Supplementary Data for:

An internal loop region is responsible for inherent target specificity of bacterial Cold-shock proteins

Satoshi Hasegawa[1,2,†], Rerina Inose[1,†], Mizuki Igarashi[1,3], Megumi Tsurumaki[1], Motofumi Saito[1], Tatsuo Yanagisawa[4], Akio Kanai[1,2,3] and Teppei Morita[1,3,*]

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, 997-0019, Japan

[2]Faculty of Environment and Information studies, Keio University, Fujisawa, Kanagawa, 252-0882, Japan

[3]Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa, 252-0882, Japan

[4]RIKEN Center for Biosystems Dynamics Research, Yokohama, Kanagawa, 230-0045, Japan

*Corresponding author. Tel: +81-235-29-0575; Email: morita-t@ttck.keio.ac.jp

†These authors contributed equally to this work

List of Supplementary Files

Supplementary Figures:

Supplementary Tables:

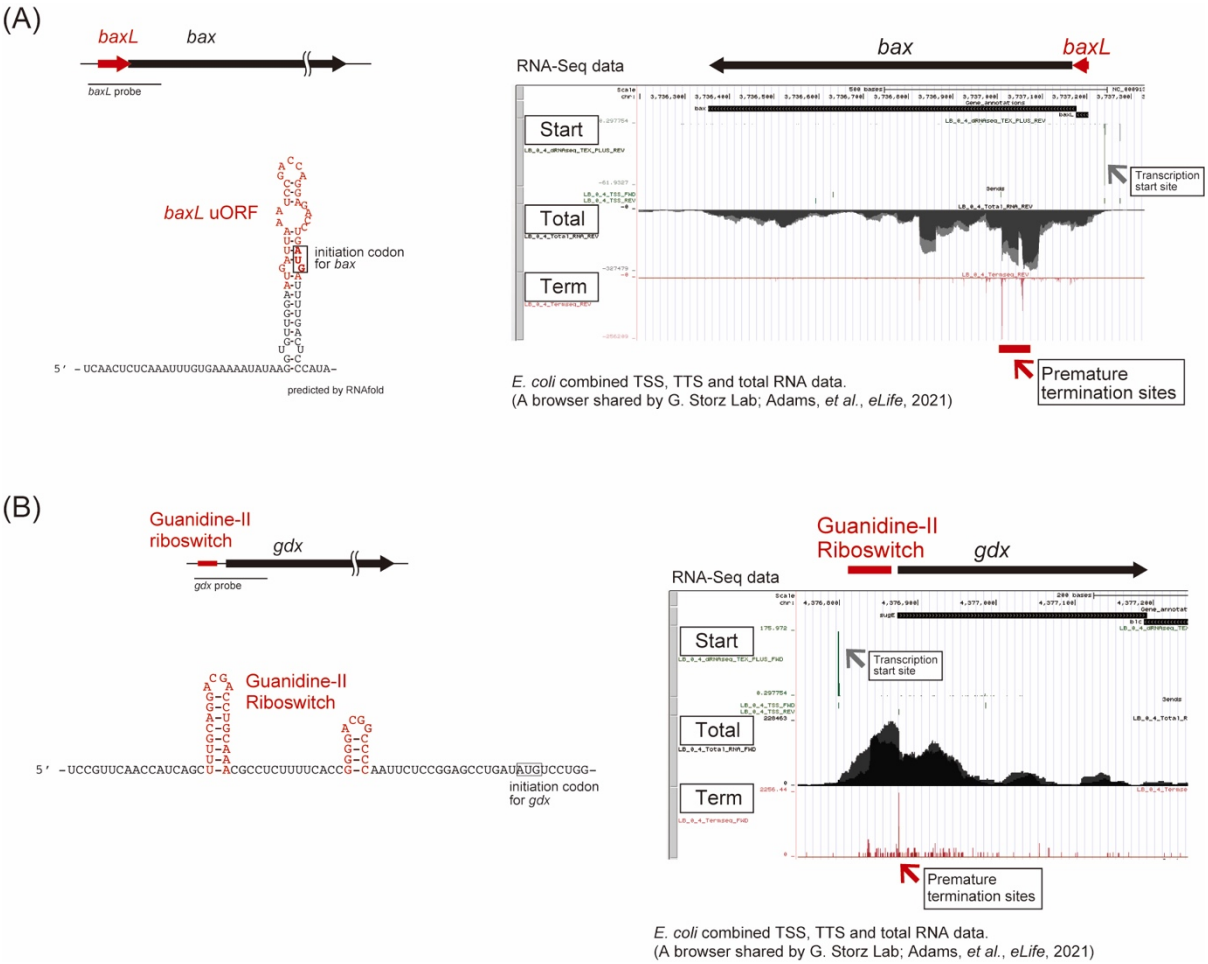

Figure S1. RNA structures adopted by the 5' region of *baxL-bax* mRNA (A) or *gdx* mRNA (B). The predicted structure of *baxL-bax* leader region was obtained by RNAfold WebServer (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). The structure of *gdx* riboswitch was shown based on Sherlock, et al. (2017). Lines below the genes indicate the positions of the probes used for Northern