Divergent evolution of insecticidal protein families across bacteria and ferns

Haruka Endo[1]* and Hokto Kazama[1,2]

[1]RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[2]Graduate School of Arts and Sciences, The University of Tokyo, Meguro-ku, Tokyo 153-8902, Japan
*Correspondence: haruka.endo@riken.jp

Abstract

Insecticidal Cry proteins produced by the bacterial insect pathogen *Bacillus thuringiensis* possess conserved three domains (3D). Two other families of insecticidal proteins—Prb proteins from bacterial symbionts of entomopathogenic nematodes and a marine bacterial pathogen, and IPD113 proteins from ferns—are two-domain (2D) proteins structurally resembling domains I and II of Cry proteins. Despite the structural resemblance, the amino acid sequence identity among these families is only ~10%, supporting the prevailing view that they evolved independently through convergent evolution. However, the possibility of remote homology and ancient horizontal gene transfer (HGT) has remained largely unexplored. Here, we present evidence supporting divergent evolution of these protein families. We found that proteins containing Cry-like two domains are distributed far more broadly across bacteria, eukaryotes, and archaea than previously recognized. Phylogenetic analyses separated the structurally similar proteins into five clades including Cry, IPD113, Prb, and 2D/3D-mixed clades. Domain-level trees revealed intermixing within domain II, with nematocidal Cry proteins frequently clustering with IPD113 and 2D/3D-mixed clades, suggesting domain swapping events across these clades. Conserved amino acid motifs and domain core structures across clades further support a shared evolutionary origin. Phylogenetic and structural analyses suggest a bacterial origin of the fern IPD113 gene and its horizontal transfer via fungal symbionts, as well as multiple HGT events into diverse terrestrial and marine eukaryotes and archaea. This work redefines the evolutionary landscape of the insecticidal protein families, revealing their unexpectedly wide taxonomic distribution, possibly shaped through bacterial ecological networks and close biological interactions such as symbiosis.

Introduction

The bacterial pathogen *Bacillus thuringiensis* (Bt) produces insecticidal and nematocidal Cry proteins, which have been widely used in agriculture for pest control for a century. The Cry protein family is highly diverse, consisting of 53 subfamilies that share less than 45% amino acid identity with each other (Crickmore et al. 2021). These proteins kill their specific targets including lepidopteran caterpillars, mosquitoes, beetles, and nematodes (van Frankenhuyzen 2009). Cry proteins are α-pore-forming toxins that share three conserved domains in their active form (Fig. 1A) (Grochulski et al. 1995). Domain I is a pore-forming domain composed of seven α-helices. Domain II adopts a β-prism structure that recognizes target membrane proteins or glycolipids in the host gut and plays a major role in determining host specificity. Domain III is a β-sandwich structure thought to contribute to structural stability and host recognition, although its precise role remains unclear.

Interestingly, two other insecticidal protein families, Prb and IPD113, exhibit striking structural similarity to domains I and II of Cry proteins (Fig. 1A) (Wei et al. 2023; Prashar et al. 2023; Lee et al. 2015). Prb proteins (*Photorhabdus* insect-related toxin B component) are produced by bacterial symbionts of entomopathogenic nematodes, such as *Photorhabdus* and *Xenorhabdus*, as well as by the marine pathogen *Vibrio parahaemolyticus* (Waterfield et al. 2005; Duchaud et al. 2003; Lee et al. 2015). Prb proteins couple with another component Pra to exhibit toxicity toward lepidopteran caterpillars (*Photorhabdus* and *Xenorhabdus*) and shrimp (*Vibrio*). In contrast, IPD113 proteins are found in polypodiopsida ferns and are toxic to lepidopteran caterpillars as single-component toxins, without the need for additional protein partners (Wei et al. 2023). Despite their structural resemblance, Cry, Prb, and IPD113 proteins share only ~10% amino acid identity. Such low amino acid identity is often referred to as the "midnight zone", where most protein pairs are unrelated in both structure and function (Rost 1997). In this zone, it is generally difficult to determine whether proteins have originated from a common ancestor through divergent evolution, or independently acquired similar structures through convergent evolution, or whether the similarity is merely coincidental. Given the low sequence identity and the presence of introns in IPD113 genes, the original report on IPD113 favored the hypothesis that these proteins represent a case of convergent evolution (Wei et al. 2023). However, the possibility of remote homology and ancient horizontal gene transfer (HGT) has remained largely unexplored.

In this study, we aimed to examine whether the three insecticidal protein families (Cry, Prb, and IPD113) are evolutionarily linked. By aligning AlphaFold 3 (AF3)-predicted structures, we were able to accurately extract amino acid residues corresponding to the two domains even among proteins with very low amino acid identity. Although the extracted sequences remain highly divergent, pairwise alignment-based phylogenetic analysis developed for protein superfamilies enabled reliable tree

construction. The results indicate that Cry, Prb, and IPD113 proteins share a common evolutionary origin, and that their genes have been unexpectedly disseminated across diverse bacterial, archaeal, and eukaryotic lineages via HGTs, some of which may have been mediated by close ecological interactions such as symbiosis.

Results

Proteins containing Cry-like two domains are broadly distributed across taxa

To explore the possibility of remote homology, we performed BLASTp searches using IPD113 and Prb sequences as queries. Putative IPD113 homologues (E-value < 1e-5) were identified in ferns, mosses, bacteria, and animals, including frogs and newts (Fig. 1B and Table S1). These bacterial and animal proteins share approximately 30% amino acid identity with the query sequence (GenBank: UDN67319.1 from *Leptochilus wrightii*) (Fig. 1B), and their AF3-predicted structures resemble those of fern IPD113 proteins (Fig. 1C). A subsequent BLASTp search using a bacterial IPD113-like protein (GenBank: WP_119147942.1 from *Cohnella faecalis*) as the query revealed additional IPD113-like proteins in fungi, archaea, and protists, as well as in plants, bacteria, and animals (Fig. S1A and Table S1). Similarly, BLASTp searches with Prb sequences from *Photorhabdus akhurstii* and *Vibrio parahaemolyticus* (GenBank: AEQ33638.1 and 3X0U_A) identified putative homologues in both bacteria and insects, such as beetles and aphids (Fig. S1B and Table S1). Eukaryotic genes encoding these proteins often possess introns (Table S2), suggesting that they are not the result of bacterial contamination in sequencing but are integrated and functional components of the eukaryotic genome, similar to fern IPD113 proteins. Together, these findings suggest an unexpectedly broad taxonomic distribution of proteins containing Cry-like two domains.

Phylogenetic analysis identifies the 2D/3D-mixed clade

To reconstruct the phylogenetic relationships among proteins containing two Cry-like domains in the midnight zone, we predicted their structures using AlphaFold3 (Abramson et al. 2024), aligned them with FoldMason (Gilchrist, Mirdita, and Steinegger 2024), and extracted the amino acid residues corresponding to domains I and II (see Methods). We selected 557 non-redundant sequences with ≥95% identity that fully contain the two domains. These include 268 Cry proteins with quaternary rank 1, i.e., the canonical isoform denoted by the terminal "1" in Cry names (e.g., Cry1Aa1, Cry3Bb1), listed in the Bacterial Pesticidal Protein Resource Center (www.bpprc.org), as well as six sequences identified by BLASTp using Cry1Aa1 (GenBank: AAA22353) as a query.

Since the extracted sequences are highly divergent, conventional multiple sequence alignment (MSA)-based phylogenetic methods, such as maximum likelihood, were not applicable. Instead, we employed

99 the Graph Splitting (GS) method, a graph-based approach suitable for reconstructing superfamily-scale
100 phylogenetic trees (Matsui and Iwasaki 2020). The resulting GS tree grouped the proteins containing
101 the Cry-like two domains into five clades (Fig. 2A). Based on known protein families, we designated
102 three of these clades as Cry, IPD113, and Prb, and the remaining two as the 2D/3D-mixed and
103 *Photobacterium* clades (Fig. 2A). The Cry clade comprises only 3D Cry proteins and includes the
104 majority of known Cry proteins (Fig. 2B). In contrast, the 2D/3D-mixed clade contains both the
105 remaining 3D Cry proteins (Cry2, Cry11, Cry18, Cry63, and some members of Cry5 and Cry21) and
106 proteins from diverse non-Bt bacteria including Pseudomonadota, as well as from fungi and archaea
107 (Fig. 2A and Fig. S2A). While most non-Bt members of this clade are Cry-like two domain proteins,
108 some harbor one or two additional domains (Fig. 2B). The IPD113 clade includes members from non-
109 Bt bacteria, fungi, plants, animals including frogs and newts, archaea, and protists (Fig. 2B and Fig.
110 S2B). Some members of this clade possess additional domains beyond the canonical two. The Prb
111 clade includes 2D proteins from *Photorhabdus* and *Xenorhabdus* as well as from Pseudomonadota
112 (e.g., *Serratia*, *Yersinia*), along with 3D proteins from beetles and aphids (Fig. S2C). These insect-
113 derived 3D proteins have a Pra-like N-terminal domain followed by two Prb-like domains (Fig. S1C).
114 The Photobacterium clade contains only four 2D proteins, all from *Photobacterium* species, while
115 another *Photobacterium*-derived sequence belongs to the 2D/3D-mixed clade (Fig. 2B). These results
116 suggest an evolutionary link between 3D Cry proteins and the 2D proteins in the 2D/3D-mixed clade.
117 The phylogeny also indicates that the Prb and IPD113 families are closely related.
118
119 **Domain II swapping across clades except for Prb**
120 We next generated domain-level phylogenetic trees. While the domain I tree largely recapitulated the
121 topology of the two-domain tree (Fig. 3A), the domain II tree revealed extensive intermixing among
122 Cry, IPD113, 2D/3D-mixed, and Photobacterium clades (Fig. 3B), suggesting that domain II swapping
123 events occurred across these four clades. Specifically, several nematocidal Cry proteins (Cry5A,
124 Cry5D, Cry5E, and Cry12), as well as Cry31 and Cry70, clustered far from other Cry proteins and
125 instead grouped with members of other clades, such as IPD113. Since domain swapping typically
126 requires homologous recombination, it is likely that the genes encoding these proteins contained
127 complementary sequences flanking the domain II coding region. These findings support the possibility
128 that proteins with Cry-like domains are remote homologues.
129
130 **Proteins containing Cry-like two domains share amino acid motifs across clades**
131 To examine remote homology at a resolution beyond the domain level, we performed a search for
132 conserved amino acid motifs. Among the 60 motifs identified by MEME (Bailey et al. 2015) (Fig.
133 S3A), 26 motifs were shared across at least two clades (Fig. 4A). They locate within domain I α-helices,

134 the linker region between domains I and II, and domain II β-strands, with their relative positions being
135 largely conserved (Fig. S4A). t-SNE visualization of motif patterns revealed partial intermixing
136 between clades in all cases, whether motifs were derived from domain I alone, domain II alone, or the
137 domain I–II region (Fig. 4B–D).
138
139 On average, the proteins in each clade possess more than four shared domain I motifs, each
140 approximately 20 amino acids in length (Fig. S4B), suggesting that domain I is unlikely to have been
141 acquired through convergent evolution. In line with this, the α5 motif, located at the center of the
142 domain I α-helical bundle, shows a gradual transition across clades (Fig. 4A and E). In Cry proteins,
143 the α5 motif corresponds to block 1, one of the five conserved sequence blocks (Höfte and Whiteley
144 1989). 79% of the 2D/3D-mixed clade members including non-Cry proteins retain the Cry-type α5
145 motif. Furthermore, 50% and 37% of these members also share the α7 and the domain I–II linker–β1
146 motifs, respectively, both of which correspond to Cry block 2 (Fig. 4A and E). Most members of the
147 IPD113 clade possess an α5 motif that is similar, though not identical, to the Cry-type α5 motif, while
148 some Cry proteins in both the Cry and 2D/3D-mixed clades carry the IPD113-type α5 motif. Members
149 of the Prb clade, on the other hand, have a longer α5 motif that is distinct from both the Cry- and
150 IPD113-types. Notably, Cry5Ca1 and Cry21Da1 classified in the Cry clade as well as some members
151 of the IPD113 clade, contain the Prb-type α5 motif. Since the motifs flanking the IPD113-type α5 motif
152 in Cry5Ea1 and Cry21Ga1 and the Prb-type α5 motif in Cry5Ca1 and Cry21Da1 are similar to those
153 of closely related Cry5 and Cry21 proteins (Fig. S4C), it is plausible that either a subdomain-level
154 swapping event occurred, or the surrounding motifs emerged after acquiring the IPD113-type and Prb-
155 type α5 motif. This observation further suggests an evolutionary relationship of domain I across Cry,
156 IPD113 and Prb proteins.
157
158 In contrast, shared motifs are less prevalent in domain II, with 20–40% of sequences in the IPD113,
159 2D/3D-mixed, and Photobacterium clades lacking any such motifs. Although only Prb proteins showed
160 no signatures of domain swapping in domain II (Fig. 3B), they nevertheless possess two long
161 conserved motifs, each corresponding to a pair of β-strands—one shared with non-Cry 2D/3D-mixed
162 proteins and the other shared with both non-Cry 2D/3D-mixed and IPD113 proteins (Fig. 4A). These
163 findings suggest a common ancestry of domain II across these clades.
164
165 **Domain core structures are highly conserved beyond clades**
166 To further investigate signatures of divergent evolution, we performed structural analyses, as protein
167 structures are more conserved than amino acid sequences (Illergård, Ardell, and Elofsson 2009). For
168 structural analyses, we selected 498 predicted models with high confidence, defined as an average

