Diverse ATPase Proteins in Mobilomes Constitute a Large Potential Sink for Prokaryotic Host ATP

Hyunjin Shim[1], Haridha Shivram[2], Shufei Lei[1], Jennifer A. Doudna[2] and Jillian F. Banfield[1,2,3,4,5,6]*

[1] Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, United States, [2] Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, United States, [3] Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, United States, [4] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [5] Chan Zuckerberg Biohub, San Francisco, CA, United States, [6] School of Earth Sciences, University of Melbourne, Melbourne, VIC, Australia

Prokaryote mobilome genomes rely on host machineries for survival and replication. Given that mobile genetic elements (MGEs) derive their energy from host cells, we investigated the diversity of ATP-utilizing proteins in MGE genomes to determine whether they might be associated with proteins that could suppress related host proteins that consume energy. A comprehensive search of 353 huge phage genomes revealed that up to 9% of the proteins have ATPase domains. For example, ATPase proteins constitute ~3% of the genomes of Lak phages with ~550 kbp genomes that occur in the microbiomes of humans and other animals. Statistical analysis shows the number of ATPase proteins increases linearly with genome length, consistent with a large sink for host ATP during replication of megaphages. Using metagenomic data from diverse environments, we found 505 mobilome proteins with ATPase domains fused to diverse functional domains. Among these composite ATPase proteins, 61.6% have known functional domains that could contribute to host energy diversion during the mobilome infection cycle. As many have domains that are known to interact with nucleic acids and proteins, we infer that numerous ATPase proteins are used during replication and for protection from host immune systems. We found a set of uncharacterized ATPase proteins with nuclease and protease activities, displaying unique domain architectures that are energy intensive based on the presence of multiple ATPase domains. In many cases, these composite ATPase proteins genomically co-localize with small proteins in genomic contexts that are reminiscent of toxin-antitoxin systems and phage helicase-antibacterial helicase systems. Small proteins that function as inhibitors may be a common strategy for control of cellular processes, thus could inspire future biochemical experiments for the development of new nucleic acid and protein manipulation tools, with diverse biotechnological applications.

Keywords: ATPase protein, metagenome, mobilome, host energy, genome editing, antibiotic
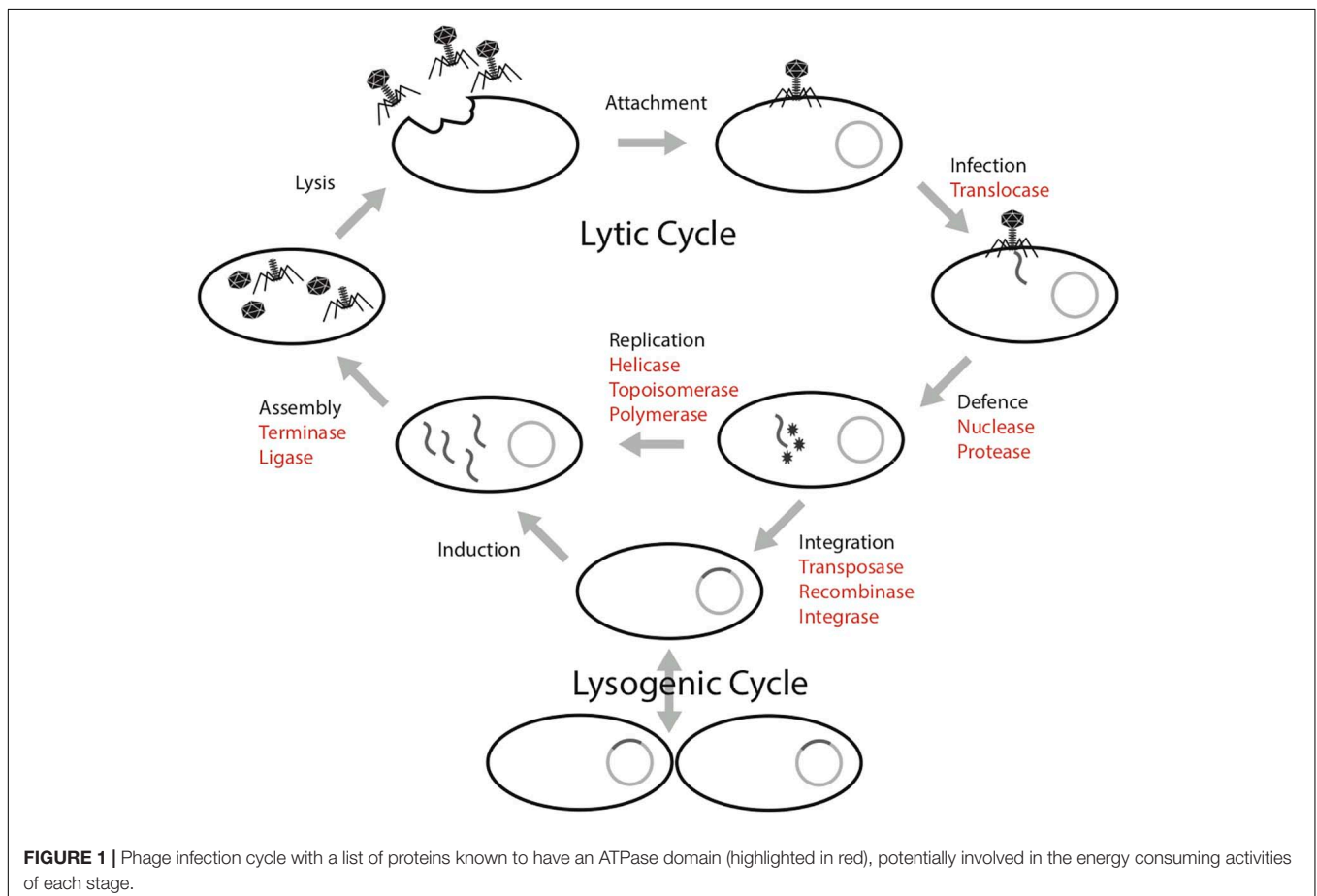
INTRODUCTION

The mobilome in prokaryotes includes plasmids, bacteriophages (phages), viruses of archaea and transposons that can move between host genomes. These mobile genetic elements are dependent on host machineries for transcription and replication (Frost et al., 2005; Gogarten and Townsend, 2005; Thomas and Nielsen, 2005). This parasitic propagation strategy necessitates the diversion of host energy, as these mobile genetic elements lack energy generating machinery (e.g., ATP synthase) and structures (e.g., the cell membrane). The infection cycle of a phage requires several energy intensive steps, both in the lytic and lysogenic cycles (**Figure 1**; Knowles et al., 2016). In the lytic cycle, phages have a DNA or RNA genome that is encapsulated within particles that are released and go on to infect new cells. In the lysogenic cycle, the phage can integrate into the host genome. These intra-genomic conflicts between hosts and mobilomes constitute a huge ATP sink.

