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

25 phyla-specific analyses of this fundamental biological process, and is an important example of
26 how research in bacterial DNA replication is necessary in organisms beyond *E. coli* and *B. subtilis.*
27
28 **IMPORTANCE**
29 Despite the fundamental importance of DNA replication for life, this process remains understudied
30 in bacteria outside of *Escherichia coli* and *Bacillus subtilis*. In particular, most bacteria do not
31 encode the helicase loading proteins that are essential in *E. coli* and *B. subtilis* for DNA replication.
32 Instead, most bacteria encode a DciA homolog that likely constitutes the predominant mechanism
33 of helicase operation in bacteria. However, it is still unknown how DciA structure and function
34 compares across diverse phyla that encode DciA proteins. In this study, we perform a
35 computational evolutionary analysis that uncovers tremendous diversity amongst DciA homologs.
36 These studies provide a significant advance in our understanding regarding an essential
37 component of the bacterial DNA replication machinery.
38
39

40 **INTRODUCTION**

41 DNA replication is a process critical to life for all organisms The current paradigm for the process
42 of DNA replication in bacteria has primarily been based on studies in *Escherichia coli* and *Bacillus*
43 *subtilis*. Bacterial DNA replication begins with the binding of the replication initiation protein DnaA
44 to specific sequences referred to as DnaA boxes at the origin of replication (oriC) (1–7). DnaA
45 binding to double-stranded DNA (dsDNA) triggers DNA unwinding at an AT-rich region of DNA
46 called the DNA unwinding element (DUE), leaving a bubble of single-stranded DNA (ssDNA) (2,
47 3, 6, 8, 9). The ssDNA bubble is coated by single-stranded binding protein (SSB) (10), followed
48 by the concerted loading of two hexameric replicative helicases onto the SSB coated replication
49 fork. The two helicases translocate along the two sides of the replication fork, unwinding the
50 dsDNA as they move (1, 3, 6, 11–14).

51 Bacterial replicative helicases (DnaB in *E. coli* and DnaC in *B. subtilis*) are Superfamily IV
52 type helicases, which are defined as hexameric RecA ATPases (3, 15, 16) that translocate in the
53 5'-3' direction (11, 12, 17). The bacterial replicative helicase translocates on ssDNA using a
54 "hand-over-hand" mechanism, which is driven by nucleotide hydrolysis (18, 19) (reviewed in
55 (17)). The C-terminus of the bacterial replicative helicase contains the RecA-like fold that is
56 responsible for the ATPase activity, and is connected to an N-terminal scaffolding domain via a
57 linker region (6, 20, 21). The replicative helicase must oligomerize into a double-layered
58 hexameric ring to be active during replication, with one layer made up of the N-termini and the
59 other layer comprised of the C-termini (6, 22, 23). In *E. coli* and *B. subtilis*, the loading of the
60 replicative helicase is performed with the help of a helicase loader, termed DnaC in *E. coli* and
61 DnaI in *B. subtilis* (3, 24–27). *dnaC* and *dnaI* were acquired by *E. coli* and *B. subtilis*, respectively,
62 via domestication of related but distinct phage ATPase-containing genes (28). DnaC and DnaI
63 are both in the ATPases Associated with diverse cellular Activities (AAA+) ATPase family, and

64 the ATPase activity of DnaC is required for its helicase loading function at the origin of replication
65 (29, 30).

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

80 Beyond the presence of the DUF721, DciA domain architecture and the relationship
81 between DciA homologs across diverse bacterial phyla are yet to be investigated. To address
82 these open questions, we took a computational evolutionary approach and analyzed the
83 phylogenic distribution, domain architecture, and sequence conservation for DciA homologs. We
84 have discovered low sequence similarity between DciA homologs from different phyla, lineage-
85 specific domain architectures, and divergent evolution of specific DciA homologs, all of which
86 likely have functional consequences. This study provides an evolutionary picture of DciA, reveals
87 key differences between homologs, and generates the framework for mechanistic investigation
88 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

112 both genome sequencing and publication bias, proteobacteria and actinobacteria were over-
113 represented in both our queries and recovered sequences (35–37) (**Figure 1A,B**).
114 No single DciA protein identified homologs in all other phyla (**Figure 1C**), supporting that
115 there is low sequence conservation between DciA homologs. To quantify the sequence
116 conservation across phyla, we analyzed the pairwise similarity for the 21 DciA protein homologs
117 used as our query set (**Figure 2**). We found a wide range of similarities (~20–60%), with the
118 majority of homologs showing 30-40% similarity. Our query sequence dataset contained multiple
119 species within each class of proteobacteria, so we were able to compare the conservation
120 between DciA homologs within proteobacteria as well as between proteobacteria and other phyla.
121 Similarity was high between DciA proteins within alphaproteobacteria (62.3% between *Brucella*
122 *abortus* and *Mesorhizobium australicum*) and gammaproteobacteria (52.3% between *Proteus*
123 *mirabilis* and *Vibrio cholerae*). Overall, DciA protein queries within proteobacteria have an
124 average of ~30% similarity, and proteobacterial homologs have an average of 31.12% similarity
125 to homologs outside of their phyla **(Figure 2)**. Actinobacterial and proteobacterial DciA homologs
126 share between 21–34% sequence similarity. For example, the *M. tuberculosis* DciA protein shares
127 28.3% identity with the *Pseudomonas aeruginosa* DciA (**Figure 2**). DciA shares this trait of low
128 sequence conservation across phyla with its interaction partner, the replicative helicase DnaB,
129 where DnaB proteins in *M. tuberculosis* and *P. aeruginosa* only share 20.4% similarity. Bacterial
130 replication proteins in general have low to moderate sequence conservation across phyla (11–
131 49% similarity across replisome proteins between *E. coli* and *B. subtilis*) (38). Therefore, our data
132 showing low DciA sequence conservation is consistent with other replication initiation proteins.
133 Given the low-level conservation across divergent DciA proteins, it was unsurprising that
134 individual DciA query proteins never returned homologs from all other phyla and emphasizes the
135 need for multiple starting points for analysis (**Figure 1C**).

136 While we identified a diverse set of DciA homologs, including those with moderate to low
137 sequence conservation from most lineages (**Figure 1**), we were still missing multiple phyla
138 previously reported to contain DciA homologs. Therefore, we expanded our search further by
139 including 66 DciA query proteins from 20 bacterial phyla as query sequences (**Figure 3A**; **Table**
140 **1**). This hugely diversified our results identifying >13K unique DciA homologs from 22 bacterial
141 phyla (**Figure 3B,C**). Consistent with the previous smaller set of DciA query proteins (**Figure 1**),
142 we found that most DciA proteins identified homologs within their corresponding phylum as well
143 as ones within actinobacteria and proteobacteria (**Figure 3C**). However, a few DciA proteins
144 recovered homologs only within their own phylum. Specifically, most actinobacterial (10/12) and
145 a few proteobacterial (10/30) DciA proteins only recovered homologs within their respective
146 phylum (**Figure 1C, 3C**). In addition, the DciA proteins from nitrospirae, gemmatimonadetes,
147 fibrobacteres, chlorobi, chlamydiae, and bacteroidetes identified homologs from actinobacteria
148 and not proteobacteria, suggesting that these DciA proteins are closer evolutionarily to
149 actinobacterial DciA homologs than proteobacterial DciA. In contrast, acidobacteria,
150 cyanobacteria, and thermodesulfobacteria recovered homologs from proteobacteria and not
151 actinobacteria. The evolution of DciA seems to mirror the phylogenetic distance of these species;
152 gemmatimonadetes, chlamydiae, and bacteroidetes are more closely evolutionary related to
153 actinobacteria than proteobacteria based on the 16S rRNA gene (28, 39). Together these data
154 suggest that the DciA homologs in actinobacteria and proteobacteria likely represent the most
155 evolutionarily divergent repertoire of DciA homologs. This analysis also indicates that DciA
156 proteins from distinct phyla carry lineage-specific signatures, likely co-evolving with other phylum-
157 specific protein families involved in DNA replication.
158

159 **DciA domain architecture varies in a lineage-specific manner**

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

171 In Group 1 DciA proteins, the DUF721 spans at least 70% of the protein sequence, with
172 ≤25 amino acids on either side of the DUF721 (**Figures 4, 5A, S2**; **Table 1**, See Group 1 example:
173 *R. rickettsii* DciA). Group 1 DciA proteins are present in acidobacteria, bacteroidetes, chlamydiae,
174 chlorobi, gemmatimonadetes, elusimicrobia, fibrobacteres, fusobacteria, thermotogae,
175 planctomycetes, spirochaetes, and proteobacteria (**Table 1**). The DUF721 domain spanning the
176 entire protein in this group suggests that the DUF721 is likely sufficient for DciA function in these
177 bacteria.

