Cell Reports

Thousands of small, novel genes predicted in global phage genomes

Graphical abstract

Authors
Brayon J. Fremin, Ami S. Bhatt, Nikos C. Kyripides, Global Phage Small Open Reading Frame (GP-SmORF) Consortium

Correspondence
bfremin@lbl.gov (B.J.F.), nckyripides@lbl.gov (N.C.K.)

In brief
Fremin et al. use comparative genomics to predict more than 40,000 small-gene families in phage from diverse environments. Small genes are approximately 3-fold more prevalent in phage than prokaryotic genomes. This resource includes more than 5,000 anti-CRISPR small-gene families and more than 9,000 secreted or transmembrane small-gene families.

Highlights
- More than 40,000 small gene families predicted in phages from diverse environments
- More than 5,000 small gene families predicted to encode anti-CRISPR proteins
- More than 9,000 small gene families predicted to encode secreted or transmembrane proteins
- Identified novel core phage proteins like baseplate proteins and phage tail proteins

Fremin et al., 2022, Cell Reports 39, 110984
June 21, 2022
https://doi.org/10.1016/j.celrep.2022.110984
Resource

Thousands of small, novel genes predicted in global phage genomes

Brayon J. Fremin,1,2,* Ami S. Bhatt,3,4 Nikos C. Kyprides,1,2,5,* and Global Phage Small Open Reading Frame (GP-SmORF) Consortium

1Department of Energy, Joint Genome Institute, Berkeley, CA, USA
2Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
3Department of Medicine (Hematology; Blood and Marrow Transplantation) and Genetics, Stanford University, Stanford, CA, USA
4Department of Genetics, Stanford University, Stanford, CA, USA
5Lead contact
*Correspondence: bfremin@lbl.gov (B.J.F.), nckyrpides@lbl.gov (N.C.K.)

SUMMARY

Small genes (<150 nucleotides) have been systematically overlooked in phage genomes. We employ a large-scale comparative genomics approach to predict >40,000 small-gene families in ~2.3 million phage genome contigs. We find that small genes in phage genomes are approximately 3-fold more prevalent than in host prokaryotic genomes. Our approach enriches for small genes that are translated in microbiomes, suggesting the small genes identified are coding. More than 9,000 families encode potentially secreted or transmembrane proteins, more than 5,000 families encode predicted anti-CRISPR proteins, and more than 500 families encode predicted antimicrobial proteins. By combining homology and genomic-neighborhood analyses, we reveal substantial novelty and diversity within phage biology, including small phage genes found in multiple host phyla, small genes encoding proteins that play essential roles in host infection, and small genes that share genomic neighborhoods and whose encoded proteins may share related functions.

INTRODUCTION

Viruses infect cells from every domain of life and are the most abundant biological entities on Earth. By no surprise, viruses encode substantial genetic diversity. Metagenomic sequencing recently expanded the known viral diversity (Emerson et al., 2018; Gregory et al., 2019, 2020; Paez-Espino et al., 2016). Several thousand metagenomic samples across various ecosystems have already been sequenced and assembled into contigs, and implementation of various computational tools have predicted that millions of these contigs are viral (Kieft et al., 2020; Ren et al., 2017; Roux et al., 2015).

The first step for understanding the roles that phages play from diverse global ecosystems is to identify their genes and other elements in their genomes. Although substantial progress has been made predicting viral genes, most studies systematically overlook small genes (Duval and Cossart, 2017; Storz et al., 2014; Su et al., 2013). We define such genes as small open reading frames (sORFs) that code for proteins that are fewer than 50 amino acids in length (Garai and Blanc-Potard, 2020; Ramamurthi and Storz, 2014; Storz et al., 2013). We use a large-scale comparative genomics approach to identify small-gene families in phages, the majority of which were novel. We refer to these small-gene families as the Fremin global phage 40K dataset.
For these small-gene families, we provide taxonomic classification for phages and their predicted microbial hosts, ecosystems where the families are found (Ivanova et al., 2010; Mukherjee et al., 2019), protein domains of the encoded small-protein families and proteins near them, predicted anti-CRISPR-encoded proteins, and predicted cellular localizations of the encoded proteins. Additionally, we performed more in-depth analyses by searching for homology between the Fremin gp40K and itself, the Sberro hm4K (Sberro et al., 2019), and the RefSeq non-redundant (nr) database (Pruitt et al., 2007). We additionally determined whether these small genes were co-localized in the genome, which would suggest novel systems of small, encoded proteins. We integrated these results to reveal substantial diversity in small genes and phage biology.

RESULTS

Identification of ~40,000 small-gene families in phage contigs

To predict novel small genes in phages, we first downloaded IMG/VR (Paez-Espino et al., 2017a; Roux et al., 2021), which contains 2,377,994 viral contigs for a combined total of over 48 billion bases of DNA. This database represents a large collection of viral data-sets (Bushman et al., 2019; Espinola et al., 2018; Garcia et al., 2020; Gregory et al., 2019, 2020; Mehrshad et al., 2021; Mobilian et al., 2020; Nayfach et al., 2021a; Paez-Espino et al., 2017b, 2019; Roux et al., 2019; Schulz et al., 2020). From these viral contigs, we predicted all ORFs using MetaProdigal (Hyatt et al., 2010), including those as short as 15 bases (Figure 1A). This resulted in 2,290,724 possible sORFs coding for proteins of fewer than 50 amino acids in length. Using CD-Hit (Li and Godzik, 2006), we clustered these putative sORFs based on at least 50% shared amino acid identity spanning at least 95% of their alignment lengths; this resulted in 633,684 clusters, generating 41,150 small-gene families.

