DciA helicase operators exhibit diversity across bacterial phyla

Helen C. Blaine[1]*, Joseph T. Burke[2]*, Janani Ravi[2#], and Christina L. Stallings[1#]

[1]Department of Molecular Microbiology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA.

[2]Departments of Pathobiology and Diagnostic Investigation, Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA.

*Contributed equally.

[#]Correspondence to stallings@wustl.edu and janani@msu.edu.

ABSTRACT

A fundamental requirement for life is replication of an organism's DNA. Studies in *Escherichia coli* and *Bacillus subtilis* have set the paradigm for how DNA replication occurs in bacteria. During replication initiation in *E. coli* and *B. subtilis*, the replicative helicase is loaded onto the DNA at the origin of replication by an ATPase helicase loader. However, most bacteria do not encode homologs to the helicase loaders in *E. coli* and *B. subtilis*, raising the question of how helicase activity is facilitated in other bacteria during DNA replication initiation. Recent work has identified the DciA protein as a predicted helicase operator that may perform a function analogous to the helicase loaders in *E. coli* and *B. subtilis*. DciA proteins are defined by the presence of a DUF721 domain and are conserved in most bacteria. However, we find that the sequence conservation between DciA proteins across different phyla is very low. Therefore, to comprehensively define the DciA protein family, we took a computational evolutionary approach. These analyses identified diversity in sequence features and domain architectures amongst DciA homologs that were associated with specific phylogenetic lineages. The diversity of DciA proteins elucidated here represents the evolution of helicase operation in bacterial DNA replication, highlights the need for

phyla-specific analyses of this fundamental biological process, and is an important example of how research in bacterial DNA replication is necessary in organisms beyond *E. coli* and *B. subtilis.*

IMPORTANCE

Despite the fundamental importance of DNA replication for life, this process remains understudied in bacteria outside of *Escherichia coli* and *Bacillus subtilis*. In particular, most bacteria do not encode the helicase loading proteins that are essential in *E. coli* and *B. subtilis* for DNA replication. Instead, most bacteria encode a DciA homolog that likely constitutes the predominant mechanism of helicase operation in bacteria. However, it is still unknown how DciA structure and function compares across diverse phyla that encode DciA proteins. In this study, we perform a computational evolutionary analysis that uncovers tremendous diversity amongst DciA homologs. These studies provide a significant advance in our understanding regarding an essential component of the bacterial DNA replication machinery.

INTRODUCTION

DNA replication is a process critical to life for all organisms The current paradigm for the process of DNA replication in bacteria has primarily been based on studies in *Escherichia coli* and *Bacillus subtilis*. Bacterial DNA replication begins with the binding of the replication initiation protein DnaA to specific sequences referred to as DnaA boxes at the origin of replication (oriC) (1–7). DnaA binding to double-stranded DNA (dsDNA) triggers DNA unwinding at an AT-rich region of DNA called the DNA unwinding element (DUE), leaving a bubble of single-stranded DNA (ssDNA) (2, 3, 6, 8, 9). The ssDNA bubble is coated by single-stranded binding protein (SSB) (10), followed by the concerted loading of two hexameric replicative helicases onto the SSB coated replication fork. The two helicases translocate along the two sides of the replication fork, unwinding the dsDNA as they move (1, 3, 6, 11–14).

Bacterial replicative helicases (DnaB in *E. coli* and DnaC in *B. subtilis*) are Superfamily IV type helicases, which are defined as hexameric RecA ATPases (3, 15, 16) that translocate in the 5'-3' direction (11, 12, 17). The bacterial replicative helicase translocates on ssDNA using a "hand-over-hand" mechanism, which is driven by nucleotide hydrolysis (18, 19) (reviewed in (17)). The C-terminus of the bacterial replicative helicase contains the RecA-like fold that is responsible for the ATPase activity, and is connected to an N-terminal scaffolding domain via a linker region (6, 20, 21). The replicative helicase must oligomerize into a double-layered hexameric ring to be active during replication, with one layer made up of the N-termini and the other layer comprised of the C-termini (6, 22, 23). In *E. coli* and *B. subtilis*, the loading of the replicative helicase is performed with the help of a helicase loader, termed DnaC in *E. coli* and DnaI in *B. subtilis* (3, 24–27). *dnaC* and *dnaI* were acquired by *E. coli* and *B. subtilis*, respectively, via domestication of related but distinct phage ATPase-containing genes (28). DnaC and DnaI are both in the ATPases Associated with diverse cellular Activities (AAA+) ATPase family, and

the ATPase activity of DnaC is required for its helicase loading function at the origin of replication (29, 30).

E. coli and *B. subtilis* have long represented the paradigm of helicase loading during bacterial replication. However, the majority of bacteria do not encode ATPase helicase loader homologs to DnaC or DnaI. Instead, most bacteria encode the ancestral protein, DciA (DnaC/I Antecedent) (28, 31), which is defined by the presence of a Domain of Unknown Function (DUF) 721. Despite the prevalence of DUF721-containing DciA homologs in bacteria (28), DciA has only been studied in actinobacterial (*Mycobacterium tuberculosis and Mycobacterium smegmatis)* and gammaproteobacterial (*Pseudomonas aeruginosa* and *Vibrio cholerae*) species (28, 31–33). DciA homologs interact with the replicative helicase DnaB and are essential for *M. tuberculosis, M. smegmatis,* and *P. aeruginosa* DNA replication and viability (28, 31). Based on DciA's interaction with the replicative helicase and requirement for DNA replication, DciA has been proposed to perform a function analogous to that of the DnaC/I helicase loaders. However, DciA does not have a predicted ATPase domain and, therefore, cannot be considered a helicase loader like DnaC/I. Instead, DciA is referred to as a predicted helicase operator, although the mechanism of DciA helicase operation is still unknown (28, 31).

Beyond the presence of the DUF721, DciA domain architecture and the relationship between DciA homologs across diverse bacterial phyla are yet to be investigated. To address these open questions, we took a computational evolutionary approach and analyzed the phylogenic distribution, domain architecture, and sequence conservation for DciA homologs. We have discovered low sequence similarity between DciA homologs from different phyla, lineage-specific domain architectures, and divergent evolution of specific DciA homologs, all of which likely have functional consequences. This study provides an evolutionary picture of DciA, reveals key differences between homologs, and generates the framework for mechanistic investigation into different classes of DciA proteins.

RESULTS

DciA proteins vary considerably in sequence across bacterial phyla

Most bacteria encode a DciA homolog, defined by the presence of the DUF721 domain (28). Based on studies in *M. tuberculosis*, the DUF721 is predicted to contain a region of structural homology to the N-terminus of DnaA, which was thus named the DnaA N-terminal-like (DANL) domain (31). The presence of the DANL domain has subsequently been confirmed in *V. cholerae* DciA (32). Our analysis of other DciA homologs indicates that this predicted structural domain is conserved, suggesting that it is important for DciA function. However, beyond the annotation of the DUF721 (henceforth referred to as the DciA domain), it is unclear how different DciA homologs relate to each other. A protein BLAST search for *M. tuberculosis* DciA homologs based on primary amino acid sequence only identifies closely related homologs, all of which are in actinobacteria (**Figure S1**). We found a similar pattern when retrieving homologs with *Pseudomonas aeruginosa*, all homologs are proteobacterial (**Figure S1**). This suggests that the DciA homologs in different phyla are diverse with low conservation in their primary amino acid sequence. We, therefore, needed a more extensive method to investigate the relationship between DciA homologs across different phyla. To achieve this, we used the MolEvolvR web application to comprehensively identify and characterize DciA homologs across all bacterial lineages using molecular evolution and phylogeny (34). Since individual DciA proteins from specific lineages were not successful in retrieving homologs from other distant phyla, we selected a much wider range of starting points. We started with 21 DciA proteins from 11 diverse phyla (28), including representatives from actinobacteria, proteobacteria, and cyanobacteria **(Table S1; Figure 1A)**, as query sequences to identify diverse DciA homologs across additional bacterial phyla **(Figure 1A)**. Our homology search resulted in identifying ~9K DciA homologs from 15 bacterial phyla **(Figure 1B)**. In line with

both genome sequencing and publication bias, proteobacteria and actinobacteria were over-represented in both our queries and recovered sequences (35–37) (**Figure 1A,B**).

No single DciA protein identified homologs in all other phyla (**Figure 1C**), supporting that there is low sequence conservation between DciA homologs. To quantify the sequence conservation across phyla, we analyzed the pairwise similarity for the 21 DciA protein homologs used as our query set (**Figure 2**). We found a wide range of similarities (~20–60%), with the majority of homologs showing 30-40% similarity. Our query sequence dataset contained multiple species within each class of proteobacteria, so we were able to compare the conservation between DciA homologs within proteobacteria as well as between proteobacteria and other phyla. Similarity was high between DciA proteins within alphaproteobacteria (62.3% between *Brucella abortus* and *Mesorhizobium australicum*) and gammaproteobacteria (52.3% between *Proteus mirabilis* and *Vibrio cholerae*). Overall, DciA protein queries within proteobacteria have an average of ~30% similarity, and proteobacterial homologs have an average of 31.12% similarity to homologs outside of their phyla **(Figure 2)**. Actinobacterial and proteobacterial DciA homologs share between 21–34% sequence similarity. For example, the *M. tuberculosis* DciA protein shares 28.3% identity with the *Pseudomonas aeruginosa* DciA (**Figure 2**). DciA shares this trait of low sequence conservation across phyla with its interaction partner, the replicative helicase DnaB, where DnaB proteins in *M. tuberculosis* and *P. aeruginosa* only share 20.4% similarity. Bacterial replication proteins in general have low to moderate sequence conservation across phyla (11–49% similarity across replisome proteins between *E. coli* and *B. subtilis*) (38). Therefore, our data showing low DciA sequence conservation is consistent with other replication initiation proteins. Given the low-level conservation across divergent DciA proteins, it was unsurprising that individual DciA query proteins never returned homologs from all other phyla and emphasizes the need for multiple starting points for analysis (**Figure 1C**).

While we identified a diverse set of DciA homologs, including those with moderate to low sequence conservation from most lineages (**Figure 1**), we were still missing multiple phyla previously reported to contain DciA homologs. Therefore, we expanded our search further by including 66 DciA query proteins from 20 bacterial phyla as query sequences (**Figure 3A**; **Table 1**). This hugely diversified our results identifying >13K unique DciA homologs from 22 bacterial phyla (**Figure 3B,C**). Consistent with the previous smaller set of DciA query proteins (**Figure 1**), we found that most DciA proteins identified homologs within their corresponding phylum as well as ones within actinobacteria and proteobacteria (**Figure 3C**). However, a few DciA proteins recovered homologs only within their own phylum. Specifically, most actinobacterial (10/12) and a few proteobacterial (10/30) DciA proteins only recovered homologs within their respective phylum (**Figure 1C, 3C**). In addition, the DciA proteins from nitrospirae, gemmatimonadetes, fibrobacteres, chlorobi, chlamydiae, and bacteroidetes identified homologs from actinobacteria and not proteobacteria, suggesting that these DciA proteins are closer evolutionarily to actinobacterial DciA homologs than proteobacterial DciA. In contrast, acidobacteria, cyanobacteria, and thermodesulfobacteria recovered homologs from proteobacteria and not actinobacteria. The evolution of DciA seems to mirror the phylogenetic distance of these species; gemmatimonadetes, chlamydiae, and bacteroidetes are more closely evolutionary related to actinobacteria than proteobacteria based on the 16S rRNA gene (28, 39). Together these data suggest that the DciA homologs in actinobacteria and proteobacteria likely represent the most evolutionarily divergent repertoire of DciA homologs. This analysis also indicates that DciA proteins from distinct phyla carry lineage-specific signatures, likely co-evolving with other phylum-specific protein families involved in DNA replication.

