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

Jeroen Meijer[1,4*], 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, Ecology, and Evolution, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University Jena, 07743, Jena, Germany

* Corresponding authors: jeroen.meijer@uni-jena.de; b.e.dutilh@uni-jena.de

Competing Interest Statement: The authors declare no competing interest.

Abstract

Bacteriophages are important drivers of microbial ecosystems, yet their dynamics in complex natural communities remain poorly understood compared to simplified laboratory systems. To address this gap, we analyzed viral dynamics across 20 compost-derived microbial communities propagated for one year in mesocosms. Communities clustered into two distinct types, each dominated by different cellulose-degraders and comprising hundreds of genera. In one type, we observed massive, parallel outbreaks of Theomophage, a previously undescribed bacteriophage, reaching 74% of metagenomic sequencing reads, the largest bacteriophage outbreak documented to date. Theomophage populations in isolated communities were composed of a single genotype that showed striking evolutionary stability throughout the experiment. In contrast, following experimental viral migration between mesocosms, all Theomophage populations showed rapid evolution via recombination between preexisting genotypes, replacement of ancestral lineages, and, upon successful migration to mesocosms of the alternate community type where Theomophage was initially absent, rapid acquisition of novel mutations that swept local populations and then spread to other mesocosms. Our study reveals the spatial and temporal scales at which bacteriophage microdiversity evolves in complex communities. It further shows that mixing of viral communities — likely common in natural systems — can rapidly accelerate bacteriophage evolution.

Introduction

Viruses that infect bacteria (bacteriophages or phages) are ubiquitous and ecologically influential components of all microbial ecosystems, from oceans and soils to host-associated microbiomes. They regulate bacterial populations, drive nutrient cycling, and shape bacterial evolution by facilitating horizontal gene transfer[1]. Phages may also shape community structure and function, though evidence is mixed[2]. These ecological and evolutionary effects arise from highly specific interactions between phages and their bacterial hosts, which are typically shaped by antagonistic coevolution[3–6]. Identifying the biotic and abiotic factors that govern phage eco-evolutionary dynamics is essential for predicting microbial community diversity, stability, and function, and for guiding applied interventions such as bacteriophage therapy[7].

Current knowledge of phage ecology and evolution is largely shaped by two contrasting study designs: simplified *in vitro* experimental systems in nutrient rich media, typically consisting of one or a few phage-host pairs, and observational surveys of complex natural ecosystems using metagenomics or large-scale isolation and phenotyping[8–10]. Experiments with single phage-host pairs show coevolution proceeding via series of selective sweeps, in which bacterial receptor modification or loss confers resistance to infection, countered by compensatory changes in phage tail proteins[4,5,11]. In contrast, studies of natural ecosystems showed that bacterial resistance is shaped more by variation in defense systems than by receptor evolution[3,12,13], and often reveal patterns of microdiversity suggesting long-term coexistence of resistant and susceptible hosts[14,15]. These differences indicate that selective pressures in simplified laboratory conditions differ fundamentally from those in complex natural environments. More broadly, our ability to extrapolate from laboratory models to predict viral dynamics in complex environments remains limited[9,10,16], prompting calls for experimental models that incorporate greater ecological complexity while retaining experimental tractability[8–10,17].

Several features of natural ecosystems likely contribute to these differences[9,10,17]. Natural communities harbor extensive macro- (taxonomic) and micro- (strain-level) diversity, in contrast to the isogenic populations typical of lab experiments. In diverse communities, genetic novelty arises not only through mutation, but also recombination and horizontal gene transfer (HGT). Increasing macrodiversity beyond a single phage-host pair can change both ecological and evolutionary trajectories (recently reviewed in [9,17]). Natural ecosystems are also spatially heterogeneous, comprising fragmented patches (e.g. skin pores, gut crypts, marine snow, soil particles) connected by restricted dispersal, which produces local populations linked intermittently by migration[18]. Such metapopulation dynamics are inherently different from those in homogeneous laboratory systems[19–21].

Microbial communities derived from natural ecosystems and propagated under laboratory conditions provide a promising middle ground between simplified lab systems and the full complexity of natural environments[22–27]. When grown on complex carbohydrates they maintain high taxonomic and functional diversity yet are amenable to longitudinal sampling and controlled perturbation. These systems have been used to

83 investigate bacterial community assembly[22–26], function[24], and strain-level interactions[27],
84 but applications to viruses remain limited[28,29]. Here, using compost-derived microbial
85 communities grown on paper, we investigate how standing microdiversity, bacteriophage
86 migration, and community context shape bacteriophage eco-evolutionary dynamics. By
87 integrating genotype-resolved viromics with a tractable yet complex system, we show
88 that bacteriophage evolution can range from stasis to rapid change, depending on
89 opportunities for migration.

Results

Compost communities cluster in two distinct and stable community types

To track the eco-evolutionary dynamics of bacteriophages in a complex system at high temporal resolution, we analyzed 184 shotgun metagenomes from a long-term mesocosm experiment with compost communities[30]. For clarity, we summarize the experimental setup. Ten independent founding communities were established by sampling 1 g of garden compost from different locations on a compost heap. Each community was incubated in mesocosms containing 20 mL nitrogen-limited minimal M9 medium supplemented with cellulose (4 cm^2 paper) as the sole carbon source (**Fig.1a**). Following two weeks of incubation, 1 mL slurry was transferred to 19 mL fresh medium and cellulose paper, then incubated for another two weeks. After this acclimatization period (week 0), two serial transfer regimes were created from each mesocosm leading to 2x10 paired communities. In the closed regime, every 2 weeks 1 mL of cellulosic slurry was transferred to 19 mL fresh medium with cellulose paper. In the open regime, transfers were conducted identically, except that at every transfer, 1 mL of 0.2 μm-filtered viral fractions (hereafter the Mobile Genetic Element, MGE cocktail) were collected, pooled, and redistributed among all 10 open mesocosms. This design allowed migration of bacteriophages and other mobile genetic elements between open mesocosms, repeatedly exposing them to new hosts and potentially affecting their eco-evolutionary dynamics. Shotgun sequencing was performed at weeks 0, 2, 4, 6, 8, 20, 32, 40, and 48 for both regimes, with additional sequencing of the MGE cocktail at weeks 2, 4, 6, and 8. This enabled us to track viral migration via the MGE cocktail, and compare cellular and viral dynamics in closed mesocosms with their paired open counterparts.

We began by assembling the metagenomes and characterizing community composition, to provide context for viral eco-evolutionary dynamics. Taxonomic annotation of contigs and bins was performed using CAT[31] and taxon relative abundances were calculated with RAT[32] (**Fig.1b**, median 96.1% of sample reads mapped, **Fig.S1-2**; Methods). Consistent with previous studies[25,30], this revealed diverse communities with mesocosms containing on average 150.3 ± 36.9 genera (mean ± standard deviation, mean 58.2% of sample reads annotated at genus rank, **Table S1; Fig.S1**). Principal component analysis (PCA) of genus-rank community composition revealed two distinct community types separated along the first component (**Fig.1bd**). Mesocosms closed_1, closed_2, closed_5, and closed_6 were dominated by genera including *Rhodanobacter* and *Nitrosomonas* (hereafter community type A), whereas the remaining 6 closed mesocosms were enriched in *Cellvibrio* and *Azospirillium* (community type B; **Fig.S4-5**). These community types were also apparent at higher taxonomic ranks and remained stable over 48 weeks, with dynamics primarily along the second PC (**Fig.S6-7**).

The two distinct community types were consistent across all paired closed and open mesocosms, suggesting a strong founder effect (**Fig.1cd**). This may reflect local sub-communities already present in different regions of the compost heap at the time of sampling, or alternatively, two distinct ecological attractors that emerged during the acclimatization period preceding the first sequenced time point. Stochastic diversity loss

during acclimatization of natural communities to lab conditions can result in divergent communities and multistability[24–26]. Biopolymer degrading communities are structured by hierarchical cross-feeding and substrate preferences, with different degraders producing distinct metabolite profiles that influence downstream community assembly of specialized byproduct consumers[23]. We hypothesize that community type A and B are structured around different cellulose degraders, as *Cellvibrio* and *Rhodanobacter* are both associated with cellulose degradation and MAGs of these genera assembled from this dataset contain cellulases[33–35]. However, since cellulases are a highly heterogeneous enzyme group and their predicted genomic presence does not always indicate cellulolytic activity[36], this remains to be experimentally validated.

