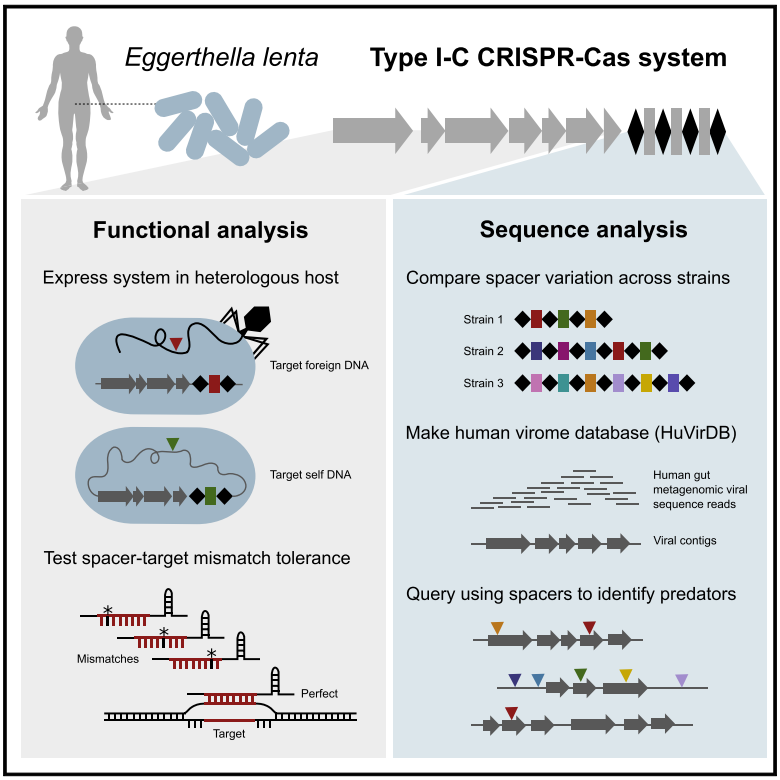
Article

Cell Host & Microbe

CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog

Graphical Abstract



Highlights

- *Eggerthella lenta*, a human gut Actinobacterium, encodes a functional CRISPR-Cas system
- Strain-level variations exist in system presence, *cas* gene sequence, and spacer content
- HuVirDB is a generalizable human virome database to search for CRISPR targets
- Hyper-targeted phages that harbor multiple protospacers were discovered

Authors

Paola Soto-Perez, Jordan E. Bisanz, Joel D. Berry, Kathy N. Lam, Joseph Bondy-Denomy, Peter J. Turnbaugh

Correspondence

joseph.bondy-denomy@ucsf.edu (J.B.-D.), peter.turnbaugh@ucsf.edu (P.J.T.)

In Brief

Soto-Perez, Bisanz, et al. focus on a type I-C CRISPR-Cas system encoded by *Eggerthella lenta*, a prevalent human gut Actinobacterium implicated in metabolism and pathogenesis. Through computational and experimental approaches, they determine the system's activity and strain-level variation while also generating a human virome database to identify common phage predators.

Soto-Perez et al., 2019, Cell Host & Microbe *26*, 1–11
September 11, 2019
https://doi.org/10.1016/j.chom.2019.08.008

CellPress

Cell Host & Microbe
Article

CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog

Paola Soto-Perez,[1,4] Jordan E. Bisanz,[1,4] Joel D. Berry,[1] Kathy N. Lam,[1] Joseph Bondy-Denomy,[1,2,*] and Peter J. Turnbaugh[1,3,5,*]
[1]Department of Microbiology & Immunology, University of California, San Francisco, San Francisco, CA 94143, USA
[2]Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA 94158, USA
[3]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: joseph.bondy-denomy@ucsf.edu (J.B.-D.), peter.turnbaugh@ucsf.edu (P.J.T.)
https://doi.org/10.1016/j.chom.2019.08.008

SUMMARY

Bacteriophages are abundant within the human gastrointestinal tract, yet their interactions with gut bacteria remain poorly understood, particularly with respect to CRISPR-Cas immunity. Here, we show that the type I-C CRISPR-Cas system in the prevalent gut Actinobacterium *Eggerthella lenta* is transcribed and sufficient for specific targeting of foreign and chromosomal DNA. Comparative analyses of *E. lenta* CRISPR-Cas systems across (meta)genomes revealed 2 distinct clades according to *cas* sequence similarity and spacer content. We assembled a human virome database (HuVirDB), encompassing 1,831 samples enriched for viral DNA, to identify protospacers. This revealed matches for a majority of spacers, a marked increase over other databases, and uncovered "hyper-targeted" phage sequences containing multiple protospacers targeted by several *E. lenta* strains. Finally, we determined the positional mismatch tolerance of observed spacer-protospacer pairs. This work emphasizes the utility of merging computational and experimental approaches for determining the function and targets of CRISPR-Cas systems.

INTRODUCTION

CRISPR-Cas are adaptive immune systems, comprised of RNA-guided nucleases, that protect prokaryotes against infection from parasitic genetic elements by cleaving foreign DNA (Barrangou and Horvath, 2017; Barrangou et al., 2007). A variety of these systems (spanning the mechanistically distinct types I–VI) have been identified in bacterial and archaeal genomes (Koonin et al., 2017) and function by storing the memory of past exposure to foreign elements as ~30-nt spacers in a clustered regularly interspaced short palindromic repeats (CRISPR) array between direct repeat sequences (Levy et al., 2015; McGinn and Marraffini, 2019). This memory element is subsequently processed, generating RNA guides (crRNA), which are packaged into complexes with CRISPR-associated (Cas) proteins (Brouns et al., 2008) to surveil the cell and mediate the recognition and cleavage of complementary sequences (Garneau et al., 2010). The outcome of these interactions is a limitation of horizontal gene transfer and prevention of phage replication (Bikard et al., 2012).

Most identified spacers cannot be assigned a target, suggesting a ubiquity of unobserved phage and mobile element diversity (Shmakov et al., 2017), especially within the human gut microbiome. Moreover, the relationship between environmental fitness in the gut and CRISPR-Cas remains to be determined, given that they defend against phages but also limit horizontal gene transfer encoding beneficial traits (Barrangou et al., 2007; Bikard et al., 2012; Palmer and Gilmore, 2010).

To date, CRISPR-Cas research in human-associated bacteria has focused on computational analyses (Tajkarimi and Wexler, 2017; Zhang et al., 2014). These studies can both over- and underestimate CRISPR-Cas prevalence (Zhang and Ye, 2017), motivating the need for experimental demonstration of CRISPR-Cas expression, array processing, and target cleavage. Here, we leverage the use of robust genetic tools in an evolutionary distant bacterium, *Pseudomonas aeruginosa*, to express *cas* genes and crRNA, utilizing a generalizable strategy for studying CRISPR-Cas in genetically intractable gut bacteria. We focus on *Eggerthella lenta* because of (1) its high prevalence in the human gut (81.6%) (Koppel et al., 2018); (2) broad impact on the metabolism of drugs (Haiser et al., 2013; Koppel et al., 2018), dietary bioactives (Bess et al., 2018), and endogenous compounds (Harris et al., 2018; Maini Rekdal et al., 2019); and (3) links to infectious (Chan and Mercer, 2008) and chronic (Qin et al., 2012) disease.

Our work highlights the presence and functionality of a prevalent CRISPR-Cas system in an understudied bacterium and host habitat. In order to identify targets of this immune system, we constructed a specialized database that allowed us to uncover putative phages repeatedly targeted by diverse *E. lenta* strains. These results serve as a strong foundation for the discovery and mechanistic dissection of phage-bacterial interactions within the gut.

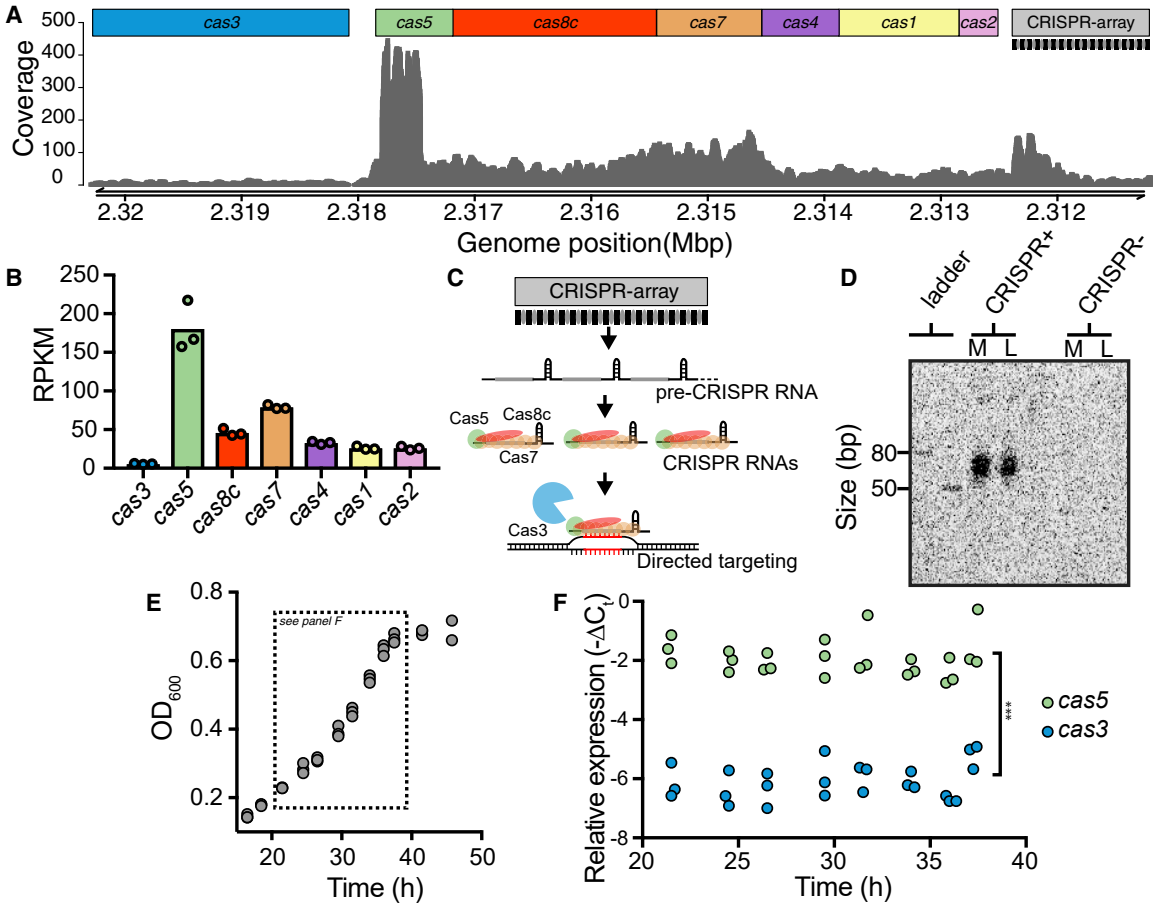


Figure 1. *E. lenta* DSM 2243 Has a Transcriptionally and Catalytically Active CRISPR-Cas System

(A) Base coverage of RNA-seq reads to the CRISPR-Cas locus in DSM 2243 and indicates active transcription.

(B) Expression levels of *cas* genes during exponential growth measured in reads per kilobase per million mapped reads (RPKM).

(C) Transcription from the CRISPR array generates a pre-CRISPR RNA that is processed by the Cas enzymes to form crRNAs that direct targeting and cleavage of foreign DNA.

(D) Northern blot demonstrates the presence of short RNA species (crRNA) in a CRISPR-positive strain (DSM 2243) but not in a CRISPR-negative strain (Valencia). Growth phase is indicated above the blot: M = mid-exponential (24 h) and L = late-exponential (37 h).

(E and F) Growth kinetics (n = 3) (E) and *cas* expression levels (F) demonstrate the stability of *cas3* and *cas5* across growth phases (n = 3, ***p < 0.001 two-way ANOVA).