blots. Transcriptional profiles and positions of promoters and termination sites were obtained from the browser of *E. coli* combined TSS, TTS and total RNA data (https://www.nichd.nih.gov/research/atNICHD/Investigators/storz/data-protocols/browsers#term-seq) and the EcoCyc database (Keseler et al. 2017).

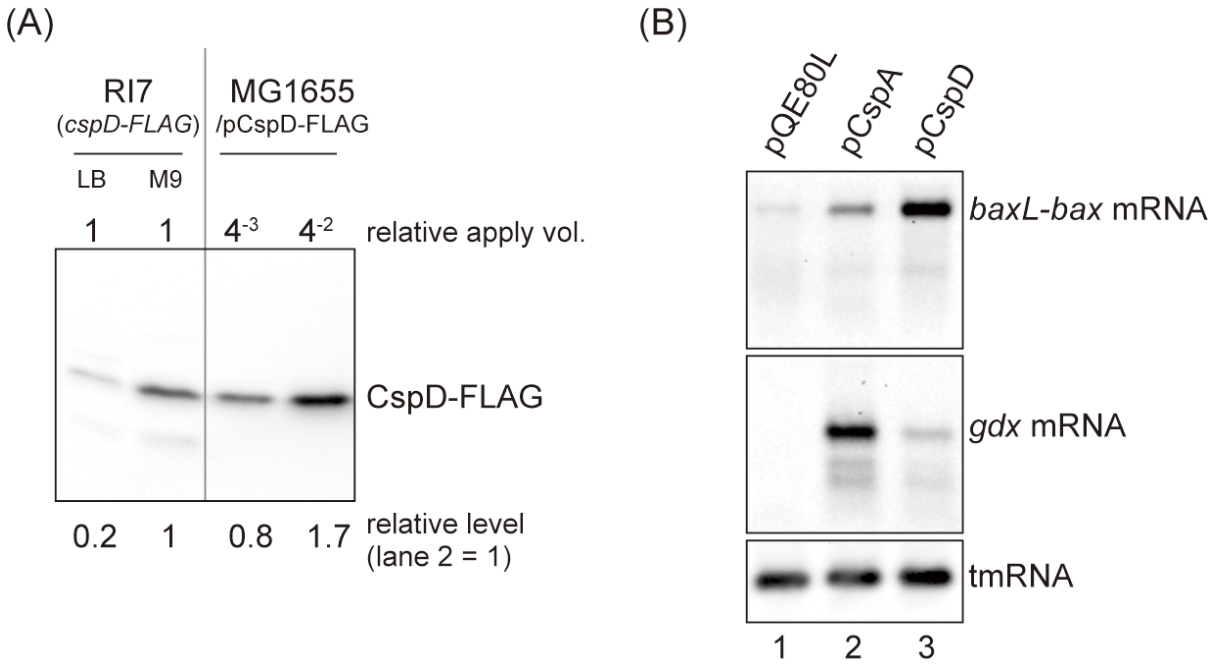

Figure S2. (A) Determination of CspD levels from the plasmid. MG1655 harboring pCspD-FLAG cells were grown in LB-ampicillin medium. At an OD_{600} of 0.3, 0.2 mM IPTG was added to cultures to induce CspD-FLAG. Incubation was continued for 30 min. RI7 (*cspD-FLAG*) cells were grown in M9 medium with 10 mM sodium acetate (OD_{600} 0.4). The protein sample was subjected to Western blotting using anti-FLAG monoclonal antibody. Relative protein levels were calculated, with the protein sample of RI7_M9 (lane 2) set to 1. The data shown are representative of the two independent experiments. (B) Effect of non-tagged Csps. MG1655 cells harboring the indicated plasmids were grown in LB-ampicillin medium. At an OD_{600} of 0.3, 0.2 mM IPTG was added to cultures to induce Csps. Incubation was continued for 30 min. The RNA sample was subjected to Northern blotting using probes for *baxL-bax* mRNA, *gdx* mRNA or tmRNA.