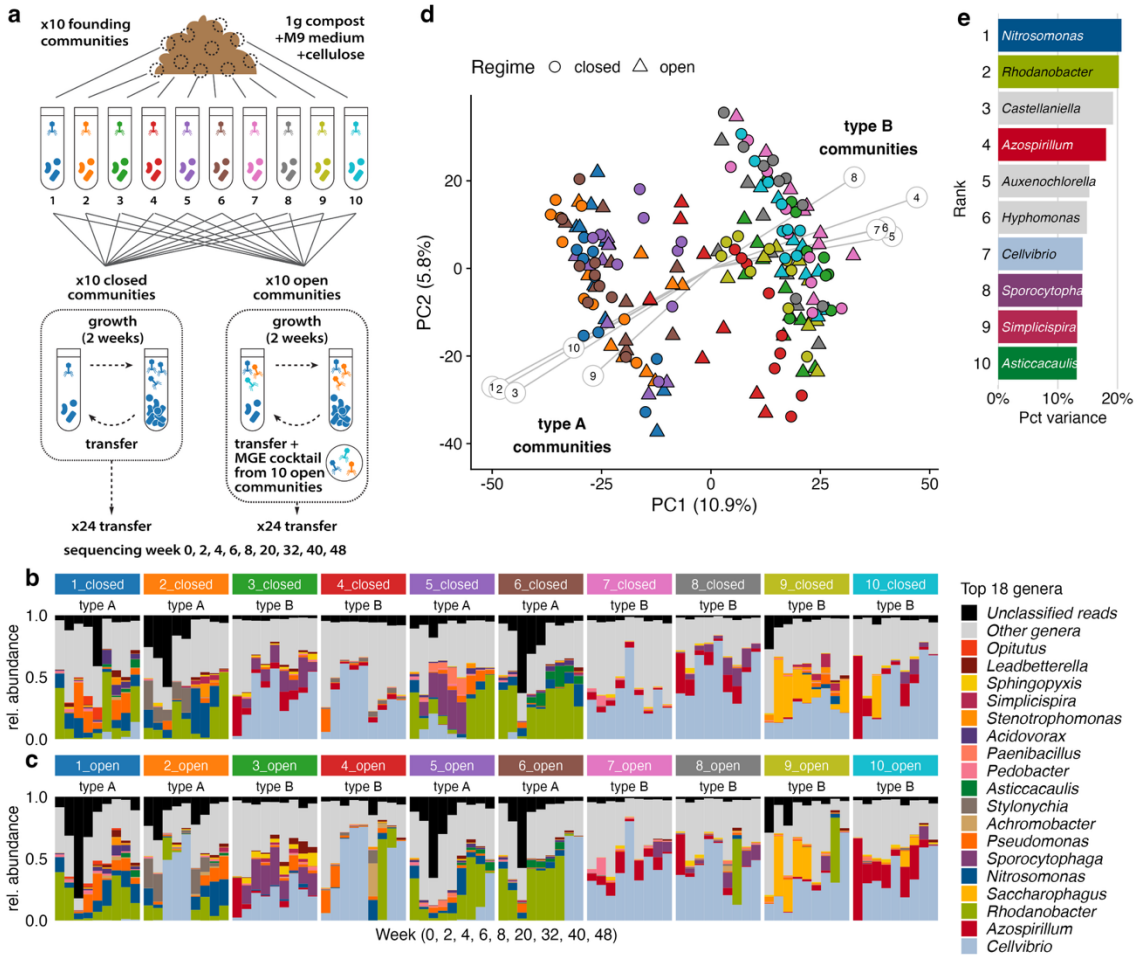


Figure 1. Compost communities cluster in two stable, distinct community types. a. *Experimental design of Quistad et al[30]. Ten founding communities were established from garden compost and propagated on cellulose paper by serial transfer (2 weeks) for 48 weeks. In the closed regime communities were propagated in isolation. In the open regime, a mobile genetic element cocktail was added at every transfer, consisting of pooled, 0.2 μm-filtered material from all open communities (Methods).* ***b*** *Relative*

*abundance (fraction of mapped metagenome reads) of the 18 most abundant genera in closed mesocosms (annotated with RAT). Unclassified reads (black) often represent viral reads (see **Fig.2**). **c** same as (b), for open mesocosms. **d** Principal component analysis (PCA) of centered log-ratio transformed genus-level relative abundances reveals two distinct community types along PC1: type A, rhodanobacter-dominated and type B, cellvibrio-dominated. Community type of mesocosms is stable over time and consistent across paired closed and open mesocosms, suggesting a strong founder effect. Grey lines indicate the top 10 genera contributing most to PC1 and PC2, numbered as in (e). For separate PCAs of closed and open regimes see Fig.S8. **e** Proportion of variance in PC1 and PC2 explained by the top 10 contributing genera. Numbers correspond to (d) and colors match (b) and (c).*

Massive, parallel outbreaks of a novel *Schitovirus* in type A communities

To investigate the viral dynamics in the compost mesocosms, we screened the assembled metagenomes using WhatThePhage[37], a pipeline that integrates twelve virus identification tools (Methods). This analysis identified 140,378 predicted viral contigs, including 33 complete genomes, 131 high-quality (>90% complete), and 234 medium-quality genomes (50-90% complete, estimated with CheckV[38]). Among these, one 63,535 bp viral contig (estimated 100% complete) stood out due to its high abundance. Further analysis identified this contig as a representative of a new subfamily in the *Schitoviridae* bacteriophage family, which we named Theomophage after its place of origin, Place Théodore-Monod in Paris, France[30] (see section "Theomophage represents a new *Schitoviridae* subfamily"). Between weeks 4-8 of the experiment, massive outbreaks of Theomophage occurred in all closed and open type-A mesocosms, with no outbreak detected in any type-B mesocosm (**Fig.2a**). These outbreaks reached abundances of up to 74.3% of total community reads (**Fig.2a**). Notably, outbreaks in the open mesocosms were often larger or occurred 2-6 weeks earlier than in their paired closed mesocosm (**Fig.2**), suggesting that Theomophage migration between open mesocosms accelerated and intensified outbreaks. Note that unclassified reads, including Theomophage reads, were not included in the PCA that delineated type-A and type-B communities, so Theomophage outbreaks did not drive the separation between community types.

To determine whether Theomophage was present in type-B communities, we analyzed its genome coverage across all samples (**Fig.2b**; Methods). In closed type-B mesocosms Theomophage was absent or very rare, with no or very low genome coverage (type A: 69.2% ± 38.9%; type B 2.44% ± 2.18% genome breadth of coverage ≥1 read; mean ± SD; **Fig.S9, Table S2**). In contrast, Theomophage was detected in 5 of 6 open type-B communities, where it persisted until week 40 and reached up to 4.5% of total community reads (mesocosms closed_3, closed_4, closed_7-9, **Fig.2d**). This indicates that Theomophage, originally present in type-A communities, successfully colonized open type-B mesocosms via migration. Supporting this, the complete Theomophage genome was assembled from and detected in all sequenced MGE cocktail samples (**Fig.S10**). Despite successful colonization, Theomophage abundances were an order of magnitude lower in type-B communities than in its the native type-A context, indicating a strong

effect of community type on ecological success. This observed lower abundance could be due to several factors. Host availability may have been lower in type-B mesocosms, or hosts may have exhibited higher levels of resistance. It is also possible that the original host of Theomophage was absent from type B, forcing it to infect a new, sub-optimal host. Although unlikely, due to the compositional nature of metagenomics, higher abundances of other taxa in type-B mesocosms could have reduced the relative abundance estimates of Theomophage, despite similar absolute abundances.

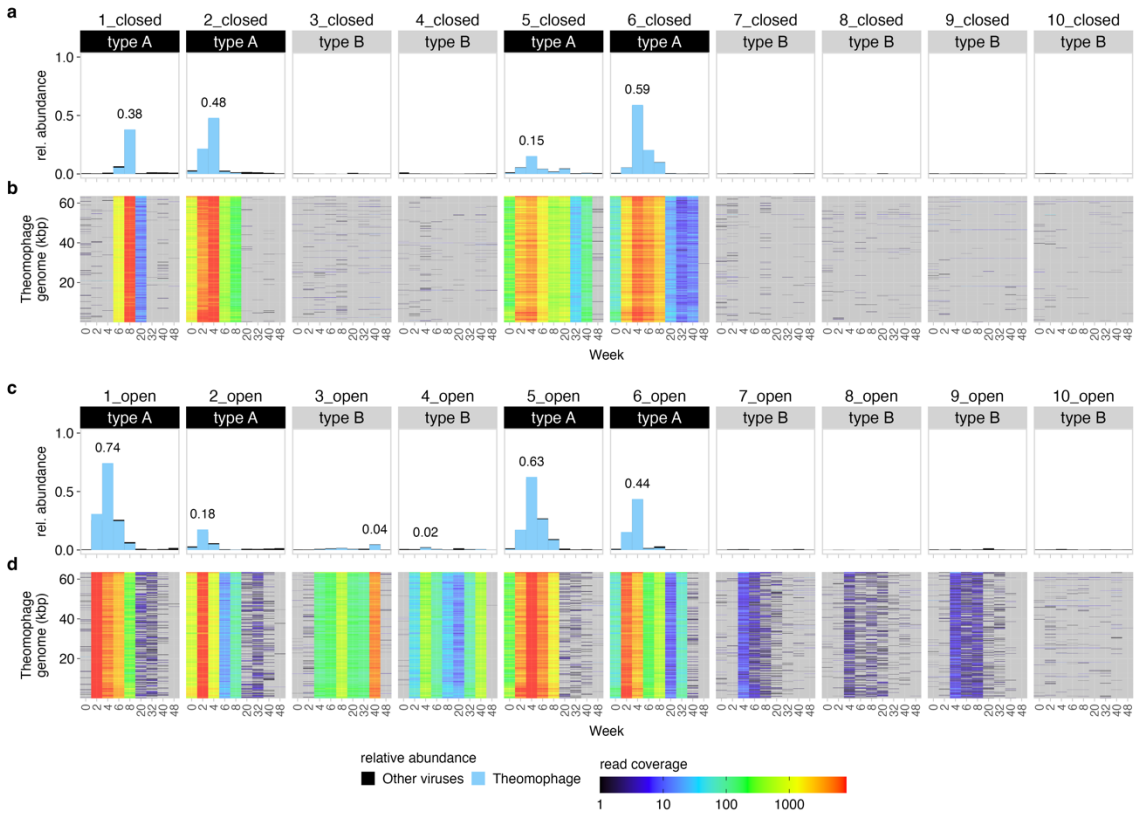


Figure 2. Massive outbreaks of a Theomophage, a novel Schitoviridae bacteriophage, are linked to community background. ***a*** *Relative abundance of all predicted viral sequences (black) and Theomophage (blue), a highly abundant, previously undescribed bacteriophage belonging to the Schiroviridae family. Theomophage accounted for up to 59,1% and 74,3% of total community reads in closed (a) and open (c) samples, respectively. Read fractions >1% are labeled.* ***b*** *Genome coverage profiles show the percentage of sample nucleotides mapped to each position of the Theomophage genome. Theomophage was natively present in all closed type-A communities (1_closed, 2_closed, 5_closed, 6_closed), but absent or very rare in all closed type-B mesocosms. Grey indicates no mapped reads.* ***c*** *same as (b) for 10 open mesocosms.* ***d*** *same as (b) but for open mesocosms. Theomophage was detected in 5 out of 6 type-B mesocosms in the open regime, where it was absent in the corresponding paired closed mesocosms, indicating successful invasion in the open regime.*

