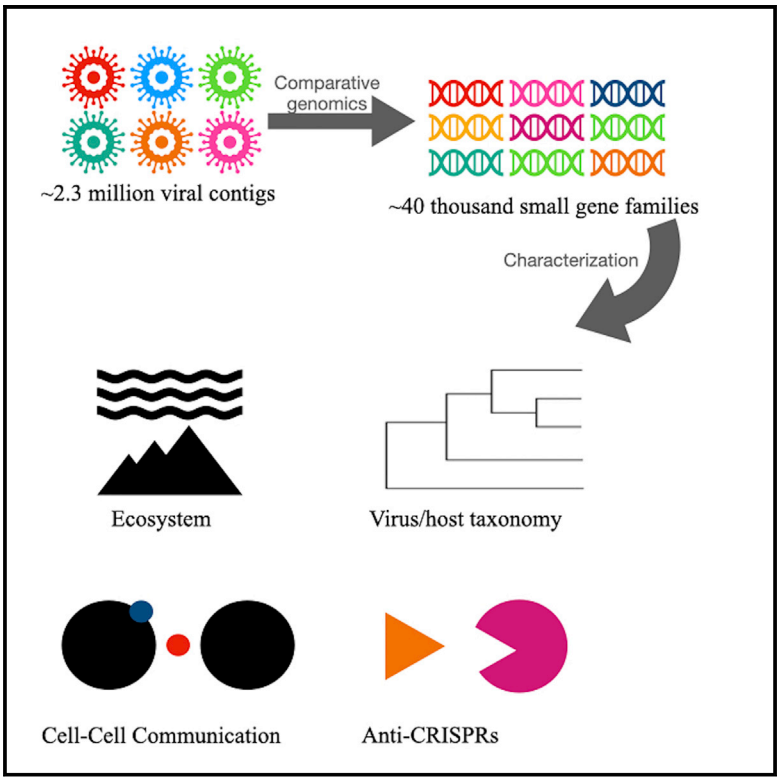
Cell Reports

Thousands of small, novel genes predicted in global phage genomes

Graphical abstract



Authors

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

Correspondence

bfremin@lbl.gov (B.J.F.), nckyrpides@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

Cell Reports

CellPress
OPEN ACCESS

Resource

Thousands of small, novel genes predicted in global phage genomes

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

[1]Department of Energy, Joint Genome Institute, Berkeley, CA, USA
[2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]Department of Medicine (Hematology; Blood and Marrow Transplantation) and Genetics, Stanford University, Stanford, CA, USA
[4]Department of Genetics, Stanford University, Stanford, CA, USA
[5]Lead contact
*Correspondence: bfremin@lbl.gov (B.J.F.), nckyrpides@lbl.gov (N.C.K.)
https://doi.org/10.1016/j.celrep.2022.110984

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., 2014). Small phage genes play diverse biological roles (Duval and Cossart, 2017). For example, bacteriophages can encode quorum-sensing systems that are similar to those of bacteria. In one case, a bacteriophage-encoded small secreted protein, AimP (43 aa), promotes host lysogeny by binding to a receptor protein, AimR (Erez et al., 2017). The underlying challenge with small-gene prediction is that in-frame start and stop codons often occur near one another by chance; thus, it is challenging to determine which subset of these possible sORFs represents true coding regions. Gene prediction tools typically set minimum ORF length thresholds by default because they will otherwise inaccurately predict these sORFs (Hyatt et al., 2010).

One way to improve the accuracy of sORF predictions is to use comparative genomics. Possible sORFs can be clustered on the basis of amino acid similarity of their encoded proteins. The variation among the small-gene homologs within these clusters can be evaluated for evolutionary signatures. For example, synonymous and conservative mutations within a small-gene family supports that the family is coding. This concept was previously applied at a large-scale to human microbiomes to predict 4,539 small-gene families, the majority of which were novel (Sberro et al., 2019); herein, we refer to this dataset as the Sberro human microbiome 4K ("Sberro hm4K"). This analysis in human microbiomes revealed a diversity of previously overlooked genes, including those that were horizontally transferred or that encoded small proteins essential for housekeeping, cross talk, or phage defense. With the wealth of viral genome data currently available (Roux et al., 2021), a similar approach can be extended to phages at a large scale from diverse global ecosystems.

In this work, we employed comparative analysis on 2.3 million phage genome contigs available through the IMG/VR v3 resource (Roux et al., 2021) to reveal 41,150 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

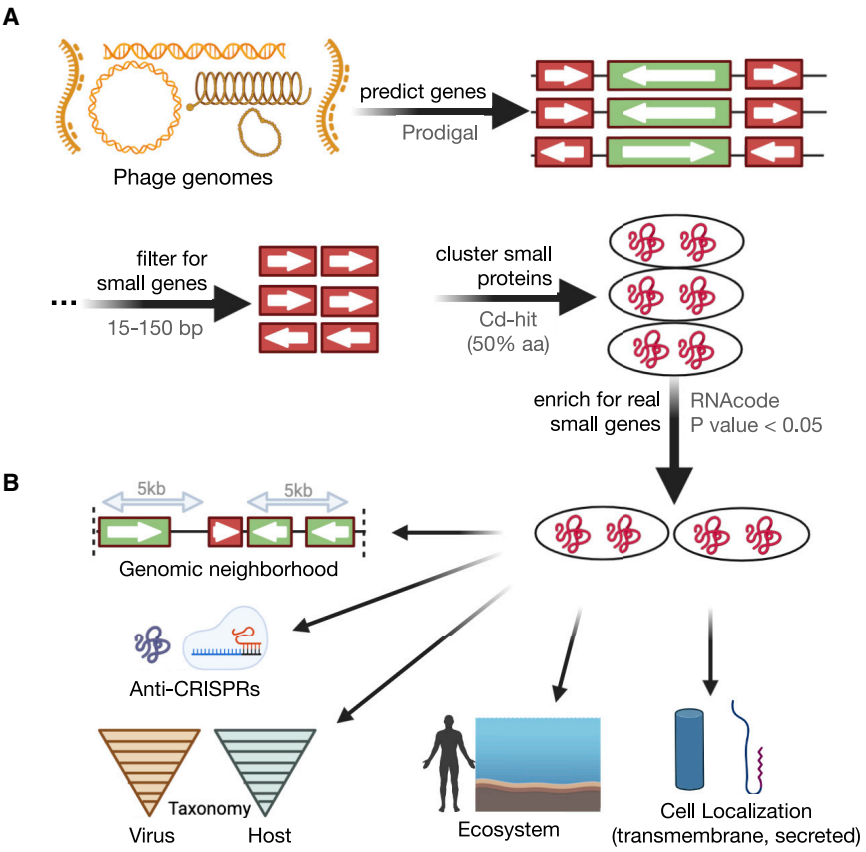


Figure 1. Pipeline to identify and characterize small genes in phages

(A) Identifying small genes in phages: 2,377,994 phage contigs were annotated using MetaProdigal, with a lower gene length cutoff of 15 bp. Proteins encoded by these small genes were clustered at 50% aa identity using CD-Hit. A comparative-genomics approach using RNAcode was applied to the resulting 633,684 clusters, generating 41,150 small-gene families.

(B) Characterizing small genes in phages. Several analyses were performed on these 41,150 small-gene families, including genomic-neighborhood analysis, prediction of anti-CRISPRs, taxonomic classification of both viruses and possible hosts containing these small genes, and prediction of cellular localization of proteins encoded by small genes.

("Fremin gp40K"). 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 datasets (Bushman et al., 2019; Espínola 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 or families of small genes. Using RNAcode (Washietl et al., 2011), a comparative genomics approach that predicts the likelihood that aligned genomic regions are coding, we filtered this set of 633,684 clusters to 41,150 higher-confidence small-gene families (herein referred to as the Fremin gp40K families). Specifically, the 41,150 small-gene families all contained at least three sequences and were assigned RNAcode p values less than 0.05 in the expected (i.e., first) reading frame. These small-gene families and encoded proteins were thoroughly characterized in terms of phage taxonomy, host taxonomy, protein domains, genomic neighborhood, and ecosystem of origin (Figure 1B, Table S1 and S2).

