Comprehensive Genomic and Evolutionary Analysis of Biofilm Matrix Clusters and Proteins in the *Vibrio* Genus

Yiyan Yang[1], Jing Yan[2,3], Rich Olson[4], Xiaofang Jiang[1,*]

[1]Intramural Research Program, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA

[3]Quantitative Biology Institute, Yale University, New Haven, CT, USA

[4]Department of Molecular Biology and Biochemistry, Molecular Biophysics Program, Wesleyan University, Middletown, CT, USA

[*] Corresponding author: E-mail: xiaofang.jiang@nih.gov

Abstract

Vibrio cholerae pathogens cause cholera, an acute diarrheal disease resulting in significant morbidity and mortality worldwide. Biofilm formation by *V. cholerae* enhances its survival in natural ecosystems and facilitates transmission during cholera outbreaks. Critical components of the biofilm matrix are the *Vibrio* polysaccharide (VPS) produced by the *vps*-1 and *vps*-2 gene clusters, and biofilm matrix proteins encoded in the *rbm* cluster. However, the biofilm matrix clusters and associated matrix proteins in other Vibrio species remain under investigated, and their evolutionary patterns are largely unknown. In this study, we systematically annotated the biofilm matrix clusters across 6,121 Vibrio genomes, revealing their distribution, diversity, and evolution. We found that biofilm matrix clusters not only exist in *V. cholerae* but also in phylogenetically distant Vibrio species. Additionally, *vps*-1 clusters tend to co-locate with *rbmABC* genes, while *vps*-2 clusters are often adjacent to *rbmDEF* genes in various *Vibrio* species, which helps explain

the separation of these clusters by the *rbm* cluster in well-characterized *V. cholerae* strains. Evolutionary analysis of RbmC and Bap1 reveals that these two major biofilm matrix proteins are sequentially and structurally related and have undergone domain/modular alterations during their evolution. *RbmC* genes are more prevalent, while *bap1* likely resulted from an ancient duplication event of *rbmC* and is only present in a major clade of species containing *rbmC* counterparts. Notably, a novel loop-less Bap1 variant, identified in two subspecies clades of *V. cholerae*, was found to be associated with altered biofilm formation and the loss of antibiotic efflux pumps and chemotaxis. Another *rbm* cluster gene, *rbmB*, involved in biofilm dispersal, was found to share a common ancestor with *Vibrio* prophage pectin lyase-like tail proteins, indicating its functional and evolutionary linkages to Vibriophage proteins. In summary, our findings establish a foundational understanding of the proteins and gene clusters that contribute to Vibrio biofilm formation from an evolutionary perspective across a broad taxonomic scale. This knowledge paves the way for future strategies aimed at engineering and controlling biofilms through genetic modification.

Introduction

Vibrio cholerae, the pathogen responsible for cholera, causes an acute diarrheal disease that can lead to hypotonic shock and death. Annually, it infects 3-5 million people, resulting in 100,000–120,000 deaths (Charles and Ryan, 2011). *V. cholerae* forms biofilms—surface-associated communities encased in a matrix—which enhance survival in ecosystems and transmission during outbreaks (Donlan and Costerton, 2002; Colwell *et al.*, 2003), while providing protection from environmental stresses like nutrient scarcity, disinfectants, antimicrobial agents, predation by unicellular eukaryotes and attack by phages (Gupta *et al.*, 2018; Matz *et al.*, 2005; Beyhan and Yildiz, 2007).

The biofilm matrix is primarily comprised of Vibrio polysaccharide (VPS), making up approximately half its mass and essential for biofilm 3D structural development (Yildiz and Schoolnik, 1999; Yildiz *et al.*, 2014; Fong *et al.*, 2010). Genes involved in VPS production are organized into two *vps* clusters, *vps*-1 and *vps*-2. The *Vps*-1 cluster contains 12 genes (*vpsU* and *vpsA-K*) while the *vps*-2 cluster is relatively shorter only containing 6 genes (*vpsL-Q*) (Yildiz and Schoolnik, 1999; Fong *et al.*, 2010).

Meanwhile, biofilm matrix proteins, such as RbmA, RbmC and Bap1, are crucial for preserving the structural integrity of the wild-type biofilm (Fong *et al.*, 2006; Fong and Yildiz, 2007), among which RbmA and RbmC are encoded in a *rbm* (rugosity and biofilm structure modulator) cluster separating the two *vps* clusters. Genes encoding Bap1 are distant from the *rbm* cluster, yet they also modulate the development of corrugated colonies and are crucial for biofilm formation (Fong and Yildiz, 2007; Berk *et al.*, 2012). RbmA, as a biofilm scaffolding protein involved in cell-cell and cell-biofilm adhesion, is required for rugose colony formation and biofilm structure integrity in *V. cholerae* (Berk *et al.*, 2012; Fong and Yildiz, 2007; Absalon *et al.*, 2011; Maestre-Reyna *et al.*, 2013; Fong *et al.*, 2006). The other two major biofilm matrix proteins, RbmC and Bap1, are homologues sharing 47% sequence similarity and contain overlapping domains to facilitate their

robust adhesion to diverse surfaces (Fong and Yildiz, 2007; Huang *et al.*, 2023). Both proteins have a conserved β-propeller domain with eight blades and at least one β-prism domain. RbmC, however, is characterized by two β-prism domains and additional tandem β/γ crystallin domains, known as M1M2 (De *et al.*, 2018; Huang *et al.*, 2023). Most notably, Bap1's β-prism contains an additional 57-amino acid (aa) sequence which promotes *V. cholerae* biofilm adhesion to lipids and abiotic surfaces while RbmC mainly mediates binding to host surfaces through recognition of N- and O-glycans (Huang *et al.*, 2023). Another interesting gene in the *rbm* cluster is *rbmB*, which encodes a putative polysaccharide lyase and has been proposed to have a role in VPS degradation and cell detachment (Teschler *et al.*, 2015; Fong and Yildiz, 2007; Bridges *et al.*, 2020; Díaz-Pascual *et al.*, 2019). Together, the *vps*-1, *rbm* and *vps*-2 clusters comprise a functional genetic module — the *V. cholerae* biofilm matrix cluster (*V. cholerae* BMC or VcBMC) (Teschler *et al.*, 2015).

The biofilm matrix cluster has primarily been investigated in the commonly studied *V. cholerae* strains and a few other Vibrio species (Lilburn *et al.*, 2010; Guo and Rowe-Magnus, 2011; Chodur and Rowe-Magnus, 2018; Gao *et al.*, 2021). However, it has not yet been systematically studied at the strain level within *V. cholerae* or more extensively across the Vibrio genus. Since the biofilm matrix cluster encodes proteins for VPS synthesis and matrix proteins, which are the major components of Vibrio biofilms, a systematic genomic analysis of this cluster and the identification of relevant genes across the Vibrio genus can provide a prospective and comprehensive view of the phenotypes related to VPS production and biofilm formation in Vibrio.

In this study, we comprehensively annotated the genes involved in the biofilm matrix cluster to explore their distribution, diversity and gene synteny by conducting large-scale comparative genomics and phylogenetic analyses on 6,121 Vibrio genomes spanning 210 species across the entire Vibrio genus as well as within the *V. cholerae* species. We observed not only a prevalent presence of this cluster in *V. cholerae* but also in other distantly related species. Our analysis reveals a distinct evolutionary pattern for the *vps*-1 and *vps*-2 clusters: genes in the *vps*-2 cluster often co-located with *rbmDEF* genes, while *vps*-1 cluster genes are commonly adjacent to *rbmABC* genes. This suggests a functional relatedness between them and explains why these two *vps* clusters are separated by a *rbm* cluster in contemporary *V. cholerae* strains. Additionally, we inferred that the *bap1* genes originated as an ancient duplicate of *rbmC* in a clade of species closely related to *V. cholerae*, while *rbmC* genes are present in two major clades and may have undergone structural domain alterations throughout their evolutionary history. We identified a unique loop-less variant of the Bap1 protein, which lacks the adhesive 57aa loop. This variant is primarily found in two of the *V. cholerae* subspecies clades and is potentially associated with altered biofilm formation as well as the loss of antibiotic efflux pumps and chemotaxis towards chitin. Finally, our findings suggest that RbmB, a putative VPS degradation enzyme, are evolutionarily related to Vibriophage pectin lyase-like tail proteins. The systematic and accurate curation of biofilm matrix clusters and their proteins not only enhances our understanding of Vibrio biofilm formation from a genomic view but also offers insights for developing strategies to engineer and control biofilms.

Results

Biofilm matrix clusters are found in phylogenetically distant Vibrio species

Leveraging over 6,000 genomes from Genome Taxonomy Database (GTDB r214) (Parks *et al.*, 2022) across the Vibrio genus, we systematically annotated the proteins within the biofilm matrix clusters and depicted an overview of the cluster's gene occurrences spanning 209 Vibrio species and seven *V. cholerae* subspecies (Fig.1A). We defined a full biofilm matrix cluster if it contains the 12 *vps* genes (namely *vpsAB*, *vpsDEF*, *vpsIJK*, and *vpsLMNO*) whose deletions have been shown to cause a dramatic reduction in VPS production and biofilm formation (Fong *et al.*, 2010) and all of the *rbm* genes. We reconstructed a Vibrio species tree, which shares a similar topology to that in a previous study (Lin *et al.*, 2018), and mapped the presence and absence of the 12 *vps* genes and all *rbm* genes to the tree tips. It is interesting to discover that, under this criterion, the full biofilm matrix clusters not only exist in *V. cholerae* and closely related species (such as *V. metoecus* and *V. mimicus*) but are also sporadically distributed across the Vibrio genus in distant species like *V. anguillarum*, *V. ordalii*, *V. aestuarianus*, *V. coralliilyticus*, *V. neptunius* and *V. cortegadensis*. Among all genes, *vps* genes in the *vps-2* cluster are the most prevalent genes with *vpsL* existing in 50% of the species, *vpsM* in 41.2%, *vpsN* in 58.3% and *vpsQ* in 64.4% following by *vps*-1 cluster genes *vpsA* (33.3%) and *vpsB* (33.8%). The higher prevalence of *vps*-2 cluster genes is due to the identification of *vps*-2 similar loci in our data, such as the *cps* (capsular polysaccharide) locus in *Vibrio parahaemolyticus*, the *wcr* (capsular and rugose polysaccharide) locus in *Vibrio vulnificus*, and *vps*-2-like loci in *Aliivibrio fischeri*, all of which contain homologs of *vpsLMNO* (Supplementary Figure 1) (Smith and Siebeling, 2003; Güvener and McCarter, 2003; Grau *et al.*, 2008; Darnell *et al.*, 2008; Yildiz and Visick, 2009). It is important to note that these loci contain genes associated with functions other than VPS production in biofilms, such as capsular polysaccharide synthesis. Therefore, they are less likely to represent true *vps*-2 clusters and are instead designated as *vps*-2 similar clusters in this study.

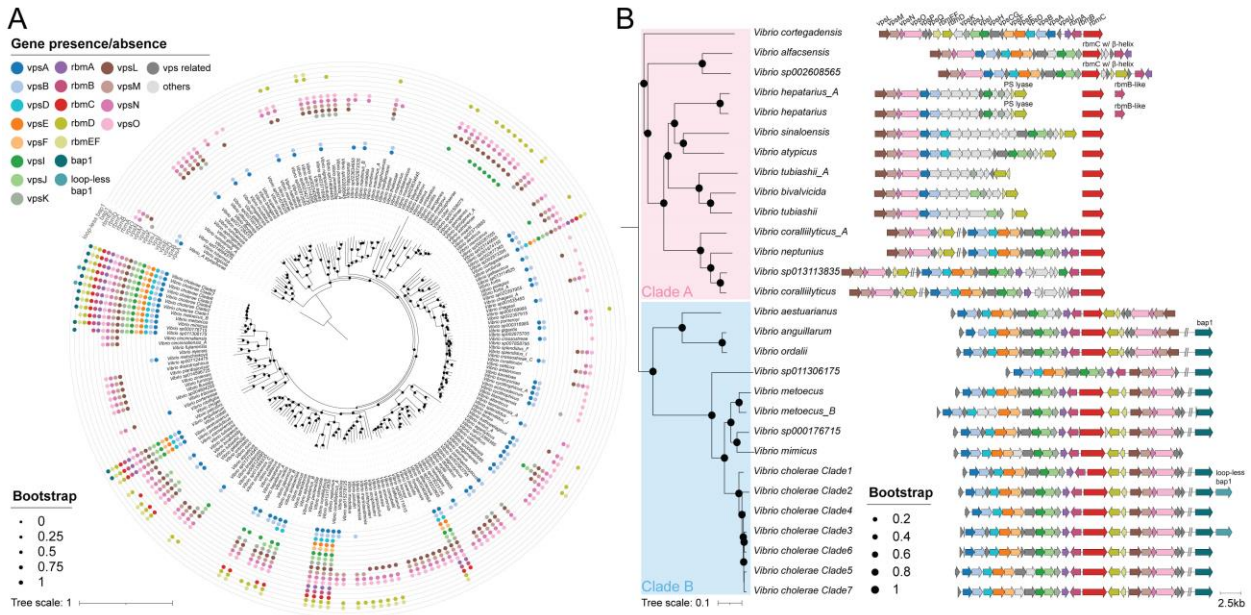


Figure 1. The distribution of biofilm matrix clusters across the Vibrio genus. (A) The phylogenomic tree with the presence and absence of important genes in biofilm matrix clusters mapped to tips which represent 216 Vibrio (sub)species. The tree was rooted with the representative genome of *Vibrio_A stylophorae* species (NCBI Assembly accession=GCA_921293875.1). (B) Gene syntenies for biofilm matrix clusters in 29 (sub)species that possess biofilm matrix protein encoding genes (*rbmC* and/or *bap1*) are illustrated using the same color palette as in panel A and the phylogenomic tree displayed is a subtree derived from the tree in panel A. The clusters are aligned with each other using *rbmC* gene as the anchor. Genes not connected with a horizontal line are found in different contigs, whereas genes separated by the "//" symbol are found in the same contig but are hundreds of genes away from each other. The *rbmE* and *rbmF* genes are combined under the single gene name *rbmEF* due to overlaps in their gene sequences and frequent annotations as a single gene. Similarly, the *vpsC* and *vpsG* genes are merged into one gene name, *vpsCG*, as they both share a highly similar domain. PS: Polysaccharide.