Biochemically, protein activities that involve conformation changes in the structure require energy processing modules such as ATPase domains to bind and hydrolyze ATP to ADP. For example, ATP-dependent proteins such as helicases use energy released from ATP hydrolysis to unwind DNA segments during replication (Liu et al., 2004; Hood and Berger, 2016). At this time, the mobile genetic elements use host machinery either to replicate their genomes prior to cell lysis or to integrate into the host genome. Machineries involved in replication that are ATP-dependent include nucleases, transposases and polymerases (Neuwald et al., 1999; Iyer et al., 2004; Castaing et al., 2013; Krishnan et al., 2020). Further, the capsid packaging step is energetically expensive, approximately one ATP is required by the terminase to package two base pairs into the capsid for active DNA packaging (Yang and Catalano, 2003).

Prokaryotes have evolved several resistance mechanisms that are often ATP-dependent processes to disable the invasion of foreign genetic elements, notably restriction-modification systems (Krishnan et al., 2020) and some CRISPR-Cas systems (Sinkunas et al., 2011). Restriction modification systems consist of restriction endonucleases that cleave and degrade specific DNA sequences that are not modified by methylases (Vasu and Nagaraja, 2013; Loenen and Raleigh, 2014; Ershova et al., 2015). Studies have shown restriction modification systems (including restriction-only or modification-only systems) are energetically expensive (Seidel et al., 2008; Roberts et al., 2011). For example, it was estimated that *Escherichia coli* could consume ∼0.2% of its total ATP pool for effective restriction against invading foreign genomes (Roberts et al., 2011). CRISPR-Cas systems are an adaptive immune system that can acquire and degrade foreign genetic elements (Makarova et al., 2006, 2019; Barrangou et al., 2007; Tyson and Banfield, 2008). Recent discovery of

FIGURE 1 | Phage infection cycle with a list of proteins known to have an ATPase domain (highlighted in red), potentially involved in the energy consuming activities of each stage.

transposon-encoded CRISPR-Cas variants Tn6677 of *Vibrio cholerae* in *E. coli* reveals that the system requires an ATPase protein for transposition (Peters et al., 2017; Klompe et al., 2019; Strecker et al., 2019).

As a counter defense, some phages possess small anti-restriction proteins that impede or inhibit the activities of restriction enzymes (Iida et al., 1987; Kelleher and Raleigh, 1994; Wilkins, 2002; Tock and Dryden, 2005). A recent study revealed the phage-encoded small protein that inhibits the ATPase domain of the *Staphylococcus aureus* helicase loader (Hood and Berger, 2016). This could allow the phage to substitute its own helicase but it also prevents ATP consumption by the bacterial helicase, conserving ATP for phage functions. Even more recently, it was shown that tiny phage-encoded hypothetical proteins block bacterial quorum sensing machinery (Shah et al., 2021). From these few cases, and given their substantial energy requirements, we deduce that phages probably have a variety of proteins that have evolved to block bacterial systems that consume ATP.

Here, we sought to identify ATPase proteins in the genomes of mobile genetic elements (mobilome) assembled from metagenomic sampling. This bioinformatic analysis used a dataset of huge phage genomes to survey the relative abundance of ATP-requiring proteins and a large metagenomic dataset from various environments to discover new types of ATPase proteins in mobilome genomes. Particularly, we focus on proteins containing domains involved in nucleic acid and protein interactions, with the long-term aim of discovering new potential candidates for nucleic acid or protein manipulation tools. Furthermore, we investigated small proteins co-localized and co-occurring with ATP-requiring proteins to generate a database of protein candidates that may inhibit host ATPases, enabling mobile elements to hijack bacterial energy resources.

RESULTS

ATPase Genes Constitute Up to 9% of the Huge Phage Genomes

The ATPase proteins predicted from huge phage genomes account for, on average, ∼3% of the proteins (average 12.59 per genome with at least one ATPase domain; **Table 1**). Some of these ATPase proteins have multiple ATPase domains (1.70 per genome, 14.29 ATPase domains per genome). The variance in the number of ATPase proteins is wide, as shown as the standard deviation in **Table 1**, with one phage genome of 706 kbp (M01_PHAGE_CU_48_59) containing 47 ATPase proteins, whereas another phage genome of 220 kbp (FFC_PHAGE_43_1208) contains only 4 ATPase proteins. In one genome, ATPase proteins constitute 9% of the gene content (F31_PHAGE_39_18, 336 kbp). As the phage genome size varies from 104 kbp to 735 kbp, the number of ATPase proteins was plotted against the genome size (**Supplementary Figure 1A**). The regression shows a linear relationship between genome length (p-value = 2×10^{-16}) and the number of ATPase proteins (4.435 per 100 kb).

As an example, we used a well sampled clade of Lak phages of ∼550 kbp genomes to conduct a detailed analysis of ATPase

proteins in a specific clade. Lak phages are currently the largest genomes identified in human and animal microbiomes (Devoto et al., 2019; Crisci et al., 2021). As shown in **Figure 2**, the predicted ATPase proteins are distributed throughout the genome and have diverse functions, including topoisomerase, transferase, protease, kinase, ligase, and helicase, while at least 7 of them have unknown functional domains or no functional domains.

