Ecological stochasticity and phage induction diversify bacterioplankton communities at the microscale

Rachel E. Szabo[a,b], Sammy Pontrelli[c], Jacopo Grilli[d], Julia A. Schwartzman[b], Shaul Pollak[b], Uwe Sauer[c], Otto X. Cordero[b,*]

[a]Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[b]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[c]Institute of Molecular Systems Biology, ETH Zürich, 8093 Zürich, Switzerland.
[d]Quantitative Life Sciences, The Abdus Salam International Centre for Theoretical Physics, 34151 Trieste, Italy.

[*]To whom correspondence may be addressed. Email: ottox@mit.edu

Author Contributions: RE Szabo designed and performed experiments; analyzed and interpreted data; and wrote the manuscript. S Pontrelli performed metabolomics and the related analysis. J Grilli developed the population dynamics model and interpreted data. JA Schwartzman designed and performed the multi-particle incubation experiment and analyzed particle images. S Pollak developed the analysis pipeline for classifying MAGs into ecological roles. OX Cordero designed experiments, interpreted data, and wrote the manuscript. All authors edited and reviewed the manuscript.

Competing Interest Statement: The authors declare no competing interest.

This PDF file includes:

Main Text
Figures 1 to 5
Main References
Supplementary Text
Figures S1 to S19
Legends for Tables S1 to S7
Supplementary References

Separate supplementary materials for this manuscript include:

Tables S1 to S7

Abstract

In many natural environments, microorganisms self-assemble around heterogeneously distributed resource patches. The growth and collapse of populations on resource patches can unfold within spatial ranges of a few hundred micrometers or less, making such microscale ecosystems hotspots of biological interactions and nutrient fluxes. Despite the potential importance of patch-level dynamics for the large-scale evolution and function of microbial communities, we have not yet been able to delineate the ecological processes that control natural populations at the microscale. Here, we addressed this challenge in the context of microbially-mediated degradation of particulate organic matter by characterizing the natural marine communities that assembled on over one thousand individual microscale chitin particles. Through shotgun metagenomics, we found significant variation in microscale community composition despite the similarity in initial species pools across replicates. Strikingly, a subset of particles was highly populated by rare chitin-degrading strains; we hypothesized that their conditional success reflected the impact of stochastic colonization and growth on community assembly. In contrast to the conserved functional structures that emerge in ecosystems at larger scales, this taxonomic variability translated to a wide range of predicted chitinolytic abilities and growth returns at the level of individual particles. We found that predation by temperate bacteriophages, especially of degrader strains, was a significant contributor to the variability in the bacterial compositions and yields observed across communities. Our study suggests that initial stochasticity in assembly states at the microscale, amplified through biotic interactions, may have significant consequences for the diversity and functionality of microbial communities at larger scales.

Significance Statement

The biogeochemical consequences of the degradation of particulate organic matter by microorganisms represent the cumulative effect of microbial activity on individual microscale

resource patches. The ecological processes controlling community dynamics in these highly localized microenvironments remain poorly understood. Here, we find that complex marine communities growing on microscale resource particles diverge both taxonomically and functionally despite assembling under identical abiotic conditions from a common species pool. We show that this variability stems from bacteriophage predation and history-dependent factors in community assembly, which create stochastic dynamics that are spatially structured at the microscale. This microscale stochasticity may have significant consequences for the coexistence, evolution, and function of diverse bacterial and viral populations in the global ocean.

Main Text

Introduction

A central challenge in microbial ecology is to connect the microscale world experienced by microbial cells to observations of large-scale community functions (1, 2). In many environments – ranging from soils (3) and sediments (4) to bioreactors (5) and hosts (6) – microbes live not in homogeneous, well-mixed cultures, but rather in diverse, spatially-structured assemblages, attached to surfaces and other cells in nutrient-dense patches on the order of 100μm in size. Patches often exist in otherwise nutrient-limiting environments, creating hotspots of ecological interactions and nutrient fluxes (7, 8).

A well-known example of micron-scale ecological hotspots is marine particulate organic matter (POM), which is degraded by complex communities of bacteria, archaea, viruses, and eukaryotes (9) with global biogeochemical consequences (10) (Fig. 1a). These interacting community members can be broadly classified as primary degraders (that produce extracellular enzymes to hydrolyze particle biopolymers), exploiters and scavengers (that are facilitated by primary degraders) (11–14), and predators (such as bacteriophages (15) and grazers (16)). Although these assemblages are often ephemeral, with organisms migrating through seawater

from patch to patch, cells can undergo multiple generations of growth while residing on a single patch of nutrient-rich POM. This implies that the evolution and ecological functions of POM-associated microbes are heavily influenced by their dynamics and interactions on microscale particles. However, little is known about the processes governing community assembly at these scales.

A major obstacle to understanding the factors that control populations at the patch level is the difficulty of characterizing natural microscale communities with high replication. Microbial communities are usually sampled at spatial scales orders of magnitude larger than those relevant for microbial life (2), which homogenizes their inherent patchiness and results in inconsistent inferences about ecological interactions (1, 17). Recent technological advances now permit the sequencing of only thousands of cells (18, 19), presenting an opportunity to systematically characterize microbial populations in units more closely approximating in scale the ecological contexts experienced by microbes.

Here, we leveraged high-replicate sequencing of individual microscale communities to evaluate the outcomes of assembly processes without the confounding effects of standard sampling procedures. We employed a hybrid natural-laboratory approach that paired the complexity of environmental microbial species pools with the controllability of synthetic microparticles as discrete resource patches (11, 12). We immersed 1222 individual hydrogel particles (85.0±24.0 μm in diameter) made of chitin – a highly abundant biopolymer in marine POM (9) – in samples of seawater containing microbes in their native states, which were then enriched on particle surfaces. By incubating single particles separately under identical abiotic conditions, each one became a microenvironment harboring a replicate community assembled from initially similar species pools. We performed a comparative analysis across these microscale ecosystems to investigate the natural variability in community composition and function among particles and to identify biological processes that contribute to particle-level variability.

Results

Bacterial community composition varies significantly across individual particles.

To quantify the variation in community states across replicate microscale ecosystems, we separately incubated single chitin particles in coastal seawater sampled from a common reservoir (Fig. 1b; Methods). Assembly outcomes were assessed by removing particles from the seawater at 13 time points over the course of 167 hours, a duration that aligns with previous measurements (12, 20) of particle lifetimes (Methods). Shotgun metagenomic sequencing of individual particle-attached communities was used to construct metagenome-assembled genomes (MAGs), which were annotated to infer strains' potential ecological roles in a chitin-degrading community as primary degraders, chitooligosaccharide exploiters, or metabolic byproduct scavengers (Fig. 1b; Table S1; Methods). These MAGs served as the references for characterizing the community composition of each particle.

We found a remarkable degree of compositional variability across individual particle communities at the end of the time course (n = 149, after 154-167 hours of incubation). The distributions of taxon (MAG) relative abundances across these late-stage particles spanned more than three orders of magnitude (Fig. 2; Fig. S1a) and were approximately lognormal with a skew towards high frequencies (Fig. S2). As a result, the community states observed at the single-particle level diverged so significantly that the relative success of taxa across particles was poorly explained by their average abundances (Fig. S1b; SI Methods). To assess whether non-ecological factors, such as sampling bias in initial species pools, could have contributed to this compositional divergence, we compared the variability in communities across late-stage particles to that across unincubated aliquots of the seawater used as the inoculum (SI Methods). Inter-sample variation was significantly higher across particles than across seawater samples (Mann-Whitney U test on Aitchison distances: $p = 1.3\times10^{-13}$; Fig. S3), indicating that the observed variability stemmed more from the community assembly process than from differences across inocula. Because other

technical sources of noise (Methods) also did not significantly impact the measured particle compositions (Fig. S4) and all particles were chemically identical, we concluded that the variation in taxon relative abundances across particles was due to biological and ecological factors that amplified stochasticity in the initial assembly states of these communities.

The skew towards high frequencies in the relative abundance distributions implied that taxa that were rare on average became dominant on a small number of particles (Fig. 2). As a result, those particles harbored low-complexity communities (Fig. S5a-b) that diverged highly from the average particle taxonomic composition (Fig. S5c). We termed the species that displayed this phenomenon "jackpot taxa" for their simultaneous local success and global rarity (Methods). The strains in this phylogenetically broad group of organisms included members of the *Enterobacterales*, *Cytophagales*, *Pseudomonadales*, *Flavobacteriales*, *Rhodobacterales*, *Fibrobacterales*, and *Chitinophagales* orders (Fig. S1, Table S1) and were mostly (87.9%) classified as chitin degraders. Jackpot taxa were more prevalent across late-stage particles than other taxa that were equally rare across inocula (Mann-Whitney *U* test: $p = 7.1\times10^{-3}$; Fig. S5d), indicating that the probability of their success on particles, while influenced by their scarcity in seawater, was also determined by ecological factors during community assembly. Notably, while taxon-specific interactions did not explain the abundance patterns observed across particles (Fig. S6; SI Methods), the most variable strains were likely to be degraders enriched in genes encoding chitinases (Fig. 2; Fig. S7; coefficient of variation vs. chitinase copy number, Spearman's $\rho = 0.44$, $p = 8.5\times10^{-7}$). These observations indicated that the conditional success of specialized degraders from a diverse initial species pool contributed to the differentiation of the many rare community states found at the single-particle level.

Taxonomic variability translates to divergent community-level productivity.

In contrast to the functionally similar gene content profiles predicted when microbial ecosystems are characterized at the macroscale (21, 22), we found that the communities formed on particles in individual microscale ecosystems were highly functionally divergent (Fig. 3a; Fig. S8). By the end of the time course, most particles (63.8%) – and especially particles dominated by jackpot taxa

– harbored majority-degrader communities (Fig. 3a), highlighting the importance of degraders for establishing and maintaining chitin-associated communities. However, the percentage of putative degraders on each late-stage particle was as low as 13.1% and as high as 97.3%, indicating that chitin degrading communities did not self-assemble to "optimized" or conserved ratios of ecological roles after a fixed incubation period. Read mapping to chitinase protein sequences rather than MAGs supported our interpretation that variability in the estimated proportion of degraders was not due to the use of MAGs as reference genomes (Fig. S9; SI Methods). We hypothesized that this extensive variability in community composition, primed by stochasticity in assembly processes, could have had significant consequences for overall community function.

Consistent with this hypothesis, individual particles sustained highly variable particle-attached biomass levels that were correlated with their community compositions (Fig. S10). The number of bacterial cells in each late-stage community, estimated using qPCR of the 16S rRNA region (Methods), ranged from approximately 1,000 to nearly 200,000 cells (Fig. 3b) and was strongly correlated with the overall frequency of degraders (Spearman's ρ = 0.45, p = 1.6×10^{-8}). Accordingly, particles that displayed the jackpot phenomenon had significantly higher cell counts (Fig. 3b; Mann-Whitney *U* test: p = 2.3×10^{-7}), revealing that jackpot taxa were dominant not only in terms of relative abundances but also absolute abundances. The distribution of cells per particle was approximately lognormal with a skew towards low cell numbers, indicating that some particles were highly productive while others harbored small populations even by the end of the incubation, as corroborated by visualizing particle-attached cells using a DNA stain (Fig. 3c; Methods). Importantly, the initial colonization of single particles incubated together in the same volume of seawater, rather than individually, resulted in particle-associated cell biomass that also spanned several orders of magnitude (Fig. S11; SI Methods). This variability in initial particle colonization was observed across a range of particle densities 15-140 times more concentrated than the conditions of the individual particle incubations, indicating that phenomena such as jackpot colonization are not specific to the environmental regime established in our separate microscale ecosystems. Collectively, these results suggested that a strain's growth was highly influenced by

its assembly context, raising the question of which biological or ecological factors could explain the large variance in species compositions and consequent yields across replicate particles.

Predation by bacteriophages contributes to variability in community composition and yield. Our observation that most (63.7%) of our MAGs contained sequences homologous to those of bacteriophages led us to investigate whether these entities impacted the abundances of bacteria on single particles. Bacteriophages (or phages, i.e. viruses that infect bacteria) are ubiquitous and abundant in marine ecosystems, making predation by phages one of the primary forms of top-down control of bacterial populations (23). High viral densities have been measured on marine particles relative to ambient seawater (15), but it is unknown to what extent this represents passive adsorption as opposed to active proliferation with impacts on bacterial growth in a natural, particle-associated context. Therefore, we sought to identify populations of actively replicating phages within the single particle communities to determine if heterogeneous phage predation could explain the variability in community composition and yield.

To detect replicating phages, we first classified contigs in our metagenomic dataset as phage-derived or bacteria-derived using tools (24, 25) that annotate phages from mixed metagenomes (Methods). We reasoned that contigs classified as phage-derived, especially those belonging to the genomes of temperate phages, were likely to be binned into the MAGs of their bacterial hosts. Phage *k*-mer signatures tend to be more similar to those of their specific hosts than to those of random bacteria (25, 26), and phages in a lysogenic cycle will have the same sequencing read coverage patterns as their hosts across samples. Therefore, phages that were lysogenic in most single-particle communities would tend to be binned with their hosts and have similar coverage levels, reflected in an inferred virus-to-microbial cell ratio (VMR) close to 1 (Fig. 4a, top left). In contrast, phages in a productive cycle (lytic or chronic) would have higher coverage than their hosts because of the multiple virion copies produced per bacterial cell (27, 28) (Fig. 4a, top right). Therefore, we considered a phage-derived contig to be productive in a sample if it was one of the most highly covered elements of its MAG (Methods).

Through this pipeline, we identified 256 phage contigs with coverage patterns consistent with lysogenic infections in all samples and 263 phage contigs with coverage patterns consistent with productive infections in a subset of samples (Table S2). Because our approach relied on comparisons between co-binned phages and MAGs, phages that exclusively employ a lytic cycle were unlikely to be detected. The VMRs of three representative examples of lysogenic and productive phage contigs are shown for each particle in Figure 4b. Comparing the coverage patterns of phage- and bacteria-derived contigs provided evidence that variable phage coverage was not due to sequencing noise, lending confidence to our estimates of VMRs for specific phages (Fig. S12; SI Methods). Using the VMRs of individual productive phages, we calculated the total productive VMR per particle as a measure of overall phage replication in each community (Methods).

The total productive VMRs of particle-associated communities sharply increased during the early stages of particle incubation in seawater (Fig. 4c), consistent with the phenomenon of rapid bacterial growth and high host densities driving the lysogeny-lysis switch in some phages (29–33). The mean productive VMR was lowest for the initial seawater inocula and rose sharply until the middle of the incubation period (59 hours), suggesting that phages became induced as their particle-associated hosts began to grow. Concomitant with this increase in productive VMRs, we observed the accumulation of metabolites in the seawater surrounding each particle until 59 hours of incubation, followed by a decrease in metabolite concentrations (Fig. 4d; Fig. S13; Table S3; Methods). These observations could be explained by metabolite release upon the initiation of bacterial growth (34) or lysis by phages (35) and by subsequent metabolite consumption by the remaining viable bacteria (36). The coinciding timescales of metabolite liberation and rising VMRs are consistent with our hypothesis that a particle-associated lifestyle among bacteria promoted phage proliferation; therefore, we sought to assess the impact of variable phage induction on each community's composition and consequent yield.

There was a striking negative relationship between cell counts and productive VMRs on late-stage particles (Fig. 4e main, red data: Spearman's ρ = -0.56, p = 3.3×10^{-13}), suggesting that phage predation impacted bacterial growth success on particles upon induction. The degrader

220 populations contributed the most to this signal, indicating that strains among this trophic level may
221 have been especially prone to phage activation (Fig. 4e inset; Fig. S14). Importantly, jackpot
222 degrader taxa had lower productive VMRs than non-jackpot degraders (Fig. S15a; Mann-Whitney
223 *U* test: $p = 1.3\times10^{-49}$). This translated to jackpot particles having significantly lower productive VMRs
224 than non-jackpot particles (Mann-Whitney *U* test: $p = 4.3\times10^{-8}$), even controlling for differences in
225 cell counts between these groups of communities (ANCOVA: $F(1,139) = 16.92$, $p = 4.1\times10^{-4}$, partial
226 $\eta^2 = 0.09$; Fig. S15b). Therefore, jackpot degraders may have been locally successful on a minority
227 of particles in part because they experienced less predation, supporting the hypothesis that top-
228 down population control by phages contributed to the large variability in the bacterial compositions
229 and thus yields observed across communities.

230 While cell counts were significantly correlated with both phage abundances and community
231 compositions, these features explained, respectively, 23% (Fig. 4e) and 34% (Fig. S10) of the
232 observed variation in yields, indicating that other factors also contributed to variable growth returns.
233 Therefore, we sought a more general framework in which to understand the key quantitative
234 features of the data – namely, the lognormal-like distributions of relative taxon abundances (with
235 right skews consistent with jackpot taxa) and of absolute cell abundances (with a left skew
236 corresponding to low-biomass communities). Incorporating (i) stochastic cell arrival on particles, (ii)
237 degraders as population founders, and (iii) noisy growth rates into a simple mathematical model of
238 community development on single particles was sufficient to reproduce these features (Figs. S16-
239 S19; SI Text). Taken together with our experimental data, this model indicates that the biological
240 processes which contribute to the stochasticity of particle colonization and growth rates – and
241 especially those processes that affect degraders – will result in variable growth returns for strains
242 across particles.
243
244
245 **Discussion**
246

While there is an abundance of evidence showing that the marine environment as experienced by microbial cells is biologically, chemically, and physically heterogeneous (7), characterizing the ecological processes controlling community assembly and development at these scales remains a fundamental challenge, particularly *in situ*. Our study takes a step toward addressing this problem using a hybrid natural-laboratory experiment that monitored the assembly outcomes of complex marine communities across hundreds of individual chitin-based resource particles. In accordance with prior work demonstrating small-scale heterogeneity on aquatic resource particles (37), we found that bacterial compositions and absolute abundances varied to such an extent across replicate particles that key community features – namely, species composition and functional potential – were not conserved. Our results contrast with those of previous studies (11, 13) that describe rapid ecological successions within particle systems that are reproducible across batches. Despite this apparent reproducibility, biomass distributions in our single-particle and our multi-particle incubations suggest that particle colonization is likely heterogeneous in both systems. Thus, the reproducible dynamics previously observed in particle systems could reflect the increasingly homogenizing effect of exchange between particles over time or the mean of a process that is highly variable on the individual-particle scale. Future work is required to determine the effect of dispersal and "cross-colonization" on the dynamics of particle systems.

Stochastic factors are anticipated to strongly influence community assembly for populations that are localized to small scales (38), such as in the microscale ecosystems on resource particles (12, 20). The first step in community assembly – the arrival of cells to a particle – is an intrinsically random process dependent on encounter probabilities. Our population dynamics model demonstrated that historical contingencies (created by stochastic arrival times and the growth dependency of non-degraders on degraders) magnified through noisy growth rates were sufficient to reproduce the distributions of bacterial abundances observed across individual particles. Because this chitin microparticle ecosystem is subject to conditions that have been shown to promote strong priority effects (e.g. a large regional species pool, rapid local growth dynamics, high resource overlap, and a dependence of late-arriving organisms on early-arriving ones) (38),

we hypothesize that biotic factors amplified this initial stochasticity in each assembly context and influenced subsequent community development.

One key biological contributor to noisy growth returns may have been variable predation by temperate bacteriophages. Phages became increasingly and differentially activated during community development on particles, with elevated virus-to-microbial cell ratios (VMRs) in low-biomass communities implicating phage-mediated lysis as one factor explaining the biomass variability on late-stage particles. These results align with those of previous studies documenting extensive variation in VMRs at small spatial scales (23) and an inverse relationship between VMRs and cell densities in marine environments (39). Because phage induction was significantly less associated with jackpot degrader strains, we hypothesize that the jackpot phenomenon – characterized by globally rare yet locally productive degraders – was partially a reflection of lower levels of phage-driven population collapse in those community contexts. Therefore, top-down control by phages may link the highly variable community compositions and yields observed among particles.

A synthesis of our mathematical model with our observations of bacterial and phage abundances suggests a conceptual framework for key processes promoting variability in microscale community composition and function (Fig. 5). We posit that stochastic arrival on particles diversifies initial assembly states; that the timescale and magnitude of degrader colonization determine the extent to which scavengers and exploiters are supported; and that phage induction and subsequent host lysis, primarily among degraders, contribute to noisy growth returns. Therefore, in this conceptual framework, the high-biomass jackpot particles are those in which degraders arrive early and resist phage induction, leading to high relative and absolute degrader abundances (Fig. 5, top). By contrast, low-biomass particles are those in which degraders are not able to proliferate, either because phage induction leads to their population collapse (Fig. 5, middle) or because they become established on a particle relatively late (Fig. 5, bottom).

In addition to the growing body of evidence that marine aggregates can stimulate the production of virulent phages (i.e. phages that exclusively engage in lytic cycles) (40), our study suggests that resource particles may be replication hotspots also for temperate phages (i.e. those

303 that conditionally employ both lytic and lysogenic cycles). In marine environments, lysogeny is
304 promoted under conditions that limit bacterial growth while the lytic cycle is favored during periods
305 of high bacterial activity (29, 41, 42), indicating that rapid host growth and abundance can regulate
306 the lysogeny-lysis switch in some temperate phages (30–33). Therefore, in a patchy nutrient
307 landscape, temperate phages may employ lysogeny as a survival strategy when their bacterial
308 hosts are at low densities and are foraging for nutrients, hitchhiking with their hosts onto resource
309 particles. Robust bacterial growth on particles may induce prophages at a time when abundant
310 host resources can be co-opted and many susceptible cells are nearby, resulting in lytic
311 suppression of the bacterial population and the release of virions into the surrounding seawater.
312 Factors such as the variable presence of prophages in the flexible genomes of strains growing on
313 different particles (43), the co-occurrence of bacterial competitors that trigger induction (44, 45),
314 and phenotypic heterogeneity resulting in differential induction (46, 47) may all contribute to the
315 varying levels of phage activation observed on individual particles in our microscale ecosystems.
316 Further research is required on the mechanisms underlying prophage induction in complex
317 communities in order to understand how lysogeny and lysis on particle hotspots shape the
318 dynamics of marine microbial communities.

