Eco-evolutionary dynamics of massive, parallel bacteriophage outbreaks in compost communities

Jeroen Meijer[1*], Petros Skiadas[1], Paul B. Rainey [2,3], Paulien Hogeweg[1], Bas E. Dutilh[1,4]

1. Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands
2. Department of Microbial Population Biology, Max Planck Institute for Evolutionary Biology, Plön, Germany
3. Laboratory of Biophysics and Evolution, CBI, ESPCI Paris, Université PSL, CNRS, Paris, France
4. Institute of Biodiversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, 07743, Jena, Germany

* Corresponding author: Jeroen Meijer

Email: jeroen.meijer@protonmail.com

Author Contributions: JM wrote the paper with input from PH and BED. PS wrote code used for variant calling and advised with the variant calling analysis. PBR provided the original experimental design and ensuing data and commented on the manuscript. JM, PH, BED contributed equally to design of the analysis and interpretation of the results. JM performed the analysis.

Competing Interest Statement: The authors declare no competing interest.

Keywords: Bacteriophage, experimental evolution, microbial communities, metagenomics

This PDF file includes:

Main Text
Figures 1 to 5
Table 1

Abstract

Bacteriophages are important drivers of microbial ecosystems, but their influence and dynamics in terrestrial biomes remain poorly understood compared to aquatic and host-associated systems. To investigate this, we analyzed shotgun metagenomics datasets from ten compost-derived microbial communities propagated over 48 weeks. We found that the communities clustered into two distinct types consisting of hundreds of microbial genera, and in one community type identified Theomophage, a lytic bacteriophage representing a new *Schitoviridae* subfamily, which accounted for up to 74.3% of the total community metagenome, indicating massive viral outbreaks. We tracked molecular evolution of Theomophage and found that isolated communities were dominated by a single strain that showed little molecular evolution during outbreaks. However, when experimental manipulation allowed phages to migrate between communities, we observed transient coexistence of strains followed by genomic recombination that underpinned replacement of the ancestral strains. Additionally, when Theomophage colonized mesocosms where it was originally absent, new mutations evolved that fixed and spread to other communities. Our study describes the largest bacteriophage outbreak reported to date and reveals the spatial and temporal scales at which terrestrial bacteriophage microdiversity evolves. It also demonstrates that mixing of viral communities, which may be frequent in natural systems, promotes rapid bacteriophage evolution.

Significance Statement

Terrestrial viral ecology and evolution is an active research area, but current knowledge on soil viruses still lags behind that of other biomes, and the terrestrial microbiome harbors untapped viral diversity. This study describes parallel, massive outbreaks of a novel bacteriophage in a 48-week evolution experiment with compost-derived microbial communities. The unprecedented abundance of this bacteriophage highlights the importance of accounting for viral sequences – which may be challenging to identify and annotate – when analyzing community sequence data. Moreover, we show how dispersal accelerates the dynamics of molecular evolution on a timescale of weeks. These results advance understanding of the spatiotemporal scales at which bacteriophage eco-evolutionary dynamics play out in terrestrial biomes.

Main Text

Introduction

Through top-down control of bacterial population and metabolic reprogramming of their hosts, bacteriophages impact the structure, functioning, assembly, dynamics, and evolution of microbial communities[1–4]. Much of our original knowledge about bacteriophages in natural ecosystems comes from metagenomics studies in marine and host-associated biomes. Terrestrial environments such as soil and compost have remained relatively understudied, e.g. because technical challenges prevent application of protocols developed for other biomes[5,6]. For example, up to 90% of bacteriophage virions adsorb to the soil matrix, which reduces viral yield from terrestrial samples[5]. Moreover, terrestrial ecosystems are extremely diverse which means that less abundant genomes such as viruses, which typically account for less than 2% of assembled reads[5], might not be recovered by metagenomic sequencing[7–9].

Recently, updated sample preparation protocols specifically for high resolution metagenomics of virus-like particles ('viromics') have begun to reveal that bacteriophages are pervasive and have a broad impact in soil ecosystems[5,6,10–14]. As in other biomes, bacteriophages change nutrient cycling in soil communities[15,16] and can carry genes that are beneficial to their host[14,18]. Examples include auxiliary metabolic genes (AMGs) involved in degradation of the naturally abundant carbon polymer chitin[11,17,18], and genes that facilitate adaptation to stress from soil pollutants such as heavy metals[19].

Due to their high spatiotemporal and physicochemical heterogeneity, terrestrial biomes have been described as "representing the 'perfect storm' of diversity drivers"[6] and "evolutionary incubators"[20]. Studies in lab conditions and natural communities have shown bacteriophage diversity to be driven by coevolution with local hosts, with rapid dynamics – where infectivity and resistance are gained and lost on short timescales – driving diversification into multiple, locally coexisting phage genotypes and bacteria[21,22]. Indeed, terrestrial biomes are predicted to contain highly diverse and dynamic viral communities[6] with soil phages showing evidence of local adaptation and rapid coevolution with hosts[23–25]. However, eco-evolutionary dynamics of soil bacteriophages in complex ecosystems remain poorly characterized[6].

Besides diversity, a critical observable for understanding bacteriophage ecology and its impact on the community is abundance. An often used method to estimate viral abundance in natural communities is counting virus-like particles (VLPs) with epifluorescence or transmission electron microscopy[26]. VLP counts indicate that virus abundances in terrestrial biomes can exceed those of other biomes with estimates up to 10^9 viruses per gram of soil[6,27]. However, VLP counting can result in mislabeling of non-viral particles and is unable to detect ecologically relevant intracellular viral particles or prophages integrated in host genomes. VLP counting also overlooks important information regarding the identity and genomic functions of the measured viruses. To address these issues and gain insight into the ecological and evolutionary dynamics of bacteriophage species or strains requires a sequence-based assessment[28].

Here, we used metagenomics to quantify the relative abundance and the genomic sequence information of uncultivated viruses[29], thus allowing precise taxonomic classification. We report massive outbreaks of a novel bacteriophage detected in metagenomes from domesticated microbial communities derived from garden compost. In addition, we characterized the dynamics of molecular evolution and showed that this was fueled by migration of phages among communities.

Results

Genus-level profiling reveals stable, distinct community types

We analyzed 90 previously published metagenomes from 10 microbial communities derived from garden compost, that were propagated on cellulose and M9 medium over a period of 48 weeks (**Fig.1A**)[30]. The high complexity of compost ecosystems makes metagenome assembly difficult, potentially leading to short contigs and a large fraction of unassembled reads. To improve assembly and increase the chance of recovering rare genomes such as bacteriophages, samples from each mesocosm were cross-assembled by pooling reads from all time points[6,31]. After assembly, communities were taxonomically profiled by mapping shotgun reads from individual

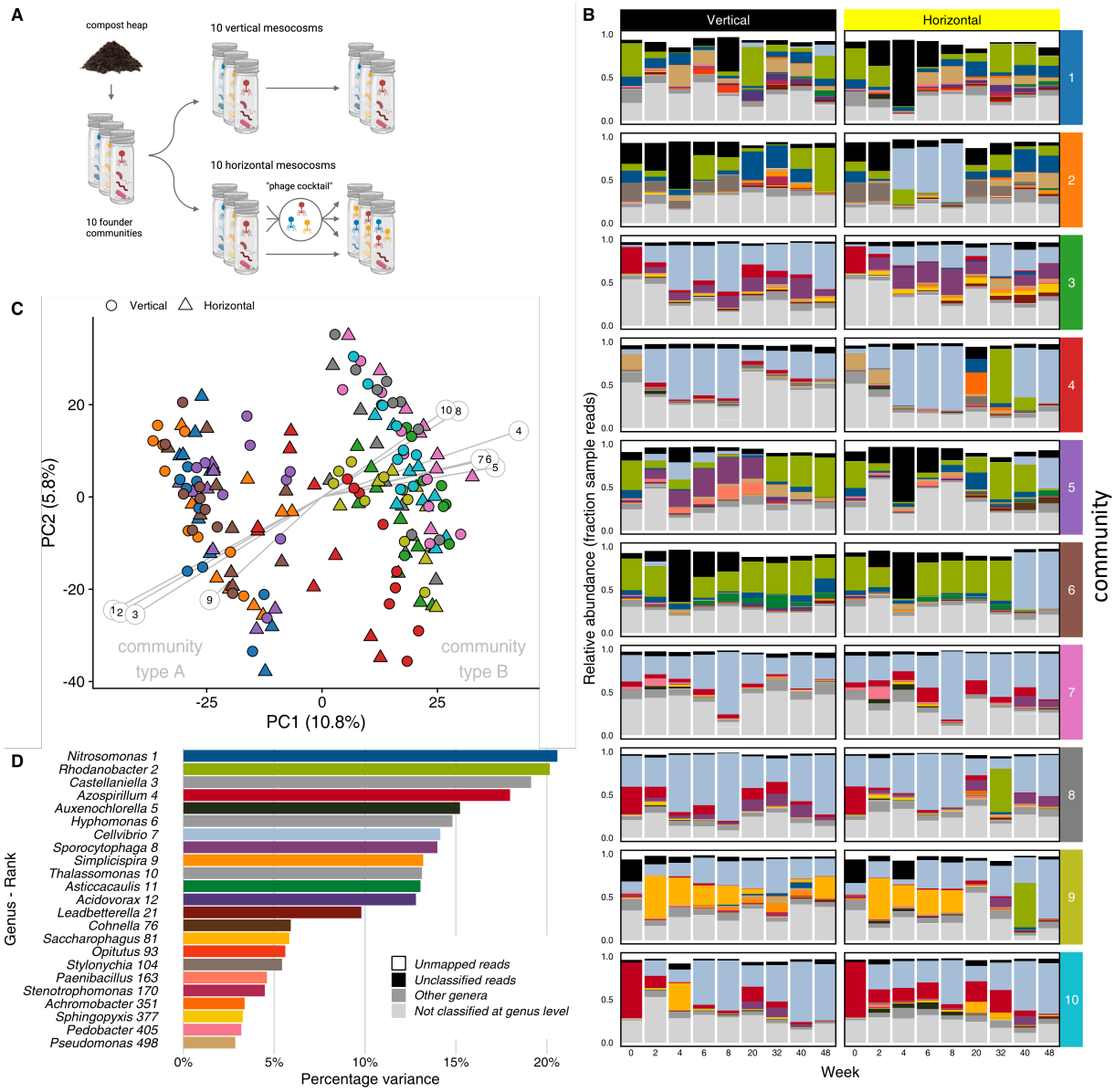


Figure 1. Compost-derived microbial communities cluster in two distinct community types. **(A)** Experimental design of Quistad et al[30]. Ten founding communities were established from gardening compost and propagated on cellulose by serial transfer (2 weeks) for 48 weeks. Vertical communities were propagated in isolation. In the horizontal treatment 'phage cocktails' were collected, pooled and redistributed between every community at every transfer (see Methods). (**B)** Relative abundance (fraction of total metagenome reads) of 20 most abundant genera, annotated with CAT[32]. Note that unannotated reads (black) often represent viral reads (see **Fig. 2**). Genus colors correspond to D. (**C)** Principal component analysis of genus relative abundances separated the 10 communities in two distinct types along the first PC (type A, rhodanobacter-dominated and type B, cellvibrio-dominated), which is robust to the horizontal treatment. Grey lines indicate the top 10 genera with highest contributions to PC1 and PC2, numbers correspond to D. For individual PCAs for H and V regimes see Fig. S8. (**D)** Amount of variance in PC1 and PC2

explained by the top 10 genera and a further 13 highly abundant genera. Numbers correspond to C. and colors to B.

time points to the corresponding cross-assembled contigs (median 96.2% of sample reads mapped to contigs, **Fig. S1**) and annotated these at different taxonomic ranks using CAT, BAT, and RAT[32] (**Fig. S2**). This allowed us to estimate abundances of taxa at different ranks. On average, 58.2% and 71.4% of reads per sample could be annotated at the genus and family ranks, respectively, and mesocosms contained 150.3 ± 36.9 genera (mean ± standard deviation, **Table S1, Fig. S3**). Based on genus-level community composition, principal component analysis partitioned the ten mesocosms into two distinct community types along the first PC: mesocosms 1, 2, 5, 6 were dominated by genera including *Rhodanobacter* and *Nitrosomonas* (hereafter "community type A"), whereas the other communities were characterized by genera including *Cellvibrio* and *Azospirillium* (mesocosms 3, 4, 7-10, hereafter "community type B", **Fig.1B-D, Fig. S4-5**). The second PC corresponded loosely to time in the experiment (**Fig. S6A**). The separation in two community types remained visible when profiling at higher taxonomic ranks (family, order, class, **Fig. S7**), and community types were stable over 48 weeks of propagation. Possibly, types A and B reflect different local sub-communities that were already established when the ten samples were taken from the compost heap. It is also possible that A and B are alternate ecological attractors that were reached during the four-week acclimatization period preceding the first sequenced time point[30] (t = 0). *Cellvibrio* and *Rhodanobacter* MAGs recovered from our samples carried genes involved in cellulose degradation[33]. *Cellvibrio* is a known cellulose degrader[34] and *Rhodanobacter* has been previously tied to cellulose degradation[34,35]. We therefore hypothesize that these two community types are assembled around different primary degraders. However, because cellulases are a highly heterogeneous group of enzymes and the presence of predicted cellulases in genomes is not always predictive of cellulolytic activity[36], this remains to be experimentally validated.

Read mapping shows massive, parallel bacteriophage outbreaks in type-A communities

We next screened total community metagenomes for viral sequences with "What the Phage", a framework incorporating twelve different virus identification tools[37] (see Methods). We recovered 140,378 predicted viral contigs, of which 33 represented predicted complete, 131 high quality (>90% complete), and 234 medium quality (50-90% complete) genomes, as estimated with CheckV (**Fig. S9A**). At 4-8 weeks into the experiment, massive outbreaks of a single contig (63,535 base pairs (bp), estimated 86.92% complete) occurred in four out of ten mesocosms (**Fig. 2A**). Closer inspection showed that this contig represents a complete bacteriophage genome (see below), which we named "Theomophage" after its place of origin, a compost heap at Square Théodore-Monod in Paris, France[30]. Theomophage outbreaks occurred exclusively in microbial community type A where they reached an abundance of up to 59,1% of total community reads (**Fig. 2A**, vertical regime). To check whether the phage was present at lower abundances in type-B communities, we mapped reads to the Theomophage sequence and inspected read coverage for all samples (**Fig. 2C;** Methods). This showed that the phage was absent or very rare in type-B samples, suggesting that it is strongly tied to the type-A communities (horizontal genome coverage: type A 69.2% ± 38.9%, type B 2.44% ± 2.18%, mean ± standard deviation, **Fig. S10**, **Table S2**).

Theomophage represents a new schitovirus subfamily

To identify the taxonomic affiliation of Theomophage we constructed a gene sharing network using vContTACT2[38] and a reference set of 19,197 bacterial and archaeal viral genomes downloaded from millardlab.org/phage-genome-dec-2022 (**Fig. 3A**). Based on shared gene

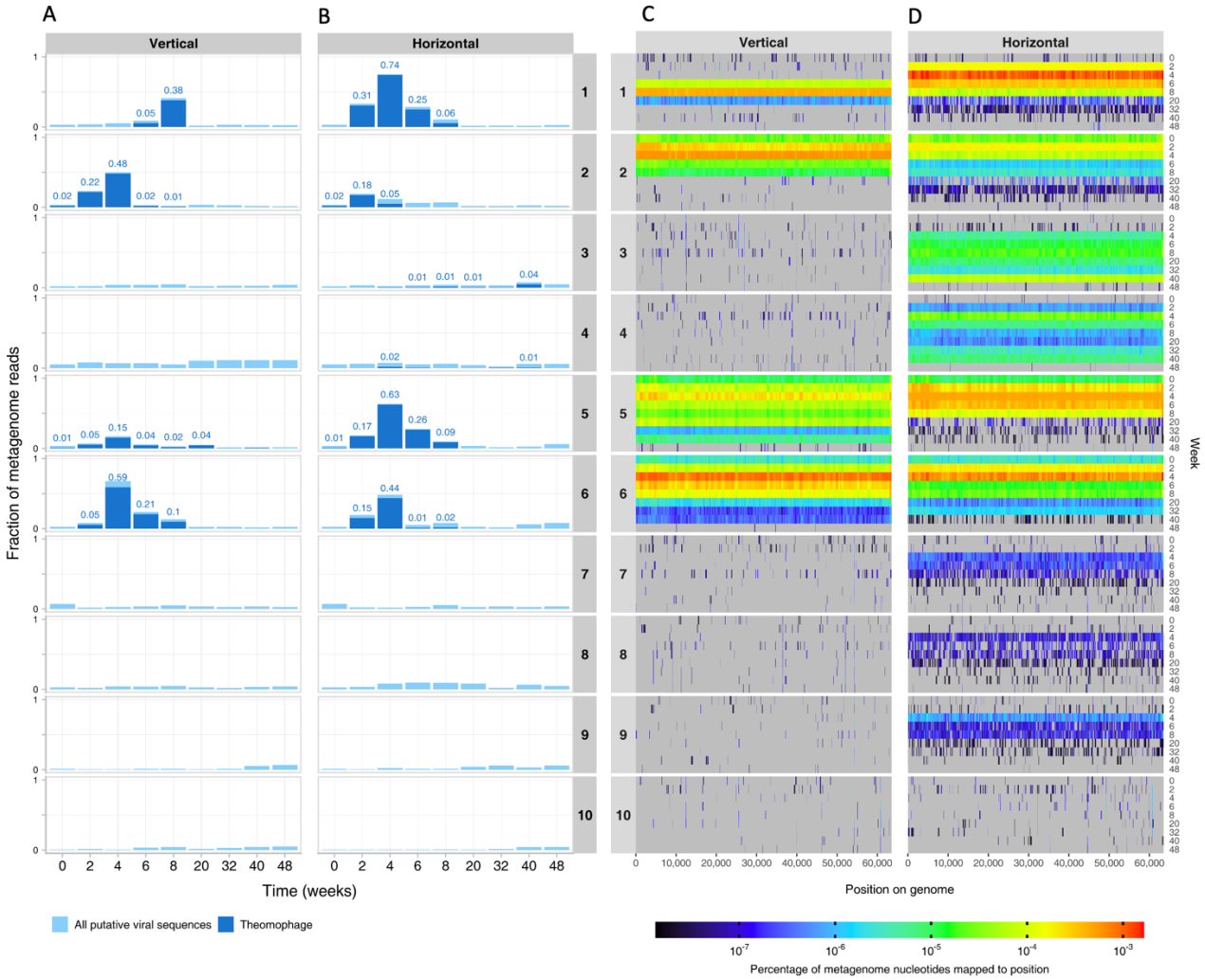


Figure 2. Massive outbreaks of a Theomophage, a novel schitoviridae bacteriophage, are linked to community type. (A) Relative abundances of predicted viral sequences (What the Phage). A new schitovirus accounted for up to 59,1% and 74,3% of total community reads in vertical and **(B)** horizontal regimes, respectively. Theomophage read factions >1% indicated with numbers. (**C)** Percentage of sample reads mapped to each position of the Theomophage genome (grey = no mapped reads). Theomophage is present in type-A communities V1, V2, V5, V6. **(D)** In horizontal communities Theomophage was detected in an additional 5 type-B communities (H3, H4, H7-9), indicating dispersion through the phage cocktail.

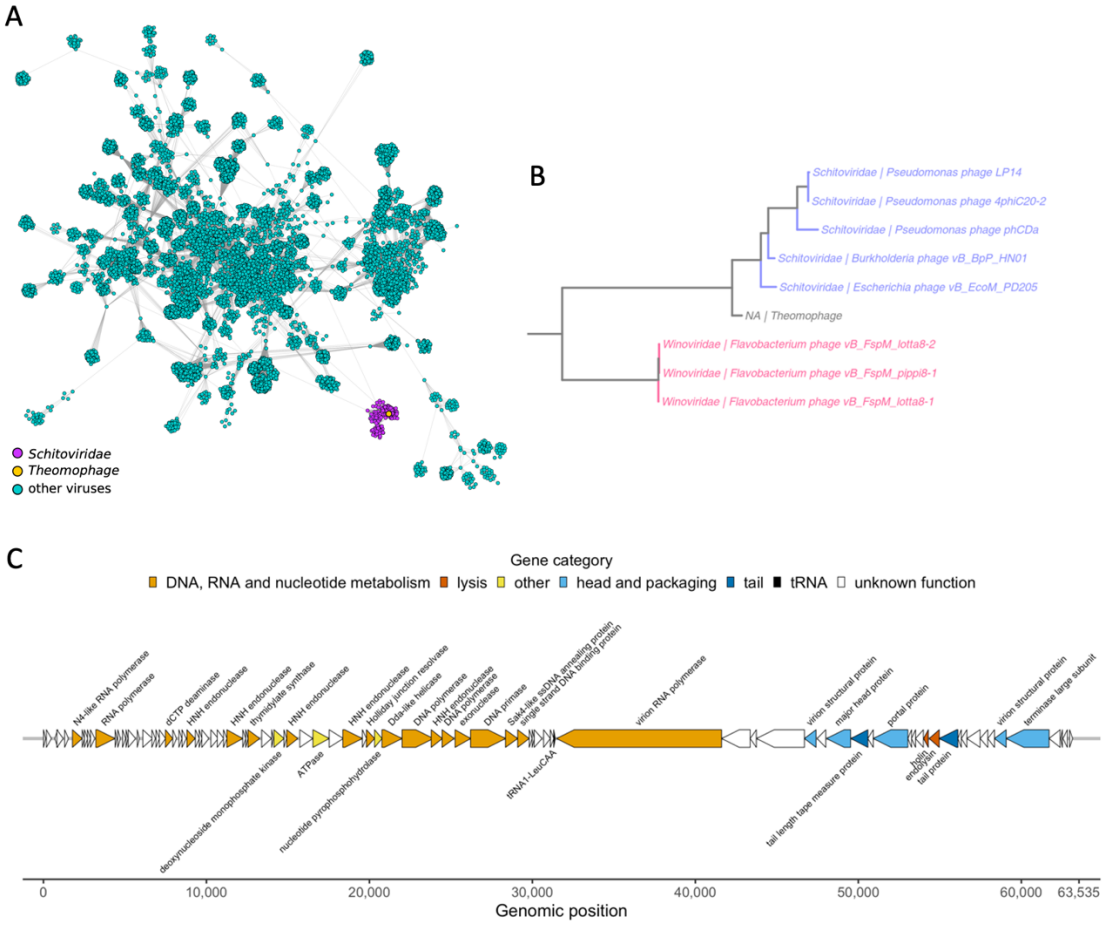


Figure 3. Phylogeny and genomic architecture of Theomophage. (A) Network based analysis of gene-sharing between 19,197 bacterial and archaeal viruses and Theomophage MAG NODE244 (yellow). Circles represent viral genomes and edges indicate the strength of connectivity based on shared protein clusters. Theomophage falls in a supercluster of Schitoviridae genomes (purple) but could not be clustered at genus rank with reference genomes, indicating no genus-level relatedness to known sequences (**Fig. S13**). Only the largest network containing Theomophage shown. **(B)** Phylogeny of TERL. The tree shown is a clade of the full tree in Fig. S11. **(C)** Architecture of the 63,535 bp Theomophage genome that encodes 88 genes and one tRNA. Gene function for 30 genes was predicted with PHROG HMM profiles.

content Theomophage clustered within *Schitoviridae*, a family of double-stranded DNA bacteriophages that until a recent update to viral taxonomy were known as *N4*-like phages[39]. *Schitoviridae* are lytic bacteriophages with linear genomes that infect gram negative bacteria, and are found globally in various habitats[39]. Individual gene phylogenies of the tailed phage marker genes Terminase Large Subunit (TerL), Major Capsid Protein (MCP), and Portal protein also place Theomophage among the *Schitoviridae*, further supporting this taxonomic assignment (**Fig. 3B; Fig. S11**).

