-1708 (2018).

45 Dong, X. *et al.* Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nature communications* **10**, 1816 (2019).

46 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).

47 Zhang, X. *et al.* Metagenomics Reveals Microbial Diversity and Metabolic Potentials of Seawater and Surface Sediment From a Hadal Biosphere at the Yap Trench. *Frontiers in microbiology* **9** (2018).

48 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).

49 Vavourakis, C. D. *et al.* Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. *BMC biology* **17**, 69 (2019).

50 Ji, M. *et al.* Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* **552**, 400-403 (2017).

51 Xu, J. *et al.* The structure and function of the global citrus rhizosphere microbiome. *Nature communications* **9**, 4894 (2018).

52 Wilbanks, E. G. *et al.* Microscale sulfur cycling in the phototrophic pink berry consortia of the Sippewissett Salt Marsh. *Environmental microbiology* **16**, 3398-3415 (2014).

53 Saghaï, A. *et al.* Comparative metagenomics unveils functions and genome features of microbialite-associated communities along a depth gradient. *Environmental microbiology* **18**, 4990-5004 (2016).

54 Rebollar, E. A. *et al.* The Skin Microbiome of the Neotropical Frog Craugastor fitzingeri: Inferring Potential Bacterial-Host-Pathogen Interactions From Metagenomic Data. *Frontiers in microbiology* **9** (2018).

55 Regan, T. *et al.* Characterisation of the British honey bee metagenome. *Nature communications* **9**, 4995-4995 (2018).

56 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66 (2017).

57 Vangay, P. *et al.* US Immigration Westernizes the Human Gut Microbiome. *Cell* **175**, 962-972.e910 (2018).

58 Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).

59 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114-2120 (2014).

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

61 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods (San Diego, Calif.)* **102**, 3-11 (2016).

62 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* **30**, 2068-2069 (2014).

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

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

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

66 Hulo, C. *et al.* ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res* **39**, D576-582 (2011).

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

68 Jacoby, W. G. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies* **19**, 577-613 (2000).

69 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

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

71 R: A language and environment for statistical computing. v. http://www.r-project.org (R Foundation for Statistical Computing, Vienna, Austria, 2017).

72 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**, 3150-3152 (2012).

73 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 informatic support by P.D. 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

LIST OF SUPPLEMENTARY FIGURES AND TABLES

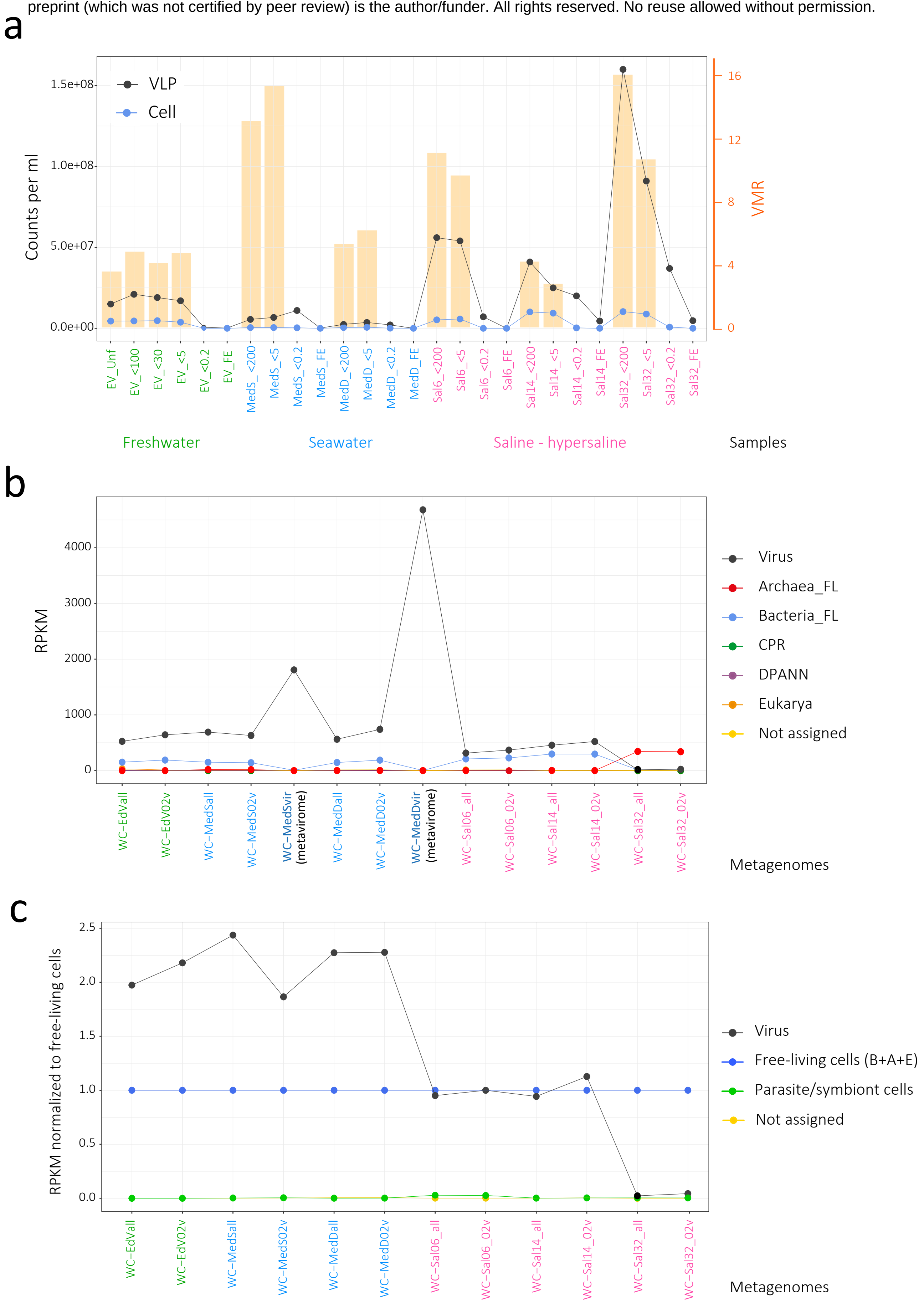


Fig.1. Epifluorescence- and normalized metagenome-based counts and ratios of cells and viruses in selected aquatic ecosystems across a salinity gradient. a, Flow-cytometry (FCM) counts of cells and viral-like particles (VPLs) and estimated virus-to-microbe ratios (histogram bars). Sample names indicate 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**, Normalized counts (RPKM) of viruses (DNApols) and different cellular types (USCGs) in global (cells + viruses) metagenomes from the same ecosystems. 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.1 and Supplementary Tables 2-3). Two metaviromes (fraction 50 kDa-0.2 µm) were sequenced for Mediterranean (MedS, MedD) plankton samples as internal control. **c**, Normalized ratios of all viruses (virion, lytic, lysogenic) and symbiotic/parasitic cells (DPANN+CPR) per free-living cell (B+A+E, bacteria, archaea, eukaryotes). RPKM, reads per kilo base per million mapped metagenome reads.