319 Our observations of wild marine communities, though made in a laboratory setting, may
320 provide insights on the ecosystem-level consequences of microscale stochastic assembly
321 dynamics. First, the stochasticity in bacterial growth, amplified through spatial structuring at the
322 microscale, may promote the maintenance of a diverse regional species pool. This is because the
323 variability in growth returns can effectively offset differences in relative fitness between competing
324 strains or species (48). Second, the variability in microscale community states could be reflected
325 in larger-scale biogeochemical patterns since the cumulative process of POM degradation can be
326 approximated as the sum of degradation events on individual particles. We found that late-stage
327 communities did not converge to a fixed proportion of chitin degraders or to a fixed amount of
328 biomass per particle; both measures are positively correlated with the rate of particle degradation
329 (12), suggesting that historical contingencies in community assembly promote functional
330 divergence (38, 49). These results contrast with those of previous studies on the replicability of

331 microbial community assembly at the functional level (21, 22) likely because of the homogenizing
332 effect of macroscale sampling. Intriguingly, the lognormal-like distribution of biomass on individual
333 particles aligns with observations and predictions of lognormally-distributed global marine organic
334 matter export and remineralization rates; these distributions may repeatedly emerge as a reflection
335 of the multiplicative effects of stochastic variables in ecological settings (50–52). Although our
336 experimental system significantly simplified the process of POM degradation in the ocean, our
337 approach provides a quantitative link between the microscale and larger-scale processes,
338 highlighting the importance of considering local variability when investigating the mechanisms
339 behind microbial community development in a spatially structured environment.
340
341
342 **Materials and Methods**
343
344 Abridged Methods are provided below; details and additional information are provided in SI
345 Methods.
346
347 **Seawater collection and individual chitin particle incubation.** Nearshore coastal seawater was
348 collected from Nahant, MA; filtered (63μm) to remove large particulate matter; gently concentrated
349 via centrifugation at 4000 × *g* for 5 minutes; and aliquoted for incubations and sequencing. Chitin
350 magnetic particles (New England Biolabs, #E8036L) were washed in sterile artificial seawater
351 (Sigma-Aldrich, #S9883) and individually selected beneath a dissecting microscope in a laminar
352 flow hood. Single chitin particles (85.0±24.0 μm in diameter) were transferred to sterile 96-well
353 plates (Thermo Fisher, #AB0600L), with one chitin particle per well. Plates were inoculated
354 consecutively with 175μL of filtered, centrifuged seawater per well; sealed (VWR, #89092-056);
355 and rotated end-over-end (7.5rpm) at room temperature. The particles in an entire plate were
356 harvested at each time point (after 12, 22.75, 34.5, 46, 59, 69, 82, 92, 103, 116.75, 113, 153.5, and
357 166.5 hours of incubation) by inspection and pipetting under a dissecting microscope in a laminar
358 flow hood. Each particle was transferred into sterile 96-well plates (Thermo Fisher, #AB0600L)

359 containing TE buffer and stored at -20°C. The seawater surrounding each harvested particle was
360 also saved in 96-well plates and stored at -20°C.
361
362 **Mock communities and negative controls.** To quantify the technical error associated with
363 creating metagenomic libraries from low DNA inputs, mock communities were simulated by
364 combining the DNA of two strains previously isolated from a chitin particle enrichment (11).
365 Libraries from three technical replicates of mock communities totaling 50pg or 5pg of DNA (SI
366 Methods), as well as from six negative controls (containing only nuclease-free water), were
367 prepared and analyzed with the same protocols used for individual chitin particle-attached
368 communities.
369
370 **DNA extraction and metagenomic sequencing.** DNA extractions were performed for twelve
371 175µL-volume aliquots of the initial, unincubated seawater and for particles harvested after 34.5,
372 59, 103, 116.75, 113, 153.5, and 166.5 hours of incubation (see Table S5 for sample metadata).
373 DNA was extracted from all samples with the Agencourt DNAdvance Genomic DNA Isolation Kit
374 (Beckman Coulter; modifications noted in SI Methods). Metagenomic libraries were prepared with
375 the Nextera XT DNA Library Prep Kit and index primers (Illumina) using the protocol developed by
376 Rinke *et al.* (18) for low DNA inputs (SI Methods). Libraries were quantified on an Agilent 4200
377 TapeStation system with High Sensitivity D5000 ScreenTapes (Agilent Technologies) and pooled
378 by time point in equimolar amounts. Sequencing was performed on an Illumina HiSeq 2500
379 machine (250bp paired-end reads) at the Whitehead Institute for Biomedical Research (Cambridge,
380 MA).
381
382 **Metagenome-assembled genome (MAG) generation, taxonomic assignment, and role**
383 **classification.** Raw sequencing reads were quality trimmed with Trimmomatic v0.36 (53). Reads
384 mapping to the PhiX and human genomes were filtered out using BBDuk v38.16 (54) and BBMap
385 v38.16, respectively (SI Methods). Trimmed, filtered reads that were error-corrected using
386 BayesHammer (55) were pooled within each time point and co-assembled using MEGAHIT v1.2.9

(56). Bins were generated with MaxBin v2.2.7 (57) and CONCOCT v1.1.0 (58); consolidated and filtered using DAS Tool v1.1.1 (59); and evaluated for completeness and contamination using CheckM v1.1.2 (60). The resulting 251 bins were used as reference MAGs (≥50% complete, ≤10% contaminated; median completeness 93.7%, median contamination 3.9%; Table S1). Highly similar MAGs obtained from separate co-assemblies were grouped into 132 clusters (SI Methods). MAG taxonomic classifications were made using GTDB-Tk v1.1.1 (61). MAGs were functionally annotated using a custom database of profile hidden Markov models (HMMs) of proteins involved in growth on chitin (SI Methods; Table S6). Ecological roles for MAGs (as degraders, chitooligosaccharide exploiters, or metabolic byproduct scavengers) were defined based on the gene content patterns observed for sequenced and phenotyped (14) strains previously isolated (11, 13) from particle enrichments (SI Methods).

Read mapping to MAGs for relative abundance estimation. Trimmed, filtered reads were mapped competitively against the MAGs generated from sequencing particle-attached communities, initial seawater samples, and negative controls. Read mapping was performed using the approach described in Leventhal *et al*. (62) (SI Methods). Reads that best mapped to predicted contaminant MAGs (SI Methods) were removed from consideration. MAG relative abundances were calculated for each sample by (1) tallying the hits to all MAGs in each MAG cluster; (2) normalizing the tally by the average genome length of all MAGs in each MAG cluster; and (3) dividing the normalized tallies for each MAG cluster by their sum for each sample. Therefore, for MAGs clustered together based on similarity, their relative abundances are represented in that of the entire MAG cluster to which they belong; this calculation circumvents the artificial underestimation of MAG relative abundances that would otherwise be obtained with a non-dereplicated reference set. The relative abundances of organisms occupying the three ecological roles (degrader, exploiter, scavenger) on each particle were calculated by summing the relative abundances of MAGs classified into each role.

Definitions of jackpot MAGs and jackpot particles. A jackpot score was calculated for each MAG cluster to quantitatively reflect the properties of rarity across most particles and dominance on a few particles (SI Methods) such that MAGs with high scores strongly displayed the jackpot phenomenon. Each particle's jackpot score was calculated as the weighted average of MAG jackpot scores (i.e. the sum of the relative abundance of each MAG cluster multiplied by its jackpot score). Particles with high jackpot scores and low Pielou's evenness were categorized as "jackpot particles" (SI Methods).

Bacteriophage analyses. Binned contigs were classified as phage-derived or bacteria-derived using VirSorter v1.0.3 with its RefSeqABVir database (24) and VirFinder v1.1 (25), two tools designed to detect phage sequences among mixed metagenomes (SI Methods). We used a read coverage-based approach to categorize phage-derived contigs as productive or lysogenic in particle-attached communities (Table S2; see SI Methods for analysis controls). Based on read mapping to MAGs, per-base coverage values for all binned contigs were computed with BEDTools v2.27.0 (63) and were used to calculate contig-wide average coverage values. For each MAG and for each sample, a phage-derived contig was considered to be productive if its coverage was greater than the coverage of the 95th percentile bacteria-derived contig in the same MAG; otherwise, it was considered to be lysogenic in that sample. The VMR of an individual phage contig in one sample is defined as the phage contig coverage divided by average coverage of the MAG with which it is binned (which was calculated using only the bacteria-derived contigs). Total VMRs – i.e. the total number of phage copies relative to the total number of bacterial MAG copies in an entire sample – were calculated separately for productive and lysogenic phage contigs. The total productive VMR for a sample was defined as:

$$\sum_{i}^{n}\left[\left(\frac{average\ coverage\ of\ productive\ phage\ contigs\ in\ MAG_i}{average\ MAG_i\ coverage}\right)\times\left(MAG_i\ relative\ abundance\right)\right]$$

$$=\frac{total\ \#\ phage\ copies\ (due\ to\ productive\ infections)}{total\ \#\ bacterial\ genome\ copies}$$

where *n* is the number of MAGs found in a sample. This calculation is equivalent to

$$\frac{\sum_i^n (average\ coverage\ of\ productive\ phage\ contigs\ in\ MAG_i)}{\sum_i^n (average\ MAG_i\ coverage)}$$

where *n* is the number of MAGs found in a sample. Total lysogenic VMRs were calculated using the same formula while considering only lysogenic-annotated contigs. VMRs for each ecological role (i.e. for the subpopulation in a community that belongs to one of the three roles of degrader, exploiter, or scavenger) were calculated using the same formula considering only the MAGs of each role and their associated phages.

Cell count estimation. Bacterial DNA extracted from individual particle-attached communities was quantified through qPCR of the 16S rRNA gene using the Femto Bacterial DNA Quantification Kit (Zymo Research), which has a lower limit of detection of 20fg. Two sets of standards and negative controls were included in each qPCR run. The number of bacterial cells for each particle was estimated from the absolute DNA amounts based on measurements indicating a mean of 2.5fg DNA per bacterial cell in seawater samples (64).

Imaging of chitin particle colonization. Subsets of chitin particles incubated individually in seawater were stained at time points by adding the DNA stain SYTO9 (Invitrogen, #S34854) at a final concentration of 500nM directly to the particle incubations. Particles were incubated in the dark at room temperature for 15 minutes before being mounted separately on microscope slides and imaged with a Zeiss epifluorescence microscope at 100X magnification.

Metabolomics. We performed untargeted metabolomics of the seawater that surrounded each harvested chitin particle and of the initial, unincubated seawater (SI Methods). We used a binary LC pump (Agilent Technologies) and an MPS2 Autosampler (Gerstel) coupled to an Agilent 6520

time-of-flight mass spectrometer (Agilent Technologies) operated in negative mode, at 2GHz, extended dynamic range, with an *m/z* (mass/charge) range of 50-1000. Ions (Table S3) were annotated against a curated library of metabolites present in marine microbes, based on the BioCyc database (65). For metabolites that exceeded the limit of detection (SI Methods), the intensities of each ion were normalized between 0 (the limit of detection) and 1 (the highest measured intensity of a given ion). Weighted ion intensities for each timepoint were calculated by taking the sum of all normalized intensities of ions in all samples for each timepoint.

Data Sharing Plans

All data will be made publicly available before publication. Sequencing data will be deposited to the National Center for Biotechnology Information (NCBI) as a BioProject, with raw reads uploaded to the Sequence Read Archive (SRA) and metagenome-assembled genomes (MAGs) uploaded to the Whole Genome Shotgun (WGS) database. All mass spectra files from the metabolomics will be accessible from MassIVE (ftp://MSV000087936@massive.ucsd.edu) before publication. MAG relative abundances for each sample and metadata for samples, MAGs, phages, and detected metabolites are provided as Supplementary Tables. All code and files used to generate figures will be made available at personal GitHub pages before publication.

Acknowledgments

We extend our gratitude to all past and present members of the Cordero lab for their support and critical feedback, as well as members of the Simons Collaboration on Principles of Microbial Ecosystems for stimulating discussions. In particular, we would like to thank: Manoshi S. Datta, for contributing to the genesis of this project; José T. Saavedra, for developing the DNA extraction and metagenomic library preparation pipeline for single particle-attached communities; Gabriel E.

Leventhal and Jakob Russel for bioinformatic mentorship and assistance; Matti Gralka, for feedback on analyses and this manuscript; Elise Ledieu, for quantifying chitin particle sizes; Anthony Gaca, for advice on metagenomic library preparation; Fatima Aysha Hussain, for feedback on the bacteriophage analysis; Emily Zakem, for insights on global particle remineralization rates; Akshit Goyal, for comments on this manuscript; and Sara Szabo and William Mandella, for assistance with seawater sampling. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. #174530. This project was supported by the Simons Collaboration: Principles of Microbial Ecosystems (PriME) award number 542395. S Pontrelli was supported by a grant from the Simons Foundation (ID608247) as part of PriME. S Pollak was supported by the EMBO ALTF Grant No. #800-2017.

Main References

1. O. X. Cordero, M. S. Datta, Microbial interactions and community assembly at microscales. *Curr Opin Microbiol* 31, 227–234 (2016).
2. D. R. Nemergut, *et al.*, Patterns and Processes of Microbial Community Assembly. *Microbiol Mol Biol R* 77, 342–356 (2013).
3. A. G. O'Donnell, I. M. Young, S. P. Rushton, M. D. Shirley, J. W. Crawford, Visualization, modelling and prediction in soil microbiology. *Nat Rev Microbiol* 5, 689–699 (2007).
4. S. E. McGlynn, G. L. Chadwick, C. P. Kempes, V. J. Orphan, Single cell activity reveals direct electron transfer in methanotrophic consortia. *Nature* 526, 531–535 (2015).
5. G. Gonzalez-Gil, C. Holliger, Aerobic granules: Microbial landscape and architecture, stages, and practical implications. *Appl Environ Microb* 80, 3433–3441 (2014).
6. J. L. M. Welch, B. J. Rossetti, C. W. Rieken, F. E. Dewhirst, G. G. Borisy, Biogeography of a human oral microbiome at the micron scale. *Proc National Acad Sci* 113, 791–800 (2016).
7. R. Stocker, Marine microbes see a sea of gradients. *Science* 338, 628–633 (2012).

8. L. M. Dann, *et al.*, Microbial micropatches within microbial hotspots. *Plos One* 13, 1–22 (2018).
9. H. Dang, C. R. Lovell, Microbial Surface Colonization and Biofilm Development in Marine Environments. *Microbiol Mol Biol R* 80, 91–138 (2016).
10. N. Jiao, *et al.*, Microbial production of recalcitrant dissolved organic matter: Long-term carbon storage in the global ocean. *Nat Rev Microbiol* 8, 593–599 (2010).
11. M. S. Datta, E. Sliwerska, J. Gore, M. F. Polz, O. X. Cordero, Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nat Commun* 7, 1–7 (2016).
12. T. N. Enke, G. E. Leventhal, M. Metzger, J. T. Saavedra, O. X. Cordero, Microscale ecology regulates particulate organic matter turnover in model marine microbial communities. *Nat Commun* 9, 2743 (2018).
13. T. N. Enke, *et al.*, Modular Assembly of Polysaccharide-Degrading Marine Microbial Communities. *Curr Biol* 29, 1528-1535.e6 (2019).
14. S. Pontrelli, *et al.*, Hierarchical control of microbial community assembly by specialists. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.06.22.449372 (accessed 24 September 2021).
15. M. G. Weinbauer, *et al.*, Viral ecology of organic and inorganic particles in aquatic systems: Avenues for further research. *Aquat Microb Ecol* 57, 321–341 (2009).
16. M. Simon, H. P. Grossart, B. Schweitzer, H. Ploug, Microbial ecology of organic aggregates in aquatic ecosystems. *Aquat Microb Ecol* 28, 175–211 (2002).
17. D. W. Armitage, S. E. Jones, How sample heterogeneity can obscure the signal of microbial interactions. *Isme J* 13, 2639–2646 (2019).
18. C. Rinke, *et al.*, Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *Peerj* 2016, 1–28 (2016).
19. R. U. Sheth, *et al.*, Spatial metagenomic characterization of microbial biogeography in the gut. *Nat Biotechnol* 37, 877–883 (2019).

549 20. M. H. Iversen, H. Ploug, Temperature effects on carbon-specific respiration rate and
550 sinking velocity of diatom aggregates – potential implications for deep ocean export
551 processes. *Biogeosciences* 10, 4073–4085 (2013).
552 21. S. Louca, *et al.*, High taxonomic variability despite stable functional structure across
553 microbial communities. *Nat Ecol Evol* 1, 1–12 (2016).
554 22. C. Huttenhower, *et al.*, Structure, function and diversity of the healthy human microbiome.
555 *Nature* 486, 207–214 (2012).
556 23. M. Breitbart, C. Bonnain, K. Malki, N. A. Sawaya, Phage puppet masters of the marine
557 microbial realm. *Nat Microbiol* 3, 754–766 (2018).
558 24. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: Mining viral signal from microbial
559 genomic data. *Peerj* 2015, 1–20 (2015).
560 25. J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, F. Sun, VirFinder: a novel k-mer based tool
561 for identifying viral sequences from assembled metagenomic data. *Microbiome* 5, 69
562 (2017).
563 26. N. A. Ahlgren, J. Ren, Y. Y. Lu, J. A. Fuhrman, F. Sun, Alignment-free d2∗ oligonucleotide
564 frequency dissimilarity measure improves prediction of hosts from metagenomically-
565 derived viral sequences. *Nucleic Acids Res* 45, 39–53 (2017).
566 27. K. Kieft, K. Anantharaman, Deciphering active prophages from metagenomes. bioRxiv
567 [Preprint] (2021). https://doi.org/10.1101/2021.01.29.428894 (accessed 24 September
568 2021).
569 28. R. F. von Boeselager, E. Pfeifer, J. Frunzke, Cytometry meets next-generation sequencing
570 – RNA-Seq of sorted subpopulations reveals regional replication and iron-triggered
571 prophage induction in Corynebacterium glutamicum. *Sci Rep-uk* 8, 1–13 (2018).
572 29. J. H. Paul, Prophages in marine bacteria: Dangerous molecular time bombs or the key to
573 survival in the seas? *Isme J* 2, 579–589 (2008).
574 30. J. R. Brum, B. L. Hurwitz, O. Schofield, H. W. Ducklow, M. B. Sullivan, Seasonal time
575 bombs: Dominant temperate viruses affect Southern Ocean microbial dynamics. *Isme J*
576 10, 437–449 (2016).

31. M. Touchon, A. Bernheim, E. P. C. Rocha, Genetic and life-history traits associated with the distribution of prophages in bacteria. *Isme J* 10, 2744–2754 (2016).
32. J. E. Silpe, B. L. Bassler, A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell* 176, 268-280.e13 (2019).
33. L. Laganenka, *et al.*, Quorum sensing and metabolic state of the host control lysogeny-lysis switch of bacteriophage T1. *Mbio* 10, 3–8 (2019).
34. B. E. Noriega-Ortega, *et al.*, Does the chemodiversity of bacterial exometabolomes sustain the chemodiversity of marine dissolved organic matter? *Front Microbiol* 10, 1–13 (2019).
35. N. Y. D. Ankrah, *et al.*, Phage infection of an environmentally relevant marine bacterium alters host metabolism and lysate composition. *Isme J* 8, 1089–1100 (2014).
36. S. Blasche, *et al.*, Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. *Nat Microbiol* 6, 196–208 (2021).
37. M. Bizic-Ionescu, D. Ionescu, H.-P. Grossart, Organic particles: heterogeneous hubs for microbial interactions in aquatic ecosystems. *Front Microbiol* Accepted, 1–15 (2018).
38. T. Fukami, Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and Priority Effects. *Annu Rev Ecol Evol Syst* 46, 1–23 (2015).
39. C. H. Wigington, *et al.*, Re-examination of the relationship between marine virus and microbial cell abundances. *Nat Microbiol* 1, 4–11 (2016).
40. L. Riemann, H. P. Grossart, Elevated lytic phage production as a consequence of particle colonization by a marine Flavobacterium (Cellulophaga sp.). *Microbial Ecol* 56, 505–512 (2008).
41. M. G. Weinbauer, I. Brettar, M. G. Höfle, Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. *Limnol Oceanogr* 48, 1457–1465 (2003).
42. J. P. Payet, C. A. Suttle, To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol Oceanogr* 58, 465–474 (2013).
43. B. C. M. Ramisetty, P. A. Sudhakari, Bacterial “grounded” prophages: Hotspots for genetic renovation and innovation. *Frontiers Genetics* 10, 1–17 (2019).

44. M. Jancheva, T. Böttcher, A Metabolite of Pseudomonas Triggers Prophage-Selective Lysogenic to Lytic Conversion in Staphylococcus aureus. *J Am Chem Soc* 143, 8344–8351 (2021).

45. J. E. Silpe, J. W. H. Wong, S. V. Owen, M. Baym, E. P. Balskus, The gut bacterial natural product colibactin triggers induction of latent viruses in diverse bacteria. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.05.24.445430 (accessed 24 September 2021).

46. J. J. Dennehy, I. N. Wang, Factors influencing lysis time stochasticity in bacteriophage. *BMC Microbiol* 11, 174 (2011).

47. L. Imamovic, E. Ballesté, A. Martínez-Castillo, C. Garćia-Aljaro, M. Muniesa, Heterogeneity in phage induction enables the survival of the lysogenic population. *Environ Microbiol* 18, 957–969 (2016).

48. A. Melbinger, M. Vergassola, The Impact of Environmental Fluctuations on Evolutionary Fitness Functions. *Sci Rep* 5, 1–11 (2015).

49. L. S. Bittleston, M. Gralka, G. E. Leventhal, I. Mizrahi, O. X. Cordero, Context-dependent dynamics lead to the assembly of functionally distinct microbial communities. *Nat Commun* 11, 1–10 (2020).

50. B. B. Cael, K. Bisson, C. L. Follett, Can Rates of Ocean Primary Production and Biological Carbon Export Be Related Through Their Probability Distributions? *Global Biogeochem Cy* 32, 954–970 (2018).

51. E. J. Zakem, B. B. Cael, N. M. Levine, A unified theory for organic matter accumulation. *Proc National Acad Sci* 118, e2016896118 (2021).

52. E. Limpert, W. A. Stahel, M. Abbt, Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* 51, 341–352 (2001).

53. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).