The Fremin gp40K genes encoded proteins that ranged from 12 to 49 amino acids in length (Figure 2A). The number of sequences in each family ranged from 3 to 4,434 (Figure 2B, Table S1). Most families included small genes that were assigned ribosome-binding sites (RBS); nearly 74% of families contained a collection of small genes in which over 60% were assigned a RBS (Figure 2C). The average family size was 21 sequences, and the median was nine sequences. These families were associated with diverse ecosystems (Figure 2D), including marine (23,655 families, 57.5%), freshwater (11,149 families, 27.1%), and digestive system (15,374 families, 37.4%). We identified 16,753 (40.7%) small-gene families in two or more ecosystems and 269 families in five or more ecosystems (Table S1). The frequencies that MetaProdigal predicted putative small genes were 41.5, 34.0, and 56.6 small genes per megabase (Mb) for marine, freshwater, and digestive system contigs, respectively. If only small genes within the Fremin gp40K were counted, the
frequencies were 22.8, 15.8, and 24.3 small genes per Mb for marine, freshwater, and digestive system contigs, respectively. This suggests that the prevalence of predicted small genes varies by ecosystem. For example, we identified approximately 1.5 times as many small genes from digestive system contigs as we did from freshwater contigs when normalized by the number of bases that we predicted for each ecosystem. Overall, we identified tens of thousands of small-gene families in phages across diverse habitats. Additionally, 8,579 of these small-gene families were associated with human hosts (Table S1).

Improved accuracy of small-gene predictions in phages
To determine whether the Fremin gp40K contained more accurate predictions of small-gene families, we examined whether genes in these families’ encoded proteins with known protein domains and had evidence supporting their translation. For each of the 2,290,724 possible sORF-encoded proteins, we used RPS-blast (Altschul et al., 1997) against the CDD (Marchler-Bauer et al., 2005) to annotate protein domains. We found that 1,356 (0.21%) of the 633,684 possible small-gene family-encoded proteins with known protein domains compared with 359 (0.87%) of Fremin gp40K-encoded known protein domains (Figure 3A, Table S1). Importantly, the Fremin gp40K and Sberro hm4K datasets were predicted using an identical pipeline with only one exception: the minimum unique sequences per family was set to 3 for the Fremin gp40K instead of 8 for the Sberro hm4K. Within the Fremin gp40K, 10,749 small-gene families contained at least eight unique sequences, and 158 (1.47%) of these encoded proteins with known protein domains (Table S1).

We then determined whether small genes from both Fremin gp40K and Sberro hm4K (Sberro et al., 2019) could be supported with ribosome-profiling sequence data. Ribosome profiling sequences mRNA transcripts that are associated with ribosomes and thus can be used to identify transcribed genes that are likely to be translated to proteins (Ingolia et al., 2009). Therefore, ribosome profiling serves as an orthogonal approach (Clauwaeert et al., 2019; Ndah et al., 2017) and validation strategy (Durrant and Bhatt, 2021) to aid in small-gene prediction. We used MetaRibo-Seq datasets generated from four metagenomic assemblies of human fecal microbiome samples (Fremin and Bhatt, 2020; Fremin et al., 2020, 2021) that were independent of the IMG/VR (Roux et al., 2021) and the HMP2 (Lloyd-Price et al., 2017) datasets, representing a non-overlapping validation dataset for the two resources. Using MetaProdigal (Hyatt et al., 2010), we predicted 869,737 genes across these four assemblies. We
also predicted possible sORFs along these contigs with modified settings for MetaProdigal (Hyatt et al., 2010). Using BLASTp (Altschul et al., 1997), we identified sORFs that shared similarity to Fremin gp40K and Sberro hm4K. MetaRibo-Seq reads were aligned to these metagenome assemblies, and reads per kilobase million (RPKM) were calculated for all genes, including possible small genes. We found that, relative to all sORFs predicted by MetaProdigal (Hyatt et al., 2010), those that shared homology with either Fremin gp40K or Sberro hm4K were significantly more likely to be translated (RPKM > 0.5; Fisher exact test \( \times 10^{-16} \)), further suggesting that our comparative genomics analysis using RNAcode enriched for translated coding regions (Figure 3B). Even if MetaProdigal was run using default settings (i.e., including encoded proteins between 30 and 49 aa), most of the same predictions were output and the set was similarly depleted in MetaRibo-Seq signal. This suggests that even if MetaProdigal is run with default settings, it performs relatively poorly on encoded proteins below 50 aa in length.

Among all possible sORFs predicted from the MetaRibo-Seq dataset, 1,583 were homologous to Fremin gp40K and 3,841 were homologous to Sberro hm4K. We found that 1,099 of the small proteins predicted in these MetaRibo-Seq-associated assemblies were homologous to both Fremin gp40K and Sberro hm4K, suggesting substantial overlap between the two datasets of small genes (Table S3). Specifically, 407 homologs (25.7%) of Fremin gp40K and 972 homologs (25.3%) of Sberro hm4K were translated compared with 4,388 (11.7%) of all putative small genes predicted by MetaProdigal (Hyatt et al., 2010). Of the 1,583 Fremin gp40K homologs, 802 were homologous to small-gene families with at least eight unique sequences in a family; however, only 205 (25.6%) of these families were translated. Thus, we chose to use a cutoff of three unique sequences per family in the Fremin gp40K because being more conservative did not improve accuracy based on this orthogonal MetaRibo-Seq analysis.