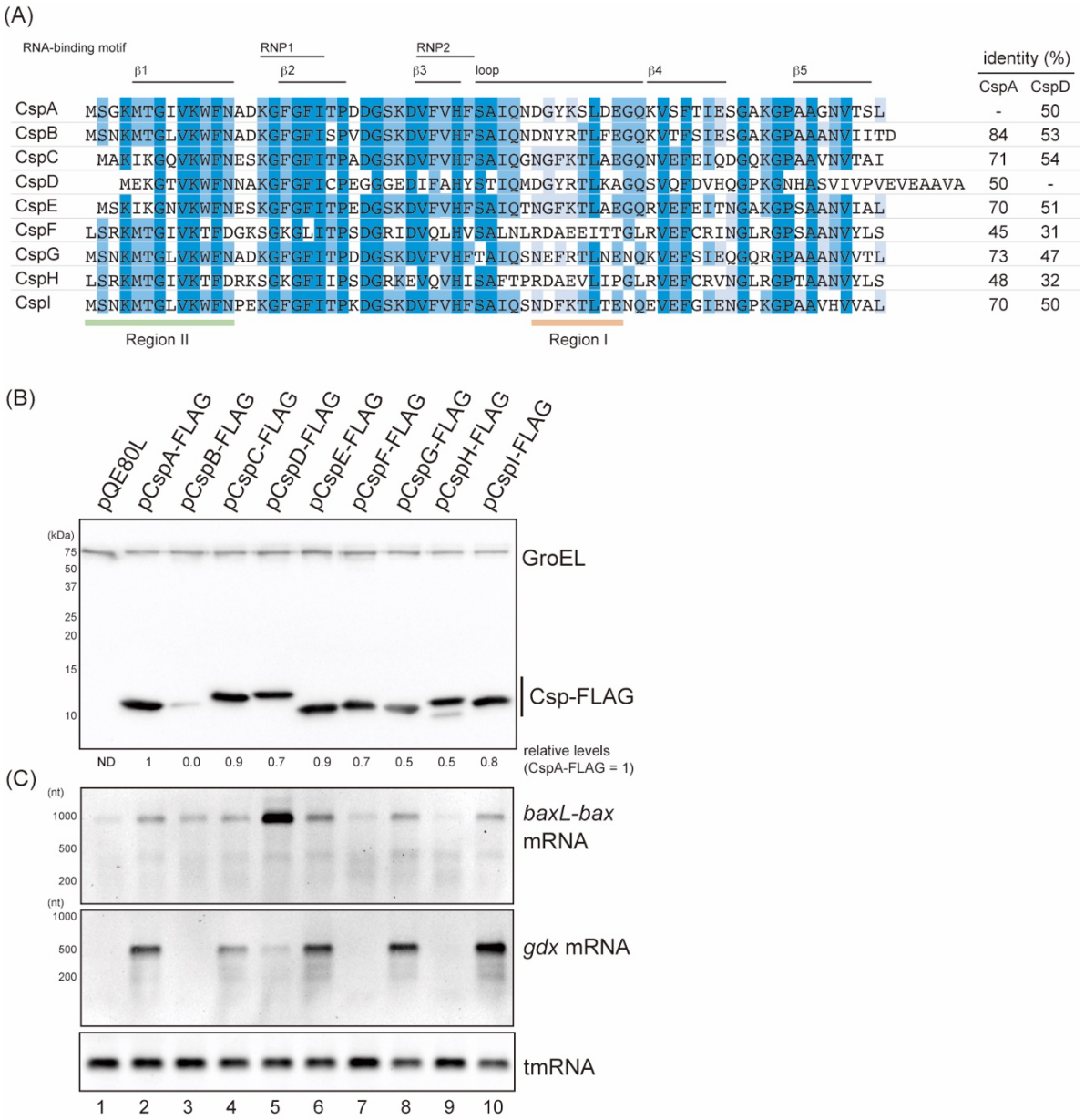

Figure S3. Properties of *E. coli* Csps. (A) Sequence alignment of *E. coli* Csps. Full-length of amino acid sequences are shown. Cold-shock domain and RNA binding motifs, RNP1 and RNP2, are indicated by lines on top of the sequence; Region I and Region II are indicated by lines below the sequence. Amino acid residues with more than 80% identity are marked in dark blue; those with identities of 60-80% and 40-60% are marked in blue and light blue, respectively. Percent identities of Csps with CspD and CspA based on comparison by the NCBI BLAST are indicated to the right. (B, C) Expression and properties of *E. coli* Csps. MG1655 cells harboring the indicated plasmids were grown in LB-ampicillin medium. At an OD_{600} of 0.3, 0.2 mM IPTG was added to cultures to induce Csps. Incubation was continued for 30 min. (B) The protein sample was subjected to Western blotting using anti-FLAG monoclonal antibody and anti-GroEL polyclonal antibodies. (C) The RNA sample was subjected to Northern blotting using probes for *baxL-bax* mRNA, *gdx* mRNA or tmRNA.

