The DUF328 family member YaaA is a DNA-binding protein with a novel fold

Janani Prahlad1*, Yifeng Yuan2*, Jiusheng Lin1, Chou-Wei Chang4, Dirk lwata-Reuyl1, Yilun Liu4, Valérie de Crécy-Lagard2,5+, Mark A. Wilson1+

1Department of Biochemistry and Redox Biology Center
University of Nebraska, Lincoln, NE 68588
2Department of Microbiology and Cell Science
University of Florida, Gainesville, FL 32611
3Department of Chemistry
Portland State University, Portland, OR 97201
4Department of Cancer Genetics and Epigenetics
Beckman Research Institute, City of Hope, Duarte, CA 91010-3000
5 University of Florida Genetics Institute, Gainesville, FL 32610

*These authors contributed equally to this work
+To whom correspondence may be addressed: vcrecy@ufl.edu, mwilson13@unl.edu

Running Title: DUF328 members are novel DNA binding proteins

Keywords: DNA binding, comparative genomics, X-ray crystallography, bacterial stress response, UPF0246

Abstract
DUF328 family proteins are present in many prokaryotes, however their molecular activities are unknown. The *Escherichia coli* DUF328 protein YaaA is a member of the OxyR regulon and is protective against oxidative stress. Because uncharacterized proteins involved in prokaryotic oxidative stress response are rare, we sought to learn more about the DUF328 family. Using comparative genomics, we found a robust association between the DUF328 family and genes involved in DNA recombination and the oxidative stress response. In some proteins, DUF328 domains are fused to other domains involved in DNA binding, recombination, and repair. Co-fitness analysis indicates that DUF328 family genes associate with recombination-mediated DNA repair pathways, particularly the RecFOR pathway. Purified recombinant YaaA binds to double-stranded DNA, duplex DNA containing bubbles of unpaired nucleotides, and Holliday junction constructs *in vitro* with dissociation equilibrium constants of 200-300 nM. YaaA binds DNA with positive cooperativity, forming multiple shifted species in electrophoretic mobility shift assays. The 1.65 Å resolution X-ray crystal structure of YaaA reveals that the protein possesses a new fold that we name the cantaloupe fold. YaaA has a positively charged cleft and a helix-hairpin-helix (HhH) DNA binding motif found in other DNA repair enzymes. Our results demonstrate that YaaA is a new type of DNA-binding protein associated with the oxidative stress response and that this molecular function is likely conserved in other DUF328 family members.

Abbreviations: BSA; bovine serum albumin, DTT; dithiothreitol, EDTA; ethylenediaminetetraacetic acid, HEPES; 4-(2-hydroxyethyl)-1-piperazinediethanesulfonic acid, IPTG; isopropy β-D-1-thiogalactopyranoside, NTA; nitritolriacetic acid, PEG; polyethylene glycol, Tris; tris(hydroxymethyl)aminomethane
Introduction

Bacteria confront diverse environmental stresses and mount defensive responses to ensure survival and growth. The accumulation of reactive oxygen species (ROS) is a common stressor, causing cellular oxidative stress and eliciting a robust and multifaceted response in bacteria. In Escherichia coli, one of the primary sensors for oxidative stress is the OxyR transcription factor, which regulates the expression of approximately 20-40 genes in response to oxidation of its regulatory cysteine residues (1,2). Hydrogen peroxide is the dominant OxyR effector and therefore most of the genes controlled by OxyR are involved in the peroxide stress response (3). Although many of these downstream peroxide-responsive proteins have been extensively characterized in bacteria, some are surprisingly poorly understood.

One peroxide-responsive protein whose function is unclear is E. coli YaaA (YaaAEc; locus b0006), a member of the DUF328/UPF0246 family. YaaAEc is upregulated in response to increased H₂O₂ levels (1) and it decreases intracellular Fe²⁺ levels in E. coli by an unknown mechanism (4). Fe²⁺ is cytotoxic in the presence of hydrogen peroxide due to the production of highly reactive hydroxyl radicals via Fenton chemistry (5). Therefore, the tight control of Fe²⁺ is important during peroxide stress and helps explain why YaaAEc is under the control of the OxyR regulon.

Gene expression array analysis showed that the transcription level of yaaA in E. coli is reduced in stationary phase and in anaerobic conditions, and increased during recovery in LB broth from a stationary phase inoculum (6). This observation suggests that YaaA is important for exponential phase aerobic growth. In addition, the transcription levels of yaaA, yaaJ (a putative transport protein next to yaaA) and ten genes (fhuA, fhuF, fhu, cirA, entC, entB, exbD, fecI, fecB and fepD) that are responsible for iron acquisition were considerably lower in exponential phase growth of an rpoS mutant compared to wild-type E. coli MG1655 (7), indicating that RpoS regulates YaaA in E. coli. The connection between iron, peroxide, and YaaA was bolstered by the recent report that a yaaA mutant of Klebsiella pneumonia has impaired survival in the presence of H₂O₂ in iron-replete culture conditions, while there is no difference in survival of wild-type and yaaA mutant K. pneumonia in oxidative stress under low-iron conditions (8).

Bioinformatic analyses of the DUF328 family have indicated a link between these proteins and nucleic acid metabolism, particularly to the metabolism of the 7-deazaguanosine modified nucleosides (9). A DUF328 domain is found in DpdC proteins involved in the formation of 2′-deoxy-7-amido-7-deazaguanosine (dADG) from 2′-deoxy-7-cyano-7-deazaguanosine (dPreQ₀) in vivo (10). The 7-deazaguanosine derivatives dADG and dPreQ₀ are recently discovered DNA modifications encoded by the dpd cluster found in a diverse set of bacteria. Of the 11 genes found in the Salmonella serovar Montevideo dpd cluster (dpdA-dpdK), dpdC probably encodes an enzyme that hydrolyzes dPreQ₀ to dADG (10). E. coli K12 MG1655 has the queuosine-tRNA (Q-tRNA) pathway but no dpd genes, suggesting that YaaAEc is likely to have a distinct function.