DciA domain architecture varies in a lineage-specific manner

The evolutionary divergence in DciA proteins prompted us to take a closer look at the sequence-structure features of these homologs, including sequence alignment and domain architectures. As a first step, we aligned the 66 DciA starting point protein sequences from 20 phyla (**Figure 3**; **Table 1**) using the MolEvolvR web application (34) and examined their domain architectures **(Figure 4**; *see Methods***)**. Tracing the phylogenetic tree of DciA sequences confirms that the evolution of DciA proteins roughly corresponds to bacterial phylogeny, where DciA homologs tend to cluster with their own phyla (*e.g.*, proteobacteria and actinobacteria) **(Figure 4 center)**. Only DciA homologs that encoded the signature DUF721 domain were included in our alignment, and could be classified into one of four distinct groups based on where the DUF721 domain occurs within the protein **(Figures 4, 5, and S2)**. We describe the group memberships and descriptions of our 66 DciA query proteins in the following sections (**Table 1**).

In Group 1 DciA proteins, the DUF721 spans at least 70% of the protein sequence, with ≤25 amino acids on either side of the DUF721 (**Figures 4, 5A, S2**; **Table 1**, See Group 1 example: *R. rickettsii* DciA). Group 1 DciA proteins are present in acidobacteria, bacteroidetes, chlamydiae, chlorobi, gemmatimonadetes, elusimicrobia, fibrobacteres, fusobacteria, thermotogae, planctomycetes, spirochaetes, and proteobacteria (**Table 1**). The DUF721 domain spanning the entire protein in this group suggests that the DUF721 is likely sufficient for DciA function in these bacteria.

Group 2 DciA homologs have ≤25 amino acids C-terminal to the DUF721 and an N-terminal extension of >25 amino acids and no known domains (**Figures 4, 5A, S2**; **Table 1**, See Group 2 example: *M. tuberculosis* DciA). Group 2 DciA homologs are present in actinobacteria and verrucomicrobia (**Figure 5A**; **Table 1**). Group 2 DciA queries in the mycobacteriales order are predicted to have intrinsically disordered regions by MobiDBLite software (also noted in (40)) in the N-terminal extension (**Figure 4, left**). This predicted intrinsically disordered region is not

present in bifidobacteriales, another order within actinobacteria, or in other Group 2 proteins. Expression of the *M. tuberculosis* DciA DUF721 alone is not sufficient to support viability in mycobacteria (31), suggesting that the N-terminal extension is essential for DciA function in mycobacteria, although its function remains unknown. The requirement for the sequence N-terminal to the DUF721 in mycobacteria and the absence of this N-terminal sequence in Group 1 and 3 DciA proteins demonstrates divergent evolution of DciA homologs in bacteria and the potential for functional variation. In addition to the N-terminal extension in all Group 2 DciA homologs, 4 actinobacterial proteins also encode a short predicted region of disorder C-terminal to the DUF721 (19aa in *M. tuberculosis*; **Figure 4, left, Table 1**).

Group 3 DciA homologs have ≤25 amino acids N-terminal to the DUF721 and a C-terminal extension >25 amino acids long (**Figures 4, 5A, S2**; **Table 1**, See Group 3 example: *V. cholerae* DciA). Group 3 DciA homologs are present in acidobacteria, spirochaetes, deferribacteres, cyanobacteria, dictyoglomi, nitrospirae, thermodesulfobacteria, and proteobacteria. The DciA homologs in *T. palladium*, *Nitrospirae moscoviensis,* and *V. parahemolyticus* are the only Group 3 proteins that are predicted to be intrinsically disordered in the region C-terminal to the DUF721 domain using the MobiDBLite predictor **(Figure 4, left**; **Table 1**). However, the C-terminal sequence of *V. cholerae* DciA was predicted to be intrinsically disordered in prior studies using PONDR and IUPred2A software, which was further confirmed by small-angle X-ray scattering (32). When we analyzed the *V. cholerae* sequence using the JRONN disorder prediction track, we were able to identify an intrinsically disordered region C-terminal to the DUF721 **(Figure 4)**. Therefore, it is possible that other Group 3 proteins could have intrinsically disordered domains that are not predicted by the tools we are using. The C-terminal intrinsically disordered sequence of *V. cholerae* DciA is required for its interaction with DnaB as well as for the enhancement of the association between *V. cholerae* DnaB and ssDNA *in vitro* (32). The absence of the C-terminal extension in Groups 1 and 2 DciA proteins further supports evolutionary divergence in DciA

protein sequence and structure, which likely impacts specialized lineage-specific protein function or mechanism of interaction with the replicative helicase. The one actinobacterial DciA protein from our query set that falls into Group 3 is encoded by *S. seoulensis*. This protein encodes a YspA domain (Pfam: YAcAr/PF10686) C-terminal to the DUF721 (**Figure 4**, **5A, S2; Table 1**). YspA domains within proteins typically have fusions to domains that process nucleotide-derived ligands such as ADP-ribose and may function as sensors of these ligands and nucleic acids (41).

In Group 4 DciA proteins, the DUF721 domain falls > 25 amino acids away from both termini **(Figure 4, 5A, S2; Table 1,** See Group 4 example: *C. mirabilis* DciA). In our query set, Group 4 consists of DciA proteins in proteobacteria, actinobacteria, and synergistetes. The Group 4 DciA homologs in *F. fastidiosum, S. coelicor, and S. avermitilis* have regions of disorder both N-terminal and C-terminal to the DUF721 **(Figure 4, left)**.

Analysis of how the DciA groups are distributed in different bacterial phyla for our starting set of DciA homologs (**Figure 4**) reveals that 75% of homologs we queried from actinobacteria fall into Group 2, and 66% of proteobacterial DciA proteins from our starting set fall into Group 3. **(Figure 5A**; **Table 1)**. These two phyla had the most representatives in our set of query proteins, and they form distinct clusters on the phylogenetic tree **(Figure 4, right)**. Further examining the characteristics of each group revealed that 75% of DciA proteins from gram-positive bacteria included in the query fall into Group 2, and 54% of DciA starting points from gram-negative bacteria fall into Group 3 **(Figure 5B)**. By contrast, no Group 1 DciA proteins are from gram-positive bacterial queries, and Group 1 makes up 37% of gram-negative bacterial queries **(Figure 5B)**. In addition to classifying each DciA protein from our query set into one of four groups based on the position of the DUF721, we found that the alphaproteobacteria DciA proteins in the Rickettsiales and Hyphomicrobiales orders harbored an insertion within the DUF721 that is not present in any other phyla **(Figure 4, right; Figure S2)**. This could indicate further functional divergence of the DciA homologs in these orders.

Overall, these analyses reveal the diversity of domain architectures amongst our selected 66 DciA homologs. A more comprehensive analysis of all DciA homologs would be required to fully understand the distribution of diverse domain architectures in different phyla. Nonetheless, these data demonstrate that DciA homologs have diverged significantly in sequence structure based on the position and sequence of the DUF721, which may impact their function and interaction with the DNA replication machinery.

The DciA domain proximity network

To further investigate the possible functions of each of the ~13K putative DciA homologs that we have identified using our diverse query set, we interrogated their domain architectures (*see Methods*). A striking majority of these homologs showed no variation, carrying a single DciA domain (the Pfam DUF721 domain) (**Figure 6**). Less than 1% of the proteins exhibited novel fusions with the DciA domain. For example, the DciA protein from *S. seoulensis* identified other YAcAr/PF10686 Pfam members in Streptomyces, as noted above in **Figure 4**. These actinobacterial homologs carry the YspA/YAcAr-like domain known to be associated with NAD utilization and ADP-ribosylation domains (41) (**Figure 6**). We find only a few streptomyces DciA homologs that share this domain architecture (**Figure 4, S3**), suggesting that DciA in this genus might have evolved this unique function. In addition, some proteobacteria also showed variation in domains associated with the DUF721 domain: i) Reyranella species carry a C-terminal thioredoxin domain, with possible redox function, and ii) the pseudomonas genus has a rare instance of a DciA dyad (two DciA domains) (**Figure 6**). We also note that there are ABC transporter-like proteins within proteobacteria with ~30% similarity to the query DciA protein in acidobacteria, but none of the other 65 query proteins (**Figure 6**). Finally, we found that one query DciA from deferribacteres identifies peptidase-like proteins in proteobacteria, again with no similarity to any of the other divergent DciA proteins. The novel fusions and alternate domain

architectures identified within the bacterial DciA homologs have been summarized in the form of a network of domain architectures reconciling all DciA homologs, with domains as nodes and co-occurrence within proteins as edges, and their co-occurrences have been further quantified (**Figure 6**). The association of the DUF721 and DciA proteins with other functional domains could shed light on the activities for DciA in these bacteria.

DISCUSSION

The recent discovery of DciA as a predicted helicase operator in bacteria (28, 31) has begun to shed light on a long-standing open question of how the majority of bacteria facilitate helicase activity during DNA replication in the absence of the ATPase helicase loaders expressed by *E. coli* and *B. subtilis*. The wide distribution of DciA in diverse bacterial phyla indicates that these proteins likely represent the predominant paradigm for helicase operation in bacteria, despite not being conserved in *E. coli* and *B. subtilis,* the organisms typically used as a model for bacterial replication. DciA proteins are defined by the presence of the DUF721 domain and prior phylogenetic analysis indicates that *dnaC* and *dnaI* homologs were acquired through evolution at the expense of *dciA* (named for *dna[CI]* antecedent) (28), suggesting that DciA and DnaC/DnaI perform a common function. In addition, it has been shown that DciA interacts with the replicative helicase and is required for DNA replication and viability in the limited organisms it has been studied in (28, 31, 33). However, the mechanism by which DciA mediates replication initiation is still unknown. Our study has revealed immense diversity in DciA proteins, where there is low sequence similarity between homologs in different phyla **(Figures 1–3)**, there are at least 4 distinct classes based on the positioning of the DUF721 domain in the protein **(Figures 4–5)**, there exist lineage-specific insertion sequences in the DUF721 domain of some proteobacterial species **(Figures 4, S2)**, and some DciA proteins have evolved as fusions to other functional

domains **(Figure 6)**. These data suggest that DciA proteins have been divergently evolving and the mechanism of helicase operation conferred by DciA may have distinct features dependent on the bacteria.

Biochemical and genetic studies have only been performed with *M. tuberculosis, P. aeruginosa*, and *V. cholerae* DciA proteins (28, 31–33). Our analyses demonstrate that the DciA homologs from these species have a number of distinct features, in addition to low sequence similarity, raising the question of how conserved their mechanisms of action will be. In particular, mycobacterial DciA is a Group 2 DciA protein with sequences predicted to be intrinsically disordered both N-terminal and C-terminal to the DUF721 **(Figures 4 and 5)**. In contrast, *V. cholerae* and *P. aeruginosa* DciA proteins are classed as Group 3, with a long sequence extension C-terminal of the DUF721 **(Figure S2)**.

