Article

Insights into the Phylogeny and Evolution of Cold Shock Proteins: From Enteropathogenic *Yersinia* and *Escherichia coli* to Eubacteria

Tao Yu [1,2,*], **Riikka Keto-Timonen** [2], **Xiaojie Jiang** [2], **Jussa-Pekka Virtanen** [2] and **Hannu Korkeala** [2]

1   Department of Life Science and Technology, Xinxiang University, Xinxiang 453003, China
2   Department of Food Hygiene and Environmental Health, University of Helsinki, P.O. Box 66, FI-00014 Helsinki, Finland
*   Correspondence: yu.tao@helsinki.fi

check for updates

Abstract: Psychrotrophic foodborne pathogens, such as enteropathogenic *Yersinia*, which are able to survive and multiply at low temperatures, require cold shock proteins (Csps). The Csp superfamily consists of a diverse group of homologous proteins, which have been found throughout the eubacteria. They are related to cold shock tolerance and other cellular processes. Csps are mainly named following the convention of those in *Escherichia coli*. However, the nomenclature of certain Csps reflects neither their sequences nor functions, which can be confusing. Here, we performed phylogenetic analyses on Csp sequences in psychrotrophic enteropathogenic *Yersinia* and *E. coli*. We found that representative Csps in enteropathogenic *Yersinia* and *E. coli* can be clustered into six phylogenetic groups. When we extended the analysis to cover Enterobacteriales, the same major groups formed. Moreover, we investigated the evolutionary and structural relationships and the origin time of Csp superfamily members in eubacteria using nucleotide-level comparisons. Csps in eubacteria were classified into five clades and 12 subclades. The most recent common ancestor of Csp genes was estimated to have existed 3585 million years ago, indicating that Csps have been important since the beginning of evolution and have enabled bacterial growth in unfavorable conditions.

1. Introduction

Psychrotrophic foodborne pathogens, such as enteropathogenic *Yersinia,* which are able to survive and multiply at low temperatures, pose a risk in modern food production, where cold chains are used to increase the shelf lives of food products [1–3]. Cold-induced proteins (Cips) are important for growth at low temperatures and are involved in RNA metabolism, protein folding and the synthesis of membrane lipid A [4]. Cips include, e.g., homologous Csps that are classified together in the Csp family [5]. An abrupt temperature downshift (cold shock) causes a cold shock response in bacteria, during which only a limited number of proteins are induced [6]. Many *csp* genes are highly induced after a temperature downshift, and Csps are known to play a key role in survival after cold shock and in adaptation to low growth temperature [4,6–14]. Recent studies have shown that Csps may also aid in other stress responses [5,15–17]. Csps are small nucleic acid-binding proteins that have been found in a variety of eubacteria, including psychrophiles, mesophiles and thermophiles [18]. Very few research papers have been published concerning the need for *csp* genes in psychrotrophic *Yersinia* [5,19,20]. Most Csp studies have been conducted using mesophilic *E. coli* [6,10–13,21–25]. Understanding how the Csps of psychrotrophic bacteria, such as enteropathogenic *Yersinia*, differ from those of mesophilic

bacteria is important, for this may represent a common feature of psychrotrophs that distinguishes them from mesophiles. Furthermore, a deeper understanding of the Csps of psychrotrophs could contribute to preventing their growth in refrigerated food or in optimizing the production of their enzyme for biotechnological purposes [26].

The Csps have two conserved RNA-binding motifs, RNP1 and RNP2 [10,27,28], which may be responsible for selective RNA-binding activity [21]. CspA was first identified in *E. coli* [6,11]. So far, a total of nine homologous Csps (CspA-CspI) have been found in *E. coli*, defined by their protein sequences [12]. However, all Csps in *E. coli* are not related to the cold shock response. Only CspA, CspB, CspE, CspG and CspI are generally recognized as cold induced [6,10,11,13,22,29]. CspC and CspE are involved in the regulation of alternative sigma factor σ^S and universal stress protein UspA, and in growth [23,24,30]. CspD is induced at the early stationary phase during nutritional starvation [25].

Csps are mainly named following the convention of *E. coli* Csps. However, the nomenclature of certain Csps reflects neither their sequences nor functions [5]. For example, of the ten Csps in *Yersinia enterocolitica* subsp. *enterocolitica* 8081, CspA2 (*YE3823*) has the highest amino acid similarity with CspI of *E. coli* [5]. In *Yersinia pseudotuberculosis* IP32953, CspA2 (*YPTB3587*) is also the most similar Csp to CspI of *E. coli* [5]. Moreover, unlike the CspD in *E. coli*, CspD in *Bacillus subtilis* is cold induced [31]. The disconnected nomenclature of Csps in various bacteria can be confusing. With increasing numbers of sequenced bacterial genomes, certain Csps have been found to have no high sequence similarity to any of those in *E. coli* [5,32,33].

In this study, we performed phylogenetic analyses on Csp sequences in enteropathogenic *Yersinia*, *E. coli* and eventually all Enterobacteriales. Based on the variation in amino acid residues, Csps in Enterobacteriales were divided into six major groups. Moreover, we performed evolutionary analyses on the *csp* genes, and clarified the phylogeny and origin time of Csps in eubacteria. Five clades and twelve subclades were identified in eubacterial Csps. The estimation of the timing of Csp evolution, based on the time to the most recent common ancestor (tMRCA) of *csp* genes in eubacteria, suggests that Csps emerged 3585 million years ago (MYA), with a mean mutation rate of 4.392×10^{-4} substitutions per site per million years.

2. Results

2.1. Phylogeny and Consistency of Csps in Enteropathogenic Yersinia Enterocolitica and Yersinia Pseudotuberculosis

A total of 447 Csp sequences with explicit gene symbols in *Y. enterocolitica* (*n* = 13) and *Y. pseudotuberculosis* (*n* = 45) strains (Table S1) were retrieved from the Pathosystems Resource Integration Center (PATRIC) database. Twenty-three Csp sequence patterns were identified and separated into five phylogenetic clusters (Figure 1, Table S2). Csps in the same pattern shared 100% amino acid sequence identity with each other. Out of these patterns, 10 patterns were specific for *Y. enterocolitica*, 10 patterns were only found in *Y. pseudotuberculosis* and both species harbored patterns 12, 13 and 14 (Figure 1, Tables S3 and S4).

The pairwise evolutionary distances between respective Csps of all Csp sequence patterns in *Y. enterocolitica* and *Y. pseudotuberculosis* were calculated based on the phylogenetic tree (Figure 1) and visualized using a heat map (Figure S1). Corresponding homologues of the representative Csps for 12 *Y. enterocolitica* patterns could be identified in the representative Csps of *Y. pseudotuberculosis* patterns. Particularly, Csps in cluster A, in which 12 of 23 patterns were included, were homologous, with a maximum distance of 0.14. The maximum distances of cluster B (four patterns included), cluster D (three patterns included) and cluster E (three patterns included) were 0.27, 0.08 and 0.20 respectively. However, pattern 17 represented by CspE2 (*YE1546*) forms a monophyletic branch (cluster C) separated from the other patterns, and no identical pattern was found in *Y. pseudotuberculosis*.