Within *Schitoviridae*, the marker protein phylogenies did not robustly link Theomophage to known subfamilies or genera (**Fig. S12**). Similarly, in the gene sharing network Theomophage did not cluster with other sequences in vConTACT2 "viral clusters", which are concordant with ICTV genus level taxonomy[38] (**Fig. S13; Methods)**. To quantify the relatedness of Theomophage to known *Schitoviridae* we calculated the intergenomic similarity (nucleotide identity normalized for genome length) to 563 *Schitoviridae* genomes and MAGs described in two studies[39,40] using VIRIDIC[41]. With a maximum sequence identity of 9.15% to this reference set, Theomophage falls below the demarcation cutoff for subfamily rank (40%). As it clearly falls within the *Schitoviridae* family, we suggests that it represents a new subfamily[39].

Theomophage is not closely related to cellulose degrading *Schitovirus* S6

Terrestrial viruses can influence complex carbon degradation by encoding auxiliary metabolic genes[11], but may also contribute directly through enzymatic activity of virions. A recent study revealed that *Schitoviridae* phage S6 virions degrade cellulose into glucose, and depend on cellulose production by their *Erwinia amylovora* hosts for successful infection[42]. Many bacteria produce exopolysaccharides to restrict phage access to the cell surface and prevent infection, and S6 presumably co-evolved cellulase activity using a phage-encoded cellulase Gp95 as a countermeasure. Theomophage is only distantly related to vB_EamP-S6 (2.39% intergenomic similarity with VIRIDIC), but because cellulose was the only provided carbon source in the compost mesocosms studied here, this prompted a search of the gene content of Theomophage for S6 cellulases. Specifically, we queried the Theomophage genome with S6 Gp95 and potential cellulases Gp94 and Gp96 using tBLASTn and default parameters[43]. This showed no significant sequence similarity to the Theomophage genome (**Table S7**).

Annotation, genomic architecture, lifestyle

We predicted 88 genes and a single tRNA (Leu aag) organized in two large blocks on opposing strands, with functionally related genes clustering together (**Fig. 3C**; **Table S3**). Theomophage has the characteristic genome architecture of *Schitoviridae*, which contains three RNA polymerases including the hallmark large virion-associated RNA polymerase (vRNAP) and n4-like RNA polymerase. The 63,535 bp genome size of Theomophage falls within within the 59-80kb range reported for this family[39]. We suggest that the genome is complete, since the contigs that were assembled from several independent mesocosms are nearly identical (46/54 assembled Theomophage MAGs were 63,535 bp length, mean 99.98% identical), were circularly permuted (**Fig. S14-15**) and have direct terminal repeats. These results are consistent with assembly of shotgun reads derived from concatemers that form during the replication phase of a phage with a linear genome, or linear genomes with direct terminal repeats[39]. Next, we used BACPHLIP to predict Theomophage lifestyle[44]. This suggests that it is a lytic phage (score 0.9625), in accordance with all known *Schitoviridae* phages[39].

Theomophage colonizes non-native type-B communities via the horizontal virome transfer protocol

The clear partitioning in community types raised the question as to whether Theomophage, which was assembled repeatedly from type-A samples but not from type-B, could also establish itself in type-B communities. The unique experimental design of Quistad *et al* allowed us to address this question. So far, we have discussed results from 10 mesocosms that were propagated in isolation and formed the vertical or "V-regime" of the experiment. A second set of 10 mesocosms was derived from the same 10 founder communities and formed the horizontal or "H-regime" (**Fig. 1A**). These mesocosms were propagated in the same way as in the V-regime, with the addition that at each transfer 1 ml size-filtered fractions (<0.2μm) were collected, pooled, and redistributed between all 10 horizontal mesocosms[32]. The distribution of these "Selfish Genetic Element (SGE) cocktails" allowed migration of bacteriophages and other SGEs between communities – a treatment that repeatedly exposes SGEs to new hosts, potentially affecting their eco-evolutionary dynamics. Moreover, this manipulation allowed us to question the extent to which Theomophage and its potential host(s) were community-type specific.

Assembly and taxonomic profiling of the 80 metagenomes of the horizontal treatment confirmed that community types were consistent between the V- and H-regimes (**Fig. 1AB, Fig. S8**). We then assembled and profiled SGE cocktails from weeks 2, 4, 6 and 8 (i.e., all that were available[32]), which confirmed the presence of Theomophage in all cocktail samples where it accounted for up to 3.3% of cocktail reads (**Fig. S16**). Next, we searched contigs assembled from horizontal communities for viral sequences. On average, more viral sequences were recovered from horizontal compared to vertically transferred communities, suggesting that the horizontal treatment increased viral diversity of communities (**Fig. S9).**

In addition to the four native type-A communities, Theomophage was also detected in 5 out of 6 type-B communities in the horizontal regime, where it was present up until week 40 (H2, 3, 7-9; **Fig. 2BD**). The phage could thus establish itself in type-B communities. However, the abundances observed in the type-B community background were an order of magnitude lower compared to type-A communities, with a maximum of 4.5% of community reads (**Fig. 2B**). This indicates that while Theomophage could establish itself, it was less successful than in the type-A communities. There are several possible reasons: for example, it may infect the same host in both community types, but the host might have a lower abundance or higher resistance to Theomophage infection in type-B communities. Another possibility is that the original host is not available, and that Theomophage switched to infecting a sub-optimal host in these communities. Finally, since metagenomes only show relative abundances, it is conceivable that the relative abundance of the phage appears lower while its absolute abundance is similar in both community types, if type-B communities had higher biomass overall.

In the native type-A communities, Theomophage outbreaks were even larger in the horizontal than in the vertical treatment, with an outbreak of 74.3% of metagenome reads in community H1 **(Fig. 2A)**. Additionally, in native communities H1 and H2, the outbreak occurred two to four weeks earlier compared to the vertical regime. This indicates that Theomophage, by migrating via the cocktail to other mesocosms, was able to create earlier and more severe outbreaks. In community H10 the phage was not detected.

Host prediction and community impact

To elucidate Theomophage's ecological role we attempted to assign it to a host. All currently known *Schitoviridae* bacteriophages infect proteobacteria[39], but beyond that little information regarding the possible host is provided by its taxonomic affiliation. First, we searched the IMG/VR crispr database with blastn with using the Theomophage genome as a query[45]. This returned no matches. Next, we used iPHoP[46], a machine-learning framework for genus-rank phage-host prediction that integrates multiple tools[47–50]. Using the default minimum confidence score of 90 (corresponding to a 10% estimated false discovery rate) iPHoP did not predict a host. The best match returned an iPHoP confidence score of 81: Candidatus *Saccharibacteria* bacterium UBA5150 (**Table S5-6**), a nanobacterium recovered from a soil metagenome. While multiple *Saccharibacteria* sequences are present in the compost communities and we recently discovered that these bacteria were also transferred in the SGE cocktail[33], *Saccharibacteria* belong to a phylum very different than that of the expected *Proteobacteria*, which would mean that Theomophage would have undergone a major host switch. Moreover, saccharibacterial abundance patterns did not correspond with Theomophage presence across communities **(Fig. S17)**, and we deem it an unlikely host prediction. Interestingly, the top-scoring match of Prokaryotic virus Host Predictor[50] (one of the tools used by iPHoP) is a *Nitrosomonadaceae* bacterium (GCA_014237605.1, PHP-score 1445.116). *Nitrosomonas* is a defining genus for the native type-A communities **(Fig. 1D)** and present in all type-A communities, but notably also in type-B communities H3 and H4, where Theomophage invaded and reached a relative abundance of up to 4.5% in the horizontal protocol (**Fig. 2B, Fig. S18**).

We next investigated the impact of Theomophage outbreaks on the communities by analyzing shifts in microbiome composition. Specifically, we compared the change in genus abundances before and after outbreaks to samples with a similar interval but not spanning a Theomophage outbreak. Outbreaks did not cause greater shifts in community composition (**Fig. S6B**), though some individual genera were strongly impacted. *Nitrosomonas* was one of several genera impacted negatively by the outbreaks (**Fig. S19**), which further suggests it is a possible host. We thus predict that *N. europaea* is a possible Theomophage host, although this remains tentative and awaits further validation.

Isolated communities are dominated by a single Theomophage strain and display little molecular evolution

While Theomophage successfully invaded five out of six communities of its non-native type B, its lower relative abundance in these communities suggests that it was less successful in this new context. We hypothesized that prolonged existence in a new community background and possibly host switching would require adaptation. To investigate any evolutionary dynamics during the 48-week experiment and pre-existing diversity, we searched for polymorphisms in Theomophage by mapping reads from each sample to the reference Theomophage genome (see Methods). In total we identified 76 Single Nucleotide Polymorphisms (SNPs) that reached frequencies >10% in the population based on read mapping (**Fig. 4AB, Fig. S20**). The majority of SNPs (80.3%) exclusively appeared following invasion of type-B communities H3 and H4 (**Fig. 4AB, Fig. S20-22**), suggesting adaptation to the new community background. In addition, a group of 9 SNPs was detected in both type-A and type-B communities, with frequencies near fixation in founding communities 1, 5 and 6 (**Fig.4A-5A**). These results indicate evidence that new mutations arose during the course of the experiment and standing genetic diversity in the Theomophage founder populations, respectively.

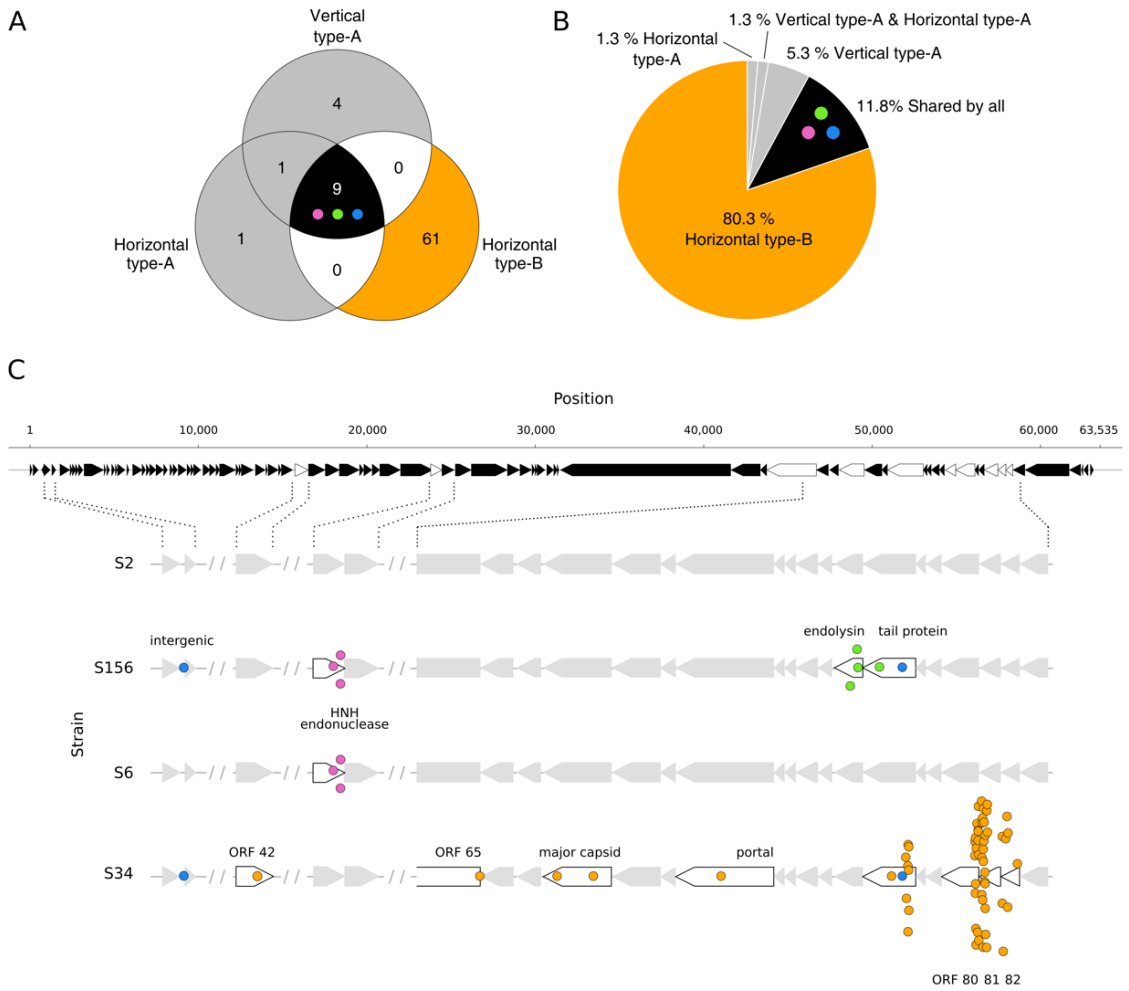


Figure 4. Single Nucleotide Polymorphisms in Theomophage populations. (A, B) A total of 76 SNPs were detected across type-A and type-B communities in the vertical and horizontal regimes. The vast majority (80.3%) was only detected in the novel type-B community background, following invasion by Theomophage. **(C)** Genomic positions of SNPs that differentiated strains S156, S6 and S34 from reference (S2). S156 differed by nine SNPs (blue, pink, green, see Table 1). During the experiment S6 and S34 evolved, most likely by recombination of S2 and S156. S34 also fixed 61 new SNPs (yellow) that were mostly located in tail protein and ORFs 80, 81 and 82 of unknown function (**Table S3**). For visual clarity variants are jittered in the y-direction. Colors correspond to Fig. 5A and Table 1 and are based on similarity of allele trajectories over time (see Fig.5A, Methods).

Table 1. Position and effect of nine genomic variants defining 4 strains. Nine SNPs that distinguish strain S156 from S2. Colors match Fig. 4 and 5. Inferred genotypes for strains S2, S156, S6 and S34 are indicated.

Position	ORF	Predicted function	Ref (S2)	Alt (S156)	Effect	S6	S34	Color
1,276	-	-	T	G	Intergenic	T	G	Blue
24,224	0051	HNH endonuclease	T	C	Missense	C	T	Pink
24,375	0051	HNH endonuclease	G	A	Synonymous	A	G	Pink
24,378	0051	HNH endonuclease	T	C	Synonymous	C	T	Pink
54,689	0076	endolysin	T	C	Missense	T	T	Green
54,853	0076	endolysin	T	C	Synonymous	T	T	Green
54,856	0076	endolysin	C	T	Synonymous	C	C	Green
55,320	0077	tail fiber protein	A	C	Missense	A	A	Green
55,814	0077	tail fiber protein	C	T	Synonymous	C	T	Blue

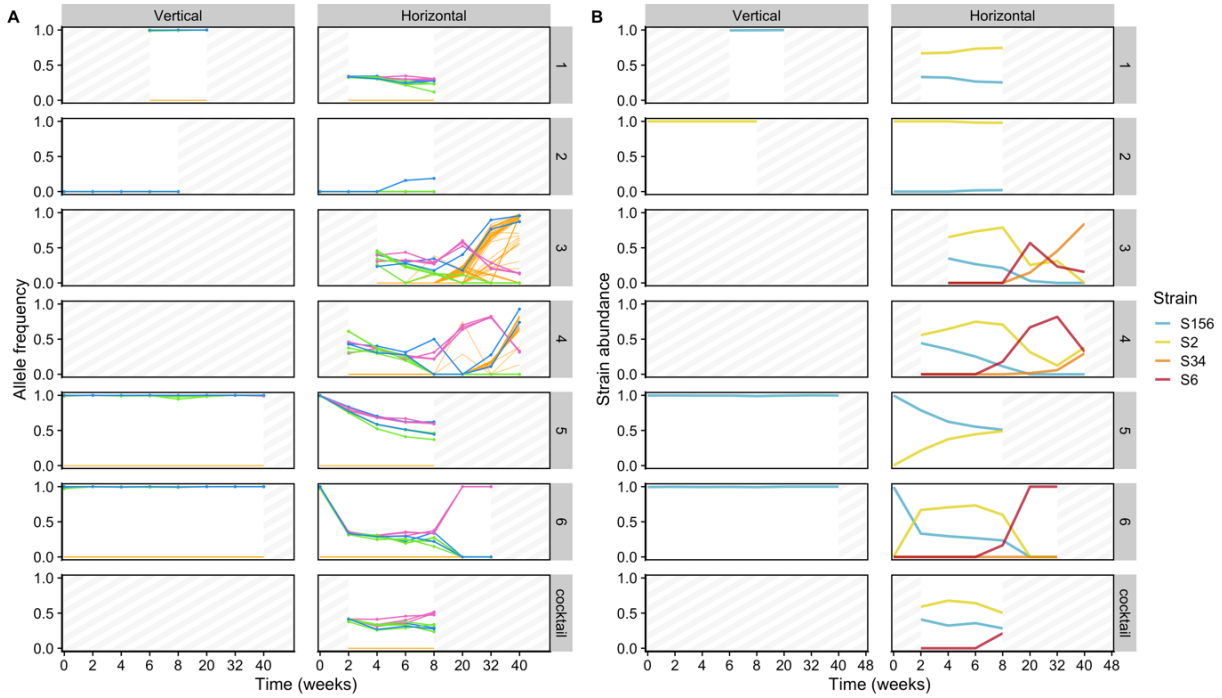


Figure 5. Dynamics of molecular evolution. (A) Allele frequency trajectories of SNPs that differed from the reference strain S2. Nine SNPs (blue, pink, green, see Table 1) were present in founding communities 1, 5 and 6, and 61 new variants (orange) that appeared first in week 8 in community H3 and spread to H4. **(B)** Simplest scenario of strain dynamics compatible with SNP trajectories in A. Theomophage populations in type-A communities were initially monomorphic for all variant positions, meaning they were composed of a single strain: S2 in community V2, which was identical to reference, and S156 in communities V1, V5, and V6, which differed from the reference by 9 SNPs as indicated by frequencies of ~1. In V-regime communities, no new variants emerged that fixed in the population during the 40 weeks over which the bacteriophage was detected. In the H-regime communities, S2 and S156 both spread via the cocktail and cross-invaded type-A communities H1, H2, H5, and H6. Both also invaded type-B communities H3 and H4, where the phage was not previously present. After week 8 in communities H3, H4 (type-A), and H6 (type-B) the trajectories of the nine SNPs that differentiate S2 and S156 became uncoupled, showing that they were no longer genetically linked which suggests genomic recombination between S2 and S156. This led to new strains S34 and S6. Additionally, 61 new SNPs (orange) appear in S34. Variants are colored based on similarity of trajectories (see Methods). Shading indicates samples without sufficient genome coverage for variant calling (>30% of the Theomophage genome with >10x read coverage) or where Theomophage was not detected. Samples from starting communities 7, 8, and 9 are not shown for this reason, although Theomophage was sparsely detected there (see Fig. 2).

Next, we assessed Theomophage strains by linking individual variants to recover haplotypes and estimate relative abundances. Unless SNPs are located on the same read, resolving genetic linkage from variants called from shotgun sequencing is challenging, as the number of strains is unknown and even SNPs with similar intermediate frequencies can potentially belong to different strains. Several bioinformatic and experimental techniques have been developed to address this issue[51]. Here, for resolving genotypes we instead exploited the observation that many samples were monomorphic for all variant sites (*e.g.,* have frequencies close to 0 or 1), indicating genetic linkage of these alleles. For example, the frequency of all 76 variants was close or equal to zero in community V2 at t = 0, indicating that this population consisted of a single strain that was identical to the reference genome (hereafter strain S2, **Fig. 4C-5C**). Similarly, founding communities V1, V5 and V6 were also monomorphic, but contained a different strain with nine SNPs in the tail fiber protein, HNH endonuclease, endolysin and an intergenic position (**Fig. 4C; Table 1**; hereafter strain "S156"). Having defined strain S2 and S156, it becomes possible to analyze their eco-evolutionary dynamics during the time course of the experiment. In the vertical communities, both S2 and S156 underwent large fluctuations in population size, but no new mutations emerged that fixed in the population during the 40 weeks over which the phage was detected (**Fig. 5A)**, although 6 SNPs provide evidence for limited, pre-existing microdiversity within S156 that was community-specific (**Fig. S21**). This suggests that both strains are well adapted to their hosts in the native communities in the V-regime.

Strain migration results in transient coexistence, followed by recombination and strain replacement

Next, we investigated the dynamics of S2 and S156 in the horizontal regime. Intermediate frequencies of the nine SNPs in the cocktail and native type-A communities indicate that strains S2 and S156 cross-invaded between these communities, with the higher relative abundance of S2 suggesting that it had a competitive advantage over S156 (**Fig. 5AB**). Both S2 and S156 also colonized non-native type-B communities H3 and H4, where they coexisted for up to eight weeks. From week eight onwards, the trajectories of the nine SNPs that differentiate S2 and S156 strongly diverged, indicating they were no longer genetically linked. They followed three different trajectories (pink, blue and green), with SNPs in closely located positions on the genome, often within the same gene, typically following similar trajectories (**Fig. 4-5**, Table 1). For example, in community H6 where initially strain S2 was dominant, from week 20 onwards the three alleles in HNH endonuclease (pink) from strain S156 were fixed in the population, while the blue and green alleles from S156 were lost. The resulting population (after 20 weeks) was monomorphic for these nine SNP positions, indicating that the population consisted of a new strain (S6) with a genome that combined alleles from both S2 and S156 (Table 1). This indicates that after at least eight weeks of coexistence the two strains recombined, with the new chimeric strain outcompeting the two founding strains.