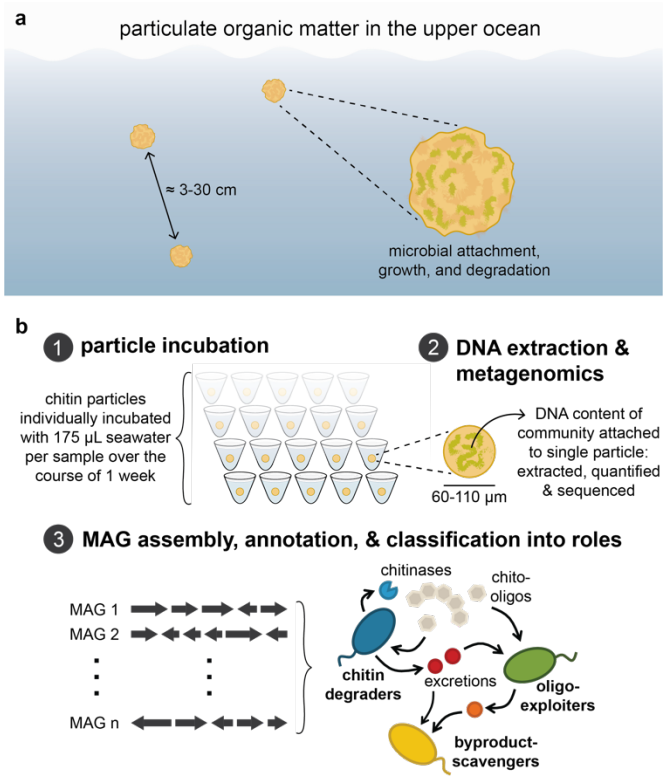
54. B. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner (2014).

55. S. I. Nikolenko, A. I. Korobeynikov, M. A. Alekseyev, BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *Bmc Genomics* 14, S7 (2013).

633 56. D. Li, *et al.*, MEGAHIT v1.0: A fast and scalable metagenome assembler driven by
634 advanced methodologies and community practices. *Methods* 102, 3–11 (2016).
635 57. Y. W. Wu, B. A. Simmons, S. W. Singer, MaxBin 2.0: An automated binning algorithm to
636 recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607 (2016).
637 58. J. Alneberg, *et al.*, Binning metagenomic contigs by coverage and composition. *Nat*
638 *Methods* 11, 1144–1146 (2014).
639 59. C. M. K. Sieber, *et al.*, Recovery of genomes from metagenomes via a dereplication,
640 aggregation and scoring strategy. *Nat Microbiol* 3, 836–843 (2018).
641 60. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM:
642 Assessing the quality of microbial genomes recovered from isolates, single cells, and
643 metagenomes. *Genome Res* 25, 1043–1055 (2015).
644 61. P. A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: A toolkit to classify
645 genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927 (2020).
646 62. G. E. Leventhal, *et al.*, Strain-level diversity drives alternative community types in
647 millimetre-scale granular biofilms. *Nat Microbiol* 3, 1295–1303 (2018).
648 63. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic
649 features. *Bioinformatics* 26, 841–842 (2010).
650 64. D. K. Button, B. R. Robertson, Determination of DNA Content of Aquatic Bacteria by Flow
651 Cytometry. *Appl Environ Microb* 67, 1636–1645 (2001).
652 65. P. D. Karp, *et al.*, The BioCyc collection of microbial genomes and metabolic pathways.
653 *Brief Bioinform* 20, 1085–1093 (2017).

654 **Figures**

655



656

657 **Figure 1. Modeling particulate organic matter degradation with a laboratory system of**
658 **enriching of marine microbes on chitin particles. (a)** Microscale marine particles are spatially-
659 separated nutrient-rich habitats dynamically populated and degraded by complex communities of
660 heterotrophic bacteria. The interparticle distance range is estimated from data reported in Simon
661 *et al.* (16). **(b)** Schematic depicting experimental design and analysis. Microscale chitin particles
662 were individually incubated in seawater, and the DNA content of particle-attached communities
663 was quantified and submitted for shotgun metagenomic sequencing. Communities were
664 characterized using metagenome-assembled genomes (MAGs), which were classified into three
665 predicted ecological roles for this ecosystem: chitin degraders, chitooligosaccharide exploiters,
666 and metabolic byproduct scavengers.

667

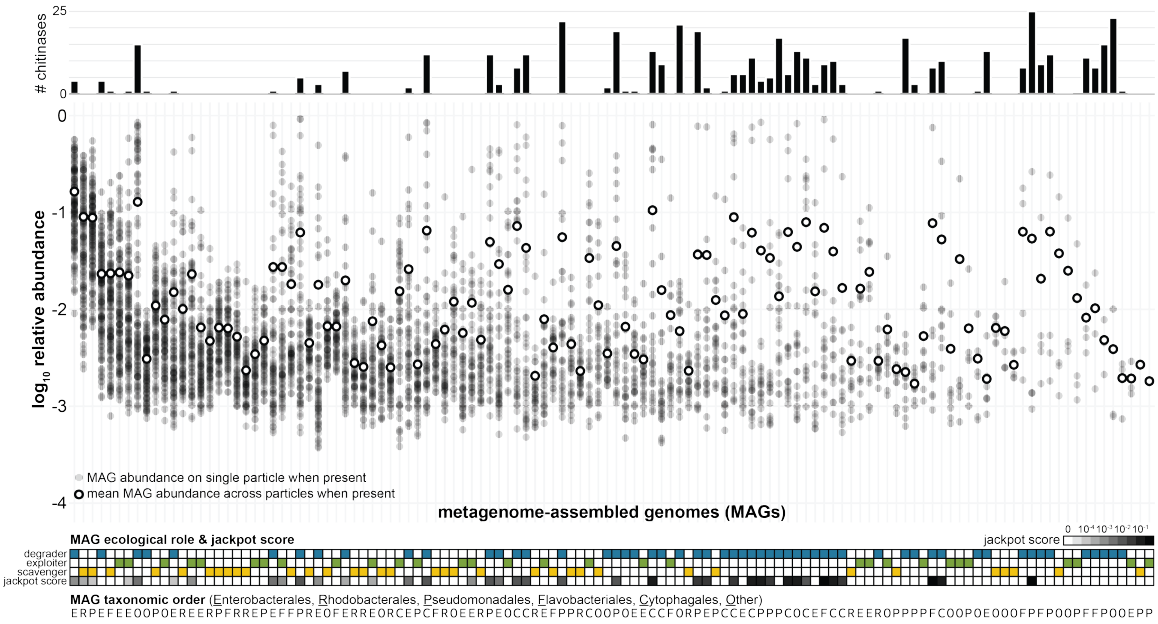


Figure 2. High compositional variability across replicate late-stage particles is driven by conditionally rare degrader taxa. Relative abundances of metagenome-assembled genomes (MAGs; *n* = 120) across late-stage particles. Smaller black dots indicate the relative abundance of each MAG per particle (*n* = 149). Larger white dots indicate the $\log_{10}$[mean relative abundance] across the particles on which the MAG was found. MAGs are sorted from left to right by their prevalence across particles (i.e. the number of particles on which they are detected). The bars above show the number of chitinases encoded in each MAG. The annotations below show each MAG's predicted ecological role (heatmap: blue = degrader, green = exploiter, yellow = scavenger); jackpot score (heatmap: white = low, black = high); and taxonomic order (E = Enterobacterales, R = Rhodobacterales, P = Pseudomonadales, F = Flavobacteriales, C = Cytophagales, O = Other). See Fig. S1 for additional details.

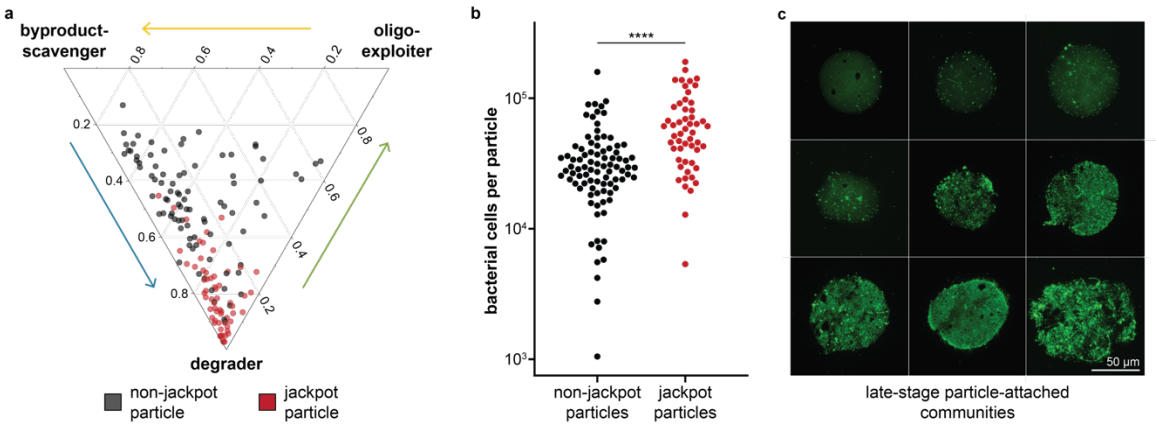


Figure 3. Late-stage particles diverge in community-level functional potential and biomass. **(a)** Ternary plot of the relative abundances of organisms occupying the three ecological roles (degrader, exploiter, scavenger) on each late-stage particle (*n* = 149), calculated by summing the relative abundances of MAGs classified into each role. Red dots represent jackpot particles, and black ones represent non-jackpot particles. Jackpot particles harbored significantly higher degrader populations than non-jackpot particles (79.8% vs. 47.4% on average; Mann-Whitney *U* test: $p < 2.2\times10^{-16}$). **(b)** Estimates of absolute bacterial cell counts on late-stage particles through qPCR of the 16S rRNA gene in DNA extracted from particle-attached communities. Jackpot particles (red dots) harbored significantly higher numbers of cells (Mann-Whitney *U* test: $p = 2.3\times10^{-7}$) than non-jackpot particles (black dots). **(c)** Representative images of late-stage particles that were harvested after 167 hours of incubation in seawater and stained with the DNA-intercalating dye SYTO 9 (scale bar, 50µm). Particle-attached communities spanned a range of growth states, from sparsely to densely populated.

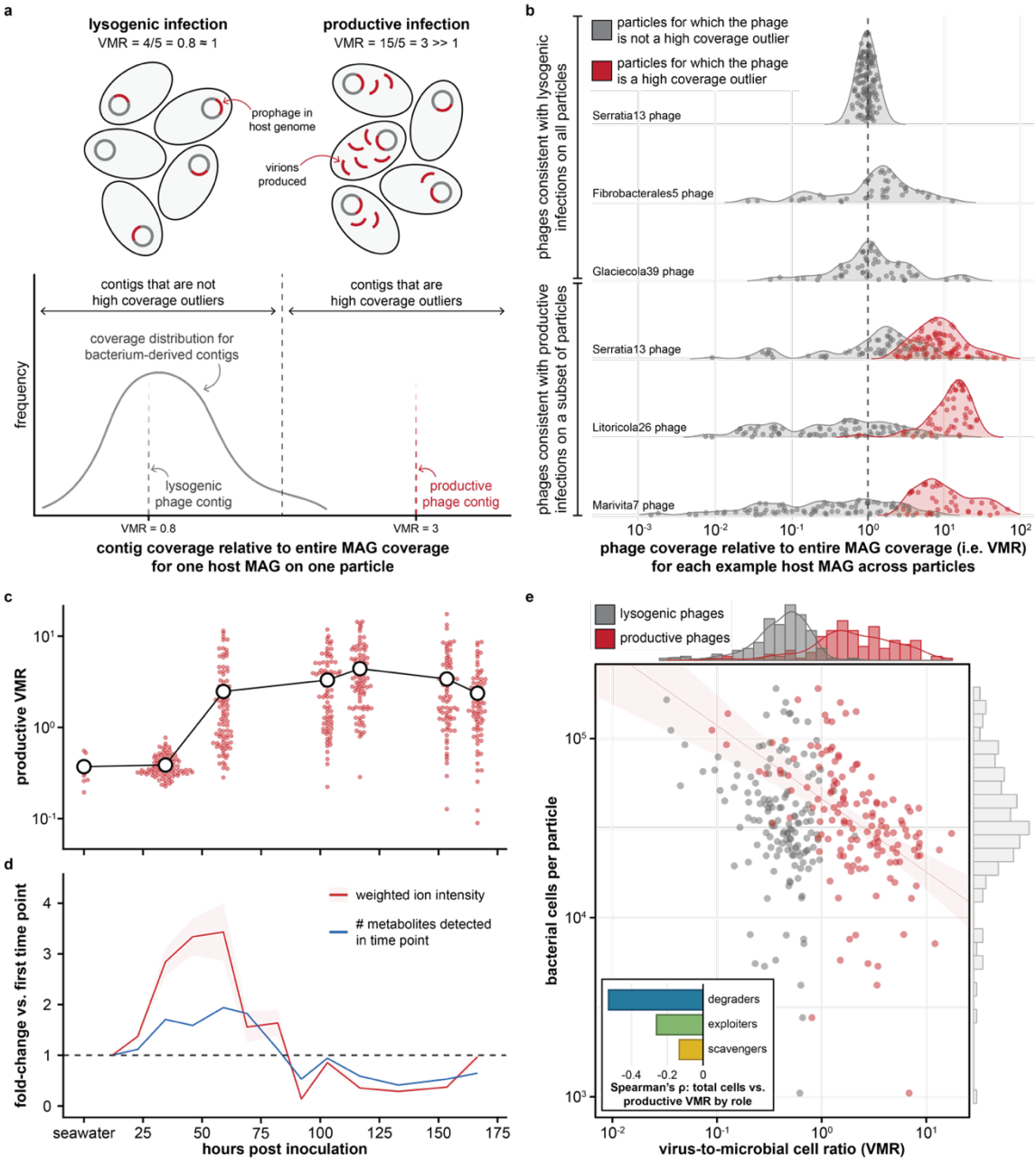


Figure 4. Bacteriophages become increasingly activated during community development and contribute to variability in bacterial abundances on late-stage particles. (a) Schematic of approach to detect productive phage infections from metagenomic data. Left: during lysogenic infections, prophages replicate with their bacterial hosts (virus-to-microbe ratio, VMR, ≈ 1, top); lysogenic phage contigs have read coverage values similar to those of most bacterial contigs of their host MAG (bottom). Right: during productive infections, prophages replicate much more than

their hosts (VMR >> 1, top); productive phage contigs have read coverage values much higher than those of most bacterial contigs of their host MAG (bottom). **(b)** Representative examples of phages with lysogenic coverage patterns on all late-stage particles (top three rows), and of phages with productive coverage patterns on a subset of particles (bottom three rows). For each phage contig, VMR is shown across late-stage particles on which each MAG is present. Gray dots, particles on which the phage contig is not a coverage outlier; red dots, particles on which the phage is a high coverage outlier. Dashed line: VMR = 1. **(c)** Total VMRs for productive phages over time. The first time point shows productive VMRs of initial seawater samples; subsequent time points show productive VMRs for chitin particle-attached communities incubated in seawater. Smaller red dots, values for individual samples; larger white dots, mean VMR for each time point. **(d)** Metabolomic profiles of the seawater surrounding chitin particles as a function of incubation duration. Values are depicted in terms of fold-change at each time point relative to the first time point (dashed line: no change). Red line (and shading): mean (±1 standard deviation) weighted ion intensity (Methods). Blue line: number of unique metabolites. **(e)** Main: Absolute bacterial cell counts on late-stage particles (n = 142), estimated through qPCR, vs. each particle's total VMR for lysogenic phages (gray dots) and productive phages (red dots). Cell counts were negatively correlated with productive VMRs (Spearman's ρ = -0.56, p = 3.3×10^{-13}; red line and shading: log-log linear regression and 95% confidence interval, R^2 = 0.23, p = 1.3×10^{-9}). Productive and lysogenic VMRs were decoupled (red vs. gray data: Spearman's ρ = 0.11, p = 0.18). Marginal histograms: distributions of productive VMRs (red), lysogenic VMRs (dark gray), and bacterial cell counts (light gray). Inset: Bar plot of values of Spearman's ρ between cell counts and productive VMRs of bacterial populations by ecological role (blue = degraders, green = exploiters, yellow = scavengers; see Fig. S14b for details).

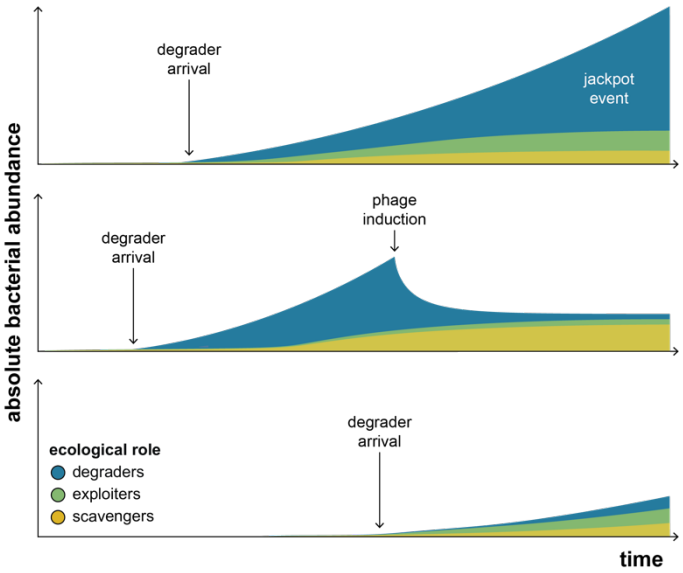


Figure 5. Conceptual model of key processes contributing to the diversification of communities on microscale particles. Schematics of community development over time are shown for three example particles, with the absolute abundances depicted for bacterial populations by ecological role (blue = degraders, green = exploiters, yellow = scavengers). Based on our conceptual model (see Discussion), high-biomass jackpot particles are those on which degraders arrive early and resist phage induction, leading to high relative and absolute degrader abundances (top). By contrast, low-biomass particles are those on which degraders are not able to proliferate, either because phage induction leads to their population collapse (middle) or because they become established on a particle relatively late (bottom).

734 **Supplementary Information Text**

735

736 **Models of abundance fluctuations**

737

738 **Notation and context**

739

740 We considered a system with M MAGs and P particles. Let x_i be the abundance of MAG i on a
741 particle and $X = \sum_i x_i$ be the total abundance. The probability distribution $p(\underline{x})$ is the probability of
742 observing a given vector of abundance $\underline{x}$, while $p_i(x)$ is the probability that species i has abundance
743 x and $P(X)$ is the probability that the total abundance is X. We also define $y_i = log x_i$ and $Y = log X$
744 (log means natural log everywhere).

745

746 Models #1-3 are reasonable models that nevertheless do not recapitulate the observed trends (i.e.
747 the right-skewed distributions of relative taxon abundances [Fig. S2] and the left-skewed
748 distribution of absolute cell abundances [Fig. 3b]), which model #4 (referenced in the main text)
749 does reproduce.

750

751 **Model #1: Stochastic arrival and exponential growth**

752

753 We assume that MAGs arrive stochastically to a particle and grow exponentially with a fixed MAG-
754 specific growth rate r_i. The log-abundance of MAG i at time t will therefore be $y_i = r_i(t - t_i^a))$,
755 where t_i^a is the arrival time of MAG i.

756

757 The only source of variation across particles is the intrinsic randomness in the arrival time, which
758 is exponentially distributed with (migration) rate λ_i. If we are considering only particles where i is
759 present, the probability should be normalized between 0 and the duration of the experiment t, which
760 leads to

761

762 $$\rho_i(t^a) = \frac{\lambda_d e^{-\lambda_i t^a}}{1 - e^{-\lambda_i t}}.$$

763

764 One can obtain the probability of observing a MAG with log-abundance y_i at time t simply by
765 inverting the relationship $y_i = r_i(t - t_i^a)$):

766

767 $$p_i(y) = \frac{\lambda_i}{r_i} \frac{exp\left(-\lambda_i t + \lambda_i \frac{y}{r_i}\right)}{1 - e^{-\lambda_i t}},$$

768

769 and, therefore, the probability of the abundance (conditioned on being present) reads

770

771 $$p_i(x) = \frac{\lambda_i}{r_i} \frac{exp(-\lambda_i t)}{1 - e^{-\lambda_i t}} x^{\frac{\lambda_i}{r_i} - 1}.$$

772

773 Note that this distribution is normalized between 0 and $\tilde{x}_i = e^{r_i t}$. We can therefore rewrite this
774 expression as

775

776 $$p_i(x) = \frac{\lambda_i}{r_i} \tilde{x}_i^{\frac{\lambda_i}{r_i}} x^{\frac{\lambda_i}{r_i} - 1}.$$

777

778 Both the arrival rate λ_i and the growth rate m_i differ across MAGs. We set their values by drawing
779 them from two independent lognormal distributions. In particular, each λ_i for $i = 1, \ldots, M$ was drawn
780 from a lognormal distribution with mean $\bar{\lambda}$ and log-variance s_λ^2. Similarly, each r_i was drawn from a
781 lognormal with mean $\bar{r}$ and variance s_r^2.

782

783 Fig. S16 shows the distribution of collapsed MAG relative abundances and the distribution of total
784 abundances obtained with this model. Model #1 always predicts a relative log-abundance
785 distribution with negative skewness and a total log-abundance distribution with non-negative
786 skewness (contrarily to what observed in the data; see Fig. 3b).

Model #2: Stochastic arrival and exponential growth with demographic stochasticity

Model #2 assumes that MAGs arrive on particles with rate λ_i. The population growth that follows is determined by a birth-death process with constant per-capita birth and death rates (b_i and d_i, respectively). The (average) growth rate r_i equals $b_i - d_i$.

Similar to the procedure of model #1, we assumed that the values of migration, growth, and death rates of each MAG were initialized as lognormal random variables with means $\bar{\lambda}$, $\bar{r}$, and $\bar{d}$ and log-variances s_λ^2, s_r^2, and s_d^2.

Fig. S17 shows that model #2 always predicts a total log-abundance distribution with positive skewness, therefore failing in reproducing the empirical shape of the total abundance distribution.