ATPase Proteins Are Diverse and Numerous in Mobilome Genomes

To further analyze the function of ATPase proteins in mobilomes, we searched for composite ATPase proteins that have an ATPase domain fused with one or more functional domains. From the mobilome dataset, we found 505 composite proteins that had an ATPase domain fused with at least one functional domain. These were categorized by analyzing the annotations on the function domains. Surprisingly, 194 of the 505 composite ATPase proteins (38.4%) have domains with no match to functional annotations.
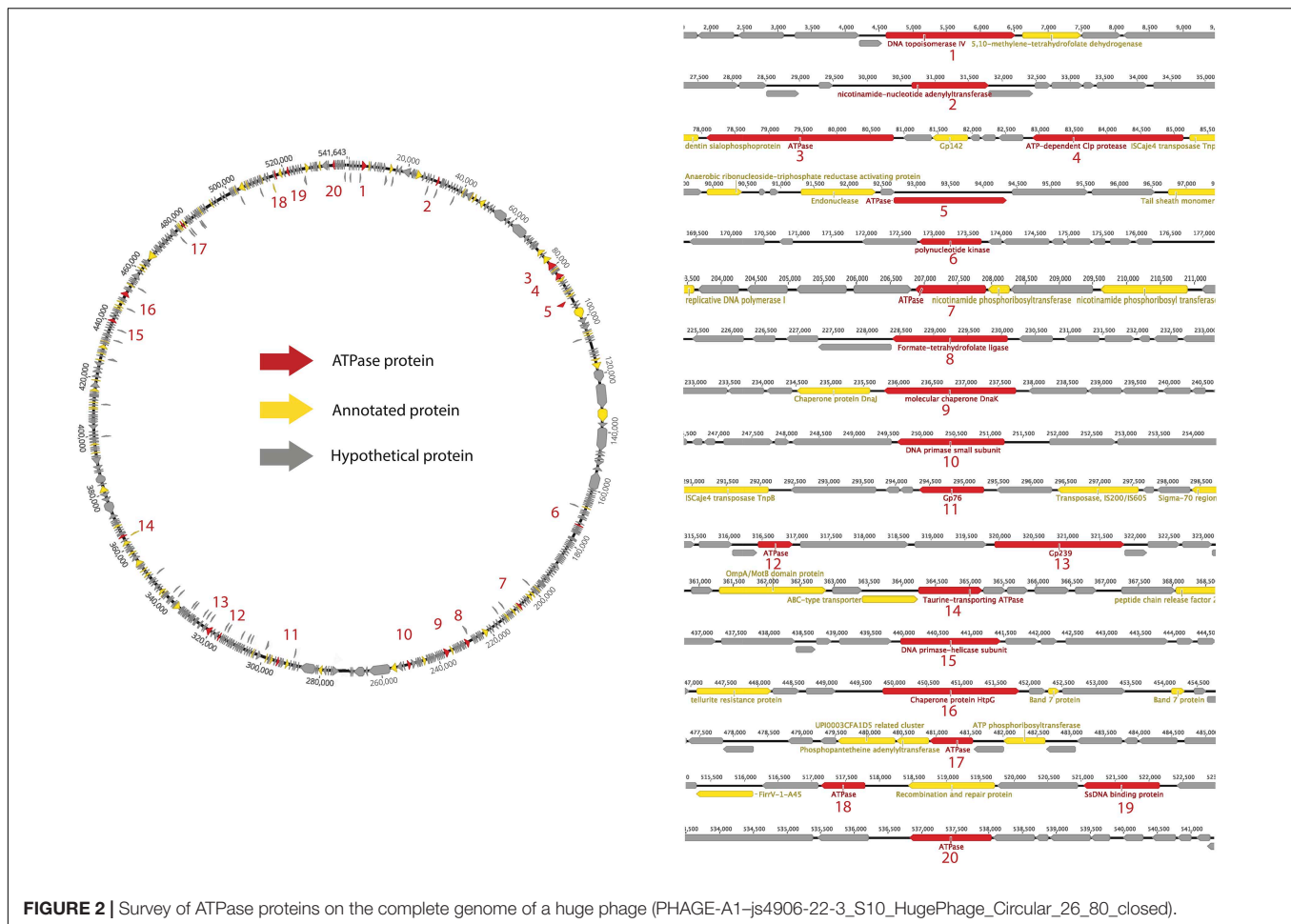
The other 311 composite ATPase proteins (61.6%) had domains with annotations related to diverse functions associated with various stages of the mobilome infection cycle, including nucleic acid and amino acid interacting enzymes, and packaging and transport related proteins. Among these, we found 90 composite ATPase proteins with annotations that are related to later stages of the mobilome infection cycle such as packaging, transport and secretion. The remaining 221 composite ATPase proteins are nucleic acid and amino acid interacting enzymes, and thus are potentially involved in replication and defense. As shown in **Figure 3**, this category includes nucleic acid interacting proteins such as nucleases and helicases, and amino acid interacting proteins such as proteases. The most abundant functions of these composite ATPase proteins are proteases, nucleases, helicases and DnaK chaperones. Other annotations indicate functions related to membrane-bound proteins for transport and secretion, which have recently been suggested to be associated with defense systems (Doron et al., 2018; Millman et al., 2020).

Some ATPases Involved in Replication Are Not Yet Functionally Characterized

We found several composite ATPase proteins with domains of unknown function (DUF) or uncharacterized function. These proteins are annotated in the search database with some

TABLE 1 | Summary statistics of 353 huge phage genomes and associated ATPase proteins.

Summary statistics	Genome length (kbp)	Total proteins	GC content (%)	ATPase domains	ATPase proteins	Multiple ATPase proteins
Mean	291	397.2	39.20	14.29	12.59	1.70
Standard deviation	89	160.68	8.99	6.14	5.28	1.30
Min	104	106	21.98	3	3	0
Max	735	1144	63.98	60	47	13

FIGURE 2 | Survey of ATPase proteins on the complete genome of a huge phage (PHAGE-A1–js4906-22-3_S10_HugePhage_Circular_26_80_closed).
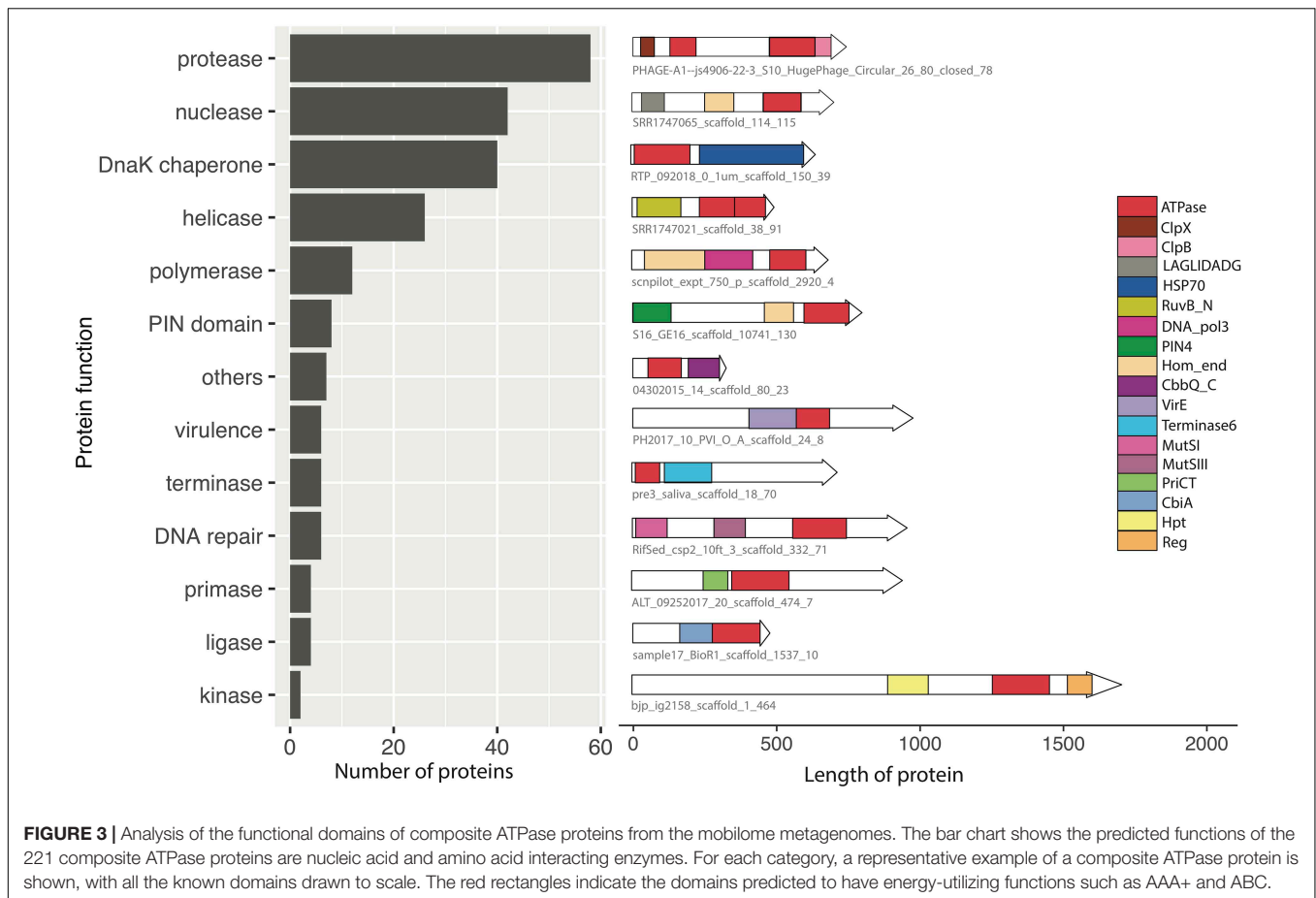
superfamily domains based on sequence or structural similarity, but the exact biological function of each family has not been characterized. For instance, a few composite ATPase proteins have PIN domains that belong to a superfamily of nucleases with distinct structural similarities of a common Rossmanoid fold (Arcus et al., 2011; Matelska et al., 2017). They are known to be broadly involved in central cellular processes, but each family participates in diverse central cellular processes such as DNA replication and repair, transcription regulation, mRNA degradation and ncRNA maturation. As shown in **Supplementary Figure 2**, the PIN-like domains found in these composite ATPase proteins belong to the PIN_4 family. The PIN_4 family was reported to occur in bacteria and eukaryotes, but the exact function is unknown (Matelska et al., 2017). Interestingly, two of these proteins also have another nuclease superfamily of the LAGLIDADG homing endonucleases fused as a domain (**Supplementary Figure 2**). These "meganucleases" are DNA cleaving enzymes in microbial genomes and cut target sites with high specificities to initiate the lateral transfer of mobile genetic elements (Taylor et al., 2012).