We next investigated the gene synteny within the biofilm matrix cluster to gain insights on how the *vps-1*, *vps-2* and *rbm* clusters have evolved during Vibrio speciation. Figure 1B and Supplementary Figure 2 illustrates the gene syntenies of full and partial biofilm matrix clusters that contain at least one *rbmC* or *bap1* gene in 29 Vibrio (sub)species representative genomes. The Vibrio (sub)species clearly form two major clades, Clades A and B, each of which is featured with distinct patterns in the biofilm matrix clusters. The examination of the isolation sources and potential hosts of Vibrio species in these clades indicates that Clade A species are primarily isolated from marine water and from healthy or diseased invertebrates such as prawns, corals, and bivalve mollusks like clams and oysters. In contrast, species in Clade B are mostly found in seawater and brackish waters, inhabiting both invertebrate and vertebrate hosts, including fish (such as *V. aestuarianus*, *V. ordalii*, and *V. anguillarum*) and humans (such as *V. metoecus*, *V. mimicus*, and *V. cholerae*), often acting as pathogens (Supplementary Table 1).

First, we observed that *rbmA* genes are absent in seven Vibrio species from Clade A (namely *V. hepatarius_A*, *V. hepatarius*, *V. sinaloensis*, *V. atypicus*, *V. tubiashii_A*, *V. tubiashii*, and *V. bivalvicida*) despite the presence of *rbmD* and *rbmEF* genes in the same operon and the presence of distant *rbmC* genes. Although these species are phylogenetically distant, we observed conservation in the neighborhoods of their *rbmC*. These *rbmC* genes are often immediately adjacent to a gene containing a methyl-accepting chemotaxis domain and are close to an operon encoding a system for the uptake and metabolism of disaccharides, suggesting their potential involvement in sugar binding process (Supplementary Figure 3 and Supplementary Table 2). These species typically possess several, but not all, *vps*-2-like and *vps*-1-like genes. For genes not annotated as *vps*-like genes, most of them are glycosyltransferases, acyltransferases and polysaccharide biosynthesis proteins, which are responsible for the synthesis, modification and export of VPS (Supplementary Figure 2 and Supplementary Table 3).

Secondly, we noticed that *vps*-1 clusters tend to co-locate with *rbmABC* genes, while *vps*-2 clusters consistently pair with *rbmDEF* genes. Although *vps*-1 clusters are much less prevalent than *vps*-2 clusters, any species having a full or nearly full *vps*-1 cluster tend to have the full set of *rbmABC* genes. In addition, it is intriguing to discover *vps*-2 cluster and *rbmDEF* genes underwent a process from separation to co-location. For majority of the species in Clade A, *vps*-2 clusters and *rbmDEF* genes are disconnected. However, in a subclade containing *V. coralliilyticus*, *V. coralliilyticus_A*, *V. neptunius*, and *V. sp013113835* species, they are joined but separate from *vps*-1 clusters and *rbmABC* genes (Fig.1B). In contrast, in Clade B, the *vps*-2 cluster and *rbmDEF* genes are adjacent to each other while also connecting with the *vps*-1 cluster and *rbmABC* genes, forming an intact biofilm matrix cluster. However, the *vps*-2 cluster is inverted in *V. aestuarianus*, *V. anguillarum*, and *V. ordalii* compared to the other species in this clade. Taken together, the co-locations of *vps*-1 cluster with *rbmABC* and *vps*-2 cluster with *rbmDEF* in several Clade A species suggest their respective functional connections. This may explain the organization of the intact biofilm matrix clusters commonly observed in Clade B, where two *vps* clusters are separated by a *rbm* cluster with all *rbm* genes closely clustered together.

Lastly, we observed that *bap1* genes are exclusively found in *V. cholerae* and its closely related species within Clade B. Upon examining the neighboring genes of *bap1*, we identified a duplicate *bap1* gene directly adjacent to the standard *bap1*, only existing in two clades within the *V. cholerae* species (Fig.1B).

Biofilm matrix proteins RbmC and Bap1 experienced structural domain alterations during evolution

Functioning as two of the major biofilm matrix proteins in Vibrio biofilms and sharing 47% identity in sequences and overlapping domains in their structures, RbmC and Bap1 are functionally and evolutionarily related (Fong and Yildiz, 2007; Huang *et al.*, 2023). We compiled a data set consisting of 2,004 *rbmC* and 2,062 *bap1* genes identified across the Vibrio genus and examined their origin and divergence. For the standard Bap1 protein, it contains an 8-bladed β-propeller domain and a β-prism domain, while for the standard RbmC protein, it has an 8-bladed β-propeller,

two β-prism and two extra β/γ-crystallin domains (i.e., M1 and M2 based on the nomenclature used in our recent work) (Huang *et al.*, 2023). It is worth noting that Bap1's β-prism domain contains a 57aa loop that has been shown to function in nonspecific adherence to abiotic surfaces and/or lipid membranes (Huang *et al.*, 2023). Therefore, in this study, we first extracted an initial set of genes with sequence similarity ≥ 40% and bit score ≥ 250 with the standard *rbmC* (GenBank accession: WP_000200580.1) and *bap1* (GenBank accession: WP_001881639.1) amino acid sequences from all genes annotated in the Vibrio genus genomes. Genes possessing only a single β-propeller and a single β-prism domain were categorized as encoding putative Bap1 proteins. Conversely, genes containing additional domains beyond a β-propeller and a β-prism were categorized as encoding putative RbmC proteins. Five genes from the genomes of *V. alfacsensis* and *V. sp002608565* species that failed initial classification drew our attention. Their protein sequences exhibit approximately 43% and 52% similarity with the standard *rbmC* and *bap1*, respectively. Notably, their predicted structures contain only a single β-propeller and a β-helix domain. Investigation of these genomes' gene syntenies showed that they possess nearly complete biofilm matrix clusters, where the genes with β-helix domains are located in positions typically associated with *rbmC* genes (thus labeled as "*rbmC* w/ β-helix" in Fig.1B). Consequently, the encoded proteins are classified as a RbmC variant in this study.

Through a deeper examination of these genes' structural features, we have identified two extra RbmC variants as well as one Bap1 variant (Supplementary Figures 4 and 5). The RbmC variants differ from the standard RbmC protein by having none or only one of the two mucin-binding domains and are therefore called M1M2-less or partial M1M2 RbmC, respectively. Most of the M1M2-less RbmC (59%) and partial M1M2 RbmC (85%) proteins were found to have signal peptides, indicating that they indeed lost the domains rather than being truncated proteins. The Bap1 variant is the protein encoded by the *bap1* duplicate we previously mentioned. It is surprising to observe that the Bap1 variant shares all the domains with standard Bap1 but lacks the 57aa loop in the β-prism domain and was named loop-less Bap1. Taken together, we identified a total of six structural groups representing different protein variants: RbmC with β-helix, M1M2-less RbmC, partial M1M2 RbmC, standard RbmC, standard Bap1 and loop-less Bap1 (Fig.2A). Next, after sequence redundancy removal, a codon-based phylogenetic tree was constructed. The phylogeny indicates that the RbmC and Bap1 form two distinct clades, and the long branch connecting them suggests their distant divergence. Protein sequences from the same structural group typically cluster together, although there are exceptions. For instance, a group of genes encoding M1M2-less RbmC is exclusively found in *V. cholerae* and nested within the largest standard RbmC clade, while genes for loop-less Bap1 fall into a subclade within the standard Bap1 clade (Fig.2A). Taking this phylogenetic information into consideration, we have further divided all the protein sequences into eight protein groups: RbmC with β-helix, M1M2-less RbmC, M1M2-less RbmC in *V. cholerae*, partial M1M2 RbmC, RbmC clade 1, RbmC clade 2, Bap1 clade, and loop-less Bap1 (Fig.2A).

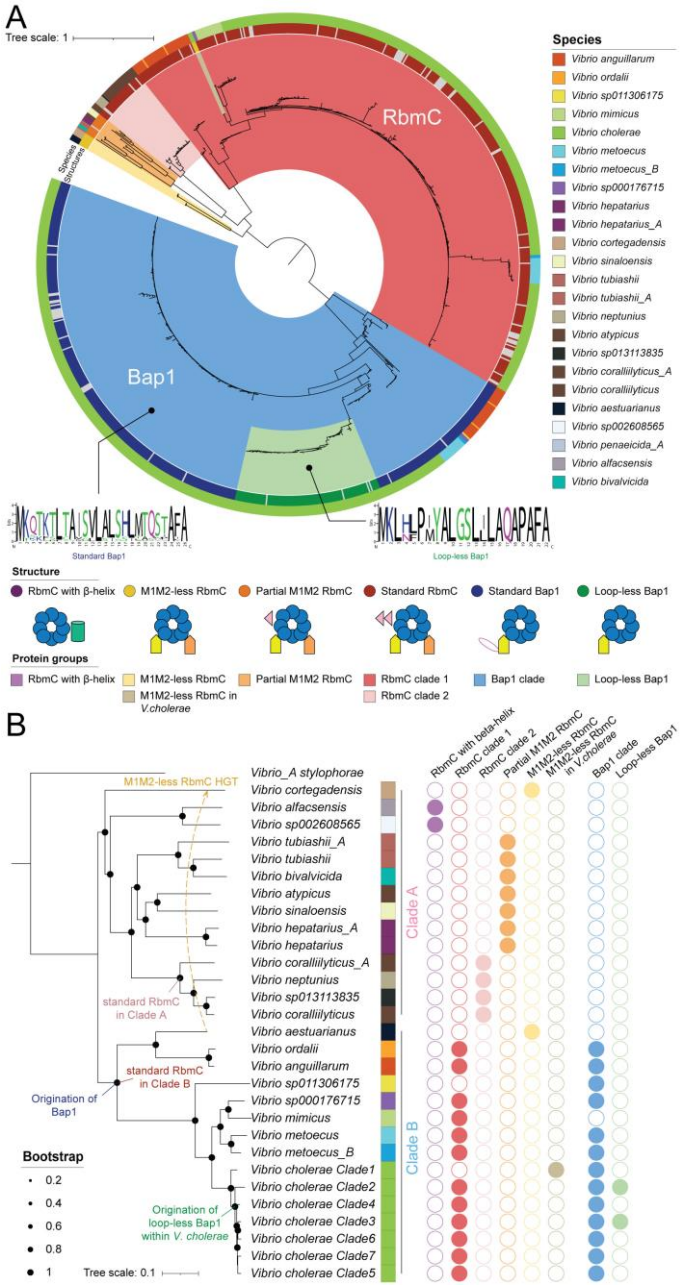


Figure 2. The gene tree and evolutionary analysis for RbmC and Bap1 proteins. (A) The gene tree was built with non-redundant codon sequences of 514 RbmC and 483 Bap1 proteins, which is rooted at the midpoint. The outer circle indicates the species of origin, while the inner circle indicates the protein structural features with grey representing truncated proteins. The cartoons at the bottom demonstrate the domain composition for the corresponding structures. Color ranges indicate different protein groups based on both structural features and phylogenetic relationships, whose legend was put under the corresponding structural features. Note that the RbmC with a β-helix domain was omitted from the gene tree due to it causing a poor multiple sequence alignment. The sequence logos for the signal peptides are shown for Bap1 clade and loop-less Bap1 clade. (B) The distribution of 9 protein groups along the phylogenomic tree suggests the evolutionary events

for *rbmC* and *bap1* genes. The tree replicates the one in Fig.1B while retaining the outgroup species. The species and protein group colors are consistent with those in panel A.

Next, we mapped these protein groups onto the Vibrio species tree tips to infer their evolutionary events. The eight protein groups demonstrated distinct patterns between Clades A and B (Fig.2B). Genes encoding all kinds of RbmC variants are observed across the species in Clade A, but no Bap1 encoded genes are found, suggesting that RbmC have undergone a series of alterations in the M1M2 domains and a β-helix domain replacing the original M1M2 and β-prisms domains during evolution. Genes encoding standard RbmC are prevalent in Clade B, in contrast to their restricted presence in a subclade of Clade A. Genes for Bap1 are also found exclusively in Clade B, suggesting that Bap1 genes originated at the ancestral node of this clade. The phylogenetic analysis of the β-propeller domains suggests that Bap1 may have diverged from the ancestor of standard RbmC in both Clade A and Clade B (Supplementary Figure 6). Additionally, it has been reported that the sequence of Bap1's β-prism diverges from the β-prisms in RbmC (De *et al.*, 2018), and our analysis further shows that Bap1's β-prism domains are closer to RbmC's first β-prism domain (β-prism C1) than to the second (β-prism C2), sharing the most recent common ancestor with RbmC's first β-prism domains exclusively in Clade A (Supplementary Figure 7). In addition, the genes encoding loop-less Bap1 are likely to originate from a *V. cholerae* lineage within Clade B. A horizontal gene transfer event (HGT) of genes encoding M1M2-less RbmC was observed from *V. cortegadensis* species in Clade B to *V. aestuarianus* species in Clade A, both of which can live in marine environments and use bivalve mollusks, such as clams and oysters, as hosts, thus possibly facilitating the HGT (Supplementary Table 1). We inferred this to be a result of horizontal gene transfer because the genes encoding M1M2-less RbmC, while phylogenetically closest (Fig. 2A), are found in two distantly related species in the Vibrio species tree (Fig. 2B). Interestingly, the biofilm matrix clusters in the genomes of these two species are similar yet slightly differ in the direction and location of the *rbmABC* genes relative to other genes in this cluster (Fig. 1B). The *V. cortegadensis* species is likely to be the HGT recipient because its gene synteny of the biofilm matrix cluster is quite different from those in other species of Clade B, indicating that this species may have acquired the gene cluster from an external source outside Clade B. Additionally, the lack of M1M2 domains in RbmC proteins from *Vibrio cholerae* Clade 1 is likely the result of a domain loss event in standard RbmC proteins, as indicated by their formation of a distinct subclade within RbmC clade 1 in the gene tree (Fig. 2A).