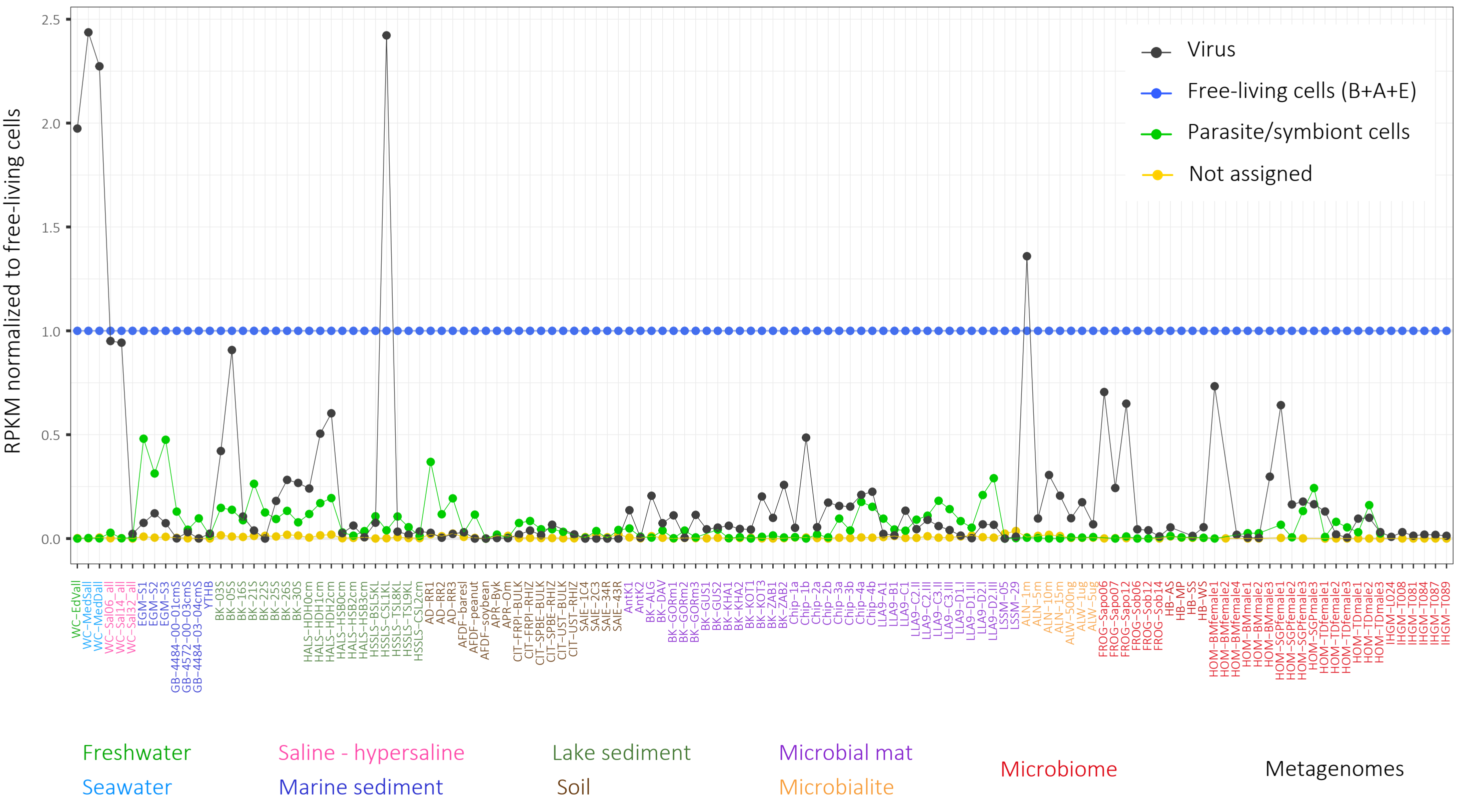


Fig.2. Metagenome-based normalized ratios of viruses and symbiotic/parasitic cells per free-living cell across ecosystem types. Viruses and cells were counted using DNApols and USCGs as proxies (see text). All the metagenomes were analysed using the same pipeline from raw reads. Symbiont/parasite cells correspond to DPANN archaea and CPR bacterial counts. The rest of identified bacteria, archaea and eukaryotes (B+A+E) were counted as free-living cells. RPKM, reads per kilo base per million mapped metagenome reads.