A further cellular connection between YaaAEc and DNA repair was reported by Liu et al., who observed enhanced mutation rates in ΔyaaA E. coli and severe growth defects for ΔyaaA recA56 and ΔyaaA polA1 double mutant strains that had intact oxidative stress response enzymes when grown in an aerobic atmosphere (4). RecA is a single-stranded DNA binding protein that plays an important role in DNA recombination and repair (11). The recA56 mutant is defective in ATP-regulated formation of nucleoprotein filaments and can partially impair recombination by wild-type RecA (12). PolA (PolI) is an E. coli DNA polymerase that is involved primarily in DNA repair; the polA1 mutation results in
increased mutagenesis and diminished base mismatch repair (13,14). The ΔyaaA recA56 and ΔyaaA polA1 double mutants displayed growth defects in aerobic growth conditions that otherwise required a severely impaired oxidative stress response to be observed in the ΔyaaA single mutant *E. coli* (4). In aggregate, these results suggest a functional link between YaaAEc, oxidative stress, and DNA repair. Despite a growing body of data about the role of YaaA and other DUF328 proteins in bacterial oxidative stress defense, little is known about any DUF328 protein at the molecular level, including their structures or functions.

Here, we provide a direct molecular connection between the DUF328 family, specifically YaaAEc, with DNA repair and the oxidative stress response in bacteria. Bioinformatic analysis shows a strong association between DUF328 proteins and genes involved in DNA recombination and oxidative stress response. We find that YaaAEc directly binds to DNA *in vitro* and the 1.65 Å resolution X-ray crystal structure of YaaAEc shows that the protein possesses a new fold with a positively charged cleft and a helix-hairpin-helix (HhH) DNA binding motif found in other proteins that bind DNA in a non-sequence-specific manner. As the first member of the DUF328 family to be structurally characterized, YaaAEc establishes a new protein fold family that we propose be named the cantaloupe fold. Our results show that the DUF328 family comprises DNA-binding proteins that play important roles in DNA repair during oxidative stress, assigning a molecular function to this family of proteins.

**Results**

Comparative genomic analyses show that DUF328 genes are widespread and are fused or physically clustered with genes involved in oxidative stress and DNA repair

YaaAEc is a member of the DUF328 family. Analysis of the taxonomic distribution of members of this family shows that it is widespread in Bacteria (Fig. 1), with homologs found in ~45.0% (10,565/23,458) of the reference organisms in the Genome Taxonomy Database (GTDB) (15). Phyla that harbor most DUF328 family members are *Campylobacterota* (~87.4% ; 222/254), *Actinobacteriota* (~83.9%; 2617/3118), *Proteobacteria* (~59.1%; 4512/7630), and *Bacteroidota* (~55.3%;1573/2843), where the values in the parenthesis indicate the fraction of total reference organisms that contain DUF328 members. In contrast, DUF328 members are much less widespread in Archaea, with homologs found in only ~1.3% (31/2,392) of the reference organisms in the GTDB (Fig. S1).

Genes with related function often cluster together in bacterial genomes. To capture gene neighborhood associations of the DUF328 family, we performed a sequence similarity network (SSN) analysis combined with Genome Neighborhood Network (GNN) analysis (16,17). Using a stringent alignment score threshold (AST) of 90, YaaAEc (UniProt ID P0A8I3) is partitioned into the largest cluster (#1) that contains no sequence with an associated gene ontology terms (GO) (Fig. S2), indicating that YaaAEc is not closely associated with any sequence of annotated function.

The GNN analysis of genes encoding proteins from the major DUF328/PF03883 SSN clusters shows extensive clustering with genes involved in oxidative stress response (Fig. S3-S7). Out of the nine extracted neighborhoods, four contain genes encoding alkyl hydroperoxide reductase/thiol-specific antioxidant proteins of the AhpC family (PF00578). Three contain genes encoding for Ruberythrin, a non-heme iron-containing metallocprotein involved in oxidative stress tolerance in anaerobic bacteria (PF02915). Two contain genes encoding oxidative dealkylation repair proteins of the AlkB family (PF13640) and for members the DUF3501 family (PF12007), previously linked to oxidative stress (18). No other
functional area is as highly represented among the DUF328 GNN proteins, even though we see quite a few RNA metabolism and translation proteins (PF00270, PF04073, PF00587, PF00849, PF10150, PF14489, PF14819).

GNN analysis links members of the DUF328 family to oxidative stress, and analysis of Rosetta stone-type protein domain fusions (19) indicates an association between DUF328 proteins and DNA maintenance and repair. Several DUF328 proteins are fused to domains from the GIY-YIG endonuclease superfamily (IPR000305) (Fig. 2). Nucleases of the GIY-YIG family are involved in DNA repair and recombination, transfer of mobile genetic elements, and restriction of incoming foreign DNA (20-22). The GIY-YIG proteins are so-named because they contain a domain of typically ~100 amino acids with two short motifs, "GIY" and "YIG", in their N-terminal regions. Additionally, DUF328 domains are found fused to multiple recombinase domains, a Zn$^{2+}$-beta ribbon domain (Zn), a PadR helix-turn-helix (HTH) motif and an AlbA_2 DNA-binding domain (Fig. 2), underscoring the connection between DUF328 and DNA recombination and repair. The gene fusion and neighborhood analysis corroborate the function of DUF328 proteins in oxidative stress first reported for the YaaAEc (4), but add a link to DNA repair that was not obvious from previous studies.