Another category of interest has domains that are related to the virulence-associated protein E (virE), which is known to be part of a prophage region that is associated with virulence

in several microbes (Ma et al., 2013; Ji et al., 2016). However, the exact mechanism by which the virE affects virulence has not been characterized. Other domains of interest include those involved in DNA repair (e.g., MutS) and nucleotide binding (e.g., CobQ/CobB/MinD/ParA domain).

ATPase Domains Cluster by Protein Function in Multiple Sequence Alignments and in Tree Phylogenies

To visualize the patterns, the ATPase domains of the composite ATPase proteins were aligned by sequence similarity (**Supplementary Figure 3**). **Supplementary Figure 3** shows the ATPase domains are a 300 amino acid sequences of high variability, with some conserved motifs such as the ATP binding motif (Walker-A) and the ATP hydrolysis motif (Walker-B). The Walker-A motif in the ATPase domains has a primary amino acid sequence of GxxGxGKT or GxxGxGKS, where the letter x represents any amino acid (Schneider and Hunke, 1998). The Walker-B motif has a primary sequence of hhhhD, where the letter h represents any hydrophobic amino acid (including Valine, Isoleucine, Leucine, Phenylalanine, and Methionine). Structurally, the Walker-A motif is an α-helix followed by

FIGURE 3 | Analysis of the functional domains of composite ATPase proteins from the mobilome metagenomes. The bar chart shows the predicted functions of the 221 composite ATPase proteins are nucleic acid and amino acid interacting enzymes. For each category, a representative example of a composite ATPase protein is shown, with all the known domains drawn to scale. The red rectangles indicate the domains predicted to have energy-utilizing functions such as AAA+ and ABC.

a glycine-rich loop, and the Walker-B motif is a β-strand connected to the Walker-A by a 100-peptide sequence.

The multiple sequence alignment shows that the ATPase domains cluster by the biological or enzymatic activity of the functional domain. The upper clusters tend to be the biological activities related to the later stages of infection, such as assembly and lysis, including chaperones and transport proteins. The lower clusters tend to be the enzymatic activities related to the earlier stages of infection, such as replication and defense, including nuclease, polymerase, protease, and helicase.
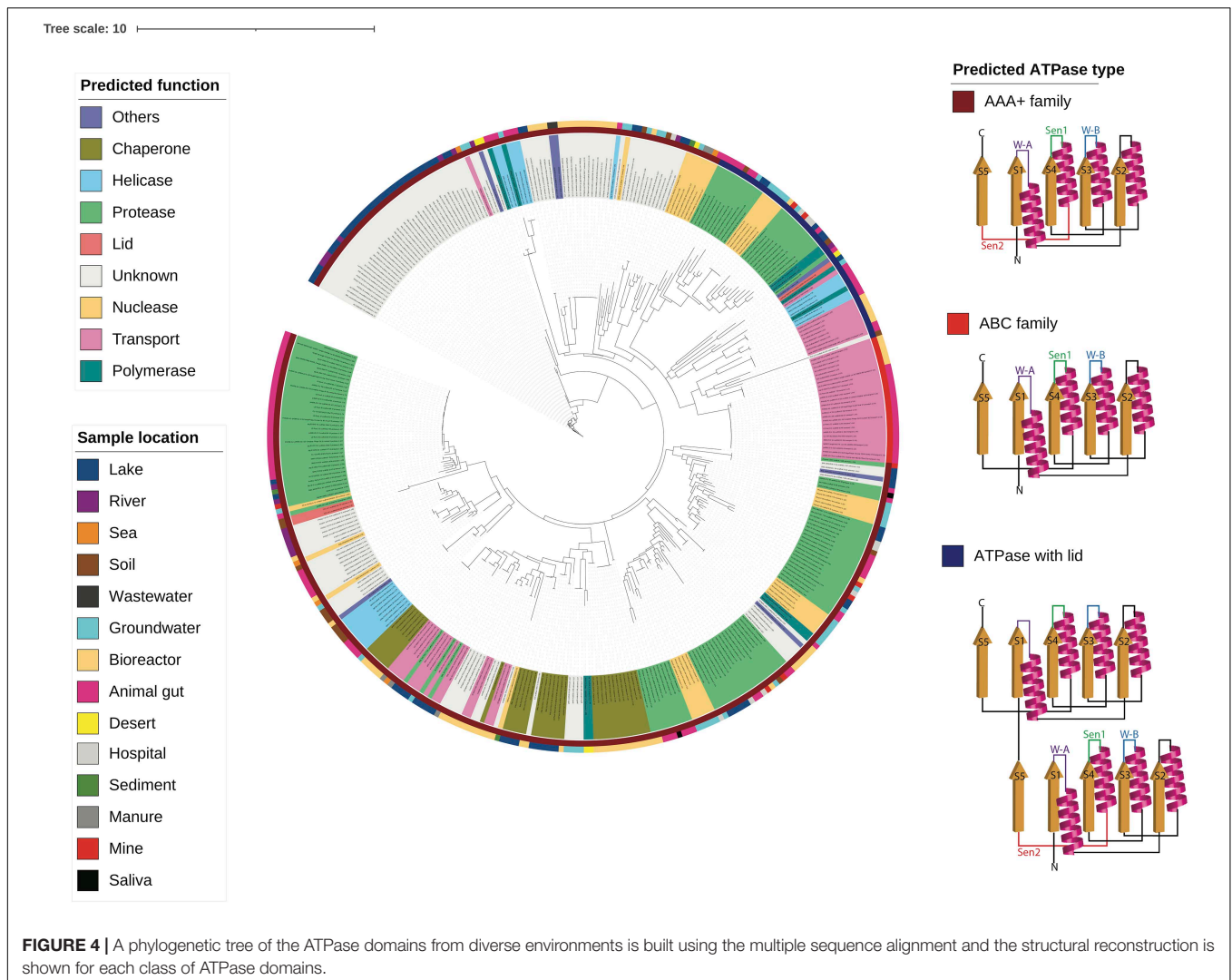
Using the multiple sequence alignment (**Supplementary Figure 3**), a phylogenetic tree of the ATPase domains was reconstructed as shown in **Figure 4**. The tree shows that the ATPase domains cluster by the predicted function. Interestingly, the family of an ATPase domain and the predicted function of its associated domain classified according to the secondary structure (Neuwald et al., 1999; Iyer et al., 2004; Castaing et al., 2013; Krishnan et al., 2020) closely aligns with the predicted ATPase function. For example, most ATPase domains associated with transport belong to the ABC family, while many ATPase domains associated with helicases and polymerases belong to the helicase superfamily (Krishnan et al., 2020). The helicase superfamily consists of one active ATPase domain fused with a lid that is an inactive ATPase domain. Furthermore, most ATPase DUF belong