222
223 ## Theomophage represents a new *Schitoviridae* subfamily
224 To determine the taxonomic affiliation of Theomophage, we constructed a gene sharing
225 network using vContTACT2[39] and the INPHARED viral reference set containing 26,302
226 bacterial and archaeal viral genomes (**Fig.3b**; Methods). Based on shared gene content,
227 Theomophage grouped with *Schitoviridae* (formerly known as N4-like phages), a family of
228 double-stranded DNA bacteriophages with podovirus-like morphology that are found
229 globally in various habitats[40]. The Theomophage genome is organized in two large blocks
230 on opposing strands, as is typical for *Schitoviridae*, and has 88 predicted genes and a single
231 tRNA, including all seven proposed *Schitoviridae* hallmark genes[41] (**Fig.3a**; **Table S3**, based
232 on Pharokka[42] and Phold[43]; Methods). Individual gene phylogenies of TerL, MCP, and
233 portal protein also place Theomophage among the *Schitoviridae*, further supporting this
234 taxonomic assignment (**Fig.S11-13**).
235
236 To assess the phylogenetic placement of Theomophage within *Schitoviridae*, we analyzed
237 573 high-quality (>90% complete) *Schitoviridae* genomes and MAGs from three recent
238 studies[40,41,44]. Using VIRIDIC[45] to calculate intergenomic similarity (Methods), we found
239 that the closest relative of Theomophage is IMGVR_UViG_3300045988_071216, a 72 kbp
240 MAG co-assembled from human fecal samples, with only 8.98% intergenomic similarity
241 (**Fig.S14**). This falls below the ≥20% threshold proposed for *Schitoviridae* subfamilies[41],
242 indicating that Theomophage represents a new subfamily. Gene-sharing analysis further
243 supported this classification. Using vConTACT2, we found that Theomophage did not
244 group with any of the 573 *Schitoviridae* genomes in vConTACT2 subclusters, which align
245 with genus-level taxonomy as defined by the International Committee on Taxonomy of
246 Viruses[39] (**Fig.S15**; Methods). These results show that Theomophage is divergent from
247 previously characterized *Schitoviridae* and represents a novel subfamily. To predict
248 Theomophage lifestyle we used BACPHLIP[46], which classified it as lytic (score 0.9625),
249 which is common for Schitoviridae[41]. This is consistent with CheckV, which reported no
250 bacterial flanking regions for any assembled Theomophage contig.
251
252 Cellulase activity has been reported for *Schitovirus* vB_EamP-S6 virions[47], prompting us to
253 investigate whether the high abundance of Theomophage in the cellulose-degrading
254 communities might be related to cellulolytic capability. We queried the Theomophage
255 genome using tblastn with vB_EamP-S6 cellulase genes Gp94, Gp95, and Gp96. No
256 significant matches were found (**Table S11**), suggesting that the ecological success of
257 Theomophage is not linked to cellulose degradation.
258

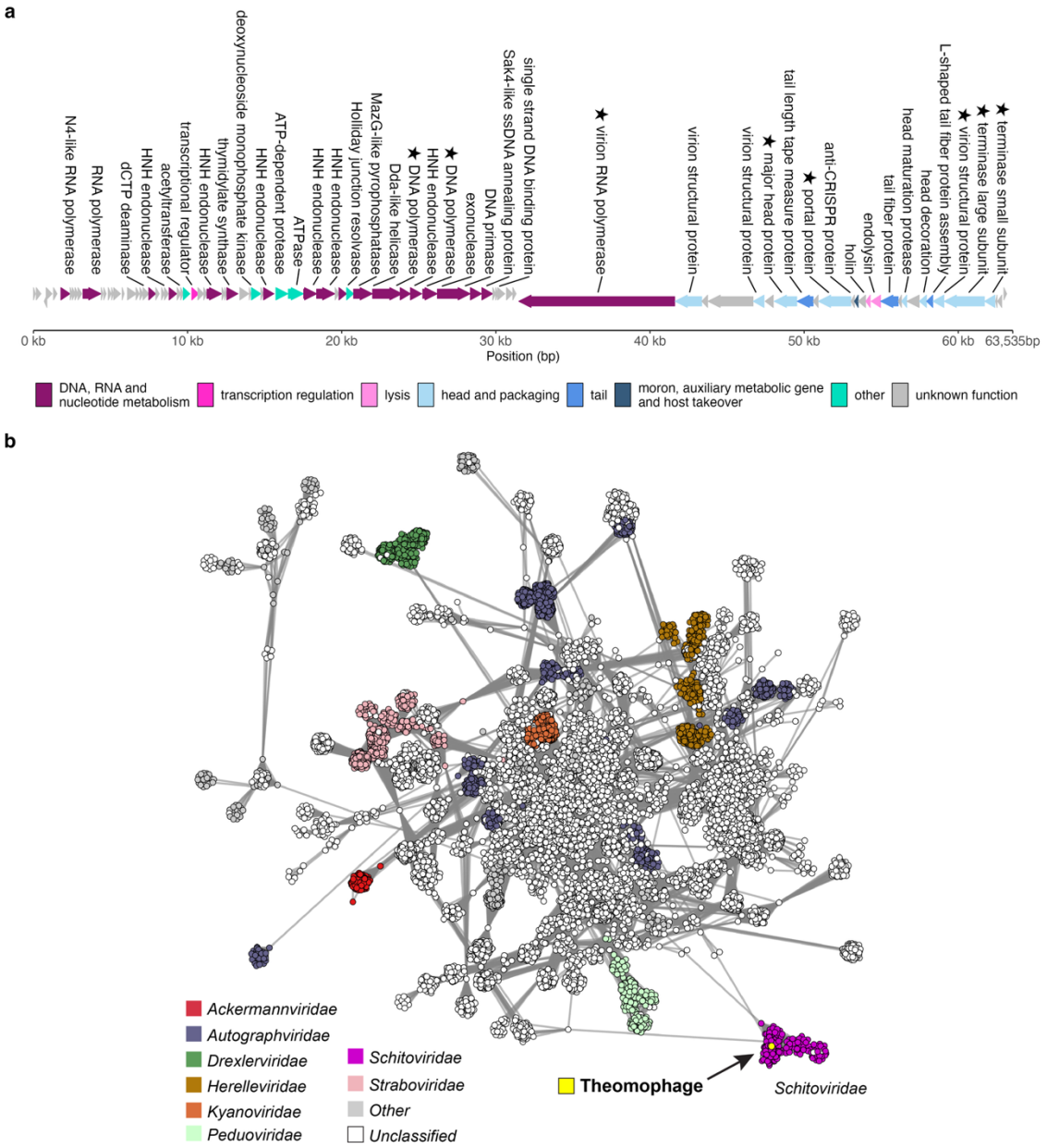


Figure 3. Genomic architecture and gene-sharing network of Theomophage, a novel Schitoviridae bacteriophage. a *Theomophage genome (63,535 bp), encoding 88 genes and one tRNA, including all 7 Schitoviridae hallmark genes (indicated with stars). Gene functions for 40 genes were predicted with Phold[43] and hallmark genes identified with HMM searches.* ***b*** *vConTACT2[39] gene-sharing network of Theomophage and 26,302 bacterial and archaeal viruses from the INPHARED database[48]. Circles represent viral genomes, with edges indicating degree of connectivity based on shared protein clusters. Theomophage groups with bacteriophages of the Schitoviridae family (purple) but is only distantly related to known Schitoviridae sequences, suggesting it represents a new subfamily (**Fig.S14-S15**). Only the subnetwork containing Theomophage is shown.*

Community composition remains stable despite massive Theomophage outbreaks

Viral predation can regulate microbial populations and drive major shifts in community composition and function[28,49]. For example, viral outbreaks following induced coccolithophore blooms can shift the balance between eukaryotic and bacterial organic matter recyclers[49], and bacteriophage outbreaks are implicated in restoring gut microbial diversity following antibiotic treatment[50]. We hypothesized that the massive Theomophage outbreaks in type-A mesocosms would similarly impact community composition. However, Bray-Curtis dissimilarity between samples taken before and after Theomophage outbreaks was no greater than shifts over similar intervals in mesocosms without outbreaks (**Fig.S16**), suggesting that the communities were resilient to the Theomophage outbreaks. This observation may support the emerging view of a potentially limited impact of phages on bacterial community structure[2].

Candidate hosts of Theomophage

The high abundance of Theomophage in several compost mesocosms suggests that it likely infects a dominant community member. To identify potential hosts, we first searched for CRISPR spacers matching the Theomophage genome in all the compost metagenomes using Spacerextractor[51] (Methods), recovering 298,447 spacers across mesocosm and cocktail datasets. None matched the Theomophage genome (up to 3 mismatches). A subsequent search against the external SpacerDB[52] database revealed six spacers with 2–3 mismatches from *Nitrosomonas* MAGs (**Table S10**). While the high number of mismatches lowers confidence of a phage-host relationship, *Nitrosomonas* was abundant and strongly associated with type-A communities where Theomophage outbreaks occurred, and was also detected in 2 out of 5 open mesocosms that were invaded by Theomophage (open_3 and open_4; **Fig.1b–e; Fig.S17a**). Notably, Theomophage reached higher abundance in these two mesocosms than in the other three, providing further ecological support for a *Nitrosomonas* host. Although no direct spacer evidence (*i.e.* matching *Nitrosomonas* spacers) was found within the compost metagenomes, the spatial co-occurrence and spacer matches together suggest a possible host-phage association.