**DUF328 genes show strong co-fitness with the RecFOR pathway for DNA repair.**

A measure of the joint contribution of two or more genes to organism survival is given by the co-fitness, which is the Pearson correlation coefficient of their contributions to organismal fitness under different physiological conditions. Co-fitness analysis derived from an extensive Tn-Seq analysis of 32 organisms in dozens of conditions (23) further supports an association between DUF328 genes, DNA recombination, and hydrogen peroxide detoxification. In the case of *E. coli* the top co-fitness associations are with peptidoglycan synthesis and cell division genes (*alr, ftsN* and *amiA*) with scores between 0.45 and 0.48, and then with a few DNA repair genes such as *recG* and *uvrD* (scores of 0.44 and 0.43 respectively). However, none of these scores were extremely high. The highest co-fitness scores (>0.75) were found in other species, indicating that DUF328 genes are likely to function in the same pathway with the DNA replication and repair protein coding gene *recF* in *Shewanella sp. ANA-3*, with Catalase/Peroxidase coding genes in *Acidovorax sp. GW101-3H11*, and with unknown genes in *Pseudomonas syringae pv. syringae B728a DmexB and Dechlorosoma suillum PS* (Table 1). The strong conserved co-fitness (co-fitness >0.6 and ortholog co-fitness >0.6) indicates functional relationships of DUF328 genes with genes coding DNA repair and recombination proteins, such as *recA* and *recN* (Table 1). DUF328 genes from at least seven bacteria show strong conserved co-fitness (ortholog co-fitness >0.6 in Table 1 and Table S1) with components of the RecFOR pathway that initiates recombination-mediated DNA repair by processing single-stranded DNA gaps and loading RecA onto the recombinogenic ends (24-26). The highest ortholog co-fitness scores (>0.75) include genes encoding the DNA replication and repair protein *recF* in *Caulobacter crescentus* and *Pseudomonas fluorescens FW300-N2E2* (Table S1). Several other genes in the *recFOR* pathway (*recA, recJ, recN, recO*, and *recR*) also display high co-fitness with DUF328 genes, pointing to a strong joint contribution to organism survival (Table 1 and Table S1). The recurrent co-fitness association with exonuclease I gene (*schB*) further corroborates the association between DUF328 proteins and single-stranded DNA involved in recombination. Prokaryotic exonuclease I possesses a 3’-5’ single stranded DNA exonuclease activity that is stimulated by single-stranded DNA binding protein (SSB) and plays an important role in DNA repair (27). Also recurring in the DUF328 co-fitness analysis is catalase, which defends against H$_2$O$_2$ stress by converting H$_2$O$_2$ to H$_2$O and O$_2$, reinforcing
the connection between DUF328 proteins and peroxide stress.

**Recombinant YaaAEc binds DNA with nanomolar affinity**

Consistent with strong bioinformatic evidence connecting DUF328 family proteins to DNA maintenance and repair, YaaAEc is associated with a large amount of nucleic acid during purification of the recombinant protein from E. coli. We determined that the associated nucleic acid was DNA based on its selective degradation by DNasel. Ultimately, nearly all of the DNA was removed from YaaAEc by hydroxyapatite chromatography (see Experimental Procedures), but strong anion exchange chromatography with a Q resin could not separate YaaAEc from the DNA. The persistence of DNA in association with YaaAEc suggests direct and tight binding.

YaaAEc binding to DNA was measured using electrophoretic mobility shift assay (EMSA) with a variety of defined DNA structures. YaaAEc binds to double-stranded, bulge-containing, and Holliday junction DNA with comparable dissociation constants ($K_D$) of ~200-300 nM (Fig. 3). YaaAEc also binds single-stranded DNA, although with lower apparent affinity (Fig. 3A). Unlike many other DNA binding proteins, YaaAEc shows no strong preference for specific DNA structures. Multiple shifted DNA bands are observed at higher concentrations of YaaAEc in the EMSA, suggesting that these DNA constructs are capable of binding multiple copies of YaaAEc (Fig. 3A-C). The fraction of the DNA duplex with a 12 nt bubble bound by YaaAEc was well-fitted by a cooperative binding model, giving a $K_D$=265 nM and a Hill coefficient of ~3.1 (Fig. 3D). All the EMSAs show evidence of positive cooperativity in YaaAEc binding, and several distinct bands are seen at higher concentrations of YaaAEc. Either multiple YaaAEc molecules can bind a single DNA molecule directly or one YaaAEc binds to the DNA fragment and additional YaaAEc molecules interact with the bound YaaAEc in the YaaAEc:DNA complex. Our data do not discriminate between these two possibilities, although the Hill coefficient of 3.1 suggests a cooperative YaaAEc recruitment mechanism.

**YaaAEc role is preQ$_0$-independent**

DUF328 domains are present in DpdC proteins that are predicted dPreQ$_0$ nitrile hydratases (10) and DUF328 genes physically cluster with queF genes encoding PreQ$_0$ reductase (Fig. 1B). A parsimonious prediction is that YaaAEc may be involved in the recognition and repair of preQ$_0$ that has been incorporated into DNA by mistake. Indeed, tRNA guanine transglycosylase (TGT), normally involved in the synthesis of Queosine (Q) in tRNA, can also insert the preQ$_0$ derivative preQ$_1$ in DNA in vitro if the thymine base is replaced with uracil (28). The presence of uracil in DNA does occur and is usually corrected by a specific repair machinery (29). However, the addition of exogenous preQ$_0$ did not affect growth of the ΔyaaA mutant in LB (Fig. S8). In addition, the growth defect caused by the deletion of yaaA in Hpx$^-$ background (4) seemed to be improved by the addition of preQ$_0$ and exacerbated by the deletion of queD gene involved in preQ$_0$ synthesis pathway. Even if the difficulty of working with the Hpx$^-$ strain makes this last result within the margin of error (Fig. S9), it does not fit with a role of YaaA in repairing potential misincorporations of preQ$_0$ in DNA. Additional studies in which uracil levels are increased and preQ$_0$ levels are measured in DNA would be needed to totally rule out this hypothesis however.

**YaaAEc possesses a new fold**

We determined the X-ray crystal structure of YaaAEc to 1.65 Å resolution using single wavelength anomalous diffraction (SAD) phasing of the selenomethionine (SeMet)-substituted protein. YaaAEc is a single domain protein possessing a new fold comprising 12 α-helices and 14 β-strands with a core defined by a three-stranded parallel β-sheet (Fig. 4A,B). Several of the
secondary structural elements are short (e.g. αD, αG, αI, and β2, β3, β4, β8, β9) and may be differently classified by various secondary structure detection algorithms. Overall, YaaAEc is wedge-shaped with an apical cleft, resembling a slice of cantaloupe (Fig. S10). Helices αB and αC compose a helix-hairpin-helix (HHH) DNA binding motif (see below) that is positioned opposite a β-strand motif comprising β11-14. The β-strand motif has an unusual abundance of solvent-exposed aromatic amino acids as well as several lysine residues. The cleft region that lies between the HHH and β-strand motifs is approximately 20 Å wide and is rich in basic residues, resulting in a highly positive electrostatic potential as calculated by the Adaptive Poisson-Boltzmann Solver (APBS) (30)(Fig. 4C). Residues in the cleft region display an approximate dyad symmetry (Fig. 4D) and are among the most highly conserved residues in the DUF328 family, including a K209KARG213 motif that binds a chloride ion from the crystallization buffer. The positive electrostatic potential of the cleft, the rough dyad symmetry of several conserved residues in the region, and a width that matches the diameter of B-form DNA make this cleft a plausible contact surface for DNA.