Model #3: Stochastic arrival and exponential growth with environmental stochasticity

Model #3, similarly to model #1, assumes that MAGs arrive stochastically to a particle with arrival rate λ_i and then grow exponentially. When a MAG arrives on a particle, it starts growing exponentially. The growth rate of MAG i is not fixed, equal to r_i across all particles, but is itself a random variable. In particular, the growth rates of MAG i across particles are normally distributed with mean r_i and variance σ_i^2 proportional to the mean squared: $\sigma_i^2 = c_r^2 r_i^2$, where c_r is the coefficient of variation.

As for model #1, the arrival rate λ_i and the average growth rates r_i are lognormally distributed with means $\bar{\lambda}$ and $\bar{r}$ and log-variances s_λ^2 and s_r^2.

Fig. S18 shows that model #3 always predicts a total log-abundance distribution with positive skewness, therefore failing in reproducing the empirical shape of the total abundance distribution.

Model #4: Exponential growth with environmental stochasticity conditioned on degrader presence

In the previous models, the growth of all MAGs was only conditioned on arrival. This assumption inevitably led to total log-abundance distributions with positive skewness, contrarily to the empirical observation of negative skewness.

Model #4 assumes that, for a particle to become viable for growth, the presence of a degrader MAG is required first. The arrival rate of a degrader is λ^d. All the cells that arrive to the particle after the first arrival of the degrader are able to grow. The time at which the population of MAG i on a particle will start to grow will be $t_g^i = t_a^i + t_d$, where t_d is the time of arrival of the degrader (an exponential random variable with rate λ^d) and t_a^i is the time between arrival of the degrader and the arrival of the MAG i (an exponential random variable with rate λ_i).

Starting at t_g^i, MAG i will start to grow exponentially with a random, normally distributed, growth rate with mean r_i and coefficient of variation c_r. Similar to the previous models, the arrival rate λ_i and the average growth rates r_i are lognormally distributed with means $\bar{\lambda}$ and $\bar{r}$ and log-variances s_λ^2 and s_r^2.

Fig. S19 shows that the predictions of model #4 agrees with the empirical observations. The total log-abundance distribution has a negative skewness, while the distribution of MAG relative abundances has a positive skewness. The shape of the patterns is robust across different parameters values. Only when the variation across MAGs is comparable to the growth rate fluctuations across particles ($c_r \sim s_r \sim 1$) does the total log-abundance distribution display a positive skewness.

Extended methods

Sample collection and incubation with individual chitin particles

Seawater sampling and treatment. Nearshore coastal ocean surface water samples were collected on July 15, 2017 from Canoe Beach, Nahant, MA, USA (42°25'11.5'' N, 70°54'26.0'' W). The seawater was immediately transported to Parsons Laboratory (MIT, Cambridge, MA, USA) for processing. In order to decrease the degree of dissimilarity between seawater aliquots used in incubations with chitin particles, large particulate matter was removed (using a 63μm filter), and the flow-through was concentrated via gentle centrifugation in 1L batches at 4000 × *g* for 5 minutes. The lower 100mL of each 1L batch was saved and pooled; aliquots of this water in 175μL volumes were either used for particle incubations or stored at -20°C for downstream DNA extraction and metagenomic sequencing.

Seawater incubation with individual chitin particles. Artificial seawater (ASW), used for washing and storing chitin particles, was prepared by dissolving 40g/L sea salts (Sigma-Aldrich, #S9883) in Milli-Q deionized water and filtering the solution through a 0.22-μm filter. Chitin magnetic particles (New England Biolabs, #E8036L) stored in 20% ethanol were washed three times (2mL particles resuspended in 50mL ASW) using a magnet to pull down the particles. Aliquots of washed chitin particles were further diluted in ASW in sterile petri dishes and individually selected beneath a dissecting microscope in a laminar flow hood. Single chitin particles were transferred in 3μL volumes of ASW into the wells of 96-well plates (Thermo Fisher, #AB0600L; UV-sterilized; free from DNase, RNase, and human DNA), with one chitin particle per well. The individual particles selected had a diameter of 85.0±24.0 μm, which was quantified from a set of 60 particles on an ImageXpress Micro Confocal (Molecular Devices). Plates containing individual particles were stored at 4°C until they were inoculated consecutively with 175μL of filtered, centrifuged seawater per well. The plates were sealed

(VWR, #89092-056) and rotated end-over-end at 7.5 revolutions/minute at room temperature. The particles in an entire plate were harvested at each time point (after 12, 22.75, 34.5, 46, 59, 69, 82, 92, 103, 116.75, 113, 153.5, and 166.5 hours of incubation) by pipetting the contents of each well onto a sterile petri dish and inspecting the water under a dissecting microscope in a laminar flow hood. Each particle was transferred in 1µL volumes into 96-well plates (Thermo Fisher, #AB0600L) pre-filled with 100µL of TE buffer; plates with harvested particles were stored at -20°C until downstream processing. The seawater surrounding each harvested particle was also saved in 96-well plates (Thermo Fisher, #AB0600L) and stored at -20°C until downstream processing.

DNA extraction and metagenomic sequencing. DNA extractions were performed for twelve 175µL-volume aliquots of the initial, unincubated seawater, as well as for particles harvested after 34.5, 59, 103, 116.75, 113, 153.5, and 166.5 hours of incubation. DNA was extracted from all samples with the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter) using reagent volumes 0.5X relative to those specified in the manufacturer's protocol, except for the elution buffer, of which only 30µL was used for each sample to avoid over-diluting low DNA yields. Metagenomic libraries were prepared with the Nextera XT DNA Library Prep Kit and index primers (Illumina) using the protocol developed by Rinke *et al.* (1) for low DNA input samples. While the results from the protocol in Rinke *et al.* were reproducible with as little as 100fg of input DNA, the authors recommend using a minimum of 1pg as input. Based on our qPCR measurements of DNA extracted from individual particle-attached communities (as described in the Methods section "Cell count estimation"), only one of our libraries (with 0.44pg input) was created with less than 1pg DNA. The modifications to the manufacturer's library preparation protocol included (i) diluting the Amplicon Tagment Mix 1:10 in non-DEPC-treated nuclease-free water, and (ii) increasing the number of PCR amplification cycles of the tagmented DNA from 12 to 20 cycles. Amplified libraries were purified with 0.6X AMPure XP beads. Each library was quantified on an Agilent 4200 TapeStation system with High Sensitivity D5000 ScreenTapes (Agilent Technologies) following the manufacturer's protocol, and successfully amplified libraries were pooled by time point in equimolar

amounts. Sequencing was performed on an Illumina HiSeq 2500 machine (250bp paired-end reads) at the Genome Technology Core of the Whitehead Institute for Biomedical Research (MIT, Cambridge, MA, USA). See Table S5 for all sample metadata.

Metagenomic analyses

Read pre-processing. Raw sequencing reads were clipped (to remove adapter sequences) and trimmed for quality with Trimmomatic v0.36 (2) (parameters: LEADING:3, TRAILING:3, SLIDINGWINDOW:10:20, MINLEN:36). Reads mapping to the PhiX genome were filtered out with BBDuk v38.16 (3) (parameters: k=31, hdist=1) and those mapping to the human genome (masked by Brian Bushnell at the Joint Genome Institute to prevent false positives) were identified and removed using BBMap v38.16 (parameters: minid=0.95 maxindel=3 bwr=0.16 bw=12 minhits=2 qtrim=rl trimq=10 untrim; reference genome: hg19_main_mask_ribo_animal_allplant_allfungus.fa.gz).

Metagenome assembly, binning, and MAG taxonomic assignment. Default parameters were used for all tools unless otherwise specified. Trimmed, filtered reads were error-corrected using BayesHammer (4) (a component of the SPAdes v3.13.0 pipeline) in order to improve contig assembly. Reads within each time point were pooled and co-assembled using MEGAHIT v1.2.9 (5). Assembled contigs at least 1kb in length were binned using two complementary tools – MaxBin v2.2.7 (6) and CONCOCT v1.1.0 (7). To provide CONCOCT with coverage estimates, error-corrected reads were mapped to contigs using Bowtie 2 v2.3.4.1 (8) with the parameters and approach described in Leventhal *et al*. (9). Bins generated with MaxBin and CONCOCT were consolidated and filtered using DAS Tool v1.1.1 (10) and evaluated for completeness and contamination with CheckM v1.1.2 (11). The resulting 251 bins that were at least 50% complete and at most 10% contaminated were used as reference MAGs, with median completeness and contamination values of 93.7% and 3.9%, respectively, across this set of MAGs (Table S1). Taxonomic classifications from the Genome Taxonomy Database

(GTDB) (12) were assigned to MAGs using GTDB-Tk v1.1.1 (13). Highly similar MAGs obtained from separate co-assemblies were identified and clustered through a pipeline developed by Dr. Jakob Russel for performing whole-genome comparisons of each MAG against all others with BLAT v36x2 (14). Briefly, a similarity score was calculated for each MAG relative to another by dividing the combined length of its contigs at least 98% identical to those in the compared MAG by the combined length of all its contigs. A threshold for distinguishing high similarity scores from low ones was determined using Otsu's method (15) with code derived from the R (16) package EBImage (17). 132 clusters of MAGs with mutually high similarity scores were identified, and all MAGs in each cluster had consistent GTDB-based taxonomic assignments. For one MAG cluster, one of the MAGs was classified as a different genus from the other MAGs; this MAG was separated from the cluster. We chose to consider clustered MAGs as a unit, rather than to dereplicate them, in order to retain potential strain-level microdiversity in our reference set.

MAG ecological role assignments. For each MAG, protein-coding genes were predicted and translated using Prodigal v2.6.3 (18). Predicted protein sequences were compared to a custom database of profile hidden Markov models (HMMs) of proteins involved in growth on chitin using the *hmmsearch* function of HMMER v3.3 with default parameters (19). Publicly-available HMMs were downloaded from the Pfam v33.1 (20) or TIGRFAM v15.0 databases (21) (see Table S6 for accession numbers). Custom HMMs were made by identifying experimentally-verified proteins of interest (22, 23), finding their homologs in the UniProtKB/Swiss-Prot v2020_06 database (24), creating a seed alignment using MAFFT v7 with default parameters (25, 26), and building the profile HMMs using the *hmmbuild* function of HMMER with default parameters (see Table S6 for details on each custom HMM). Protein-coding sequences were annotated based on the *hmmsearch* results if the protein length was at least 100 amino acids, the independent E-value was less than 1×10^{-9}, and the domain score was greater than 30. Only the most significant annotation was used for each protein sequence. Gene copy numbers were calculated for each MAG by tallying the number of annotations made for each protein group

(Table S1). Ecological roles (as degraders, chitooligosaccharide exploiters, or metabolic byproduct scavengers) for MAGs were defined based on the gene content patterns observed for strains previously isolated from particle enrichments (27, 28), fully sequenced, and phenotyped according to their abilities to grow on colloidal chitin, chitobiose, and GlcNAc (29). MAGs were classified as degrader genomes if they encoded at least 1 chitinase and at least 1 copy of any of the following genes: GlcNAc-specific methyl-accepting chemotaxis protein (MCP), GlcNAc-specific phosphotransferase system IIBC component (PTS), GlcNAc-specific TonB-dependent transporter (TBDT), N,N'-diacetylchitobiose phosphorylase, beta-N-acetylhexosaminidase, or GlcNAc kinase. MAGs were classified as exploiter genomes if they encoded 0 chitinases and had at least one of the following characteristics: more than 1 copy of beta-N-acetylhexosaminidase or at least 1 copy of MCP, PTS, TBDT, or N,N'-diacetylchitobiose phosphorylase. MAGs were classified as scavenger genomes if they encoded 0 chitinase, MCP, PTS, TBDT, and N,N'-diacetylchitobiose phosphorylase copies, and 1 or fewer copies of beta-N-acetylhexosaminidase. If MAGs clustered by similarity were assigned different ecological roles by these heuristics, then either (i) the role assigned to all MAGs defaulted to the role of the MAG with the lowest contamination and/or highest completeness (which occurred for 4 MAG clusters), or (ii) the MAG cluster was split into two subclusters (which occurred for 5 MAG clusters); these discrepancies are indicated in Table S1. Following this MAG cluster curation, there were a total of 138 MAG clusters.

Read mapping to MAGs for relative abundance estimation. All trimmed, filtered reads were mapped competitively against the MAGs created from sequencing particle-attached communities; the initial, unincubated seawater; and the negative controls (see the Methods section “Mock communities and negative controls”). Samples with fewer than 100,000 trimmed, filtered reads were excluded from analyses. Read mapping was performed using Bowtie 2 v2.3.4.1 (8) with the parameters and approach described in Leventhal *et al*. (9) and post-processed using SAMtools v1.7 (30). Reads that best mapped (based on alignment scores) to MAGs obtained from the negative controls (which were contaminants from laboratory reagents)

and to MAGs obtained from particle sequences that were also likely environmental contaminants (indicated in Table S1; determined through a literature search of each strain's taxonomy in studies of the marine environment) were removed from consideration when estimating community compositions. To avoid artifactually double-counting hits from paired reads, only the best hit of the forward read was considered for read pairs that survived trimming and quality filtering. Hits to completely bacteriophage-derived contigs (as opposed to prophages integrated into bacterial genome contigs) were also excluded from estimates of MAG relative abundances (see the Methods section "Bacteriophage analysis"). To minimize spurious detection, MAGs were considered to be "present" in a sample if they recruited at least 0.05% of the reads in a sample; for MAGs that recruited reads below this threshold in a sample, their abundance was set to 0 for that sample. MAG relative abundances for MAGs above this threshold were calculated for each sample by (1) tallying the hits to all MAGs in each MAG cluster; (2) normalizing the tally by the average genome length of all MAGs in each MAG cluster; and (3) dividing the normalized tallies for each MAG cluster by their sum for each sample. Therefore, for MAGs clustered together based on similarity (see the Methods section "Metagenome assembly, binning, and MAG taxonomic assignment"), their relative abundances are represented in that of the entire MAG cluster to which they belong; this calculation circumvents the artificial underestimation of MAG relative abundances that would otherwise be obtained with a non-dereplicated reference set. The relative abundances of organisms occupying the three ecological roles (degrader, exploiter, scavenger) on each particle were calculated by summing the relative abundances of MAGs classified into each role. Based on information gathered from relative abundance estimation, particles harvested at 113 hours post-inoculation were excluded from analyses because of a clear batch effect at that time point characterized by high abundances of MAG Serratia_liquefaciens93 (96.7% of particles on which Serratia_liquefaciens93 was at least 10% abundant were from t=113h, which included 98.9% of particles from that time point; this MAG was also the most abundant MAG on 81.1% of particles from t=113h and was not the most abundant MAG on any particles from other time points; see Table S7).

Comparison of variability in seawater vs. particle-associated communities. Inter-sample variability was estimated as the Aitchison distance between the community compositions of pairs of samples (i.e. the Euclidian distance between center log-ratio-transformed MAG relative abundance vectors). Aitchison distances were calculated between aliquots of the initial, unincubated seawater and between late-stage particle communities separately, and the distributions of distances between all pairs of samples were compared to each other.

Definitions of jackpot MAGs and jackpot particles. A jackpot score was calculated for each MAG cluster to quantitatively reflect the properties of rarity across most particles and dominance on a few particles. Based on relative abundances across late-stage particles, each MAG's jackpot score was defined as:

$$\frac{(coefficient\ of\ variation\ of\ relative\ abundances) * (\#\ particles\ on\ which\ MAG\ is\ the\ most\ abundant) * (highest\ relative\ abundance\ achieved)}{(\#\ particles\ on\ which\ MAG\ is\ present)^2}$$

Therefore, MAGs with high scores strongly display the jackpot phenomenon, whereas MAGs with low scores do not. The jackpot score for each particle was calculated as the weighted average of MAG jackpot scores (i.e. the sum of the relative abundance of each MAG cluster multiplied by its jackpot score). Each particle's jackpot score was compared to its species evenness (calculated as Pielou's evenness, i.e. the Shannon diversity index divided by the natural logarithm of species richness) with the expectation that particles that most strongly display the jackpot phenomenon have low species evenness. Particles were defined as "jackpot particles" if they have jackpot scores that exceed the threshold value above which log-transformed values of species evenness drop sharply (Fig. S5a); this value corresponds to the 60th percentile of jackpot particle scores. For comparing the binary categories of "jackpot degraders" and "non-jackpot degraders" (Fig. S15a), "jackpot degraders" were those MAGs that had jackpot scores greater than zero and that were present on less than 75% of late-stage particles; this thresholding was done in order to exclude the MAG clusters Serratia13 and

Fibrobacterales5 that had very low yet non-zero jackpot scores because of their high relative abundances on many particles (see Tables S1 and S7).

Calculation of the percent variance explained in MAG abundances on individual particles by the MAG abundances theoretically obtained by sequencing particles in bulk. To evaluate the extent to which community compositions at the single particle level diverged from that of a "bulk" measurement theoretically obtained by sequencing all particles together, we calculated "bulk" MAG abundances by (1) normalizing the mapped read counts to each MAG cluster by the total number of read counts for each sample; (2) summing the counts for each MAG cluster across samples; (3) normalizing the sum across samples by the average genome length of all MAGs in each MAG cluster; and (4) dividing the length-normalized counts for each MAG cluster by their sum. (These "bulk" MAG relative abundances are equivalent to the mean MAG relative abundances calculated across all particles, including those particles where MAGs are absent.) The percent variance in the abundance ranks of MAGs on single particles explained by the abundance ranks for the theoretical bulk measurement was calculated for each particle as the square of the Pearson correlation coefficient (between each individual vs. the bulk abundance rank), multiplied by 100.

Multivariate analysis. We inferred the number of conditional dependencies between MAGs from the estimated inverse covariance matrix of center log-ratio-transformed MAG relative abundances, repeating this process for 1000 randomizations of the data in which we permuted particle labels for each MAG but retained their abundance distributions. The inverse covariance matrices were estimated using a graphical lasso approach with the R package glasso (31) for several values of the regularization parameter (ρ = 0.005, ρ = 0.001, ρ = 0.0005, and ρ = 0.0001).

Read mapping to chitinases and calculating the chitinase-weighted means. To evaluate whether the use of MAGs as reference genomes could have biased our estimate of the

degrader population relative abundances in particle-attached communities, reads were also mapped to a reference set of chitinase genes (regardless of binning). All assembled contigs (binned and unbinned) were annotated for chitinase genes using the HMM-based approach described in the Methods section "MAG ecological role assignments." A custom DIAMOND database of 3,370 translated chitinase genes was created using the *makedb* function of DIAMOND v0.9.10.111 (32) with default parameters. Because of the high sequence diversity of chitinase genes, we chose to make this custom database so that the chitinase sequences used as references would be representative of those found in this experiment. Trimmed, quality-filtered reads were mapped to this database using the *blastx* function of DIAMOND with default parameters. To avoid artifactually double-counting hits from paired reads, only the best hit of the forward read was considered for read pairs that survived trimming and quality filtering. Only the most significant hit was counted for each read and only if the E-value was less than or equal to 1×10^{-25}. The number of such hits was tallied for each sample and divided by the number of trimmed, quality-filtered reads used in the mapping step to yield the percent of reads in each sample mapping to chitinase genes (Table S5). If the degrader population relative abundance estimated by MAGs were a consistent approximation of the true degrader population abundance, then the wide range in the number of chitinases encoded in each degrader MAG (Table S1) would be reflected in the percent of reads in each community mapping to chitinase genes. Therefore, the community-weighted mean (CWM) for chitinases was calculated as another comparison to the percent of reads mapping to chitinases. The chitinase CWM was calculated by multiplying the relative abundance of each degrader MAG (or MAG cluster) by the number of chitinases encoded in it (or the mean number of chitinases for a MAG cluster), and finally by summing these values.

Bacteriophage analysis.

Identifying phage-derived contigs. Binned contigs assembled from our metagenomic dataset were first classified as phage-derived or bacteria-derived using tools designed to

detect phage sequences among mixed metagenomes – namely, (i) VirSorter v1.0.3 (33) via the CyVerse platform (www.cyverse.org; National Science Foundation Awards DBI-0735191, DBI-1265383, DBI-1743442) using its RefSeqABVir database and default parameters; and (ii) VirFinder v1.1 (34) with default parameters. Contigs were classified as phage-derived if they met one of the following standards as employed in Gregory *et al*. (35): (i) they were classified by VirSorter as Category 1 or 2 (complete phage contig, higher confidence); (ii) they were classified by VirFinder with a score ≥ 0.9 and p-value < 0.05; or (iii) they were classified both by VirSorter as Category 3 (complete phage contig, lower confidence) and by VirFinder with a score ≥ 0.7 and p-value < 0.05.

Identifying productive vs. lysogenic phage-derived contigs. We used a read coverage-based approach to categorize phage-derived contigs as productive or lysogenic in particle-attached communities. Phages in a productive cycle in a particular sample would have a higher coverage than the bacterial contigs of the MAG with which they were binned because of the multiple virion copies produced per bacterial cell. In contrast, phages in a lysogenic cycle would have coverage values comparable to those of the bacterial contigs of the MAG with which they were binned. We reasoned that contigs classified as phage-derived, especially those belonging to the genomes of temperate phages, were likely to be binned into the MAGs of their bacterial hosts because: i) phage *k*-mer signatures tend to be more similar to those of their specific hosts than to those of random bacteria (34, 36, 37); ii) phages in a lysogenic cycle will have the same sequencing read coverage patterns as their hosts across samples; and iii) accordingly, both of the binning algorithms we employed clustered contigs based on their tetranucleotide frequencies and their coverage levels across multiple samples. Because our approach relied on comparisons between co-binned phages and MAGs, we considered in our analyses only phage-classified contigs at least 5kb in length, since the likelihood of mis-binning decreases with increasing contig length.

Based on read mapping to MAGs (see the Methods section "Read mapping to MAGs for relative abundance estimation"), per-base coverage values for all binned contigs were computed with the *genomecov* function of BEDTools v2.27.0 (38) and were used to calculate contig-wide average coverage values. For each MAG and for each sample, a phage-derived contig was considered to be productive if its coverage was greater than the coverage of the 95th percentile bacteria-derived contig in the same MAG. A phage derived-contig was considered to be lysogenic in a sample if its coverage did not exceed the coverage of the 95th percentile bacteria-derived contig in the same MAG. Through this pipeline, we identified 263 phage contigs with coverage patterns consistent with productive infections in a subset of samples and 256 phage contigs with coverage patterns consistent with lysogenic infections in all samples (Table S2).

