



Evolution of the Natural Transformation Protein, ComEC, in Bacteria

Zachary T. Pimentel and Ying Zhang*

Department of Cell and Molecular Biology, College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI, United States

OPEN ACCESS

Edited by:

Baolei Jia,
Chung-Ang University, South Korea

Reviewed by:

R. Thane Papke,
University of Connecticut,
United States
Jens Kreth,
Oregon Health & Science University,
United States

*Correspondence:

Ying Zhang
yingzhang@uri.edu

Specialty section:

This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 03 September 2018

Accepted: 19 November 2018

Published: 12 December 2018

Citation:

Pimentel ZT and Zhang Y (2018)
Evolution of the Natural
Transformation Protein, ComEC,
in Bacteria. *Front. Microbiol.* 9:2980.
doi: 10.3389/fmicb.2018.02980

Natural transformation enables the incorporation of exogenous DNA into host genomes and plays a fundamental role in the evolution of microbial populations. At the center of the natural transformation machinery, the ComEC protein mediates DNA import and serves potential functions in DNA recognition and single strand degradation. Despite its importance, the evolution of ComEC is not fully understood. Here, we aim to fill this knowledge gap by surveying putative ComEC proteins across 5,574 bacteria that span diverse phyla. We first derived the presence of a universal, core *Competence* domain through the analysis of ComEC proteins from known naturally competent species. Then, we followed this observation to identify *Competence* domain containing proteins (CDCPs) from all bacteria and used CDCPs as putative ComEC proteins for evolutionary analysis. A near universal presence of CDCPs was revealed, with 89% of the proteomes and 96% of the genomes encoding a single CDCP or a CDCP-like fragment. Two domains, *DUF4131* and *Lactamase_B*, were found to commonly co-occur with the *Competence* domain. Ancestral state reconstruction of CDCPs over the bacterial species phylogeny suggested an origin of a *Competence*-only domain profile, while multiple gains and losses of the *DUF4131* and *Lactamase_B* domains were observed among diverse bacterial lineages.

Keywords: natural transformation, natural competence, ComEC, horizontal gene transfer, *Competence*, *DUF4131*, *Lactamase_B*

INTRODUCTION

Natural transformation (NT) is a process by which exogenous DNA is imported and incorporated into a genome. It is an important mechanism of horizontal gene transfer and as such enables DNA repair and the acquisition of new functions, such as antibiotic resistance and nutrition utilization (Ambur et al., 2016). Due to the diverse regulatory mechanisms involved in NT, natural competence is often difficult to initiate in laboratory conditions (Dubnau, 1991; Meibom et al., 2005). To date, only around 80 naturally competent bacterial strains have been confirmed in the literature (Lorenz and Wackernagel, 1994; Johnsborg et al., 2007; Johnston et al., 2014).

The NT process is mediated by a complex machinery of proteins via multiple steps. In gram-negative bacteria, initiation of outer membrane double-stranded DNA (dsDNA) uptake into the periplasmic space involves proteins related to components of the type IV pilus, the type II secretion system (Chen and Dubnau, 2004), or the type IV secretion system

(Hofreuter et al., 2001). The dsDNA then binds to ComEA, which is hypothesized to pull the transforming DNA into the periplasm and is present in both gram-positive and gram-negative species (Bergé et al., 2002; Matthey and Blokesch, 2016). One strand of the dsDNA is then subject to degradation, for example, by the nuclease EndA in *Streptococcus pneumoniae* (Puyet et al., 1990) or by a putative nuclease domain of the ComEC protein in *Bacillus subtilis* (Baker et al., 2016). The single-stranded DNA (ssDNA) retained from degradation of its complement strand is then transported by a conserved ComEC protein into the cytosol (Draskovic and Dubnau, 2005; Baker et al., 2016). In a following step, the cytosolic ssDNA is recognized by DprA and recombined into the genomic DNA by RecA (Mortier-Barrière et al., 2007).

The composition of NT machinery varies among naturally competent species (Johnston et al., 2014). Several proteins, such as RecA and PilA, have known functions outside of the NT process (Taha et al., 1992; Cox, 2007). The DNA Processing Protein A (DprA) and the Competence protein EC (ComEC), however, are shown to be essential for NT (Friedrich et al., 2001; Yeh et al., 2003; Wilharm et al., 2013; Mell and Redfield, 2014). While DprA is responsible for binding ssDNA in the cytosol (Mortier-Barrière et al., 2007; Quevillon-Cheruel et al., 2012), ComEC is predicted to mediate multiple steps including the DNA binding, single strand degradation (to expose the other strand), and ssDNA membrane translocation (Baker et al., 2016). Further, while homologs of DprA and RecA are present in archaea, ComEC appears to be absent among the few naturally transformable archaeal species, indicating distinct DNA import mechanisms between bacteria and archaea (Lipscomb et al., 2011; van Wolferen et al., 2016).

ComEC proteins have been identified among diverse bacterial phyla, including candidate phyla that lack cultured representatives (Kantor et al., 2013; Mell and Redfield, 2014). The ComEC protein of *B. subtilis* contains three Pfam domains, a domain of unknown function (*DUF4131*), a transmembrane competence domain (*Competence*), and a metallo-beta-lactamase domain (*Lactamase_B*), ordered accordingly from the N-terminus to the C-terminus of the protein (Baker et al., 2016). The *DUF4131* domain contains a putative OB fold that is predicted to have functions in nucleic acids binding (Baker et al., 2016), the *Competence* domain includes a set of core transmembrane helices that mediate the uptake of ssDNA (Draskovic and Dubnau, 2005), and the *Lactamase_B* domain belongs to a broad family of DNA and RNA nucleases that could be responsible for the degradation of a single strand in the dsDNA (Callebaut, 2002; Dominski, 2007; Baker et al., 2016).

While studies have been performed on the structure, function, and regulation of ComEC proteins among a small number of competent bacteria, little is known about the taxonomic distribution and domain evolution of this protein across a broader range of bacterial species. To fill this knowledge gap, here we report the identification of putative ComEC proteins (referred to as CDCPs) among 5,574 bacteria and provide a first ancestral state reconstruction of the protein among diverse bacterial lineages.

