AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Check for updates

| Bacteriophages | Research Article

Host translation machinery is not a barrier to phages that interact with both CPR and non-CPR bacteria

Jett Liu,[1,2] Alexander L. Jaffe,[1,3] LinXing Chen,[4,5] Batbileg Bor,[2,6] Jillian F. Banfield[4,5,7]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT Within human microbiomes, Gracilibacteria, Absconditabacteria, and Saccharibacteria, members of Candidate Phyla Radiation (CPR), are increasingly correlated with human oral health and disease. We profiled the diversity of CRISPR-Cas systems in the genomes of these bacteria and sought phages that are capable of infecting them by matching their spacer inventories to large phage sequence databases. Gracilibacteria and Absconditabacteria recode the typical TGA stop codon to glycine and are putatively infected by phages that share their host's alternate genetic code. Unexpectedly, however, other predicted phages of Gracilibacteria and Absconditabacteria do not use an alternative genetic code. Some of these phages may infect both alternatively coded CPR bacteria and standard-coded bacteria. These phages typically rely on other stop codons besides TGA and thus should be capable of producing viable gene products in either bacterial host type. By avoiding the acquisition of in-frame stop codons, these phages may have a broadened host range. Interestingly, we additionally predict that some phages of Saccharibacteria are targeted by spacers encoded in Actinobacteria, a phylum that includes known hosts for episymbiotic Saccharibacteria.

IMPORTANCE Here, we profiled putative phages of Saccharibacteria, which are of particular importance as Saccharibacteria influence some human oral diseases. We additionally profiled putative phages of Gracilibacteria and Absconditabacteria, two Candidate Phyla Radiation (CPR) lineages of interest given their use of an alternative genetic code. Among the phages identified in this study, some are targeted by spacers from both CPR and non-CPR bacteria and others by both bacteria that use the standard genetic code as well as bacteria that use an alternative genetic code. These findings represent new insights into possible phage replication strategies and have relevance for phage therapies that seek to manipulate microbiomes containing CPR bacteria.

KEYWORDS CPR bacteria, CRISPR-Cas systems, bacteriophage evolution, bacteriophage genetics, bioinformatics

Interest in human microbiome-associated Saccharibacteria, Gracilibacteria, and Absconditabacteria (hereafter referred to as SGA) has increased, in part due to their association with disease (1). SGA are lineages within the Candidate Phyla Radiation (CPR), a monophyletic radiation within the domain Bacteria, characterized in part by consistently reduced genomes, small cell sizes, and limited metabolic capabilities (2). CPR bacteria adhere to lifestyles dependent upon other cells, either by episymbiotic attachment—whereby CPR cells attach to and obtain nutrients from a larger host bacterium—or by deriving essential compounds such as lipids (3) from the surrounding microbial community. In most cases, the hosts of CPR bacteria are unknown, but in the case of certain oral and environmental Saccharibacteria, the hosts have been experimentally established to be species of Actinobacteria (4–8). The attachment by Saccharibacteria can have a profound impact on the Actinobacteria host, leading to

Invited Editor Natalya Yutin, National Institutes of Health, Bethesda, Maryland, USA

Editor Igor B. Zhulin, The Ohio State University, Columbus, Ohio, USA

Address correspondence to Jillian F. Banfield, jbanfield@berkeley.edu.

J.F.B. is a founder of Metagenomi.

See the funding table on p. 14.

Received 2 August 2023
Accepted 12 October 2023

Published 27 November 2023

cycles of rapid host evolution and drastic changes in host physiology (4, 6, 9). The Saccharibacteria-host-bacteria relationship in the human oral cavity has recently been evaluated *in vivo*, demonstrating that Saccharibacteria reduces the inflammatory effects of periodontitis and the pathogenicity of their host Actinobacteria (10). These studies have catalyzed a paradigm shift from the previous characterization of Saccharibacteria as a likely pathogen (11, 12).

In contrast to an episymbiotic lifestyle, one Saccharibacteria species and several Gracilibacteria and Absconditabacteria species are thought to live predatory lifestyles, whereby they feed on specific non-CPR bacteria (6, 8, 13, 14). Predatory bacteria are an emerging area of research garnering interest as an antibiotic alternative with narrow, targeted effects (15, 16). The predatory Saccharibacteria *Ca. M. amalyticus*, for instance, has been proposed as a tool to precisely consume mycolata bacteria that are recalcitrant to antibiotic and phage treatments (8).

Another intriguing feature of Gracilibacteria and Absconditabacteria is that they employ an alternative genetic code in which the canonical stop codon, TGA, is instead recognized as glycine (genetic code 25) (17–19). While the alternative genetic code of Absconditabacteria and Gracilibacteria is well-established, very little is known about the genetic code of their phages. It has become clear that phages can adopt a genetic code that is distinct from that of their hosts (20–23). For example, phages that have reassigned the TAG stop codon to be translated as glutamine infect *Prevotella* that use the standard bacterial code (21). These alternatively coded phages encode in-frame stop codons within late-stage phage genes to likely prevent premature production of structural and lytic proteins (21, 23). To enable the production of these proteins in bacteria that use the standard code, these phage genomes must utilize "code-switching" machinery. These findings raise the possibility that standard-coded phages can replicate in bacteria with alternatively coded genomes, but this question has not been comprehensively investigated to date. Here, we explored the diversity and genomic features, including the genetic codes, of phages that are predicted to infect SGA bacteria. In addition to expanding our knowledge of fundamental biology, phages of SGA bacteria could have practical importance, as phages can be used to alter the composition of microbiomes with species or strain specificity (24, 25).

RESULTS

CRISPR-Cas systems within SGA

As CRISPR spacers are fragments of phage genomes stored within CRISPR-Cas systems, a common technique used to link phages to their bacterial hosts is via spacer-phage matching (26–29). To find CRISPR-Cas systems encoded within SGA bacteria, we began with a previously compiled database that contained 861 genomes from the SGA lineages (30) (See Tables S1 and S2 at https://doi.org/10.5281/zenodo.8422333). SGA bacteria in this database are from a wide array of environments, including human microbiome, non-human animal microbiome, soil, freshwater, and marine ecosystems. We de-replicated the database at 99% average nucleotide identity (ANI) to form a non-redundant set of 391 genomes (318 Saccharibacteria genomes, 44 Gracilibacteria genomes, and 27 Absconditabacteria genomes; See Table S1 at https://doi.org/10.5281/zenodo.8422333).

To survey the incidence of complete CRISPR-Cas systems within our genome set, we searched for Cas loci using the full suite of TIGRFAM HMM profiles (see Materials and Methods) within the genomes that contained high-confidence CRISPR arrays predicted by CRISPRCasFinder (CCF). We manually examined scaffolds that contained *cas* gene annotations to ensure that they originated from one of our three SGA lineages (see Table S3 at https://doi.org/10.5281/zenodo.8422333). We identified 43 CRISPR-Cas systems present in our non-redundant database (Fig. 1A; see also Table S4 at https://doi.org/10.5281/zenodo.8422333). Encoding at least one CRISPR-Cas system were: 16 Gracilibacteria genomes (among 44 genomes—36.3% prevalence), 22 Saccharibacteria genomes (among 318 genomes—7.9% prevalence), and 2 Absconditabacteria genomes (among 27 genomes—7.4% prevalence). The 36.3% prevalence of CRISPR-Cas systems among