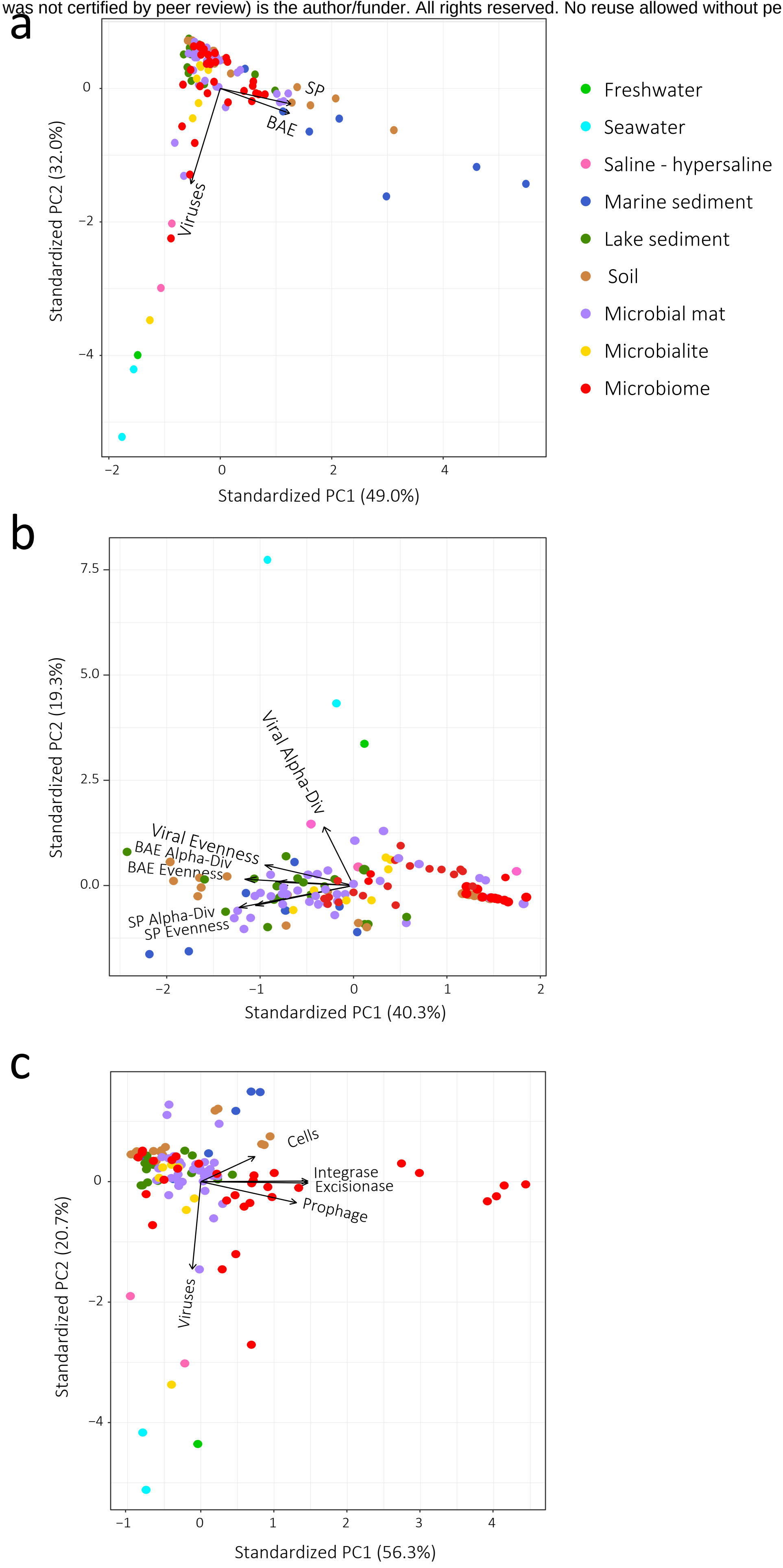
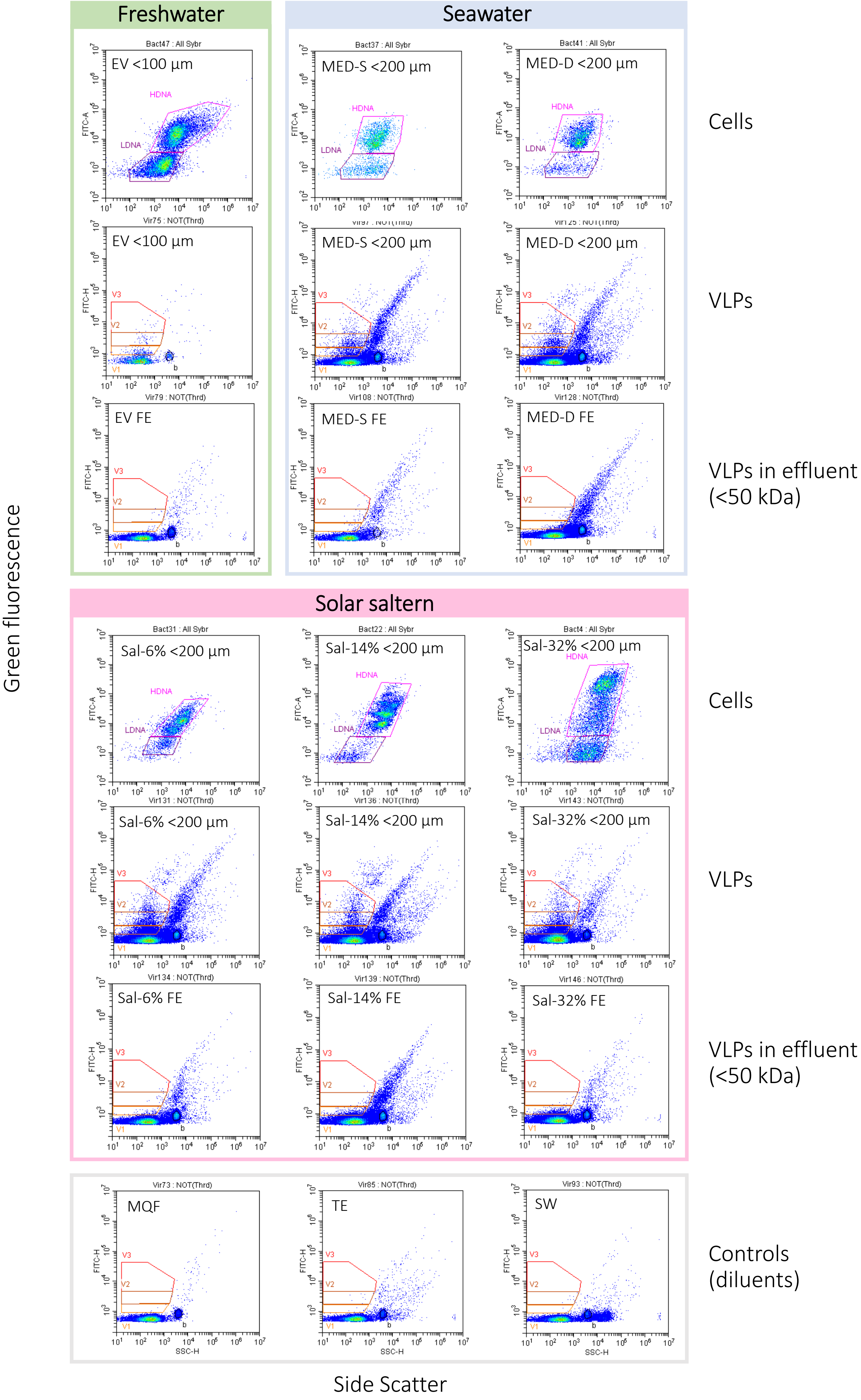