169 pLDDT ≥ 80 and PAE ≤ 10 Å in both domains I and II. We compared the domains I and II of *P.*
170 *akhurstii* Prb and *L. wrightii* IPD113 with those of other predicted structures. Although sequences from
171 other clades exhibited low sequence scores, some of them still fell within the root-mean-square
172 deviation (RMSD) range observed for comparisons within the Prb or IPD113 clade (Fig. 5A-D). When
173 aligned pairwise, we found that even among structures from other clades, approximately 25–50% of
174 the Cα atoms in each domain could be superimposed within 1.5 Å (Fig. 5E–H). In domain I, the central
175 helix α5 together with several adjacent helices showed strong structural conservation (Fig. 5I and J).
176 In domain II, one or more of the antiparallel β-sheets forming the β-prism overlapped almost
177 completely (Fig. 5K and L). The domain-level GS tree based on the 3D interaction (3Di) alphabet, a
178 type of structural alphabet that describes tertiary interactions in FoldMason (van Kempen et al. 2023),
179 revealed partial intermixing among clades (Fig. S5). These cross-clade conservations of domain core
180 structures support the notion that these domains have undergone divergent evolution while maintaining
181 their functional architecture, rather than having arisen through convergent evolution.
182
183 **Cry domain III and Pra are not evolutionarily related**
184 To test the homology among Cry domain III, Pra, and the Pra-like domains of insect Prb-containing
185 proteins, we performed the same analyses on these domains. The GS tree based on amino acid
186 sequences separated Cry domain III from Pra and the insect Pra-like domains, while Pra and the insect
187 Pra-like domains formed a single clade (Fig. S6A). Motif analysis revealed that Cry domain III did not
188 share motifs with Pra or the insect Pra-like domains (Figs. S6C and D). Meanwhile, Pra and the insect
189 Pra-like domains shared two motifs (Fig. S6C and D). The GS tree based on structure (3Di) also
190 showed intermixing between Pra and the insect Pra-like domains, but not with Cry (Fig. S6B). In
191 pairwise structural alignments with *P. akhurstii* Pra, some Cry domain III showed pruning of
192 approximately 30% of the domain length; however, no conservation of the core structures observed in
193 domain I or II was detected (Fig. S6E-G). These results suggest that Cry domain III and Pra are not
194 remote homologues, whereas Pra and the insect Pra-like domains share a common evolutionary origin.
195 *Pra* and *Prb* genes are located adjacent to each other in the *Photorhabdus* genome (Waterfield et al.
196 2005), suggesting that the two proteins have fused into a single protein at some point.
197
198 **Many horizontal gene transfers shaped broad taxonomic distribution**
199 As our results suggested a common evolutionary origin of Cry, IPD113, and Prb, we hypothesized that
200 ferns, as well as other eukaryotes and archaea, may have acquired these genes through ancient HGT.
201 In the GS tree shown in Fig. 1C, the 2D/3D-mixed and IPD113 clades include a mixture of bacterial
202 and non-bacterial clades, and the Prb clade contains a single bacterial member nested within an animal
203 clade (Fig. S7). Since the GS method does not estimate evolutionary distances, we additionally

204 constructed maximum-likelihood (ML) trees for each clade to obtain amino acid distances under ML
205 substitution models. In this analysis, we used only the sequences that were subjected to structural
206 comparisons. As expected, approximately 30–40% of all branches in ML trees lacked strong support
207 (Fig. 6A and S9). To narrow down the possible scenarios of ancient HGT, we compared amino acid
208 (AA) distances with structural distances (Fig. S8). We found that a number of pairs between bacteria
209 and non-bacterial taxa exhibit a lower 3Di/AA distance ratio than the regression curves derived from
210 bacteria–bacteria pairs within the same clade. This indicates that, for the same amino acid distance,
211 structures of those pairs are closer than expected for typical bacterial pairs. Such structural closeness
212 supports the placement of non-bacterial members within bacterial clades and thus suggests ancient
213 HGT events. In particular, pairs that also show shorter AA distances (below the bacterial mean) can be
214 regarded as stronger candidates, because lower sequence divergence combined with high structural
215 similarity between bacteria and non-bacterial taxa is more parsimoniously explained by relatively
216 recent HGT than by independent evolution.
217
218 In the IPD113 clade, we found bacteria–fern pairs that met the criteria (Fig. 6), suggesting a bacterial
219 origin for IPD113. Specifically, seven bacterial sequences from *Brevibacillus*, *Aneurinibacillus*,
220 *Clostridia*, *Cohnella*, *Salinicola*, and *Bermanella*, showed a low 3Di/AA distance ratio and a short AA
221 distance (both below the bacterial mean minus one standard deviation) (Fig. 5B, large solid yellow-
222 green circles), making them a strong candidate HGT donor. Interestingly, most of the bacteria that
223 emerged as potential HGT donors to ferns also exhibited a lower 3Di/AA distance ratio with arbuscular
224 mycorrhizal (AM) fungi, known as plant endosymbionts, implying a scenario of bacteria–AM fungi–
225 fern HGT. Likewise, bacteria–symbiont–sea anemone HGTs were suggested, as well as HGTs to frogs,
226 newts, sturgeons, and *Fusarium* (Fig. 6). In the 2D/3D-mixed clade, additional HGT events were
227 suggested toward *Fusarium* and other fungal plant pathogens and parasites, nematode-trapping fungi,
228 and marine archaea (Figs. S9A–B). In the Prb clade, HGTs from bacteria, including the insect
229 endosymbiont *Rickettsia*, to aphids and beetles were also suggested (Figs. S9C–D). However,
230 compared with the distribution among bacteria–bacteria pairs, the bacteria–animal pairs tended to
231 show larger AA and 3Di distances in this clade (Fig. S8C and F), suggesting that the actual HGT donors
232 may include yet-unidentified bacterial lineages harboring Pra+Prb three-domain proteins. These results
233 suggest that, in addition to HGT among diverse bacteria, close ecological interactions such as
234 symbiosis have contributed to shaping the current broad taxonomic distribution of these genes.
235
236 **Discussion**
237 Since Cry proteins have so far been identified only in *B. thuringiensis* and several other Bacillota
238 bacteria including *Paenibacillus* and *Lysinibacillus*, they have generally been considered to represent

239 a lineage uniquely acquired and diversified within these bacteria. Here, however, we show that the first
240 two domains are broadly distributed across diverse organisms, with the Cry protein family representing
241 one terminal outcome of this broader diversification. Although our results cannot determine whether
242 the three-domain Cry proteins evolved by gaining an additional domain from two-domain ancestors or
243 two-domain proteins emerged by domain loss from three-domain ancestors, the phylogenetic and motif
244 analyses (Figs. 2–4) suggest that the 2D/3D-mixed clade has played an intermediate, bridging role
245 linking the Cry, IPD113, and Prb lineages. In the 2D/3D-mixed clade, both 2D proteins and canonical
246 3D Cry proteins were found in *Paenibacillus* species (Table S2), suggesting that this clade represents
247 a site of domain-3 gain or loss. In addition, certain *Lysinibacillus* species produce 2D proteins that
248 belong to the IPD113 clade (Table S2).
249
250 Domain swapping has been proposed to underlie the diversification of Cry proteins (de Maagd, Bravo,
251 and Crickmore 2001). Our results extend this concept beyond the Cry family, encompassing proteins
252 in the 2D/3D-mixed, IPD113, and Photobacterium clades as well. Many Cry proteins that form
253 intermixed clades are nematocidal (Fig. 3B), suggesting that ecological interactions may have occurred
254 across bacteria including Bt strains that use nematodes as hosts. In contrast, domain II swapping
255 appears not to have occurred with the Prb clade. This may be explained by the fact that *Prb* genes are
256 chromosomally encoded in *Photorhabdus*, *Xenorhabdus*, *Sodalis*, and *Yersinia* (Wang et al. 2023)
257 whereas Cry proteins are typically plasmid-borne. Furthermore, *Photorhabdus* and *Xenorhabdus* are
258 endosymbionts of nematodes, which may limit their opportunities for horizontal gene exchange with
259 other bacteria after the establishment of symbiosis.
260
261 Our results indicate the bacterial origin of fern IPD113. Surprisingly, AM fungi also possess IPD113
262 genes. Symbiotic relationships with AM fungi are common in ferns. Among the ferns analyzed in this
263 study, *Adiantum aethiopicum*, *Blechnum occidentale*, *Lygodium flexuosum*, and *Christella dentata*
264 have been reported to exhibit AM colonization (Kumari et al. 2024). The hyphosphere of AM fungi
265 harbors a microbiome distinct from that of bulk soil and the rhizosphere. This microbiome includes
266 *Brevibacillus*, *Cohnella*, *Tahibacter*, *Undibacterium*, *Hymenobacter*, and *Novosphingobium* (Zhang et
267 al. 2022)—all of which have Cry-like two domain proteins and were identified as potential bacterial
268 donors for HGT to AM fungi and ferns (Fig. 6). This is therefore consistent with a scenario of bacteria–
269 AM fungi–fern HGT. Despite the widespread symbiosis of AM fungi with vascular plants, only ferns
270 and bryophytes seem to possess this gene. Understanding the evolutionary significance and underlying
271 mechanisms of this unique distribution in plants will require further investigation.
272
273 Ferns appear to have retained IPD113 for its original insecticidal function. However, the physiological

roles of proteins containing Cry-like two domains in many bacteria, eukaryotes, and archaea remain unclear. Previous studies have noted that Pra and Prb proteins resemble a beetle protein, which we refer to in this study as the insect Pra+Prb protein (Duchaud et al. 2003; Waterfield et al. 2005). The insect Pra+Prb protein was originally isolated from the hemolymph of the Colorado potato beetle *Leptinotarsa decemlineata* and predicted to function as a juvenile hormone esterase (Vermunt et al. 1997, 1998), although its true function remains unknown. The western corn rootworm *Diabrotica virgifera virgifera* possesses four copies of the gene encoding the three-domain protein, one of which lacks introns (Table S2), suggesting that HGT events have occurred repeatedly, including a relatively recent one. The frequent occurrence of HGT may be attributed not only to close biological interactions but also to the potential contribution of transposable elements. Future studies are needed to elucidate both the functions of proteins containing Cry-like two domains and the genetic mechanisms underlying their broad taxonomic distribution.

Material and methods

BLAST searches

Non-redundant protein sequences (nr) were searched using BLASTp (Camacho et al. 2009). The raw output data is shown in Table S1. Taxonomic information was obtained from the NCBI Taxonomy database (Federhen 2012). For downstream analyses, sequences showing ≥95% amino acid identity were removed using CD-HIT (Li and Godzik 2006).

Structure prediction and alignment

Tertiary structures of full-length BLAST-hit proteins and Cry proteins were predicted using AlphaFold 3 (Abramson et al. 2024) under template-free conditions. Unless otherwise specified, all other parameters followed the default settings of AlphaFold3. For Cry proteins, we selected Cry proteins with quaternary rank 1 from the Bacterial Pesticidal Protein Resource Center (www.bpprc.org). Cry1If1, Cry1Jc1, Cry8La1, and Cry53Aa1 were excluded from the analysis because their amino acid sequences contained unknown residues (X). The top-ranked model among five predictions was used for further analyses. Predicted structures that did not contain complete Cry-like two domains were excluded.

Structural alignments were performed using FoldMason (Gilchrist, Mirdita, and Steinegger 2024). To ensure precise domain trimming, four separate alignments were generated: (1) Cry, (2) BLAST hits obtained using *L. wrightii* IPD113 and *C. faecalis* IPD113-like proteins as queries, (3) BLAST hits obtained using *P. akhurstii* and *V. parahaemolyticus* Prb as queries, and (4) BLAST hits obtained using *P. akhurstii* and *V. parahaemolyticus* Pra as queries. Domain trimming was guided by the reference structures (Cry1Aa1, *L. wrightii* IPD113, *P. akhurstii* Prb, and Pra). For domain I, residues from the N-terminal amino acid forming the first α-helix (α1) to the C-terminal amino acid forming the seventh α-helix (α7) were retained. For domain II, residues from the N-terminal amino acid forming the first β-strand (excluding those included in domain I) to the C-terminal amino acid forming the last β-strand were extracted. The same criteria were applied to Cry domain III and Pra domain. For full-length sequences of the domain I–II region, the linker region between domains I and II was included. After trimming, we evaluated qualities of the trimmed predicted structures using AF3 confidence metrics, pLDDT and predicted aligned error (PAE). For phylogenetic and motif analysis, we used 557 AF3 models with mean pLDDT ≥60 (a per-atom confidence index) to extract amino acid sequences corresponding to domain I, domain II, and the domain I–II region. For structural analysis, we used 498 predicted models with high confidence, defined as an average pLDDT ≥ 80 and PAE ≤ 10 Å in both domains I and II.

322 **Phylogenetic analysis**

323 Phylogenetic trees were reconstructed using the Graph Splitting method (Figs. 1, 2, S4 and S6) (Matsui
324 and Iwasaki 2020) and the maximum likelihood method (Figs. 5 and S8). For GS trees, we used Graph
325 Splitting v2.4. The Edge Perturbation (EP) values were computed with 100 replicates. For ML trees,
326 the amino acid sequences were aligned using MAFFT v7.526 (Katoh and Standley 2013) and the trees
327 were generated with IQ-TREE2 v2.3.6 (Minh et al. 2020) using the ModelFinder option with 1,000
328 ultrafast bootstrap and SH-aLRT replicates. The WAG+F+I+R4, pfam+F+R5, and yeast+R5 models
329 were selected for the 2D/3D-mixed, IPD113, and Prb clades, respectively. All trees were visualized
330 using the ggtree R package (Yu et al. 2017).

331

332 **Motif analysis**

333 Amino acid motifs were identified using MEME v5.5.7 (Bailey et al. 2015). Preliminary searches
334 suggested that 60 and 20 motifs were sufficient to cover the entire sequences of the two-domain and
335 domain III/Pra/insect Pra-like protein datasets, respectively, with the identified motifs evenly
336 distributed across the alignments. To evaluate whether the identified motifs contained sufficient
337 information and evolutionary relevance, we applied two criteria: (1) average information content ≥ 1.0,
338 and (2) at least two positions with information content ≥ 1.5. Only one motif (motif 40) identified in
339 the domain I–II sequences failed to meet these criteria and was excluded from subsequent analyses. A
340 binary matrix representing the presence or absence of each motif across sequences was constructed
341 and subjected to dimensionality reduction using t-SNE implemented in the Rtsne R package. The
342 perplexity parameter was set to one-third of the number of sequences.

343

344 **Structural analysis**

345 Protein structures were visualized using UCSF ChimeraX v1.8 (Pettersen et al. 2004). Pairwise
346 structural alignments were performed in UCSF ChimeraX v1.8 using the MatchMaker command
347 (sequence alignment algorithm, Needleman–Wunsch; Matrix, BLOSUM-62), which provided
348 alignment parameters including sequence score, RMSD (all and pruned pairs), and number of aligned
349 residues.