**Novelty within the ~40,000 small-gene families**

To better characterize the overlap between the Fremin gp40K and Sberro hm4K datasets, we used BLASTp (e value \( \leq 0.05 \) and length between 0.9 and 1.1) querying Fremin gp40K representative sequences against Sberro hm4K representative sequences (Figure 4A). We found that 3,344 families from Fremin gp40K (8%) were homologous to 1,961 families (43%) from Sberro hm4K, suggesting that many Sberro hm4K small-gene families contain homologs in phages (Table S4). Of the 359 (57%) small-gene families encoding proteins in the Fremin gp40K with known protein domains, 204 were homologous to families from the Sberro hm4K set. The most common shared protein domain, which is typically found on small phage proteins (Figure 4B). These 154 families included genes that were all homologous to the small genes in the Sberro hm4K. Several known protein domains were identified in proteins encoded by the Fremin gp40K dataset that were not found...
in the Sberro hm4K dataset. For example, pqqa, found typically on small proteins required for coenzyme pyrroloquinoline quinone (PQQ) biosynthesis, was identified in 24 families encoded in the Fremin gp40K with no homologs encoded in the Sberro hm4K. A putative high-light-inducible protein (PHA02337) was encoded in 19 families in the Fremin gp40K with no homologs encoded in the Sberro hm4K. A domain of unknown function (DUF1127) recently characterized to play roles in phosphate and carbon metabolism in *Agrobacterium tumefaciens* (Kraus et al., 2020), was encoded in 14 families in Fremin gp40K with no Sberro hm4K encoded homologs (Figure 4B).

Among the 37,806 small-gene families that were not homologous to Sberro hm4K, 37,651 (99.6%) of encoded proteins could not be assigned a known protein domain, suggesting that of these families were novel. Taxonomically, these small-gene families were difficult to classify. Of the 41,150 small-gene families, 7,180 (17.4%) could not be classified at the kingdom level (Figure 4C). We classified 32,796 small-gene families (79.7%) to the kingdom Heunggongvirae and 1,209 families (2.9%) to Bamfordvirae. All other classifications included less than 1% of small-gene families (Figure 4C).

To test whether encoded small proteins in the Fremin gp40K were homologous to larger proteins, we used BLASTp (Altschul et al., 1997) comparing the encoded small proteins with encoded proteins predicted by MetaProdigal (Hyatt et al., 2010) that were 150 aa or greater. We found that 4,411 encoded small proteins were homologous to larger encoded proteins (Table S4). This could suggest that some of these encoded proteins might contain functions or protein domains also found in larger proteins. However, these 4,411 small genes are likely enriched in false positives; stop codon reassignments and frameshifting are known to occur in phages (Baranov et al., 2001; Ivanova et al., 2014). We found that 2,623 small-gene families shared similarity with larger proteins non-randomly; for example, these families shared homology with only the first half of the larger
proteins or the last half exclusively. We found that 1,370 families always shared either the same start or same stop as those larger proteins. For example, family #52 always shared the same start as larger terminases, which suggests that it is a false positive (Table S4).

In order to identify small-gene families that are likely phage specific as opposed to also common in core host genomes, we first built Hidden Markov models (HMMs) using hmmbuild (Eddy, 2009) of all small-gene families in Fremin gp40K (Data File S1). We predicted possible small genes by using MetaProdigal in IMG/VR and GTDB (Parks et al., 2020), a database that contains 47,894 species clusters including bacteria and archaea. Although prophages exist in the GTDB, they represent a minority of the database and prevalent, phage-specific genes should be strongly enriched in the IMG/VR. Using hmmssearch (Eddy, 2009), we identified which possible small genes in IMG/VR and GTDB were part of the Fremin gp40K. The median enrichment in the IMG/VR compared with the GTDB for these small-gene families was 14-fold. We found that 4,264 of the small-gene families were over 100-fold more likely to be found in the IMG/VR than in the GTDB (Data S1), suggesting that these families are prevalent small-gene families in phages that are less commonly found in core host genomes.

### Small genes are more prevalent in phages than host genomes

We then tested whether small genes were more prevalent in the genomes of phages or their host bacteria. To do this, we first tested how many families in the Sberro hm4K were found in the IMG/VR database (Roux et al., 2021). Using BLASTp of encoded proteins in the Sberro hm4K against the 2,290,724 possible sORFs predicted in the IMG/VR, 2,494 of the 4,539 small-gene families (54.9%) in Sberro hm4K had small-gene homologs, revealing that most of those in Sberro hm4K were found in phages (Table S5). Given this, we hypothesized that phage genomes are much more likely to encode small genes than microbial genomes. To test this hypothesis, we predicted small genes by using MetaProdigal within the GEM (Genomes from Earth’s Microbiomes) dataset (Nayfach et al., 2021b), containing 52,515 metagenome assembled genomes (MAGs). The GEM dataset contained 129,930,639,550 nucleotides, which are mostly associated with prokaryotic genomes, and we predicted 1,975,235 possible sORFs from these genomes in total (15.2 possible sORFs/Mb). Within the GEM dataset, there were 686,959,122 nucleotides predicted to be prophages, which contained 27,678 possible sORFs (40.3 sORFs/Mb). Thus, we were approximately 2.7 times more likely to predict small genes in prophages within the GEM dataset than across all the GEM dataset microbial contigs. The IMG/VR dataset contained 48,566,528,056 nucleotides, and we predicted 2,290,724 possible sORFs (47.2 possible sORFs/Mb), suggesting that sORFs were over 3-fold more likely to be called by MetaProdigal in these IMG/VR phage genomes than in GEM microbial genomes. Together, 116,135 of the small genes predicted from the GEM dataset were homologous to the Sberro hm4K (0.894 sORFs/Mb), while 142,478 small genes predicted from the IMG/VR were homologous to the Sberro hm4K (2.9 sORFs/Mb), overall suggesting that IMG/VR phage sequences were roughly 3.3-fold more likely than GEM microbial sequences to contain small genes from the Sberro hm4K (Table S5).