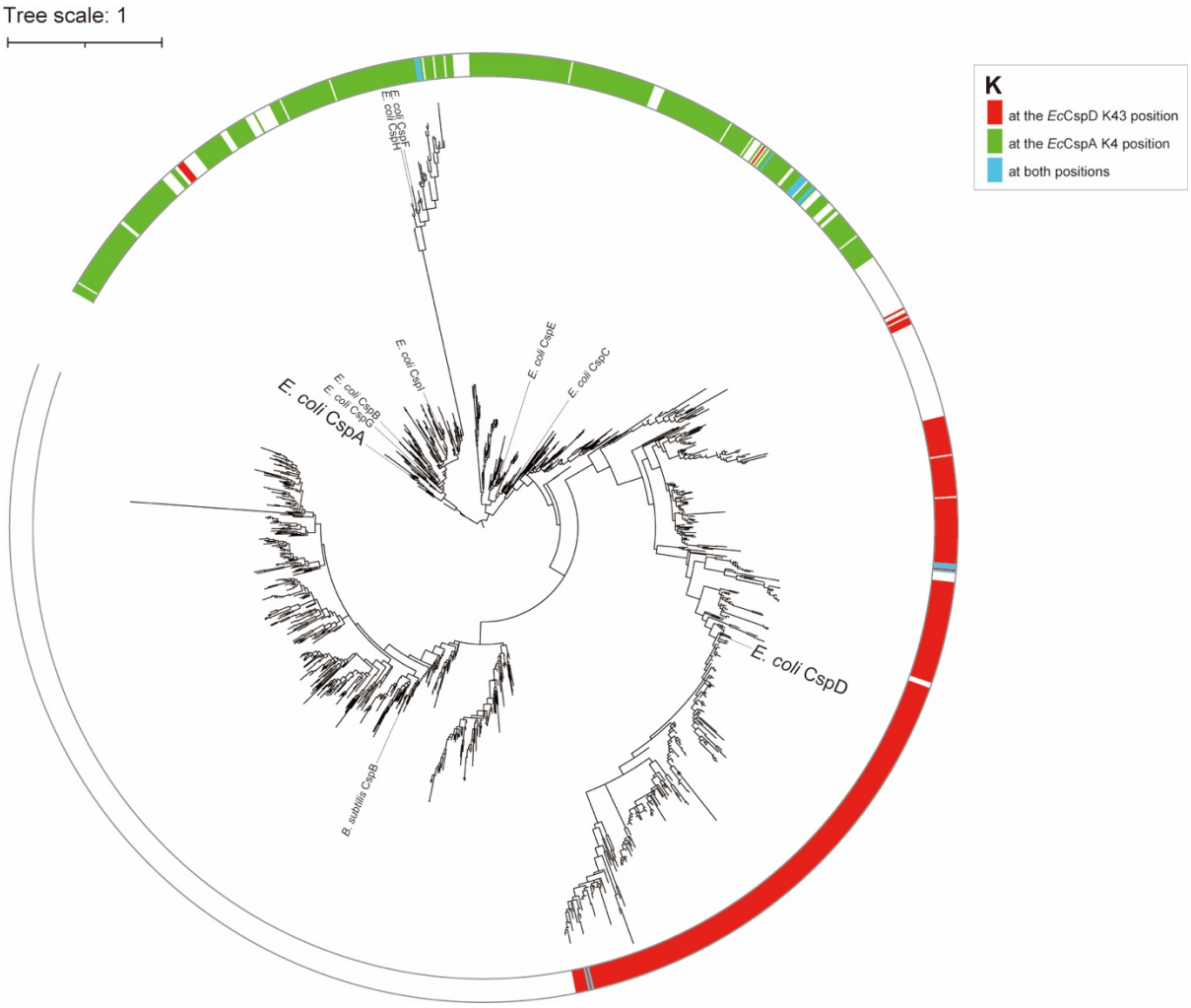

Figure S4. Proteins that contain a lysine residue at the same position as *Ec*CspD K43 and as *Ec*CspA K4 are represented by red and green, respectively, on the ring. The Csps that were used as queries for the PSI-BLAST search are shown; two *Salmonella* Csps [GCF_000210855.2: CBW17860.1, CBW16717.1] were not indicated because they were not included in the RefSeq database (October 2022 data set). Proteins that contain both lysine residues are represented by blue. Three proteins (WP_119936521.1, WP_075204774.1, WP_236108833.1) having both K-P in the loop and K4 are included in the K4 group.