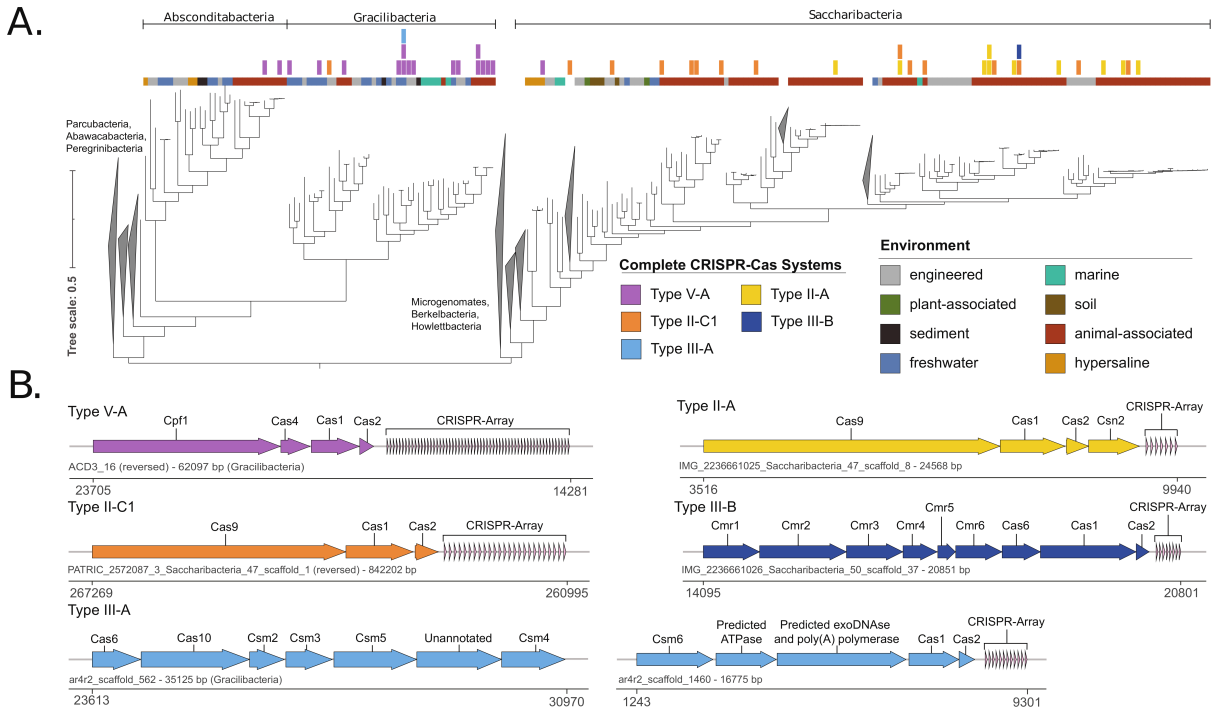

FIG 1 Distribution of CRISPR-Cas Systems in SGA bacteria. (A) Maximum-likelihood tree based on 16 concatenated ribosomal proteins (see Materials and Methods). The identified CRISPR-Cas systems and the environmental origin of genomes are overlaid above. (B) Gene architecture of representative CRISPR-Cas types identified in our SGA database. Color corresponds to the system types displayed in panel A. Below the gene diagrams are the name and size of the scaffold encoding the system, along with the chromosome coordinates of the system.

Gracilibacteria genomes is substantially higher than reported rates for other CPR bacteria (31) and closer to the typical CRISPR-Cas system prevalence across the domain Bacteria (~39%) (32, 33).

When comparing the environmental origin of the genomes containing CRISPR-Cas systems, there is an apparent discrepancy in system distribution between the three SGA lineages. In Saccharibacteria, as has been previously observed (30, 34, 35), CRISPR-Cas systems are abundant in human and other mammal microbiomes and scarce in other environments. Only 3 of the 25 CRISPR-Cas systems in Saccharibacteria from our database are from non-animal-associated environments. In contrast, 11 of the 16 Gracilibacteria CRISPR-Cas systems belong to genomes from non-animal-associated environments (Fig. 1A).

Despite the streamlined nature of CPR genomes, we also identified five genomes that encode multiple CRISPR-Cas systems. Remarkably, a Gracilibacteria genome (ALUM-ROCK_MS4_BD1-5_24_33_curated) encodes three distinct *cas* loci, including a CRISPR array with 80 spacers. When *cas* genes and CRISPR arrays are taken together, this specific genome dedicates 24,788 bp of its 2,138,004 bp genome (1.16%) to CRISPR-Cas defense systems.

We examined the architecture of our complete CRISPR-Cas systems and categorized the systems, based on previous classifications (32, 36), into five distinct CRISPR-Cas subtypes: type II-A, type II-C1, type III-A, type III-B, and type V-A (Fig. 1B). The distribution of the CRISPR-Cas subtypes in relation to SGA lineage is as follows: Saccharibacteria encode type II-A, type II-C1, and type III-B systems; Gracilibacteria encode type V-A and type III-A systems; Absconditabacteria encode type V-A systems. There were two exceptions to these generalizations: (i) one Saccharibacteria genome encodes a type V-A system and (ii) one Gracilibacteria genome encodes a type II-C1 system. Four of the five subtypes have been previously identified in CPR bacteria, type II-A (37), type II-C1 (34, 37), type III-A (38), and type V-A (13). To our knowledge, subtype III-B has not been

previously reported in CPR bacteria. While we expected to find primarily class 2 CRISPR-Cas systems, which are typically more compact and utilize a single effector gene (32, 36), the type-III systems (class 1) we identified utilize a multisubunit effector complex. The type III-A and type III-B systems we identified, for instance, contained nine identifiable *cas* genes together in a single operon. The targets of these subtypes are known to vary; type II and type V-A systems are thought to target double-stranded DNA, while type III-A and type III-B systems are capable of targeting both DNA and RNA (32, 36). This may indicate that some Saccharibacteria are capable of targeting both DNA and RNA phages. Interestingly, we also found that Gracilibacteria and Absconditabacteria almost exclusively rely on type V CRISPR-Cas systems despite the system's rarity among bacteria (<2% of all CRISPR-Cas systems identified in bacteria) (36). Furthermore, we compared the system architecture within each subtype based on average amino acid identity (AAI) of component proteins and noted a mostly uniform architecture within each subtype (see Fig. S1 to S3 at https://doi.org/10.5281/zenodo.8422333). Among the systems, we found 10 variants of the canonical CRISPR-Cas subtypes that contained unannotated open reading frames (ORFs) in the interior of a *cas* operon. These variants may represent novel subtypes within the broader system classification. One of these variants appears to be a type II-A system (see Fig. S1 at https://doi.org/10.5281/zenodo.8422333), and nine appear to be type V-A systems (see Fig. S3 at https://doi.org/10.5281/zenodo.8422333). The functions of these unannotated ORFs and whether they participate in concert with their respective CRISPR-Cas systems remain topics of future study.

To evaluate the novelty of the annotated genes within the complete CRISPR-Cas systems, we compared each Cas amino acid sequence to NCBI's nr database (see Fig. S4 at https://doi.org/10.5281/zenodo.8422333). While most Saccharibacteria and Absconditabacteria Cas proteins are well-represented in Genbank, there were a number of our Gracilibacteria Cas proteins with less than 50% AAI to known sequences. One such protein is a Cas9 that only displays a 34% AAI to the best match in Genbank.

Putative SGA-infecting phages

To identify candidate phages that potentially infect the SGA bacteria of our database, we extracted 1,296 non-identical spacers from our quality-controlled, high-confidence arrays from the complete genome data set (147 arrays encoded in 119 scaffolds; see Table S3 at https://doi.org/10.5281/zenodo.8422333). We also searched metagenomic reads for variant sequences that are not reflected in the consensus metagenomic assembly (see Materials and Methods). We recovered an additional 344 unique spacers from 10 SGA genomes.

We compared our set of 1,640 spacers to two large phage databases, IMG/VR (26) and GVD (39), using the thresholds of at least 95% coverage and less than two mismatches (see Tables S5 and S6 at https://doi.org/10.5281/zenodo.8422333). After de-replicating hits at 99% ANI, we identified 547 distinct phage scaffolds that putatively infect SGA bacteria (see Table S7 at https://doi.org/10.5281/zenodo.8422333). Based on spacer-matching, 440, 57, and 50 of our identified phages were predicted to infect Saccharibacteria, Gracilibacteria, and Absconditabacteria, respectively. Additionally, 26 of the 547 phage genomes were circularized (Fig. 2A; see also Table S7 at https://doi.org/10.5281/zenodo.8422333). We further identified 147 integrated prophages from the same set of de-replicated SGA genomes: 120, 15, and 12 prophages within Saccharibacteria, Absconditabacteria, and Gracilibacteria, respectively (see Table S7 at https://doi.org/10.5281/zenodo.8422333).