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

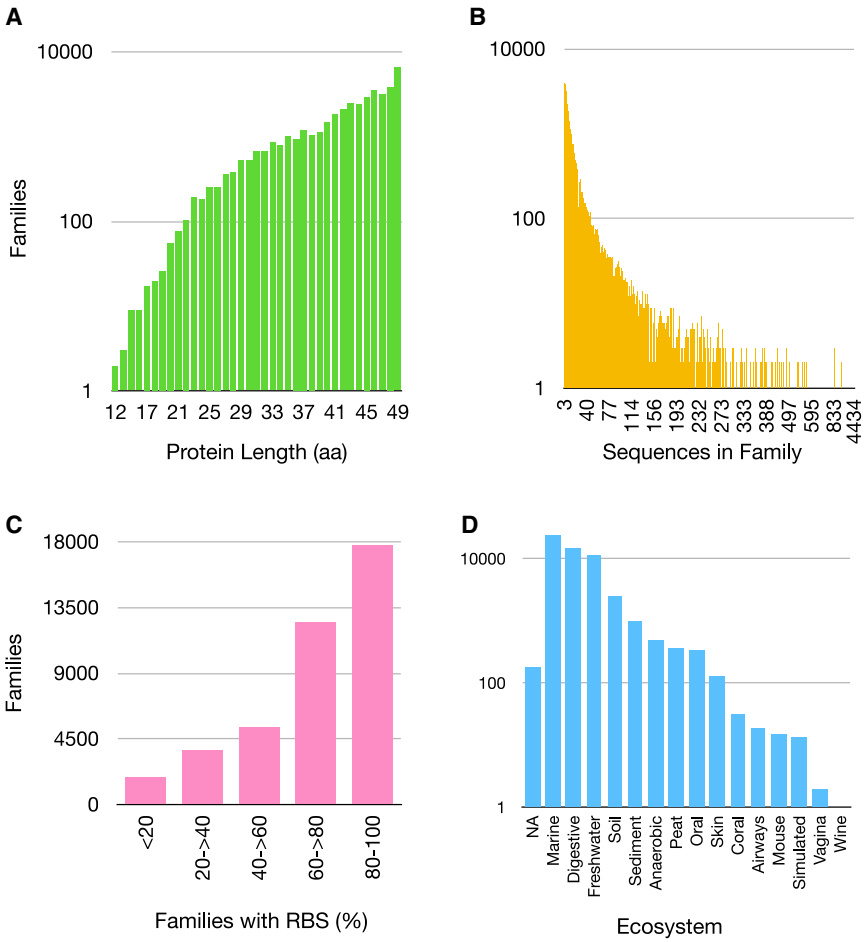


Figure 2. Summary statistics on the Fremin gp40K

(A) Histogram showing the distribution of protein lengths among families in the Fremin gp40K.

(B) Histogram displaying the number of sequences present in each family in the Fremin gp40K.

(C) Histogram showing the number of families in which members were assigned ribosome-binding sites.

(D) Histogram displaying the number of families found in specific ecosystems.

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), indicating a significant enrichment of known protein domains in the Fremin gp40K compared with all possible small-gene families (Fisher exact test $p < 2.2 \times 10^{-16}$). This enrichment was also observed in a previous study (Sberro et al., 2019), which identified small proteins in human microbiomes (Figure 3A). Notably, of the 997 possible small-gene families encoding proteins that contained known domains but were not included in the Fremin gp40K, only 38 (3.8%) contained more than two unique sequences, suggesting that most of these small-gene families contained insufficient numbers of members to be retained following our comparative genomics analysis. Moreover, only 0.87% of Fremin gp40K gene families contained known protein domains compared with the 4.5% in the Sberro hm4K, suggesting that small proteins in phages are an especially unknown and novel sequence space to explore (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 (Clauwaert et al., 2019; Ndah et al., 2017) and validation strategy (Durrant and Bhatt, 2021) to aid in small-gene prediction. We used Meta-Ribo-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

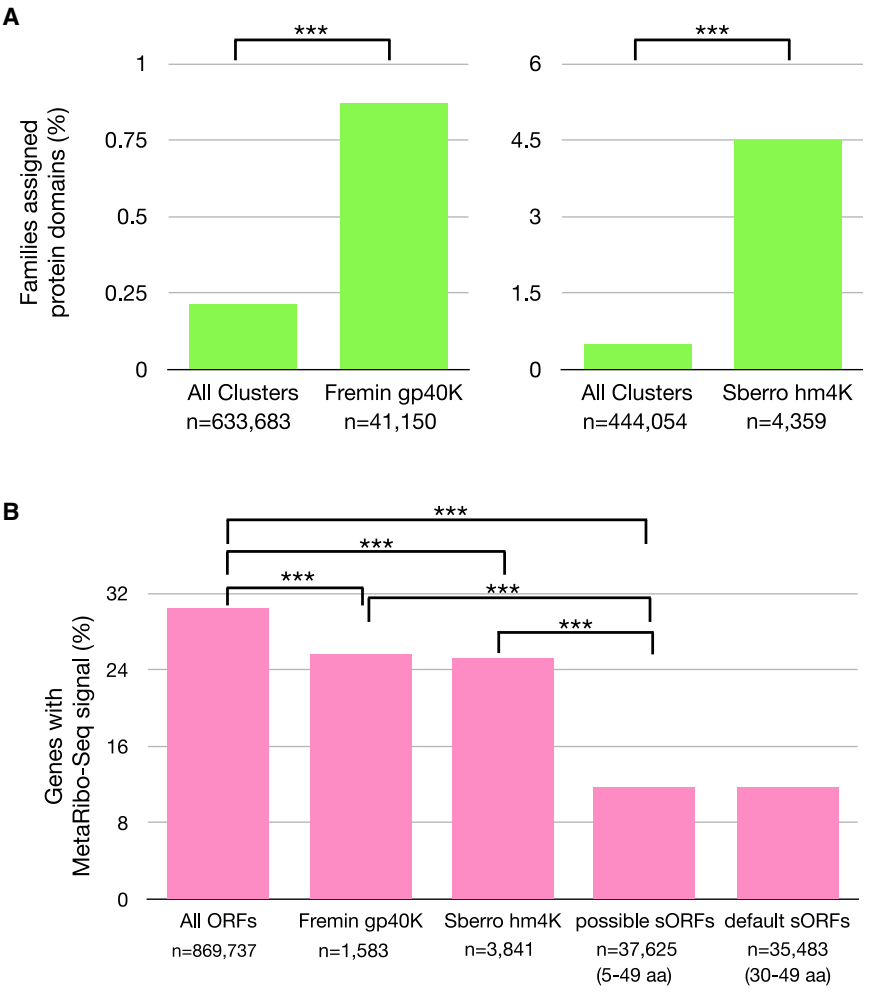


Figure 3. Comparative genomics enriches for real small genes

(A) Enriching for small genes encoding proteins with known protein domains. The bar plot shows the percentage of small-gene clusters encoding proteins that contain known protein domains, including all phage clusters (n = 633,683), the Fremin gp40K (n = 41,150), all human microbiome clusters from Sberro et al. (2019) (n = 444,054), and the Sberro hm4K (n = 4,539).

(B) Enriching for small genes that are translated in human microbiomes. The bar plot shows the percentage of genes with a MetaRibo-Seq signal, including all genes (annotated using default MetaProdigal), small genes homologous to the Fremin gp40K, small genes homologous to the Sberro hm4K, and all small genes. Fisher's exact test was used to compare between groups. (***$p < 0.0001$).