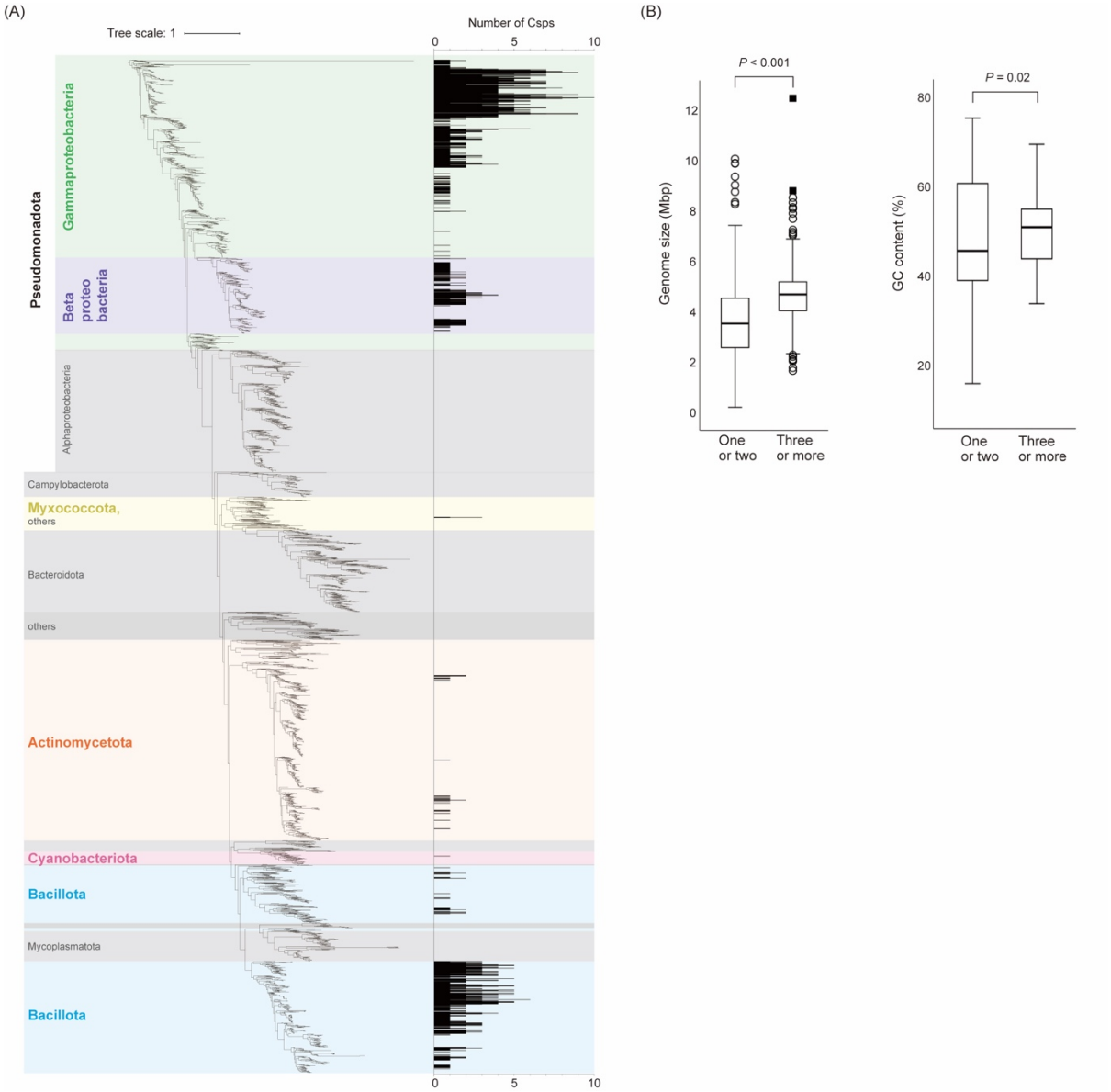

Figure S5. Distribution of Csps in bacterial genomes. (A) A phylogenetic tree was constructed from 4,031 bacterial genomes that were used for large-scale similarity search. 2,440 Csps were mapped to the owner genomes. Bacterial phyla and classes are shown to the left. The horizontal axes represent the total number of the contained Csps. The detailed method is described in Materials and Methods. (B) Box plots showing the genome traits comparison between OneOrTwoCsps (n = 683) and ThreeOrMoreCsps (n = 336). The *p*-values were computed using the Mann-Whitney U test, a nonparametric two-sample test that does not assume normality in the datasets. The detailed method is described in Materials and Methods.