### Small-gene families potentially involved in host-cell interactions

We explored whether small-gene families in phages encoded proteins that might be secreted by host cells or exposed on the cell surface of bacteria (i.e., transmembrane) and thus would be more likely to be involved in host cell communication. To identify potentially secreted and transmembrane proteins, we used SignalP-5.0 (Almagro Armenteros et al., 2019) and TMHMM (Krogh et al., 2001), respectively, with the requirement that 80% of the members of the family shared the same prediction (Table S1). We found that 9,742 of the 41,150 small-gene families (23.7%) encoded proteins that were predicted to be potentially secreted and/or transmembrane. Specifically, 539 families were predicted to encode potentially secreted proteins only, 8,257 were predicted to encode transmembrane proteins only, and 946 were predicted to encode both potentially secreted and transmembrane proteins (Figure S1). Additionally, we determined which small-gene families encoded proteins with antimicrobial properties. We found that 560 (1.4%) small-gene families could potentially represent novel antimicrobial proteins by using AmPEP (Bhadra et al., 2018). We also found that 15 of these predicted antimicrobial proteins were also predicted to encode potentially secreted proteins, suggesting that these may be viral exotoxins (Figure S2). For example, family #91442 encoded potentially secreted antimicrobial proteins that are found in environmental and plant-associated samples and is found in phages predicted to infect Pseudomonas species. Family #4483 encoded potentially secreted antimicrobial proteins found mostly in freshwater (Figure S2, Table S1). Given that anti-CRISPRs are typically small phage proteins, we used PaCRISPR (Wang et al., 2020) on representative sequences from each family to predict whether these small-gene families encoded anti-CRISPR proteins. We found that 5,419 small-gene families were predicted to encode anti-CRISPR proteins (Table S1) and thus might be involved in counter-defense against CRISPR-Cas systems (Wang et al., 2020). Moreover, we found that 539 small-gene families were, on average, found within 5 kb of 10 or more previously proposed anti-CRISPR proteins or anti-CRISPR-associated proteins in AcrDB, a database of anti-CRISPR operons (Huang et al., 2021). Of these 539 small-gene families, 81 were also predicted to be anti-CRISPR proteins by using PaCRISPR (Table S1).

### Multi-host small-gene families in phages

Host ranges of phages containing these small-gene families were predicted, with a particular focus on small-gene families found in multiple hosts. Within the IMG/VR, contigs are assigned to hosts where applicable (Roux et al., 2021). We defined multi-host small-gene families as those that were found in phage genomes predicted to infect four or more host phyla, suggesting non-clade-specific roles. We underestimated the number of multi-host small-gene families in this work because 74.4% of small-gene families could not be classified to host phyla. Nonetheless, there were 27 small-gene families that were found in phages that infect four or more different host phyla (Figure 5).
Most of these multi-host small-gene families were encoded by phages within the phylum Uroviricota and were found across diverse ecosystems (Figure 5). Four of these small-gene families encoded proteins that were assigned a protein domain. Families #1, #57, and #9 were assigned PHA02324, annotated as a hypothetical protein. Family #12 was assigned PHA02337, annotated as a putative high-light-inducible protein and found in marine, freshwater, soil, and sediment but not the digestive system. Families #12 and #1051 were predicted to be transmembrane proteins (Table S1).

**Homology within the Fremin gp40K**

To characterize homology within the Fremin gp40K, we analyzed all pairwise comparisons among small-gene family-encoded proteins from an all-versus-all BLASTp of the Fremin gp40K. This revealed that 22,998 of the 41,150 families (55.9%) were homologous to at least one other family in the Fremin gp40K. Furthermore, 468 (1.1%) of the families were homologous to five or more other families in the Fremin gp40K. This suggests that the majority in the Fremin gp40K are homologous with at least one other family in the dataset and that these small genes...
have substantially diverged and evolved over time to the extent that they clustered independently in our analysis (Table S4). For reference, 1,989 of the 4,539 families (43.8%) were homologous to at least one other family in the Sberro hm4K (Table S4). Although we identified 27 multi-host small-gene families (Figure 5), all of these families were homologous to at least three other families within the Fremin gp40K and collectively were homologous to 834 families (Table S4). For example, family #27 was homologous to 113 other families in the Fremin gp40K (Figure S3).