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

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

193 Group 3 DciA homologs have ≤25 amino acids N-terminal to the DUF721 and a C-terminal
194 extension >25 amino acids long (**Figures 4, 5A, S2**; **Table 1**, See Group 3 example: *V. cholerae*
195 DciA). Group 3 DciA homologs are present in acidobacteria, spirochaetes, deferribacteres,
196 cyanobacteria, dictyoglomi, nitrospirae, thermodesulfobacteria, and proteobacteria. The DciA
197 homologs in *T. palladium*, *Nitrospirae moscoviensis,* and *V. parahemolyticus* are the only Group
198 3 proteins that are predicted to be intrinsically disordered in the region C-terminal to the DUF721
199 domain using the MobiDBLite predictor **(Figure 4, left**; **Table 1**). However, the C-terminal
200 sequence of *V. cholerae* DciA was predicted to be intrinsically disordered in prior studies using
201 PONDR and IUPred2A software, which was further confirmed by small-angle X-ray scattering
202 (32). When we analyzed the *V. cholerae* sequence using the JRONN disorder prediction track,
203 we were able to identify an intrinsically disordered region C-terminal to the DUF721 **(Figure 4)**.
204 Therefore, it is possible that other Group 3 proteins could have intrinsically disordered domains
205 that are not predicted by the tools we are using. The C-terminal intrinsically disordered sequence
206 of *V. cholerae* DciA is required for its interaction with DnaB as well as for the enhancement of the
207 association between *V. cholerae* DnaB and ssDNA *in vitro* (32). The absence of the C-terminal
208 extension in Groups 1 and 2 DciA proteins further supports evolutionary divergence in DciA

209 protein sequence and structure, which likely impacts specialized lineage-specific protein function
210 or mechanism of interaction with the replicative helicase. The one actinobacterial DciA protein
211 from our query set that falls into Group 3 is encoded by *S. seoulensis*. This protein encodes a
212 YspA domain (Pfam: YAcAr/PF10686) C-terminal to the DUF721 (**Figure 4**, **5A, S2; Table 1**).
213 YspA domains within proteins typically have fusions to domains that process nucleotide-derived
214 ligands such as ADP-ribose and may function as sensors of these ligands and nucleic acids (41).
215 In Group 4 DciA proteins, the DUF721 domain falls > 25 amino acids away from both
216 termini **(Figure 4, 5A, S2; Table 1,** See Group 4 example: *C. mirabilis* DciA). In our query set,
217 Group 4 consists of DciA proteins in proteobacteria, actinobacteria, and synergistetes. The Group
218 4 DciA homologs in *F. fastidiosum, S. coelicor, and S. avermitilis* have regions of disorder both
219 N-terminal and C-terminal to the DUF721 **(Figure 4, left)**.
220 Analysis of how the DciA groups are distributed in different bacterial phyla for our starting
221 set of DciA homologs (**Figure 4**) reveals that 75% of homologs we queried from actinobacteria
222 fall into Group 2, and 66% of proteobacterial DciA proteins from our starting set fall into Group 3.
223 **(Figure 5A**; **Table 1)**. These two phyla had the most representatives in our set of query proteins,
224 and they form distinct clusters on the phylogenetic tree **(Figure 4, right)**. Further examining the
225 characteristics of each group revealed that 75% of DciA proteins from gram-positive bacteria
226 included in the query fall into Group 2, and 54% of DciA starting points from gram-negative
227 bacteria fall into Group 3 **(Figure 5B)**. By contrast, no Group 1 DciA proteins are from gram-
228 positive bacterial queries, and Group 1 makes up 37% of gram-negative bacterial queries **(Figure**
229 **5B)**. In addition to classifying each DciA protein from our query set into one of four groups based
230 on the position of the DUF721, we found that the alphaproteobacteria DciA proteins in the
231 Rickettsiales and Hyphomicrobiales orders harbored an insertion within the DUF721 that is not
232 present in any other phyla **(Figure 4, right; Figure S2)**. This could indicate further functional
233 divergence of the DciA homologs in these orders.

234 Overall, these analyses reveal the diversity of domain architectures amongst our selected
235 66 DciA homologs. A more comprehensive analysis of all DciA homologs would be required to
236 fully understand the distribution of diverse domain architectures in different phyla. Nonetheless,
237 these data demonstrate that DciA homologs have diverged significantly in sequence structure
238 based on the position and sequence of the DUF721, which may impact their function and
239 interaction with the DNA replication machinery.
240

241 **The DciA domain proximity network**

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

259 architectures identified within the bacterial DciA homologs have been summarized in the form of
260 a network of domain architectures reconciling all DciA homologs, with domains as nodes and co-
261 occurrence within proteins as edges, and their co-occurrences have been further quantified
262 (**Figure 6**). The association of the DUF721 and DciA proteins with other functional domains could
263 shed light on the activities for DciA in these bacteria.
264
265
266 **DISCUSSION**
267 The recent discovery of DciA as a predicted helicase operator in bacteria (28, 31) has begun to
268 shed light on a long-standing open question of how the majority of bacteria facilitate helicase
269 activity during DNA replication in the absence of the ATPase helicase loaders expressed by *E.*
270 *coli* and *B. subtilis*. The wide distribution of DciA in diverse bacterial phyla indicates that these
271 proteins likely represent the predominant paradigm for helicase operation in bacteria, despite not
272 being conserved in *E. coli* and *B. subtilis,* the organisms typically used as a model for bacterial
273 replication. DciA proteins are defined by the presence of the DUF721 domain and prior
274 phylogenetic analysis indicates that *dnaC* and *dnaI* homologs were acquired through evolution at
275 the expense of *dciA* (named for *dna[CI]* antecedent) (28), suggesting that DciA and DnaC/DnaI
276 perform a common function. In addition, it has been shown that DciA interacts with the replicative
277 helicase and is required for DNA replication and viability in the limited organisms it has been
278 studied in (28, 31, 33). However, the mechanism by which DciA mediates replication initiation is
279 still unknown. Our study has revealed immense diversity in DciA proteins, where there is low
280 sequence similarity between homologs in different phyla **(Figures 1–3)**, there are at least 4
281 distinct classes based on the positioning of the DUF721 domain in the protein **(Figures 4–5)**,
282 there exist lineage-specific insertion sequences in the DUF721 domain of some proteobacterial
283 species **(Figures 4, S2)**, and some DciA proteins have evolved as fusions to other functional

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

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

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

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

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

330 A lot of unknowns still remain regarding DciA proteins and bacterial DNA replication. The
331 computational evolutionary analysis described herein highlights the complexities and diversity
332 that have evolved in the fundamental process of DNA replication, where no single species of
333 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

384 **Figure 3. Retrieving DciA homologs using the extended query set of 66 DciA proteins. A,**
385 **B. Lineages of query and homologous DciA proteins.** *See Figure 1* for details. The three phyla
386 excluded in the query searches were Deinococcus-Thermus, Chrysiogenetes, and Firmicutes. **C.**
387 **Phyletic spread of the DciA homologs by query.** *See Figure 1* for details. Deinococcus-
388 Thermus and Firmicutes were both recovered in the resulting set of homologs. Our queries did
389 not recover the DciA homologs in chrysiogenetes, suggesting that the homologs in this phylum
390 are the most divergent in sequence from the query sequences.
391
392 **Figure 4. Characterizing the full list (66) of DciA query proteins.**
393 **The domain architectures and disorder predictions (left panel) of the 66 DciA query**
394 **proteins are overlaid with the multiple sequence alignment (right), and phylogenetic tree**
395 **(middle).** Each DciA protein is marked with the kingdom (B, bacteria), phylum (first 6 letters),
396 Genus, and species (represented as '*Gspecies*'). The Pfam and MobiDB annotations for each
397 domain prediction are shown in the legend (top). The colors in the multiple sequence alignment
398 depiction correspond to different amino acids (bottom legend). The data was generated and
399 visualized using the MolEvolvR web application. In addition, JronnWS (60) disorder predictions
400 were performed within Jalview (58) (not shown here), where other DciA proteins such as *V.*
401 *cholerae* show disorder regions.
402
403 **Figure 5. DciA groups and their distribution within our query sequences.**
404 **A. Example domain architectures of each of the 4 groups of DciA homologs.** Group 1 DciA
405 homologs have ≤25 aa on either side of the DUF721 (top, blue), Group 2 homologs have ≤25aa
406 C-terminal to the DUF721 (second, pink), Group 3 DciA homologs have ≤25aa amino acids N-
407 terminal to the DUF721 (third, orange), and Group 4 DciA homologs have >25aa both N- and C-
408 terminal to the DUF721 (bottom, teal). Graphics created using BioRender.com. **B. Distribution**