MATERIALS AND METHODS

Collection and Domain Analysis of Known ComEC Proteins

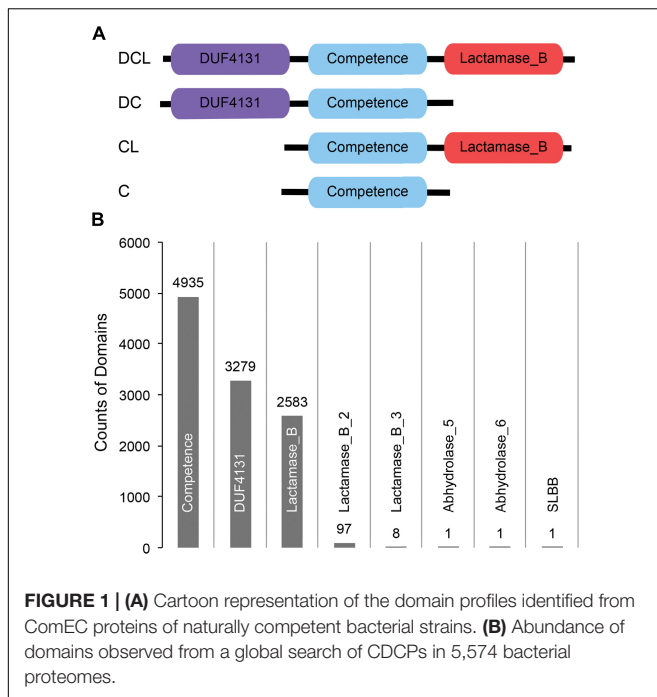
A collection of ComEC proteins from strains experimentally confirmed to be capable of NT were obtained from the literature (**Supplementary Table S1**). The ComEC proteins were compared against the Pfam-A database, version 27.0 (Finn et al., 2016) using the *hmmScan* function in the HMMER package, version 3.1b2 (Eddy, 1998). Domain hits with an e-value of greater than 1×10^{-5} and a coverage of less than 70% of the Pfam-A models were removed. The remaining domain mappings were used to reconstruct the profiles of domain abundances for each ComEC protein.

Identification and Domain Profiling of CDCPs

A set of 5,574 complete bacterial genomes and their corresponding proteomes were downloaded from the NCBI RefSeq database (O'Leary et al., 2016) on October 25, 2016. Proteins containing the *Competence* domain (Pfam accession number: PF03772) were identified as CDCPs following the above-mentioned criteria of domain mapping, and the CDCPs were further analyzed to enumerate the co-occurrence of additional domains. Detailed information of all CDCPs were enlisted as individual rows in **Supplementary Table S2**. The occurrence of different domains was examined among all proteomes analyzed (**Figure 1**). In order to decrease biases associated with over sampled species when calculating the taxonomic distribution of the different profile types (**Figure 2**), each domain profile (i.e., combination of domains in a protein) was counted proportionally based on their frequencies among different strains of a species. For example, three distinct profiles were observed among different strains of *Escherichia coli* with an occurrence of 9, 13, and 177 strains for each profile. Hence, a relative count of 0.045, 0.065, and 0.889, respectively, were assigned to each of the profiles in *E. coli*.

Genomic Analysis and Identification of Missing CDCPs

Genomic sequences of all bacterial strains were analyzed to search for potential CDCP coding regions or degenerated CDCP fragments that were not captured in the proteomic analysis. First, a query database was constructed by extracting nucleotide sequences that correspond to the CDCPs identified from the proteome-wide analysis. Then, subsequences of the genes that specifically encoded the *Competence* domains were searched against the database of 5,574 complete genomes using the *blastn* function in the BLAST+ package, version 2.2.31 (Camacho et al., 2009). A threshold of e-value less or equal to 1×10^{-5} , query coverage 30% or greater, and identity greater than or equal to 60% were used in the database search. Hits were identified among genomes that did not contain a CDCP in their corresponding proteomes (**Supplementary Table S2**).



Reconstruction of the Bacterial Species Phylogeny

AMPHORA2 (Wu and Scott, 2012) was used to identify marker genes from the 2,373 representative proteomes selected from the reduction of oversampling biases (described in an above section). Species of the archaeal order Thermococcales were used as an outgroup for the species phylogeny (39 proteomes). In cases where multiple copies of a marker were predicted in one genome, the protein with the lowest e-value was retained. Each identified marker gene was aligned and trimmed using the '-Trim' option in the MarkerAlignTrim.pl script of AMPHORA2. The individual protein alignments were then concatenated into one alignment. The hybrid mode of RAxML version 8.2.1 (Stamatakis, 2014) was used to infer a phylogeny using the JTT substitution model and CAT model of evolution. The internal nodes of the phylogeny were labeled based on the lowest common ancestor (LCA) of the underlying leaves. The tree was collapsed manually in iTol (Letunic and Bork, 2016) and taxa that contained a single representative in the dataset were removed to facilitate the visualization of major taxonomic groups (**Supplementary Data Sheet S1**). The distribution of different CDCP domain profiles in **Figure 2** was calculated based on the fraction of relative counts among all strains represented in each collapsed branch.

Ancestral State Reconstruction

The evolution of domains in CDCPs was inferred through reconstructing ancestral domain compositions across the species phylogeny. The three most abundant Pfam domains, *Competence*, *DUF4131* (Pfam accession PF13567), and *Lactamase_B* (Pfam accession PF00753), were considered in the ancestral state reconstruction. A total of five different states were assigned to each leaf of the species phylogeny based on the presence or

absence of each domain, these included (1) No CDCP (*NC*); (2) *Competence*-only (*C*); (3) *Competence* and *Lactamase_B* (*CL*); (4) *DUF4131* and *Competence* (*DC*); and (5) *DUF4131*, *Competence*, and *Lactamase_B* (*DCL*). Ancestral state reconstruction was performed with the 'ace' function from the APE package, version 5.1 (Paradis et al., 2004) in R, version 3.4.2 (R Core Team, 2017) with the 'discrete' type and equal rates ('ER') model. Subsequently, the posterior probabilities of each state at each internal node was overlaid onto the species phylogeny using the Phytools package, version 0.6-44 (Revell, 2012) in R, version 3.4.2 (R Core Team, 2017).

Reconstruction of Protein Domain Phylogenies

Protein sequences of the *Competence*, *DUF4131*, and *Lactamase_B* domains were extracted from entire CDCP sequences according to alignment positions of the respective Pfam domains. The extracted sequences were aligned using MUSCLE, version 3.8.31 (Edgar, 2004). Individual domain phylogenies were constructed using the hybrid mode of RAxML, version 8.2.1 (Stamatakis, 2014) with the JTT substitution model and the CAT model of evolution. The trees were midpoint rooted with taxonomic information mapped to the leaves. Due to the redundant nature of NCBI protein identifiers across proteomes, if one protein identifier was associated with multiple proteomes the corresponding leaf identifier was randomly assigned to one of the lineages. Additional verifications were performed to ensure that such redundant protein identifiers only occurred in closely related bacterial strains and hence did not influence the derivation of LCAs.

