Comparative proteomics enables identification of non-annotated cold shock proteins in *E. coli*

Nadia G. D'Lima[†‡1], Alexandra Khitun[†‡1], Aaron D. Rosenbloom[†], Peijia Yuan[†‡], Brandon M. Gassaway[§∥], Karl W. Barber[§∥], Jesse Rinehart[§∥], Sarah A. Slavoff[†‡⊥*]

[†] Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

[‡] Chemical Biology Institute, Yale University, West Haven, Connecticut 06516, United States

[§] Department of Cellular and Molecular Physiology, Yale University, New Haven, Connecticut 06520

[∥] Systems Biology Institute, Yale University, West Haven, Connecticut 06511

[⊥] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06529, United States

[1] Indicates equal contribution

* Correspondence to: sarah.slavoff@yale.edu

Supporting Information

Figure S1. Analysis of control proteins for quantitative proteogenomics in E. coli .
Figure S2. Mass spectrometric evidence for the predicted protein YpaA.
Figure S3. MS/MS spectra of additional tryptic peptides detected from (A) YmcF and (B) YnfQ supporting protein identifications.
Figure S4. Validation of knock-in strains via integration check PCR (icPCR).
Figure S5. Nucleotide sequences upstream of and including (A) ymcF and (B) ynfQ.
Figure S6. Expanded stop codon mutagenesis scanning of candidate near-cognate YmcF start codons.
Figure S7. YnfQ may also initiate at an ATT start codon.
Figure S8. Structure prediction and structural homology for YmcF.
Table S1. Peptide-spectral match metrics.
Table S2. Primer sequences utilized for genomic knock-ins and integration check PCR (Supporting Fig. 4).
Worksheet S1. Key for proteomic analyses.
Worksheet S2. Replicate 1 cold shock peptide level evidence.
Worksheet S3. Replicate 1 cold shock protein level evidence.
Worksheet S4. Replicate 1 control peptide level evidence.
Worksheet S5. Replicate 1 control protein level evidence.
Worksheet S6. Replicate 2 cold shock peptide level evidence.
Worksheet S7. Replicate 2 cold shock protein level evidence.
Worksheet S8. Replicate 2 control peptide level evidence.
Worksheet S9. Replicate 2 control protein level evidence.
Worksheet S10. Replicate 3 cold shock peptide level evidence.
Worksheet S11. Replicate 3 cold shock protein level evidence.
Worksheet S12. Replicate 4 cold shock peptide level evidence.
Worksheet S13. Replicate 4 cold shock protein level evidence. (XLSX)