YaaAEc has multiple extended stretches of polypeptides adopting non-standard structures from residues 6-22, 67-78 and 123-135, interrupted by a short β-strand from residues 10-12. These distinctive and atypical polypeptide structures define a broad area on the exterior of YaaAEc that includes part of the cleft and also penetrate into the core of the protein (Fig. 5A). These regions are not loops as typically defined, as they are not confined to the surface of the protein and meander through the core of the protein structure. Unlike loops, these atypically structured residues make extensive contacts with neighboring residues and they are well-ordered judging from both the quality of the electron density (Fig. S11) and their low refined atomic displacement parameters (ADPs). The majority of the peptide atoms in these non-standard secondary structural regions make hydrogen bonds with solvent or nearby amino acid sidechains rather than other peptides, again differentiating them from peptide groups in standard secondary structural elements such as α-helices and β-strands. Among these unusual contacts, a chain of Tyr-mediated hydrogen bonds extends across the conserved core of the region. This Tyr-rich hydrogen bond network includes a highly conserved SGXYG motif (S112GLYG116 in YaaAEc) that is located at a sharp turn between β-strands in the core of YaaAEc (Fig. 5B) and makes extended contacts with residues 67-72 and 124-131. The high degree of conservation (Fig. S12) and clear structural importance of the SGXYG motif indicates that the surrounding unusually structured regions are likely conserved features of the DUF328 family.

YaaA contains a helix-hairpin-helix (HHH) DNA binding motif

Although YaaAEc possesses a new fold, the Phyre2 fold recognition server (31) identifies an HHH motif from residues 35-66 that is found in several DNA binding proteins (Fig. 6A). HHH motifs bind DNA in Hef-domain RuvA domain 2-like proteins (PDB 2AQ0, 1X2I, 2BGW)(32-34), excinuclease abc subunit c (PDB 1KFT)(35), DNA/RNA-binding 3-helical bundle; GerE-like (LuxR/UhpA family of transcriptional regulators) (PDB 1FSE)(36), DNA excision repair protein XPF-ERCC1 (PDB 6SXB)(37), and mitochondrial transcription elongation factor 2 (PDB 5OL9)(38), among others. In the structures that include bound DNA (6SXB and 2BGW), the HHH motif directly contacts the minor groove of DNA, suggesting that this is a possible DNA binding mode for YaaAEc as well (Fig. 6B). Minor groove binding is consistent with the lack of sequence-specific DNA binding by HHH domains (39). The putative DNA binding regions in the apical cleft of YaaAEc, including parts of the HHH motif, show the highest sequence conservation in the DUF328 family and strongly support the functional significance of these regions (Fig. S13).
Discussion

In this study we show that DUF328 proteins, including the *E. coli* representative YaaAEC, are DNA-binding proteins involved in the oxidative stress response. Prior work from the Imlay group showed that YaaAEC is important for regulating ferrous iron (Fe^{2+}) levels in oxidatively-stressed *E. coli*. Fe^{2+} is dangerous in the presence of H\textsubscript{2}O\textsubscript{2} because Fenton chemistry can generate the highly reactive hydroxyl radical (\cdotOH), which is indiscriminately destructive (40). \cdotOH is thought to result in bacterial death predominantly through DNA damage (41). Although our results do not explain how YaaAEC regulates Fe^{2+} levels, they demonstrate that a key molecular activity of YaaAEC is DNA binding, indicating that it plays an important role in DNA maintenance and repair under oxidative stress conditions. This is consistent with prior observations that *yaaA* deletion resulted in a mutator phenotype and a filamentous *E. coli* cell morphology (4), which is a common cellular manifestation of extensive DNA damage in bacteria (42). Moreover, the multiple and strong comparative genomic connections between YaaAEC homologs and DNA repair suggest that this function is likely conserved throughout the DUF328 family.

Co-fitness analysis reveals a strong connection between DUF328 proteins and the RecFOR pathway of single stranded gap DNA repair. While this indicates a connection between DUF328 proteins and DNA repair, it does not require that YaaAEC operate in the RecFOR pathway. Co-fitness indicates DUF328 and RecFOR genes make joint contributions to organismal survival, which may be because they operate in the same pathway or, alternatively, because they operate in different but functionally intersecting DNA repair pathways. The other major DNA repair pathway in bacteria is the RecBCD pathway, which primarily targets and repairs double stranded breaks in DNA (43). Although genes in this pathway did not associate with DUF328 genes in our comparative genomics study, it is possible that YaaA is involved in this or other DNA repair pathways that partially overlap with the RecFOR pathway (44). A circumstantial argument against DUF328 proteins being direct participants in the RecFOR pathway is that this DNA repair pathway is present in archaea but DUF328 proteins are rare in that kingdom. Moreover, defects in the RecFOR pathway in *E. coli* result in phenotypes that do not depend on oxidative stress, whereas a phenotype is seen in ΔyaaA *E. coli* only in oxidative stress conditions (4). However, the selective protective role of YaaAEC during oxidative stress may be driven by its increased expression under oxidative stress, rather than specificity for repair of oxidative DNA damage. Although the details of DUF328 proteins’ role in DNA protection during oxidative stress remain to be determined, co-fitness analysis indicates a probable contribution from the RecFOR pathway in diverse bacteria.

The crystal structure of YaaAEC reveals a new fold for the DUF328 family (45). Given its resemblance to a slice of cantaloupe (Fig. S10), we propose that this be called the cantaloupe fold. This fold features a distinctive abundance of structured peptide stretches that are neither α-helix nor β-strand and an apical cleft demarcated on one end by an HhH DNA binding domain and on the other by a β-strand motif. HhH domains are non-sequence-specific DNA binding modules that are commonly found in proteins that digest, synthesize, or repair DNA (39). The apical cleft is highly basic, enriched in conserved residues, and is also a plausible site for DNA binding. It is unclear if DNA could simultaneously bind both regions, and if such bound DNA would be partially unwound or otherwise structurally perturbed. The minor groove binding preference of HhH motifs suggests that at least some portion of the bound DNA should be a B-form double helix. The presence of two candidate DNA binding sites may explain the presence of multiple shifted bands at higher concentrations of protein in EMSA of YaaAEC-DNA complexes (Fig. 3). These multiple bands indicate a DNA:YaaAEC...
binding stoichiometry of greater than 1:1 at higher YaaA\textsubscript{Ec} concentrations, consistent with either with multiple YaaA\textsubscript{Ec} proteins binding cooperatively to the DNA. The Hill coefficient of 3.1 for YaaA\textsubscript{Ec} binding indicates strong positively cooperative DNA binding, suggesting either that protein-protein interactions enhance YaaA\textsubscript{Ec} affinity for DNA or that binding-induced perturbations to DNA structure recruit additional YaaA\textsubscript{Ec} molecules. Determining how YaaA\textsubscript{Ec} interacts with DNA will be an important future direction for elucidating the molecular basis of DNA recognition by this new protein fold.