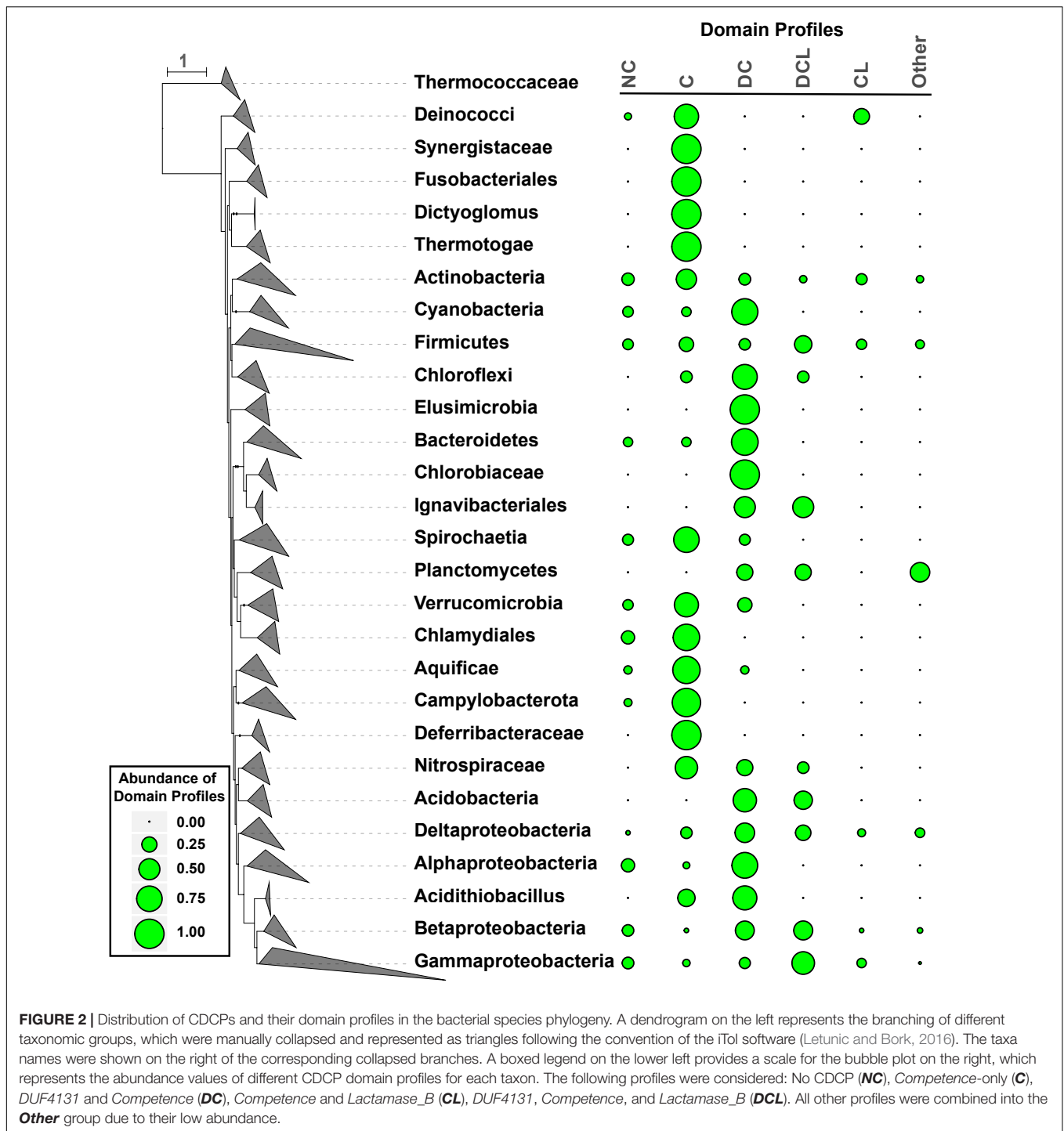
Comparison of Evolutionary Distances

Pairwise phylogenetic distances were computed for leaves of the species phylogeny and the phylogenies of the *Competence*, *DUF4131*, and *Lactamase_B* domains with the 'cophenetic' function in R, version 3.4.2 (R Core Team, 2017). The distances in the domain phylogenies were normalized by the maximum pairwise distances found in each tree. The distances in the species phylogeny were similarly normalized, but the maximum pairwise distance among all CDCP-containing strains were used as the denominator for the normalization. Pairs of strains that both have the *C*, *DC*, *CL*, or *DCL* profiles from the *Competence* tree along with pairs of strains that both have the *DC* or *DCL* and *CL* or *DCL* profiles in the *DUF4131* and *Lactamase_B* trees, respectively, were identified (**Figure 3**). Linear models were constructed using the 'lm' function in R, version 3.4.2 (R Core Team, 2017) to correlate the species distances with the domain distances for each subplot in **Figure 3**.