350

351 **Exponential fitting of AA–3Di distance**

352 AA distances were derived from the distance matrices of the ML trees generated using IQ-TREE2.
353 Pairwise distances between MAFFT-aligned 3Di sequences were calculated as *p*-distance, defined as
354 the proportion of differing positions between two sequences after excluding gap positions. Specifically,
355 for each sequence pair i, j, *p*-distance was calculated as:

$$p_{ij} = 1 - \frac{N_{\text{match}}}{N_{\text{compared}}}$$

where N_{match} is the number of aligned positions with identical symbols (excluding gaps), and N_{compared} is the total number of positions compared (i.e., sites without gaps in either sequence).

The relationship between sequence and structural similarity is strongly nonlinear, and structural divergence becomes particularly pronounced at low sequence identity; therefore, exponential models have been used to approximate this divergence (Chothia and Lesk 1986). To obtain the bacteria–bacteria regression, we fit an asymptotic exponential model of the form

$$y = A - B\exp\{-C(x - D)\}$$

where x is the AA distance, y is the 3Di distance, and A, B, C, D are free parameters. Nonlinear least-squares fitting was performed using the minpack.lm R package. Outliers beyond ±3 standard deviations of the residuals were removed, and the model was refit using the remaining bacteria–bacteria pairs.

Acknowledgements

The authors thank Keisuke Shoji (Tokyo University of Agriculture and Technology) for fruitful discussion. This research was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (25K22485).

Author Contributions

H.E. designed research, performed analysis and wrote the manuscript. H.K. contributed to data interpretation and critically reviewed the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Data availability

All data generated and used for the analyses are deposited in the CBS Data Sharing Platform (URL and password are provided for reviewers and it will be made public upon acceptance)

Code availability

Custom code used in the analysis is deposited in the CBS Data Sharing Platform (URL and password are provided for reviewers and it will be made public upon acceptance). Protein structure modeling and structural alignment were performed using AlphaFold 3 (https://github.com/google-deepmind/alphafold3) and FoldMason (https://github.com/steineggerlab/foldmason). Phylogenetic analysis was conducted using the Graph Splitting method (https://github.com/MotomuMatsui/gs) and IQ-TREE2 (https://github.com/iqtree/iqtree2). Amino acid motif analysis was performed using MEME (https://github.com/cinquin/MEME). Sequence alignment and processing were performed using MAFFT (https://github.com/GSLBiotech/mafft) and CD-HIT (https://github.com/weizhongli/cdhit). For plotting and data processing, we used Python v.3.10.18 (https://www.python.org/), R v.4.4.1 (https://www.r-project.org/), and R studio v2024.12.1+563.

Figure legends

Fig. 1 | Putative IPD113 homologues identified by BLASTp.

(A) Domain architecture of Cry, Prb, and IPD113 proteins. Cry (PDB: 1CIY), Prb (PDB: 7FDP), Pra (PDB: 7FCN), and IPD113 (PDB: 8D2J).

(B) Scatter plots of amino acid identity versus E-value from BLASTp searches. *Leptochilus wrightii* IPD113 protein. Dots indicate putative homologues (E-value < 1e−5). Raw BLASTp results are provided in Table S1.

(C) AF3-predicted structures of putative IPD113 homologues.

Fig. 2 | The GS tree of domain I–II region of **proteins containing Cry-like two domains.**

(A) Graph Splitting tree of the domain I proteins containing Cry-like two domains. Black and gray branches indicate high and low reliability, supported by edge perturbation (EP) values ≥ 0.9 and < 0.9, respectively. Red, light blue, and green node circles indicate known Cry proteins, Prb proteins from *Photorhabdus*, *Xenorhabdus*, and *Vibrio*, and fern IPD113 proteins.

(B) Sankey plots showing the taxonomy and the number of domains, based on AF3-predicted structures, in each clade.

Fig. 3 | The domain-level GS tree of proteins containing Cry-like two domains.

GS trees of domain I (A) and domain II (B) sequences. Black and gray branches indicate high and low reliability, supported by edge perturbation (EP) values ≥ 0.9 and < 0.9, respectively. Node circle colors correspond to the clades defined in the two-domain GS tree shown in Fig. 1C. Cry proteins shown in bold indicate nematocidal Cry proteins.

Fig. 4 | Amino acid motif patterns.

(A) Raster plot of amino acid motif patterns. Twenty-six motifs shared across at least two clades were mapped onto the two-domain GS tree shown in Fig. 2A.

(B–D) t-SNE visualization of motif patterns. The motif presence/absence matrix from (A) was dimensionally reduced to visualize motifs located in domain I (B), domain II (C), and the domain I–II region (D).

(E) α5 motifs of each clade. Cry proteins carrying IPD113-type or Prb-type α5 motifs are listed.

Fig. 5 | Domain-level structural comparisons.

The domain I and II structures of *Photorhabdus akhurstii* Prb and *Leptochilus wrightii* IPD113 were compared with those of other proteins containing Cry-like two domains.

(A–D) Sequence similarity scores and whole-domain RMSD values obtained from pairwise alignments and structural superpositions. Dot colors correspond to the same clade color scheme shown in Fig. 1C.

(E–H) Pruned residue pairs and their pruned RMSD values derived from structural superpositions. Dot colors correspond to the same clade color scheme shown in Fig. 1C.

(I–L) Representative structural superpositions with proteins from different clades. Pruned regions are highlighted, and non-pruned regions are shown with increased transparency.

Fig. 6 | Potential HGT pairs in the IPD113 clade.

(A) Maximum likelihood tree of the IPD113 clade. Black branches indicate high reliability (Ultrafast Bootstrap ≥ 95% and SH-aLRT ≥ 80%); other branches are shown in gray. Node circle colors indicate taxonomy. The scale bar represents the number of substitutions per site.

(B) Pairs with a low 3Di/AA distance ratio, falling below the bacteria–bacteria regression. The vertical position of each circle corresponds to the sequence labels in the phylogenetic tree shown in (A). Each column corresponds to a red-highlighted bacterial sequence (potential HGT donor), with which sequences of non-bacterial sequences (rows) are compared. Proteins plotted in the same column have a 3Di/AA distance ratio that falls below the bacteria–bacteria regression in Fig. S8, indicating potential HGT recipients. The fill opacity of each circle reflects its amino acid (AA) distance relative to the mean of bacteria–bacteria pairs: circles with AA distances above the mean are outlined with high transparency, those between the mean and −1 SD are semi-transparent, and those below −1 SD are shown as fully opaque. Large circles indicate proteins below −1 SD from the regression, while small circles indicate those between 0 and −1 SD. In particular, large opaque circles are more likely to represent candidate HGT pairs involving bacterial sequences.

457 **Fig. S1 | Putative IPD113 and Prb homologues identified by BLASTp.**
458 (A, B) Scatter plots of amino acid identity versus E-value from BLASTp searches. *Cohnella faecalis*
459 IPD113-like protein (A) and *Photorhabdus akhurstii* Prb protein (B) were used as queries. Dots
460 indicate putative homologues (E-value < 1e−5). Raw BLASTp results are provided in Table S1. (C)
461 AF3-predicted structures of a putative Prb homologue.
462
463 **Fig. S2 | Sankey plots showing the taxonomy of proteins in the 2D/3D-mixed (A), IPD113 (B),**
464 **and Prb (C) clades.** Full taxonomic assignments derived from NCBI are listed in Table S2.
465
466 **Fig. S3 | Amino acid motifs identified by MEME.**
467 Amino acid motifs of (A) Cry-like 2D proteins and (B) domain III, Pra, insect Pra-like domain
468 proteins.
469
470 **Fig. S4 | Positions and counts of amino acid motifs.**
471 (A) Distribution of motif positions mapped onto the structural alignment generated by FoldMason.
472 Each row corresponds to one detected motif. Labels show the number of occurrences (*n*) and the
473 mean starting position of the motif ± standard deviation. Motif 40 was excluded from the analysis
474 because it did not meet two information-content criteria (see Methods). Light blue boxplots show the
475 median (vertical black line), first and third quartiles (box), and whiskers extending to 1.5× the
476 interquartile range. Gray dots represent individual occurrences of each motif. For most motifs, the
477 interquartile range is extremely small, making the light blue boxes barely visible.
478 (B) Counts of motifs in domain I, domain II, and the domain I–II region among members of each
479 clade. Boxplots show the median (horizontal line), first and third quartiles (box), and whiskers
480 extending to 1.5× the interquartile range (outliers hidden); diamonds indicate the mean.
481 (C) Motif locations in Cry5 and Cry21 proteins. Each rectangle represents a single motif, with colors
482 indicating the combination of clades in which the motif is shared.
483
484 **Fig. S5 | Domain-level GS trees based on 3Di sequences.**
485 (A) Domain I tree. (B) Domain II tree. Black and gray branches indicate high and low reliability,
486 supported by edge perturbation (EP) values ≥ 0.9 and < 0.9, respectively. Node circle colors
487 correspond to the clades defined in the two-domain GS tree shown in Fig. 1C.
488
489 **Fig. S6 | Analyses of Cry domain III, Pra, and insect Pra-like domains.**
490 (A, B) GS trees of Cry domain III, Pra, and insect Pra-like domains based on amino acid (A) and 3Di
491 (B) sequences. Black and gray branches indicate high and low reliability, supported by edge

492 perturbation (EP) values ≥ 0.9 and < 0.9, respectively. Node circle colors correspond to the clades.
493 (C) Raster plot of amino acid motif patterns. Twenty motifs were mapped onto the GS tree shown in
494 (A).
495 (D) t-SNE visualization of motif patterns. The motif presence/absence matrix from (A) was
496 dimensionally reduced. Sequences without motifs (Cry11Ba1, Cry18Aa, Cry18Ba1, Cry18Ca1) were
497 excluded from this analysis. Dot colors correspond to the same clade color scheme as in (A) and (B).
498 (E–G) Structural comparisons of *P. akhurstii* Pra with other members shown in (A). Colors
499 correspond to the same clade color scheme as in (A).
500 (E) Sequence similarity scores and whole-domain RMSD values obtained from pairwise alignments
501 and structural superpositions.
502 (F) Pruned residue pairs and their pruned RMSD values derived from structural superpositions.
503 (G) Representative structural superpositions with proteins from different clades. Pruned regions are
504 highlighted, and non-pruned regions are shown with increased transparency.
505
506 **Fig. S7 | The two-domain GS tree colored by taxonomy.**
507 The two-domain GS tree shown in Fig. 1C is colored by taxonomy for the IPD113 (A), 2D/3D-
508 mixed (B), and Prb (C) clades. Black and gray branches indicate high and low reliability, supported
509 by edge perturbation (EP) values ≥ 0.9 and < 0.9, respectively.
510
511 **Fig. S8 | Distributions of AA and 3Di distances.**
512 (A–C) Distributions of amino acid (AA) and 3Di distances among bacteria–bacteria pairs in the
513 IPD113 (A), 2D/3D-mixed (B), and Prb (C) clades. AA distances were derived from the distance
514 matrices of the ML trees shown in Figs. 5 and S9A and C. 3Di distances were calculated as pairwise
515 distances (p-distances) based on 3Di alignments. Black solid lines indicate the exponential fit; dotted
516 lines indicate fit –1 SD. Red dots represent outliers.
517 (D–F) Distributions of AA and 3Di distances among all pairs in the IPD113 (D), 2D/3D-mixed (E),
518 and Prb (F) clades. Black solid lines indicate the exponential fit; dotted lines indicate fit –1 SD. Dots
519 representing pairs between bacteria and non-bacterial taxa are color-coded.
520
521 **Fig. S9 | Potential HGT pairs in the 2D/3D-mixed and Prb clades.**
522 (A, C) Maximum likelihood tree of the 2D/3D-mixed and Prb clades. Black branches indicate high
523 reliability (Ultrafast Bootstrap ≥ 95% and SH-aLRT ≥ 80%); other branches are shown in gray. Node
524 circle colors indicate taxonomy. The scale bar represents the number of substitutions per site.
525 (B, D) Pairs with a low 3Di/AA distance ratio, falling below the bacteria–bacteria regression. The
526 vertical position of each circle corresponds to the sequence labels in the phylogenetic tree shown in

(A). Each column corresponds to a red-highlighted bacterial sequence (potential HGT donor). Proteins plotted in the same column have a 3Di/AA distance ratio that falls below the bacteria–bacteria regression in Fig. S8, indicating potential HGT recipients. The fill opacity of each circle reflects its amino acid (AA) distance relative to the mean of bacteria–bacteria pairs: circles with AA distances above the mean are outlined with high transparency, those between the mean and −1 SD are semi-transparent, and those below −1 SD are shown as fully opaque. Large circles indicate proteins below −1 SD from the regression, while small circles indicate those between 0 and −1 SD. In particular, large opaque circles are more likely to represent candidate HGT pairs involving bacterial sequences.

Table S1 | Raw BLASTp results

Table S2 | List of proteins containing Cry-like two domains used in this study.

References

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, et al. 2024. "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3." *Nature*, May, 1–3.

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. "The MEME Suite." *Nucleic Acids Research* 43 (W1): W39-49.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421.

Chothia, C., and A. M. Lesk. 1986. "The Relation between the Divergence of Sequence and Structure in Proteins." *The EMBO Journal* 5 (4): 823–26.

Crickmore, Neil, Colin Berry, Suresh Panneerselvam, Ruchir Mishra, Thomas R. Connor, and Bryony C. Bonning. 2021. "A Structure-Based Nomenclature for Bacillus Thuringiensis and Other Bacteria-Derived Pesticidal Proteins." *Journal of Invertebrate Pathology* 186 (November): 107438.

Duchaud, Eric, Christophe Rusniok, Lionel Frangeul, Carmen Buchrieser, Alain Givaudan, Séad Taourit, Stéphanie Bocs, et al. 2003. "The Genome Sequence of the Entomopathogenic Bacterium Photorhabdus Luminescens." *Nature Biotechnology* 21 (11): 1307–13.

Federhen, Scott. 2012. "The NCBI Taxonomy Database." *Nucleic Acids Research* 40 (Database issue): D136-43.