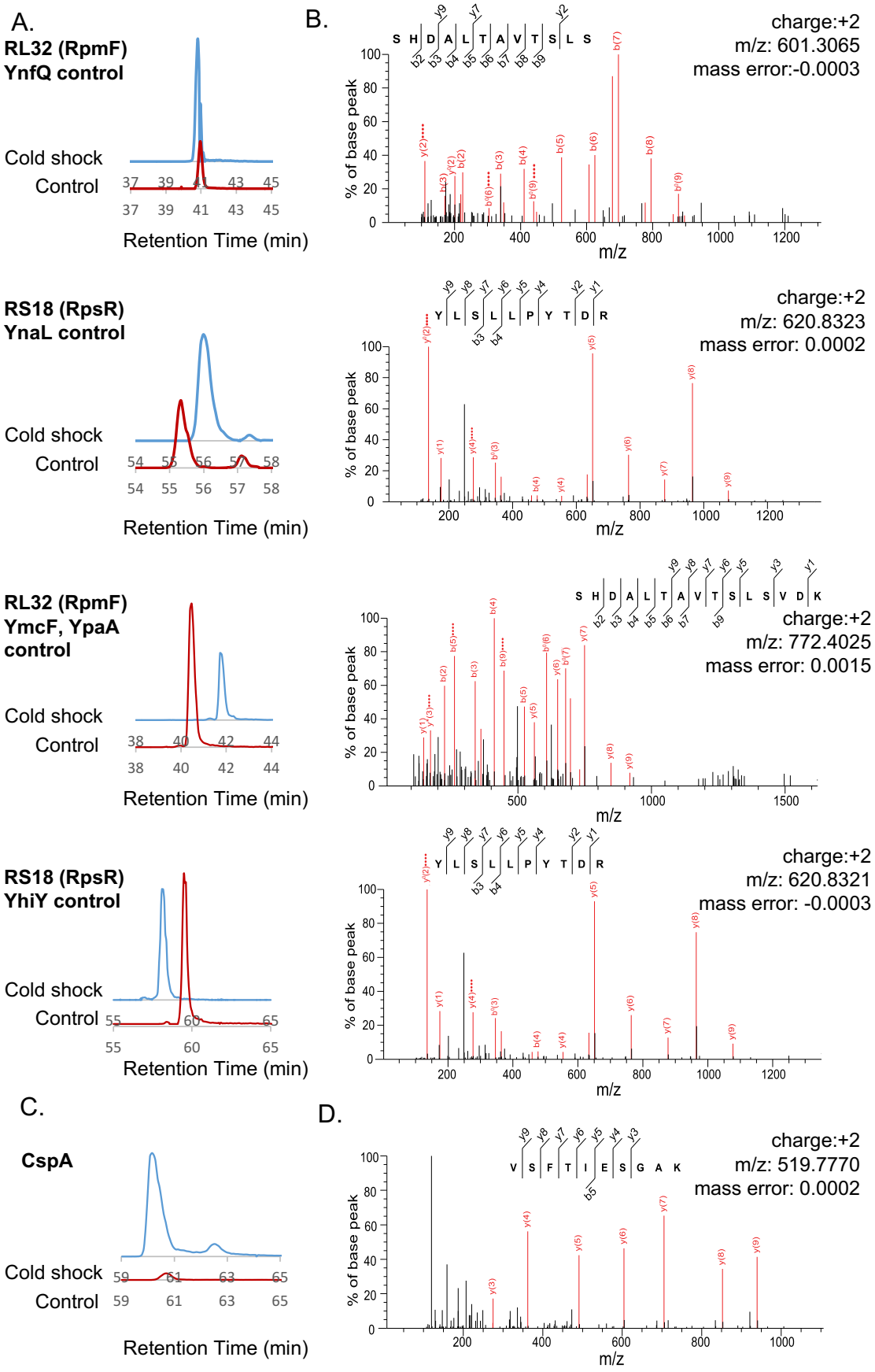

A.
RL32 (RpmF)
YnfQ control
Cold shock
Control
37 39 41 43 45
Retention Time (min)
B.
S H D A L T A V T S L S
charge:+2
m/z: 601.3065
mass error:-0.0003
% of base peak
m/z
RS18 (RpsR)
YnaL control
Y L S L L P Y T D R
charge:+2
m/z: 620.8323
mass error: 0.0002
RL32 (RpmF)
YmcF, YpaA
control
S H D A L T A V T S L S V D K
charge:+2
m/z: 772.4025
mass error: 0.0015
RS18 (RpsR)
YhiY control
charge:+2
m/z: 620.8321
mass error: -0.0003
C.
CspA
D.
V S F T I E S G A K
charge:+2
m/z: 519.7770
mass error: 0.0002

Figure S1. Analysis of control proteins for quantitative proteogenomics in *E. coli*. For each ERLIC fraction in which a non-annotated peptide of interest was identified, a control ribosomal protein that should not exhibit cold-shock regulation was analyzed with the same semi-quantitative methodology. For YnfQ YmcF, and YpaA, the controls were two peptides from RpmF; for YnaI and YhiY, controls were two peptides from RpsR. (A) Extracted ion chromatograms (EICs) from MS_1 spectra corresponding to (B) MS/MS spectra of tryptic peptides from control ribosomal proteins. The EIC intensity at the same retention time for a 1-Da window around the parent ion mass was compared for the control (red) vs. cold shock (blue) samples. Each matched EIC pair is presented on the same y-axis scale. Because the analysis is semi-quantitative, substantial intensity in both samples was taken to indicate similar expression. MS/MS spectra (right) presented correspond to the experimental EICs shown (left). Y- and b-ions are shown in red and indicated on the matched peptide scores above each spectrum. m/z, mass to charge ratio. As an additional control, we confirmed that our method accurately recapitulated expected cold-shock upregulation of the major cold shock protein CspA. (C) EICs are shown for the corresponding tryptic fragment MS/MS spectrum shown in (D