We characterized our candidate SGA-infecting phages, and their hosts by generating a protein-sharing network in which the proteomes of phages and SGA hosts are clustered based on similarity (Fig. 2B). The proteomes of the Saccharibacteria phages and Absconditabacteria phages predicted in this study cluster with those of their predicted host bacteria and those of phages previously identified to infect the same host bacteria, strongly supporting host inference based on our spacer targeting analyses. When clustered in a separate network with non-SGA reference phages, the predicted SGA

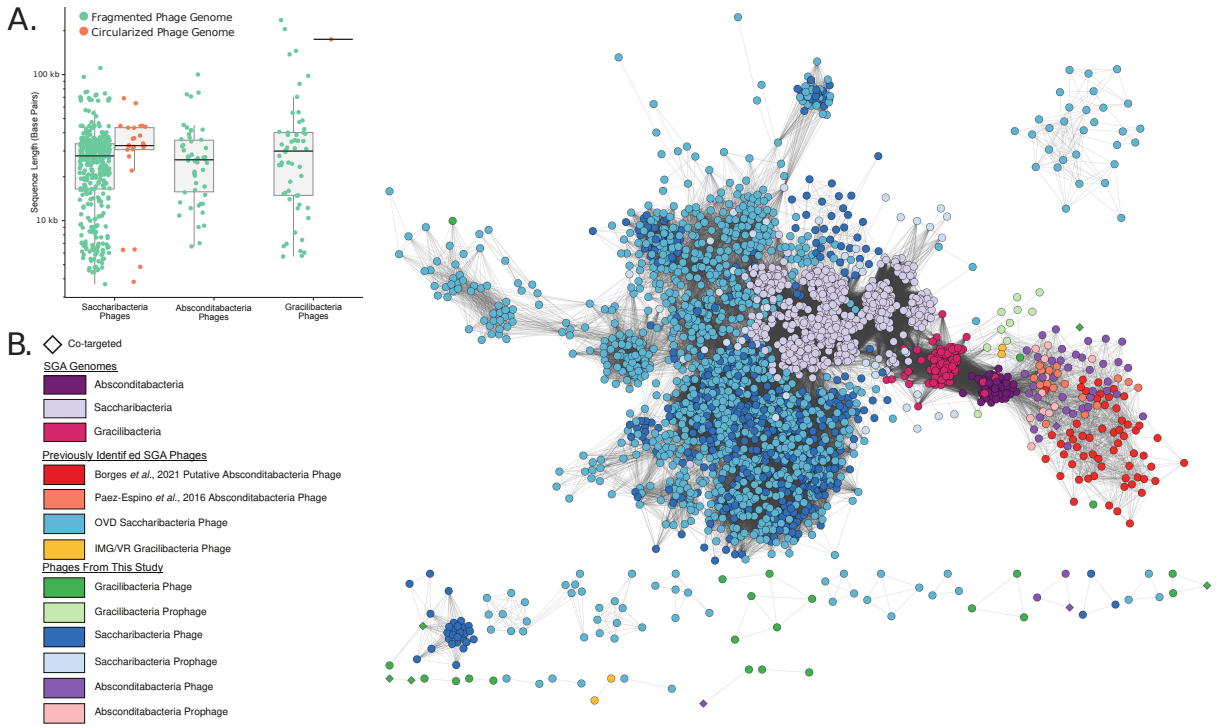

FIG 2 SGA phage genome size and protein-sharing network analysis. (A) Size and completeness of the putative SGA phages. (B) Protein-sharing network of the putative SGA phages and SGA bacterial genomes, where each node represents a phage or bacterial genome. Nodes are clustered together based on protein similarity and a number of shared proteins. Previously identified SGA phages (23, 40, 41) were included in the network. Nodes are colored based on the predicted host of the phages or the SGA genome taxonomy. Co-targeted phages indicate those targeted by spacers from CRISPR-Cas arrays of both SGA and non-SGA bacteria.

phages from this study tended to cluster apart from the non-SGA reference phages (see Fig. S5 at https://doi.org/10.5281/zenodo.8422333). This includes several phages newly predicted to infect Saccharibacteria, which form distinct clusters apart from previously identified Saccharibacteria phages or non-SGA reference phages and are thus inferred to be novel lineages. Within both networks, our putative Gracilibacteria-infecting phages did not form a singular cluster. Within the network between SGA bacteria and their predicted phages, three putative Gracilibacteria phages predicted by spacer-matching cluster with Gracilibacteria prophages or Absconditabacteria phages (Fig. 2B). Within the protein-sharing network containing non-SGA reference phages, a number of putative Gracilibacteria phages place within a sparse network that includes reference phages predicted to infect bacteria from either the Bacteroidota or the Firmicutes phylum (see Fig. S5 at https://doi.org/10.5281/zenodo.8422333).

Diverse coding strategies among predicted Absconditabacteria phages and Gracilibacteria phages

To investigate the genetic codes of our candidate phages, we predicted ORFs for each phage genome larger than 20 kb in both the alternative code 25 (the genetic code of Gracilibacteria and Absconditabacteria) and the standard code 11. Using these predictions, we calculated the coding density (a portion of the genome dedicated to protein-coding genes) in each genetic code. Differences in coding densities between code 25 and code 11 were negligible for putative Saccharibacteria phages, indicating that they share genetic code 11 with their predicted hosts (Fig. 3A; see also Table S7 at https://doi.org/10.5281/zenodo.8422333). Contrary to our expectations, 34 of the 38 putative Gracilibacteria-infecting phages larger than 20 kb displayed small changes in coding density between the two genetic codes, indicating that they are

not clearly alternatively coded (Fig. 3B; see also Table S7 at https://doi.org/10.5281/zenodo.8422333). Most putative Absconditabacteria-infecting phages displayed a much higher coding density in code 25 compared to code 11, indicating that they mainly share their predicted host's alternative genetic code (Fig. 3C; see also Table S7 at https://doi.org/10.5281/zenodo.8422333). However, 6 of the 50 predicted Absconditabacteria phages are not clearly alternatively coded (less than a 10% change between code 25 and code 11 coding densities). Notably, Gracilibacteria prophages and Absconditabacteria prophages displayed a much higher coding density in code 25, indicating they preferentially adhere to the alternative genetic code 25 (see Table S9 at https://doi.org/10.5281/zenodo.8422333).

To further assess the genetic code of the putative Gracilibacteria phages and Absconditabacteria phages, we annotated and visualized the predicted ORFs of each phage in both code 11 and code 25. Examination of 40 putative Gracilibacteria and Absconditabacteria phage genomes that were not clearly alternatively coded showed that they had very similar gene annotations and genome architectures in both genetic codes (Fig. 4A, see also Table S8 at https://doi.org/10.5281/zenodo.8422333). Furthermore, their genes displayed an absence of in-frame TGA codons and the presence of multiple, different stop codons in close proximity at gene termini (Fig. 4A; see also Table S8 at https://doi.org/10.5281/zenodo.8422333). These phage genomes are therefore likely compatible with both code 11 and code 25. They contrast with the clearly alternatively coded Gracilibacteria phages and Absconditabacteria phages, which contained high densities of in-frame TGA codons and displayed almost no gene annotations in code 11 (Fig. 4B).

Notably, all Gracilibacteria prophages and Absconditabacteria prophages contained ORFs with high densities of in-frame TGA codons. The classification of these genome regions as prophage rather than novel portions of bacteria genomes is supported by the identification of canonical phage genes that produce a portal protein, tail-related protein, terminase, or integrase (Fig. 4C; see also Table S10 at https://doi.org/10.5281/

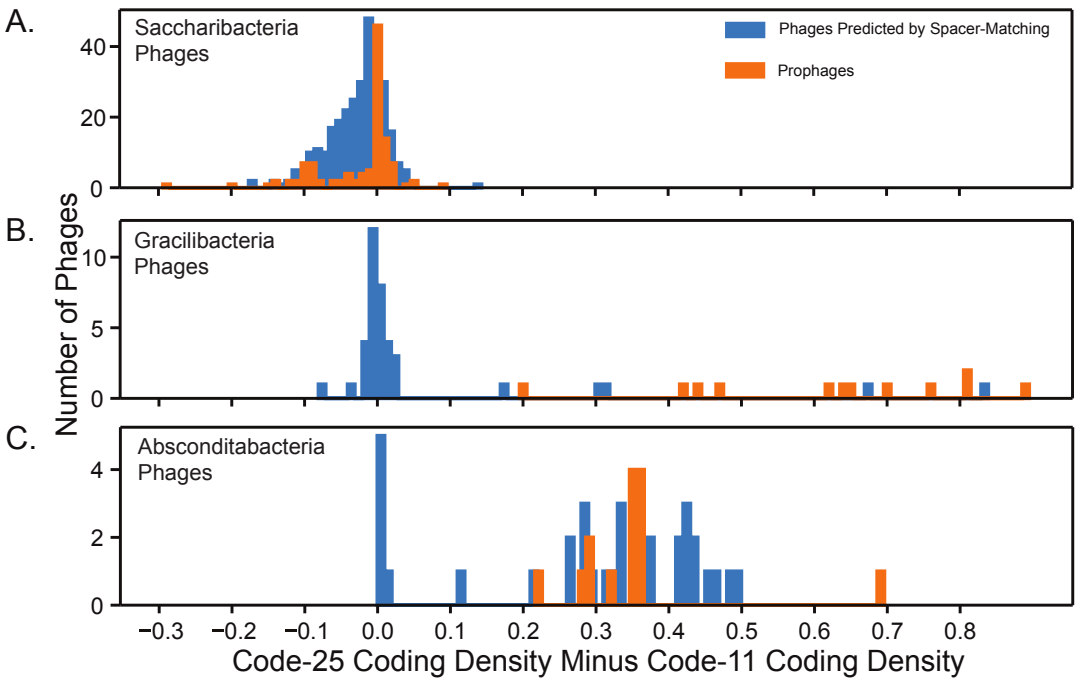

FIG 3 Phage genetic code analysis. Histogram of phages displaying the change in coding density between code-25 and code-11 predictions for predicted phages of (A) Saccharibacteria, (B) Gracilibacteria, and (C) Absconditabacteria. A larger *x*-value indicates a higher likelihood of adhering to genetic code 25, while an *x*-value near zero indicates the likely usage of genetic code 11. Only phages larger than 20 kb were included in this analysis.