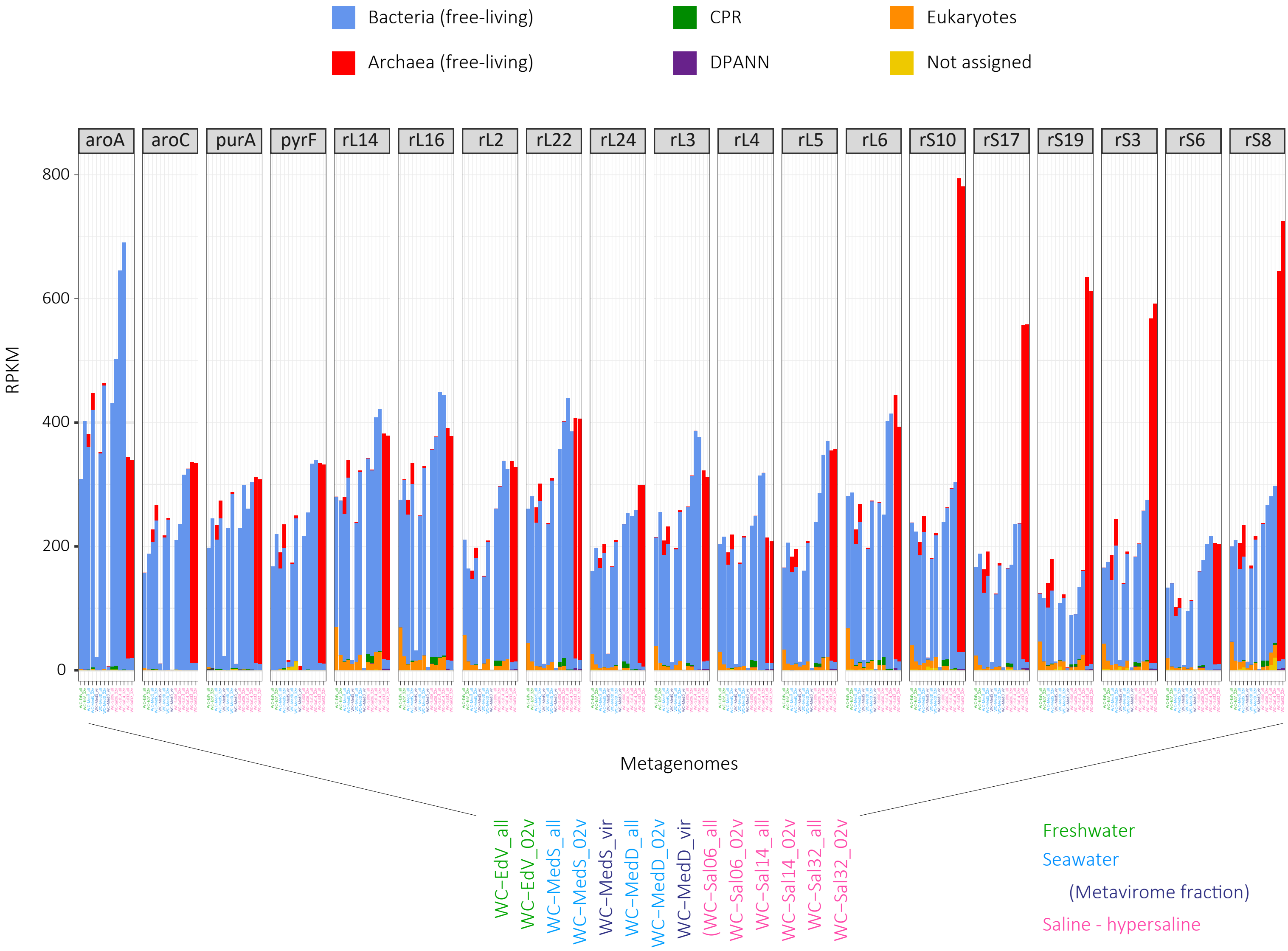
Fig.3. Principal component analyses of viruses, symbiont/parasite and free-living cells, diversity indices and lysogeny-related genes. a, PCA of different metagenomes maximizing the variance explained by viruses, symbiont/parasite cells (SP, DPANN archaea + CPR bacteria) and free-living cells (BAE, other bacteria, archaea and eukaryotes). **b**, PCA of different metagenomes as a function of viral, symbiont/parasite (SP) and free-living (BAE) cells alpha-diversity and evenness (diversity indices are detailed in Supplementary Fig.5). **c**, PCA of metagenomes maximizing the variance explained by viruses, cells and lysogeny-related genes.

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

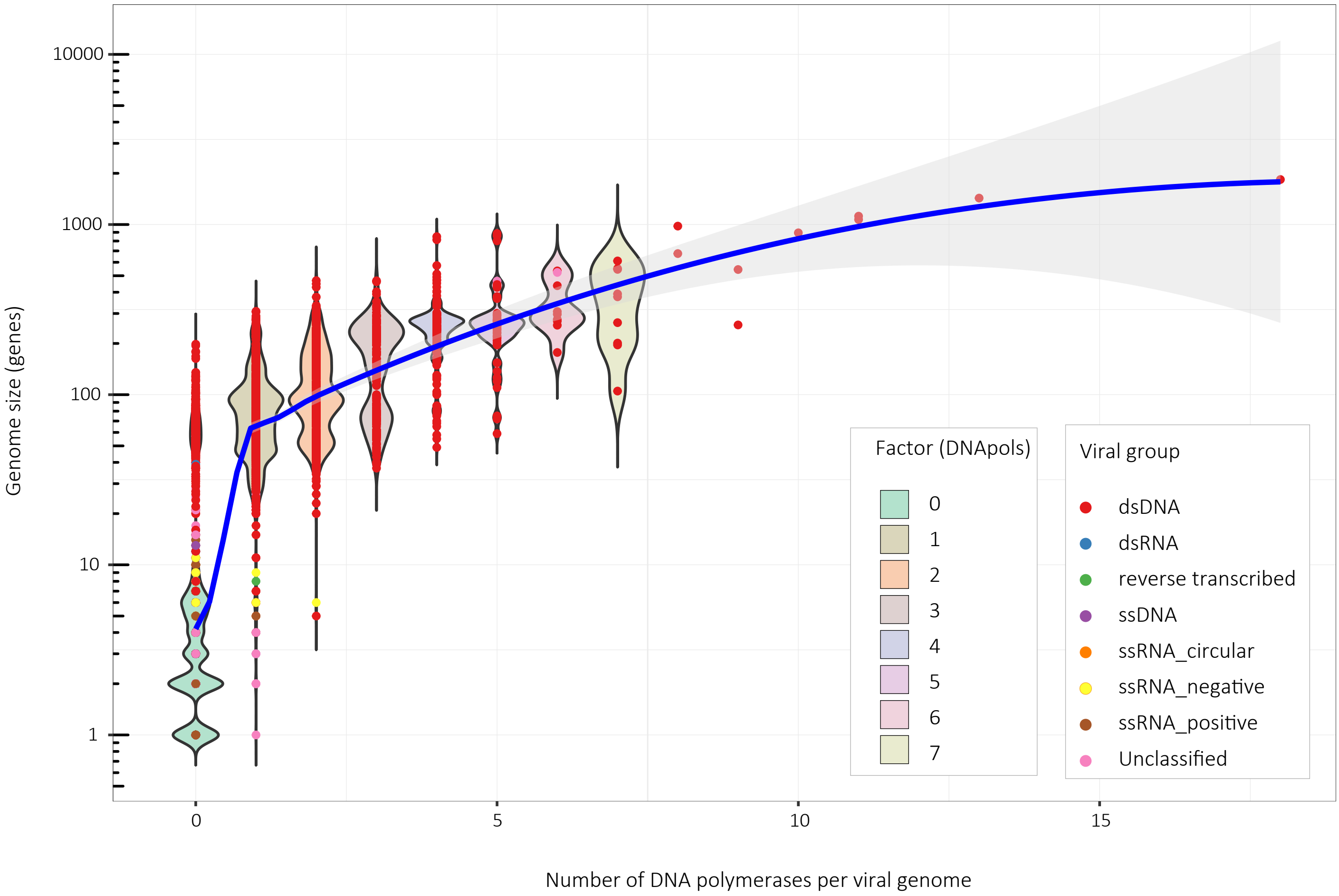
Extended Data Fig.1. Origin and main characteristics of analyzed metagenomes. Extended information about metagenome origin and statistics are given in Supplementary Table 3.
*metagenomes generated in our laboratory (this and previous or ongoing studies).