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 $p < 2.2 \times 10^{-16}$), 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 ≤ 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, also present across 154 families in the Fremin gp40K, was Phage_XkdX, 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

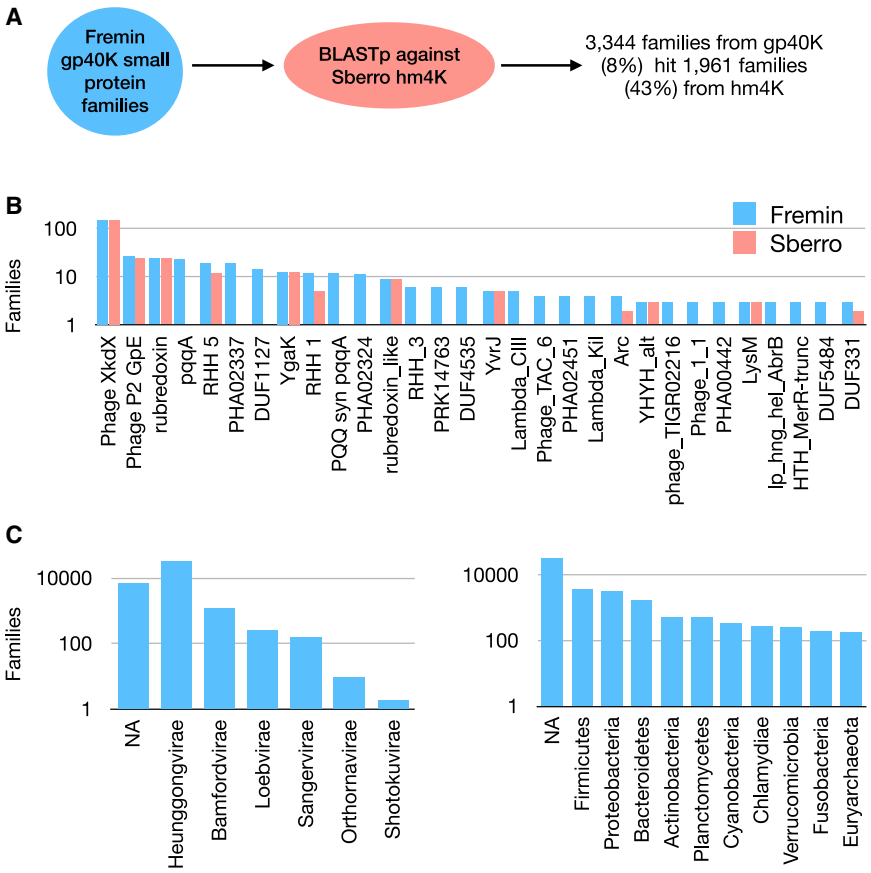


Figure 4. Comparing the Fremin gp40K with the Sberro hm4K

(A) Overlap of Fremin gp40K and Sberro hm4K datasets. The flowchart displays the use of BLASTp to determine that 3,344 of the 40K families were homologous to 1,961 of the Sberro hm4K families.

(B) Families encoding proteins with known protein domains. The histogram shows the number of families in the 40K encoding proteins that were annotated with specific protein domains and which of those were homologous to the Sberro hm4K for the top 30 most commonly assigned protein domains.

(C) Taxonomy of the Fremin gp40K. The histogram shows the number of families that were classified at various taxonomic levels and the taxonomic classifications of predicted hosts of families. Families with no taxonomic assignment were classified as "NA."

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). Of the 33,970 small-gene families that were classified to a kingdom, only 506 were classified to more than two kingdoms (Table S1). Only three small-gene families, families #0, #16208, and #67, were classified to more than three kingdoms (Table S1). Host classification of the viral genomes containing the Fremin gp40K gene families was more challenging, with the predicted hosts for 30,610 small-gene families (74.4% of total) lacking classification at the phylum level. Because viral genomes can be connected to multiple hosts, they may be counted multiple times in host classification. The most common classified hosts were Firmicutes, Proteobacteria, and Bacteroides, at 3,699 (9.0%), 3,113 (7.6%), and 1,590 (3.9%) small-gene families, respectively (Figure 4C).

To determine whether we can directly detect proteins encoded by small genes, we inspected a previously generated dataset that extracted small proteins and performed proteomics on *Bacteroides thetaiotaomicron* (Sberro et al., 2019). MetaProdigal predicted 35 possible sORFs in *B. thetaiotaomicron*. Four of these small genes encoded proteins that were detected by mass spectrometry in this dataset. By use of BLASTp, three of these four detected proteins were homologous to encoded proteins in the Sberro hm4K set, including a predicted novel ribosomal protein (Sberro et al., 2019). Interestingly, all four of these small proteins were homologous to encoded proteins in the Fremin gp40K (Table S4), suggesting that homologs of all four families are detectable at the protein level.

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 hmmsearch (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 were 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 families 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-CRISPRs 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).

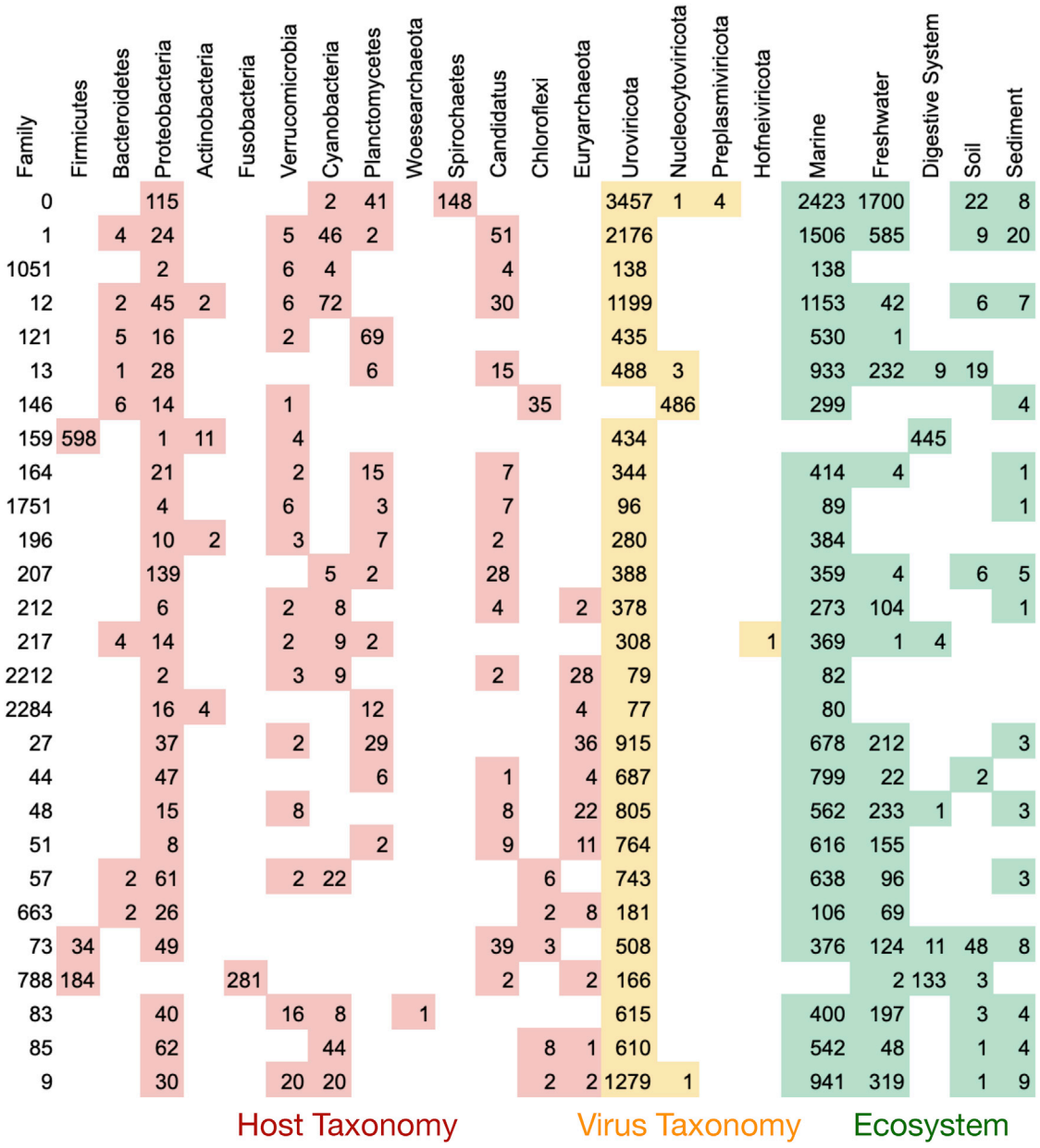


Figure 5. Multi-host small-gene families

Homology between multi-host families and the Fremin gp40K. Visual representing homology and ecosystem metadata between the multi-host small-gene families and other small-gene families within the Fremin gp40K. We indicate the number of small genes in each family that belongs to a specific taxa or ecosystem.