The one feature conserved in all DciA homologs is the presence of the DUF721, which contains the DANL domain and is predicted to structurally resemble the N-terminus of DnaA (31). The N-terminus of DnaA is critical for the interaction of DnaA with the helicase and other regulators (42, 43), however, the role of the DciA DANL domain in the interaction with DnaB has yet to be established. A tryptophan residue conserved in the DANL domains of many DciA homologs has structural similarity to a phenylalanine residue in the DnaA N-terminus that has been predicted to have a key role in making contacts between DnaA and its interacting partners, including DnaB (31, 44). Mutation of the conserved tryptophan in the DANL domain of *M. tuberculosis* DciA results in slow growth and decreased DNA replication (31). This supports that the conserved tryptophan within the DANL domain plays a key role for DciA function *in vivo*, however that precise role has yet to be elucidated. It is also important to note that not all DciA homologs encode this tryptophan residue within their DANL domain (32) **(Figure S2)**. In fact, there is considerable diversity in the DUF721 sequences between DciA homologs from different phyla, including an insertion in the DUF721 of some alphaproteobacteria DciA proteins, which is

not observed in the DciA proteins analyzed here from other classes **(Figure 4, Figure S2)**. Therefore, even the defining DUF721 feature of DciA proteins has evolved, likely reflecting either lineage-specific adaptation in mechanism of action or mode of interaction with the replicative helicase.

There are multiple DciA homologs predicted to encode intrinsically disordered regions N-terminal and/or C-terminal to the DUF721 **(Figure 4)**. The intrinsically disordered sequence C-terminal to the DUF721 in the *V. cholerae* Group 3 DciA protein enhances the association between DnaB and ssDNA and truncation of this intrinsically disordered sequence results in loss of the interaction between *V. cholerae* DciA and the DnaB helicase (32). Although it is unknown how much of this mechanism will be conserved in other bacterial species encoding DciA proteins with divergent domain architectures and many DciA proteins do not encode predicted intrinsically disordered regions, there is precedent for roles of intrinsically disordered domains in other bacterial DNA replication proteins. For example, the intrinsically disordered linker (IDL) within the C-terminus of SSB is important for its cooperative ssDNA binding, as well as the displacement of SSB from ssDNA (45, 46) (reviewed in (47)). The IDL has also been proposed to be important for SSB protein-protein interactions, such as the interaction between SSB and the DNA repair protein RecG (48). The intrinsically disordered C-terminus of the replication restart helicase Rep is also important for the interaction between Rep and its regulator PriC, as well as between Rep and the replicative helicase DnaB (49, 50). In addition, the helicase loaders DnaC and DnaI as well as the replication initiation protein DnaA have been predicted to encode intrinsically disordered domains, currently of unknown function (32).

A lot of unknowns still remain regarding DciA proteins and bacterial DNA replication. The computational evolutionary analysis described herein highlights the complexities and diversity that have evolved in the fundamental process of DNA replication, where no single species of bacteria will be able to represent a central dogma that holds true throughout the Kingdom. These

334 studies provide a framework for researchers to consider the evolutionary variation while dissecting
335 the mechanistic basis for helicase operation in bacteria.
336
337 **METHODS**
338 **Query selection**
339 We selected our small and full set of DciA query proteins based on the DUF721 location defined
340 by on Pfam annotation, using a variety of DciA containing phyla with annotated DciA sequences.
341 The DciA domain was annotated using Pfam annotation and subsequently confirmed using a
342 multiple sequence alignment **(Figure S2)**. The full list of starting points is listed in **Table 1**. The
343 only DciA-containing phyla excluded from our set of 66 query proteins were Deinococcus-
344 Thermus, Chrysiogenetes, and Firmicutes. Firmicutes and Deinococcus-Thermus were
345 subsequently recovered in our MolEvolvR searches.
346
347 **Analysis using MolEvolvR**
348 We used MolEvolvR (34) to determine and characterize all DciA query proteins and their
349 homologs across the bacterial kingdom. We first identified all the homologs for each of the query
350 proteins in RefSeq (51) genomes, and reconciled the comprehensive set of DciA homologous
351 proteins. Next, we characterized each of the query proteins and their homologs in terms of domain
352 architectures (including Pfam (52), Gene3D (53)), localization (using Phobius (54), TMHMM (55)),
353 and disorder (using MobiDB (56)). The domain architectures of these homologs were analyzed
354 by lineage, quantified with Upset plots, and reconciled using domain proximity networks. We then
355 performed phylogenetic analysis including phyletic spreads (sunburst, heatmap), multiple
356 sequence alignment, and tree construction using MolEvolvR and custom R scripts. The MSA for
357 subset of the sequences with representatives from the 4 DciA groups shown in Figure SY was
358 generated using Kalign (57) and Jalview (58). All our data, analyses, and visualizations

359 summarizing the DciA homologs across the bacterial kingdom along with their domain
360 architectures and phyletic spreads are available at https://github.com/JRaviLab/dcia_evolution.
361
362 **Pairwise Similarity Analysis**
363 The similarity matrix was designed using the MatGat application (59). We compared each DciA
364 query protein of the 21 starting points (**Figure 1**) to each other in order to calculate pairwise
365 percent similarities. DnaB from *M. tuberculosis* and *P. aeruginosa* sequence similarity were
366 compared using the EMBOSS Needle pairwise similarity tool (57).
367
368 **FIGURE LEGENDS**
369 **Figure 1. Query of DciA homologs using 20 DciA protein starting points reveals diversity**
370 **across bacterial phyla. A. Lineages of query DciA proteins.** Sunburst plot showing the
371 lineages of the 21 query DciA proteins. In each plot, the inner ring corresponds to the kingdom
372 (bacteria, in this case), and the outer ring represents the distribution of phyla. **B. Lineages of**
373 **DciA homologs.** Sunburst plot showing the phyletic spread of all the DciA homologs generated
374 using the 21 starting points. **C. Phyletic spread of the DciA homologs by query.** The heatmap
375 shows the presence/absence of homologs of DciA across bacterial lineages (columns) for each
376 query DciA (rows). The color gradient indicates the highest number of homologs in a particular
377 lineage. *Note: The sunburst plots only display lineages of >0.1% fraction of total proteins. The
378 heatmap gives the full picture.
379
380 **Figure 2. Pairwise similarity analysis of DciA proteins.** Pairwise percentage similarities for 21
381 query DciA proteins across 11 phyla were computed using MatGat (59) and the standard
382 BLOSUM62 matrix for similarity metric calculation.
383

Figure 3. Retrieving DciA homologs using the extended query set of 66 DciA proteins. A, B. Lineages of query and homologous DciA proteins. *See Figure 1* for details. The three phyla excluded in the query searches were Deinococcus-Thermus, Chrysiogenetes, and Firmicutes. **C. Phyletic spread of the DciA homologs by query.** *See Figure 1* for details. Deinococcus-Thermus and Firmicutes were both recovered in the resulting set of homologs. Our queries did not recover the DciA homologs in chrysiogenetes, suggesting that the homologs in this phylum are the most divergent in sequence from the query sequences.

Figure 4. Characterizing the full list (66) of DciA query proteins.

The domain architectures and disorder predictions (left panel) of the 66 DciA query proteins are overlaid with the multiple sequence alignment (right), and phylogenetic tree (middle). Each DciA protein is marked with the kingdom (B, bacteria), phylum (first 6 letters), Genus, and species (represented as '*Gspecies*'). The Pfam and MobiDB annotations for each domain prediction are shown in the legend (top). The colors in the multiple sequence alignment depiction correspond to different amino acids (bottom legend). The data was generated and visualized using the MolEvolvR web application. In addition, JronnWS (60) disorder predictions were performed within Jalview (58) (not shown here), where other DciA proteins such as *V. cholerae* show disorder regions.

Figure 5. DciA groups and their distribution within our query sequences.

A. Example domain architectures of each of the 4 groups of DciA homologs. Group 1 DciA homologs have ≤25 aa on either side of the DUF721 (top, blue), Group 2 homologs have ≤25aa C-terminal to the DUF721 (second, pink), Group 3 DciA homologs have ≤25aa amino acids N-terminal to the DUF721 (third, orange), and Group 4 DciA homologs have >25aa both N- and C-terminal to the DUF721 (bottom, teal). Graphics created using BioRender.com. **B. Distribution**

of groups within Gram-positive and Gram-negative DciA query proteins. Pie charts comparing the number of gram-positive and gram-negative bacteria that have DciA homologs in one of the 4 groups. Gram-positive (left), gram-negative (right). Percentages rounded to the nearest whole number. Details of group and Gram stain assignments of each DciA homolog are found in **Table 1**.

Figure 6. DciA partners. A. Domain proximity network. The network visualizes co-occurring domains within all bacterial DciA homologs generated with our 66 starting points (**Figure 1**; **Table 1**). The nodes and edges correspond to domains and co-occurrence of domains within a protein; the size corresponds to the frequency of occurrence with a minimum scaling factor. The full data can be accessed at https://github.com/JRaviLab/dcia_evolution. **B. Frequencies of co-occurring domains in DciA homologs.** Upset plot of the DciA homologs are shown. Blue histogram: Distribution of the predominant domains. Dots and connections: Combinations in which these domains come together in DciA domain architectures. Red histogram: Frequency of occurrences of domain architectures. Of these only the DciA containing domain architectures were used for alignments and phylogenetic trees.

SUPPLEMENTARY FIGURES

Figure S1. Mycobacterium tuberculosis and Pseudomonas aeruginosa DciA proteins only identify DciA homologs in their respective phylum. Heatmap description as in Figure 1.

Figure S2. Multiple sequence alignment of select DciA proteins. Alignment of the 66 DciA proteins used in the full query set. Numbering of residues across the top of the alignment is based on the annotation *M. tuberculosis* DciA. Based on the numbering of *M. tuberculosis* residues, The

DUF721 domain falls between 75–162 amino acids (red box), the conserved tryptophan residue in the DANL domain is at position 133 (red star), and the insertion present in some alphaproteobacterial DciA proteins occurs after position 118 on the alignment. The multiple sequence alignment was generated with Kalign (57, 61) and visualized using Jalview (58); (color scheme: Clustalx).

Figure S3. Full list representative homolog characterization (with DciA)

The domain architectures, multiple sequence alignment, and phylogenetic tree were generated using representative DciA homologs (one per domain architecture per lineage). See Figure 4 for details.

TABLES

Table 1. DciA query proteins used to identify homologs across the bacterial kingdoms. Protein, domain architecture, group, and lineage-related metadata for each of the 66 diverse starting points of DciA proteins across the bacterial kingdom are shown in this table. The full homolog data, analyses, and figures can be found here: https://github.com/jravilab/dcia_evolution.

ACKNOWLEDGEMENTS

We are very grateful to the Midwest Microbial Pathogenesis Conference (MMPC) 2021 organizers for providing HCB a travel award and opportunity to present the DciA story, and for providing an interactive venue for JR and JTB to start this collaboration with CLS and HCB.

FUNDING

CLS is supported by a Burroughs Wellcome Fund Investigator in the Pathogenesis of Infectious Disease Award. HCB is supported by the Sondra Schlesinger Student Fellowship in Molecular Microbiology. JR is supported by Michigan State University (MSU) College of Veterinary Medicine Endowed Research Funds and MSU start-up funds.

DATA AVAILABILITY AND REUSE

All the data, analyses, and visualizations are available in our GitHub repository, https://github.com/JRaviLab/dcia_evolution. Text, figures, and data are licensed under Creative Commons Attribution CC BY 4.0.

REFERENCES

1. Kaguni JM. 2011. Replication initiation at the Escherichia coli chromosomal origin. Current opinion in chemical biology, 2011/08/18 ed. 15:606–613.