2.2. Phylogeny and Consistency of Csps in E. coli

A total of 1013 Csp sequences in 135 *E. coli* strains (Supplementary Data 1) were retrieved from the PATRIC database. The phylogenetic tree showed that Csps were aggregated together into nine discrete regions of the tree, corresponding to nine Csp members in *E. coli*, CspA to CspI (Figure 2). Csps within each paralog were highly conserved across all the *E. coli* strains. In 135 *E. coli* strains, 20 Csp sequence patterns were identified, and the amino acid sequences were highly conserved in each Csp member (Figure S2, Table S5). The sequences of CspA in 135 representative *E. coli* strains were identical, as were the sequences of CspC, CspD and CspE. CspG had two different sequence patterns, whereas CspF, CspH and CspI all had three different patterns. Five sequence patterns were identified in CspB and it therefore appeared to be more polymorphic than other Csp members in *E. coli*. However, only four polymorphic sites at the most were identified between pairs of CspB sequences, and the substitutions were limited in structurally similar amino acids such as phenylalanine and tyrosine, or isoleucine and valine (Figure S2). For convenience, nine Csps in *E. coli* str. K-12 substr. MG1655 were considered representative Csps of *E. coli* species in the following analysis.

The pairwise evolutionary distances between nine representative Csps in *E. coli* were estimated according to the phylogenetic tree presented in Figure 2 and displayed in a distance matrix (Table S6). Of the nine Csp members, CspA, CspB, CspG and CspI were more similar to each other, with a maximum distance of 0.43. CspC and CspE, along with CspF and CspH were homologous Csps with evolutionary distances of 0.23 and 0.25, respectively. However, the CspF/CspH cluster was phylogenetically far from the other Csp members. All distances between CspF/CspH and other Csp members were larger than 1.2, indicating that the CspF/CspH cluster evolves independently. Similar to the CspF/CspH cluster, the minimum distance between CspD and other Csp members was 0.89, suggesting that CspD is also a Csp member evolving independently.

2.3. Comparison of Csps in Enteropathogenic Yersinia and E. coli

We constructed a phylogenetic tree combining representative Csps in enteropathogenic *Yersinia* and *E. coli* (Figure 3). Based on amino acid sequences, representative Csps in enteropathogenic *Yersinia* and *E. coli* were clustered into six phylogenetic groups (Figure 3 and Figure S3, Table S7). Cold-inducible Csps of *E. coli*, such as CspA, CspB, CspG and CspI, are included in group I, but Csps of *Yersinia* in group I form a monophyletic group separated from *E. coli* Csps. However, the groups still have a close phylogenetic relationship with each other, suggesting that these Csps should belong to one monophyletic group. All CspC and CspE of *Yersinia* and *E. coli* except CspE2 (*YE1546*) formed monophyletic group II, indicating that these Csp members are conserved in enteropathogenic *Yersinia* and *E. coli*. Similarly, in group V, CspDs also formed a monophyletic group. Group III consisted of only one CspE2 pattern identified in *Y. enterocolitica*. CspF and CspH patterns were identified only in *E. coli* and formed a monophyletic group VI far apart from other Csp patterns, with a specific RNP2 sequence (consensus VQ$\frac{V}{L}$H$\frac{I}{V}$). Similar to CspF and CspH in *E. coli*, three CspB patterns only identified in *Yersinia* (*YE1547*, *YE105_C2577* and *YPTB1423*) also had a specific RNP2 sequence (consensus VYV$\frac{S}{N}$K), forming monophyletic group IV.