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 hub assembly catalyst (gp51) from Pelagibacter phage Mosig EXVCO30M (QOI69098.1), 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 share 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_ICEBs1_C_like and 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

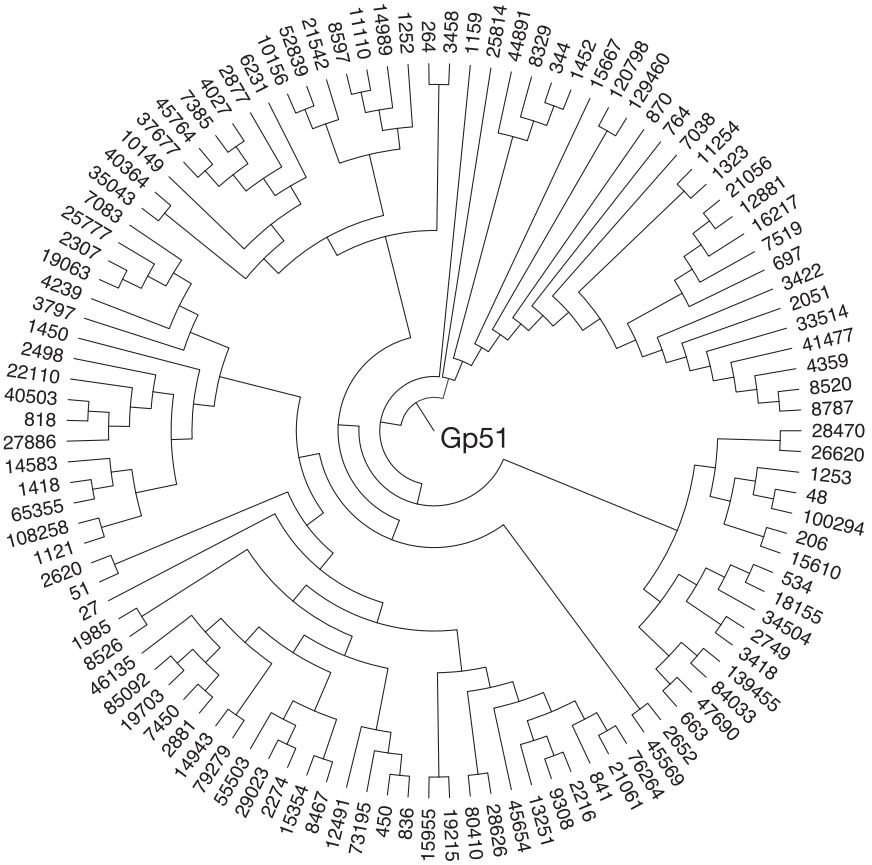


Figure 6. Novel gp51 small-gene families

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).

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 counteracting 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

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 Mecsas, 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 pseudogenization could potentially have resulted in false-positive 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

Members of the Global Phage Small Open Reading Frame (GP-SmORF) Consortia are Aditi Sengupta, Alexander Sczyrba, Aline Maria da Silva, Alison Buchan, Amelie Gaudin, Andreas Brune, Ann M. Hirsch, Anthony Neumann, Ashley Shade, Axel Visel, Barbara Campbell, Brett Baker, Brian P. Hedlund, Byron C. Crump, Cameron Currie, Charlene Kelly, Chris Craft, Christina Hazard, Christopher Francis, Christopher W. Schadt, Colin Averill, Courtney Mobilian, Dan Buckley, Dana Hunt, Daniel Noguera, David Beck, David L. Valentine, David Walsh, Dawn Sumner, Despoina Lymperopoulou, Devaki Bhaya, Donald A. Bryant, Elise Morrison, Eoin Brodie, Erica Young, Erik Lilleskov, Eva Högfors-Rönnholm, Feng Chen, Frank Stewart, Graeme W. Nicol, Hanno Teeling, Harry R. Beller, Hebe Dionisi, Hui-Ling Liao, J. Michael Beman, James Stegen, James Tiedje, Janet Jansson, Jean VanderGheynst, Jeanette Norton, Jeff Dangl, Jeffrey Blanchard, Jennifer Bowen, Jennifer Macalady, Jennifer Pett-Ridge, Jeremy Rich, Jérôme P. Payet, John D. Gladden, Jonathan D. Raff, Jonathan L. Klassen, Jonathan Tarn, Josh Neufeld, Kelly Gravuer, Kirsten Hofmockel, Ko-Hsuan Chen, Konstantinos Konstantinidis, Kristen M DeAngelis, Laila P. Partida-Martinez, Laura Meredith, Ludmila Chistoserdova, Mary Ann Moran, Matthew Scarborough, Matthew Schrenk, Matthew Sullivan, Maude David, Michelle A. O'Malley, Monica Medina, Mussie Habteselassie, Nicholas D. Ward, Nicole Pietrasiak, Olivia U. Mason, Patrick O. Sorensen, Paulina Estrada de los Santos, Petr Baldrian, R. Michael McKay, Rachel Simister, Ramunas Stepanauskas, Rebecca Neumann, Rex Malmstrom, Ricardo Cavicchioli, Robert Kelly, Roland Hatzenpichler, Roman Stocker, Rose Ann Cattolico, Ryan Ziels, Rytas Vilgalys, Sara Blumer-Schuette, Sean Crowe, Simon Roux, Steven Hallam, Steven Lindow, Susan H. Brawley, Susannah Tringe, Tanja Woyke, Thea Whitman, Thomas Bianchi, Thomas Mock, Timothy Donohue, Timothy Y. James, Udaya C. Kalluri, Ulas Karaoz, Vincent Denef, Wen-Tso Liu, William Whitman, and Yang Ouyang.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data download
  - Clustering sORFs
  - Identifying possible sORF families with comparative genomics
  - Protein domain assignment
  - Identifying small proteins in other datasets
  - HMM database

- Visualizations
- MetaRibo-Seq analysis
- Functional analyses
- Genomic neighborhood analysis

- QUANTIFICATION AND STATISTICAL ANALYSIS

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 #AI148623–01, A Sloan 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

Conceptualization, B.J.F. and N.C.K.; Methodology, B.J.F. and N.C.K.; Software, B.J.F.; Formal analysis, B.J.F. Investigation, B.J.F., A.S.B., and N.C.K.; Resources, B.J.F., G.C., A.S.B., and N.C.K.; Data curation, B.J.F.; Writing – original draft, B.J.F. and N.C.K.; Writing – reviewing & editing, B.J.F., A.S.B., and N.C.K.; Visualization, B.J.F.; Supervision, A.S.B. and N.C.K.; Project administration, B.J.F. and N.C.K.; Funding acquisition, A.S.B. and N.C.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 15, 2021
Revised: February 14, 2022
Accepted: May 27, 2022
Published: June 21, 2022

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. *55*, 539–552. https://doi.org/10.1080/10635150600755453.

Arisaka, F., Yap, M.L., Kanamaru, S., and Rossmann, M.G. (2016). Molecular assembly and structure of the bacteriophage T4 tail. Biophys. Rev. *8*, 385–396. https://doi.org/10.1007/s12551-016-0230-x.

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. *37*, 420–423. https://doi.org/10.1038/s41587-019-0036-z.

Baranov, P.V., Gurvich, O.L., Fayet, O., Prère, M.F., Miller, W.A., Gesteland, R.F., Atkins, J.F., and Giddings, M.C. (2001). RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucleic Acids Res. *29*, 264–267. https://doi.org/10.1093/nar/29.1.264.

Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S.W.I. (2018). AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci. Rep. *8*, 1697. https://doi.org/10.1038/s41598-018-19752-w.

Bushman, T.J., Akob, D.M., Bohu, T., Beyer, A., Woyke, T., Shapiro, N., Lapidus, A., Klenk, H.-P., and Küsel, K. (2019). Draft genome sequence of Mn(II)-Oxidizing bacterium *Oxalobacteraceae* sp. Strain AB_14. Microbiol. Resour. Announc. *8*, e01024-19. https://doi.org/10.1128/mra.01024-19.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Chevenet, F., Brun, C., Bañuls, A.L., Jacq, B., and Christen, R. (2006). TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinf. *7*, 439. https://doi.org/10.1186/1471-2105-7-439.

Clauwaert, J., Menschaert, G., and Waegeman, W. (2019). DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. Nucleic Acids Res. *47*, e36. https://doi.org/10.1093/nar/gkz061.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., et al. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. *36*, W465–W469. https://doi.org/10.1093/nar/gkn180.