Gene Name	Figure	Peptide Sequence	Ions Score	Peptide Expectation Score	Charge	Mass error (Da)	Result File
YmcF	2B	TDHAPLDFTK	17	0.51	3+	0.0059	rep1_CS
		TDHAPLDFTK	55	0.00026	3+	-0.0011	CS3
		TSAFDVTER	48	0.0041	2+	0.0039	CS4
		TDHAPLDFTK	48	0.00098	3+	0.005	CS4
	S3A	TSAFDVTER	52	0.017	2+	-0.0016	CS4
YnfQ	2D	TSSFDVSDMNPFGAK	41	0.0041	2+	0.0008	rep1_CS
		TSSFDVSDMNPFGAK	86	2.00E-07	2+	0.0033	CS3
	S3B	SMMITFDNISQYLNASR	94	4.40E-08	2+	0.0074	CS3
YnaL		PQPMPDPPPDEEPIKL	45	0.0024	2+	0.0073	rep1_CNTRL
		PQPMPDPPPDEEPIK	30	0.045	2+	0.0042	rep1_CS
	2F	PQPMPDPPPDEEPIKL	38	0.016	2+	0.0001	rep1_CS
YhiY		PSPLALNALR	48	0.00032	2+	-0.0015	CS3
	2J	PSPLALNALR	31	0.017	2+	-0.0021	rep1_CS
YpaA	S2	DQVLAATQLSEADLAANNH	71	8.30E-06	3+	0.0056	rep1_CS

Table S1. Peptide-spectral match metrics. All sequenced MS/MS spectra that were matched to peptides derived from the non-annotated and predicted proteins YmcF, YnfQ, YnaL, YhiY and YpaA in our shotgun profiling experiments are reported in this table alongside the figure number in which they appear. MASCOT ions scores, expect values, precursor charge state, precursor mass error (Da) are provided, along with the condition in which they were detected and sequenced (each referenced worksheet is explained in the supplementary worksheets legend).

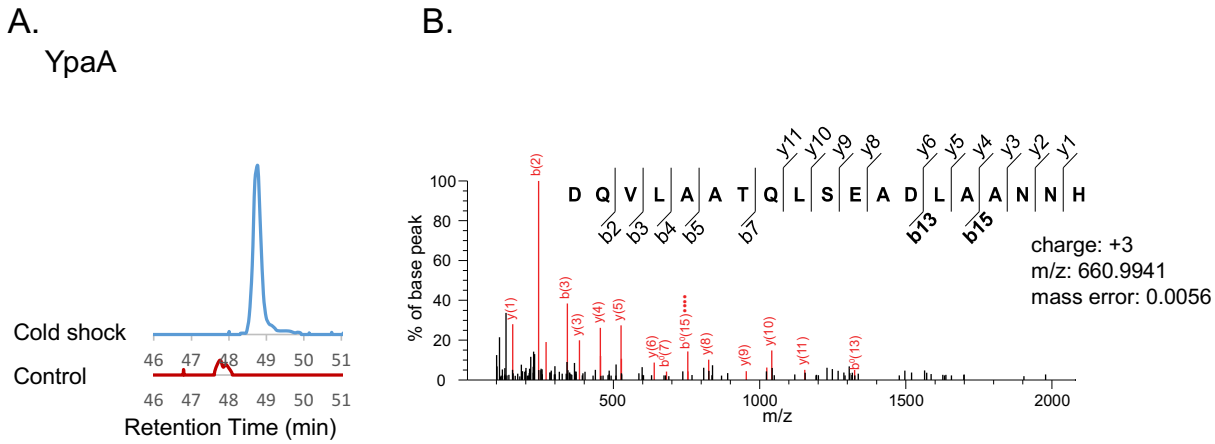


Figure S2. Mass spectrometric evidence for the predicted protein YpaA. (A) Extracted ion chromatograms (EICs) from MS_1 spectra corresponding to (B) MS/MS spectra of non-annotated tryptic peptide identified in our shotgun profiling experiments.