A similar pattern of transient coexistence followed by recombination and strain replacement by the newly formed genotype was observed in communities H3 and H4. Here, the resulting genotype (S34, table 1) appeared first in H3 and was formed when two alleles from strain S156 (TerL gene and intergenic position, in blue) fixed in the population, while the remaining seven S156 alleles were lost (endolysin and HNH endonuclease, green and pink). This new genotype S34 then spread to H4, and a more sensitive SNP analysis also indicated invasion of H6 at week 32 (Methods, **Fig. S22**). The divergent trajectories of pink alleles in communities H3 and H4 suggest

that before the rise of S34, recombinant variant S6 also invaded there from H6 via the SGE cocktail (**Fig. 5**).

Accompanying the rise and spread of S34, the 61 SNPs observed exclusively in type-B communities mentioned earlier (yellow) appeared in H3 and H4, with their trajectories indicating they are located on the S34 genome. Most of these variants were in tail protein and three adjacent genes of unknown function that are predicted to encode structural proteins, based on their genomic location (ORFs 80-82; **Table S3-4, Fig.3C**). Additionally, we identified four missense SNPs in the Major Capsid Protein and Tail Protein among these variants.

The 61 new variants that appeared in H3 and H4 followed colonization of the new community type B, suggesting that they may be adaptive. As they were not detected in any sample before their appearance in H3 and H4, these variants were either so rare in the Theomophage founding populations that they fell below the metagenomic detection threshold, or alternatively, truly absent from the Theomophage populations. In the latter case, they could be the result of recombination with another bacteriophage genome, which seems likely given the novel variants were mostly located in close proximity of each other. To test for recombination we queried all assembled contigs with BLASTn for ORFs 80-82[43]. This resulted in no significant matches, apart from Theomophage MAGs and genome fragments.

Another possibility is that these SNPs evolved *de novo*. Many phages carry diversity-generating retroelements (DGRs), domesticated retroelements that use error prone reverse transcription to generate hypervariation in target genes[52]. DGRs are suggested to help phages evolve to new hosts or co-evolve with their host as they mainly target structural proteins involved in host attachment. Since the majority of SNPs was detected in genes that are predicted to encode structural proteins, the Theomophage genome was interrogated for DGRs using MyDGR[53]. No reverse transcriptase or DGRs were found. The source of these mutations thus remains unclear.

Summarizing, coexistence of the two strains in native communities (H3, H6) lead to at least two recombination events during the experiment. Due to the high abundance of Theomophage, coinfection by different strains seems likely, contributing to their evolution. In the new ecological background (H3 and H4), further molecular evolution occurred that included 61 new SNPs.

Discussion

In this study we have reported massive outbreaks of the novel schitovirus Theomophage, which reproducibly occurred in laboratory-cultured compost communities after 4-8 weeks. The extreme abundance of the phage reads in the metagenomic datasets are, to our knowledge, an unparalleled observation. While viruses typically comprise a small part (~5%) of total metagenome datasets[54], previous studies showed that bacteriophages sometimes account for a considerable fraction of DNA sequence reads, even though their genomes are relatively small compared to those of cellular organisms. For example, crAssphage accounted for up to 22% of sequencing reads in total community shotgun metagenomes from human stool samples[55], and in another study, viruses accounted for 4-17% of the total community metagenome reads, also from human stool[56]. Similar numbers were found in cheese starter cultures, simple microbial communities that can consist of as few as two bacteria and three associated phages, where a single lysogenic *Streptococcus* phage accounted for approximately 20% of the reads in a total community shotgun metagenome[57]. Here we report that a single bacteriophage can reach an abundance of 74,3% of the total reads in complex compost communities.

Bacteriophage abundance is influenced by many factors, and several plausible mechanisms for the abundance of Theomophage are possible. For crAss-like phages, an unconventional infection mode has been suggested as the mechanism driving the high abundance and long persistence with their host[58]. A pseudolysogenic or carrier state-like phenotype was observed for crAss001 in dynamic subpopulations of sensitive and resistant *Bacteroides* hosts, with delayed release of phage progeny by infected cells. Such non-canonical infection cycles of virulent phages that challenge the lytic/lysogenic binary have been reported for a wide range of viruses[59–64]. It is possible that a similar non-lethal infection mechanism drives the high abundance of Theomophage in the compost-derived communities. A continual shedding of viral particles by a subpopulation of infected host cells could lead to the accumulation of phage particles in mesocosms during the two-week transfer cycle. In addition, changes in the relative abundance of sensitive and resistant subpopulations could explain the persistence at lower abundances of this lytic phage following massive outbreaks. In all communities, the phage ultimately dropped below the detection limit.

Potential biases

While our results support a massive Theomophage outbreak in parallel compost communities, several potential biases could affect the observed abundance in the metagenomic sequencing data. First, metagenomics can be biased for specific genomes, including biases in library preparation or sequencing[65,66]. Notably, rolling circle amplification[67] was not used to create the current dataset. Second, biases in the Nextera XT kit used in library preparation may result in reduced coverage of low-GC regions[68]. However, as Theomophage has a low GC content (34%) compared to the mean GC of the assembled metagenomes, this bias would result in an underestimation rather than overestimation of relative abundance. Third, many tailed bacteriophages bind exopolysaccharides[69], and it is possible that Theomophage binds to the cellulose paper (the sole carbon source), causing it to trap virions and thus inflating its estimated abundance.

Eco-evolutionary dynamics

We observed rapid evolution involving recombination and the spread of 61 new mutations in Theomophage populations following its invasion into non-native type-B communities, but little to

no molecular evolution in native type-A communities. Increased mutation accumulation by bacteriophage populations was previously linked to adaptation to sub-optimal hosts[70]. In that study, experimental evolution on different hosts showed that cyanophage populations accumulated fewer mutations when infecting an optimal host with high infection efficiency, whereas infection of suboptimal hosts resulted in higher rates of mutation accumulation, including synonymous mutations in structural genes that allowed use of host tRNAs. This suggests that the higher mutation accumulation and diversification we observed may be partially driven by a switch to a sub-optimal host in these non-native communities, which is further supported by the low abundance of Theomophage in these communities, which may indicate lower infection efficiency.

On the one hand, dispersal among communities can accelerate viral evolution because migration increases the local genetic variation on which selection can act, but on the other hand it can also disrupt local co-evolution between bacteriophages and hosts. A recent study showed that coevolution of a single phage-host pair in spatially structured communities increased diversity compared to mixed populations, through local diversification in multiple coexisting ecotypes[22]. The H-regime of the Quistad *et al*. experiment mixed viral communities while leaving the microbial communities isolated. It thus provided an intermediate situation between structured and mixed conditions. On shorter timescales (up to 8 weeks) this increased viral diversity on both local (single mesocosm) and metacommunity (across mesocosms) level. This was a consequence of the coexistence of different Theomophage strains, but contributions also came from molecular events involving genomic recombination and the appearance of new mutations. In contrast, in the V-regime, which maintained both viral and microbial community structure, little to no molecular evolution of Theomophage was observed. This suggests that in natural ecosystems, events that disturb local community structure, such as seasonal rainfall in soil communities[13] or cycles of community assembly and dispersal (*e.g.,* on oceanic particles), act as drivers of bacteriophage evolution.

Conclusion

We have described a massive outbreak of a single bacteriophage in high diverse and complex compost communities. Theomophage consistently dominated only community type A, except when an experimental horizontal virome transfer protocol allowed the bacteriophage to spread. In the face of migration and exposure to new hosts, genomic recombination and new mutations arose in the non-native community type B. These results contribute a fascinating peek into the complex eco-evolutionary dynamics of a bacteriophage in mesocosms, a tractable setup that mimics the complexity of microbial systems in nature.

Materials and Methods

Experimental setup and sequencing

We analyzed previously published shotgun metagenomes from compost derived microbial communities[30]. We here briefly recapitulate the experimental setup for clarity. Ten different initial founding mesocosms were established from 1 g compost by sampling different parts of a compost heap, followed by a 4-week acclimatization period. From these 10 communities (t = 0) two new series were created: a 'vertical' regime that was serially transferred every 2 weeks to a new mesocosm consisting of 19 ml nitrogen-limited minimal M9 medium plus cellulose (4 cm^2 paper) as the sole carbon source, and a 'horizontal' regime which was treated the same as the vertical mesocosm, with the addition that at each transfer 1 ml size-filtered (0.2μm) viral fractions were collected, pooled and redistributed between all 10 horizontal mesocosms, to allow migration of bacteriophages and other selfish genetic elements between communities (Fig.1A).

Sequencing was performed for the initial founding communities (t = 0), and after 2, 4, 6, 8, 20, 32, 40 and 48 weeks, allowing us to compare the eco-evolutionary dynamics of 'closed' ecosystems (vertical regime) with 'open' ecosystems that receive viral influx from the 9 other communities in the horizontal regime every 2 weeks. In addition, size-filtered viral fractions were sequenced after 2, 4, 6, and 8 weeks, allowing us to assess the contents of the viral cocktail during part of the experiment.

Assembly and binning

Read quality control was done with Prinseq v.0.20.4 [71] using parameters `-derep 14 -lc method dus -c threshold 20`. Adapters were trimmed using Flexbar v.3.5.0 [72] with `--adapter-preset Nextera -ap ON`.

To increase the chances of assembling low-abundant genomes such as bacteriophages, we combined reads from all 9 time points for each community and cross-assembled these (*i.e.* one cross-assembly for each community). Additionally, for Theomophage genome annotation and variant calling we also assembled shotgun reads from each sampled time point separately (*i.e.* 170 mesocosms samples) reasoning that this would preserve sample-level diversity. Samples from the viral cocktails were assembled in the same way.

All assemblies were done with metaSPAdes v.3.14.0 [73]**.** Read mapping was done with BWA-mem2 v.0.7.17-r1188[74]. Coverage calculation was done with SAMtools v.1.7 [75]. Contig depths were calculated with `jgi_summarize_bam_contig_depths` and binned with Metabat2 v.2.12.1[76].

Taxonomic profiling of communities

Taxonomic classification was assigned to assembled scaffolds and bins using CAT v5.3[32] with `--index_chunk 1 --block_size 5 --top 11 --I_know_what_Im_doing`, using CAT database and taxonomy constructed on 2021-04-30. To estimate abundances at a given taxonomic rank, reads from individual samples were mapped against the corresponding cross-assembly using RAT and read counts were summed for each taxon. Taxa that were not classified by RAT or only classified at other taxonomic ranks were dropped. For dimension reduction summed reads were center-logratio transformed accounting for zero values by adding 1 to all read counts. To estimate community richness, the number of genera classified by RAT that reached a chosen abundance cutoff were counted (**Fig. S3, Table S1**). Reads from contigs that could not be annotated at genus

level by RAT, for example due to conflicting information from ORFs or incomplete taxonomic annotation of best matches, were either counted as a separate genus, or removed. Putative viral sequences were identified with WhatThePhage v.0.9.0. While counting any contig marked as viral by this suite of virus identification tools increased the number of false positives, our subsequent analysis focused on a single bacteriophage sequence identified with high confidence by manual curation (see main text). Estimates of total viral abundance are thus an upper boundary, meaning the abundance of Theomophage relative to all viral sequences is a very conservative estimate. To estimate viral abundances (including Theomophage), we randomly selected a cross-assembly where Theomophage was completely assembled and used that as a target for read mapping.

vMAG completeness

Completeness of viral contigs was estimated with CheckV[77]. Theomophage was independently assembled in 54 single-time points assemblies. Genome size was consistently 63,535 bp (46/54) and MAGs shared 99.98 ANI, strongly suggesting complete assembly of the viral genome. Although CheckV[77] reports DTR (Direct Terminal Repeats) of 55 bp length in the assembled contig indicative of a completely assembled genome, it assigns it medium-quality due to the low estimated completeness (AAI-based 88,92% with >90% required for High-quality) because the contig length does not match the expected length.

Gene calling, function prediction, lifestyle prediction

Theomophage MAG NODE_224_length_63535_cov_5745_934389 from single time point assembly H3T2 was annotated with Pharokka v1.2.0[78]. Specifically, coding sequences (CDS) were predicted with PHANOTATE[79], tRNAs were predicted with tRNAscan-SE 2.0[80], tmRNAs were predicted with Aragorn[81] and CRISPRs were predicted with CRT[82]. Functional annotation was generated by matching each CDS to the PHROGs[83], VFDB[84] and CARD[85] databases using MMseqs2[86]. Fig. 3C and 4C were drawn in R v.4.3.1 with gggenes v.0.4.1 (https://github.com/wilkox/gggenes). Theomophage lifestyle was predicted with BACPHLIP v.0.9.6[44].

vConTACT2 analysis

19,197 bacterial and archaeal viral reference genomes were obtained from millardlab.org (release 1 December 2022) [87]. To this set predicted proteins from *Theomophage* MAG NODE244 were added and VconTACT2 v.0.9.19 was used to create a gene sharing network (–rel-mode 'Diamond' –pcs-mode MCL –vcs-mode ClusterONE). Networks were visualized with Cytoscape v.3.9.1[88] and processed using annotation files from millardlab.org and Inkscape 0.92.5 (www.inkscape.org).

Protein and gene phylogenies

For tailed bacteriophages three conserved proteins - the Terminase large subunit (terl), major capsid protein (MCP) and portal protein (portal) can be used to infer taxonomic relationships [89]. We gathered proteins annotated as "terminase large" or "terminase NOT small", "portal", or "major capsid" or "large capsid" from the Millard dataset and added the corresponding protein from Theomophage MAG NODE_244. Genes were aligned with Clustal Omega v.1.2.4 [90] using default parameters and positions with gaps in more than 50% of the sequences were removed with trimAl v.1.4.rev15 [91]. Gene trees were made using FastTree version 2.1.10 [92] with default parameters and plotted with R v.4.3.1 using the ggtrees package [93].

For the *Schitoviridae* gene trees, fasta files were downloaded from Genbank for the 119 *Schitoviridae* described in Wittmann *et al*[39]. These genomes were annotated with Prokka and PHROG hmm profiles as described in the previous section, and parsed PHROG annotations for "portal", "terminase large", and "major capsid" OR "large capsid". Trees were then constructed as described above and rooted with corresponding genes from two randomly selected *Ackermannviridae* genomes.

VIRIDIC analysis

Intergenomic similarities (pairwise nucleotide identity normalized for genome lengths) were calculated from MAG NODE244 and 119 *Schitoviridae* sequences obtained from Wittmann *et al*[39] using the VIRIDIC webserver (BLASTN parameters "-word_size 7 -reward 2 -penalty -3 -gapopen 5 -gapextend -2")[41]. The same analysis was done for an additional set of 444 *Schitoviridae* sequences recovered from metagenomes described in [40].

Host prediction

iPHoP[46] was run with default parameters and the most recent database (release Sept_2021_pub) and additionally, with a lowered confidence cutoff (-m 80). iPHoP predictions and raw output per bioinformatic tool are available as **Table S5, 6.** In addition, we queried the IMG CRISPR spacer database with the Theomophage genome using blastn and the default parameters on the IMG/VR website[45], which returned no matches.

Impact of outbreak on individual genera

To assess the impact of the Theomophage outbreaks on individual genera, genus abundances were compared before and after the eight observed Theomophage outbreaks. Specifically, for the 55 genera that reached a threshold of least 1% of sample reads, genus abundances were normalised for each sample to the maximum abundance of that genus in the experiment, and the difference in abundance between time points directly before and after the peak of the Theomophage outbreak was compared to changes in abundance across the same interval for time points without an outbreak (**Fig. S19**).

Variant calling

To reduce the chance of 'read stealing', QC-filtered reads were competitively mapped from 174 metagenomes against all contigs in a randomly selected single-sample assembly that included the complete 63,535 bp Theomophage MAG (sample H3T2 containing NODE_224_length_63535_cov_5745_934389). Variants were called for the Theomophage MAG using Haplotypecaller, GenomicsDBImport and GenotypeGVCFs from GATK v.4.2.3.0[94]. To filter sequencing artifacts and read mapping artifacts, only alleles that reached a frequency of 10% in the population and were supported by a minimum of 4 reads, coming from samples where ≥30% of the genome was covered by ≥10 reads were considered for downstream analysis. This resulted in 76 SNPs. To further rule out read mapping artifacts, we inspected read depth coverage profiles from BAM files with Integrative Genomics Viewer V.2.13.2[95], and confirmed any variants in genomic areas with suspicious coverage by (1) re-assembling mapped read pairs for that sample, and (2) separately assembling the complete metagenome from that sample, and comparing the resulting assembled Theomophage contigs to the reference.

Resolving strains and estimating strain abundance

To resolve strains, genotypes were inferred in samples where variant frequencies for all variants were close to either 0 or 1, indicating monomorphic populations. S2 and S156 were inferred from

V2T0, and V1T0, V5T0 and V6T0 respectively. Similar reasoning allowed the identity of S6 to be inferred from samples H6T20 and H6T32, and S34 from sample H3S40. For Fig 5.B we then assumed the simplest scenario that could explain the evolution of S6 and S34 from S2 and S156 that was compatible with the variant trajectories in Fig.5A, namely two separate recombination events and fixation of 61 new variants in S34. We approximated strain abundances for time points 8, 20, 32 and 40 in communities H3,H4, H6 and the cocktail as follows: S34 abundance as the average relative abundance of orange and blue variants; S6 abundance as the average of pink variants; S156 as the average of the blue and green variants, and S2 as 1 minus the abundance of S156, S24 and S6. For other time points the abundance of S156 was estimated as the average of blue, pink, and green SNPs and S2 as 1 minus the abundance of S156. The four strains inferred above accounted for 92.1% of the reported SNPs and we conclude that they captured most of the relevant dynamics of genetic diversity within the Theomophage population. The remaining 6 SNPs represent further community-specific microdiversity within strain S156 that was limited to communities V5, V6 and H5 **(Fig. S21)**. For visual clarity we gave SNPs with similar trajectories from cluster 1 different colors in Fig. 5-6.

The effects of variants were predicted with SNPEff v.5.0e[96] (-noLog -noStats -no-downstream -no-upstream -no-utr) using a custom SNPEff database built with NODE_224_length_63535_cov_5745_934389 as reference, a .gff file generated using Prokka and the "Bacterial_and_Plant_Plastid" codon table. MAGs were interrogated for Diversity Generating Retroelements (DGRs) with the webserver MyDGR[53].

Unless noted otherwise, figures and analysis was done in RStudio v.2022.12.0 with R v.4.2.0 using cowplot v1.1.1 (github.com/wilkelab/cowplot), Tidyverse v.1.3.2[97], data.table v1.14.6, ggbeeswarm v.0.7.1. Fig. 1A was created with BioRender.com.

Acknowledgments

The authors gratefully acknowledge Jan Kees van Amerongen for running the local computer cluster. JM is supported by the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, www.nwo.nl) grant 022.005.023. BED is supported by European Research Council (ERC) Consolidator grant 865694: DiversiPHI, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860, the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship founded by German Federal Ministry of Education and Research, and the European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF). PBR acknowledges core support from the Max Planck Society and funding from the Deutsche Forschungsgemeinschaft (DFG) Collaborative Research Center 1182 'Origin and Function of Metaorganisms' (grant no. SFB1182, Project C4).

References

1. Duerkop, B. A., Clements, C. V., Rollins, D., Rodrigues, J. L. M. & Hooper, L. V. A composite bacteriophage alters colonization by an intestinal commensal bacterium. *Proc. Natl. Acad. Sci.* **109**, 17621–17626 (2012).

2. Szabo, R. E. *et al.* Historical contingencies and phage induction diversify bacterioplankton communities at the microscale. *Proc. Natl. Acad. Sci.* **119**, e2117748119 (2022).

3. Vincent, F. *et al.* Viral infection switches the balance between bacterial and eukaryotic recyclers of organic matter during coccolithophore blooms. *Nat. Commun.* **14**, 510 (2023).

4. Pal, C., Maciá, M. D., Oliver, A., Schachar, I. & Buckling, A. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature* **450**, 1079–1081 (2007).

5. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics. *Soil Syst.* **4**, 23 (2020).

6. Roux, S. & Emerson, J. B. Diversity in the soil virosphere: to infinity and beyond? *Trends Microbiol.* (2022) doi:10.1016/j.tim.2022.05.003.

7. Taş, N. *et al.* Metagenomic tools in microbial ecology research. *Curr. Opin. Biotechnol.* **67**, 184–191 (2021).

8. Nesme, J. *et al.* Back to the Future of Soil Metagenomics. *Front. Microbiol.* **7**, (2016).

9. Baldrian, P. The known and the unknown in soil microbial ecology. *FEMS Microbiol. Ecol.* **95**, fiz005 (2019).

10. Pratama, A. A. & van Elsas, J. D. The 'Neglected' Soil Virome – Potential Role and Impact. *Trends Microbiol.* **26**, 649–662 (2018).

11. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).

12. Lee, S. *et al.* Soil pH influences the structure of virus communities at local and global scales. *Soil Biol. Biochem.* 108569 (2022) doi:10.1016/j.soilbio.2022.108569.

13. Santos-Medellín, C. *et al.* Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *Proc. Natl. Acad. Sci.* **119**, e2209132119 (2022).

14. Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F. & Firestone, M. K. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci.* **116**, 25900–25908 (2019).

15. Albright, M. B. N. *et al.* Experimental evidence for the impact of soil viruses on carbon cycling during surface plant litter decomposition. *ISME Commun.* **2**, 1–8 (2022).

16. Trubl, G. *et al.* Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems* **3**, e00076-18 (2018).

17. Wu, R. *et al.* DNA Viral Diversity, Abundance, and Functional Potential Vary across Grassland Soils with a Range of Historical Moisture Regimes. *mBio* **12**, e02595-21 (2021).

18. Wu, R. *et al.* Structural characterization of a soil viral auxiliary metabolic gene product – a functional chitosanase. *Nat. Commun.* **13**, 5485 (2022).

19. Huang, D. *et al.* Enhanced mutualistic symbiosis between soil phages and bacteria with elevated chromium-induced environmental stress. *Microbiome* **9**, 150 (2021).

20. Rillig, M. C., Muller, L. A. & Lehmann, A. Soil aggregates as massively concurrent evolutionary incubators. *ISME J.* **11**, 1943–1948 (2017).

21. Piel, D. *et al.* Phage–host coevolution in natural populations. *Nat. Microbiol.* **7**, 1075–1086 (2022).

22. Shaer Tamar, E. & Kishony, R. Multistep diversification in spatiotemporal bacterial-phage coevolution. *Nat. Commun.* **13**, 7971 (2022).

23. Buckling, A. & Rainey, P. B. Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. Lond. B Biol. Sci.* **269**, 931–936 (2002).

24. Gómez, P. & Buckling, A. Bacteria-Phage Antagonistic Coevolution in Soil. *Science* **332**, 106–109 (2011).

25. Vos, M., Birkett, P. J., Birch, E., Griffiths, R. I. & Buckling, A. Local Adaptation of Bacteriophages to Their Bacterial Hosts in Soil. *Science* **325**, 833–833 (2009).

26. Bhat, T., Cao, A. & Yin, J. Virus-like Particles: Measures and Biological Functions. *Viruses* **14**, 383 (2022).

27. Williamson, K. E., Fuhrmann, J., Wommack, K. & Radosevich, M. Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annu. Rev. Virol.* (2017) doi:10.1146/annurev-virology-101416-041639.