DiMaio, D. (2014). Viral miniproteins. Annu. Rev. Microbiol. *68*, 21–43. https://doi.org/10.1146/annurev-micro-091313-103727.

Durrant, M.G., and Bhatt, A.S. (2021). Automated prediction and annotation of small open reading frames in microbial genomes. Cell Host Microbe. *29*, 121–131.e4. https://doi.org/10.1016/j.chom.2020.11.002.

Duval, M., and Cossart, P. (2017). Small bacterial and phagic proteins: an updated view on a rapidly moving field. Curr. Opin. Microbiol. *39*, 81–88. https://doi.org/10.1016/j.mib.2017.09.010.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform *23*, 205–211.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., Singleton, C.M., Solden, L.M., Naas, A.E., Boyd, J.A., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. Nat. Microbiol. *3*, 870–880. https://doi.org/10.1038/s41564-018-0190-y.

Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., et al. (2017). Communication between viruses guides lysis–lysogeny decisions. Nature *541*, 488–493. https://doi.org/10.1038/nature21049.

Espínola, F., Dionisi, H.M., Borglin, S., Brislawn, C.J., Jansson, J.K., Mac Cormack, W.P., Carroll, J., Sjöling, S., and Lozada, M. (2018). Metagenomic analysis of Subtidal sediments from polar and Subpolar coastal environments highlights the relevance of anaerobic hydrocarbon degradation processes. Microb. Ecol. *75*, 123–139. https://doi.org/10.1007/s00248-017-1028-5.

Federici, S., Nobs, S.P., and Elinav, E. (2021). Phages and their potential to modulate the microbiome and immunity. Cell. Mol. Immunol. *18*, 889–904. https://doi.org/10.1038/s41423-020-00532-4.

Fremin, B.J., and Bhatt, A.S. (2020). Structured RNA contaminants in bacterial ribo-Seq. mSphere *5*, e00855-20. https://doi.org/10.1128/msphere.00855-20.

Fremin, B.J., Sberro, H., and Bhatt, A.S. (2020). MetaRibo-Seq measures translation in microbiomes. Nat. Commun. *11*, 3268. https://doi.org/10.1038/s41467-020-17081-z.

Fremin, B.J., Nicolaou, C., and Bhatt, A.S. (2021). Simultaneous ribosome profiling of hundreds of microbes from the human microbiome. Nat. Protoc. *16*, 4676–4691. https://doi.org/10.1038/s41596-021-00592-4.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Garai, P., and Blanc-Potard, A. (2020). Uncovering small membrane proteins in pathogenic bacteria: regulatory functions and therapeutic potential. Mol. Microbiol. *114*, 710–720. https://doi.org/10.1111/mmi.14564.

Garcia, M.O., Templer, P.H., Sorensen, P.O., Sanders-DeMott, R., Groffman, P.M., and Bhatnagar, J.M. (2020). Soil microbes trade-Off biogeochemical cycling for stress tolerance traits in response to year-round climate change. Front. Microbiol. *11*, 616. https://doi.org/10.3389/fmicb.2020.00616.

Green, E.R., and Mecsas, J. (2016). Bacterial secretion systems: an overview. Microbiol. Spectr. *4*, 213–239. https://doi.org/10.1128/microbiolspec.vmbf-0012-2015.

Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from Pole to Pole. Cell *177*, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

Gregory, A.C., Zablocki, O., Zayed, A.A., Howell, A., Bolduc, B., and Sullivan, M.B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe *28*, 724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology *59*, 307–321.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704. https://doi.org/10.1080/10635150390235520.

Huang, L., Yang, B., Yi, H., Asif, A., Wang, J., Lithgow, T., Zhang, H., Minhas, F.U.A.A., and Yin, Y. (2021). AcrDB: a database of anti-CRISPR operons in prokaryotes and viruses. Nucleic Acids Res. *49*, D622–D629. https://doi.org/10.1093/nar/gkaa857.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

Hyatt, D., LoCascio, P.F., Hauser, L.J., and Uberbacher, E.C. (2012). Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics *28*, 2223–2230. https://doi.org/10.1093/bioinformatics/bts429.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223. https://doi.org/10.1126/science.1168978.

Ivanova, N., Tringe, S.G., Liolios, K., Liu, W.-T., Morrison, N., Hugenholtz, P., and Kyrpides, N.C. (2010). A call for standardized classification of metagenome projects. Environ. Microbiol. *12*, 1803–1805. https://doi.org/10.1111/j.1462-2920.2010.02270.x.

Ivanova, N.N., Schwientek, P., Tripp, H.J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N.C., and Rubin, E.M. (2014). Stop codon reassignments in the wild. Science *344*, 909–913. https://doi.org/10.1126/science.1250691.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90. https://doi.org/10.1186/s40168-020-00867-0.

Kraus, A., Weskamp, M., Zierles, J., Balzer, M., Busch, R., Eisfeld, J., Lambertz, J., Nowaczyk, M.M., and Narberhaus, F. (2020). Arginine-rich small proteins with a domain of unknown function, DUF1127, play a role in phosphate and carbon metabolism of Agrobacterium tumefaciens. J. Bacteriol. *202*, e00309-20. https://doi.org/10.1128/jb.00309-20.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J. Mol. Biol. *305*, 567–580. https://doi.org/10.1006/jmbi.2000.4315.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., et al. (2017). Strains, functions and dynamics in the expanded human microbiome Project. Nature *550*, 61–66. https://doi.org/10.1038/nature23889.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. *47*, W636–W641. https://doi.org/10.1093/nar/gkz268.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., et al. (2005). CDD: a conserved domain database for protein classification. Nucleic Acids Res. *33*, D192–D196. https://doi.org/10.1093/nar/gki069.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res *39*, D225–D229.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10. https://doi.org/10.14806/ej.17.1.200.

Mehrshad, M., Lopez-Fernandez, M., Sundh, J., Bell, E., Simone, D., Buck, M., Bernier-Latmani, R., Bertilsson, S., and Dopson, M. (2021). Energy Efficiency and Biological Interactions Define the Core Microbiome of Deep Oligotrophic Groundwater. Nat. Commun. *12*, 4253.

Mobilian, C., Wisnoski, N.I., Lennon, J.T., Alber, M., Widney, S., and Craft, C.B. (2020). Differential effects of press vs. pulse seawater intrusion on microbial communities of a tidal freshwater marsh. Limnol. Oceanogr. Lett., lol2.10171. https://doi.org/10.1002/lol2.10171.

Moreno-Gámez, S., Sorg, R.A., Domenech, A., Kjos, M., Weissing, F.J., van Doorn, G.S., and Veening, J.-W. (2017). Quorum sensing integrates environmental cues, cell density and cell history to control bacterial competence. Nat. Commun. *8*, 854. https://doi.org/10.1038/s41467-017-00903-y.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H.Y., Mojica, A., Chen, I.-M.A., Kyrpides, N.C., and Reddy, T. (2019). Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res. *47*, D649–D659. https://doi.org/10.1093/nar/gky977.

Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., and Kyrpides, N.C. (2021a). Metagenomic compendium of 189, 680 DNA viruses from the human gut microbiome. Nature Microbiol. *6*, 960–970. https://doi.org/10.1038/s41564-021-00928-6.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. (2021b). Author Correction: a genomic catalog of Earth's microbiomes. Nat. Biotechnol. *39*, 521. https://doi.org/10.1038/s41587-021-00898-4.

Ndah, E., Jonckheere, V., Giess, A., Valen, E., Menschaert, G., and Van Damme, P. (2017). REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. Nucleic Acids Res. *45*, e168. https://doi.org/10.1093/nar/gkx758.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425–430. https://doi.org/10.1038/nature19094.

Paez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., et al. (2017a). IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. *45*, D457–D465. https://doi.org/10.1093/nar/gkw1030.

Paez-Espino, D., Pavlopoulos, G.A., Ivanova, N.N., and Kyrpides, N.C. (2017b). Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat. Protoc. *12*, 1673–1682. https://doi.org/10.1038/nprot.2017.063.

Paez-Espino, D., Zhou, J., Roux, S., Nayfach, S., Pavlopoulos, G.A., Schulz, F., McMahon, K.D., Walsh, D., Woyke, T., Ivanova, N.N., et al. (2019). Diversity, evolution, and classification of virophages uncovered through global metagenomics. Microbiome *7*, 157. https://doi.org/10.1186/s40168-019-0768-5.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. *38*, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8.