Small-gene families involved in phage function
Because family #27 was homologous to the most families in the Fremin gp40K, we inspected the genomic neighborhoods of these families to infer function. We discovered that the genes for all but 3 of these 114 homologous small-gene families were found near T4 baseplate protein domains (Table S7). The baseplate of a bacteriophage T4 controls host recognition, attachment to host, tail sheath contraction, and viral DNA ejection into the host (Arisaka et al., 2016; Taylor et al., 2016). The formation of the baseplate hub is controlled by six genes, gp5, gp27, gp26, gp28, gp29, and gp51. Gene gp51 encodes a protein that functions catalytically to form the dome-shaped baseplate (Snustad, 1968). We used BLASTp to query family #27 against the nr database, and the top hit was putative baseplate assembly catalyst (gp51) from Pelagibacter phage Mosig EXVC030M (Q0169098.1), with a 75% identity and 85% query coverage (Table S6). Of these 111 potentially novel gp51 families near baseplate proteins, two (families #22110 and #41447) were assigned the PHA02078 protein domain (Table S1). Although this is annotated as a hypothetical protein in the CDD, it is also consistently found near baseplate proteins and is homologous to other gp51 families. Overall, the proteins encoded by these 111 homologous families whose genes were located near baseplate proteins had lengths that ranged from 28 to 49 aa and collectively represented 8,505 total and 3,429 unique sequences in IMG/VR (Table S1). Only four of these families, #10156, #79279, #45764, and #8467, did not hit any gp51 sequences upon BLASTp to the nr database (e value >0.05), and over 80% of these hits were to gp51 proteins that were greater than 50 aa in length, suggesting that these small-protein families were especially divergent from previously characterized gp51 sequences (Table S6). As an example of using homology within the Fremin gp40K as well as genomic neighborhood analyses to assign functions to novel small genes, we identified substantial diversity within novel gp51 small-gene families, which encode proteins that are essential for baseplate formation and host infection (Figure 6).

Using homology within the Fremin gp40K, together with homology with the nr database, and genomic neighborhood analysis, we explored the functions of several other small-gene families. We identified 76 small-gene families that encoded proteins homologous to phage tail proteins in the nr database (Table S6). Among these 76 families, 26 were assigned the Phage_P2_GpE protein domain, which is closely related to the gpE phage tail protein. One of these 76 families, family #2109, was homologous to 29 other families within the Fremin gp40K (Table S4). Of these 30 families, 29 contained genes that were found near genes encoding proteins with phage minor tail and tape measure protein domains, which is the expected genomic neighborhood for phage tail proteins (Figure S4). These 29 families were not assigned to known protein domains. We found that only 19 of these 29 phage tail protein families were homologous to known phage tail proteins in the nr database, suggesting that the other 10 families were divergent and novel small-protein families (Table S6).

Although integrating various approaches provides confidence in assigning functions, simply using the homology between Fremin gp40K and the nr database is invaluable to prioritizing small-gene families of interest. We found that 16,352 families shared significant similarity to proteins in the nr database, with 3,981 of these families being homologous to proteins that were not annotated as hypothetical, uncharacterized, or unknown (Table S6). For example, we found that 86 families shared homology to antitoxins, 62 were homologous to peptidases, 12 were homologous to ribosomal proteins, 30 were homologous to stabilization proteins, 8 were homologous to multidrug transporters, 7 were homologous to inhibitory peptide Kil, and 18 were homologous to entericidins (Table S6). Of the 12 families encoding proteins homologous to ribosomal proteins, only two were assigned protein domains, which were ribosomal. Those small-gene families encoding proteins that were homologous to antitoxins are particularly interesting, given that phages have been shown to encode antitoxins to inhibit host toxins and preserve the host (Song and Wood, 2020). Many of the hits were to proteins that were larger than these small-gene families (66% of hits were to proteins greater than 50 aa). In the future, these nr database results should be strengthened using additional lines of evidence.

To identify small-gene families encoding proteins that might directly interact, we determined which families had small genes that were found within 500 bp of other small genes. We found that 10,824 of the small-gene families included genes that were within 500 bp of a small gene from another family at least twice. For example, families #905 and #309 had genes that were typically found near each other and also consistently found near genes encoding proteins with HTH-XRE and Bro-N protein domains. Though it is unclear what roles the proteins encoded by these small-gene families perform, family #905 was predicted to encode potentially secreted proteins (Figure S5). Additionally, families #1753 and #1755 included genes that were typically found near each other and that encoded proteins containing signal peptides. Genes from these families were found near genes encoding proteins with INT выше_семейства и XerC domains. Perhaps these represent novel systems of small proteins containing one or multiple potentially secreted signaling proteins. Genes within these families did not encode proteins with known domains, nor were they homologous to proteins in the nr database; however, they are intriguing based on their co-occurrence with genes from other small-gene families, genomic neighborhood, and encoded proteins with predicted signal peptides (Figure S5).

DISCUSSION
Although small genes play critical roles in phages (Duval and Cossart, 2017), they are difficult to predict accurately and are overlooked systematically as a result. Substantial progress has...
recently been made to predict small-gene families within the human microbiome (Sberro et al., 2019). Our work has now revealed that over one-half of the small genes in the Sberro hm4K, the human microbiome dataset, were homologous to those identified in phages. This perhaps is unsurprising, given the shared functions between hosts and phage, the abundance of phages in the human microbiome (Federici et al., 2021), and the fact that viruses employ several strategies to maintain small genome sizes, one of which is utilization of small genes (DiMaio, 2014). However, substantial diversity of small genes in phages has remained elusive, due both to methodological limitations in identifying small genes from metagenomic datasets and to the lack, until recently, of adequate viral genome diversity captured from metagenomic approaches. In this work, we exploited the largest publicly available resource in viral genomics and interrogated over 2.3 million viral genome contigs uncovered from a large diversity of ecosystems for the presence of small genes (Roux et al., 2021). We employed a comparative-genomics approach to predict 41,150 small-gene families in phages. By observing enrichment in both known protein domains within these encoded small proteins and translation within these 41,150 small-gene families relative to all possible small-gene families, we increased our confidence that our approach indeed enriched for coding regions.