Loop-less Bap1 positive *V. cholerae* strains are associated with altered biofilm formation and the loss of antibiotic efflux pumps and chemotaxis towards chitin

In previous sections, we described a Bap1 variant which is deficient in the 57aa sticky loop in the β-prism domain and encoded by a duplicated gene located directly adjacent to the standard *bap1* gene. The comparison of the predicted structures and sequences between Bap1 and the loop-less variant demonstrated that these two proteins are highly similar in both structures (TM-score=0.8020) and sequences (identity=78.5%) (Supplementary Figure 4E-F). Despite of the lack of a loop, the loop-less Bap1 is thought to remain functional and likely to be a secretory protein

282 due to a 22aa signal peptide found at its N-terminus. The signal peptide differs in sequence pattern
283 and peptide length from that of the standard Bap1, whose signal peptide is 26aa (Fig.2A).

284 To explore potential functions of loop-less Bap1, we analyzed its gene distribution in the *V.*
285 *cholerae* subspecies tree (Fig.3A). The phylogeny shows that *V. cholerae* is divided into seven
286 distinct sub-species clades, where the loop-less Bap1 encoded genes are enriched in Clades 2 and
287 3, and few of them are scattered in Clade 5. Given that the genes are distributed across multiple
288 interspersed clades, the presence of loop-less Bap1 in these clades may not be the result of simple
289 sub-speciation. Instead, it could reflect independent strategies adopted by individual clades to
290 enhance their fitness. Since there are no significant differences in the habitats of these two clades
291 compared to others, we treated the presence or absence of loop-less Bap1 proteins in genomes as
292 distinct phenotypes and subsequently conducted genotype-phenotype association analysis to
293 uncover gene groups related to these phenotypes, aiming to understand the functional impacts of
294 introducing loop-less Bap1.

295

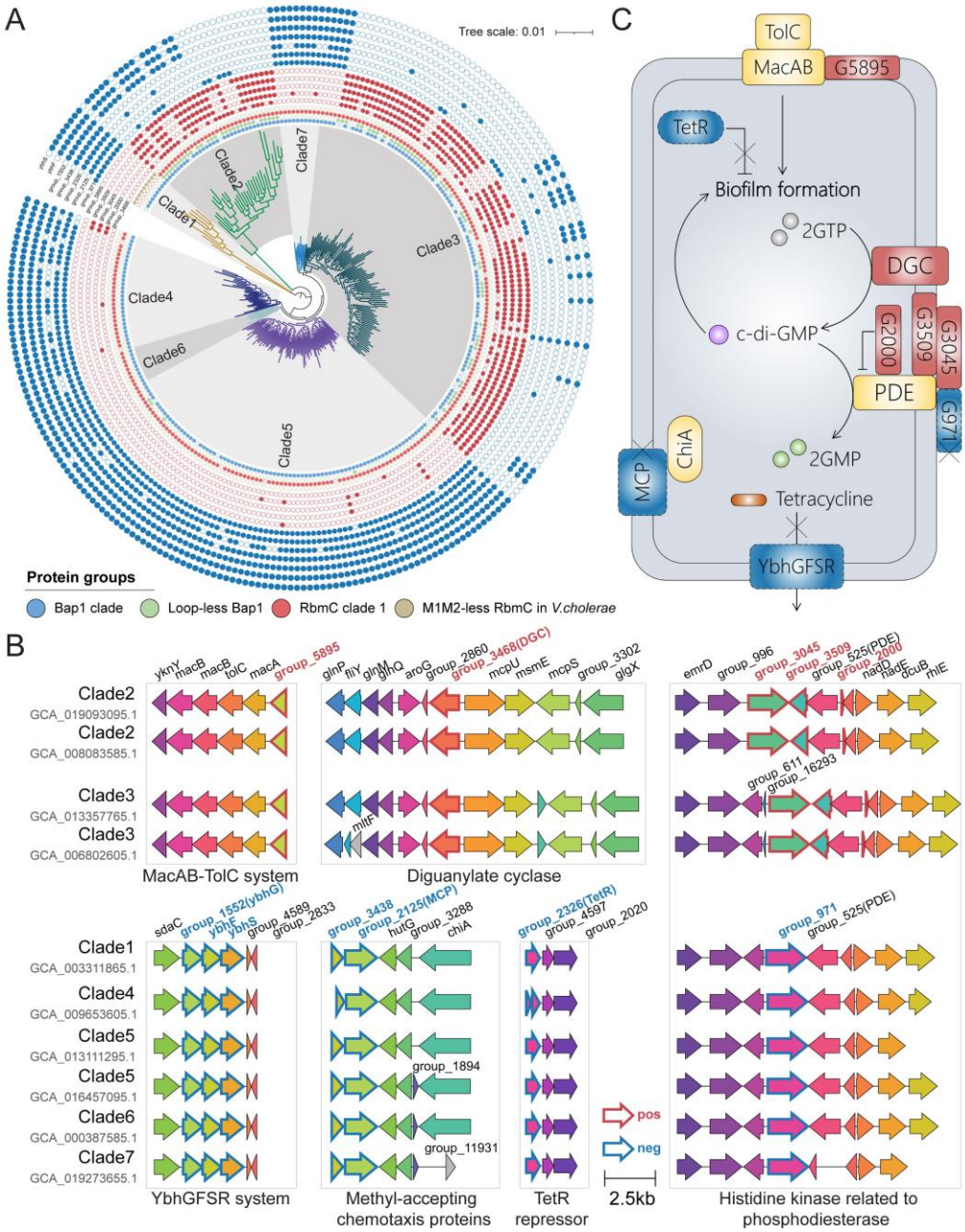


Figure 3. Loop-less Bap1 encoded genes are enriched in two *V. cholerae* clades, which are associated with the presence of gene groups related to biofilm formation and the absence of genes groups related to antibiotic efflux pumps and chemotaxis. (A) The phylogenomic tree for *V. cholerae* species was built with protein sequences from the core genes found by Roary (Page *et al.*, 2015). The tree was rooted at Clade 1. The presence and absence of RbmC/Bap1 variants (inner circles, using the same palette in Fig.2) and gene groups either positively (red) or negatively (blue) associated with loop-less Bap1-positive strains (outer circles) are mapped to the tips. (B) Gene syntenies for associated gene groups in ten genomes selected from seven clades. They are highlighted by thicker red or blue borders to indicate their positive or negative associations, respectively. Genes in the same boxes are colored by gene clusters sharing more than 80%

sequence similarities. (C) A schematic-diagram proposed to demonstrate the positively (colored in red with solid lines as borders) and negatively (colored in blue with dashed lines as borders) associated gene groups in the loop-less Bap1-positive strains. Relevant gene groups are colored in yellow. G5895 stands for group_5895, and G2000 stands for group_2000 and so on. DGC: Diguanylate cyclase; PDE: Phosphodiesterase; MCP: Methyl-accepting chemotaxis protein; GTP: Guanosine-5'-triphosphate; GMP: Guanosine monophosphate.

As a result, we identified five positively and seven negatively associated gene groups, which demonstrate nearly identical and opposite presence/absence patterns to that of the loop-less Bap1, respectively (Fig.3A). Ten of the 12 gene groups, except for *ybhF* and *yhbS*, are under-studied and assigned unknown groups by the pan genome analysis software Roary (Page *et al.*, 2015). To annotate these gene groups, we predicted their domains and examined their neighboring genes (Supplementary Table 4). Among the positively associated gene groups, group_5895 is a gene group potentially annotated as a sensor domain in periplasmic binding protein-like II family (SUPERFAMILY: SSF53850), which is often located immediately upstream a MacAB-TolC-like operon containing two *macB* genes (Fig.3B). As another positively associated gene group, group_3468 is annotated as diguanylate cyclase (DGC) with a GGDEF domain (Pfam: PF00990) and often flanked by a glutamate transporter operon and methyl-accepting chemotaxis-related proteins in Clades 2 and 3 genomes (Fig.3B).

Compared to positively associated gene groups, we identified more negatively associated ones, among which group_1552, *ybhF* and *ybhS* are frequently organized together in the genomes and may function in a gene cluster. Because group_1552 is predicted to encode an HlyD family secretion protein (Pfam: PF13437, SUPERFAMILY: SSF111369) and close to *ybhF* and *ybhS* genes, it is highly likely that group_1552 is a gene group representing the *ybhG* genes, which also belong to the HlyD_D23 protein family (Yamanaka *et al.*, 2016). Meanwhile, *sdaC* gene group, often located directly upstream, is annotated as tryptophan/tyrosine permease family (Pfam: PF03222), thus potentially acting similarly to *ybhR* as a multidrug ABC transporter permease (Feng *et al.*, 2020). Group_3438 and group_2125 are also co-localized in an operon and negatively associated with the presence of loop-less Bap1 (Fig.3B). Although no protein domain is detected for group_3438, group_2125 is predicted to have a methyl-accepting chemotaxis protein (MCP) signaling domain (Pfam: PF00015). These two gene groups are located next to an operon encoding a chitinase (*chiA*), an enzyme to degrade chitin which is often found in the exoskeleton of zooplankton and other crustaceans and serves as a sole carbon source for *V. cholerae* (Li and Roseman, 2004; Meibom *et al.*, 2004; Drescher *et al.*, 2014). Although more than one or no chitinase has been found in about half of the genomes, in the remaining genomes, the only existing chitinase is the one close to group_3438 and group_2125, indicating that these gene groups are associated with the main functional chitinase. Group_2326 is a gene group predicted to possess bacterial regulatory proteins, TetR family (Pfam: PF00440) and tetracyclin repressor-like, C-terminal domain (Pfam: PF14514), probably functioning as a TetR repressor (Fig.3B).

Lastly, an operon that captured our attention includes three gene groups and one gene group exclusively found in loop-less Bap1-positive and negative strains, respectively, while the synteny of other genes in the operon remain largely unchanged (Fig.3B). Positively associated group_3045

348 and negatively associated group_971 are both predicted as putative histidine kinases since they
349 have histidine kinase-/DNA gyrase B-/HSP90-like ATPase domain (Pfam: PF02518) and a
350 periplasmic sensor domain often found in signal transduction proteins (Pfam: PF17149). However,
351 group_3045 is accompanied by two other positively associated gene groups, group_3509
352 (SUPERFAMILY: SSF53850, Periplasmic binding protein-like II) and group_2000 (no domain
353 found). These three gene groups are positioned around group_525, which is annotated as a c-di-
354 GMP phosphodiesterase (PDE) (PANTHER: PTHR45228). This enzyme functions to break down
355 c-di-GMP, thereby reducing its levels and inhibiting the biofilm formation process (Christen *et al.*,
356 2005; Hengge, 2009). Meanwhile, the negatively associated group_971, also located adjacent to
357 group_525, exclusively occurs in loop-less Bap1 negative strains, suggesting a different role in
358 regulating c-di-GMP phosphodiesterase activity (Fig.3B).

359 To explain the functional changes associated with gene groups in loop-less Bap1 positive strains,
360 we propose a model (Fig.3C). Our model suggests that these strains preferentially retain genes that
361 regulate the MacAB-TolC-like system and c-di-GMP levels, leading to altered Vibrio biofilm
362 formation. A recent study showed that MacAB-TolC system is involved in the envelope stress
363 response and adaptation to deleterious conditions occurring in mature biofilms of *Acinetobacter*
364 *baumannii* (Robin *et al.*, 2022), suggesting a similar role in *Vibrio cholerae* which also belongs to
365 the Pseudomonadota phylum. We therefore conjecture that group_5895, often located next to the
366 operon and potentially functioning as a sensor, collaborates with the MacAB-TolC system to
367 facilitate biofilm formation in *V. cholerae*. As for group_3468, it may function as a DGC which is
368 responsible for the synthesis of c-di-GMP (Whiteley and Lee, 2015), and elevated levels of this
369 molecule are well known for suppressing motility and promoting sessility and biofilm formation
370 in bacteria (Russell *et al.*, 2013; Liu *et al.*, 2022). Although no domain has been predicted for
371 group_2000, it represents a set of genes encoding small proteins, each around 36 amino acids in
372 length, and predicted to fold into an α-helix shape. Small proteins have been shown to associate
373 with larger membrane proteins to regulate their levels or activities. Examples include the 30-amino
374 acid PmrR protein found in *Salmonella* and the more broadly distributed 49-amino acid AcrZ and
375 37-amino acid SgrT proteins (Gray *et al.*, 2022; Yadavalli and Yuan, 2022). Potential protein-
376 protein interaction predicted by AlphaFold-Multimer (Abramson *et al.*, 2024) suggest that
377 group_2000 may interact with the HD-GYP domain of PDE (group_525) (Supplementary Figure
378 8). Taken together, we hypothesize that group_3045, functioning as histidine kinases, group_3509
379 as periplasmic binding proteins, and group 2000 as small proteins, collectively replace another
380 negatively associated histidine kinase gene group, group_971, in loop-less Bap1 positive strains.
381 This replacement may regulate PDE (group_525) activity in a different way, subsequently affecting
382 c-di-GMP levels in *V. cholerae* cells.

383 On the other hand, the model hypothesizes that strains may lose redundant capabilities, such as
384 antibiotic resistance, chemotaxis towards carbon sources, and biofilm suppression, particularly
385 when bacterial cells are protected by an altered biofilm involving loop-less Bap1. Group_1552,
386 *ybhF*, *ybhS* along with their adjacent *sdaC* genes are often co-located in an operon encoding a
387 YbhGFSR-like efflux pump, which has been recently characterized to export tetracycline
388 antibiotics, including tetracycline, oxytetracycline, chlortetracycline, and doxycycline, in *E. coli*
389 (Feng *et al.*, 2020). Notably, the tetracycline antibiotic class has long been the most effective for

treating cholera despite the increasing and prevalent antimicrobial resistance to this class of antibiotics in *Vibrio cholerae* (Dengo-Baloi *et al.*, 2017; Yildiz and Schoolnik, 1999; Kumar *et al.*, 2012). However, bacterial biofilms have been reported to enhance bacterial cells' tolerance to antibiotics (Gupta *et al.*, 2018; Høiby *et al.*, 2010). Therefore, the absence of these tetracycline-resistant gene groups in loop-less Bap1-positive strains may suggest that the altered biofilm matrix exhibits different structures that further enhance the cells' resistance to tetracycline antibiotics, thus resulting in the loss of tetracycline resistance-related efflux pumps in these strains. Similarly, the absence of MCP signaling-related groups, group_3438 and group_2125, located near the chitinases, indicates a lack of chemotaxis towards carbon sources and suggests that the cells are more likely to remain in a sessile state. Meanwhile, group_2326 is likely to act as Tet repressors, playing a role in the transcriptional control of several cellular processes, including biofilm formation and antibiotic resistance in bacteria (Teschler *et al.*, 2015). A previous study reported that the deletion of a TetR repressor named *brpT* resulted in a significant increase in biofilm formation in *Streptococcus sanguinis* (Liu *et al.*, 2017). Consequently, we suggest that the absence of group_2326 in loop-less Bap1 positive strains may lead to enhanced, unrestrained biofilm formation. It is noted that these are proposed hypotheses based on our observation while the real situations might be much more complicated and need further investigation and experimental validation.