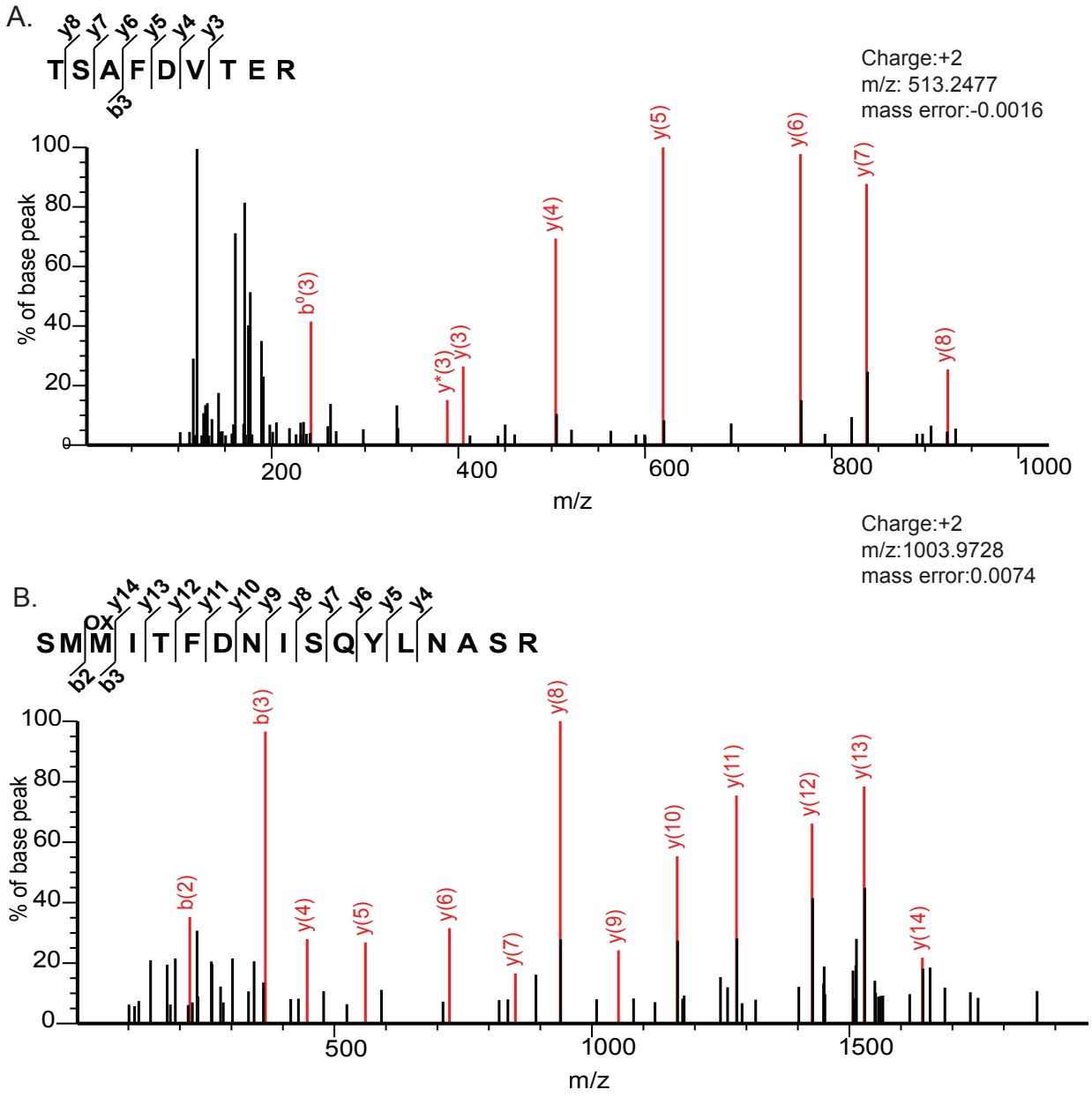


Figure S3. MS/MS spectra of additional tryptic peptides detected from (A) YmcF and (B) YnfQ supporting protein identifications. "Ox" indicates methionine oxidation.