to the AAA+ family that are associated with diverse cellular activities (Neuwald et al., 1999).

Genomic Context of Composite ATPase Proteins Is Conserved

The analysis of protein lengths and genomic context shows that the composite ATPase proteins are often co-localized with small proteins (**Supplementary Figure 4**). It is notable that the average length of the composite ATPase proteins is more than double than the average length of the co-localized proteins. To identify if these small proteins are conserved, homologs for each protein in the genomic context of three proteins downstream and upstream of the composite ATPase proteins were identified (**Supplementary Figure 5**). As shown in **Supplementary Figure 5**, the proteins in the genomic context of the composite ATPase proteins have many close homologs. Hierarchical clustering reveals that genomic contexts have three main types: the first with many homologs across the genomic context (bottom), the second with many homologs in the proteins adjacent to the composite ATPase proteins (top), and the third with few homologs in the genomic context (middle).

The first and the second clusters were investigated further for a conserved genomic context (**Supplementary Table 1**).

FIGURE 4 | A phylogenetic tree of the ATPase domains from diverse environments is built using the multiple sequence alignment and the structural reconstruction is shown for each class of ATPase domains.

The analyses reveal that the order of protein homologs was better conserved when the protein has more homologs, which is consistent with the hypothesis that the two factors are correlated. Some genomic contexts of the composite ATPase proteins from the first cluster and the second cluster are shown in **Figures 5A,B**, respectively. It is notable that most of the co-localized proteins are hypothetical, but yet conserved in this genomic context. **Figure 5A** shows the highly conserved genomic context of the composite ATPase proteins, whose function is mainly related to proteolytic activities. These composite ATPase proteins with multiple ATPase domains belong to the family of Lak megaphages. **Figure 5B** shows the genomic contexts with many homologs in the proteins adjacent to the composite ATPase proteins. Interestingly, these composite ATPase proteins display diverse functions, such as nuclease, protease, terminase, primase and helicase, and they overlap in sequence with the adjacent proteins. These genes that overlap in sequence with the composite ATPase proteins are more conserved than non-overlapping genes in the genomic context.

DISCUSSION

To replicate, mobile genetic elements require ATP available within the host cell. For each step of the mobilome infection cycle, we identified many ATPase proteins that are potentially involved in energy-requiring activities. A comprehensive survey of 353 huge phage genomes revealed that, on average, 3% of the proteins are ATP-binding, although some genomes encode up to 9% ATP-binding proteins. Given that statistical analysis shows the number of ATPase proteins increases linearly with the genome length, phages with large genomes represent a huge sink for bacterial ATP. Lak megaphages, whose genomes are >20% of the length of their host *Prevotella* genomes, have diverse ATPase proteins, including topoisomerases, transferases, proteases, kinases, ligases, helicases, and proteins of unknown functions. Thus, these phages invest substantial host energy in the form of ATP in nucleic acid, as well as protein transformations.

We analyzed the functions of ATPase proteins in mobile genetic elements from diverse environments. Of the 505 composite ATPase proteins with additional domains, 38.4% had
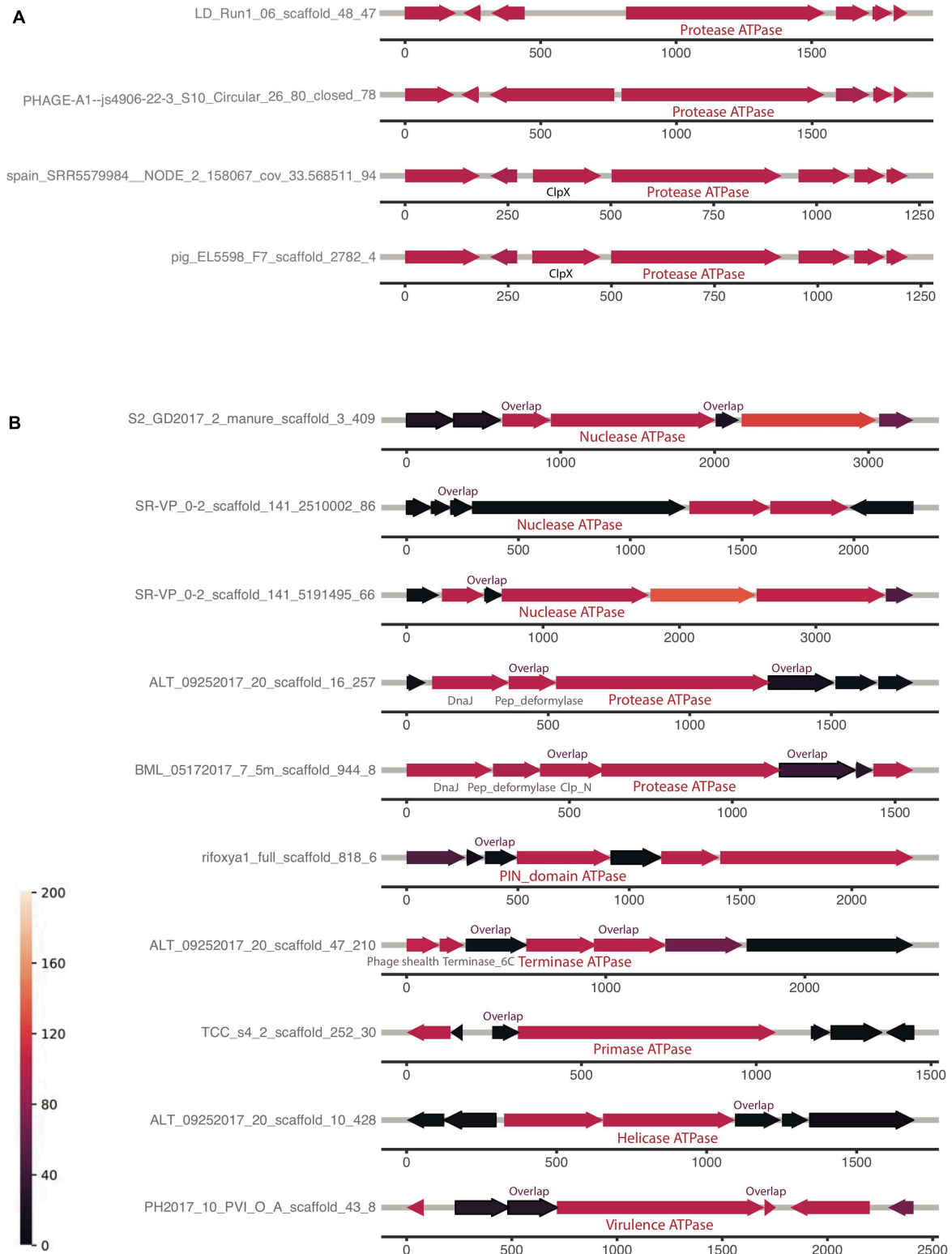
FIGURE 5 | Examples of genomic neighborhood of composite ATPase proteins, with the proteins localized ±3 of each composite ATPase protein indicated in the middle. Most proteins are unannotated and hypothetical proteins, except those with labels. Identical to the genomic neighborhood heatmap, the color represents the number of blast hits in the ggKbase database. (A) Composite ATPase proteins that are found to have highly conserved operons (indicated with green stars in **Supplementary Figure 5B**). (B) Composite ATPase proteins that are found to have highly conserved pairs with a sequence overlap (indicated with blue stars in **Supplementary Figure 5B**).