596 Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt,
597 Elaine C. Meng, and Thomas E. Ferrin. 2004. "UCSF Chimera--a Visualization System for
598 Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13): 1605–12.
599 Prashar, Arpit, Omkar U. Kinkar, Ashwani Kumar, Ashok B. Hadapad, Ravindra D. Makde, and
600 Ramesh S. Hire. 2023. "Crystal Structures of PirA and PirB Toxins from Photorhabdus
601 Akhurstii Subsp. Akhurstii K-1." *Insect Biochemistry and Molecular Biology* 162 (104014):
602 104014.
603 Rost, B. 1997. "Protein Structures Sustain Evolutionary Drift." *Folding & Design* 2 (3): S19-24.
604 Vermunt, A. M., A. B. Koopmanschap, J. M. Vlak, and C. A. de Kort. 1997. "Cloning and Sequence
605 Analysis of CDNA Encoding a Putative Juvenile Hormone Esterase from the Colorado Potato
606 Beetle." *Insect Biochemistry and Molecular Biology* 27 (11): 919–28.
607 ———. 1998. "Evidence for Two Juvenile Hormone Esterase-Related Genes in the Colorado Potato
608 Beetle." *Insect Molecular Biology* 7 (4): 327–36.
609 Wang, Hao-Ching, Shin-Jen Lin, Han-Ching Wang, Ramya Kumar, Phuoc Thien Le, and Jiann-Horng
610 Leu. 2023. "A Bacterial Binary Toxin System That Kills Both Insects and Aquatic Crustaceans:
611 Photorhabdus Insect-Related Toxins A and B." *PLoS Pathogens* 19 (5): e1011330.
612 Waterfield, Nicholas, Shizuo George Kamita, Bruce D. Hammock, and Richard ffrench-Constant.
613 2005. "The Photorhabdus Pir Toxins Are Similar to a Developmentally Regulated Insect
614 Protein but Show No Juvenile Hormone Esterase Activity." *FEMS Microbiology Letters* 245
615 (1): 47–52.
616 Wei, Jun-Zhi, Amy Lum, Eric Schepers, Lu Liu, Ross T. Weston, Bruce S. McGinness, Matthew J.
617 Heckert, et al. 2023. "Novel Insecticidal Proteins from Ferns Resemble Insecticidal Proteins
618 from *Bacillus Thuringiensis*." *Proceedings of the National Academy of Sciences of the United*
619 *States of America* 120 (44): e2306177120.
620 Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017.
621 "Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their
622 Covariates and Other Associated Data." *Methods in Ecology and Evolution* 8 (1): 28–36.
623 Zhang, Lin, Jiachao Zhou, Timothy S. George, Erik Limpens, and Gu Feng. 2022. "Arbuscular
624 Mycorrhizal Fungi Conducting the Hyphosphere Bacterial Orchestra." *Trends in Plant Science*
625 27 (4): 402–11.

Frankenhuyzen, Kees van. 2009. “Insecticidal Activity of Bacillus Thuringiensis Crystal Proteins.” *Journal of Invertebrate Pathology* 101 (1): 1–16.

Gilchrist, Cameron L. M., Milot Mirdita, and Martin Steinegger. 2024. “Multiple Protein Structure Alignment at Scale with FoldMason.” *BioRxiv*. https://doi.org/10.1101/2024.08.01.606130.

Grochulski, P., L. Masson, S. Borisova, M. Pusztai-Carey, J. L. Schwartz, R. Brousseau, and M. Cygler. 1995. “Bacillus Thuringiensis CryIA(a) Insecticidal Toxin: Crystal Structure and Channel Formation.” *Journal of Molecular Biology* 254 (3): 447–64.

Höfte, H., and H. R. Whiteley. 1989. “Insecticidal Crystal Proteins of Bacillus Thuringiensis.” *Microbiological Reviews* 53 (2): 242–55.

Illergård, Kristoffer, David H. Ardell, and Arne Elofsson. 2009. “Structure Is Three to Ten Times More Conserved than Sequence--a Study of Structural Response in Protein Cores.” *Proteins* 77 (3): 499–508.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Kempen, Michel van, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. 2023. “Fast and Accurate Protein Structure Search with Foldseek.” *Nature Biotechnology*, May. https://doi.org/10.1038/s41587-023-01773-0.

Kumari, Pratibha, Meenam Bhatia, Priti Giri, and Prem Lal Uniyal. 2024. “Mycorrhizal Association and Its Relation with Pteridophytes.” *Frontiers in Microbiology* 15 (July): 1406891.

Lee, Chung-Te, I-Tung Chen, Yi-Ting Yang, Tzu-Ping Ko, Yun-Tzu Huang, Jiun-Yan Huang, Ming-Fen Huang, et al. 2015. “The Opportunistic Marine Pathogen *Vibrio Parahaemolyticus* Becomes Virulent by Acquiring a Plasmid That Expresses a Deadly Toxin.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (34): 10798–803.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics (Oxford, England)* 22 (13): 1658–59.

Maagd, R. A. de, A. Bravo, and N. Crickmore. 2001. “How Bacillus Thuringiensis Has Evolved Specific Toxins to Colonize the Insect World.” *Trends in Genetics: TIG* 17 (4): 193–99.

Matsui, Motomu, and Wataru Iwasaki. 2020. “Graph Splitting: A Graph-Based Approach for Superfamily-Scale Phylogenetic Tree Reconstruction.” *Systematic Biology* 69 (2): 265–79.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34.

Fig. 1

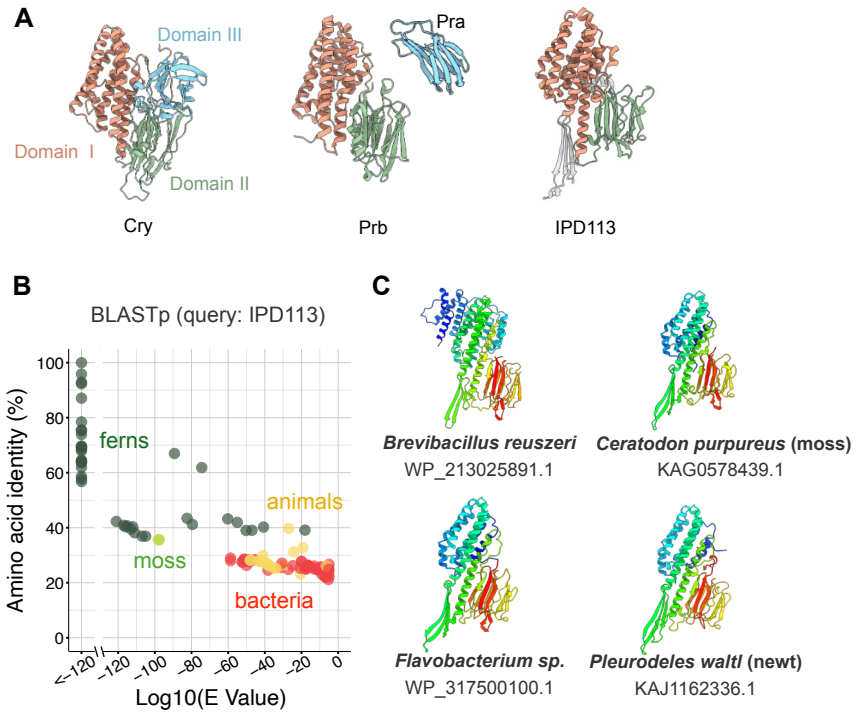

A
Domain III
Domain I
Domain II
Cry
Pra
Prb
IPD113
B
BLASTp (query: IPD113)
Amino acid identity (%)
100
80
60
40
20
0
ferns
animals
moss
bacteria
<-120
-120
-100
-80
-60
-40
-20
0
Log10(E Value)
C
Brevibacillus reuszeri
WP_213025891.1
Ceratodon purpureus (moss)
KAG0578439.1
Flavobacterium sp.
WP_317500100.1
Pleurodeles waltl (newt)
KAJ1162336.1

Fig. 2

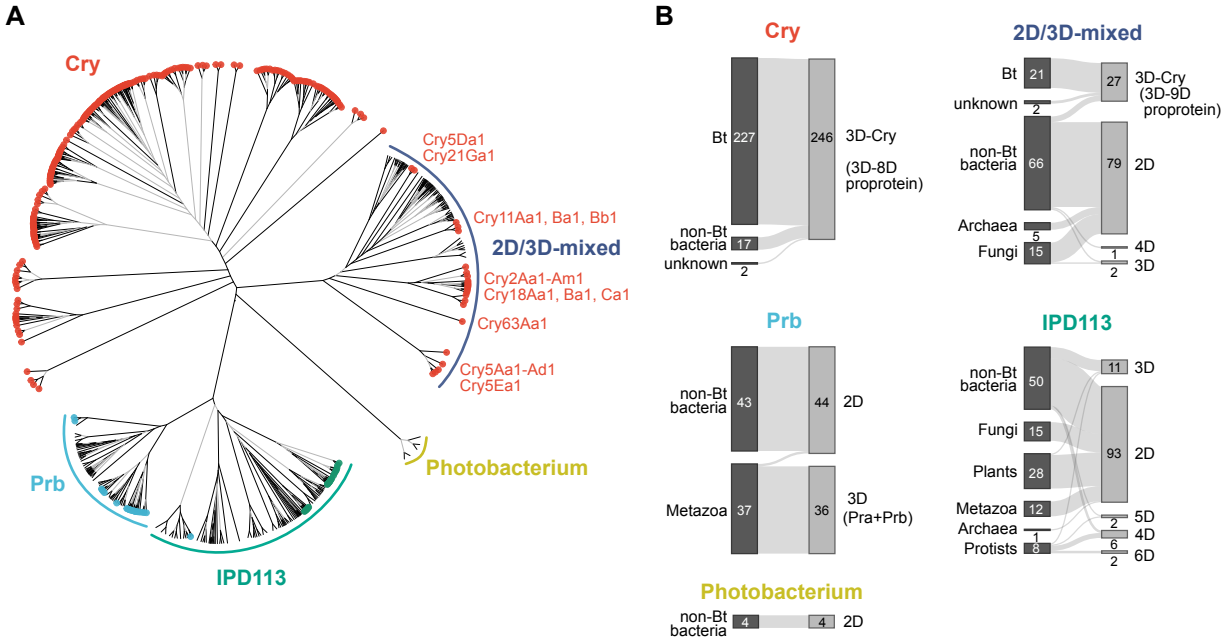

A
Cry
Cry5Da1
Cry21Ga1
Cry11Aa1, Ba1, Bb1
2D/3D-mixed
Cry2Aa1-Am1
Cry18Aa1, Ba1, Ca1
Cry63Aa1
Cry5Aa1-Ad1
Cry5Ea1
Photobacterium
Prb
IPD113
B
Cry
Bt 227
non-Bt bacteria 17
unknown 2
246 3D-Cry (3D-8D proprotein)
2D/3D-mixed
Bt 21
unknown 2
non-Bt bacteria 66
Archaea 5
Fungi 15
27 3D-Cry (3D-9D proprotein)
79 2D
1 4D
2 3D
Prb
non-Bt bacteria 43
Metazoa 37
44 2D
36 3D (Pra+Prb)
IPD113
non-Bt bacteria 50
Fungi 15
Plants 28
Metazoa 12
Archaea 1
Protists 8
11 3D
93 2D
2 5D
1 4D
6 6D
2
Photobacterium
non-Bt bacteria 4
4 2D

Fig. 3

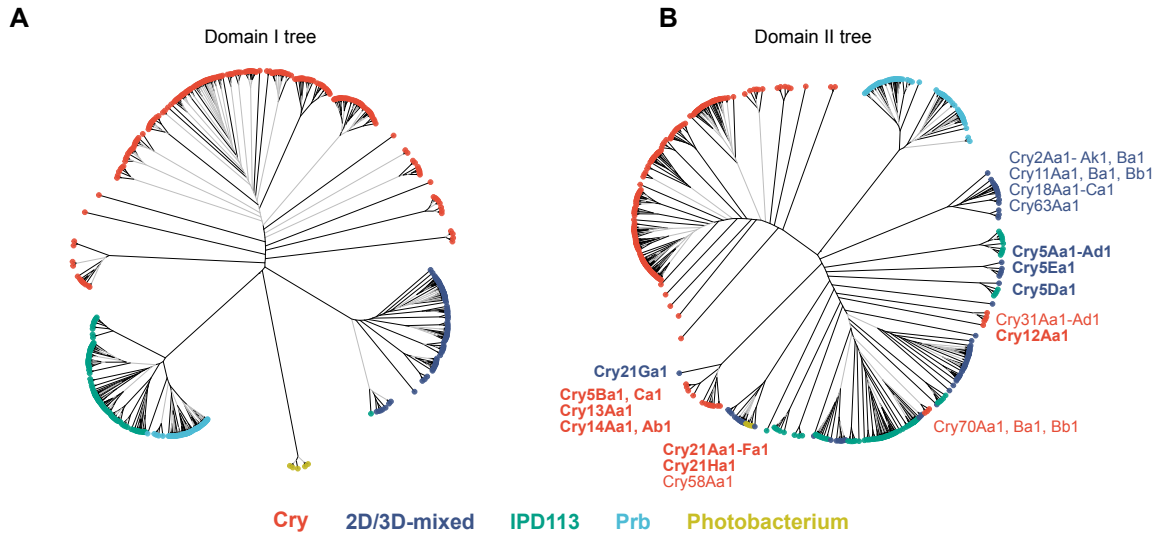

A
Domain I tree
B
Domain II tree
Cry2Aa1- Ak1, Ba1
Cry11Aa1, Ba1, Bb1
Cry18Aa1-Ca1
Cry63Aa1
Cry5Aa1-Ad1
Cry5Ea1
Cry5Da1
Cry31Aa1-Ad1
Cry12Aa1
Cry21Ga1
Cry5Ba1, Ca1
Cry13Aa1
Cry14Aa1, Ab1
Cry21Aa1-Fa1
Cry21Ha1
Cry58Aa1
Cry70Aa1, Ba1, Bb1
Cry
2D/3D-mixed
IPD113
Prb
Photobacterium