409 **of groups within Gram-positive and Gram-negative DciA query proteins.** Pie charts
410 comparing the number of gram-positive and gram-negative bacteria that have DciA homologs in
411 one of the 4 groups. Gram-positive (left), gram-negative (right). Percentages rounded to the
412 nearest whole number. Details of group and Gram stain assignments of each DciA homolog are
413 found in **Table 1**.
414
415 **Figure 6. DciA partners. A. Domain proximity network.** The network visualizes co-occurring
416 domains within all bacterial DciA homologs generated with our 66 starting points (**Figure 1**; **Table**
417 **1**). The nodes and edges correspond to domains and co-occurrence of domains within a protein;
418 the size corresponds to the frequency of occurrence with a minimum scaling factor. The full data
419 can be accessed at https://github.com/JRaviLab/dcia_evolution. **B. Frequencies of co-**
420 **occurring domains in DciA homologs.** Upset plot of the DciA homologs are shown. Blue
421 histogram: Distribution of the predominant domains. Dots and connections: Combinations in
422 which these domains come together in DciA domain architectures. Red histogram: Frequency of
423 occurrences of domain architectures. Of these only the DciA containing domain architectures
424 were used for alignments and phylogenetic trees.
425
426
427 **SUPPLEMENTARY FIGURES**
428 **Figure S1. Mycobacterium tuberculosis and Pseudomonas aeruginosa DciA proteins only**
429 **identify DciA homologs in their respective phylum.** Heatmap description as in Figure 1.
430
431 **Figure S2. Multiple sequence alignment of select DciA proteins.** Alignment of the 66 DciA
432 proteins used in the full query set. Numbering of residues across the top of the alignment is based
433 on the annotation *M. tuberculosis* DciA. Based on the numbering of *M. tuberculosis* residues, The

434 DUF721 domain falls between 75–162 amino acids (red box), the conserved tryptophan residue
435 in the DANL domain is at position 133 (red star), and the insertion present in some
436 alphaproteobacterial DciA proteins occurs after position 118 on the alignment. The multiple
437 sequence alignment was generated with Kalign (57, 61) and visualized using Jalview (58); (color
438 scheme: Clustalx).
439
440 **Figure S3. Full list representative homolog characterization (with DciA)**
441 The domain architectures, multiple sequence alignment, and phylogenetic tree were generated
442 using representative DciA homologs (one per domain architecture per lineage). See Figure 4 for
443 details.

444 **TABLES**

445 **Table 1. DciA query proteins used to identify homologs across the bacterial kingdoms.**
446 Protein, domain architecture, group, and lineage-related metadata for each of the 66 diverse
447 starting points of DciA proteins across the bacterial kingdom are shown in this table. The full
448 homolog data, analyses, and figures can be found here:
449 https://github.com/jravilab/dcia_evolution.
450
451

452 **ACKNOWLEDGEMENTS**
453 We are very grateful to the Midwest Microbial Pathogenesis Conference (MMPC) 2021 organizers
454 for providing HCB a travel award and opportunity to present the DciA story, and for providing an
455 interactive venue for JR and JTB to start this collaboration with CLS and HCB.
456

457 **FUNDING**

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

462

463 **DATA AVAILABILITY AND REUSE**

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

467

468 **REFERENCES**

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

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

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

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

478 5. Fuller RS, Funnell BE, Kornberg A. 1984. The dnaA protein complex with the E. coli
479 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.

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

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

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

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

510 20. Nakayama N, Arai N, Kaziro Y, Arai K. 1984. Structural and functional studies of the dnaB
511 protein using limited proteolysis. Characterization of domains for DNA-dependent ATP
512 hydrolysis and for protein association in the primosome. The Journal of biological chemistry
513 259:88–96.

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

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

519 23. Arias-Palomo E, Puri N, O'Shea Murray VL, Yan Q, Berger JM. 2019. Physical Basis for the
520 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.

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

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

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

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

554 35. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, Thomson NR, Iqbal Z.
555 2021. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA
556 sequences. PLoS biology 19:e3001421–e3001421.

557 36. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G,
558 Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome
559 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.

624 2021. MobiDB: intrinsically disordered proteins in 2021. Nucleic acids research 49:D361–
625 D367.

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

629 58. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—
630 a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–
631 1191.

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

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

636 61. Lassmann T. 2020. Kalign 3: multiple sequence alignment of large datasets. Bioinformatics
637 36:1928–1929.