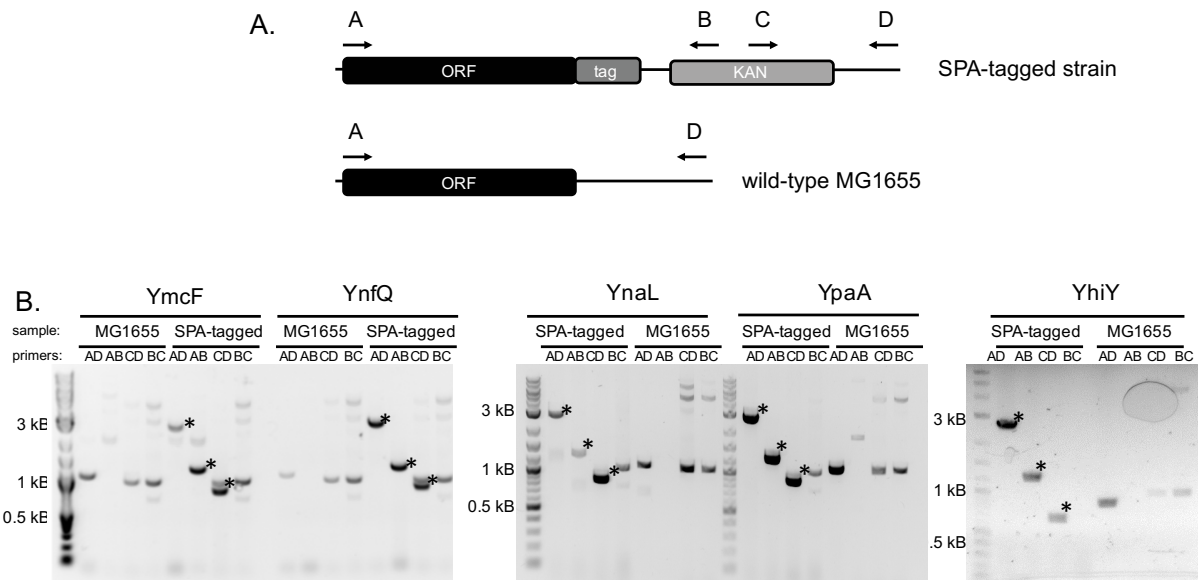


Figure S4. Validation of knock-in strains via integration check PCR (icPCR). (A) Schematic of gene loci and icPCR primer annealing sites. For each open reading frame (ORF), primers are designed to anneal to the 5' end of the coding sequence, as well as the genomic DNA downstream. Primers specific to the knock-in construct (selection marker) are also designed. (B) icPCR results showing specific products at the expected sizes (asterisks) in the SPA-tagged strains. Non-specific bands consistently appear in the untagged MG1655 strain in the CD and BC lanes. Correct integration was confirmed by Sanger sequencing of PCR products.

Integration Check Primers	
B	TTCTACGTGTTCCGCTTC
C	CTATCAGGACATAGCGTTG
YmcF A	TTCACAGTACCGCACATC
YmcF D	CAGTAAGGTGTTGGTTCC
YnfQ A	GGTTCCCAGTATAGAACATC
YnfQ D	GCCATTACAGTAAGGTGTG
YnfQ A	CTACCCTTTAAGAGTCCACC
YnfQ D	TTCTACGTGTTCCGCTTC
YpaA A	CTGATGCTGATATTCCGAAC
YpaA D	ACGATGCTCTACAGCGAT
YhiY A	CACACTTATGAGTATGCGTG
YhiY D	AGTAACCCGACAACCCAC
SPA-tag Knock-In Primers	
YmcF Forward	GCGCTGGTCGTATTTGC
YmcF Reverse	ATAATCCCTGTAAGCAAACGAC
YnfQ Forward	ATATTTCACAATACTTAAATGCCAGCCGTCTGTCGTTGGATTTAAAAAAGTCCATGGAAAAGAGAA G
YnfQ Reverse	ATCCATCGAATAGACACCAAGCAAAAAAGCTCCCGAAGGAGCCTTCATTTCATATGAATATCCTCC TTAG
YnaLForward	CGATTAAATTGTCGCATCGTGAGCGTAGATCTGCGAGGATACGCGCCTGCTCCATGGAAAAGAGA AG
YnaL Reverse	CAGGGTAGAAAAAGCGGTCACAATCTATTCTCGTGGTCATCGACGCAAAGCATATGAATATCCTC CTTAG
YpaA Forward	TGCTCGCGGCCACCCAGCTAAGCGAAGCCGATCTGGCAGCGAATAACCACTCCATGGAAAAGAG AAG
YpaA Reverse	TAGTTATTGATATCAAAGGCCCATGGGGATCGGCTGTGGGCCTGTGTTAACATATGAATATCCTCC TTAG
YhiY Forward	ACGTCGCGCTGCAAA
YhiY Reverse	TGAAGAAAAATAAGGCAGATATAAA

Table S2. Primer sequences utilized for genomic knock-ins and integration check PCR (Fig. S4).

