1 Long-run bacteria-phage coexistence dynamics under natural habitat conditions in an 2 environmental biotechnology system

3 Leandro D. Guerrero^a (lguerrero@dna.uba.ar), María V. Pérez^{a,b} (mvperez.ingebi@gmail.com),
4 Esteban Orellana^a (esteban.89@gmail.com), Mariana Piuri^c (mpiuri@qb.fcen.uba.ar), Cecilia
5 Quiroga^d (ceciliaquiroga@conicet.gov.ar), Leonardo Erijman^{a,e}

6

7 ^aInstituto de Investigaciones en Ingeniería Genética y Biología Molecular “Dr Héctor N. Torres”
8 (INGEBI-CONICET) Vuelta de Obligado 2490 - C1428ADN; Buenos Aires, Argentina

9 ^bAgua y Saneamientos Argentinos S.A. Tucumán 752 - C1049APP; Buenos Aires, Argentina

10 ^cDepartamento de Química Biológica, Facultad de Ciencias Exactas y Naturales. Universidad de
11 Buenos Aires. Intendente Güiraldes 2160 - C1428EGA; Buenos Aires, Argentina

12 ^dUniversidad de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Tecnológicas,
13 Instituto de Investigaciones en Microbiología y Parasitología Médica (IMPAM), Facultad de
14 Medicina. Paraguay 2155- C1121ABG; Buenos Aires, Argentina

15 ^eDepartamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y
16 Naturales. Universidad de Buenos Aires. Intendente Güiraldes 2160 - C1428EGA; Buenos Aires,
17 Argentina

18

19 *Running title: Host-phage dynamics in an engineered environment*

20

21 #Corresponding author

22 Leonardo Erijman

23 Tel: + 54 11 4783 2871 ext. 36

24 Fax: + 54 11 4786 8578

25 Email: erijman@dna.uba.ar

Abstract

Bacterial viruses are widespread and abundant across natural and engineered habitats. They influence ecosystem functioning through interactions with their hosts. Laboratory studies of phage-host pairs have advanced our understanding of phenotypic and genetic diversification in bacteria and phages. However, the dynamics of phage-host interactions has been seldom recorded in complex natural environments. We conducted an observational metagenomic study of the dynamics of interaction between *Gordonia* and their phages using a three-year data series of samples collected from a full-scale wastewater treatment plant. The aim was to obtain a comprehensive picture of the coevolution dynamics in naturally evolving populations at relatively high time resolution. Co-evolution was followed by monitoring changes over time in the CRISPR loci of *Gordonia* metagenome-assembled genome, and reciprocal changes in the viral genome. Genome-wide analysis indicated low strain variability of *Gordonia*, and almost clonal conservation of the trailer-end of the CRISPR loci. Incorporation of newer spacers gave rise to multiple coexisting bacterial populations. A host population containing a CRISPR array variant, which did not contain spacers against the coexisting phages, accounted for more than half of the total host abundance in the majority of samples. Phages genome co-evolved by introducing directional changes, with no preference for mutations within the protospacer and PAM regions. Metagenomic reconstruction of time-resolved variants of host and virus genomes revealed how selection operates at the population level. In activated sludge, it differed from the arms-race observed in nutrient rich media and resembled the fluctuating selection dynamics observed in natural environments.

47 **Keywords**

48 Virus-host co-evolution, CRISPR, Genome-centric metagenomics, Wastewater treatment,

49 Mycolata, fluctuating selection dynamics

50 Introduction

51 Bacterial viruses (also bacteriophages or phages) are present in large number and are widely
 52 distributed in naturally functioning microbial ecosystems, including engineered environments,
 53 where they have a recognized key role in controlling microbial abundance and composition [1–
 54 3]. Well-designed experiments examining single host-phage pairs in the laboratory have
 55 advanced significantly our understanding of phenotypic and genetic diversification in bacteria
 56 and phages [4–6]. Longitudinal studies revealed that bacterial hosts under phage attack
 57 respond by evolving continuously, using a number of mechanisms to evade phage predation. In
 58 turn, genome rearrangements, mutations and antibacterial defense systems allow phages to
 59 overcome these barriers, leading to an evolutionary arms race [7–11].

60 However, the responses of bacterial cultures following phage infection in laboratory settings do
 61 not necessarily reflect the more complex interactions experienced in real world ecosystems,
 62 where bacterial hosts can be targeted by multiple phages [12], phages can infect more than one
 63 host [13], and density-dependent mechanisms may affect phage-bacteria coexistence [14] and
 64 immunity [15]. In addition, it has been recently demonstrated that bacteria-phage co-
 65 evolutionary interactions in natural environments are strongly affected by the complexity of
 66 biotic networks [16] and by the spatial structure of the environment [17].

67 Therefore, there is a strong need to complement lab-scale studies with field evaluation of the
 68 dynamics of phage-host interactions in naturally occurring microbial communities [18]. Even
 69 with the current advancement of metagenomics, monitoring the coevolutionary dynamics of
 70 phages and bacteria in the environment is not straightforward. Particularly challenging is the

assignment of phages to their microbial hosts [19, 20]. Possibly one of the most suitable tool for the identification of phage–host pairs is the genomic region of clustered regularly interspaced short palindromic repeats (CRISPR), present in about half of bacterial and in most archaeal genomes. Within CRISPR arrays, bacteria and archaea accumulate short sequences of invading DNA, including phages, as a genetic memory of past invasions [21]. The presence of these foreign DNA sequences, named spacers, provides prokaryotes a molecular mechanism of immunity, which allows further defense against new attacks from the same entity [22]. Because a spacer sequence matches a protospacer sequence in the infecting phage, it can be used to pinpoint phage sequences within a complex metagenomic dataset and thus identify phage–host pairs. In addition, since new spacers are added at one end of the locus (the leader end), CRISPR loci provide a reverse chronological record of previous infections. The usefulness of this concept was demonstrated in a pioneer study, where phage and host microbial genome sequences were reconstructed from natural acidophilic biofilms [23]. Metagenomics was later successfully applied to address the role of CRISPR–Cas immunity in shaping microbial communities in a variety of natural environments [24–27].