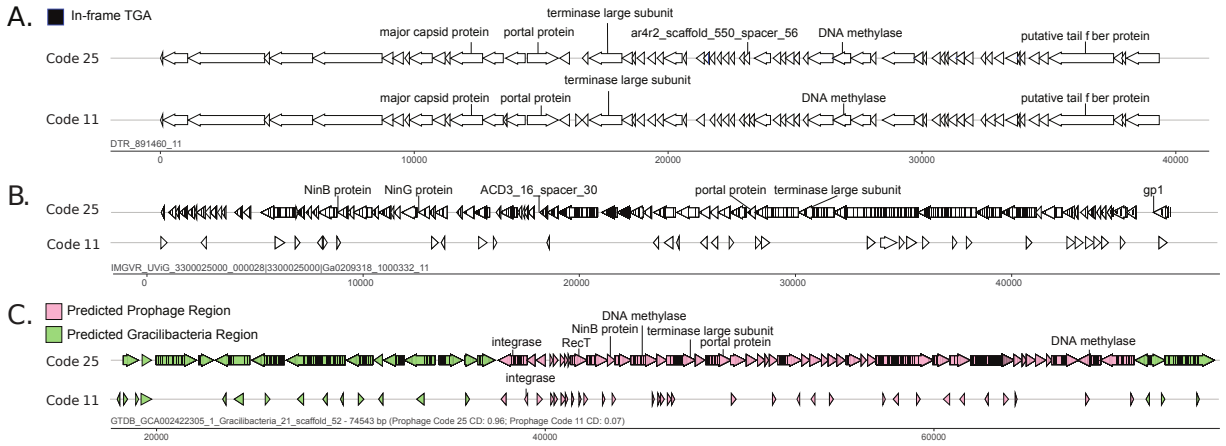

FIG 4 In-frame TGA codon usage among putative Gracilibacteria phages and prophages. (A) Genome diagrams in code 25 and code 11 of a code 11-compatible phage predicted to infect Gracilibacteria. In-frame TGA codons are marked by a black line. (B) Genome diagrams in code 25 and code 11 of a clearly alternatively coded (code 25) phage predicted to infect Gracilibacteria. (C) Genome diagrams in code 25 and code 11 of a Gracilibacteria prophage. Predicated prophage regions and genome regions are colored pink and green, respectively.

zenodo.8422333). We conclude that the Gracilibacteria prophages and Absconditabacteria prophages are likely exclusively compatible with code 25.

As it was surprising to find code 11-compatible phages that were targeted by the code-25 Gracilibacteria or Absconditabacteria, we sought to further verify that these phages infect their presumed alternatively coded hosts. Three putative code 11-compatible Gracilibacteria phages and Absconditabacteria phages in the protein-sharing network (Fig. 2B) cluster with clearly alternatively coded Gracilibacteria prophages or Absconditabacteria phages (see Fig. S7 at https://doi.org/10.5281/zenodo.8422333). Additionally, we predicted the taxonomic affiliation of each gene within our identified phages. Most putative Absconditabacteria phages contained genes with taxonomic affiliations matching their host (see Fig. S6; Table S10 at https://doi.org/10.5281/zenodo.8422333), including one possible Absconditabacteria phage compatible with code 11. Five of the 57 putative Gracilibacteria phages contained genes predicted to originate from Gracilibacteria, including one predicted Gracilibacteria phage that was compatible with code 11 (see Fig. S6; Table S10 at https://doi.org/10.5281/zenodo.8422333). These two analyses, in tandem with the spacer-phage matching, strongly suggest that there are, indeed, some Gracilibacteria phages and Absconditabacteria phages that are compatible with code 11.

As many of these code 11-compatible phages did not cluster coherently within the protein-sharing networks (Fig. 2B; see also Fig. S5 and S7 at https://doi.org/10.5281/zenodo.8422333) and did not contain any genes predicted to originate from their predicted host bacteria, we considered the possibility that some of the Gracilibacteria and Absconditabacteria spacer-to-phage hits might be artifactual matches within the large phage databases. Such spurious matches would be most probable if the spacer length is unusually short. Thus, to constrain this probability, we examined the median spacer length that matched with predicted code 11-compatible Gracilibacteria phages and Absconditabacteria phages. We found that these spacers, at 26 bp, were only slightly smaller than those that matched Saccharibacteria phages (30 bp) and those that matched clearly alternatively coded Absconditabacteria phages (28 bp), for which predicted phages generally clustered as expected (see Tables S5 and S6 at https://doi.org/10.5281/zenodo.8422333). Additionally, when compared to spacers extracted from across the domain Bacteria, a spacer length of 26 bp is within the range of a typical spacer length (29) (see Fig. S8 at https://doi.org/10.5281/zenodo.8422333).

Phages that interact with SGA and non-SGA bacteria

We next explored the host range of our putative SGA-infecting phages by comp'aring them to a large spacer database from a wide diversity of bacterial genomes (see Materials and Methods). As members of Actinobacteria are known hosts of Saccharibacteria, we augmented this database with spacers from diverse Actinobacteria genomes (see Materials and Methods). In comparing these spacers to our predicted SGA phages, we identified 23 probable SGA phages also targeted by spacers from non-SGA bacteria (see Tables S11 and S12 at https://doi.org/10.5281/zenodo.8422333). We considered that these spacers may have been acquired in either the SGA bacteria or in the non-SGA bacteria by horizontal transfer. In comparing the spacer inventories of the co-targeting bacteria, we did not find evidence that they shared identical spacers, likely ruling out the possibility that these spacer matches can be attributed to horizontal transfer.

These 23 co-targeted phages (i.e., phages targeted by both SGA and non-SGA bacteria) included nine predicted to infect Saccharibacteria, five predicted to infect Absconditabacteria, and nine predicted to infect Gracilibacteria. Seven of the nine putative Saccharibacteria phages were co-targeted by bacteria from the phylum Actinobacteria, including *Corynebacterium sp. NML130628*, *Actinomyces oris*, *Actinomyces. sp. HMSC075C01*, *Actinomyces naeslundii*, and *Actinomyces viscosus*. These species are particularly notable as a majority of cultured Saccharibacteria attach to host bacteria from the *Actinomyces* genus (5, 42). When placed in the context of our two protein-sharing networks, many putative Saccharibacteria phages co-targeted by Actinobacteria are situated within a dense cluster of previously identified Saccharibacteria infecting phages (representative example in Fig. 5A and 2B; and see also Fig. S5 at https://doi.org/10.5281/zenodo.8422333).