Given that prophages have been reported to influence biofilm formation in pathogens including species in *Vibrio* genus (Rice *et al.*, 2009; Wang *et al.*, 2023; Tan *et al.*, 2020), we investigated prophage integration in the genomes of *V. cholerae* subspecies. Interestingly, apart from the differences in gene groups, we observed a significantly smaller number of detected prophage regions in the genomes of loop-less Bap1-positive strains compared to negative ones (one-sided, two-sample Wilcoxon rank sum test, p-value = 2.9e-09) (Supplementary Figure 9). This reveals an unprecedented correlation between prophage integration and the presence of loop-less Bap1, suggesting that prophages may also play a role in the formation of altered biofilms in these strains.

RbmB is evolutionarily related to Vibrio prophage pectin lyase-like tail proteins

RbmB, a gene flanked by *rbmA* and *rbmC* but with a different transcriptional direction in the *rbm* cluster, encodes a putative polysaccharide lyase, RbmB, that plays an important part in VPS degradation and cell detachment (Fong and Yildiz, 2007; Díaz-Pascual *et al.*, 2019; Bridges *et al.*, 2020). Given its great potential in biofilm dispersal and control, the identification of RbmB proteins is crucial and can improve our understanding of how and when *V. cholerae* cells disperse from a biofilm. By integrating both gene synteny and structural information, we confidently identified *rbmB* genes when a gene is predicted to have a single-stranded right-handed β-helix/pectin lyase domain (SUPERFAMILY: SSF51126) and is within an 8-gene distance from either a *rbmC* or *rbmA* gene. It turns out that *rbmB* genes make up only 23.4% of the 7,532 genes encoding the pectin lyase-like domain across the Vibrio genus, raising our curiosity about the source and relationships of the other genes with *rbmB*. Particularly, given the well-documented role of pectin lyase-like domains in breaking down polysaccharides (Burnim *et al.*, 2024) and their presence in some Vibriophage tail depolymerases, which facilitate the degradation of Vibrio

biofilms (Cevallos-Urena *et al.*, 2023), we are exploring the possibility that RbmB is evolutionarily related to Vibriophage proteins. To address the abovementioned questions, we constructed a gene tree for all Vibrio proteins predicted to have the single-stranded right-handed β-helix/pectin lyase-like domains (Fig.4A). We observed that more than half of the genes (56.1%) are unidentified non-RbmB-encoded genes, and 28.2% are putative pectate lyases. The third largest gene group comprises RbmB-encoded genes (N=319, highlighted in pink), forming a monophyly in the gene tree. The top five species to which these genes belong are *V. cholerae* (N=225), *V. mimicus* (N=20), *V. coralliilyticus* (N=19), *V. metoecus* (N=15) and *V. anguillarum* (N=12) species. Genes in this group have a median length of 408 amino acids and possess signal peptides. This group is closely related to a sister group consisting of 21 non-RbmB-encoded genes (highlighted in yellow). Together, the two groups are part of a larger clade that includes a large outgroup of 143 non-RbmB-encoded genes (highlighted in blue) (Fig.4B). Both groups of 21- and 143-non-RbmB-encoded genes exhibit high structural similarity and moderate sequence similarity to those of the RbmB group, suggesting their close evolutionary relationship (Fig.4C).

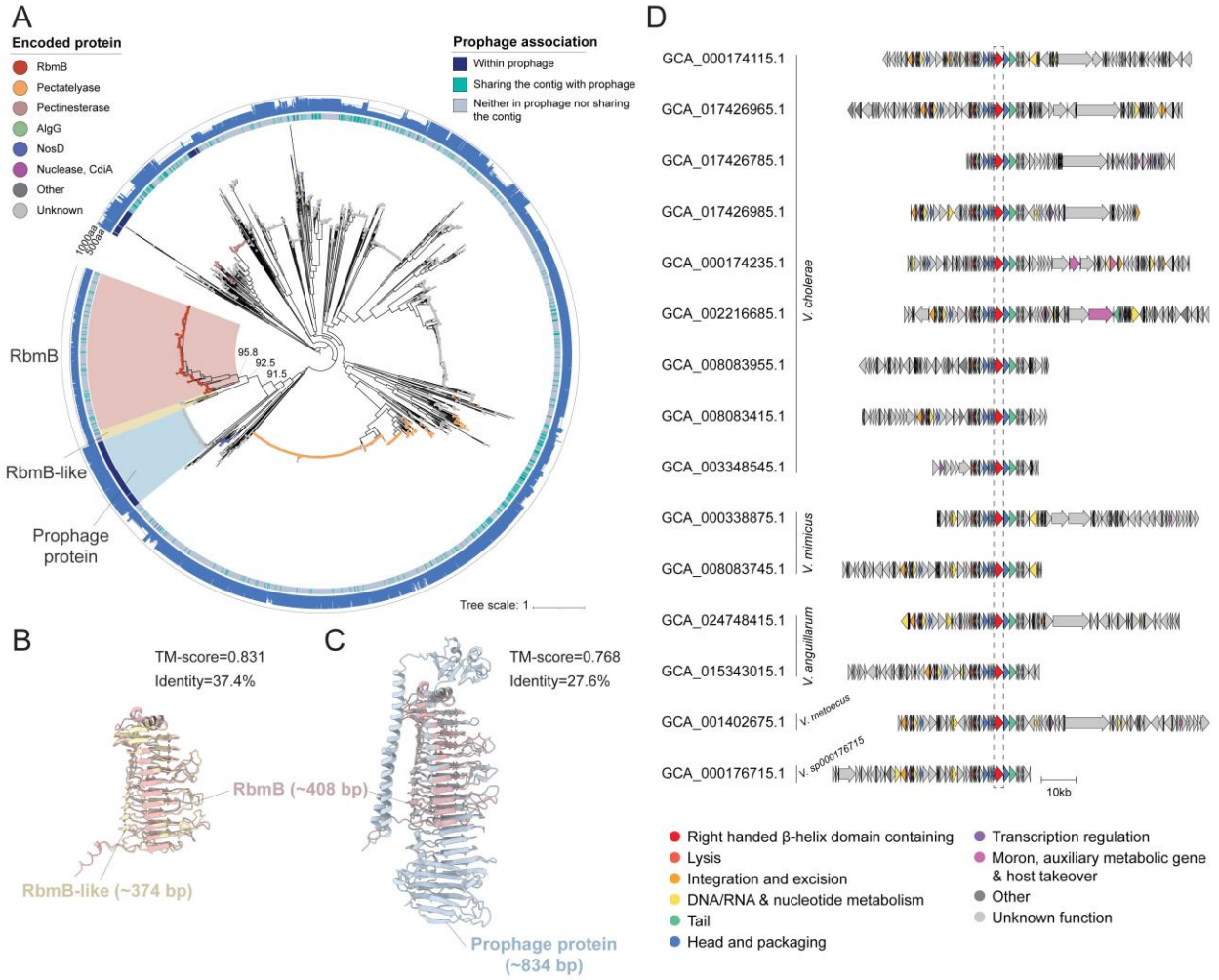


Figure 4. Single-stranded right-handed β-helix domain containing gene tree suggests an association between RbmB and prophage proteins. (A) The gene tree was built with non-

redundant protein sequences containing single stranded right-handed β-helix domains (SUPERFAMILY: SSF51126) and was rooted at the midpoint. Encoded proteins are annotated as colored dots at tips. The inner circle represents the associations of the genes with the prophages found in the same contigs, while the outer circle represents the gene lengths. Bootstrap values are shown at three key internal nodes. The color ranges highlight the clades for RbmB encoded genes (pink), RbmB-like encoded genes (yellow) and prophage-related genes (blue). (B-C) Pairwise superimposition of predicted protein structures. The structures displayed are for RbmB (colored pink, gene accession: GCA_013111535.1_02619), RbmB-like (colored yellow, gene accession: GCA_002284395.1_03257), and prophage proteins (colored blue, gene accession: GCA_002097735.1_02038). The signal peptides were removed from RbmB and RbmB-like proteins and the structures were predicted by AlphaFold3 (Abramson *et al.*, 2024). (D) Gene syntenies for the 15 representative prophages that possess single-stranded right-handed β-helix domain containing genes. Each gene synteny is accompanied by the genome accessions from which the prophage fragment was found. Genes encoding the single-stranded right-handed β-helix domain are colored red, while other genes are colored according to phage functional categories. AlgG: Mannuronan C5-epimerase; NosD: Putative ABC transporter binding protein.

The 21 non-RbmB encoded genes belong to *V. cholerae* (N=9), *V. anguillarum* (N=6), *V. hepatarius* (N=2), *V. hepatarius_A* (N=2) and *V. mimicus* (N=2) species, with a median gene length of 374 amino acids and possessing signal peptides. Thirteen out of the 21 genomes containing these genes also host confidently curated *rbmB* genes, located hundreds of genes away, and all these genomes additionally contains *rbmC* genes. Taken together, we believe that these genes encode secretory proteins that are functionally different from the real *rbmB* and are named *rbmB*-like genes in this study. To explore their possible functions, we further investigated their gene neighbors. The 21 genes showed distinct roles in different species. In *V. hepatarius* and *V. hepatarius_A*, *rbmB*-like genes are located immediately downstream of a gene encoding a peptidase family C69-like protein, a nuclease complex SbcCD operon, and a CAI-1 autoinducer sensor kinase (CqsS) (see *rbmB*-like genes in Fig.1B and Supplementary Table 5). It is also interesting to find that, although *V. hepatarius* and *V. hepatarius_A* species don't have real *rbmB* genes near their biofilm matrix clusters, they instead include putative polysaccharide lyases with β-jelly roll domains in the cluster, which might serve as RbmB alternatives for biofilm dispersal (Supplementary Figure 2 and Supplementary Table 5). As for *V. anguillarum*, the *rbmB*-like genes are flanked by ectABC and proVWX operons, which are responsible for the synthesis and transporter system of ectoine – a cyclic amino acid essential for the growth of *V. anguillarum* under cold stress (Ma *et al.*, 2017). For the remaining species, *V. cholerae* and *V. mimicus*, these genes are mostly surrounded by unknown genes but are sometimes accompanied by genes encoding N-acyltransferases such as *lpxD*, *yiaC*, and *aaaT* (Supplementary Table 5). However, this hasn't thoroughly surveyed and thus required further studies in the future.

On the other hand, the majority of the 143 non-RbmB encoded genes are from *V. cholerae* (N=124), while the remaining are from *V. mimicus* (N=8), *V. anguillarum* (N=6), *V. metoecus* (N=4) and *V. sp000176715* (N=1) species, with a median gene length of 834 amino acids and lacking signal peptides. One hundred and twenty-six of the 143 genomes containing these genes possess confidently curated *rbmB* genes, which are far from these genes, and all the genomes, except for

one, also host *rbmC* or *bap1* genes. Strikingly, we found that 142 of 143 the genes are in the prophage regions. For the only one gene not detected in any prophage regions in the same contig, it is likely due to that this gene is the sole gene in the contig, which is relatively short and only 2,667 base pairs long. Gene synteny analysis demonstrated the similarity in the locations of the genes in the 15 representative prophage genomes, where they are situated between two head and packing function-related genes and close to a tail protein (Fig.4D). In addition, BLASTp results showed that all of the 143 genes' best hits (Camacho *et al.*, 2009) share around 30% identity with the tail fiber protein in Vibrio phage vB_VchM_Kuja (GeneBank accession: MN718199) when queried against the Infrastructure for a PHAge Reference Database (INPHARED, accessed on August 15th, 2024) (Cook *et al.*, 2021), suggesting these genes may also function as part of the phage tail fibers (Supplementary Table 6). Based on the phylogenetic relationships between RbmB, RbmB-like, and prophage pectin lyase-like proteins, we infer that they are derived from a common ancestor, with the prophage proteins diverging before the split of the RbmB and RbmB-like proteins. The longer branches of prophage proteins also indicate their faster evolution, a typical feature of phage proteins. Overall, our finding marks the first time that RbmB has been demonstrated to evolutionarily related to Vibriophage pectin lyase-like tail proteins, thus expanding our understanding of their genetic and functional connections.

Discussion

Bacterial biofilms play a vital role as a lifestyle niche for bacteria in natural environments. They also represent a significant health hazard due to their contribution to persistent infections and the contamination of medical equipment (Donlan, 2016; Hall-Stoodley *et al.*, 2004; Costerton *et al.*, 1999; Flemming *et al.*, 2016). Despite their importance in bacterial survival and the challenges they pose in clinical settings, the organization and evolution of the genes encoding the components in biofilm-related clusters have not been extensively studied. A deeper genomic and phylogenetic understanding of these clusters and genes is crucial for the development of innovative genetic engineering strategies that target biofilm-surface interactions and offer alternatives to antibiotic treatments. In this study, using *Vibrio cholerae*—the causative agent of pandemic cholera and a model organism for biofilm studies (Nelson *et al.*, 2009; Teschler *et al.*, 2015) as well as other related species in the Vibrio genus as examples, we propose a framework that integrates comparative genomics, phylogeny, gene synteny analysis and structure prediction to thoroughly characterize biofilm matrix clusters and related proteins, a methodology that can be extended to the study of the biofilm associated clusters and proteins in other bacterial species including important pathogens. This approach has also allowed us to identify domain and modular changes in proteins across their evolutionary timelines, revealing the commonality of domain alterations in Vibrio biofilm matrix proteins and their potential implications for biofilm development.