DpdC proteins are members of the DUF328 family that possess nitrile hydratase activity involved in preQ\textsubscript{0} metabolism (9,10). The involvement of DpdC proteins in preQ\textsubscript{0}-containing DNA is broadly consistent with our results connecting YaaA\textsubscript{Ec} to DNA repair, although we find that YaaA\textsubscript{Ec} is not involved in queosine metabolism in \textit{E. coli}. Regardless, the nitrile hydratase activity of DpdC proteins demonstrates that the DUF328 cantaloupe fold can support enzymatic activity, leaving open the possibility that YaaA\textsubscript{Ec} may have an undiscovered enzymatic function. An important avenue for future research will be to determine the mechanism by which YaaA\textsubscript{Ec} (and other DUF328) proteins regulate Fe\textsuperscript{2+} levels to protect cells from oxidative stressors (4), including if YaaA\textsubscript{Ec} simply binds DNA or acts on it as a substrate. This report provides a molecular activity for YaaA\textsubscript{Ec} and other DUF328 proteins that will inform additional research into the protective mechanisms of this new family and fold class of DNA binding protein.

**Experimental Procedures**

**Bioinformatics Analyses**

The protein sequences were retrieved from the National Center for Biotechnology Information (NCBI) using the following accession numbers: YaaA\textsubscript{Ec}, Uniprot ID P0A8I3; DpdC, AJQ72467.1. Protein domain analysis was performed using HHpred by the MPI Bioinformatics Toolkit (46,47) against the Pfam database (48) using default parameters.

319 reviewed sequences of the DUF328/IPR005583/PF03883 family were retrieved from InterPro (https://www.ebi.ac.uk/interpro/beta/entry/pfam/PF03883/). Multiple sequence alignment was built using MAFFT (https://mafft.cbrc.jp/alignment/server/) (49). WebLogo (50) was used to generate sequence logo. Domain architecture was analyzed by the Conserved Domain Architecture Retrieval Tool (CDART) (51).

SSNs were generated with the Enzyme Function Initiative (EFI) suite of webtools (16,17). SSNs were visualized using Cytoscape (52). The parameters used for the generation of the DUF328/IPR005583/PF03883 SSNs were as follows: for the whole family SSN, the input method “FASTA” (Option C) was used, using YaaA\textsubscript{Ec} and IPR005583, UniRef90 with the minimum length filter of 100 and maximum length filter of 500. The Alignment Score Threshold (AST) was 90. Sequences that share 90% or more identity are collapsed together to a single node in order to reduce the complexity for visualization. The obtained SSN was subjected to EFI Genome Neighborhood Tool (GNT) analysis, in order to obtain Genome Neighborhood Diagrams (GND).

Fitness data was retrieved from the fitness browser (http://fit.genomics.lbl.gov/cgi-bin/myFrontPage.cgi) (23,53) using YaaA\textsubscript{Ec} as the query sequence. Genes show strong co-fitness (co-fitness >0.75) and strong conserved co-fitness (co-fitness >0.6 and ortholog co-fitness >0.6) with DUF328 genes are placed in Table 1. Genes with ortholog co-fitness >0.6 and co-fitness × ortholog co-fitness >0.3 with DUF328 genes are placed in Table S1. The AnnoTree v1.2.0 tool (54) was used to analyze the distribution of members of the PF03883 family in the reference
species set from GTDB database Bacteria Release RS89.

**Bacterial Strains**

All strains used in this study are listed in Table 2. *E. coli* strains were routinely grown in LB media (Tryptone 10 g/L, yeast extract 5 g/L, NaCl 5 g/L) at 37°C. When antibiotic selection was required, media were supplemented with 100 μg/ml ampicillin, 25 μg/ml chloramphenicol, or 50 μg/ml kanamycin. PreQ₀ was purchased from ArkPharm (Libertyville, IL, USA AK-32535).

Growth studies were adapted from those previously described (4). Anaerobic overnight cultures were diluted into anaerobic LB and grown for four to five generations to early log phase (optical density at 600 nm (OD₆₀₀) of 0.15 to 0.20). The cultures were then diluted into aerobic LB of the same composition to an OD₆₀₀ of ≈0.005. All cultures were grown at 37°C. LB medium was made one day prior to culturing and stored in the dark to avoid the photochemical generation of H₂O₂. It was transferred immediately after autoclaving to an anaerobic chamber (Bactron anaerobic chamber), where it was stored under an atmosphere of 5% CO₂, 10% H₂, 85% N₂ overnight or longer prior to use.

**Expression and Purification of YaaAEc**

The YaaAEc gene (b0006) was PCR amplified from *E. coli* XL1-blue genomic DNA using Platinum II Taq DNA polymerase (ThermoFisher) and primers that introduced 5’ NdeI and 3’ XhoI restriction sites. The YaaAEC gene was subcloned between the NdeI and XhoI sites of pET15b (Novagen) and sequence-verified by dideoxy DNA sequencing. The validated YaaAEc-pET15b construct was transformed by heat shock into chemically competent BL21(DE3) (Novagen) *E. coli* for protein expression. This construct expresses YaaAEc with a thrombin-cleavable N-terminal hexahistidine tag, although the tag was difficult to remove from the purified protein with thrombin and was retained in the final purified protein. BL21(DE3) cells containing YaaAEc-pET15b were grown in LB medium with 100 μg/ml ampicillin at 37°C with shaking at 270 rpm to an OD₆₀₀ of 0.2-0.3, at which point the culture was transferred to 20 °C and incubated with shaking at 150 rpm for an additional two hours. YaaAEc expression was induced with the addition of IPTG (Calbiochem) to a final concentration of 0.2 mM and the culture was incubated at 20 °C with shaking overnight. Chloramphenicol was added to a final concentration of 100 μg/ml two hours prior to harvest to enhance protein solubility (56). Cells were harvested by centrifugation and cell pellets were frozen on liquid N₂ and stored at -80°C.

