Bioinformatic discovery of type 11 secretion system (T11SS) cargo across the Proteobacteria

Authors: Alex S. Grossman[1,2], Nicholas C. Mucci[1], Heidi Goodrich-Blair[1,*]

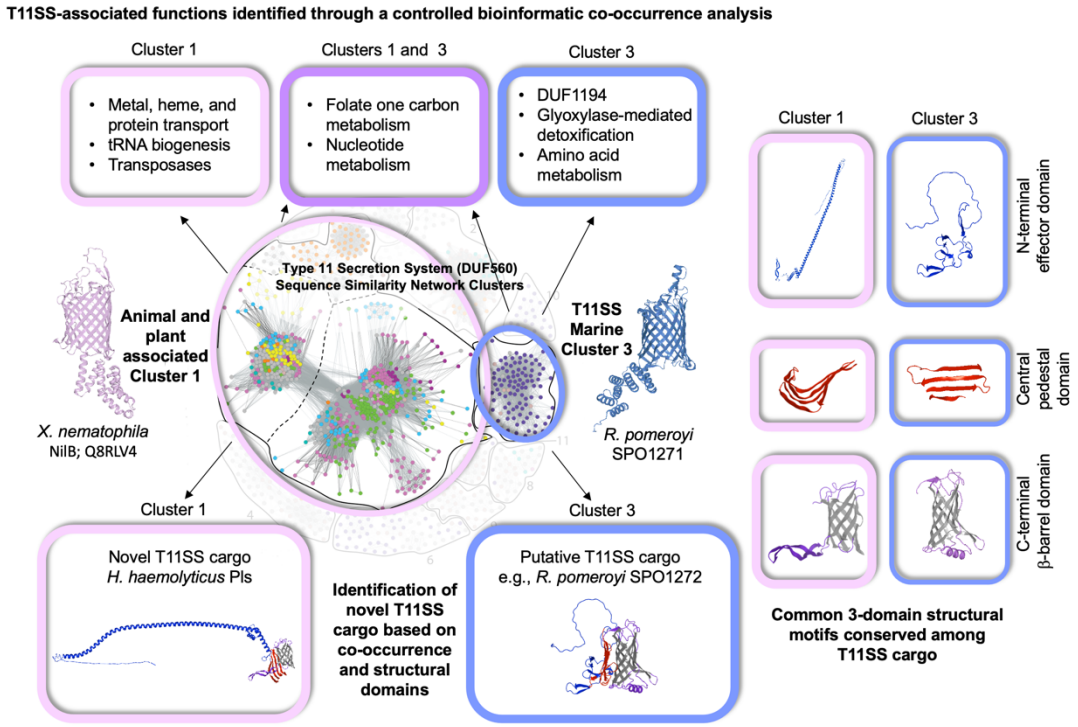
[1]Department of Microbiology, University of Tennessee, Knoxville, Knoxville, TN 37996-0845

[2]Current address: The ADA Forsyth Institute, 245 First St, Cambridge, MA 02142

[*]To whom correspondence should be addressed: hgblair@utk.edu

Keywords: Co-occurrence; T11SS, Host-association; Cargo detection; Plasmin-sensitive protein; DUF1194.

Graphical abstract:



Abstract:

Type 11 secretion systems (T11SS) are broadly distributed among proteobacteria, with more than 3000 T11SS family outer membrane proteins (OMPs) comprising 10 major sequence similarity network (SSN) clusters. Of these, only 7, all from animal-associated cluster 1, have been experimentally verified as secretins of cargo, including adhesins, hemophores, and metal binding proteins. To identify novel cargo of a more diverse set of T11SS, we identified gene families co-occurring in gene neighborhoods with either cluster 1 or marine microbe-associated cluster 3 T11SS OMP genes. We developed bioinformatic controls to ensure perceived co-occurrences are specific to T11SS, and not general to OMPs. We found that both cluster 1 and cluster 3 T11SS OMPs frequently co-occur with single carbon metabolism and nucleotide synthesis pathways, but that only cluster 1 T11SS OMPs had significant co-occurrence with metal and heme pathways, as well as with mobile genetic islands, potentially indicating diversified function of this cluster. Cluster 1 T11SS co-occurrences included 2556 predicted cargo proteins, unified by the presence of a C-terminal β-barrel domain, which fall into 141 predicted UniRef50 clusters and approximately 10 different architectures: 4 similar to known cargo and 6 uncharacterized types. We experimentally demonstrate T11SS-dependent secretion of an uncharacterized cargo type with homology to Plasmin sensitive protein (Pls). Unexpectedly, genes encoding marine cluster 3 T11SS OMPs only rarely co-occurred with the C-terminal β-barrel domain and instead frequently co-occurred with DUF1194-containing genes. Overall, our results show that with sufficiently large-scale and controlled genomic data, T11SS-dependent cargo proteins can be accurately predicted.

Impact Statement: We describe a novel method for controlling genomic neighborhood analyses and use this controlled co-occurrence technique to examine two distinct clusters of genes from the recently described type 11 secretion system (T11SS): Cluster 1, predominantly encoded by host-associated microbes and Cluster 3, encoded by marine microbes. We found that both clusters of predicted T11SS frequently co-occur with single carbon metabolism and nucleotide synthesis pathways, but only the host-associated cluster co-occurred with iron uptake pathways. Using these datasets, we predicted 2687 T11SS dependent cargo with approximately ten unique architectures, six of which have not previously been linked to T11SS. Finally, we validate our results by demonstrating T11SS-dependent secretion of a cargo protein with one of the novel architectures, plasmin-sensitive surface protein Pls from *Haemophilus parahaemolyticus*.

Introduction

Identifying protein function remains a difficult task in the postgenomic age. As of the Pfam 2021 update v33.1, there were 4244 domains of unknown function (DUF) families, which constitutes around 23% of the database (1). This means roughly one quarter of the Pfam database remains uninvestigated, limiting our understanding of the breadth of biological processes occurring on Earth. Functional characterization of proteins with DUF domains enables their annotation and removal from the DUF list, improving our datasets permanently. *In silico* approaches applied to protein domains are an important component of this process, as they can focus attention on the likely functions of a particular protein family (2). Recently, a combination of *in silico* and experimental approaches has yielded insight into the DUF560 protein family of outer membrane proteins (OMPs), defining it as the type 11 bacterial secretion system (T11SS), represented by 10 distinct sequence-level clusters (3). T11SS homologs are present among a diverse array of Proteobacteria, including human pathogens in the *Neisseria*, *Haemophilus*, and *Moraxella* genera (4–6). In the gamma-proteobacterium *Xenorhabdus nematophila*, a T11SS system encoded within the

58 nematode intestinal localization locus (*nil*) is necessary for colonization of the mutualistic nematode
59 host *Steinernema carpocapsae* (7). In this system the T11SS, NilB, regulates the surface exposure of a
60 surface lipoprotein, NilC, both of which are necessary for colonization. Other known lipidated
61 T11SS-dependent cargo include co-receptors for metal-containing compounds heme, transferrin,
62 and lactoferrin (4–6,8,9). These cargo are composed of two types of domains, variable N-terminal
63 effector domains and 8-stranded C-terminal β-barrel domains (such as TbpBBD and Lipoprotein C).
64 T11SS can transport unlipidated, soluble cargo proteins across the outer membrane. For example,
65 T11SS-dependent hemophilin homologs have been characterized from *X. nematophila* (HrpC),
66 *Acinetobacter baumannii* (HphA), and *Haemophilus haemolyticus* (Hpl) (3,8,10,11). However, the majority
67 of T11SS OMPs do not have a known or predicted cognate cargo. A sequence similarity network
68 analysis suggested that experimentally characterized T11SS only capture a small portion of the
69 functional diversity present in this family (3).

70 The availability and accessibility of sequencing data have skyrocketed since the first bacterial
71 genome was sequenced in 1995, with over 200,000 bacterial and archaeal genomes currently
72 deposited on public databases (12). This tremendous repository of genomic sequence data allows
73 computational biologists to probe these genomes for patterns in genetic co-occurrence, as measured
74 through proximity to a gene of interest. Genes within a common functional pathway often cluster
75 together within a genome. Such patterns can be useful when forming hypotheses about the function
76 of uncharacterized proteins. Gene neighborhood studies analyze the genomic regions upstream and
77 downstream of any given gene of interest with the assumption that nearby genetic partners could
78 illuminate the role(s) of the unknown gene (13).

79 However, one flaw in the current methodology used to derive co-occurrence lists is the lack
80 of bioinformatic controls capable of removing non-specific co-occurring genes. For example,

81 commonly inherited, deeply ancestral domains, such as transcription and translation machinery,
82 often appear as the top result of a co-occurrence analysis, but do not yield useful information about
83 the specific query gene since their appearance is a coincidence of ubiquity. Also, meaningful results
84 can be obscured by co-occurrence based on structural or other common features present within the
85 query protein, such as those that target the gene product to specific cellular locations, which are
86 expected to interact commonly with the relevant protein trafficking machinery. Herein, we use
87 computational approaches to elucidate potential functional roles played by T11SS-dependent cargo
88 and predict novel cargo. Using these results, we then generated a co-expression system in *Escherichia*
89 *coli* to experimentally demonstrate the secretion of one novel, bioinformatically predicted cargo
90 family using the plasmin sensitive protein (Pls) from *Haemophilus parahaemolyticus* and the type eleven
91 Pls Secretor (TepS).

92 **Methods**

93 **Gene neighborhood analysis**

94 The T11SS genome neighborhoods (±6 genes around a T11SS gene) from cluster 1 (1111
95 queries) and cluster 3 (145 queries) (3) were analyzed. T11SS genome neighborhoods were generated
96 through Rapid ORF Detection & Evaluation Online (RODEO) (14). As described in the main text,
97 a control was developed to identify and remove non-specific genes from among those flagged as co-
98 inherited or syntenic with T11SS (Fig. 2). The primary control list was generated by extracting all Go
99 term outer membrane proteins from proteobacteria within 1 amino acid of the median T11SS length
100 and removing all DUF560 containing proteins. For both T11SS cluster analyses a random subset of
101 control sequences were chosen equal to the number of queries being analyzed. Controlled protein
102 sequences were downloaded from UniProt (15) in February 2021. Co-occurring protein domains
103 were annotated and summed up to determine frequency among queries and controls. False

104 Discovery Rate was calculated by dividing frequency among T11SS queries and control sequences.
105 All domains with and FDR > 0.1 were excluded from analysis. For cluster 1 all domains with fewer
106 than 25 occurrences were excluded. For cluster 3 all domains with fewer than 5 co-occurrences were
107 excluded.

108 The Joint Genome Institute Integrated Microbial Genomes and Microbiomes platform
109 (https://img.jgi.doe.gov/) (16,17) was used to examine genomic regions surrounding the T11SS
110 OMPs-encoding genes of select cluster 3 bacteria. Briefly, the "find genes" function was used to
111 identify T11SS OMP homologs from selected representatives of the Roseobacteriaceae and
112 Rhodobacteraceae. The "gene neighborhoods" function was then used to visualize the regions
113 surrounding the T11SS OMP homologs and to identify syntenic loci (see Neighborhood Alignment
114 Tab of Supplemental File 3 for an example). Predicted structures were downloaded as .pdb files
115 from AlphaFold 2.0 (18,19) except in one instance (*Amylibacter cionae* CGMCC 1.15880
116 IEX10_RS12250 for which a Phyre 2.0 was used to predict the structure, and a .pdb file of the top ranked
117 prediction was used for Fig. 8G (18,20). This prediction was ranked with 98.51 confidence only over
118 16% identity over the C-terminal portion of the predicted protein, on the crystal structure of
119 lactoferrin binding protein b (LbpB) template 7jrd. Using each .pdb file, structural motifs were
120 highlighted using Protean 3D®, Version 17.4.3, DNASTAR, Madison, WI.

121

122 **Functional and structural analyses of co-occurring gene neighborhood domains**

123 Protein sequences that contain domains that co-occur with T11SS from the gene
124 neighborhood analysis were obtained via UniProt (15). KofamKOALA (21) and BlastKOALA (22)
125 were used to assess function. KofamKOALA assigns KEGG orthology to user sequence data by
126 HMMER/HMMSEARCH against the KEGG database. BlastKOALA is similar but uses a BLAST

127 search to assign KEGG orthology. Genes were matched to brite hierarchies to assess general cellular
128 function and KEGG pathway to assess potential interactions.

129 To identify potential cargo proteins all downloaded protein sequences were matched to
130 UniRef50 clusters to compress potential homologs (23). The reference sequences from these
131 UniRef50 clusters where then subjected to a BLAST search against known T11SS cargo β-barrel and
132 handle domains (nematode intestine localization protein (NilC) in *Xenorhabdus nematophila*, transferrin
133 binding protein (TbpB) N-terminal and C-terminal in *Xenorhabdus nematophila*, heme receptor protein
134 (HrpC) in *Xenorhabdus nematophila*, Cobalt receptor protein (CrpC) in *Xenorhabdus cabanillasii*,
135 Hemophilin (Hpl) in *Haemophilus haemolyticus*, hemophilin (HphA) in *Acinetobacter baumanii*, factor H
136 binding protein (fHbp) *Neisseria meningitidis*, and lactoferrin binding protein (LbpB) N-terminal and
137 C-terminal in *Neisseria meningitidis* to search for homologs of T11SS cargo with a low stringency E-
138 value filter of <0.05 in order to find even distantly related homologs, generating a list of 141
139 potential cargo proteins from cluster 1 and a single potential cargo from cluster 3. MSA alignment
140 was used to group the predicted cargo proteins into distinct architectures. 10 groups emerged,
141 TbpB-like, LbpB-like, Hemophilin-like 1 and 2, fHbp-like, Pls-like, highly disordered N-terminus 1
142 and 2, *Psychrobacter* surface protein, and *Sphingomonas* surface protein. Structural data predictions for
143 the 9 architectures which contained more than one sequence were generated through PSIPRED (24)
144 for secondary structure and AlphaFold2 (18,19) for tertiary structure.

145

146 **Co-expression of Pls and its cognate T11SS protein TepS**

147 All cultures were grown in glucose minimal media supplemented with 1% LB. Plate based
148 cultures were grown on glucose minimal plates (25). For plasmid-based expression, chemically
149 competent *Escherichia coli* strain BL21-DE3 (C43) were chosen for ease of transformation and their

150 ability to non-toxically express membrane proteins (26,27). Strains of *E coli* were grown at 37°C.
151 Where appropriate media was supplemented with ampicillin at a concentration of 150 μg/ml.
152 Protein expression was induced at the midlog point of bacterial growth via addition of isopropyl β-
153 d-1-thiogalactopyranoside (henceforth IPTG) at a concentration of 0.5mM.
154 Using sequences from *H. parahaemolyticus* strain NCTC10794, GenScript synthesized genes
155 for Plasmin sensitive surface protein (Pls-STO63610.1) with an attached C-terminal FLAG tag and
156 its genomically associated T11SS OMP (TepS-STO63613.1) with an N-terminal 6xHis tag. Both
157 genes were codon optimized for expression in *E. coli* Bl21DE3. GenScript cloned Pls-FLAG into
158 multiple cloning site 1 of pETDuet-1 and 6xHis-TepS into multiple cloning site 2, resulting in
159 expression plasmids pETDuet-1/Pls-FLAG and pETDuet-1/6xHis-TepS/Pls-FLAG. Expression
160 plasmids were transformed into *E. coli* via electroporation. Strains were grown in defined medium
161 with 150 μg/ml ampicillin (28). Bacteria were subcultured into 100 ml of broth at an initial optical
162 density at 600 nm (OD_{600}) of 0.1, grown for 6 h at 37°C to reach late logarithmic growth, and
163 induced with 500 μM isopropyl-β-d-thiogalactopyranoside (IPTG) for 1 hour. To monitor surface
164 exposure of Pls whole cells normalized to an OD600 of 1, rinsed 3x, and spotted in technical
165 triplicate on nitrocellulose membranes. The remaining cells were lysed open via 30 seconds of
166 sonication at ~500 root mean square volts (V_{rms}). Cell lysates were also spotted in technical triplicate
167 on nitrocellulose membranes to enumerate total cellular Pls. To monitor extracellular secretion of
168 Pls supernatant was collected from cultures after induction and filtered sterilized to remove all cells.
169 Supernatant samples were ultracentrifuged at 150,000 relative centrifugal force (RCF) for 3 h to
170 separate the soluble fraction from the insoluble component, composed of cell membrane
171 components and membrane vesicles. The membrane vesicle fraction was suspended in protein
172 sample buffer. 700 μl of each soluble supernatant fraction was precipitated via 10% trichloroacetic
173 acid (TCA) and the resulting pellet was suspended in protein sample buffer (29). Samples were

174 analyzed by 10% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and
175 immunoblotting. All dot blots and Western blots were probed with rat anti-FLAG primary antibody
176 and goat anti-rat secondary antibody conjugated to a 680nm fluorophore. Intensities were recorded
177 for FLAG reactive bands. For every supernatant sample, a band from the Coomassie blue-stained
178 gel was used as a loading control to normalize intensities of supernatant samples prior to analysis.
179 Fold change of secretion was determined by dividing the amount of supernatant Pls detected in the
180 Pls/TepS co-expression treatment by the amount detected in the Pls alone treatment. A Tukey's
181 honestly significant difference (HSD) test was used for comparing fold change of secretion (30).

182 **Results:**

183 **Establishing a control for domain of unknown function (DUF) gene neighborhood co-**
184 **occurrence analysis**

185 Characterized T11SS-cargo pairs encoded by Proteobacteria typically have conserved roles in
186 host-metal acquisition and immune evasion (3,6,8,10). Only a small portion of predicted T11SS-
187 cargo pairs has been experimentally characterized, leaving open the possibility that as-yet
188 uncharacterized systems may be participating in diverse symbiotic and metabolic pathways. To gain
189 insights into the potential breadth of biological processes in which T11SS-cargo pairs may function,
190 we collected all T11SS sequences identified as part of the largest T11SS sequence similarity network
191 cluster described in Grossman *et al.,* 2021 (cluster 1, containing 1111 genes) and utilized the Rapid
192 ORF Description & Evaluation Online (RODEO) software to identify all protein domains encoded
193 within six open reading frames upstream or downstream of the query genes, resulting in a data set of
194 1504 domains, co-occurring between 2 and 928 times (Cluster1Co tab of Supplemental file 1) (14).
195 To avoid spurious correlations, we chose to exclude any domains that co-occurred fewer than 25
196 times (~2.7% of the maximum frequency co-occurrence). This number was arrived at by generating

197 a histogram of co-occurrence frequency and identifying the inflection point between the rare co-
198 occurring sequences which made up the bulk of the dataset (1331 or 88.5%) and those domains
199 whose co-occurrence frequency was greater than expected randomly (173 or 11.5%).