Among our significant findings is the identification of a Bap1 variant lacking the 57aa loop, referred to as loop-less Bap1. This variant has garnered interest due to its predicted association with altered biofilm formation, decreased antibiotic efflux, and reduced mobility. Additionally, strains positive for loop-less Bap1 contain significantly fewer prophages compared to negative

529 strains. This observation may follow a similar mechanism reported in *Vibrio anguillarum*, where
530 enhanced biofilm formation and a reduced number of prophages are coupled at low cell density,
531 mediated by quorum-sensing signaling (Tan *et al.*, 2020). While these findings are currently
532 based on computational analysis, we anticipate future experimental studies to validate them. For
533 instance, investigating how the deletion of loop-less *bap1* genes impacts biofilm morphology,
534 antibiotic susceptibility, and prophage induction in Vibrio species will further deepen our
535 understanding of biofilm dynamics, resistance mechanisms and phage-host interactions in these
536 bacteria.

537 As an alternative to combating antibiotic resistance and biofilm formation in Vibrio pathogens,
538 phage therapies are increasingly attracting attention. Notably, phage host-receptor binding proteins,
539 typically tail fibers or tailspikes, are recognized for their ability to cleave polysaccharides such as
540 VPS of biofilms (Yen *et al.*, 2017; Jensen *et al.*, 2006; Bhandare *et al.*, 2019; Barman *et al.*, 2022;
541 Yang *et al.*, 2024). Concurrently, *rbmB* genes, encoding RbmB proteins involved in biofilm
542 disassembly, demonstrate significant potential for controlling biofilms and potentially serve as a
543 promising approach to combat Vibrio infections. Interestingly, RbmB proteins and phage tail
544 proteins both feature a common domain—the single-stranded right-handed β-helix/pectin lyase-
545 like domain—suggesting a potential functional link. However, the evolutionary relationship
546 between these proteins has remained elusive. Here, we reveal for the first time that RbmB proteins,
547 along with a group of RbmC-like proteins, share a more recent common ancestor with prophage
548 pectin lyase-like tail proteins than with other pectin lyase-like domain containing proteins. This
549 comprehensive annotation of RbmB in Vibrio species, coupled with the Vibrio prophage pectin
550 lyase-like tail proteins, could establish a foundation for a biofilm degrader pool, paving the way
551 for novel protein-based therapies to effectively and precisely target biofilms in emerging Vibrio
552 pathogens.

553 Our findings clarify numerous aspects of the *Vibrio* biofilm matrix cluster while also raising new
554 questions. We have conducted a comprehensive search for this cluster in the existing genomes
555 across the Vibrio genus, yet for those species with only partial biofilm matrix clusters, it remains
556 uncertain whether there are other gene clusters co-function to produce VPS – or if they produce
557 VPS at all. Similarly, the proteins involved in species lacking this cluster, and their organizational
558 structures, are yet to be fully understood. It is also interesting to explore whether there are
559 polysaccharide lyases or glycosidic hydrolases, aside from RbmB, that could help bacterial cells
560 escape from the biofilm during dispersal. For instance, while RbmB-like proteins are present in *V.*
561 *hepatarius_A* and *V. hepatarius*, their effectiveness in biofilm disassembly is questionable due to
562 their remote location from other *vps* and *rbm* genes. Instead, polysaccharide lyases containing β-
563 jelly roll domains within the biofilm matrix-like cluster may assume this role. Further experimental
564 work is needed to understand how variations in RbmC and Bap1 influence biofilm assembly and
565 the extent to which changes in a single domain/module can impact Vibrio phenotypes.

566

Methods

Curation of the biofilm matrix cluster

We downloaded 6,121 genomes classified by GTDB r214 (Genome Taxonomy Database) (Parks *et al.*, 2022) as Vibrio and Vibrio_A species from NCBI assembly database (Kitts *et al.*, 2016) (accessed on February 18th, 2024) (Supplementary Data 1). Genomes were annotated by Prokka v1.14.6 (Seemann, 2014) with default parameters. KofamScan (https://github.com/takaram/kofam_scan) (Aramaki *et al.*, 2020) and InterProScan v5.63-95.0 (Jones *et al.*, 2014) (with options "-t p -iprlookup --goterms --pathways" and chunksize of 400) were applied to assign KEGG ortholog and predict domains for the genes with default parameters. These genomes along with their gene protein files (.faa), annotation files (.gff) and kofam annotation files (.kofam.tsv) were used as input for ProkFunFind (https://github.com/nlm-irp-jianglab/ProkFunFind) (Dufault-Thompson and Jiang, 2024) to detect potential biofilm matrix clusters. To prepare the queries for the biofilm matrix protein encoded genes, we have collected a set of KEGG orthologs (i.e. KOfam) covering all *vps* genes as well as the *rbmA* gene from Kyoto Encyclopedia of Genes and Genomes (KEGG) database (https://www.genome.jp/kegg/) (Kanehisa *et al.*, 2017). We have also composed a hmm profile for all the *rbm* genes. Any clusters of genes containing more than four of the *vps* or *rbm* cluster genes and having less than 18 genes between the furthest gene pair were assigned a cluster ID as a potential biofilm matrix-associated cluster. The *rbmA*, *rbmB*, *rbmC* and *bap1* as well *vpsE* and *vpsF* genes in an output gene annotation file (.gff) was further recognized and curated in the following section, to generate a refined gene annotation file. The configuration file for ProkFunFind, KOFam list and hmm profile files are provided at https://github.com/nlm-irp-jianglab/ProkFunFind and https://zenodo.org/doi/10.5281/zenodo.11509588. The refined gene annotation output obtained from ProkFunFind is available in Supplementary Data 2.

Curation and classification of the biofilm matrix proteins RbmC and Bap1

Since Bap1 shares over 40% sequence identity with RbmC, traditional sequence-based computational approaches often perform poorly to distinguish them. Furthermore, these two proteins are usually annotated as hemolysin-like proteins by NCBI genome annotation pipeline, yet they only share less than 40% identity in the single β-prism domain with hemolysins. Another example lies in the initial scanning of ProkFunFind where both *rbmC* and *bap1* genes have been identified as *rbmC* using hmm profile-based search. Nevertheless, RbmC and Bap1 consist of well-studied domains, which inspires us to leverage structural information to distinguish them. First, 4,066 potential RbmC and Bap1 encoded sequences were obtained by querying WP_000200580.1 (RbmC) and WP_001881639.1 (Bap1) against all protein sequences in *Vibrio* genomes using BLASTp v2.15.0+ (Camacho *et al.*, 2009), with criteria of > 40% identity, > 250 bit score, and > 200 amino acids in aligned length. Next, to better perform multiple sequence alignment (MSA), after removing sequence redundancy we excluded the five RbmC with β-helix encoded genes and only selected high-quality RbmC and Bap1 encoded genes. High-quality genes are genes with ≥

80% identity with a Bap1 query and ranging from 650-700aa in length or with ≥ 80% identity with a RbmC query and ranging from 950-1000aa in length, both with bit scores > 900, while the remaining are classified as low-quality genes. We applied MAFFT v7.475 (Katoh, 2002) to align high-quality protein sequences with options "--maxiterate 1000 --localpair" and aligned low-quality protein sequences by adding them to the previously aligned high-quality genes using MAFFT with option "-add". The aligned protein sequences were mapped back to the nucleotide sequences to align by codons using PAL2NAL v14 (Suyama *et al.*, 2006). Finally, a codon-based phylogenetic tree was built with the aligned nucleotide sequences using RAxML v8.2.12 (Stamatakis, 2006) by providing a partition file ("-m GTRGAMMA -q dna12_3.partition.txt"), based on which the genes were initially classified as RbmC or Bap1 encoded. The detailed structural classification was performed according to the presence and absence of domains in both sequences and structures (Supplementary Data 3-4). The domain boundaries were manually determined by investigating the MSA in Geneious Prime v2023.1.2 (https://www.geneious.com) and double checked with the predicted structures obtained from ESMfold v2.0.0 (Lin *et al.*, 2023) (Supplementary Data 5). All gene syntenies were annotated using Clinker v0.0.28 (Gilchrist and Chooi, 2021).

Curation of RbmB, RbmA, VpsE and VspF proteins

We composed a confident set of *rbmB* genes by first including any genes within an eight-gene distance of either a curated *rbmC* or a putative *rbmA* gene that possess a single-stranded right-handed β-helix domain (SUPERFAMILY: SSF51126) or are annotated as *rbmB* by hidden Markov model (HMM) search. Since *rbmA* genes haven't been thoroughly curated, the neighboring *vps* and *rbm* genes of identified *rbmB* genes adjacent only to a putative *rbmA* gene were manually reviewed to determine if they are real *rbmB* genes. Additionally, ten *rbmB* genes were added to the set because they share over 60% sequence identity and cover more than 90% of the alignment with *rbmB* genes in the confident set. The gene context and the presence of *rbmC* in the same genomes were examined to support the likelihood that these genes are real *rbmB* genes but are not connected to other *rbm* genes due to poor genome assembly and sequencing quality.

Likewise, we curated genes as *rbmA* genes if they are within a nine-gene distance of either a curated *rbmB* or a curated *rbmC* gene, as confirmed in previous sections, that possess two fibronectin type III domains (Gene3D: 2.60.40.3880) or are annotated as *rbmA* by hidden Markov model (HMM) search. For genes located distantly from any *rbmB* or *rbmC* genes but having two fibronectin type III domains, we only included them to the *rbmA* gene set if they, as well as the *rbmB* or *rbmC* genes in the same genomes, are on the edge of contigs, indicating a break in the contig. Regarding genes possessing fewer than two fibronectin type III domains but close to a *rbmB* or *rbmC*, we annotated them as *rbmA* only if they are split into multiple smaller genes or fragmented due to poor genome assembly.

We have cautiously annotated *vpsE* and *vpsF*, as they encode the Wzy-polymerase (VpsE) and Wzx-flippase (VpsF) in the *vps*-1 cluster (Schwechheimer *et al.*, 2020), indicating their important roles in the Wzy/Wzx-dependent VPS synthesis pathway. Any genes within a *vps* gene context that are predicted to be polysaccharide biosynthesis proteins (Pfam: PF13440) and have a

polysaccharide biosynthesis C-terminal domain (Pfam: PF14667) or are identified as VpsF family polysaccharide biosynthesis proteins (NCBIfam: NF038256), are regarded as *vpsE* or *vpsF*, respectively. Split and fragmented genes, which only have part or none of the domains, were manually annotated and added if they are close to a well-annotated *vpsF*/*vpsE*.

The gene sequences and typing information in this section are provided as Supplementary Data 6-9.

Selection of Vibrio species representative genomes

We didn't simply use the GTDB representative genomes for the 210 Vibrio species in this study. Although the representative genomes generally have high completeness and low contamination, they might have fragmented biofilm matrix clusters and don't necessarily have the matrix proteins due to genome assembly issues. To take this into consideration, we developed a strategy to pick representative genomes which have maximally reflected the biofilm matrix cluster status at the Vibrio species levels. For the 23 species whose genomes possess *rbmC* and/or *bap1* genes, we manually selected the representative genomes to have the most intact biofilm matrix proteins as well as the untruncated RbmC/Bap1 proteins and are representative of the gene synteny of the biofilm matrix cluster in the species. For 73 species in which no biofilm matrix cluster associated proteins was detected, their GTDB representative genomes were used. For the remaining 114 species, 76 of them have multiple genomes. We ranked the genomes in each species higher if they have 1) fewer contigs, implying they have less fragmented contigs, 2) more key *vps*-1 and *vps*-2 genes in the same gene cluster, and 3) more curated *rbm* or *bap1* genes. The genomes meeting these criteria best were selected as the representatives, while the genomes in the 38 single-genome species were picked as species representatives. The final 216 representative genomes for Vibrio species and *V. cholerae* subspecies are provided as Supplementary Data 10.

Pan-genome analysis of *Vibrio cholerae*

A total of 194 core genes were detected and aligned in 1663 *V. cholerae* genomes by pan-genome analysis using the Roary v3.13.0 with options "-i 90 -cd 90 -g 500000 -s -e --mafft" (Page *et al.*, 2015). The core gene alignment of a subset of 273 representative genomes with completeness > 90% and contamination < 5% was leveraged to build a phylogenomic tree using FastTree v2.1.11 with default options (Price *et al.*, 2010) (Supplementary Data 11). The seven clade representative genomes within *V. cholerae* species, which have intact biofilm matrix clusters and *rbmC*/*bap1* genes, were randomly picked for the corresponding clades.

Construction of phylogenomic *Vibrio* species tree

We applied PIRATE v1.0.5 to the 209 Vibrio species representative genomes (excluding *V. cholerae*) and seven *V. cholerae* subspecies representative genomes to obtain genus-wise marker genes (with options "-k '--diamond'") (Bayliss *et al.*, 2019). PIRATE can rapidly create pangenomes from coding sequences over a wide range of amino acid identity thresholds, thus

recognizing the most robust set of core genes. The core gene nucleotide alignment provided by PIRATE was used to build the Vibrio species tree using FastTree v2.1.11 with options "-gtr -nt" (Supplementary Data 12).

Identification of loop-less Bap1 positive strains associated gene groups

Given the *V. cholerae* phylogenomic tree, the presence and absence of the gene groups defined by Roary (Supplementary Data 13) and the existence of loop-less Bap1 as the positive phenotype for genomes, we ran Evolink (https://github.com/nlm-irp-jianglab/Evolink) (Yang and Jiang, 2023) to find three positively and six negatively associated gene groups related to loop-less Bap1 presence. Extra two positively and one negatively associated gene groups were further added since they usually co-function in the same operons with the significantly associated gene groups.