Extended Data Fig. 2. 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.1 and Supplementary Tables 1 and 3.

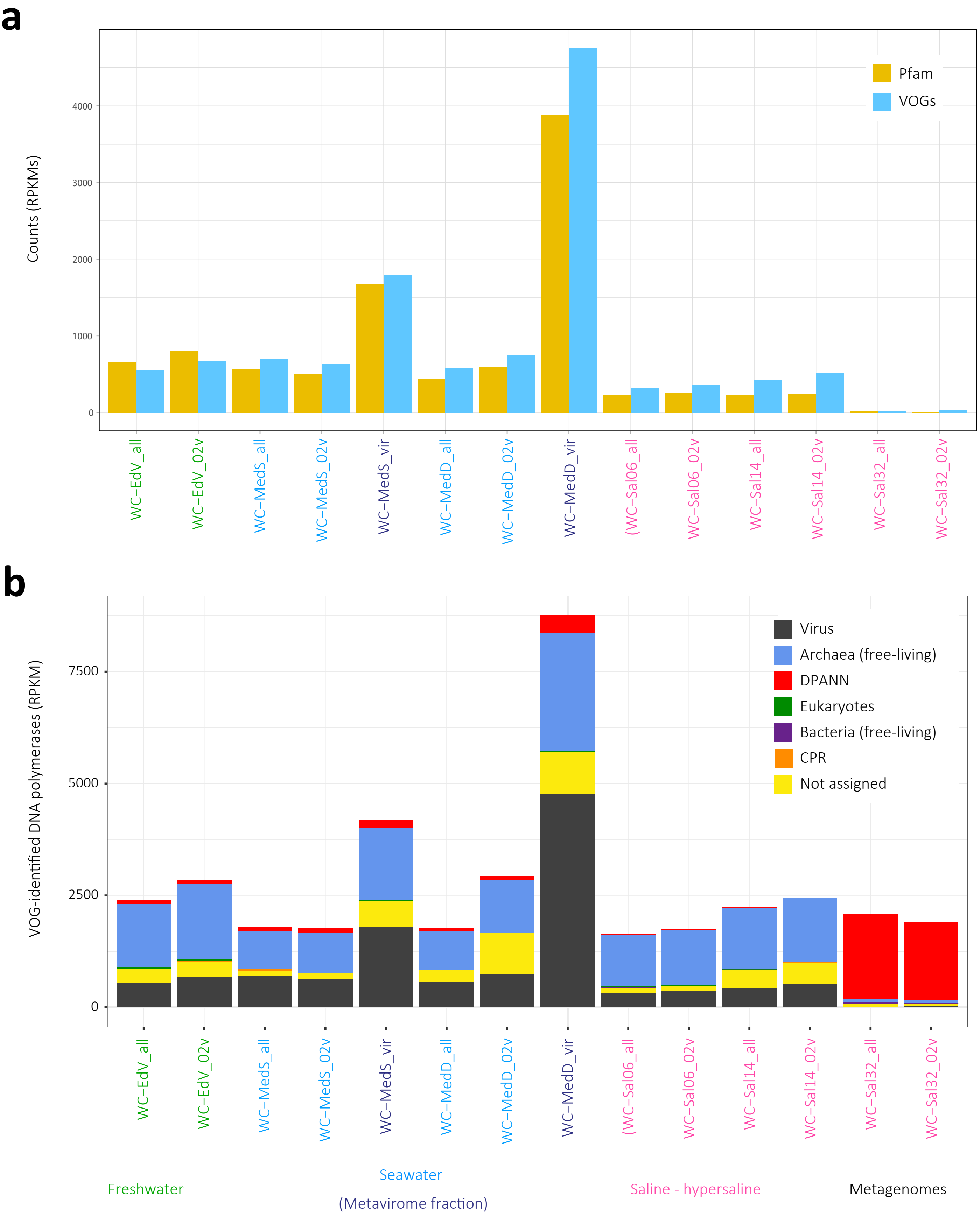


Extended Data Fig. 3. Normalized counts of selected universal single copy genes (USCGs) in aquatic metagenomes of coexisting virus and cell fractions. Suffixes in sample names indicate the plankton size fraction: _all, >50 kDa - 200 µm (100 µm in EdV); _02V, >50 kDa - 5 µm; _vir, >50 kDa - 0.2 µm (’metavirome’). 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.