There remain many unknowns about the dynamics of real populations of phages and bacterial in the environment [28]. A key challenge is to understand how phages and bacteria manage to coexist in complex natural environments, avoiding complete host and virus eradication [29]. To help bridge this gap, we conducted a detailed observational study of the dynamics of phage–bacteria interaction under natural habitat conditions, using a three-year data series of samples collected from a municipal full-scale activated sludge plant [30].

92 In this study, we focused on a bacterial host belonging to the genus *Gordonia*. Member of this
 93 genus have allegedly a prevailing role in biological foaming, a common operational problem
 94 affecting wastewater treatment plants (WWTPs) worldwide [31–34]. We used metagenomic
 95 data to reconstruct CRISPR-Cas loci of *Gordonia* populations. We then searched for
 96 corresponding protospacers in the metavirome, which allowed us to match phages to their
 97 specific bacterial host. The dynamics of phage-host interaction was followed by monitoring
 98 changes over time in the CRISPR loci of *Gordonia* populations, and the reciprocal changes in the
 99 viral genomes. The aim of the study was to obtain a comprehensive picture of phage-host
 100 coevolution in naturally evolving populations within an engineered environment at relatively
 101 high time resolution.

Methods

Wastewater treatment plant

Characteristics of the wastewater treatment plant, and full details of the sampling, nucleic acid extraction, and metagenomic analysis are provided in our previous article [30]. Briefly, samples were obtained from a full-scale municipal WWTP that provides sewage treatment for a population of ca. 600,000 residents of the metropolitan area of Buenos Aires (Argentina). A total of 60 samples of activated sludge collected biweekly over a period of 3 years from one aeration basin were stored at -80°C until analysis.

DNA extraction and sequencing

DNA from sludge samples was extracted by a direct lysis procedure involving physical disruption of cells and a CTAB (cetyltrimethylammonium bromide) purification step [30]. The DNA extracted from the 60 sludge samples were sent to INDEAR (Rosario, Argentina) for Nextera DNA library preparation and sequencing. A rapid-run sequencing on two lanes was performed in a HiSeq 1500 Illumina, generating paired-end (PE) reads of 250 bp.

Gordonia MAG assembly and characterization

Reads were recovered from the *Gordonia* MAG assembled in Perez 2019 [30]. We confirmed that all scaffolds belonged to the same MAG by calculating the abundance and tetranucleotide frequency correlation of each scaffold in the MAG. Relatedness of the MAG to known *Gordonia* species was evaluated on the basis of whole genome pairwise average nucleotide identity (ANI) [35].

The number and abundance of strains in the core *Gordonia* MAG that were present over the whole sampling period was inferred using DESMAN [36]. DESMAN pipeline was applied over marker genes extracted from checkM and filtered by COG annotation (only annotated genes), protein minimum length (330 aa) and total coverage (range 280X to 375X).

Targeted metagenomic CRISPR array reconstruction

CRISPR-Cas system identification was done by detection and analysis of the *cas* operons using CRISPRCasFinder [37] and CRISPRDetect [38]. The leader regions were defined as the sequences spanning from the last *cas* gene to the first repeat.

After identified the repetitive sequence of each CRISPR, bbdduck (bbmap software) [39] was used to extract all the reads from the metagenomic samples containing that sequence allowing 1 mismatch. These reads were used to reconstruct all the detectable variants of that particular CRISPR array as follow:

For each read from the previous step, the position of the repeat sequence was identified and used to trim the read into small sequences compose of seqL-repeat-seqR (s-r-s), where seqL and seqR were the left and right flanking sequences of the repeat. These s-r-s sequences were used as building blocks to reconstruct the CRISPR array network by pairwise alignment of the flanking sequences (s part of the sequence). If two s-r-s aligned by one of their extremes, both sequences were contiguous ($s_1-r-s_2 + s_3-r-s_4 \Rightarrow s_1-r-s_{2=3}-r-s_4$). Therefore, if a sequence was flanked by repeats it had to be a spacer. The leader sequence was identified as the sequence that more often appeared as a “leaf” in the network. Its presence was later checked in the CRISPR-Cas locus of the *Gordonia* MAG. As a result, we obtained a network of all possible

spacers (nodes) connected by repeats (edges). A series of parameters for nodes and edges, such as % mismatch, alignment length, coverage, sequence quality, were calculated and used to manually curate the network. For the analysis, a series of python scripts were created and used in combination with other available scripts from Mothur [40] (**Supplemental Fig S1**).

Phage genomes assembly and annotation

The reconstructed CRISPR were used to predict phage-host associations [20]. We used the sequence of the CRISPR spacers to match phage protospacers in the metagenome obtained from the same longitudinal study. Matching contigs with identical temporal abundance profile were extracted and reassembled using Spades [41]. Base on this approach, two complete or nearly complete phage genomes were reconstructed. These genomes were used as templates to estimate their relative abundance mapping reads using bowtie2 [42] and the JGI script from Metabat2 [43]. When possible, reads from individual samples were assembled independently to identify potential genomics rearrangement.

Phages annotation was performed using DNAMaster [44] and curated manually.

Protospacers were identified in the phage genomes sequences using blastn, allowing for up to 5 mismatches, which were subsequently inspected manually. To identify the PAM sequence, 6 bp at both end of the protospacers were extracted, aligned and visualized using WebLogo [45].

Analysis of sequence variants in phage DC-56 genome

Phage SNVs were identified in samples with a minimum coverage of 10X. The tool *Samtools mpileup* [46] was used to assess the proportion of each variant per position in the phage genome using the read mapping information (BAM files). Each nucleotide position was

considered as an SNV when a mismatch to the consensus nucleotide at that position was supported for at least 5 reads. To reduce the possibility of including false positive SNV due to unspecific recruitment of reads, positions with coverage in excess of two-fold greater than the average coverage of the genome were excluded.

Phylogenetic signal analysis

Differences in phage SNV composition among samples were identified using *bcftools gtcheck* [46] and represented as an error rates matrix. Phylogenetic signal analysis was performed with the R package phylosignal [47], using a cluster of samples based on the number of discordant SNV (error rate matrix) and time as trait. Random generated values and sample time randomization were used as control traits.