200 To ensure that the observed co-occurrences were specific to the T11SS protein family and
201 not due to common features of genomic loci encoding outer membrane proteins, we developed and
202 implemented a control process. Briefly, RODEO co-occurrence neighborhoods were generated
203 using randomly selected proteins with biophysical characteristics similar to our DUF560 query
204 proteins such as protein length and outer membrane localization (GO term: outer membrane
205 GO0019867), and then used as a point of comparison for gene neighborhoods from T11SS genes.
206 The control sequences obtained were limited to the Proteobacteria phylum, where T11SS
207 predominantly occur (3). Due to a left-skewed distribution, the median length of all DUF560
208 domain containing proteins in Pfam (accessed February 2021) was used to reduce bias from
209 pseudogenes and gene fragments (Fig. S1) (1). The median size of all predicted T11SS proteins was
210 481 amino acids, so we extracted all outer membrane ORFs between 480 and 482 amino acids in
211 size. Sequences that appeared in the T11SS database were removed from the control database to
212 ensure unbiased selection. A random subsample of this control database was taken equal to the
213 number of sequences in the original query database (ControlQueries tab of Supplemental file 1) (Fig.
214 1).

215 OMP sequences from the subsampled control dataset were then submitted as a query to the
216 genome neighborhood network function of RODEO using the same parameters as the T11SS query
217 list. The output of this control co-occurrence analysis was used to generate a false discovery rate
218 (FDR) for any given domain by dividing frequency of co-occurrence with random OMPs by
219 frequency of co-occurrences with T11SS genes. We then gated the dataset for domains with an FDR

220 less than 0.1 (10%) to exclude non-specific correlations, resulting in a final set of 51 significantly co-
221 occurring domains. (Cluster1FDR_Threshold tab of Supplemental file 1). The most common
222 significant co-occurring domains alongside cluster 1 T11SS OMPs were other DUF560 proteins
223 (928/1111), the TbpBBD β barrel domain (503/1111), the TonB C-terminal domains (364/1111),
224 heme oxygenase (137/1111), and DUF454 (118/1111). Next, all genes that co-occurred with a
225 cluster 1 T11SS gene and possessed one of the identified domains were extracted resulting in 3405
226 significantly co-occurring genes (Cluster1SigGenes tab of Supplemental file 1). These genes were
227 submitted to KofamKOALA and BlastKOALA in parallel for annotation into molecular pathways
228 and brite hierarchies (21,22).

229

230 **Functional assessment of T11SS co-occurring proteins identifies a role of iron uptake,**
231 **protein export, and single carbon metabolism**

232 We examined the co-occurrence gene set for those with annotations in the curated Kyoto
233 Encyclopedia of Genes and Genomes (KEGG) ortholog (KO) database (21,31). A hidden Markov
234 model search using KofamKOALA (HMMER/HMMSEARCH) was used to annotate the genes
235 according to the KO database. The software found matches for 1488 of the 3405 significant co-
236 occurring genes (Cluster1KofamBrite tab of Supplemental file 2). When matched to cellular
237 functions (brite hierarchies) the most common categories were transporters (periplasmic TonB,
238 ExbD, Signal peptidase II, TolA, etc.), enzymes (heme oxygenase, formate dehydrogenase,
239 exopolyphosphatase, dihydrofolate reductase, etc.), and tRNA biogenesis (tRNA modifying GTPase,
240 tryptophanyl-tRNA synthase, Ribonuclease P, aminoacyl tRNA synthase) (Fig. 2A). Additionally,
241 cluster 1 T11SS genes were found to significantly co-occur with transposase genes. Of the 1488
242 matched sequences, 259 had known pathway association, most commonly biosynthesis of secondary

243 metabolites (heme oxygenase, cobalamin-dependent methionine synthase, vitamin K biosynthesis
244 protein MenH, etc.), porphyrin metabolism (heme oxygenase), and protein export (signal peptidase
245 II and YidC insertase) (Fig. 2B) (Cluster1KofamPath tab of Supplemental file 2). The list of
246 pathways also included ribosome components (RpmB, RpmG, RpmH) and aminoacyl tRNA
247 biosynthesis (TyrS, TrpS). A high frequency of co-occurrence with transposases, ribosome
248 components, and tRNA synthesis genes is characteristic of pathogenicity islands, phage regions, and
249 other mobile genetic elements, since many mobile genetic elements use universally conserved
250 ribosome and tRNA genes as anchors for genomic insertion (32). The frequency at which T11SS
251 genes are encoded within mobile genetic element islands suggests that T11SS may frequently be
252 horizontally acquired. Finally, the list included several carbohydrate biosynthesis pathways
253 (hsa00670, one carbon pool by folate; map00230, purine metabolism).

254 To further functionally annotate co-occurrences, we used BLAST alignment, rather than
255 hidden Markov searching, to establish sequence similarity using BlastKOALA, which compares
256 query sequences to the non-redundant KEGG dataset (22). The software found matches for 1387 of
257 the 3405 significantly co-occurring genes. Categorization into cellular functions (brite hierarchies)
258 again revealed the largest categories to be enzymes, transporters, and tRNA biogenesis (Fig. S2A)
259 (Cluster1BlastBrite tab of Supplemental file 2). BlastKOALA identified more co-occurring amino
260 acid importers (arginine, lysine, tyrosine, tryptophan, and serine transporters) in the dataset than
261 KofamKOALA. Of the 1387 identified genes, 260 had known pathways. This analysis identified
262 similar pathways to those identified by KofamKOALA, including secondary metabolite biosynthesis,
263 protein export, porphyrin metabolism, and carbohydrate biosynthesis (Fig. S2B) (Cluster1BlastPath
264 tab of Supplemental file 2).

265

266 **Predicted structure analysis of putative cluster 1 T11SS cargo reveals variable N-termini and**
267 **shared C-terminal domains**

268 To bioinformatically identify potential cluster 1 T11SS-dependent cargo we searched co-
269 occurring proteins for homology to domains of known cargo. The 3405 significantly cluster 1 co-
270 occurring sequences were first consolidated using UniProt ID mapper into 666 UniRef50 clusters
271 with 50% or greater identity at an amino acid level (Cluster1Uniref tab of Supplemental file 1) (23).
272 The reference sequences of these UniRef50 were next searched using BLASTp for sequence
273 similarity with N-terminal domains and C-terminal β-barrel domains from experimentally verified
274 T11SS cargo proteins: NilC, TbpB, LbpB, HrpC, CrpC, Hpl, and Factor H binding protein, with a
275 low stringency E-value cutoff of 0.05 (3,5,6). No matches were found for NilC or the N-terminal
276 ligand binding domain of fHbp, but all other queries matched at least one cluster. This analysis
277 revealed 141 clusters of potential T11SS-dependent cargo proteins representing 2656 ORFs across
278 1048 species/taxa (Cluster1PutativeCargo tab of Supplemental file 1). Each sequence was also
279 annotated with SignalP 6 (33,34) which predicted 81 Sec secreted lipoproteins, 33 Sec secreted
280 soluble proteins, and 27 proteins with no predicted signal peptide. Several of the proteins with no
281 signal peptide were annotated as fragments or pseudogenes, supporting the previous observations of
282 T11SS transporting Sec dependent lipoproteins and soluble proteins.

283 We utilized multiple sequence alignment to separate the predicted cargo protein clusters into
284 distinct, annotated architectures (Cluster1PutativeCargo tab of Supplemental file 1). This analysis
285 identified two distinctive groups of Hemophilin-like proteins based on homology to known
286 hemophilins (Hpl, HrpC, HsmA), the slightly larger and predominantly soluble group 1 (96
287 lipoproteins, 912 soluble) and the predominantly lipidated group 2 (65 lipoproteins, 7 soluble), a
288 group of TbpB-like proteins (638 lipoproteins, 7 soluble), a group of proteins annotated as fHbp-

289 like proteins by virtue of their Lipoprotein C domain (54 lipoproteins, 285 soluble), and a group of
290 lactoferrin-binding protein-like proteins (105 lipoproteins). In addition, lipoproteins with an
291 elongated, disordered N-terminal region occurred as two groups: disordered N-terminus 1 (196
292 lipoproteins, 8 soluble) and 2 (1 lipoprotein only). Another group had an elongated N-terminal
293 domain with repeat rich (EEARKA motif) α-helical regions homologous to the Pls surface
294 glycoprotein in *Staphylococcus aureus*, termed Pls-like proteins (173 lipoproteins, 20 soluble). Two
295 phylogenetically restricted architectures were observed with multiple TbpBBD domains,
296 *Sphingomonas* surface proteins, which have an N-terminal sequence of proline-threonine repeats (18
297 lipoproteins, 12 soluble) and *Psychrobacter* surface proteins which have an N-terminal sequence of
298 ~12 proline-threonine-aspartic acid repeats (PTD motif) (55 lipoproteins, 4 soluble). Proline-
299 threonine repeats occur in *Xanthomonas campestris* endoglucanase, where they provide elasticity and
300 facilitate enzyme positioning around its substrate (35). The PTD motif is found in a family of
301 *Gracilibacteria* (BD1-5) surface proteins and in the eukaryotic microbe *Plasmodium falciparum* and are
302 hypothesized to act as adhesins facilitating interactions with hosts (36,37).

303 Using PsiPred 4.0 (24) and AlphaFold2 (18) we predicted the secondary and tertiary
304 structures of representative proteins from the largest UniRef50 cluster from each architecture with
305 more than one representative (Fig. 3 and Fig. 4). All the architectures were predicted to have at least
306 one domain, generally at the C-terminus, composed of β-strands linked by disordered loops that
307 form a C-terminal β-barrel domain similar to the TbpBBD domain or the Lipoprotein C domain.
308 The N-terminal domains of most cargo had β-strands predicted to form handle domains like those
309 seen in hemophilin or TbpB (3,5). When present, the disordered regions and long α-helical repeat
310 regions preceded this handle domain.

311 When examining secondary and tertiary structures in parallel we noticed that the predicted
312 cargo form three distinctive structural regions: a common C-terminal 8-stranded β-barrel domain, a
313 common central region formed by β-strands, which we have termed a "pedestal" region based on its
314 positioning as a platform for a variable N-terminal domain "effector" domain, predicted to serve the
315 functional role specific to each cargo protein type. Using solved crystal structures where available
316 (4,8,38), and Alphafold2 or Phyre 2.0 models for representative cargo proteins (18,20), we divided
317 the predicted cargo structures into these three regions and categorized them by recurrent patterns
318 (Fig. 5). The predicted central pedestal-domains fell into six visibly distinguishable categories; 4-
319 stranded groove, wing, and sheet, 5-stranded sheet, and 6-stranded/extended sheets (Fig. 5). None
320 of the examined pedestal regions featured complex, variable loops, suggesting they may play a
321 structural role (Fig. 5). In contrast, the C-terminal 8-stranded β-barrel core structures are
322 embellished within each protein by diverse inter-strand loops including disordered regions, extended
323 β-strands, and α-helices that could be playing functional roles, as do the loops of the N-lobe
324 TbpBBD domain of TbpB (39).

325 The N-terminal effector domains displayed a variety of structural motifs including
326 disordered loops, α-helices, β-strands, and β-barrels (Fig. 5). The examined effector regions ranged
327 in size, with some barely extending beyond the pedestal domain and others including duplications of
328 the 8-stranded β-barrel domain. The latter is typified by the solved structure of TbpB, which
329 comprises two lobes, each containing the hemophilin handle-β-barrel pair architecture. The TbpB
330 N-lobe pair comprises the effector region, while the C-lobe handle comprises the pedestal and the
331 C-lobe barrel is the putative T11SS targeting domain (38,39). This domain duplication and
332 functional specialization may also explain the architecture of *Sphingomonas* and *Psychrobacter* surface
333 proteins that feature 2-3 copies of the handle-barrel pair. However, the long, disordered loops
334 present in those proteins likely generate a distinctive structural architecture relative to TbpB.

335

336 **Demonstrating T11SS-dependent secretion of Plasmin sensitive surface protein**

337 To assess the accuracy of our bioinformatic predictions of T11SS cargo, we experimentally
338 tested levels of secretion in *E. coli* of a representative novel cargo protein with or without its
339 predicted cognate T11SS protein. Several of our predicted cargo possess a large α-helix repeat region
340 absent among known T11SS cargo. We chose to test a representative of this group, the 525 amino
341 acid Plasmin sensitive protein homolog (Pls) (STO63610.1) and its genomically associated putative
342 type eleven Pls secretor (TepS) (STO63613.1) from *H. parahaemolyticus* strain NCTC10794. The *H.*
343 *parahaemolyticus* Pls α-helix repeat region shares between 49 and 62% identity with the repeat regions
344 of surface glycoprotein Pls from *Staphylococcus aureus* (WP_256928426, WP_256927786,
345 WP_258808513, WP_256933640, and WP_257570445) that contributes to virulence by stimulating
346 biofilm formation (40). *Staphylococcus* and *H. parahaemolyticus* Pls differ in that homologs of the former
347 are larger and lack the C-terminal β-barrel domain.

348 To experimentally test for TepS-mediated secretion of *H. parahaemolyticus* Pls, we used
349 addition of a FLAG-tag epitope to allow Pls detection and expressed this protein with or without
350 TepS using pETDuet-1 based plasmids in *E. coli* BL21 DE3 C43. After protein expression
351 induction, anti-FLAG immunodotblots were performed on intact (to measure surface exposed
352 protein) and lysed cells (to measure total protein) to compare among treatments. Co-expression with
353 TepS did not significantly impact surface exposure of Pls or the intracellular concentration of Pls
354 (Fig. 6 and S3). This result was initially surprising, since Pls was predicted to be a surface anchored
355 lipoprotein. However, since *S. aureus* Pls can be cleaved from the cell surface spontaneously and by
356 proteinases such as plasmin (40,41) we next monitored free and extracellular-vesicle fractions of cell-
357 free supernatants for the presence of Pls via SDS-PAGE. These data revealed that Pls was present in

358 the supernatant, but only when co-expressed with the T11SS protein TepS. Some Pls was located in
359 extracellular vesicles independent of TepS expression, but this amount was higher in the presence of
360 TepS (Fig. 6 and Fig. S3). In the immunoblot, FLAG-tagged Pls is visible as a single clear band in
361 individual replicates, but varied in apparent size between biological replicates from 82.3 to 121.3 kD
362 with an average of 103kD. This is almost twice the value predicted from sequence alone (56.8kD).
363 This large and variable size may suggest that Pls is either being significantly slowed by glycan
364 moieties as it travels through polyacrylamide, or that it exists as a dimer when expressed in *E. coli*.

365

366 **Bioinformatic analysis of cluster 3 T11SS proteins from marine environments reveals distinct**
367 **co-occurrence patterns**

368 To test the robustness of our analysis we then utilized our co-occurrence technique to
369 explore cluster 3 T11SS proteins, predominantly containing sequences from marine microbes. This
370 cluster almost exclusively contains sequences from the families *Roseobacteriaceae* and *Rhodobacteraceae*
371 (42), including both pelagic microbes and symbionts of algae (*Silicimonas algicola*), mollusks
372 (*Aliiroseovarius crassostreae*), tunicates (*Ascidiaceihabitans donghaensis*), echinoderms (*Sulfitobacter delicatus*),
373 and corals (*Roseivivax isoporae*) (43–47). All genes from cluster 3 (145 sequences) were used as queries
374 in the RODEO genome neighborhood network function, using the same parameters as the cluster 1
375 analysis, resulting in 203 co-occurring domains with frequencies between 2 and 102 (Cluster3Co tab
376 of Supplemental file 1). Domains which co-occurred fewer than 5 times were excluded from analysis
377 (122 or ~60.0%). Co-occurring domains were compared to another random subset of the non-
378 specific control co-occurrence dataset to assign FDR values. All domains with an FDR greater than
379 0.1 were excluded from analysis, resulting in 42 significantly co-occurring domains
380 (Cluster3FDR_Threshold tab of Supplemental file 1). The most common co-occurring domains

381 were additional DUF560/T11SS domains (102/145 loci), DUF1194 (102/145 loci), glyoxalase
382 domains (89/145 loci), and thymidylate synthase/thymidylate synthase complementing protein
383 (77/145 loci). The TbpBBD domain passed the FDR threshold, however relative to the analysis of
384 cluster 1 it was exceptionally rare (8/145 loci). All genes that co-occurred with a marine cluster 3
385 T11SS OMP and contained any of the significant co-occurring domains were extracted, resulting in
386 860 significant co-occurring genes which we submitted to KofamKOALA and BlastKOALA for
387 functional analysis (21,22).

388 KofamKOALA found matches for 595 of the 860 significant co-occurring genes
389 (Cluster3KofamBrite tab of Supplemental file 2). When matched to cellular functions (brite
390 hierarchies) the most common categories were enzymes (lactoylglutathione lyase, thymidylate
391 synthase complementing proteins, aspartate aminotransferase, etc.), prokaryotic defense systems
392 (antitoxins CptB and HigA-1, topoisomerase IV B), and transcription factors (Cell division repressor
393 DicA, cold shock protein) (Fig. 7A). Of the 595 matched sequences, 270 had known pathway
394 association, the most common pathways being biosynthesis of one carbon pool by folate
395 (dihydrofolate reductase, thymidylate synthase), pyrimidine/nucleotide metabolism (thymidylate
396 synthase), pyruvate metabolism (lactoylglutathione lyase), and biosynthesis of amino acids (aspartate
397 aminotransferase) (Cluster3KofamPath tab of Supplemental file 2). Unlike cluster 1, transposases
398 were not among the significantly co-occurring genes of cluster 3. However, aminoacyl-tRNA
399 synthesis functions, which can be associated themselves with mobile genetic elements (32,48), do
400 appear on the co-occurrence list, including alanyl-tRNA, prolyl-tRNA, and threonyl-tRNA synthases
401 (Fig. 7B). Parallel analysis with BlastKOALA identified 514 of the 860 significant co-occurring
402 genes, of which 267 had known pathway association (Cluster3BlastBrite and Path tabs of
403 Supplemental file 2). BlastKOALA did not identify any additional cellular functions or pathways not
404 already revealed by KofamKOALA (Fig. S4).

405

406 **Bioinformatic investigation of potential cluster 3 T11SS cargo**

407 Initial attempts to identify new cargo proteins within the co-occurrence datasets from cluster
408 3 used the same methods used for cluster 1. UniRef50 clustering with ≥50% identity resulted in 218
409 homologous groups (Cluster3Uniref tab of Supplemental file 1) that were assessed with BLASTp for
410 sequence level homology to experimentally characterized T11SS cargo domains. Only a few sets of
411 homologs were identified, even when including all TbpBBD-domain-containing UniRef50 clusters.
412 Homologs of six UniRef50 clusters, specific to the *Rhodobacteraceae*/*Roseobacteriaceae* families,
413 resembled the Hemophilin-like 1 architecture, and homologs from another UniRef50 resembled the
414 disordered N-terminus 1 architecture noted previously (see Figs. 3 and 4). Collectively these 7
415 UniRef50 clusters encompass a total of only 31 ORFs across 27 species/taxa (Cluster3PutativeCargo
416 tab of Supplemental file 1). The ratio of predicted cargo per T11SS OMP was far lower in cluster 3
417 (0.21 Cargo/input OMP) than it was in cluster 1 (2.39 Cargo/input OMP). This dearth of co-
418 occurring putative cargo proteins suggests that marine cluster 3 T11SS OMPs seldom co-occur with
419 their cognate cargo proteins, utilize cargo proteins with below-threshold sequence similarity to those
420 of cluster 1 T11SS, or have evolved distinct functions and no longer have partner cargo proteins.