Calculating virus-to-microbial cell ratios (VMRs). The VMR of an individual phage contig in one sample is defined as the phage contig coverage divided by average coverage of the MAG with which it is binned (which was calculated using only the bacteria-derived contigs). Total VMRs – i.e. the total number of phage copies relative to the total number of bacterial MAG copies in an entire sample – were calculated separately for productive and lysogenic phage contigs. The total productive VMR for a sample was defined as:

$$\sum_{i}^{n}\left[\left(\frac{average\ coverage\ of\ productive\ phage\ contigs\ in\ MAG_i}{average\ MAG_i\ coverage}\right)\right.$$

$$\left.\times\ (MAG_i\ relative\ abundance)\right]$$

$$=\frac{total\ \#\ phage\ copies\ (due\ to\ productive\ infections)}{total\ \#\ bacterial\ genome\ copies}$$

where *n* is the number of MAGs found in a sample. This calculation is equivalent to

$$\frac{\sum_i^n (average\ coverage\ of\ productive\ phage\ contigs\ in\ MAG_i)}{\sum_i^n (average\ MAG_i\ coverage)}$$

where *n* is the number of MAGs found in a sample. Similarly, the total lysogenic VMR for a sample was defined as

$$\sum_i^n \left[\left(\frac{average\ coverage\ of\ lysogenic\ phage\ contigs\ in\ MAG_i}{average\ MAG_i\ coverage} \right) \times (MAG_i\ relative\ abundance) \right] = \frac{total\ \#\ phage\ copies\ (due\ to\ lysogenic\ infections)}{total\ \#\ bacterial\ genome\ copies}$$

where *n* is the number of MAGs found in a sample. VMRs for each ecological role (i.e. for the subpopulation in a community that belongs to one of the three roles of degrader, exploiter, or scavenger) were calculated using the same formulas as above while considering only the MAGs of each role and their associated phages. When calculating total VMRs, we used the average coverage value of all phage contigs in each MAG (rather than the sum of the coverage values for all phage contigs in each MAG) to obtain a more conservative estimate of phage copy number. For example, if two phage contigs belonged to the same phage genome but did not overlap in sequence, they would appear to be two separate phages; thus, using their sum would double the apparent phage copy number, while using their average would provide a more accurate representation of their abundance.

Analysis controls. Given that read coverage from metagenomic data is often noisy, it is conceivable that phage contigs identified as “productive” have high coverage relative to their associated bacterial MAGs simply due to sequencing noise. We performed two analyses to examine this possibility. Firstly, we considered that because productive phages

are identified based on coverage, there is a chance that more productive phages would be found in samples with more reads. (Ensuring that this is not the case is one of the controls used in Kieft *et al.* (39), which also employs a coverage-based method for finding productive phages in mixed metagenomes.) Therefore, we calculated the Spearman's correlation coefficient between the number of reads in a sample and the number of phage-derived contigs with coverage values above the 95th percentile for their MAG (as described in the Methods section "Identifying productive vs. lysogenic phage-derived contigs"). We calculated these correlations for samples within time points to avoid spurious correlations created by systematic differences in the number of reads obtained across time points. The phage contigs from the MAGs that showed a significant correlation ($p < 0.05$) were excluded from analyses. Secondly, for each sample, we compared the average coverage of all phage-derived contigs (≥ 5kb) with coverage values above the 95th percentile for their MAG with the average coverage of all bacteria-derived contigs (≥ 5kb) with coverage values above the 95th percentile for their MAG. If the high coverage phage contigs have comparable average coverage to the bacterial contigs, that would indicate that the phage contigs had high coverage only due to sequencing noise. The average bacterial coverage is larger than the average phage coverage in only 2.0% (3/149) of late-stage particles (Fig. S12). Therefore, for samples with high total productive VMRs, phage contigs with high coverage values likely represent phages that were replicating more than their bacterial hosts, rather than representing contigs with randomly higher coverage values.

Mock communities and negative controls. In order to quantify the technical error associated with creating metagenomic libraries from low DNA inputs, mock communities were simulated by combining the DNA of two strains isolated from a previous chitin particle enrichment experiment using seawater from the same location sampled for this project (27). The total genomic DNA of *Vibrio splendidus* strain 1A01 (BioProject #PRJNA414740, Accession #PDUR00000000) and *Maribacter sp*. 6B07 (BioProject #PRJNA414740, Accession #PDUT00000000) was extracted using the MasterPure DNA Purification Kit (Epicentre), and double-stranded DNA content was

quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). The DNA of each strain was mixed in equimolar amounts and serially diluted to either 50pg or 5pg total (to include a range of expected DNA input concentrations from extractions of communities attached to single chitin particles). Metagenomic libraries from three technical replicates of each concentration, as well as from six negative controls (containing only nuclease-free water), were prepared using the same protocol used for individual chitin particle-attached communities (as described in the Methods section "DNA extraction and metagenomic sequencing"). The results from the mock community sequencing are shown in Fig. S4. Of the six negative control libraries, only three amplified; the MAGs recovered from these samples included *Delftia acidivorans* and a *Brevundimonas sp.*, which belong to taxonomic groups previously found as contaminants in laboratory reagents used in DNA extractions and sequencing (40, 41). These MAGs were included as references for the metagenomic analysis, and the reads best mapping to them (based on alignment scores) were removed from consideration when estimating community compositions.

Cell count estimation. Bacterial DNA extracted from individual particle-attached communities was quantified through qPCR of the 16S rRNA gene using the Femto Bacterial DNA Quantification Kit (Zymo Research), which has a lower limit of detection of 20fg. Two sets of standards and negative controls were included in each qPCR run. The number of bacterial cells for each particle was estimated from the absolute DNA amounts based on measurements indicating a mean of 2.5fg DNA per bacterial cell in seawater samples (42).

Metabolomics experiments and analyses. We performed untargeted metabolomics of the seawater that surrounded each chitin particle (after removing the chitin particles at each time point) and of the initial, unincubated seawater (t=0). All samples were first diluted 1:100 in nuclease-free water (in two serial 1:10 dilutions). We used a binary LC pump (Agilent Technologies) and an MPS2 Autosampler (Gerstel) coupled to an Agilent 6520 time-of-flight mass spectrometer (Agilent Technologies) operated in negative mode, at 2GHz, extended dynamic range, with an *m/z* (mass/charge) range of 50-1000. The mobile phase consisted of isopropanol:water (60:40, v/v) pH

9, with the addition of 5mM ammonium fluoride and a flow rate of 150 µl/min. Raw data were processed and analyzed using preprocessing raw mass spectrometry data functions contained in the bioinformatics toolbox of MATLAB (43, 44). We detected 5714 ions, of which 121 were annotated against a curated library of metabolites that are present in marine microbes, based on the BioCyc database (45). Certain ions were matched with multiple isomeric or isobaric compounds (as noted in Table S3). Detectable metabolites were those with ion intensities that passed the detection threshold above the inoculum [sample ion intensity > (mean ion intensity at t=0) + (3*standard deviation of ion intensity at t=0)]. For metabolites that exceeded the limit of detection, the intensities of each ion were normalized between 0 and 1, where 0 is the limit of detection and 1 is the highest intensity measured of a given ion. Weighted ion intensities for each timepoint were calculated by taking the sum of all normalized intensities of ions in all samples for each timepoint.

Sample collection and incubation with many chitin particles. Seawater was collected on the day of the experiment from Canoe Beach, Nahant, MA, USA (42°25'11.5'' N, 70°54'26.0'' W), the same source as seawater used elsewhere in this study. Chitin magnetic particles (New England Biolabs, #E8036S) were collected on a 40µm cell strainer then passed through a 100µm cell strainer to restrict the size range of the particles to 40-100µm (Corning). The size selected particles were then resuspended in 0.2µm-filtered natural seawater to create three suspensions: 807±99 particles/mL (± indicates standard deviation, *n* = 3), 182±28 particles/mL (sd, *n* = 3), or 88±3 particles/mL (sd, *n* = 3). Unfiltered natural seawater, containing microbes, was left undiluted, or diluted 1:10, or diluted 1:100 into 0.2µm-filtered natural seawater to create three different initial densities of bacterioplankton. All combinations of particles and cells were combined by adding 5mL particle mixture to 10mL cell mixture to create a matrix of 9 separate conditions. Particle/cell mixtures were incubated in 15mL polystyrene tubes (Falcon) with end-over end rotation at a rate of 8 revolutions/minute on a Stuart SB3 rotator at room temperature (21-25°C).

Imaging and quantification of chitin particle colonization by natural seawater bacteria. For the experiment incubating chitin particles individually in seawater (see the Methods section

"Seawater incubation with individual chitin particles"), at each time point, the communities on a subset of particles (that were not sequenced) were stained with the DNA stain SYTO9 (Invitrogen, #S34854) at a final concentration of 500nM. STYO9 was added directly to the wells containing the particles and seawater, which were subsequently incubated in the dark at room temperature for 15 minutes before the individual chitin particles were harvested (as described in the Method section "Seawater incubation with individual chitin particles") and mounted separately on microscope slides. Particles were imaged with a Zeiss epifluorescence microscope at 100X magnification. For the experiment incubating many particles together in seawater (see the Methods section "Sample collection and incubation with many chitin particles"), after 24 hours of incubation, 200µl samples of each condition were stained with SYTO9 at a final concentration of 5µM. The SYTO9-stained samples were transferred to a black-walled Greiner Bio-One µClear 96-well plate. Samples were imaged on an ImageXpress Micro Confocal (Molecular Devices) in widefield mode using a Nikon 10x Plan Apo lambda objective (NA 0.45) and FITC filter (ex 482/35, em 536/40, dichroic 506 nm) with blue LED illumination from a Lumencore Light Engine. Nine fields of view capturing the entire well were acquired to quantify all particles present in each well. For chitin particles incubated both individually and in bulk, a custom analysis script was written in MATLAB vR2019a (The Mathworks) to quantify the area of each chitin particle colonized by cells. The code defines chitin particle area, and the area of each particle covered by cells using intensity-based thresholds. Code and original data will be publicly available before publication at the following GitHub page: https://github.com/jaschwartzman/seawater_colonize

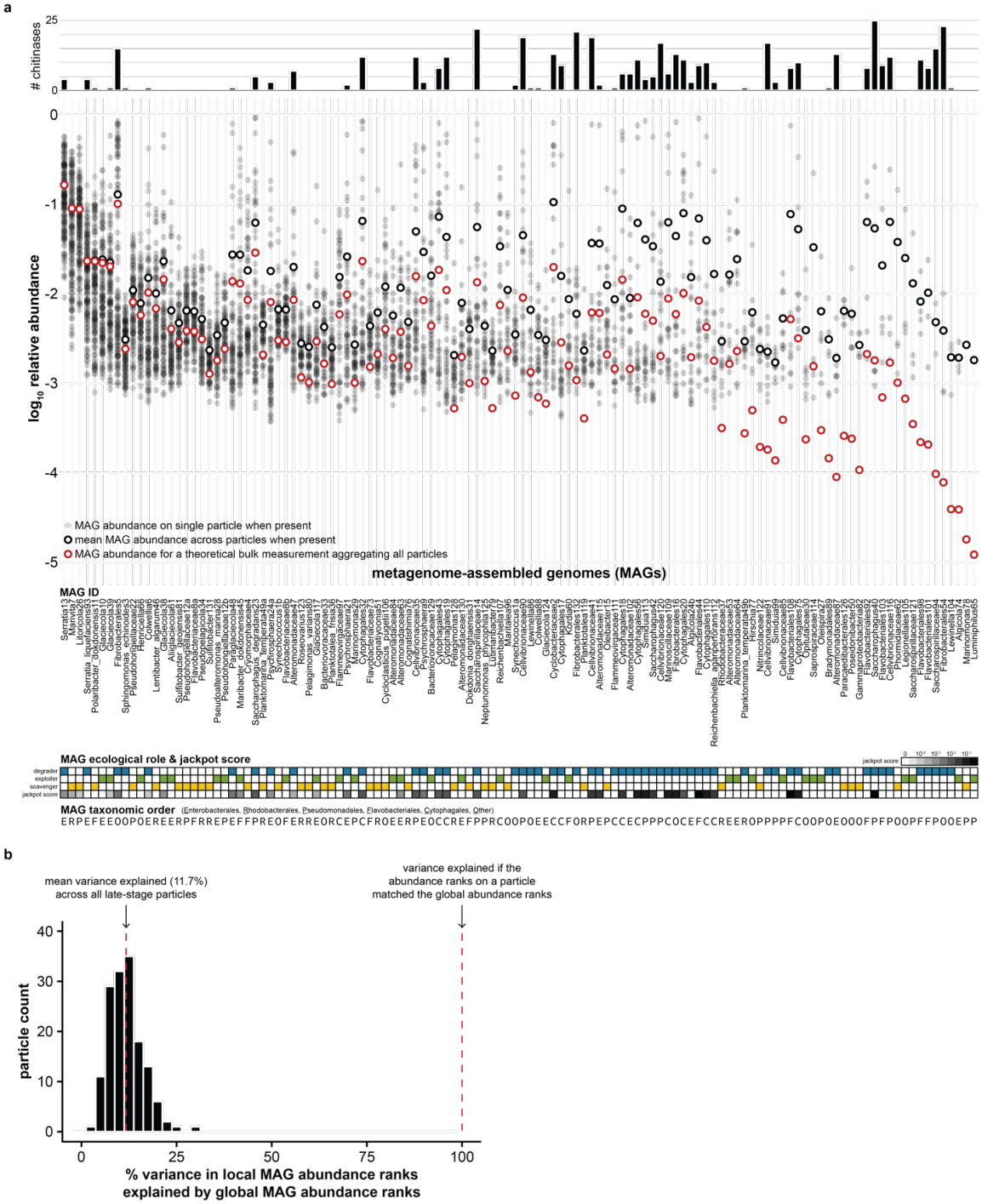
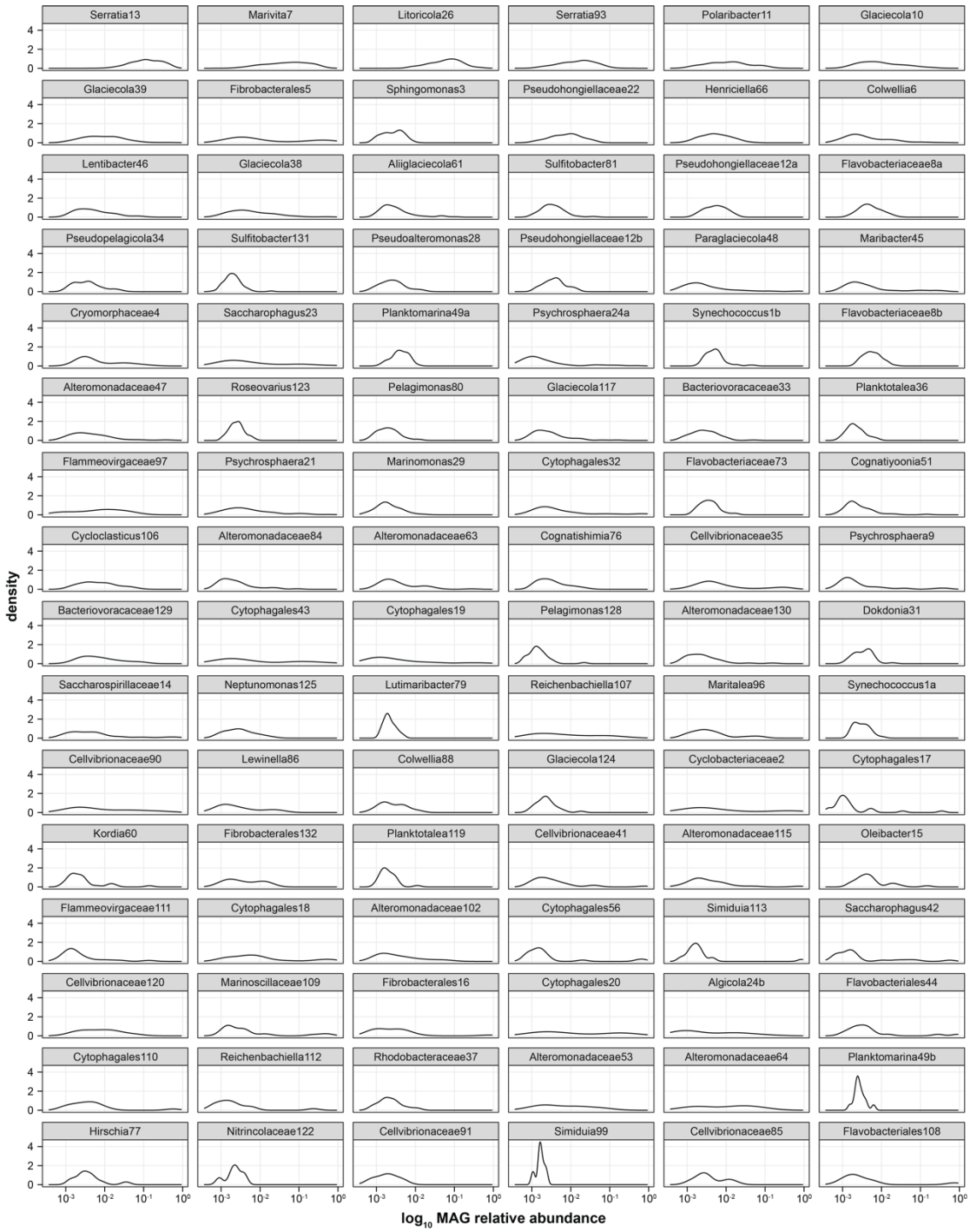


Fig. S1. Taxonomic abundances averaged across particles do not represent the compositions of communities on individual particles. (a) Extended version of Fig. 2. Relative abundances of metagenome-assembled genomes (MAGs; *n* = 120) across late-stage particles. Smaller black dots indicate the relative abundance of each MAG per particle (*n* = 149). Larger black

circles indicate the $\log_{10}$[mean relative abundance] across the particles on which the MAG was found. Larger red circles indicate the $\log_{10}$[mean relative abundance] across all the particles (i.e. the MAG abundances for a theoretical bulk measurement aggregating all particles). MAGs are sorted from left to right by their prevalence across particles (i.e. the number of particles on which they are detected). The bars above show the average number of chitinases encoded in each cluster of highly similar MAGs (see Methods). The annotations below show each MAG's taxonomic ID (matching Table S1); predicted ecological role (heatmap: blue = degrader, green = exploiter, yellow = scavenger); jackpot score (heatmap: white = low, black = high); and taxonomic order (E = Enterobacterales, R = Rhodobacterales, P = Pseudomonadales, F = Flavobacteriales, C = Cytophagales, O = Other). (**b)** Histogram of the percent variance explained in the abundance ranks of MAGs on each late-stage particle by the abundance ranks for a theoretical bulk measurement aggregating all particles (which is equivalent to the average abundance across all particles). If the MAG abundance rank of a single particle's community matched that of the theoretical average, the percent variance explained would be 100% (right dashed line); however, the ensemble scale explained only an average of 11.7% (left dashed line) of the variance in abundance ranks at the single particle level.



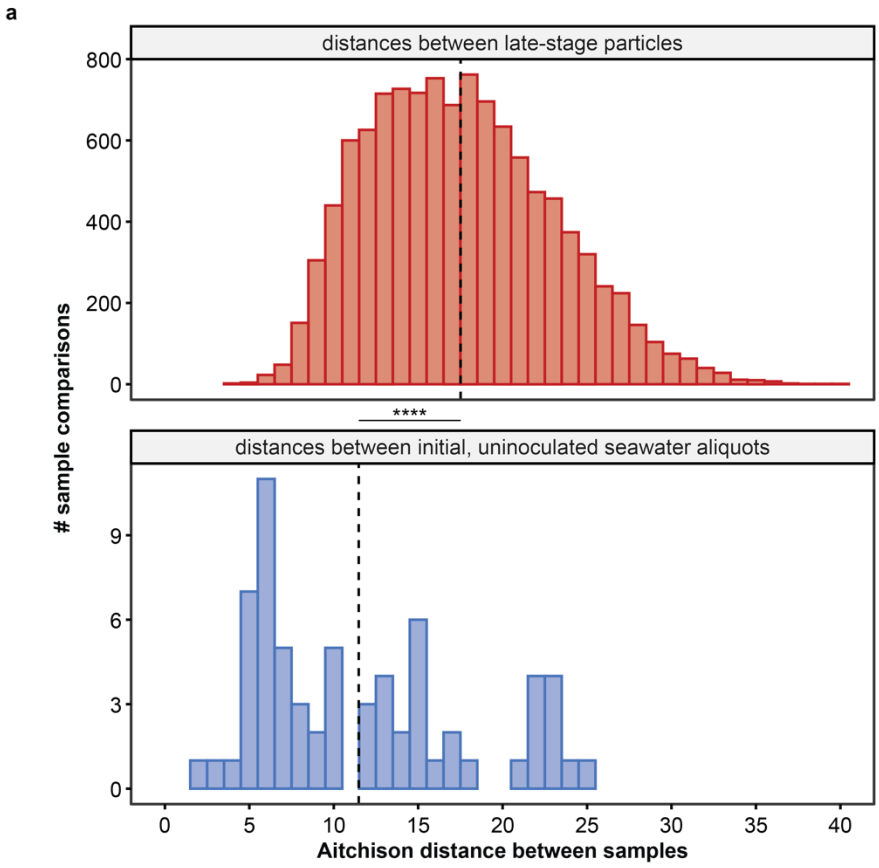
1296

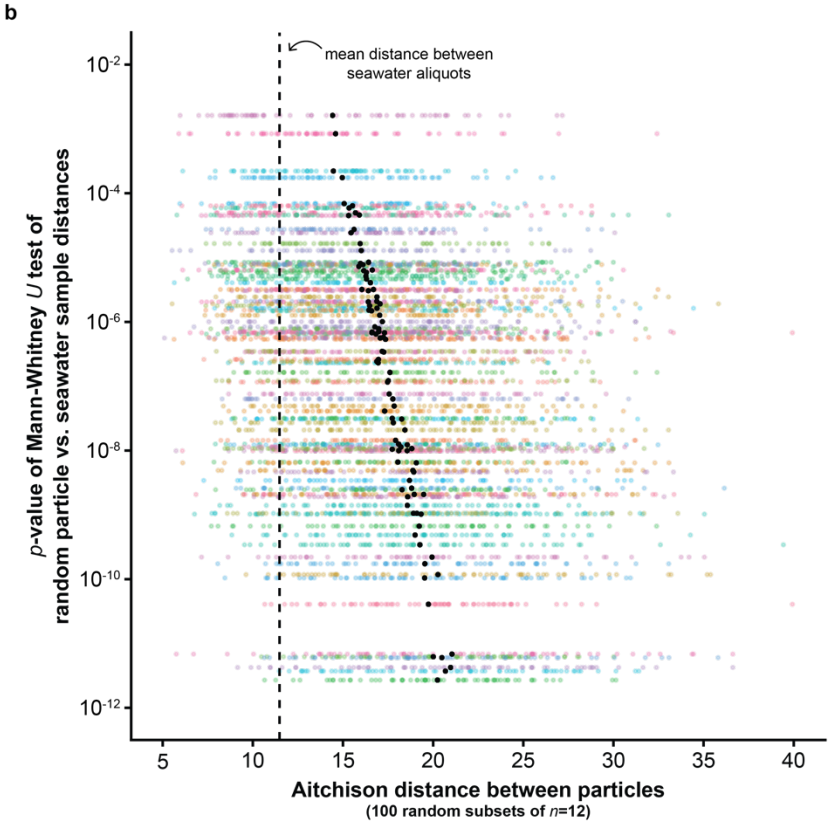
1297 **Fig. S2. Distributions of MAG relative abundances on late-stage particles are approximately**

1298 **lognormal and right-skewed (i.e. towards high frequencies).** Distributions are shown as

1299 Gaussian kernel density estimates for MAGs present on at least 10 late-stage particles. The area

1300 under each curve equals one.