638

Figure 1

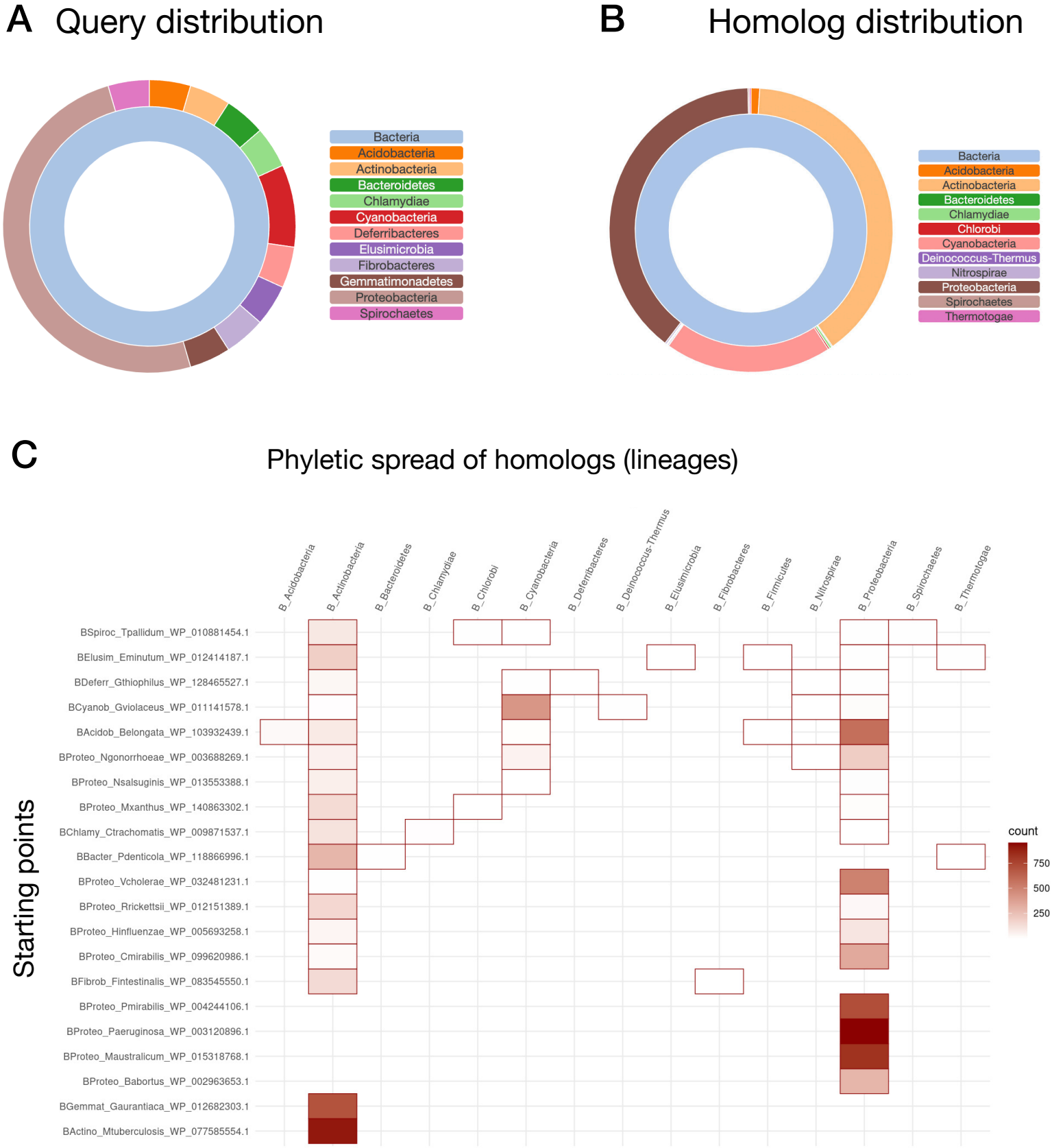

A
Query distribution
Bacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Chlamydiae
Cyanobacteria
Deferribacteres
Elusimicrobia
Fibrobacteres
Gemmatimonadetes
Proteobacteria
Spirochaetes
B
Homolog distribution
Bacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Chlamydiae
Chlorobi
Cyanobacteria
Deinococcus-Thermus
Nitrospirae
Proteobacteria
Spirochaetes
Thermotogae
C
Phyletic spread of homologs (lineages)
B_Acidobacteria
B_Actinobacteria
B_Bacteroidetes
B_Chlamydiae
B_Chlorobi
B_Cyanobacteria
B_Deferribacteres
B_Deinococcus-Thermus
B_Elusimicrobia
B_Fibrobacteres
B_Firmicutes
B_Nitrospirae
B_Proteobacteria
B_Spirochaetes
B_Thermotogae
Starting points
BSpiroc_Tpallidum_WP_010881454.1
BElusim_Eminutum_WP_012414187.1
BDeferr_Gthiophilus_WP_128465527.1
BCyanob_Gviolaceus_WP_011141578.1
BAcidob_Belongata_WP_103932439.1
BProteo_Ngonorrhoeae_WP_003688269.1
BProteo_Nsalsuginis_WP_013553388.1
BProteo_Mxanthus_WP_140863302.1
BChlamy_Ctrachomatis_WP_009871537.1
BBacter_Pdenticola_WP_118866996.1
BProteo_Vcholerae_WP_032481231.1
BProteo_Rrickettsii_WP_012151389.1
BProteo_Hinfluenzae_WP_005693258.1
BProteo_Cmirabilis_WP_099620986.1
BFibrob_Fintestinalis_WP_083545550.1
BProteo_Pmirabilis_WP_004244106.1
BProteo_Paeruginosa_WP_003120896.1
BProteo_Maustralicum_WP_015318768.1
BProteo_Babortus_WP_002963653.1
BGemmat_Gaurantiaca_WP_012682303.1
BActino_Mtuberculosis_WP_077585554.1
count
750
500
250