421 To further examine this potential transition in function, we comparatively annotated a
422 syntenic T11SS OMP-encoding locus in 11 representative *Rhodobacteraceae*/*Roseobacteriaceae* strains
423 (Clus3SyntenicPairs tab of Supplemental file 3). Each locus encodes all four of the top co-occurring
424 domains from our co-occurrence analysis, including flavin-dependent thymidylate synthase,
425 lactoylglutathione lyase, and one or more DUF1194 (NeighborhoodAlignment tab of Supplemental
426 file 3 and Fig. 8A). All but one (*Pseudooceanicola marinus* LMG23705) of the loci also encodes a protein
427 that is predicted to be a T11SS cargo lipoprotein (282-375 aa) based on the presence of an N-

428 terminal SPII type signal sequence and a C-terminal 8-stranded β-barrel domain (Fig. 5F; Fig. 8B-K)
429 (18,20). The presence of both cluster 1 and cluster 3 co-occurring domains indicates that this
430 conserved locus is a hybrid of the two. It is tempting to speculate that this locus represents a
431 transitional state in the divergence of these systems, with at least one analyzed genome appearing to
432 have lost the cluster 1 co-occurring, 8-stranded β-barrel domain.

433 We next examined the cluster 3 marine T11SS co-occurrence dataset to consider the
434 possibility of alternative putative cargo types. In contrast to cluster 1 T11SS, co-occurrence with
435 TbpBBD/lipoprotein C domains, TonB-dependent receptors, and TonB domains are rare within
436 the marine cluster 3 T11SS loci. Instead, the DUF1194 domain (PF06707) was highly prevalent
437 (102/145 queries) (Supplemental file 4). Of the 102 DUF1194-containing sequences found in the
438 cluster 3 co-occurrence list, 7 are predicted to be lipoproteins, 76 are predicted to be soluble Sec-
439 secreted proteins, and 19 have no detectable signal peptide according to SignalP 6 (33,34). Many
440 cluster 3 T11SS loci encoded 2 or 3 distinct DUF1194 genes in close proximity (e.g., Fig. 8),
441 predicted to encode both lipidated and non-lipidated proteins (e.g., algal symbiont *Rhodobacteraceae*
442 PD-2 ETA49263.2 and ETA49262.2, respectively) (Supplemental file 4). The predicted
443 RoseTTAFold structure (49) of the DUF1194 domain is a mixed β-strand/α-helix. According to the
444 Pfam database, the DUF1194 domain occurs in 31 distinct protein architectures with other domains,
445 including C-terminal autotransporter domain (PF03797) predicted to function in T5SS secretion, the
446 CARDB domain (PF07705), which adopts a 7 β-strand structure, DUF11 (PF01345), predicted to
447 have an 8-stranded β-barrel structure, and PF18911, the known crystal structures of which also
448 adopt a β-barrel (e.g., 1wgo and 2y72).

449 Of the co-occurring DUF1194 sequences, 6 are annotated as having homology to von
450 Willebrand factor type A domain or vWA (PF00092). This homology is reflected in the structural

451 similarity between these two domains, including the presence of a Rossman fold, which may hint at
452 similar molecular roles. vWA protein architectures can include divalent-cation-binding metal ion-
453 dependent adhesion sites (MIDAS), with a characteristic DxSxS…T…D motif (50,51). Almost all
454 the DUF1194 proteins that co-occurred with DUF560 proteins contained at least one MIDAS motif
455 (98/102). To examine potential structural features of cluster 3 T11SS co-occurring DUF1194-
456 containing homologs, we focused on three adjacently encoded DUF1194 genes from
457 *Ascidiaceihabitans donghaensis* (SPH20589/SPH20588/SPH20587) for structural analysis. PsiPred 4.0
458 predicted that these proteins have at least 5 predominantly hydrophobic β-strands, all separated by
459 α-helix regions in a structure that also appears to be a variation on the classical α/β Rossman fold
460 (Fig. 9A). RoseTTAFold predictions of the tertiary structure of these DUF1194 proteins reveal
461 globular proteins wherein 6 hydrophobic β-strands form a β-sheet at the core of the protein, which
462 is protected by three α-helices per side. All three representatives had a MIDAS motif starting 10-11
463 residues after the predicted signal peptide. For each protein, the last 11-13 C-terminal residues
464 appeared to be disordered, and their positions could not be estimated with a high degree of
465 confidence (Fig. 9B).

466 **Discussion**

467 The controlled bioinformatics analyses we report here reinforce the idea that cluster 1
468 T11SS-dependent cargo have roles in metal uptake, single carbon metabolism, and nutrient
469 provision. Further, our data extend the experimentally verified T11SS cargo to include *Haemophilus*
470 *parahaemolyticus* Pls and lend credence to the concept that T11SS cargo can be bioinformatically
471 predicted. The structural prediction that this cargo protein is a fusion between a functional Pls
472 domain, typical of Gram-positive bacteria, and a C-terminal 8-stranded β-barrel domain lends
473 further evidence of the function of the barrel domain in directing cargo to the T11SS, and for the
474 function of T11SS in facilitating transport across an outer membrane. However, our analysis of the

475 more distantly related cluster 3 T11SS OMPs revealed that known barrel domains (TbpBBD and
476 lipoprotein C) are rare among the significantly co-occurring genes identified, raising the intriguing
477 possibility that novel structural domains may exist that mediate targeting to cluster 3 T11SS.

478 Our bioinformatic approach relied upon gene co-occurrence analysis, which uses genomic
479 proximity as a proxy for coordinate regulation and potential interactions, since genes within a
480 common functional pathway often cluster together within a genome. While this is not a foolproof
481 method, the use of large genomic datasets increases the power of co-occurrence analysis to identify
482 potential co-interacting partners. One of the methodological refinements we developed was the use
483 of *in silico* controls for genome neighborhood co-occurrence to reduce false positives and noise in
484 the output datasets of putative T11SS co-occurring protein families. For instance, activities
485 associated with protein translation and RNA metabolism, which are essential cellular processes
486 expected to be present in every cell, are not always useful co-occurrence signatures for elucidating
487 T11SS function and components. Such general functions are not indicative of the specific functions
488 of T11SS relative to any other OMP and can be considered a form of false negative, Type II error;
489 such co-occurrences are not biologically relevant, but are not being excluded from the analysis
490 because they do co-occur frequently with the query gene. However, removal of such co-occurrences
491 can also lead to false positive, Type I error, because they can also result in removal of some
492 biologically relevant co-occurrences. For example, there could be protein families that are important
493 functional partners of the query protein, but that are removed from the dataset if they were
494 important for both query and control proteins. This type of error occurred in our analysis. For
495 example, TonB-dependent receptors, with which many T11SS-dependent cargo interact, were
496 excluded from the cluster 1 T11SS analyses since these domains associate with other outer
497 membrane proteins unrelated to T11SS. In a sense, this control narrows the focus from total co-
498 occurrence to unique co-occurrence, and as such it is best suited towards applications in which the

499 researcher seeks to differentiate protein families. Additionally, the technique can be further adapted
500 and refined, depending on application. For instance, multiple non-specific gene neighborhood
501 controls could be generated by randomly subsampling the list of proteins that are biophysically
502 similar to the query proteins, allowing for increased stringency and measures of variance.

503 The cluster 1 T11SS significantly co-occurring genes identified here are consistent with
504 previous observations (3,5). Iron uptake from the environment is a common function among
505 experimentally verified T11SS-dependent cargo proteins and this theme was extended within the
506 cluster 1 T11SS co-occurrence dataset, which included heme oxygenases and heme transporters that
507 facilitate iron uptake (52). Similarly, the protein export genes identified as co-occurring with cluster 1
508 T11SS included TonB and ExbD, which are essential to energize uptake of nutrients from known
509 T11SS-dependent cargo (6,8,53). TolA is essential to bring periplasmic nutrients into the cytoplasm
510 of the cell (54). Signal peptidase II is essential for the generation of bacterial lipoproteins, which
511 constitute a substantial percentage of T11SS-dependent cargo (55). Finally, the observed co-
512 occurrence with transposases, ribosome components, and tRNA-synthases is consistent with the
513 hypothesis that T11SS are regularly associated with mobile genetic islands and can be horizontally
514 acquired and contribute to fitness within a host environment (32).

515 Significantly co-occurring pathways were commonly identified for both cluster 1 and cluster
516 3 T11SS, and represent pathways not previously associated with T11SS. These included folate
517 (vitamin B9)-related pathways, including folate biosynthesis, folate-dependent enzymes, and one
518 carbon metabolism via folate. Folate, like heme, is an enzyme cofactor required for many diverse
519 biological pathways. It is essential for the function of thymidylate synthase complementing protein,
520 also known as flavin-dependent thymidylate synthase, in nucleotide biosynthesis (56). This protein is
521 the third most common cluster 3 marine T11SS co-occurring gene, suggesting coordinate regulation

522 or inheritance. Since, unlike heme, folate does not incorporate a metallic ion and most bacteria
523 synthesize their folate instead of scavenging it from the environment (57), we do not favor the
524 hypothesis that folate itself is acquired through a T11SS-dependent uptake mechanism. Instead, the
525 close association of folate metabolism and T11SS may reflect the fact that folate biosynthesis, as well
526 as iron-sulfur cluster biosynthesis, and methylotrophic metabolism by formate dehydrogenase
527 (which also significantly co-occurred with T11SS) depend on the metal-bearing enzyme cofactors,
528 cobalamin (vitamin B12) and molybdopterin cofactor (MoCo) (58–61). Our data suggest that some
529 T11SS OMPs may function in cobalamin or molybdopterin uptake pathways. Regardless, our data
530 clearly demonstrate an association between T11SS and single carbon metabolism and methyltropism.

531 Cluster 1 T11SS co-occurrences featured TonB energization or heme/iron uptake that were lacking
532 from cluster 3 T11SS co-occurrences. In turn, the most frequent cluster 3 T11SS co-occurrences were
533 DUF1194 and lactoylglutathione lyase (GloA), neither of which appeared in the Cluster 1 T11SS co-
534 occurrence datasets, but both of which have connections to metal ions. The function of DUF1194 is
535 currently unknown but it is predicted to fold into a globular protein with an α/β Rossman fold.
536 Several of the DUF1194 proteins (6/103) we detected also had homology to a structurally similar
537 α/β Rossman fold domain called von Willebrand factor A (vWA), which is most often associated
538 with serum glycoproteins and large multimeric protein complexes in mammalian systems (62,63). In
539 *Myxococcus xanthus* bacteria, the vWA domain-containing surface lipoprotein CglB acts as a surface
540 adhesin essential for gliding motility (64). Upon binding a ligand, vWA domain-containing adhesins
541 such as CglB can stabilize their bond via a conformational change induced by MIDAS motif
542 allosteric binding of a divalent metal cation, often Mg^{2+}(51,64). The presence of a MIDAS motif in
543 the vast majority of DUF1194 identified within our dataset suggests that these proteins may also
544 play a role in adhesion or may participate in metal homeostasis in previously unpredicted ways. An
545 exciting future direction will be to determine if DUF1194 proteins are a type of T11SS-dependent

546 cargo, or if they contribute to a multimeric protein complex that interacts with T11SS proteins.
547 Lactoylglutathione lyase is an enzyme with a coordinated metal ion (typically nickel) and functions in
548 the regeneration step of the glyoxalase-mediated aldehyde detoxification system (65). It is possible
549 that the nickel requirement of this enzyme reflects a role for cluster 3 T11SS in facilitating nickel
550 uptake. Alternatively, this co-occurrence may indicate that cluster 3 T11SS are more generally
551 participating in an aldehyde producing process. Notably, aldehydes like methylglyoxal are
552 exceptionally cytotoxic and can damage DNA and proteins, so the cluster 3 T11SS co-occurrence of
553 nucleotide and amino acid biosynthesis may reflect a more general relationship between T11SS and
554 the aldehyde repair response (65,66).

555 In addition to providing insights into functional pathways associated with T11SS, our
556 analysis also allowed us to identify 148 homology groups of putative T11SS-dependent cargo,
557 including those with domain architectures distinct from those found in known cargo. We verified
558 that at least one of these, the plasmin-sensitive surface protein (Pls) from the Gram-negative
559 organism *H. parahaemolyticus,* is a bona fide T11SS-dependent cargo protein, demonstrating the
560 efficacy of our bioinformatics approach. Pls represents a novel N-terminal effector domain among
561 T11SS cargo. This domain is homologous to glycosylated Pls proteins found on or spontaneously
562 cleaved from the surfaces of Gram-positive organisms such as *Staphylococcus aureus* and *S. epidermidis*
563 where function in biofilm formation and modulating adhesion to host proteins (40,41). Given that
564 the function of Pls requires surface exposure, its presence in a Gram-negative organism is
565 complicated by the existence of the outer membrane. In *H. parahaemolyticus*, this hurdle seems to
566 have been overcome via fusion of the Pls repeat domain to the C-terminal 8-stranded β-barrel
567 domain necessary for cluster 1 T11SS-dependent secretion across the outer membrane (3,5).

568 When co-expressed in *E. coli* with its T11SS partner, TepS, Pls predominantly localizes to the
569 supernatant and is not tethered to the cell surface as would be expected of a lipoprotein. Our
570 working model is that some Pls can reach the cell surface in the absence of T11SS, as has been
571 observed for factor H binding protein (67,68), and that in the presence of the T11SS TepS, Pls is
572 secreted but then cleaved, releasing it into the extracellular milieu Given that our experiments were
573 conducted in a protease-deficient strain of *E. coli*, it will be exciting in future studies to determine if
574 TepS alters the topology of Pls in order to make it susceptible to proteolysis, or if it is directly
575 responsible for Pls cleavage, which would be a novel function for T11SS. The possibility that Gram-
576 negative Pls are secreted via T11SS and may be released from the cell surface in the process will be
577 an important factor to consider in ongoing efforts to target Pls homologs for vaccine development,
578 such as that being pursued for *Actinobacillus pleuropneumoniae* Pls (YP_001652736.1) (69). Also, the
579 Pls-like proteins have repeat rich regions and the possible glycosylation of *H. parahaemolyticus* Pls is
580 supported by our observations of its aberrant mobility in SDS-PAGE. It is possible that other
581 T11SS-dependent cargo, including lipoproteins, may also be glycosylated. If so, such glycoproteins
582 tethered to the cell surface would display glycans as surface antigens that may modulate host-
583 microbe recognition and immune evasion (70–72).

584 In summary, our findings have enhanced our understanding of potential effectors of the
585 novel T11SS and expanded the list of experimentally verified T11SS secreted cargo. We also have
586 expanded the field of genomic co-occurrence analysis by establishing a protocol that can serve as a
587 basis for more controlled and specific experimental designs. Exciting avenues of future research
588 could focus on generating an automated pipeline for controlling co-occurrences analyses,
589 determining if Pls is required for host-colonization by *Haemophilus parahaemolyticus*, investigating if
590 DUF1194 represents a novel T11SS cargo type, and exploring the relationship between one carbon
591 metabolism and T11SS.

592

593 **Conflicts of interest**

594 The author(s) declare that there are no conflicts of interest.

595

596 **Funding information**

597 This work was funded, in part, by funds awarded to H.G.-B. from the University of Tennessee-
598 Knoxville (David and Sandra White Endowment) and the National Science Foundation (IOS-
599 2128266).

600

601 **Author contributions**

602 ASG: Conceptualization, data curation, formal analysis, investigation, methodology, validation,
603 writing original draft and review/editing

604 NCM: Conceptualization, data curation, investigation, formal analysis, methodology, writing original
605 draft

606 HGB: Conceptualization, data curation, funding acquisition, investigation, methodology, project
607 administration, resources, supervision, visualization, writing original draft and review/editing

608

609 **Acknowledgements**

610 We thank past and present members of the Goodrich-Blair lab for feedback on the development of
611 controls for bioinformatic co-occurrence analyses.

612 **References:**

613 1. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar G, Sonnhammer ELL, et al. Pfam:
614 The protein families database in 2021. Nucleic Acids Res [Internet]. 2021 Jan
615 8;49(D1):D412–9. Available from: https://doi.org/10.1093/nar/gkaa913

616

617 2. Mudgal R, Sandhya S, Chandra N, Srinivasan N. De-DUFing the DUFs: Deciphering distant
618 evolutionary relationships of Domains of Unknown Function using sensitive homology
619 detection methods. Biol Direct [Internet]. 2015;10(1):38. Available from:
620 https://doi.org/10.1186/s13062-015-0069-2

621

622 3. Grossman AS, Mauer TJ, Forest KT, Goodrich-Blair H, Comstock L. A widespread bacterial
623 secretion system with diverse substrates. mBio [Internet]. 2022 Apr 19;12(4):e01956-21.
624 Available from: https://doi.org/10.1128/mBio.01956-21

625

626 4. Shin HE, Pan C, Curran DM, Bateman TJ, Chong DHY, Ng D, et al. Prevalence of Slam-
627 dependent hemophilins in Gram-negative bacteria. J Bacteriol [Internet]. 2024 May
628 30;206(6):e00027-24. Available from: https://doi.org/10.1128/jb.00027-24

629

630 5. Hooda Y, Lai CCL, Moraes TF. Identification of a large family of Slam-dependent surface
631 lipoproteins in Gram-negative bacteria. Front Cell Infect Microbiol [Internet]. 2017;7.
632 Available from: http://dx.doi.org/10.3389/fcimb.2017.00207

633

634 6. Hooda Y, Lai CCL, Judd A, Buckwalter CM, Shin HE, Gray-Owen SD, et al. Slam is an outer
635 membrane protein that is required for the surface display of lipidated virulence factors in
636 Neisseria. Nat Microbiol [Internet]. 2016;1(4). Available from:
637 http://dx.doi.org/10.1038/nmicrobiol.2016.9

638

639 7. Cowles CE, Goodrich-Blair H. The Xenorhabdus nematophila nilABC genes confer the
640 ability of Xenorhabdus spp. to colonize Steinernema carpocapsae nematodes. J Bacteriol
641 [Internet]. 2008 Jun 15;190(12):4121–8. Available from: https://doi.org/10.1128/JB.00123-
642 08

643

644 8. Bateman TJ, Shah M, Ho TP, Shin HE, Pan C, Harris G, et al. A Slam-dependent hemophore
645 contributes to heme acquisition in the bacterial pathogen Acinetobacter baumannii. Nat
646 Commun [Internet]. 2021;12(1):6270. Available from: https://doi.org/10.1038/s41467-021-
647 26545-9

9. Latham RD, Torrado M, Atto B, Walshe JL, Wilson RK, Guss JM, et al. A heme-binding protein produced by Haemophilus haemolyticus inhibits non-typeable Haemophilus influenzae. Mol Microbiol [Internet]. 2020 Feb 1;113(2):381–98. Available from: https://doi.org/10.1111/mmi.14426

10. Grossman AS, Gell DA, Wu DG, Carper DL, Hettich RL, Goodrich-Blair H. Bacterial hemophilin homologs and their specific type eleven secretor proteins have conserved roles in heme capture and are diversifying as a family. J Bacteriol [Internet]. 2024 Mar 20;206(6):e00444-23. Available from: https://doi.org/10.1128/jb.00444-23

11. Sam F, Brianna A, Arianna M, J HK, S RJ, Elan E, et al. Heme sequestration by hemophilin from Haemophilus haemolyticus reduces respiratory tract colonization and infection with non-typeable Haemophilus influenzae. mSphere [Internet]. 2024 Feb 21;9(3):e00006-24. Available from: https://doi.org/10.1128/msphere.00006-24

12. Zhang Z, Wang J, Wang J, Wang J, Li Y. Estimate of the sequenced proportion of the global prokaryotic genome. Microbiome. 2020 Sep;8(1):134.

13. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, et al. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. 2002 May;30(10):2212–23.

14. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai HC, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol. 2017 May;13(5):470–8.

15. The UniProt Consortium. UniProt: The universal protein knowledge base in 2021. Nucleic Acids Res [Internet]. 2021 Jan 8;49(D1):D480–9. Available from: https://doi.org/10.1093/nar/gkaa1100

16. Chen IMA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The IMG/M data management and analysis system v.7: content updates and new features. Nucleic Acids Res [Internet]. 2023 Jan 6;51(D1):D723–32. Available from: https://doi.org/10.1093/nar/gkac976

683 17. Mukherjee S, Stamatis D, Li CT, Ovchinnikova G, Bertsch J, Sundaramurthi JC, et al.
684 Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in
685 v.9. Nucleic Acids Res [Internet]. 2023 Jan 6;51(D1):D957–63. Available from:
686 https://doi.org/10.1093/nar/gkac974

687

688 18. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate
689 protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

690

691 19. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making
692 protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679–82.

693

694 20. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for
695 protein modeling, prediction and analysis. Nat Protoc [Internet]. 2015;10(6):845–58. Available
696 from: https://doi.org/10.1038/nprot.2015.053

697

698 21. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al.
699 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score
700 threshold. Bioinformatics [Internet]. 2020 Apr 1;36(7):2251–2. Available from:
701 https://doi.org/10.1093/bioinformatics/btz859

702

703 22. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for
704 functional characterization of genome and metagenome sequences. J Mol Biol. 2016
705 Feb;428(4):726–31.

706

707 23. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-
708 redundant UniProt reference clusters. Bioinformatics [Internet]. 2007 May 15;23(10):1282–8.
709 Available from: https://doi.org/10.1093/bioinformatics/btm098

710

711 24. Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. Nucleic
712 Acids Res. 2019 Jul;47(W1):W402–7.

713

714 25. Bhasin A, Chaston JM, Goodrich-Blair H. Mutational analyses reveal overall topology and
715 functional regions of NilB, a bacterial outer membrane protein required for host association
716 in a model of animal-microbe mutualism. J Bacteriol [Internet]. 2012 Apr 1;194(7):1763–76.
717 Available from: https://doi.org/10.1128/JB.06711-11

26. Miroux B, Walker JE. Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J Mol Biol. 1996 Jul;260(3):289–98.

27. Dumon-Seignovert L, Cariot G, Vuillard L. The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). Protein Expr Purif. 2004 Sep;37(1):203–6.

28. Orchard SS, Goodrich-Blair H. Identification and functional characterization of a Xenorhabdus nematophila oligopeptide permease. Appl Environ Microbiol [Internet]. 2004 Sep 1;70(9):5621–7. Available from: https://doi.org/10.1128/AEM.70.9.5621-5627.2004

29. Koontz L. Chapter one - TCA precipitation. In: Lorsch JBTM in E, editor. Laboratory methods in enzymology [Internet]. Academic Press; 2014. p. 3–10. Available from: http://www.sciencedirect.com/science/article/pii/B978012420119400001X

30. Tukey JW. Comparing individual means in the analysis of variance. Biometrics [Internet]. 1949 May 25;5(2):99–114. Available from: http://www.jstor.org/stable/3001913

31. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan;45(D1):D353–61.

32. Dobrindt U, Reidl J. Pathogenicity islands and phage conversion: evolutionary aspects of bacterial pathogenesis. International Journal of Medical Microbiology [Internet]. 2000;290(6):519–27. Available from: https://www.sciencedirect.com/science/article/pii/S143842210080017X

33. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol [Internet]. 2019;37(4):420–3. Available from: http://dx.doi.org/10.1038/s41587-019-0036-z

34. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat

753 Biotechnol [Internet]. 2022;40(7):1023–5. Available from: https://doi.org/10.1038/s41587-
754 021-01156-3

755

756 35. Skaf MS, Polikarpov I, Stanković IM. A linker of the proline-threonine repeating motif
757 sequence is bimodal. J Mol Model. 2020 Jun;26(7):178.

758

759 36. Swearingen KE, Lindner SE, Shi L, Shears MJ, Harupa A, Hopp CS, et al. Interrogating the
760 Plasmodium sporozoite surface: identification of surface-exposed proteins and demonstration
761 of glycosylation on CSP and TRAP by mass spectrometry-based proteomics. PLoS Pathog.
762 2016 Apr;12(4):e1005606.

763

764 37. Sieber CMK, Paul BG, Castelle CJ, Hu P, Tringe SG, Valentine DL, et al. Unusual
765 Metabolism and Hypervariation in the Genome of a Gracilibacterium (BD1-5) from an Oil-
766 Degrading Community. mBio. 2019 Nov;10(6).

767

768 38. Calmettes C, Alcantara J, Yu RH, Schryvers AB, Moraes TF. The structural basis of
769 transferrin sequestration by transferrin-binding protein B. Nat Struct Mol Biol [Internet].
770 2012;19(3):358–60. Available from: https://doi.org/10.1038/nsmb.2251

771

772 39. Ostberg K, DeRocco A, Mistry S, Dickinson Mary K, Cornelissen CN. Conserved regions of
773 gonococcal TbpB are critical for surface exposure and transferrin iron utilization. Infect
774 Immun [Internet]. 2013 Sep 1;81(9):3442–50. Available from:
775 https://doi.org/10.1128/IAI.00280-13

776

777 40. Bleiziffer I, Eikmeier J, Pohlentz G, McAulay K, Xia G, Hussain M, et al. The Plasmin-
778 Sensitive Protein Pls in Methicillin-Resistant Staphylococcus aureus (MRSA) is a glycoprotein.
779 PLoS Pathog [Internet]. 2017 Jan 12;13(1):e1006110. Available from:
780 https://doi.org/10.1371/journal.ppat.1006110

781

782 41. Savolainen K, Paulin L, Westerlund-Wikström B, Foster TJ, Korhonen TK, Kuusela P.
783 Expression of pls, a gene closely associated with the mecA gene of methicillin-resistant
784 Staphylococcus aureus, prevents bacterial adhesion in vitro. Infect Immun. 2001
785 May;69(5):3013–20.

786

787 42. Liang KYH, Orata FD, Boucher YF, Case RJ. Roseobacters in a Sea of Poly- and Paraphyly:
788 Whole Genome-Based Taxonomy of the Family Rhodobacteraceae and the Proposal for the
789 Split of the "Roseobacter Clade" Into a Novel Family, Roseobacteraceae fam. nov. Front
790 Microbiol [Internet]. 2021 [cited 2024 Aug 18];12. Available from:
791 https://doi.org/10.3389/fmicb.2021.683109

792

793 43. Chen MH, Sheu SY, Chen CA, Wang JT, Chen WM. Roseivivax isoporae sp. nov., isolated
794 from a reef-building coral, and emended description of the genus Roseivivax. Int J Syst Evol
795 Microbiol. 2012 Jun;62(Pt 6):1259–64.

796

797 44. Crenn K, Serpin D, Lepleux C, Overmann J, Jeanthon C. Silicimonas algicola gen. nov., sp.
798 nov., a member of the Roseobacter clade isolated from the cell surface of the marine diatom
799 Thalassiosira delicatula. Int J Syst Evol Microbiol. 2016 Nov;66(11):4580–8.

800

801 45. Kessner L, Spinard E, Gomez-Chiarri M, Rowley DC, Nelson DR. Draft genome sequence
802 of Aliiroseovarius crassostreae CV919-312, the causative agent of Roseovarius Oyster Disease
803 (formerly Juvenile Oyster Disease). Genome Announc. 2016 Mar;4(2).

804

805 46. Kim YO, Park S, Nam BH, Lee C, Park JM, Kim DG, et al. Ascidiaceihabitans donghaensis
806 gen. nov., sp. nov., isolated from the golden sea squirt Halocynthia aurantium. Int J Syst Evol
807 Microbiol. 2014 Dec;64(Pt 12):3970–5.

808

809 47. Ivanova EP, Gorshkova NM, Sawabe T, Zhukova N V, Hayashi K, Kurilenko V V, et al.
810 Sulfitobacter delicatus sp. nov. and Sulfitobacter dubius sp. nov., respectively from a starfish
811 (Stellaster equestris) and sea grass (Zostera marina). Int J Syst Evol Microbiol. 2004 Mar;54(Pt
812 2):475–80.

813

814 48. Alamos P, Tello M, Bustamante P, Gutiérrez F, Shmaryahu A, Maldonado J, et al.
815 Functionality of tRNAs encoded in a mobile genetic element from an acidophilic bacterium.
816 RNA Biol. 2018;15(4–5):518–27.

817

818 49. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate
819 prediction of protein structures and interactions using a three-track neural network. Science.
820 2021 Aug;373(6557):871–6.

821

822 50. Bergelson JM, Hemler ME. Integrin- Ligand Binding: Do integrins use a 'MIDAS touch' to
823 grasp an Asp? Current Biology [Internet]. 1995;5(6):615–7. Available from:
824 https://www.sciencedirect.com/science/article/pii/S0960982295001242

825

826 51. Cantí C, Nieto-Rostro M, Foucault I, Heblich F, Wratten J, Richards MW, et al. The metal-
827 ion-dependent adhesion site in the Von Willebrand factor-A domain of alpha2delta subunits
828 is key to trafficking voltage-gated Ca2+ channels. Proc Natl Acad Sci U S A. 2005
829 Aug;102(32):11230–5.

830

831 52. Hantke K. Is the bacterial ferrous iron transporter FeoB a living fossil? Trends Microbiol.
832 2003 May;11(5):192–5.

833

834 53. Ollis AA, Postle K. ExbD mutants define initial stages in TonB energization. J Mol Biol.
835 2012 Jan;415(2):237–47.

836

837 54. Levengood SK, Beyer WFJ, Webster RE. TolA: a membrane protein involved in colicin
838 uptake contains an extended helical region. Proc Natl Acad Sci U S A. 1991 Jul;88(14):5939–
839 43.

840

841 55. Tjalsma H, Kontinen VP, Prágai Z, Wu H, Meima R, Venema G, et al. The role of
842 lipoprotein processing by signal peptidase II in the Gram-positive eubacterium bacillus
843 subtilis. Signal peptidase II is required for the efficient secretion of alpha-amylase, a non-
844 lipoprotein. J Biol Chem. 1999 Jan;274(3):1698–707.

845

846 56. Agrawal N, Lesley SA, Kuhn P, Kohen A. Mechanistic studies of a flavin-dependent
847 thymidylate synthase. Biochemistry [Internet]. 2004 Aug 1;43(32):10295–301. Available from:
848 https://doi.org/10.1021/bi0490439

849

850 57. Dawadi S, Kordus SL, Baughn AD, Aldrich CC. Synthesis and analysis of bacterial folate
851 metabolism intermediates and antifolates. Org Lett [Internet]. 2017 Oct 6;19(19):5220–3.
852 Available from: https://doi.org/10.1021/acs.orglett.7b02487

853

854 58. Luk LYP, Javier Ruiz-Pernía J, Dawson WM, Roca M, Loveridge EJ, Glowacki DR, et al.
855 Unraveling the role of protein dynamics in dihydrofolate reductase catalysis. Proc Natl Acad
856 Sci U S A. 2013 Oct;110(41):16344–9.

857

858 59. Mendel RR. The molybdenum cofactor. J Biol Chem. 2013 May;288(19):13165–72.

859

860 60. Zhong Q, Kobe B, Kappler U. Molybdenum enzymes and how they support virulence in
861 pathogenic bacteria [Internet]. Vol. 11, Frontiers in Microbiology. 2020. Available from:
862 https://www.frontiersin.org/articles/10.3389/fmicb.2020.615860

863

864 61. Van Bibber M, Bradbeer C, Clark N, Roth JR. A new class of cobalamin transport mutants
865 (btuF) provides genetic evidence for a periplasmic binding protein in Salmonella
866 typhimurium. J Bacteriol. 1999 Sep;181(17):5539–41.

867

868 62. Peyvandi F, Garagiola I, Baronciani L. Role of von Willebrand factor in the haemostasis.
869 Blood Transfus. 2011 May;9 Suppl 2(Suppl 2):s3-8.

870

871 63. Von Willebrand EA. Hereditary pseudohaemophilia. Haemophilia. 1999 May;5(3):223–31;
872 discussion 222.

873

874 64. Islam ST, Jolivet NY, Cuzin C, Belgrave AM, My L, Fleuchot B, et al. Unmasking of the von
875 Willebrand A-domain surface adhesin CglB at bacterial focal adhesions mediates
876 myxobacterial gliding motility. Sci Adv [Internet]. 2023 May 8;9(8):eabq0619. Available from:
877 https://doi.org/10.1126/sciadv.abq0619

878

879 65. Lee C, Park C. Bacterial responses to glyoxal and methylglyoxal: Reactive electrophilic
880 species. Int J Mol Sci. 2017 Jan;18(1).

881

882 66. Anaya-Sanchez A, Feng Y, Berude JC, Portnoy DA. Detoxification of methylglyoxal by the
883 glyoxalase system is required for glutathione availability and virulence activation in Listeria
884 monocytogenes. PLoS Pathog. 2021 Aug;17(8):e1009819.

885

886 67. Fantappiè L, Irene C, De Santis M, Armini A, Gagliardi A, Tomasi M, et al. Some Gram-
887 negative lipoproteins keep their surface topology when transplanted from one species to
888 another and deliver foreign polypeptides to the bacterial surface. Molecular & Cellular
889 Proteomics [Internet]. 2017 Jul 1;16(7):1348 LP – 1364. Available from:
890 http://www.mcponline.org/content/16/7/1348.abstract

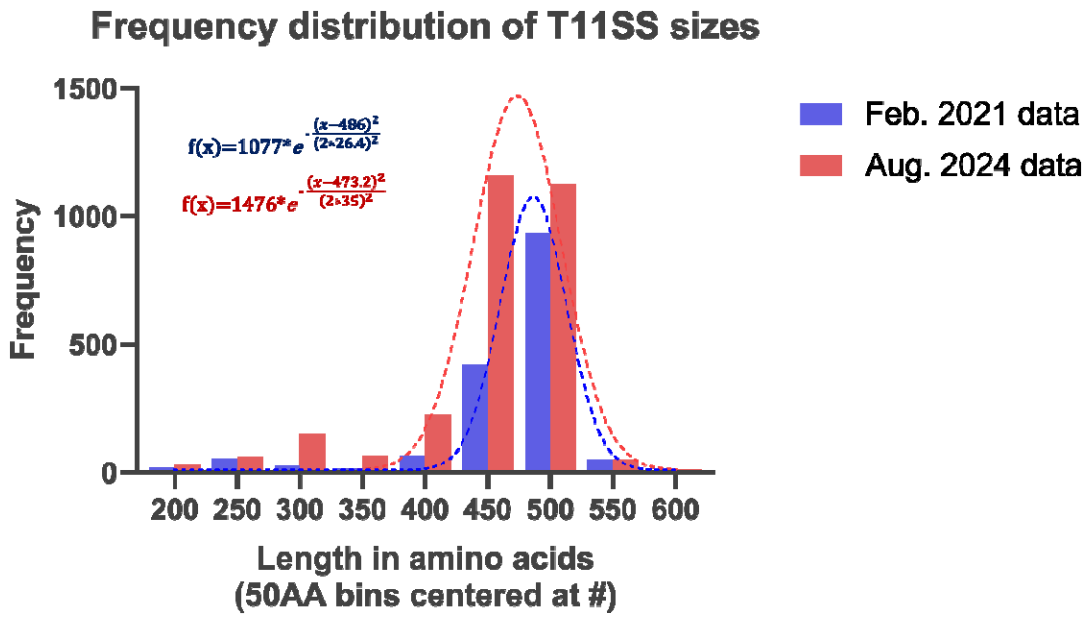
68. Konar M, Rossi R, Walter H, Pajon R, Beernink PT. A mutant library approach to identify improved Meningococcal Factor H Binding Protein vaccine antigens. PLoS One [Internet]. 2015 Jun 9;10(6):e0128185. Available from: https://doi.org/10.1371/journal.pone.0128185

69. Cao Y, Gao L, Zhang L, Zhou L, Yang J, Deng L, et al. Genome-wide screening of lipoproteins in Actinobacillus pleuropneumoniae identifies three antigens that confer protection against virulent challenge. Sci Rep. 2020 Feb;10(1):2343.

70. Schmidt MA, Riley LW, Benz I. Sweet new world: glycoproteins in bacterial pathogens. Trends Microbiol. 2003 Dec;11(12):554–61.

71. Tortorelli G, Rautengarten C, Bacic A, Segal G, Ebert B, Davy SK, et al. Cell surface carbohydrates of symbiotic dinoflagellates and their role in the establishment of cnidarian-dinoflagellate symbiosis. ISME J. 2022 Jan;16(1):190–9.

72. Zhou M, Wu H. Glycosylation and biogenesis of a family of serine-rich bacterial adhesins. Microbiology (Reading). 2009 Feb;155(Pt 2):317–27.



Supplemental Figure 1. **T11SS (DUF560) protein size distribution (in amino acids).** After examination of the size distribution, median was chosen as an appropriate measure of representative protein size over mean due to left-skewed distribution. The blue dataset was used to perform the T11SS co-occurrence analysis described in this manuscript. In the intervening time more homologs have been added to the Pfam database, the red dataset reflects these additions and their impact on the size distribution.