2.4. Phylogenetic Diversity of Csps in Enterobacteriales

We assessed an alignment of 322 non-redundant Csp sequences in 104 Enterobacteriales strains (Table S8) to gain a more comprehensive view on Csp phylogeny. Tree reconstructions reinforced our previous results of clustering Csps into six groups, at least in Enterobacteriales (Figure 4). We also investigated the phylogeny of the 322 Csps using just their cold shock domain (CSD) sequences. The underlying topology of the latter tree (Figure S4) very closely matched the topology of the former tree (Figure 4), and showed similar phylogenetic grouping, suggesting that the phylogeny of Csps could be determined by their CSD sequences in Enterobacteriales.
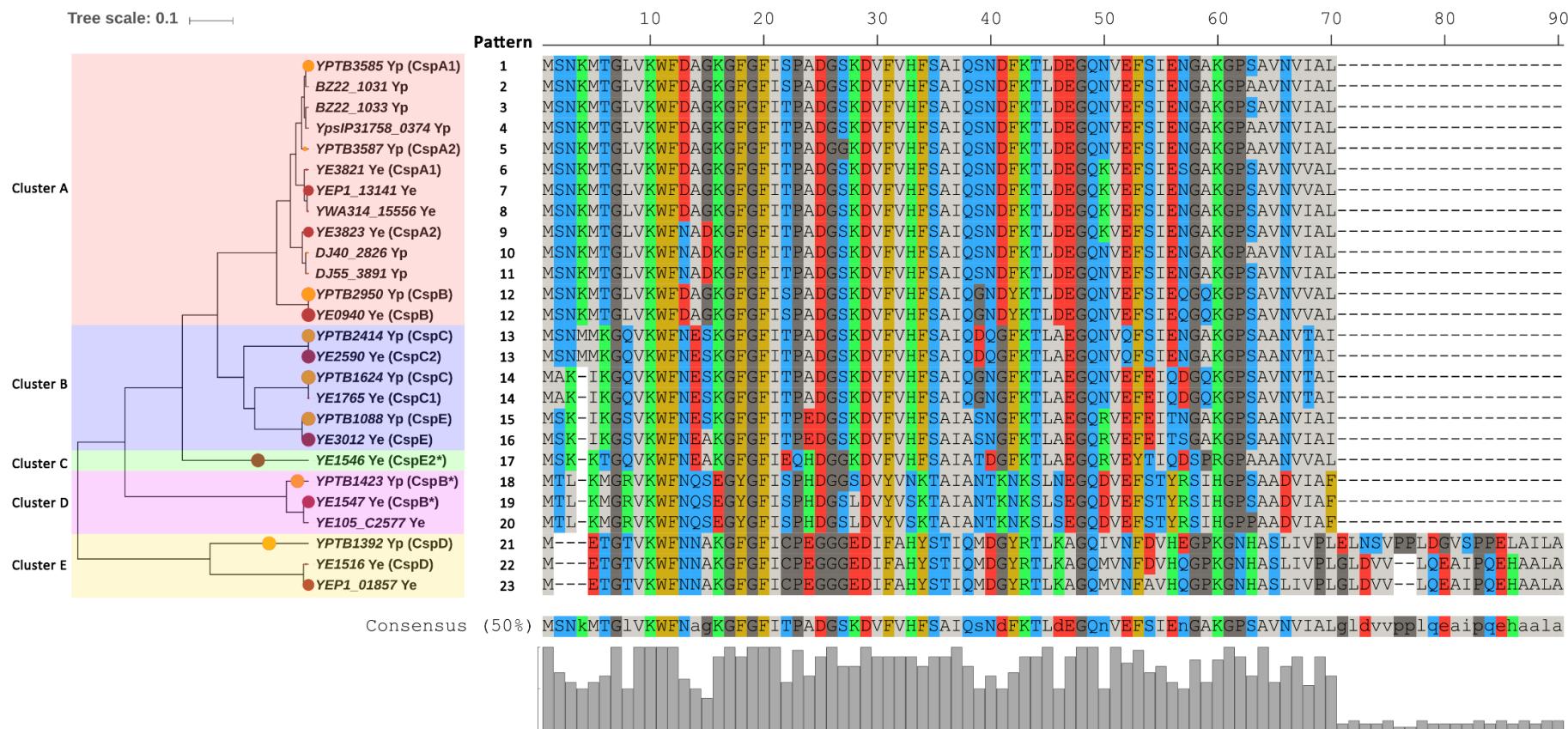
Figure 1. Phylogenetic cluster of Csp sequence patterns in enteropathogenic *Yersinia*. A phylogenetic tree composed of representative Csps of sequence patterns in enteropathogenic *Yersinia*. Circle radii represent the abundance ratios of each Csp pattern in *Y. pseudotuberculosis* (orange) or *Y. enterocolitica* (brown). Most Csps were named using the convention of *E. coli*, if any, except those marked with an asterisk. Histogram shows conservation of Csp sequence.
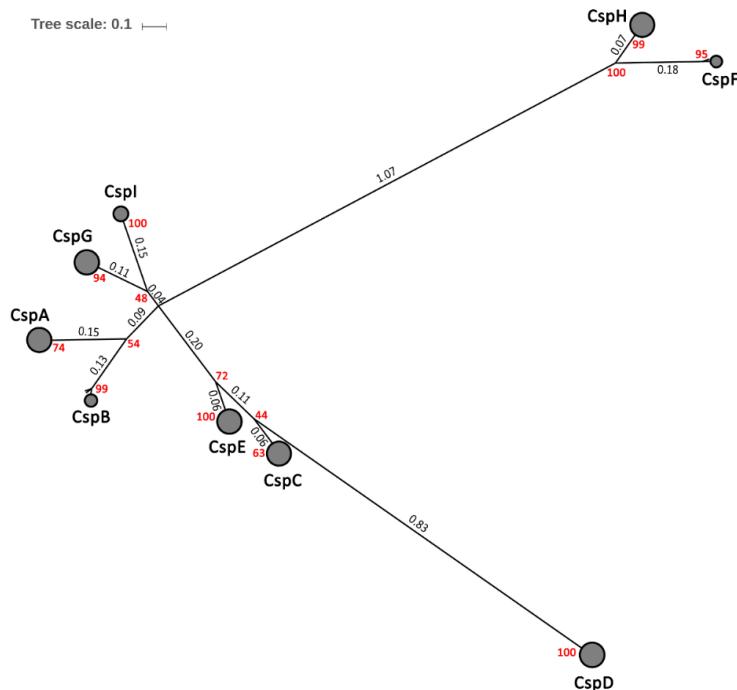
Figure 2. Phylogenetic tree for Csps in *Escherichia coli*. A phylogenetic tree composed on 1013 Csps in 135 *E. coli* strains. Monophyletic nodes have been collapsed and are represented by a circle. The number of Csps in a collapsed node is indicated by the circle radius. Black numbers indicate the evolutionary distances between the main nodes, and red numbers indicate the bootstrap values.
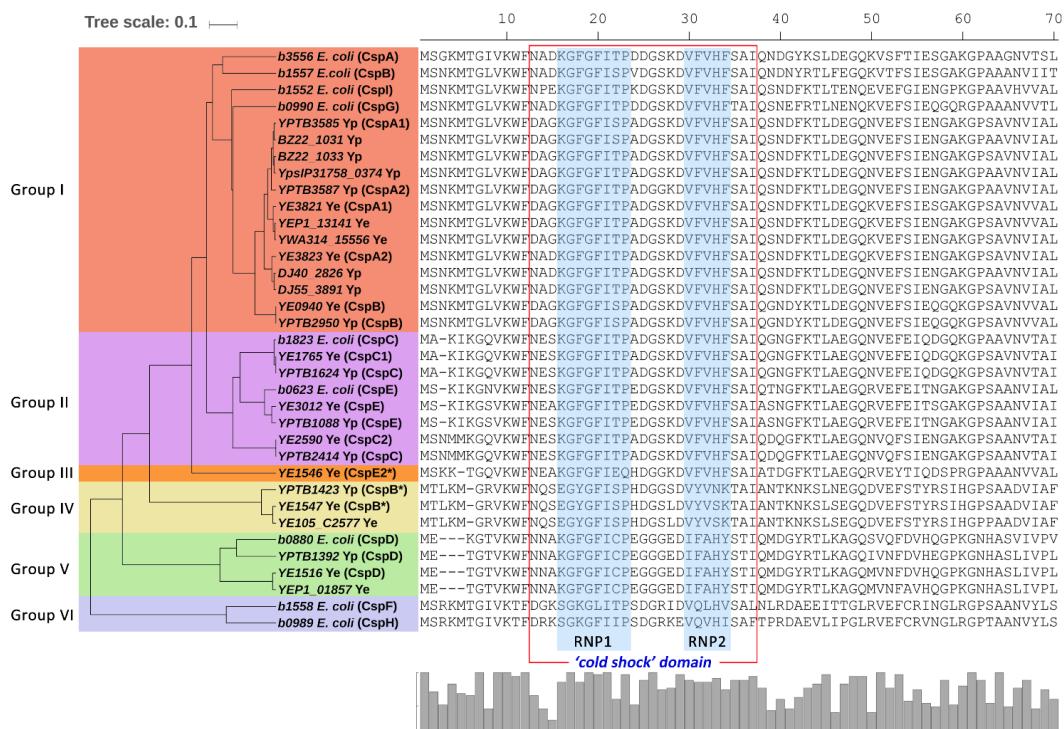
Figure 3. Phylogenetic group of representative Csps in *Escherichia coli* and enteropathogenic *Yersinia*. A phylogenetic tree and alignment of representative Csps in *E. coli* and enteropathogenic *Yersinia*. The phylogenetic groups of representative Csps were represented by color. Conserved RNP motifs of Csps were colored light blue. Most Csps were named using the convention of *E. coli*, if any, except those marked with an asterisk. The alignment is trimmed to the 70th site for better viewing. Histogram shows Csp sequence conservation.

2.5. Clarification of Phylogeny and Origin Time of Csp Genes in Eubacteria

To evaluate the phylogeny and origin time of Csp in eubacteria, we constructed maximum clade credibility (MCC) tree in BEAST for 89 *csp* gene sequences from 26 bacterial taxa (Figure 5). Phylogenetic analysis of these 89 Csps showed that Csps in eubacteria could be classified into five clades and 12 subclades (Table 1).