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

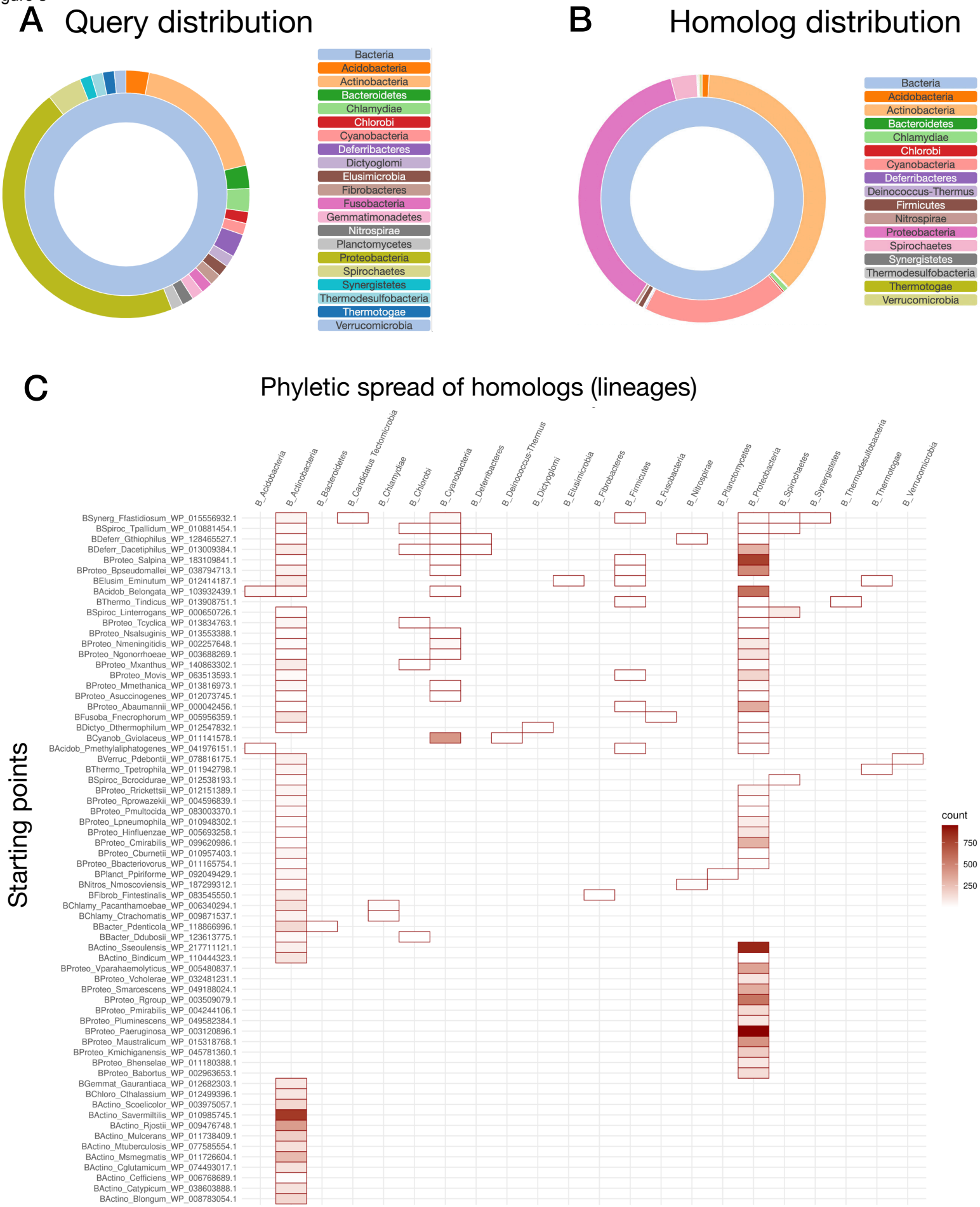

A
Query distribution
Bacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Chlamydiae
Chlorobi
Cyanobacteria
Deferribacteres
Dictyoglomi
Elusimicrobia
Fibrobacteres
Fusobacteria
Gemmatimonadetes
Nitrospirae
Planctomycetes
Proteobacteria
Spirochaetes
Synergistetes
Thermodesulfobacteria
Thermotogae
Verrucomicrobia
B
Homolog distribution
Bacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Chlamydiae
Chlorobi
Cyanobacteria
Deferribacteres
Deinococcus-Thermus
Firmicutes
Nitrospirae
Proteobacteria
Spirochaetes
Synergistetes
Thermodesulfobacteria
Thermotogae
Verrucomicrobia
C
Phyletic spread of homologs (lineages)
B_Acidobacteria
B_Actinobacteria
B_Bacteroidetes
B_Candidatus Tectomicrobia
B_Chlamydiae
B_Chlorobi
B_Cyanobacteria
B_Deferribacteres
B_Deinococcus-Thermus
B_Dictyoglomi
B_Elusimicrobia
B_Fibrobacteres
B_Firmicutes
B_Fusobacteria
B_Nitrospirae
B_Planctomycetes
B_Proteobacteria
B_Spirochaetes
B_Synergistetes
B_Thermodesulfobacteria
B_Thermotogae
B_Verrucomicrobia
Starting points
BSynerg_Ffastidiosum_WP_015556932.1
BSpiroc_Tpallidum_WP_010881454.1
BDeferr_Gthiophilus_WP_128465527.1
BDeferr_Dacetiphilus_WP_013009384.1
BProteo_Salpina_WP_183109841.1
BProteo_Bpseudomallei_WP_038794713.1
BElusim_Eminutum_WP_012414187.1
BAcidob_Belongata_WP_103932439.1
BThermo_Tindicus_WP_013908751.1
BSpiroc_Linterrogans_WP_000650726.1
BProteo_Tcyclica_WP_013834763.1
BProteo_Nsalsuginis_WP_013553388.1
BProteo_Nmeningitidis_WP_002257648.1
BProteo_Ngonorrhoeae_WP_003688269.1
BProteo_Mxanthus_WP_140863302.1
BProteo_Movis_WP_063513593.1
BProteo_Mmethanica_WP_013816973.1
BProteo_Asuccinogenes_WP_012073745.1
BProteo_Abaumannii_WP_000042456.1
BFusoba_Fnecrophorum_WP_005956359.1
BDictyo_Dthermophilum_WP_012547832.1
BCyanob_Gviolaceus_WP_011141578.1
BAcidob_Pmethylaliphatogenes_WP_041976151.1
BVerruc_Pdebontii_WP_078816175.1
BThermo_Tpetrophila_WP_011942798.1
BSpiroc_Bcrocidurae_WP_012538193.1
BProteo_Rrickettsii_WP_012151389.1
BProteo_Rprowazekii_WP_004596839.1
BProteo_Pmultocida_WP_083003370.1
BProteo_Lpneumophila_WP_010948302.1
BProteo_Hinfluenzae_WP_005693258.1
BProteo_Cmirabilis_WP_099620986.1
BProteo_Cburnetii_WP_010957403.1
BProteo_Bbacteriovorus_WP_011165754.1
BPlanct_Ppiriforme_WP_092049429.1
BNitros_Nmoscoviensis_WP_187299312.1
BFibrob_Fintestinalis_WP_083545550.1
BChlamy_Pacanthamoebae_WP_006340294.1
BChlamy_Ctrachomatis_WP_009871537.1
BBacter_Pdenticola_WP_118866996.1
BBacter_Ddubosii_WP_123613775.1
BActino_Sseoulensis_WP_217711121.1
BActino_Bindicum_WP_110444323.1
BProteo_Vparahaemolyticus_WP_005480837.1
BProteo_Vcholerae_WP_032481231.1
BProteo_Smarcescens_WP_049188024.1
BProteo_Rgroup_WP_003509079.1
BProteo_Pmirabilis_WP_004244106.1
BProteo_Pluminescens_WP_049582384.1
BProteo_Paeruginosa_WP_003120896.1
BProteo_Maustralicum_WP_015318768.1
BProteo_Kmichiganensis_WP_045781360.1
BProteo_Bhenselae_WP_011180388.1
BProteo_Babortus_WP_002963653.1
BGemmat_Gaurantiaca_WP_012682303.1
BChloro_Cthalassium_WP_012499396.1
BActino_Scoelicolor_WP_003975057.1
BActino_Savermiltilis_WP_010985745.1
BActino_Rjostii_WP_009476748.1
BActino_Mulcerans_WP_011738409.1
BActino_Mtuberculosis_WP_077585554.1
BActino_Msmegmatis_WP_011726604.1
BActino_Cglutamicum_WP_074493017.1
BActino_Cefficiens_WP_006768689.1
BActino_Catypicum_WP_038603888.1
BActino_Blongum_WP_008783054.1
count
750
500
250

Figure 4

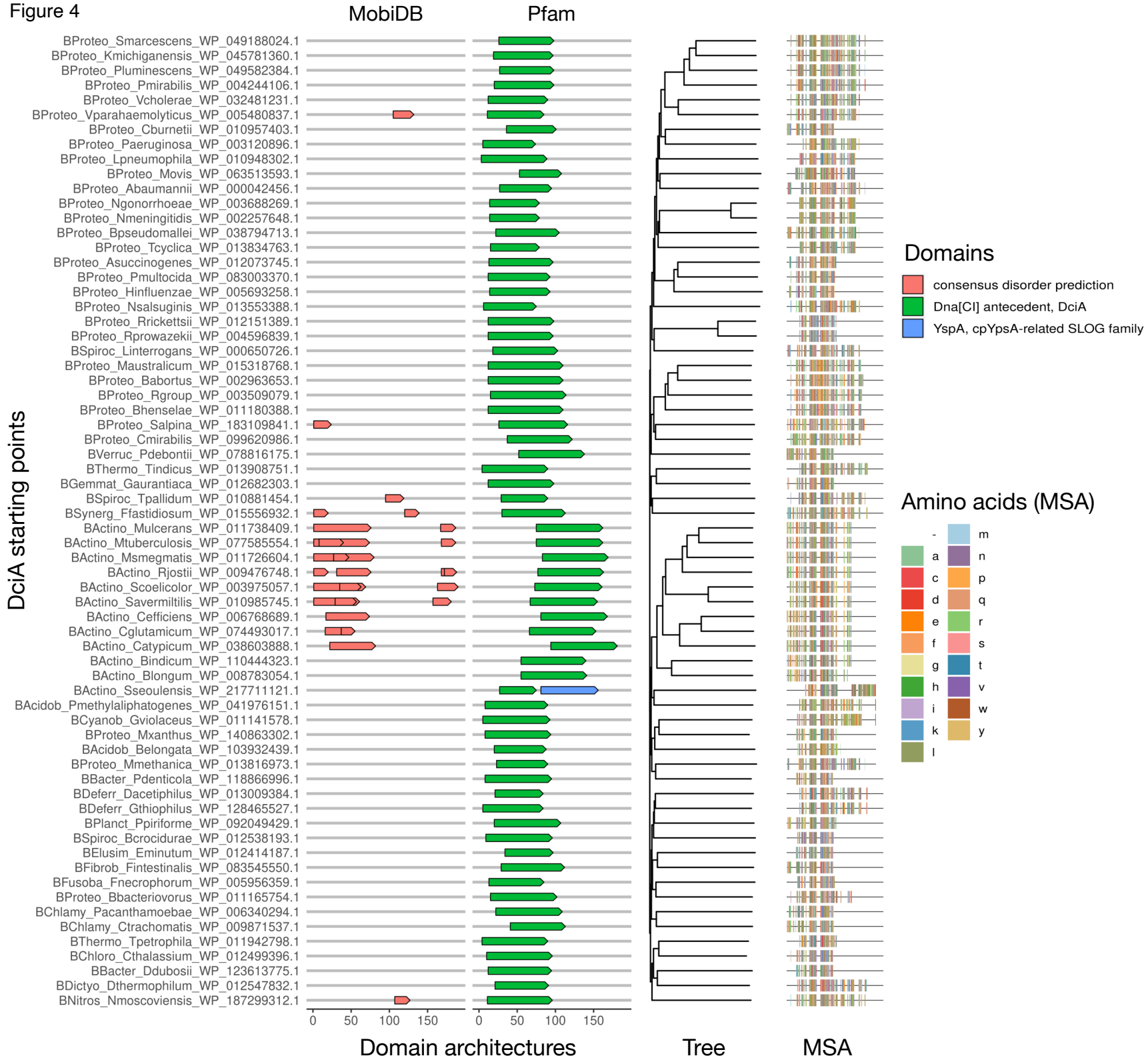

MobiDB
Pfam
BProteo_Smarcescens_WP_049188024.1
BProteo_Kmichiganensis_WP_045781360.1
BProteo_Pluminescens_WP_049582384.1
BProteo_Pmirabilis_WP_004244106.1
BProteo_Vcholerae_WP_032481231.1
BProteo_Vparahaemolyticus_WP_005480837.1
BProteo_Cburnetii_WP_010957403.1
BProteo_Paeruginosa_WP_003120896.1
BProteo_Lpneumophila_WP_010948302.1
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