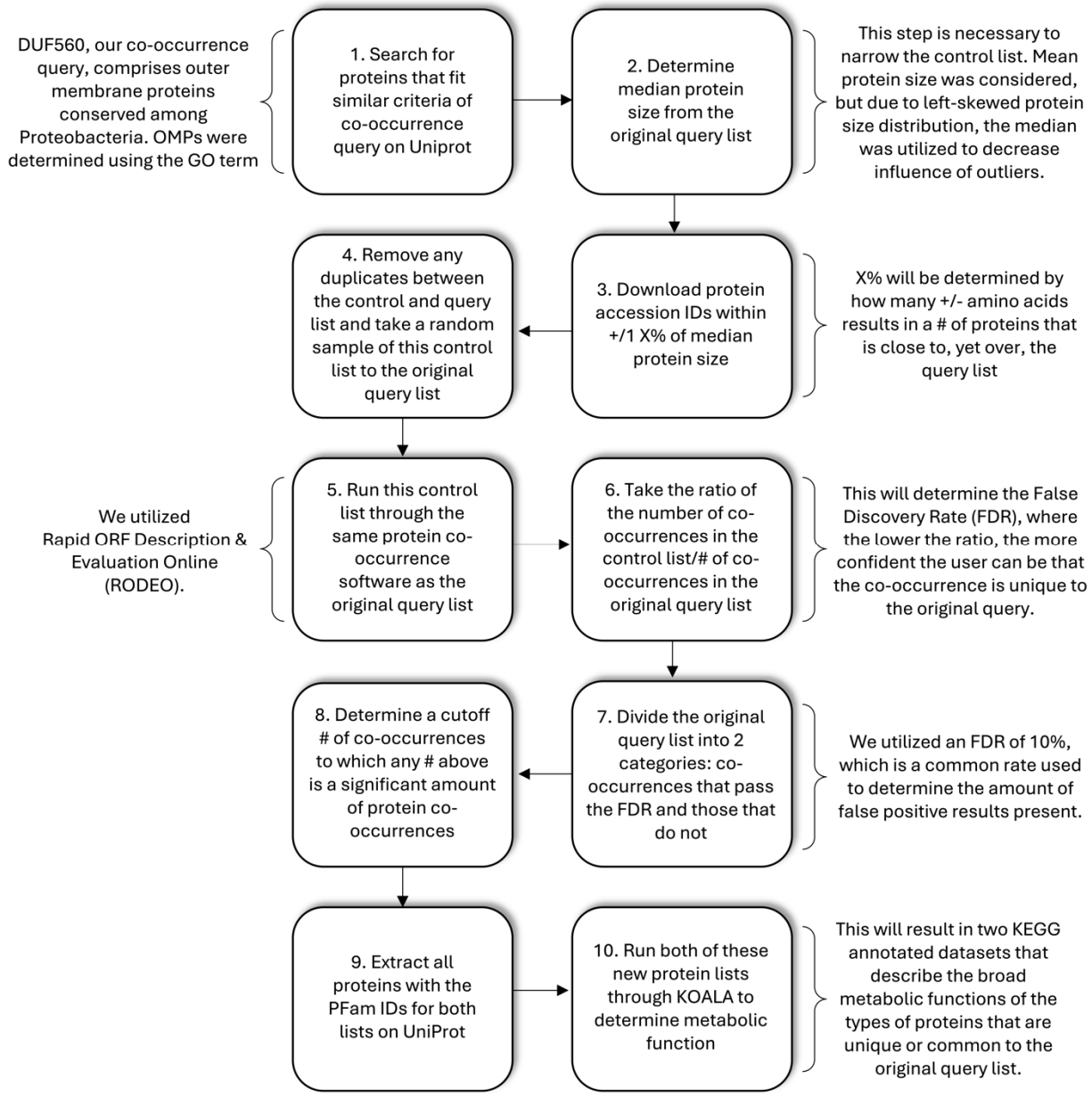


Figure 1. **Design for a novel control technique for genomic neighborhood co-occurrence analysis.** Schematic diagram outlining the process we used to remove nonspecific co-occurrences from a genomic neighborhood analysis, with commentary highlighting how we used this outline to refine our T11SS co-occurrence database.

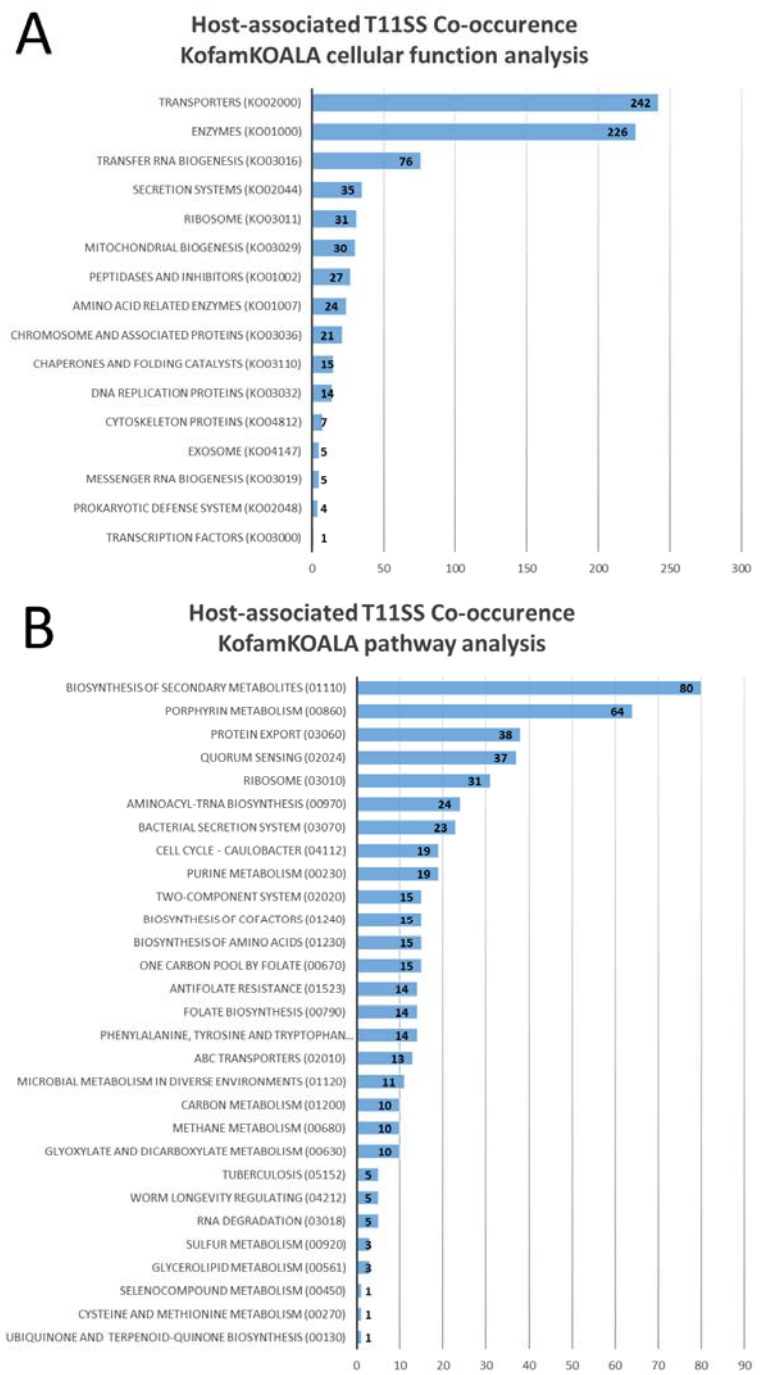
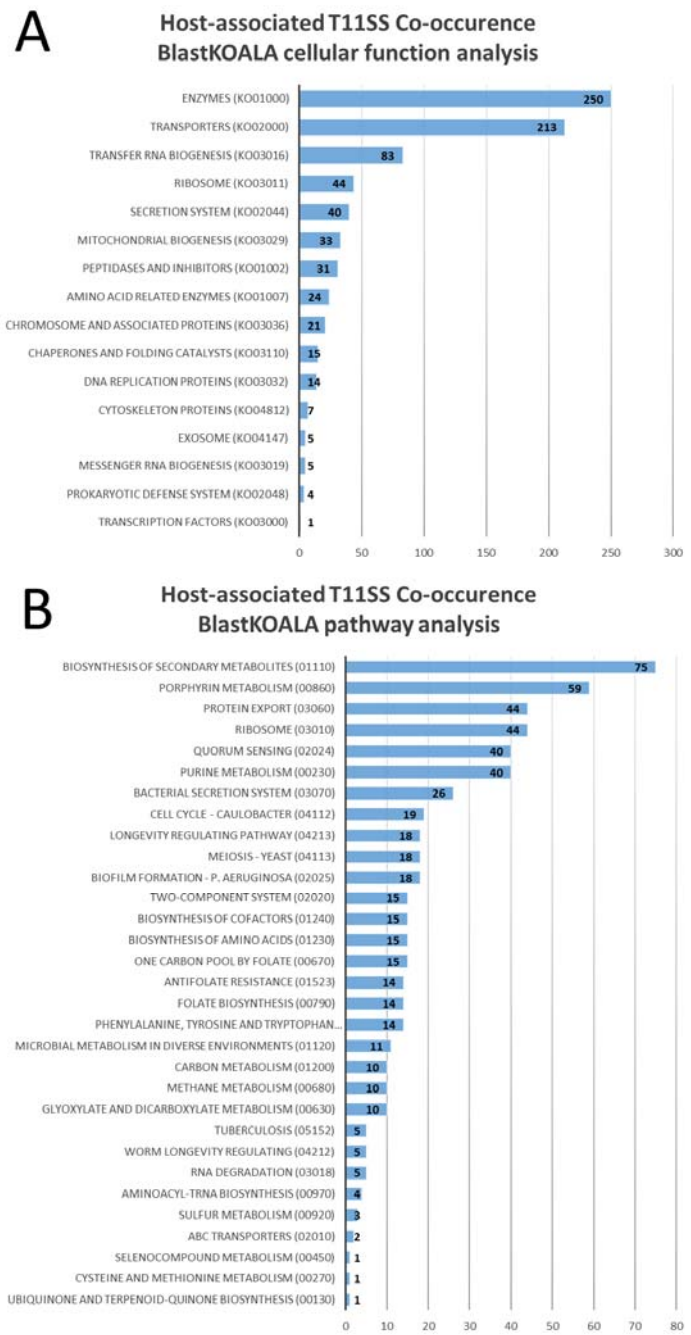


Figure 2. **Host-associated T11SS genome neighborhoods reveal conserved association with iron/heme uptake, export, and one-carbon metabolism**. KofamKOALA uses hidden Markov models to assign functions to query sequences and reveal shared pathways. Cellular functions were estimated using brite hierarchies and assigned to known pathways.



Supplemental Figure 2. **BlastKOALA supports T11SS association with iron/heme uptake, protein export, and one-carbon metabolism**. BlastKOALA uses the BLAST alignment algorithm to assign functions to query sequences and reveal shared pathways. Cellular functions were estimated using brite hierarchies and assigned to known pathways.

Figure 3. **Secondary structure of predicted T11SS-dependent cargo reveals diverse and novel gene architectures and effector domains**. PsiPred 4 and DisoPred 3 use amino acid sequence to annotate secondary structures highlighting a diversity of N-terminal domains with proline rich repeats, intrinsically disordered regions, α-helical repeats, and ligand binding handles. For each of the detected cargo architectures the representative sequence from the largest UniRef50 cluster was trimmed of it signal peptide to reflect a mature protein.

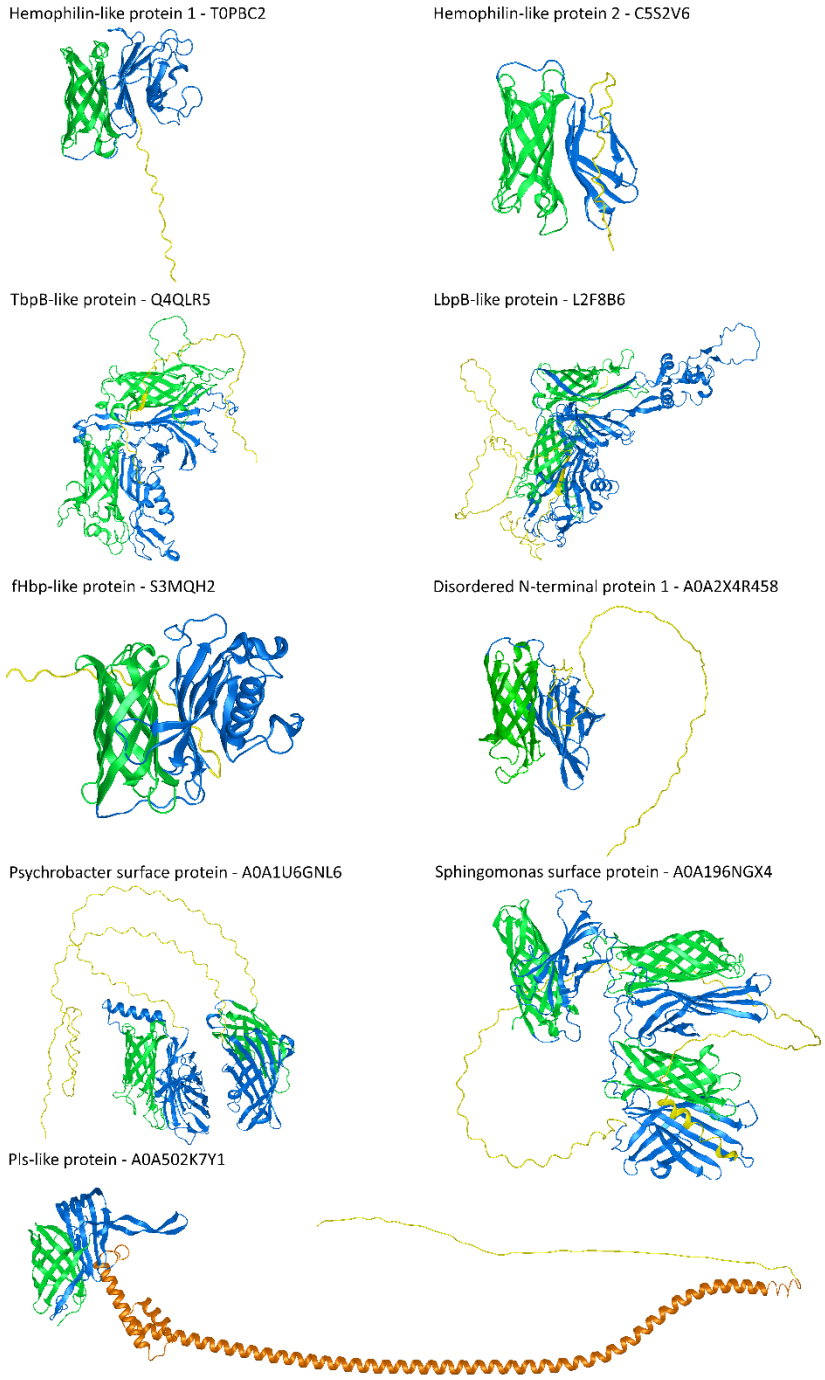


Figure 4. **AlphaFold2 annotation of predicted T11SS-dependent cargo reveals conservation of C-terminal or centrally located β-barrel domains**. For each of the detected cargo architectures the representative sequence from the largest UniRef50 cluster was used to predict a representative structure. Green residues indicate C-terminal β-barrel domains. Blue residues indicate putative ligand binding handle domains. Orange residues indicate glycosylation motif repeats within the Pls-like protein architecture. Yellow residues indicate regions predicted to be disordered.

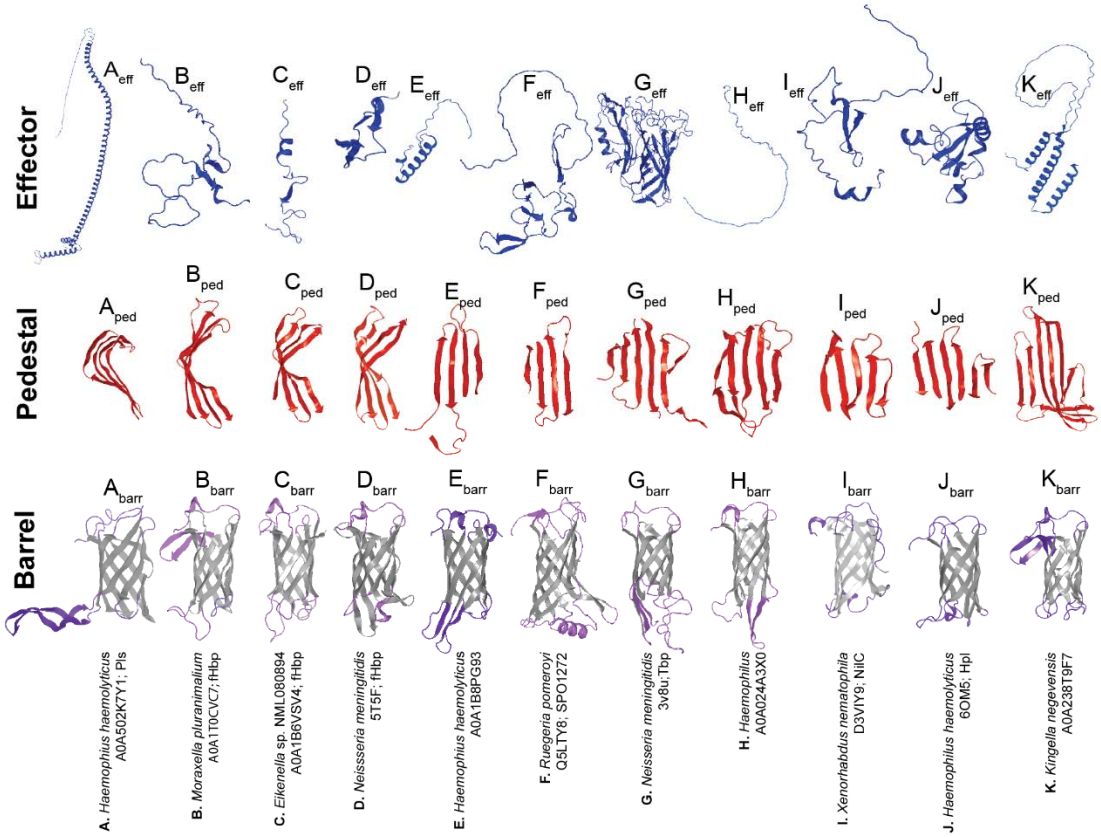
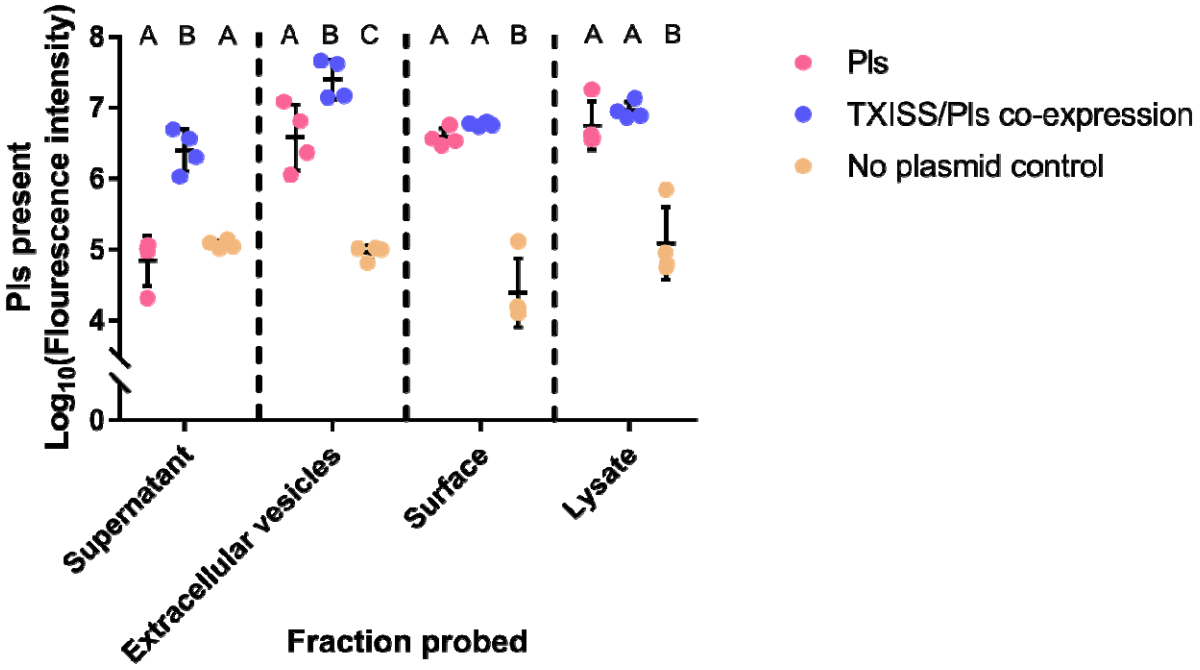
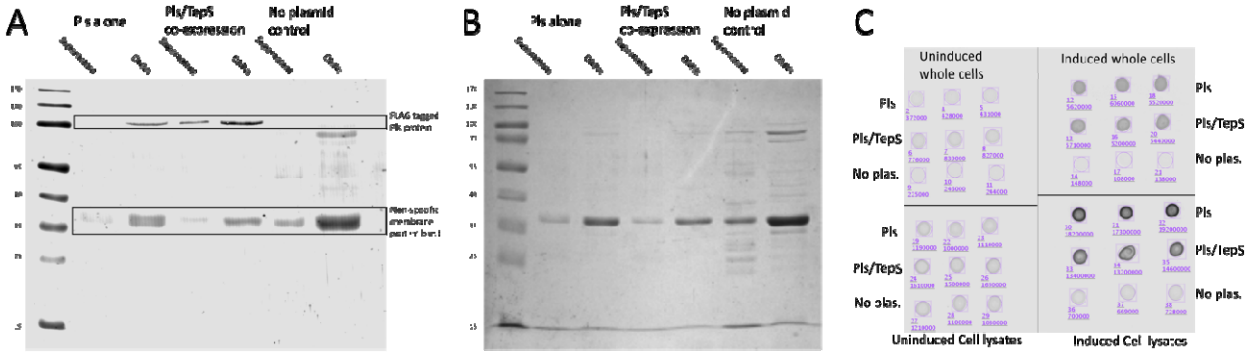