RESULTS

The *E. lenta* CRISPR-Cas System Is Transcriptionally Active

Analysis of the *E. lenta* DSM 2243 genome revealed a putative CRISPR-Cas system in the type I-C subgroup (Figure 1A). Given the evidence for type I-E *cas* transcriptional repression during *in vitro* growth (Pul et al., 2010), we examined transcriptional data from *E. lenta* DSM 2243 in the mid-exponential phase and detected expression of all *cas* genes (Figures 1A and 1B). We observed heterogeneity across the locus, ranging from *cas3* (5.6 ± 0.6 RPKM ± SD) to *cas5* (181.0 ± 32.2) (Figure 1B), both higher than intragenic expression (0.0747 ± 0.006). The depth of mapped reads (Figure 1A) and predicted transcriptional start sites (Figure S1A) both suggested that this locus produces at least 2 distinct transcripts. We experimentally confirmed this by performing a nested PCR of cDNA using primer pairs that span the junction of each gene pair (Figure S1A). These results, shown in Figure S1B, are consistent with the presence of a monocistronic *cas3* transcript and at least one additional polycistronic transcript encompassing the genes from *cas5* to *cas2*.

CRISPR array transcription generates a precursor transcript (pre-crRNA) (Figure 1C) whose expression was supported by RNA sequencing (Figure 1A). Consistent with prior reports, the 5′ end of the array, where new spacers are acquired, was more highly transcribed (Rollie et al., 2015). We sought to test if the pre-crRNA is processed into the short active CRISPR RNA species (crRNA), which are essential for the formation of the interference complex that recruits the endonuclease Cas3 to cleave targets (Figure 1C). Through northern blot analysis, we detected mature crRNAs, between 50 and 80 nt, during both mid- and late-exponential growth (Figure 1D), which are generated by Cas5 (Hochstrasser et al., 2016). No bands were observed using a control *E. lenta* strain lacking a CRISPR-Cas system (Figure 1D).

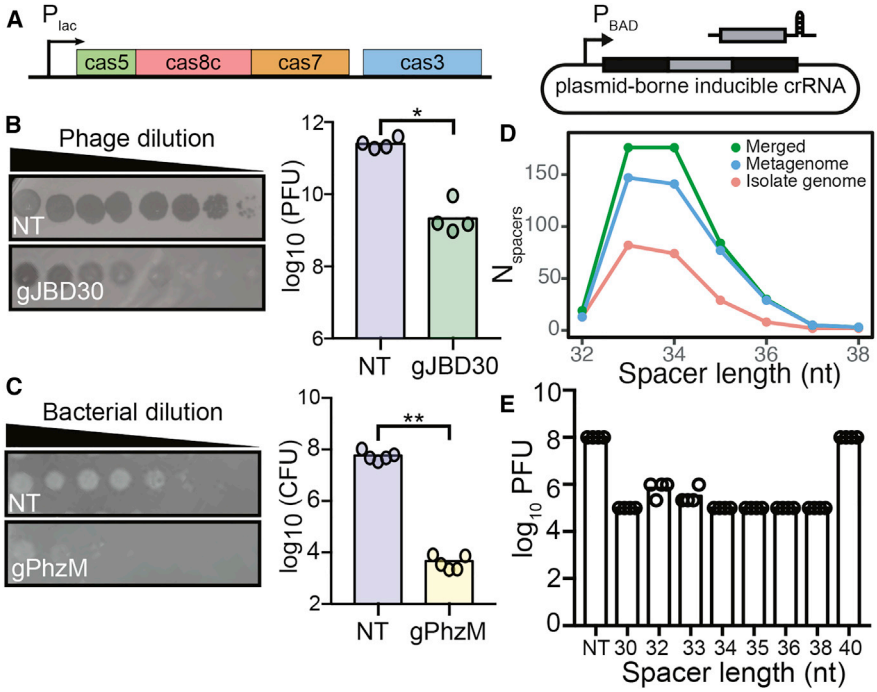


Figure 2. Heterologous Expression in *P. aeruginosa* Demonstrates the Ability to Target Phage and Chromosomal DNA

(A) *P. aeruginosa* strain (PA01 tn7::lentaIC) constructed to inducibly express the minimal *cas* genes required for interference and a plasmid containing a minimal CRISPR array.

(B) Expression of gJBD30 (phage-targeting) causes a 120-fold reduction in the number of plaque-forming units (PFUs) when compared to a NT control (n = 4, *p = 0.0286, Mann-Whitney U test).

(C) Expression of gPhzM (chromosome-targeting) decreases the colony-forming units (CFUs) by 13,450-fold (n = 5, **p = 0.0079, Mann-Whitney U test).

(D) Distribution of spacer lengths found in the *E. lenta* isolate genomes, metagenomes, and merged datasets.

(E) Variable length crRNAs decrease the number of PFUs with the exception of a 40-nt crRNA (n = 4).

Consistent with these results, the *cas* genes were also stably transcribed throughout exponential growth (Figures 1E and 1F). The relative expression level between *cas5* and *cas3* was also stable over time; *cas5* was expressed at 17.3- ± 1.2-fold higher levels than *cas3* (p_{gene} < 0.001, two-way ANOVA, Figure 1F). This transcriptional control of *cas3* has been proposed to keep low but sufficient levels of the protein in order to provide immunity while avoiding off-target nuclease activity (Majsec et al., 2016). Together, these results indicate that the type I-C CRISPR-Cas system of *E. lenta* DSM 2243 is transcriptionally active and that mature crRNAs are generated during *in vitro* growth.

The *E. lenta* CRISPR-Cas System Is Sufficient to Target Phage and Chromosomal DNA

To definitively demonstrate targeting by the *E. lenta* CRISPR-Cas system, and to circumvent the lack of genetic tools available, we designed a heterologous expression system in *P. aeruginosa* PA01, which lacks an endogenous system. The resulting strain (PA01 *tn7::lentaIC*) expresses the minimal machinery from the *E. lenta* system required for targeting and cleavage (*cas5*, *cas8c*, *cas7*, and *cas3*) (Figure 2A). To complete the interference complex, we constructed a plasmid expressing a minimal CRISPR array (Figure 2A). To target sequences of interest, we used the type I-C canonical protospacer adjacent motif (PAM), responsible for identifying non-self-DNA sequence (TTC).

We tested the system's ability to target foreign DNA by providing a 34-nt spacer targeting the phage JBD30 (gJBD30). When challenged with JBD30, there was a 120-fold reduction in plaque formation compared to a non-targeting (NT) control (p = 0.0286, Mann-Whitney U test, Figure 2B). Phage targeting was also evident from plaque morphology: individual plaques became smaller and less opaque, indicative of inhibited lytic activity (Figure 2B). Next, to determine if the system was capable of targeting self-DNA, we designed a 34-nt spacer to target the region upstream of the pyocyanin pigment biosynthetic gene (*phzM*) in the host genome. Expression of this crRNA (gPhzM) resulted in a >10,000-fold reduction in colony formation compared to the NT control (p = 0.0079, Mann-Whitney U, Figure 2C). Together, these results demonstrate that the *E. lenta* type I-C CRISPR-Cas effector complex is sufficient for the specific recognition and cleavage of foreign and self-DNA.

We designed both spacers, gJBD30 and gPhzM, to be 34 nt long; however, spacers within *E. lenta* isolates naturally vary from 32 to 38 nt with 74.2% of the spacers being 33 or 34 nt (Figure 2D). This spacer length variation has been observed in soil bacteria with a type I-C system (Lee et al., 2018). To determine the effect of spacer length on targeting efficiency, we designed multiple spacers varying from 30 to 40 nt against a single JBD30 protospacer. We observed similar plaquing efficiencies for all spacer lengths with the exception of the 40-nt spacer

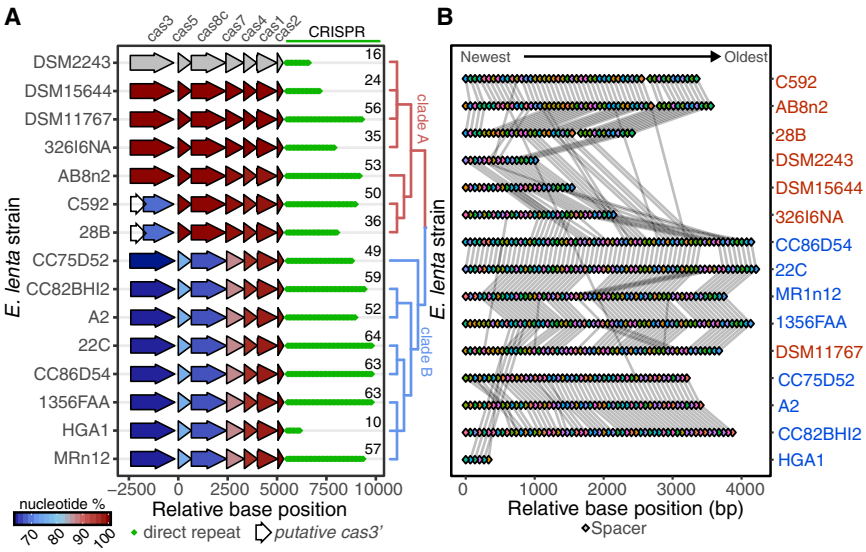


Figure 3. Strain-Level Variation in the *E. lenta* Type I-C CRISPR-Cas System

(A) 15 sequenced isolate genomes contain a type I-C system that clusters into two distinct clades based on an alignment of the *cas* genes. The global nucleotide identity to the DSM 2243 ortholog is shown. Diamonds indicate the number of direct repeats, and the exact number of spacers in each array is displayed.

(B) CRISPR spacer conservation between strains. Spacers are indicated as colored diamonds with identical spacers linked by a gray line.

(Figure 2E), demonstrating that all of the naturally occurring spacer lengths are efficient at phage targeting.

The Presence of CRISPR-Cas Systems Varies within the *E. lenta* Species

Multiple studies have emphasized strain-level variation in gut microbial metabolism (Koppel et al., 2018), immune interactions (Belkaid and Hand, 2014), and pathogenesis (Britton and Young, 2014). To assess if CRISPR-Cas presence in *E. lenta* is similarly strain specific, we expanded our analysis to include a collection of human-associated *E. lenta* strains (Bisanz et al., 2018). These genomes have a mean size of 3.53 Mb, a minimum contig length covering 50% of the genome (N_{50}) of 431,316 bp, and $N_{contigs}$ = 59. Of the 24 *E. lenta* genomes analyzed, 15 had a type I-C CRISPR-Cas system (Figure 3A) and no other complete CRISPR-Cas system types were observed. For CRISPR-Cas-encoding strains, the genomic context was conserved, and phylogenetic analysis based on *cas* alignment revealed 2 distinct clades: A and B (Figure 3A). The number of spacers per CRISPR array ranged from 10 to 64 (median 52, Figure 3A) with a total of 210 unique spacers across the 15 *E. lenta* genomes.

Strains C592 and 28B were annotated as having a 5′ truncated *cas3* (Figure 3A). We observed and confirmed a single base insertion in the 28B *cas3* sequence that caused a premature stop codon, leading to an internal ATG sequence being identified as the *cas3* translational start (Figure S2A). Prediction of functional domains revealed that the insertion separated the helicase and endonuclease domains into 2 separate coding sequences (Figures S2B and S2C). More work is necessary to determine if this leads to inactivation or if these open reading frames are still able to generate functional polypeptides carrying out their respective activities, as shown in other systems (Makarova et al., 2011; Plagens et al., 2012).

The spacers found on the 3′ end of the array were more conserved even across *cas* clades (Figure 3B). In most instances, exemplified by strains DSM 11767 and DSM 15644, unique spacers are found near the 5′ end of the array, consistent with acquisition of spacers over time. Spacers interrupting stretches of highly correlated spacer order could be due to loss via recombination or low-frequency spacer acquisition in the middle rather than the start of the array (Deveau et al., 2008).

To enrich our sampling of *E. lenta* CRISPR diversity, we leveraged metagenomic data and the nature of the CRISPR direct repeat. An alignment of the direct repeat sequences from our reference genomes revealed a highly conserved 33-nt motif (Figure 4A). This appears to be unique to *E. lenta*, as it is absent from the CRISPR-Cas systems of other members of the Coriobacteriia with the nearest homologous direct repeat observed in *Bifidobacterium thermophilum* RBI67 (5 mismatches) (Grissa et al., 2007). Because of the low abundance of *E. lenta* within the human gut microbiota (Bisanz et al., 2018), we utilized a select set of 96 gut metagenomes that we previously found to have high *E. lenta* genome coverage (Koppel et al., 2018) to identify spacers by retrieving and assembling reads containing the direct repeat and then extracting spacers flanked by repeats containing no more than 3 mismatches from our consensus motif (Figure S3).