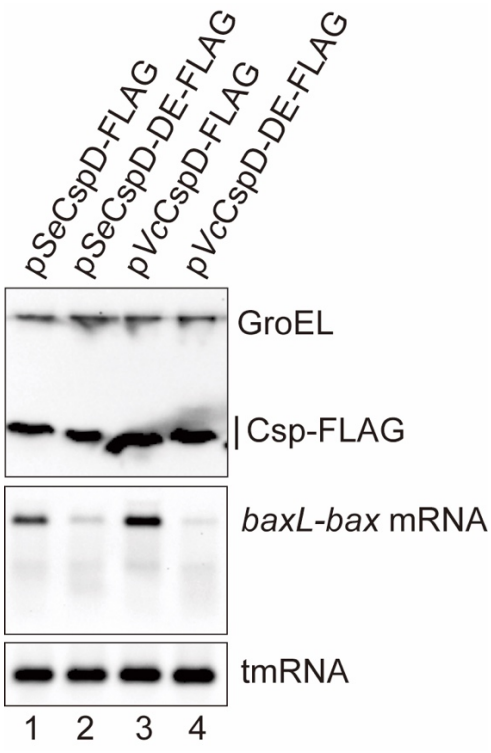

Figure S6. Effect of mutation in the loop of *Se*CspD and *Vc*CspD. MG1655 cells harboring the indicated plasmids were grown in LB-ampicillin medium. At an OD_{600} of 0.3, 0.2 mM IPTG was added to cultures to induce Csps. Incubation was continued for 30 min. The protein sample was subjected to Western blotting using anti-FLAG monoclonal antibody and anti-GroEL polyclonal antibodies. The RNA sample was subjected to Northern blotting using probes for *baxL-bax* mRNA or tmRNA.