In addition, we estimated the tMRCA of all known *csp* genes in eubacteria to be 3585 MYA (95% highest posterior density (HPD) interval: 2132 to 5554 MYA), with a mean mutation rate of 4.392×10^{-4} substitutions per site per million years (95% HPD interval: 4.056×10^{-4} to 4.747×10^{-4}). To validate the reliability of this estimation, we also collected 26 gene sequences of 16S RNA from the 26 bacterial taxa, and calculated the tMRCA (3224 MYA with a 95% HPD interval from 2127 to 4696 MYA) and mean mutation rate (1.333×10^{-4} substitutions per site per million years with a 95% HPD interval from 9.550×10^{-5} to 1.710×10^{-4} substitutions per site per million years) of 16S rRNA genes in eubacteria using the same method. Our results are consistent with previous studies [34,35].
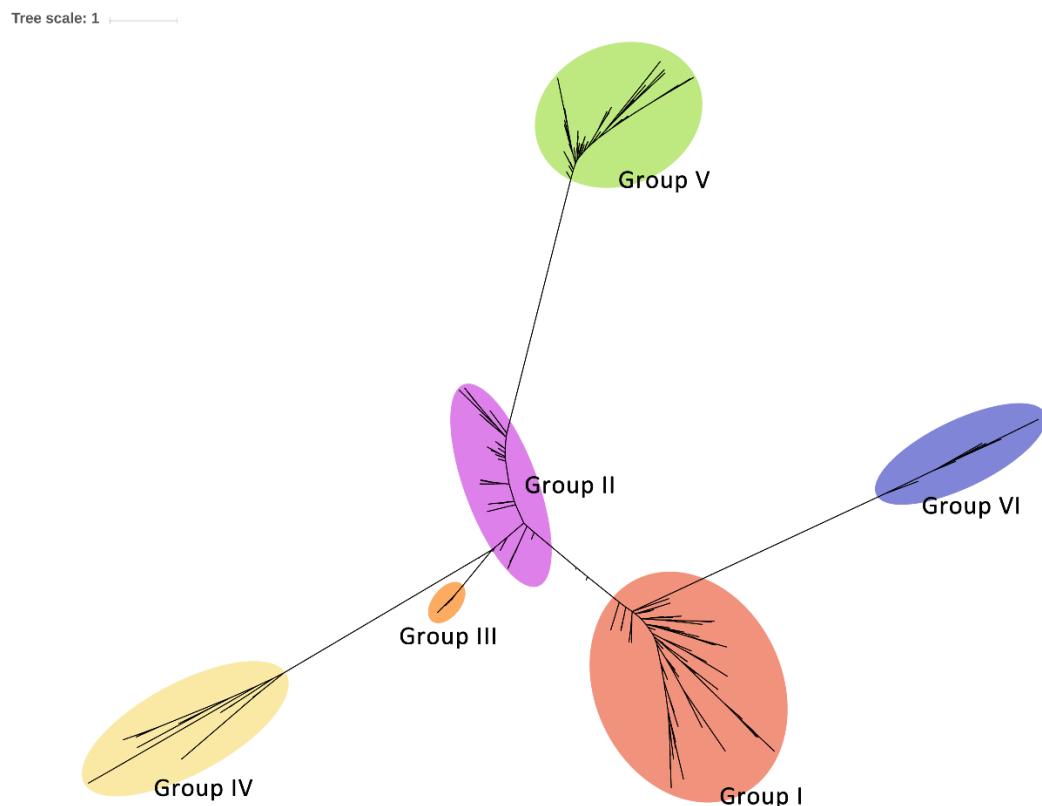
Figure 4. Phylogenetic tree for Csps in Enterobacteriales. A phylogenetic tree composed on 322 non-redundant Csps in 104 Enterobacteriales strains. The phylogenetic groups of Csps are represented by color. The complete tree with full bootstrap values is available in Newick format as Supplementary Data 2.
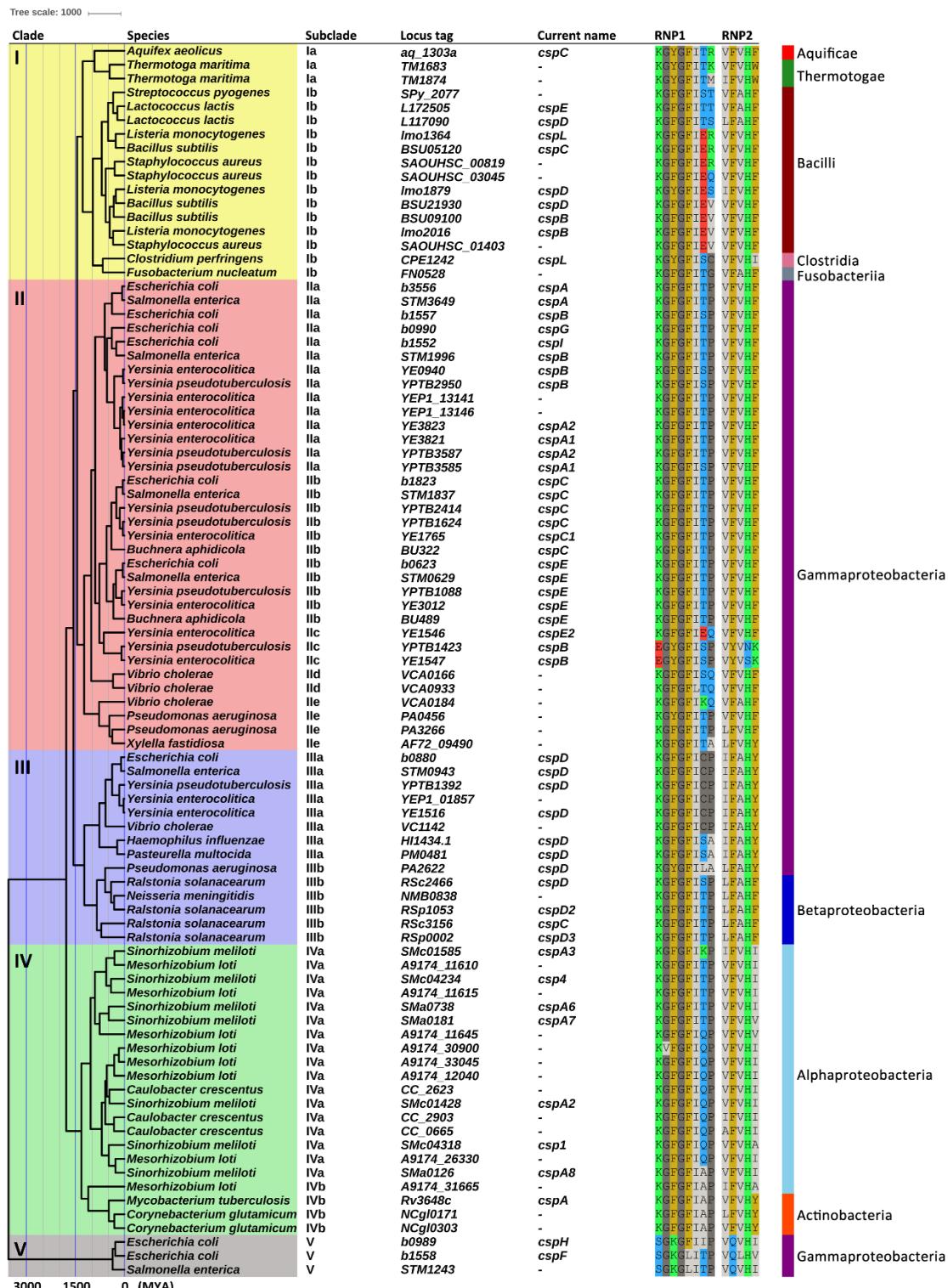
Figure 5. A maximum clade credibility tree of *csp* genes in eubacteria. Branch lengths are scaled to years. Branch lengths are given as millions of years ago (MYA). The phylogenetic clades of Csps are represented by color. The bacterial class of the organism containing each *csp* gene sequence is represented by the color of the strip.