This analysis increased the total number of *E. lenta*-derived spacers (210 to 493; 2.3-fold). Consistent with our reference genomes, spacer length varied from 32 to 38 nt in metagenomes with 69.4% being 33 and 34 nt. When both isolate and metagenomes are combined and dereplicated, it is apparent that both datasets display a similar length distribution (Figure 2D). No assembled arrays were detected in a control set of 96 randomly selected metagenomes that contain *E. lenta* below the limit of detection (Nayfach et al., 2015). We next looked at shared spacers across reference genomes and metagenomes observing correspondence between spacer content and *cas* clade (Figure 4B). Metagenome-assembled CRISPR arrays were interwoven between clades, suggesting strains of *E. lenta* representing both clades can be found within the human gastrointestinal tract. The correspondence between spacer content clade was correlated with strain phylogeny (Figure S4), consistent with the idea that these sequences at least partially reflect evolutionary history. We also detected evidence for horizontal gene transfer: strain AB8n2 phylogenetically clusters with strains from clade B but contains a clade A system. While a common set of 47 spacers was observed across clades A, B, and metagenomes, each had a unique set with considerably

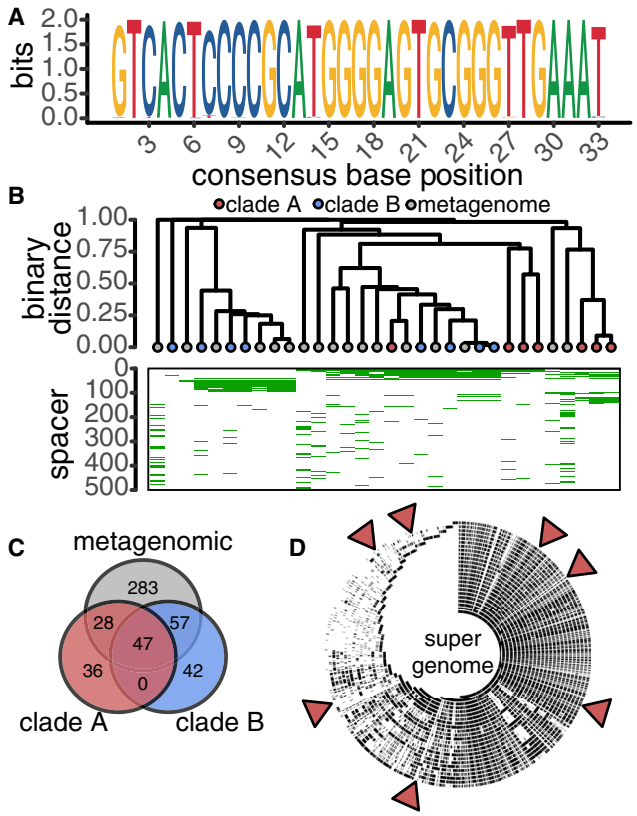


Figure 4. Direct Repeat and Spacer Conservation across Reference Genomes and Metagenomic Datasets

(A) The 33-nt *E. lenta* direct repeat was found to be highly conserved in all 15 CRISPR-positive isolate genomes.

(B) Analysis of shared spacer content between strains provides evidence of a clade-specific pattern of conservation. Spacers were numbered 1–493 ordered by frequency of occurrence.

(C) Venn diagram of shared spacer content between isolate genome clades and metagenomic data.

(D) Evaluation of self- (targeting of the genome by a spacer encoded within genome) and inter-strain targeting. Alignment of spacers to the *E. lenta* "super genome": a 7-Mbp non-redundant sequence representing the aggregate genomes of this bacterial species. Red triangles indicate spacer matches within putative prophage and mobile elements.

higher diversity in the metagenomic data (57.4% of unique spacers, Figure 4C).

To determine the extent to which CRISPR targeting occurs within the *E. lenta* pangenome, spacers were compared to a non-redundant representation of the *E. lenta* species genome. We found 60 putative protospacers present in 18 strains targeting a limited number of loci (Figure 4D). Of these protospacers, 8 occur within the genome encoding the spacer, which may suggest self-targeting (the remaining 52 were inter-strain and intra-species). Closer inspection of the protospacers revealed that 6 occur in a putative prophage observed in 2 of these strains (Bisanz et al., 2018), 1 in a suspected integrated plasmid, and another in a region adjacent to a tetracycline resistance gene (Table S1). Neither perfect alignments nor a flanking sequence indicative of a PAM was observed, suggesting that the *E. lenta* system does not actively target these sites.

Protospacer Identification Reveals Undescribed *E. lenta* Phages

Most spacers found in sequenced prokaryotic genomes lack a predictable target, emphasizing that many mobile genetic elements and phages remain unknown (Shmakov et al., 2017). To identify potential parasitic elements targeted by CRISPR, we queried 3 publicly available databases for matches to previously characterized plasmids or viruses; however, no significant matches were found. The NCBI non-redundant database allowed us to assign 1.6% of the spacers to chromosomal genes of cryptic function and origin (Figure 5A). These results are consistent with the vast viral diversity within humans and its limited representation in established databases.

To enable a more comprehensive platform for the identification of protospacers, we built a custom human virome database (HuVirDB) that integrates data from 18 publicly available virome studies representing 1,831 samples from 730 humans from 9 countries (Figure 5B; Tables S2 and S3). We assembled 19.4 Gbp of sequence from 1,783 samples recovering 3,386 putative protospacers representing 249/493 (50.5%) of spacers. These protospacers were observed across 218 human samples, 161 individuals, and 14 studies representing a broad geographical distribution (Table S4). Furthermore, we used this data to determine the PAM sequence through motif analysis of the protospacer adjacent regions. This revealed the canonical 5′ type I-C PAM "TTC" (Figure 5C) with no strong conservation in the 3′ region. In recovering protospacers, HuVirDB outperformed the NCBI environmental non-redundant database (NCBI env nt), which is 6.4-fold larger (124.1 Gbp) but resulted in half the matches (25.0%) with higher computational overhead and without easily accessible metadata (Figure 4A). We similarly contrasted our recovery of protospacers against the Integrated Microbial Genome/Virus 2.0 (IMG VR) (Paez-Espino et al., 2019), finding >2-fold increased protospacer identification with HuVirDB for *E. lenta*.

To examine the utility of this approach for the study of other gut bacterial species, we extracted spacers from the Human Microbiome Project (HMP) reference genomes, Pathosystems Resource Integration Center (PATRIC) genomes of human gastrointestinal origin, and a subset thereof belonging to 28 strains of *Akkermansia muciniphila*. Similar to *E. lenta*, *A. muciniphila* showed improved protospacer identification in HuVirDB compared against all other databases (Figure S5A). However, the overall HMP and PATRIC datasets had increased protospacer identification with IMG VR and the two NCBI databases, likely due to the presence of data from pathogens and bacteria from other body habitats in these other databases. Consistent with these observations, network analysis of CRISPR-array-containing genomes linked through common targets revealed the presence of strong clade specificity of CRISPR targeting (Figure S5B). These results emphasize the value of having complementary databases for protospacer identification depending on the specific bacterial host of interest.

To examine *E. lenta* target diversity, we clustered the protospacer-containing scaffolds at 80% global nucleotide identity into 13 non-singleton phage genomes (Figure 6A). Analysis of a representative sequence for each of these families revealed as many as 96 distinct protospacers, suggesting that *E. lenta*

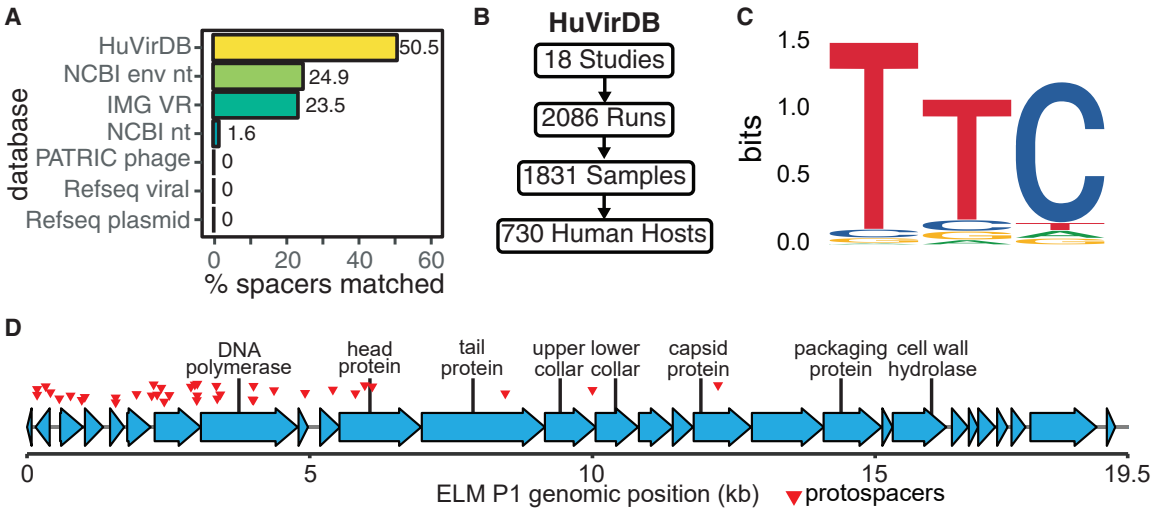


Figure 5. Discovering *E. lenta* Predators Based on Protospacer Enrichment

(A) Comparison of protospacer matches within HuVirDB (249/493) versus other publicly accessible databases, including isolated and sequenced plasmids and phages from RefSeq.

(B) To facilitate phage discovery, public virome sequencing data were collected and assembled for our HuVirDB.

(C) The 5′ flanking sequence was enriched for the canonical type I-C protospacer adjacent motif (PAM) TTC.

(D) Detailed annotation of a representative phage (ELM P1), identified based on a high frequency of matching *E. lenta* spacers.

has repeatedly been exposed to these "hyper-targeted" phage (Figure 6A). A representative phage, referred to as *Eggerthella lenta* metagenomic phage 1 (ELM P1), with a genome size of 19,474 bp, contains 37 distinct protospacers (Figure 5D). This phage possesses genes homologous to *Actinomyces* phage AV-1 and *Bacillus* phage phi29, which are small double-stranded DNA (dsDNA) phages of the podophage families with a genome size in the 17–22 kbp range (Delisle et al., 2006; Meijer et al., 2001). The protospacer sequences were concentrated within discrete portions of phage genomes, which may indicate bias toward the sequence injected earliest (Modell et al., 2017) and/or primed acquisition (Fineran et al., 2014; Künne et al., 2016). In almost all of the targeted phage sequences, we found annotated genes that suggest the presence of tail, collar, and head proteins (Figure 6A; Table S5).

To better understand the taxonomy and phylogeny of these ELM phages, we began by clustering previously described phages with approved taxonomies by the International Committee for the Taxonomy of Viruses (ICTV) (Bin Jang et al., 2019). We found that a subset (7/13) of ELM phages formed a subcluster that could not be assigned even a family-level taxonomy while the remaining phages were singletons (Figure 6B). We next built a phylogenetic tree that grouped 6/7 of these related ELM phages into a single cluster (Figure 6C), supporting their close taxonomic and phylogenetic relationship. The remaining ELM phages were clustered into at least two additional groups. Of note, ELM P1 grouped together with the other 6 phages by taxonomy but was in a distinct cluster based on phylogeny, potentially due to the mosaic nature of phage genomes. These results suggest that the metagenomic *E. lenta* phages we have observed represent a previously undescribed branch of phage diversity targeting gut Actinobacteria.

The Type I-C CRISPR-Cas System Can Accommodate Common Protospacer Mismatches

Further examination revealed that only 40.2% of protospacers were a perfect match to the spacer before dereplication into seed sequences (Table S4). Partial matches could indicate accumulated mutations that allow the phage to evade CRISPR-mediated immunity (Semenova et al., 2011) or that this system can accommodate mismatches, as has been demonstrated in other CRISPR-Cas system types (Pyenson et al., 2017). In order to distinguish between these two alternatives, we examined mismatch frequency as a function of the spacer-protospacer nucleotide position. The most commonly observed mismatched positions occur, in order of frequency, at nucleotides 18, 1, 33, and 9 (Figure 7A). We designed a series of spacers against JBD30-containing point mutations and examined their efficiency (Figures 7B and 7C).