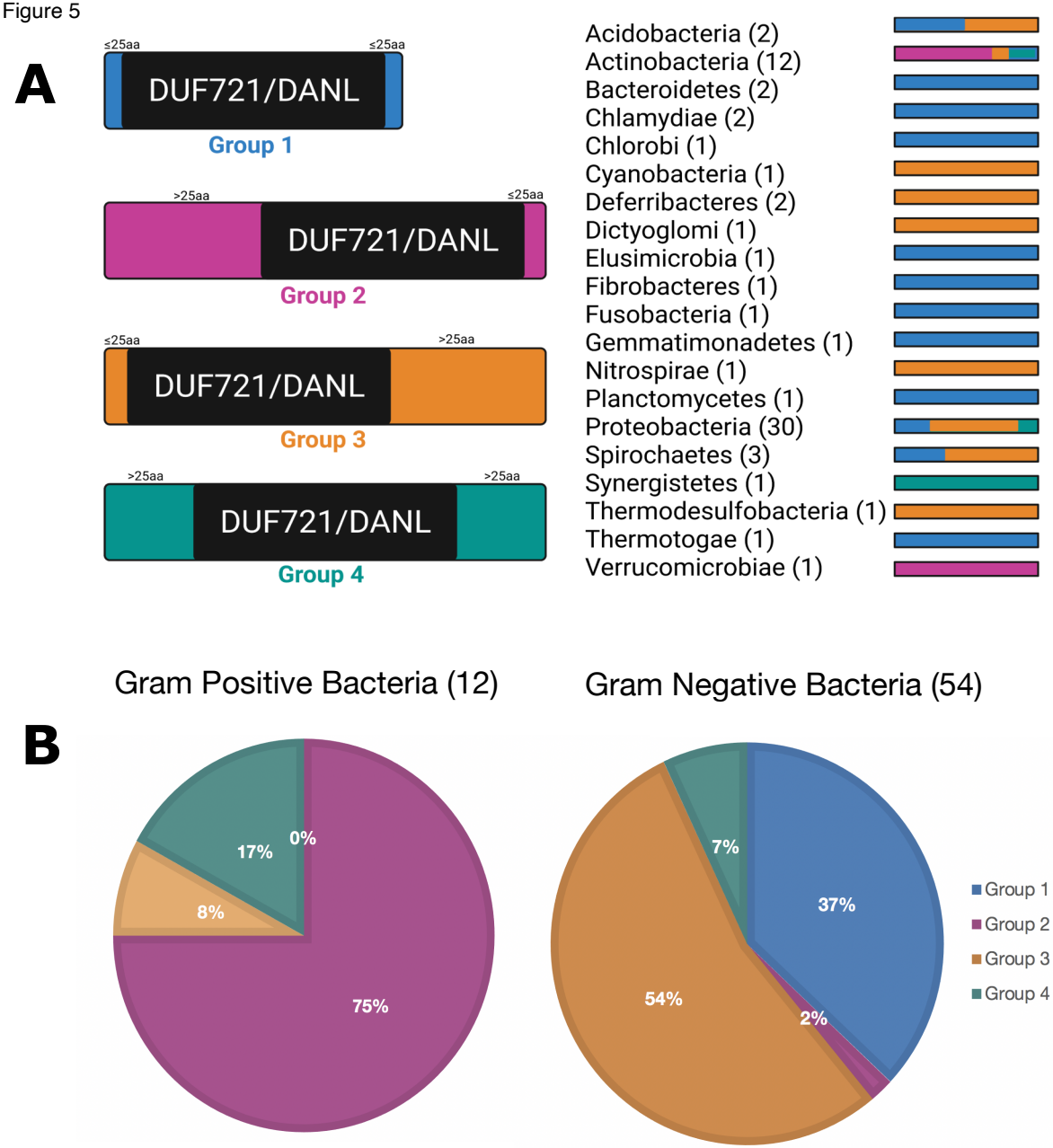
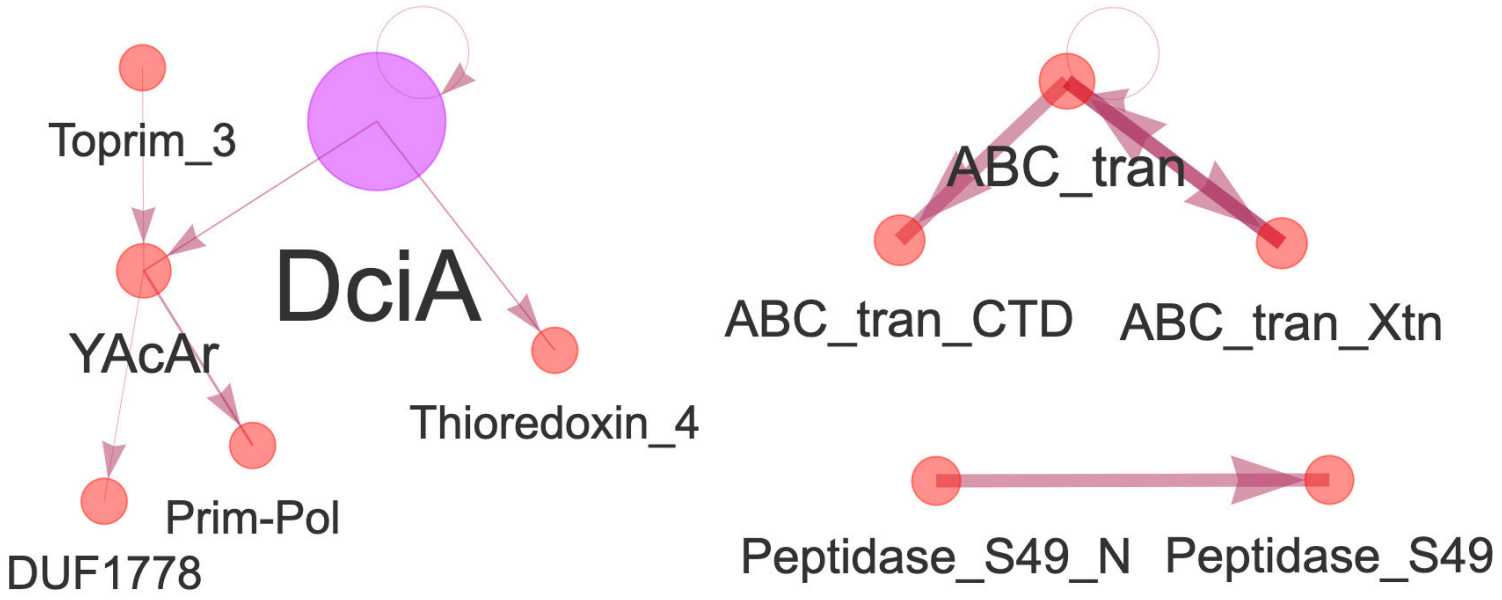

A
≤25aa
≤25aa
DUF721/DANL
Group 1
>25aa
≤25aa
DUF721/DANL
Group 2
≤25aa
>25aa
DUF721/DANL
Group 3
>25aa
>25aa
DUF721/DANL
Group 4
Acidobacteria (2)
Actinobacteria (12)
Bacteroidetes (2)
Chlamydiae (2)
Chlorobi (1)
Cyanobacteria (1)
Deferribacteres (2)
Dictyoglomi (1)
Elusimicrobia (1)
Fibrobacteres (1)
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