Fig. 4

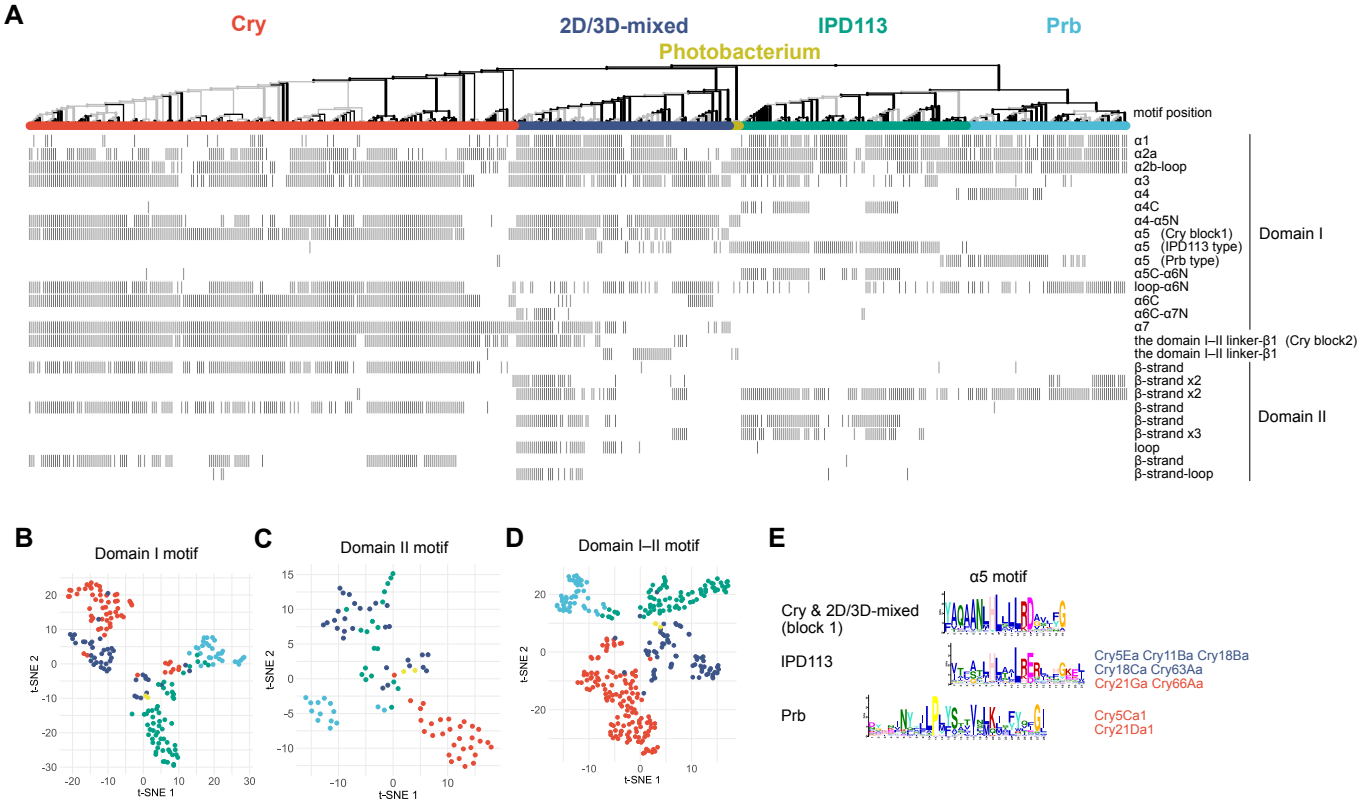

A
Cry
2D/3D-mixed
IPD113
Prb
Photobacterium
motif position
α1
α2a
α2b-loop
α3
α4
α4C
α4-α5N
α5 (Cry block1)
α5 (IPD113 type)
α5 (Prb type)
α5C-α6N
loop-α6N
α6C
α6C-α7N
α7
the domain I–II linker-β1 (Cry block2)
the domain I–II linker-β1
β-strand
β-strand x2
β-strand x2
β-strand
β-strand
β-strand x3
loop
β-strand
β-strand-loop
Domain I
Domain II
B
Domain I motif
t-SNE 1
t-SNE 2
C
Domain II motif
t-SNE 1
t-SNE 2
D
Domain I–II motif
t-SNE 1
t-SNE 2
E
α5 motif
Cry & 2D/3D-mixed (block 1)
IPD113
Prb
Cry5Ea Cry11Ba Cry18Ba
Cry18Ca Cry63Aa
Cry21Ga Cry66Aa
Cry5Ca1
Cry21Da1

Fig. 5

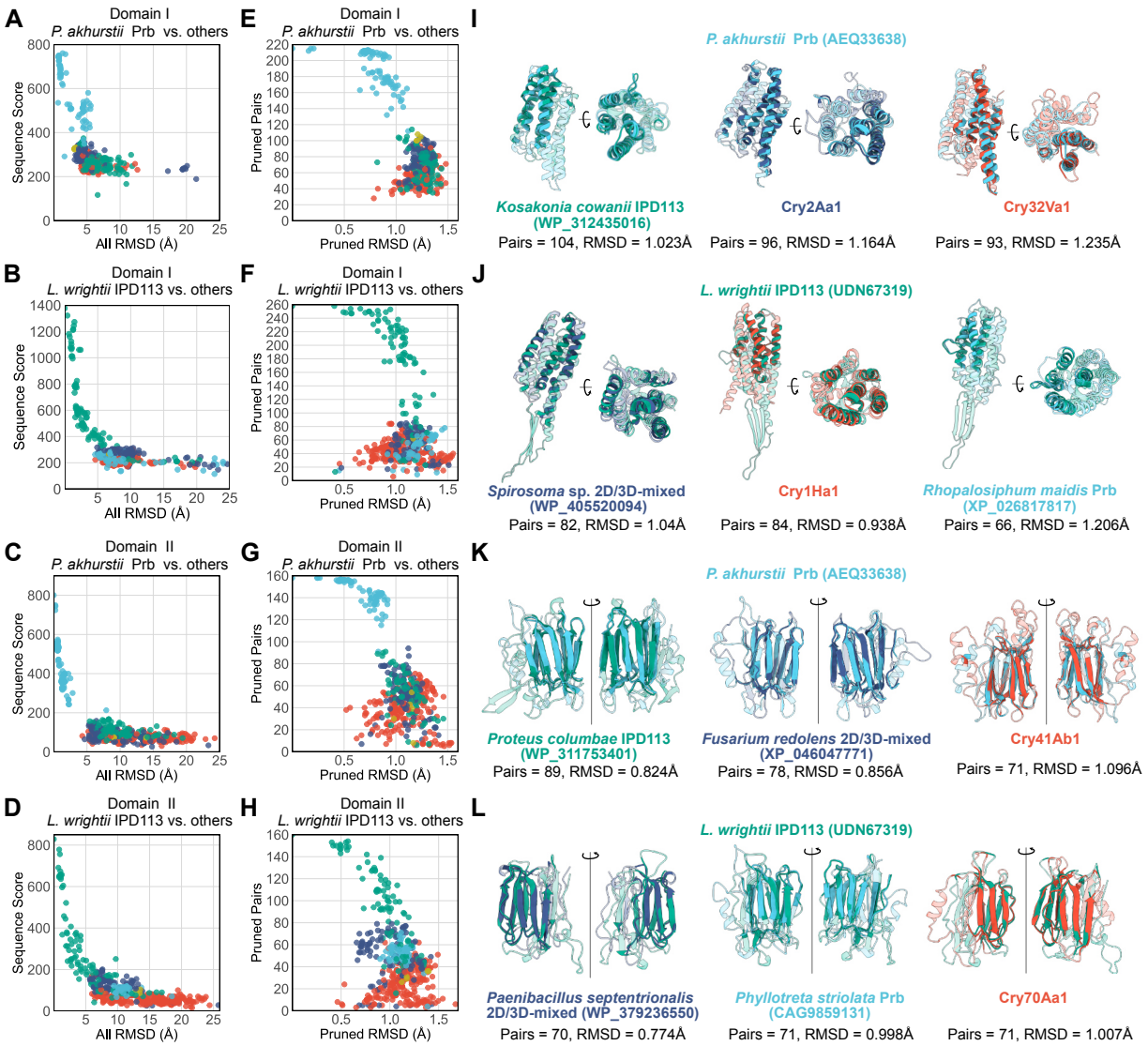

A
Domain I
P. akhurstii Prb vs. others
Sequence Score
All RMSD (Å)
B
Domain I
L. wrightii IPD113 vs. others
Sequence Score
All RMSD (Å)
C
Domain II
P. akhurstii Prb vs. others
Sequence Score
All RMSD (Å)
D
Domain II
L. wrightii IPD113 vs. others
Sequence Score
All RMSD (Å)
E
Domain I
P. akhurstii Prb vs. others
Pruned Pairs
Pruned RMSD (Å)
F
Domain I
L. wrightii IPD113 vs. others
Pruned Pairs
Pruned RMSD (Å)
G
Domain II
P. akhurstii Prb vs. others
Pruned Pairs
Pruned RMSD (Å)
H
Domain II
L. wrightii IPD113 vs. others
Pruned Pairs
Pruned RMSD (Å)
I
P. akhurstii Prb (AEQ33638)
Kosakonia cowanii IPD113 (WP_312435016)
Pairs = 104, RMSD = 1.023Å
Cry2Aa1
Pairs = 96, RMSD = 1.164Å
Cry32Va1
Pairs = 93, RMSD = 1.235Å
J
L. wrightii IPD113 (UDN67319)
Spirosoma sp. 2D/3D-mixed (WP_405520094)
Pairs = 82, RMSD = 1.04Å
Cry1Ha1
Pairs = 84, RMSD = 0.938Å
Rhopalosiphum maidis Prb (XP_026817817)
Pairs = 66, RMSD = 1.206Å
K
P. akhurstii Prb (AEQ33638)
Proteus columbae IPD113 (WP_311753401)
Pairs = 89, RMSD = 0.824Å
Fusarium redolens 2D/3D-mixed (XP_046047771)
Pairs = 78, RMSD = 0.856Å
Cry41Ab1
Pairs = 71, RMSD = 1.096Å
L
L. wrightii IPD113 (UDN67319)
Paenibacillus septentrionalis 2D/3D-mixed (WP_379236550)
Pairs = 70, RMSD = 0.774Å
Phyllotreta striolata Prb (CAG9859131)
Pairs = 71, RMSD = 0.998Å
Cry70Aa1
Pairs = 71, RMSD = 1.007Å

Fig. 6

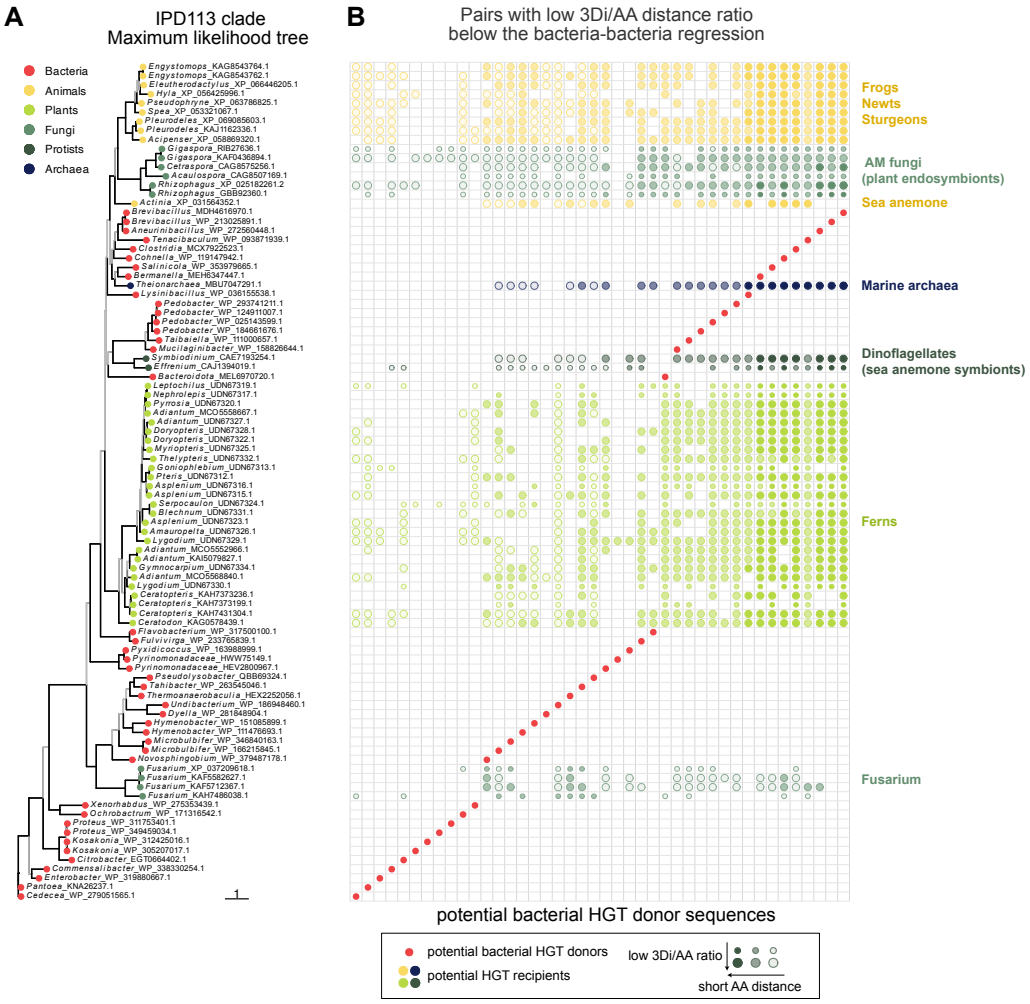

A
IPD113 clade
Maximum likelihood tree
Bacteria
Animals
Plants
Fungi
Protists
Archaea
B
Pairs with low 3Di/AA distance ratio
below the bacteria-bacteria regression
Frogs
Newts
Sturgeons
AM fungi
(plant endosymbionts)
Sea anemone
Marine archaea
Dinoflagellates
(sea anemone symbionts)
Ferns
Fusarium
potential bacterial HGT donor sequences
potential bacterial HGT donors
potential HGT recipients
low 3Di/AA ratio
short AA distance