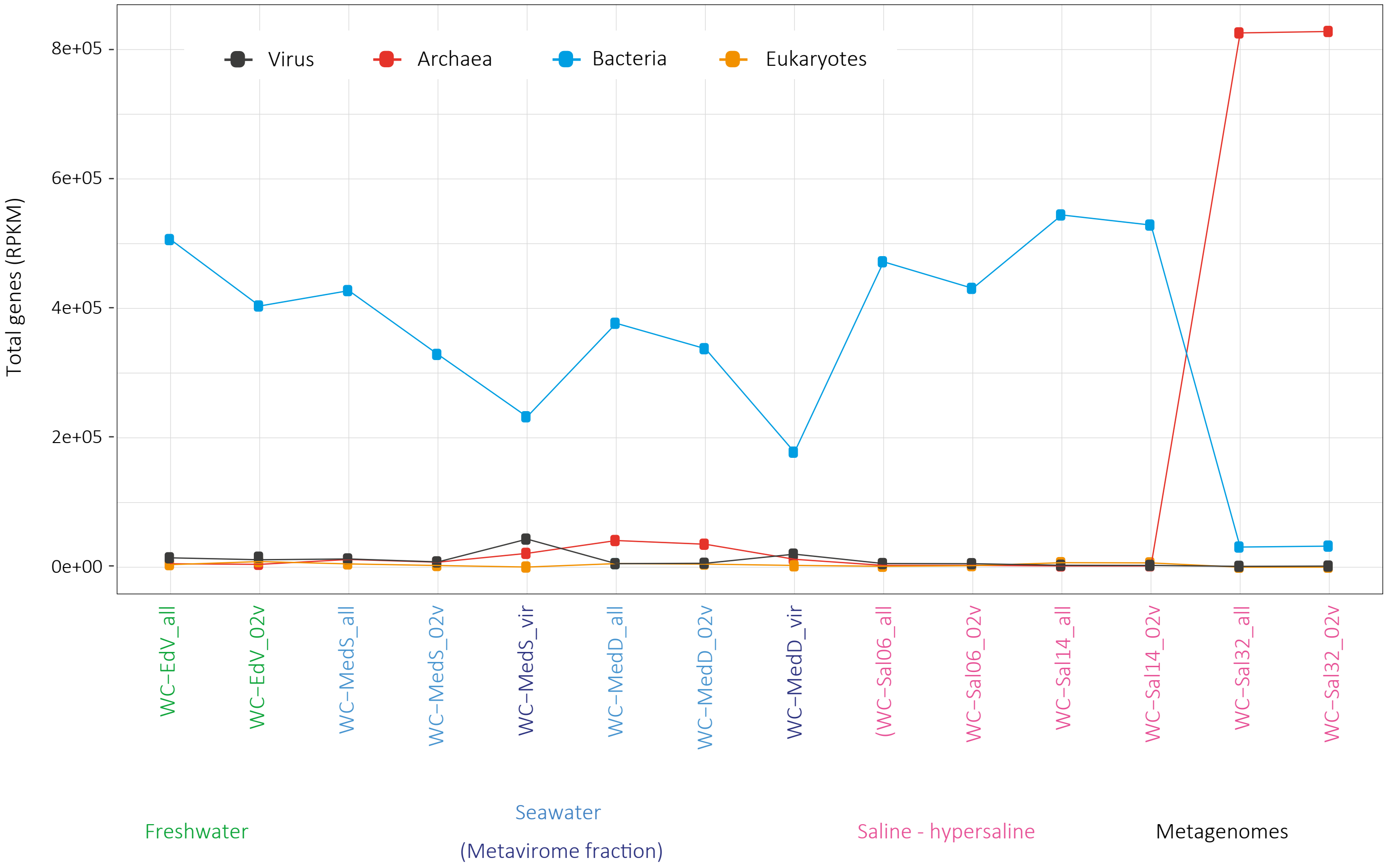


	All viruses	dsDNA viruses
% viruses 0 DNApol	71.34	32.36
% viruses 1 DNA pol	13.34	31.08
% viruses > 2 DNA pol	15.32	36.57
Average DNApol number/viral genome	0.58	1.36

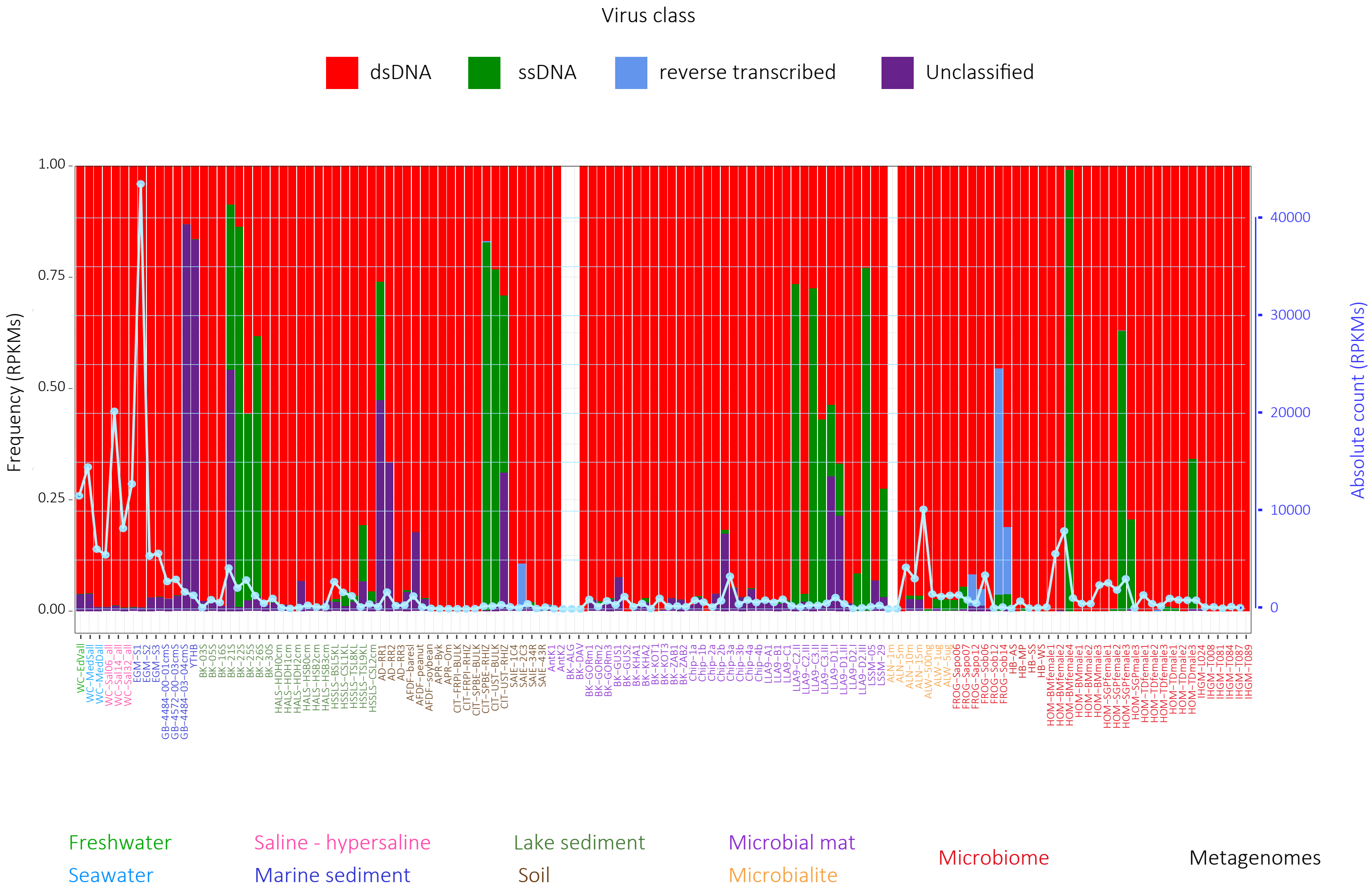
Extended Data Fig. 4. Number and proportion of DNA polymerases per viral genome. The scatter plot represents 6,569 viral genomes retrieved from the NCBI Viral Ref Seq database as a function of their genome size (expressed in number of genes), viral class and the number of (DNA) polymerases identified in each genome using VOG-based HMMs (see Methods). Some polymerases in RNA viruses are identified using these HMMs. The Y-axis is displayed in log10 scale for better visualization (dots more scattered). Violin plots show a non-normal distribution of polymerases as a function of viral genome size; the regression line was plotted using the non-parametric Loess method.