Data availability

Sequencing data are available at NCBI BioProject under accession no. PRJNA484416. CRISPR reconstruction scripts are available at <https://github.com/GuerreroCRISPR/Gordonia-CRISPR>

Results

Virus assembly and annotation

Draft genomes of two phages, designated phage DC-56 and phage DS-92, were assembled each into one contig of 56 274 bp and 92 107 bp, respectively (**Supplemental Fig S2**). Structural variants between genomes assembled from individual samples were not detected. Therefore, reconstructed phage genomes were composite assemblies from all reads collected from samples in which phages were present.

G+C content of phages DC-56 and phage DS-92 were 65.2 and 52.4, respectively. Both phages belonged to the *Siphoviridae* family. Phage DC-56 was classified into cluster DC and phage DS-92 into cluster DS, according to the similarity with other *Gordonia* phages and members of the corresponding mycobacteriophage clusters [48, 49].

Phage DC-56 has 67 putative open reading frames (ORFs), 64 of which are transcribed in one direction and only three in the opposite direction (**Supplemental Fig S2A**). Like other *Gordonia* phages, phage DC-56 encodes two integrase genes, a tyrosine integrase (Int-Y, gene 46) and a serine integrase (Int-S, gene 47) near the center of the genome. Phage DS-92 encodes 111 potential ORFs, of which 46 are transcribed rightwards and 65 leftwards. We were not able to identify integrase genes in this genome.

Gordonia metagenome-assembled genome

The *Gordonia* metagenome-assembled genome (MAG) was one of the 173 MAGs obtained in a previous study using data from a time-series analysis of a municipal full-scale activated sludge plant [30]. The refined *Gordonia* sp. MAG comprised 59 contigs, with a size of 5.1 Mbp, N50 of

158 040 and a G+C content of 66.35%. The analysis using single-copy marker genes [50] indicated that the assembly was 98.6% complete with less than 0.3% contamination. The MAG contained a total of 4 512 gene-coding regions, 58 tRNAs, and the complete sequences of 5S, 16S and 23S ribosomal RNAs. Therefore, the MAG complies with all standards of a high-quality draft [51].

Scores of genome pairwise average nucleotide identity computed with 87 genomes of recognized species of the genus *Gordonia* deposited in the NCBI database were <86%, well below the value of 95% that is considered the threshold for novel species definition (**Supplemental Table S1**).

Since *Gordonia* MAG was co-assembled using multiple samples across the time-series, it was important to assess the strain heterogeneity of the resulting assembly. Single-nucleotide variants (SNV) in 216 marker genes and co-occurrence across samples from the longitudinal study were used to investigate strain-level variations in the *Gordonia* MAG. Only 178 significant SNVs were distributed across *Gordonia* MAG, allowing the differentiation into four co-existing strains, with one strain dominating throughout the whole investigated period (**Supplemental Fig S3**). The very low genetic diversity between the four strains suggests that they derived from the same ancestor and most likely did not differ in their functional traits.

CRISPR loci

Gordonia genome contained 2 CRISPR-Cas loci, classified as type I-E and type III-B-like systems, based on the architecture of the Cas operon (**Supplemental Fig S4**). Automatic prediction of CRISPR arrays from MAGs may underestimate the diversity of spacers because of the limited

capacity to resolve the multiple CRISPR variants present in a mixed population. To overcome this problem we used a specific pipeline to reconstruct all observed CRISPR variants, based on the overlapping of every possible connection between spacers identified in raw reads across the complete time-resolved metagenomic dataset (see M&M) [52–54]. The reconstruction of the type I-E CRISPR array, designated CRISPR-1, can be visualized in the network shown in **Fig 1**. The spacer arrangement of CRISPR-1 showed one particular spacer (node circled with a thick line in **Fig 1**) acting as a breaking point, after which the diversity of CRISPR variants increased markedly in the direction of the leader end. Thus, spacers were highly conserved in the direction of the trailer-end of the CRISPR locus, whereas new spacer acquisition led to a considerable increase in the CRISPR population.

The low-coverage, fan-like-structure at the leader end ($\geq 1X$, **Fig 1**) confirmed the active incorporation of new spacers. The majority of these newly acquired spacers perfectly matched sequences of one of the coexisting phages (DC-56). Few spacers also matched sequences of the phage DS-92 genome, notably in the short-branched variants, suggesting simultaneous adaptation to multiple phages. The proximal end of most CRISPR-1 variants contained leader sequences, which are essential for the acquisition of new spacers. Spacers within the consensus region did not match phage DC-56 genome, and no leader sequences could be detected downstream of the breaking point node.

The average coverage of spacers within the consensus was similar to the average coverage of contigs from the complete *Gordonia* MAG, suggesting that the majority of *Gordonia* cells carried the CRISPR-1 array (**Supplemental Fig S5**). Upon filtering the low coverage nodes, 11

main variants of the CRISPR-1 were identified upstream the breaking point node (**Fig 2A**). Most of the spacers within these main branches matched to the co-existing phage DC-56, but four of them showed 1 and 2 mismatches (**Fig 1**). The abundance distribution of the main variants over time is shown in **Fig 2B**. The CRISPR-1 variant with a leader sequence immediately upstream the consensus region (named L1 in **Fig 2**) accounted for more than half of the total abundance in the majority of samples (**Fig 2C**).

The reconstructed type III CRISPR locus, designated CRISPR-2, contained only 8 spacers (**Supplemental fig S4, Supplemental fig S6A**). The fact that the coverage of this locus matched the coverage of the *Gordonia* MAG (**Supplemental fig S6B**) indicates that it was likely also present in most members of the *Gordonia* population. Nevertheless, the lack of spacers matching coexisting phages and the absence of “branching” suggested that this system was either inactive or at least was not actively incorporating new spacers during the time span of this study.

Protospacer adjacent motif sequences

The protospacer adjacent motif (PAM) sequence was searched by alignment of short stretches of sequences (up to 6 bp) immediately flanking the protospacer regions of phage DC-56 sequence. The trinucleotide GTT was highly conserved immediately at the 5' end of those protospacers that matched spacers in CRISPR-1 (**Supplemental fig S7**), which suggests that this is the PAM sequence required for incorporation of new spacers [55].