Similarly, we also examined putative Absconditabacteria phages and Gracilibacteria phages that matched spacers from non-SGA bacteria. Three of the five putative Absconditabacteria phages matched spacers from arrays in the genomes of Firmicutes. One such phage is situated in the primary Absconditabacteria cluster within the protein-sharing network (Fig. 5B). Among the nine putative Gracilibacteria phages, five

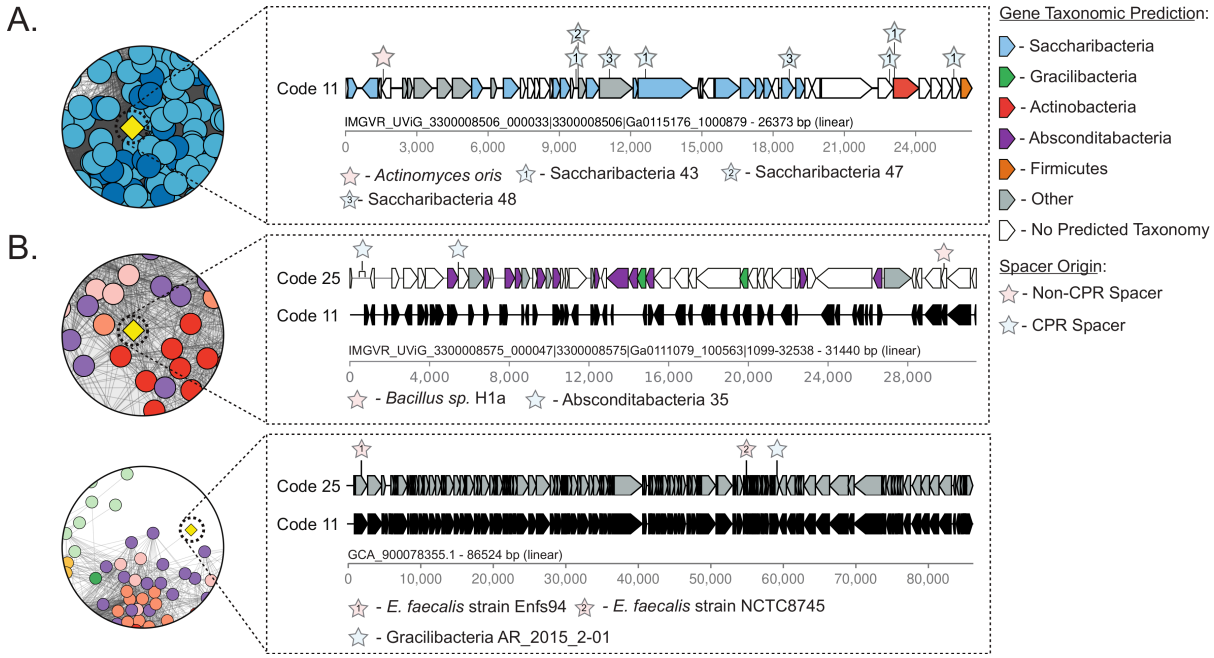

FIG 5 Representative phages targeted by both SGA and non-SGA bacteria. The leftmost circular windows display a cutout of the protein-sharing network in Fig. 2, with colors listed in the Fig. 2 legend and the co-targeted phage spotlighted in yellow. The rightmost panels display a genome diagram of the highlighted co-targeted phage. Overlaid are gene taxonomic predictions and the location of spacer matches. (A) Putative Saccharibacteria phage co-targeted by *A. oris*. (B) Putative Absconditabacteria phage co-targeted by *Bacillus sp*. H1a. (C) Putative Gracilibacteria phage co-targeted by *Enterococcus faecalis*.

have matches to spacers from arrays within Bacteroidetes species and two matched spacers from arrays within Actinobacteria species. Notably, one predicted Gracilibacteria phage was targeted by multiple spacers from *Enterococcus faecalis* strains and was linked to the Absconditabacteria phage cluster in the protein-sharing network (Fig. 5C).

By examining the genetic code of the nine candidate Gracilibacteria phages that matched spacers from non-SGA bacteria, all nine had similar genome architectures and gene annotations in both code 11 and code 25 (representative example in Fig. 5C; see also Table S8 at https://doi.org/10.5281/zenodo.8422333). Thus, if they indeed infect Gracilibacteria and another standard-coded bacterium, they are likely capable of producing viable gene products in both their alternatively coded Gracilibacteria host and their standard-coded non-SGA host. Two of the five putative Absconditabacteria phages that matched spacers from non-SGA bacteria, however, contained ORFs dense with in-frame TGA codons and clearly use code 25 (representative example in Fig. 5B).

DISCUSSION

Here, we examined the genetic code of predicted SGA phages and observed that some share the genetic code of their putative hosts. This analysis required us to link phages to host bacteria, which we did primarily via CRISPR-Cas spacer targeting. This has been done many times previously (26–29) and is believed to be generally robust given that the spacers in a CRISPR locus of a host bacterial genome derive directly from the genomes of phages that infect them (28, 43, 44). A recent study, however, proposed several alternative methods by which a spacer targeting a phage may be acquired in non-viable bacterial host cells: (i) by uptake or entry of phage DNA into physically proximal bacterial cells; (ii) horizontal gene transfer of spacer arrays into non-viable bacterial cells, and (iii) the host-range of the phage is altered after spacer acquisition (45).

To further support our CRISPR spacer-based links, we performed a number of additional analyses. We observed highly similar phage and putative bacterial host genes. Phages are well known to acquire genes from their hosts, so the most likely explanation is that these phage genes are derived directly from the genome of their host bacterium (46–48). Furthermore, in multiple protein-sharing networks, we observed strong clustering between bacteria and many of their predicted phages identified by spacer-matching. While it is possible that these linkages may be the result of horizontal gene transfer, we observed that many of our predicted SGA phages cluster exclusively with other identified SGA phages when placed in a protein-sharing network with reference phages. Finally, given that only a tiny subset of microbial community members use an alternative genetic code (17, 19), our linkage of alternatively coded phages to alternatively coded host bacterial groups using spacer-phage matching, as has been shown previously (20, 23), suggests that spacer-phage matches are very likely not coincidental. Thus, a variety of methods reinforce our confidence in the spacer-matching approach to identify hosts of phages. It is important to note, however, that our host-prediction methods do not fully indicate the capacity for identified phages to replicate in predicted host cells. To confirm the predicted host range of phages identified in this study, isolation, and experimental validation are still necessary.

In addition to identifying alternatively coded phages targeted by alternatively coded Gracilibacteria and Absconditabacteria, we were surprised to identify phages without in-frame TGA usage targeted by either Gracilibacteria or Absconditabacteria. These phages appear compatible with the standard code 11. The phenomenon wherein phages utilize a genetic code that is different than that of their host bacteria is not without precedent (20, 21, 23). For example, some Lak phages, despite infecting standard-coded bacteria of the genus *Prevotella*, have alternatively coded genomes in which the canonical stop codon TAG is reassigned to glutamine (21). Alternatively coded phages that infect standard coded hosts may use an alternative code in part to prevent premature production of proteins that are important in late-stage infection (e.g., in-frame TGA codons within ORFs annotated as structural or lysis-related proteins) (23). To enable the translation code shift needed to produce these proteins, phage

genomes often encode a suppressor tRNA that recognizes a canonical stop codon as a sense codon and incorporates a specific amino acid (20–23, 49, 50). Some alternatively coded phage genomes also encode tRNA synthetases that can charge suppressor tRNAs with amino acids (23) and release factors that terminate translation at only two of the three canonical stop codons (20, 21, 23). For example, some alternative code 4 phage genomes, in which UGA is interpreted as tryptophan, encode release factor 1 (RF1), which only recognizes UAA and UAG, a suppressor tRNA that decodes the UGA stop codon as tryptophan, and a tryptophanyl tRNA-synthetase which charges the suppressor tRNA with tryptophan (23). In combination, these previous observations underline the conclusion that phage genomes that use fewer stop codons than their host genomes require specific adaptations in the form of code shift machinery.

The situation with code 11-compatible phages, such as the predicted Gracilibacteria phages and Absconditabacteria phages we identify in this study, is different because their lack of in-frame canonical stop codons presents no issue for translation. Where TGA is used as a stop codon, it is followed by alternative stop codons in close proximity to terminate translation. In general, this backup stop codon strategy is not uncommon in bacterial genomes and likely evolved to reduce the impact of accidental stop codon read-through (51, 52). Thus, phages that employ three stop codons should generally produce viable gene products even if the bacterial translation system only recognizes two stop codons. Unlike alternatively coded phages that infect standard-coded host bacteria, phages that use the standard genetic code generally do not need to alter the translation environment of their hosts.

An intriguing finding of this study is that all identified integrated phage sequences (prophages) in Gracilibacteria and Absconditabacteria genomes were clearly alternatively coded (contained ORFs dense with in-frame TGA codons). This observation suggests that there is an advantage for a prophage to share the alternative genetic code of its host. This contrasts with the finding that some prophages adopt an alternative genetic code yet reside in bacterial genomes that use the standard bacterial code (23). One potential explanation may be that, akin to codon optimization, higher levels of the alternative code tRNAs are expressed within alternatively coded host bacteria compared to canonical tRNAs, allowing phages with dense in-frame TGAs a more efficient translation of their gene products. The codon optimization hypothesis is supported by the high usage of TGA as a glycine codon within code 25 host bacteria (17), that the use of rare codons can lead to various translation errors (53), and that competition over rare tRNAs can incur lower expression of gene products (54, 55).