A.

1 ACGAAAATCAGAAAGTTGAATTTTCTATTGAGCAGGGGCAACGTGGCCCCGCGGCAGCGAACGTTGTTACGCTCTAAGGTTGCCATTATTA
92 CTCAACATCTCCATTTCCGCTGTCCATGTTGTCATGGTTCACAGTACCGCACATCGGCATTCGATGTGACGGAGCGAAACCCTTTGGGCGCT
184 AAGTGTATTTTTTTGTAAATCGACGATGATCACCTTTGATAACGTCGCGCTGCAAATACGCACTGACCATGCGCCGCTGGATTTCACAAAATAA

B.

1 TGGCGCTTTGAGGTAGACAATATTACAAACCATATTCACTTTAGATGCCCGTGTTGTCATGGTTCCCAGTATAGAACATCATCTTTTGATGTTT
95 CTGACATGAATCCTTTCGGGGCAAAATGTATCTTTTGTAAATCAATGATGATTACATTTGATAATATTTCACAATACTTAAATGCCAGCCGTCTG
190 TCGTTGGATTTAAAAAAGTGA

Figure S5. Nucleotide sequences upstream of and including (A) *ymcF* and (B) *ynfQ.* Sequences and numbering are provided for the nucleotide sequence beginning immediately 3' to the stop codon of *cspG* and *cspI,* respectively. Tryptic fragments of detected peptides are encoded by the sequences highlighted in blue and in-frame stop codons are highlighted in red. All start codon candidates mutated to TAG in our experiments are highlighted in green. The correct start codons implicated in our results are $A_{88}TT$ and $A_{21}TT$, respectively.