Figure 5. **AlphaFold2 annotation of predicted T11SS-dependent cargo reveals three distinct domains ubiquitous to cargo.** Solved **(D, G, J)** or predicted **(A, B, C, E, F, H, I, K)** structures from representative UniRefF50 cluster sequences of predicted **(A, B, D, E, F, H, K)** or known cargo **(D, G, I, J)**. Each structure was separated into three sections representing common structural domains: N-terminal effector (blue), pedestal (red), and C-terminal β-barrel (gray barrel residues and purple loop residues). N-terminal effector domains varied in their predicted structural motifs and included disordered **(B, C, D, H, I)**, α-helical **(A, E, K)**, β-strand **(D, F, J)**, and barrel **(G)**. Pedestal domains comprised 4-stranded groove **(A)**, 4-stranded "wing" **(B-D),** 4-stranded sheet **(E, F)**, 5-stranded sheet **(G-I)**, or 6-stranded/extended β-sheet **(J, K)**. Some of these β-strands are donated by primary sequences from either the N-terminus or the C-terminus of the protein. Barrel domains exhibited variation in the presence of extended loops, helices, or strands in predicted surface- or periplasmic-exposed regions, as indicated by purple coloration.



963

964 Figure 6. **Localization of FLAG tagged Pls in the absence and presence of its cognate T11SS**
965 **secretor.** *E. coli* cells expressing Pls-FLAG alone or Pls-FLAG alongside its cognate T11SS (TepS)
966 were probed using dot blots (cell surfaces and lysates) or Western blotting (supernatant and
967 extracellular vesicles) to see how expression of the T11SS protein impacted protein localization.
968 Supernatant secretion of Pls was only detectable in the presence of the T11SS TepS. The
969 extracellular vesicle fraction collected via ultracentrifugation showed some Pls localization in the
970 absence of TepS, however that localization was significantly increased by TepS. Pls was detected on
971 the surface of both treatments equally, and cell lysates showed that the level of Pls expression
972 between the treatments was equivalent.
973



Supplemental Figure 3. **Localization of FLAG tagged Pls in the absence and presence of its cognate T11SS secretor.** *E. coli* cells expressing Pls-FLAG alone or Pls-FLAG alongside its cognate T11SS (TepS) were probed using dot blots (cell surfaces and lysates) or Western blotting (supernatant and extracellular vesicles) to see how expression of the T11SS protein impacted protein localization. **(A)** a representative Western blot demonstrating TepS dependent secretion of Pls. **(B)** A representative Coomassie stain showing total supernatant protein content. **(C)** Immuno-dot blots demonstrating that TepS was not required for surface exposure of Pls when expressed in *E. coli.*

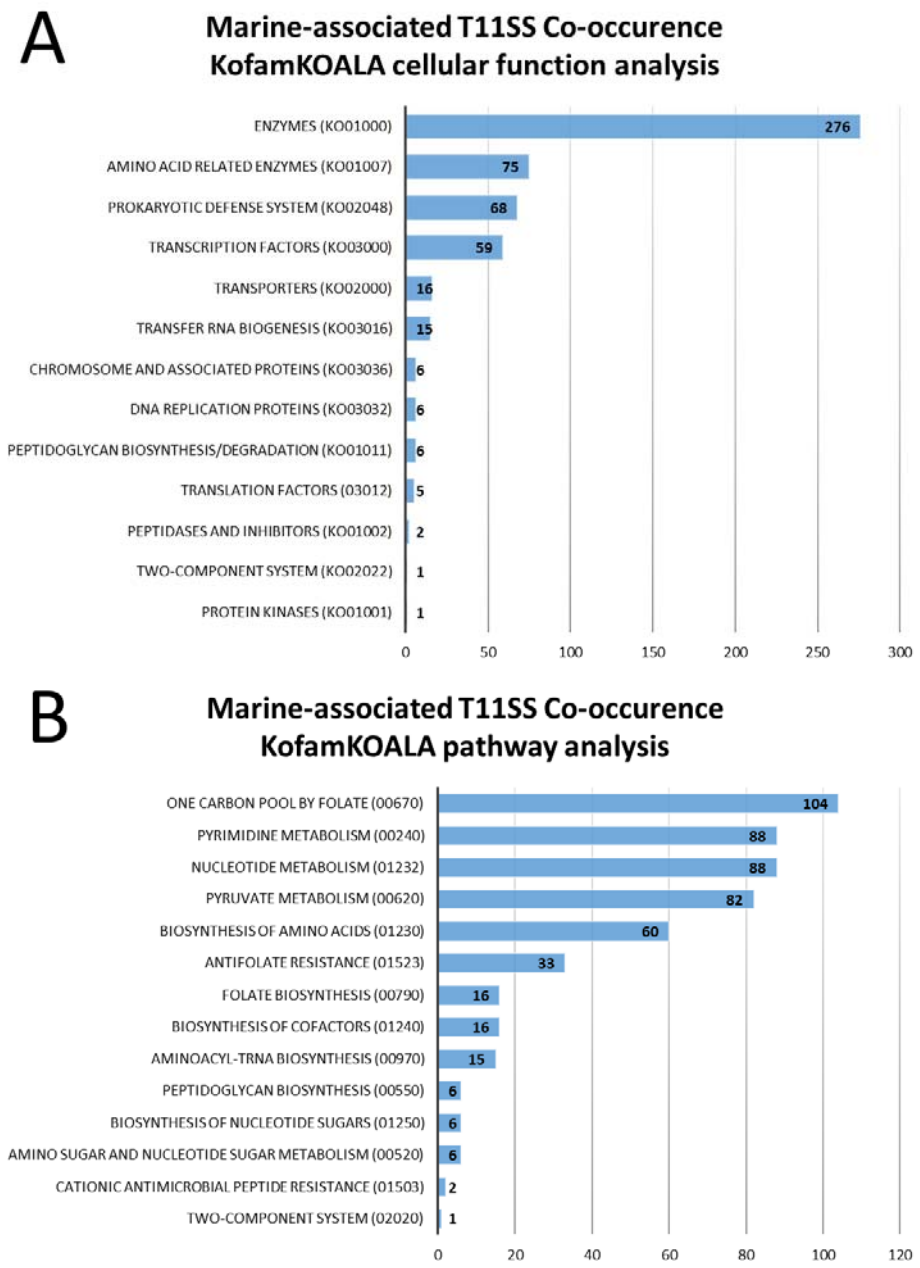


Figure 7. **Marine T11SS genome neighborhoods reveal association with one-carbon metabolism, nucleotide metabolism, and the glyoxalase-detoxification pathway**. KofamKOALA uses hidden Markov models to assign functions to query sequences and reveal shared pathways. Cellular functions were estimated using brite hierarchies and assigned to known pathways.

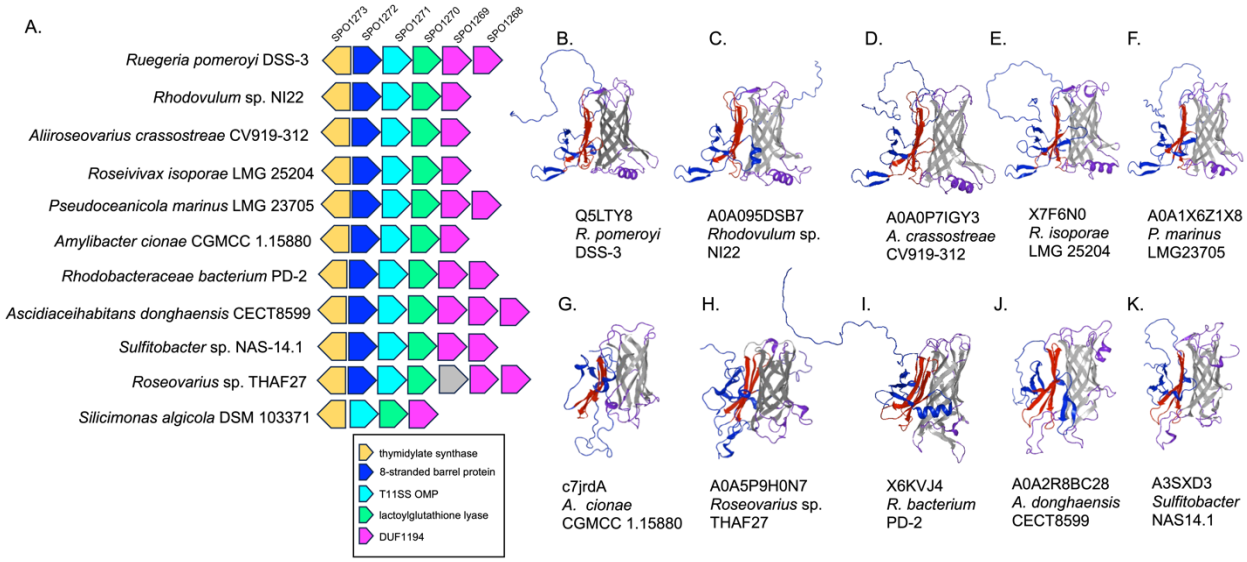
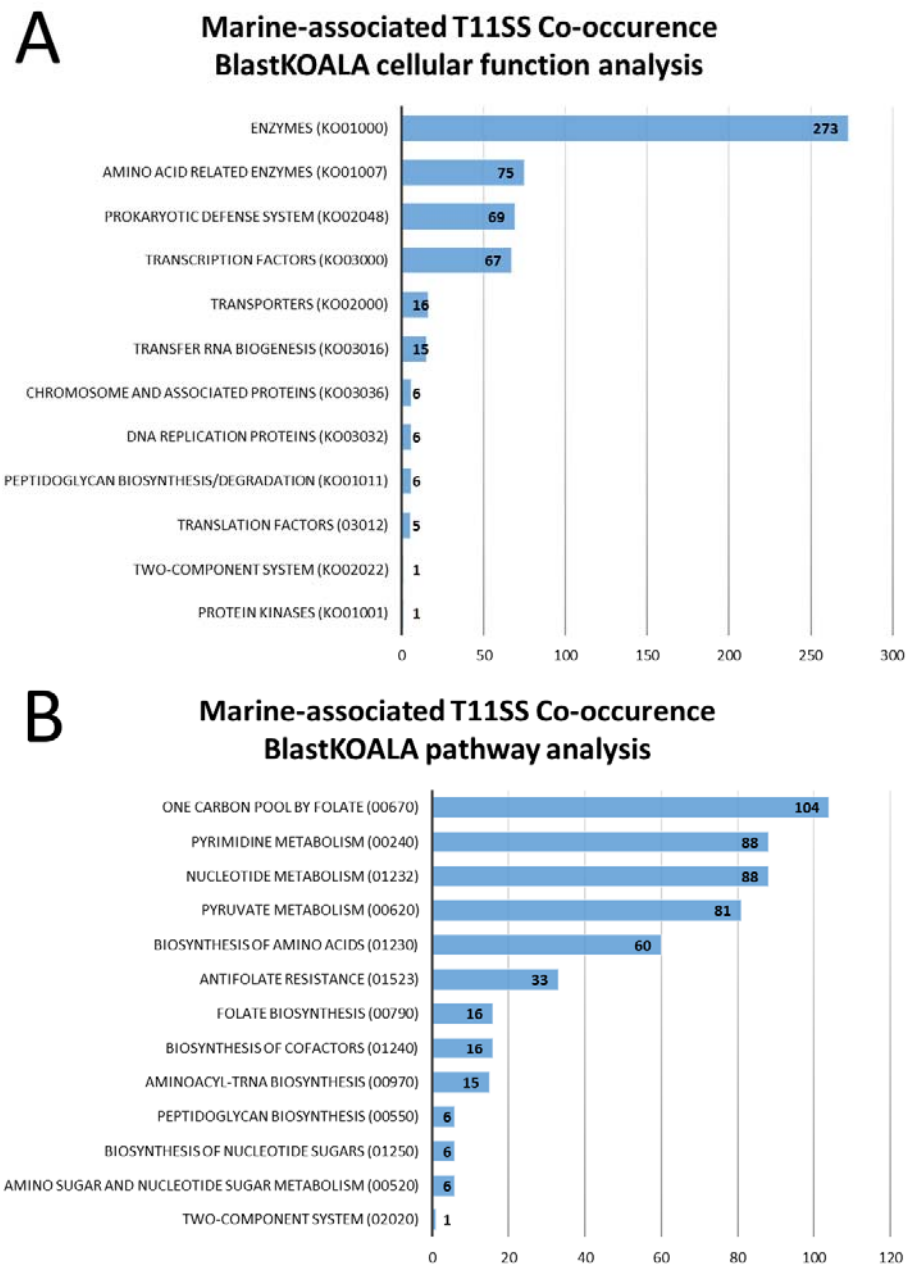


Fig. 8. **A conserved marine cluster 3, T11SS OMP locus encodes predicted cargo proteins.** A schematic representation of select examples of a T11SS OMP-encoding, syntenic locus found with genomes of indicated strains of *Rhodobacteraceae* and *Roseobacteriaceae* **(A)**. Each gene is represented by a block arrow (not to scale), and the *Ruegeria pomeroyi* locus tags of each homolog are indicated at the top. In all 11 genomes examined, the syntenic locus included genes predicted to encode a T11SS OMP (turquoise arrows) as well as the top cluster 3 co-occurring functions: flavin-dependent thymidylate synthase (gold arrows), lactoylglutathione lyase (neon green arrows), and one or more DUF1194/von Willebrand factor proteins (pink arrows). In all but one, the locus includes a putative T11SS cargo protein (light blue arrows), identified based on the presence of a predicted C-terminal 8-stranded β-barrel domain. Predicted Alphafold2 or Phyre2 predicted structures of putative cargo proteins encoded within these loci is shown in **(B-K)**. Color coding is used to indicate the putative effector (dark blue), pedestal (red), barrel (gray), and surface or periplasmic loops (purple) of each protein. The top five appear to have similar structural motifs, including an α-helix in one of the β-barrel loops, antiparallel β-strands in the N-terminal domain (blue), and a 4-stranded sheet pedestal domain (red). The remaining five structures show the overall domains of N-terminal (blue), pedestal (red) and β-barrel (gray), but the N-terminal domains are predicted to have different structural motifs relative to the others examined.



Supplemental Figure 4. **BlastKOALA supports T11SS association with one-carbon metabolism, nucleotide metabolism, and the glyoxalase-detoxification pathway**. BlastKOALA uses the BLAST alignment algorithm to assign functions to query sequences and reveal shared pathways. Cellular functions were estimated using brite hierarchies and assigned to known pathways.