Phage-host dynamics

Time course of phages and *Gordonia* coexistence is shown in **Fig 2D**. There was considerably overlap in peaks of abundance of bacteria and phage DC-56 over the time of the study (Pearson correlation $r=0.42$, $p=0.0008$; **Supplemental Table S2**), suggesting that bacteria have evolved towards phage resistance. At the same time, it is inferred that phage replication and survival was allowed by the presence of populations of susceptible bacterial cells. This is supported by the fact that peaks of phage DC-56 coincided especially with time points of higher abundance of *Gordonia* bacterial cells lacking spacers against the virus (**L1, Fig 2** Pearson correlation $r=0.50$; $p=4.10^{-5}$; **Supplemental Table S2**). Phage DS-92 peaked with different abundance at two different times, in which *Gordonia* was also present in high abundance (**Fig 2D**).

Adaptation of phage DC-56

To analyze the viral response to CRISPR-1 diversification, we focused on the sequence of phage DC-56 genome, whose abundance mirrored *Gordonia*'s MAG. The 285 spacers that targeted the phage DC-56 genome were distributed across the entire phage genome (**Fig 3**). There is an indication of a moderate spacer acquisition bias, although we found no evidence that the location and strand specificity of protospacers were influenced by previously acquired spacers. We also could not relate genomic regions with less density of protospacers with the presence of higher frequency of the canonical *E. coli* Chi sequence [56]. Nevertheless, with the available data we cannot ruled out that low protospacer density regions are protected by a different Chi sequence that is recognized by *Gordonia*'s RecBCD system, or the involvement of other processes such as interference-driven or primed spacer acquisition [57].

Changes in phage DC-56 genome over three years were investigated through viral SNV analysis. Each nucleotide position was marked as an SNV when a mismatch to the consensus nucleotide at that position had a minimum coverage of 5. A total of 610 SNVs satisfied this criterion. To test for differences in the efficacy of variants selection, we examined the ratios of nonsynonymous to synonymous polymorphisms in protospacers and elsewhere in the phage genome and found that the ratio was not significantly different (two-tailed Fisher's exact test $P=0.66$). Eighty-two SNVs were located within 54 protospacers and 5 PAM regions (**Fig 4**). Fisher's exact test revealed that SNVs do not occur in protospacers more frequently than in the rest of the genome ($p=0.11$).

Phylogenetic signal was used to investigate the pattern of longitudinal population dynamics of phage DC-56. We looked at the relationship between DC-56 sequence in each sample and time [58]. The significant temporal autocorrelation between the DC-56 phage population point to the higher similarity of samples closer in time than samples separated by longer time intervals, an indication of a directional trend in the variation of phage DC-56 genome sequence (**Fig 5**, **Supplemental Fig S8**).

Discussion

This work provides a comprehensive picture of the complex dynamics of naturally evolving populations of a bacterial host and its infecting phage within an engineered environment. Metagenomic reconstruction of time-resolved variants of host and virus genomes revealed how selection operates at the population level. Diversification of the bacterial CRISPR locus, which occurs as a result of phage infection, gives rise to multiple coexisting populations of a *Gordonia*

strain. The majority of newly acquired spacers matched sequences of one of the coexisting phages. These observations coincide with simulations [59], experiments performed under laboratory settings [60, 61] and observation from natural environments [27], confirming that CRISPR array evolution is driven by co-occurring phages and suggesting that the synergistic effect of spacer diversity established at the host population level aim at reducing the probability of phages to escape mutations [60, 61].

Phage-host dynamics

Under the arms race dynamics, phages would co-evolve to escape CRISPR by preferentially selecting variants with polymorphic positions at protospacer regions. For example, directional changes in the phage genome were observed upon long-term coinubation of *S. thermophilus* DGCC7710 and phage 2972 in milk medium [10]. This effect, which has also been observed in other phages, helps them to overcome sequence-specific CRISPR immunity [62, 63]. In contrast, in our study we observed directional changes, yet accumulation of mutations in the PAM and protospacer regions was not favored (**Fig 4**). This is not surprising, given that phages that escape bacterial immunity by mutating their target protospacers within one host will not be successful against coexisting bacterial populations that possess different spacers [64]. Yet, as long as no individual bacterial population achieves a much higher density than their competitors, phages would still find excess of susceptible hosts, reducing the selection pressure to mutate their targeted protospacers or PAM regions. As a matter of fact, the most abundant *Gordonia* population observed throughout this study contained a CRISPR array with no spacers matching protospacer sequences of phage DC-56 (L1 in **Fig 2**). The continuous presence of a

large population of susceptible host may act as a propagating host for the phage DC-56, granting its endurance [11, 65]. Nevertheless, we cannot tell whether this is a non-immune population, associated with the fitness cost of maintaining a fully active CRISPR-Cas system [66, 67] or that the population maintains active different defense mechanisms or it is protected from phage exposure by the spatial heterogeneity of the environment [68].

It follows that in the long run neither the host nor the virus are driven extinct, in opposition to the results of laboratory coevolution experiments, where spacer diversity of CRISPR-mediated resistance caused rapid phage extinction [60]. Rather than a typical unidirectional arms race dynamics, the outcome of phage-*Gordonia* coevolution in activated sludge most closely resembled a fluctuating selection dynamics (FSD) [69]. FSD has been reported to occur in environments with low resource availability, such as soil [70] and the ocean [71], where the costs of phage resistance are expected to be high [72–74]. This is not entirely unexpected considering that *Gordonia* thrives in activated sludge, where bacteria grow under carbon limitation conditions, with typical loading rates in the range of 0.05 to 0.4 kg Biochemical Oxygen Demand per kg dry biomass and per day [30, 75, 76]. Yet, we cannot confirm whether the fluctuation of phage and bacterial genotypes is associated to a negative frequency-dependent selection of virus and hosts with different infectivity and resistance profiles [18].

CRISPR-Cas systems are one among several mechanisms that have evolved in bacteria to fight phages [73, 77]. In its natural habitat, *Gordonia* would be protected by several lines of defense, including the protecting effect of EPS and spatial heterogeneity of the matrix. However, the fact that phage DC-56 appears to evolve relatively slowly in this environment suggests that it is the

continuous presence of large number of susceptible cells that makes it possible to maintain the stable phage-host coexistence.