RESULTS

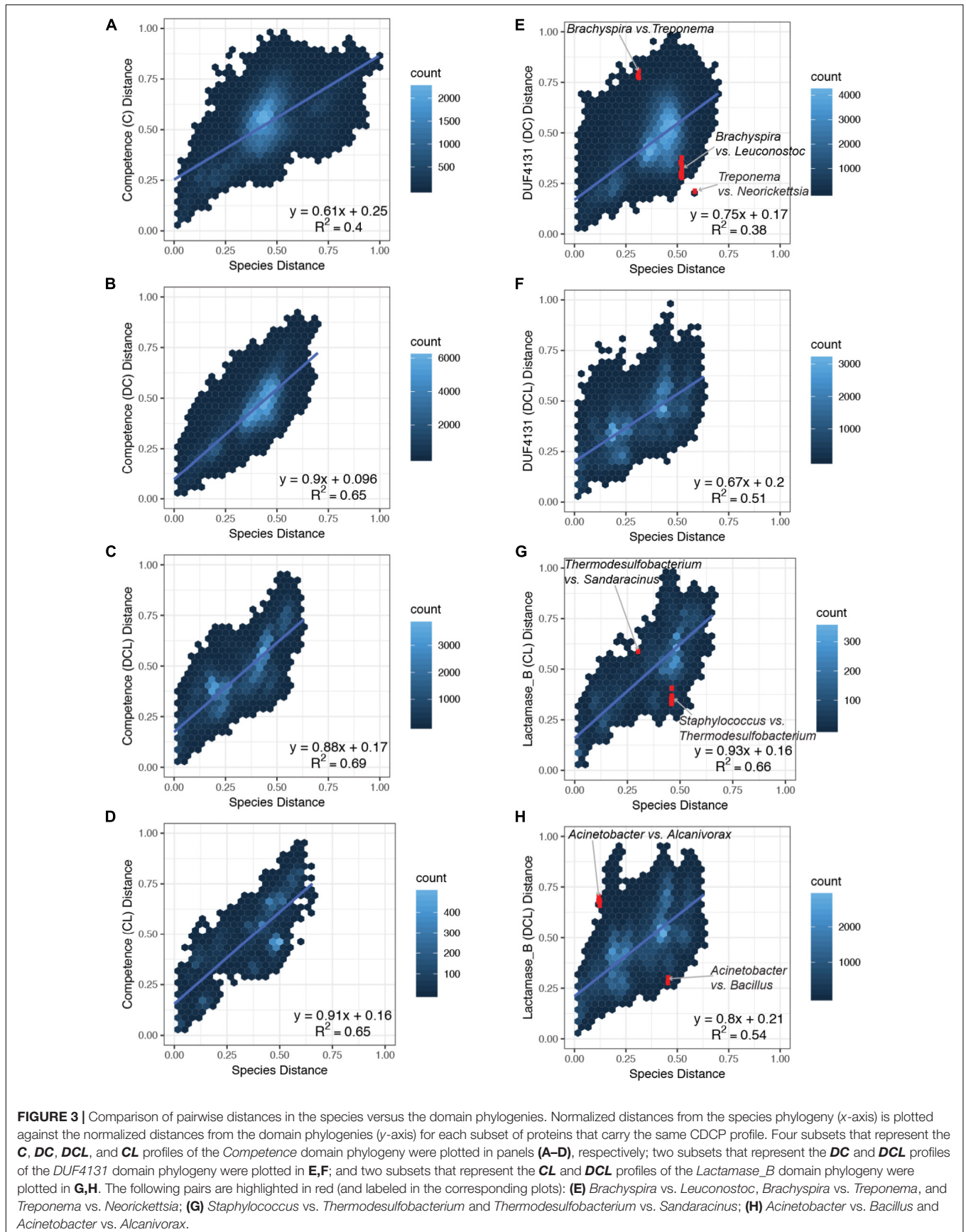
Domain Composition of Known ComEC Proteins

A list of 51 naturally competent bacterial strains were identified from the literature (Johnston et al., 2014 and references



therein). These strains span eight different phyla, and each strain encodes a single ComEC protein varying in length from 417 to 872 amino acids (**Supplementary Table S1**). Three distinct protein domains, *DUF4131*, *Competence*, and *Lactamase_B*, were observed based on sequence comparisons of the ComEC proteins to the Pfam database (Materials and Methods). While the *Competence* domain was universally conserved, the *DUF4131* and *Lactamase_B* domains appeared

in some but not all ComEC proteins. Overall, four domain profiles were identified (**Figure 1A**). The most abundant was the **DCL** profile (24 occurrences), followed by the **DC** profile (14 occurrences). The least abundant were the **CL** (7 occurrences) and the **C** profiles (6 occurrences). The great diversity in domain compositions and sequence lengths of ComEC proteins prevents the global identification of the protein from comparison of full length sequences alone. However, due to the universal presence of



the *Competence* domain among all known ComEC proteins, here we study the *Competence* domain containing proteins (CDCPs) and use them as a proxy to infer the distribution, domain profiles, and evolution of putative ComEC proteins among a diverse range of fully sequenced bacterial genomes.

Abundance and Distribution of CDCPs in Bacteria

Complete proteomes of 5,574 bacteria were analyzed, where CDCPs were found in 4,934 proteomes. Of the 640 proteomes for which a CDCP was not identified, 409 were predicted to encode either complete or fragmented *Competence* domains in the genomic DNA (**Supplementary Table S2**). Hence, a near universal presence of the protein was observed among 89% of the proteomes and 96% of the genomes analyzed in this study. The remaining 231 strains for which a *Competence* domains was missing included representatives from the *Buchnera*, *Rickettsia*, *Blattabacterium*, *Mycoplasma*, and *Prochlorococcus* genera. These genera are typically obligate endosymbionts, intracellular pathogens, or small marine phytoplanktons, and they are known to have highly reduced genomes (Wernegreen and Moran, 1999; Moran, 2002; Kuo et al., 2009).

When present, only a single CDCP was found in every proteome except for the *Pseudarthrobacter sulfonivorans* strain Ar51, which contains two identified CDCPs in its proteome. However, one of the proteins (WP_058932413.1) appeared to be a truncated copy of the other (WP_058932395.1), as they share a sequence identity of 94% and the former is 330 amino acids shorter than the latter. In addition, all identified CDCPs were encoded in the chromosomes except for the strains *Xanthomonas sacchari* R1 and *Prevotella* sp. oral taxon 299 F0039, where the CDCPs were encoded in plasmids. The occurrence of CDCPs varied among different taxa. For example, in the phyla Thermotogae, Fusobacteria, Chlorobi, and Chloroflexi, a CDCP was found in every strain. In contrast, a CDCP was missing among 10.7% and 36.6% of the strains in Actinobacteria and Tenericutes, respectively.

Domain Variations in CDCPs

Besides the *DUF4131* and *Lactamase_B* domains found in known ComEC proteins, five additional domains were found to co-occur with the *Competence* domain through the global search of bacterial proteomes (**Figure 1B**). These included members of the beta-lactamase superfamilies *Lactamase_B_2* (Pfam accession PF12706) and *Lactamase_B_3* (Pfam accession PF13483), the alpha/beta hydrolase families *Abhydrolase_5* (Pfam accession PF12695) and *Abhydrolase_6* (Pfam accession PF12697), and the *SLBB* domain (Pfam accession PF10531) of the beta-grasp fold (Burroughs et al., 2007). The *DUF4131* and *Lactamase_B* remained to be the most dominant among all domains: of the 4,935 CDCPs identified, the *DUF4131* domain was in 66.4% (3,279 occurrences) and the *Lactamase_B* domain was in 52.3% (2,583 occurrences). In contrast, all other domains only had from one to 97 occurrences (**Figure 1B**).