Perhaps some of the most promising small-gene families are those that were predicted to encode proteins that are transmembrane and/or potentially secreted, because these represent proteins that are more likely to be involved in cross talk between phages and microbes, or even between phages and free-living taxa (Moreno-Gámez et al., 2017). We highlighted several of these, including multi-host small-gene families predicted to encode proteins that are transmembrane and potentially novel systems involving one or multiple secreted small proteins. Moreover, we identified 560 small-protein families with predicted antimicrobial properties (Almagro Armenteros et al., 2019), 15 of which were also predicted to be potentially secreted and may act as exotoxins. We found that 5,419 small-protein families were predicted to be anti-CRISPRs, suggesting that thousands of these small proteins may play roles in countering CRISPR-Cas systems (Wang et al., 2020).

It is difficult to fully address tens of thousands of diverse small-gene families thoroughly in a single study, especially given that even less is known about small genes in phages than about those in bacteria. Because taxonomic classifications of these phages are difficult and host classifications are even more challenging, it is premature to estimate whether individual small-gene families are widespread across phages or whether they are more phage specific. Nonetheless, these small-gene families may prove useful in assigning viral taxonomy based on viral-protein families, which has been successful for other gene families (Pons et al., 2021). We identified 27 small-gene families found on phages that infect four or more phyla; however, this was an underestimate of multi-host families. Additionally, we observed substantial homology across these small-gene families within the Fremin gp40K, suggesting that it may be more useful to view families as homologous groups instead of individual families. Following this logic of focusing on groups of homologous small-gene families, much of this work centered around small-gene families that encoded proteins essential for host cell infection, such as small proteins involved in baseplate formation (gp51) and phage tail proteins. These small-gene families were some of the largest and most convincing results supported by our homology analyses and further supported by genomic-neighborhood analyses. In general, a substantial amount of diversity exists across phages (Roux et al., 2021), and, unsurprisingly, our top small-gene family hits in these analyses were

Figure 6. Novel gp51 small-gene families
Cladogram showing 111 proteins encoded by novel small-gene families homologous to family #27. All genes in these families were found near genes encoding other baseplate proteins. Each family contained at least three unique homologs (not shown in the tree).
predicted to serve as essential and core components of phages, displaying a wide diversity of lengths and amino acid similarities.

Limitations of the study
This work has several limitations. First, we did not consider small-gene families that included fewer than three different sequences. Consequently, we ignored small genes either that happened to be rare in phages or that were divergent and distributed across multiple small-gene families. This limitation was especially obvious considering that we ignored 959 small-gene families with known protein domains because they contained fewer than two unique sequences. Second, our comparative-genomics approach likely produced false-positive small-gene families that are difficult to quantify. Although we showed that our comparative-genomics approach significantly enriched for predicted small genes that were actively translated in fecal microbiomes, we still found that a greater proportion of larger genes were being translated. This suggests that we successfully enriched for coding regions, but perhaps we did not predict them as well as larger genes. Third, our prediction of small proteins that are potentially secreted was likely underestimated given the lack of knowledge in signal peptides among phages. Other mechanisms of secretion exist, and proteins without signal peptides can still be secreted (Green and Mecasas, 2016). Additionally, our predictions of transmembrane proteins were likely overestimated, given the overlap between secreted and transmembrane proteins. Signal peptides contain hydrophobic regions that are sometimes mistaken for transmembrane regions (Krogh et al., 2001). Fourth, the phage contigs from which we predicted small-gene families were of variable completeness, which can affect the genomic-neighborhood analyses we performed. Fifth, longer genes undergoing pseudo-readthrough in phages may have resulted in false positives in small-gene predictions. Sixth, small genes within DGR (Nayfach et al., 2021a) systems may result in false positives. Seventh, small-gene families encoded by phages using alternative genetic codes were not represented in this resource. Eighth, stop codon readthrough in phages may have resulted in false positives in which we would have mistaken longer genes for smaller genes. Ninth, host taxonomic assignments are incomplete and biased to prophage and hosts with CRISPR spacer matches, since these are the methods used to assign hosts to phage in IMG/VR.

Follow-up studies are necessary to understand functions of the proteins encoded by these small-gene families as well as to alleviate several of the limitations described above. In the cases of phages where host information was available, follow-up experiments within these hosts would likely be informative. The most translational follow-up work would involve studying the 8,579 small-gene families that were human host associated, exploring their function, and predicting their abilities to interact with human proteins. For example, small genes can be overexpressed in relevant host bacteria as well as in knockdown/knockout experiments to assess function. Other targeted follow-up experiments could involve testing which of the antimicrobial predicted gene families are toxic to hosts and which families encode secreted proteins that affect host expression. Overall, our comparative-genomics approach enriched for tens of thousands of novel, small genes in phages and our “guilt-by-association” approach using several downstream analyses has substantially expanded upon previously unknown and core proteins involved in phage biology.

CONSORTIA
SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.110984.