1301

Fig. S3. Taxonomic variability in the initial seawater does not significantly account for variability observed across late-stage particles. (a) Distributions of the Aitchison distances (Methods) calculated between all pairs of communities on late-stage particles (*n* = 149, red histogram) and between all pairs of aliquots of unincubated, initial seawater (*n* = 12, blue histogram). Dashed vertical lines represent the means of each distribution. Late-stage particles were significantly more dissimilar from one another than initial seawater samples (Mann-Whitney *U* test: $p = 1.3 \times 10^{-13}$). **(b)** The amount of inter-sample variability detected could depend on sample size, and many more pairs of particles than pairs of seawater samples were assessed in **(a)**. Therefore, we calculated the Aitchison distances between random subsets of 12 late-stage particles and compared those distributions to that of the seawater samples. Small points represent inter-particle Aitchison distances calculated for 100 random subsets (each with its own point color), and black dots indicate the mean value for each subset. The inter-particle distances for each subset are plotted against the *p*-value from a Mann-Whitney *U* test comparing the particle and seawater distributions. The dashed vertical line indicates the mean Aitchison distance between seawater samples (the same value as in the blue histogram in **(a)**). For all particle subsets, inter-particle distances were significantly higher than inter-seawater distances.

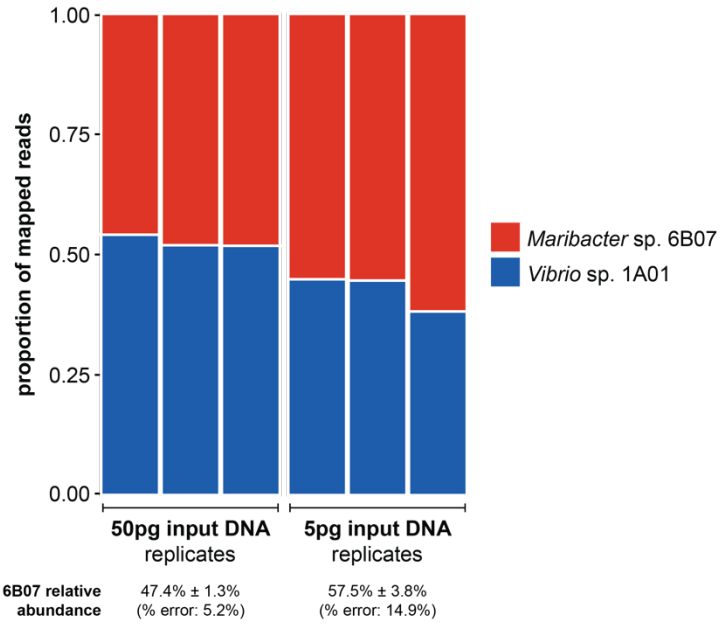


Fig. S4. Mock communities sequenced with same protocols as particle-attached communities show relatively little deviation from expected strain abundances. See Methods for details on the preparation of the mock communities, which contained equal proportions of *Marinobacter* sp. 6B07 genomic DNA and *Vibrio* sp. 1A01 genomic DNA. Relative abundances estimated from metagenomic libraries prepared using 50 pg of input DNA showed 5.2% error, whereas libraries prepared with 5 pg of input DNA showed 14.9% error.

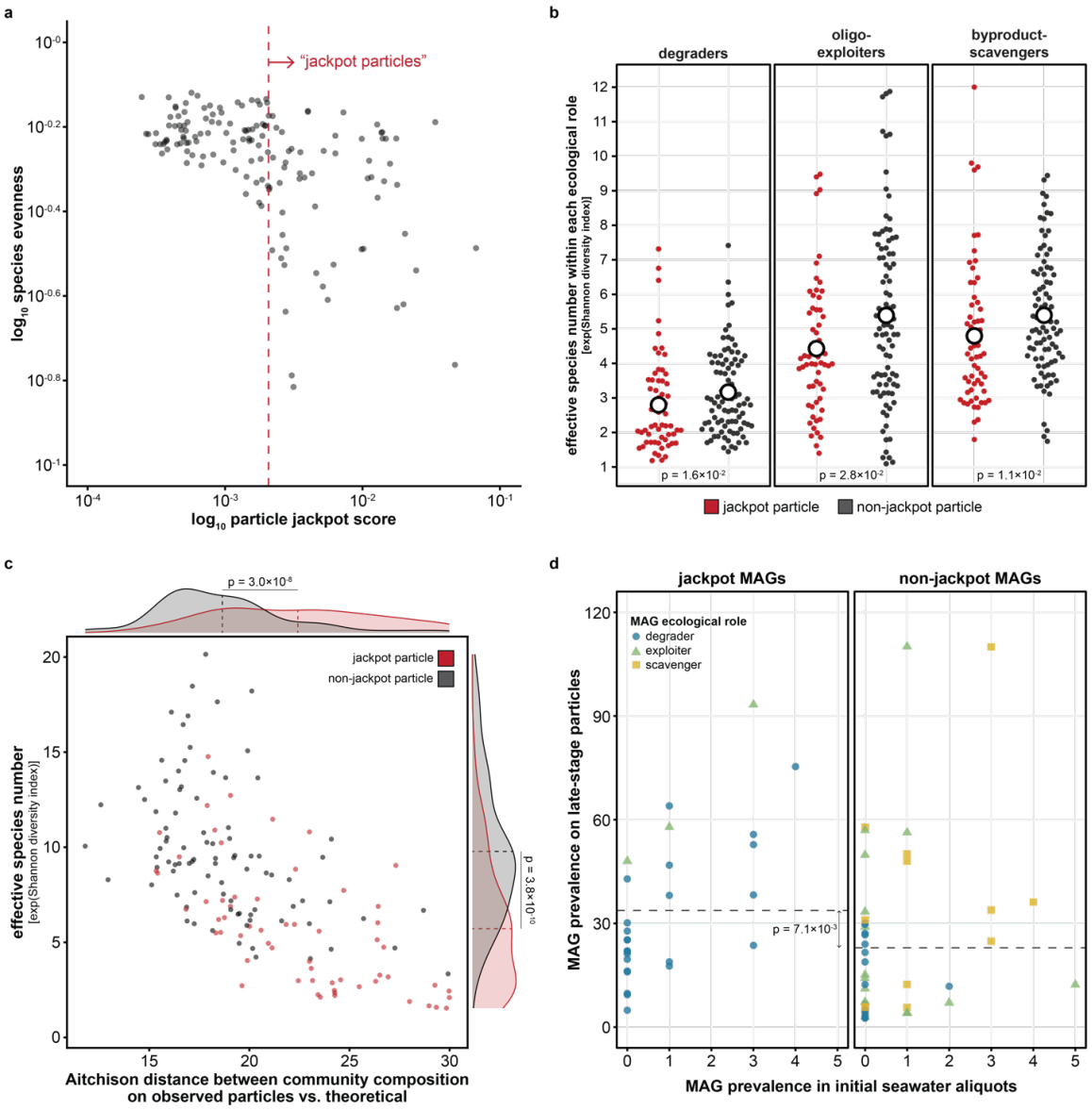


Fig. S5. Communities on jackpot particles are dominated by globally rare and locally abundant strains. (a) Particles were defined as "jackpot particles" if they had high jackpot scores (indicating high relative abundances of jackpot taxa; see Methods for details). Each dot represents one late-stage particle (n = 149), and the red dashed line indicates the particle jackpot score threshold above which log-transformed values of Pielou's species evenness drop sharply. **(b)** The effective species numbers (calculated from the Shannon diversity index) within each ecological role. Each smaller dot represents a late-stage particle (n = 149), and dot color indicates whether the particle was a jackpot particle (red) or a non-jackpot particle (black). Larger white dots represent

the mean of each distribution. The diversity on jackpot particles was significantly lower than on non-jackpot particles for each of the roles (Mann-Whitney *U* test: degraders $p = 1.6\times10^{-2}$, exploiters $p = 2.8\times10^{-2}$, scavengers $p = 1.1\times10^{-2}$). **(c)** Community diversity (represented as effective species number, calculated from the Shannon diversity index) was inversely correlated (Spearman's $\rho = -0.68$, $p < 2.2\times10^{-16}$) with the Aitchison distance between the community composition observed on each particle and the composition of the theoretical average particle (see larger red circles in Fig. S1a). Each dot represents a late-stage particle ($n = 149$), and dot color indicates whether the particle was a jackpot particle (red) or a non-jackpot particle (black; see Methods for definitions). Jackpot particle communities were significantly less diverse (Mann-Whitney *U* test: $p = 3.8\times10^{-10}$) and more divergent from the theoretical average particle (Mann-Whitney *U* test: $p = 3.0\times10^{-8}$) than non-jackpot particles. **(d)** Jackpot taxa (left panel) were significantly more prevalent across late-stage particles (Mann-Whitney *U* test: $p = 7.1\times10^{-3}$) than non-jackpot taxa (right panel) that were equally rare across aliquots of the initial, unincubated seawater (prevalence in seawater samples, Mann-Whitney *U* test: $p = 0.33$; mean abundance in seawater samples, Mann-Whitney *U* test: $p = 0.49$). Each point represents a MAG that was detected on fewer than half of the seawater aliquots, with the point color and shape indicating its predicted ecological role (blue circle = degrader, green triangle = exploiter, yellow square = scavenger).

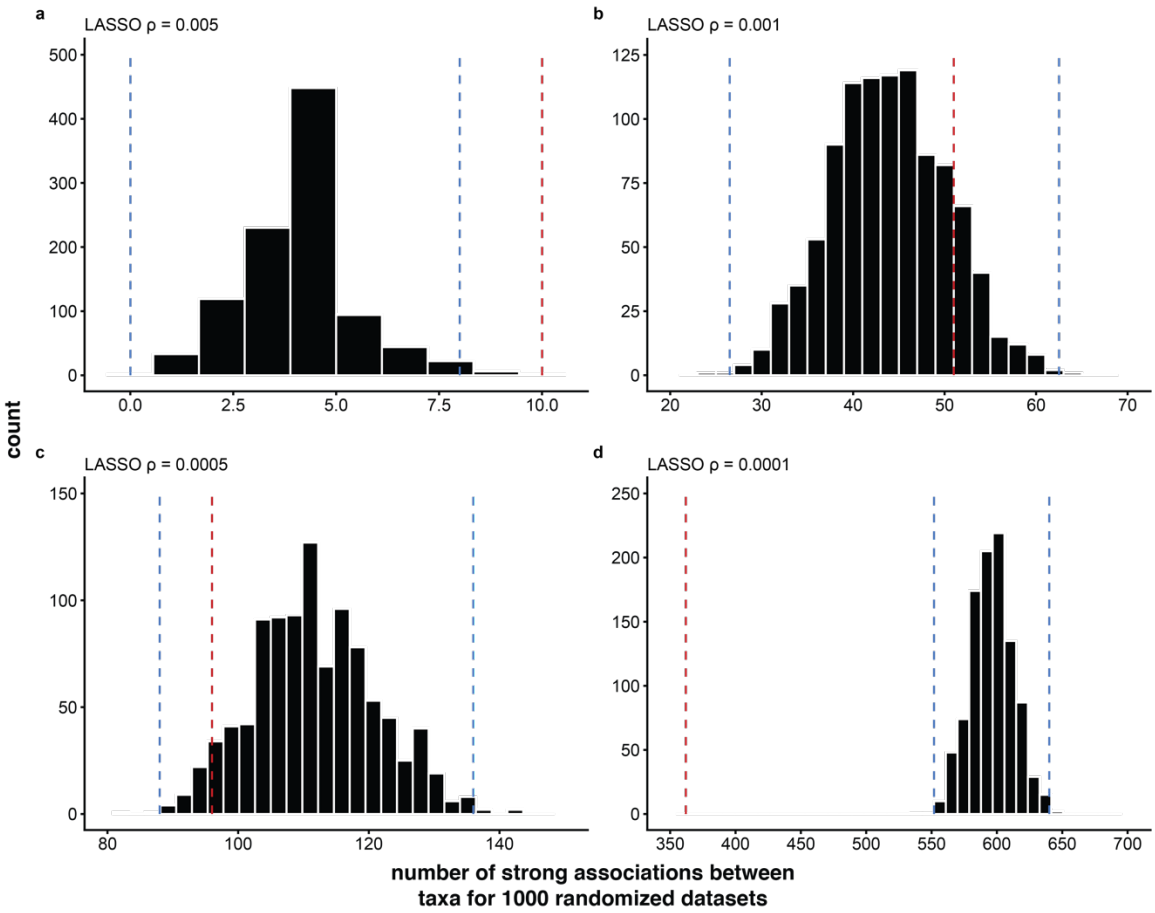
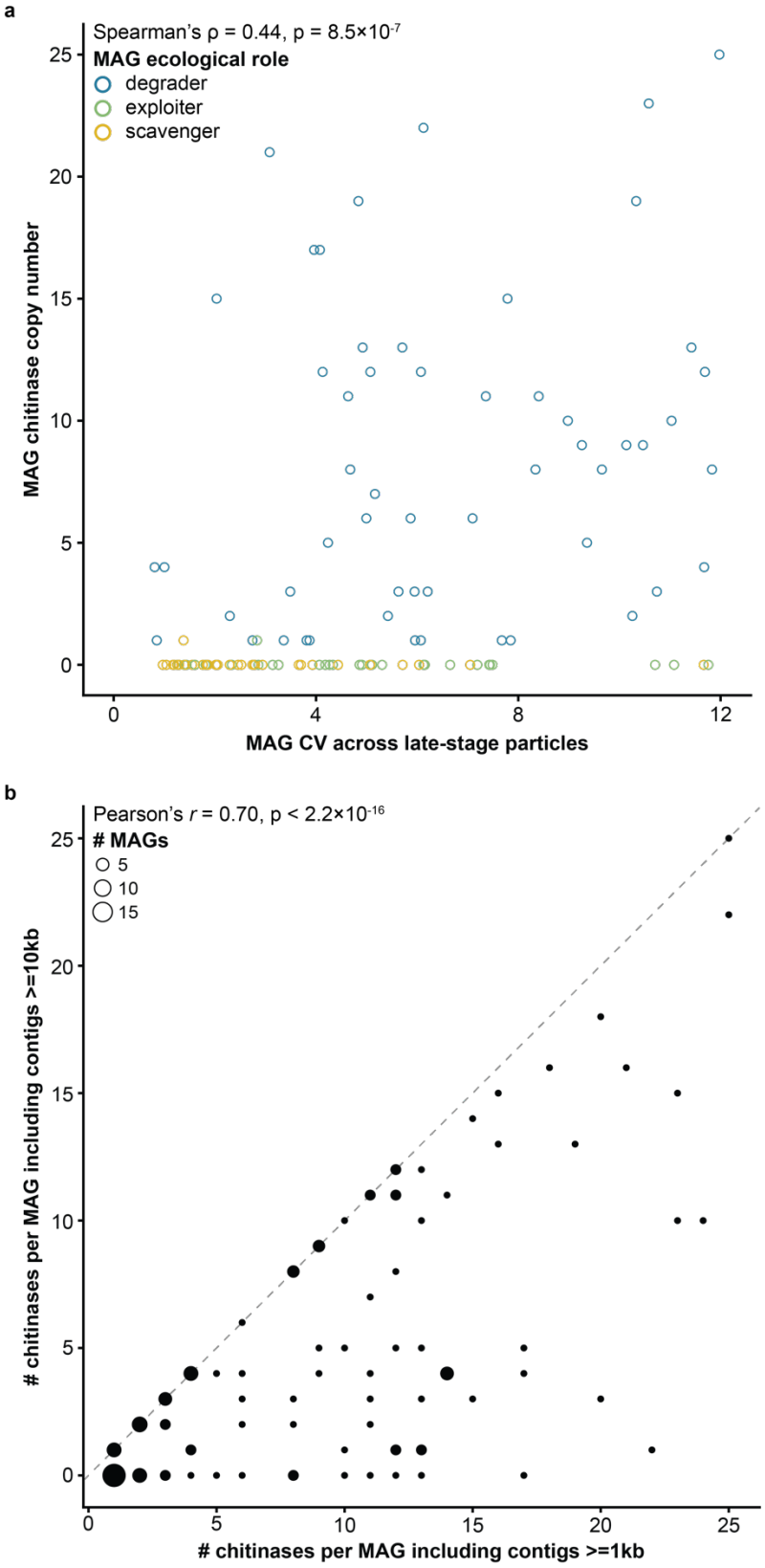


Fig. S6. Late-stage particles exhibit little specific taxonomic structure. For the observed data, as well as for 1000 randomizations of the data, we inferred the number of conditional dependencies between taxa from the estimated inverse covariance matrix of center log-ratio-transformed relative abundances across late-stage particles. The inverse covariance matrices were calculated using a graphical lasso approach with a regularization parameter of **(a)** ρ = 0.005, **(b)** ρ = 0.001, **(c)** ρ = 0.0005, or **(d)** ρ = 0.0001. Each plot shows the distribution of the number of conditional dependencies (with strengths ≥ 0.2 or ≤ -0.2) between MAGs inferred for the randomizations of the data. The red lines indicate the number of conditional dependencies (with strengths ≥ 0.2 or ≤ -0.2) inferred from the observed data. The blue lines indicate thresholds beyond which values are considered outliers relative to the distribution calculated for the randomized datasets (using the interquartile range [IQR] method – left lines indicate the value of [Q_1 – 1.5×IQR], and right lines indicate the value of [Q_3 + 1.5×IQR]). The choice of the regularization parameter in the analyses

used to estimate the number conditional dependencies between taxa resulted in more strong associations being inferred for the observed communities than the randomized ones only when a trivially small number of associations were inferred (panel **a**). Therefore, in terms of the number of strain-specific associations, the observed particles were either indistinguishable from, or less structured than, random communities.



1369

Fig. S7. Highly variable taxa are often degrader strains encoding many chitinase genes. (a) There was a significant positive correlation between the coefficients of variation of MAG relative abundances across late-stage particles and the number of chitinase genes encoded in MAGs (Spearman's ρ = 0.44, p = 8.5×10^{-7}; calculated for 120 MAGs across 149 particles). Each open dot represents a MAG, with the color indicating its predicted ecological role (blue = degrader, green = exploiter, yellow = scavenger). **(b)** Comparison of the number of chitinase genes encoded by each MAG when considering contigs ≥ 10kb (which are binned more reliably than shorter contigs) vs. considering contigs ≥ 1kb (the minimum length of binned contigs). There was a strong correlation between chitinase copy numbers when considering contigs ≥ 10kb vs. contigs ≥ 1kb (Pearson's r = 0.70, p < 2.2×10^{-16}), lending confidence to estimates of high chitinase copy numbers in certain bins. Dot sizes indicate the number of MAGs at each coordinate.

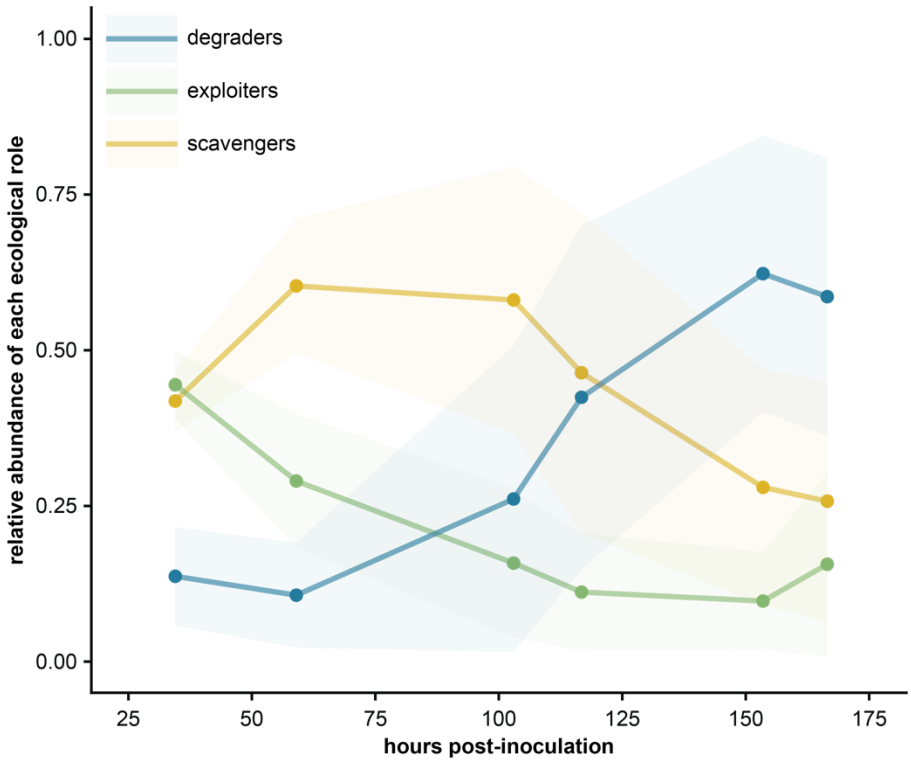


Fig. S8. The proportion of predicted degraders on particles increases and becomes more variable over time. The relative abundances of the predicted ecological roles (degrader = blue, exploiter = green, scavenger = yellow) on particles harvested after varying incubation durations were calculated by summing the relative abundances of MAGs classified into each role. Points indicate mean values across particles at each time point, with shading representing ±1 standard deviation. The number of particles considered at each time point were (in order from early to late) 88, 88, 80, 90, 76, and 73.