28. López-García, P. *et al.* Metagenome-derived virus-microbe ratios across ecosystems. *ISME J.* 1–12 (2023) doi:10.1038/s41396-023-01431-y.

29. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).

30. Quistad, S. D., Doulcier, G. & Rainey, P. B. Experimental manipulation of selfish genetic elements links genes to microbial community function. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20190681 (2020).

31. Dutilh, B. E. *et al.* Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **28**, 3225–3231 (2012).

32. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).

33. Dijk, B. van *et al.* Identifying and tracking mobile elements in evolving compost communities yields insights into the nanobiome. 2023.02.02.526783 Preprint at https://doi.org/10.1101/2023.02.02.526783 (2023).

34. Lewin, G. R. *et al.* Long-Term Cellulose Enrichment Selects for Highly Cellulolytic Consortia and Competition for Public Goods. *mSystems* **7**, e01519-21 (2022).

35. López-Mondéjar, R., Zühlke, D., Becher, D., Riedel, K. & Baldrian, P. Cellulose and hemicellulose decomposition by forest soil bacteria proceeds by the action of structurally variable enzymatic systems. *Sci. Rep.* **6**, 25279 (2016).

36. Větrovský, T., Steffen, K. T. & Baldrian, P. Potential of Cometabolic Transformation of Polysaccharides and Lignin in Lignocellulose by Soil Actinobacteria. *PLOS ONE* **9**, e89108 (2014).

37. Marquet, M. *et al.* What the Phage: a scalable workflow for the identification and analysis of phage sequences. *GigaScience* **11**, giac110 (2022).

38. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).

39. Wittmann, J. *et al.* From Orphan Phage to a Proposed New Family–The Diversity of N4-Like Viruses. *Antibiotics* **9**, 663 (2020).

40. Zheng, K. *et al.* Identification of hidden N4-like viruses and their interactions with hosts in global metagenomes. Preprint at https://doi.org/10.21203/rs.3.rs-539338/v1 (2022).

41. Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses* **12**, 1268 (2020).

42. Knecht, L. E. *et al.* Bacteriophage S6 requires bacterial cellulose for Erwinia amylovora infection. *Environ. Microbiol.* **24**, 3436–3450 (2022).

43. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

44. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).

45. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).

46. Roux, S. *et al.* iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLOS Biol.* **21**, e3002083 (2023).

47. Coutinho, F. H. *et al.* RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* **2**, 100274 (2021).

48. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).

49. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2017).

50. Lu, C. *et al.* Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **19**, 5 (2021).

51. Roodgar, M. *et al.* Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. *Genome Res.* **31**, 1433–1446 (2021).

52. Roux, S. *et al.* Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* **12**, 3076 (2021).

53. Sharifi, F. & Ye, Y. MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res.* **47**, W289–W294 (2019).

54. Pinto, Y., Chakraborty, M., Jain, N. & Bhatt, A. S. Phanta: Phage-inclusive profiling of human gut metagenomes. 2022.08.05.502982 Preprint at https://doi.org/10.1101/2022.08.05.502982 (2022).

55. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, ncomms5498 (2014).

56. Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).

57. Somerville, V. *et al.* Functional strain redundancy and persistent phage infection in Swiss hard cheese starter cultures. *ISME J.* 1–12 (2021) doi:10.1038/s41396-021-01071-0.

58. Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biol.* **19**, 163 (2021).

59. Weiss, M. *et al.* In vivo replication of T4 and T7 bacteriophages in germ-free mice colonized with Escherichia coli. *Virology* **393**, 16–23 (2009).

60. Cenens, W. *et al.* Expression of a Novel P22 ORFan Gene Reveals the Phage Carrier State in Salmonella Typhimurium. *PLOS Genet.* **9**, e1003269 (2013).

61. Latino, L., Midoux, C., Hauck, Y., Vergnaud, G. & Pourcel, C. Pseudolysogeny and sequential mutations build multiresistance to virulent bacteriophages in Pseudomonas aeruginosa. *Microbiology* **162**, 748–763 (2016).

62. Siringan, P., Connerton, P. L., Cummings, N. J. & Connerton, I. F. Alternative bacteriophage life cycles: the carrier state of Campylobacter jejuni. *Open Biol.* **4**, 130200 (2014).

63. Pourcel, C., Midoux, C., Vergnaud, G. & Latino, L. A carrier state is established in Pseudomonas aeruginosa by phage LeviOr01, a newly isolated ssRNA levivirus. *J. Gen. Virol.* **98**, 2181–2189 (2017).

64. Mäntynen, S., Laanto, E., Oksanen, H. M., Poranen, M. M. & Díaz-Muñoz, S. L. Black box of phage–bacterium interactions: exploring alternative phage infection strategies. *Open Biol.* **11**, 210188 (2021).

65. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).

66. Benler, S. & Koonin, E. V. Fishing for phages in metagenomes: what do we catch, what do we miss? *Curr. Opin. Virol.* **49**, 142–150 (2021).

67. Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *eLife* **9**, e51971 (2020).

68. Sato, M. P. *et al.* Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Res.* **26**, 391–398 (2019).

69. Bertozzi Silva, J., Storms, Z. & Sauvageau, D. Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363**, fnw002 (2016).

70. Enav, H., Kirzner, S., Lindell, D., Mandel-Gutfreund, Y. & Béjà, O. Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. *Nat. Commun.* **9**, 4698 (2018).

71. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

72. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology* **1**, 895–905 (2012).

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

74. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.

75. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

76. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

77. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).

78. Bouras, G. *et al.* Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, btac776 (2023).

79. McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinforma. Oxf. Engl.* **35**, 4537–4542 (2019).

80. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).

81. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).

82. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).

83. Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics Bioinforma.* **3**, lqab067 (2021).

84. Chen, L. *et al.* VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325-328 (2005).

85. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).

86. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

87. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE* (2021) doi:10.1089/phage.2021.0007.

88. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

89. Dutilh, B. E. *et al.* Perspective on taxonomic classification of uncultivated viruses. *Curr. Opin. Virol.* **51**, 207–215 (2021).

90. Sievers, F. *et al.* Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

91. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

92. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

93. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

94. Poplin, R. *et al. Scaling accurate genetic variant discovery to tens of thousands of samples*. https://www.biorxiv.org/content/10.1101/201178v3 (2018) doi:10.1101/201178.

95. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

96. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

97. Journal of Open Source Software: Welcome to the Tidyverse. https://joss.theoj.org/papers/10.21105/joss.01686.

Supporting Information for

Eco-evolutionary dynamics of massive, parallel bacteriophage outbreaks in compost communities

Jeroen Meijer[1*], Petros Skiadas[1], Paul B. Rainey [2,3], Paulien Hogeweg[1], Bas E. Dutilh[1,4]

1. Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands
2. Department of Microbial Population Biology, Max Planck Institute for Evolutionary Biology, Plön, Germany
3. Laboratory of Biophysics and Evolution, CBI, ESPCI Paris, Université PSL, CNRS, Paris, France
4. Institute of Biodiversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, 07743, Jena, Germany

* Corresponding author: Jeroen Meijer

Email: jeroen.meijer@protonmail.com

This PDF file includes:

Figures S1 to S22

Tables S1 to S7

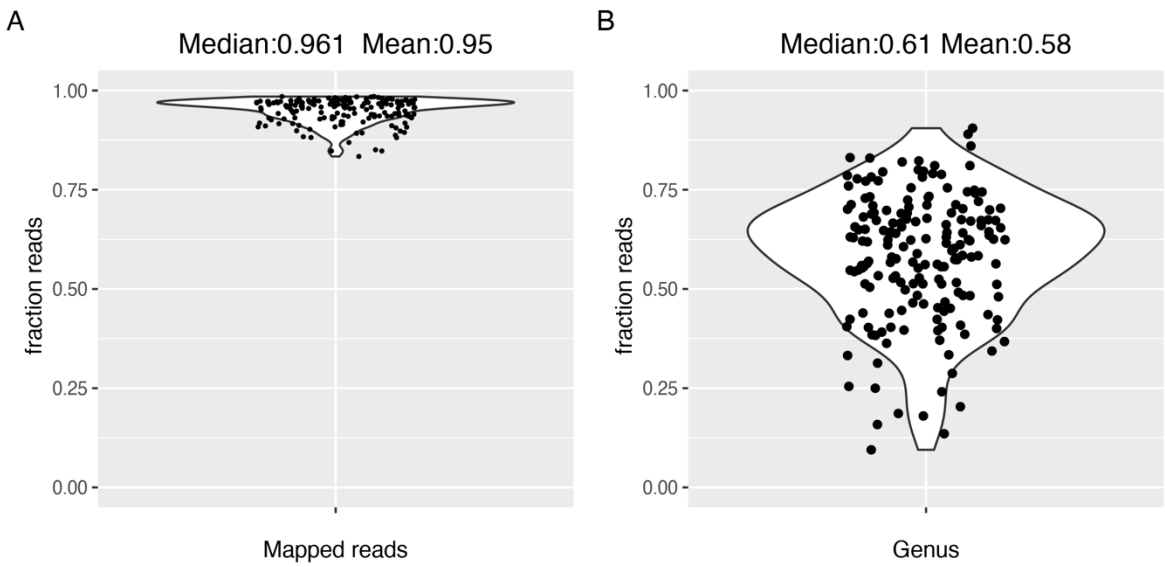


Figure S1. (**A)** Fraction of sample reads mapped to contigs (170 samples). **(B)** Fraction of reads mapping to contigs annotated at genus rank.

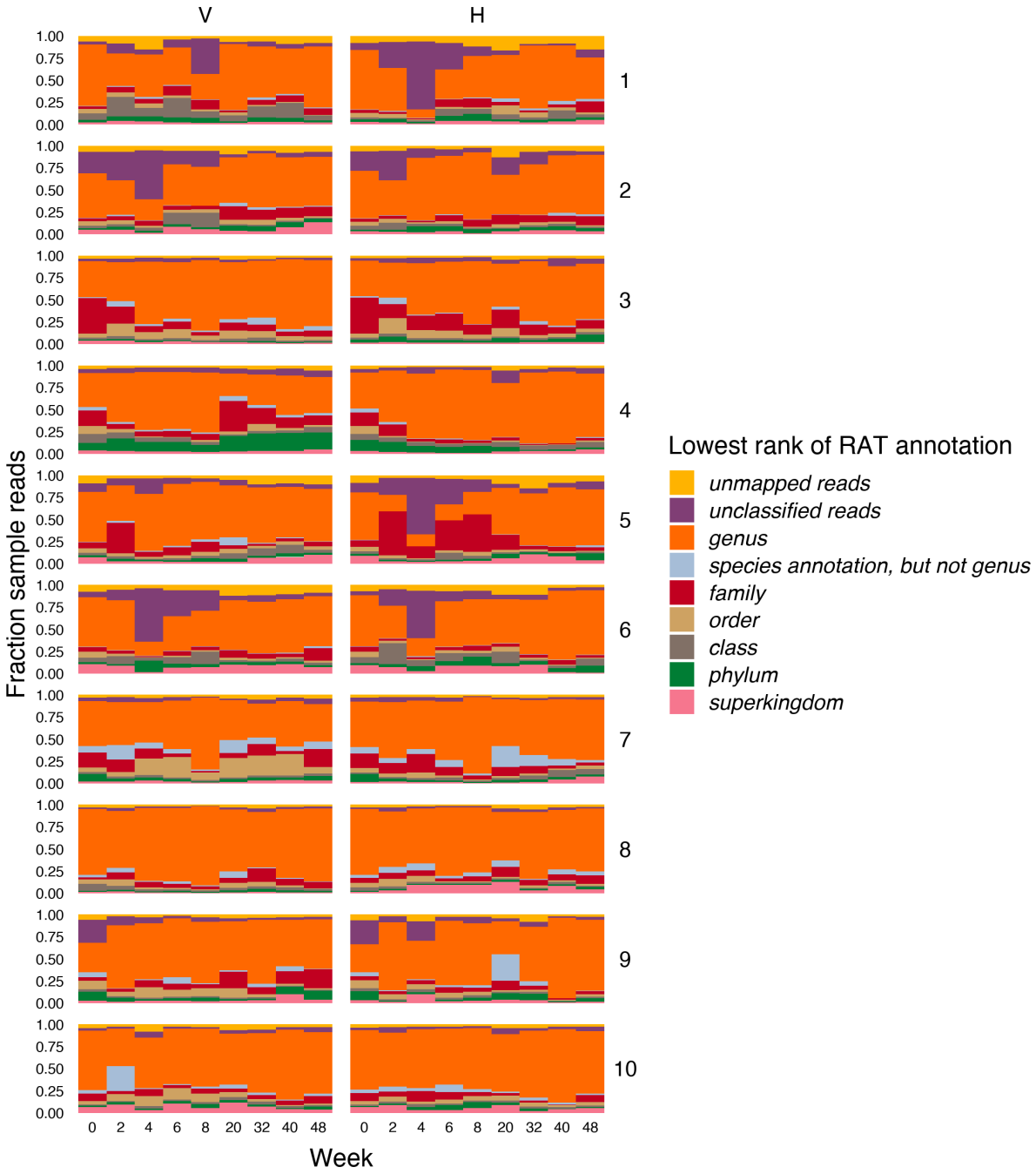


Figure S2. Reads annotated at different taxonomic ranks.

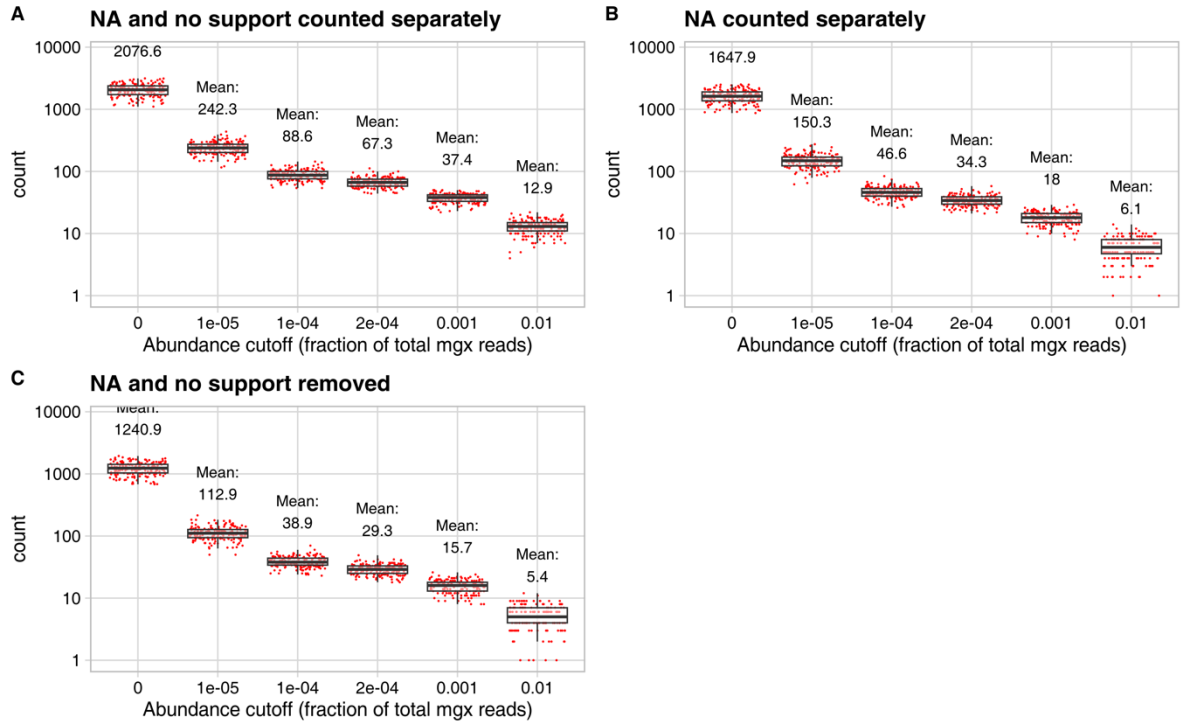


Figure S3. Community richness estimates per sample for different abundance cutoffs and filtering methods. RAT does not support a taxonomic rank at genus level when ORFs give conflicting genera ("no support"), or when the best reference genome is not annotated at genus rank but might have annotation at species level ("NA"). **(A)** Both cases counted separate genera each. **(B)** Only including NA in genus counts **(C)** Only lineages that could be annotated at genus rank were counted, "no support" and NA filtered out. For SD see Table S1.

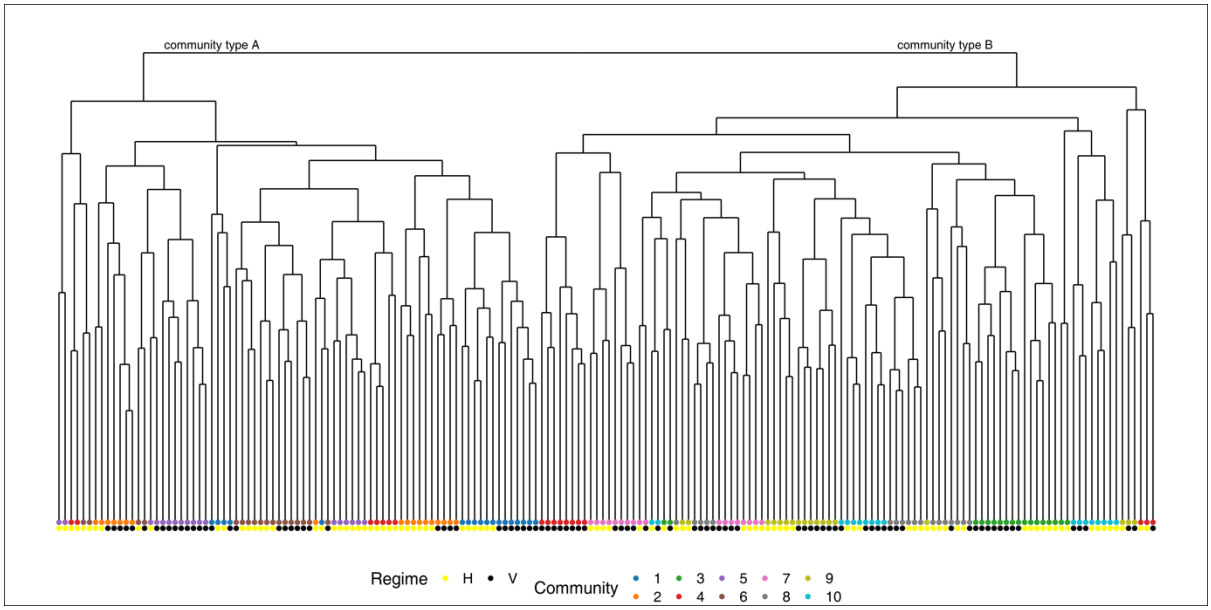


Figure S4. Hierarchical clustering of communities based on Aitchison distance of genus abundances (see Methods). Communities cluster in two types, and community type is stable to treatment regime. Only samples from community 4 (red) in the horizontal regime change cluster affiliation, suggesting the community composition is impacted by the influx of sequences through the SGE cocktail.

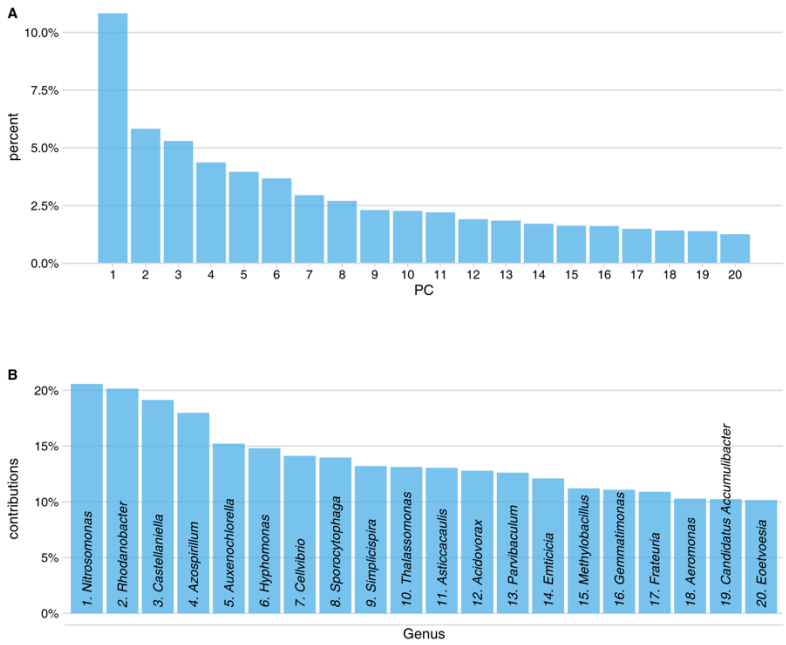


Figure S5. (**A**) Percentage explained by top 20 components of PCA in Fig.1c of the main text. (**B)** Top 20 genera contributing to PC1 and PC2.