Two additional candidate hosts – *Moraxella,* and JACPRH01, a family within the *Nitrososphaerales* order of soil ammonia-oxidizing archaea[53] – were identified through computational predictions. A single spacer with 1 mismatch was found in SpacerDB belonging to a *Moraxella* genome from a wastewater sample, while JACPRH01 was the highest-confidence prediction of iPHoP, a machine-learning framework that integrates multiple phage-host prediction tools (iPHoP score 85.80, corresponding to a 14.2% false discovery rate)[54–58]. Both taxa were detected in the compost metagenomes, but at very low abundance near the metagenomics detection threshold (a maximum of 182 reads or 0.0005**%** per sample) and showed no association with Theomophage presence (**Fig.S17bc**). If JACPRH01 is the true host, it would suggest an unexpected host range expansion of the *Schitoviridae* family to the archaeal domain, as all known *Schitoviridae* phages infect *Proteobacteria*[40]. Given the limited evidence and low abundance of these

taxa, our best prediction remains a *Nitrosomonas* host, although further validation is required.

Migration drives Theomophage evolution by exposure to new community contexts

The strong association of Theomophage with type-A communities, along with its successful colonization of type-B mesocosms, enabled us to investigate how migration and community context shape bacteriophage evolution. Specifically, we examined three conditions: (i) closed type-A mesocosms (no viral migration, native context), (ii) open type-A mesocosms (viral migration, native context), and (iii) open type-B mesocosms (viral migration, new context). We hypothesized that each condition would impose distinct selection pressures and result in different evolutionary signatures. In particular, we hypothesized that following colonization of non-native type-B communities, molecular evolution of Theomophage would be accelerated because its lower relative abundance suggests local maladaptation or a possible host switch. To test this, we identified Single Nucleotide Polymorphisms (SNPs) in Theomophage populations by mapping sample reads to reference assemblies (Methods). In total, we identified 76 SNPs with frequencies >10% in Theomophage populations. Open type-B mesocosms contained the greatest number of SNPs (n=70), and notably, 80.3% (n=61) of these variants were unique to type-B communities (**Fig.4ab, Fig.S18a**). This pattern supports the idea that exposure to new community contexts accelerates bacteriophage evolution and diversification.

Theomophage evolution is constrained in closed, native communities

Next, to explore how these SNPs emerged and spread in more detail, we reconstructed Theomophage genotypes and tracked them over time and across mesocosms. Reconstructing genotypes from short-read data is challenging because SNPS typically lie on different reads, making it hard to resolve linkage. Here we leveraged the observation that many samples were monomorphic at all variant positions, indicating clonal populations (i.e. all allele frequencies near 0 or 1; Methods). Using this approach, we identified two genetically distinct Theomophage genotypes in the founding communities. Mesocosm 2 contained a genotype identical to the assembled vMAG used for variant calling (hereafter genotype G1), while mesocosms 1, 5, and 6 harbored a second genotype (G2) which differed from G1 by nine SNPs located in genes associated with host interaction and replication, including the tail fiber protein, HNH endonuclease, endolysin, and one intergenic region (**Fig.4c,g**; **Table 1**). Each founding mesocosm contained a single genotype, with minimal within-population diversity.

With the identity and initial distribution of genotypes established, we next tracked how these genotypes changed over time within each mesocosm. In closed mesocosms, no new SNPs appeared in either genotype G2 or G1 over the 40 weeks in which Theomophage was detected, though genotype G1 contained limited preexisting microdiversity in mesocosms closed_5 and closed_6 (**Fig.4c, S18**). The absence of novel mutations is striking given the massive phage outbreaks and the long duration over which Theomophage was observed, which suggest extensive viral replication. In experimental

evolution studies in simplified phage–host systems[4,19] and dairy starter cultures with limited taxonomic diversity[59], phage evolution is typically rapid and driven by coevolution. In contrast, we observed prolonged genetic stability under ecologically complex conditions. Consistent with previous observations in natural environments[14,15,60], our results suggest that phage evolution in complex microbial communities may proceed only slowly, with long-term, stable coexistence between phage and host.

Migration accelerates evolution via recombination between preexisting Theomophage genotypes

In contrast to the genetic stability observed in closed mesocosms, Theomophage populations in open mesocosms exhibited rapid and complex evolutionary dynamics, driven by genotype migration, recombination, and the acquisition of new SNPs following colonization of the type-B mesocosms. While closed mesocosms only contained either genotype G1 or G2, in open mesocosms both genotypes were often detected immediately following the first serial transfer, indicating rapid dissemination between mesocosms. (**Fig.4ef**). For example, mesocosm open_5 initially contained only genotype G2, but by week 2 genotype G1 had invaded and rose to ~50% frequency by week 8, as indicated by the coupled frequency decline of the nine SNPs that characterized G2. Similar cross-invasions were observed in all other open mesocosms with sufficient coverage for variant calling, including colonization of the novel type-B mesocosms by both Theomophage genotypes. Importantly, both migrating genotypes were also detected in the MGE cocktail used for transfers between mesocosms, providing further evidence that these patterns resulted from viral migration rather than parallel evolution (**Fig.4ef**).

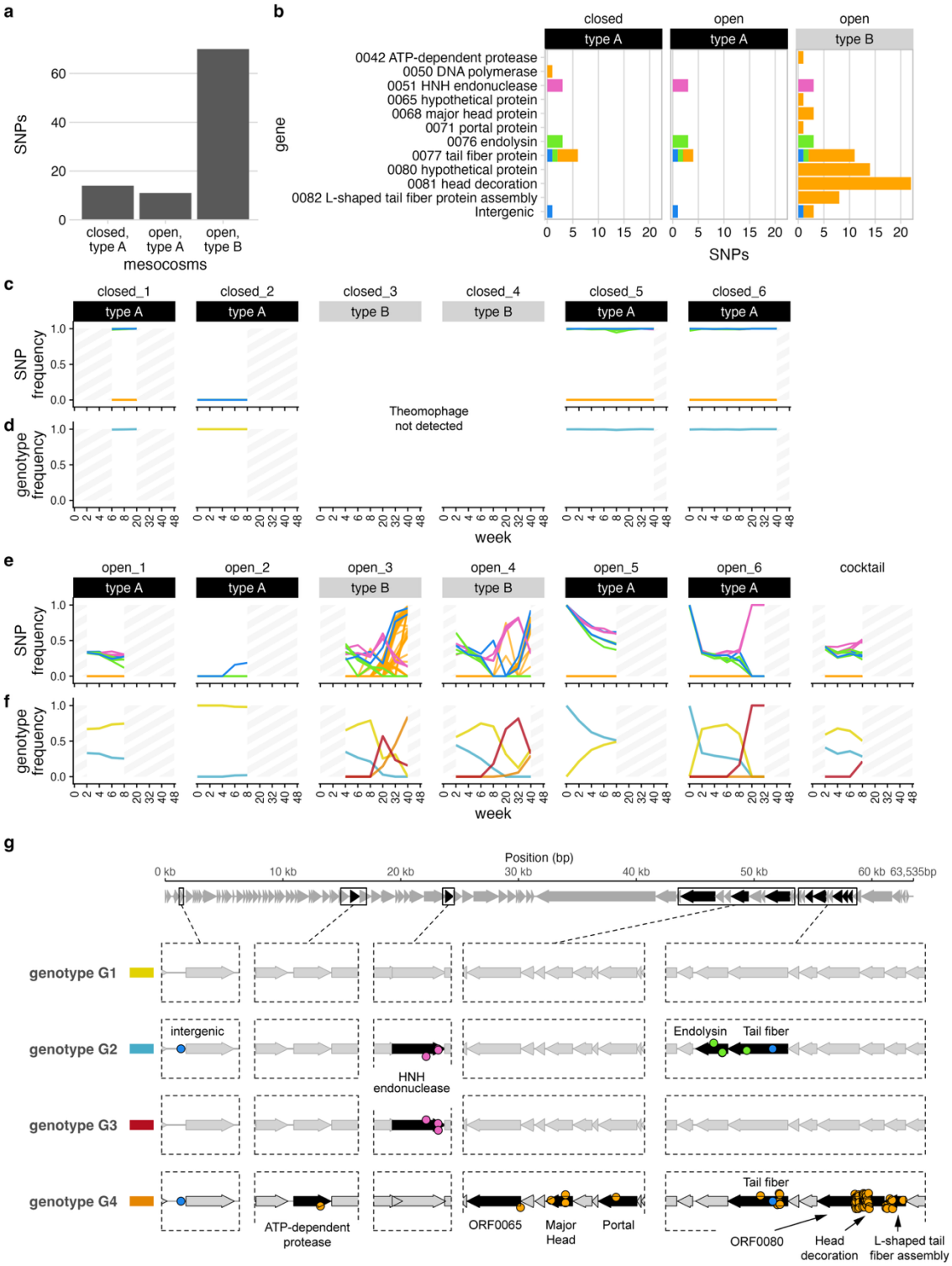
In all open mesocosms where G1 and G2 coexisted beyond week 8, the frequency trajectories of their characteristic alleles became decoupled, indicating recombination between the genotypes. For example, in open_6, genotype G1 and G2 coexisted for 8 weeks, after which a recombinant genotype (G3) emerged, characterized by three G1-derived alleles in HNH endonuclease and six G2-derived alleles (**Fig.4ef**; Table 1). Similarly, in open_3 a second recombinant genotype (G4) arose that combined alleles from both founder genotypes. Both recombinants replaced their parental genotypes and spread to other mesocosms via continued migration. G3, originating in open_6, was detected in open_3 and open_4, while G4 spread from open_3 to open_4 and eventually to open_6 by week 32, as revealed by a more sensitive SNP analysis (**Fig.S19**).