Mutations in the 5′ spacer region, called the seed sequence, have been shown to provide phages an opportunity for escaping CRISPR immunity (Semenova et al., 2011). In accordance with this, we observed a high efficiency of plaquing for crRNAs with a mutation at the 3rd position or insertion at the 2nd position ($p = 0.021$, Kruskal-Wallis with Dunnett's post-test, Figure 7C). Interestingly, we noted that a mutation in the 31st nucleotide also allowed the phage to evade CRISPR interference ($p = 0.021$, Figures 7B and 7C). In contrast, single, double, and triple mutations in the middle of the spacer were tolerated, thus still providing immunity. Together, these results demonstrate that most naturally occurring mismatches still allow for efficient targeting of the invading sequence.

DISCUSSION

Here, we report the characterization of an active type I-C CRISPR-Cas system in a prevalent member of the human gut

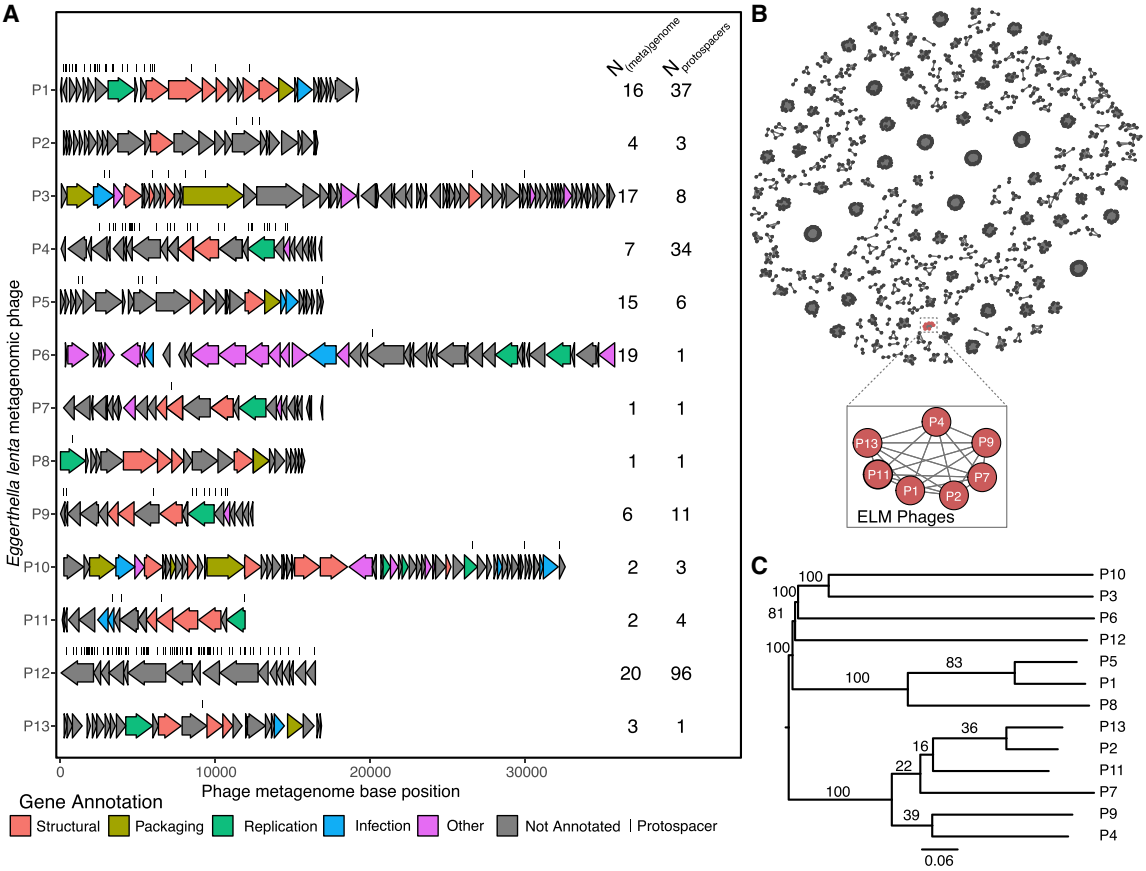


Figure 6. *E. lenta* Metagenomic (ELM) Phage Genomes

(A) Genomes are presented with annotated genes colored by high-level function, and protospacer locations are indicated by dashes. Protospacers were allowed to have up to 4 mismatches to the spacer sequence. The number of unique (meta)genomes targeting the seed phage and the number of unique spacers are shown.

(B) Clustering for taxonomic annotation of ELM phages with prokaryotic viral genomes from Viral RefSeq v.85 based on gene sharing (Bin Jang et al., 2019) demonstrates a unique clade formed by 7 ELM phages. Only non-singleton clusters are depicted.

(C) A phylogenetic tree of the ELM phages based on genome-wide BLAST distances (Meier-Kolthoff and Göker, 2017). Numbers on the tree represent pseudo-bootstrap support values from 100 replications.

microbiota, revealing undescribed hyper-targeted phages that infect gut Actinobacteria, which have eluded isolation despite their prevalence. By combining a systematic meta-analysis of virome datasets, metagenomics, and comparative genomics, we were able to uncover putative targets for >50% of *E. lenta* spacers. These results support the critical role of CRISPR-Cas systems in adaptive immunity to bacteriophages while also raising the question as to whether or not the remaining spacers target bacteriophages that remain to be discovered, mobile genetic elements, and/or as-of-yet unknown novel targets. These spacer-protospacer matches provide more definitive evidence for the host range of phages identified in virome datasets, as exemplified by the discovery of hyper-targeted phages that appear to have been repeatedly encountered and targeted by geographically diverse *E. lenta* CRISPR-Cas systems. The identified hyper-targeted phages are likely major determinants of *E. lenta* fitness, and their isolation or synthetic reconstitution would provide a major step forward in understanding the biology of this neglected bacterial species and determining whether or not the presence of multiple spacers within a single array is necessary for robust immunity.

Despite common mismatches detectable in gut viromes, we found that the *E. lenta* CRISPR-Cas system could tolerate single and even double or triple mutations within the middle of the spacer, as described in other types of systems (Pyenson et al., 2017). This suggests that phages may have a limited ability to escape targeting by mutation, requiring a mismatch in the first few nucleotides of the spacer or the PAM motif (both of which we detected in our computational analysis). Surprisingly, we also found a significant impact of point mutations in nucleotide 31, more work is necessary to determine why this particular nucleotide matters, either through disrupting complex formation, target binding, and/or nuclease activity.

Our results emphasize the critical importance of providing experimental support for CRISPR-Cas system function. In addition to the previously described false positives driven by incomplete systems (Zhang and Ye, 2017) and other types of genomic repeats (Zhang and Ye, 2017) in other environments, we found

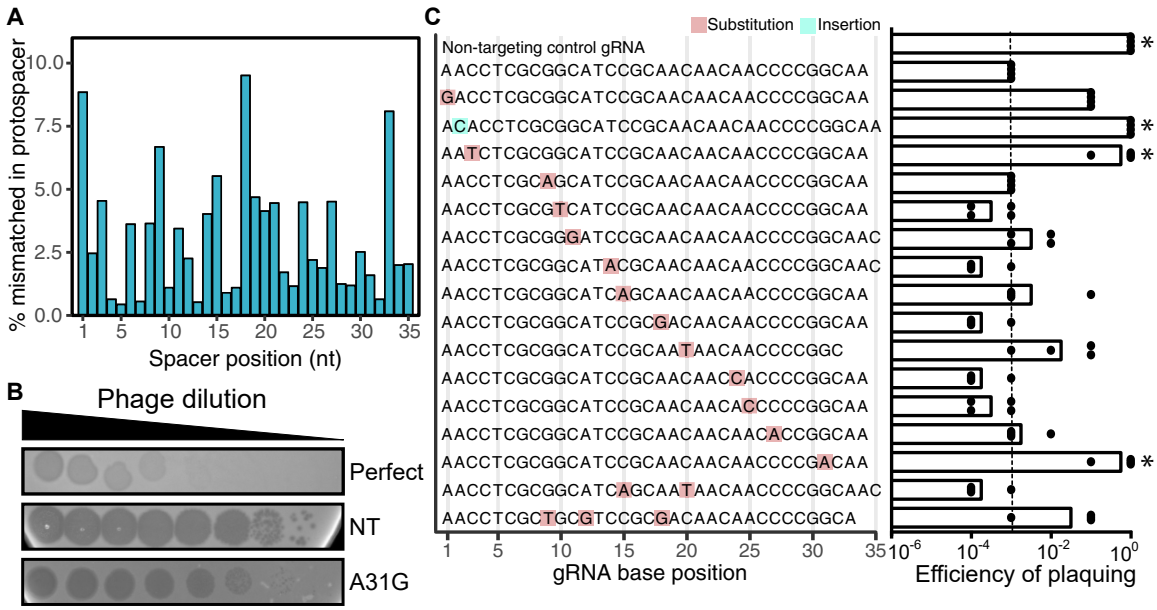


Figure 7. Most Mismatches between the Spacer and Protospacer Sequences Still Provide Immunity against Phages

(A) The occurrence of mismatches throughout the length of the spacer was calculated using HuVirDB.

(B) Plaque assays revealed two phenotypes: opaque plaques (efficient targeting) and clearer plaques (poor targeting). Controls include: a perfect match positive control and a NT negative control. A representative mismatch (A31G) is shown.

(C) Plaquing efficiency (log estimation in tested gRNA divided by log estimation in NT control) reveals that mutations in the seed sequence of the crRNA allow the phage to escape CRISPR-Cas immunity (n = 4). The dashed line at 10^{-5} denotes the average value for the perfect match control gRNA. *$p < 0.05$ Kruskal-Wallis with uncorrected Dunnett's (compared to targeting control).

that of the 24 *E. lenta* genomes analyzed, 9 lacked the entire type I-C system. The 15 strains that encoded a complete system could be binned into two distinct clades based on *cas* gene homology and spacer content, emphasizing the strain-level variation of these systems. Given this strain-level heterogeneity, our results emphasize the challenges in predicting bacterial interactions with phages based only on species abundance and the need for continued progress toward the functional characterization and mechanistic dissections of these systems within their natural host bacteria and physiological context.

These results also emphasize the utility of combining the computational and functional dissection of CRISPR-Cas systems in bacterial reference genome and metagenomic datasets to gain insights into the bacterial and viral components of the human microbiome. The approaches we have used do not require genetic tools in the target microorganism, enabling mechanistic insights into the vast majority of human-associated bacteria that remain genetically intractable (Burstein et al., 2017). More broadly, our development of HuVirDB provides a useful resource with rich metadata, enabling the study of predator-prey relationships across the human microbiome. While our current studies have focused on the *E. lenta* type I-C system, this database could be readily queried for matches to spacers from other human gut bacteria of interest. To facilitate the rapid adoption of this tool in the microbiome and CRISPR-Cas community, we have made all of the data publicly accessible and have integrated it into a widely used graphical tool for spacer matching (Biswas et al., 2013).

Finally, our work provides fundamental biological insights into endogenous CRISPR-Cas systems found within the human gut microbiome, an essential prerequisite for efforts to reprogram these systems to impact the structure and function of complex host-associated microbial communities more precisely. Continued progress in this area will require the development of approaches for gene delivery within the gastrointestinal tract, robust methods to engineer bacteriophage or other vectors, and the identification of bacterial targets with readily quantifiable impacts on host pathophysiology.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Bacterial and Phage Culture
- METHOD DETAILS
  - RNA Extraction
  - RT-qPCR
  - Northern Blot
  - RNA-Sequencing Analysis
  - Construction of the *Pseudomonas aeruginosa* Strain Carrying the *Eggerthella lenta cas* Genes
  - crRNA Cloning
  - Transformation of *Pseudomonas aeruginosa*
  - Phage Plaque Assays
  - Metagenomic CRISPR Spacer Arrays

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2019.08.008.

ACKNOWLEDGMENTS

We are grateful to Stephen Nayfach for providing the E. *lenta*-enriched metagenomes, Elizabeth Bess for providing RNA sequencing data, Chris Brown and Peter Fineran for including HuVirDB as a database for CRISPRTarget, and all of the members of the Turnbaugh and Bondy-Denomy labs for helpful discussion. This work was supported by the National Institutes of Health (R01HL122593), the Searle Scholars Program (SSP-2016-1352), and the UCSF Program for Breakthrough Biomedical Research (partially funded by the Sandler Foundation). P.J.T. holds an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund, is a Chan Zuckerberg Biohub investigator, and is Nadia's Gift Foundation Innovator supported, in part, by the Damon Runyon Cancer Research Foundation (DRR-42-16). Fellowship support was provided by the Canadian Institutes of Health Research (K.N.L.), the Natural Sciences and Engineering Research Council (J.E.B.), and the UCSF Discovery Fellows (P.S.-P.). J.B.-D. was supported by an NIH Office of the Director Early Independence Award (DP5-OD021344).