By default, the collection of proteomes included an over-representation of certain species, such as *E. coli* (199 distinct

proteomes), *Helicobacter pylori* (83 distinct proteomes), and *B. subtilis* (45 distinct proteomes). Substrains of these species may contain CDCPs with distinct domain profiles. For example, of the 199 distinct *E. coli* proteomes analyzed, nine encoded no CDCP (*NC*), 13 encoded CDCPs of the *CL* profile, and 177 encoded CDCPs of the *DCL* profile (**Supplementary Table S2**). To avoid biases introduced by oversampling of such species while still account for the diversity of domain profiles, a non-redundant set of 2,373 representative proteomes were identified that proportionally represents domain profiles among bacterial species (Materials and Methods).

A bacterial species phylogeny was constructed from protein marker sequences identified in the representative proteomes (**Figure 2**). The frequency of CDCP domain profiles within any given taxon was calculated by dividing the relative counts of each profile by the sum of relative counts for all proteomes in the taxon (Materials and Methods). Six distinct profiles were considered. Besides the dominant *C*, *DC*, *DCL*, and *CL* profiles, strains that encode any other domains were classified into a group named *Other*, and strains that encode no CDCP were classified as *NC*.

The analysis of domain distributions (**Figure 2**) revealed a great diversity among different taxa. While the Actinobacteria, Firmicutes, Betaproteobacteria, and Deltaproteobacteria included a relatively even proportion of all profiles, other taxa demonstrated preferences to one or a few profiles. Among all CDCPs, the *C* profile appeared to be the most widely distributed across a diverse range of bacterial phyla. It was dominant (with a frequency of over 0.50) in Aquificae, Campylobacterota, Thermotogae, Deinococcus-Thermus, Chlamydiae, and Spirochaetes. The *DC* profile, one of the most abundant among all profiles, was the main representative in Cyanobacteria, Bacteroidetes, and Alphaproteobacteria. The *DCL* profile occurred in multiple taxa, but was only dominant in the Gammaproteobacteria. Additionally, a dominance of the *CL*, *NC*, and *Other* profiles was not observed in any taxa, and the main occurrences of these profiles were in Actinobacteria, Firmicutes, and Gammaproteobacteria.

Evolution of Domain Compositions in CDCPs

An ancestral state reconstruction of the domains indicated the *C* profile as the most parsimonious ancestral state among all bacteria, while it also revealed multiple gains and losses of the *DUF4131* and *Lactamase_B* domains (**Supplementary Figure S1**). For example, in the Deinococcus-Thermus branch, the *Lactamase_B* domain appeared to be gained in the family Thermaceae but remained absent in the Deinococcaceae (**Supplementary Figure S2A**). In Spirochaetes, the *C* profile dominated all major subclades except for the *Brachyspira* genus and *Treponema azotonutricium* ZAS-9, where a gain of the *DUF4131* domain was observed (**Supplementary Figure S2B**). In Bacteroidetes, the *DUF4131* domain was potentially gained in its common ancestor and retained among the majority of its subclades except for the family Porphyromonadaceae, which is dominated by the *C* profile. Additionally, the Bacteroidetes also included a subclade that demonstrated a loss of entire

CDCPs. This subclade is composed solely of the endosymbionts of the *Blattabacterium* genus (**Supplementary Figure S2C**). In Proteobacteria (**Supplementary Figures S2D–G**), the *DUF4131* domain was observed in the common ancestor of this phylum, followed by a secondary acquisition of the *Lactamase_B* domain in the *Burkholderia* genus and the Neisseriales order of Betaproteobacteria, the Desulfuromonadales order of Deltaproteobacteria, and the Gammaproteobacteria. Several subsequent losses were also observed throughout the different classes of Proteobacteria. Specifically, entire CDCPs were lost in the endosymbionts *Kinetoplastibacterium*, *Buchnera*, and *Wigglesworthia*, as well as in most of the intracellular pathogens of the *Rickettsia*, *Francisella*, and *Coxiella* genera.

The Firmicutes phylum had both the *DUF4131* and the *Lactamase_B* domains in its ancestral state, while it experienced multiple losses and gains of either or both domains in some subclades (**Supplementary Figure S2H**). For example, both domains were lost in the families Clostridiaceae and Planococcaceae, the *DUF4131* domain was lost in substrains of the genus *Lactobacillus*, and the *Lactamase_B* domain was lost in substrains of the genus *Listeria*, *Desulfosporosinus*, and *Desulfotobacterium*. While the family Staphylococcaceae may have originated with a loss of both domains, the *Lactamase_B* domain was regained in some species of the *Staphylococcus* genus. Similarly, the family Leuconostocaceae also had an initial loss of both domains, while the *DUF4131* domain was regained in the *Leuconostoc* genus.

Subclades of the Tenericutes phylum had an intermingled branching with the Erysipelotrichaceae family of Firmicutes, which is an important species in the human gut microbiota (Kaakoush, 2015) and has emerged between Acholeplasmataceae and other families of Tenericutes (**Supplementary Figure S2H**). While entire CDCPs were lost in Erysipelotrichaceae, Acholeplasmataceae, and the genus *Mycoplasma*, the rest of this clade was dominated by CDCPs of the C profile.

Evolution of Individual Domains in CDCPs

Phylogenies of the *Lactamase_B*, *DUF4131*, and *Competence* domains (**Supplementary Figures S3–S5**) provided further insights into the multiple domain gains and losses throughout the species phylogeny. For example, the *Thermodesulfobacterium* genus, while shown to form a clade within Deltaproteobacteria in the species phylogeny (**Supplementary Figure S2**), clustered with the *Staphylococcus* genus in the *Lactamase_B* domain phylogeny (**Supplementary Figure S3**). Further examination of the neighboring branches of this clade in the *Lactamase_B* phylogeny revealed a similarity to the Clostridia in Firmicutes, suggesting an acquisition of the *Lactamase_B* domain from Clostridia into both *Staphylococcus* and *Thermodesulfobacterium*. Similarly, the *DUF4131* domain of *Leuconostoc* is a close neighbor of the *Brachyspira* genus in Spirochaetes, and the broader context of the domain phylogeny suggested a potential acquisition from Clostridia to *Leuconostoc* and *Brachyspira* (**Supplementary Figure S4**).