2. Mott ML, Berger JM. 2007. DNA replication initiation: mechanisms and regulation in bacteria. Nature Reviews Microbiology 5:343–354.

3. Jameson KH, Wilkinson AJ. 2017. Control of Initiation of DNA Replication in Bacillus subtilis and Escherichia coli. Genes 8:22–22.

4. Fuller RS, Kornberg A. 1983. Purified dnaA protein in initiation of replication at the Escherichia coli chromosomal origin of replication. Proceedings of the National Academy of Sciences 80:5817 LP – 5821.

5. Fuller RS, Funnell BE, Kornberg A. 1984. The dnaA protein complex with the E. coli chromosomal replication origin (oriC) and other DNA sites. Cell 38:889–900.

480 6. Chodavarapu S, Kaguni JM. 2016. Replication Initiation in Bacteria. Enzymes 39:1–30.

481 7. Schaper S, Messer W. 1995. Interaction of the initiator protein DnaA of Escherichia coli with
482 its DNA target. The Journal of biological chemistry 270:17622–17626.

483 8. O'Donnell M, Langston L, Stillman B. 2013. Principles and concepts of DNA replication in
484 bacteria, archaea, and eukarya. Cold Spring Harbor perspectives in biology 5:a010108–
485 a010108.

486 9. Bramhill D, Kornberg A. 1988. Duplex opening by dnaA protein at novel sequences in
487 initiation of replication at the origin of the E. coli chromosome. Cell 52:743–755.

488 10. Meyer RR, Laine PS. 1990. The single-stranded DNA-binding protein of Escherichia coli.
489 Microbiological reviews 54:342–380.

490 11. LeBowitz JH, McMacken R. 1986. The Escherichia coli dnaB replication protein is a DNA
491 helicase. The Journal of biological chemistry 261:4738–4748.

492 12. Baker TA, Funnell BE, Kornberg A. 1987. Helicase action of dnaB protein during replication
493 from the Escherichia coli chromosomal origin in vitro. Journal of Biological Chemistry
494 262:6877–6885.

495 13. Lewis JS, Jergic S, Dixon NE. 2016. The E. coli DNA Replication Fork. The Enzymes 39:31–
496 88.

497 14. Oakley AJ. 2019. A structural view of bacterial DNA replication. Protein science: a
498 publication of the Protein Society 28:990–1004.

499 15. Gorbalenya AE, Koonin EV. 1993. Helicases: amino acid sequence comparisons and
500 structure-function relationships. Current Opinion in Structural Biology 3:419–429.

16. Leipe DD, Aravind L, Grishin NV, Koonin EV. 2000. The bacterial replicative helicase DnaB evolved from a RecA duplication. Genome research 10:5–16.

17. Fernandez AJ, Berger JM. 2021. Mechanisms of hexameric helicases. Critical Reviews in Biochemistry and Molecular Biology 56:621–639.

18. Spinks RR, Spenkelink LM, Stratmann SA, Xu Z-Q, Stamford NPJ, Brown SE, Dixon NE, Jergic S, van Oijen AM. 2021. DnaB helicase dynamics in bacterial DNA replication resolved by single-molecule studies. Nucleic acids research 49:6804–6816.

19. Itsathitphaisarn O, Wing RA, Eliason WK, Wang J, Steitz TA. 2012. The hexameric helicase DnaB adopts a nonplanar conformation during translocation. Cell 151:267–277.

20. Nakayama N, Arai N, Kaziro Y, Arai K. 1984. Structural and functional studies of the dnaB protein using limited proteolysis. Characterization of domains for DNA-dependent ATP hydrolysis and for protein association in the primosome. The Journal of biological chemistry 259:88–96.

21. Sakamoto Y, Nakai S, Moriya S, Yoshikawa H, Ogasawara N. 1995. The Bacillus subtilis dnaC gene encodes a protein homologous to the DnaB helicase of Escherichia coli. Microbiology (Reading, England) 141 (Pt 3:641–644.

22. Liu B, Eliason WK, Steitz TA. 2013. Structure of a helicase–helicase loader complex reveals insights into the mechanism of bacterial primosome assembly. Nat Commun 4:2495.

23. Arias-Palomo E, Puri N, O'Shea Murray VL, Yan Q, Berger JM. 2019. Physical Basis for the Loading of a Bacterial Replicative Helicase onto DNA. Molecular cell 74:173-184.e4.

521 24. Kobori JA, Kornberg A. 1982. The Escherichia coli dnaC gene product. II. Purification,
522 physical properties, and role in replication. Journal of Biological Chemistry 257:13763–
523 13769.

524 25. Wahle E, Lasken RS, Kornberg A. 1989. The dnaB-dnaC replication protein complex of
525 Escherichia coli. II. Role of the complex in mobilizing dnaB functions. J Biol Chem
526 264:2469–2475.

527 26. Koonin EV. 1992. DnaC protein contains a modified ATP-binding motif and belongs to a
528 novel family of ATPases including also DnaA. Nucleic acids research 20:1997–1997.

529 27. Bruand C, Farache M, McGovern S, Ehrlich SD, Polard P. 2001. DnaB, DnaD and DnaI
530 proteins are components of the Bacillus subtilis replication restart primosome. Molecular
531 microbiology 42:245–255.

532 28. Brézellec P, Vallet-Gely I, Possoz C, Quevillon-Cheruel S, Ferat J-L. 2016. DciA is an
533 ancestral replicative helicase operator essential for bacterial replication initiation. Nat
534 Commun 7.

535 29. Davey MJ, Fang L, McInerney P, Georgescu RE, O'Donnell M. 2002. The DnaC helicase
536 loader is a dual ATP/ADP switch protein. EMBO J 21:3148–3159.

537 30. Ioannou C, Schaeffer PM, Dixon NE, Soultanas P. 2006. Helicase binding to DnaI exposes
538 a cryptic DNA-binding site during helicase loading in Bacillus subtilis. Nucleic Acids Res
539 34:5247–5258.

31. Mann KM, Huang DL, Hooppaw AJ, Logsdon MM, Richardson K, Lee HJ, Kimmey JM, Aldridge BB, Stallings CL. 2017. Rv0004 is a new essential member of the mycobacterial DNA replication machinery. PLoS Genet 13:e1007115.

32. Chan-Yao-Chong M, Marsin S, Quevillon-Cheruel S, Durand D, Ha-Duong T. 2020. Structural ensemble and biological activity of DciA intrinsically disordered region. Journal of Structural Biology 212:107573.

33. Marsin S, Adam Y, Cargemel C, Andreani J, Baconnais S, Legrand P, Li de la Sierra-Gallay I, Humbert A, Aumont-Nicaise M, Velours C, Ochsenbein F, Durand D, Le Cam E, Walbott H, Possoz C, Quevillon-Cheruel S, Ferat J-L. 2021. Study of the DnaB:DciA interplay reveals insights into the primary mode of loading of the bacterial replicative helicase. Nucleic Acids Research gkab463.

34. Joseph T Burke*, Samuel Z Chen*, Lauren M Sosinski*, John B Johnson, Janani Ravi. MolEvolvR: A web-app for characterizing proteins using molecular evolution and phylogeny. In preparation.

35. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, Thomson NR, Iqbal Z. 2021. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS biology 19:e3001421–e3001421.

36. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Functional & integrative genomics 15:141–161.

560 37. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes
561 database: new representation and annotation strategy. Nucleic Acids Research 42:D553–
562 D559.

563 38. Robinson A, Causer RJ, Dixon NE. 2012. Architecture and conservation of the bacterial
564 DNA replication machinery, an underexploited drug target. Current drug targets 13:352–
565 372.

566 39. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
567 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R,
568 Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nature Microbiology 1:16048–
569 16048.

570 40. Necci M, Piovesan D, Dosztányi Z, Tosatto SCE. 2017. MobiDB-lite: fast and highly specific
571 consensus prediction of intrinsic disorder in proteins. Bioinformatics (Oxford, England)
572 33:1402–1404.

573 41. Burroughs AM, Zhang D, Schäffer DE, Iyer LM, Aravind L. 2015. Comparative genomic
574 analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts,
575 immunity and signaling. Nucleic acids research, 2015/11/20 ed. 43:10633–10654.

576 42. Sutton MD, Carr KM, Vicente M, Kaguni JM. 1998. Escherichia coli DnaA protein. The N-
577 terminal domain and loading of DnaB helicase at the E. coli chromosomal origin. J Biol
578 Chem 273:34255–34262.

579 43. Abe Y, Jo T, Matsuda Y, Matsunaga C, Katayama T, Ueda T. 2007. Structure and function
580 of DnaA N-terminal domains: specific sites and mechanisms in inter-DnaA interaction and
581 in DnaB helicase loading on oriC. The Journal of biological chemistry 282:17816–17827.

582 44. Keyamura K, Abe Y, Higashi M, Ueda T, Katayama T. 2009. DiaA dynamics are coupled
583 with changes in initial origin complexes leading to helicase loading. The Journal of
584 biological chemistry 284:25038–25050.

585 45. Tan HY, Wilczek LA, Pottinger S, Manosas M, Yu C, Nguyenduc T, Bianco PR. 2017. The
586 intrinsically disordered linker of E. coli SSB is critical for the release from single-stranded
587 DNA. Protein science : a publication of the Protein Society 26:700–717.

588 46. Kozlov AG, Weiland E, Mittal A, Waldman V, Antony E, Fazio N, Pappu RV, Lohman TM.
589 2015. Intrinsically Disordered C-Terminal Tails of E. coli Single-Stranded DNA Binding
590 Protein Regulate Cooperative Binding to Single-Stranded DNA. Journal of Molecular
591 Biology 427:763–774.

592 47. Antony E, Lohman TM. 2019. Dynamics of E. coli single stranded DNA binding (SSB)
593 protein-DNA complexes. Seminars in cell & developmental biology 86:102–111.

594 48. Bianco PR, Pottinger S, Tan HY, Nguyenduc T, Rex K, Varshney U. 2017. The IDL of E. coli
595 SSB links ssDNA and protein binding by mediating protein-protein interactions. Protein
596 science : a publication of the Protein Society 26:227–241.

597 49. Guy CP, Atkinson J, Gupta MK, Mahdi AA, Gwynn EJ, Rudolph CJ, Moon PB, van
598 Knippenberg IC, Cadman CJ, Dillingham MS, Lloyd RG, McGlynn P. 2009. Rep Provides a
599 Second Motor at the Replisome to Promote Duplication of Protein-Bound DNA. Molecular
600 Cell 36:654–666.

601 50. Nguyen B, Shinn MK, Weiland E, Lohman TM. 2021. Regulation of E. coli Rep helicase
602 activity by PriC. Journal of Molecular Biology 433:167072–167072.

603 51. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse
604 B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V,
605 Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher
606 E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy
607 MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz
608 SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W,
609 Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016.
610 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
611 functional annotation. Nucleic acids research 44:D733-45.

612 52. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto
613 SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families
614 database in 2021. Nucleic Acids Research 49:D412–D419.

615 53. Nowotny J, Ahmed S, Xu L, Oluwadare O, Chen H, Hensley N, Trieu T, Cao R, Cheng J.
616 2015. Iterative reconstruction of three-dimensional models of human chromosomes from
617 chromosomal contact data. BMC bioinformatics 16:338–338.

618 54. Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal
619 peptide prediction method. Journal of molecular biology 338:1027–1036.

620 55. Möller S, Croning MD, Apweiler R. 2001. Evaluation of methods for the prediction of
621 membrane spanning regions. Bioinformatics (Oxford, England) 17:646–653.