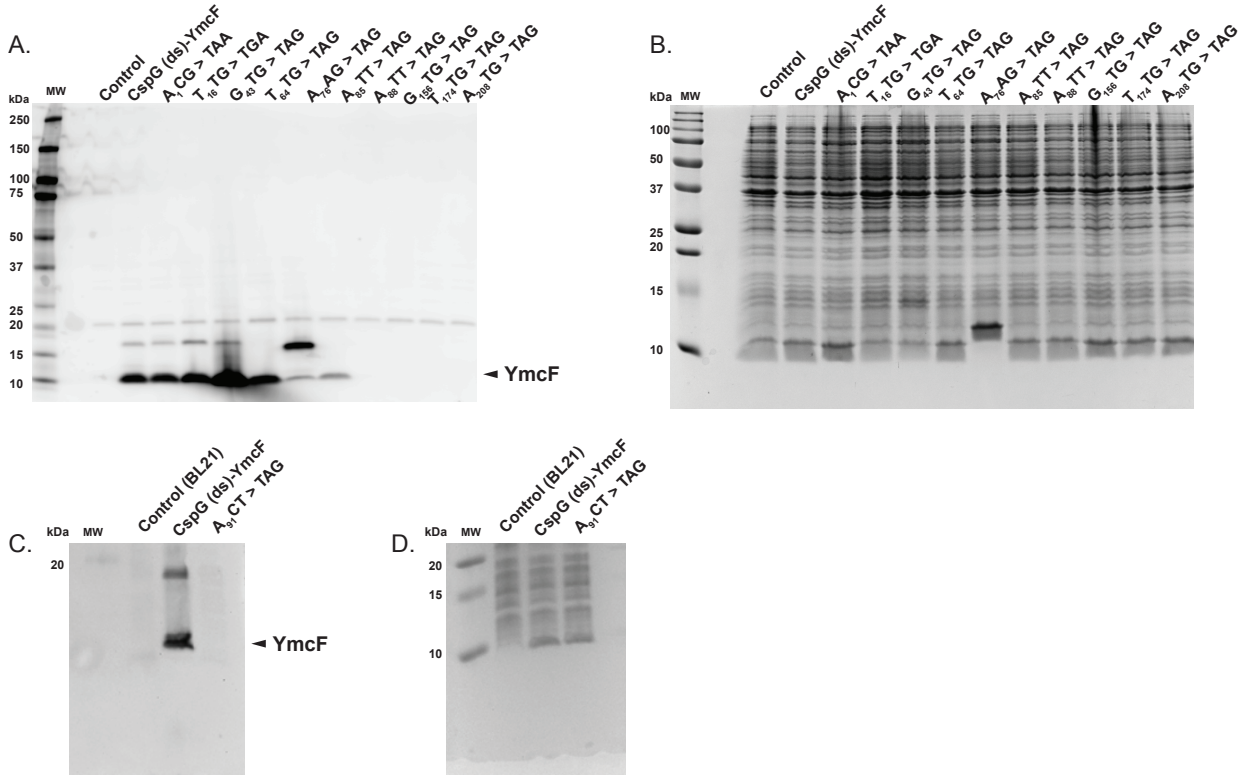


Figure S6. Expanded stop codon mutagenesis scanning of candidate near-cognate YmcF start codons. As in Fig. 5, candidate near-cognate *ymcF* start codons were mutated to stop codons in the *cspG(ds) ymcF* plasmid. Again, as a negative control, a stop codon was inserted before the His_6 tag in the *cspG(ds) ymcF* plasmid. Two downstream codons, G(156)TG and T(174)TG, were mutated to stop codons to confirm the reading frame. Finally, the codon immediately 3' to the putative ATT start codon, $A_{91}CT$, was mutated to a stop codon in order to confirm that YmcF protein translation remains "off". Nucleotide numbering starts immediately after the stop codon of *cspG* and sequences are provided in Supporting Fig. 5. To observe expression, the expression constructs were introduced into BL21 cells, and IPTG induced cell lysates were subjected to SDS-PAGE followed by blotting against an antibody to the His_6 tag. (A, C) Western blotting against the His_6 tag in this expanded stop codon scanning library, and (B, D) Coomassie staining of lysates as a loading control.

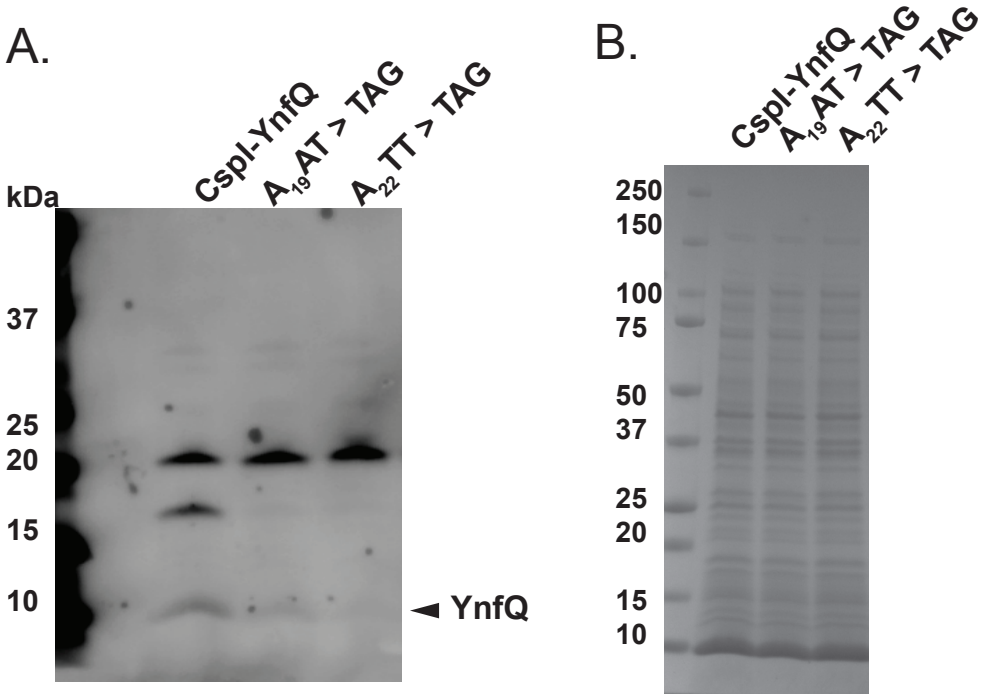
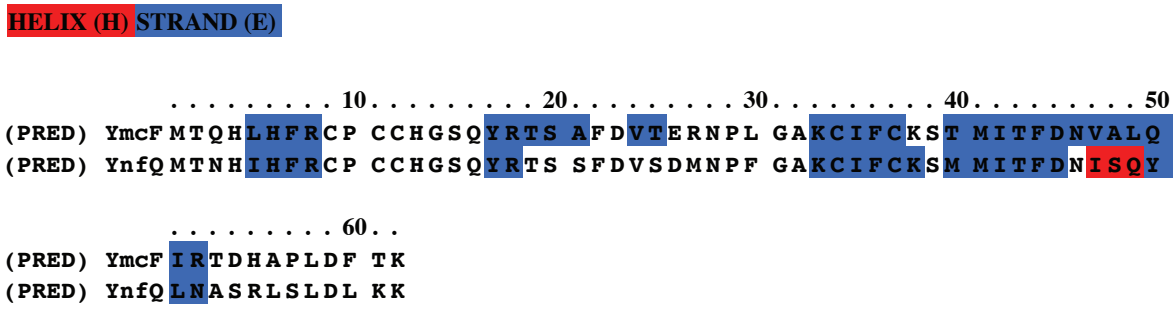
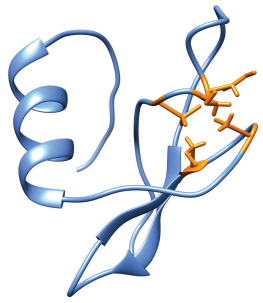