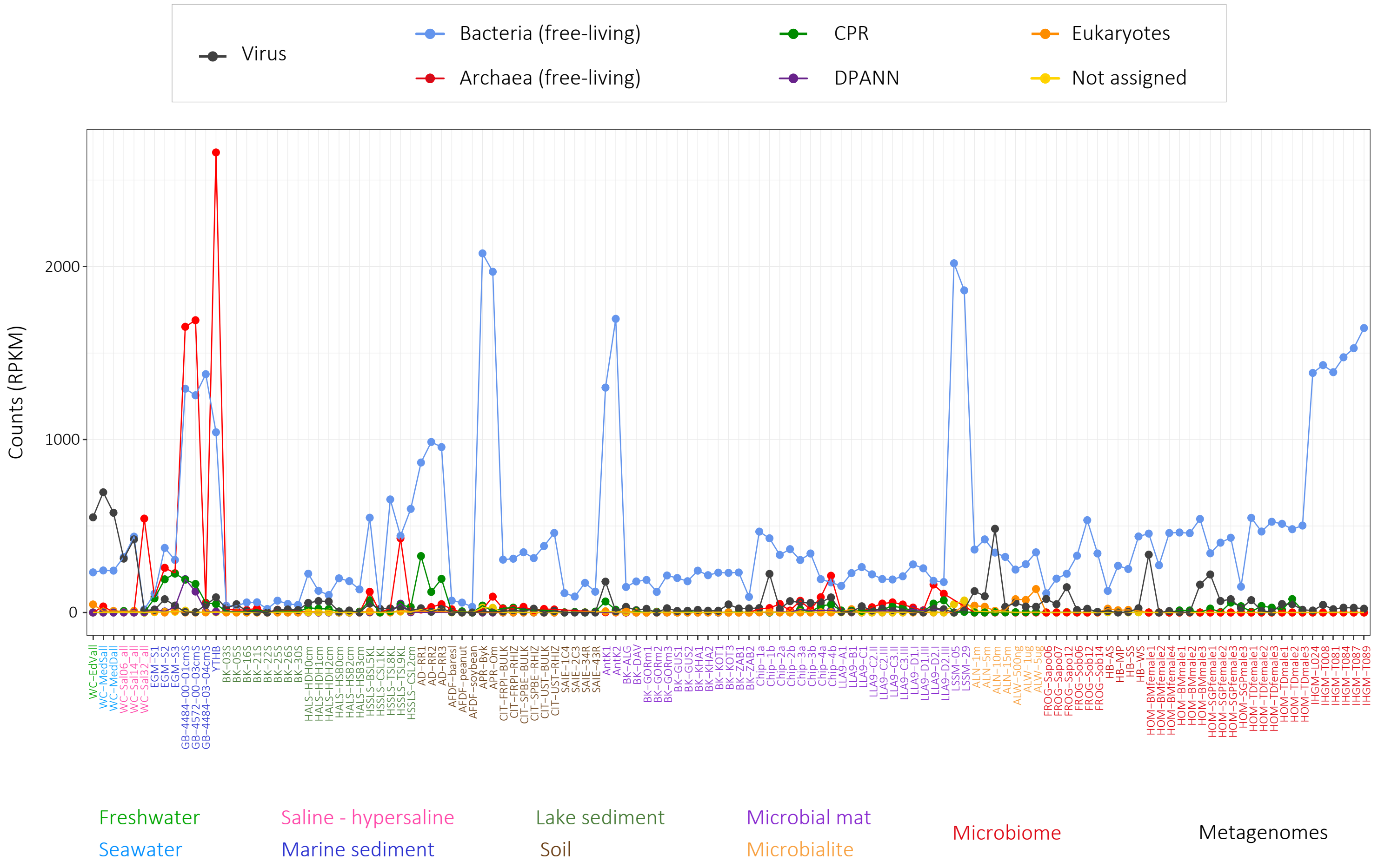
Extended Data Fig. 5. Normalized counts of viral DNA polymerases in metagenomes from different planktonic fractions in aquatic ecosystems along a salinity gradient. a, normalized numbers of virus-specific DNA polymerases identified using Pfam or VOG databases. **b**, total number of VOG-identified DNA polymerases in metagenomes of aquatic ecosystems. Only virus-specific polymerases are used for VMR estimates. Sample names are as in Extended Data Figs. 1 and 3. RPKM, reads per kilo base per million mapped metagenome reads.



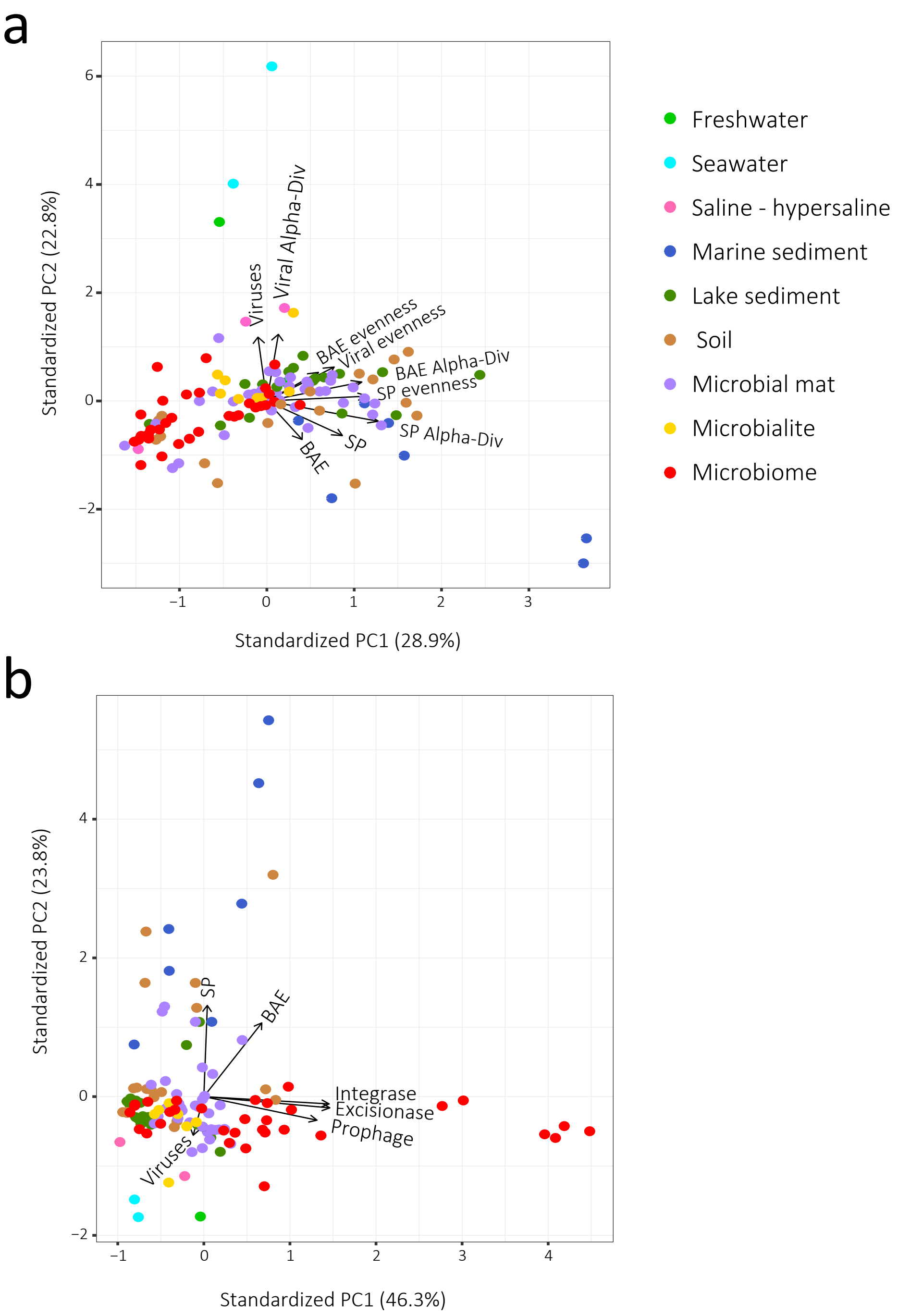
Extended Data Fig. 6. Normalized counts of viral and cellular genes in aquatic ecosystems following a salinity gradient. RPKM, reads per kilo base per million mapped metagenome reads. Sample names are as in Extended Data Figs. 1 and 3. Lines connecting points are shown to ease visualization.