Fig. S1

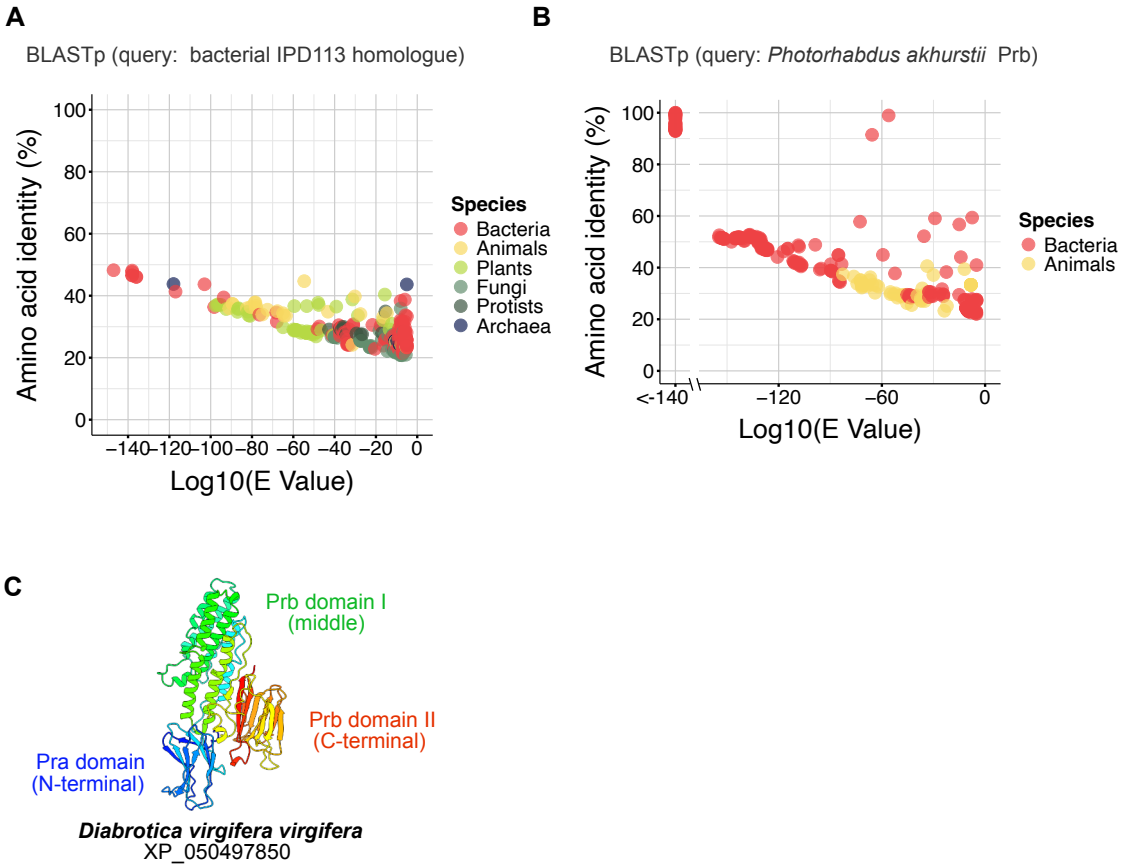
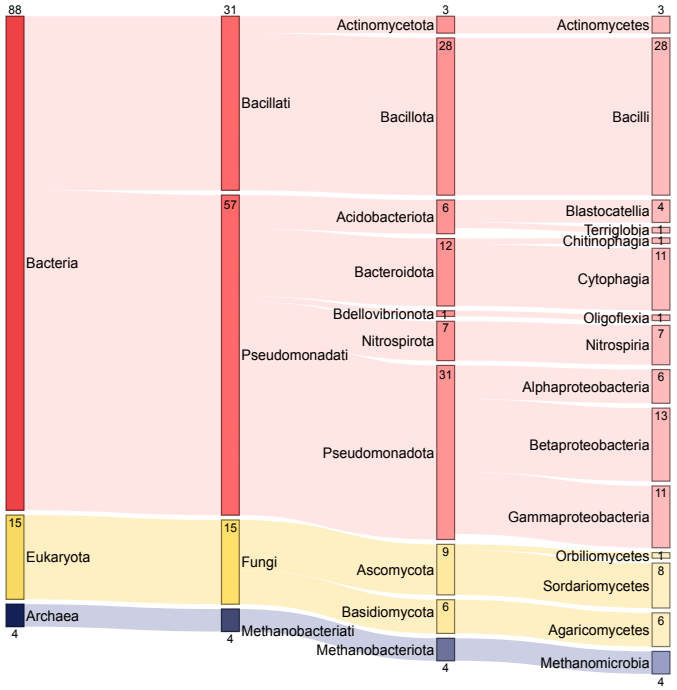

A
BLASTp (query: bacterial IPD113 homologue)
Amino acid identity (%)
100
80
60
40
20
0
−140 −120 −100 −80 −60 −40 −20 0
Log10(E Value)
Species
Bacteria
Animals
Plants
Fungi
Protists
Archaea
B
BLASTp (query: *Photorhabdus akhurstii* Prb)
Amino acid identity (%)
100
80
60
40
20
0
<-140
−120
−60
0
Log10(E Value)
Species
Bacteria
Animals
C
Prb domain I
(middle)
Prb domain II
(C-terminal)
Pra domain
(N-terminal)
Diabrotica virgifera virgifera
XP_050497850

Fig. S2

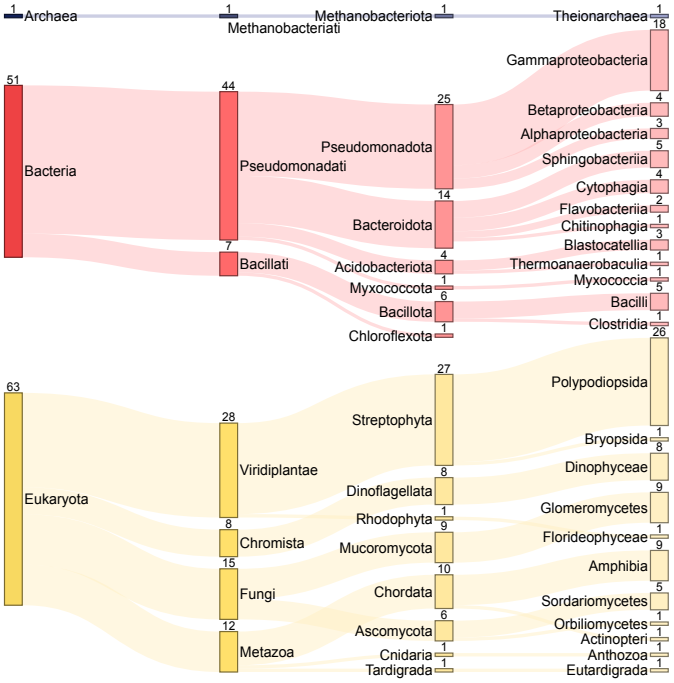
A

2D/3D-mixed clade



B

IPD113 clade



C

Prb clade

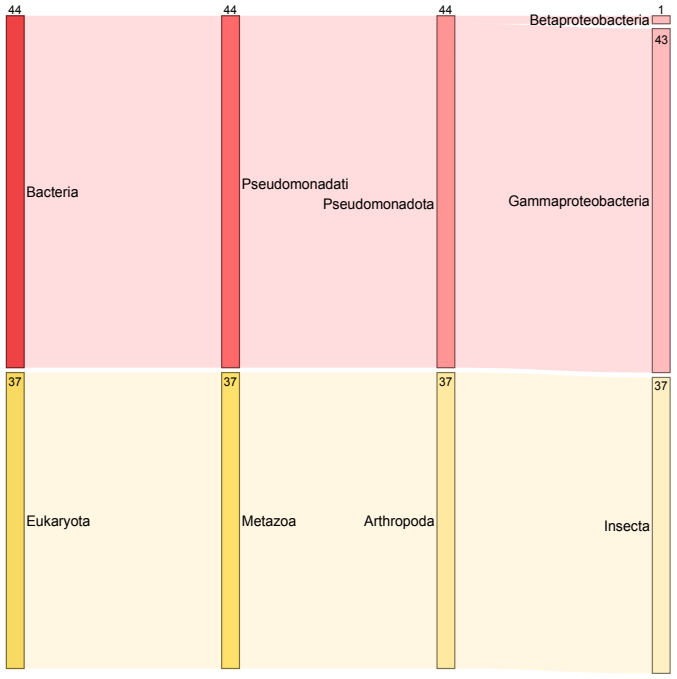
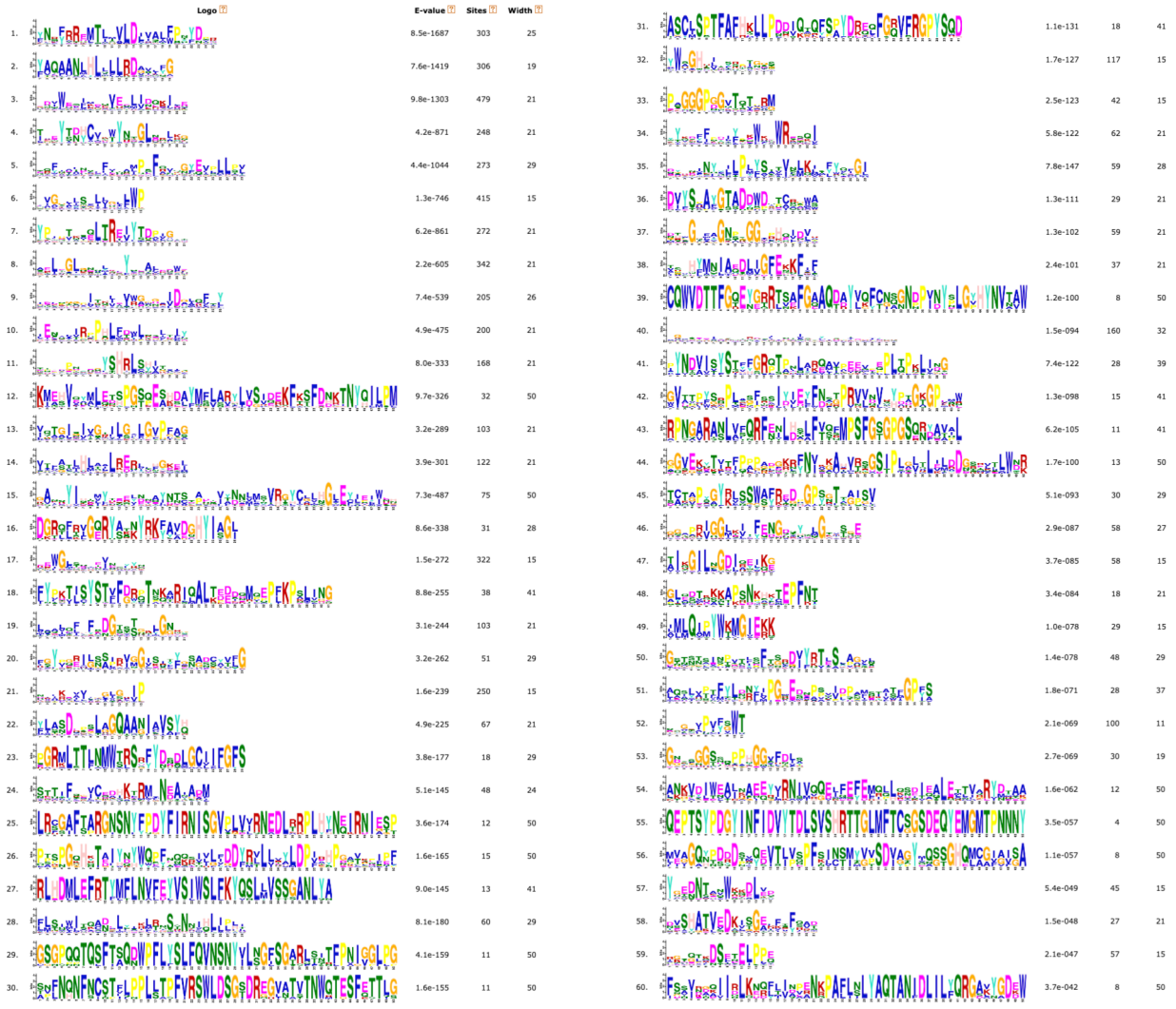