Rapid SNP accumulation following migration to new community type

At week 8, 61 new SNPs appeared in open_3 that were not detected in any founding or closed communities (**Fig.4ef**). The close similarity in frequency trajectories between these novel SNPs and the defining alleles of G4 indicates that they were located on the same genome. These mutations occurred in several predicted structural proteins, including Major Capsid and Tail proteins, but were mostly clustered (46/61) in a region spanning three consecutive genes: ORF0080, head decoration protein, and L-shaped tail fiber assembly protein (**Fig.4bg, Table S5**).

402 To investigate the mechanistic origin of these mutations, we first tested whether these
403 mutations arose *de novo* via diversity-generating retroelements (DGRs)—error-prone
404 reverse transcriptase systems that target specific genomic loci, typically structural genes,
405 to promote rapid adaptation to new hosts[61]. However, screening the Theomophage
406 genome with MyDGR[62] revealed no DGRs or reverse transcriptase genes, making this
407 mechanism unlikely. Another major driver of viral evolution is recombination[12]. The tight
408 genomic clustering of mutations suggests that recombination, rather than independent
409 point mutations, may underlie the emergence of these variants. However, a blastn[63]
410 query of all assembled contigs with the Theomophage region spanning the three mutated
411 genes yielded no significant matches. This suggests that recombination may have
412 occurred with a rare sequence not captured in our assemblies. Rare viral genotypes (e.g.
413 below the metagenomics detection threshold) can act as viral seed banks in natural
414 ecosystems and harbor genetic diversity that can become adaptive when selection
415 pressures change[64]. We hypothesize that following migration to type-B mesocosm
416 open_3, Theomophage recombined with a member of the rare virosphere giving rise to
417 genotype G4. While the precise mechanism remains unresolved, the timing, localization,
418 and spread of these mutations from open_3 to open_4 and open_6 suggests they
419 conferred selective advantage in the type-B community context.
420
421 ***Figure 4. (next page) Molecular evolution and migration of Theomophage in closed and***
422 ***open mesocosms. a*** *Total SNPs detected in Theomophage populations in type-A and type-*
423 *B mesocosms under closed and open regimes. Most SNPs (80.3%) appeared only in the*
424 *novel type-B communities, which Theomophage colonized in the open regime.* ***b*** *Genomic*
425 *location of SNPs in (A).* ***c*** *Allele frequency trajectories of SNPs that differed from the*
426 *reference genotype G1 in closed mesocosms. Theomophage populations in type-A*
427 *communities were initially monomorphic for all variant positions, meaning they were*
428 *composed of a single genotype: G1 in community closed_2, which was identical to*
429 *reference, and G2 in communities closed_1, closed_5, and closed_6, which differed from*
430 *the reference by 9 SNPs as indicated by frequencies of ~1. No new variants emerged that*
431 *fixed in the population during the 40 weeks over which the bacteriophage was detected.*
432 *Shading indicates samples with insufficient coverage for variant calling or where*
433 *Theomophage was not detected.* ***d*** *Inferred genotype dynamics from (C); each mesocosm*
434 *contained either G1 or G2.* ***e*** *Same as (C), but for open mesocosms and MGE cocktail. Both*
435 *genotypes spread via the MGE cocktail, cross-invading type-A and previously*
436 *Theomophage-free type-b mesocosms. Decoupled SNP trajectories at week 6-8 in*
437 *mesocosms open_3 and open_6 indicate two independent recombination events between*
438 *G1 and G2, giving rise to new genotypes G3 and G4. G4 also acquired 61 additional SNPs*
439 *(orange).* ***f*** *Genotype dynamics in open mesocosms. Samples from open_7, open_8 and*
440 *open_9 not shown due to insufficient read coverage for variant calling, although*
441 *Theomophage was sparsely detected there (Fig.2).* ***g*** *Genomic positions of SNPs*
442 *distinguishing genotypes G1, G2, G3 and G4. Most (46/61) new mutations in G4 occurred*
443 *in ORFs 80, head decoration protein and L-shaped tail fiber assembly protein (**Table 1,***
444 ***Table S5**). SNPs are jittered vertically for clarity. SNP colors correspond to Fig.4b and Table*
445 *1 and are based on similarity of allele trajectories (Methods).*

446


a
SNPs
mesocosms
closed, type A
open, type A
open, type B
b
closed
open
type A
type B
gene
0042 ATP-dependent protease
0050 DNA polymerase
0051 HNH endonuclease
0065 hypothetical protein
0068 major head protein
0071 portal protein
0076 endolysin
0077 tail fiber protein
0080 hypothetical protein
0081 head decoration
0082 L-shaped tail fiber protein assembly
Intergenic
SNPs
c
closed_1
closed_2
closed_3
closed_4
closed_5
closed_6
SNP frequency
Theomophage not detected
d
genotype frequency
week
e
open_1
open_2
open_3
open_4
open_5
open_6
cocktail
f
g
Position (bp)
0 kb
10 kb
20 kb
30 kb
40 kb
50 kb
60 kb
63,535bp
genotype G1
genotype G2
genotype G3
genotype G4
intergenic
HNH endonuclease
Endolysin
Tail fiber
ATP-dependent protease
ORF0065
Major Head
Portal
ORF0080
Head decoration
L-shaped tail fiber assembly

447
448

449 **Table 1. Genomic variants defining four Theomophage genotypes.** For each Single
450 Nucleotide Polymorphism (SNPs), the position, effect, associated open reading frame
451 (ORF), predicted gene function, and inferred genotype for each genotype are shown.
452 Nine SNPs differentiate genotype G2 from G1. Details for SNPs arising in G4 are
453 provided in Supplementary Table 5. Colors match Fig.4b-g.
454

Variant	Effect	ORF	Predicted Function	G1	G2	G3	G4	Color in Fig.4
1276 T → G	intergenic	–	–	–	√	–	√	
24224 T → C	Ile → Thr	0051	HNH endonuclease	–	√	√	–	
24375 G → A	synonymous	0051	HNH endonuclease	–	√	√	–	
24378 T → C	synonymous	0051	HNH endonuclease	–	√	√	–	
54689 T → C	Glu → Gly	0076	endolysin	–	√	–	–	
54853 T → C	synonymous	0076	endolysin	–	√	–	–	
54856 C → T	synonymous	0076	endolysin	–	√	–	–	
55320 A → C	Val → Gly	0077	tail fiber protein	–	√	–	–	
55814 C → T	Synonymous	0077	tail fiber protein	–	√	–	√	
16296 T → A	Leu → Ile	0042	ATP-dependent protease	–	–	–	√	
22232 A → G	Synonymous	0050	DNA polymerase	–	–	–	√	
46679 C → T	Gly → Ser	0065	hypothetical protein	–	–	–	√	
3 SNPs	see Table S5	0068	major head protein	–	–	–	√	
51895 G → T	Leu → Ile	0071	portal protein	–	–	–	√	
14 SNPs	see Table S5	0077	tail fiber protein	–	–	–	√	
14 SNPs	see Table S5	0080	hypothetical protein	–	–	–	√	
2 SNPs	intergenic	–	–	–	–	–	√	
22 SNPs	see Table S5	0081	head decoration	–	–	–	√	
8 SNPs	see Table S5	0082	L-shaped tail fiber protein assembly	–	–	–	√	

455

456 Discussion

457 The factors that drive bacteriophage evolution in complex natural ecosystems, such as
458 local microdiversity, community context, and migration, cannot be fully captured by
459 laboratory studies that reduce the genomic and spatiotemporal complexity of natural
460 environments[10,16]. Previous studies have examined the effect of migration using single
461 phage-host pairs[65,66], but its influence in taxonomically rich communities has not been
462 explored.
463 Our results show that bacteriophage population dynamics and evolution in complex
464 communities are highly contingent on community background and migration. In isolated,
465 native mesocosms, Theomophage showed no detectable molecular evolution over 40
466 weeks, despite repeated massive outbreaks. In contrast, migration to new communities
467 sparked rapid evolution via recombination between preexisting genotypes and SNP
468 accumulation, producing new genotypes that spread across mesocosms.
469

Our results indicate that in complex microbial communities, phage–host dynamics can remain evolutionarily static for 40 weeks unless disrupted by migration. This contrasts with results from simplified lab systems, where antagonistic coevolution often produces rapid phage evolution via sequential selective sweeps[8–10]. Instead, our findings match diversity patterns in natural microbial ecosystems that indicate long-term coexistence of susceptible and resistant hosts without rapid turnover[14,15]. They further suggest that in natural ecosystems, disturbances to local community structure such as seasonal rainfall in soils[18,67] or cycles of dispersal and community assembly on marine organic particles[68], may act as major drivers of bacteriophage diversification.

Viral abundance is a key parameter in microbial ecology but remains challenging to measure in complex microbial communities[69]. Two methods are commonly used: microscopy-based counting of virus-like particles (VLPs), which estimates free viral particles but cannot detect ecologically relevant intracellular viral stages or provide taxonomic or functional insight; and metagenomics, which captures both free and intracellular viruses and enables genome-level characterization. Each method has limitations[69,70], but metagenomics is uniquely suited to quantify specific viral taxa in community context[71]. Here, using metagenomics, we have captured evidence of massive, parallel outbreaks of Theomophage, a previously undescribed bacteriophage, across different independent compost communities, where it reached unprecedented abundances of up to 74.3% of metagenome reads. Such dominance by a single virus is rare: viral reads typically constitute ~5% of metagenomic sequences[72], though higher values have been reported. For example, crAssphage can reach 22% in human gut metagenomes[73] and a *Streptococcus* phage comprised ~20% of reads in cheese starter cultures composed of two bacteria and three associated phages[74]. Our findings far exceed these cases, showing that a single phage can dominate even highly diverse microbial communities in terms of genomic information.