A potential limitation of this study is that we cannot rule out the existence of additional hosts of the phages. Nevertheless, the dynamics of *Gordonia*-DC-56 phage coexistence shows a pattern in which the most abundant phages replicate actively as long as the host populations remain susceptible to their attack, as predicted by the “killing the winner” model [78]. The appearance of peaks of infection by a second phage are in turn consistent with the seed bank model, which postulates that minor virus populations are maintained in a bank fractions and can potentially become active when the environment changes [13].

Conservation of trailer-end spacers

This work confirms earlier observations that spacers at the trailer-end of the CRISPR locus is highly conserved despite the fact that they would not provide immunity against coexisting viruses [24, 79, 80]. Simulations have predicted that trailer-ends may become clonal across time due to selective sweeps of highly immune CRISPR lineages [81]. More generally, the purge of genetic heterogeneity at the trailer-end of the CRISPR locus appears to conform to the genome-wide sweep predicted by the ecotype model [82], where a single population (ecotype) acquires an advantageous trait that helps it to survive, reducing overall population diversity. The occurrence of genome-wide selective sweeps has already been revealed in freshwater lake microbial communities, where with few exceptions, overall genetic heterogeneity of most bacterial populations was relatively low over a period of several years [83]. Nonetheless, the low bacterial strain diversity and the existence of a conserved trailer-end is puzzling because it

suggests that the entire *Gordonia* population observed over 3 years appears to share a common ancestry, despite the fact that *Gordonia* are widespread in the environment and activated sludge is an open ecosystem, in which immigration plays an important role in community assembly [84–86].

Implications for the application of phage therapy in environmental processes

Gordonia belongs to a group of Actinobacteria that have relatively high content of mycolic acids on their cell surface, called thereafter Mycolata. The highly hydrophobic cell surface of Mycolata in combination with their ability to excrete extracellular polymers, which may act as biosurfactants, favor the formation and stabilization of foams in wastewater treatment processes [87, 88]. Despite major efforts to find universal strategies to prevent or to control the unwanted proliferation of these bacteria, foaming remains a common operational problem in activated sludge. In recent years, considerable research has been directed towards the use of bacteriophages as an alternative approach to chemicals to reduce the levels of filamentous bacteria in wastewater treatment, including foam forming *Gordonia* [34, 88, 89]. The results obtained in this work suggest that the presence of an active lytic phage might not be sufficient to control *Gordonia* abundance. Successful application of phage therapy at the full-scale will still require a better understanding of virus-host interactions in the complex real-world context.

Conclusions

In activated sludge, the pattern of coevolution of *Gordonia* and its phages differed from the arms-race dynamic observed in nutrient rich media and resembled the fluctuating selection

dynamics observed in other natural environments with low resource availability. This hypothesis can be tested in further investigation on phage infectivity and host range and bacterial resistance [5, 27], as well as the protection effect due to spatial refuges for susceptible bacteria [90].

More than fifteen years ago, there was a call proposing wastewater treatment as model systems to address fundamental ecological questions in realistic environmental conditions [91, 92]. We renew this call, proposing that wastewater treatment, in conjunction with metagenomic time-series studies, are suitable means to gain further insight into bacteria-phage dynamics and interaction in real populations.

Acknowledgments

This work was supported by grants from AySA-CONICET (Res. 3816/11 and 1371/15) and FONCyT (PICT 0746/15). The funders had no role in study design, data analysis, decision to publish, or preparation of the manuscript.

The authors declare no competing financial interests.

References

1. Liu R, Qi R, Wang J, Zhang Y, Liu X, Rossetti S, et al. Phage-host associations in a full-scale activated sludge plant during sludge bulking. *Appl Microbiol Biotechnol* 2017; **101**: 6495–6504.
2. Bouvier T, del Giorgio PA. Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environ Microbiol* 2007; **9**: 287–297.
3. Brown MR, Baptista JC, Lunn M, Swan DL, Smith SJ, Davenport RJ, et al. Coupled virus - bacteria interactions and ecosystem function in an engineered microbial system. *Water Res* 2019; **152**: 264–273.
4. Shapiro OH, Kushmaro A, Brenner A. Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J* 2010; **4**: 327–336.
5. Buckling A, Rainey PB. Antagonistic coevolution between a bacterium and a bacteriophage. *Proc R Soc B Biol Sci* 2002; **269**: 931–936.
6. Fineran PC, Gerritzen MJH, Suárez-Diez M, Kuñne T, Boekhorst J, Van Hijum SAFT, et al. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A* 2014; **111**.
7. Stern A, Sorek R. The phage-host arms race: Shaping the evolution of microbes. *BioEssays* 2011; **33**: 43–51.
8. Rostøl JT, Marraffini L. (Ph)ighting Phages: How Bacteria Resist Their Parasites. *Cell Host Microbe* . 2019. Cell Press. , **25**: 184–194
9. McLaughlin RN, Malik HS. Genetic conflicts: The usual suspects and beyond. *J Exp Biol* . 2017. Company of Biologists Ltd. , **220**: 6–17
10. Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, et al. CRISPR immunity drives rapid phage genome evolution in streptococcus thermophilus. *MBio* 2015; **6**: 1–9.
11. Common J, Morley D, Westra ER, Van Houte S. CRISPR-Cas immunity leads to a coevolutionary arms race between Streptococcus thermophilus and lytic phage. *Philos Trans R Soc B Biol Sci* 2019; **374**.