An overview of the correlation between species distances and domain distances is summarized in **Figure 3**. Pairs of bacterial strains were divided into subsets based on the CDCP domain profiles encoded in their proteomes. Domain distances were considered among the *Competence* (**Figures 3A–D**), *DUF4131* (**Figures 3E,F**), and *Lactamase_B* (**Figures 3G,H**) phylogeny, and a linear regression was applied to each subplot in order to establish the quantitative association of domain distances as compared to the species distances. For the *Competence* domain, the slope of the linear models ranged from 0.61 to 0.91 for the C, DC, DCL, and CL profiles, with the C profile carrying the lowest slope while the CL profile carrying the highest slope. For the *DUF4131* domain, only two profiles (DC and DCL) were available, and the slope of their linear models are 0.75 and 0.67, respectively. Finally, for the *Lactamase_B* domain, slopes of 0.93 and 0.80, respectively, were observed for the CL and DCL profiles. All domain distances had a positive offset as compared to the species distances. The offset ranges from 0.16 to 0.25, with only an exception in the DC profile of the *Competence* domain, which had an offset of 0.096 (**Figure 3B**). Overall, linear regression revealed a positive correlation between distances of the species phylogeny and distances of the domain phylogenies. While three profiles of the *Competence* domain (**Figures 3B–D**) and all profiles of the *Lactamase_B* domain (**Figures 3G,H**) had a near 1:1 ratio with the species distances, the C profile of *Competence* (**Figure 3A**) and all profiles of the *DUF4131* domain (**Figures 3E,F**) had a lower ratio when compared to the species distances.

Despite capturing the overall trends of species and domain distance correlations, the linear models were insufficient in explaining all variations in the domain distances (R-squared values between 0.38 and 0.69). Some of the variations indicated the lateral acquisition of particular domains. For example, in the above-mentioned case studies, the acquisition of *DUF4131* domain in *Brachyspira* and *Leuconostoc*, and the acquisition of *Lactamase_B* domain in *Staphylococcus* and *Thermodesulfobacterium* resulted in smaller domain distances among taxa that are more distant in the species phylogeny. Hence, coordinates of these pairs were below the linear regression lines (**Figures 3E,H**).

Additional variations were identified through a close examination of outliers in domain distances. The termite gut bacterium, *T. azotonutricium* ZAS-9 (Graber et al., 2004), is related to the *Brachyspira* of Spirochaetia while distant from the *Neorickettsia* and *Wolbachia* of Alphaproteobacteria in the species phylogeny. However, in the *DUF4131* domain phylogeny, *T. azotonutricium* ZAS-9 had a shorter distance to *Neorickettsia* and *Wolbachia*, while it had a longer distance to the *Brachyspira* (**Figure 3E**). Similarly, the *Thermodesulfobacterium* and the *Sandaracinus amylolyticus* of Deltaproteobacteria are closely related in the species phylogeny, but they are distant in the *Lactamase_B* domain phylogeny (**Figure 3G**). Finally, *Acinetobacter baumannii* and *Acinetobacter equi* of Gammaproteobacteria are distant from *Bacillus pumilus* of Firmicutes in the species phylogeny, but the *Lactamase_B* domain distances between *Acinetobacter* and *B. pumilus* are shorter than expected from their species distances. In contrast, the *Lactamase_B*

domains of *Acinetobacter* and *Alcanivorax* are distant, despite the taxa being close neighbors in the species phylogeny (Figure 3H).

DISCUSSION

Natural transformation is a mechanism of horizontal gene transfer and has fundamental roles in bacterial genome evolution (Ambur et al., 2016). The ComEC protein mediates the critical step of ssDNA membrane translocation, and as such it serves as an essential component of the NT machinery. Despite extensive studies of ComEC among selected strains that are known to be naturally competent, little is known about the taxonomic distribution and domain evolution of ComEC in diverse lineages of Bacteria. The current study has provided a stepping stone into bridging this knowledge gap through a global survey of putative ComEC proteins (referred to as CDCPs) and a reconstruction of their ancestral states. Similar analyses can be performed on other proteins that compose the NT machinery to enhance our understanding of its diversity and evolution.

A global presence of CDCPs has been identified among 89% of proteomes and 96% of genomes analyzed in this study. This is surprising considering that only less than 1% of these bacterial strains are known to be naturally competent under laboratory settings (Johnsborg et al., 2007; Johnston et al., 2014). Several considerations could help reconcile the apparent discrepancy between the abundance of CDCPs and the lack of detectable natural competence in many bacteria: (1) the NT process is mediated by many proteins besides ComEC, so variations in other proteins could inhibit the competence phenotype; (2) the NT process is subject to complex regulatory mechanisms, which may prevent the expression of essential proteins in the NT machinery (including ComEC) even when all the proteins are encoded in the genome (Sinha and Redfield, 2012); and (3) besides being a competence protein, ComEC could have other functions. For example, it may be involved in the virulence of *Listeria monocytogenes* (Rabinovich et al., 2012) and *A. baumannii* (Wilharm et al., 2013), and it may contribute to the twitching motility in *Thermus thermophilus* (Salzer et al., 2016).

For all but one strain, only a single CDCP has been found in each proteome (Supplementary Table S2). The two closely related copies of CDCPs in *P. sulfonivorans* Ar51 are likely to be a result of a recent gene duplication event, and the fact that one protein appears to be a truncated copy of the other suggests the former is under the process of being degenerated from the genome. The broad presence of single copy CDCP-encoding genes suggests its key roles in the cell machinery, although its function and regulation in strains that are not known to undergo NT is still unclear.

Despite the near universal presence, CDCPs are missing from 231 bacterial strains using search criteria defined in this study. These strains are enriched with obligate endosymbionts and intracellular pathogens. Considering the significant genomic reductions in these organisms due to genetic drift

(Wernegreen and Moran, 1999; Moran, 2002; Kuo et al., 2009), ComEC proteins may have been lost from these strains, either in the ancestral state of an entire genera (e.g., *Buchnera*) or among selected subgroups (e.g., *Mycoplasma*, Supplementary Figure S2).

Two domains, *DUF4131* and *Lactamase_B*, are commonly present in CDCPs. While five additional domains have also been observed, these domains occur in low frequencies and hence could have minor roles in the evolution of CDCPs. Of the dominant domain profiles, the *Competence*-only (*C*) profile appears to be the most ancient and most widely distributed among bacterial phyla (Figure 2). This observation has been supported with an ancestral state reconstruction (Supplementary Figures S1, S2), which predicted that the protein has emerged from the *C* profile, and subsequently modified through multiple gains and losses of the *DUF4131* and *Lactamase_B* domains over diverse lineages.