The extraordinary abundance of Theomophage highlights the need to account for viral sequences when analyzing metagenomes. State-of-the-art taxonomic profilers used in metagenome analysis struggle to identify viral sequences[75], requiring the use of specialized viral identification tools[76,77]. Comprehensive viral profiling remains limited to well-described biomes[72], meaning that the most abundant members of microbial communities may be missed in metagenome profiling.

Viral abundance estimates using metagenomics can be influenced by technical biases during DNA extraction, library preparation, and sequencing[69,70,78]. In particular, rolling circle amplification is known to skew viral abundances[79], but this was not used here. The Nextera XT library preparation kit used in this study may underrepresent low-GC regions[80], but given that Theomophage has a lower GC content (34%) than the average of the assembled metagenomes, this bias would underestimate rather than inflate its abundance. Thus, our Theomophage abundance estimates may be conservative.

One potential explanation for the exceptional abundance of Theomophage is a carrier state-like infection mode, in which cells shed virions without undergoing lysis. This non-canonical infection mode has been observed in crAss-like phage crAss001 where it was proposed to underlie high abundance and stable coexistence with a *Bacteroides* host[81]. Similar modes have been reported for a wide range of virulent phages[82–87], and it is possible that a comparable mechanism enabled the high abundances and long-term persistence of Theomophage reported here.

Finally, we highlight potential limitations of our study. First, we cannot rule out that the coarse temporal resolution in later stages of the experiment caused us to miss additional Theomophage outbreaks. The more frequent sequencing earlier in the experiment showed that outbreaks took 6 to 8 weeks, suggesting similar events may have occurred undetected later on. The temporal resolution also limited our ability to assess the impact of Theomophage outbreaks on community composition. Second, a confirmed host would provide valuable context for interpreting the evolutionary dynamics of Theomophage. Host identification could clarify whether the observed genotype-level diversity in native type-A communities reflects local adaptation, and whether the increased evolutionary rates following migration to new community types are mirrored in host evolution, or a signature of adaptation to a new, suboptimal host[88]. If *Nitrosomonas* were confirmed as the host, this could suggest a host switch following migration to mesocosms open_7, open_8 and open_9, where *Nitrosomonas* was sparsely detected.

Together, our study has leveraged genotype-resolved analysis of existing shotgun metagenomic data to reach new conclusions about the ecology and evolution of a previously unknown bacteriophage in a complex microbial ecosystem. We identified two alternative, stable community types, observed massive and reproducible bacteriophage outbreaks confined to one type, and showed that phage migration and colonization of the second community type accelerated bacteriophage evolution. These findings demonstrate that genotype tracking of laboratory-propagated natural microbial communities can reveal viral eco-evolutionary dynamics with genomic precision, providing a powerful approach for studying viruses in complex yet tractable ecosystems.

Materials and Methods

Note that the closed and open treatments were previously referred to as "vertical" and "horizontal", respectively[30,34].

Assembly and binning

Read quality control was done with Prinseq v.0.20.4[89] using parameters -derep 14 -lc method dus -c threshold 20. Adapters were trimmed using Flexbar v.3.5.0[90] with --adapter-preset Nextera -ap ON. The high complexity of compost ecosystems makes metagenome assembly difficult, potentially leading to short contigs and a large fraction of unassembled reads[91]. To improve assembly and increase the chance of recovering low-abundant genomes such as bacteriophages[92,93], reads from all timepoints for each mesocosm were pooled and co-assembled with metaSPAdes v.3.14.0[94]. To preserve sample-level diversity for Theomophage genome annotation and variant calling, all samples (n=170) were also assembled separately. Samples from the viral cocktails were assembled in the same way. Reads were mapped using BWA-mem2 v0.7.17-r1188[95] and coverage was calculated with SAMtools v1.7[96]. Contig depths were calculated with jgi_summarize_bam_contig_depths and binned using Metabat2 v.2.12.1[97].

Taxonomic profiling

Taxonomic classification of assembled scaffolds and bins was performed with CAT v5.3[31] with --index_chunk 1 --block_size 5 --top 11 --I_know_what_Im_doing, using CAT NCBI database and taxonomy v2021-04-30. Taxon abundances (fraction of total sample reads) were calculated using RAT version Sept 2021[32] and database v2021-04-30. For dimension reduction, unclassified taxa or taxa that were only classified at other taxonomic ranks were dropped, and relative abundances were center-log-ratio (CLR) transformed, with zero values adjusted by adding 1 to all read counts. To estimate community richness the number of detected genera was counted with different abundance cutoffs (**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 v0.9.0. While counting any contig marked as viral by this panel of virus identification tools likely increased the number of false positives, our subsequent analysis focused on a single bacteriophage sequence identified with high confidence by manual curation. To estimate viral abundances (including Theomophage), we randomly selected a co-assembly where Theomophage was completely assembled and used that as a target for read mapping.

Viral completeness estimation, annotation, lifestyle prediction

Viral contig completeness was estimated using CheckV v1.0.3 with database v1.5[38]. Theomophage contig NODE_224 was estimated 100% complete based on the presence of 55bp direct terminal repeats (DTRs). Complete Theomophage genomes were independently assembled in 54 independent, single-timepoint assemblies, with a consistent contig length of 63,535bp (46/54 contigs, **Fig.S20)** and high sequence similarity

588 (99.98 ANI). The genomes were circularly permuted and contained DTRs, consistent with
589 complete genome assembly from shotgun reads derived from phage replication
590 concatemers or linear genomes with DTRs (**Fig.S21**).
591 Theomophage MAG NODE_224_length_63535 assembled from sample 3_open-week4
592 was annotated with Pharokka v1.7.5[42] and Phold v0.2.0[43] using default settings.
593 Specifically, coding sequences (CDs) were predicted with PHANOTATE[98], tRNAs were
594 predicted with tRNAscan-SE 2.0[99], tmRNAs were predicted with Aragorn[100] and CRISPRs
595 were predicted with CRT[101]. Functional annotation was generated by translating protein
596 amino acid sequences to the 3Di token alphabet with ProstT5[102], and Foldseek[103] was then
597 used to search these against the Phold phage protein structures database, which contains
598 PHROGs[104], VFDB[105] and CARD[106] structures predicted with Colabfold[107]. To search for the
599 seven *Schitoviridiae* hallmark genes, a protein hidden Markov model (HMM) was
600 constructed for each from protein sequences from REF[41] by aligning them with MAFFT
601 v7.526[108] and hmmbuild (HMMER v3.4, www.hmmer.org). The resulting models were
602 used to search the predicted Theomophage CDs using hmmsearch (HMMER v3.4) with
603 default parameters. Output is available as **Table S4. Fig.3a and 4g** were drawn in R v4.3.1
604 with gggenes v0.4.1 (https://github.com/wilkox/gggenes) and ggmagnify v0.4.1
605 (https://github.com/hughjonesd/ggmagnify). Theomophage lifestyle was predicted with
606 BACPHLIP v.0.9.6[46].
607

608 ## Theomophage taxonomy

609 The gene sharing network was created by combining predicted proteins of 26,302
610 bacterial and archaeal viral reference genomes from the INPHARED reference set
611 (1Oct2023)[48] with predicted proteins from Theomophage NODE_224_length_63535,
612 using vConTACT2 v.0.9.19 (–rel-mode 'Diamond' –pcs-mode MCL –vcs-mode
613 ClusterONE). For **Fig.S15** a gene sharing network was created for Theomophage and 573
614 *Schitoviridae* reference sequences (see below). Networks were visualized with Cytoscape
615 v.3.9.1[109] and Adobe Illustrator 2025).
616 A reference set of 573 high quality (predicted >90% complete) *Schitoviridae* sequences
617 was created by downloading sequences described as "high quality" or "complete" in
618 Wittmann *et al*[40] and Zheng et al[41] from Genbank and IMG/VR v3, complemented by
619 adding sequences annotated as "Schitoviridae" and "high quality" or "isolate" from
620 IMG/VR v4[44], finally followed by dereplication at 100% ANI with CD-HIT v.4.8.1[110]. For
621 gene sharing networks, protein prediction was performed as described above.
622 For tailed bacteriophages three conserved proteins - the Terminase large subunit (terl),
623 major capsid protein (MCP) and portal protein (portal) can be used to infer taxonomic
624 relationships[111]. We gathered proteins annotated as "terminase large", "portal", or
625 "major capsid" or "large capsid" from the INPHARED dataset and added the
626 corresponding protein from Theomophage MAG NODE_244. Genes were aligned with
627 Clustal Omega v.1.2.4[112] using default parameters and positions with gaps in more than
628 50% of the sequences were removed with trimAl v.1.4.rev15[113]. Gene trees were made
629 using FastTree version 2.1.10[114] with default parameters and plotted with R v.4.3.1 using
630 the ggtrees package[115].