- 434 12. Flores CO, Valverde S, Weitz JS. Multi-scale structure and geographic drivers of cross-
435 infection within marine bacteria and phages. *ISME J* 2013; **7**: 520–532.
- 436 13. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends*
437 *Microbiol* 2005; **13**: 278–284.
- 438 14. Heilmann S, Snepken K, Krishna S. Coexistence of phage and bacteria on the boundary of
439 self-organized refuges. *Proc Natl Acad Sci U S A* 2012; **109**: 12828–12833.
- 440 15. Patterson AG, Jackson SA, Taylor C, Evans GB, Salmond GPC, Przybilski R, et al. Quorum
441 Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas
442 Systems. *Mol Cell* 2016; **64**: 1102–1108.
- 443 16. Alseth EO, Pursey E, Luján AM, McLeod I, Rollie C, Westra ER. Bacterial biodiversity drives
444 the evolution of CRISPR-based phage resistance. *Nature* 2019; **574**: 549–552.
- 445 17. Sousa JAM de, Rocha EPC. Environmental structure drives resistance to phages and
446 antibiotics during phage therapy and to invading lysogens during colonisation. *Sci Rep*
447 2019; **9**: 1–13.
- 448 18. Scanlan PD. Bacteria–Bacteriophage Coevolution in the Human Gut: Implications for
449 Microbial Diversity and Functionality. *Trends Microbiol* 2017; **25**: 614–623.
- 450 19. Dang VT, Sullivan MB. Emerging methods to study bacteriophage infection at the single-
451 cell level. *Front Microbiol* 2014; **5**: 724.
- 452 20. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict
453 bacteriophage-host relationships. *FEMS Microbiol Rev* 2016; **40**: 258–272.
- 454 21. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR
455 provides acquired resistance against viruses in prokaryotes. *Science* 2007; **315**: 1709–
456 1712.
- 457 22. McGinn J, Marraffini LA. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat*
458 *Rev Microbiol* 2019; **17**: 7–12.
- 459 23. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in
460 natural microbial communities. *Science* 2008; **320**: 1047–1050.
- 461 24. Sun CL, Thomas BC, Barrangou R, Banfield JF. Metagenomic reconstructions of bacterial
462 CRISPR loci constrain population histories. *ISME J* 2016; **10**: 858–870.
- 463 25. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. CRISPR targeting reveals a reservoir of common
464 phages associated with the human gut microbiome. *Genome Res* 2012; **22**: 1985–94.

- 465 26. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, et al. Virus-host
466 and CRISPR dynamics in archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia.
467 *Archaea* 2013; **2013**.
- 468 27. Laanto E, Hoikkala V, Ravantti J, Sundberg LR. Long-term genomic coevolution of host-
469 parasite interaction in the natural environment. *Nat Commun* 2017; **8**: 111.
- 470 28. Westra ER, Van Houte S, Gandon S, Whitaker R. The ecology and evolution of microbial
471 CRISPR-Cas adaptive immune systems. *Philos Trans R Soc B Biol Sci* 2019; **374**.
- 472 29. Fernández L, Rodríguez A, García P. Phage or foe: an insight into the impact of viral
473 predation on microbial communities. *ISME J* 2018; **12**: 1171–1179.
- 474 30. Pérez MV, Guerrero LD, Orellana E, Figuerola EL, Erijman L. Time Series Genome-Centric
475 Analysis Unveils Bacterial Response to Operational Disturbance in Activated Sludge.
476 *mSystems* 2019; **4**: e00169-19.
- 477 31. de los Reyes MF, de los Reyes FL, Hernandez M, Raskin L. Quantification of *Gordonia*
478 *amarae* strains in foaming activated sludge and anaerobic digester systems with
479 oligonucleotide hybridization probes. *Appl Environ Microbiol* 1998; **64**: 2503–12.
- 480 32. Guo F, Wang ZP, Yu K, Zhang T. Detailed investigation of the microbial community in
481 foaming activated sludge reveals novel foam formers. *Sci Rep* 2015; **5**: 1–9.
- 482 33. Kragelund C, Remesova Z, Nielsen JL, Thomsen TR, Eales K, Seviour R, et al. Ecophysiology
483 of mycolic acid-containing Actinobacteria (Mycolata) in activated sludge foams. *FEMS*
484 *Microbiol Ecol* 2007; **61**: 174–184.
- 485 34. Liu M, Gill JJ, Young R, Summer EJ. Bacteriophages of wastewater foaming-associated
486 filamentous *Gordonia* reduce host levels in raw activated sludge. *Sci Rep* 2015; **5**: 13754.
- 487 35. Rodriguez-R LM, Konstantinidis KT. Bypassing Cultivation To Identify Bacterial Species,
488 Culture-independent genomic approaches identify credibly distinct clusters, avoid
489 cultivation bias, and provide true insights into microbial species. *Microbe* 2014; **9**: 111–
490 117.
- 491 36. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al. DESMAN: a
492 new tool for de novo extraction of strains from metagenomes. *Genome Biol* 2017; **18**:
493 181.
- 494 37. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, et al.
495 CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced

496 performance and integrates search for Cas proteins. *Nucleic Acids Res* 2018; **46**: W246–
497 W251.

498 38. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible
499 algorithm to define CRISPR arrays. *BMC Genomics* 2016; **17**: 356.

500 39. Bushnell B. BBTools - DOE Joint Genome Institute. [https://jgi.doe.gov/data-and-](https://jgi.doe.gov/data-and-tools/bbtools/)
501 [tools/bbtools/](https://jgi.doe.gov/data-and-tools/bbtools/). .

502 40. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing
503 mothur: open-source, platform-independent, community-supported software for
504 describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**:
505 7537–7541.

506 41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new
507 genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*
508 2012; **19**: 455–77.

509 42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;
510 **9**: 357–359.

511 43. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
512 reconstructing single genomes from complex microbial communities. *PeerJ* 2015; **3**:
513 e1165.

514 44. Pope WH, Jacobs-Sera D. Annotation of bacteriophage genome sequences using DNA
515 master: An overview. *Methods in Molecular Biology*. 2018. pp 217–229.

516 45. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator.
517 *Genome Res* 2004; **14**: 1188–1190.

518 46. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
519 population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;
520 **27**: 2987–2993.

521 47. Keck F, Rimet F, Bouchez A, Franc A. PhyloSignal: An R package to measure, test, and
522 explore the phylogenetic signal. *Ecol Evol* 2016; **6**: 2774–2780.

523 48. Pope WH, Mavrich TN, Garlena RA, Guerrero-Bustamante CA, Jacobs-Sera D,
524 Montgomery MT, et al. Bacteriophages of *Gordonia* spp. Display a spectrum of diversity
525 and genetic relationships. *MBio* 2017; **8**.

526 49. Russell DA, Hatfull GF. PhagesDB: The actinobacteriophage database. *Bioinformatics*
527 2017; **33**: 784–786.

528 50. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the
529 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
530 *Genome Res* 2015; **25**: 1043–1055.

531 51. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al.
532 Minimum information about a single amplified genome (MISAG) and a metagenome-
533 assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017; **35**: 725–731.