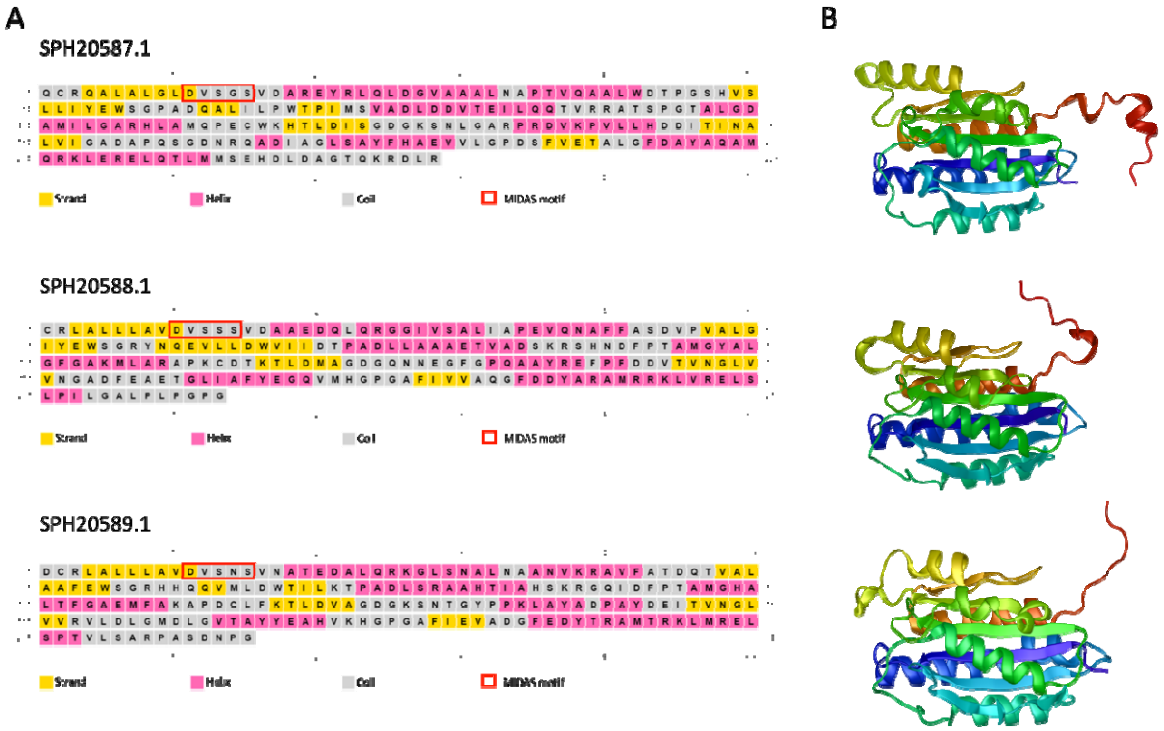
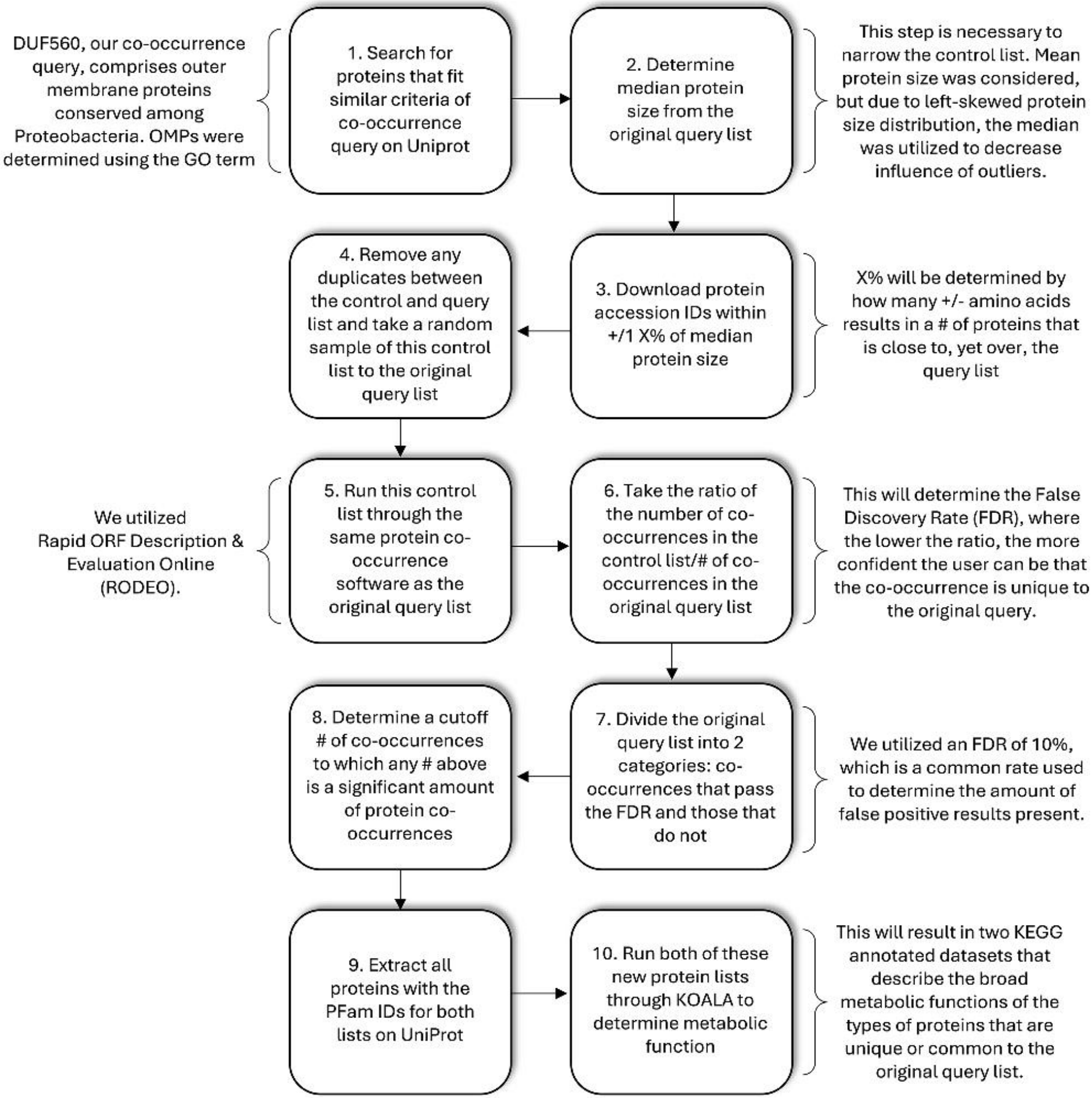


Figure 9 **Representative structures of DUF1194 proteins located within a T11SS genomic neighborhood in *Ascidiaceihabitans donghaensis***. These three DUF1194 homologs are encoded adjacent to a T11SS OMP-encoding gene within a microbe isolated from a tunicate host. These proteins appear to be paralogs of each other based on sequence similarity. Signal peptides were predicted with SignalP 6 and trimmed from the sequences prior to submission to PsiPred **(A)** and RoseTTAFold **(B).**

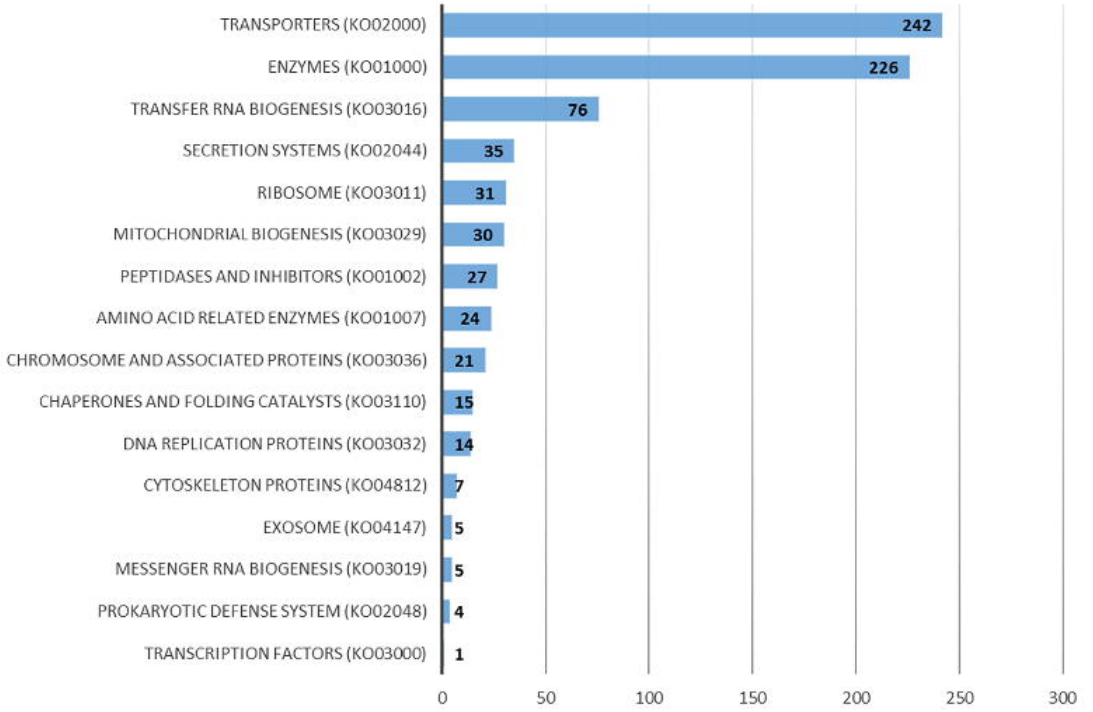
1023 Supplemental Files:

1024 Supplemental File 1: **T11SSGenomicCo-occurrence.xlsx**
1025 Controlled co-occurrence analysis of all domains and genes significantly co-occurring with T11SS in
1026 a cluster of host-associated organisms (cluster 1) or a cluster of marine organisms (cluster 3).
1027 Significant genes were then clustered into UniRef50 clusters to combine close homologs and mined
1028 for predicted T11SS cargo proteins via homology to known cargo proteins.
1029
1030 Supplemental File 2: **Co-occurrenceFunctionalAnnotations.xlsx**
1031 KEGG based analysis genes significantly co-occurring with host associated T11SS (cluster 1) and a
1032 marine T11SS (cluster 3).
1033
1034 Supplemental File 3: **Cluster3CargoLoci.xlsx**
1035 Selected Cluster 3 predicted cargo proteins co-occurring with T11SS.
1036
1037 Supplemental File 4: **T11SS-DUF1194Co-occurrence.xlsx**
1038 Co-occurrence of Marine associated T11SS alongside DUF1194 domain encoding genes, color
1039 coded to indicate host-associated organisms.


DUF560, our co-occurrence query, comprises outer membrane proteins conserved among Proteobacteria. OMPs were determined using the GO term
1. Search for proteins that fit similar criteria of co-occurrence query on Uniprot
2. Determine median protein size from the original query list
This step is necessary to narrow the control list. Mean protein size was considered, but due to left-skewed protein size distribution, the median was utilized to decrease influence of outliers.
3. Download protein accession IDs within +/1 X% of median protein size
X% will be determined by how many +/- amino acids results in a # of proteins that is close to, yet over, the query list
4. Remove any duplicates between the control and query list and take a random sample of this control list to the original query list
We utilized Rapid ORF Description & Evaluation Online (RODEO).
5. Run this control list through the same protein co-occurrence software as the original query list
6. Take the ratio of the number of co-occurrences in the control list/# of co-occurrences in the original query list
This will determine the False Discovery Rate (FDR), where the lower the ratio, the more confident the user can be that the co-occurrence is unique to the original query.
7. Divide the original query list into 2 categories: co-occurrences that pass the FDR and those that do not
We utilized an FDR of 10%, which is a common rate used to determine the amount of false positive results present.
8. Determine a cutoff # of co-occurrences to which any # above is a significant amount of protein co-occurrences
9. Extract all proteins with the PFam IDs for both lists on UniProt
10. Run both of these new protein lists through KOALA to determine metabolic function
This will result in two KEGG annotated datasets that describe the broad metabolic functions of the types of proteins that are unique or common to the original query list.

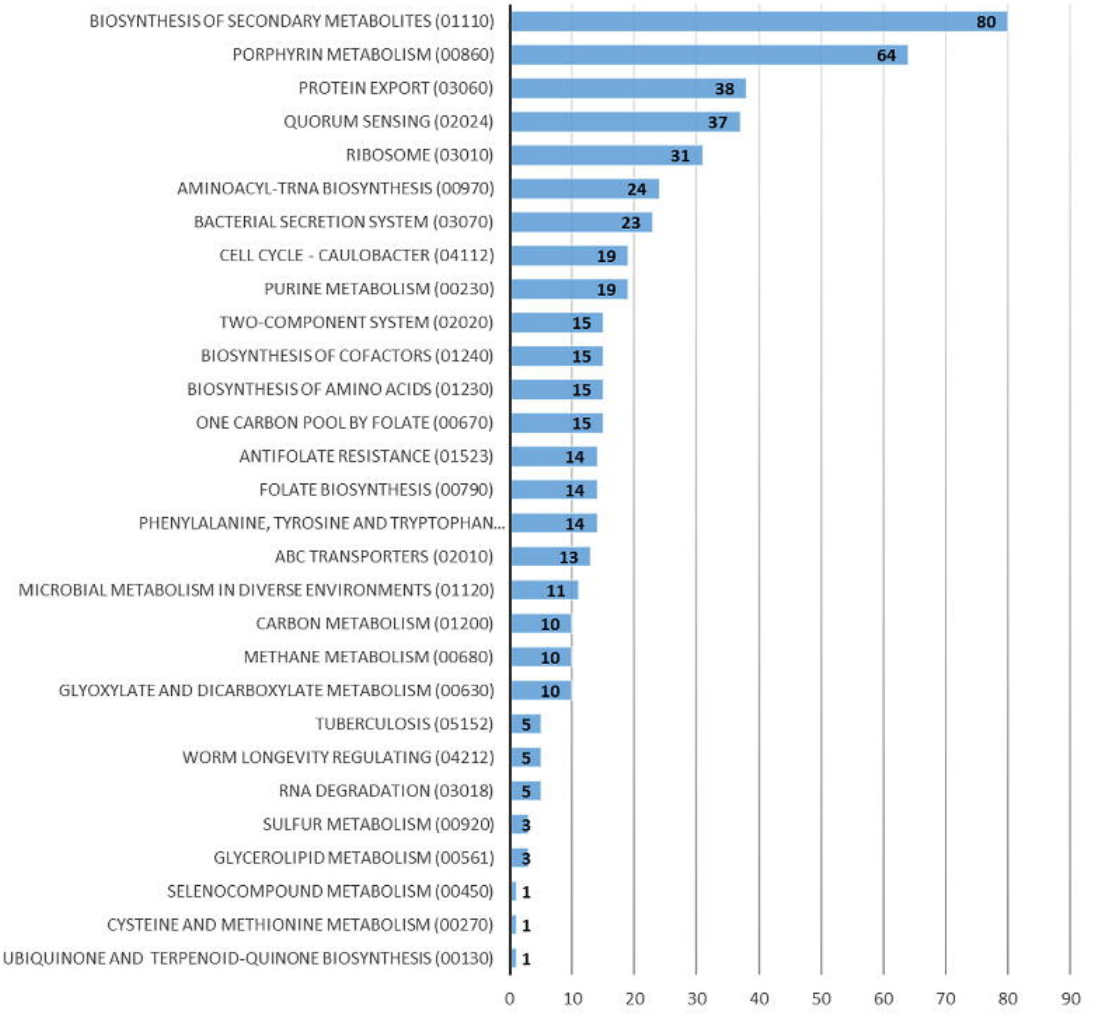
A

Host-associated T11SS Co-occurence KofamKOALA cellular function analysis



B

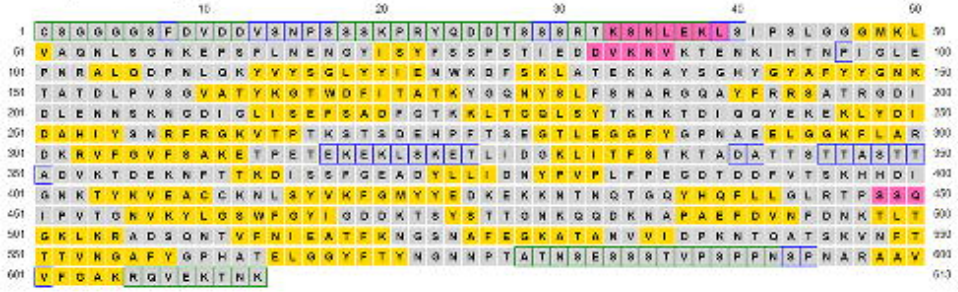
Host-associated T11SS Co-occurence KofamKOALA pathway analysis



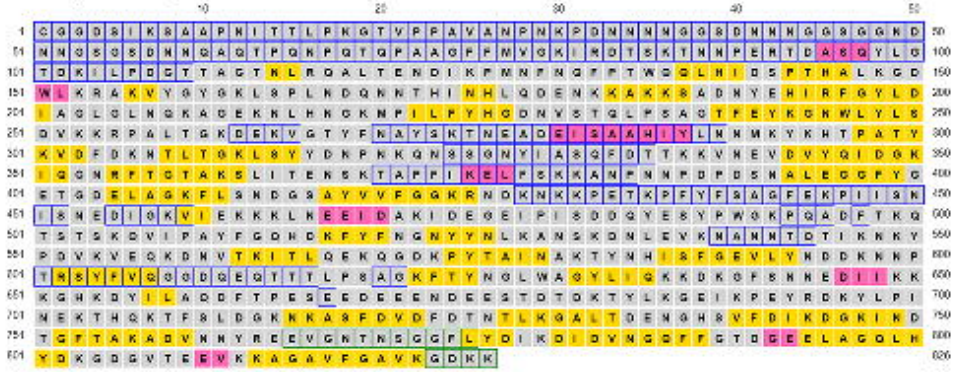
Hemophilin-like protein 1 - T0PBC2

Hemophilin-like protein 2 - C5S2V6

TbpB-like protein - Q4QLR5



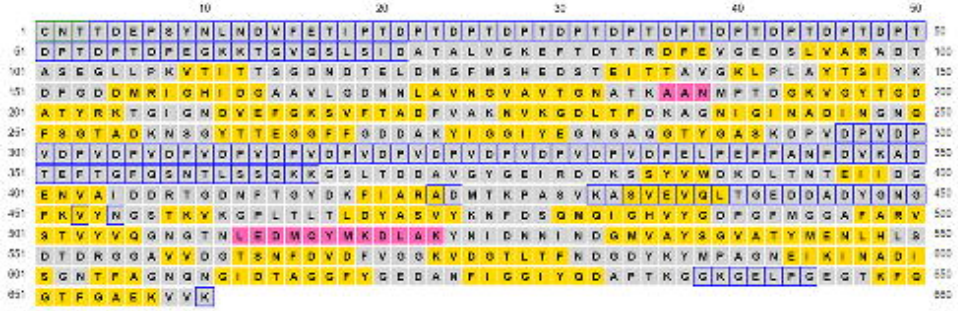
LbpB-like protein - L2F8B6



fHbp-like protein - S3MQH2

Disordered N-terminal protein 1 - A0A2X4R458

Psychrobacter surface protein - A0A1U6GNL6



Sphingomonas surface protein - A0A196NGX4

Pls-like protein - A0A502K7Y1

Strand

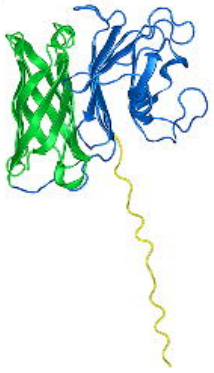
Helix

Coil

Disordered

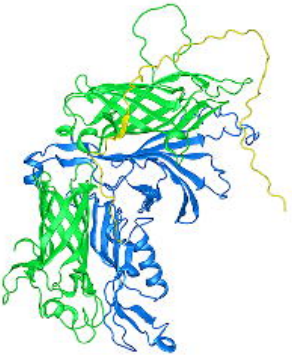
Disordered, protein binding

Hemophilin-like protein 1 - T0PBC2

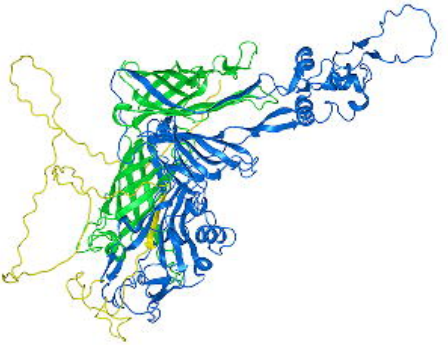


Hemophilin-like protein 2 - C5S2V6

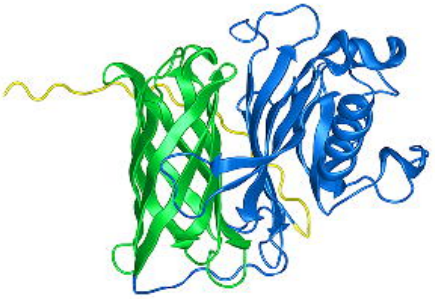
TbpB-like protein - Q4QLR5



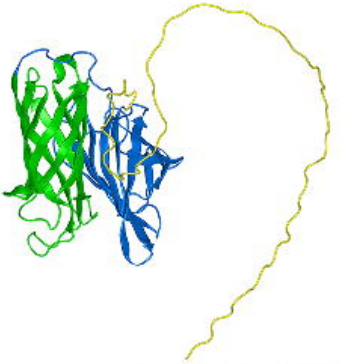
LbpB-like protein - L2F8B6



fHbp-like protein - S3MQH2



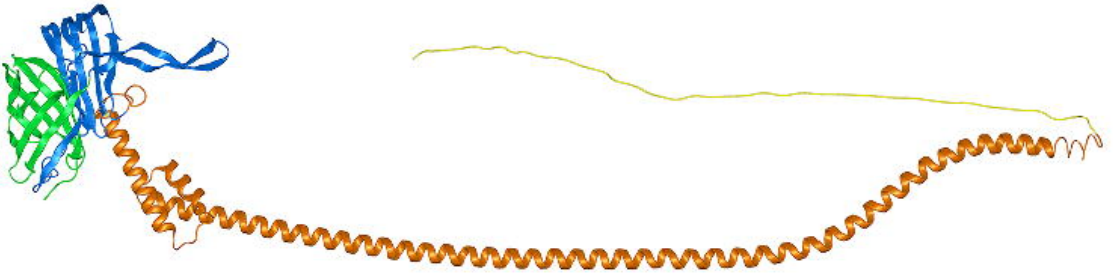
Disordered N-terminal protein 1 - A0A2X4R458