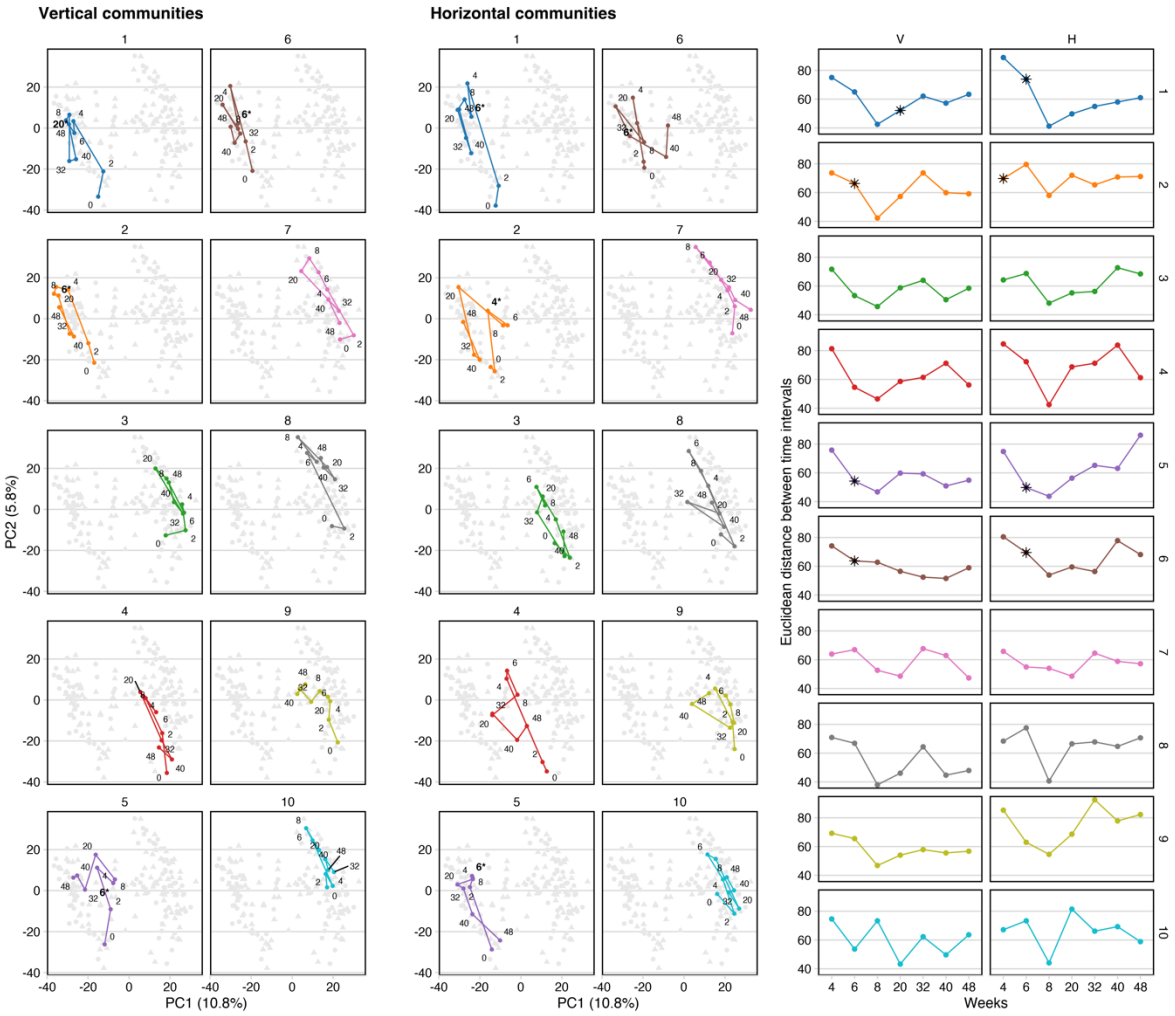


Figure S6. (A) PCA of communities 1-10 taxonomically profiled at genus rank, plotted per mesocosm. Consecutive samples are connected by lines and numbers indicate week of sampling. In communities 1,2,5 and 6 samples directly after the peak of a Theomophage outbreak are in bold and marked with an asterisk. **(B)** Euclidean distance in the PCA between community samples (lag 2 intervals). Distances between samples directly before and after a Theomophage outbreak are marked with an asterisk, showing there is no observable effect of outbreaks compared to other sampling intervals.

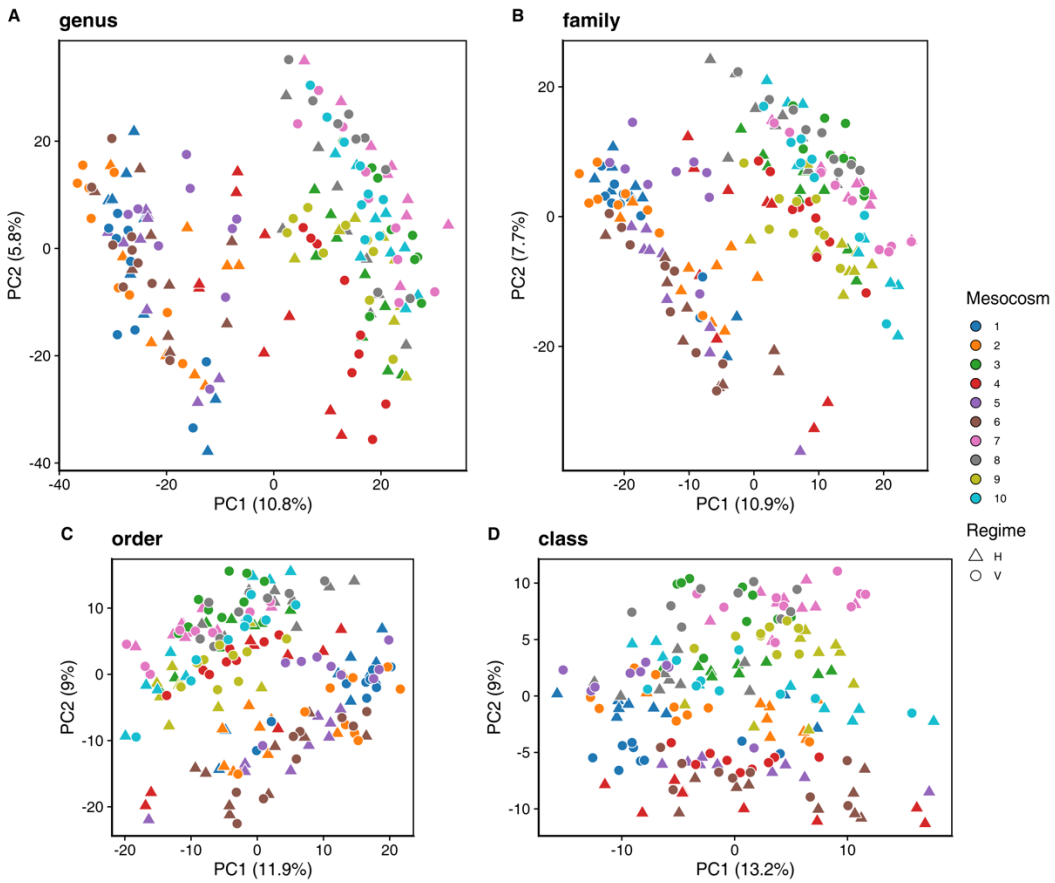


Figure S7. Increasing the fraction of metagenome reads considered in dimension reduction by profiling at higher taxonomic ranks maintains clustering in 2 community types. **(A)** identical to Fig1D in main text, included here a guide. **(B)** reads classified at family rank (*i.e.* orange plus red fractions in Supplementary Figure 2, annotated at family level). PC1 separates the samples in 2 community types. **(C)** Same, but for order. **(D)** Same, but for class.

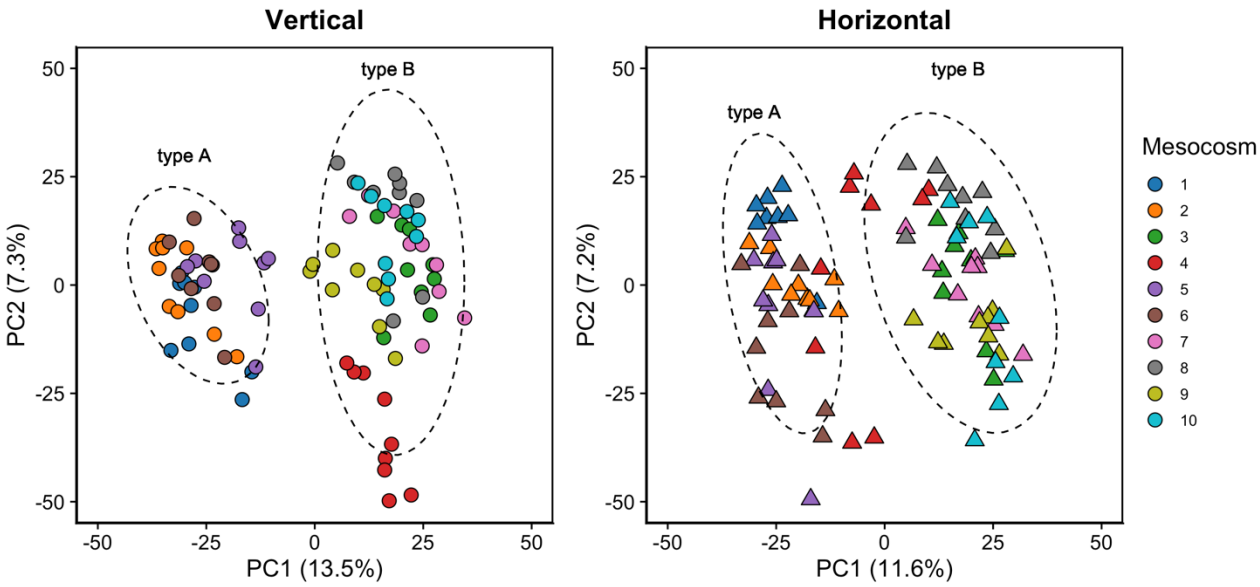


Figure S8. Individual Principal Component Analyses of samples from vertical and horizontal regimes. Separation of community types along the first component is largely maintained for horizontal samples. Ellipses indicate a 95% confidence level of manually selected type-A (mesocosms 1,2,5,6) and type-B communities (3,4,7-10).

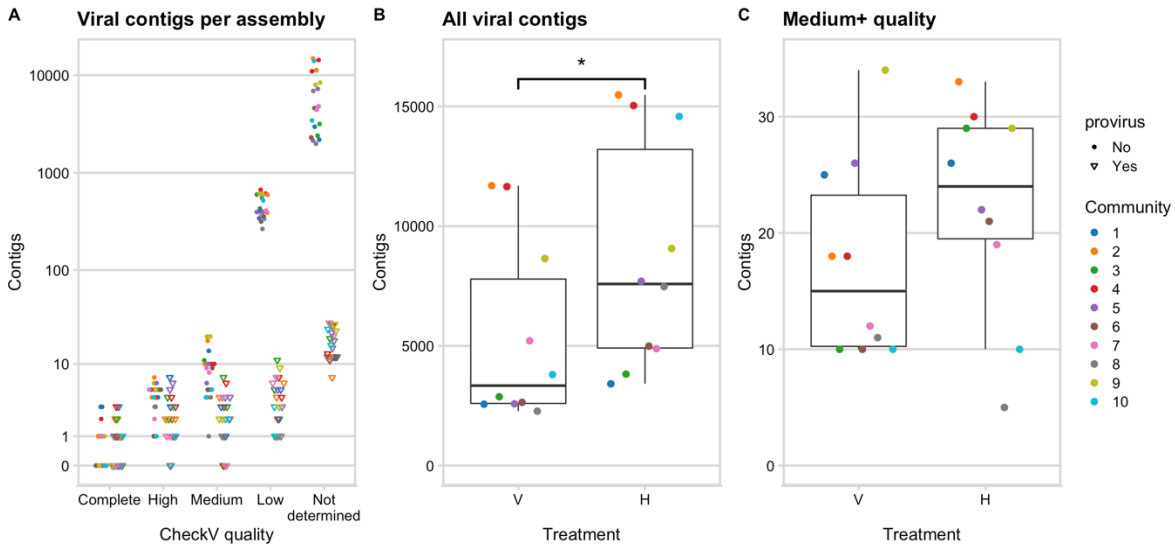


Figure S9. (A) Putative viral sequences identified by What the Phage from 20 cross-assemblies, split by MAG completeness estimated by CheckV and whether the sequence is predicted to be proviral. Colors indicate cross-assembly. **(B)** Total number of predicted viral sequences assembled per community, split per treatment. Two-tailed paired t-test indicates more viral sequences were recovered from horizontal than vertical communities ($p = 0.01211$) **(C)** As in B but only including viral contigs of estimated medium quality or higher.

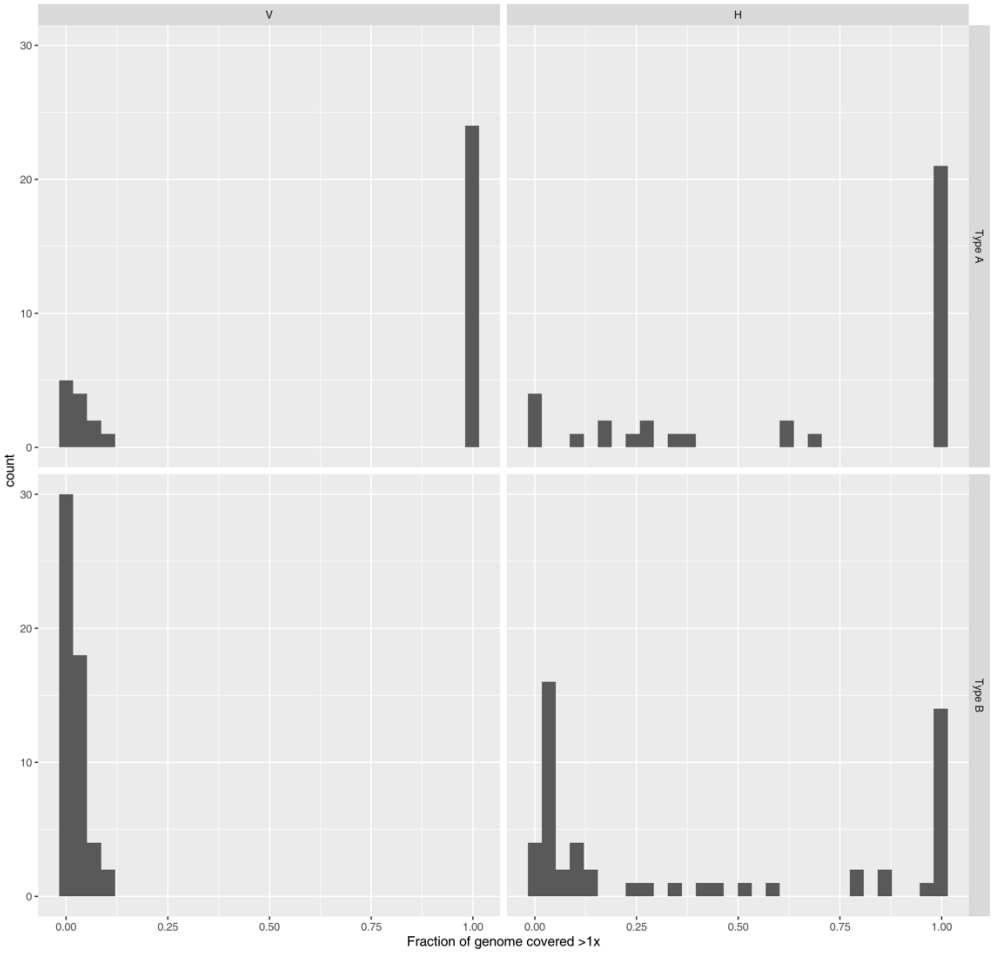
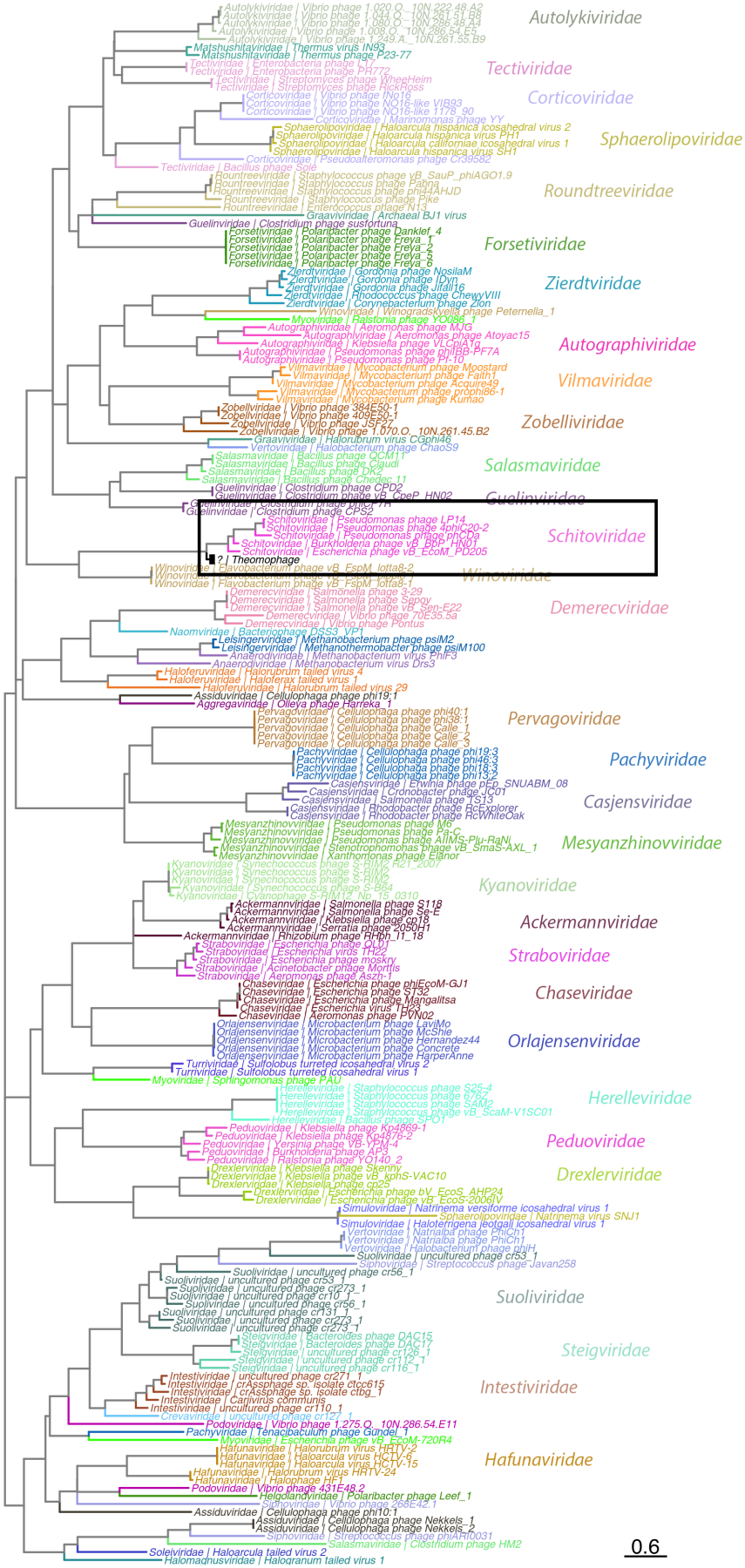
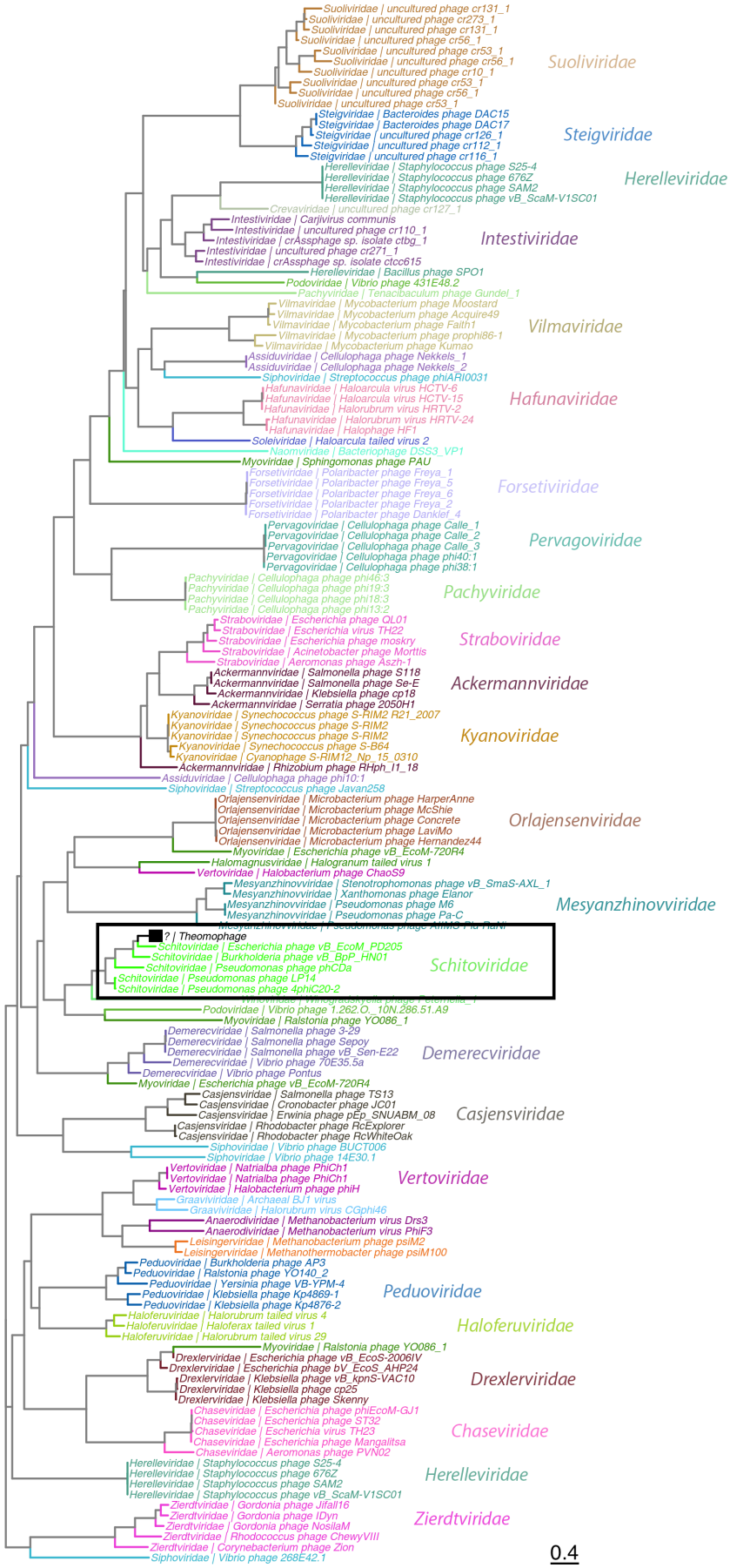


Figure S10. Histograms showing the fraction of the Theomophage genome covered at least by 1 read in horizontal and vertical communities, split for community type. In the vertical regime Theomophage was not detected in type-B communities.