ACKNOWLEDGMENTS

We would like to thank Stephen Nayfach for help in identifying resources and suggesting ideas to enhance the manuscript. We also thank Heather Maughan for critical reading of the manuscript. We thank Timothy James for feedback on the manuscript. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. Computing costs were also supported via NIH S10 Shared Instrumentation Grant (1S10OD02014101), NIH R01 #1A148623–01, A Simons Foundation Fellowship, and Damon Runyon Clinical Investigator Award to A.S.B. Sample collection costs were supported by NSF grants 1826734 and 1441717 as well as Simons Foundation grant 827839.

AUTHOR CONTRIBUTIONS


DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES


**STAR METHODS**

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deposited data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMG/VR (version 3)</td>
<td>Roux et al. (2021)</td>
<td><a href="https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html">https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html</a></td>
</tr>
<tr>
<td>Raw Sequencing Reads</td>
<td>Fremin et al. (2020)</td>
<td>PRJNA510123</td>
</tr>
<tr>
<td>Software and algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prodigal (version 2.6.3)</td>
<td>Hyatt et al. (2010)</td>
<td><a href="https://github.com/hyattpd/Prodigal">https://github.com/hyattpd/Prodigal</a></td>
</tr>
<tr>
<td>CD-Hit</td>
<td>Fu et al. (2012)</td>
<td><a href="http://weizhong-lab.ucsd.edu/cdhit_suite">http://weizhong-lab.ucsd.edu/cdhit_suite</a></td>
</tr>
<tr>
<td>RNAcode (version 0.3)</td>
<td>Washietti et al. (2011)</td>
<td><a href="https://wash.github.io/macode/">https://wash.github.io/macode/</a></td>
</tr>
<tr>
<td>trim galore (version 0.4.0)</td>
<td></td>
<td><a href="https://github.com/FelixKrueger/TrimGalore">https://github.com/FelixKrueger/TrimGalore</a></td>
</tr>
<tr>
<td>bedtools (version 2.27.1)</td>
<td>Quinlan and Hall (2010)</td>
<td><a href="https://sourceforge.net/projects/bedtools/">https://sourceforge.net/projects/bedtools/</a></td>
</tr>
<tr>
<td>SignalP-5.0</td>
<td>Almagro Armenteros et al., 2019</td>
<td><a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a></td>
</tr>
<tr>
<td>TMHMM (version 2)</td>
<td>Krogh et al. (2001)</td>
<td><a href="http://www.cbs.dtu.dk/services/TMHMM/">http://www.cbs.dtu.dk/services/TMHMM/</a></td>
</tr>
<tr>
<td>PaCRISPR</td>
<td>Wang et al. (2020)</td>
<td><a href="https://pacrispr.erc.monash.edu/">https://pacrispr.erc.monash.edu/</a></td>
</tr>
<tr>
<td>AcrDB</td>
<td>Huang et al. (2021)</td>
<td><a href="https://bcb.unl.edu/AcrDB/">https://bcb.unl.edu/AcrDB/</a></td>
</tr>
<tr>
<td>GEM</td>
<td>Nayfach et al. (2021a, 2021b)</td>
<td><a href="https://portal.nersc.gov/GEM/">https://portal.nersc.gov/GEM/</a></td>
</tr>
</tbody>
</table>

### RESOURCE AVAILABILITY

**Lead contact**

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Brayon Fremin (bfremin@lbl.gov).

**Materials availability**

This study did not generate new material.

**Data and code availability**

- Required data reported in this paper will be shared by the Lead contact upon request.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the Lead contact upon request.

### METHOD DETAILS

**Data download**

All phage contigs used to predict small gene families were downloaded from IMG/VR (Roux et al., 2021). The metadata for these contigs is also publicly available in IMG/VR, which we used to assign phage taxonomy, host taxonomy, and ecosystem information. The GEM database (Nayfach et al., 2021b) can be downloaded from [https://portal.nersc.gov/GEM/genomes/](https://portal.nersc.gov/GEM/genomes/). From supplemental tables, we downloaded representative protein sequences of 4K small proteins from human microbiomes (Sberro et al., 2019). To validate
translation of a subset of small proteins, we downloaded metagenomic and MetaRibo-Seq data from bioproject: PRJNA510123 (Fremin et al., 2020).

Clustering sORFs
Across the 2,377,994 contigs from IMG/VR (Roux et al., 2021), ORFs were predicted using MetaProdigal (Hyatt et al., 2010); however, parameters files were modified to include ORFs as small as 15 bp. We considered only small ORFs (15–150 bp) that contained both a start and stop codon, resulting in a total of 2,290,724 possible sORFs. These sORFs were clustered at a 50% amino acid similarity level using CD-Hit (Li and Godzik, 2006) with the following parameters: -n 2 -p 1 -c 0.5 -d 200 -M 50000 -l 5 -s 0.95 -aL 0.95 –g 1. This generated 633,684 clusters of possible small gene families.

Identifying possible sORF families with comparative genomics
Among the 633,684 possible small gene families, 152,170 families contained at least 3 unique sequences. We applied RNAcode (Washietl et al., 2011) to these 152,170 possible small gene families and 41,150 of these families were assigned a p value of < 0.05 within the correct reading frame. These 41,150 small gene families were represented by 880,213 gene sequences.

Protein domain assignment
The Conserved Domain Database (CDD) (Marchler-Bauer et al., 2005) was downloaded in February 2021. All 2,290,724 possible sORFs were searched against CDD (Marchler-Bauer et al., 2005) using RPS-blast (Altschul et al., 1997). If the e value of a hit was ≤0.01 and at least 80% of the PSSM’s length was covered by the small gene, the hit was considered significant.