AUTHOR CONTRIBUTIONS

Conceptualization, P.S.-P., J.E.B., J.B.-D., and P.J.T.; Methodology, P.S.-P., J.E.B., J.D.B., and K.N.L.; Software, J.E.B.; Investigation, P.S.-P., J.E.B., J.D.B., and K.N.L.; Resources, J.B.-D. and P.J.T.; Writing – Original Draft, P.S.-P. and J.E.B.; Writing – Review & Editing, P.S.-P., J.E.B., K.N.L., J.B.-D., and P.J.T.; Supervision, J.B.-D. and P.J.T.; Funding Acquisition, J.B.-D. and P.J.T.

DECLARATION OF INTERESTS

P.J.T. is on the scientific advisory board for Kaleido, Seres, SNIPRbiome, uBiome, and WholeBiome, and J.B.-D. is on the scientific advisory board for SNIPRbiome and Excision Biotherapeutics and is a co-founder of Acrigen Biosciences; there is no direct overlap between the current study and these consulting duties.

Received: March 21, 2019
Revised: June 20, 2019
Accepted: August 13, 2019
Published: September 3, 2019

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. *44*, W16–W21.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Barrangou, R., and Horvath, P. (2017). A decade of discovery: CRISPR functions and applications. Nat. Microbiol. *2*, 17092.

Belkaid, Y., and Hand, T.W. (2014). Role of the microbiota in immunity and inflammation. Cell *157*, 121–141.

Bess, E.N., Bisanz, J.E., Spanogiannopoulos, P., Ang, Q.Y., Bustion, A., Kitamura, S., Alba, D.L., Wolan, D.W., Koliwad, S.K., and Turnbaugh, P.J. (2018). The genetic basis for the cooperative bioactivation of plant lignans by a human gut bacterial consortium. bioRxiv. https://doi.org/10.1101/357640.

Bikard, D., Hatoum-Aslan, A., Mucida, D., and Marraffini, L.A. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during *in vivo* bacterial infection. Cell Host Microbe *12*, 177–186.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639.

Bisanz, J.E., Soto-Perez, P., Lam, K.N., Bess, E.N., Haiser, H.J., Allen-Vercoe, E., Rekdal, V.M., Balskus, E.P., and Turnbaugh, P.J. (2018). Illuminating the microbiome's dark matter: a functional genomic toolkit for the study of human gut Actinobacteria. bioRxiv. https://doi.org/10.1101/304840.

Biswas, A., Gagnon, J.N., Brouns, S.J.J., Fineran, P.C., and Brown, C.M. (2013). CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. RNA Biol. *10*, 817–827.

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., and Hochreiter, S. (2015). msa: an R package for multiple sequence alignment. Bioinformatics *31*, 3997–3999.

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

Bondy-Denomy, J., Pawluk, A., Maxwell, K.L., and Davidson, A.R. (2013). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature *493*, 429–432.

Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., et al. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci. Rep. *5*, 8365.

Britton, R.A., and Young, V.B. (2014). Role of the intestinal microbiota in resistance to colonization by *Clostridium difficile*. Gastroenterology *146*, 1547–1553.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR–Cas systems from uncultivated microbes. Nature *542*, 237–241.

Chan, R.C., and Mercer, J. (2008). First Australian description of *Eggerthella lenta* bacteraemia identified by 16S rRNA gene sequencing. Pathology *40*, 409–410.

Choi, K.H., and Schweizer, H.P. (2006). Mini-Tn7 insertion in bacteria with single attTn7 sites: example *Pseudomonas aeruginosa*. Nat. Protoc. *1*, 153–161.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147.

Delisle, A.L., Barcak, G.J., and Guo, M. (2006). Isolation and expression of the lysis genes of *Actinomyces naeslundii* phage Av-1. Appl. Environ. Microbiol. *72*, 1110–1117.

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J. Bacteriol. *190*, 1390–1400.

Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S.A.F.T., Staals, R.H.J., and Brouns, S.J.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proc. Natl. Acad. Sci. USA *111*, E1629–E1638.

Garneau, J.E., Dupuis, M.È, Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Göker, M., García-Blázquez, G., Voglmayr, H., Tellería, M.T., and Martín, M.P. (2009). Molecular taxonomy of phytopathogenic fungi: a case study in *Peronospora*. PLoS One *4*, e6319.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics *8*, 172.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Hahne, F., and Ivanek, R. (2016). Visualizing genomic data using Gviz and Bioconductor. Methods Mol. Biol. *1418*, 335–351.

Haiser, H.J., Gootenberg, D.B., Chatman, K., Sirasani, G., Balskus, E.P., and Turnbaugh, P.J. (2013). Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. Science *341*, 295–298.

Harris, S.C., Devendran, S., Méndez-García, C., Mythen, S.M., Wright, C.L., Fields, C.J., Hernandez, A.G., Cann, I., Hylemon, P.B., and Ridlon, J.M. (2018). Bile acid oxidation by Eggerthella lenta strains C592 and DSM 2243T. Gut Microbes *9*, 1–17.

Hochstrasser, M.L., Taylor, D.W., Kornfeld, J.E., Nogales, E., and Doudna, J.A. (2016). DNA targeting by a minimal CRISPR RNA-guided cascade. Mol. Cell *63*, 840–851.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

Koppel, N., Bisanz, J.E., Pandelia, M.E., Turnbaugh, P.J., and Balskus, E.P. (2018). Discovery and characterization of a prevalent human gut bacterial enzyme sufficient for the inactivation of a family of plant toxins. eLife *7*, e33953.

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. Mol. Cell *63*, 852–864.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018). Cas4-dependent prespacer processing ensures high-fidelity programming of CRISPR arrays. Mol. Cell *70*, 48–59.

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol. Biol. Evol. *32*, 2798–2800.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAM tools. Bioinformatics *25*, 2078–2079.

Maini Rekdal, V.M., Bess, E.N., Bisanz, J.E., Turnbaugh, P.J., and Balskus, E.P. (2019). Discovery and inhibition of an interspecies gut bacterial pathway for levodopa metabolism. Science *364*, eaau6323.

Majsec, K., Bolt, E.L., and Ivančić-Baće, I. (2016). Cas3 is a limiting factor for CRISPR-Cas immunity in *Escherichia coli* cells lacking H-NS. BMC Microbiol. *16*, 28.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Marino, N.D., Zhang, J.Y., Borges, A.L., Sousa, A.A., Leon, L.M., Rauch, B.J., Walton, R.T., Berry, J.D., Joung, J.K., Kleinstiver, B.P., et al. (2018). Discovery of widespread type I and type V CRISPR-Cas inhibitors. Science *362*, 240–242.

McGinn, J., and Marraffini, L.A. (2019). Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat. Rev. Microbiol. *17*, 7–12.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics *14*, 60.

Meier-Kolthoff, J.P., and Göker, M. (2017). VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics *33*, 3396–3404.

Meier-Kolthoff, J.P., Hahnke, R.L., Petersen, J., Scheuner, C., Michael, V., Fiebig, A., Rohde, C., Rohde, M., Fartmann, B., Goodwin, L.A., et al. (2014). Complete genome sequence of DSM 30083T, the type strain (U5/41T) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. Stand. Genomic Sci. *9*, 2.

Meijer, W.J., Horcajadas, J.A., and Salas, M. (2001). Phi29 family of phages. Microbiol. Mol. Biol. Rev. *65*, 261–287.

Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101–104.

Nayfach, S., Fischbach, M.A., and Pollard, K.S. (2015). MetaQuery: a web server for rapid annotation and quantitative analysis of specific genes in the human gut microbiome. Bioinformatics *31*, 3368–3370.

Paez-Espino, D., Roux, S., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Huntemann, M., Reddy, T.B.K., Pons, J.C., Llabrés, M., et al. (2019). IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res. *47*, D678–D686.

Palmer, K.L., and Gilmore, M.S. (2010). Multidrug-resistant *enterococci* lack CRISPR-cas. mBio *1*, e00227.

Plagens, A., Tjaden, B., Hagemann, A., Randau, L., and Hensel, R. (2012). Characterization of the CRISPR/Cas subtype IA system of the hyperthermophilic crenarchaeon *Thermoproteus tenax*. J. Bacteriol. *194*, 2491–2500.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.

Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N., and Wagner, R. (2010). Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. Mol. Microbiol. *75*, 1495–1512.

Pyenson, N.C., Gayvert, K., Varble, A., Elemento, O., and Marraffini, L.A. (2017). Broad targeting specificity during bacterial type III CRISPR-Cas immunity constrains viral escape. Cell Host Microbe *22*, 343–353.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Qiu, D., Damron, F.H., Mima, T., Schweizer, H.P., and Yu, H.D. (2008). PBAD-based shuttle vectors for functional analysis of toxic and highly regulated genes in *Pseudomonas* and *Burkholderia* spp. and other bacteria. Appl. Environ. Microbiol. *74*, 7422–7426.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L., and White, M.F. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. eLife *4*, e08716.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. USA *108*, 10098–10103.

Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V., and Koonin, E.V. (2017). The CRISPR spacer space Is dominated by sequences from species-specific mobilomes. mBio *8*, e01397.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., et al.

(2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature *406*, 959–964.

Tajkarimi, M., and Wexler, H.M. (2017). CRISPR-Cas systems in *Bacteroides fragilis*, an important pathobiont in the human gut microbiome. Front. Microbiol. *8*, 2234.

Wagih, O. (2017). Ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647.

Zhang, Q., Doak, T.G., and Ye, Y. (2014). Expanding the catalog of *cas* genes with metagenomes. Nucleic Acids Res. *42*, 2448–2459.

Zhang, Q., and Ye, Y. (2017). Not all predicted CRISPR-Cas systems are equal: isolated *cas* genes and classes of CRISPR like elements. BMC Bioinformatics *18*, 92.