631 Intergenomic similarities (pairwise nucleotide identity normalized for genome lengths)
632 were calculated from MAG NODE244 and 573 high-quality *Schitoviridae* sequences using
633 VIRIDIC standalone v1.1 (blastn parameters "-word_size 7 -reward 2 -penalty -3 -gapopen
634 5 -gapextend -2")[45].
635

636 ## Host prediction

637 iPHoP v1.3.3, database Aug_2023_pub_rw[54] was run with default parameters and
638 additionally with a lowered confidence cutoff (-m 80). iPHoP predictions and raw output
639 per bioinformatic tool are available as **Table S6-S7.**
640 Crispr spacers were extracted from mesocosm and cocktail reads and filtered and
641 denoised with Spacerextractor[51] v.0.9.5 using default parameters. Resulting spacers were
642 mapped against the four inferred Theomophage genotypes (MAG
643 NODE_224_length_63535_cov_5745_934389, genotypes G2, G3 and G4) using
644 Spacerextractor with default parameters, which resulted in no matches. Spacers from the
645 global SpacerDB database v.2025-05-02[52] were mapped in the same way against the four
646 Theomophage genotypes, only considering spacers that mapped with 0 or 1 mismatch
647 across the full spacer length. Results are available as **Table S8-S10.**
648

649 ## Variant calling

650 To reduce the chance of spurious read mapping, QC-filtered reads were competitively
651 mapped against all contigs from a randomly selected single-sample assembly containing
652 the complete Theomophage MAG (sample open_3-week_4 containing
653 NODE_224_length_63535_cov_5745_934389) with BWA-mem2[95]. Variants were called
654 for the Theomophage MAG using Haplotypecaller, GenomicsDBImport and
655 GenotypeGVCFs from GATK v.4.2.3.0[116] following the joint variant calling pipeline and the
656 best practices for germline short-variant discovery. To exclude sequencing and read
657 mapping artifacts, only alleles that reached a frequency of 10% in any sample and were
658 supported by a minimum of 4 reads, coming from samples where ≥30% of the genome
659 was covered by ≥10 reads were considered for downstream analysis. This resulted in a
660 total of 76 unique SNPs across all samples. To further exclude read mappings artifacts,
661 read depth coverage profiles were inspected with Integrative Genomics Viewer
662 V.2.13.2[117]. Variants located in genomic areas with irregular coverage (e.g. abrupt depth
663 changes) were further evaluated by separately reassembling mapped read pairs for these
664 samples and a second round of variant calling using the resulting contig as reference. All
665 SNPs were confirmed and no variants were removed.
666 Variant effects were predicted with SNPEff v.5.0e[118] (-noLog -noStats -no-downstream -
667 no-upstream -no-utr) using a custom SNPEff database built with
668 NODE_224_length_63535_cov_5745_934389 as reference, a .gff file generated using
669 Prokka and the Bacterial_and_Plant_Plastid codon table. MAGs were interrogated for
670 Diversity Generating Retroelements (DGRs) with the webserver MyDGR[62].
671

Resolving genotypes and estimating genotype abundance

Genotypes were inferred from samples where variant frequencies for all 76 variants were close to either 0 or 1, indicating monomorphic populations. For example, genotype G1 was inferred from sample closed_2 week 0, where the frequency of all 76 variants was close or equal to 0, indicating that this population consisted of a single genotype that was identical to the reference genome (**Fig.4c,d,g; Table 1**). Similarly, genotype G2 was inferred from monomorphic samples closed_1, closed_5, closed_6 week 0, which had nine SNPs in the tail fiber protein, HNH endonuclease, endolysin and an intergenic position with frequencies near 1. Note that Theomophage populations were not monomorphic in open samples, but the 9 SNPs that differentiated S1 and G1 showed similar frequency dynamics over time, further suggesting genetic linkage in these samples in accordance with the pigeonhole principle. Similar reasoning allowed the identity of G3 to be inferred from samples open_6 Week 20 and 32, and G4 from sample open_3 week 40. For **Fig 4.D** and **F** we then assumed the simplest scenario that could explain the evolution of G3 and G4 from G1 and G2 that was compatible with the variant trajectories, namely two separate recombination events and fixation of 61 new variants in G4. We approximated genotype abundances for weeks 8, 20, 32 and 40 in communities open_3, open_4, open_6 and the cocktail as follows: G4 abundance as the average relative abundance of orange and blue variants; G3 abundance as the average of pink variants; G2 as the average of the blue and green variants, and G1 as 1 minus the abundance of G2, G3 and G4. For other time points the abundance of G2 was estimated as the average of blue, pink, and green SNPs and G1 as 1 minus the abundance of G2. The four genotypes 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 genotype G2 that was limited to communities closed_5, closed_6 and open_5 **(Fig.S21)**. For visual clarity we gave SNPs with similar trajectories from cluster 1 different colors.

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[119], data.table v1.14.6, ggbeeswarm v.0.7.1.

Author Contributions

JM wrote the paper with input from PH and BED. JM, PH, BED contributed equally to design of the analysis and interpretation of the results. JM performed the data analysis. 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.

Data availability

All data and code will be made publicly available before publication. Code for analysis and figure generation will be made available on github.com/MGXlab/theomophage. The

Theomophage genome will be deposited in Genbank. Raw sequencing data and assemblies will be made available through online repositories.

Acknowledgments

The authors gratefully acknowledge Jan Kees van Amerongen for maintaining the computational facilities. JM is supported by the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, www.nwo.nl) grant 022.005.023 and the Utrecht University BINF fund. 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. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
2. Castledine, M. & Buckling, A. Critically evaluating the relative importance of phage in shaping microbial community composition. *Trends Microbiol.* **0**, (2024).
3. Hussain, F. A. *et al.* Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* **374**, 488–492 (2021).
4. Borin, J. M. *et al.* Rapid bacteria-phage coevolution drives the emergence of multiscale networks. *Science* **382**, 674–678 (2023).
5. Buckling, A. & Rainey, P. B. Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. Lond. B Biol. Sci.* **269**, 931–936 (2002).
6. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
7. Kortright, K. E., Chan, B. K., Koff, J. L. & Turner, P. E. Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host Microbe* **25**, 219–232 (2019).
8. Tecon, R. *et al.* Bridging the Holistic-Reductionist Divide in Microbial Ecology. *mSystems* **4**, 10.1128/msystems.00265-18 (2019).
9. Koskella, B., Hernandez, C. A. & Wheatley, R. M. Understanding the Impacts of Bacteriophage Viruses: From Laboratory Evolution to Natural Ecosystems. *Annu. Rev. Virol.* (2022) doi:10.1146/annurev-virology-091919-075914.
10. Chevallereau, A., Pons, B. J., van Houte, S. & Westra, E. R. Interactions between bacterial and phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49–62 (2022).
11. Meyer, J. R. *et al.* Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* **335**, 428–432 (2012).
12. Kauffman, K. M. *et al.* Resolving the structure of phage–bacteria interactions in the context of natural diversity. *Nat. Commun.* **13**, 1–20 (2022).
13. Piel, D. *et al.* Phage–host coevolution in natural populations. *Nat. Microbiol.* **7**, 1075–1086 (2022).
14. Schwartz, D. A. & Lindell, D. Genetic hurdles limit the arms race between Prochlorococcus and the T7-like podoviruses infecting them. *ISME J.* **11**, 1836–1851 (2017).
15. Guerrero, L. D. *et al.* Long-run bacteria-phage coexistence dynamics under natural habitat conditions in an environmental biotechnology system. *ISME J.* **15**, 636–648 (2021).
16. Dey, R. *et al.* Emergent higher-order interactions enable coexistence in phage-bacteria community dynamics. 2025.05.15.651590 Preprint at https://doi.org/10.1101/2025.05.15.651590 (2025).
17. Blazanin, M. & Turner, P. E. Community context matters for bacteria-phage ecology and evolution. *ISME J.* **15**, 3119–3128 (2021).
18. Bickel, S. & Or, D. Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nat. Commun.* **11**, 116 (2020).
19. Shaer Tamar, E. & Kishony, R. Multistep diversification in spatiotemporal bacterial-phage coevolution. *Nat. Commun.* **13**, 7971 (2022).
20. Dubey, M. *et al.* Environmental connectivity controls diversity in soil microbial communities. *Commun. Biol.* **4**, 492 (2021).
21. Wu, F. *et al.* Modulation of microbial community dynamics by spatial partitioning. *Nat. Chem. Biol.* 1–9 (2022) doi:10.1038/s41589-021-00961-w.
22. Bittleston, L. S., Gralka, M., Leventhal, G. E., Mizrahi, I. & Cordero, O. X. Context-dependent dynamics lead to the assembly of functionally distinct microbial communities. *Nat. Commun.* **11**, 1440 (2020).

23. Pontrelli, S. *et al.* Metabolic cross-feeding structures the assembly of polysaccharide degrading communities. *Sci. Adv.* **8**, eabk3076 (2022).
24. Estrela, S. *et al.* Functional attractors in microbial community assembly. *Cell Syst.* **13**, 29-42.e7 (2022).
25. Silverstein, M. R., Bhatnagar, J. M. & Segrè, D. Metabolic complexity drives divergence in microbial communities. *Nat. Ecol. Evol.* **8**, 1493–1504 (2024).
26. Prabhakara, K. H., Zhao, Y., Farr, A. D. & Rainey, P. B. Diverse microbial communities assemble on both recalcitrant and labile carbon sources. 2025.08.07.669103 Preprint at https://doi.org/10.1101/2025.08.07.669103 (2025).
27. Goyal, A., Bittleston, L. S., Leventhal, G. E., Lu, L. & Cordero, O. X. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *eLife* **11**, e74987 (2022).
28. Szabo, R. E. *et al.* Historical contingencies and phage induction diversify bacterioplankton communities at the microscale. *Proc. Natl. Acad. Sci.* **119**, e2117748119 (2022).
29. Zhao, Y. *et al.* Jumbo phage-mediated transduction of genomic islands. 2024.08.29.610337 Preprint at https://doi.org/10.1101/2024.08.29.610337 (2025).
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. 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).
32. Hauptfeld, E. *et al.* Integrating taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes. *Nat. Commun.* **15**, 3373 (2024).
33. Lewin, G. R. *et al.* Long-Term Cellulose Enrichment Selects for Highly Cellulolytic Consortia and Competition for Public Goods. *mSystems* **7**, e01519-21 (2022).
34. van Dijk, B. *et al.* Identifying and tracking mobile elements in evolving compost communities yields insights into the nanobiome. *ISME Commun.* **3**, 1–13 (2023).
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. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
39. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
40. Wittmann, J. *et al.* From Orphan Phage to a Proposed New Family–The Diversity of N4-Like Viruses. *Antibiotics* **9**, 663 (2020).
41. Zheng, K. *et al.* Identification of hidden N4-like viruses and their interactions with hosts. *mSystems* **0**, e00197-23 (2023).
42. Bouras, G. *et al.* Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, btac776 (2023).
43. Bouras, G. *et al.* Protein Structure Informed Bacteriophage Genome Annotation with Phold. 2025.08.05.668817 Preprint at https://doi.org/10.1101/2025.08.05.668817 (2025).