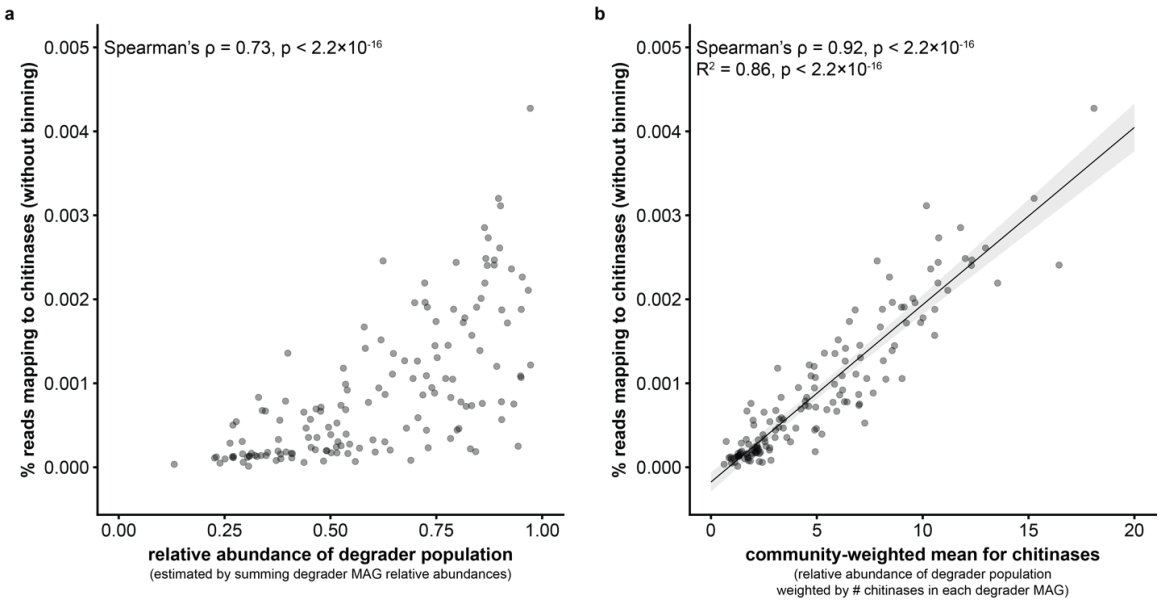


Fig. S9. The wide range of degrader population relative abundances estimated for late-stage particles holds when genes are used as read mapping references rather than MAGs. Conceivably, the use of MAGs as reference genomes could have biased our estimate of the degrader population abundance; therefore, we also mapped reads to a reference set of chitinase genes (Methods). **(a)** There was a strong correlation (Spearman's ρ = 0.73, p < 2.2×10^{-16}) between the degrader population relative abundances estimated using MAGs and the percent of reads that mapped to chitinase genes. Each dot represents one late-stage particle (n = 149). **(b)** If the degrader population relative abundance estimated by MAGs were a consistent approximation of the true degrader population abundance, then the wide range in the number of chitinases encoded in each degrader MAG (Fig. S6, Table S1) would be reflected in the percent of reads in each community mapping to chitinase genes. Therefore, we weighted the degrader population relative abundances by the number of chitinases in each MAG to calculate the community-weighted mean for chitinases of each late-stage particle (Methods). There was an even stronger correlation (Spearman's ρ = 0.92, p < 2.2×10^{-16}) between the chitinase community-weighted mean estimated using MAGs and the percent of reads that mapped to chitinase genes. Each dot represents one late-stage particle (n = 149). The black line represents the linear regression line (R^2 = 0.86, p < 2.2×10^{-16}; shading indicates the 99% confidence interval). Therefore, reference MAGs captured a

1407 representative subsample of the degraders within particle-attached communities, and predictions

1408 of chitinolytic potential were consistent between MAG- and gene-based approaches.

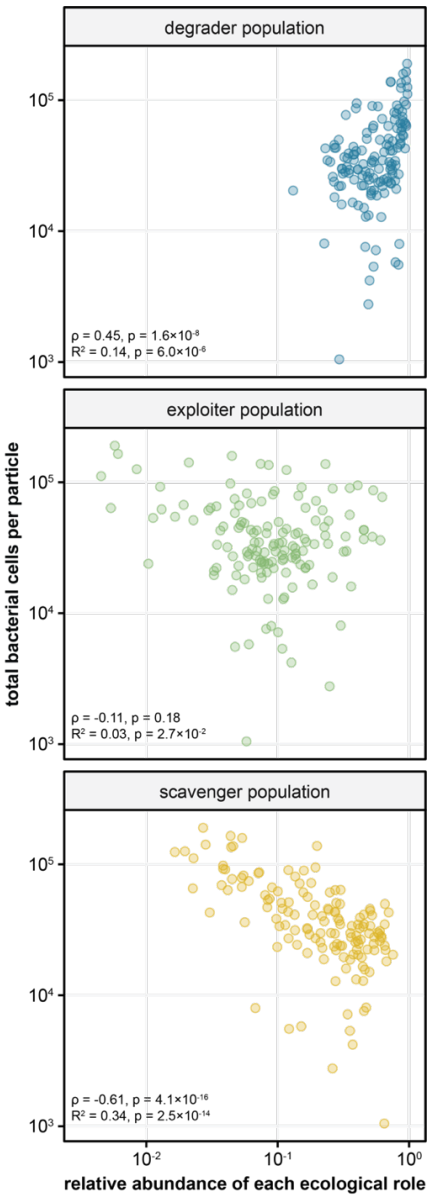


Fig. S10: The overall yield of late-stage particles is correlated with community composition. There was a strong negative correlation between the proportion of scavengers and the number of bacterial cells in late-stage communities, estimated through qPCR (yellow dots, n = 142; Spearman's ρ = -0.61, p = 4.1×10^{-16}; log-log linear regression: R^2 = 0.34, p = 2.5×10^{-14}). There was a less strong, though still highly significant, positive correlation between biomass and the proportion of degraders (blue dots; Spearman's ρ = 0.45, p = 1.6×10^{-8}; log-log linear regression: R^2 = 0.14, p = 6.0×10^{-6}), and there was no correlation with the exploiter population (green dots; Spearman's ρ = -0.11, p = 0.18; log-log linear regression: R^2 = 0.03, p = 2.7×10^{-2}).

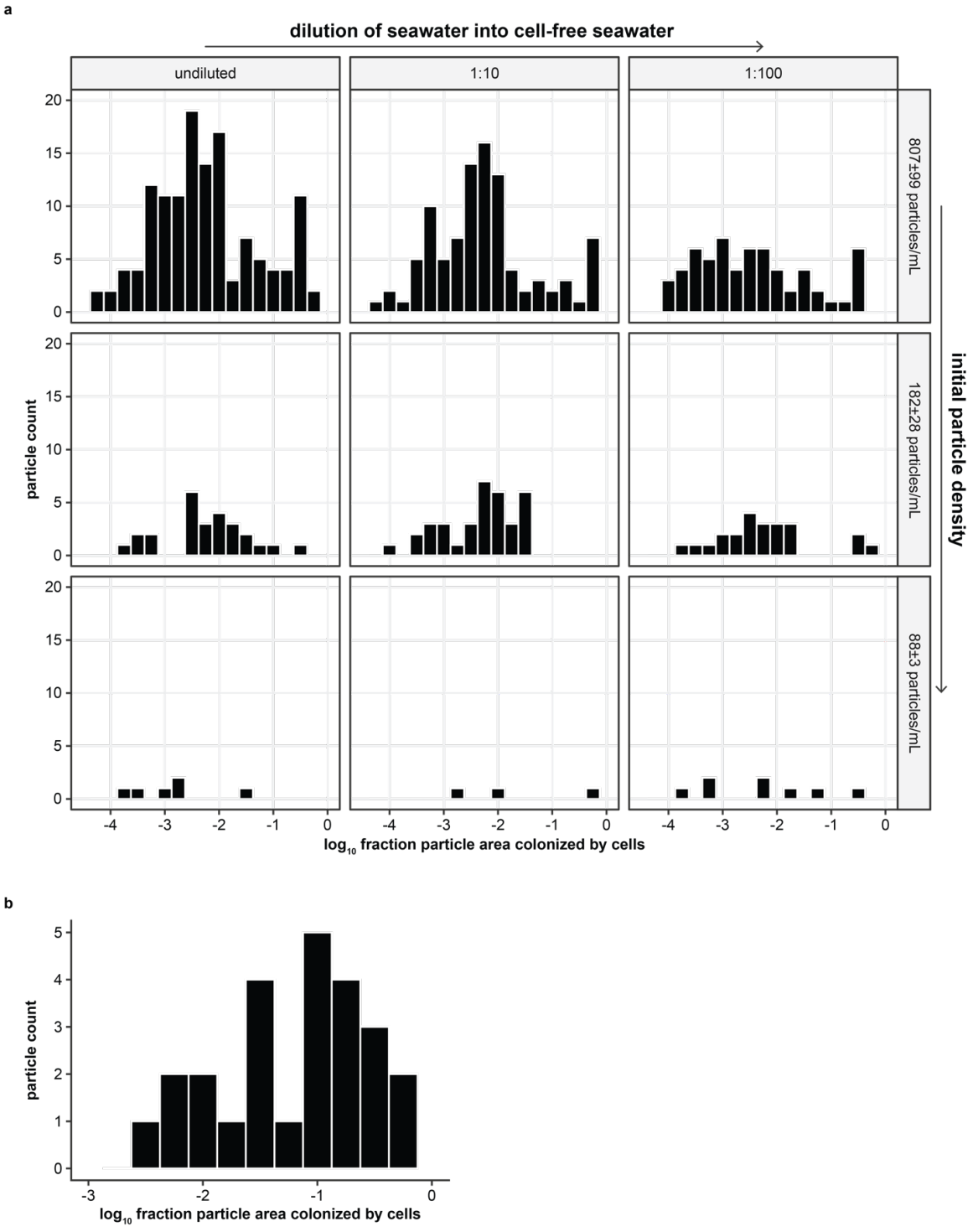


Fig. S11. The cell counts on single particles incubated in the same volume of seawater span several orders of magnitude, matching the range estimated for single particles incubated separately. All plots show the distributions of the proportion of a particle's area occupied by cells

(transformed on a $\log_{10}$ scale), estimated by visualizing particles stained with the DNA intercalating dye SYTO 9. **(a)** Cell count distributions for communities on single particles incubated in seawater together for 24 hours at various initial particle concentrations (top row: 807±99 particles/mL; middle row: 182±28 particles/mL; bottom row: 88±3 particles/mL; , ± indicates 1 standard deviation for *n* = 3 replicates throughout) and at various initial cell concentrations (left column: undiluted natural seawater; middle column: seawater inoculum diluted 1:10 into 0.2μm-filtered natural seawater; right column: diluted 1:100). **(b)** Cell count distributions for communities on single particles incubated in seawater separately and harvested after 154-167 hours (i.e. late-stage communities).

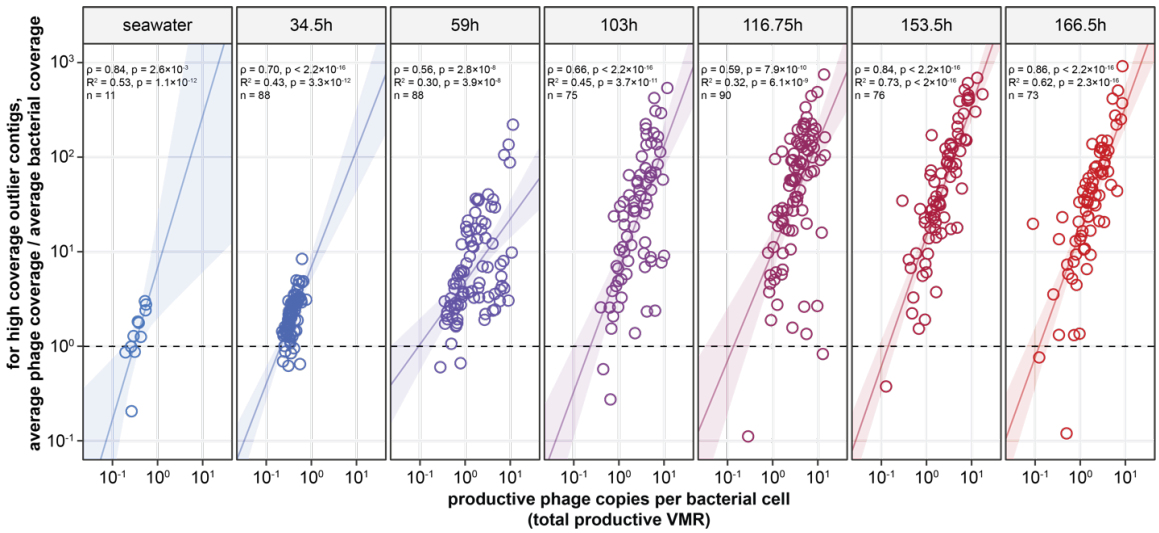


Fig. S12. Phage-derived contigs that are coverage outliers have much higher average read coverage than bacteria-derived contigs that are coverage outliers. Given that read coverage values from metagenomic data are often noisy, it is conceivable that productive phage contigs had unusually high coverage simply due to sequencing noise. However, for contigs that were high coverage outliers, the ratio for each particle (open dots) of the average coverage of phage contigs to the average coverage of bacterial contigs was often much greater than 1 (the black horizontal dashed line). Notably, these coverage ratios were overall lowest in the initial seawater inocula and rose during the incubation period, coinciding with the timescale of increasing mean productive VMRs (Fig. 4c). This indicates that phage contigs with high coverage values represented phages that were replicating more than their bacterial hosts, rather than representing contigs with randomly higher coverage values. Furthermore, there were strong positive relationships between this coverage ratio and the total productive VMR at each time point (see each subplot for significance values; the solid lines represent the log-log linear regression lines, and shading indicates the 95% confidence intervals). Thus, as expected, the particles on which high outlier phage contig coverage was indistinguishable from high outlier bacterial contig coverage were mostly those with low productive VMRs.

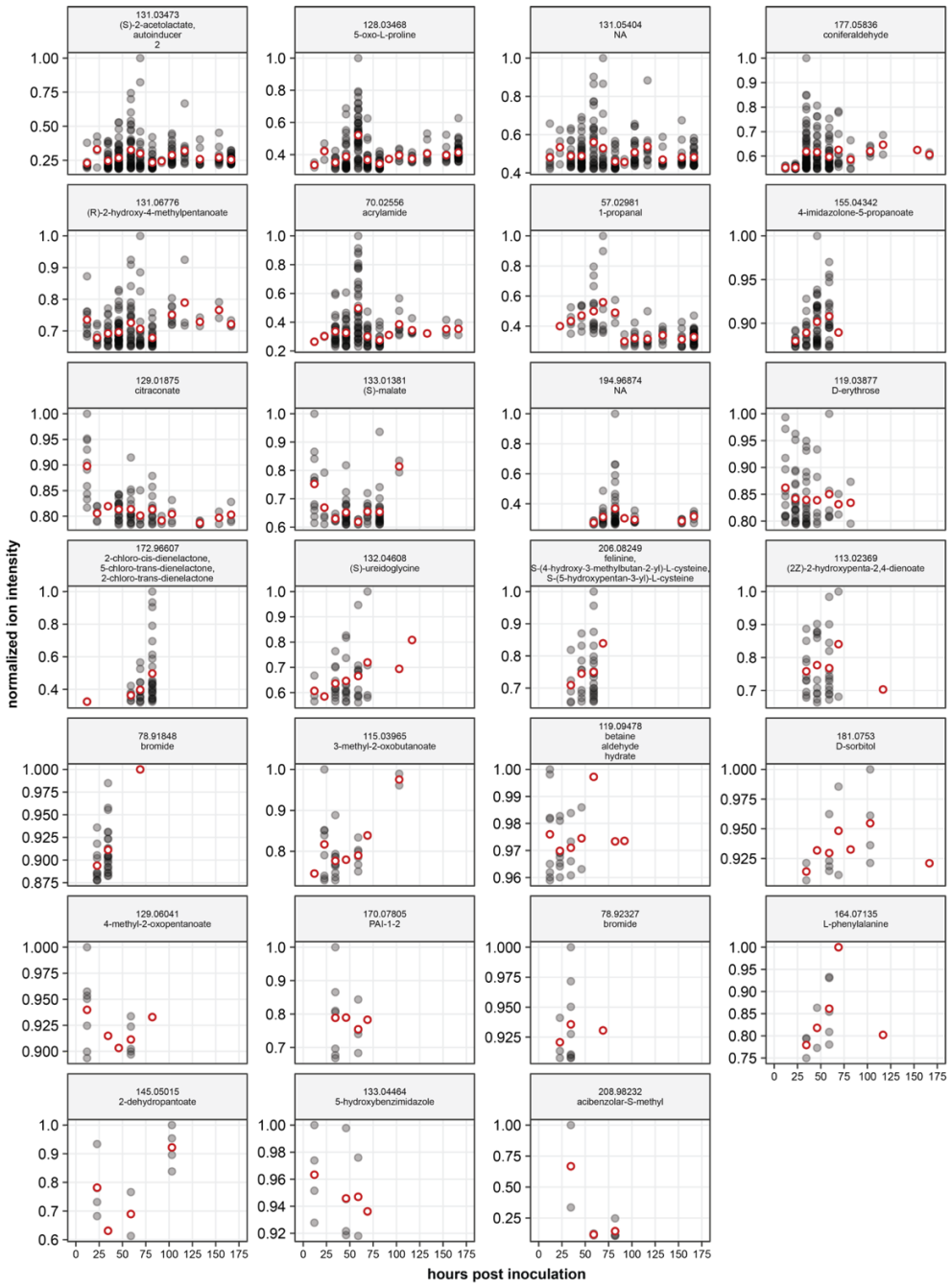


Fig. S13. Normalized intensities of individual ions over time. The normalized intensities of ions across particles harvested after varying incubation durations are shown for ions that were significantly enriched (relative to the initial seawater) on at least 10 particles (see Methods). Gray dots indicate measurements for individual particles, and red circles represent the mean normalized intensities at each time point. Panel labels include the *m/z* ratio and predicted annotation for each ion (see Methods).

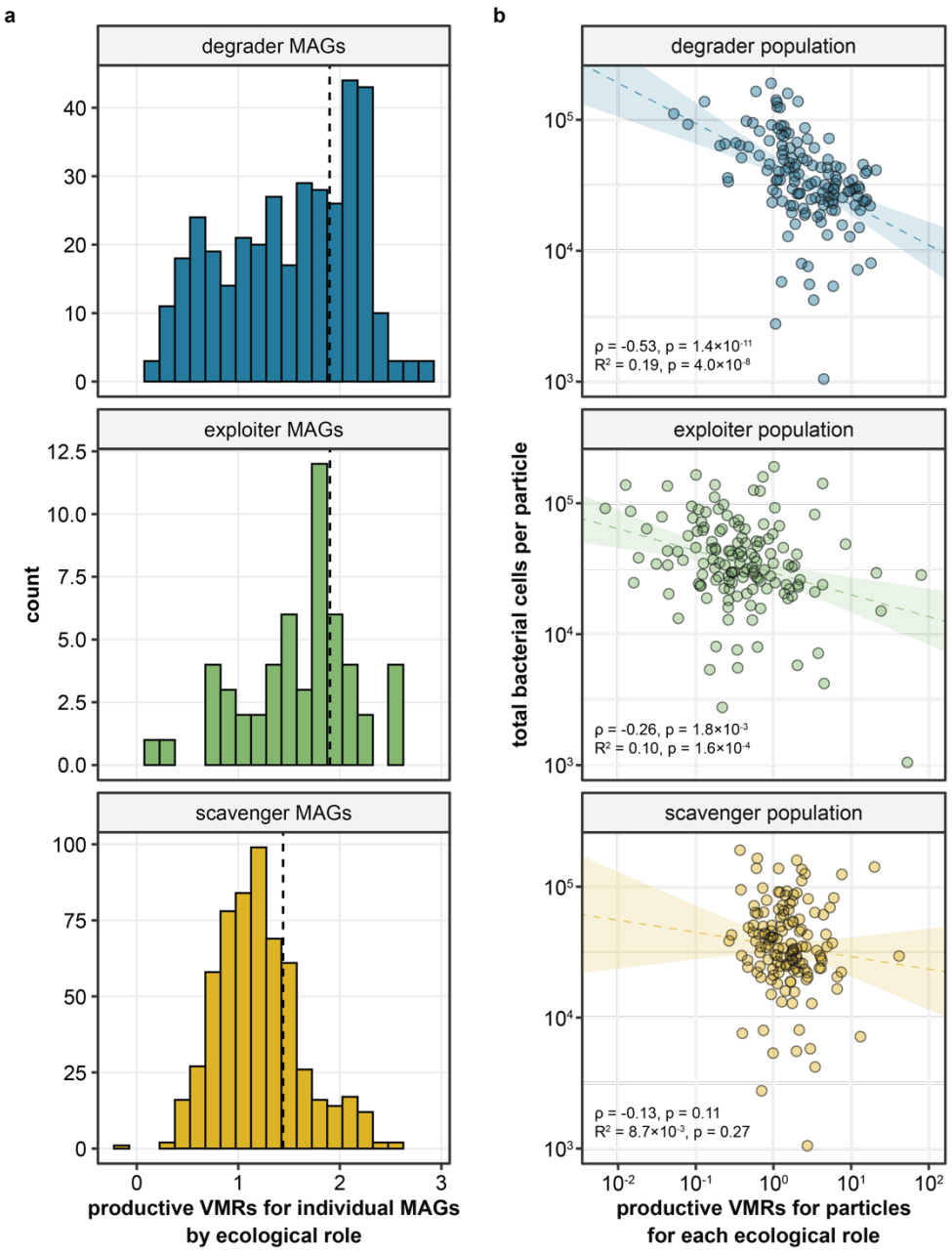


Fig. S14. The degrader population contributes significantly to particle-level productive VMRs. (a) Degrader MAGs (top panel, blue distribution) and exploiter MAGs (middle panel, green distribution) had significantly higher productive VMRs across late-stage particles than scavenger MAGs (bottom panel, yellow distribution) when phages were productive. Distributions depict the non-zero VMRs for each group of MAGs (dashed lines represent the means of each distribution, with degraders having a mean VMR of 76.8, exploiters 78.6, and scavengers 27.3; one-way

ANOVA: $F(998,2) = 47.4$, $p = 2.1\times10^{-20}$; Tukey's HSD test: degrader-exploiter $p = 0.40$; degrader-scavenger $p < 1.0\times10^{-7}$; exploiter-scavenger $p = 1.0\times10^{-7}$). When instances of VMRs equaling zero are included in the distributions, degraders had a mean VMR of 5.9, exploiters 1.3, and scavengers 2.6 (one-way ANOVA: $F(14049,2) = 42.0$, $p = 6.3\times10^{-19}$; Tukey's HSD test: degrader-exploiter $p < 1.0\times10^{-7}$; degrader-scavenger $p < 1.0\times10^{-7}$; exploiter-scavenger $p = 3.5\times10^{-2}$). This suggests that degraders overall experienced the most phage activation. **(b)** Absolute bacterial cell counts on late-stage particles (n = 142), estimated through qPCR, vs. each particle's productive VMR for the MAGs in each ecological role. Cell counts were negatively correlated with productive VMRs most strongly and significantly for degraders (top panel, blue dots; Spearman's $\rho = -0.53$, $p = 1.4\times10^{-11}$) and less so for exploiters (middle panel, green dots; Spearman's $\rho = -0.26$, $p = 1.8\times10^{-3}$), and there was no correlation between cell counts and productive VMRs for scavengers (bottom panel, yellow dots; Spearman's $\rho = -0.13$, $p = 0.11$). Dashed lines represent the log-log linear regression lines between cell counts and productive VMR (degraders: $R^2 = 0.19$, $p = 4.0\times10^{-8}$; exploiters: $R^2 = 0.10$, $p = 1.6\times10^{-4}$; scavengers: $R^2 = 8.7\times10^{-3}$, $p = 0.27$; shading indicates the 95% confidence intervals). The productive VMRs for each ecological role were also significantly different from each other, with degraders having the highest mean VMR (one-way ANOVA: $F(423,2) = 96.6$, $p = 2.9\times10^{-35}$; Tukey's HSD test: degrader-exploiter $p < 1.0\times10^{-7}$; degrader-scavenger $p = 1.7\times10^{-3}$; exploiter-scavenger $p = < 1.0\times10^{-7}$). This suggests that the effect of phage activation on particle yield was largely driven by the degrader trophic level.