Table 1. Cold shock protein gene clades in eubacteria.

Clade [1]	Subclade [1]	Representative *csp* Gene(s)	Currently Used Name(s) [2]	RNP1 [3]	RNP2 [3]	Known Function of Csps and Occurrence in Bacteria
I	Ia	*aq_1303a*	*cspC*	KGYGFITX	$\frac{V}{I}$FVH$\frac{W}{F}$	Exists in the hyperthermophilic bacterium.
	Ib	*Lmo2016, BSU05120, lmo1879, L172505, CPE1242*	*cspB, cspC, cspD, cspE, cspL*	KG$\frac{F}{Y}$GFIXX	XF$\frac{V}{A}$H$\frac{F}{I}$	Involves in regulation of cold and osmotic stress tolerance, virulence, cellular aggregation and flagella-based motility [7,15]; mainly exists in gram-positive bacteria.
II	IIa	*b3556, b1557, b0990, b1552*	*cspA, cspB, cspG, cspI*	KGFGFI$\frac{T}{S}$P	VFVHF	Cold-inducible [6,11,18]; exists in Gammaproteobacteria.
	IIb	*b1823, b0623*	*cspC, cspE*	KGFGFITP	VFVHF	Involves in cold adaptation (CspE) [29], transcriptional regulation and/or chromosome condensation [12,23,36]; exists in Gammaproteobacteria.
	IIc	*YPTB1423, YE1546*	*cspB, cspE2*	$\frac{K}{E}$G$\frac{F}{Y}$GFI$\frac{E}{S}$$\frac{Q}{P}$	V$\frac{F}{Y}$VX$\frac{F}{K}$	Involves in stress response *in vivo* [17]; mainly exists in *Yersiniaceae*.
	IId	*VCA0166*	-	KGFGF$\frac{I}{L}$$\frac{S}{T}$Q	VFVHF	Exists in *Vibrionaceae*.
	IIe	*PA0456*	-	KG$\frac{F}{Y}$GFI$\frac{K}{T}$X	$\frac{V}{L}$F$\frac{V}{A}$H$\frac{F}{Y}$	Exists in Gammaproteobacteria (not including Enterobacteriales).
III	IIIa	*b0880*	*cspD*	KGFGFI$\frac{C}{S}$$\frac{P}{A}$	IFAHY	Plays a role in the nutrient-stress response [25,37]; exists in Gammaproteobacteria.
	IIIb	*RSp1053, RSc3156*	*cspD2, cspC*	KG$\frac{F}{Y}$GFIX$\frac{P}{A}$	LFAH$\frac{F}{Y}$	Mainly exists in Betaproteobacteria.
IV	IVa	*SMc04318, SMc04234*	*csp1, csp4*	K$\frac{G}{V}$FGFIXP	XFVHX	Exists in Alphaproteobacteria.
	IVb	*Rv3648c*	*cspA*	KGFGFIAP	XFVH$\frac{Y}{A}$	Mainly exists in Actinobacteria.
V	V	*b0989, b1558*	*cspF, cspH*	SGKG$\frac{F}{L}$I$\frac{I}{T}$P	VQ$\frac{V}{L}$H$\frac{I}{V}$	Function unknown [10]; only exists in *Enterobacteriaceae*.

[1] Clustering of Csps in eubacteria is based on the maximum clade credibility tree presented in Figure 5. [2] Hyphen indicates a *csp* gene with no current name. [3] RNA-binding motif.

3. Discussion

In the present study, we investigated the diversity and consistency of Csp sequences in enteropathogenic *Yersinia, E. coli* and Enterobacteriales. A total of 23 Csp sequence patterns were identified by analyzing 447 Csps from 58 enteropathogenic *Yersinia* strains. Pattern 17 represented by CspE2 (*YE1546*) formed a monophyletic branch separated from other patterns, and no identical pattern was found in *Y. pseudotuberculosis*, suggesting this Csp pattern may perform certain specific functions in *Y. enterocolitica.* According to Figure 1, CspD patterns (*YE1516* and *YEP1_01857* in *Y. enterocolitica* and *YPTB1392* in *Y. pseudotuberculosis*) formed a monophyletic clade separate from other Csp members in enteropathogenic *Yersinia*, indicating that CspD is an independent Csp member. Noteworthily, three CspB patterns (*YE1547*, *YE105_C2577* in *Y. enterocolitica* and *YPTB1423* in *Y. pseudotuberculosis*) also formed a monophyletic clade, which separated from other CspB and CspA patterns in enteropathogenic *Yersinia*. Likewise, CspE2 (*YE1546* in *Y. enterocolitica*) also separated from other CspE patterns (patterns 15 and 16). These results confirm the perception that the currently used nomenclature is not appropriate for all Csps.

Moreover, 23 Csp sequence patterns in enteropathogenic *Yersinia* and nine representative Csps in *E. coli* were classified into six phylogenetic groups due to the alignment of their amino acid sequences (Figure 3). In group I, Csps from enteropathogenic *Yersinia* (patterns one to 12 in Figure 1) were phylogenetically closer to cold-inducible Csps (CspA, CspB, CspG and CspI) [18] than other Csps in *E. coli*. One of the unique features of *cspA*, *cspB*, *cspG* and *cspI* in *E. coli* is the unusually long (159, 160, 155 and 145 bases, respectively) 5′ untranslated region (5′-UTR), which contains a highly conserved unique sequence called the cold box [14,18]. Interestingly, the highly similar sequences were also identified in the 5′-UTR of enteropathogenic *Yersinia* Csps in group I. Another unique feature of cold-inducible *csp* genes of *E. coli* in group I is a highly conservative sequence located 14 bases downstream of the initiation codon, which has been termed the downstream box [10,18]. This element was also present in the *csp* genes of enteropathogenic *Yersinia* that were phylogenetically close to the cold-inducible group I Csps of *E. coli* (Figure 3 and Figure S5). These findings indicate that Csps from psychrotrophic enteropathogenic *Yersinia* in group I may also connect with cold-adaptation functions. However, the ability of enteropathogenic *Yersinia* to grow at near-zero temperatures does not appear to depend on these *csp* genes. In mesophilic *E. coli*, although CspE can complement the cold-sensitive phenotype of the quadruple △*cspA*△*cspB*△*cspG*△*cspE* deletion strain [29], CspE was classified as group II together with CspC (Figure 3), considering high similarity in the protein sequence and roles in transcription regulation [23,36]. Furthermore, no cold box or downstream box was identified in *cspE* in *E. coli*, demonstrating it should not be classified as group I.

Our study defined two *Yersinia*-specific Csp groups. Group III is represented by *YE1546* in *Y. enterocolitica*, and Group IV is represented by *YE1547* and *YE105_C2577* in *Y. enterocolitica* and *YPTB1423* in *Y. pseudotuberculosis*. *YPTB1423* is upregulated during growth at low temperatures in *Y. pseudotuberculosis* IP32953 [38]. However, their role in the cold adaptation of enteropathogenic *Yersinia* should be clarified.