domains of unknown function. A small subset of membrane-bound proteins appears to be involved in defense (Doron et al., 2018; Millman et al., 2020). Most of the remainder of the composite ATPase proteins have functional domains that suggest roles during the early stages of infection related to replication or defense, such as nuclease, protease, polymerase, helicase and transposase. Consistent with these indications of energetically expensive stages of the mobilome activity, we discovered some composite ATPase proteins with up to three ATPase domains per protein. These proteins are mostly predicted to be nucleases, proteases, polymerases and chaperones. Some of the most interesting proteins in this category are predicted to be site-specific nucleases of the PIN domain and LAGLIDADG domain fused with multiple ATPase domains. The explanation for nucleases that require multiple ATPase domains is not known. These may be candidates of interest for experimental evaluation.

Mobilome genomes encode large inventories of small proteins, the functions of which remain a mystery. We analyzed the genomic contexts of some small proteins encoded adjacent to ATPase domain proteins, with the objective of developing hypotheses regarding their roles. We discovered genes for several small hypothetical proteins that overlap with genes encoding composite ATPase proteins. This colocalization pattern is conserved in many examples. This is reminiscent of the genomic architecture of a previously validated system in which the ATPase is a toxin and the small protein inhibitor an antitoxin (**Supplementary Figure 6**). This study found an inhibitory role for a small phage protein, the gene for which overlapped with a gene for an ATPase protein (Liu et al., 2004; Hood and Berger, 2016). Specifically, a phage-encoded small protein was shown to inhibit the ATPase domain of the *S. aureus* helicase loader (Hood and Berger, 2016). Based on the combination of our results and these findings, we suspect that small proteins associated with ATP-binding proteins may inhibit host homologs and substitute for their function, simultaneously eliminating an undesirable sink for host ATP. In fact, we speculate that many phage-borne small proteins have evolved to block host bacterial machinery, a concept that is supported by the discovery of small anti-CRISPR proteins that protect phages by inhibiting the function of CRISPR-Cas bacterial immune systems. An additional indication that small proteins are part of a widespread phage strategy to control host cells is the very recent discovery that small phage-encoded proteins bind to and block the regulator of the bacterial quorum sensing pathway. These proteins also bind to the pilus assembly ATPase protein to prevent superinfection by other phages (Shah et al., 2021).

From the analysis of genomic context, we found three clusters of distinct features neighboring the composite ATPase proteins. The first cluster with the highly conserved genomic context is potentially a mobilome operon with proteolytic activity. Given that the second cluster had ATPase genes with highly conserved context we could derive hypotheses regarding the roles of ATPases based on the predicted roles of the adjacent proteins. The third cluster is of interest because the proteins lacked conserved genomic context. Some of these ATPase proteins are predicted to have diverse functions, including as helicases, ligases, nucleases, and replication regulatory elements. Future

evolutionary biology computational analyses studies of these seemingly isolated ATPase proteins may determine whether the lack of conserved genomic context is because they function without partners or is due to rapid evolution of the adjacent genes (McInerney et al., 2017; Shim, 2019).

It is known from experimental studies that phage-encoded ATPase inhibitors bind specifically the host ATPase proteins and not the phage ATPase homolog (Liu et al., 2004; Hood and Berger, 2016). Thus, an interesting question relates to why small proteins block the ATP binding sites of the bacterial protein but not that of the mobilome homolog. We found that the ATPase domains in the mobilome display highly varying degrees of sequence conservation. All have functionally critical motifs (Walker-A and Walker-B) and variations in other parts of the sequences, including Motif-C and other sensing regions. We predict that small proteins that bind and inhibit the ATP binding site interact with these more variable regions. This would enable the specificity required to avoid blocking of the phage protein. If correct, this will open new opportunities to discover and design small protein inhibitors, with broad applications in areas such as antimicrobial therapy and cancer-suppression.

MATERIALS AND METHODS

ATPase Domain Prediction in the Huge Phage Dataset

The huge phage dataset contains 353 near complete genomes from diverse environments, including rivers, lakes, groundwater, sea, soil and human/animal microbiomes (Al-Shayeb et al., 2020). The dataset consists of 140,206 proteins. The genes of some huge phages such as Lak phages were predicted using Prodigal and genetic code 15, as the stop codon (TAG) has been reprogrammed to encode glutamine (Devoto et al., 2019). ATPase domains from the huge phage dataset were predicted with the *e*-value cut-off of 1×10^{-3} based on the accession of their Hmmsearch match against the Pfam database (version 33.1) (Sonnhammer et al., 1998). The Pfam domain search was compared and supplemented with the Hmmsearch match against the Supfam database for structural information (Wilson et al., 2009). The number of ATPase proteins was plotted against the phage genome size (**Supplementary Figure 1A**). A linear regression was fitted after the two distributions were evaluated with the Normal Q-Q plot to check they are linearly related and normally distributed (**Supplementary Figure 1B**).

Big Mobilome Dataset and Curation

The initial mobilome dataset contains 91,205 contigs from ggKbase that were identified based on their gene content (last accessed in August 2019). ggKbase is a web application designed to assist in aggregating, visualizing, exploring and analyzing metagenomic data. The dataset encompasses 9,23,411 proteins in total; 7,31,372 phage proteins, 44,234 virus proteins, 112,101 plasmid proteins; and 35,695 others which includes proteins from the contigs that were manually binned as mobilome (**Table 2**; **Supplementary Figure 7**). This dataset was used to investigate ATP-requiring proteins. Each protein was first annotated using

TABLE 2 | Percentage of ATPase per protein by category in the mobilome metagenomes.