Recombinant hexahistidine-tagged YaaAEc was purified using Ni²⁺-NTA metal affinity chromatography as previously described (57). Briefly, the cell pellet was lysed by the addition of lysozyme to a final concentration of 1 mg/ml followed by sonication. The crude lysate was clarified by centrifugation at 12,000xg and the supernatant was incubated with HIS-select Ni²⁺-NTA resin (Sigma), washed with wash buffer (25 mM HEPES pH 7.5, 300 mM NaCl, 25 mM imidazole), and eluted with elution buffer (25 mM HEPES pH 7.5, 300 mM NaCl, 250 mM imidazole). Fractions containing the YaaAEc protein were determined by Coomassie-stained SDS-PAGE and the purest fractions were pooled. Despite bearing a thrombin cleavage site downstream of the N-terminal hexahistidine tag, incubation of the purified protein with thrombin did not efficiently cleave the hexahistidine tag. Therefore, the final protein retains the tag and the thrombin cleavage site, adding the N-terminal sequence MGSSHHHHHHSSGLVPRGSH-before the first methionine of YaaAEc.

YaaAEc co-purified with large amounts of nucleic acid that was identified as DNA based on its sensitivity to DNaseI on GelRed (Biotium) stained agarose gel electrophoresis. The contaminant DNA was present at approximately 0.1 mg/mg of
protein based on 260 nm/280 nm absorbance ratio and could not be effectively removed by passage over a Hi-Q anion exchange column (Bio-Rad Laboratories, Hercules, CA, USA). The large majority of the DNA could be removed using hydroxyapatite chromatography (ceramic hydroxyapatite Type II resin, Bio-Rad) with gradient elution from 20 mM to 500 mM KPO₄ buffer, pH 7.2 over 10 column volumes. A small amount of residual nucleic acid remained in YaaAEc even after hydroxyapatite chromatography as determined by GelRed-stained agarose gel electrophoresis of the purified protein. After hydroxyapatite chromatography, the final YaaAEc protein was dialyzed into storage buffer (25 mM Tris-HCl pH=8.8, 100 mM KCl) and was concentrated to 28 mg/mL (ε₂₈₀ = 25900 M⁻¹ cm⁻¹) with a 10 kDa molecular weight cutoff regenerated cellulose membrane spin concentrator at 4°C (Millipore). Before storage, 10 mM EDTA (final concentration) was added to enhance protein stability and the protein was flash-frozen in 50-200 uL aliquots in liquid N₂ and stored at -80 °C until needed.

YaaAEc-DNA binding assay

Synthetic Holliday junction (HJ X0), single-stranded, and double-stranded oligonucleotide substrates were generated as previously described (58,59). DNA binding assays were performed by incubating YaaAEc protein with 4 pg of ³²P-labeled oligonucleotide substrates in DNA binding buffer (30 mM HEPES pH 7.5, 1 mM DTT, 100 μg/ml BSA) on ice for 15 min. The protein-DNA complexes were analyzed on 5% native polyacrylamide gels. To compare YaaAEc binding to either circular, blunt-end linear DNA, or linear DNA with single-stranded overhangs, 0.5μg of pUC19 plasmid (ThermoFisher, cat# SD00661) with or without Smal or Sall digestion was incubated with the indicated amount of YaaAEc in DNA binding buffer on ice for 30 min prior to separation of protein-DNA complexes on 0.6% agarose gels. ImageJ (60) was used to quantify the band intensities to determine the relative amount of DNA that was free or in complex with YaaAEc. These fractional binding values were plotted as a function of total YaaAEc concentration and fit to a single-site binding model with positive cooperativity in Prism (GraphPad) to determine K_D and Hill coefficient. DNA binding was assayed for multiple independent preparations of recombinant YaaAEc to ensure consistency.

X-ray crystallographic structure determination of YaaAEc

Crystallization conditions for purified hexahistidine-tagged YaaAEc at 28 mg/ml in storage buffer were screened using commercial sparse matrix screens in sitting drop 96-well plates. Protein and reservoir solutions were dispensed in 0.5 μL drops using a Gryphon liquid handling robot (Art Robbins Instruments). Initial needle-shaped crystals were optimized by manual sitting drop vapor equilibration and improved crystals grew in 100 mM NaCl, 100 mM sodium citrate pH 4.6, 100 mM Na₂HPO₄, 75 mM NaH₂PO₄ and 15% PEG 8000. Further optimization with the Hampton additive screen identified 3% benzamidine-HCl as improving crystal morphology and size. An ordered benzamidine molecule makes hydrogen bonds to two Pro73 residues related by crystallographic symmetry in the final structure. In addition, the phenyl ring of benzamidine participates in cation-π interactions with two symmetry-related Arg77 residues, explaining why this additive aided the growth of diffraction-quality crystals.

Because YaaAEc has no homologs of known structure, experimental phasing using single-wavelength anomalous diffraction (SAD) was used to calculate initial electron density maps. The pET15b-YaaAEc expression construct was transformed into the methionine auxotroph E. coli strain B834(DE3) (Novagen) by heat shock of chemically competent cells. YaaAEc was expressed in M9 minimal medium supplemented with 42 mg/L of each L-amino acid except methionine and cysteine, 125
mg/L each of adenine, guanosine, thymine, uracil, 4 mg/L thiamine, 4 mg/L D-biotin, and 30 mg/L L-selenomethionine (AcrosOrganics). SeMet-YaaAEc was purified as described above and crystallized in the same condition as the native protein. Diffraction-quality crystals of SeMet-YaaAEc were cryoprotected by serial transfer and brief soaking in reservoir solution supplemented with ethylene glycol in 5% increments to a final concentration of 15%. Crystals were mounted in nylon loops and cryocooled by rapid immersion in liquid nitrogen.