If code-11 compatible phages can, in fact, proliferate in Gracilibacteria and Absconditabacteria, there are two possible explanations for why the predicted code 11-compatible Gracilibacteria phages and Absconditabacteria phages do not need to incorporate in-frame TGA codons. First, these standard-coded phages may have recently evolved to infect alternatively coded hosts. However, if this were true, we would expect such phages to be rare. As standard code compatibility is apparently not uncommon among Gracilibacteria and Absconditabacteria phages (Fig. 3 and 4), we infer that there is an advantage for these phages to retain their standard code. Second, use of the standard code may broaden their host range, a possibility that is supported by our finding that some standard code compatible Gracilibacteria phages and Absconditabacteria phages are targeted by spacers encoded within standard code non-SGA bacteria. Phages capable of replicating in hosts across phyla have been reported in a previous experimental study by Malki et al. (56), but have not been fully characterized or definitively confirmed.

Some predicted Saccharibacteria phages are also targeted by Actinobacteria spacers. For cases where Actinobacteria are hosts for episymbiotic Saccharibacteria, these phages may infect both partners. CPR bacteria thus may serve as a decoy to protect their larger bacterial symbionts from phage infection, as has been suggested previously (31, 57).

The phages reported here expand known phage diversity. Our results suggest that some of them may infect both standard and alternatively coded host bacteria, and we

deduce that there is no fundamental barrier to this phenomenon. Given interest in the use of phages as therapeutics, this finding raises the possibility of producing phages to infect SGA bacteria in standard code bacteria, which may be substantially easier to cultivate than SGA bacteria themselves. Furthermore, this may provide a path by which SGA phages can be generated for morphological characterization.

MATERIALS AND METHODS

Absconditabacteria, Saccharibacteria, and Gracilibacteria database preparation

We began with a database of 861 CPR genomes derived from a previous publication (30) that contained bacteria from three different lineages: Absconditabacteria, Saccharibacteria, and Gracilibacteria. We de-replicated the database using dRep (58) at 99% ANI clustering and default alignment fraction (10%). For each genome, we predicted protein sequences using the "single" mode of Prodigal (59). For Saccharibacteria genomes, genes were predicted in genetic code 11. As Gracilibacteria and Absconditabacteria adhere to a non-standard genetic code, code 25 (17, 18), Gracilibacteria and Absconditabacteria genes were predicted in genetic code 25. Gene taxonomic predictions were performed using USEARCH (60) with the UniRef100 (61) database.

A phylogenetic tree of the nonredundant genomes was constructed, as previously described (30), using a concatenated set of 16 syntenic ribosomal proteins. Briefly, sequences were individually aligned using MAFFT (62), trimmed using BMGE (63), and concatenated. A maximum-likelihood tree was then inferred for the concatenated alignments using IQ-tree (64) (ultrafast bootstrap, -bb 1000, -m MFP) and visualized with iTOL (65).

CRISPR-Cas array prediction and curation

To search for CRISPR arrays in the SGA genome database, we ran CRISPRCas Finder (66) (CCF) on all genomes. We then selected scaffolds containing CRISPR arrays designated as evidence level 3 or 4—arrays deemed highly likely candidates by CCF—for further curation.

We manually curated the scaffolds containing high evidence-level CRISPR arrays to ensure they did not originate from misbinning. Our manual curation considered three complementary metrics: we considered a scaffold to be from SGA bacteria if (i) the majority of predicted proteins appeared to have the closest taxonomic hits to SGA bacteria, (ii) if individual, phylogenetically informative proteins appeared to have the closet taxonomic hits to SGA bacteria, and (iii) if scaffolds displayed high coding density in code 25 relative to code 11.

Identification of complete CRISPR-Cas systems

To identify complete CRISPR-Cas systems present in our database, among the genomes containing high-confidence, manually curated CRISPR arrays, we searched for Cas proteins using the full suite of TIGRFAM HMMs (67) (hmmsearch, model-specific noise cutoff). We additionally manually curated all scaffolds containing *cas* gene annotations to ensure they were from SGA bacteria using the metrics described above. We defined complete CRISPR-Cas systems based on previously published descriptions of various CRISPR-Cas systems (32, 36). For each array, (i) if *cas9*, *csn2*, *cas1*, and *cas2* genes were also encoded within the same genome, we categorized it as a type II-A system; (ii) if *cas9*, *cas1*, *and cas2* genes and no *csn2* genes were encoded within the same genome, we categorized it as a type II-C1 system; (iii) if *cpf1*, *cas1*, *cas4*, and *cas2* genes were encoded in the same genome, we categorized it as a type V-A system; (iv) if *cas10*, *cas7*, *cas5*, and *csm2* genes were encoded in the same genome, we considered it a complete type III-A system; and (v) if *cas10*, *cas7*, *cas5*, and *cmr5* genes were encoded in the same genome, we considered it a complete type III-B system. Complete CRISPR-Cas systems

were visualized using gggenes (https://github.com/wilkox/gggenes). For CRISPR-Cas systems containing all Cas proteins on the same scaffold, AAI similarities and *cas* operon architectures were visualized using Clinker (68) at default parameters.

To assess the novelty of identified Cas proteins, we compared the Cas proteins within each complete CRISPR-Cas system to the NCBI nr database using BLASTp (evalue ≥ 1e−3, coverage ≥ 0.75) and retained the best hit per gene product based on percent identity.

Compiling a Saccharibacteria, Absconditabacteria, and Gracilibacteria spacer database

To compile a spacer database, we extracted all spacers from high evidence-level arrays on scaffolds from our redundant SGA database (861 genomes) that passed our manual curation step. In addition, we ran a previously described spacer array expansion step to gather additional spacers from variant sequences that are not reflected in the consensus metagenomic assembly (38). Briefly, if available, we gathered the metagenomic reads originally used to assemble each genome. We then reassembled the reads using MEGAHIT at default parameters (69), mapped reads back to assembled contigs using Bowtie2 at default parameters (70), and predicted proteins using the "meta" flag of Prodigal. We compared the newly assembled scaffolds to the original, publicly available scaffolds. If a newly assembled scaffold matched an original, manually curated scaffold above the thresholds of 95% coverage and 90% ANI, we predicted CRISPR arrays in the newly assembled scaffold using CCF, extracted spacers from high-evidence level arrays, and added extracted spacers to our spacer database. We de-replicated the spacer database at 100% ANI using USEARCH.

Identification of putative SGA phages

To search for phages putatively infecting SGA bacteria, we compared each unique spacer in our spacer database to two phage databases: IMG/VRv3 (26) and GVD Human Gut Virome (39). BLASTn parameters were set to at least 95% coverage of the spacer and one or less allowed mismatch with the specific flags: -task "blastn-short" -word_size 7 -gapopen 10 -gapextend 2 -penalty −1. Prophages within the CPR genomes were predicted using VIBRANT at default parameters (71).

Absconditabacteria, Saccharibacteria, and Gracilibacteria phage characterization

To de-replicate the putative SGA-infecting phages, we ran dRep at 99% ANI clustering and default alignment fraction (10%). We additionally predicted phage genome circularization using VIBRANT.

To identify the likely genetic code of putative SGA-infecting phages larger than 20 kb, we used the Prodigal "single" flag to calculate the coding density of each phage in genetic codes 11 and 25. Furthermore, genome diagrams of Gracilibacteria and Absconditabacteria prophages and phages greater than 20 kb were generated using the Prodigal ORF predictions in code 11 and code 25. In-frame TGA codons were additionally located within the ORF predictions. The genome diagrams of the phages in both code 11 and code 25 were visualized using gggenes.

To annotate phage proteins, we used the Prodigal gene predictions in genetic code 11 for putative Saccharibacteria phages and the Prodigal gene predictions in genetic code 25 for putative Gracilibacteria phages and Absconditabacteria phages. We annotated predicted proteins using pVOG (72) HMM profiles with hmmsearch from HMMER3. Gene taxonomic predictions were performed using DIAMOND (73) with the UniRef100 database.

To compare the putative SGA-infecting phages to reference phages and their predicted host bacteria, we constructed two protein-sharing networks using vContact2 (74) (--rel-mode Diamond, –vcs-mode ClusterONE, and --pcs-mode MCL). One network linked the proteomes of the putative SGA phages (including prophages) identified in this

study, previously identified SGA phages (23, 26, 40, 41), and SGA bacteria. The second network linked the SGA phages identified in this study, the previously identified SGA phages, and non-SGA reference phages (--db "ProkaryoticViralRefSeq201-Merged"). The resulting protein-sharing networks and their associated metadata were visualized in Cytoscape (75).