534 52. Skennerton CT, Imelfort M, Tyson GW. Crass: Identification and reconstruction of CRISPR
535 from unassembled metagenomic data. *Nucleic Acids Res* 2013; **41**: e105.

536 53. Lei J, Sun Y. Assemble CRISPRs from metagenomic sequencing data. *Bioinformatics* 2016;
537 **32**: i520–i528.

538 54. Lam TJ, Ye Y. Long reads reveal the diversification and dynamics of CRISPR reservoir in
539 microbiomes. *BMC Genomics* 2019; **20**.

540 55. Leenay RT, Maksimchuk KR, Slotkowski RA, Agrawal RN, Gomaa AA, Briner AE, et al.
541 Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol Cell*
542 2016; **62**: 137–147.

543 56. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, et al. CRISPR adaptation biases
544 explain preference for acquisition of foreign DNA. *Nature* 2015; **520**: 505–510.

545 57. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJJ. CRISPR-Cas:
546 Adapting to change. *Science* 2017; **356**.

547 58. Ellington MJ, Heinz E, Wailan AM, Dorman MJ, de Goffau M, Cain AK, et al. Contrasting
548 patterns of longitudinal population dynamics and antimicrobial resistance mechanisms in
549 two priority bacterial pathogens over 7 years in a single center. *Genome Biol* 2019; **20**:
550 184.

551 59. Childs LM, England WE, Young MJ, Weitz JS, Whitaker RJ. CRISPR-induced distributed
552 immunity in microbial populations. *PLoS One* 2014; **9**.

553 60. Van Houte S, Ekroth AKE, Broniewski JM, Chabas H, Ashby B, Bondy-Denomy J, et al. The
554 diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* 2016;
555 **532**: 385–388.

- 556 61. Jackson SA, Birkholz N, Malone LM, Fineran PC. Imprecise Spacer Acquisition Generates
557 CRISPR-Cas Immune Diversity through Primed Adaptation. *Cell Host Microbe* 2019; **25**:
558 250-260.e4.
- 559 62. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, et al. Phage
560 response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* 2008;
561 **190**: 1390–1400.
- 562 63. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, et al.
563 Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is
564 governed by a seed sequence. *Proc Natl Acad Sci U S A* 2011; **108**: 10098–10103.
- 565 64. Chabas H, Nicot A, Meaden S, Westra ER, Tremblay DM, Pradier L, et al. Variability in the
566 durability of CRISPR-Cas immunity. *Philos Trans R Soc B Biol Sci* 2019; **374**.
- 567 65. Chabas H, Lion S, Nicot A, Meaden S, van Houte S, Moineau S, et al. Evolutionary
568 emergence of infectious diseases in heterogeneous host populations. *PLoS Biol* 2018; **16**:
569 e2006738.
- 570 66. Levin SA, Udovic JD. A Mathematical Model of Coevolving Populations. *Am Nat* 1977;
571 **111**: 657–675.
- 572 67. Lenski RE. Coevolution of bacteria and phage: Are there endless cycles of bacterial
573 defenses and phage counterdefenses? *J Theor Biol* 1984; **108**: 319–325.
- 574 68. Bull JJ, Christensen KA, Scott C, Jack BR, Crandall CJ, Krone SM. Phage-bacterial dynamics
575 with spatial structure: Self organization around phage sinks can promote increased cell
576 densities. *Antibiotics* 2018; **7**: 8.
- 577 69. Hall AR, Scanlan PD, Morgan AD, Buckling A. Host-parasite coevolutionary arms races
578 give way to fluctuating selection. *Ecol Lett* 2011; **14**: 635–642.
- 579 70. Gómez P, Buckling A. Bacteria-phage antagonistic coevolution in soil. *Science* 2011; **332**:
580 106–109.
- 581 71. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. Long-term stability and Red Queen-like
582 strain dynamics in marine viruses. *Nat Microbiol* . 2020. Nature Research. , **5**: 265–271
- 583 72. Lopez Pascua L, Hall AR, Best A, Morgan AD, Boots M, Buckling A. Higher resources
584 decrease fluctuating selection during host-parasite coevolution. *Ecol Lett* 2014; **17**: 1380–
585 1388.

- 586 73. Hampton HG, Watson BNJ, Fineran PC. The arms race between bacteria and their phage
587 foes. *Nature* . 2020. Nature Research. , **577**: 327–336
- 588 74. Brockhurst MA, Chapman T, King KC, Mank JE, Paterson S, Hurst GDD. Running with the
589 Red Queen: The role of biotic conflicts in evolution. *Proc R Soc B Biol Sci* 2014; **281**.
- 590 75. Vuono DC, Benecke J, Henkel J, Navidi WC, Cath TY, Munakata-Marr J, et al. Disturbance
591 and temporal partitioning of the activated sludge metacommunity. *ISME J* 2015; **9**: 425–
592 435.
- 593 76. Kuai L, Verstraete W. Ammonium removal by the oxygen-limited autotrophic
594 nitrification- denitrification system. *Appl Environ Microbiol* 1998; **64**: 4500–4506.
- 595 77. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of
596 antiphage defense systems in the microbial pangenome. *Science* 2018; **359**: eaar4120.
- 597 78. Winter C, Bouvier T, Weinbauer MG, Thingstad TF. Trade-Offs between Competition and
598 Defense Specialists among Unicellular Planktonic Organisms: the ‘Killing the Winner’
599 Hypothesis Revisited. *Microbiol Mol Biol Rev* 2010; **74**: 42–57.
- 600 79. Tyson GW, Banfield JF. Rapidly evolving CRISPRs implicated in acquired resistance of
601 microorganisms to viruses. *Environ Microbiol* 2008; **10**: 200–7.
- 602 80. Pauly MD, Bautista MA, Black JA, Whitaker RJ. Diversified local CRISPR-Cas immunity to
603 viruses of *Sulfolobus islandicus*. *Philos Trans R Soc B Biol Sci* 2019; **374**: 20180093.
- 604 81. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin E V. Viral diversity threshold
605 for adaptive immunity in prokaryotes. *MBio* 2012; **3**.
- 606 82. Cohan FM, Perry EB. A Systematics for Discovering the Fundamental Units of Bacterial
607 Diversity. *Curr Biol* 2007; **17**: R373–R386.
- 608 83. Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-
609 wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J*
610 2016; **10**: 1589–1601.
- 611 84. Sloan WT, Lunn M, Woodcock S, Head IM, Nee S, Curtis TP. Quantifying the roles of
612 immigration and chance in shaping prokaryote community structure. *Environ Microbiol*
613 2006; **8**: 732–740.
- 614 85. Ofițeru ID, Lunn M, Curtis TP, Wells GF, Criddle CS, Francis CA, et al. Combined niche and
615 neutral effects in a microbial wastewater treatment community. *Proc Natl Acad Sci U S A*
616 2010; **107**: 15345–15350.