Signal peptide detection

Signal peptides were predicted for RbmC and Bap1-related proteins using SignalP6.0 server (https://services.healthtech.dtu.dk/services/SignalP-6.0/) (Teufel *et al.*, 2022). The signal peptides were aligned with MAFFT v7.475 (Katoh, 2002) and visualized as sequence logo using WebLogo server (https://weblogo.berkeley.edu/logo.cgi) (Crooks *et al.*, 2004) (Supplementary Data 14).

Construction of gene and domain trees

After removing sequence redundancy, single-stranded right-handed β-helix domain containing protein sequences were aligned using MAFFT-DASH (Rozewicki *et al.*, 2019) to take structural alignment into consideration. The multiple sequence alignment was next trimmed using TrimAl v 1.2rev59 (Capella-Gutiérrez *et al.*, 2009) to obtain cleaner MSA and used to reconstruct their phylogeny using FastTree v2.1.11 with default options (Price *et al.*, 2010).

The β-propeller and β-prism domains sequences were extracted based on domain segmentation of RbmC and Bap1 proteins. The alignment using MAFFT v7.475 (Katoh, 2002) were used to build trees using FastTree v2.1.11 with default options (Price *et al.*, 2010). All trees were visualized and annotated with iTOL v6 server (https://itol.embl.de/) (Letunic and Bork, 2024).

The tree files were provided as Supplementary Data 15-17.

Prophage regions identification

Prophage regions in genomes were detected using VirSorter v2.2.4 (Guo *et al.*, 2021) with options "--min-length 1000" (Supplementary Data 18). Phage genes within the determined prophage regions were annotated and categorized using Pharokka v1.3.2 (Bouras *et al.*, 2023).

Data and code availability

The data underlying this article can be accessed through Zenodo (https://zenodo.org/doi/10.5281/zenodo.11509588). All scripts utilized throughout the publication can be accessed through the main branch on the GitHub repository (https://github.com/YiyanYang0728/Vibrio_biofilm_matrix_cluster).

Reference

Abramson,J. *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*.

Absalon,C. *et al.* (2011) A Communal Bacterial Adhesin Anchors Biofilm and Bystander Cells to Surfaces. *PLoS Pathog*, **7**, e1002210.

Aramaki,T. *et al.* (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, **36**, 2251–2252.

Barman,R.K. *et al.* (2022) Screening of Potential Vibrio cholerae Bacteriophages for Cholera Therapy: A Comparative Genomic Approach. *Front. Microbiol.*, **13**, 803933.

Bayliss,S.C. *et al.* (2019) PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*, **8**, giz119.

Berk,V. *et al.* (2012) Molecular Architecture and Assembly Principles of *Vibrio cholerae* Biofilms. *Science*, **337**, 236–239.

Beyhan,S. and Yildiz,F.H. (2007) Smooth to rugose phase variation in *Vibrio cholerae* can be mediated by a single nucleotide change that targets c-di-GMP signalling pathway. *Molecular Microbiology*, **63**, 995–1007.

Bhandare,S. *et al.* (2019) Reviving Phage Therapy for the Treatment of Cholera. *The Journal of Infectious Diseases*, **219**, 786–794.

Bouras,G. *et al.* (2023) Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics*, **39**, btac776.

Bridges,A.A. *et al.* (2020) Identification of signaling pathways, matrix-digestion enzymes, and motility components controlling *Vibrio cholerae* biofilm dispersal. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 32639–32647.

Burnim,A.A. *et al.* (2024) The three-sided right-handed β-helix is a versatile fold for glycan interactions. *Glycobiology*, **34**, cwae037.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Capella-Gutiérrez,S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Cevallos-Urena,A. *et al.* (2023) Vibrio-infecting bacteriophages and their potential to control biofilm. *Food Sci Biotechnol*, **32**, 1719–1727.

Charles,R.C. and Ryan,E.T. (2011) Cholera in the 21st century: *Current Opinion in Infectious Diseases*, **24**, 472–477.

Chodur,D.M. and Rowe-Magnus,D.A. (2018) Complex Control of a Genomic Island Governing Biofilm and Rugose Colony Development in Vibrio vulnificus. *J Bacteriol*, **200**.

Christen,M. *et al.* (2005) Identification and Characterization of a Cyclic di-GMP-specific Phosphodiesterase and Its Allosteric Control by GTP. *Journal of Biological Chemistry*, **280**, 30829–30837.

755 Colwell,R.R. *et al.* (2003) Reduction of cholera in Bangladeshi villages by simple filtration. *Proc. Natl.*
756 *Acad. Sci. U.S.A.*, **100**, 1051–1055.
757 Cook,R. *et al.* (2021) INfrastructure for a PHAge REference Database: Identification of Large-Scale
758 Biases in the Current Collection of Cultured Phage Genomes. *PHAGE*, **2**, 214–223.
759 Costerton,J.W. *et al.* (1999) Bacterial Biofilms: A Common Cause of Persistent Infections. *Science*, **284**,
760 1318–1322.
761 Crooks,G.E. *et al.* (2004) WebLogo: A Sequence Logo Generator. *Genome Res.*, **14**, 1188–1190.
762 Darnell,C.L. *et al.* (2008) The Putative Hybrid Sensor Kinase SypF Coordinates Biofilm Formation in
763 *Vibrio fischeri* by Acting Upstream of Two Response Regulators, SypG and VpsR. *J Bacteriol*,
764 **190**, 4941–4950.
765 De,S. *et al.* (2018) Structural basis of mammalian glycan targeting by Vibrio cholerae cytolysin and
766 biofilm proteins. *PLoS Pathog*, **14**, e1006841.
767 Dengo-Baloi,L.C. *et al.* (2017) Antibiotics resistance in El Tor Vibrio cholerae 01 isolated during cholera
768 outbreaks in Mozambique from 2012 to 2015. *PLoS ONE*, **12**, e0181496.
769 Díaz-Pascual,F. *et al.* (2019) Breakdown of Vibrio cholerae biofilm architecture induced by antibiotics
770 disrupts community barrier function. *Nat Microbiol*, **4**, 2136–2145.
771 Donlan,R.M. (2016) Microbial Biofilms, Second Edition. *Emerg. Infect. Dis.*, **22**, 1142–1142.
772 Donlan,R.M. and Costerton,J.W. (2002) Biofilms: Survival Mechanisms of Clinically Relevant
773 Microorganisms. *Clin Microbiol Rev*, **15**, 167–193.
774 Drescher,K. *et al.* (2014) Solutions to the Public Goods Dilemma in Bacterial Biofilms. *Current Biology*,
775 **24**, 50–55.
776 Dufault-Thompson,K. and Jiang,X. (2024) Annotating microbial functions with ProkFunFind. *mSystems*,
777 **9**, e00036-24.
778 Feng,Z. *et al.* (2020) A Putative Efflux Transporter of the ABC Family, YbhFSR, in Escherichia coli
779 Functions in Tetracycline Efflux and Na+(Li+)/H+ Transport. *Front. Microbiol.*, **11**, 556.
780 Flemming,H.-C. *et al.* (2016) Biofilms: an emergent form of bacterial life. *Nat Rev Microbiol*, **14**, 563–
781 575.
782 Fong,J.C.N. *et al.* (2006) Identification and Characterization of RbmA, a Novel Protein Required for the
783 Development of Rugose Colony Morphology and Biofilm Structure in *Vibrio cholerae*. *J*
784 *Bacteriol*, **188**, 1049–1059.
785 Fong,J.C.N. *et al.* (2010) Role of Vibrio polysaccharide (vps) genes in VPS production, biofilm formation
786 and Vibrio cholerae pathogenesis. *Microbiology*, **156**, 2757–2769.
787 Fong,J.C.N. and Yildiz,F.H. (2007) The *rbmBCDEF* Gene Cluster Modulates Development of Rugose
788 Colony Morphology and Biofilm Formation in *Vibrio cholerae*. *J Bacteriol*, **189**, 2319–2330.
789 Gao,C. *et al.* (2021) Coral mucus rapidly induces chemokinesis and genome-wide transcriptional shifts
790 toward early pathogenesis in a bacterial coral pathogen. *The ISME Journal*, **15**, 3668–3682.
791 Gilchrist,C.L.M. and Chooi,Y.-H. (2021) clinker & clustermap.js: automatic generation of gene cluster
792 comparison figures. *Bioinformatics*, **37**, 2473–2475.
793 Grau,B.L. *et al.* (2008) Further Characterization of *Vibrio vulnificus* Rugose Variants and Identification of
794 a Capsular and Rugose Exopolysaccharide Gene Cluster. *Infect Immun*, **76**, 1485–1497.
795 Gray,T. *et al.* (2022) Small Proteins; Big Questions. *J Bacteriol*, **204**, e00341-21.
796 Guo,J. *et al.* (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and
797 RNA viruses. *Microbiome*, **9**, 37.
798 Guo,Y. and Rowe-Magnus,D.A. (2011) Overlapping and unique contributions of two conserved
799 polysaccharide loci in governing distinct survival phenotypes in *Vibrio vulnificus*. *Environmental*
800 *Microbiology*, **13**, 2888–2990.
801 Gupta,P. *et al.* (2018) Increased antibiotic resistance exhibited by the biofilm of Vibrio cholerae O139.
802 *Journal of Antimicrobial Chemotherapy*, **73**, 1841–1847.
803 Güvener,Z.T. and McCarter,L.L. (2003) Multiple Regulators Control Capsular Polysaccharide Production
804 in *Vibrio parahaemolyticus*. *J Bacteriol*, **185**, 5431–5441.

805 Hall-Stoodley,L. *et al.* (2004) Bacterial biofilms: from the Natural environment to infectious diseases. *Nat*
806 *Rev Microbiol*, **2**, 95–108.
807 Hengge,R. (2009) Principles of c-di-GMP signalling in bacteria. *Nat Rev Microbiol*, **7**, 263–273.
808 Høiby,N. *et al.* (2010) Antibiotic resistance of bacterial biofilms. *International Journal of Antimicrobial*
809 *Agents*, **35**, 322–332.
810 Huang,X. *et al.* (2023) Vibrio cholerae biofilms use modular adhesins with glycan-targeting and
811 nonspecific surface binding domains for colonization. *Nat Commun*, **14**, 2104.
812 Jensen,M.A. *et al.* (2006) Modeling the role of bacteriophage in the control of cholera outbreaks. *Proc.*
813 *Natl. Acad. Sci. U.S.A.*, **103**, 4652–4657.
814 Jones,P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**,
815 1236–1240.
816 Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic*
817 *Acids Res*, **45**, D353–D361.
818 Katoh,K. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier
819 transform. *Nucleic Acids Research*, **30**, 3059–3066.
820 Kitts,P.A. *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res*, **44**,
821 D73–D80.
822 Kumar,P. *et al.* (2012) Tetracycline resistant V. cholerae O1 biotype El Tor serotype Ogawa with classical
823 ctxB from a recent cholera outbreak in Orissa, Eastern India. *Journal of Infection and Public*
824 *Health*, **5**, 217–219.
825 Letunic,I. and Bork,P. (2024) Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree
826 display and annotation tool. *Nucleic Acids Research*, gkae268.
827 Li,X. and Roseman,S. (2004) The chitinolytic cascade in Vibrios is regulated by chitin oligosaccharides
828 and a two-component chitin catabolic sensor/kinase. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 627–631.
829 Lilburn,T.G. *et al.* (2010) Comparative genomics of the family Vibrionaceae reveals the wide distribution
830 of genes encoding virulence-associated proteins. *BMC Genomics*, **11**, 369.
831 Lin,H. *et al.* (2018) Comparative genomic analysis reveals the evolution and environmental adaptation
832 strategies of vibrios. *BMC Genomics*, **19**, 135.
833 Lin,Z. *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model.
834 *Science*, **379**, 1123–1130.
835 Liu,J. *et al.* (2017) TetR Family Regulator brpT Modulates Biofilm Formation in Streptococcus sanguinis.
836 *PLoS ONE*, **12**, e0169301.
837 Liu,X. *et al.* (2022) The Effect of the Second Messenger c-di-GMP on Bacterial Chemotaxis in
838 Escherichia coli. *Appl Environ Microbiol*, **88**, e00373-22.
839 Ma,Y. *et al.* (2017) Stationary phase-dependent accumulation of ectoine is an efficient adaptation strategy
840 in Vibrio anguillarum against cold stress. *Microbiological Research*, **205**, 8–18.
841 Maestre-Reyna,M. *et al.* (2013) Structural Insights into RbmA, a Biofilm Scaffolding Protein of V.
842 Cholerae. *PLoS ONE*, **8**, e82458.
843 Matz,C. *et al.* (2005) Biofilm formation and phenotypic variation enhance predation-driven persistence of
844 *Vibrio cholerae*. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 16819–16824.
845 Meibom,K.L. *et al.* (2004) The *Vibrio cholerae* chitin utilization program. *Proc. Natl. Acad. Sci. U.S.A.*,
846 **101**, 2524–2529.
847 Nelson,E.J. *et al.* (2009) Cholera transmission: the host, pathogen and bacteriophage dynamic. *Nat Rev*
848 *Microbiol*, **7**, 693–702.
849 Page,A.J. *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**,
850 3691–3693.
851 Parks,D.H. *et al.* (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a
852 phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic*
853 *Acids Research*, **50**, D785–D794.
854 Price,M.N. *et al.* (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.
855 *PLoS ONE*, **5**, e9490.