	Virus	Phage	Plasmid	Other
Number of proteins	44,243	731,372	112,101	35,695
Number of ATPases	361	2,654	2,216	266
ATPase/protein	0.82%	0.36%	1.98%	0.75%

the basic local alignment search tool (BLAST: version 2.10.1) against the NCBI database. The initial search for ATPase proteins was done by searching for annotations of the mobilome proteins with the keyword "ATPase" (**Table 2**; **Supplementary Figure 8**). The mobilome dataset had 5,497 proteins with the matching criteria.

Composite Gene Families Detection

Selecting composite ATPase proteins from the mobilome dataset was achieved using CompositeSearch with a two-step procedure. First, the initial ATPase dataset from the mobilome dataset was used to find homologs using BLAST against the NCBI RefSeq non-redundant database and the mobilome dataset. The resulting ATPase dataset contained 1,433,097 ATPase protein homologs. Second, this homolog dataset was used to search for composite ATPase proteins in the NCBI RefSeq non-redundant database and the mobilome dataset, where the sequence similarity networks was used to detect composite gene families. The output file from CompositeSearch contained 64,089 composite ATPase proteins in total. Only 1,003 composite ATPase proteins from the ggKbase metagenomic dataset were selected for further analyses.

Among this composite ATPase proteins from the mobilome dataset, some were from contigs that had bacterial or archaeal taxonomic profiles. To eliminate contigs that are likely to be non-mobilome, the composite ATPase proteins from the contigs that were identified to encode >50% proteins with best hits to bacteria or/and archaea were excluded from further analysis, resulting in a set of 562 candidates. Finally, after eliminating the 57 duplicate candidates generated by the gene prediction program (Prodigal version 2.6.3), the final set of composite ATPase proteins was 505.

Functional Annotation and Domain Prediction

Composite ATPase proteins were functionally annotated based on the protein signature recognition of their best InterProScan match (version 5.47-87) (Zdobnov and Apweiler, 2001). InterPro Application Programming Interface (API) was used for direct access to the InterPro database (version 81.0), with the *e*-value cut-off of 1×10^{-3}. Domains were predicted based on the accession of their best Hmmsearch match against the Pfam database (version 33.1), using the same procedure. The composite ATPase proteins with the functional domains of interest were further explored using TRee-based exploration of neighborhoods and domains (TREND) (Gumerov and Zhulin, 2020). This platform uses MAFFT algorithms for multiple protein sequence alignment (-auto option) and uses FastTree (version 2.0.0) (Price et al., 2010) to perform tree reconstruction. The resulting tree-based exploration of domains can be visualized with the domain

architecture identified from the various sources, including Pfam, CDD, NCBI and MiST.

Multiple Sequence Alignment of Predicted ATPase Domains

ATPase domains were extracted using the domain predictions from the Pfam database (version 33.1) by retaining only the domains with the annotation keywords "ATP" or "AAA." Protein sequences of the ATPase domains were extracted through the start and end predictions of the domain for each composite ATPase protein. The resulting ATPase domains had 179 sequences, which were aligned using a multiple sequence alignment program, MAFFT (version 7.471) (-auto option) (Katoh et al., 2002). Subsequently, the multiple sequence alignment of the predicted ATPase domains was visualized using Jalview (version 2.11.1.3). The residue positions are coded with the Taylor protein coloring scheme (Taylor, 1997) using the conservation visibility of 15% (**Supplementary Figure 3**).

Phylogenetic Analyses of Predicted ATPase Domains

Tree reconstructions from the multiple sequence alignments of the predicted ATPase domains were performed using IQ-TREE (version 1.6.12) (Nguyen et al., 2015). From the ModelFinder (Kalyaanamoorthy et al., 2017), the best performing substitution model was selected as LG + R5, and 1,000 ultrafast bootstrap were run (Hoang et al., 2018). The reconstructed tree was visualized with iTOL (version 4) and iTOL annotation editor (Letunic and Bork, 2019), with the following labels: protein name with predicted function, ATPase classification and sample location.

Identification of Genomic Context Conservation Near Composite ATPase Proteins

The genomic context of the composite ATPase proteins was defined as proteins co-localized three genes upstream and downstream. First, each protein within the genomic context was searched for homologs in the ggKbase database using BLAST (Altschul et al., 1990) with the *e*-value cut-off of 1×10^{-30} (**Supplementary Table 1**). The BLAST result for each composite ATPase protein and its neighboring proteins was summarized as a heatmap. The intensity of color represents the number of BLAST hits, which is equivalent to the number of homologs in the ggKbase. The heatmap was clustered hierarchically by similarity for pattern recognition (**Supplementary Figure 5**). The BLAST hits were analyzed for conservation of order in different clusters. If the order of the genomic context was conserved, these ATPase proteins were defined to have a conserved genomic context that may be functionally linked as an operon.

DATA AVAILABILITY STATEMENT

The huge phage genomes can be downloaded from ggkbase.berkeley.edu/hpae_final/organisms. The mobilome

metagenomic data can be downloaded from ggkbase.berkeley. edu/phage-plasmidvirus-protein-families/organisms. The code used to analyze the datasets is available at github.com/hjshim/ Pro_ATP.

AUTHOR CONTRIBUTIONS

JB and HSm conceived of the presented idea. HSm performed the analysis and wrote the manuscript. JD and HSv contributed to the study design. SL provided bioinformatics support. All authors contributed to writing the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb. 2021.691847/full#supplementary-material

REFERENCES

Al-Shayeb, B., Sachdeva, R., Chen, L. X., Ward, F., Munk, P., Devoto, A., et al. (2020). Clades of huge phages from across Earth's ecosystems. *Nature* 578, 425–431.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Arcus, V. L., Mckenzie, J. L., Robson, J., and Cook, G. M. (2011). The PIN-domain ribonucleases and the prokaryotic VapBC toxin-antitoxin array. *Protein Eng. Des. Sel.* 24, 33–40. doi: 10.1093/protein/gzq081

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712. doi: 10.1126/science.1138140

Castaing, J. P., Nagy, A., Anantharaman, V., Aravind, L., and Ramamurthi, K. S. (2013). ATP hydrolysis by a domain related to translation factor GTPases drives polymerization of a static bacterial morphogenetic protein. *Proc. Natl. Acad. Sci. U.S.A.* 110, E151–E160.