Fig. S3

A Cry-like 2D proteins



B DomainIII/Pra/insect Pra-like domain motif

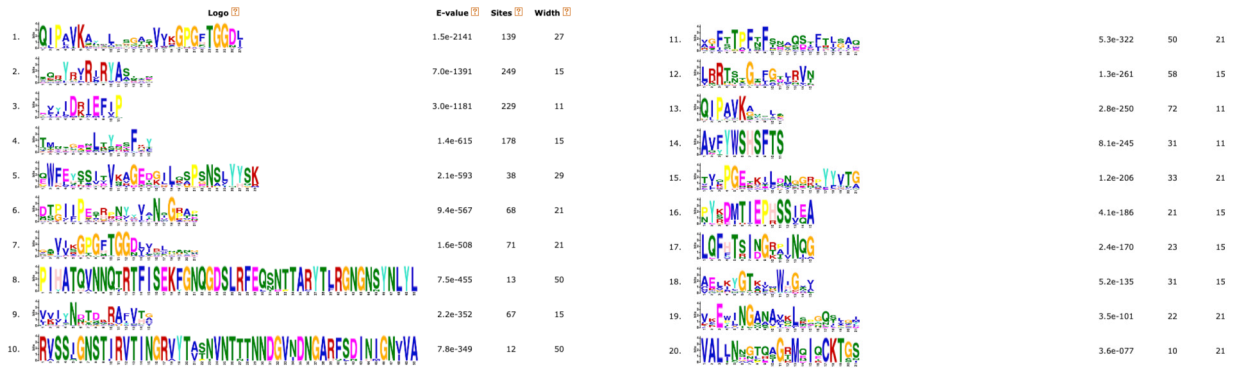


Fig. S4

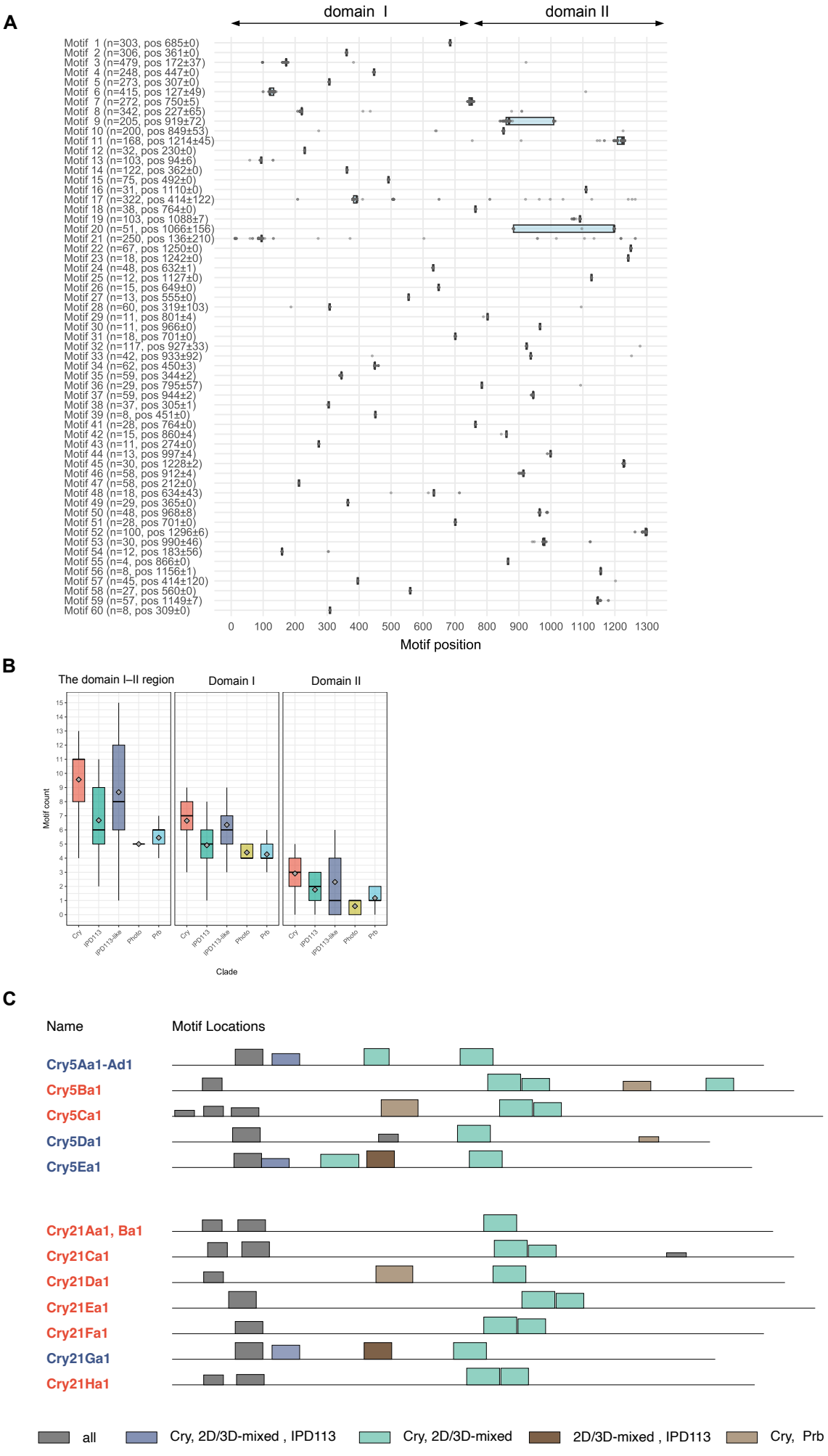

A
domain I
domain II
Motif 1 (n=303, pos 685±0)
Motif 2 (n=306, pos 361±0)
Motif 3 (n=479, pos 172±37)
Motif 4 (n=248, pos 447±0)
Motif 5 (n=273, pos 307±0)
Motif 6 (n=415, pos 127±49)
Motif 7 (n=272, pos 750±5)
Motif 8 (n=342, pos 227±65)
Motif 9 (n=205, pos 919±72)
Motif 10 (n=200, pos 849±53)
Motif 11 (n=168, pos 1214±45)
Motif 12 (n=32, pos 230±0)
Motif 13 (n=103, pos 94±6)
Motif 14 (n=122, pos 362±0)
Motif 15 (n=75, pos 492±0)
Motif 16 (n=31, pos 1110±0)
Motif 17 (n=322, pos 414±122)
Motif 18 (n=38, pos 764±0)
Motif 19 (n=103, pos 1088±7)
Motif 20 (n=51, pos 1066±156)
Motif 21 (n=250, pos 136±210)
Motif 22 (n=67, pos 1250±0)
Motif 23 (n=18, pos 1242±0)
Motif 24 (n=48, pos 632±1)
Motif 25 (n=12, pos 1127±0)
Motif 26 (n=15, pos 649±0)
Motif 27 (n=13, pos 555±0)
Motif 28 (n=60, pos 319±103)
Motif 29 (n=11, pos 801±4)
Motif 30 (n=11, pos 966±0)
Motif 31 (n=18, pos 701±0)
Motif 32 (n=117, pos 927±33)
Motif 33 (n=42, pos 933±92)
Motif 34 (n=62, pos 450±3)
Motif 35 (n=59, pos 344±2)
Motif 36 (n=29, pos 795±57)
Motif 37 (n=59, pos 944±2)
Motif 38 (n=37, pos 305±1)
Motif 39 (n=8, pos 451±0)
Motif 41 (n=28, pos 764±0)
Motif 42 (n=15, pos 860±4)
Motif 43 (n=11, pos 274±0)
Motif 44 (n=13, pos 997±4)
Motif 45 (n=30, pos 1228±2)
Motif 46 (n=58, pos 912±4)
Motif 47 (n=58, pos 212±0)
Motif 48 (n=18, pos 634±43)
Motif 49 (n=29, pos 365±0)
Motif 50 (n=48, pos 968±8)
Motif 51 (n=28, pos 701±0)
Motif 52 (n=100, pos 1296±6)
Motif 53 (n=30, pos 990±46)
Motif 54 (n=12, pos 183±56)
Motif 55 (n=4, pos 866±0)
Motif 56 (n=8, pos 1156±1)
Motif 57 (n=45, pos 414±120)
Motif 58 (n=27, pos 560±0)
Motif 59 (n=57, pos 1149±7)
Motif 60 (n=8, pos 309±0)
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300
Motif position
B
The domain I–II region
Domain I
Domain II
Motif count
Cry
IPD113
IPD113-like
Photo
Prb
Clade
C
Name
Motif Locations
Cry5Aa1-Ad1
Cry5Ba1
Cry5Ca1
Cry5Da1
Cry5Ea1
Cry21Aa1, Ba1
Cry21Ca1
Cry21Da1
Cry21Ea1
Cry21Fa1
Cry21Ga1
Cry21Ha1
all
Cry, 2D/3D-mixed , IPD113
Cry, 2D/3D-mixed
2D/3D-mixed , IPD113
Cry, Prb

Fig. S5

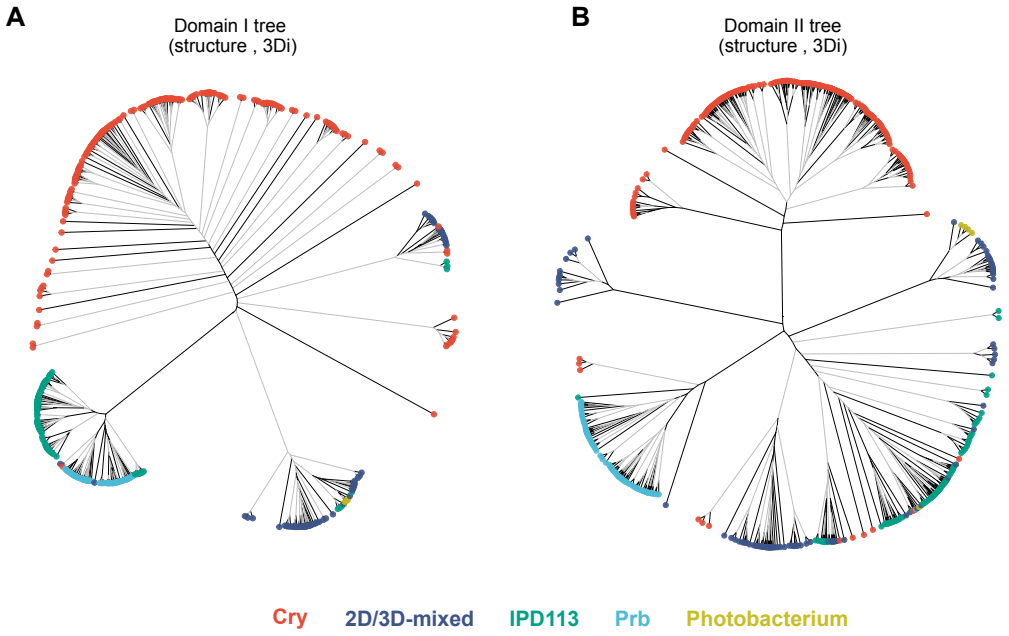

A
Domain I tree
(structure , 3Di)
B
Domain II tree
(structure , 3Di)
Cry
2D/3D-mixed
IPD113
Prb
Photobacterium

Fig. S6

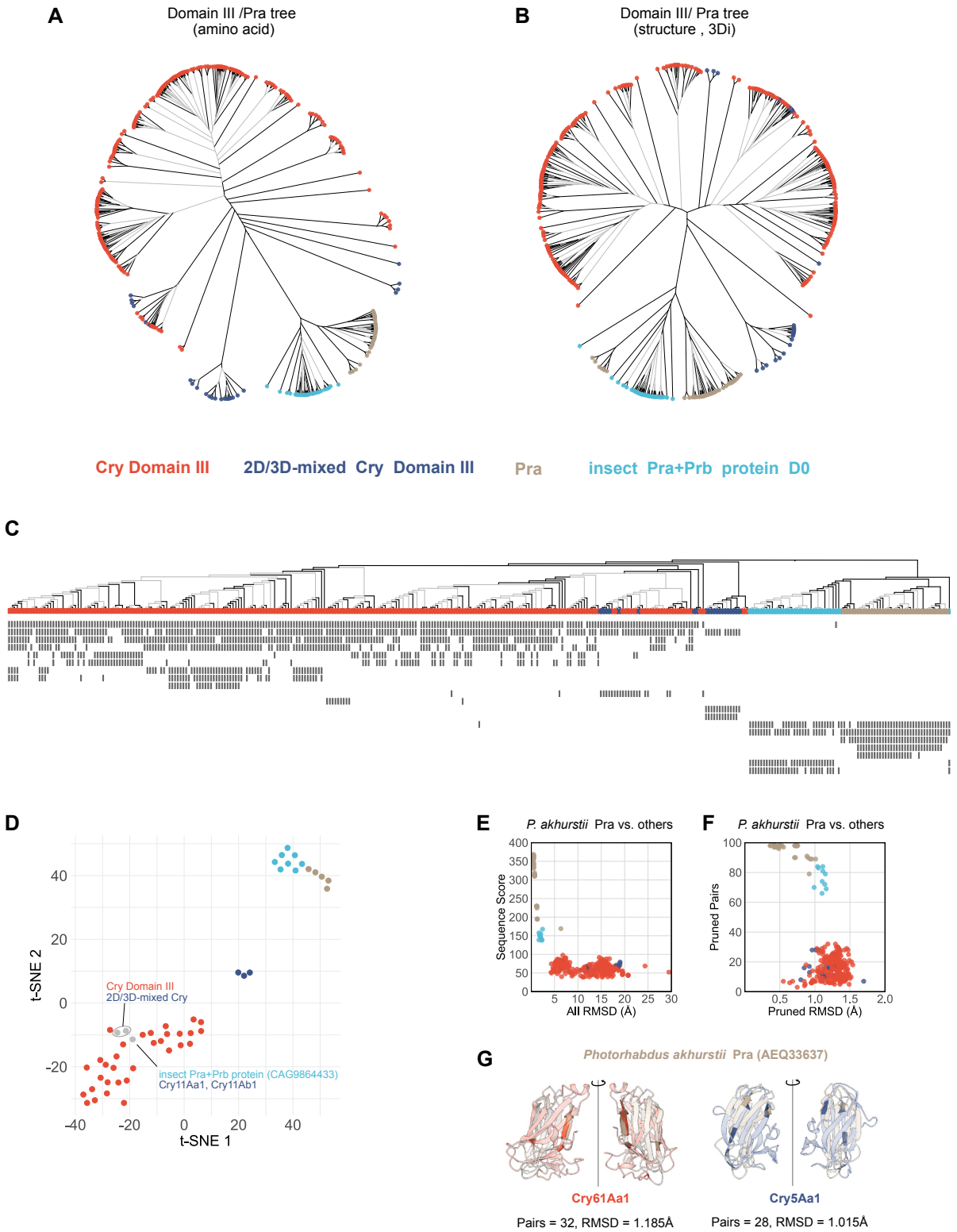

A
Domain III /Pra tree
(amino acid)
B
Domain III/ Pra tree
(structure , 3Di)
Cry Domain III
2D/3D-mixed Cry Domain III
Pra
insect Pra+Prb protein D0
C
D
t-SNE 2
t-SNE 1
Cry Domain III
2D/3D-mixed Cry
insect Pra+Prb protein (CAG9864433)
Cry11Aa1, Cry11Ab1
E
P. akhurstii Pra vs. others
Sequence Score
All RMSD (Å)
F
P. akhurstii Pra vs. others
Pruned Pairs
Pruned RMSD (Å)
G
Photorhabdus akhurstii Pra (AEQ33637)
Cry61Aa1
Pairs = 32, RMSD = 1.185Å
Cry5Aa1
Pairs = 28, RMSD = 1.015Å