86. Ayarza JM, Erijman L. Balance of Neutral and Deterministic Components in the Dynamics of Activated Sludge Floc Assembly. *Microb Ecol* 2011; **61**: 486–495.
87. de los Reyes FL. Foam in Wastewater Treatment Facilities. *Handbook of Hydrocarbon and Lipid Microbiology*. 2010. Springer Berlin Heidelberg, pp 2401–2411.
88. Petrovski S, Seviour RJ, Tillett D. Prevention of *Gordonia* and *Nocardia* stabilized foam formation by using bacteriophage GTE7. *Appl Environ Microbiol* 2011; **77**: 7864–7867.
89. Petrovski S, Seviour R. Activated sludge foaming: Can phage therapy provide a control strategy? *Microbiol Aust* 2018; **39**: 162–164.
90. Hosseinioust Z, Tufenkji N, van de Ven TGM. Formation of biofilms under phage predation: considerations concerning a biofilm increase. *Biofouling* 2013; **29**: 457–468.
91. Curtis TP, Head IM, Graham DW. Peer Reviewed: Theoretical Ecology for Engineering Biology. *Environ Sci Technol* 2003; **37**: 64A–70A.
92. Daims H, Taylor MW, Wagner M. Wastewater treatment: a model system for microbial ecology. *Trends Biotechnol* 2006; **24**: 483–489.

Legends to the Figs

Fig 1: Reconstruction of the spacer arrangement of *Gordonia*'s CRISPR-1 across the complete set of metagenomes from Perez et al. [30]. Each circle represents a spacer. Leader sequence are shown as pentagons. The lines connecting each spacer represent their position relative to other spacers, with arrows pointing to the direction of the leader sequence. Coverage is indicated by the diameter of the nodes and the width and color of the lines. Every combination of nodes connected by edges pointing in the same direction represents a possible CRISPR variant in the population. Spacers targeting phage DS-92 and phage DC-56 are colored with yellow and green, respectively. Spacer with mismatches with phages' protospacers are marked with an X.

Fig 2: Abundance of the main CRISPR-1 variants over time. A) Simplified representation of Fig 1 showing nodes with coverage > 1 of the main variants. Nodes used to estimate the average abundance of each variant are identified by colors. B) Abundance of each variant per sample. C) Sum of all variants in panel B compared to variant L1. D) Time course of the abundance of the *Gordonia* MAG and the phages DC-56 and DS-92.

Fig 3: Phage DC-56 genome map highlighting the position of protospacers. Genes are indicated by boxes, colored according to their forward (green) or reverse (yellow) orientation. When known, the name of the genes was added. The location of protospacers targeted by CRISPR-1 was indicated by verticals bars. Bar colors indicate the orientation and the similarity to the corresponding spacer sequence in the *Gordonia* MAG.

654

655 **Fig 4:** The 82 single-nucleotide variants (SNV) of phage DC-56 affecting protospacers and PAM
 656 regions. Only samples with a total coverage ≥ 10 and a minimum coverage of 5 reads per
 657 sample were analyzed. Spacers are drawn as arrows to indicate orientation. Bases within the
 658 same spacer are connected by a gray bar. Bases in spacers located in the reverse orientation
 659 were colored according to the corresponding forward sequence.

660

661 **Fig 5:** Phylogenetic signal analysis of reconstructed phage DC-56. A) Cluster of phage DC-56
 662 genome based on shared SNV in samples with coverage ≥ 10 and their correspondence to
 663 phage abundance profile (coverage) over time. B) Correlogram of the distance between
 664 samples using time as trait. Red bar indicates positive correlation.

665

666

667

Legends to Supplemental tables and Figs

Supplemental table S1. Pairwise ANI of *Gordonia* metagenome assembled genome compared to all available *Gordonia* genomes on NCBI (at March 2020)

Supplemental Table S2. Pearson correlation between the abundance of phage DC-56 and the abundance of the different CRISPR-1 variants over the time span of the study.

Fig S1. Schematic representation of the pipeline used to reconstruct the CRISPR arrays. A detailed explanation and scripts used can be found at <https://github.com/GuerreroCRISPR/Gordonia-CRISPR>.

Fig S2. Genome maps of *Gordonia* phages DC-56 (A) and DS-92 (B). Genes are indicated by boxes colored by orientation (green and red, for forward and reverse genes respectively). Ruler scale: 1000 bp.

Fig S3. Abundance of co-existing *Gordonia* strains over time.

Fig S4. CRISPR-Cas systems from *Gordonia* sp. A) Type I-E system. B) Type III-B-like system. Repeats are depicted as gray diamonds (Rn) and spacers as small vertical rectangles (Sn). Horizontal arrows represent the Cas operon, empty arrows represent putative RAMP genes and small subunit cas11. RT represents the group II introns reverse transcriptase.

689

690 **Fig S5.** Average coverage of the spacers within the consensus CRISPR-1 arrangement in

691 *Gordonia* MAG (red line) and total *Gordonia* MAG (blue line) over time.

692

693 **Fig S6.** A) Reconstruction of the spacer arrangement of *Gordonia*'s CRISPR-2. B) Average

694 coverage of the spacers within the CRISPR-2 arrangement in *Gordonia* MAG (orange line) and

695 total *Gordonia* MAG (blue line) over time.

696

697 **Fig S7.** WebLogo for the PAM consensus sequence predicted for phage DC-56 required for

698 CRISPR-1.

699

700 **Fig S8.** Phylogenetic signal analysis for phage DC-56 genome over time. Randomized time and

701 random generated values were used as control traits. A) Data traits visualization mapped along

702 the samples cluster. B) Correlograms for traits. Red bar indicates positive correlation.

703

704

705

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5