622 56. Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, Quaglia F, Paladin L,
623 Ramasamy P, Dosztányi Z, Vranken WF, Davey NE, Parisi G, Fuxreiter M, Tosatto SCE.

2021. MobiDB: intrinsically disordered proteins in 2021. Nucleic acids research 49:D361–D367.

57. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research 47:W636—W641.

58. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

59. Campanella JJ, Bitincka L, Smalley J. 2003. MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences. BMC Bioinformatics 4:29–29.

60. Troshin PV, Procter JB, Barton GJ. 2011. Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA. Bioinformatics 27:2001–2002.

61. Lassmann T. 2020. Kalign 3: multiple sequence alignment of large datasets. Bioinformatics 36:1928–1929.

Figure 1

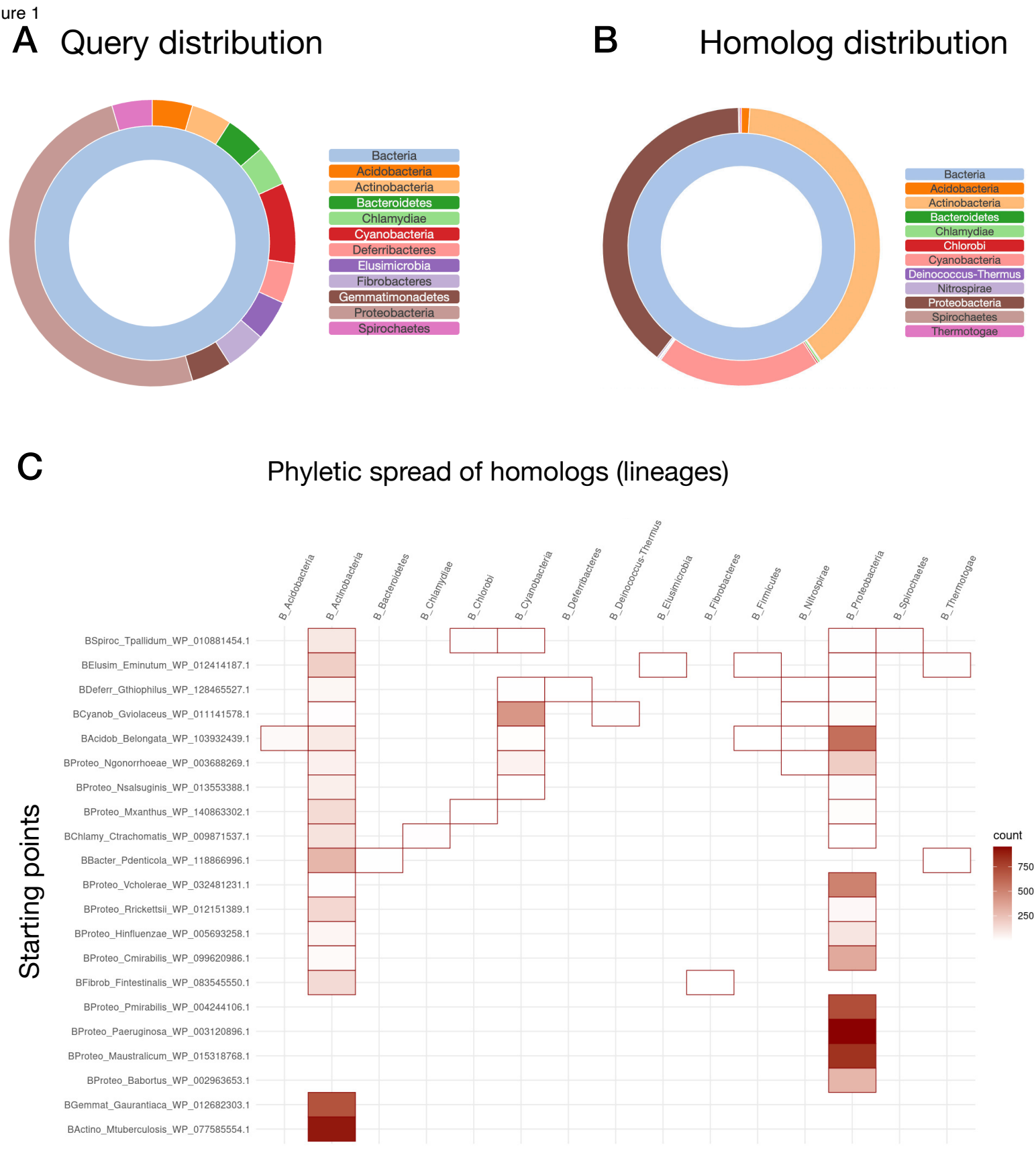


Figure 2

Phylum	*Species*	*B. elo*	*M. tub*	*P. den*	*C. tra*	*G. vio*	*G. thio*	*E. min*	*F. int*	*G. aur*	*B. abo*	*P. aer*	*N. gon*	*P. mir*	*H. inf*	*R. ric*	*N. sal*	*M. aus*	*V. cho*	*C. mir*	*M. xan*	*T. pal*
Acidobacteria	*Bryocella elongata*																					
Actinobacteria	*Mycobacterium tuberculosis*	23.0																				
Bacteroidetes	*Prevotella denticola*	42.3	26.2																			
Chlamydiae	*Chlamydia trachomatis*	31.9	28.9	41.4																		
Cyanobacteria	*Gloeobacter violaceus*	26.7	33.2	26.1	28.9																	
Deferribacteres	*Geovibrio thiophilus*	32.9	28.9	27.4	30.8	34.4																
Elusimicrobia	*Elusimicrobium minutum*	38.5	21.4	38.8	39.7	26.1	29.5															
Fibrobacteres	*Fibrobacter intestinalis*	36.3	27.3	41.6	42.2	26.7	30.8	35.4														
Gemmatimonadetes	*Gemmatimonas aurantiaca*	39.4	28.3	43.7	41.4	28.3	33.6	38.8	34.5													
Proteobacteria	*Brucella abortus*	25.1	31.0	26.3	29.1	33.3	32.6	20.6	25.7	33.1												
Proteobacteria	*Pseudomonas aeruginosa*	29.8	28.3	31.3	31.3	30.0	34.2	30.5	31.3	29.8	36.0											
Proteobacteria	*Neisseria gonorrhoeae*	30.7	29.9	29.3	29.3	32.8	37.0	25.7	28.6	29.3	34.9	39.3										
Proteobacteria	*Proteus mirabilis*	26.2	29.9	26.7	32.0	38.3	39.0	29.1	32.0	24.4	40.0	44.2	38.4									
Proteobacteria	*Haemophilus influenzae*	35.6	23.0	40.4	38.8	23.3	29.5	42.3	39.8	36.5	21.1	28.2	26.4	26.7								
Proteobacteria	*Rickettsia rickettsii*	30.8	25.7	43.9	39.7	26.1	31.5	38.3	39.8	36.4	26.9	30.5	30.0	26.7	32.7							
Proteobacteria	*Nitratifractor salsuginis*	24.8	26.7	24.8	22.9	30.6	32.7	28.8	29.4	26.1	29.7	27.5	30.7	32.6	27.5	33.3						
Proteobacteria	*Mesorhizobium australicum*	25.9	32.1	24.7	27.1	37.2	34.3	23.5	27.1	27.7	62.3	37.3	35.5	37.2	29.5	27.1	28.9					
Proteobacteria	*Vibrio cholerae*	27.4	27.8	26.1	32.5	32.2	29.9	26.1	30.6	36.3	33.1	42.0	38.9	52.3	30.6	32.5	35.7	34.9				
Proteobacteria	*Caulobacter mirabilis*	28.0	34.8	26.9	27.5	40.7	34.1	22.0	25.3	28.0	47.8	28.6	35.2	34.1	23.1	27.5	33.0	46.2	35.7			
Proteobacteria	*Myxococcus xanthus*	38.5	28.3	42.6	32.8	26.7	31.5	34.7	40.7	41.7	26.9	31.3	29.3	26.2	27.9	40.2	25.5	30.1	22.9	26.9		
Spirochaetes	*Treponema pallidum*	31.0	31.0	28.3	29.7	31.1	39.0	28.3	26.2	32.4	29.7	36.6	31.7	38.4	34.5	29.7	35.3	34.3	37.6	31.9	26.9	