Zhu, Y., Stephens, R.M., Meltzer, P.S., and Davis, S.R. (2013). SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics *14*, 19.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
Eggerthella lenta strain DSM 2243	(Bisanz et al., 2018)	DSM 2243
Eggerthella lenta strain 28B	(Bisanz et al., 2018)	28B
Eggerthella lenta strain Valencia	(Bisanz et al., 2018)	Valencia
Pseudomonas aeruginosa strain PA01	(Stover et al., 2000)	PA01
JBD30	(Bondy-Denomy et al., 2013)	JBD30
Chemicals, Peptides, and Recombinant Proteins		
DNA Polymerase I, Large (Klenow) Fragment	New England Biolabs	Cat#M0210S
T4 Polyucleotide Kinase	New England Biolabs	Cat#M0201S
T4 DNA Ligase	New England Biolabs	Cat#M0202S
Critical Commercial Assays		
Purelink™ RNA Mini Kit	Invitrogen	Cat#12183025
Tri Reagent®	Sigma	Cat#T3809
iScript™ Reverse Transcription SuperMix	Bio-Rad	Cat#1708841
SYBR Select Master Mix For CFX	Thermofisher	Cat#4472942
TURBO DNase	Thermofisher	Cat#AM2238
QIAquick PCR Purification Kit	QIAGEN	Cat#28106
Purelink Quick Gel Extraction Kit	Thermofisher	Cat#K210025
NEB® 5-alpha Competent E. coli	New Englands Biolabs	Cat#2987H
Ribo-Zero rRNA Removal Kit	Illumina	Cat#MRZB12424
NEBNext Ultra RNA Library Prep Kit	New England Biolabs	Cat#K210025
Deposited Data		
HuVirDB Assemblies	opengut.ucsf.edu/HuVirDB-1.0.fasta.gz	Version 1.0
HuVirDB Metadata	github.com/jbisanz/HuVirDB	Version 1.0
E. lenta genomes	NCBI Genomes	PRJNA412637, PRJNA384908, PRJNA21093, PRJNA46413, PRJNA40023, PRJNA59527
HMP Reference Genomes	NCBI Genome	PRJNA28331
PATRIC Genomes	patricbrc.org/view/Taxonomy/2#view_tab=genomes	Accessed 23 August 2018
IMG VR	(Paez-Espino et al., 2019)	Jan 2018
Experimental Models: Organisms/Strains		
PA01 tn7::lentaIC	this study	N/A
PA01 tn7::lentaIC cas3 after p30- 168/169	this study	N/A
PA01 tn7::IC minimal system from E. lenta, cas3 after p30 gKz	this study	N/A
PA01 tn7::IC minimal system from E. lenta, cas3 after p30 gJBD30	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 41	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 611	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 729	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 445	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 15	this study	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PA01 tn7::Lenta I-C, pJB3 + gRNA 18	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 24	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 27	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 1	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 10	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 3	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 25	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + gRNA 31	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 32	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 33	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 34	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 35	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + insert 95	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + insert 102	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + NTC insert	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + insertion +2	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 30	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 36	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 38	this study	N/A
PA01 tn7::Lenta I-C, pJB3 + control 40	this study	N/A
Oligonucleotides		
See Table S6.		
Recombinant DNA		
pHERD30T	(Qiu et al., 2008)	N/A
pJB3	this study	N/A
Software and Algorithms		
CRISPRCasFinder 1.1.0	(Couvin et al., 2018)	crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index
MINced 0.2.0	GitHub	github.com/csSkennerton/minced
Biostrings 2.48.0	Bioconductor	bioconductor.org/packages/release/bioc/html/Biostrings.html
Tidyverse 1.2.1	CRAN	https://www.tidyverse.org/
R 3.5.0	CRAN	https://www.r-project.org/
FastTree 2.1.10	(Price et al., 2009)	microbesonline.org/fasttree/
Progressive Mauve Feb 13 2015	(Darling et al., 2010)	darlinglab.org/mauve/download.html
BLAST 2.6.0+	NCBI	ftp.ncbi.nlm.nih.gov/blast/executables/blast+
SAMtools 1.9	(Li et al., 2009)	samtools.sourceforge.net/
Trimmomatic 0.32	(Bolger et al., 2014)	usadellab.org/cms/?page=trimmomatic
metaSPAdes 3.13.0	(Bankevich et al., 2012)	cab.spbu.ru/software/spades/
SPAdes 3.7.0	(Bankevich et al., 2012)	cab.spbu.ru/software/spades/
RASTtk	(Brettin et al., 2015)	http://rast.theseed.org/
PHASTER	(Arndt et al., 2016)	phaster.ca
Gggenes 0.3.1	GitHub	github.com/wilkox/gggenes
Vsearch 1.11.0	(Rognes et al., 2016)	github.com/torognes/vsearch
vConTACT2 0.9.9	(Bin Jang et al., 2019)	bitbucket.org/MAVERICLab/vcontact2/src/master/
VICTOR	(Meier-Kolthoff and Göker, 2017)	ggdc.dsmz.de/victor.php
Ggnet 0.1.0	GitHub	briatte.github.io/ggnet/
FigTree v1.4.3	(Price et al., 2009)	tree.bio.ed.ac.uk/software/figtree/

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Peter Turnbaugh (peter.turnbaugh@ucsf.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bacterial and Phage Culture

All strains used for these studies are listed in the Key Resources Table. *Pseudomonas aeruginosa* phage JBD30 was propagated on *P. aeruginosa* PA01 (Stover et al., 2000) and stored in SM buffer at 4°C. The titer of the phage was determined by performing serial dilutions and mixing 10 μl of phage with 150 μl of *P. aeruginosa* PA01 (grown overnight) in 0.7% LB agar and incubating overnight at 30°C. Routine culturing of *E. lenta* was done under anaerobic conditions (Coy Lab Products) using BHI++ media (BHI with 1% arginine, 0.05% L-cysteine-HCl, 1-μg/mL vitamin K, 5 μg/mL hemin, and 0.0001% w/v resazurin (Bisanz et al., 2018). Routine culturing of *P. aeruginosa* was done aerobically with rotation in LB media. For *P. aeruginosa* PA01 tn7::lentaIC with plasmid pJB3 carrying the crRNA construct, the cells were grown in LB supplemented with 50 μg/ml of gentamicin.

METHOD DETAILS

RNA Extraction

RNA extractions were performed as described previously (Bess et al., 2018). Briefly, a 24-hr broth culture of *Eggerthella lenta* DSM 2243 was subcultured at 1% v/v in BHI++ and allowed to grow for 24 hours (until mid-exponential, OD_{600} of ~0.3) in an anaerobic chamber (Coy Laboratory Products). The cells were spun down at max speed, 4°C, for 10 minutes. Cells were resuspended in TRI Reagent (Sigma) and lysed by a bead beater (BioSpec Products). Chloroform was added 1:5 to the mixture, incubated at room temperature for 10 minutes, and then spun down at 16,000 × *g* for 15 minutes. The top phase was placed in a clean tube and mixed 1:1 with 100% ethanol. The RNA extraction was then done using Purelink™ RNA Mini Kit (Invitrogen) according to the manufacturer's protocol, with the addition of an in-column DNase treatment. A second DNase treatment was done after eluting using TURBO-DNase (Ambion).

RT-qPCR

Reverse transcription was carried out using 500 ng of total RNA and the iScript™ Reverse Transcription SuperMix according to the manufacturer's protocol (Bio-Rad). The qPCR assays were performed using SYBR Select Master Mix for CFX (Applied Biosystems, Waltham, Massachusetts, United States of America) run in a CFX384 Real-Time System (BioRad) using 10-μl reactions according to the manufacturer's recommendations. 200-nM primers were used to quantify gene expression as listed in the Key Resources Table. All primers used are listed in Table S6.

Northern Blot

The Northern blot was carried out as previously described (Bondy-Denomy et al., 2013). Briefly, the probe was generated by amplifying a fragment spanning the first four spacers of the DSM 2243 CRISPR array, cleaning the PCR product (Qiagen PCR Purification Kit) and labeling 300 ng of the clean product with Alpha-32P dCTP using Klenow polymerase (NEB M0210L). 5 μg of total RNA from *E. lenta* DSM 2243 (grown to mid-exponential) were used to run (per lane) in a denaturing gel. The RNA was transferred to a positively charged nylon membrane (Roche) using the semi-dry setting in a Trans-Blot Turbo (Bio-Rad) and crosslinked with 10 mJ UV burst over 30 seconds (Stratagene). The membrane was blocked with pre-hybridization buffer, consisting of 50% formamide, 5x Denhardts solution, 6x SSC, and 100 μg/ml of salmon sperm DNA, at 42°C for 1 hour. Probing was done at 42°C for 16-18 hours using the probe labeled with >$4x10^5$ cpm of dCTP. Afterwards, the blot was washed with wash solution 1 (2xSSC and 1% SDS) twice for 10 minutes at 18°C, two 30 minutes washes at 65°C, and wash solution 2 (0.2x SSC and 0.1%SDS) for 10 minutes at 18C. The blot was developed using a phosphoimager.

RNA-Sequencing Analysis

The RNA-sequencing of *E. lenta* DSM 2243 was described elsewhere (Bess et al., 2018) and reads are available under Sequence Read Archive Project SRP140684. Briefly, RNA was extracted from triplicate mid-exponential cultures as described above and rRNA depletion (Illumina Ribo-Zero) was used for subsequent library construction (NEBNext Ultra RNA). Sequencing was conducted via Illumina HiSeq 2500 with single ended 51 bp chemistry. Using Bowtie2 (Langmead and Salzberg, 2012), the reads were mapped to the reference DSM 2243 assembly (GCA_000024265.1) with the following parameters: –end-to-end –sensitive –trim5 5 –trim3 5. Next counts per feature were determined using htseq-count (Anders et al., 2015) and normalized using the reads per million per kilobase (RPKM) method. Sequencing coverage over the entire CRISPR-Cas locus was visualized using Gviz (Hahne and Ivanek, 2016). The calculation of background expression levels was done by averaging the reads of intergenic regions (leaving out ±200 bp from coding sequences).

Construction of the *Pseudomonas aeruginosa* Strain Carrying the *Eggerthella lenta cas* Genes

Chromosomal insertion of *cas5-8-7-3* genes into *P. aeruginosa* was done as previously described, with insertion at the Tn7 location via a helper transposase vector (Choi and Schweizer, 2006). The *cas3* gene was cloned downstream of *cas7* to mitigate toxicity due to overexpression. Gentamicin resistant strains were selected and the insertion location confirmed via PCR. The gentamicin marker was then flipped out via FLP recombinase, generating a gentamicin sensitive strain with stably integrated and IPTG-inducible Cas proteins. To introduce crRNAs, the pHERD30T vector (Qiu et al., 2008) was used, a high copy gentamicin resistance, arabinose inducible shuttle vector. An ''entry'' array was designed containing a repeat-pseudospacer-repeat organization (pJB3). The pseudospacer possessed two BsaI sites to enable the cloning of annealed oligonucleotides as described previously (Marino et al., 2018).

crRNA Cloning

The vector pJB3 was digested using the enzyme BsaI (NEB) and the fragment was gel extracted (Invitrogen Gel Extraction Kit). The primers (IDT & Sigma), carrying the point mutations of interest, were annealed and phosphorylated in a single reaction with 10x T4 Ligation buffer (NEB) and T4 Polynucleotide kinase (NEB) by incubating at 37°C for 2 hours, 95°C for 5 minutes, and ramp down to 20°C at 5°C/minute. Afterwards, they were diluted 1:500 in water and 1 μl was used to ligate to 60 ng of digested pJB3. The ligation was carried out overnight and stopped by incubating at 65°C for 20 minutes. 2 μl of the ligation were used to transform into NEB 5-alpha competent *E. coli* following the manufacturer's protocol. The cells were grown in LB agar supplemented with 30 μg/ml of gentamicin. Cloning was verified by Sanger sequencing and plasmids were used to transform *Pseudomonas aeruginosa* tn7::lentaIC.

Transformation of *Pseudomonas aeruginosa*

A seed culture of the *P. aeruginosa* strain was subcultured 1:100 in fresh LB media and allowed to grow for 18 hours. 2 ml of the culture were spun down and washed twice with 300 mM sucrose and then re-suspended in 225 μl of 300 mM sucrose. 100 μl of the washed cells and 10-100 ng of plasmid DNA were used per transformation reaction. The cells were electroporated in a 0.2 mm cuvette using 25 μF, 200 ohm, and 2.5 kV. After the pulse, 800 μl of LB were added to the cells and then incubated at 37°C with shaking for 45 minutes. 100 μl of the reaction were used to spread in an LB agar plate supplemented with 50 μg/ml of gentamicin.

Phage Plaque Assays

Bacterial lawns were made by mixing 150 μl of an overnight culture of host bacteria with 4 ml of 0.7% LB agar with 10 mM $MgSO_4$, 50 μg/ml of gentamicin and the inducers of expression (0.5 mM IPTG and 0.1% arabinose). Phage dilutions were made by diluting the phage in SM buffer and 3 μl of each dilution were used to spot on the bacterial lawn. The plates were incubated at 30°C for 16–18 hours, after which the pfus were quantified.

Metagenomic CRISPR Spacer Arrays

As previously identified (Koppel et al., 2018), paired end sequences from 96 *E. lenta*-enriched metagenomes were retrieved from the NCBI Sequence Read Archive. Reads were filtered using Trimmomatic (Bolger et al., 2014) to remove potential adapters and trimmed using the default sliding window filter. Next reads containing the consensus direct repeat were identified using vsearch (Rognes et al., 2016) with the following parameters: –usearch_global –id 0.87 –maxgaps 1 –maxsubs 4 –mincols 32 –maxaccepts 0 –maxrejects 0 –strand both. Assembly was then carried out with SPAdes 3.7.0 (Bankevich et al., 2012). 96 *E. lenta*-deficient metagenomes, as determined from Metaquery (Nayfach et al., 2015), were also included and used as negative controls. CRISPR-array assemblies could not be generated from any of the *E. lenta*-deficient metagenomes. Spacers were extracted from these assemblies as below.