Rice,S.A. *et al.* (2009) The biofilm life cycle and virulence of *Pseudomonas aeruginosa* are dependent on a filamentous prophage. *The ISME Journal*, **3**, 271–282.
Robin,B. *et al.* (2022) MacAB-TolC Contributes to the Development of Acinetobacter baumannii Biofilm at the Solid–Liquid Interface. *Front. Microbiol.*, **12**, 785161.
Rozewicki,J. *et al.* (2019) MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*, gkz342.
Russell,M.H. *et al.* (2013) Integration of the Second Messenger c-di-GMP into the Chemotactic Signaling Pathway. *mBio*, **4**, e00001-13.
Schwechheimer,C. *et al.* (2020) A tyrosine phosphoregulatory system controls exopolysaccharide biosynthesis and biofilm formation in Vibrio cholerae. *PLoS Pathog*, **16**, e1008745.
Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
Smith,A.B. and Siebeling,R.J. (2003) Identification of Genetic Loci Required for Capsular Expression in *Vibrio vulnificus*. *Infect Immun*, **71**, 1091–1097.
Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, W609–W612.
Tan,D. *et al.* (2020) High cell densities favor lysogeny: induction of an H20 prophage is repressed by quorum sensing and enhances biofilm formation in *Vibrio anguillarum*. *The ISME Journal*, **14**, 1731–1742.
Teschler,J.K. *et al.* (2015) Living in the matrix: assembly and control of Vibrio cholerae biofilms. *Nat Rev Microbiol*, **13**, 255–268.
Teufel,F. *et al.* (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol*, **40**, 1023–1025.
Wang,S. *et al.* (2023) Temperate phage influence virulence and biofilm-forming of Salmonella Typhimurium and enhance the ability to contaminate food product. *International Journal of Food Microbiology*, **398**, 110223.
Whiteley,C.G. and Lee,D.-J. (2015) Bacterial diguanylate cyclases: Structure, function and mechanism in exopolysaccharide biofilm development. *Biotechnology Advances*, **33**, 124–141.
Yadavalli,S.S. and Yuan,J. (2022) Bacterial Small Membrane Proteins: the Swiss Army Knife of Regulators at the Lipid Bilayer. *J Bacteriol*, **204**, e00344-21.
Yamanaka,Y. *et al.* (2016) Transcription factor CecR (YbiH) regulates a set of genes affecting the sensitivity of Escherichia coli against cefoperazone and chloramphenicol. *Microbiology*, **162**, 1253–1264.
Yang,Y. *et al.* (2024) Large-scale genomic survey with deep learning-based method reveals strain-level phage specificity determinants. *GigaScience*, **13**, giae017.
Yang,Y. and Jiang,X. (2023) Evolink: a phylogenetic approach for rapid identification of genotype–phenotype associations in large-scale microbial multispecies data. *Bioinformatics*, **39**, btad215.
Yen,M. *et al.* (2017) A cocktail of three virulent bacteriophages prevents Vibrio cholerae infection in animal models. *Nat Commun*, **8**, 14187.
Yildiz,F. *et al.* (2014) Structural Characterization of the Extracellular Polysaccharide from Vibrio cholerae O1 El-Tor. *PLoS ONE*, **9**, e86751.
Yildiz,F.H. and Schoolnik,G.K. (1999) *Vibrio cholerae* O1 El Tor: Identification of a gene cluster required for the rugose colony type, exopolysaccharide production, chlorine resistance, and biofilm formation. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 4028–4033.
Yildiz,F.H. and Visick,K.L. (2009) Vibrio biofilms: so much the same yet so different. *Trends in Microbiology*, **17**, 109–118.

Supplementary Figure/Table Legends

Figure S1. Gene syntenies for *vps*-2 locus in *Vibrio cholerae*, *cps* locus in *Vibrio parahaemolyticus*, *wcr* loci in *Vibrio vulnificus*, and *vps*-2-like loci in *Alliivibrio fisheri* are depicted. Genes with more than 30% sequence similarity are color-coded. Link colors indicate sequence identities. The *cps*, *wcr*, and *vps*-2 loci all contain genes within their clusters that are similar to those found in the *vps*-2 cluster, particularly genes resembling *vpsLMNO*.

Figure S2. Detailed gene syntenies for biofilm matrix clusters in 29 (sub)species with the same color palette as in Figure 1. GT: Glycosyltransferase; PS: Polysaccharide; O-PS: O-Antigen; AT: Acyltransferase; GH: Glycoside hydrolase; GD: Glycoside deacetylase; OR: Oxidoreductase; TPP: Thiamine pyrophosphate; TR protein: Transcriptional regulatory protein.

Figure S3. Gene syntenies for *rbmC* genes and their neighboring genes in species with *rbmC* genes distant from the biofilm matrix cluster. Genes are color-coded by clusters sharing more than 80% sequence similarity, and link colors represent sequence identities. Detailed information is available in Supplementary Table 2.

Figure S4. Predicted structures by AlphaFold3 (Abramson *et al.*, 2024) for proteins representing the six RbmC and Bap1 structural variants defined in this study. (A-F) Predicted structures of GCA_019670025.1_03371, GCA_003312035.1_01787, GCA_000259295.1_03774, GCA_019048845.1_03201, GCA_024746925.1_01708 and GCA_003716425.1_01353 representing proteins of RbmC with β-helix, M1M2-less RbmC, partial M1M2 RbmC, standard RbmC, standard Bap1 and loop-less Bap1, respectively.

Figure S5. The heatmap indicating the domain presence and absence in 997 RbmC and Bap1 encoded genes. Rows represent domains. Columns represent genes and are mapped to the gene tree. The tips are annotated with the species of origin and structural types. Grey strips represent truncated proteins.

Figure S6. The domain tree for 1001 β-propeller domains of RbmC and Bap1 encoded sequences, rooted with the RbmC with β-helix encoded genes. The outer circle indicates the species of origin, while the inner circle indicates the protein structural features. Grey strips represent truncated proteins.

Figure S7. The domain tree for 1433 β-prism domains of RbmC and Bap1 encoded sequences, rooted at the midpoint. The outer circle indicates the species of origin, while the inner circle indicates the protein structural features. The color ranges indicate the domain source. Grey strips represent truncated proteins.

Figure S8. Predicted structures by AlphaFold-Multimer for the PDE dimer alone (gene accession: GCA_019093095.1_02056) and the complex of the PDE dimer with the small protein group_2000 (gene accession: GCA_019093095.1_02057). Putative signal peptide in PDE protein has been removed.

Figure S9. The boxplot (A) and histogram (B) displaying the prophage count and density in loop-less Bap1-positive and negative strains.

Table S1. The isolation source and pathogenicity information for Vibrio species.

Table S2. Gene synteny and annotation information for *rbmC* genes and their neighboring genes in species with *rbmC* genes distant from the biofilm matrix cluster. This is provided as the supporting data for Supplementary Figure 3.

Table S3. Gene synteny and annotation for genes not annotated as *vps* and *rbm* genes in nine species (*V. hepatarius_A*, *V. hepatarius*, *V. sinaloensis*, *V. atypicus*, *V. tubiashii_A*, *V. bivalvicida*, *V. tubiashii*, *V. sp013113835* and *V. coralliilyticus*) from Figure 1B.

Table S4. Gene synteny and annotation information for positively and negatively associated gene groups related to loop-less Bap1 encoding strains. This is provided as supporting data for Figure 3B.

Table S5. Gene synteny and annotation information for the 21 *rbmB*-like genes and their neighboring genes.

Table S6. BLASTp results for the best hits of the 143 single-stranded right-handed β-helix domain containing prophage genes in the INPHARED.

Supplementary Data

Data S1. NCBI Assembly accessions and GTDB species for 6,121 Vibrio genomes.

Data S2. Biofilm matrix cluster and proteins annotation with ProkFunFind.

Data S3. RbmC and Bap1 protein classification table.

Data S4. RbmC and Bap1 protein sequences in FASTA format.

Data S5. 1,007 non-redundant predicted structures of RbmC and Bap1 proteins using ESMfold in PDB format.

Data S6. RbmB protein sequences in FASTA format.

Data S7. RbmA protein sequences in FASTA format.

Data S8. VpsE protein sequences in FASTA format.

Data S9. VpsF protein sequences in FASTA format.

Data S10. 216 representative genomes for Vibrio species.

Data S11. *V. cholerae* subspecies tree in NEWICK format.

Data S12. Vibrio species tree in NEWICK format.

971 **Data S13.** Gene groups detected in *V. cholerae* pangenome analysis using Roary.

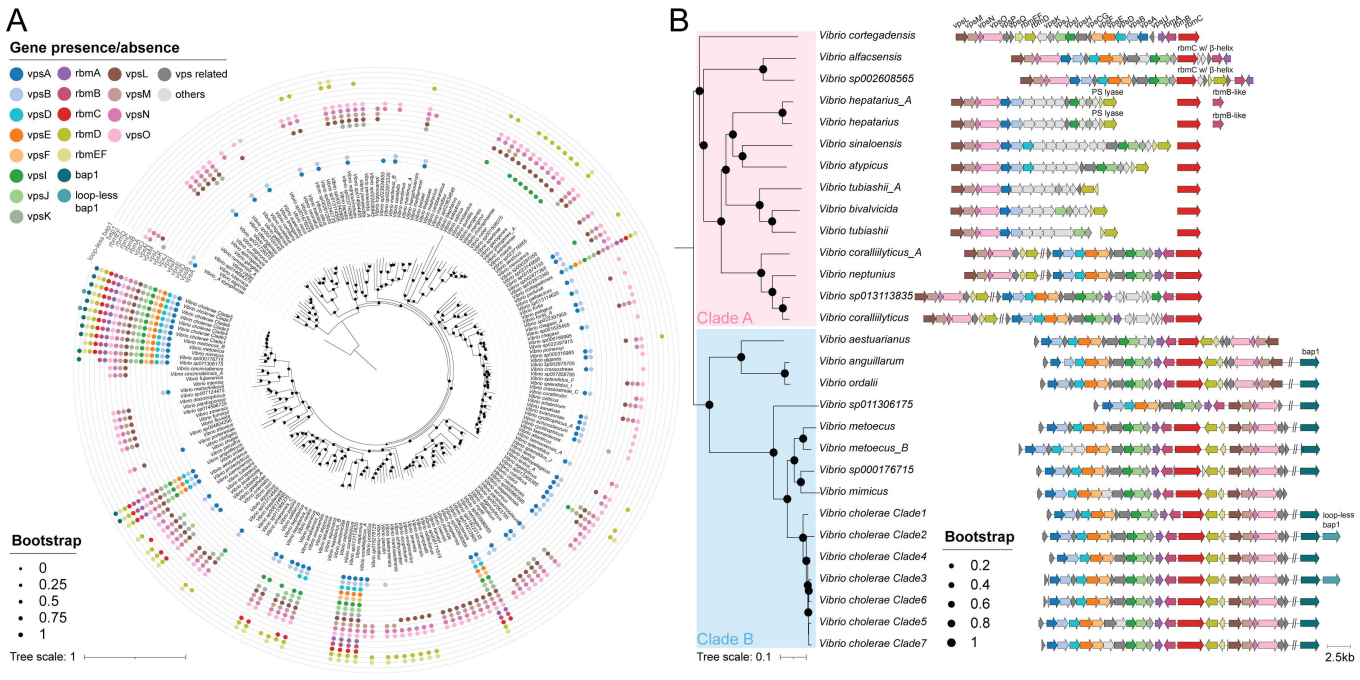
972 **Data S14.** Signal peptide positions detected for RbmC and Bap1 using SignalP6.0.

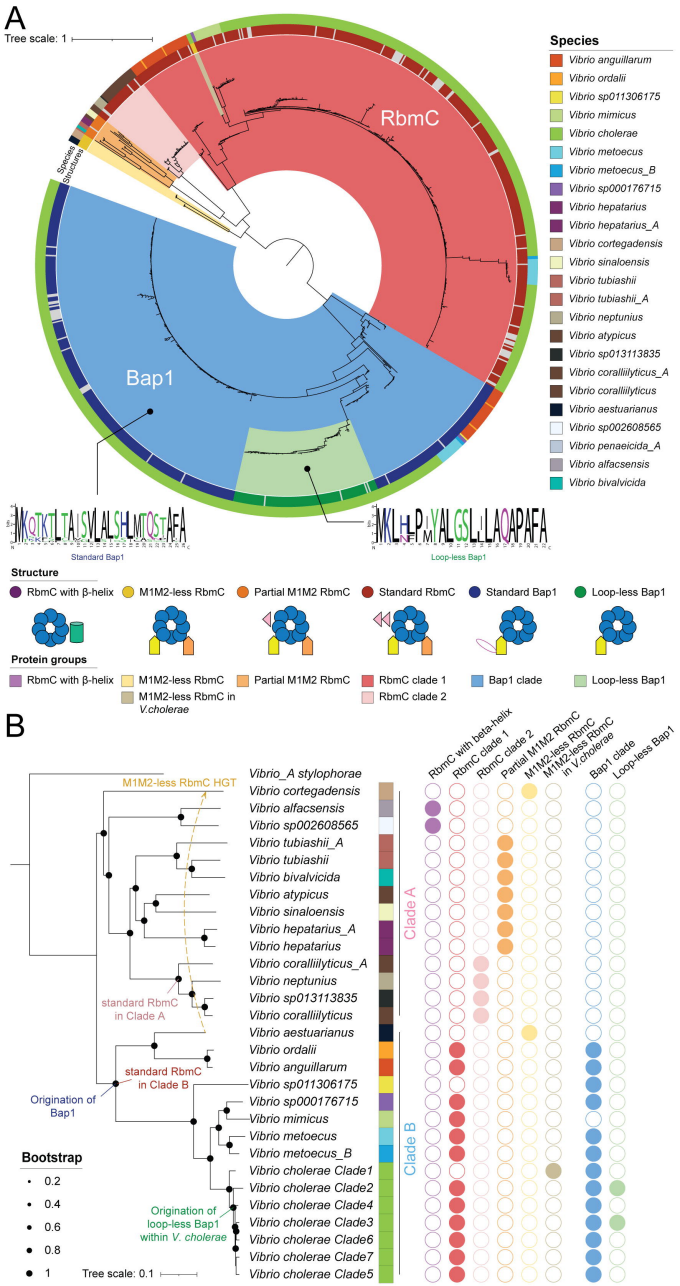
973 **Data S15.** Single-stranded right-handed β-helix domain containing gene tree in NEWICK format.