Figure 3

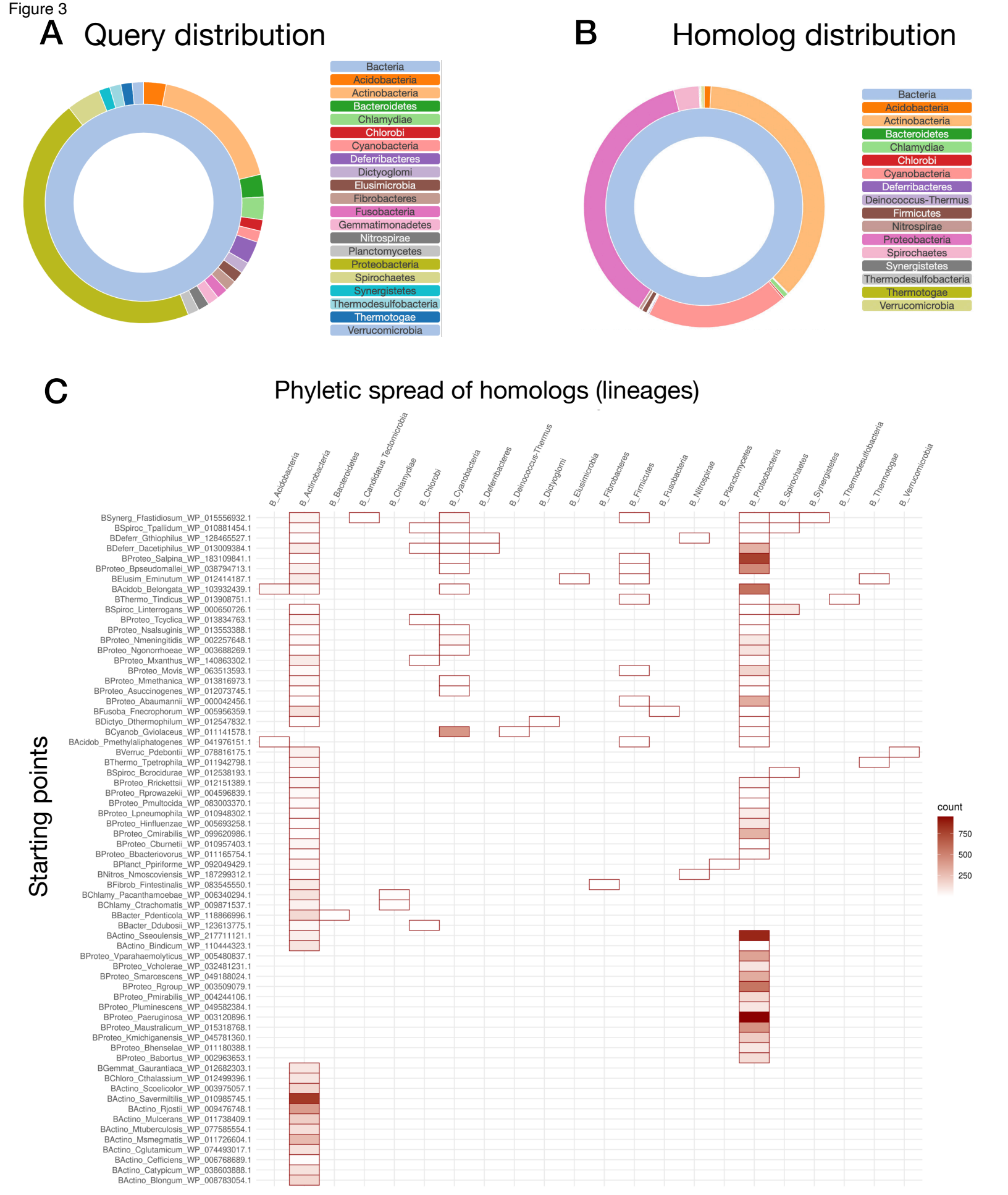


Figure 4

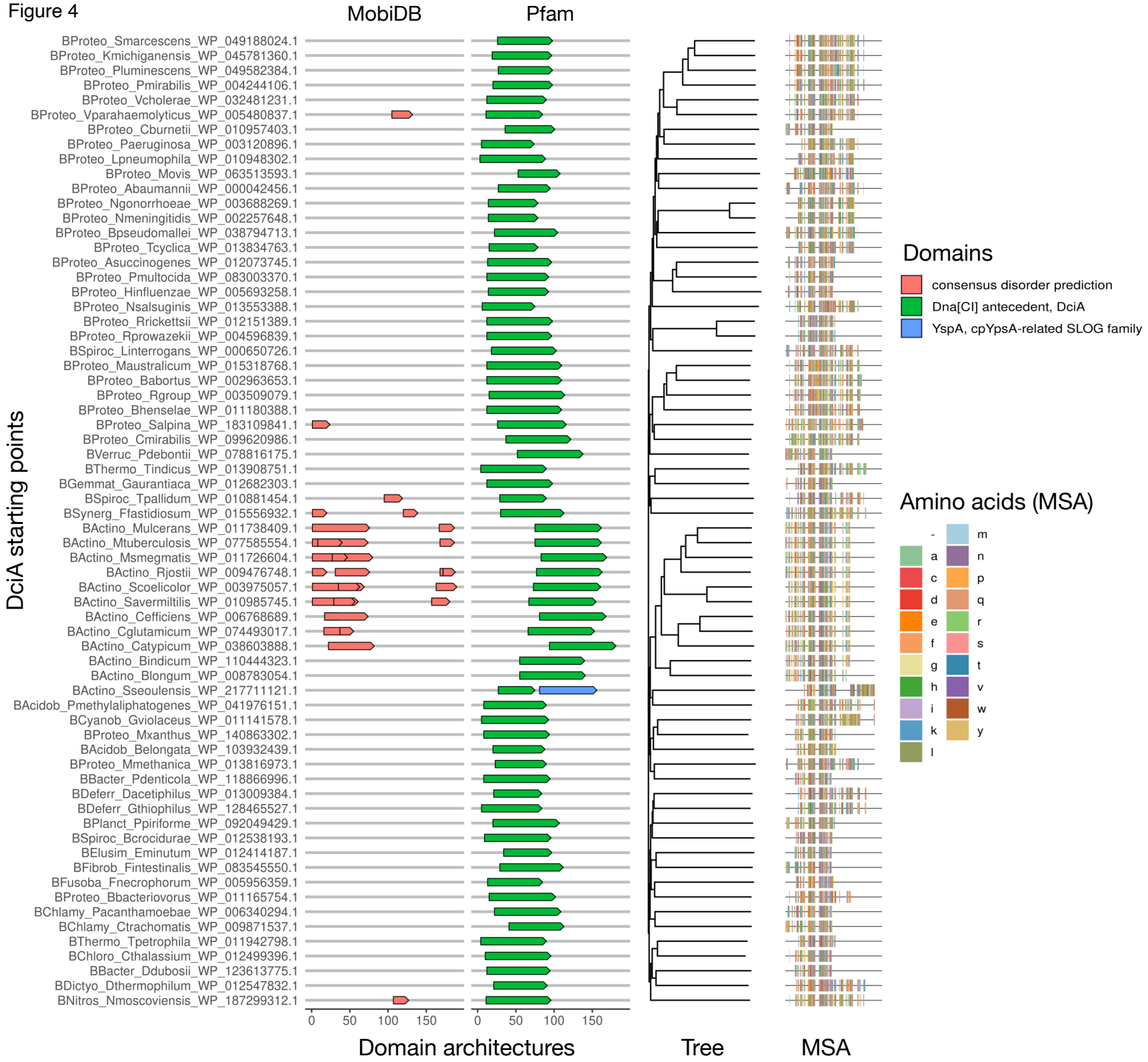
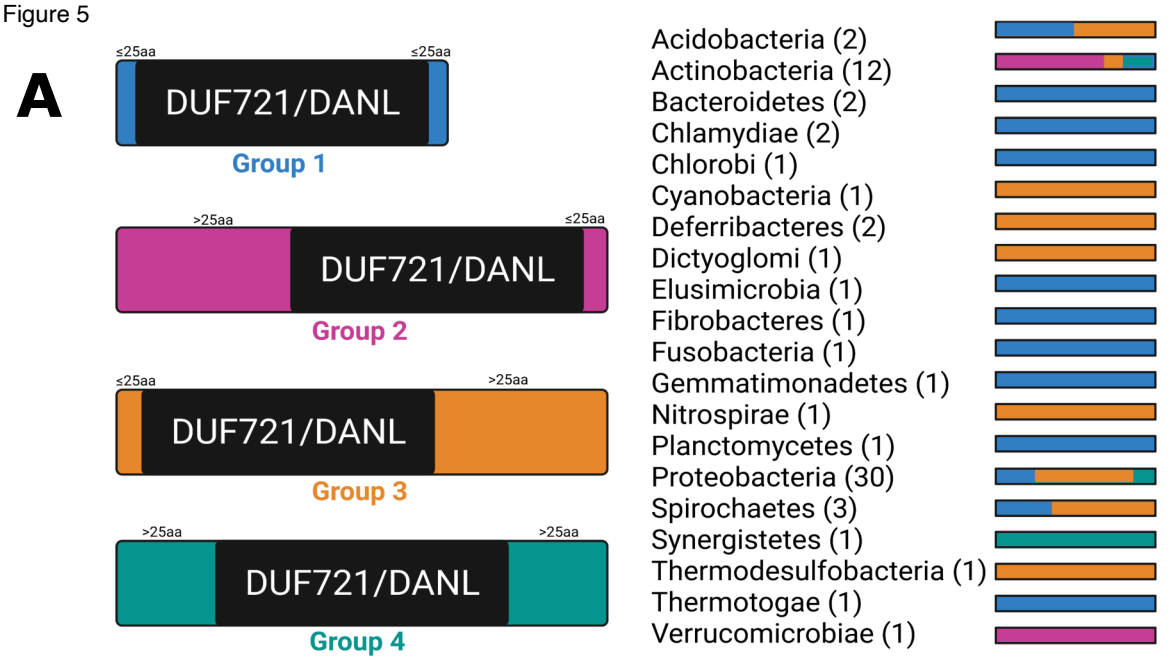

MobiDB
Pfam
BProteo_Smarcescens_WP_049188024.1
BProteo_Kmichiganensis_WP_045781360.1
BProteo_Pluminescens_WP_049582384.1
BProteo_Pmirabilis_WP_004244106.1
BProteo_Vcholerae_WP_032481231.1
BProteo_Vparahaemolyticus_WP_005480837.1
BProteo_Cburnetii_WP_010957403.1
BProteo_Paeruginosa_WP_003120896.1
BProteo_Lpneumophila_WP_010948302.1
BProteo_Movis_WP_063513593.1
BProteo_Abaumannii_WP_000042456.1
BProteo_Ngonorrhoeae_WP_003688269.1
BProteo_Nmeningitidis_WP_002257648.1
BProteo_Bpseudomallei_WP_038794713.1
BProteo_Tcyclica_WP_013834763.1
BProteo_Asuccinogenes_WP_012073745.1
BProteo_Pmultocida_WP_083003370.1
BProteo_Hinfluenzae_WP_005693258.1
BProteo_Nsalsuginis_WP_013553388.1
BProteo_Rrickettsii_WP_012151389.1
BProteo_Rprowazekii_WP_004596839.1
BSpiroc_Linterrogans_WP_000650726.1
BProteo_Maustralicum_WP_015318768.1
BProteo_Babortus_WP_002963653.1
BProteo_Rgroup_WP_003509079.1
BProteo_Bhenselae_WP_011180388.1
BProteo_Salpina_WP_183109841.1
BProteo_Cmirabilis_WP_099620986.1
BVerruc_Pdebontii_WP_078816175.1
BThermo_Tindicus_WP_013908751.1
BGemmat_Gaurantiaca_WP_012682303.1
BSpiroc_Tpallidum_WP_010881454.1
BSynerg_Ffastidiosum_WP_015556932.1
BActino_Mulcerans_WP_011738409.1
BActino_Mtuberculosis_WP_077585554.1
BActino_Msmegmatis_WP_011726604.1
BActino_Rjostii_WP_009476748.1
BActino_Scoelicolor_WP_003975057.1
BActino_Savermiltilis_WP_010985745.1
BActino_Cefficiens_WP_006768689.1
BActino_Cglutamicum_WP_074493017.1
BActino_Catypicum_WP_038603888.1
BActino_Bindicum_WP_110444323.1
BActino_Blongum_WP_008783054.1
BActino_Sseoulensis_WP_217711121.1
BAcidob_Pmethylaliphatogenes_WP_041976151.1
BCyanob_Gviolaceus_WP_011141578.1
BProteo_Mxanthus_WP_140863302.1
BAcidob_Belongata_WP_103932439.1
BProteo_Mmethanica_WP_013816973.1
BBacter_Pdenticola_WP_118866996.1
BDeferr_Dacetiphilus_WP_013009384.1
BDeferr_Gthiophilus_WP_128465527.1
BPlanct_Ppiriforme_WP_092049429.1
BSpiroc_Bcrocidurae_WP_012538193.1
BElusim_Eminutum_WP_012414187.1
BFibrob_Fintestinalis_WP_083545550.1
BFusoba_Fnecrophorum_WP_005956359.1
BProteo_Bbacteriovorus_WP_011165754.1
BChlamy_Pacanthamoebae_WP_006340294.1
BChlamy_Ctrachomatis_WP_009871537.1
BThermo_Tpetrophila_WP_011942798.1
BChloro_Cthalassium_WP_012499396.1
BBacter_Ddubosii_WP_123613775.1
BDictyo_Dthermophilum_WP_012547832.1
BNitros_Nmoscoviensis_WP_187299312.1
DciA starting points
0
50
100
150
0
50
100
150
Domain architectures
Tree
MSA
Domains
consensus disorder prediction
Dna[CI] antecedent, DciA
YspA, cpYpsA-related SLOG family
Amino acids (MSA)
-
a
c
d
e
f
g
h
i
k
l
m
n
p
q
r
s
t
v
w
y

Figure 5


A
≤25aa
≤25aa
DUF721/DANL
Group 1
>25aa
≤25aa
DUF721/DANL
Group 2
≤25aa
>25aa
DUF721/DANL
Group 3
>25aa
>25aa
DUF721/DANL
Group 4
Acidobacteria (2)
Actinobacteria (12)
Bacteroidetes (2)
Chlamydiae (2)
Chlorobi (1)
Cyanobacteria (1)
Deferribacteres (2)
Dictyoglomi (1)
Elusimicrobia (1)
Fibrobacteres (1)
Fusobacteria (1)
Gemmatimonadetes (1)
Nitrospirae (1)
Planctomycetes (1)
Proteobacteria (30)
Spirochaetes (3)
Synergistetes (1)
Thermodesulfobacteria (1)
Thermotogae (1)
Verrucomicrobiae (1)