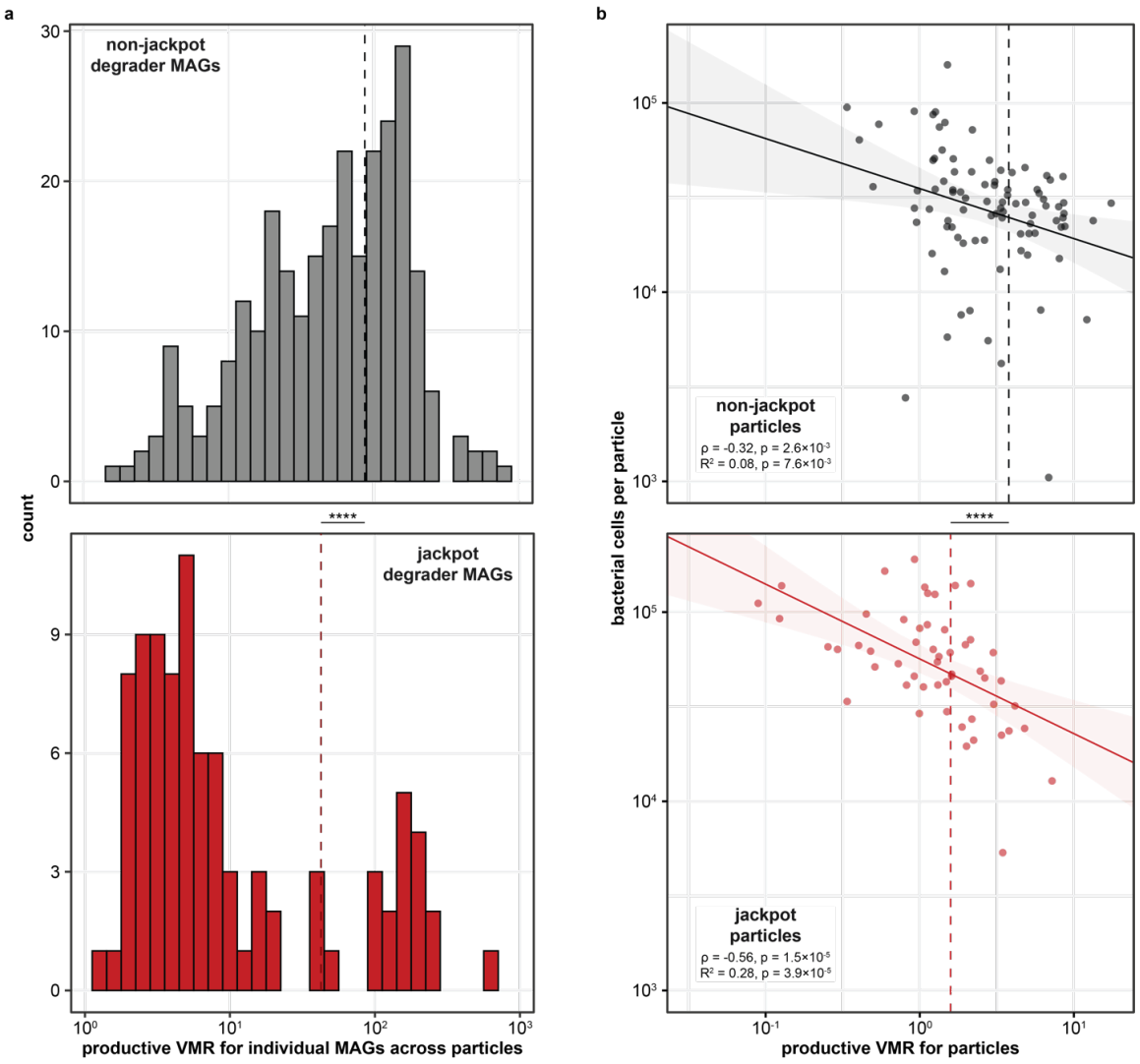


Fig. S15. The jackpot growth phenomenon is associated with less phage activation. (a) Jackpot degrader MAGs (bottom panel, red distribution) had lower productive VMRs across late-stage particles than non-jackpot degraders (top panel, grey distribution). Distributions depict the non-zero VMRs for each group of MAGs (dashed lines represent the means of each distribution, with jackpot degraders having a mean VMR of 42.4 and non-jackpot degraders having a mean of 88.0; Mann-Whitney *U* test: $p = 6.2\times10^{-15}$). When instances of VMRs equaling zero are included in the distributions, jackpot degraders still have a lower mean VMR (1.32 vs. 12.9; Mann-Whitney *U* test: $p = 1.3\times10^{-49}$). **(b)** Jackpot particles had lower productive VMRs than non-jackpot particles. Modified version of Fig. 4e in which jackpot particles (bottom panel, red dots) are shown separately from non-jackpot particles (top panel, dark gray dots). Jackpot particles had significantly lower

productive VMRs than non-jackpot particles (dashed lines represent the means of each distribution; Mann-Whitney *U* test: $p = 4.3\times10^{-8}$), even controlling for differences in biomass between these groups of particles (ANCOVA: $F(1,139) = 16.92$, $p = 4.1\times10^{-4}$, partial $\eta^2 = 0.09$). Both groups of particles showed significant negative relationships between biomass (estimated through qPCR) and productive VMR (jackpot particles: Spearman's $\rho = -0.56$, $p = 1.5\times10^{-5}$; non-jackpot particles: Spearman's $\rho = -0.32$, $p = 2.6\times10^{-3}$). The solid lines represent the log-log linear regression lines between cell counts and productive VMRs (red line for jackpot particles: $R^2 = 0.28$, $p = 3.9\times10^{-5}$; black line for non-jackpot particles: $R^2 = 0.08$, $p = 7.6\times10^{-3}$; shading indicates the 95% confidence intervals).

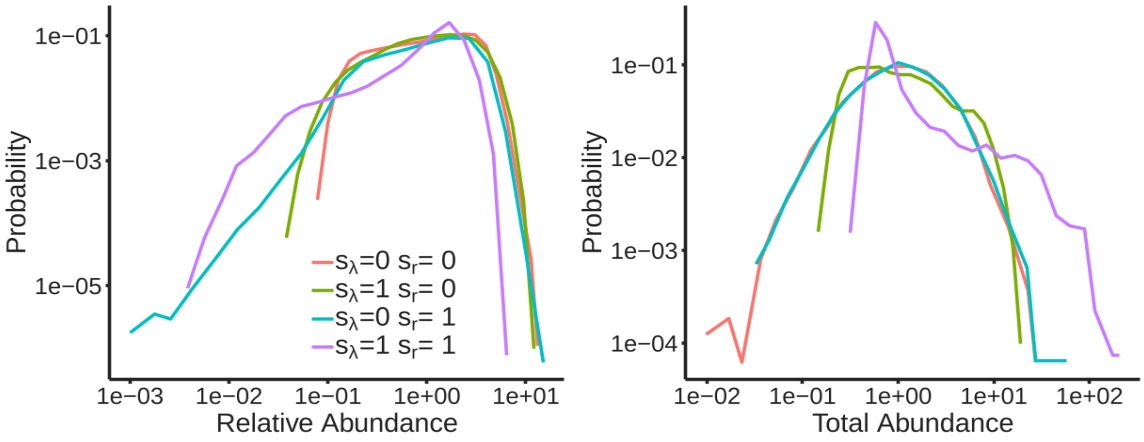


Fig. S16. Numerical simulations of mathematical model #1. The left panel shows the distribution of rescaled relative abundances averaged over MAGs. For each MAG the logarithm of the relative abundances across particles was rescaled by mean and variance, so that it had mean zero and unit variance. Lines represent averages over MAGs. Colors refer to a particular parameterization. The right plot shows the distribution of the total biomass across particles. In all the simulations we set $\bar{\lambda} = 1$ and $\bar{r} = 1$ (where the total time of the experiment was also set to be equal to 1). Different values of s_r and s_λ correspond to different colors.

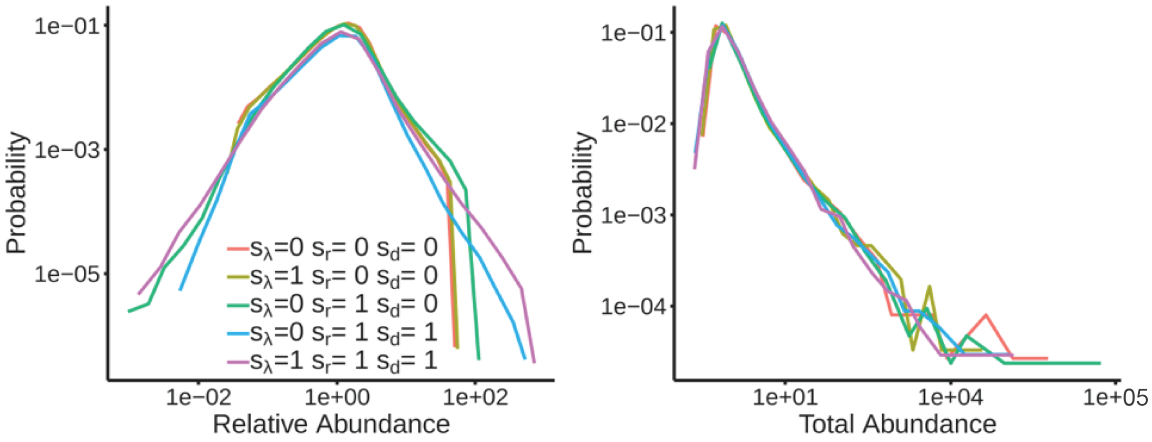


Fig. S17. Numerical simulations of mathematical model #2. The left panel shows the distribution of rescaled relative abundances averaged over MAGs. For each MAG the logarithm of the relative abundances across particles was rescaled by mean and variance, so that it had mean zero and unit variance. Lines represent averages over MAGs. Colors refer to a particular parameterization. The right plot shows the distribution of the total biomass across particles. In all the simulations we set $\bar{\lambda} = 1$ and $\bar{r} = 1$ (where the total time of the experiment was also set to be equal to 1). Different values of s_r and s_λ correspond to different colors.

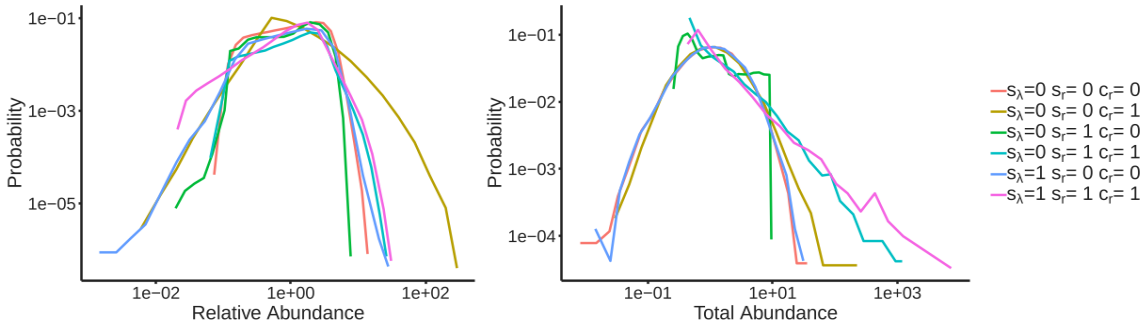
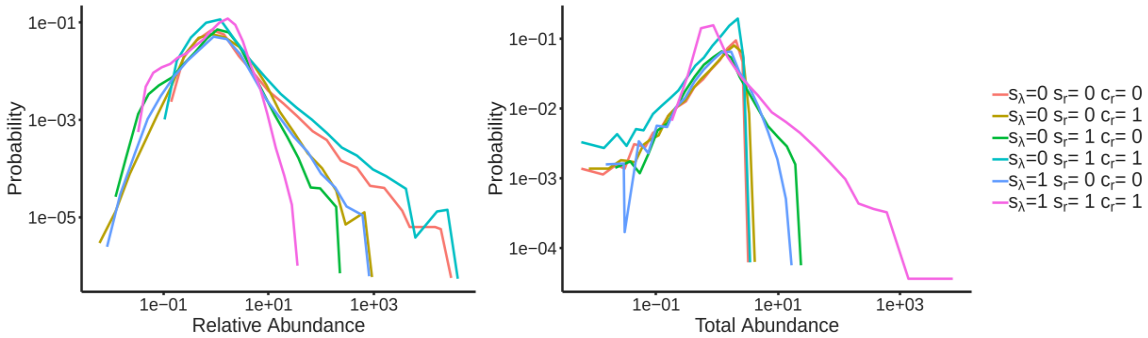


Fig. S18. Numerical simulations of mathematical model #3. Numerical simulations of model #3. The panels show the same distributions as in Fig. S16. In all the simulations we set $\bar{\lambda} = 1$ and $\bar{r} = 1$ (where the total time of the experiment was also set to be equal to 1). Different values of s_r, s_λ and c_r correspond to different colors.



1521

1522 **Fig. S19. Numerical simulations of mathematical model #4.** The panels show the same
1523 distributions as in Fig. S16. The right plot shows the distribution of the total biomass across
1524 particles. In all the simulations we set $\lambda^{d} = 1$, $\bar{\lambda} = 1$, and $\bar{r} = 1$ (where the total time of the
1525 experiment was also set to be equal to 1). Different values of s_r, s_λ and c_r correspond to different
1526 colors.

Table S1 (separate file). Metadata accompanying metagenome-assembled genomes (MAGs) from this study.

Table S2 (separate file). Metadata accompanying bacteriophage-annotated sequences from this study.

Table S3 (separate file). Metadata accompanying metabolomics performed in this study.

Table S4 (separate file). Statistics on the distributions of bacteriophage-annotated sequences in accompanying metagenome-assembled genomes (MAGs) from this study according to the predicted MAG ecological role.

Table S5 (separate file). Metadata accompanying metagenomic samples collected in this study.

Table S6 (separate file). Accession numbers and methods for the creation of custom profile hidden Markov models (HMMs) used to annotate chitin metabolism-related genes in metagenome-assembled genomes in this study.

Table S7 (separate file). Relative abundances of metagenome-assembled genomes (MAGs) in each sample collected in this study.

Supplementary References

1. C. Rinke, *et al.*, Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *Peerj* 2016, 1–28 (2016).
2. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).

3. B. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner (2014).
4. S. I. Nikolenko, A. I. Korobeynikov, M. A. Alekseyev, BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *Bmc Genomics* 14, S7 (2013).
5. D. Li, *et al.*, MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11 (2016).
6. Y. W. Wu, B. A. Simmons, S. W. Singer, MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607 (2016).
7. J. Alneberg, *et al.*, Binning metagenomic contigs by coverage and composition. *Nat Methods* 11, 1144–1146 (2014).
8. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359 (2012).
9. G. E. Leventhal, *et al.*, Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. *Nat Microbiol* 3, 1295–1303 (2018).
10. C. M. K. Sieber, *et al.*, Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 3, 836–843 (2018).
11. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25, 1043–1055 (2015).
12. D. H. Parks, *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36, 996 (2018).
13. P. A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927 (2020).
14. W. J. Kent, BLAT---The BLAST-Like Alignment Tool. *Genome Res* 12, 656–664 (2002).
15. N. Otsu, A threshold selection method from gray-level histograms. *Ieee Transactions Syst Man Cybern* 9, 62–66 (1979).
16. R. C. Team, R: A Language and Environment for Statistical Computing (2021).
17. G. Pau, F. Fuchs, O. Sklyar, M. Boutros, W. Huber, EBImage-an R package for image processing with applications to cellular phenotypes. *Bioinformatics* 26, 979–981 (2010).

18. D. Hyatt, *et al.*, Prodigal: prokaryotic gene recognition and translation.
19. S. R. Eddy, Accelerated profile HMM searches. *Plos Comput Biol* 7, e1002195 (2011).
20. J. Mistry, *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Res* 49, 1–8 (2020).
21. D. H. Haft, *et al.*, TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res* 29, 41–43 (2001).
22. K. L. Meiborn, *et al.*, The Vibrio cholerae chitin utilization program. *P Natl Acad Sci Usa* 101, 2524–2529 (2004).
23. S. Eisenbeis, S. Lohmiller, M. Valdebenito, S. Leicht, V. Braun, NagA-dependent uptake of N-acetyl-glucosamine and N-acetyl-chitin oligosaccharides across the outer membrane of Caulobacter crescentus. *J Bacteriol* 190, 5230–5238 (2008).
24. A. Bateman, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* 47, D506–D515 (2019).
25. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30, 772–780 (2013).
26. F. Madeira, *et al.*, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 47, W636–W641 (2019).
27. M. S. Datta, E. Sliwerska, J. Gore, M. F. Polz, O. X. Cordero, Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nat Commun* 7, 1–7 (2016).
28. T. N. Enke, *et al.*, Modular Assembly of Polysaccharide-Degrading Marine Microbial Communities. *Curr Biol* 29, 1528-1535.e6 (2019).
29. S. Pontrelli, *et al.*, Hierarchical control of microbial community assembly by specialists. *Biorxiv*, 1–18 (2021).
30. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
31. J. Friedman, T. Hastie, R. Tibshirani, glasso: Graphical Lasso: Estimation of Gaussian Graphical Models (2019).

1610 32. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND.
1611 *Nat Methods* 12, 59–60 (2015).
1612 33. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: Mining viral signal from microbial
1613 genomic data. *Peerj* 2015, 1–20 (2015).
1614 34. J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, F. Sun, VirFinder: a novel k-mer based tool
1615 for identifying viral sequences from assembled metagenomic data. *Microbiome* 5, 69
1616 (2017).
1617 35. A. C. Gregory, *et al.*, Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*
1618 177, 1109-1123.e14 (2019).
1619 36. N. A. Ahlgren, J. Ren, Y. Y. Lu, J. A. Fuhrman, F. Sun, Alignment-free d2∗ oligonucleotide
1620 frequency dissimilarity measure improves prediction of hosts from metagenomically-
1621 derived viral sequences. *Nucleic Acids Res* 45, 39–53 (2017).
1622 37. R. A. Edwards, K. McNair, K. Faust, J. Raes, B. E. Dutilh, Computational approaches to
1623 predict bacteriophage-host relationships. *Fems Microbiol Rev* 40, 258–272 (2016).
1624 38. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic
1625 features. *Bioinformatics* 26, 841–842 (2010).
1626 39. K. Kieft, K. Anantharaman, Deciphering active prophages from metagenomes. *bioRxiv*,
1627 2021.01.29.428894 (2021).
1628 40. M. R. Olm, *et al.*, The source and evolutionary history of a microbial contaminant identified
1629 through soil metagenomic analysis. *Mbio* 8, 1–12 (2017).
1630 41. N. Dumont-Leblond, M. Veillette, C. Racine, P. Joubert, C. Duchaine, Development of a
1631 robust protocol for the characterization of the pulmonary microbiota. *Commun Biology* 4,
1632 1–9 (2021).
1633 42. D. K. Button, B. R. Robertson, Determination of DNA Content of Aquatic Bacteria by Flow
1634 Cytometry. *Appl Environ Microb* 67, 1636–1645 (2001).
1635 43. MATLAB, *9.7.0.1190202 (R2019b)* (The MathWorks Inc.).

44. T. Fuhrer, D. Heer, B. Begemann, N. Zamboni, High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Anal Chem* 83, 7074–7080 (2011).

45. P. D. Karp, *et al.*, The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 20, 1085–1093 (2017).