Terminase large subunit (trimmed alignment)



Portal (trimmed alignment)



Major capsid protein (trimmed alignment)

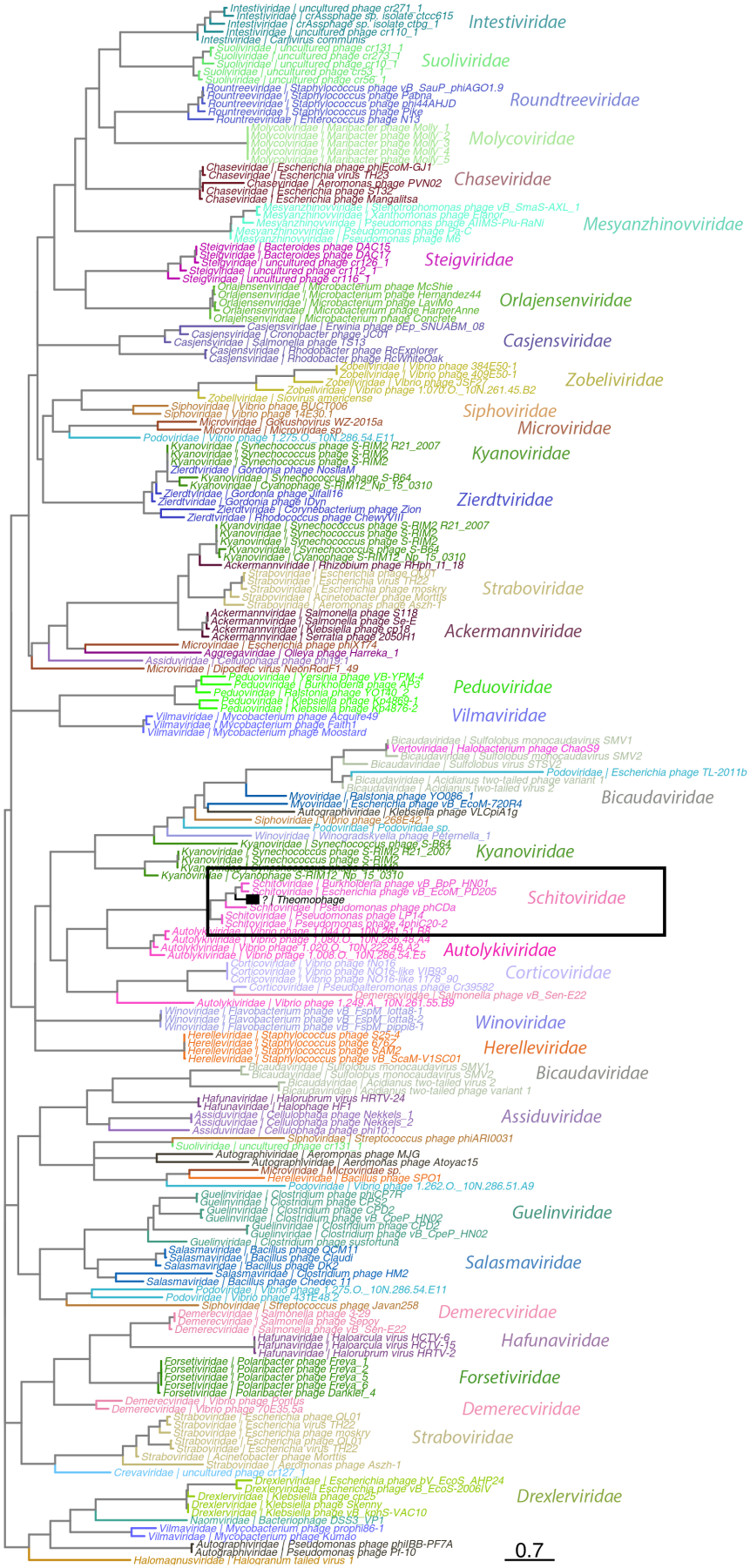
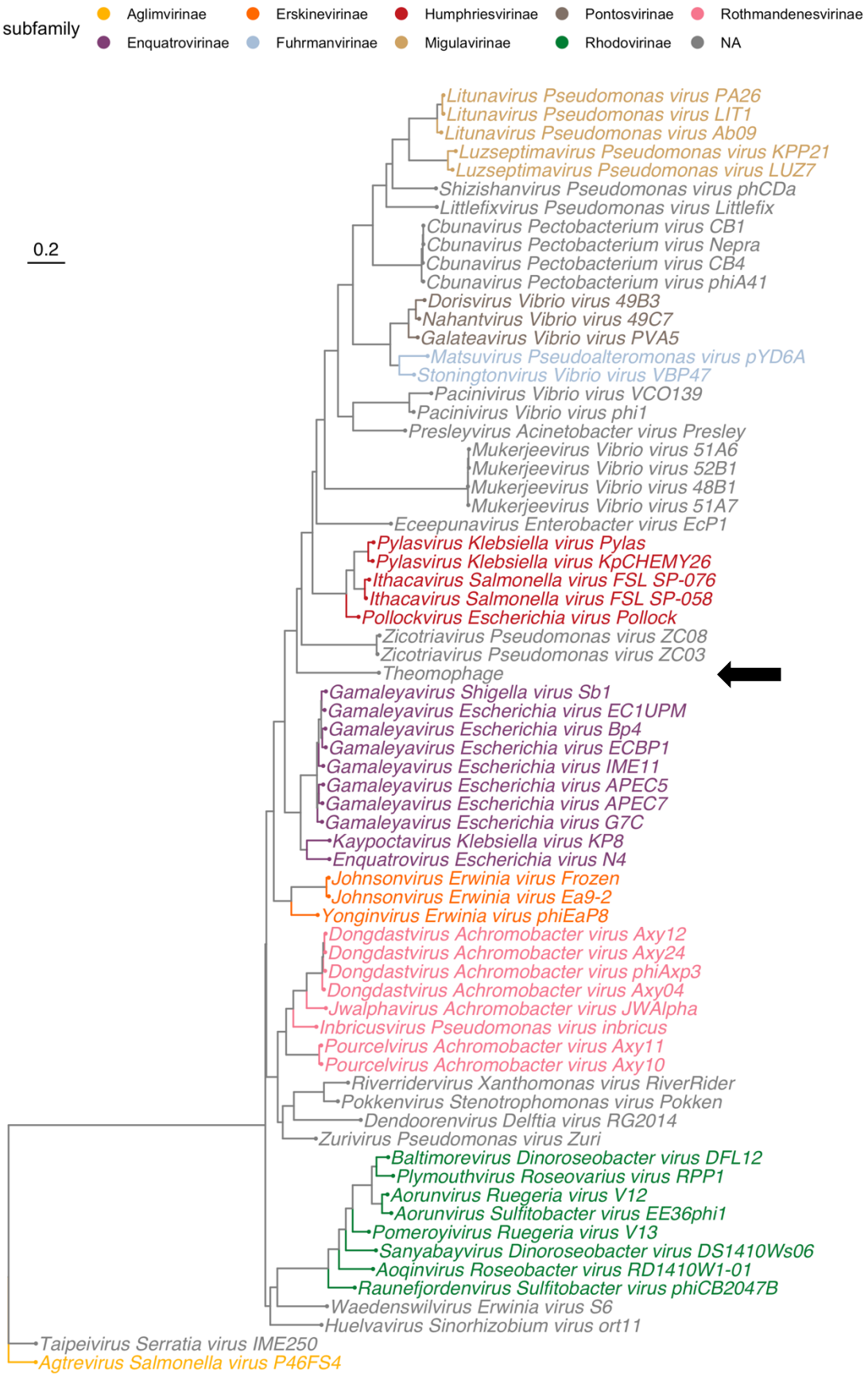


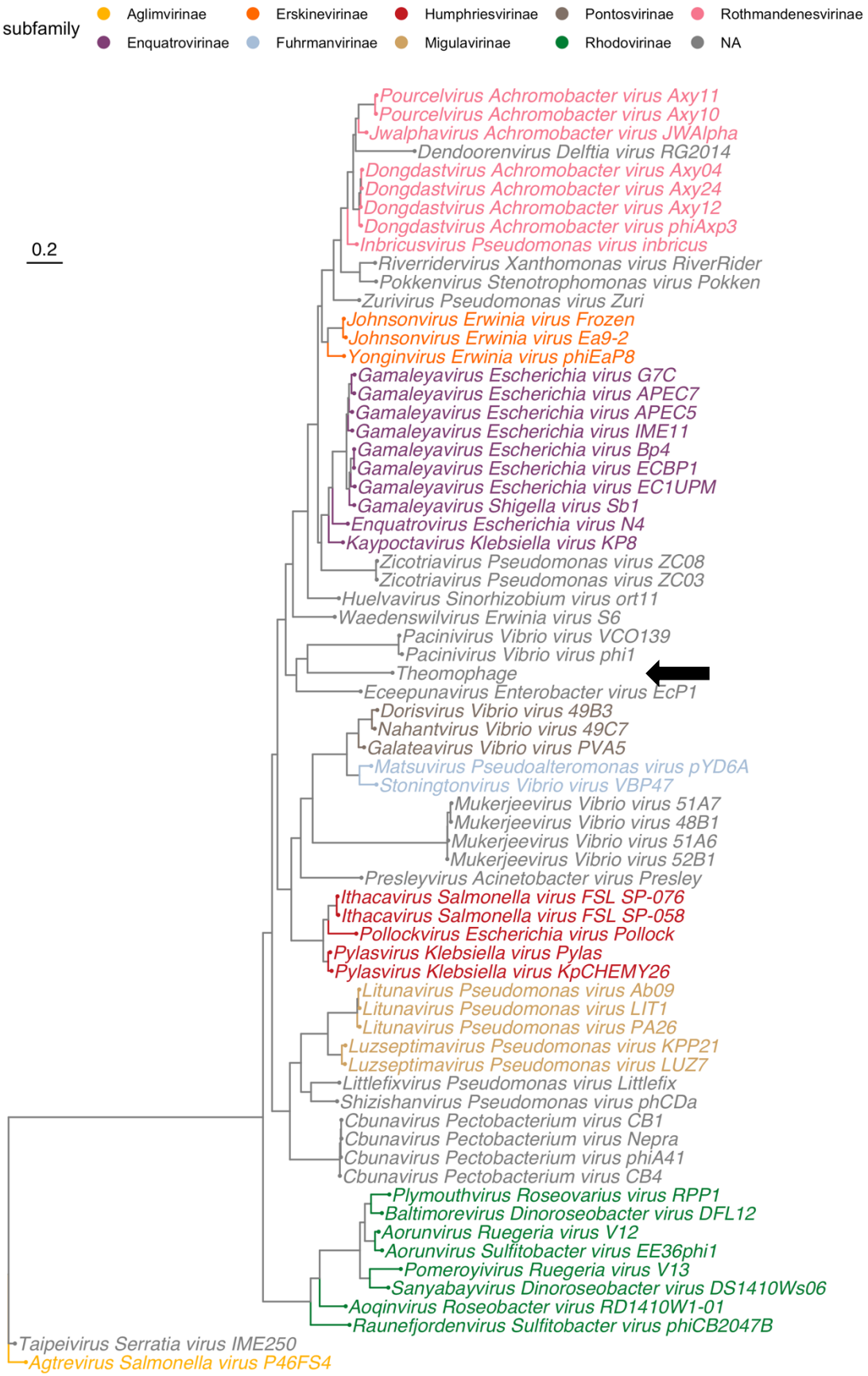
Figure S11 (previous pages)

Unrooted protein trees of **(A)** terminase large subunit, **(B)** portal and **(C)** major capsid protein for 5 randomly selected bacteriophage genomes per family in the Millard dataset (see Methods). Theomophage robustly clusters with *Schitoviridae*. Node labels are colored for family and include species name.

Terminase large Subunit (ICTV Schitoviridae genomes, rooted with 2 Ackermannviridae)



Major Capsid (ICTV Schitoviridae genomes, rooted with 2 Ackermannviridae)



Portal (ICTV Schitoviridae genomes, rooted with 2 Ackermannviridae)

Figure S12 (previous pages)

Gene trees of *Schitoviridae* **(A)** terminase large subunit, **(B)** major capsid and **(C)** portal genes, rooted with corresponding genes from 2 randomly selected *Ackermannviridae* genomes. Theomophage did not consistently cluster within a subfamily (colour) or genus (tip label), reflecting a distant relation to known *Schitoviridae*.

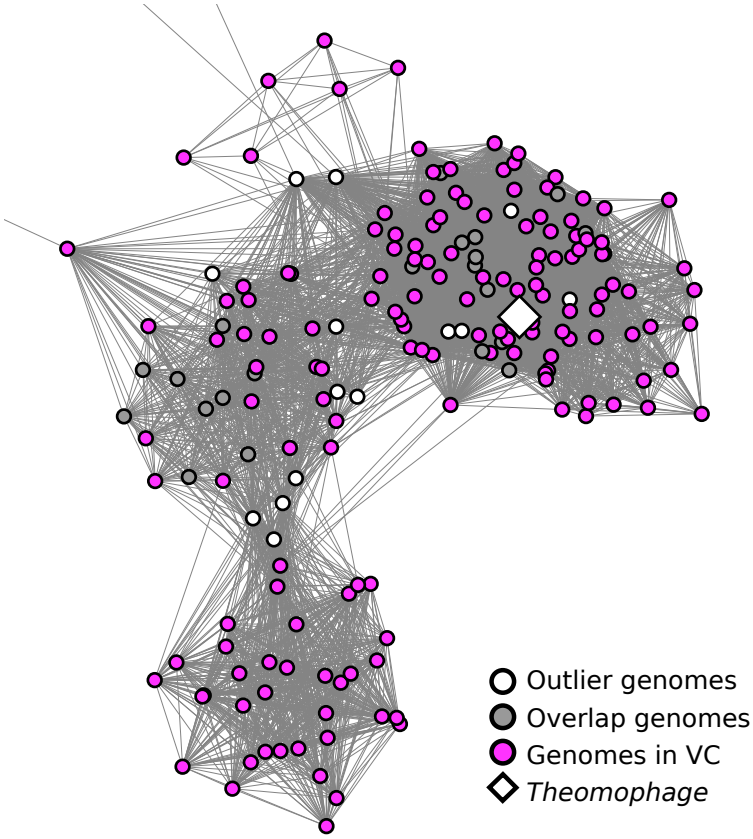


Figure S13. Enlarged view of the *Schitoviridae* supercluster (190 genomes) of Fig. 3A in the main text (gene sharing network of 19,197 bacterial and archaeal genome virus genomes and Theomophage MAG NODE244). The supercluster is formed by sub-clusters that correspond to ICTV genera. Outlier genomes (white) are not part of a genus cluster, indicating no clustering at genus level with other sequences. Overlap genomes are part of multiple VCs.

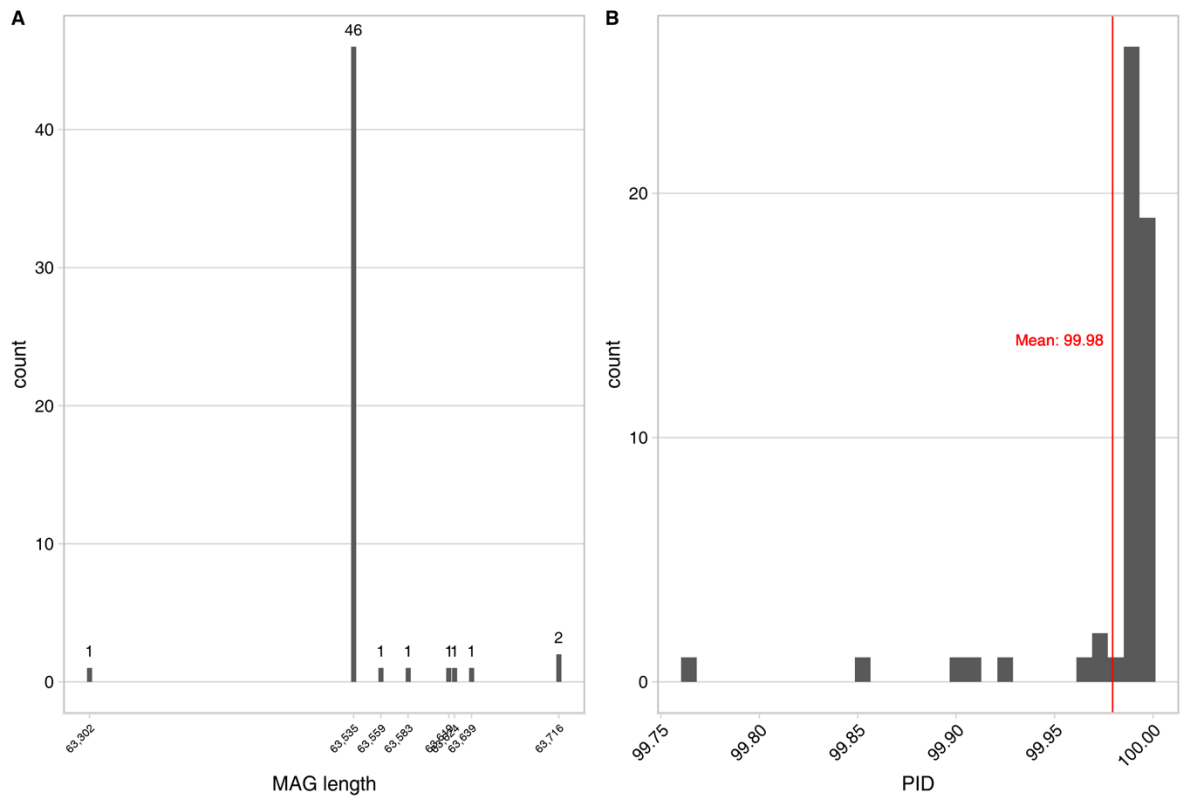


Figure S14. Theomophage MAGs. **(A)** Lengths of Theomophage MAGs. We collected Theomophage MAGs from 170 separately assembled samples by Blasting the assembly and selecting the top hit (highest PID and length), and then selecting contigs >60k bp. **(B)** Sequence similarity of contigs in A calculated with VIRIDIC webserver.

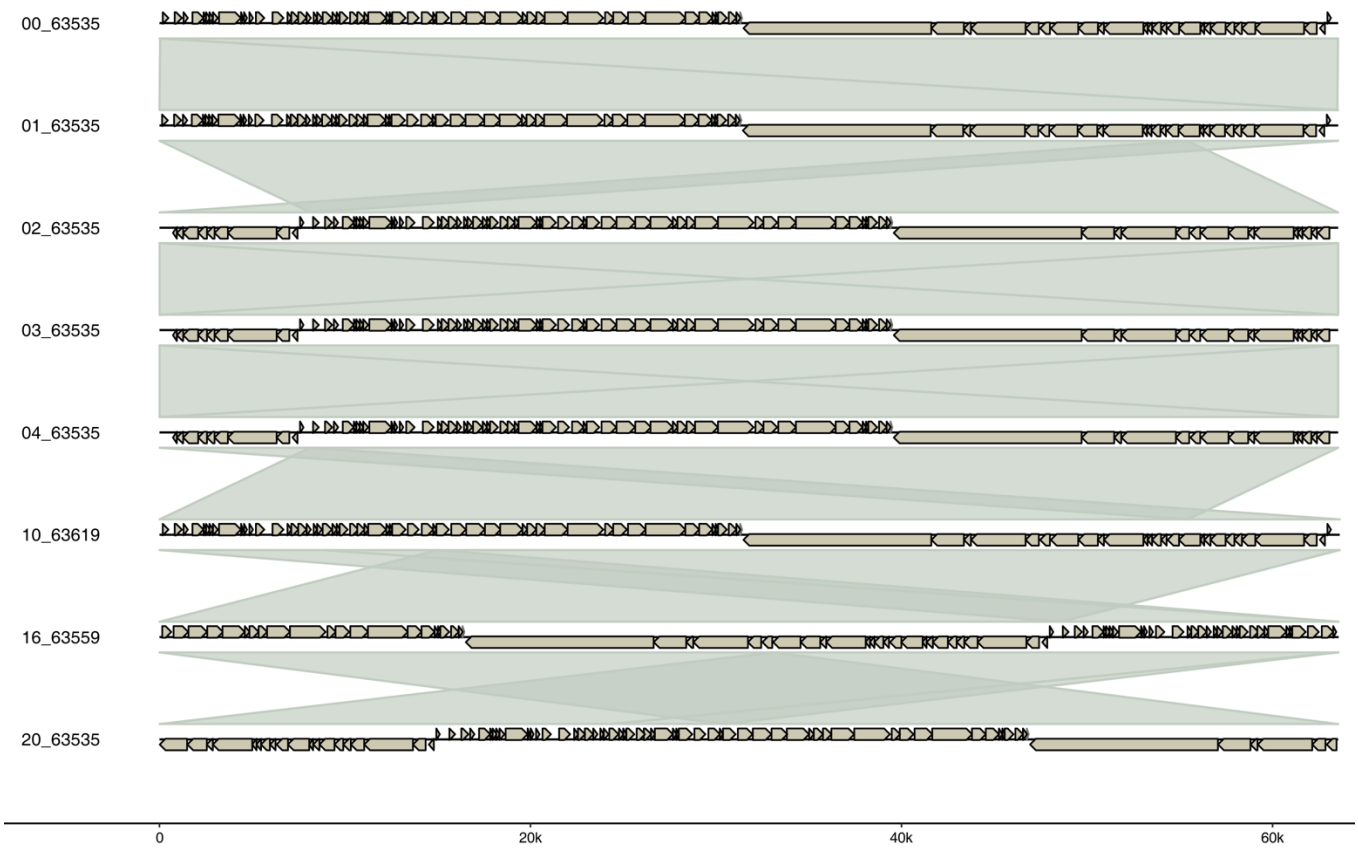


Figure S15. Theomophage vMAGs assembled from mesocosm VC5 samples at transfer 0, 1, 2, 3, 4, 10, 16 and 20, showing vMAGs are circularly rotated relative to each other and of similar length (1st nr indicates transfer (every 2 weeks), 2nd number indicates vMAG length).

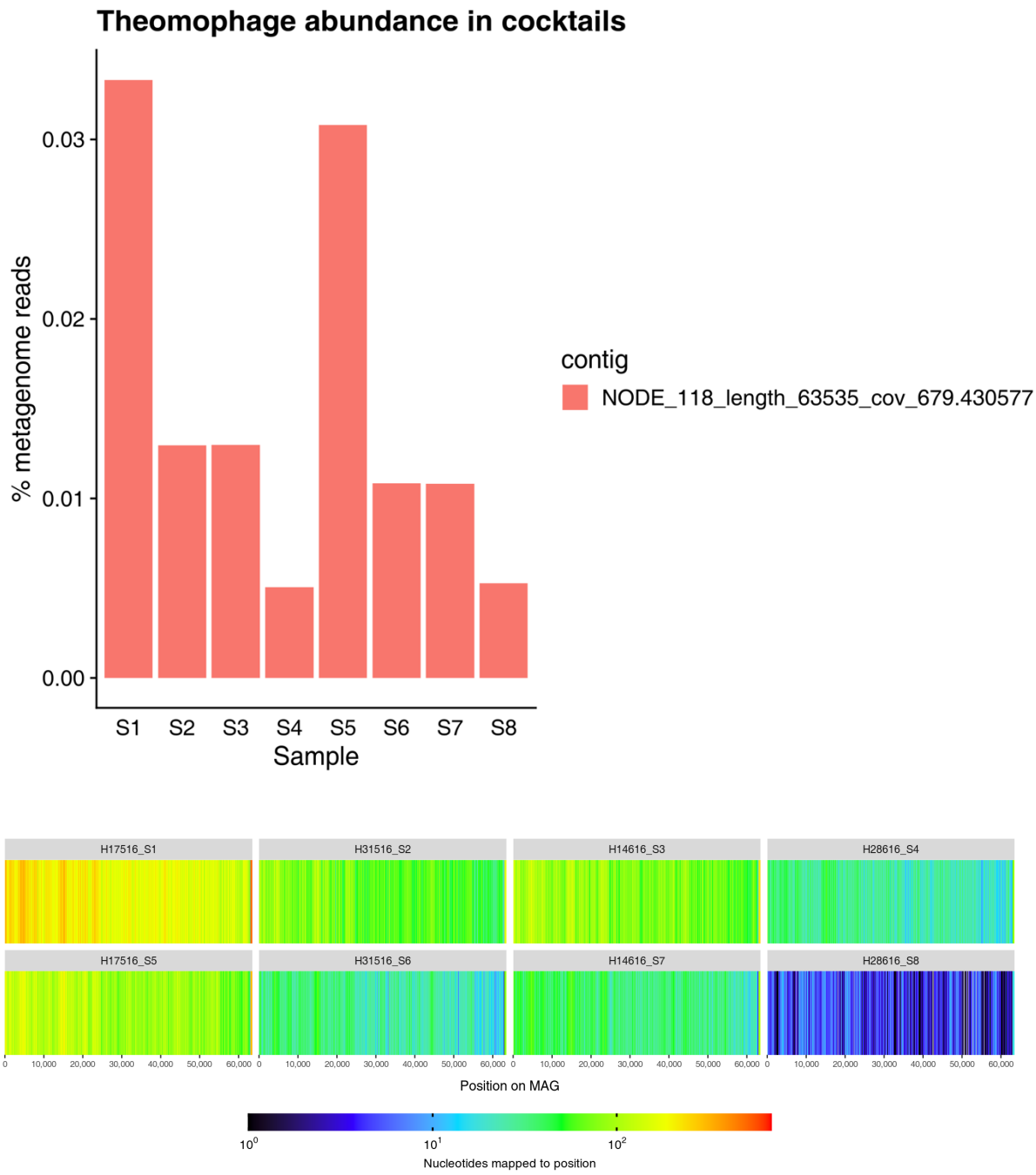


Figure S16. (A) Theomophage abundance in the SGE cocktail samples for time points 2, 4, 6 and 8 weeks. Samples were sequences twice. **(B)** genome coverage profile for Theomophage MAG indicates the full genome is present in all time points.