Figure S7. YnfQ may also initiate at an ATT start codon. A *cspI ynfQ* expression plasmid was cloned with a His_6 tag in-frame at the C-terminus of the *ynfQ* coding sequence. In this construct, we mutated the probable *ynfQ* start codon $A_{22}TT$ to a stop codon to a TAG stop codon, as well as preceding $A_{18}AT$ codon. Nucleotide numbering starts immediately after the stop codon of *cspG* and sequences are provided in Supporting Fig. 5. To observe expression, the expression constructs were introduced into BL21 cells, and IPTG induced cell lysates were subjected to SDS-PAGE followed by blotting against an antibody to the His_6 tag. (A) Western blotting with an antibody against the His_6 tag showed that expression of the major YnfQ protein product (carat) is decreased but still detectable in the $A_{18}AT$ > TAG mutant, but completely disappears in the $A_{22}TT$ > TAG mutant. (B) Total lysates were stained with Coomassie blue as a protein loading control.

A.



B.



C.

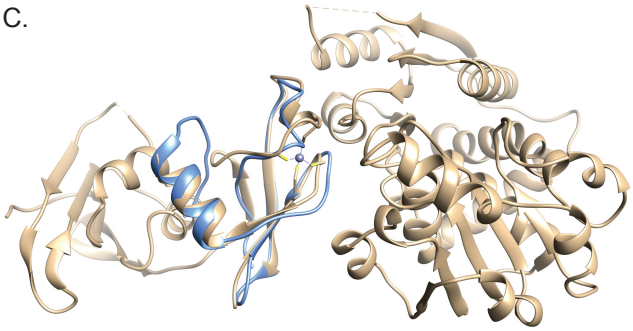


Figure S8. Structure prediction and structural homology for YmcF. (A) Secondary structure prediction using DSSP[1] and PSIPRED[2] shows these proteins may have predominately beta sheet content. (B) Predicted structure for YmcF using I-TASSER[3-5]. This model chosen was one with the highest confidence score (C-score: -1.40). UCSF Chimera[6] was used to visualize and generate the figure. The cysteine residues that we hypothesize participate in metal binding are shown in orange. (C) TM-align structural alignment program[7] showed the generated model of YmcF had best structural alignment with a region of aspartate transcarbamolyase (PDB ID 1PG5). Cysteine residues of the enzyme involved in zinc coordination are highlighted in yellow.

References

1. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-637.
2. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **1999,** *292* (2), 195-202.
3. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **2010,** *5* (4), 725-38.
4. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **2015,** *12* (1), 7-8.
5. Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008,** *9*, 40.
6. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **2004,** *25* (13), 1605-12.

7. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **2005,** *33* (7), 2302-9.