YaaAEc crystallized in space group P2₁₂₁₂₁ with two protein chains in the asymmetric unit (ASU). Diffraction data extending to 1.65 Å resolution were collected from a SeMet-YaaAEc crystal measuring 350x100x100 μm at beamline 7-1 of the Stanford Synchrotron Radiation Laboratory (SSRL) using the oscillation method. The incident X-rays were tuned to the K-edge of selenium (0.9788 Å) in order to maximize anomalous signal from the six SeMet residues in YaaAEc. Inverse beam geometry was not used. Data were processed using HKL2000 (61) and significant anomalous signal extended to 2.45 Å resolution using a CC<sub>anom</sub> cutoff of 0.15 (62) as determined by Aimless (63) in the CCP4 suite (64). See Table S2 for data statistics.

SAD phasing using unmerged input reflections and local scaling was performed in PHENIX (65). The figure of merit (FOM) for the initial experimental SAD phases was 0.39, and these phases were improved by density modification prior to autobuilding in PHENIX (66,67). The initial autobuilt model was manually improved in Coot (68), including addition of ordered waters. Five residues of the uncleaved N-terminal tag were ordered in the crystal and thus included in the model. Refinement was performed in PHENIX using a maximum likelihood target function based on anomalous amplitudes (i.e. Bijvoet mates were kept separate), optimization of the X-ray/stereochemical and atomic displacement parameter (ADP) weights, and translation-libration-screw (TLS) treatment of the ADPs (66). Benzamidine was modeled into unambiguous electron density and mediates a key crystal-packing contact, explaining why it was an effective crystallization additive. Final model validations were performed using Coot (68) and MolProbity (69). The final model has excellent stereochemical and clashscore statistics, with an overall MolProbity score of 0.95 (100<sup>th</sup> percentile). See Table S2 for model statistics.

**YaaAEc structural analysis and display**

The surface electrostatic potential of YaaAEc was calculated using the Adaptive Poisson-Boltzmann Solver (APBS) (30,70) using PDB2PQR (71) for atomic partial charge and radius assignments. Default values of solvent (78) and protein (2) dielectric constant, probe radius (1.4 Å), and temperature (298.15 K) were used. Structural figures were made with UCSF Chimera (72) and POVSscript+ (73). Sequence conservation was mapped onto the structure of YaaAEc using the ConSurf server (74).

**Data Availability**

Data for Figures 1, 2, and Supporting Information Figure S1 are available from Dr. Valérie de Crécy-Lagard, University of Florida, email: vcrecy@ufl.edu. Data for Figure 3, Table 1, Supporting Information Figures S3-S7, and S12, and Supporting Information Table S1 are contained in their entirety in the manuscript. Data for PF03883 sequence similarity network and data for growth curves are available at University of Florida Digital Collections: Data for Supporting Information Figure S2 ([https://ufdcimages.uflib.ufl.edu/IR/00/01/12/21/00001/PF03883_SSN.xlsx](https://ufdcimages.uflib.ufl.edu/IR/00/01/12/21/00001/PF03883_SSN.xlsx)), Data for Supporting Information Figure S8 ([https://ufdcimages.uflib.ufl.edu/IR/00/01/12/21/00001/Data%20for%20FigureS8.xlsx](https://ufdcimages.uflib.ufl.edu/IR/00/01/12/21/00001/Data%20for%20FigureS8.xlsx)), and Data for Supporting Information Figure S9 ([https://ufdcimages.uflib.ufl.edu/IR/00/01/12/21/00001/Data%20for%20FigureS9.xlsx](https://ufdcimages.uflib.ufl.edu/IR/00/01/12/21/00001/Data%20for%20FigureS9.xlsx)).
X-ray crystallographic structure factor data and coordinates (Figures 4-6, Supporting Information Figures S10, S11, and S13) are available for download from the Protein Data Bank (PDB) with accession code 5CAJ. Raw diffraction data are available by request from Dr. Mark Wilson, University of Nebraska, email: mwilson13@unl.edu.

Acknowledgments

We thank Professor James Imlay and Dr. Yuanyuan Liu (University of Illinois) for the gift of the YaaA deletion and Hpx-catalase and peroxidase-deficient E. coli strains and Carol Cook for photography.

Funding

Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health, National Institute of General Medical Sciences (P41GM103393). Portions of this research were also funded by the National Institutes of Health R01GM70641 to V. dC-L. and R01GM092999 to M.A.W. Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41GM103311. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.
References

1. Seo, S. W., Kim, D., Szubin, R., and Palsson, B. O. (2015) Genome-wide Reconstruction of OxyR and SoxRS Transcriptional Regulatory Networks under Oxidative Stress in Escherichia coli K-12 MG1655. *Cell Rep* **12**, 1289-1299

2. Zheng, M., Wang, X., Templeton, L. J., Smulski, D. R., LaRossa, R. A., and Storz, G. (2001) DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide. *J Bacteriol* **183**, 4562-4570

3. Dubbs, J. M., and Mongkolsuk, S. (2012) Peroxide-sensing transcriptional regulators in bacteria. *J Bacteriol* **194**, 5495-5503

4. Liu, Y., Bauer, S. C., and Imlay, J. A. (2011) The YaaA protein of the Escherichia coli OxyR regulon lessens hydrogen peroxide toxicity by diminishing the amount of intracellular unincorporated iron. *J Bacteriol* **193**, 2186-2196

5. Fenton, H. J. H. (1894) Oxidation of tartaric acid in presence of iron. *Journal of the Chemical Society, Transactions* **65**, 899-911

6. Sangurdekar, D. P., Srienc, F., and Khodursky, A. B. (2006) A classification based framework for quantitative description of large-scale microarray data. *Genome Biol* **7**, R32

7. Dong, T., Kirchhof, M. G., and Schellhorn, H. E. (2008) RpoS regulation of gene expression during exponential growth of Escherichia coli K12. *Mol Genet Genomics* **279**, 267-277

8. Paczosa, M. K., Silver, R. J., McCabe, A. L., Tai, A. K., McLeish, C. H., Lazinski, D. W., and Meccas, J. (2020) Transposon Mutagenesis Screen of Klebsiella pneumoniae Identifies Multiple Genes Important for Resisting Antimicrobial Activities of Neutrophils in Mice. *Infect Immun* **88**

9. Thiaville, J. J., Kellner, S. M., Yuan, Y., Hutinet, G., Thiaville, P. C., Jumpathong, W., Mohapatra, S., Brochier-Armanet, C., Letarow, A. V., Hillebrand, R., Malik, C. K., Rizzo, C. J., Dedon, P. C., and de Crecy-Lagard, V. (2016) Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc Natl Acad Sci U S A* **113**, E1452-1459