Comparative Genomics and Spacer Identification

The collection and sequencing of the *E. lenta* genomes is described elsewhere (Bisanz et al., 2018). Annotation of the *cas* genes of *E. lenta* strain 28B was done using CRISPRCasFinder (Couvin et al., 2018) (Figure S3C). The presence of CRISPR arrays and their direct repeats in genome assemblies was first determined using the MINced 0.2.0 (github.com/ctSkennerton/minced). The consensus direct repeat sequence was determined via the MSA package (Bodenhofer et al., 2015) and ggseqlogo (Wagih, 2017). Arrays were then recalled from both isolate and meta-genomes by extracting regions flanked by the consensus 5′-GTCACTCCCCG CATGGGGAGTGCGGGTTGAAAT-3′ allowing for up to 3 mismatches from the consensus. The uniqueness of the *E. lenta* direct repeat was determined through comparison against our Coriobacteriia collection and through the use of the CRISPRdb (Grissa et al., 2007). The locus diagram was prepared through identification of orthologous gene clusters containing the DSM 2243 *cas* genes and extracting their genomic coordinates from GenBank transfer format files (Bisanz et al., 2018). Relative base position was determined by recentering coordinates on the 5′ translational start site of *cas5*. Nucleotide identity was determined by a Needleman-Wunsch global alignment of nucleotide sequence with percent ID calculated as 100×(identical positions) / (aligned positions + internal gap positions). The *cas* gene phylogenetic tree was created by concatenating the individual alignments of the *cas genes* as before, and building a tree with FastTree (Price et al., 2009). The super genome alignment was created using the Progressive Mauve algorithm (Darling et al., 2010) and plotting hits on this set of super-coordinates.

HuVirDB

We queried the NCBI Sequence Read Archive for studies of human-associated phage communities with shotgun sequencing data available. 18 studies were identified with sufficient metadata for inclusion (Table S2). Where possible, relevant per-sample metadata

was preserved as identified in the SRA or in the original publication. A total of 1831 samples were collected for assembly and matching runs determined using the SRAdb package (Zhu et al., 2013), however 49 low-coverage samples (2.7%) failed assembly and were not pursued further. Trimmomatic was used to remove possible adapter contamination with sliding window filtering, then metaSPAdes (or SPAdes when SE reads) was used for assembly. When 454 sequencing was applied, error correction was bypassed using the –only-assembler flag. Resulting contigs were identified by their default identifier concatenated to their SRA sample accession and merged to form a single large database. Assembly statistics were generated using QUAST with a minimum contig size of 200 (Gurevich et al., 2013). To identify *E. lenta* protospacers, per-sample databases were queried through BLASTn (-task BLASTn-short) reporting 100 alignments with no more than 4 misaligned bases (qlen-nident<=4) allowed and filtered to ensure that the flanking sequences (SAMtools 1.9) (Li et al., 2009) did not contain either the *E. lenta* direct repeat, or other repetitive sequence that could indicate the hit was a component of a contaminating CRISPR array. Phages of interest were annotated using a combination of RASTtk (Brettin et al., 2015) and PHASTER (Arndt et al., 2016) and visualized using gggenes (github.com/wilkox/gggenes). Based on these annotations, the genes were manually grouped into distinct functional categories: structural, packaging, replication, infection, other, and not annotated. *E. lenta*-targeted contigs were dereplicated based on a all-versus-all global nucleotide alignment strategy with 80% identity (measured as identities over the length of the shorter sequence) used as the clustering threshold. The largest phage assembly within the cluster served as the seed sequence, and if a fragment could be assigned with equal confidence to multiple seeds, one was randomly selected. Seed phage sequences are available at github.com/jbisanz/HuVirDB. Taxonomic clustering of ELM phages was carried out using VConTACT2 v0.9.9 against the ProkaryoticViralRefSeq85-ICTV database and visualized in R using ggnet v0.1.0. To generate the phylogenetic tree of ELM phages, all pairwise comparisons of the nucleotide sequences were conducted using the Genome-BLAST Distance Phylogeny (GBDP) method (Meier-Kolthoff et al., 2013) under settings recommended for prokaryotic viruses (Meier-Kolthoff and Göker, 2017). The resulting intergenomic distances were used to infer a balanced minimum evolution tree with branch support via FASTME including SPR postprocessing (Lefort et al., 2015) for formula D0. Branch support was inferred from 100 pseudo-bootstrap replicates each. Trees were rooted at the midpoint and visualized with FigTree v1.4.3. Taxon boundaries were estimated with the OPTSIL program (Göker et al., 2009), the recommended clustering thresholds (Meier-Kolthoff and Göker, 2017) and an F value (fraction of links required for cluster fusion) of 0.5 (Meier-Kolthoff et al., 2014).

To contrast databases, HMP reference genomes and PATRIC reference genomes had CRISPR-arrays extracted as above using *MINced* which were then merged with the *E. lenta* spacers previously identified. These were queried against BLAST databases as above including a concatenated HuVirDB, NCBI environmental non-redundant (env nt), IMG VR (January 2018 release), NCBI non-redundant nucleotide (nt), PATRIC phage, and Refseq viral and plasmid databases. *Akkermansia muciniphila* spacers were identified by being encoded in a genome annotated as *A. muciniphila* according to PATRIC metadata.

QUANTIFICATION AND STATISTICAL ANALYSIS

Where applicable, statistical analysis was carried out using either Prism 7 (GraphPad Software) or R 3.5.0 using Mann-Whitney U-tests or Kruskal-Wallis one-way analysis of variance with Dunnett's multiple comparison post-hoc test. Phage plaque counts were estimated to the nearest 10-fold dilution with representative images of plaque morphology provided.

DATA AND CODE AVAILABILITY

HuVirDB metadata and related information is available at github.com/jbisanz/HuVirDB and the database itself for download at opengut.ucsf.edu/HuVirDB-1.0.fasta.gz. HuVirDB has been made available in CRISPRTarget as an available database (http://crispr.otago.ac.nz/CRISPRTarget/crispr_analysis.html) for general queries. Genome assemblies are available under the following BioProjects: PRJNA412637, PRJNA384908, PRJNA21093, PRJNA46413, PRJNA40023, PRJNA59527. HMP reference genomes were retrieved from NCBI using BioProject PRJNA28331 (retrieved 23 August 2018). PATRIC reference genomes were retrieved from NCBI on 23 August 2018 by assembly accession as identified in the PATRIC genome catalog (patricbrc.org/view/Taxonomy/2#view_tab=genomes) after filtering for host_name containing human or sapiens, and an isolation_source containing stool, faecal, faeces, fecal, feces, gastrointestinal, gut, intestine, rectal, or rectum.

Cell Host & Microbe, Volume *26*

Supplemental Information

CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog

Paola Soto-Perez, Jordan E. Bisanz, Joel D. Berry, Kathy N. Lam, Joseph Bondy-Denomy, and Peter J. Turnbaugh

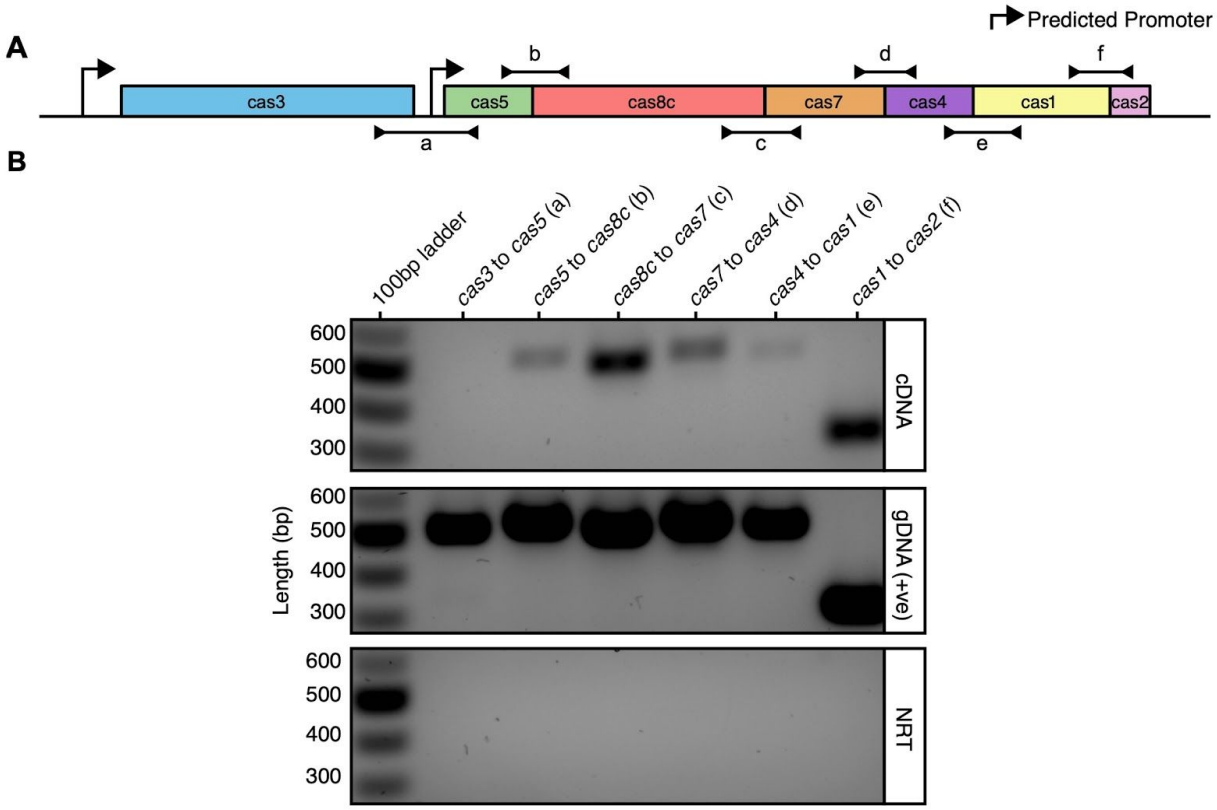


Figure S1. Mapping of the *cas* operon related to Figure 1. (**A**) Schematic of the computationally predicted promoters and primers designed to amplify regions that overlap each pair of neighboring genes. (**B**) The genes *cas3* and *cas5* are regulated by distinct promoter sequences. All of the other tested gene pairs were detected on the same transcript. A positive genomic DNA (gDNA) control is shown in addition to a negative control: NRT (no reverse transcriptase).