Pons, J.C., Paez-Espino, D., Riera, G., Ivanova, N., Kyrpides, N.C., and Llabrés, M. (2021). VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. Bioinformatics *37*, 1805–1813. https://doi.org/10.1093/bioinformatics/btab026.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*, D61–D65. https://doi.org/10.1093/nar/gkl842.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Ramamurthi, K.S., and Storz, G. (2014). The small protein floodgates are opening; now the functional analysis begins. BMC Biol. *12*, 96. https://doi.org/10.1186/s12915-014-0096-y.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 69. https://doi.org/10.1186/s40168-017-0283-5.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985. https://doi.org/10.7717/peerj.985.

Roux, S., Krupovic, M., Daly, R.A., Borges, A.L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P.B., Cheng, J.-F., Ivanova, N.N., et al. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat Microbiol *4*, 1895–1906. https://doi.org/10.1038/s41564-019-0510-x.

Roux, S., Páez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T.B.K., Nayfach, S., Schulz, F., Call, L., et al. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. *49*, D764–D775. https://doi.org/10.1093/nar/gkaa946.

Sberro, H., Fremin, B.J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M.P., Pavlopoulos, G.A., Kyrpides, N.C., and Bhatt, A.S. (2019). Large-scale Analyses of human microbiomes reveal thousands of small, novel genes. Novel Genes. Cell *178*, 1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., and Woyke, T. (2020). Giant virus diversity and host interactions through global metagenomics. Nature *578*, 432–436. https://doi.org/10.1038/s41586-020-1957-x.

Snustad, D.P. (1968). Dominance interactions in Escherichia coli cells mixedly infected with bacteriophage T4D wild-type and amber mutants and their possible implications as to type of gene-product function: catalytic vs. stoichiometric. Virology *35*, 550–563. https://doi.org/10.1016/0042-6822(68)90285-7.

Song, S., and Wood, T.K. (2020). A primary physiological role of toxin/antitoxin systems is phage inhibition. Front. Microbiol. *11*, 1895. https://doi.org/10.3389/fmicb.2020.01895.

Storz, G., Wolf, Y.I., and Ramamurthi, K.S. (2014). Small proteins can No longer Be ignored. Annu. Rev. Biochem. *83*, 753–777. https://doi.org/10.1146/annurev-biochem-070611-102400.

Su, M., Ling, Y., Yu, J., Wu, J., and Xiao, J. (2013). Small proteins: untapped area of potential biological importance. Front. Genet. *4*, 286. https://doi.org/10.3389/fgene.2013.00286.

Taylor, N.M.I., Prokhorov, N.S., Guerrero-Ferreira, R.C., Shneider, M.M., Browning, C., Goldie, K.N., Stahlberg, H., and Leiman, P.G. (2016). Structure of the T4 baseplate and its function in triggering sheath contraction. Nature *533*, 346–352. https://doi.org/10.1038/nature17971.

Wang, J., Dai, W., Li, J., Xie, R., Dunstan, R.A., Stubenrauch, C., Zhang, Y., and Lithgow, T. (2020). PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. Nucleic Acids Res. *48*, W348–W357. https://doi.org/10.1093/nar/gkaa432.

Wang, J., Dai, W., Li, J., Li, Q., Xie, R., Zhang, Y., Stubenrauch, C., and Lithgow, T. (2021). AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins. Nucleic Acids Res. *49*, D630–D638. https://doi.org/10.1093/nar/gkaa951.

Washietl, S., Findeiss, S., Müller, S.A., Müller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA *17*, 578–594. https://doi.org/10.1261/rna.2536111.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
IMG/VR (version 3)	Roux et al. (2021)	https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html
Raw Sequencing Reads	Fremin et al. (2020)	PRJNA510123
Software and algorithms		
Prodigal (version 2.6.3)	Hyatt et al. (2010)	https://github.com/hyattpd/Prodigal
CD-Hit	Fu et al. (2012)	http://weizhong-lab.ucsd.edu/cdhit_suite
RPSBlast	Marchler-Bauer et al. (2005, 2011)	ftp://ftp.ncbi.nih.gov/blast/executables/
BLASTp	Altschul et al. (1997)	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast
RNAcode (version 0.3)	Washietl et al. (2011)	https://wash.github.io/rnacode/
trim galore (version 0.4.0)		https://github.com/FelixKrueger/TrimGalore
cutadapt (version 1.8.1)	Martin (2011)	https://cutadapt.readthedocs.io/en/stable/
bowtie (version 1.1.1)	Langmead et al. (2009)	https://sourceforge.net/projects/bowtie-bio/files/bowtie
bedtools (version 2.27.1)	Quinlan and Hall (2010)	https://sourceforge.net/projects/bedtools/
SignalP-5.0	Almagro Armenteros et al., 2019	http://www.cbs.dtu.dk/services/SignalP/
TMHMM (version 2)	Krogh et al. (2001)	http://www.cbs.dtu.dk/services/TMHMM/
AmPEP	Bhadra et al. (2018)	https://cbbio.cis.um.edu.mo/software/AmPEP
PaCRISPR	Wang et al. (2020)	https://pacrispr.erc.monash.edu/
AcrDB	Huang et al. (2021)	https://bcb.unl.edu/AcrDB/
PhyML	Guindon et al. (2010)	http://www.atgc-montpellier.fr/phyml/
MUSCLE	Edgar (2004)	https://www.drive5.com/muscle/
HMMER3	Eddy (2009)	http://hmmer.org/
Other		
CDD DB	Marchler-Bauer et al. (2005, 2011)	ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/
GEM	Nayfach et al. (2021a, 2021b)	https://portal.nersc.gov/GEM/

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/. 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 $\leq$ 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 $\leq$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 $\leq$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 $\leq$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 $\leq$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 $\leq$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.

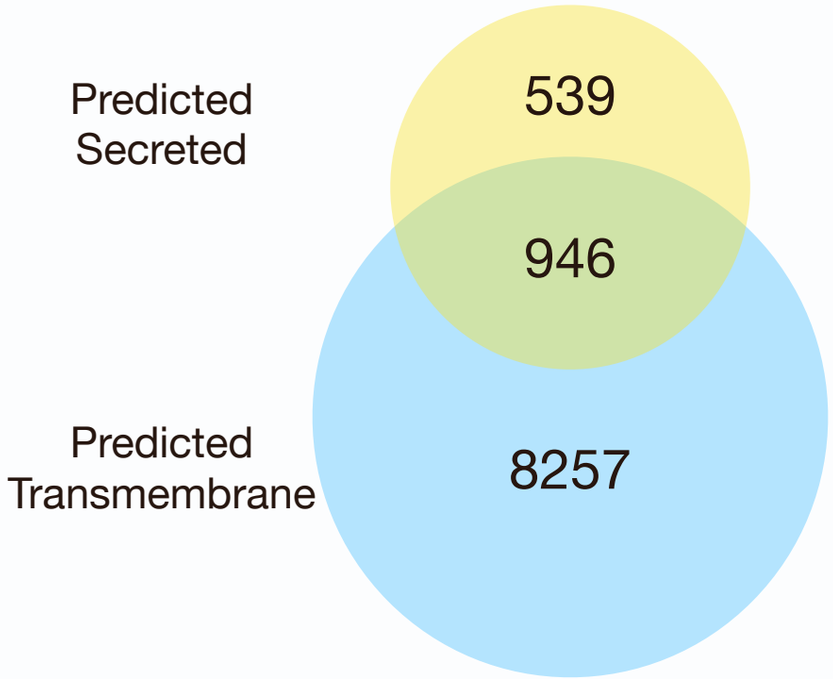
Cell Reports, Volume *39*

Supplemental information

Thousands of small, novel genes predicted in global phage genomes

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

A



B

Figure S1: Small proteins potentially involved in communication with host, related to Figure 1.

(A) Venn diagram showing the overlap between proteins predicted to be potentially secreted and those predicted to be transmembrane. Secreted proteins with signal peptides were sometimes falsely also predicted to be transmembrane proteins due to the hydrophobic region they contained.

(B) Multiple sequence alignment of representative sequences of family 12 and its homologous families, all of which were predicted to be transmembrane proteins.