An earlier study of naturally competent strains and the ComEC proteins encoded by these strains suggests the *Competence* domain evolve at a rate sixfold higher than the corresponding species (Johnston et al., 2014). This phenomenon, however, was not observed in our broader survey of putative ComEC proteins. In contrast, the *Competence* domain of *DC*, *DCL*, and *CL* proteins appeared to maintain a near 1:1 slope, and that of the *C* profile had a lower slope of 0.61 (Figures 3A–D). Further, the larger dataset provided from this study has revealed additional complexity in the evolution of ComEC proteins that are not explained by the linear models. This is reflected in the relatively low R-squared values in linear fitting. As detailed in several case studies in the Results section, this deviation from linear models could indicate potential events of horizontal domain acquisition that contributes to the gains and losses of *DUF4131* and *Lactamase_B* in CDCPs among diverse lineages. Interestingly, no obvious HGT has been found for the *DUF4131* domain among proteins that carry the *DCL* profile (Figure 3F). Since the acquisition of *DUF4131* generally proceeds the acquisition of *Lactamase_B* (Supplementary Figure S1), it suggests that proteins of the *DCL* profile are unlikely to lose the *DUF4131* domain and then regain it into the resulting *CL* proteins.

AUTHOR CONTRIBUTIONS

YZ conceived the project and led the data analysis and interpretation. ZP contributed to data acquisition, analysis and interpretation. All authors composed the manuscript.

FUNDING

This work was supported by the USDA National Institute of Food and Agriculture, Hatch fund, RI0015-H002, accession #1007170, the National Science Foundation under grant #1553211, and a subcontract from the RI Research Alliance Collaborative Research Grant. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

ACKNOWLEDGMENTS

We thank Dr. Peter Belenky and Dr. Marta Gomez-Chiarri for the constructive discussions at the beginning of the project. Thanks to Keith Dufault-Thompson and Ke Zhang for comments and suggestions on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2018.02980/full#supplementary-material>

FIGURE S1 | Ancestral reconstruction of CDCPs in the species phylogeny. Five states were considered in the reconstruction: No CDCP (**NC**), *Competence*-only (**C**), *DUF4131* and *Competence* (**DC**), *Competence* and *Lactamase_B* (**CL**), and *DUF4131*, *Competence*, and *Lactamase_B* (**DCL**). Phyla (except for the case of the Proteobacteria which were analyzed at the class level) with greater than 20 leaves were labeled. Selected groups that carry the **NC** profile were also labeled throughout the phylogeny.

REFERENCES

- Ambur, O. H., Engelstädter, J., Johnsen, P. J., Miller, E. L., and Rozen, D. E. (2016). Steady at the wheel: conservative sex and the benefits of bacterial transformation. *Philos. Trans. R. Soc. B Biol. Sci.* 371:20150528. doi: 10.1098/rstb.2015.0528
- Baker, J. A., Simkovic, F., Taylor, H. M. C., and Rigden, D. J. (2016). Potential DNA binding and nuclease functions of ComEC domains characterized in silico. *Proteins Struct. Funct. Bioinform.* 84, 1431–1442. doi: 10.1002/prot.25088
- Bergé, M., Moscoso, M., Prudhomme, M., Martin, B., and Claverys, J. P. (2002). Uptake of transforming DNA in Gram-positive bacteria: a view from *Streptococcus pneumoniae*. *Mol. Microbiol.* 45, 411–421. doi: 10.1046/j.1365-2958.2002.03013.x
- Burroughs, A. M., Balaji, S., Iyer, L. M., and Aravind, L. (2007). A novel superfamily containing the β -grasp fold involved in binding diverse soluble ligands. *Biol. Direct* 2:4.
- Callebaut, I. (2002). Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res.* 30, 3592–3601. doi: 10.1093/nar/gk470
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- Chen, I., and Dubnau, D. (2004). DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.* 2, 241–249. doi: 10.1038/nrmicro844
- Cox, M. M. (2007). Motoring along with the bacterial RecA protein. *Nat. Rev. Mol. Cell Biol.* 8, 127–138. doi: 10.1038/nrm2099
- Dominski, Z. (2007). Nucleases of the metallo-beta-lactamase family and their role in DNA and RNA metabolism. *Crit. Rev. Biochem. Mol. Biol.* 42, 67–93. doi: 10.1080/10409230701279118
- Draskovic, I., and Dubnau, D. (2005). Biogenesis of a putative channel protein, ComEC, required for DNA uptake: membrane topology, oligomerization and formation of disulphide bonds. *Mol. Microbiol.* 55, 881–896. doi: 10.1111/j.1365-2958.2004.04430.x
- Dubnau, D. (1991). Genetic competence in *Bacillus subtilis*. *Mol. Microbiol.* 55, 395–424. doi: 10.1111/j.1365-2958.1991.tb01820.x

FIGURE S2 | Detailed view of the sub-branches displayed in **Supplementary Figure S1** along with the ancestral state reconstruction results. The Deinococcus-Thermus (**A**), Spirochaetes (**B**), Bacteroidetes (**C**), Betaproteobacteria (**D**), Deltaproteobacteria (**E**), Gammaproteobacteria (**F**), Alphaproteobacteria (**G**), and Firmicutes (**H**) are displayed.

FIGURE S3 | Phylogenetic reconstruction of the *Lactamase_B* domain.

FIGURE S4 | Phylogenetic reconstruction of the *DUF4131* domain.

FIGURE S5 | Phylogenetic reconstruction of the *Competence* domain.

TABLE S1 | Domain profiles of ComEC in bacteria confirmed to be capable of natural transformation.

TABLE S2 | Taxonomic and domain presence/absence data for CDCPs predicted from 5,574 complete bacterial proteomes/genomes. A total of eight domains (including *Competence*) were observed among all of the identified CDCPs. The abundance of each domain was recorded in their corresponding columns, with zero indicating the absence of a domain in a CDCP, and non-zero values indicating the number of occurrences of a domain in a protein. Each row of the table represents a unique combination of a CDCP (if identified) and a bacterial strain that carries the protein. The columns included the protein accession number (if present), protein length, genome assembly identifier from the NCBI RefSeq database (O’Leary et al., 2016), and the taxonomy assignments from the NCBI Taxonomy databases (Sayers et al., 2009). For the strains in which no CDCP was identified, the domain profiles were filled with all zeros for each domain. In the one case where two CDCPs were identified in one strain (NCBI assembly GCF_001484605.1), two entries/rows were included with the same NCBI assembly identifier mapped to the two different CDCP protein identifiers.