Crisci, M. A., Chen, L. X., Devota, A. E., Borges, A. L., Bordlin, N., Sachdeva, R., et al. (2021). Wide distribution of alternatively coded Lak megaphages in animal microbiomes. *bioRxiv* [Preprint]. doi: 10.1101/2021.01.08.425732 bioRxiv: 2021.01.08.425732

Devoto, A. E., Santini, J. M., Olm, M. R., Anantharaman, K., Munk, P., Tung, J., et al. (2019). Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* 4, 693–700. doi: 10.1038/s41564-018-0338-9

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., et al. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359, eaar4120. doi: 10.1126/science.aar4120

Ershova, A. S., Rusinov, I. S., Spirin, S. A., Karyagina, A. S., and Alexeevski, A. V. (2015). Role of restriction-modification systems in prokaryotic evolution and ecology. *Biochemistry* 80, 1373–1386. doi: 10.1134/s0006297915100193

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–732. doi: 10.1038/nrmicro1235

Gogarten, J. P., and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687. doi: 10.1038/nrmicro1204

Gumerov, V. M., and Zhulin, I. B. (2020). TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Res.* 48, W72–W76.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hood, I. V., and Berger, J. M. (2016). Viral hijacking of a replicative helicase loader and its implications for helicase loading control and phage replication. *Elife* 5:e14158.

Iida, S., Streiff, M. B., Bickle, T. A., and Arber, W. (1987). Two DNA antirestriction systems of bacteriophage P1, darA, and darB: characterization of darA- phages. *Virology* 157, 156–166. doi: 10.1016/0042-6822(87)90324-2

Iyer, L. M., Leipe, D. D., Koonin, E. V., and Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* 146, 11–31. doi: 10.1016/j.jsb.2003.10.010

Ji, X., Sun, Y., Liu, J., Zhu, L., Guo, X., Lang, X., et al. (2016). A novel virulence-associated protein, vapE, in *Streptococcus suis* serotype 2. *Mol. Med. Rep.* 13, 2871–2877. doi: 10.3892/mmr.2016.4818

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kelleher, J. E., and Raleigh, E. A. (1994). Response to UV damage by four *Escherichia coli* K-12 restriction systems. *J. Bacteriol.* 176, 5888–5896. doi: 10.1128/jb.176.19.5888-5896.1994

Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S., and Sternberg, S. H. (2019). Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* 571, 219–225. doi: 10.1038/s41586-019-1323-z

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobián-Güemes, A. G., et al. (2016). Lytic to temperate switching of viral communities. *Nature* 531, 466–470.

Krishnan, A., Burroughs, A. M., Iyer, L. M., and Aravind, L. (2020). Comprehensive classification of ABC ATPases and their functional radiation in nucleoprotein dynamics and biological conflict systems. *Nucleic Acids Res.* 48, 10045–10075. doi: 10.1093/nar/gkaa726

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259.

Liu, J., Dehbi, M., Moeck, G., Arhin, F., Bauda, P., Bergeron, D., et al. (2004). Antimicrobial drug discovery through bacteriophage genomics. *Nat. Biotechnol.* 22, 185–191. doi: 10.1038/nbt932

Loenen, W. A. M., and Raleigh, E. A. (2014). The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res.* 42, 56–69. doi: 10.1093/nar/gkt747

Ma, Z., Geng, J., Yi, L., Xu, B., Jia, R., Li, Y., et al. (2013). Insight into the specific virulence related genes and toxin-antitoxin virulent pathogenicity islands in swine streptococcosis pathogen *Streptococcus equi* ssp. zooepidemicus strain ATCC35246. *BMC Genomics* 14:377. doi: 10.1186/1471-2164-14-377

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., and Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* 1:7.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., et al. (2019). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* 18, 67–83. doi: 10.1038/ s41579-019-0299-x

Matelska, D., Steczkiewicz, K., and Ginalski, K. (2017). Comprehensive classification of the PIN domain-like superfamily. *Nucleic Acids Res.* 45, 6995– 7020. doi: 10.1093/nar/gkx494

McInerney, J. O., McNally, A., and O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nat. Microbiol.* 2, 1–5.

Millman, A., Melamed, S., Amitai, G., and Sorek, R. (2020). Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat. Microbiol.* 5, 1608–1615. doi: 10.1038/s41564-020-0777-y

Neuwald, A. F., Aravind, L., Spouge, J. L., and Koonin, E. V. (1999). AAA+: a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* 9, 27–43.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu 300

Peters, J. E., Makarova, K. S., Shmakov, S., and Koonin, E. V. (2017). Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7358–E7366.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10. 1371/journal.pone.0009490

Roberts, G. A., Cooper, L. P., White, J. H., Su, T. J., Zipprich, J. T., Geary, P., et al. (2011). An investigation of the structural requirements for ATP hydrolysis and DNA cleavage by the EcoKI Type i DNA restriction and modification enzyme. *Nucleic Acids Res.* 39, 7667–7676. doi: 10.1093/nar/gkr480

Schneider, E., and Hunke, S. (1998). ATP-binding-cassette (ABC) transport systems: functional and structural aspects of the ATP-hydrolyzing subunits/domains. *FEMS Microbiol. Rev.* 22, 1–20. doi: 10.1111/j.1574-6976.1998.tb00358.x

Seidel, R., Bloom, J. G. P., Dekker, C., and Szczelkun, M. D. (2008). Motor step size and ATP coupling efficiency of the dsDNA translocase EcoR124I. *EMBO J.* 27, 1388–1398. doi: 10.1038/emboj.2008.69

Shah, M., Taylor, V. L., Bona, D., Tsao, Y., Stanley, S. Y., Pimentel-Elardo, S. M., et al. (2021). A phage-encoded anti-activator inhibits quorum sensing in *Pseudomonas aeruginosa. Mol. Cell* 81, 571–583.e6.

Shim, H. (2019). Feature learning of virus genome evolution with the nucleotide skip-gram neural network. *Evol. Bioinform.* 15:117693431882107. doi: 10.1177/ 1176934318821072

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* 30, 1335–1342. doi: 10.1038/ emboj.2011.41

Sonnhammer, E., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26, 320–322. doi: 10.1093/nar/26.1.320

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J. L., Makarova, K. S., Koonin, E. V., et al. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. *Science* 365, 48–53. doi: 10.1126/science.aax9181

Taylor, G. K., Petrucci, L. H., Lambert, A. R., Baxter, S. K., Jarjour, J., and Stoddard, B. L. (2012). LAHEDES: the LAGLIDADG homing endonuclease database and engineering server. *Nucleic Acids Res.* 40:W110.

Taylor, W. R. (1997). Residual colours: a proposal for aminochromography. *Protein Eng. Des. Sel.* 10, 743–746. doi: 10.1093/protein/10.7.743

Thomas, C. M., and Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–721. doi: 10.1038/nrmicro1234

Tock, M. R., and Dryden, D. T. F. (2005). The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* 8, 466–472. doi: 10.1016/j.mib.2005.06. 003

Tyson, G. W., and Banfield, J. F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* 10, 200–207.

Vasu, K., and Nagaraja, V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* 77, 53–72. doi: 10.1128/mmbr.00044-12

Wilkins, B. M. (2002). Plasmid promiscuity: meeting the challenge of DNA immigration control. *Environ. Microbiol.* 4, 495–500. doi: 10.1046/j.1462-2920. 2002.00332.x

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., et al. (2009). SUPERFAMILY – sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* 37:D380.

Yang, Q., and Catalano, C. E. (2003). Biochemical characterization of bacteriophage lambda genome packaging in vitro. *Virology* 305, 276–287. doi: 10.1006/viro. 2002.1602

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847