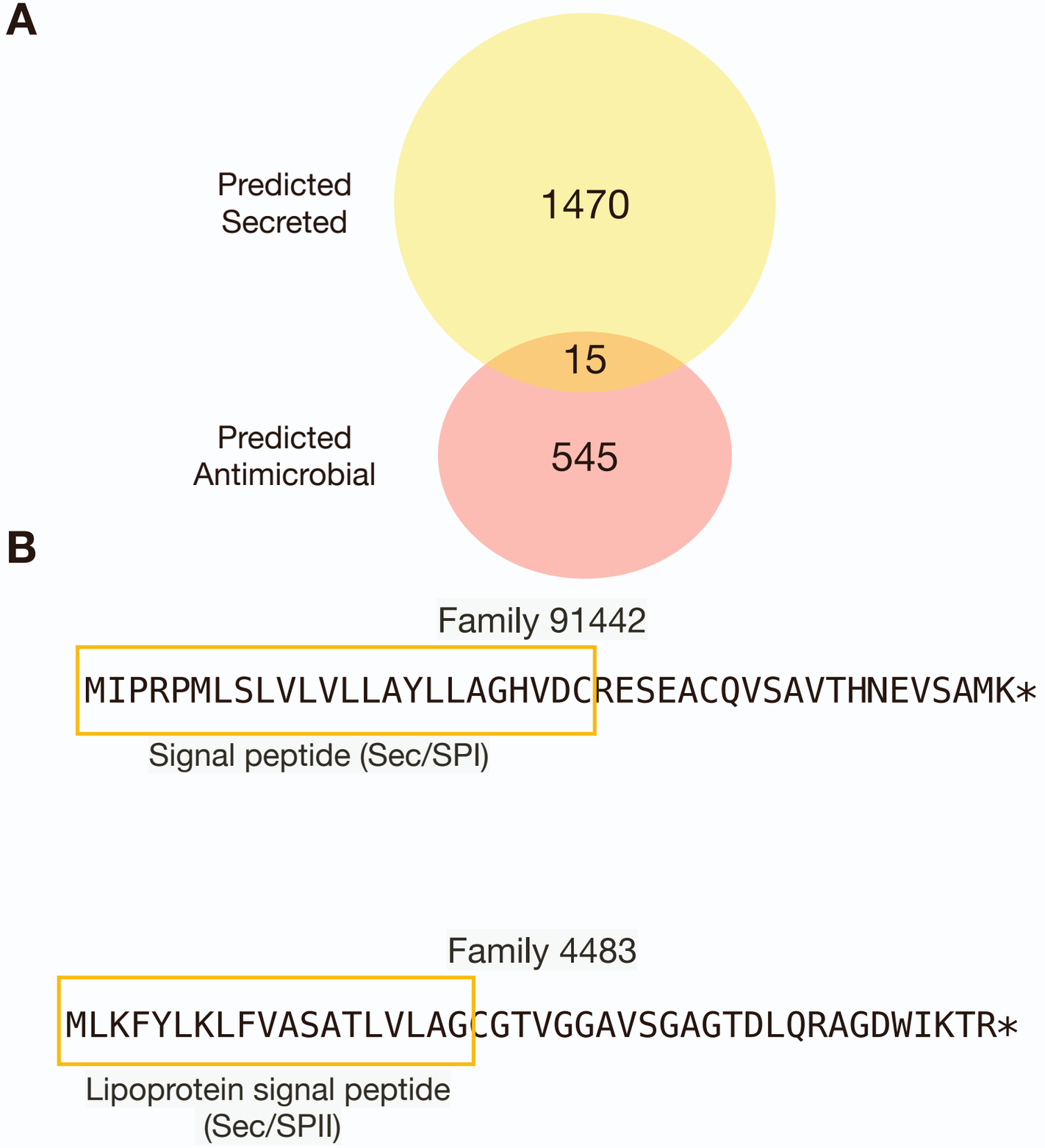


Figure S2: Small proteins potentially functional as exotoxins, related to Figure 1.

(A) Venn diagram showing the overlap between proteins predicted to be potentially secreted and those predicted to be antimicrobial.

(B) Examples of potential exotoxins containing signal peptides, including family 91442 and 4483.

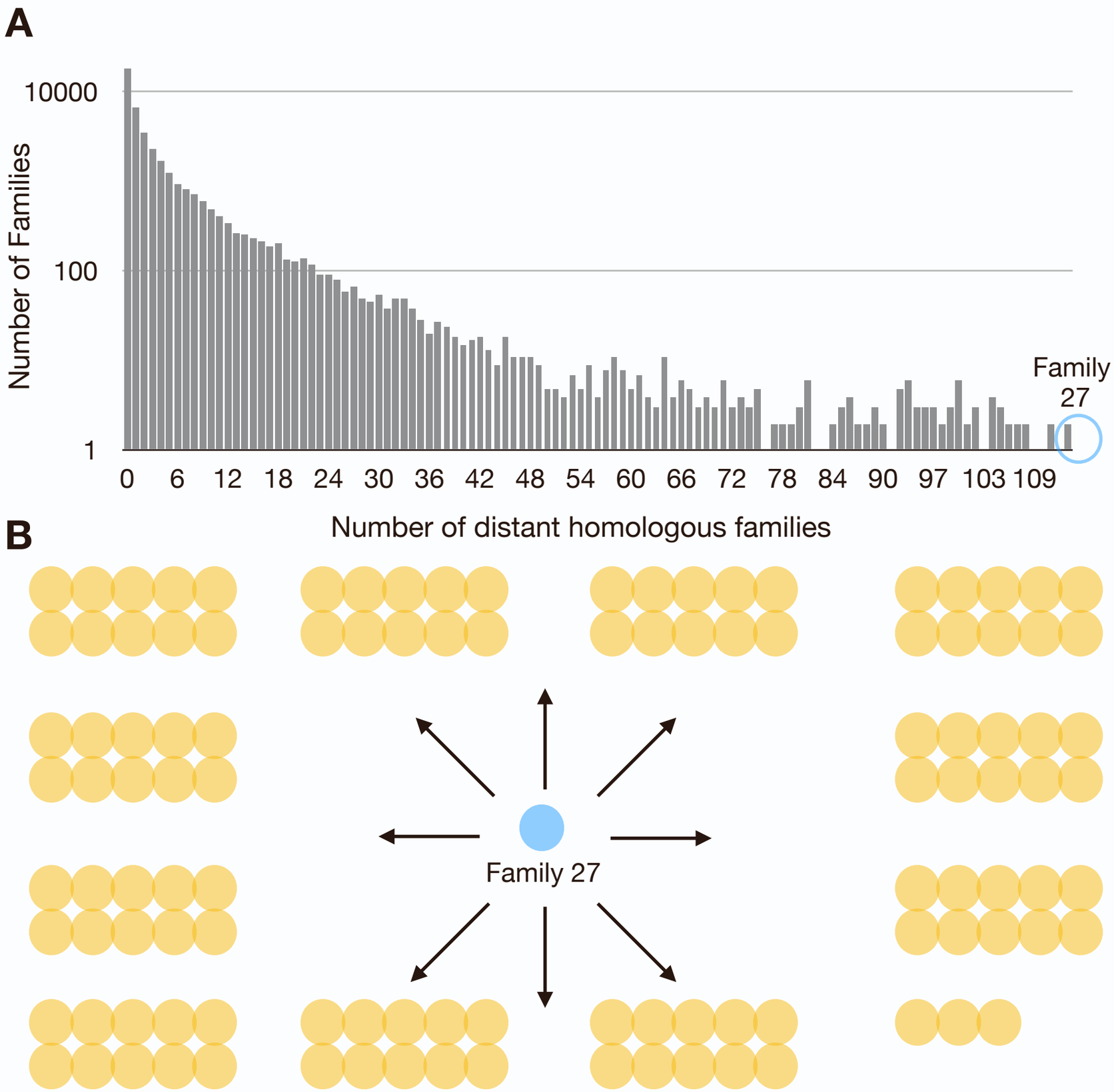


Figure S3: Homology within the Fremin gp40K, related to Figure 5.

(A) Histogram displaying the number of families that share homology with other families in the Fremin gp40K.

(B) Visualization of family 27 and the 113 homologous small gene families.