44. 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).
45. Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses* **12**, 1268 (2020).
46. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).
47. Knecht, L. E. *et al.* Bacteriophage S6 requires bacterial cellulose for Erwinia amylovora infection. *Environ. Microbiol.* **24**, 3436–3450 (2022).
48. 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.
49. 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).
50. Pfeifer, E. *et al.* Third-generation cephalosporin antibiotics induce phage bursts in the human gut microbiome. 2025.02.07.636470 Preprint at https://doi.org/10.1101/2025.02.07.636470 (2025).
51. Roux, S. SpacerExtractor. https://code.jgi.doe.gov/SRoux/spacerextractor (2025).
52. Roux, S. *et al.* Planetary-scale metagenomic search reveals new patterns of CRISPR targeting. 2025.06.12.659409 Preprint at https://doi.org/10.1101/2025.06.12.659409 (2025).
53. Bei, Q. *et al.* Metabolic potential of Nitrososphaera-associated clades. *ISME J.* **18**, wrae086 (2024).
54. 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).
55. Coutinho, F. H. *et al.* RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* **2**, 100274 (2021).
56. 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).
57. 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).
58. Lu, C. *et al.* Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **19**, 5 (2021).
59. Kupczok, A. *et al.* Rates of Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades. *Mol. Biol. Evol.* **35**, 1147–1159 (2018).
60. Rozwalak, P., Barylski, J., Wijesekara, Y., Dutilh, B. E. & Zielezinski, A. Ultraconserved bacteriophage genome sequence identified in 1300-year-old human palaeofaeces. *Nat. Commun.* **15**, 495 (2024).
61. Roux, S. *et al.* Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* **12**, 3076 (2021).
62. Sharifi, F. & Ye, Y. MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res.* **47**, W289–W294 (2019).
63. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
64. Ramos-Barbero, M. D. *et al.* Experimental evolution at ecological scales allows linking of viral genotypes to specific host strains. *ISME J.* **18**, wrae208 (2024).

880 65. Morgan, A. D., Gandon, S. & Buckling, A. The effect of migration on local adaptation in a
881 coevolving host–parasite system. *Nature* **437**, 253–256 (2005).
882 66. Kerr, B., Neuhauser, C., Bohannan, B. J. M. & Dean, A. M. Local migration promotes
883 competitive restraint in a host–pathogen 'tragedy of the commons'. *Nature* **442**, 75–78
884 (2006).
885 67. Santos-Medellín, C. *et al.* Spatial turnover of soil viral populations and genotypes overlain by
886 cohesive responses to moisture in grasslands. *Proc. Natl. Acad. Sci.* **119**, e2209132119
887 (2022).
888 68. Datta, M. S., Sliwerska, E., Gore, J., Polz, M. F. & Cordero, O. X. Microbial interactions lead to
889 rapid micro-scale successions on model marine particles. *Nat. Commun.* **7**, (2016).
890 69. Roux, S. & Brum, J. R. Counting dots or counting reads? Complementary approaches to
891 estimate virus-to-microbe ratios. *ISME J.* 1–2 (2023) doi:10.1038/s41396-023-01468-z.
892 70. Benler, S. & Koonin, E. V. Fishing for phages in metagenomes: what do we catch, what do
893 we miss? *Curr. Opin. Virol.* **49**, 142–150 (2021).
894 71. López-García, P. *et al.* Metagenome-derived virus-microbe ratios across ecosystems. *ISME J.*
895 1–12 (2023) doi:10.1038/s41396-023-01431-y.
896 72. Pinto, Y., Chakraborty, M., Jain, N. & Bhatt, A. S. Phage-inclusive profiling of human gut
897 microbiomes with Phanta. *Nat. Biotechnol.* **42**, 651–662 (2024).
898 73. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of
899 human faecal metagenomes. *Nat. Commun.* **5**, ncomms5498 (2014).
900 74. Somerville, V. *et al.* Functional strain redundancy and persistent phage infection in Swiss
901 hard cheese starter cultures. *ISME J.* 1–12 (2021) doi:10.1038/s41396-021-01071-0.
902 75. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of
903 challenges. *Nat. Methods* **19**, 429–440 (2022).
904 76. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat.*
905 *Biotechnol.* 1–10 (2023) doi:10.1038/s41587-023-01953-y.
906 77. Wijesekara, Y. *et al.* Jaeger: an accurate and fast deep-learning tool to detect bacteriophage
907 sequences. 2024.09.24.612722 Preprint at https://doi.org/10.1101/2024.09.24.612722
908 (2024).
909 78. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to
910 generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
911 79. Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *eLife* **9**,
912 e51971 (2020).
913 80. Sato, M. P. *et al.* Comparison of the sequencing bias of currently available library
914 preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Res.*
915 **26**, 391–398 (2019).
916 81. Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with
917 phase variation in Bacteroides intestinalis. *BMC Biol.* **19**, 163 (2021).
918 82. Weiss, M. *et al.* In vivo replication of T4 and T7 bacteriophages in germ-free mice colonized
919 with Escherichia coli. *Virology* **393**, 16–23 (2009).
920 83. Cenens, W. *et al.* Expression of a Novel P22 ORFan Gene Reveals the Phage Carrier State in
921 Salmonella Typhimurium. *PLOS Genet.* **9**, e1003269 (2013).
922 84. Latino, L., Midoux, C., Hauck, Y., Vergnaud, G. & Pourcel, C. Pseudolysogeny and sequential
923 mutations build multiresistance to virulent bacteriophages in Pseudomonas aeruginosa.
924 *Microbiology* **162**, 748–763 (2016).
925 85. Siringan, P., Connerton, P. L., Cummings, N. J. & Connerton, I. F. Alternative bacteriophage
926 life cycles: the carrier state of Campylobacter jejuni. *Open Biol.* **4**, 130200 (2014).

927 86. Pourcel, C., Midoux, C., Vergnaud, G. & Latino, L. A carrier state is established in
928 Pseudomonas aeruginosa by phage LeviOr01, a newly isolated ssRNA levivirus. *J. Gen. Virol.*
929 **98**, 2181–2189 (2017).
930 87. Mäntynen, S., Laanto, E., Oksanen, H. M., Poranen, M. M. & Díaz-Muñoz, S. L. Black box of
931 phage–bacterium interactions: exploring alternative phage infection strategies. *Open Biol.*
932 **11**, 210188 (2021).
933 88. Enav, H., Kirzner, S., Lindell, D., Mandel-Gutfreund, Y. & Béjà, O. Adaptation to sub-optimal
934 hosts is a driver of viral diversification in the ocean. *Nat. Commun.* **9**, 4698 (2018).
935 89. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets.
936 *Bioinformatics* **27**, 863–864 (2011).
937 90. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR—Flexible Barcode and Adapter
938 Processing for Next-Generation Sequencing Platforms. *Biology* **1**, 895–905 (2012).
939 91. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-Age Characterization of Soil
940 Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics. *Soil*
941 *Syst.* **4**, 23 (2020).
942 92. Roux, S. & Emerson, J. B. Diversity in the soil virosphere: to infinity and beyond? *Trends*
943 *Microbiol.* (2022) doi:10.1016/j.tim.2022.05.003.
944 93. Dutilh, B. E. *et al.* Reference-independent comparative metagenomics using cross-assembly:
945 crAss. *Bioinformatics* **28**, 3225–3231 (2012).
946 94. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
947 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
948 95. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of
949 BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed*
950 *Processing Symposium (IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.
951 96. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
952 97. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
953 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
954 98. McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: a novel
955 approach to gene identification in phage genomes. *Bioinforma. Oxf. Engl.* **35**, 4537–4542
956 (2019).
957 99. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and
958 functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
959 100. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes
960 in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
961 101. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of
962 clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
963 102. Heinzinger, M. *et al.* Bilingual language model for protein sequence and structure. *NAR*
964 *Genomics Bioinforma.* **6**, lqae150 (2024).
965 103. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat.*
966 *Biotechnol.* 1–4 (2023) doi:10.1038/s41587-023-01773-0.
967 104. Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote
968 homology. *NAR Genomics Bioinforma.* **3**, lqab067 (2021).
969 105. Chen, L. *et al.* VFDB: a reference database for bacterial virulence factors. *Nucleic Acids*
970 *Res.* **33**, D325-328 (2005).
971 106. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive
972 antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
973 107. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**,
974 679–682 (2022).

975 108. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:
976 Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
977 109. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of
978 biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
979 110. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
980 generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
981 111. Dutilh, B. E. *et al.* Perspective on taxonomic classification of uncultivated viruses. *Curr.*
982 *Opin. Virol.* **51**, 207–215 (2021).
983 112. Sievers, F. *et al.* Fast, scalable generation of high quality protein multiple sequence
984 alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
985 113. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
986 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
987 (2009).
988 114. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood
989 Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
990 115. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization
991 and annotation of phylogenetic trees with their covariates and other associated data.
992 *Methods Ecol. Evol.* **8**, 28–36 (2017).
993 116. Poplin, R. *et al. Scaling Accurate Genetic Variant Discovery to Tens of Thousands of*
994 *Samples*. https://www.biorxiv.org/content/10.1101/201178v3 (2018) doi:10.1101/201178.
995 117. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
996 118. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
997 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain
998 w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
999 119. Journal of Open Source Software: Welcome to the Tidyverse.
1000 https://joss.theoj.org/papers/10.21105/joss.01686.
1001