Group V is mainly formed by CspDs in *E. coli* and enteropathogenic *Yersinia*. Stationary-phase-induced CspD (b0880) in *E. coli* [25] may evolve independently due to the large phylogenetic distance of CspD from other Csps (Figure 2). Group VI is represented by CspF (b1558) and CspH (b0989) in *E. coli*. Functions of Csps in this group are still unknown. CspF and CspH in *E. coli* are probably also independent in evolution, considering the phylogenetic distance of CspF/CspH from other Csps. The orthologs of CspF or CspH in *E. coli* could not be found in Csps in *Yersinia*.

The classification of Csps into six groups was still valid when extending Csps in the order of Enterobacteriales (Figure 4), but not in eubacteria. Phylogenetic analysis of the 89 Csps from 26 taxa suggested that Csps in eubacteria could be classified into five clades and 12 subclades (Figure 5, Table 1). Homologues of CspF and CspH in *E. coli* (clade V, in which group VI of Enterobacteriales Csps is included) were most distantly related to all the other known Csp homologues. Moreover, they did not express one of the characteristic features of other Csp homologues, i.e., a high aromatic amino

acid content, which is an essential feature required for the RNA-binding role of Csps [12]. CspF and CspH in *E. coli* only have three and four aromatic residues, respectively, in contrast to approximately eight aromatic residues found in other Csp homologues [12]. In addition to the distinctive assumed RNP sequences, a substantial evolutionary divergence has occurred between CspF/CspH homologues (clade V) and other Csp homologues (clades I to IV). In clade I, subclade Ia is hyperthermophilic and bacteria-specific. Csps classified as subclade Ib mainly existed in gram-positive bacteria, and some are involved in the regulation of cold and osmotic stress tolerance, virulence, cellular aggregation and flagella-based motility [7,15]. Clade II is the largest group that includes most of the 'classical Csps', consisting of subclades IIa to IIe. A subset of the Csps in this clade are involved in cold tolerance and transcriptional regulation [6,11,12,18,23,29,36]. Most known cold-inducible Csps are classified as subclade IIa, in which group I of Enterobacteriales Csps is included. Many transcription regulation-related Csps belong to subclade IIb, in which Enterobacteriales Csp group II is included. Subclade IIc is *Yersiniaceae*-specific, including groups III and IV of Enterobacteriales Csps (Figures 4 and 5). Csp functions are unknown in subclades IId and IIe. Clade III contains *cspD* genes, which were separated from the other *csp* genes in gram-negative bacteria. The representative Csp of clade III is CspD (*b0880*) in *E. coli*, which participates in responses to nutrient stress [25,37]. Group V of Enterobacteriales Csps is included in subclade IIIa (Figures 4 and 5), and it is noteworthy that the currently named *cspD* genes in gram-positive bacteria were not included in this clade. Furthermore, this CspD clade of gram-negative bacteria only included Betaproteobacteria and Gammaproteobacteria, indicating that CspD is specific for these two classes. Generally, *csp* genes in the same bacterial class were clustered together, suggesting that *csp* genes represented in each class evolved from a single ancestral Csp homologue, except *cspD* genes in clade III. Clade IV contains Csps from Alphaproteobacteria and Actinobacteria. The functions of Csps in this clade are unclear.

Divergence time estimates for the major groups of eubacteria range between 2.5 and 3.2 billion years ago [35]. However, CSD-containing proteins have been found in all three domains of life [39–43], suggesting that a CSD-like protein was present before the divergence of bacteria and eukarya/archaea. Thus, an ancestral Csp may be found at the beginning of single-cell evolution, approximately 3.5 billion years ago [37]. This inference is supported by our evaluation (Figure 5). Csps are small and conserved proteins with large numbers of homologues. However, the nomenclature of *csp* genes is misleading and limited, due to being first found in cold shock studies. In fact, *csp* genes are probably general stress response or adaptation genes. Csps play a more general role than their sole implication in cold adaptation. They have been important tools in bacterial adaption to various, often disadvantageous, conditions. The presence of Csps in Archaea [39,40] and the earliest-diverging hyperthermophilic bacteria [42,43] also indicates that bacteria may originally have needed them to survive extreme conditions, not just cold. Compared with the synonymous mutation rate (4.5×10^{-3} substitutions per site per million years) in *E. coli* [34], the low mean mutation rate (4.392×10^{-4} substitutions per site per million years) of *csp* genes in eubacteria has limited their heterogeneity. Csps have not been adapting to change rapidly and the evolution of *csp* genes seems to be slow compared to many other genes.

4. Materials and Methods

4.1. Selection of Csp Sequences from Databases

All the Csp sequence data for the eubacteria used in our work were collected from the Pathosystems Resource Integration Center (PATRIC; https://www.patricbrc.org/) [44] and the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/). In *E. coli* and enteropathogenic *Yersinia*, Csp sequences are 65 to 75 [5,31] and 69 to 87 (this study) amino acids (aa) in length, respectively. Thus, the amino acid sequences annotated as cold shock proteins but shorter than 65 aa or longer than 87 aa were excluded from further analyses.

4.2. Protein Clustering, Alignment and Phylogeny

To generate non-redundant datasets for the phylogenetic analyses, 447 Csp sequences of enteropathogenic *Yersinia* (104 Csps from 13 *Y. enterocolitica* strains and 343 Csps from 45 *Y. pseudotuberculosis* strains), 1013 Csp sequences from 135 *E. coli* and 534 Csp sequences from 104 Enterobacteriales were clustered as three separate datasets at the 100% identity level using CD-Hit v4.5.4 [45]. Nine Csps in *E. coli* str. K-12 substr. MG1655 were considered as representative Csps of *E. coli* species in the following analyses. For Csp phylogeny, the MAFFT v7 (https://mafft.cbrc.jp/alignment/server/index.html) [46] with G-INS-i strategy (Unalignlevel 0.3, "leave gappy regions" set, and other default parameters) was used for all alignments. Maximum likelihood phylogenetic trees of aligned proteins were inferred with RAxML v8.2.9 [47], with a protein-specific amino-acid substitution model identified using RAxML (PROTGAMMAAUTO); tree topologies were checked by 100 bootstrapping replicates. iTOL v3 (http://itol.embl.de/) [48] was used to visualize the trees.

4.3. Estimation of Substitution Rate and Origin Time of Csp and 16S rRNA

The gene sequences of 89 Csps and 26 16S rRNAs from 26 taxa in eubacteria were obtained from PATRIC, respectively. The GTR+I+G4 and GTR+F+I+G4 substitution models were selected as the best-fit nucleotide substitution models for Csp and 16S rRNA respectively, by ModelFinder, which is included in IQ-TREE v1.5.4 [49]. A Bayesian framework using BEAST v2.4.7 and associated software (Beauti v2.4.7, Tracer v1.6.0, LogCombiner v2.4.7, TreeAnnotator v2.4.7) [50] were used to date the phylogeny and estimate the mutation rate for Csp and 16S rRNA in eubacteria, respectively. Analyses used the GTR model of nucleotide substitution with four discrete gamma-distributed rate categories and were run using a log-normal relaxed clock and the coalescent constant population model of speciation. The default settings were used to estimate priors for all other parameters. A run was also performed without data to assess prior distributions and allow comparisons with posterior distributions. The Markov chain Monte Carlo (MCMC) in BEAST was run for 100 million generations and sampled every 10,000 generations. Four replicate runs were performed to check for convergence of the MCMC and combined together after discarding 10 million generations from each run as burn-in. We extracted the clock rate and tMRCA estimate, and their distributions, with Tracer v1.6.0. Finally, the MCC tree was built using TreeAnnotator v1.6.0 with 10% burn-in and visualized in Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL v3.