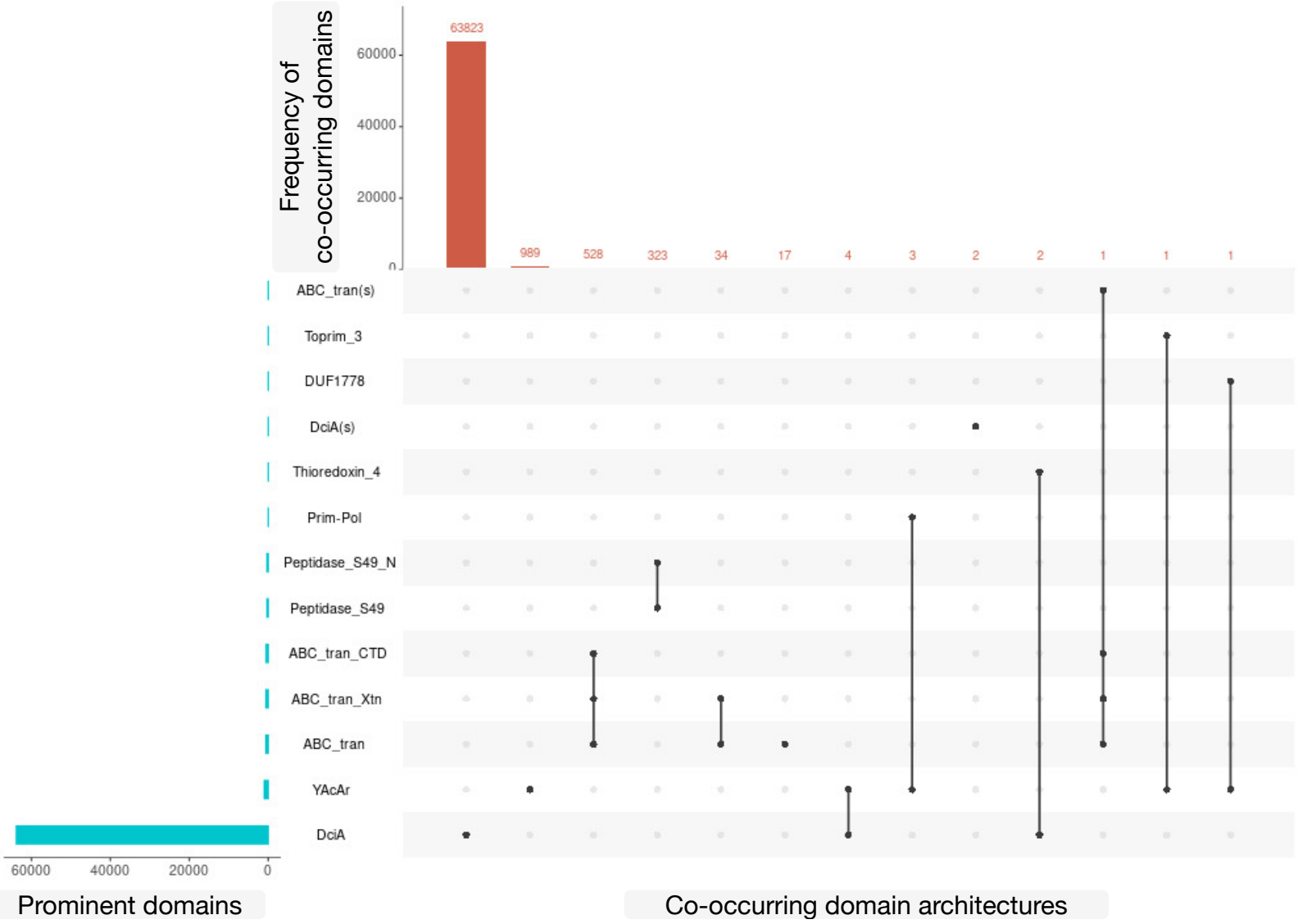


Table 1

Table 2. Hoft query proteins used to identify homology across the bacterial kingdom