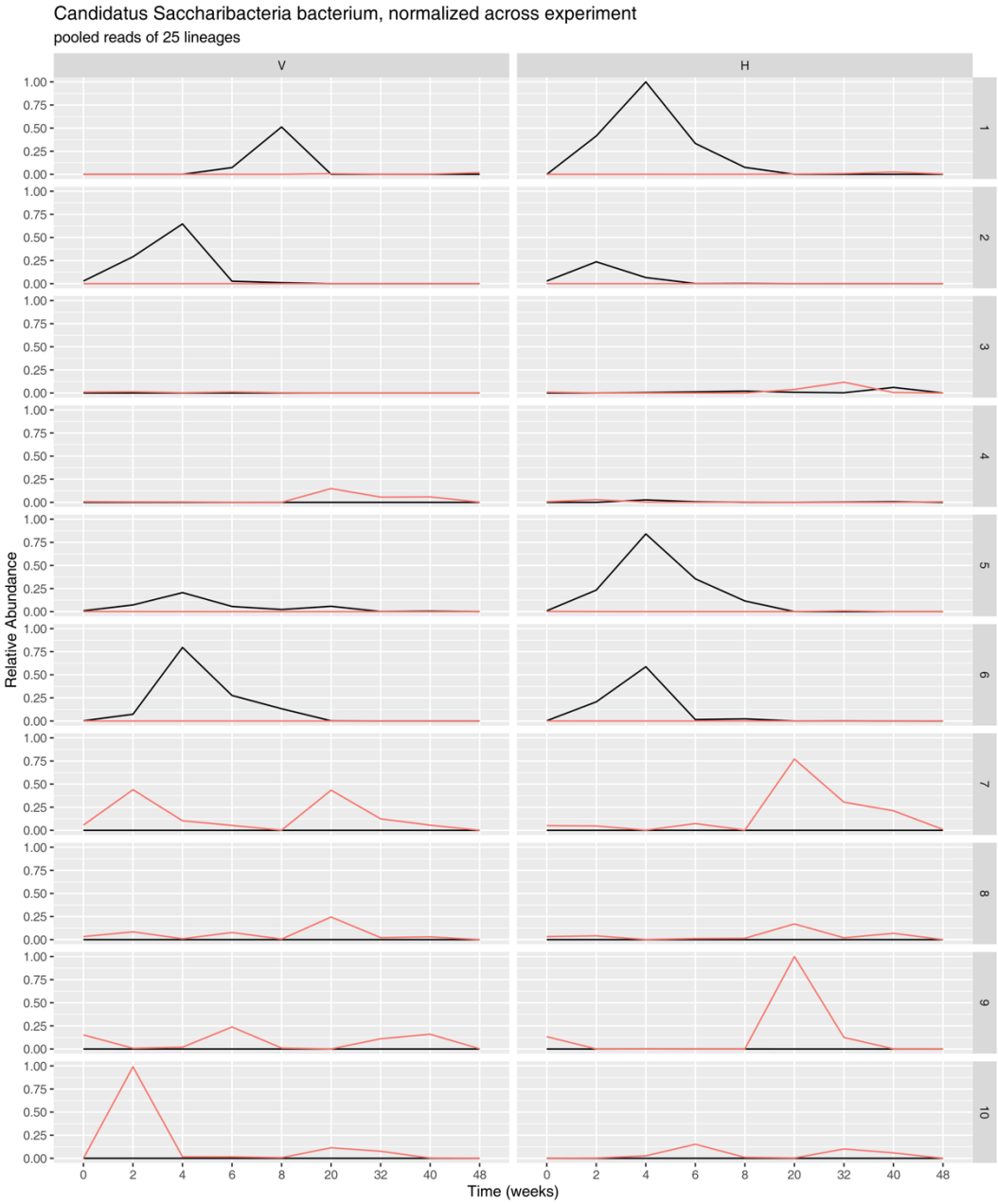


Fig. S17. Relative abundance of Candidatus *Saccharibacteria,* normalized to maximum abundance across all samples. Candidatus *Saccharibacteria* is mainly present in communities 7-10.

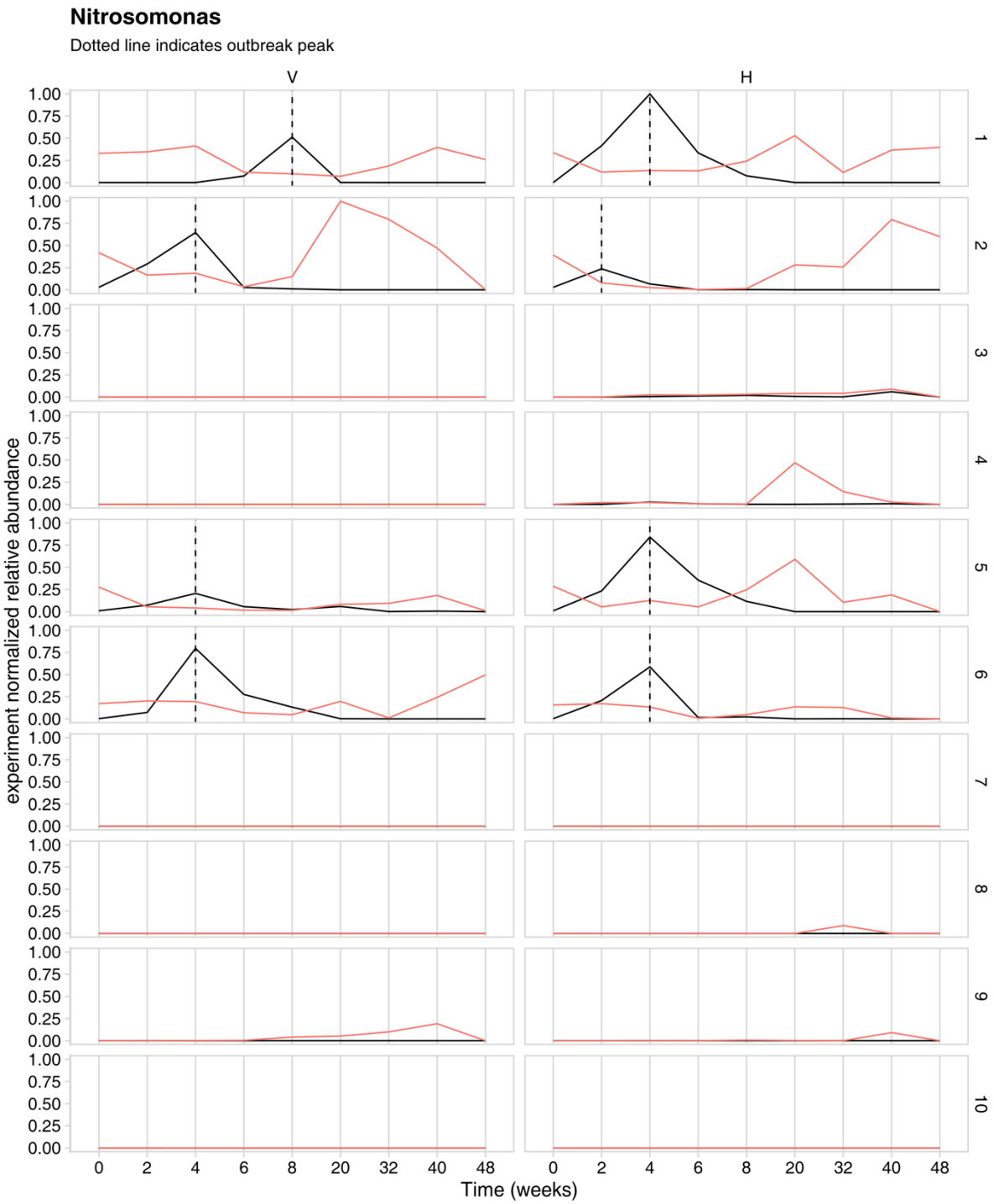


Fig. S18. Relative abundance of *Nitrosomonas*, normalized to maximum abundance across all samples. *Nitrosomonas* is abundant in communities 1,2,5 and 6, and present but less abundant in communities 3,4,7-9, where Theomophage invaded in the horizontal regime.

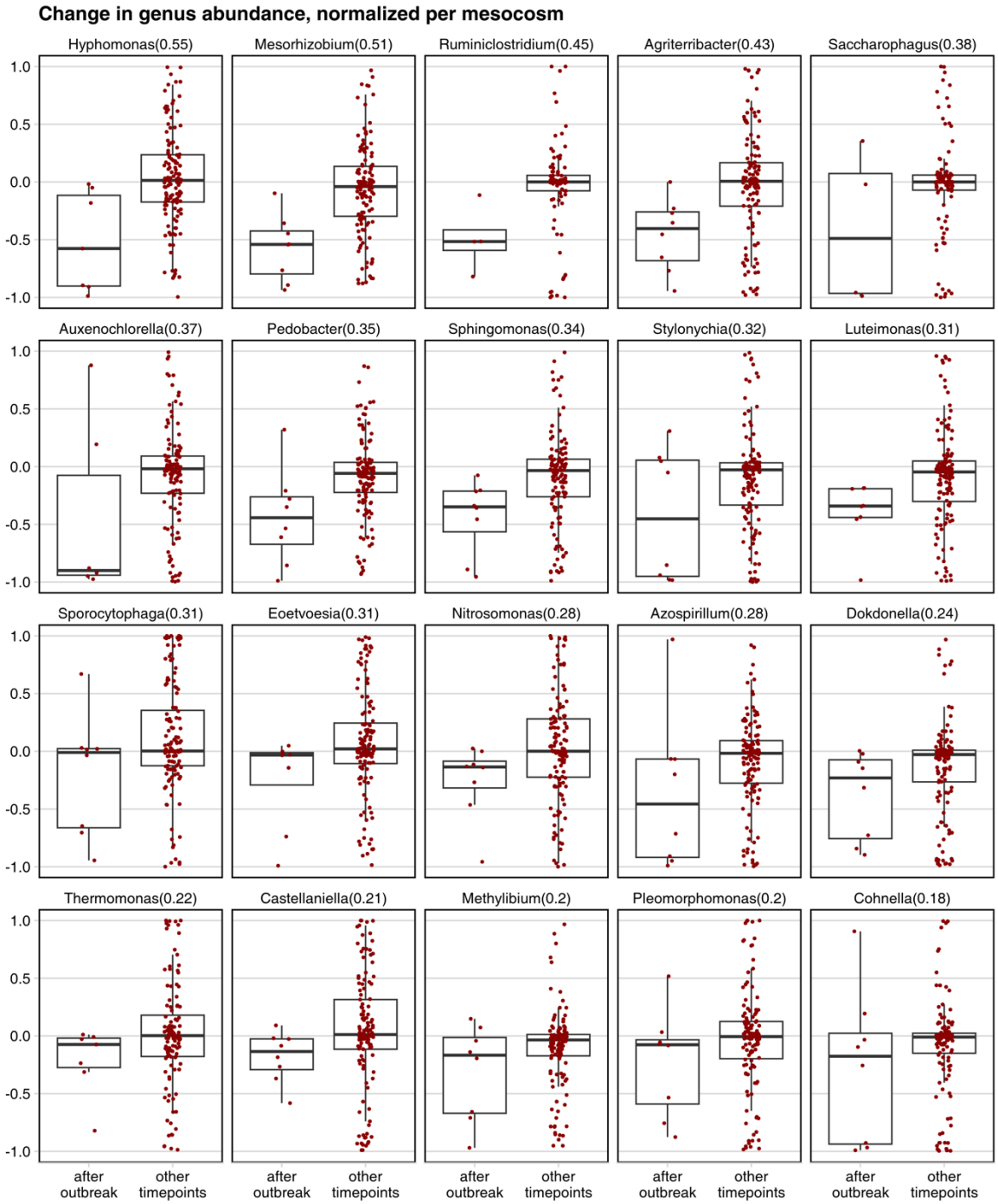


Fig. S19 Top 20 genera whose relative abundance was most negatively impacted by Theomophage outbreaks, ordered by difference in means. Number between brackets indicates difference in means between 'after outbreak' and 'other timepoints'. Genus abundance was normalized for each mesocosm. When a genus was not detected in a sample but was detected in other time points for that community, the abundance in that sample was set to 0.

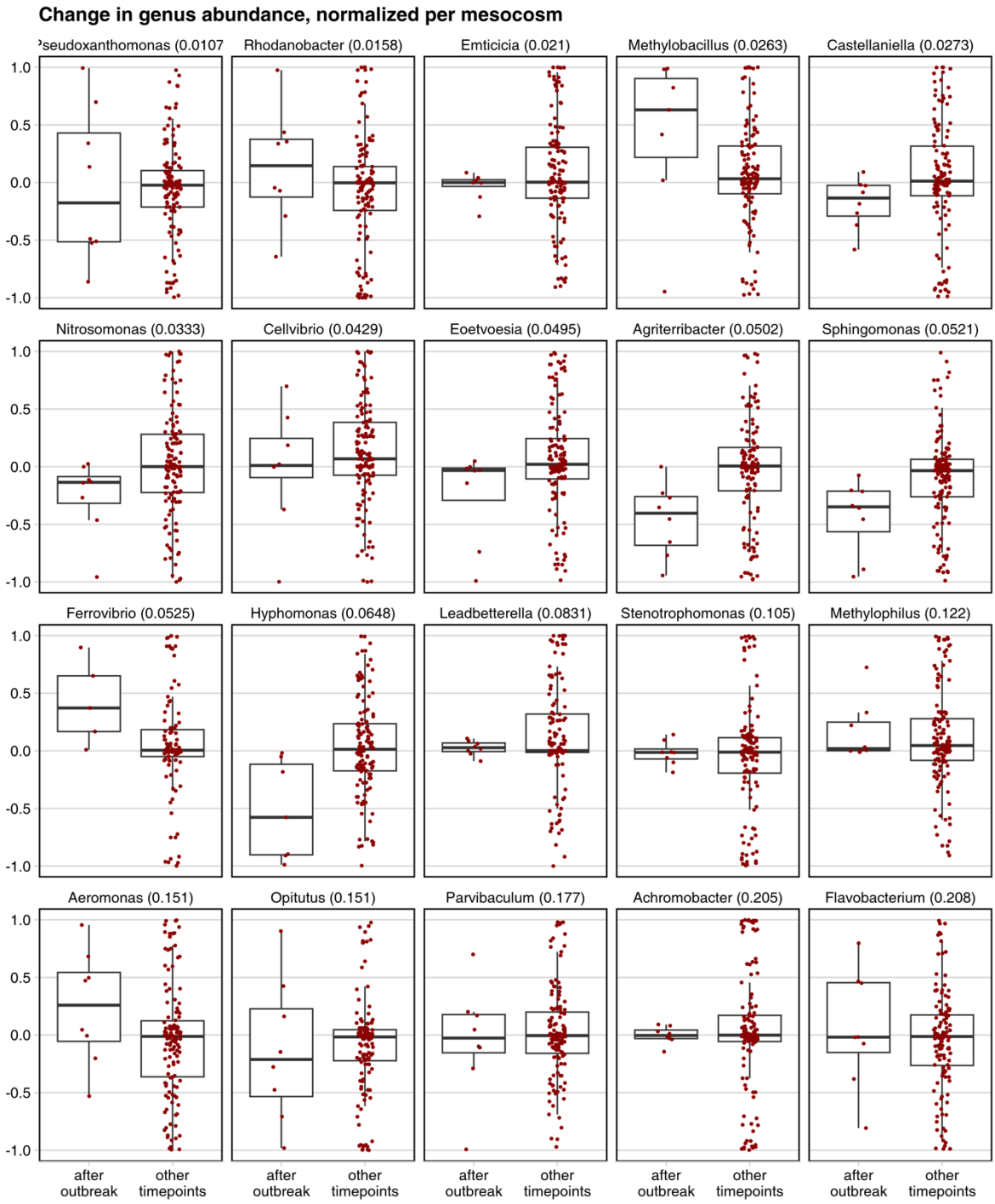


Fig. S20. Same as Fig. S19, but top 20 selected by p-value (number indicated between brackets).

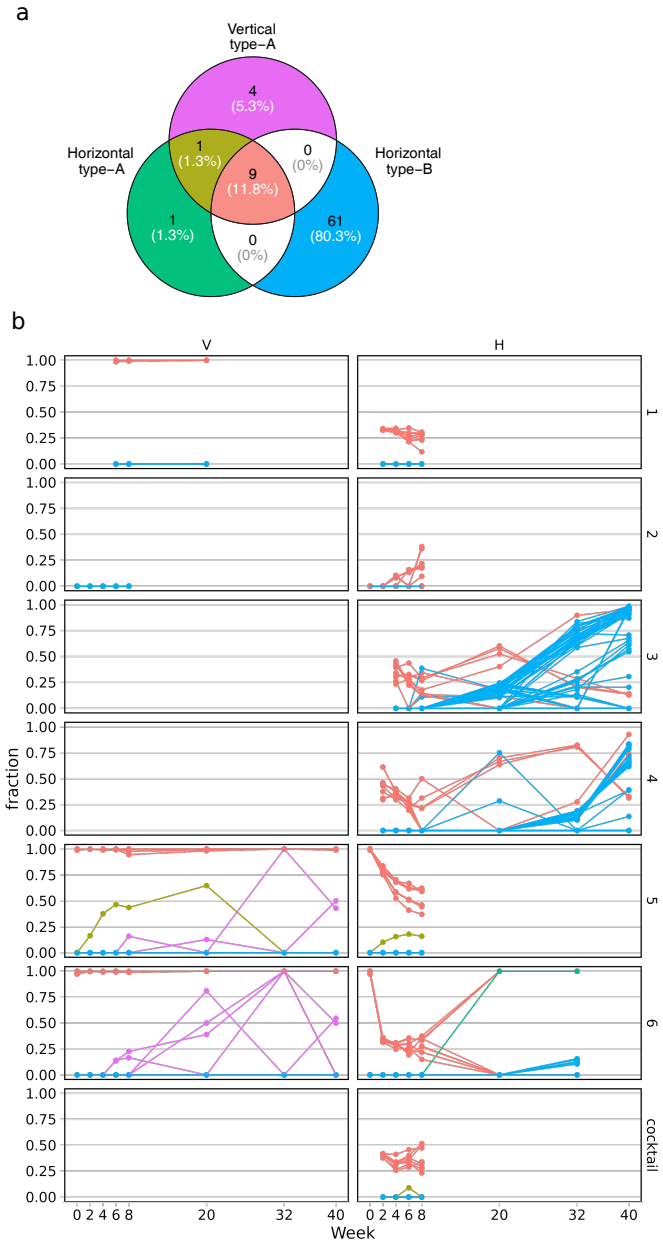


Figure S21. Single Nucleotide Polymorphisms (SNPs), colored for observation in different community types. (A) Venn diagram of 76 SNPs observed in Theomophage populations in native (type-A) and new (type-B) community backgrounds in the vertical and horizontal regimes. Theomophage was not detected in type-B communities in the vertical regime. **(B)** SNP frequency trajectories of SNPs, colored according to A. The olive SNP is unique to community 5 and observed in both the horizontal and vertical regime, and the green SNP is unique to community H6. Olive and pink SNPs represent further microdiversity within strain S156, in communities V5, V6 and H5.

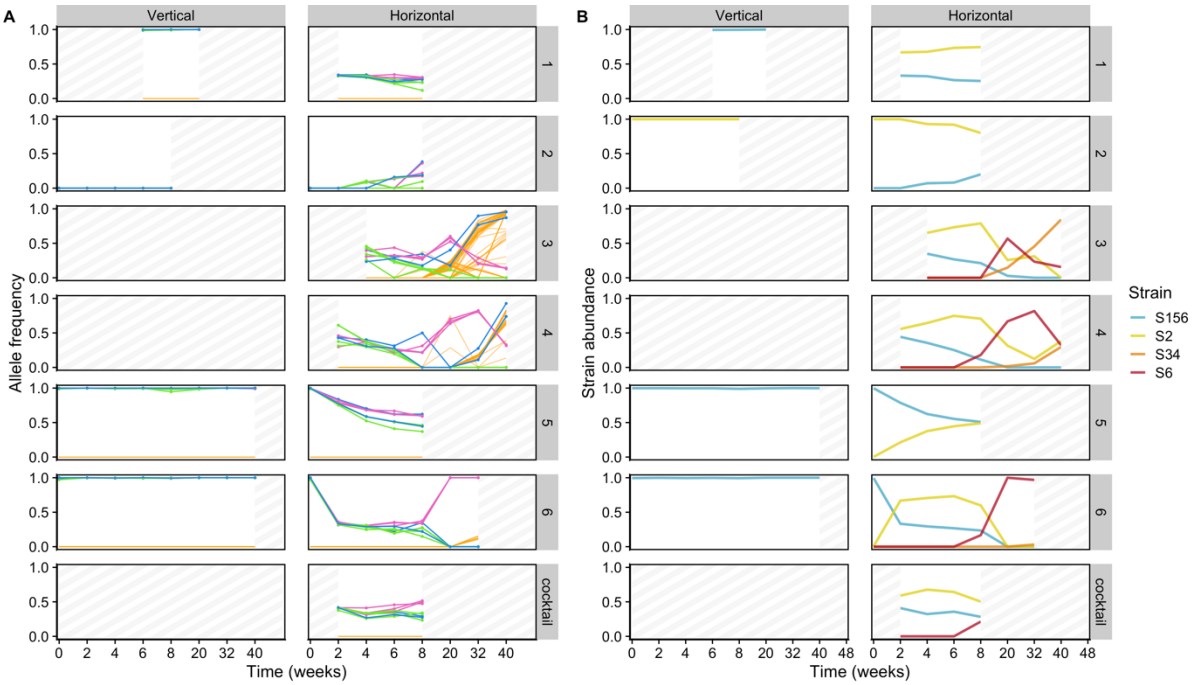


Figure S22. Same as Figure 5 of the main text but using a less stringent read coverage cutoff for alleles that characterized strains S156, S2, S34 and S6. This more sensitive analysis shows that all 9 SNPs defining strain S156 were detected in community H2, which was near the detection threshold in Figure 5 of the main text. Additionally, this analysis shows evidence that S34 also invaded community H6 at week 32.