4.4. Molecular Clock Calibration

The present phylogenetic analyses showed that CspD proteins belonging to the CspD family formed a monophyletic group separated from other Csp members in Enterobacteriales (Figure 4), assuming that CspD evolves independently of other Csps. We calibrated the node that corresponded with the presumed divergence event between *cspD* genes in *E. coli* and *Salmonella enterica* (*b0880* and *STM0943*). To ensure the reliability of this estimation, we calibrated another node, which corresponded to the divergence event of *cspD* genes in *Haemophilus influenzae* and *Pasteurella multocida* (*HI1434.1* and *PM0481*). The priors for these nodes were assigned normal distribution with standard deviation (s.d.) based on the divergence time of *E. coli* and *S. enterica*, *H. influenzae* and *P. multocida*, respectively. The divergence event of *b0880–STM0943* was calibrated by setting the prior for that node to 102 MYA with a s.d. of 3.0, while *HI1434.1–PM0481* divergence was calibrated at 220 MYA with a s.d. of 11.0 [35].

The reported credible intervals around estimates are 95% HPD intervals. Ages of various nodes were estimated from the resulting tree, using the 95% HPD as an estimate of error for the age of each node. When reporting the mutation rate from a relaxed-clock model, the mean rate (mean of the rates of each branch weighted by the time length of the branch) is given.

4.5. Data Availability

The log files generated by BEAST and phylogenetic trees are available online (https://github.com/fffish8888/Csp-2018). The authors declare that the other main data supporting the findings of this study are available within this Article and in the Supplementary Information files.

Other data that support the findings of this study are available from the corresponding author upon request.

Abbreviations

5′-UTR	5′ untranslated region
aa	amino acids
BEAST	Bayesian Evolutionary Analysis Sampling Trees
Cips	Cold-induced proteins
CSD	Cold shock domain
CSPs	Cold shock proteins
HPD	Highest posterior density
MAFFT	Multiple Alignment using Fast Fourier Transform
MCMC	Markov chain Monte Carlo
MCC	Maximum clade credibility
MYA	Million years ago
NCBI	National Center for Biotechnology Information
PATRIC	Pathosystems Resource Integration Center
RAxML	Randomized Axelerated Maximum Likelihood
s.d.	standard deviation
tMRCA	time to the most recent common ancestor

References

1. Laukkanen-Ninios, R.; Fredriksson-Ahomaa, M.; Korkeala, H. Enteropathogenic *Yersinia* in the pork production chain: Challenges for control. *Compr. Rev. Food Sci. Food Saf.* **2014**, *13*, 1165–1191. [CrossRef]
2. Fredriksson-Ahomaa, M.; Lindström, M.; Korkeala, H. Yersinia enterocolitica and *Yersinia pseudotuberculosis*. In *Pathogens and Toxins in Foods*; Juneja, V.K., Sofos, J.N., Eds.; ASM Press: Washington, DC, USA, 2010; Volume 11, pp. 164–180.
3. Keto-Timonen, R.; Pöntinen, A.; Aalto-Araneda, M.; Korkeala, H. Growth of *Yersinia pseudotuberculosis* strains at different temperatures, pH values, and NaCl and ethanol concentrations. *J. Food Prot.* **2017**, *81*, 142–149. [CrossRef] [PubMed]
4. Barria, C.; Malecki, M.; Arraiano, C.M. Bacterial adaptation to cold. *Microbiology* **2013**, *159*, 2437–2443. [CrossRef] [PubMed]
5. Keto-Timonen, R.; Hietala, N.; Palonen, E.; Hakakorpi, A.; Lindström, M.; Korkeala, H. Cold shock proteins: A minireview with special emphasis on Csp-family of enteropathogenic *Yersinia*. *Front. Microbiol.* **2016**, *7*, 1–7. [CrossRef]
6. Jones, P.G.; VanBogelen, R.A.; Neidhardt, F.C. Induction of proteins in response to low temperature in *Escherichia coli*. *J. Bacteriol.* **1987**, *169*, 2092–2095. [CrossRef]
7. Schmid, B.; Klumpp, J.; Raimann, E.; Loessner, M.J.; Stephan, R.; Tasara, T. Role of cold shock proteins in growth of *Listeria monocytogenes* under cold and osmotic stress conditions. *Appl. Environ. Microbiol.* **2009**, *75*, 1621–1627. [CrossRef]

8. Graumann, P.; Schröder, K.; Schmid, R.; Marahiel, M.A. Cold shock stress-induced proteins in *Bacillus subtilis*. *J. Bacteriol.* **1996**, *178*, 4611–4619. [CrossRef]

9. Wemekamp-Kamphuis, H.H.; Karatzas, A.K.; Wouters, J.A.; Abee, T. Enhanced levels of cold shock proteins in *Listeria monocytogenes* LO28 upon exposure to low temperature and high hydrostatic pressure. *Appl. Environ. Microbiol.* **2002**, *68*, 456–463. [CrossRef]

10. Wang, N.; Yamanaka, K.; Inouye, M. CspI, the ninth member of the CspA family of *Escherichia coli*, is induced upon cold shock. *J. Bacteriol.* **1999**, *181*, 1603–1609.

11. Goldstein, J.; Pollitt, N.S.; Inouye, M. Major cold shock protein of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 283–287. [CrossRef]

12. Yamanaka, K.; Fang, L.; Inouye, M. The CspA family in *Escherichia coli*: Multiple gene duplication for stress adaptation. *Mol. Microbiol.* **1998**, *27*, 247–255. [CrossRef] [PubMed]

13. Nakashima, K.; Kanamaru, K.; Mizuno, T.; Horikoshi, K. A novel member of the *cspA* family of genes that is induced by cold shock in *Escherichai coli*. *J. Bacteriol.* **1996**, *178*, 2994–2997. [CrossRef] [PubMed]

14. Jiang, W.; Fang, L.; Inouye, M. The role of the 5′-end untranslated region of the mRNA for CspA, the major cold-shock protein of *Escherichia coli*, in cold-shock adaptation. *J. Bacteriol.* **1996**, *178*, 4919–4925. [CrossRef] [PubMed]

15. Eshwar, A.K.; Guldimann, C.; Oevermann, A.; Tasara, T. Cold-shock domain family proteins (Csps) are involved in regulation of virulence, cellular aggregation, and flagella-based motility in *Listeria monocytogenes*. *Front. Cell. Infect. Microbiol.* **2017**, *7*, 1–15. [CrossRef] [PubMed]

16. Uppal, S.; Rao Akkipeddi, V.S.N.; Jawali, N. Posttranscriptional regulation of *cspE* in *Escherichia coli*: Involvement of the short 5′-untranslated region. *FEMS Microbiol. Lett.* **2008**, *279*, 83–91. [CrossRef] [PubMed]

17. Lathem, W.W.; Crosby, S.D.; Miller, V.L.; Goldman, W.E. Progression of primary pneumonic plague: A mouse model of infection, pathology, and bacterial transcriptional activity. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17786–17791. [CrossRef] [PubMed]

18. Phadtare, S. Recent developments in bacterial cold-shock response. *Curr. Issues Mol. Biol.* **2004**, *6*, 125–136. [PubMed]

19. Annamalai, T.; Venkitanarayanan, K. Expression of major cold shock proteins and genes by *Yersinia enterocolitica* in synthetic medium and foods. *J. Food Prot.* **2005**, *68*, 2454–2458. [CrossRef]

20. Neuhaus, K.; Francis, K.P.; Rapposch, S.; Görg, A.; Scherer, S. Pathogenic *Yersinia* species carry a novel, cold-inducible major cold shock protein tandem gene duplication producing both bicistronic and monocistronic mRNA. *J. Bacteriol.* **1999**, *181*, 6449–6455.