DATA SHEET S1 | Manual operations for the visualization of the species phylogeny.

- Eddy, S. (1998). Profile hidden markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340
- Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344
- Friedrich, A., Hartsch, T., and Averhoff, B. (2001). Natural transformation in mesophilic and thermophilic bacteria: identification and characterization of novel, closely related competence genes in acinetobacter sp. *Appl. Environ. Microbiol.* 67, 3140–3148. doi: 10.1128/AEM.67.7.3140
- Graber, J. R., Leadbetter, J. R., and Breznak, J. A. (2004). Description of *Treponema azotonutricium* sp. nov. and *Treponema primitia* sp. nov., the First Spirochetes Isolated from Termite Guts. *Appl. Environ. Microbiol.* 70, 1315–1320. doi: 10.1128/AEM.70.3.1315-1320.2004
- Hofreuter, D., Odenbreit, S., and Haas, R. (2001). Natural transformation competence in *Helicobacter pylori* is mediated by the basic components of a type IV secretion system. *Mol. Microbiol.* 41, 379–391. doi: 10.1046/j.1365-2958.2001.02502.x
- Johnsborg, O., Eldholm, V., and Håvarstein, L. S. (2007). Natural genetic transformation: prevalence, mechanisms and function. *Res. Microbiol.* 158, 767–778. doi: 10.1016/j.resmic.2007.09.004
- Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J. P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* 12, 181–196. doi: 10.1038/nrmicro3199
- Kaakoush, N. O. (2015). Insights into the role of Erysipelotrichaceae in the human host. *Front. Cell. Infect. Microbiol.* 5:84. doi: 10.3389/fcimb.2015.00084
- Kantor, R. S., Wrighton, K. C., Handley, K. M., Sharon, I., Hug, L. A., Castelle, C. J., et al. (2013). Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* 4:e00708-13. doi: 10.1128/mBio.00708-13
- Kuo, C. H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19, 1450–1454. doi: 10.1101/gr.091785.109

- Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290
- Lipscomb, G. L., Stirrett, K., Schut, G. J., Yang, F., Jenney, F. E., Scott, R. A., et al. (2011). Natural competence in the hyperthermophilic archaeon *Pyrococcus furiosus* facilitates genetic manipulation: construction of markerless deletions of genes encoding the two cytoplasmic hydrogenases. *Appl. Environ. Microbiol.* 77, 2232–2238. doi: 10.1128/AEM.02624-10
- Lorenz, M. G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* 58, 563–602.
- Matthey, N., and Blokesch, M. (2016). The DNA-uptake process of naturally competent *Vibrio cholerae*. *Trends Microbiol.* 24, 98–110. doi: 10.1016/j.tim.2015.10.008
- Meibom, K. L., Blokesch, M., Dolganov, N. A., Wu, C.-Y., and Schoolnik, G. K. (2005). Chitin induces natural competence in *Vibrio cholerae*. *Science* 310, 1824–1827. doi: 10.1126/science.1120096
- Mell, J. C., and Redfield, R. J. (2014). Natural competence and the evolution of DNA uptake specificity. *J. Bacteriol.* 196, 1471–1483. doi: 10.1128/JB.01293-13
- Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108, 583–586. doi: 10.1016/S0092-8674(02)00665-7
- Mortier-Barrière, I., Velten, M., Dupaigne, P., Mirouze, N., Piétremont, O., McGovern, S., et al. (2007). A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA. *Cell* 130, 824–836. doi: 10.1016/j.cell.2007.07.038
- O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412
- Puyet, A., Greenberg, B., and Lacks, S. A. (1990). Genetic and structural characterization of endA. A membrane-bound nuclease required for transformation of *Streptococcus pneumoniae*. *J. Mol. Biol.* 213, 727–738. doi: 10.1016/S0022-2836(05)80259-1
- Quevillon-Cheruel, S., Campo, N., Mirouze, N., Mortier-Barrière, I., Brooks, M. A., Boudes, M., et al. (2012). Structure-function analysis of pneumococcal DprA protein reveals that dimerization is crucial for loading RecA recombinase onto DNA during transformation. *Proc. Natl. Acad. Sci. U.S.A.* 109, E2466–E2475. doi: 10.1073/pnas.1205638109
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <http://www.r-project.org/>
- Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., and Herskovits, A. A. (2012). Prophage excision activates listeria competence genes that promote phagosomal escape and virulence. *Cell* 150, 792–802. doi: 10.1016/j.cell.2012.06.036
- Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169.x
- Salzer, R., Kern, T., Joos, F., and Averhoff, B. (2016). The *Thermus thermophilus* comEA/comEC operon is associated with DNA binding and regulation of the DNA translocator and type IV pili. *Environ. Microbiol.* 18, 65–74. doi: 10.1111/1462-2920.12820
- Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, et al. (2009). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 44, D7–D19. doi: 10.1093/nar/gkn741
- Sinha, S., and Redfield, R. J. (2012). Natural DNA uptake by *Escherichia coli*. *PLoS One* 7:e35620. doi: 10.1371/journal.pone.0035620
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033
- Taha, M. K., Larribe, M., Dupuy, B., Giorgini, D., and Marchal, C. (1992). Role of pilA, an essential regulatory gene of *Neisseria gonorrhoeae*, in the stress response. *J. Bacteriol.* 174, 5978–5981. doi: 10.1128/jb.174.18.5978-5981.s1992
- van Wolferen, M., Wagner, A., van der Does, C., and Albers, S.-V. (2016). The archaeal Ced system imports DNA. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2496–2501. doi: 10.1073/pnas.1513740113
- Wernegreen, J. J., and Moran, N. A. (1999). Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol. Biol. Evol.* 16, 83–97. doi: 10.1093/oxfordjournals.molbev.a026040
- Wilharm, G., Piesker, J., Laue, M., and Skiebe, E. (2013). DNA Uptake by the nosocomial pathogen *Acinetobacter baumannii* occurs during movement along wet surfaces. *J. Bacteriol.* 195, 4146–4153. doi: 10.1128/JB.00754-13
- Wu, M., and Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* 28, 1033–1034. doi: 10.1093/bioinformatics/bts079
- Yeh, Y. C., Lin, T. L., Chang, K. C., and Wang, J. T. (2003). Characterization of a ComE3 homologue essential for DNA transformation in *Helicobacter pylori*. *Infect. Immun.* 71, 5427–5431. doi: 10.1128/IAI.71.9.5427-5431.2003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Pimentel and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.