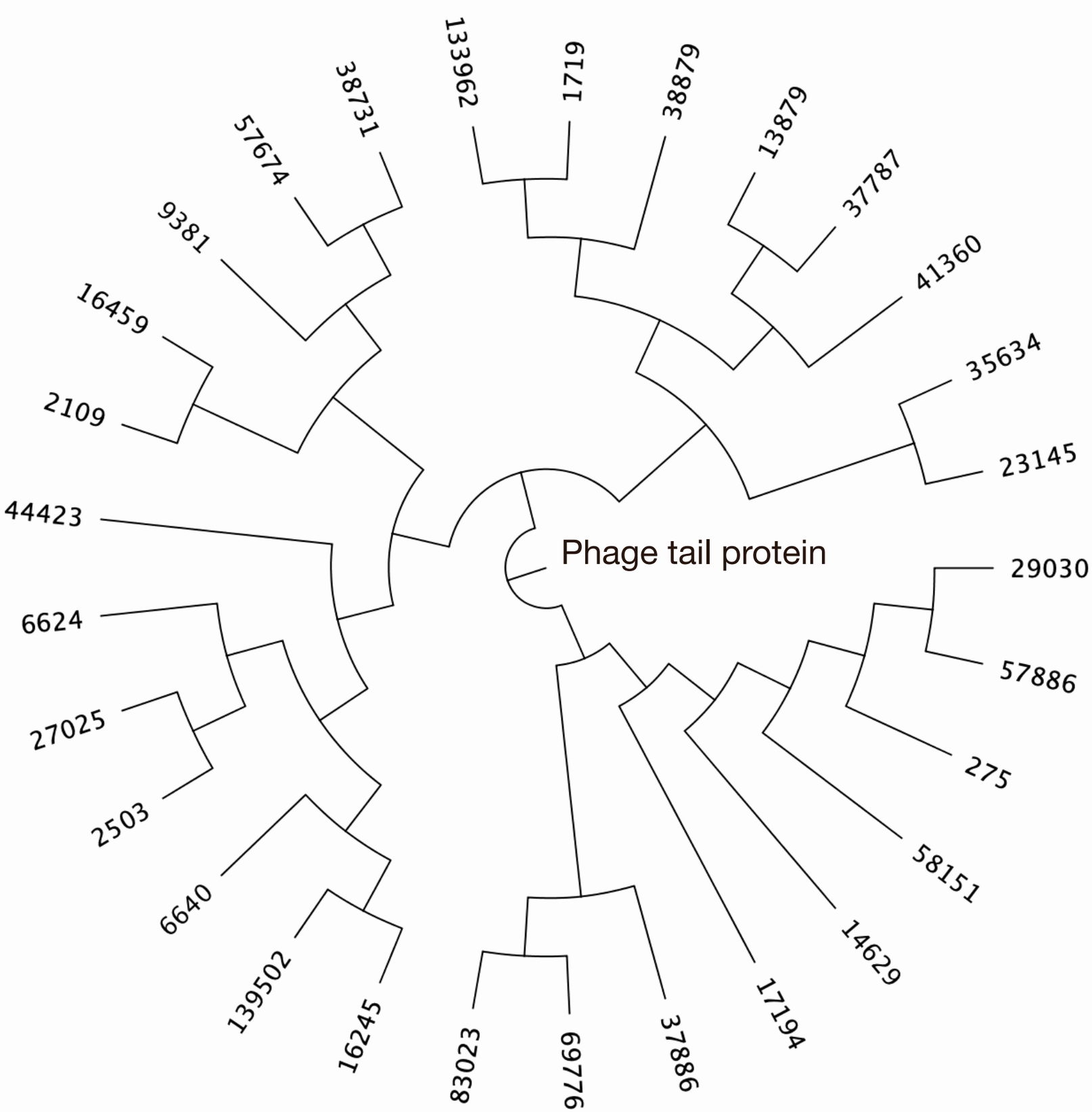


Figure S4: Novel phage tail small protein families, related to Figure 6.

Cladogram showing 29 novel small protein families that were homologous to Family 2109 and whose genes were all found near genes encoding phage minor tail and tape measure protein domains. Each family contained at least 3 unique homologs (not shown in the tree).

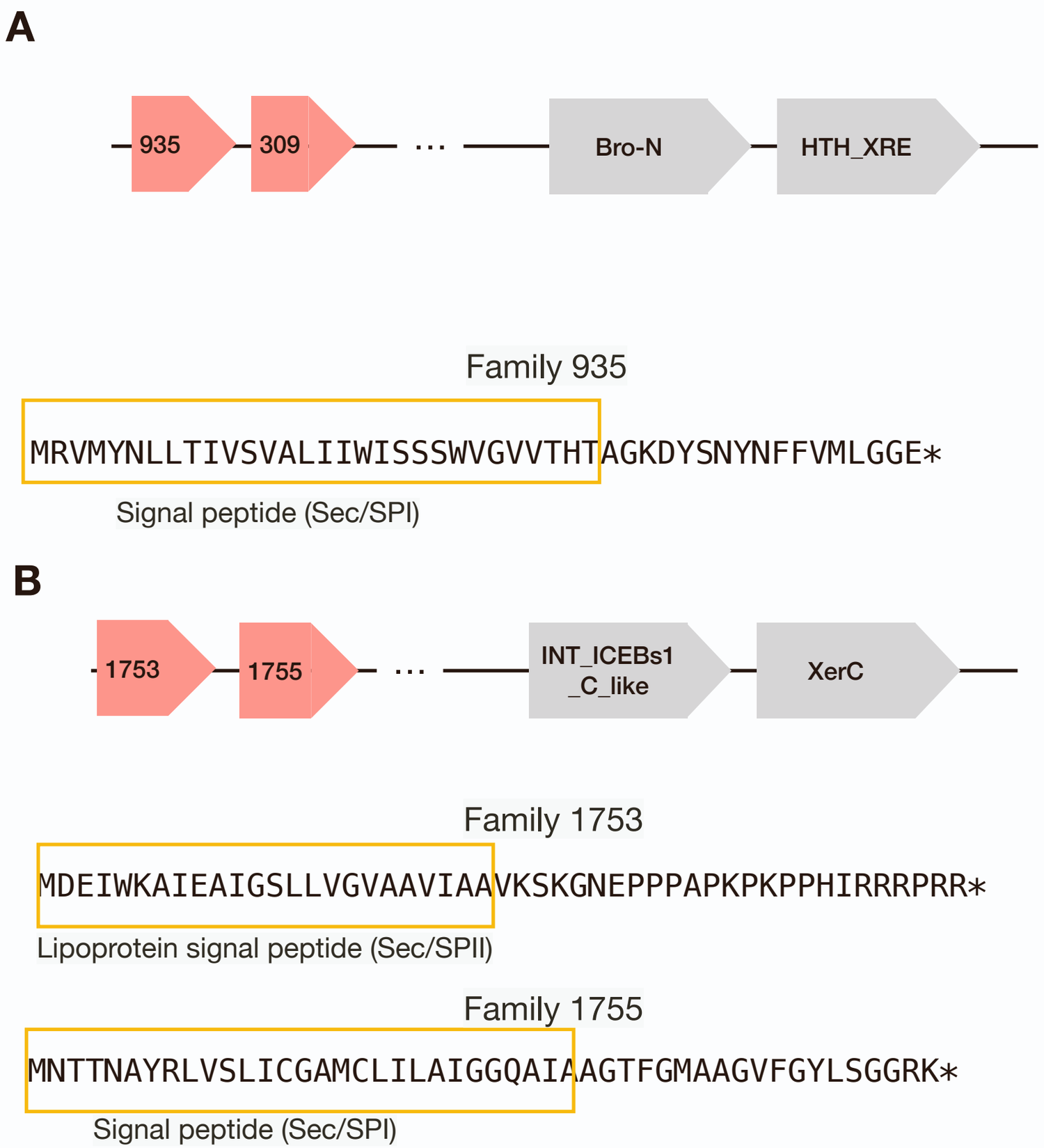


Figure S5: Novel potentially secreted families co-occurring with other families, related to Figure 1.

(A) Genomic representation showing families 935 and 303, whose genes often occurred next to each other and near genes encoding HTH-XRE and Bro-N protein domains. Annotation of signal peptide on family 935.

(B) Genomic representation showing families 1753 and 1755, whose genes often occurred next to each other and near genes encoding INT_ICEBs1_C_like and XerC protein domains. Annotation of signal peptides on families 1753 and 1755.