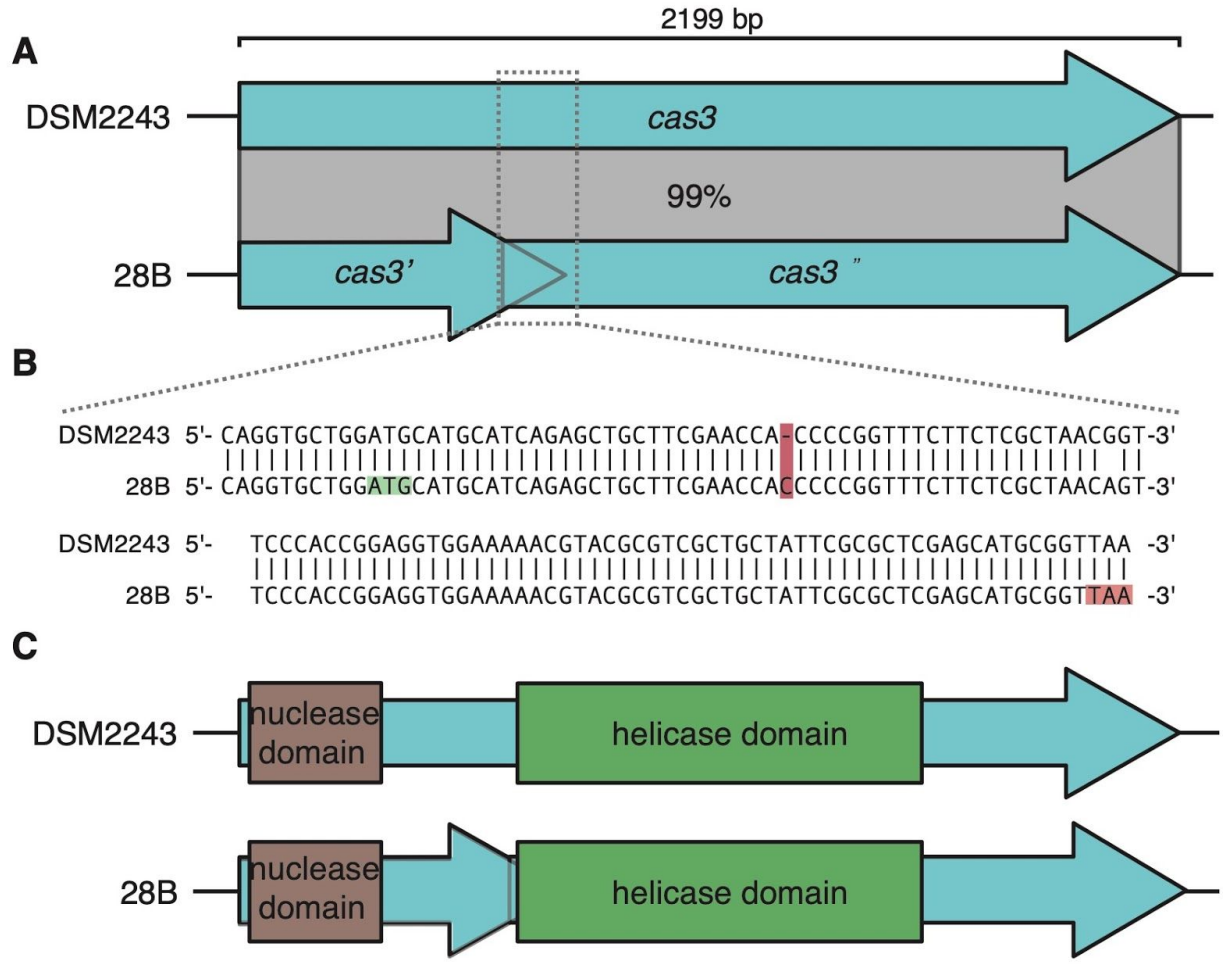


Figure S2. Truncated *cas3* in *E. lenta* 28B related to Figure 3. (**A**) The annotation of *cas3* in *E. lenta* DSM 2243 denotes a 2,199 sequence that comprises a single protein while in strain 28B, two distinct proteins are predicted: cas3' and cas3''. Nucleotide identity between these genes and the *cas3* of DSM 2243 is 99%. (**B**) Sequence alignment reveals an insertion near the start codon of the *cas3''* gene of the 28B strain. The ATG marked in green denotes the start codon of *cas3''* in 28B. The TAA marked in red denotes the stop codon of the *cas3'* gene of 28B. Sanger sequencing was used to verify this insertion. (**C**) Schematic representation of conserved domains in *cas3* reveals that nuclease and helicase domains are divided between *cas3'* and *cas3''* respectively in the strain 28B.

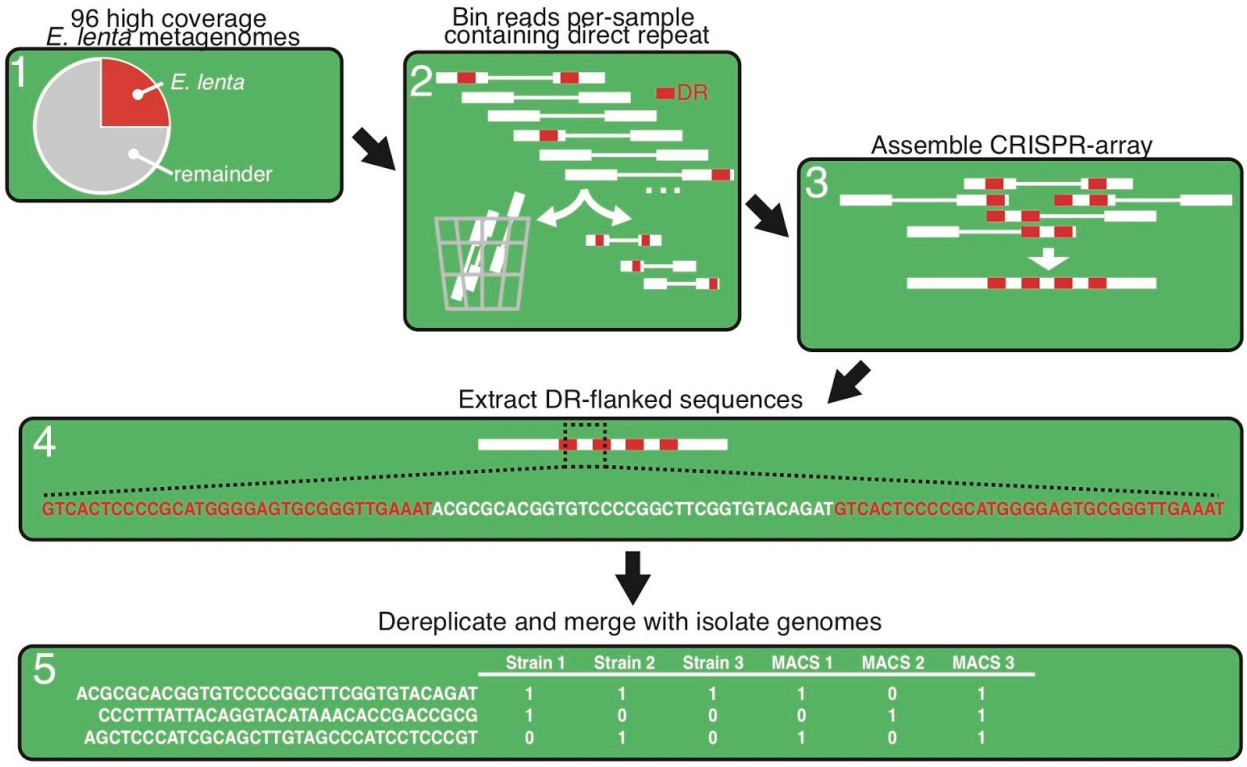


Figure S3. Assembly Strategy for uncovering metagenome-assembled CRISPR spacers relating to Figure 4. (**1**) A set of metagenomes is identified with elevated abundance of organism of interest to increase the probability of array recovery. (**2**) Paired-end reads are screened for at least one instance of the direct repeat (vsearch --usearch_global). (**3**) Resulting binned reads are assembled on a per-sample basis to recover the spacer array. (**4**) Sequences between 25 to 40 nucleotides flanked by an approximation of the direct repeat (≤3 mismatches), are extracted and (**5**) tabulated to examine spacer occurrence across isolates and metagenomes.

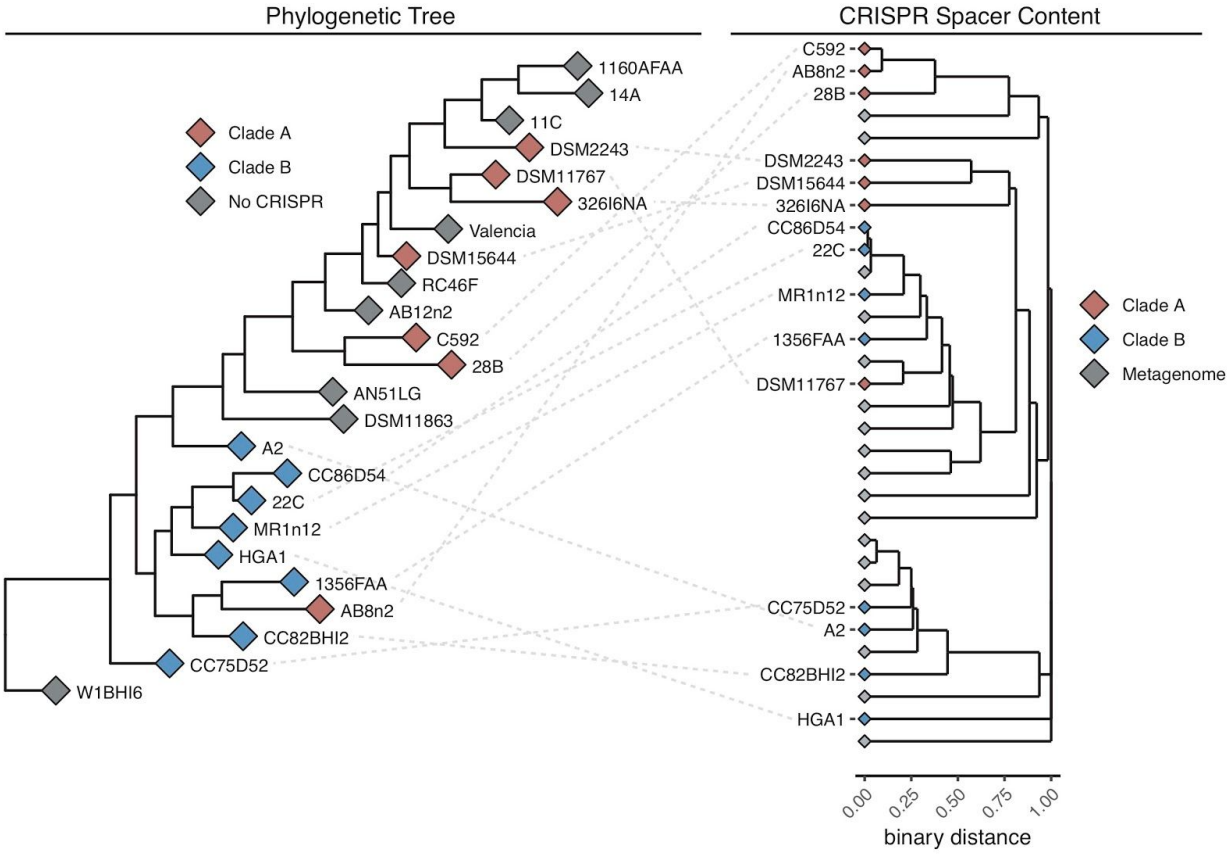


Figure S4. CRISPR spacers correlate with phylogeny related to Figure 4. With the exception of 1 strain for each metric, *cas* gene phylogeny and spacer content are consistent with the whole-genome derived phylogeny.

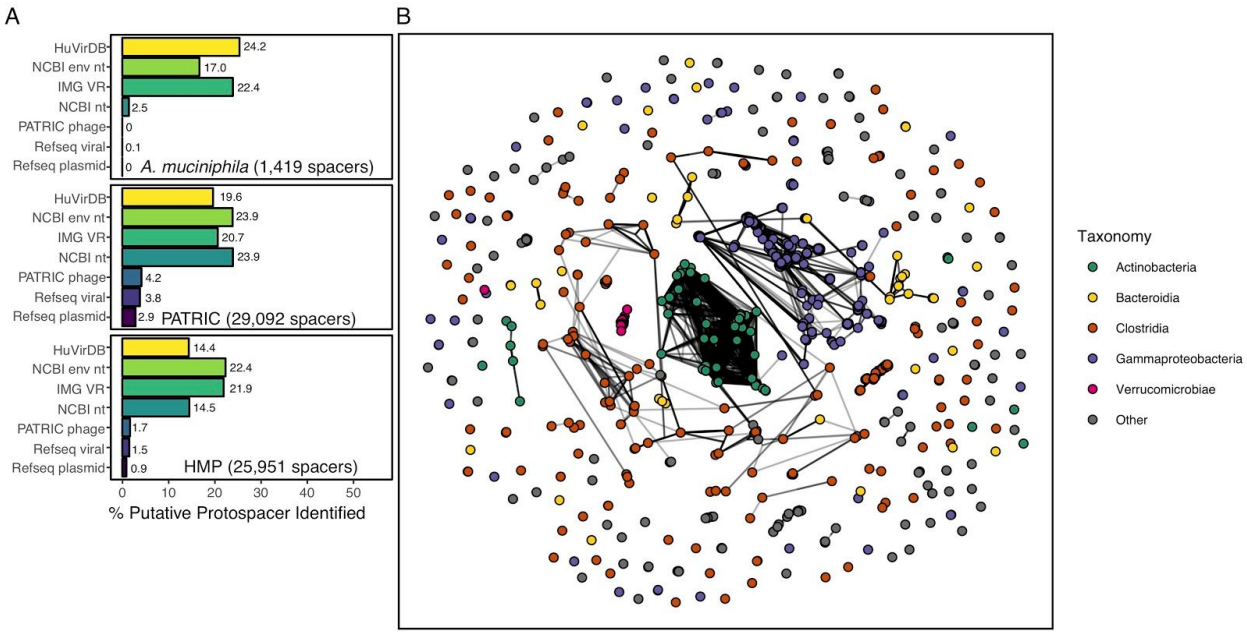


Figure S5. Protospacers across databases relating to Figure 5. (A) Number of identified protospacers as a function of input database demonstrates that highly prevalent, but under-characterized, gut bacteria have improved identification of CRISPR targets in human viromes; for example, *A. muciniphila* (shown here) and *E. lenta* (shown in **Figure 5A**). The numbers in the brackets correspond to the number of spacers extracted from the sum of all genomes in the dataset. **(B)** Network analysis of genomes (linked by shared protospacer targets) reveals that CRISPR targets are conserved within bacterial classes.