Fig. S7

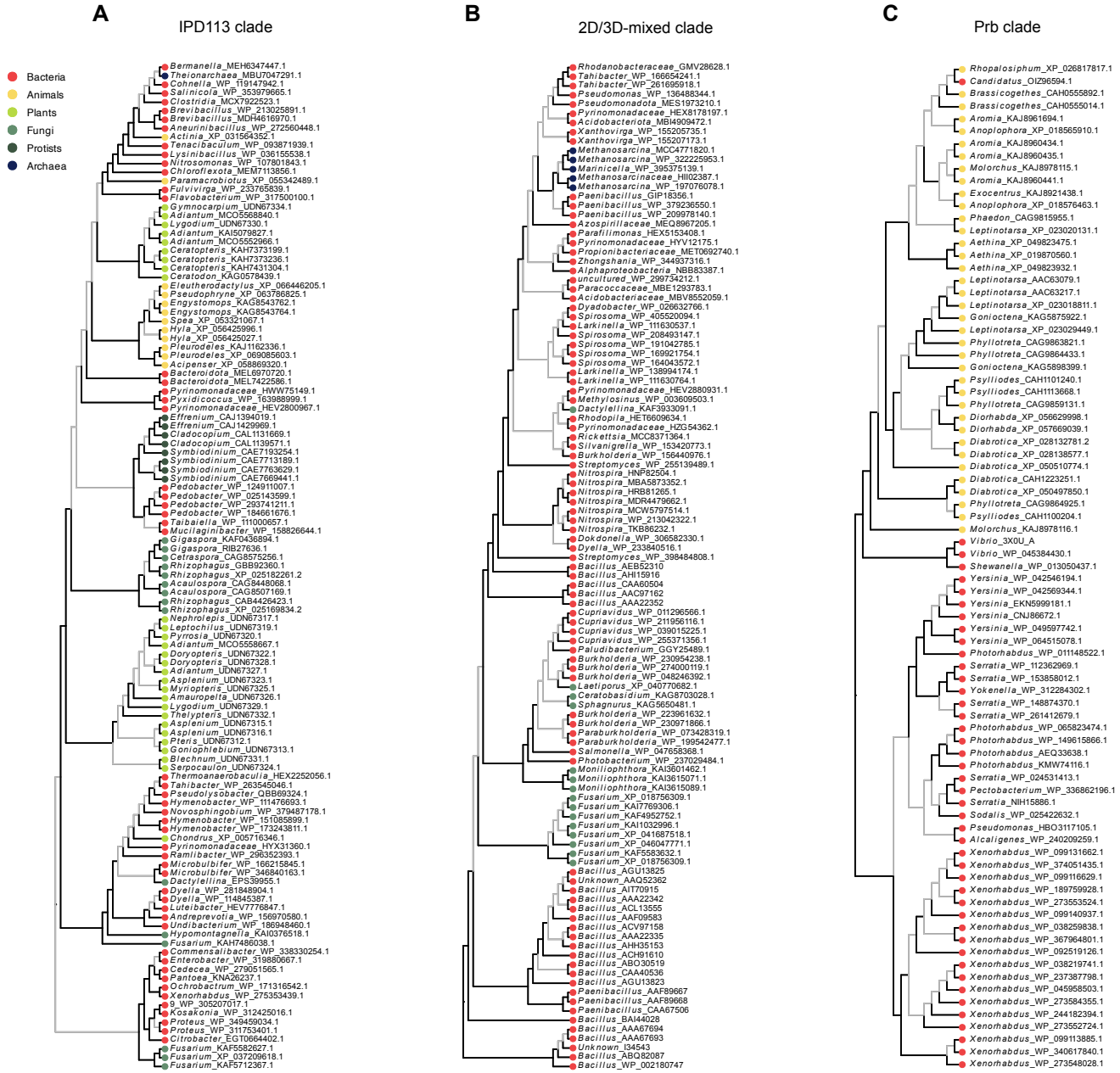

A
IPD113 clade
B
2D/3D-mixed clade
C
Prb clade
Bacteria
Animals
Plants
Fungi
Protists
Archaea

Fig. S8

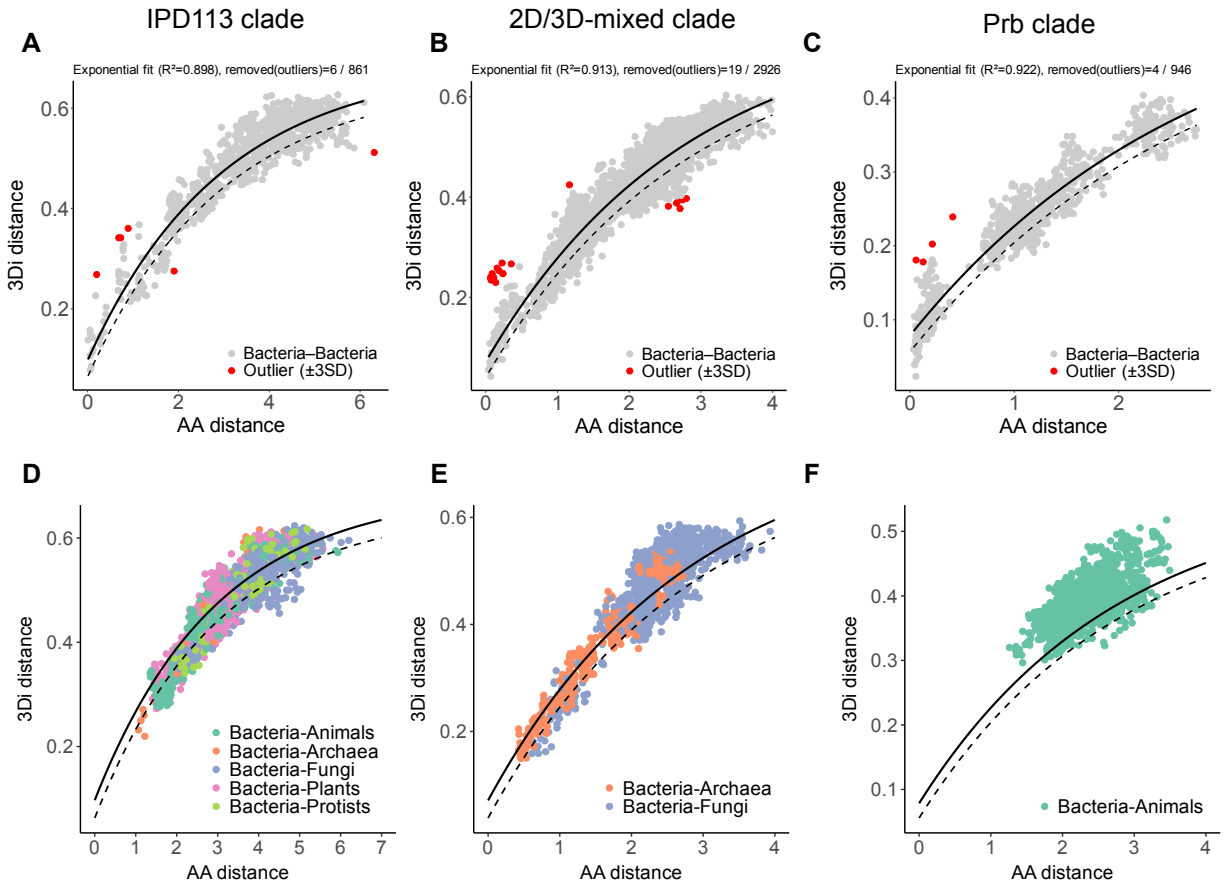
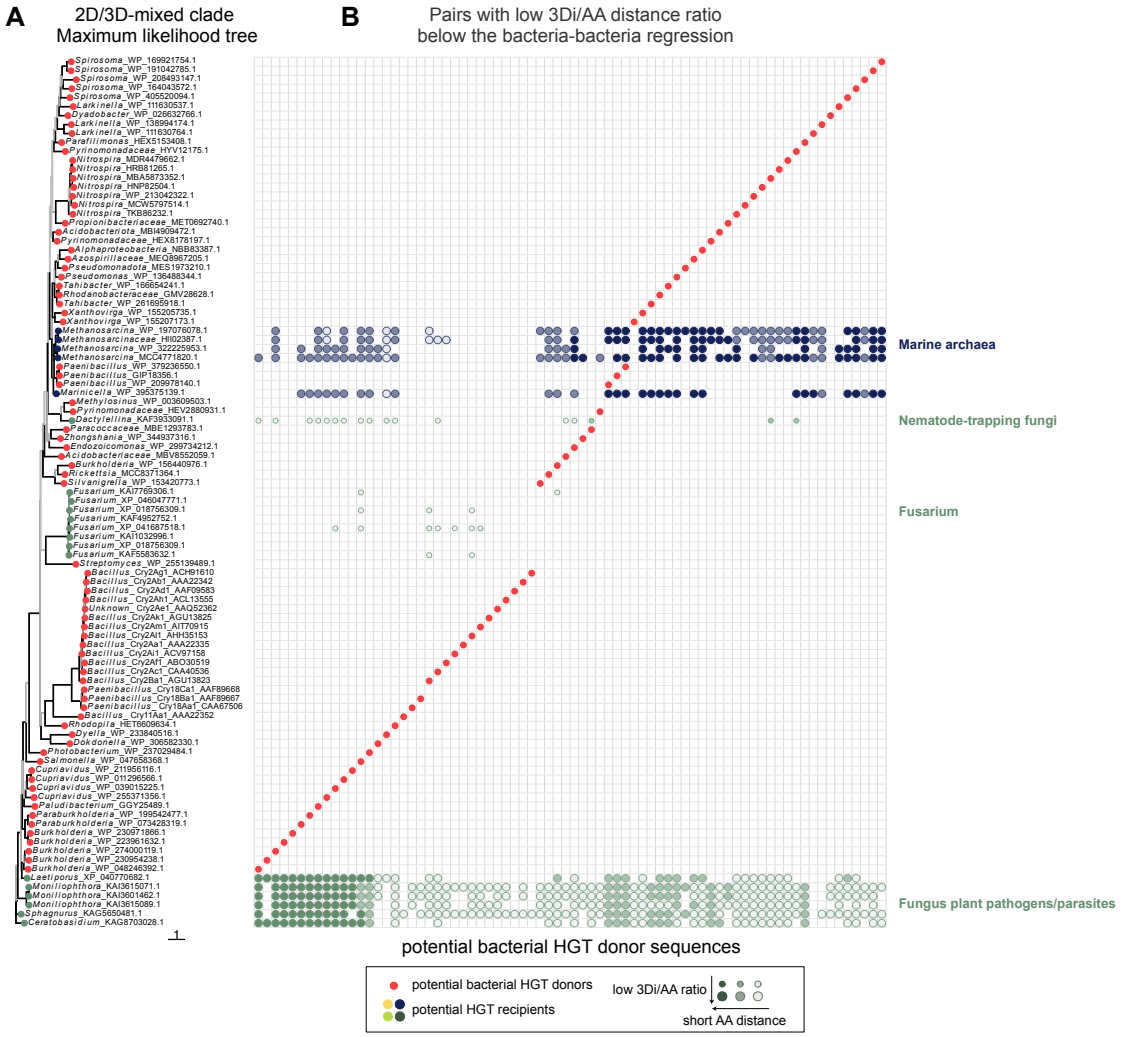
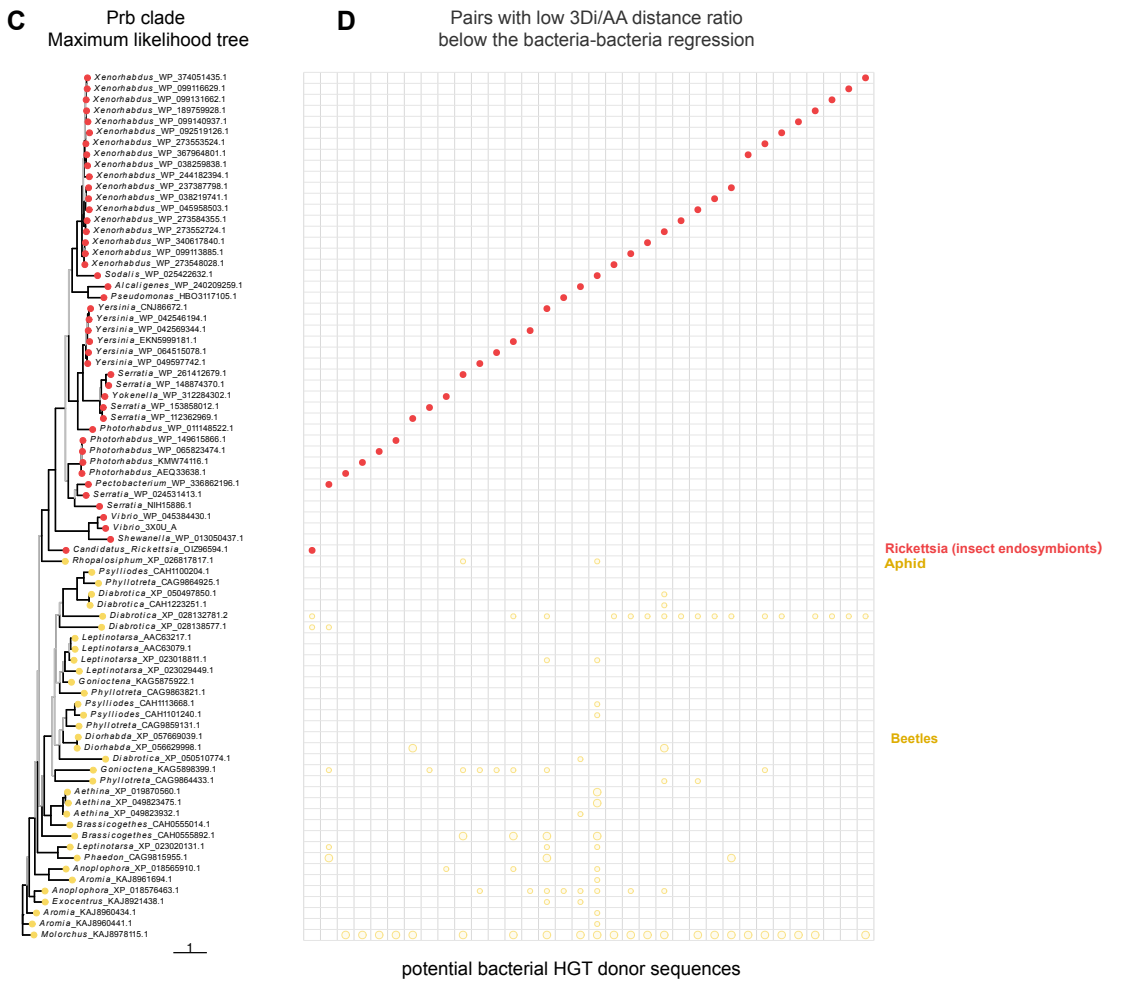

IPD113 clade
2D/3D-mixed clade
Prb clade
A
B
C
D
E
F
Exponential fit (R²=0.898), removed(outliers)=6 / 861
Exponential fit (R²=0.913), removed(outliers)=19 / 2926
Exponential fit (R²=0.922), removed(outliers)=4 / 946
3Di distance
AA distance
Bacteria–Bacteria
Outlier (±3SD)
Bacteria-Animals
Bacteria-Archaea
Bacteria-Fungi
Bacteria-Plants
Bacteria-Protists

Fig. S9


A
2D/3D-mixed clade
Maximum likelihood tree
B
Pairs with low 3Di/AA distance ratio
below the bacteria-bacteria regression
Bacteria
Fungi
Archaea
Marine archaea
Nematode-trapping fungi
Fusarium
Fungus plant pathogens/parasites
potential bacterial HGT donor sequences
potential bacterial HGT donors
potential HGT recipients
low 3Di/AA ratio
short AA distance


C
Prb clade
Maximum likelihood tree
D
Pairs with low 3Di/AA distance ratio
below the bacteria-bacteria regression
Bacteria
Animals
Rickettsia (insect endosymbionts)
Aphid
Beetles
potential bacterial HGT donor sequences