Identifying small proteins in other datasets
To determine the overlap between Fremin gp40K small protein families we predict and the Sberro hm4K, we used BLASTp (Altschul et al., 1997) with word size 2. We considered small proteins with an e value ≤0.05 and length between 0.9 and 1.1 of the small protein length. To predict small proteins in the MetaRibo-Seq dataset (Fremin et al., 2020), we first predicted all possible small genes in the metagenomic assemblies using Prodigal (Hyatt et al., 2010) with a 15 bp lower cutoff. To identify homology within the Fremin gp40K, we used BLASTp (Altschul et al., 1997) with word size 2 querying all 40K small proteins against each other in an all-vs-all BLASTp analysis. We retained a hit if its e value was ≤0.05 and length was between 0.9 and 1.1. To identify homologs of the Sberro hm4K in the GEM database, we used MetaProdigal to predict genes along all contigs within the GEM database, then used BLASTp to query these possible small proteins against the Sberro hm4K (e value ≤0.05 and length between 0.9 and 1.1). We similarly performed BLASTp querying the Fremin gp40K against the nr database. We also retained hits if they had an e value ≤0.05 with a maximum number of hits up to 20.

HMM database
Multiple sequence alignments of all 41,150 small gene families were created using MUSCLE (Edgar, 2004) and HMMs for each family were created using hmmbuild from HMMER3 (Eddy, 2009). We searched across two databases, IMG/VR and GTDB release 202 (Parks et al., 2020). The GTDB contained 47,894 species clusters of bacteria and archaea. Possible small genes were predicted from these resources using MetaProdigal. We identified 2,294,433 possible sORFs in GTDB and 2,290,724 possible sORF in IMG/VR; therefore, the databases were of near identical size for this analysis. Hmmsearch (-T 50) was used to identify which of the Fremin gp40K were found among predicted small genes in IMG/VR and GTDB. We calculated the fold enrichment (after adding 1 to all counts) of how many times a small gene family was identified in IMG/VR relative to GTDB.

Visualizations
To create trees to visualize homologs, we used PhyML (Anisimova and Gascuel, 2006; Castresana, 2000; Chevenet et al., 2006; Dereeper et al., 2008; Edgar, 2004; Guindon and Gascuel, 2003). To create alignments for visualization purposes, we used Clustal Omega (Madeira et al., 2019).

MetaRibo-Seq analysis
MetaRibo-Seq reads were trimmed using cutadapt (Martin, 2011) and mapped to associated metagenomic assemblies using bowtie (Langmead et al., 2009). MetaProdigal (Hyatt et al., 2012) was used to predict small genes along these metagenomic assemblies. The number of MetaRibo-Seq reads mapping to each gene was counted using bedtools coverage (Quinlan and Hall, 2010) only if over 70% of the read aligned to the gene and in the appropriate strand orientation. RPKM was calculated based on these counts. Genes containing a MetaRibo-Seq RPKM >0.5 were defined as translated.

Functional analyses
For all small proteins within the Fremin gp40K families, we predicted signal peptides using SignalP-5.0 (Almagro Armenteros et al., 2019) using default parameters in “gram +” and “gram -” mode. We predicted which proteins were transmembrane using TMHMM (Krogh et al., 2001). If more than 80% of the proteins within a family were predicted to contain a signal peptide or transmembrane region, we considered the entire family potentially secreted or transmembrane, respectively. Representative protein sequences
encoded by the Fremin gp40K were assessed for antimicrobial properties using AmPEP (Bhadra et al., 2018) using default settings. Using representative protein sequences of the Fremin gp40K, we predicted anti-CRISPR proteins using ACRhub (Wang et al., 2021), a web server that performed PaCRISPR (Wang et al., 2020, 2021) using default settings. To determine which small gene families were found near anti-CRISPR and anti-CRISPR associated proteins, we used BLASTp of all genes within 5 kb of each small gene family against AcrDB (Huang et al., 2021). Those with e values less than 0.05 were retained. The average number of anti-CRISPR or anti-CRISPR associated proteins within 5 kb were calculated for each family.

Genomic neighborhood analysis
All ORFs were predicted from all IMG/VR (Roux et al., 2021) contigs using MetaProdigal (Hyatt et al., 2010) with default settings. If a gene was found within 5 kb of a predicted small gene on a contig, we extracted each gene’s predicted amino acid sequence. We performed RPS-BLAST (Altschul et al., 1997) against CDD (Marchler-Bauer et al., 2005) on these amino acid sequences. We considered hits with e values less than 0.01 and alignments containing at least 80% of the PSSM’s length.

QUANTIFICATION AND STATISTICAL ANALYSIS
In Figure 2, the numbers of small gene families were quantified in terms of amino acid length, number of sequences in families, percent of members in family with RBS, and number of families found in various ecosystems.
In Figure 3A, the percentage of small gene families with protein domains were quantified. Differences between groups were determined using Fisher’s exact test.
In Figure 3B, the percentage of small gene families with MetaRibo-Seq signal were quantified. Differences between groups were determined using Fisher’s exact test.
In Figure S3, the number of small gene families that share homology to other small gene families were quantified.
In Figure 4, the number of small gene families that were assigned protein domains and taxonomically classified were quantified.
In Figure 5, the number of small gene families found in multiple host phyla were quantified by taxonomy and ecosystem.
In Figure S5, the number of times small gene families occur next to one another was quantified. Hypergeometric test were used to determine if small genes found near other small genes were occurring at a frequency greater than random chance.