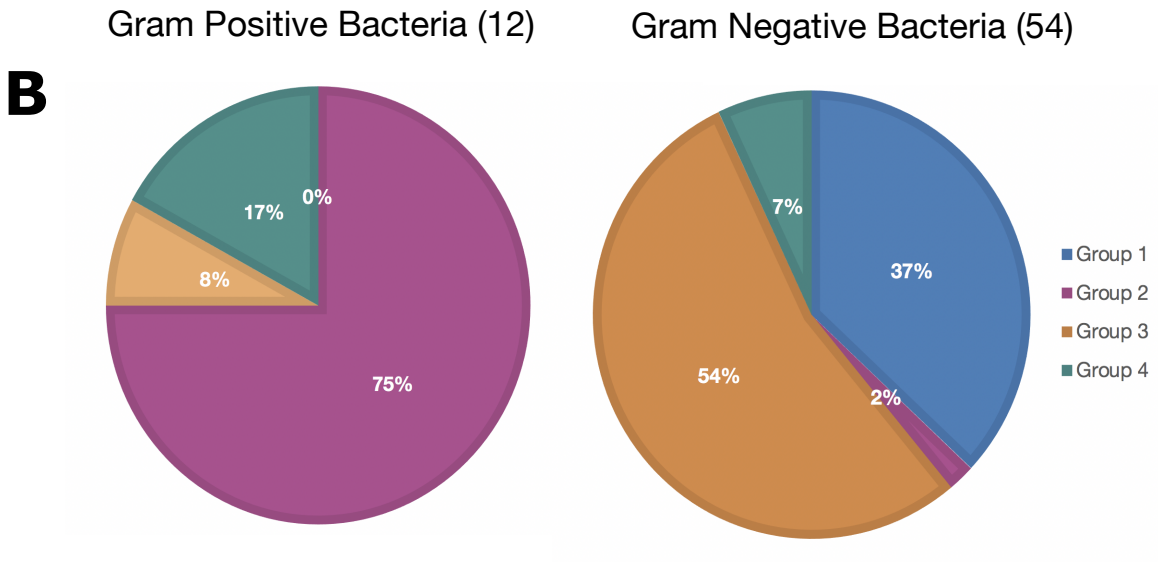
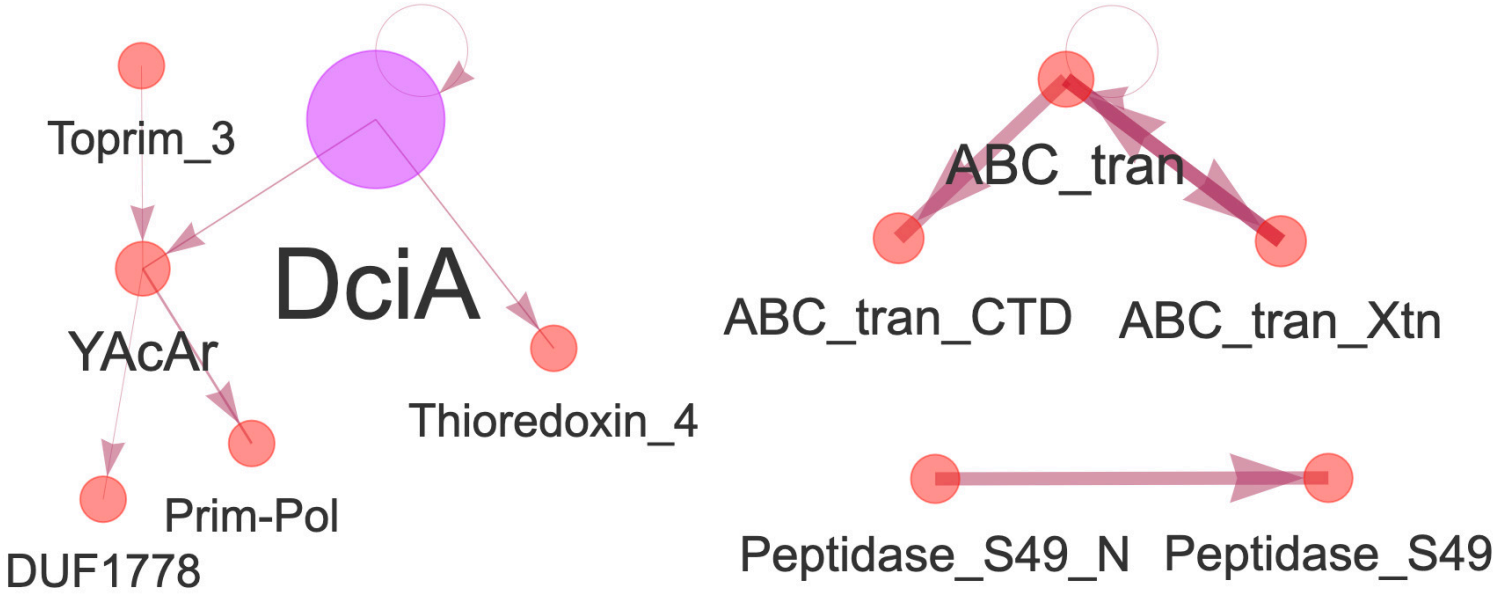

B
Gram Positive Bacteria (12)
0%
17%
8%
75%
Gram Negative Bacteria (54)
7%
37%
54%
2%
Group 1
Group 2
Group 3
Group 4

Figure 6

A Domain proximity network



B Domain co-occurrence frequency

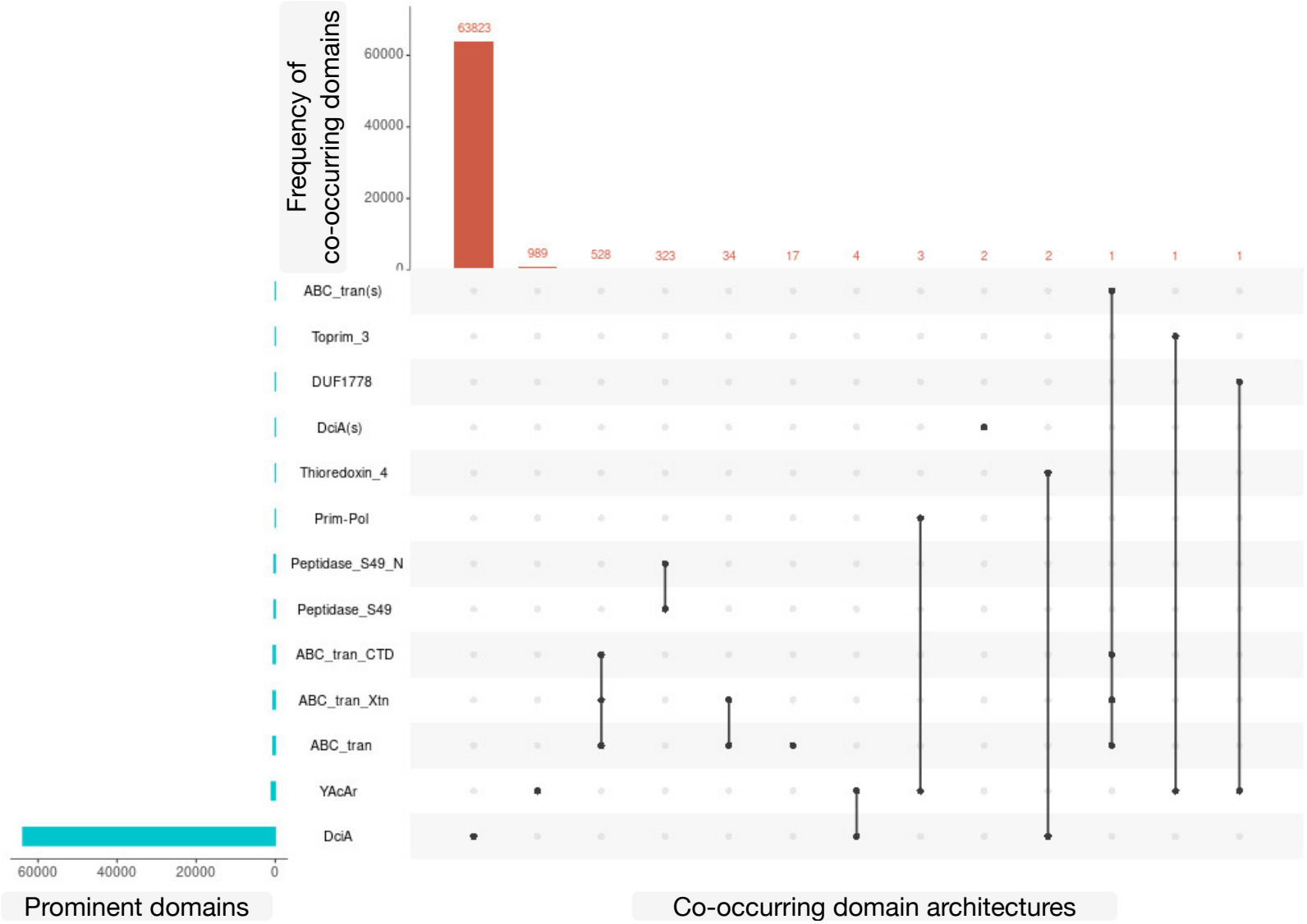


Table 1

Table 1: DciA query proteins used to identify homologs across the bacterial kingdom