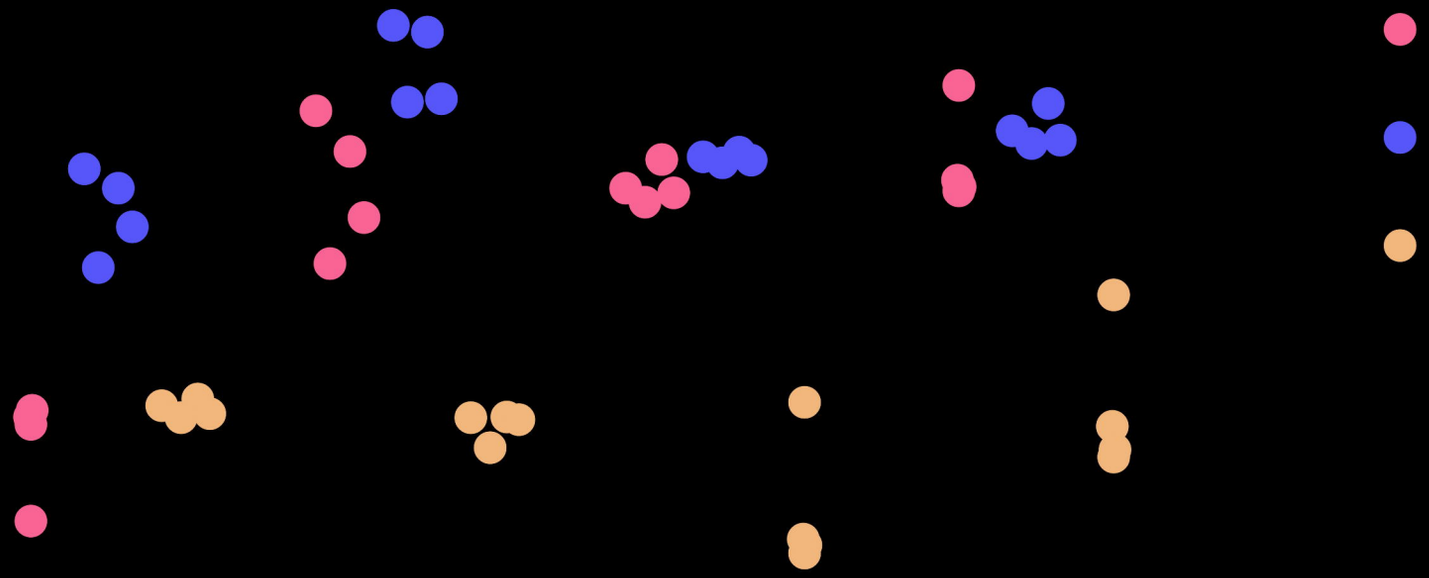
Psychrobacter surface protein - A0A1U6GNL6

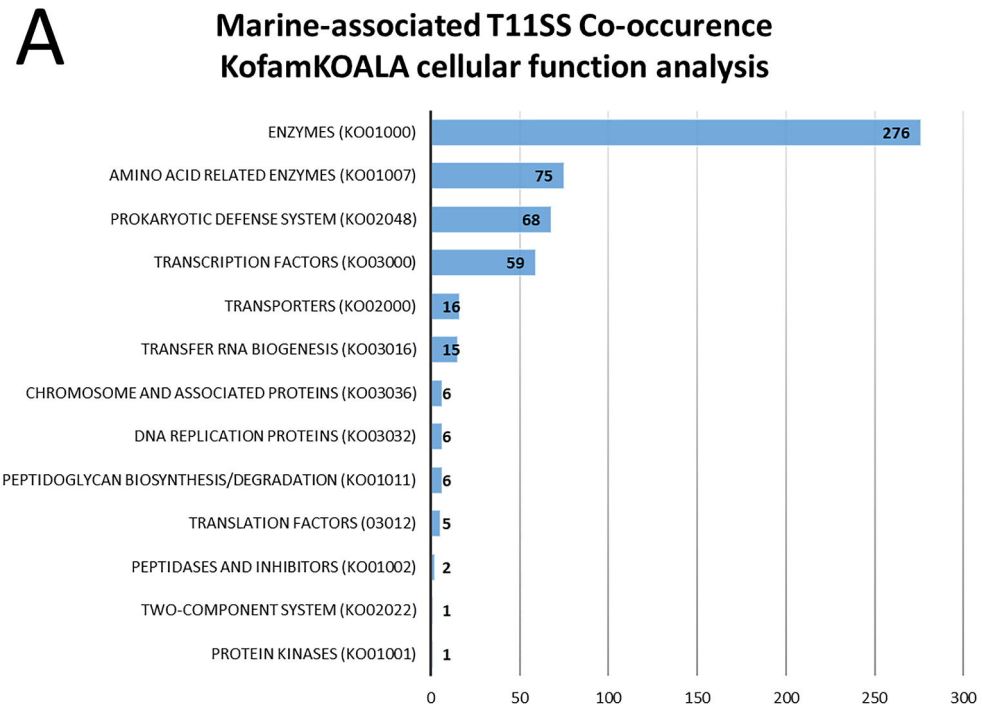
Sphingomonas surface protein - A0A196NGX4

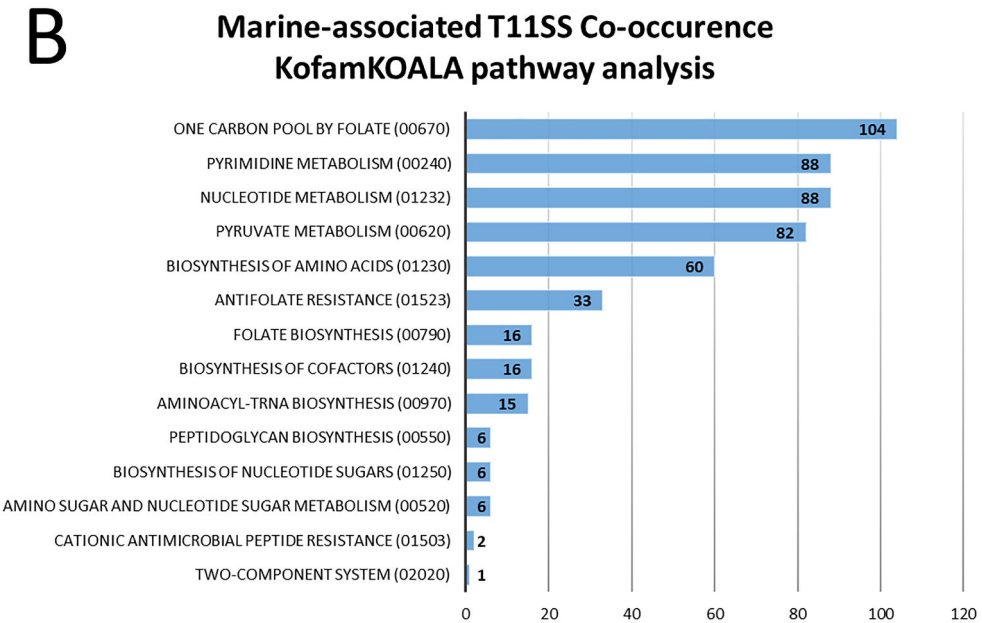
Pls-like protein - A0A502K7Y1

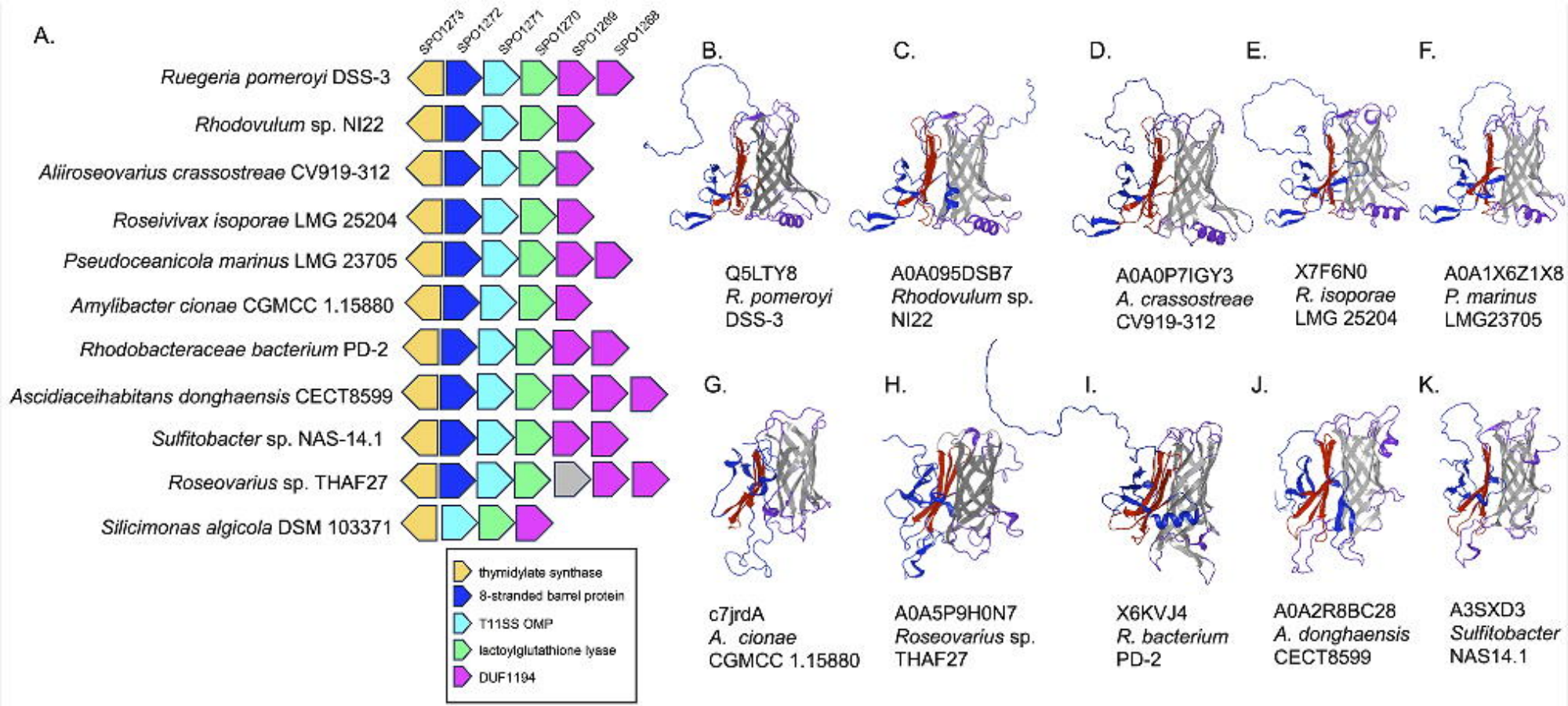


Effector
Pedestal
Barrel
A_eff
B_eff
C_eff
D_eff
E_eff
F_eff
G_eff
H_eff
I_eff
J_eff
K_eff
A_ped
B_ped
C_ped
D_ped
E_ped
F_ped
G_ped
H_ped
I_ped
J_ped
K_ped
A_barr
B_barr
C_barr
D_barr
E_barr
F_barr
G_barr
H_barr
I_barr
J_barr
K_barr
A. Haemophius haemolyticus A0A502K7Y1; Pls
B. Moraxella pluranimalium A0A1T0CVC7; fHbp
C. Eikenella sp. NML080894 A0A1B6VSV4; fHbp
D. Neissseria meningitidis 5T5F; fHbp
E. Haemophius haemolyticus A0A1B8PG93
F. Ruegeria pomeroyi Q5LTY8; SPO1272
G. Neisseria meningitidis 3v8u;Tbp
H. Haemophilus A0A024A3X0
I. Xenorhabdus nematophila D3VIY9; NilC
J. Haemophilus haemolyticus 6OM5; Hpl
K. Kingella negevensis A0A238T9F7

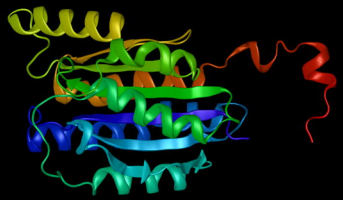



A
Marine-associated T11SS Co-occurence
KofamKOALA cellular function analysis
ENZYMES (KO01000)
276
AMINO ACID RELATED ENZYMES (KO01007)
75
PROKARYOTIC DEFENSE SYSTEM (KO02048)
68
TRANSCRIPTION FACTORS (KO03000)
59
TRANSPORTERS (KO02000)
16
TRANSFER RNA BIOGENESIS (KO03016)
15
CHROMOSOME AND ASSOCIATED PROTEINS (KO03036)
6
DNA REPLICATION PROTEINS (KO03032)
6
PEPTIDOGLYCAN BIOSYNTHESIS/DEGRADATION (KO01011)
6
TRANSLATION FACTORS (03012)
5
PEPTIDASES AND INHIBITORS (KO01002)
2
TWO-COMPONENT SYSTEM (KO02022)
1
PROTEIN KINASES (KO01001)
1
0
50
100
150
200
250
300

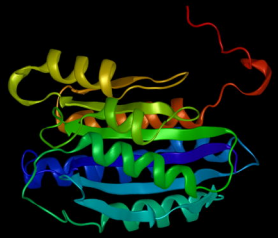
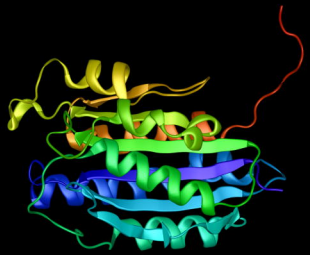

B
Marine-associated T11SS Co-occurence
KofamKOALA pathway analysis
ONE CARBON POOL BY FOLATE (00670)
104
PYRIMIDINE METABOLISM (00240)
88
NUCLEOTIDE METABOLISM (01232)
88
PYRUVATE METABOLISM (00620)
82
BIOSYNTHESIS OF AMINO ACIDS (01230)
60
ANTIFOLATE RESISTANCE (01523)
33
FOLATE BIOSYNTHESIS (00790)
16
BIOSYNTHESIS OF COFACTORS (01240)
16
AMINOACYL-TRNA BIOSYNTHESIS (00970)
15
PEPTIDOGLYCAN BIOSYNTHESIS (00550)
6
BIOSYNTHESIS OF NUCLEOTIDE SUGARS (01250)
6
AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM (00520)
6
CATIONIC ANTIMICROBIAL PEPTIDE RESISTANCE (01503)
2
TWO-COMPONENT SYSTEM (02020)
1
0
20
40
60
80
100
120


A.
SPO1273
SPO1272
SPO1271
SPO1270
SPO1269
SPO1268
Ruegeria pomeroyi DSS-3
Rhodovulum sp. NI22
Aliiroseovarius crassostreae CV919-312
Roseivivax isoporae LMG 25204
Pseudoceanicola marinus LMG 23705
Amylibacter cionae CGMCC 1.15880
Rhodobacteraceae bacterium PD-2
Ascidiaceihabitans donghaensis CECT8599
Sulfitobacter sp. NAS-14.1
Roseovarius sp. THAF27
Silicimonas algicola DSM 103371
thymidylate synthase
8-stranded barrel protein
T11SS OMP
lactoylglutathione lyase
DUF1194
B.
Q5LTY8
R. pomeroyi
DSS-3
C.
A0A095DSB7
Rhodovulum sp.
NI22
D.
A0A0P7IGY3
A. crassostreae
CV919-312
E.
X7F6N0
R. isoporae
LMG 25204
F.
A0A1X6Z1X8
P. marinus
LMG23705
G.
c7jrdA
A. cionae
CGMCC 1.15880
H.
A0A5P9H0N7
Roseovarius sp.
THAF27
I.
X6KVJ4
R. bacterium
PD-2
J.
A0A2R8BC28
A. donghaensis
CECT8599
K.
A3SXD3
Sulfitobacter
NAS14.1

```
QCRQALALGLDVSGSVDAREYRLQLDGVAAALNAPTVQAALWDTPGSHVS
LLIYEWSGPADQALILPWTPIMSVADLDDVTEILQQTVRRATSPGTALGD
AMILGARHLAMQPECWKHTLDISGDGKSNLGARPRDVKPVLLHDDITINA
LVIGADAPQSGDNRQADIAGLSAYFHAEVVLGPDSFVETALGFDAYAQAM
QRKLERELQTLMMSEHDLDAGTQKRDLR
```



```
CRLALLLAVDVSSSVDAAEDQLQRGGIVSALIAPEVQNAFFASDVPVALG
IYEWSGRYNQEVLLDWVIIDTPADLLAAAETVADSKRSHNDFPTAMGYAL
GFGAKMLARAPKCDTKTLDMAGDGQNNEGFGPQAAYREFPFDDVTVNGLV
VNGADFEAETGLIAFYEGQVMHGPGAFIVVAQGFDDYARAMRRKLVRELS
LPILGALPLPGPG
```



```
DCRLALLLAVDVSNSVNATEDALQRKGLSNALNAANVKRAVFATDQTVAL
AAFEWSGRHHQQVMLDWTILKTPADLSRAAHTIAHSKRGQIDFPTAMGHA
LTFGAEMFAKAPDCLFKTLDVAGDGKSNTGYPPKLAYADPAYDEITVNGL
VVRVLDLGMDLGVTAYYEAHVKHGPGAFIEVADGFEDYTRAMTRKLMREL
SPTVLSARPASDNPG
```