Host range of putative SGA-infecting phages

To examine the host range of putative SGA-infecting phages, we compared spacers from four comprehensive databases (41, 66, 76, 77) composed of spacers from across the domain Bacteria to the predicted SGA-infecting phages. As before, the BLASTn parameters were set to at least 95% coverage of the spacer and one or less allowed mismatch with the specific flags: -task "blastn-short" -word_size 7 -gapopen 10 -gapextend 2 -penalty −1.

We additionally constructed a database for Actinobacteria, some of which are known hosts of Saccharibacteria, by sampling one genome per species-level group from GTDB (release 95, August 2020). Using a similar workflow as with the SGA database, we searched these genomes for high evidence-level arrays with CCF, extracted spacers, and compared them to the putative CPR-infecting phages with the above parameters. Visualization of spacer hits was performed using DNA Features Viewer (78).

To assess if the co-targeting of phages by SGA bacteria and non-SGA bacteria occurred due to the horizontal transfer of CRISPR spacers, we compared the spacer inventories of SGA scaffolds to the comprehensive non-SGA spacer database. For this comparison, we used BLASTn at the parameters: 100% identity and 100% coverage.

ACKNOWLEDGMENTS

We thank Rohan Sachdeva for bioinformatic support, Adair Borges for thoughtful insights regarding alternatively coded phages, and Luis Valentin-Alvarado for support in generating genomes from the Alum Rock field site. We thank Christopher Brown for providing us access to the CRISPRbank spacer sequences and Kristopher Kieft for help with VIBRANT.

This research was funded in part by the Rausser College of Natural Resources Sponsored Projects for Undergraduate Research and the Regents' and Chancellor's Research Fellowship (J.L.). Funding was also provided by grants from the National Institutes of Health and the Moore Foundation to J.F.B. We were additionally supported by the National Institute of Dental and Craniofacial Research under awards 1R01DE023810 and 1R01DE031274 (B.B.).

AUTHOR AFFILIATIONS

[1]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA
[2]Department of Microbiology, Forsyth Institute, Cambridge, Massachusetts, USA
[3]Department of Earth System Science, Stanford University, Stanford, California, USA
[4]Innovative Genomics Institute, University of California, Berkeley, California, USA
[5]Department of Earth and Planetary Science, University of California, Berkeley, California, USA
[6]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, Massachusetts, USA
[7]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA

AUTHOR ORCIDs

Jett Liu http://orcid.org/0000-0001-7048-4248
Alexander L. Jaffe http://orcid.org/0000-0002-6903-1687

LinXing Chen http://orcid.org/0000-0003-2774-1952
Batbileg Bor http://orcid.org/0000-0002-1797-1730
Jillian F. Banfield http://orcid.org/0000-0001-8203-8771

FUNDING

Funder	Grant(s)	Author(s)
Gordon and Betty Moore Foundation (GBMF)	71785	Alexander L. Jaffe LinXing Chen Jillian F. Banfield
HHS | National Institutes of Health (NIH)	62602	Alexander L. Jaffe LinXing Chen Jillian F. Banfield
HHS | NIH | National Institute of Dental and Craniofacial Research (NIDCR)	1R01DE031274, 1R01DE023810	Batbileg Bor

AUTHOR CONTRIBUTIONS

Jett Liu, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing | Alexander L. Jaffe, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | LinXing Chen, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – review and editing | Batbileg Bor, Investigation, Project administration, Supervision, Validation, Visualization, Writing – review and editing | Jillian F. Banfield, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

Supplemental figures and tables are available on Zenodo (https://doi.org/10.5281/zenodo.8422333). Non-redundant SGA genome accessions and associated metadata are listed in Table S1. Redundant SGA genomes are additionally available on Zenodo. SGA phage accessions are listed in Tables S5 and S6, including IMG/VR UViGs and GVD scaffold names. All IMG/VR phages used in this study are publicly accessible without use restriction. All code used in this project is available on GitHub (https://github.com/jett-liu/SGA_Phages).

REFERENCES

1. Naud S, Ibrahim A, Valles C, Maatouk M, Bittar F, Tidjani Alou M, Raoult D. 2022. Candidate phyla radiation, an underappreciated division of the human microbiome, and its impact on health and disease. Clin Microbiol Rev 35:e0014021. https://doi.org/10.1128/cmr.00140-21
2. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. 2018. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol 16:629–645. https://doi.org/10.1038/s41579-018-0076-2
3. Castelle CJ, Banfield JF. 2018. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell 172:1181–1197. https://doi.org/10.1016/j.cell.2018.02.016
4. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, Dorrestein PC, Esquenazi E, Hunter RC, Cheng G, Nelson KE, Lux R, Shi W. 2015. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci U S A 112:244–249. https://doi.org/10.1073/pnas.1419038112
5. Utter DR, He X, Cavanaugh CM, McLean JS, Bor B. 2020. The Saccharibacterium TM7x elicits differential responses across its host range. ISME J 14:3054–3067. https://doi.org/10.1038/s41396-020-00736-6
6. Bor B, Collins AJ, Murugkar PP, Balasubramanian S, To TT, Hendrickson EL, Bedree JK, Bidlack FB, Johnston CD, Shi W, McLean JS, He X, Dewhirst FE. 2020. Insights obtained by culturing Saccharibacteria with their bacterial hosts. J Dent Res 99:685–694. https://doi.org/10.1177/0022034520905792
7. Xie B, Wang J, Nie Y, Chen D, Hu B, Wu X, Du W. 2021. Epicpcr-directed cultivation of a *Candidatus* Saccharibacteria symbiont reveals a type IV pili-dependent epibiotic lifestyle. bioRxiv. https://doi.org/10.1101/2021.07.08.451036
8. Batinovic S, Rose JJA, Ratcliffe J, Seviour RJ, Petrovski S. 2021. Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams. Nat Microbiol 6:703–711. https://doi.org/10.1038/s41564-021-00892-1