Name	AccNum	Species	TaxID	Lineage	DomSeq16.Phlam	DomSeq8.PfamDB.dat	Length	EGF7213 range (aa)	Hoft group	Gram stain
WPmma_Adacnum01_MP_0100435851	MP_0100435851.1	Acinetobacter baumannii	470	Bacteria\Proteobacteria	Hoft	NA	354	27-90	Group 5	-
WPgiza_Leptorrigen_MP_0100507216.1	MP_0100507216.1	Leptotrichia rotoryrgen	171	Bacteria\Spirochaetes	Hoft	NA	354	10-103	Group 5	-
WPmma_Neisseriglob_MP_0101570368.1	MP_0101570368.1	Neisseria meningitidis	487	Bacteria\Proteobacteria	Hoft	NA	340	10-79	Group 5	-
WPmma_Bakcmma_MP_0129634551.1	MP_0129634551.1	Brucella abortus	234	Bacteria\Proteobacteria	Hoft	NA	175	12-110	Group 5	-
WPmma_Pseugcmma_MP_0101329890.1	MP_0101329890.1	Pseudomonas aeruginosa	28	Bacteria\Proteobacteria	Hoft	NA	111	7-74	Group 5	-
WPmma_Sgroup_MP_0101509276.1	MP_0101509276.1	Escherichia/Shigella/enterobacter group	197390	Bacteria\Proteobacteria	Hoft	NA	171	14-114	Group 5	-
WPmma_Nigcmcrbmma_MP_0104882630.1	MP_0104882630.1	Neisseria gonorrhoeae	485	Bacteria\Proteobacteria	Hoft	NA	340	9-88	Group 5	-
Bdctma_Streptolor_MP_0097958731	MP_0097958731.1	Streptomyces coelicolor	1885	Bacteria\Actinobacteria	Hoft	shorter region+thicker region+thicker region	392	72-241	Group 6	+
WPmma_Petrabfb_MP_0102450261.1	MP_0102450261.1	Petrotoga mobilis	93	Bacteria\Proteobacteria	Hoft	NA	172	8-106	Group 5	-
WPmma_Sgroumnb_MP_0097958391	MP_0097958391.1	Rickettsia prowazekii	783	Bacteria\Proteobacteria	Hoft	NA	308	4-98	Group 5	-
WPmma_Ureaehamycldyna_MP_0091883837.1	MP_0091883837.1	Vibrio parahaemolyticus	670	Bacteria\Proteobacteria	Hoft	shorter region	251	8-94	Group 5	-
WPmma_Hbcfcamma_MP_0095855284.1	MP_0095855284.1	Haemophilus influenzae	727	Bacteria\Proteobacteria	Hoft	NA	334	13-98	Group 5	-
WPmmbc_Pneumphmma_MP_0095855391.1	MP_0095855391.1	Pseudomonas mosselii sp.	83	Bacteria\Proteobacteria	Hoft	NA	95	4-91	Group 5	-
WCfamy_Faecelbacmma_MP_0083402964.1	MP_0083402964.1	Faecalibacterium prausnitzii	853532	Bacteria\Clostridia	Hoft	NA	312	15-108	Group 1	variable (fdk), inverted sa +
Bdctma_Cellcaem_MP_0087838881.1	MP_0087838881.1	Corynebacterium efficiens	152790	Bacteria\Actinobacteria	Hoft	shorter region+thicker region	255	81-168	Group 2	+
Bdctma_Bmcgam_MP_0087839341.1	MP_0087839341.1	Bifidobacterium longum	216816	Bacteria\Actinobacteria	Hoft	NA	256	52-141	Group 2	+
Bdctma_Rpmb_MP_0094317481.1	MP_0094317481.1	Rhodococcus jostii	2027	Bacteria\Actinobacteria	Hoft	shorter region+thicker region+thicker region+thicker region	246	70-162	Group 2	+
WCfamy_Cructomb_MP_0098735337.1	MP_0098735337.1	Clostridia rectibacteria	853	Bacteria\Clostridia	Hoft	NA	126	20-112	Group 1	variable (fdk), inverted sa +
Bdctma_Tpalabcm_MP_0100814951.1	MP_0100814951.1	Treponema pallidum	160	Bacteria\Spirochaetes	Hoft	shorter region+thicker region	245	7-93	Group 5	-
WPmma_cpneumphfdu_MP_0109840932.1	MP_0109840932.1	Legionella pneumophila	446	Bacteria\Proteobacteria	Hoft	NA	340	9-89	Group 5	-
WPmma_Cturmcl_MP_0120975651.1	MP_0120975651.1	Candida tenuis	77	Bacteria\Proteobacteria	Hoft	NA	114	21-109	Group 5	-
Bdctma_Saureb0fdu_MP_0129957581.1	MP_0129957581.1	Staphylococcus aureus0fdu	1883	Bacteria\Actinobacteria	Hoft	shorter region+thicker region+thicker region	261	50-155	Group 6	+
Bcjamb_Ovisbacem_MP_0111435776.1	MP_0111435776.1	Ehrlichia/ter cirdacem	55372	Bacteria\Lyme/ancter	Hoft	NA	280	5-93	Group 5	-
WPmma_Bbacterimma_MP_0111435754.1	MP_0111435754.1	Bdellovibrio bacteriovorus	958	Bacteria\Proteobacteria	Hoft	NA	151	12-102	Group 5	-
WPmma_Bbcmcbm_MP_0111435381.1	MP_0111435381.1	Burkholdia bacteriovorus	36512	Bacteria\Proteobacteria	Hoft	NA	246	12-142	Group 5	-
Bdctma_Pneumgcmm_MP_0117256041	MP_0117256041.1	Mycobacterium smegmatis	1772	Bacteria\Actinobacteria	Hoft	shorter region+thicker region	264	42-168	Group 2	+
Bdctma_Pdcrcam_MP_0117384281.1	MP_0117384281.1	Mycobacterium abscessus	1859	Bacteria\Actinobacteria	Hoft	shorter region+thicker region	267	75-162	Group 2	+
WCfamy_Tpetrgbfdu_MP_0129627981.1	MP_0129627981.1	Thermotoga petrophila	93624	Bacteria\Thermotogae	Hoft	NA	221	9-92	Group 5	-
WPmma_Acnctmcgmm_MP_0120757381.1	MP_0120757381.1	Actinobacillus succinogenes	83963	Bacteria\Proteobacteria	Hoft	NA	132	12-97	Group 5	-
WPmma_Rcidamb2_MP_0123151881.1	MP_0123151881.1	Ralstonia solanacr	760	Bacteria\Proteobacteria	Hoft	NA	207	4-96	Group 5	-
WBmm_Entmmm_MP_0121614587.1	MP_0121614587.1	Enterococcium mimum	42560	Bacteria\Enterococcidia	Hoft	NA	96	9-87	Group 5	-
BCfmm_Cfdatamm_MP_0124995961.1	MP_0124995961.1	Chlorobaculum tepidum	160734	Bacteria\Chlorobi	Hoft	NA	97	9-94	Group 5	-
WCfmm_Bcmcbcmm_MP_0123836351.1	MP_0123836351.1	Borrelia burgdorferi	29523	Bacteria\Spirochaetes	Hoft	NA	95	8-85	Group 5	-
BCfmm_Dtbcmgcbfdu_MP_0125297651.1	MP_0125297651.1	Desulfotomaculum gingiae	33	Bacteria\Firmicutes	Hoft	NA	80	4-92	Group 5	-
WCfmmm_Gacmmbm_MP_0123825251.1	MP_0123825251.1	Gammatimonas aurantiaca	171680	Bacteria\Gammatimonadetes	Hoft	NA	215	12-98	Group 5	-
BCfafm_Zmcbpfbm_MP_0130095861.1	MP_0130095861.1	Zoobacteria anaerobica	126600	Bacteria\Saltwaterbacteria	Hoft	NA	147	8-84	Group 5	-
WPmma_Sbcdmgcm_MP_0130915861.1	MP_0130915861.1	Metallosphaera sedugcmm	280261	Bacteria\Proteobacteria	Hoft	NA	151	6-75	Group 5	-
WPmma_Pbvcafmm_MP_0130165751	MP_0130165751.1	Phaliphromas vacbacmm	411	Bacteria\Proteobacteria	Hoft	NA	154	14-92	Group 5	-
WPmma_Tcycfm_MP_0130345761.1	MP_0130345761.1	Tetrotomopria cycloa	147588	Bacteria\Proteobacteria	Hoft	NA	245	6-89	Group 5	-
BCfamm_Tmfdm_MP_0130687611.1	MP_0130687611.1	Thermodesulfobacter indicus	175480	Bacteria\Thermodesulfobacteria	Hoft	NA	259	3-91	Group 5	-
WPmma_Pbacmbcm_MP_0131525764.1	MP_0131525764.1	Phaeobacbacter gseakdcum	48240	Bacteria\Proteobacteria	Hoft	NA	182	17-53	Group 5	-
WCgmy_Plmbfdmm_MP_0135595011	MP_0135595011.1	Parlibacterium hamilbcum	651852	Bacteria\Spirochaetes	Hoft	shorter region+thicker region	176	29-112	Group 6	-
WPmma_Ucbcbmm_MP_0135812211.1	MP_0135812211.1	Ulvis chelaeae	666	Bacteria\Proteobacteria	Hoft	NA	257	8-94	Group 5	-
Bdctma_Cmgcbcm_MP_0136015881.1	MP_0136015881.1	Corynebacterium amycicum	191610	Bacteria\Actinobacteria	Hoft	shorter region	256	92-241	Group 2	+
WPmma_Rmmdcmmbm_MP_0136792711.1	MP_0136792711.1	Rubrobacter xylanophilus	26692	Bacteria\Proteobacteria	Hoft	NA	158	22-105	Group 5	-
Bdctbb_Pmrbfgbfgmmgmm_MP_0137524211.1	MP_0137524211.1	Pyrinomonas methylaliphatogenes	651456	Bacteria\Actinobacteria	Hoft	NA	156	1-93	Group 5	-
WPmma_Ntrcbbgmmm_MP_0137813651.1	MP_0137813651.1	Rickettsia rickettsiaemia	1134587	Bacteria\Proteobacteria	Hoft	NA	171	8-98	Group 5	-
WPmma_Smcmcmmm_MP_0138188241.1	MP_0138188241.1	Serratia marcescens	615	Bacteria\Proteobacteria	Hoft	NA	171	8-93	Group 5	-
WPmma_Pmmbcmmm_MP_0138921561.1	MP_0138921561.1	Photorhabdus luminescens	29488	Bacteria\Proteobacteria	Hoft	NA	172	22-100	Group 5	-
WPmma_Pmm_MP_0139213591.1	MP_0139213591.1	Pectoralis sols	29632	Bacteria\Proteobacteria	Hoft	NA	172	12-108	Group 5	-
Bdctma_Cglamcmm_MP-0139595571	MP_0139595571.1	Corynebacterium glutamicum	1718	Bacteria\Actinobacteria	Hoft	shorter region+thicker region	178	92-153	Group 2	+
Bdctma_Mbcbcmmm_MP_0177685561.1	MP_0177685561.1	Mycobacterium tuberculosis	1773	Bacteria\Actinobacteria	Hoft	shorter region+thicker region+thicker region	267	75-162	Group 2	+
WPmmm_Pmbmm1_MP_0178842171.1	MP_0178842171.1	Pseudomonas mbendi	46667	Bacteria\Proteobacteria	Hoft	NA	152	72-138	Group 2	-
WPmma_Pmcdmcmm_MP_0180025751	MP_0180025751.1	Pantoea mdocola	747	Bacteria\Proteobacteria	Hoft	NA	161	4-93	Group 5	-
WPbmb_Fmmmbm_MP_0181695551	MP_0181695551.1	Fibrobacter intestinalis	28122	Bacteria\Fibrobacteria	Hoft	NA	115	25-112	Group 5	-
BPmcn_Rpcfbmm_MP_0020496261	MP_0020496261.1	Planctomycetacium pifcbcma	1576340	Bacteria\Planctomycetes	Hoft	NA	132	12-102	Group 5	- (but stripe)
WPmma_Cmtcmfm_MP_0092399861.1	MP_0092399861.1	Catabacter mtmfdu	29940	Bacteria\Proteobacteria	Hoft	NA	152	52-122	Group 5	-
Bdctbb_Rabtmgm_MP_1035518951	MP_1035518951.1	Rycinda atmgam	803512	Bacteria\Actinobacteria	Hoft	NA	104	1-99	Group 5	-
Bdctbm_Bmbcam_MP_1104443351	MP_1104443351.1	Bifidobacterium indicum	1695	Bacteria\Actinobacteria	Hoft	NA	165	55-140	Group 2	+
BBactm_Pdmtcbm_MP_1188049961.1	MP_1188049961.1	Prosteella dentricola	28129	Bacteria\Bacteroidetes	Hoft	NA	94	8-95	Group 5	-
BBactm_Dmbcm1_MP_1216127771.1	MP_1216127771.1	Dorsomatella dubosii	2132615	Bacteria\Bacteroidetes	Hoft	NA	91	6-91	Group 5	-
Bdctbc_Ddbtgbfm_MP_1186855727.1	MP_1186855727.1	Gardnerella vaginalis	106409	Bacteria\Saltwaterbacteria	Hoft	NA	98	7-93	Group 5	-
WPmma_Pbacmbm_MP_1028833021.1	MP_1028833021.1	Hyperoxeus saetifera	34	Bacteria\Proteobacteria	Hoft	NA	215	8-94	Group 5	-
BPmma_Sabpma_MP_1031298661.1	MP_1031298661.1	Sphingomonas alpina	653951	Bacteria\Proteobacteria	Hoft	shorter region+thicker region	264	20-118	Group 5	-
WBmm_Scmmcmmm_MP_1872995111.1	MP_1872995111.1	Streptococcus mmcmmm	42293	Bacteria\Strepcoa	Hoft	shorter region	155	12-96	Group 5	-
Bdctm_Sanmbcm_MP_1077112211.1	MP_1077112211.1	Streptomyces sanmbcm	70040	Bacteria\Actinobacteria	Hoft+Hoft	NA	295	1-75	Group 5	+