Name	AccNum	Species	TaxID	Lineage	DomArch.Pfam	DomArch.MobiDBLite	Length	DUF721 range (aa)	DciA group	Gram stain
BProteo_Abaumannii_WP_000042456.1	WP_0000424[illegible]6.1	[illegible] baumannii	[illegible]70	Bacteria>Proteobacteria	DciA	NA	[illegible]	27-[illegible]	Group 4	-
BSpiroc_Linterrogans_WP_000650726.1	WP_000650726.1	Leptospira interrogans	171	Bacteria>Spirochaetes	DciA	NA	154	15-103	Group 3	-
BProteo_N[illegible]_WP_00[illegible]257648.1	WP_00[illegible]257648.1	Neisseria meningitidis	[illegible]	Bacteria>Proteobacteria	DciA	NA	[illegible]	[illegible]-79	[illegible]	-
BProteo_Babortus_WP_002963653.1	WP_002963653.1	Brucella abortus	234	Bacteria>Proteobacteria	DciA	NA	175	12-110	Group 3	-
BProteo_Paeruginosa_WP_003120896.1	WP_003120896.1	Pseudomonas aeruginosa	286	Bacteria>Proteobacteria	DciA	NA	131	5-74	Group 3	-
BProteo_Rgroup_WP_003509079.1	WP_003509079.1	Rhizobium/Agrobacterium group	227290	Bacteria>Proteobacteria	DciA	NA	173	14-114	Group 3	-
BProteo_Ngonorrhoeae_WP_003688269.1	WP_003688269.1	Neisseria gonorrhoeae	485	Bacteria>Proteobacteria	DciA	NA	140	3-88	Group 3	-
BActino_Scoelicolor_WP_003975057.1	WP_003975057.1	Streptomyces coelicolor	1883	Bacteria>Actinobacteria	DciA	disorder region+disorder region+disorder region	190	72-161	Group 4	+
BProteo_Pmirabilis_WP_004244106.1	WP_004244106.1	Proteus mirabilis	583	Bacteria>Proteobacteria	DciA	NA	172	8-100	Group 3	-
BProteo_Rprowazekii_WP_004596839.1	WP_004596839.1	Rickettsia prowazekii	782	Bacteria>Proteobacteria	DciA	NA	108	4-98	Group 1	-
BProteo_Vparahaemolyticus_WP_005480837.1	WP_005480837.1	Vibrio parahaemolyticus	670	Bacteria>Proteobacteria	DciA	disorder region	151	8-94	Group 3	-
BProteo_Hinfluenzae_WP_005693258.1	WP_005693258.1	Haemophilus influenzae	727	Bacteria>Proteobacteria	DciA	NA	104	13-98	Group 1	-
BFusoba_Fnecrophorum_WP_005956359.1	WP_005956359.1	Fusobacterium necrophorum	859	Bacteria>Fusobacteria	DciA	NA	99	4-89	Group 1	-
BChlamy_Pacanthamoebae_WP_006340294.1	WP_006340294.1	Parachlamydia acanthamoebae	83552	Bacteria>Chlamydiae	DciA	NA	112	21-109	Group 1	variable (NA), counted as -
BActino_Cefficiens_WP_006768689.1	WP_006768689.1	Corynebacterium efficiens	152794	Bacteria>Actinobacteria	DciA	disorder region+disorder region	193	81-168	Group 2	+
BActino_Blongum_WP_008783054.1	WP_008783054.1	Bifidobacterium longum	216816	Bacteria>Actinobacteria	DciA	NA	156	54-141	Group 2	+
BActino_Rjostii_WP_009476748.1	WP_009476748.1	Rhodococcus jostii	1827	Bacteria>Actinobacteria	DciA	disorder region+disorder region+disorder region+disorder region	188	76-163	Group 2	+
BChlamy_Ctrachomatis_WP_009871537.1	WP_009871537.1	Chlamydia trachomatis	813	Bacteria>Chlamydiae	DciA	NA	116	25-113	Group 1	variable (NA), counted as -
BSpiroc_Tpallidum_WP_010881454.1	WP_010881454.1	Treponema pallidum	160	Bacteria>Spirochaetes	DciA	disorder region+disorder region	145	7-93	Group 3	-
BProteo_Lpneumophila_WP_010948302.1	WP_010948302.1	Legionella pneumophila	446	Bacteria>Proteobacteria	DciA	NA	140	3-89	Group 3	-
BProteo_Cburnetii_WP_010957403.1	WP_010957403.1	Coxiella burnetii	777	Bacteria>Proteobacteria	DciA	NA	114	25-109	Group 1	-
BActino_Savermiltilis_WP_010985745.1	WP_010985745.1	Streptomyces avermiltilis	1883	Bacteria>Actinobacteria	DciA	disorder region+disorder region+disorder region	181	66-155	Group 4	+
BCyanob_Gviolaceus_WP_011141578.1	WP_011141578.1	Gloeobacter violaceus	33072	Bacteria>Cyanobacteria	DciA	NA	180	5-93	Group 3	-
BProteo_Bbacteriovorus_WP_011165754.1	WP_011165754.1	Bdellovibrio bacteriovorus	959	Bacteria>Proteobacteria	DciA	NA	131	14-102	Group 3	-
BProteo_Bhenselae_WP_011180388.1	WP_011180388.1	Bartonella henselae	38323	Bacteria>Proteobacteria	DciA	NA	166	12-110	Group 3	-
BActino_Msmegmatis_WP_011726604.1	WP_011726604.1	Mycolicibacterium smegmatis	1772	Bacteria>Actinobacteria	DciA	disorder region+disorder region	194	82-169	Group 2	+
BActino_Mulcerans_WP_011738409.1	WP_011738409.1	Mycobacterium ulcerans	1809	Bacteria>Actinobacteria	DciA	disorder region+disorder region	187	75-162	Group 2	+
BThermo_Tpetrophila_WP_011942798.1	WP_011942798.1	Thermotoga petrophila	93929	Bacteria>Thermotogae	DciA	NA	101	3-90	Group 1	-
BProteo_Asuccinogenes_WP_012073745.1	WP_012073745.1	Actinobacillus succinogenes	67854	Bacteria>Proteobacteria	DciA	NA	110	12-97	Group 1	-
BProteo_Rrickettsii_WP_012151389.1	WP_012151389.1	Rickettsia rickettsii	783	Bacteria>Proteobacteria	DciA	NA	107	4-98	Group 1	-
BElusim_Eminutum_WP_012414187.1	WP_012414187.1	Elusimicrobium minutum	423605	Bacteria>Elusimicrobia	DciA	NA	98	9-97	Group 1	-
BChloro_Cthalassium_WP_012499396.1	WP_012499396.1	Chloroherpeton thalassium	100716	Bacteria>Chlorobi	DciA	NA	97	9-96	Group 1	-
BSpiroc_Bcrocidurae_WP_012538193.1	WP_012538193.1	Borrelia crocidurae	29520	Bacteria>Spirochaetes	DciA	NA	99	8-96	Group 1	-
BDictyo_Dthermophilum_WP_012547832.1	WP_012547832.1	Dictyoglomus thermophilum	14	Bacteria>Dictyoglomi	DciA	NA	160	8-92	Group 3	-
BGemmat_Gaurantiaca_WP_012682303.1	WP_012682303.1	Gemmatimonas aurantiaca	173480	Bacteria>Gemmatimonadetes	DciA	NA	103	11-99	Group 1	-
BDeferr_Dacetiphilus_WP_013009384.1	WP_013009384.1	Denitrovibrio acetiphilus	118000	Bacteria>Deferribacteres	DciA	NA	147	4-84	Group 3	-
BProteo_Nsalsuginis_WP_013553388.1	WP_013553388.1	Nitratifractor salsuginis	269261	Bacteria>Proteobacteria	DciA	NA	153	6-75	Group 3	-
BProteo_Mmethanica_WP_013816973.1	WP_013816973.1	Methylomonas methanica	421	Bacteria>Proteobacteria	DciA	NA	154	14-96	Group 3	-
BProteo_Tcyclica_WP_013834763.1	WP_013834763.1	Thiomicrospira cyclica	147268	Bacteria>Proteobacteria	DciA	NA	143	6-89	Group 3	-
BThermo_Tindicus_WP_013908751.1	WP_013908751.1	Thermodesulfatator indicus	171695	Bacteria>Thermodesulfobacteria	DciA	NA	159	3-90	Group 3	-
BProteo_Maustralicum_WP_015318768.1	WP_015318768.1	Mesorhizobium australicum	68287	Bacteria>Proteobacteria	DciA	NA	166	12-110	Group 3	-
BSynerg_Ffastidiosum_WP_015556932.1	WP_015556932.1	Fretibacterium fastidiosum	651822	Bacteria>Synergistetes	DciA	disorder region+disorder region	176	29-113	Group 4	-
BProteo_Vcholerae_WP_032481231.1	WP_032481231.1	Vibrio cholerae	666	Bacteria>Proteobacteria	DciA	NA	157	8-94	Group 3	-
BActino_Catypicum_WP_038603888.1	WP_038603888.1	Corynebacterium atypicum	191610	Bacteria>Actinobacteria	DciA	disorder region	206	94-181	Group 2	+
BProteo_Bpseudomallei_WP_038794713.1	WP_038794713.1	Burkholderia pseudomallei	28450	Bacteria>Proteobacteria	DciA	NA	159	22-105	Group 3	-
BAcidob_Pmethylaliphatogenes_WP_041976151.1	WP_041976151.1	Pyrinomonas methylaliphatogenes	454194	Bacteria>Acidobacteria	DciA	NA	158	3-90	Group 3	-
BProteo_Kmichiganensis_WP_045781360.1	WP_045781360.1	Klebsiella michiganensis	1134687	Bacteria>Proteobacteria	DciA	NA	171	8-98	Group 3	-
BProteo_Smarcescens_WP_049188024.1	WP_049188024.1	Serratia marcescens	615	Bacteria>Proteobacteria	DciA	NA	171	8-99	Group 3	-
BProteo_Pluminescens_WP_049582384.1	WP_049582384.1	Photorhabdus luminescens	29488	Bacteria>Proteobacteria	DciA	NA	172	15-100	Group 3	-
BProteo_Movis_WP_063513593.1	WP_063513593.1	Moraxella ovis	29433	Bacteria>Proteobacteria	DciA	NA	172	53-108	Group 4	-
BActino_Cglutamicum_WP_074493017.1	WP_074493017.1	Corynebacterium glutamicum	1718	Bacteria>Actinobacteria	DciA	disorder region+disorder region	178	66-153	Group 2	+
BActino_Mtuberculosis_WP_077585554.1	WP_077585554.1	Mycobacterium tuberculosis	1773	Bacteria>Actinobacteria	DciA	disorder region+disorder region+disorder region	187	75-162	Group 2	+
BVerruc_Pdebontii_WP_078816175.1	WP_078816175.1	Prosthecobacter debontii	48467	Bacteria>Verrucomicrobia	DciA	NA	141	52-138	Group 2	-
BProteo_Pmultocida_WP_083003370.1	WP_083003370.1	Pasteurella multocida	747	Bacteria>Proteobacteria	DciA	NA	101	8-93	Group 1	-
BFibrob_Fintestinalis_WP_083545550.1	WP_083545550.1	Fibrobacter intestinalis	28122	Bacteria>Fibrobacteres	DciA	NA	113	25-112	Group 1	-
BPlanct_Ppiriforme_WP_092049429.1	WP_092049429.1	Planctomicrobium piriforme	1576369	Bacteria>Planctomycetes	DciA	NA	116	14-102	Group 1	- (but unique)
BProteo_Cmirabilis_WP_099620986.1	WP_099620986.1	Caulobacter mirabilis	69666	Bacteria>Proteobacteria	DciA	NA	182	31-122	Group 4	-
BAcidob_Belongata_WP_103932439.1	WP_103932439.1	Bryocella elongata	863522	Bacteria>Acidobacteria	DciA	NA	104	3-89	Group 1	-
BActino_Bindicum_WP_110444323.1	WP_110444323.1	Bifidobacterium indicum	1691	Bacteria>Actinobacteria	DciA	NA	165	54-140	Group 2	+
BBacter_Pdenticola_WP_118866996.1	WP_118866996.1	Prevotella denticola	28129	Bacteria>Bacteroidetes	DciA	NA	96	8-95	Group 1	-
BBacter_Ddubosii_WP_123613775.1	WP_123613775.1	Duncaniella dubosii	2518971	Bacteria>Bacteroidetes	DciA	NA	96	8-95	Group 1	-
BDeferr_Gthiophilus_WP_128465527.1	WP_128465527.1	Geovibrio thiophilus	139438	Bacteria>Deferribacteres	DciA	NA	146	2-84	Group 3	-
BProteo_Mxanthus_WP_140863302.1	WP_140863302.1	Myxococcus xanthus	34	Bacteria>Proteobacteria	DciA	NA	101	8-94	Group 1	-
BProteo_Salpina_WP_183109841.1	WP_183109841.1	Sphingomonas alpina	653931	Bacteria>Proteobacteria	DciA	disorder region+disorder region	184	25-118	Group 3	-
BNitros_Nmoscoviensis_WP_187299312.1	WP_187299312.1	Nitrospira moscoviensis	42253	Bacteria>Nitrospirae	DciA	disorder region	155	11-96	Group 3	-
BActino_Sseoulensis_WP_217711121.1	WP_217711121.1	Streptomyces seoulensis	73044	Bacteria>Actinobacteria	DciA+YAcAr	NA	206	1-75	Group 3	+