Filtering	cutoff	mean	sd
NA genera as lineage	0	1647.894	380.6
NA genera as lineage	1E-05	**150.339**	36.851
NA genera as lineage	1E-04	46.644	9.256
NA genera as lineage	2E-04	**34.306**	6.515
NA genera as lineage	5E-04	23.733	4.468
NA genera as lineage	1	17.994	4.059
NA genera as lineage	0.01	6.061	2.362
NA and unclassified genera separate	0	2076.567	458.314
NA and unclassified genera separate	1E-05	242.261	53.37
NA and unclassified genera separate	1E-04	88.589	16.671
NA and unclassified genera separate	2E-04	67.306	11.715
NA and unclassified genera separate	5E-04	48.117	7.309
NA and unclassified genera separate	1	37.4	6.061
NA and unclassified genera separate	0.01	12.95	3.326
NA and no support removed	0	1240.944	281.18
NA and no support removed	1E-05	112.889	27.199
NA and no support removed	1E-04	38.911	7.566
NA and no support removed	2E-04	29.272	5.519
NA and no support removed	5E-04	20.528	3.985
NA and no support removed	1	15.689	3.571
NA and no support removed	0.01	5.367	2.068

Table S1. Genus counts for different abundance cutoffs (fraction of total sample reads) and filtering.

Regime	Community type	Mean	sd
H	Type A	0.692	0.398
H	Type B	0.416	0.427
V	Type A	0.677	0.463
V	Type B	0.024	0.022

Table S2. Fraction of Theomophage genome covered by at least 1 read.

cds	start	stop	length	frame	phrog	category	annotation
CDS_0001	2	133	131	+	17652	unknown function	hypothetical protein
CDS_0002	181	489	308	+	33581	unknown function	hypothetical protein
CDS_0003	874	695	179	-	24049	unknown function	hypothetical protein
CDS_0004	824	1210	386	+	4099	unknown function	hypothetical protein
CDS_0005	1300	1533	233	+	No_PHROG	unknown function	hypothetical protein
CDS_0006	1780	2379	599	+	2166	DNA, RNA and nucleotide metabolism	N4-like RNA polymerase
CDS_0007	2381	2515	134	+	30689	unknown function	hypothetical protein
CDS_0008	2512	2721	209	+	16282	unknown function	hypothetical protein
CDS_0009	2699	2887	188	+	3831	unknown function	hypothetical protein
CDS_0010	2884	3150	266	+	17804	unknown function	hypothetical protein
CDS_0011	3212	4390	1178	+	2051	DNA, RNA and nucleotide metabolism	RNA polymerase
CDS_0012	4400	4504	104	+	2400	unknown function	hypothetical protein
CDS_0013	4560	4721	161	+	16495	lysis	Rz-like spanin
CDS_0014	4851	5042	191	+	No_PHROG	unknown function	hypothetical protein
CDS_0015	5070	5186	116	+	No_PHROG	unknown function	hypothetical protein
CDS_0016	5201	5653	452	+	12011	unknown function	hypothetical protein
CDS_0017	5753	5884	131	+	32061	unknown function	hypothetical protein
CDS_0018	6096	6689	593	+	32203	unknown function	hypothetical protein
CDS_0019	6653	6865	212	+	25062	unknown function	hypothetical protein
CDS_0020	6904	7116	212	+	13263	unknown function	hypothetical protein
CDS_0021	7109	7495	386	+	21856	unknown function	hypothetical protein
CDS_0022	7482	7934	452	+	1663	DNA, RNA and nucleotide metabolism	dCTP deaminase
CDS_0023	7934	8164	230	+	14379	unknown function	hypothetical protein
CDS_0024	8310	8495	185	+	25320	unknown function	hypothetical protein
CDS_0025	8500	8760	260	+	30554	unknown function	hypothetical protein
CDS_0026	8801	9298	497	+	99	DNA, RNA and nucleotide metabolism	HNH endonuclease
CDS_0027	9323	9559	236	+	17944	unknown function	hypothetical protein
CDS_0028	9552	9725	173	+	14742	unknown function	hypothetical protein
CDS_0029	9718	10149	431	+	6472	unknown function	hypothetical protein
CDS_0030	10275	10724	449	+	323	transcription regulation	transcriptional regulator, D5-like
CDS_0031	10663	11058	395	+	17693	unknown function	hypothetical protein
CDS_0032	11045	11257	212	+	29710	unknown function	hypothetical protein
CDS_0033	11254	12225	971	+	374	DNA, RNA and nucleotide metabolism	HNH endonuclease
CDS_0034	12229	12414	185	+	35026	unknown function	hypothetical protein
CDS_0035	12411	12542	131	+	9602	unknown function	hypothetical protein
CDS_0036	12547	13272	725	+	85	DNA, RNA and nucleotide metabolism	thymidylate synthase
CDS_0037	13386	13979	593	+	28714	unknown function	hypothetical protein
CDS_0038	14003	14101	98	+	No_PHROG	unknown function	hypothetical protein
CDS_0039	14140	14763	623	+	139	other	deoxynucleoside monophosphate kinase
CDS_0040	14754	14918	164	+	7051	unknown function	hypothetical protein
CDS_0041	14928	15608	680	+	99	DNA, RNA and nucleotide metabolism	HNH endonuclease
CDS_0042	15729	16520	791	+	5534	unknown function	hypothetical protein
CDS_0043	16535	17530	995	+	1297	other	ATPase
CDS_0044	17532	18392	860	+	21018	unknown function	hypothetical protein
CDS_0045	18367	19566	1199	+	1314	DNA, RNA and nucleotide metabolism	HNH endonuclease
CDS_0046	19570	19824	254	+	34165	unknown function	hypothetical protein
CDS_0047	19821	20318	497	+	1770	DNA, RNA and nucleotide metabolism	Holliday junction resolvase
CDS_0048	20315	20776	461	+	671	other	nucleotide pyrophosphohydrolase
CDS_0049	20769	22001	1232	+	325	DNA, RNA and nucleotide metabolism	Dda-like helicase

CDS_0050	22002	23924	1922	+	17	DNA, RNA and nucleotide metabolism	DNA polymerase
CDS_0051	23791	24474	683	+	59	DNA, RNA and nucleotide metabolism	HNH endonuclease
CDS_0052	24458	25198	740	+	2573	DNA, RNA and nucleotide metabolism	DNA polymerase
CDS_0053	25255	26211	956	+	2139	DNA, RNA and nucleotide metabolism	exonuclease
CDS_0054	26204	28354	2150	+	2135	DNA, RNA and nucleotide metabolism	DNA primase
CDS_0055	28347	29093	746	+	124	DNA, RNA and nucleotide metabolism	Sak4-like ssDNA annealing protein
CDS_0056	29093	29803	710	+	743	DNA, RNA and nucleotide metabolism	single strand DNA binding protein
CDS_0057	29803	29982	179	+	11455	unknown function	hypothetical protein
CDS_0058	29986	30117	131	+	No_PHROG	unknown function	hypothetical protein
CDS_0059	30110	30586	476	+	872	unknown function	hypothetical protein
CDS_0060	30687	31100	413	+	31223	unknown function	hypothetical protein
CDS_0061	31097	31315	218	+	13619	unknown function	hypothetical protein
CDS_0062	41604	31480	10124	-	2206	DNA, RNA and nucleotide metabolism	virion RNA polymerase
CDS_0063	43360	41612	1748	-	11089	unknown function	hypothetical protein
CDS_0064	43739	43362	377	-	16623	unknown function	hypothetical protein
CDS_0065	46688	43752	2936	-	2220	unknown function	hypothetical protein
CDS_0066	47401	46688	713	-	2050	head and packaging	virion structural protein
CDS_0067	47998	47468	530	-	2130	unknown function	hypothetical protein
CDS_0068	49523	48048	1475	-	1357	head and packaging	major head protein
CDS_0069	50585	49539	1046	-	2120	tail	tail length tape measure protein
CDS_0070	50910	50578	332	-	2346	unknown function	hypothetical protein
CDS_0071	53035	50912	2123	-	1216	head and packaging	portal protein
CDS_0072	53265	53038	227	-	11498	unknown function	hypothetical protein
CDS_0073	53507	53262	245	-	1048	unknown function	hypothetical protein
CDS_0074	53986	53504	482	-	17078	unknown function	hypothetical protein
CDS_0075	54289	53999	290	-	1356	lysis	holin
CDS_0076	54960	54343	617	-	7	lysis	endolysin
CDS_0077	56104	54953	1151	-	22896	tail	tail protein
CDS_0078	56315	56106	209	-	12784	unknown function	hypothetical protein
CDS_0079	56657	56319	338	-	16814	unknown function	hypothetical protein
CDS_0080	57466	56657	809	-	16219	unknown function	hypothetical protein
CDS_0081	57940	57470	470	-	644	unknown function	hypothetical protein
CDS_0082	58349	57942	407	-	27774	unknown function	hypothetical protein
CDS_0083	59068	58349	719	-	2125	head and packaging	virion structural protein
CDS_0084	61697	59070	2627	-	19299	head and packaging	terminase large subunit
CDS_0085	62380	61697	683	-	2109	unknown function	hypothetical protein
CDS_0086	62517	62422	95	-	No_PHROG	unknown function	hypothetical protein
CDS_0087	62820	62551	269	-	20421	unknown function	hypothetical protein
CDS_0088	62957	63145	188	+	10929	unknown function	hypothetical protein
tRNA	31306	31401	95	+	No_PHROG	tRNA	tRNA1-LeuCAA

Table S3. Predicted CDs and function.

Set	POS	REF	var	annotation	category	gene	phrog
shared_HA_VA	55378	G	A	tail protein	tail	77	22896
unique_HA	55482	G	C	tail protein	tail	77	22896
unique_HB	16296	T	A	hypothetical protein	unknown function	42	5534
unique_HB	46679	C	T	hypothetical protein	unknown function	65	2220
unique_HB	48350	T	C	major head protein	head and packaging	68	1357
unique_HB	49134	A	C	major head protein	head and packaging	68	1357
unique_HB	49134	A	T	major head protein	head and packaging	68	1357
unique_HB	51895	G	T	portal protein	head and packaging	71	1216
unique_HB	55582	C	T	tail protein	tail	77	22896
unique_HB	55889	A	G	tail protein	tail	77	22896
unique_HB	55904	C	T	tail protein	tail	77	22896
unique_HB	55934	C	A	tail protein	tail	77	22896
unique_HB	55937	C	T	tail protein	tail	77	22896
unique_HB	55940	T	C	tail protein	tail	77	22896
unique_HB	55949	T	C	tail protein	tail	77	22896
unique_HB	55954	T	C	tail protein	tail	77	22896
unique_HB	55955	G	A	tail protein	tail	77	22896
unique_HB	57383	T	C	hypothetical protein	unknown function	80	16219
unique_HB	57385	G	A	hypothetical protein	unknown function	80	16219
unique_HB	57394	T	C	hypothetical protein	unknown function	80	16219
unique_HB	57397	T	C	hypothetical protein	unknown function	80	16219
unique_HB	57400	T	C	hypothetical protein	unknown function	80	16219
unique_HB	57403	C	T	hypothetical protein	unknown function	80	16219
unique_HB	57407	T	C	hypothetical protein	unknown function	80	16219
unique_HB	57415	A	G	hypothetical protein	unknown function	80	16219
unique_HB	57428	A	T	hypothetical protein	unknown function	80	16219
unique_HB	57443	A	C	hypothetical protein	unknown function	80	16219
unique_HB	57452	G	A	hypothetical protein	unknown function	80	16219
unique_HB	57453	A	G	hypothetical protein	unknown function	80	16219
unique_HB	57454	T	C	hypothetical protein	unknown function	80	16219
unique_HB	57460	C	T	hypothetical protein	unknown function	80	16219
unique_HB	57530	T	G	hypothetical protein	unknown function	81	644
unique_HB	57533	C	T	hypothetical protein	unknown function	81	644
unique_HB	57538	T	C	hypothetical protein	unknown function	81	644
unique_HB	57545	C	A	hypothetical protein	unknown function	81	644
unique_HB	57548	G	A	hypothetical protein	unknown function	81	644
unique_HB	57551	T	A	hypothetical protein	unknown function	81	644
unique_HB	57560	A	G	hypothetical protein	unknown function	81	644
unique_HB	57563	C	T	hypothetical protein	unknown function	81	644
unique_HB	57566	C	A	hypothetical protein	unknown function	81	644
unique_HB	57572	A	T	hypothetical protein	unknown function	81	644
unique_HB	57578	G	A	hypothetical protein	unknown function	81	644
unique_HB	57584	A	G	hypothetical protein	unknown function	81	644
unique_HB	57593	C	T	hypothetical protein	unknown function	81	644
unique_HB	57599	G	A	hypothetical protein	unknown function	81	644
unique_HB	57602	A	T	hypothetical protein	unknown function	81	644
unique_HB	57605	A	T	hypothetical protein	unknown function	81	644
unique_HB	57611	T	A	hypothetical protein	unknown function	81	644
unique_HB	57617	G	A	hypothetical protein	unknown function	81	644
unique_HB	57638	C	T	hypothetical protein	unknown function	81	644
unique_HB	57642	A	G	hypothetical protein	unknown function	81	644
unique_HB	57653	A	G	hypothetical protein	unknown function	81	644
unique_HB	57656	A	G	hypothetical protein	unknown function	81	644
unique_HB	57975	T	A	hypothetical protein	unknown function	82	27774
unique_HB	57975	T	C	hypothetical protein	unknown function	82	27774
unique_HB	57996	G	A	hypothetical protein	unknown function	82	27774
unique_HB	58050	C	T	hypothetical protein	unknown function	82	27774
unique_HB	58077	T	A	hypothetical protein	unknown function	82	27774
unique_HB	58092	G	A	hypothetical protein	unknown function	82	27774
unique_HB	58101	C	T	hypothetical protein	unknown function	82	27774
unique_HB	58302	T	C	hypothetical protein	unknown function	82	27774
unique_HB	57467	G	A	NA	NA	-	NA
unique_HB	57468	T	G	NA	NA	-	NA

unique_VA	22232	A	G	DNA polymerase	RNA and nucleotide metabolism	50	
unique_VA	55321	C	T	tail protein	tail	77	22896
unique_VA	55362	G	T	tail protein	tail	77	22896
unique_VA	55510	C	T	tail protein	tail	77	22896

Table S4. 67 Single Nucleotide Polymorphisms observed in Theomophage populations. The first column indicates in which community type and regime the SNP is observed.

AAI to closest RaFAH reference	Host genus	Confidence score	List of methods
20.59	d Bacteria;p Patescibacteria;c Saccharimonadia;o Saccharimonadales;f UBA4665;g UBA5150	81.00	iPHoP-RF;82.20

0 **Table S5.** iPHoP prediction for Theomophage MAG NODE_224_length_63535_cov_5745.934389

1

2

Method	Host	Host taxonomy	Metric 1	Score 1	Metric 2	Score 2	Rank	Host representative
crispr	GCA_013298545.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__AKYH767;f__Palsa-965;g__GCA-2737665;s__GCA-2737665 sp013298545	N_mismatches	3.000	Spacer_length	30.0	1	GB_GCA_013298545.1
crispr	GCF_900625095.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__Acinetobacter sp900625095	N_mismatches	3.000	Spacer_length	32.0	1	RS_GCF_900625095.1
iPHoP-RF	GB_GCA_003153395.1	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__UBA4665;g__UBA5150;s__UBA5150 sp003153395	Confidence_score	0.906	FDR	0.178	1	GB_GCA_003153395.1
iPHoP-RF	GB_GCA_003280165.1	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__PCC-6307;f__Cyanobiaceae;g__Prochlorococcus_A;s__Prochlorococcus_A sp003280165	Confidence_score	0.894	FDR	0.194	2	GB_GCA_003280165.1
iPHoP-RF	3300025405_18	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__UBA4665;g__;s	Confidence_score	0.848	FDR	0.25	3	3300025405_18
iPHoP-RF	3300023184_106	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__UBA4665;g__;s__	Confidence_score	0.824	FDR	0.276	4	3300023184_106
iPHoP-RF	3300025221_12	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SAR86;f__SAR86;g__;s__	Confidence_score	0.820	FDR	0.281	5	3300025221_12
php	GCA_014237605.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Nitrosomonadaceae;g__GCA-2721545;s__GCA-2721545 sp014237605	Score	1445.116	NA	NA	1	GB_GCA_014237605.1
php	3300013138_179	d__Bacteria;p__Patescibacteria;c__Dojkabacteria;o__B142;f__;g__;s__	Score	1443.065	NA	NA	2	3300013138_179
php	GCA_013214975.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__NORP154;g__DT-38;s__DT-38 sp013214975	Score	1442.799	NA	NA	3	GB_GCA_013214975.1
php	GCA_003214355.1	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__PCC-6307;f__Cyanobiaceae;g__Prochlorococcus_A;s__Prochlorococcus_A sp003214355	Score	1442.114	NA	NA	4	GB_GCA_003214355.1
php	GCA_002730985.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__BACL11;g__BACL11;s__BACL11 sp002730985	Score	1442.057	NA	NA	5	GB_GCA_002730985.1
rafah	Escherichia	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Escherichia	Confidence_score	0.095	FDR	0.915	1	NA
rafah	Haemophilus	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Pasteurellaceae;g__Haemophilus	Confidence_score	0.055	FDR	0.956	2	NA
rafah	Moraxella	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Moraxella	Confidence_score	0.049	FDR	0.962	3	NA
rafah	Vibrio	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Vibrionaceae;g__Vibrio	Confidence_score	0.044	FDR	0.967	4	NA
rafah	Sandarakinorhabdus	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sandarakinorhabdus	Confidence_score	0.039	FDR	0.972	5	NA
vhm	GCA_001628005.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SAR86;f__SAR86;g__AEGEAN-183;s__AEGEAN-183 sp001628005	S2star_similarity	0.496	NA	NA	1	GB_GCA_001628005.1

vhm	3300013126_131	d__Bacteria;p__Patescibacteria;c__WWE3;o__UBA101185;f__UBA10185;g__;s__	S2star_similarity	0.488	NA	NA	2	3300013126_131
vhm	GCA_012965075.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SAR86;f__SAR86;g__AEGEAN-183;s__AEGEAN-183 sp012965075	S2star_similarity	0.476	NA	NA	3	GB_GCA_012965075.1
vhm	GCF_009711055.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Myroides;s__Myroides pelagicus	S2star_similarity	0.472	NA	NA	4	RS_GCF_009711055.1
vhm	3300020243_9	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SAR86;f__SAR86;g__AEGEAN-183;s__	S2star_similarity	0.469	NA	NA	5	3300020243_9
wish	GCF_900109075.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Myroides;s__Myroides marinus	Log-likelihood	-1.326	p-value	0.0398269	1	RS_GCF_900109075.1
wish	GCF_002964605.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Vallitaleaceae_A;g__Vallitalea_A;s__Vallitalea_A okinawensis	Log-likelihood	-1.328	p-value	0.0536214	2	RS_GCF_002964605.1
wish	GCF_001485415.1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Myroides;s__Myroides odoratimimus	Log-likelihood	-1.329	p-value	0.0449743	3	RS_GCF_001485415.1
wish	GCF_009928445.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__CT4;s__CT4 sp009928445	Log-likelihood	-1.329	p-value	0.0505899	4	RS_GCF_009928445.1
wish	GCF_003313305.1	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Vallitaleaceae;g__Vallitalea;s__Vallitalea guaymasensis	Log-likelihood	-1.329	p-value	0.0489162	5	RS_GCF_003313305.1

Table S6. Host predictions for individual virus-host prediction tools included in iPHoP. Top 5 hits shown for Theomophage MAG NODE_224_length_63535_cov_5745.934389. Score 1 / Score 2 columns correspond respectively to: For BLAST hits => Total number of matches in hsps / Average percentage of identity across the full alignment. For crispr hits => Number of mismatches / Length of CRISPR spacer. For WIsH => Log-likelihood / P-value. For VirHostMatcher => s2 similarity/ NA. For PHP => PHP score/ NA. For RaFAH => Confidence score / estimated FDR. For iPHoP-RF => Confidence score / estimated FDR

10

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore	Annotation
YP_007005830.1	NODE_224_length_63535_cov_5745_934389	26.667	90	51	4	140	220	62411	62662	4.0	21.9	Hypothetical protein
YP_007005830.1	NODE_224_length_63535_cov_5745_934389	41.667	24	14	0	563	586	61296	61367	6.4	21.6	Terminase large subunit
YP_007005830.1	NODE_224_length_63535_cov_5745_934389	42.105	19	11	0	392	410	35853	35797	7.2	21.2	Virion RNA polymerase
YP_007005830.1	NODE_224_length_63535_cov_5745_934389	34.615	26	17	0	403	428	60680	60603	8.4	21.2	Terminase large subunit
YP_007005830.1	NODE_224_length_63535_cov_5745_934389	34.146	41	27	0	609	649	25213	25335	9.5	20.8	DNA polymerase
YP_007005832.1	NODE_224_length_63535_cov_5745_934389	29.412	17	12	0	128	144	28653	28703	5.5	20.0	Sak4-like ssDNA annealing protein
YP_007005832.1	NODE_224_length_63535_cov_5745_934389	42.105	19	11	0	34	52	16957	16901	6.0	20.0	ATPase
YP_007005832.1	NODE_224_length_63535_cov_5745_934389	36.667	30	17	1	150	177	56035	55946	7.9	19.6	Tail protein
YP_007005832.1	NODE_224_length_63535_cov_5745_934389	35.714	28	17	1	70	97	56297	56217	8.3	19.6	Hypothetical protein

11

12 **Table S7.** tBLASTn hits of S6 cellulase proteins ORF94 (YP_007005830.1) and ORF96 (YP_007005832.1). ORF95 resulted in no significant hits.