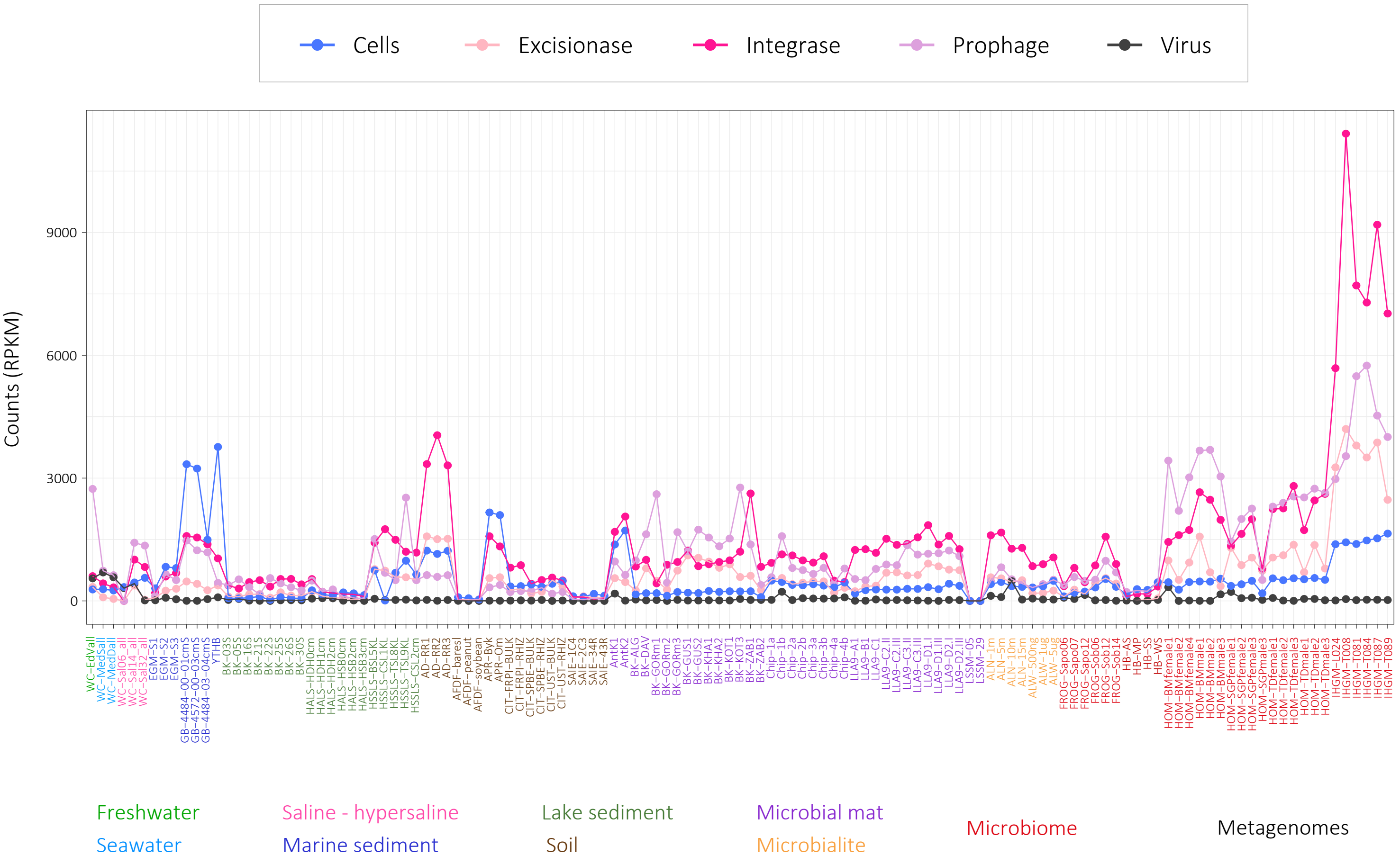
Extended Data Fig. 7. Absolute viral counts (DNApols) and frequency of viral classes identified in the studied metagenomes. The normalized viral counts per metagenome are indicated by a light blue line. White bars correspond to samples where DNApols were not detected. Metagenome samples are coloured according to ecosystem type.



Extended Data Fig. 8. Normalized metagenome counts of viruses and unicellular organisms across ecosystem types. Viruses and cells were counted using DNApols and USCGs as proxies (see text). All the metagenomes were analysed using the same pipeline from raw reads. RPKM, reads per kilo base per million mapped metagenome reads. Lines connecting points are shown to ease visualization.



Extended Data Fig. 9. Principal component analyses of viruses, symbiont/parasite and free-living cells, diversity indices and lysogeny-related genes. a, PCA of different metagenomes as a function of viral, symbiont/parasite (SP) and free-living (BAE) cells and their alpha-diversity and evenness (diversity indices are detailed in Supplementary Fig.5). **b**, PCA of different metagenomes maximizing the variance explained by viruses (DNApols), lysogeny-related genes (integrases, excisionases and prophage-related; see Supplementary Table 8), symbiont/parasite cells (SP, DPANN archaea + CPR bacteria) and free-living cells (BAE, other bacteria, archaea and eukaryotes).



Extended Data Fig. 10. Normalised metagenome counts of lysogeny-related genes, viruses and total cells across ecosystem types. Excisionases, integrases and prophage-related genes are expressed in RPKMs. Viruses and cells were counted using DNApols and USCGs. All the metagenomes were analysed using the same pipeline from raw reads. RPKM, reads per kilo base per million mapped metagenome reads. Lines connecting points are shown to ease visualization.