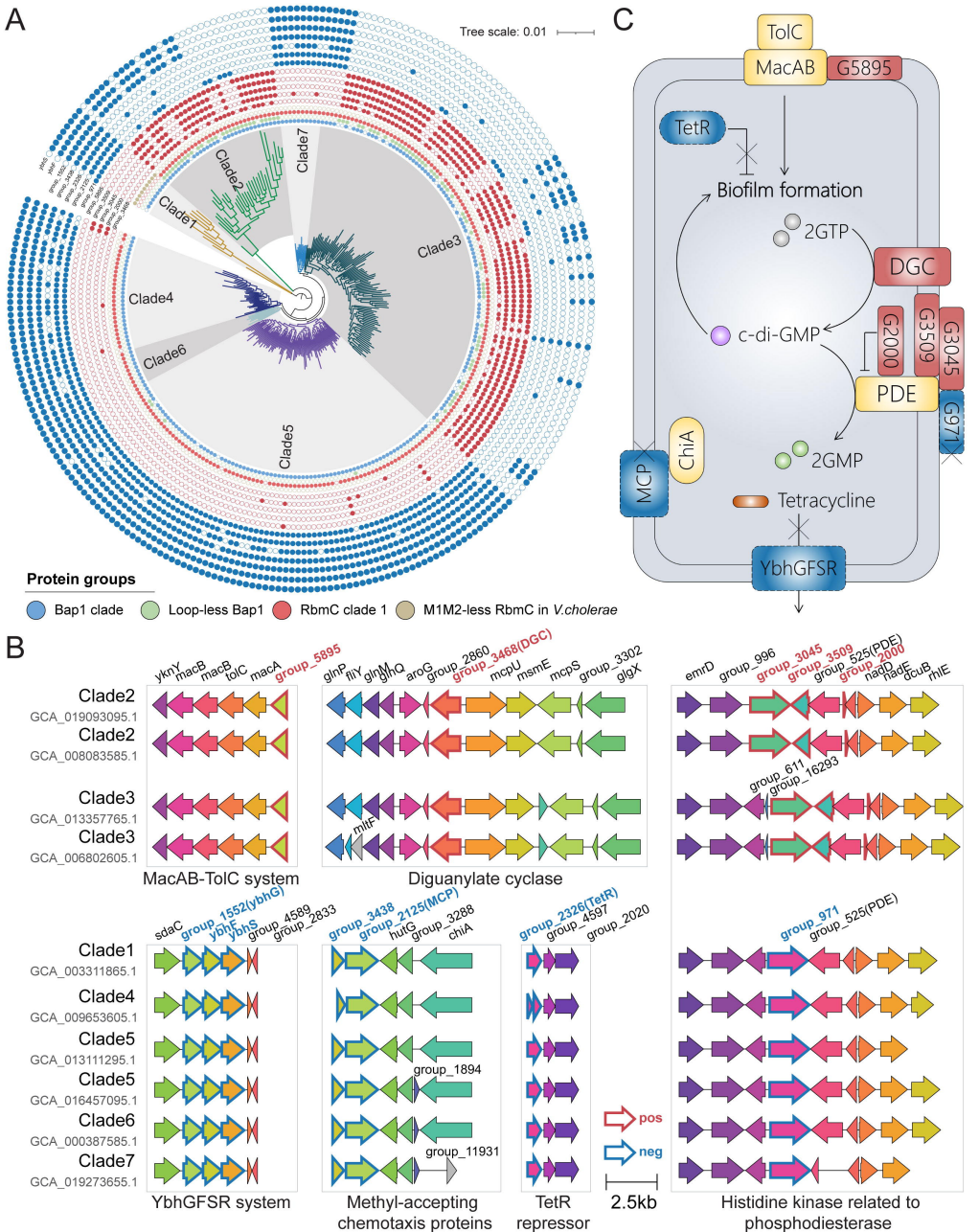
974 **Data S16.** RbmC and Bap1 proteins' β-propeller domain tree in NEWICK format.

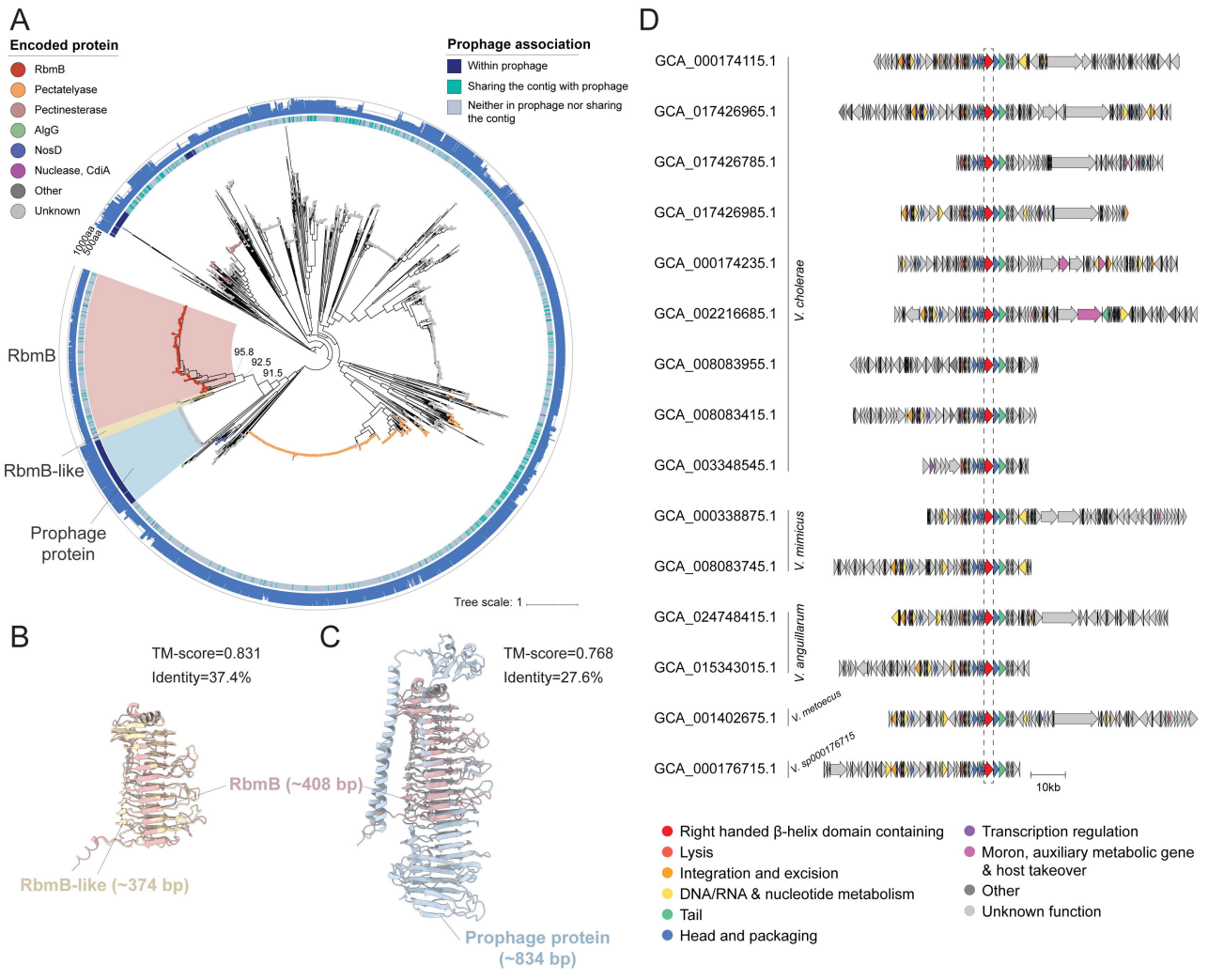
975 **Data S17.** RbmC and Bap1 proteins' β-prism domain tree in NEWICK format.

976 **Data S18.** Prophage regions detected in 1,803 genomes having single-stranded right-handed β-
977 helix domain containing genes.


A
Gene presence/absence
vpsA
vpsB
vpsD
vpsE
vpsF
vpsI
vpsJ
vpsK
rbmA
rbmB
rbmC
rbmD
rbmEF
bap1
loop-less bap1
vpsL
vpsM
vpsN
vpsO
vps related
others
Bootstrap
0
0.25
0.5
0.75
1
Tree scale: 1
B
Clade A
Clade B
Vibrio cortegadensis
Vibrio alfacsensis
Vibrio sp002608565
Vibrio hepatarius_A
Vibrio hepatarius
Vibrio sinaloensis
Vibrio atypicus
Vibrio tubiashii_A
Vibrio bivalvicida
Vibrio tubiashii
Vibrio coralliilyticus_A
Vibrio neptunius
Vibrio sp013113835
Vibrio coralliilyticus
Vibrio aestuarianus
Vibrio anguillarum
Vibrio ordalii
Vibrio sp011306175
Vibrio metoecus
Vibrio metoecus_B
Vibrio sp000176715
Vibrio mimicus
Vibrio cholerae Clade1
Vibrio cholerae Clade2
Vibrio cholerae Clade4
Vibrio cholerae Clade3
Vibrio cholerae Clade6
Vibrio cholerae Clade5
Vibrio cholerae Clade7
rbmC w/ β-helix
PS lyase
rbmB-like
bap1
loop-less bap1
Bootstrap
0.2
0.4
0.6
0.8
1
Tree scale: 0.1
2.5kb


A
Tree scale: 1
Species
Vibrio anguillarum
Vibrio ordalii
Vibrio sp011306175
Vibrio mimicus
Vibrio cholerae
Vibrio metoecus
Vibrio metoecus_B
Vibrio sp000176715
Vibrio hepatarius
Vibrio hepatarius_A
Vibrio cortegadensis
Vibrio sinaloensis
Vibrio tubiashii
Vibrio tubiashii_A
Vibrio neptunius
Vibrio atypicus
Vibrio sp013113835
Vibrio coralliilyticus_A
Vibrio coralliilyticus
Vibrio aestuarianus
Vibrio sp002608565
Vibrio penaeicida_A
Vibrio alfacsensis
Vibrio bivalvicida
RbmC
Bap1
Standard Bap1
Loop-less Bap1
Structure
RbmC with β-helix
M1M2-less RbmC
Partial M1M2 RbmC
Standard RbmC
Standard Bap1
Loop-less Bap1
Protein groups
RbmC with β-helix
M1M2-less RbmC
Partial M1M2 RbmC
RbmC clade 1
Bap1 clade
Loop-less Bap1
M1M2-less RbmC in V.cholerae
RbmC clade 2
B
M1M2-less RbmC HGT
standard RbmC in Clade A
standard RbmC in Clade B
Origination of Bap1
Origination of loop-less Bap1 within V. cholerae
Bootstrap
0.2
0.4
0.6
0.8
1
Tree scale: 0.1
Vibrio_A stylophorae
Vibrio cortegadensis
Vibrio alfacsensis
Vibrio sp002608565
Vibrio tubiashii_A
Vibrio tubiashii
Vibrio bivalvicida
Vibrio atypicus
Vibrio sinaloensis
Vibrio hepatarius_A
Vibrio hepatarius
Vibrio coralliilyticus_A
Vibrio neptunius
Vibrio sp013113835
Vibrio coralliilyticus
Vibrio aestuarianus
Vibrio ordalii
Vibrio anguillarum
Vibrio sp011306175
Vibrio sp000176715
Vibrio mimicus
Vibrio metoecus
Vibrio metoecus_B
Vibrio cholerae Clade1
Vibrio cholerae Clade2
Vibrio cholerae Clade4
Vibrio cholerae Clade3
Vibrio cholerae Clade6
Vibrio cholerae Clade7
Vibrio cholerae Clade5
Clade A
Clade B
RbmC with beta-helix
RbmC clade 1
RbmC clade 2
Partial M1M2 RbmC
M1M2-less RbmC
M1M2-less RbmC in V.cholerae
Bap1 clade
Loop-less Bap1


A
Tree scale: 0.01
Clade1
Clade2
Clade3
Clade4
Clade5
Clade6
Clade7
Protein groups
Bap1 clade
Loop-less Bap1
RbmC clade 1
M1M2-less RbmC in V.cholerae
C
TolC
MacAB
G5895
TetR
Biofilm formation
2GTP
DGC
c-di-GMP
G2000
G3509
G3045
PDE
G971
ChiA
MCP
2GMP
Tetracycline
YbhGFSR
B
yknY macB macB tolC macA group_5895
Clade2 GCA_019093095.1
Clade2 GCA_008083585.1
Clade3 GCA_013357765.1
Clade3 GCA_006802605.1
MacAB-TolC system
glnP fliY glnM glnQ aroG group_2860 group_3468(DGC) mcpU msmE mcpS group_3302 glgX
mlrF
Diguanylate cyclase
emrD group_996 group_3045 group_3509 group_525(PDE) group_1000 nadE dcuB rhlE
group_611 group_16293
sdaC group_1552(ybhG) ybhF ybhS group_4589 group_2833
Clade1 GCA_003311865.1
Clade4 GCA_009653605.1
Clade5 GCA_013111295.1
Clade5 GCA_016457095.1
Clade6 GCA_000387585.1
Clade7 GCA_019273655.1
YbhGFSR system
group_3438 group_2125(MCP) hutG group_3288 chiA
group_1894
group_11931
Methyl-accepting chemotaxis proteins
group_2326(TetR) group_4597 group_2020
TetR repressor
pos
neg
2.5kb
group_971 group_525(PDE)
Histidine kinase related to phosphodiesterase


A
Encoded protein
RbmB
Pectatelyase
Pectinesterase
AlgG
NosD
Nuclease, CdiA
Other
Unknown
Prophage association
Within prophage
Sharing the contig with prophage
Neither in prophage nor sharing the contig
100aa
50aa
RbmB
95.8
92.5
91.5
RbmB-like
Prophage protein
Tree scale: 1
B
TM-score=0.831
Identity=37.4%
RbmB (~408 bp)
RbmB-like (~374 bp)
C
TM-score=0.768
Identity=27.6%
Prophage protein (~834 bp)
D
GCA_000174115.1
GCA_017426965.1
GCA_017426785.1
GCA_017426985.1
GCA_000174235.1
GCA_002216685.1
GCA_008083955.1
GCA_008083415.1
GCA_003348545.1
V. cholerae
GCA_000338875.1
GCA_008083745.1
V. mimicus
GCA_024748415.1
GCA_015343015.1
V. anguillarum
GCA_001402675.1
V. metoecus
GCA_000176715.1
V. sp000176715
10kb
Right handed β-helix domain containing
Lysis
Integration and excision
DNA/RNA & nucleotide metabolism
Tail
Head and packaging
Transcription regulation
Moron, auxiliary metabolic gene & host takeover
Other
Unknown function