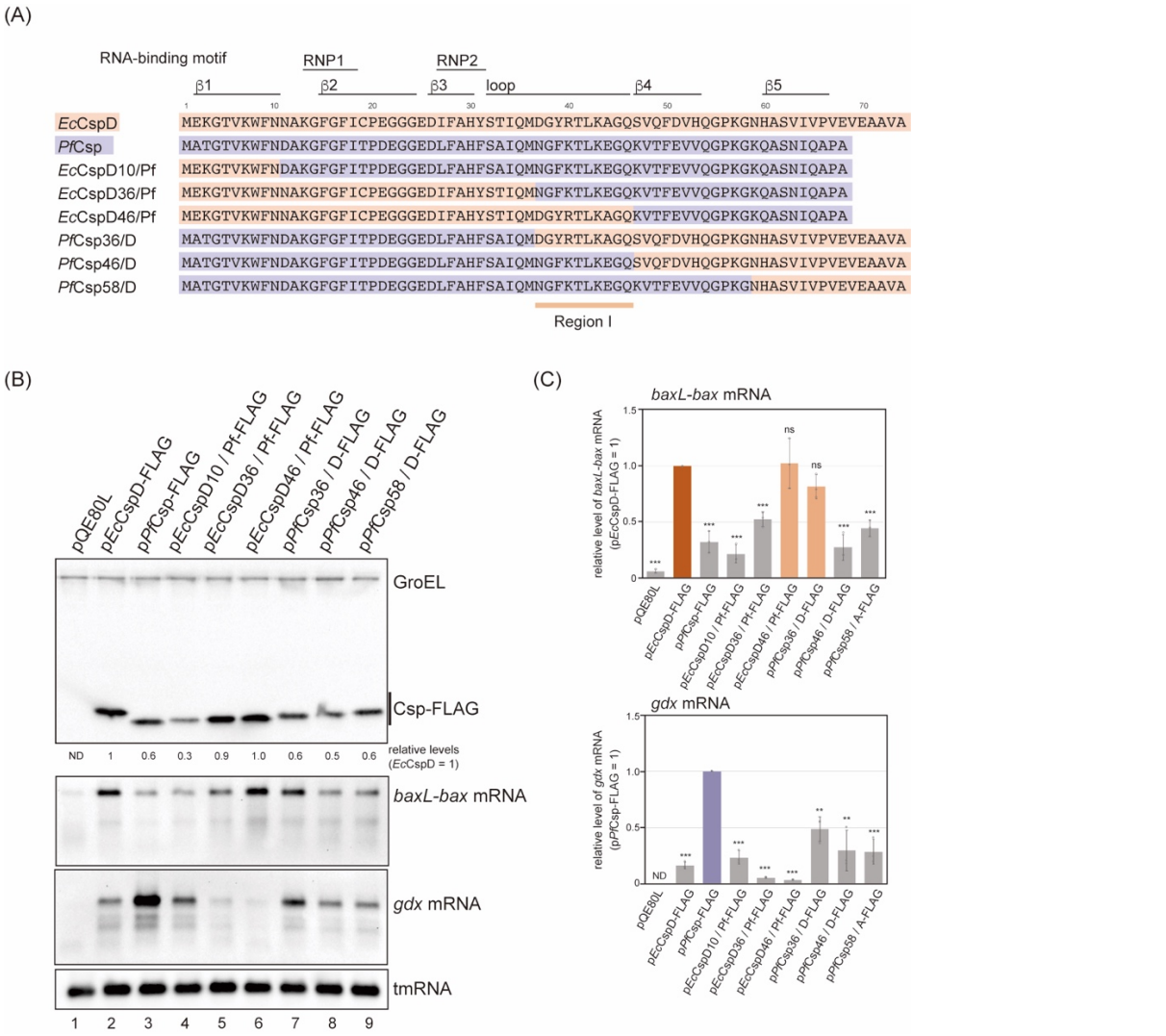

Figure S7. Chimeric variants of *Ec*CspD and *Pf*Csp. (A) Sequence alignment of *Ec*CspD, *Pf*Csp and chimeras. For chimeras, the regions originating from *Ec*CspD and *Pf*Csp are marked in orange and purple, respectively. RNA-binding motifs, RNP1 and RNP2, a loop region, and β-strands are indicated by lines on top of the sequence. (B) Expression and properties of chimeras. MG1655 cells harboring the indicated plasmids were grown in LB-ampicillin medium. At an OD_{600} of 0.3, 0.2 mM IPTG was added to cultures to induce Csps. Incubation was continued for 30 min. The protein sample was subjected to Western blotting using anti-FLAG monoclonal antibody and anti-GroEL polyclonal antibodies. The RNA sample was subjected to Northern blotting using probes for *baxL-bax* mRNA, *gdx* mRNA or tmRNA. (C) Quantitation of the Northern blotting data. Relative RNA levels were calculated, with the RNA sample of *Ec*CspD set to 1 for *baxL-bax* mRNA and *Pf*Csp set to 1 for *gdx* mRNA. The results are averages of three independent experiments, with error bars representing the standard deviations. Statistical significance was calculated using an unpaired two-tailed Student's *t* test. ns, not significant; ** $P < 0.01$; *** $P < 0.001$.