21. Phadtare, S.; Inouye, M. Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of *Escherichia coli*. *Mol. Microbiol.* **1999**, *33*, 1004–1014. [CrossRef]

22. Lee, S.J.; Xie, A.; Jiang, W.; Etchegaray, J.-P.; Jones, P.G.; Inouye, M. Family of the major cold-shock protein, CspA (CS7.4), of *Escherichia coli*, whose members show a high sequence similarity with the eukaryotic Y-box binding proteins. *Mol. Microbiol.* **1994**, *11*, 833–839. [CrossRef]

23. Phadtare, S.; Inouye, M. Role of CspC and CspE in regulation of expression of RpoS and UspA the stress response proteins in *Escherichia coli*. *J. Bacteriol.* **2001**, *183*, 1205–1214. [CrossRef]

24. Bae, W.; Phadtare, S.; Severinov, K.; Inouye, M. Characterization of *Escherichia coli cspE*, whose product negatively regulates transcription of *cspA*, the gene for the major cold shock protein. *Mol. Microbiol.* **1999**, *31*, 1429–1441. [CrossRef]

25. Yamanaka, K.; Inouye, M. Growth-phase-dependent expression of *cspD*, encoding a member of the CspA family in *Escherichia coli*. *J. Bacteriol.* **1997**, *179*, 5126–5130. [CrossRef]

26. Hébraud, M.; Potier, P. Cold shock response and low temperature adaptation in psychrotrophic bacteria. *J. Mol. Microbiol. Biotechnol.* **1999**, *1*, 211–219.

27. Landsman, D. RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain. *Nucleic Acids Res.* **1992**, *20*, 2861–2864. [CrossRef]

28. Schindelin, H.; Marahiel, M.A.; Heinemann, U. Universal nucleic acid-binding domain revealed by crystal structure of the *B. subtilis* major cold-shock protein. *Nature* **1993**, *364*, 164–168. [CrossRef]

29. Phadtare, S.; Inouye, M.; Severinov, K. The nucleic acid melting activity of *Escherichia coli* CspE is critical for transcription antitermination and cold acclimation of cells. *J. Biol. Chem.* **2002**, *277*, 7239–7245. [CrossRef]

30. Rath, D.; Jawali, N. Loss of expression of *cspC*, a cold shock family gene, confers a gain of fitness in *Escherichia coli* K-12 strains. *J. Bacteriol.* **2006**, *188*, 6780–6785. [CrossRef]

31. Graumann, P.; Marahiel, M.A. Some like it cold: Response of microorganisms to cold shock. *Arch. Microbiol.* **1996**, *166*, 293–300. [CrossRef]

32. Chapot-Chartier, M.P.; Schouler, C.; Lepeuple, A.S.; Gripon, J.C.; Chopin, C. Characterization of *cspB*, a cold-shock-inducible gene from *Lactococcus lactis*, and evidence for a family of genes homologous to the *Escherichia coli cspA* major cold shock gene. *J. Bacteriol.* **1997**, *179*, 5589–5593. [CrossRef]

33. Mayo, B.; Derzelle, S.; Fernandez, M.; Leonard, C.; Ferain, T.; Hols, P.; Suárez, J.E.; Delcour, J. Cloning and characterization of *cspL* and *cspP*, two cold-inducible genes from *Lactobacillus plantarum*. *J. Bacteriol.* **1997**, *179*, 3039–3042. [CrossRef]

34. Ochman, H.; Elwyn, S.; Moran, N.A. Calibrating bacterial evolution. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 12638–12643. [CrossRef]

35. Battistuzzi, F.U.; Feijao, A.; Hedges, S.B. A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol. Biol.* **2004**, *4*, 44. [CrossRef]

36. Shenhar, Y.; Biran, D.; Ron, E.Z. Resistance to environmental stress requires the RNA chaperones CspC and CspE. *Environ. Microbiol. Rep.* **2012**, *4*, 532–539. [CrossRef]

37. Graumann, P.L.; Marahiel, M.A. A superfamily of proteins that contain the cold-shock domain. *Trends Biochem. Sci.* **1998**, *23*, 286–290. [CrossRef]

38. Virtanen, J.-P.; Keto-Timonen, R.; Jaakkola, K.; Salin, N.; Korkeala, H. Changes in transcriptome of *Yersinia pseudotuberculosis* IP32953 grown at 3 °C and 28 °C detected by RNA sequencing shed light on cold adaptation. *Front. Cell. Infect. Microbiol.* **2018**, *8*, 416. [CrossRef]

39. Cavicchioli, R. Cold-adapted archaea. *Nat. Rev. Microbiol.* **2006**, *4*, 331. [CrossRef]

40. Giaquinto, L.; Curmi, P.M.G.; Siddiqui, K.S.; Poljak, A.; DeLong, E.; DasSarma, S.; Cavicchioli, R. Structure and function of cold shock proteins in archaea. *J. Bacteriol.* **2007**, *189*, 5738–5748. [CrossRef]

41. Mihailovich, M.; Militti, C.; Gabaldón, T.; Gebauer, F. Eukaryotic cold shock domain proteins: Highly versatile regulators of gene expression. *BioEssays* **2010**, *32*, 109–118. [CrossRef]

42. Deckert, G.; Warren, P.V.; Gaasterland, T.; Young, W.G.; Lenox, A.L.; Graham, D.E.; Overbeek, R.; Snead, M.A.; Keller, M.; Aujay, M.; et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **1998**, *392*, 353. [CrossRef]

43. Perl, D.; Welker, C.; Schindler, T.; Schröder, K.; Marahiel, M.A.; Jaenicke, R.; Schmid, F.X. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* **1998**, *5*, 229. [CrossRef]

44. Wattam, A.R.; Davis, J.J.; Assaf, R.; Boisvert, S.; Brettin, T.; Bun, C.; Conrad, N.; Dietrich, E.M.; Disz, T.; Gabbard, J.L.; et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* **2017**, *45*, D535–D542. [CrossRef]

45. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef]

46. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]

47. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]

48. Letunic, I.; Bork, P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **2016**, *44*, W242–W245. [CrossRef]

49. Trifinopoulos, J.; Nguyen, L.-T.; von Haeseler, A.; Minh, B.Q. W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **2016**, *44*, W232–W235. [CrossRef]

50. Bouckaert, R.; Heled, J.; Kühnert, D.; Vaughan, T.; Wu, C.H.; Xie, D.; Suchard, M.A.; Rambaut, A.; Drummond, A.J. BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* **2014**, *10*, 1–6. [CrossRef]