9. Bor B, McLean JS, Foster KR, Cen L, To TT, Serrato-Guillen A, Dewhirst FE, Shi W, He X. 2018. Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. Proc Natl Acad Sci U S A 115:12277–12282. https://doi.org/10.1073/pnas.1810625115
10. Chipashvili O, Utter DR, Bedree JK, Ma Y, Schulte F, Mascarin G, Alayyoubi Y, Chouhan D, Hardt M, Bidlack F, Hasturk H, He X, McLean JS, Bor B. 2021. Episymbiotic Saccharibacteria suppresses gingival inflammation and bone loss in mice through host bacterial modulation. Cell Host Microbe 29:1649–1662. https://doi.org/10.1016/j.chom.2021.09.009
11. Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson JA, Strausbaugh LD, Gamonal J, Diaz PI. 2013. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. ISME J 7:1016–1025. https://doi.org/10.1038/ismej.2012.174
12. Kumar PS, Griffen AL, Barton JA, Paster BJ, Moeschberger ML, Leys EJ. 2003. New bacterial species associated with chronic periodontitis. J Dent Res 82:338–344. https://doi.org/10.1177/154405910308200503
13. Moreira D, Zivanovic Y, López-Archilla AI, Iniesto M, López-García P. 2021. Reductive evolution and unique predatory mode in the CPR bacterium *Vampirococcus lugosii*. Nat Commun 12:2454. https://doi.org/10.1038/s41467-021-22762-4
14. Yakimov MM, Merkel AY, Gaisin VA, Pilhofer M, Messina E, Hallsworth JE, Klyukina AA, Tikhonova EN, Gorlenko VM. 2022. Cultivation of a vampire: '*Candidatus* Absconditicoccus praedator'. Environ Microbiol 24:30–49. https://doi.org/10.1111/1462-2920.15823
15. Atterbury RJ, Tyson J. 2021. Predatory bacteria as living antibiotics - where are we now? Microbiology (Reading) 167. https://doi.org/10.1099/mic.0.001025
16. Tyson J, Sockett RE. 2017. Predatory bacteria: moving from curiosity towards curative. Trends Microbiol 25:90–91. https://doi.org/10.1016/j.tim.2016.12.011
17. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Söll D, Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc Natl Acad Sci U S A 110:5540–5545. https://doi.org/10.1073/pnas.1303090110
18. Hanke A, Hamann E, Sharma R, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R, Makinwa K, Hettich RL, Banfield JF, Tegetmeyer HE, Strous M. 2014. Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. Front Microbiol 5:231. https://doi.org/10.3389/fmicb.2014.00231
19. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. https://doi.org/10.1038/nature12352
20. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. 2014. Stop codon reassignments in the wild. Science 344:909–913. https://doi.org/10.1126/science.1250691
21. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ, Seed KD, Blekhman R, Aarestrup FM, Thomas BC, Banfield JF. 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. Nat Microbiol 4:693–700. https://doi.org/10.1038/s41564-018-0338-9
22. Crisci MA, Chen L-X, Devoto AE, Borges AL, Bordin N, Sachdeva R, Tett A, Sharrar AM, Segata N, Debenedetti F, Bailey M, Burt R, Wood RM, Rowden LJ, Corsini PM, van Winden S, Holmes MA, Lei S, Banfield JF, Santini JM. 2021. Closely related Lak megaphages replicate in the microbiomes of diverse animals. iScience 24:102875. https://doi.org/10.1016/j.isci.2021.102875
23. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Penev PI, Jaffe AL, Lei S, Santini JM, Banfield JF. 2022. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. Nat Microbiol 7:918–927. https://doi.org/10.1038/s41564-022-01128-6
24. Hatfull GF, Dedrick RM, Schooley RT. 2022. Phage therapy for antibiotic-resistant bacterial infections. Annu Rev Med 73:197–211. https://doi.org/10.1146/annurev-med-080219-122208
25. Chan BK, Abedon ST, Loc-Carrillo C. 2013. Phage cocktails and the future of phage therapy. Future Microbiol 8:769–783. https://doi.org/10.2217/fmb.13.47
26. Roux S, Páez-Espino D, Chen I-M, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 49:D764–D775. https://doi.org/10.1093/nar/gkaa946
27. Andersson AF, Banfield JF. 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320:1047–1050. https://doi.org/10.1126/science.1157358
28. Zhang R, Mirdita M, Levy Karin E, Norroy C, Galiez C, Söding J. 2021. SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. Bioinformatics 37:3364–3366. https://doi.org/10.1093/bioinformatics/btab222
29. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, Tritt A. 2022. iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. bioRxiv. https://doi.org/10.1101/2022.07.28.501908
30. Jaffe AL, Thomas AD, He C, Keren R, Valentin-Alvarado LE, Munk P, Bouma-Gregson K, Farag IF, Amano Y, Sachdeva R, West PT, Banfield JF. 2021. Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in candidate phyla radiation bacteria. mBio 12:e0052121. https://doi.org/10.1128/mBio.00521-21
31. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, Thomas BC, Banfield JF. 2016. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat Commun 7:10613. https://doi.org/10.1038/ncomms10613
32. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, et al. 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18:67–83. https://doi.org/10.1038/s41579-019-0299-x
33. Bernheim A, Bikard D, Touchon M, Rocha EPC. 2020. Atypical organizations and epistatic interactions of CRISPRs and *cas* clusters in genomes and their mobile genetic elements. Nucleic Acids Res 48:748–760. https://doi.org/10.1093/nar/gkz1091
34. Dudek NK, Sun CL, Burstein D, Kantor RS, Aliaga Goltsman DS, Bik EM, Thomas BC, Banfield JF, Relman DA. 2017. Novel microbial diversity and functional potential in the marine mammal oral microbiome. Curr Biol 27:3752–3762. https://doi.org/10.1016/j.cub.2017.10.040
35. Shaiber A, Willis AD, Delmont TO, Roux S, Chen L-X, Schmid AC, Yousef M, Watson AR, Lolans K, Esen ÖC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren AM. 2020. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol 21:292. https://doi.org/10.1186/s13059-020-02195-w
36. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJM, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV. 2015. An updated evolutionary classification of CRISPR–Cas systems. Nat Rev Microbiol 13:722–736. https://doi.org/10.1038/nrmicro3569
37. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W, He X. 2020. Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to mammalian hosts. Cell Rep 32:107939. https://doi.org/10.1016/j.celrep.2020.107939
38. Chen L-X, Al-Shayeb B, Méheust R, Li W-J, Doudna JA, Banfield JF. 2019. Candidate phyla radiation Roizmanbacteria from hot springs have novel and unexpectedly abundant CRISPR-Cas systems. Front Microbiol 10:928. https://doi.org/10.3389/fmicb.2019.00928
39. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. 2020. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe 28:724–740. https://doi.org/10.1016/j.chom.2020.08.003
40. Li S, Guo R, Zhang Y, Li P, Chen F, Wang X, Li J, Jie Z, Lv Q, Jin H, Wang G, Yan Q. 2022. A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. iScience 25:104418. https://doi.org/10.1016/j.isci.2022.104418

41. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering earth's virome. 7617. Nature 536:425–430. https://doi.org/10.1038/nature19094
42. Nie J, Utter DR, Kerns KA, Lamont EI, Hendrickson EL, Liu J, Wu T, He X, McLean J, Bor B, Hallam SJ. 2022. Strain-level variation and diverse host bacterial responses in episymbiotic Saccharibacteria. mSystems 7:e0148821. https://doi.org/10.1128/msystems.01488-21
43. McGinn J, Marraffini LA. 2019. Molecular mechanisms of CRISPR–Cas spacer acquisition. Nat Rev Microbiol 17:7–12. https://doi.org/10.1038/s41579-018-0071-7
44. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev 40:258–272. https://doi.org/10.1093/femsre/fuv048
45. Hwang Y, Roux S, Coclet C, Krause SJE, Girguis PR. 2023. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nat Microbiol 8:946–957. https://doi.org/10.1038/s41564-023-01347-5
46. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A 108:E757–E764. https://doi.org/10.1073/pnas.1102164108
47. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86–89. https://doi.org/10.1038/nature04111
48. Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA. 2019. Discovery of several novel, widespread, and ecologically distinct marine *Thaumarchaeota* viruses that encode *amoC* nitrification genes. ISME J 13:618–631. https://doi.org/10.1038/s41396-018-0289-4
49. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y, et al. 2020. Clades of huge phages from across earth's ecosystems. Nature 578:425–431. https://doi.org/10.1038/s41586-020-2007-4
50. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, Antipov D, Pevzner PA, Koonin EV. 2021. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative crass-like phages with unique genomic features. Nat Commun 12:1044. https://doi.org/10.1038/s41467-021-21350-w
51. Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem 289:30334–30342. https://doi.org/10.1074/jbc.M114.606632
52. Vakhrusheva AA, Kazanov MD, Mironov AA, Bazykin GA. 2011. Evolution of prokaryotic genes by shift of stop codons. J Mol Evol 72:138–146. https://doi.org/10.1007/s00239-010-9408-1
53. Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol 6:494–500. https://doi.org/10.1016/0958-1669(95)80082-4
54. Ward NJ, Buckley SMK, Waddington SN, VandenDriessche T, Chuah MKL, Nathwani AC, McIntosh J, Tuddenham EGD, Kinnon C, Thrasher AJ, McVey JH. 2011. Codon optimization of human factor VIII cDNAs leads to high-level expression. Blood 117:798–807. https://doi.org/10.1182/blood-2010-05-282707
55. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255–258. https://doi.org/10.1126/science.1170160
56. Malki K, Kula A, Bruder K, Sible E, Hatzopoulos T, Steidel S, Watkins SC, Putonti C. 2015. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. Virol J 12:164. https://doi.org/10.1186/s12985-015-0395-0
57. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48
58. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126
59. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119
60. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461
61. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932. https://doi.org/10.1093/bioinformatics/btu739
62. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436
63. Criscuolo A, Gribaldo S. 2010. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210. https://doi.org/10.1186/1471-2148-10-210
64. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300
65. Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301
66. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCas-Finder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res 46:W246–W251. https://doi.org/10.1093/nar/gky425
67. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373. https://doi.org/10.1093/nar/gkg128
68. Gilchrist CLM, Chooi Y-H. 2021. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics 37:2473–2475. https://doi.org/10.1093/bioinformatics/btab007
69. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033
70. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
71. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90. https://doi.org/10.1186/s40168-020-00867-0
72. Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res 45:D491–D498. https://doi.org/10.1093/nar/gkw975
73. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x
74. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639. https://doi.org/10.1038/s41587-019-0100-8
75. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303
76. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. 2017. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. mBio 8:e01397-17. https://doi.org/10.1128/mBio.01397-17
77. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. 2016. CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC Genomics 17:356. https://doi.org/10.1186/s12864-016-2627-0
78. Zulkower V, Rosser S. 2020. DNA features viewer: a sequence annotation formatting and plotting library for python. Bioinformatics 36:4350–4352. https://doi.org/10.1093/bioinformatics/btaa213