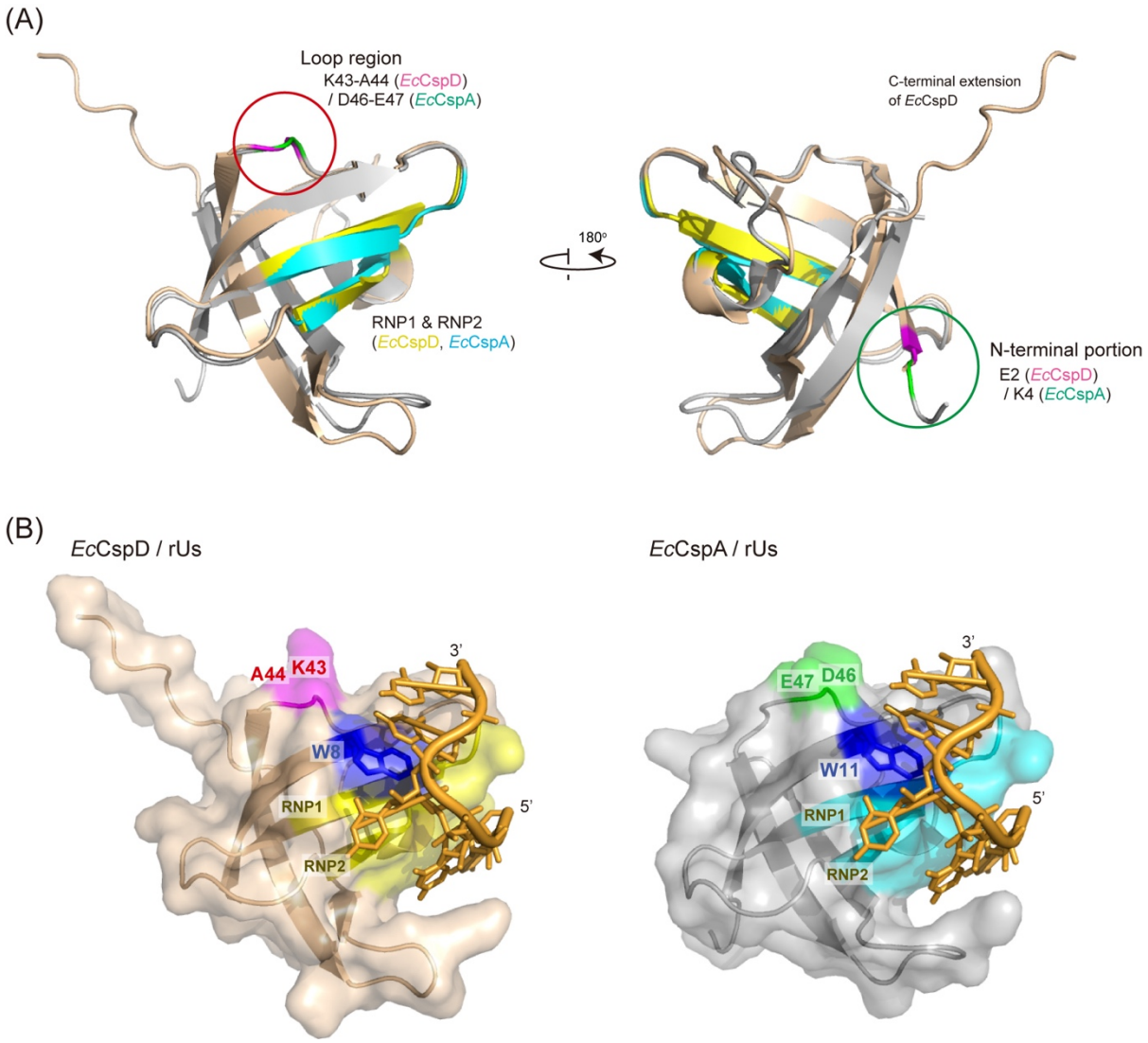

Figure S8. (A) Superimposition of the structures of *E. coli* CspD and *E. coli* CspA, represented by ribbon models. The predicted structure of the *Ec*CspD monomer (wheat) was obtained from AlphaFold Protein Structure Database (Jumper et al. 2021; Varadi et al. 2022). The structure of the *Ec*CspA monomer (gray) was obtained from Protein Data Bank (ID: 1MJC). RNA binding motifs, RNP1 and RNP2, of *Ec*CspD are colored yellow, and those of *Ec*CspA are colored cyan. The loop region and the N-terminal portion are circled in red and green, respectively. (B) Structural models of Csp/RNA complexes. The models of the *Ec*CspD (left) and *Ec*CspA (right) monomers docked with a hexauridine rU6 were constructed based on the structure of *B. subtilis* CspB/rU6 complex (PDB ID: 3PF5). The *Ec*CspD and *Ec*CspA structures are shown in translucent surface and ribbon models. The rU6 nucleotides are shown in ocher cartoon and stick models. RNA binding motifs, RNP1 and RNP2, of *Ec*CspD are colored yellow, and those of *Ec*CspA are colored cyan. The loop regions of *Ec*CspD and *Ec*CspA are highlighted in magenta and green, respectively. The tryptophan residues W8 (*Ec*CspD) and W11 (*Ec*CspA) are highlighted in blue stick models.

Table S1. List of Csps that were analyzed in this study. (A) All Csps, (B) K-A type, (C) K-E type, (D) K4 type, (E) KA/KE and K4 type and (F) Csps outside the identified types were listed. As exceptions, two proteins (GCF_000015725.1_WP_011830051.1 and GCF_900475885.1_WP_006995997.1) having K-Q and K-V in the loop, respectively, were not classified.

Table S2. (A) Strains and plasmids used in this study. (B) Oligonucleotides and gBlocks used in this study.

Supplementary References

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583-589.

Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M et al. 2017. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res* **45**: D543-d550.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**: D439-d444.