10. Yuan, Y., Hutinet, G., Valera, J. G., Hu, J., Hillebrand, R., Gustafson, A., Iwata-Reuyl, D., Dedon, P. C., and de Crecy-Lagard, V. (2018) Identification of the minimal bacterial 2'-deoxy-7-amido-7-deazaguanine synthesis machinery. *Mol Microbiol* **110**, 469-483

11. Cox, M. M. (1999) Recombinational DNA repair in bacteria and the RecA protein. *Prog Nucleic Acid Res Mol Biol* **63**, 311-366

12. Lauder, S. D., and Kowalczykowski, S. C. (1993) Negative co-dominant inhibition of recA protein function. Biochemical properties of the recA1, recA13 and recA56 proteins and the effect of recA56 protein on the activities of the wild-type recA protein function in vitro. *J Mol Biol* **234**, 72-86

13. Gross, J., and Gross, M. (1969) Genetic analysis of an E. coli strain with a mutation affecting DNA polymerase. *Nature* **224**, 1166-1168

14. Fix, D. F., Burns, P. A., and Glickman, B. W. (1987) DNA sequence analysis of spontaneous mutation in a PolA1 strain of Escherichia coli indicates sequence-specific effects. *Mol Gen Genet* **207**, 267-272

15. Oren, A., da Costa, M. S., Garrity, G. M., Rainey, F. A., Rossello-Mora, R., Schink, B., Sutcliffe, I., Trujillo, M. E., and Whitman, W. B. (2015) Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes. *Int J Syst Evol Microbiol* **65**, 4284-4287

16. Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhn, B., Slater, D. R., and Whalen, K. L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta* **1854**, 1019-1037
17. Zallot, R., Oberg, N. O., and Gerlt, J. A. (2018) 'Democratized' genomic enzymology web tools for functional assignment. *Curr Opin Chem Biol* 47, 77-85

18. Cardenas, J. P., Quatrini, R., and Holmes, D. S. (2016) Aerobic Lineage of the Oxidative Stress Response Protein Rubrerythrin Emerged in an Ancient Microaerobic, (Hyper)Thermophilic Environment. *Front Microbiol* 7, 1822

19. Date, S. V. (2008) The Rosetta stone method. *Methods Mol Biol* 453, 169-180

20. Kowalski, J. C., Belfort, M., Stapleton, M. A., Holpert, M., Dansereau, J. T., Pietrokovski, S., Baxter, S. M., and Derbyshire, V. (1999) Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings. *Nucleic Acids Res* 27, 2115-2125

21. Truglio, J. J., Rhau, B., Croteau, D. L., Wang, L., Skorvaga, M., Karakas, E., DellaVecchia, M. J., Wang, H., Van Houten, B., and Kisker, C. (2005) Structural insights into the first incision reaction during nucleotide excision repair. *EMBO J* 24, 885-894

22. Van Roey, P., Meehan, L., Kowalski, J. C., Belfort, M., and Derbyshire, V. (2002) Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nat Struct Biol* 9, 806-811

23. Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., Blow, M. J., Arkin, A. P., and Deutschbauer, A. M. (2018) Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* 557, 503-509

24. Horii, Z., and Clark, A. J. (1973) Genetic analysis of the recF pathway to genetic recombination in *Escherichia coli* K12: isolation and characterization of mutants. *J Mol Biol* 80, 327-344

25. Kuzminov, A. (1999) Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. *Microbiol Mol Biol Rev* 63, 751-813, table of contents

26. Lloyd, R. G., and Thomas, A. (1983) On the nature of the RecBC and RecF pathways of conjugal recombination in *Escherichia coli*. *Mol Gen Genet* 190, 156-161

27. Lovett, S. T. (2011) The DNA Exonucleases of *Escherichia coli*. *EcoSal Plus* 4

28. Nonekowksi, S. T., Kung, F. L., and Garcia, G. A. (2002) The *Escherichia coli* tRNA-guanine transglycosylase can recognize and modify DNA. *J Biol Chem* 277, 7178-7182

29. el-Hajj, H. H., Zhang, H., and Weiss, B. (1988) Lethality of a dut (deoxyuridine triphosphatase) mutation in *Escherichia coli*. *J Bacteriol* 170, 1069-1075

30. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., Wei, G. W., Holst, M. J., McCammon, J. A., and Baker, N. A. (2018) Improvements to the APBS biomolecular solvation software suite. *Protein Sci* 27, 112-128

31. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10, 845-858

32. Das, D., Tripsianes, K., Jaspers, N. G., Hoeijmakers, J. H., Kaptein, R., Boelens, R., and Folkers, G. E. (2008) The HhH domain of the human DNA repair protein XPF forms stable homodimers. *Proteins* 70, 1551-1563

33. Nishino, T., Komori, K., Ishino, Y., and Morikawa, K. (2005) Structural and functional analyses of an archaeal XPF/Rad1/Mus81 nuclease: asymmetric DNA binding and cleavage mechanisms. *Structure* 13, 1183-1192

34. Newman, M., Murray-Rust, J., Lally, J., Rudolf, J., Fadden, A., Knowles, P. P., White, M. F., and McDonald, N. Q. (2005) Structure of an XPF endonuclease with and without DNA suggests a model for substrate recognition. *EMBO J* 24, 895-905
35. Singh, S., Folkers, G. E., Boelens, R., Wechselberger, R., Nizatayev, A., and Kaptein, R. (2002) Solution structure and DNA-binding properties of the C-terminal domain of UvrC from E.coli. *EMBO J* **21**, 6257-6266

36. Ducros, V. M., Lewis, R. J., Verma, C. S., Dodson, E. J., Leonard, G., Turkenburg, J. P., Murshudov, G. N., Wilkinson, A. J., and Brannigan, J. A. (2001) Crystal structure of GerE, the ultimate transcriptional regulator of spore formation in Bacillus subtilis. *J Mol Biol* **306**, 759-771

37. Jones, M., Beuron, F., Borg, A., Nans, A., Earl, C. P., Briggs, D. C., Snijders, A. P., Bowles, M., Morris, E. P., Linch, M., and McDonald, N. Q. (2020) Cryo-EM structures of the XPF-ERCC1 endonuclease reveal how DNA-junction engagement disrupts an auto-inhibited conformation. *Nat Commun* **11**, 1120

38. Hillen, H. S., Parshin, A. V., Agaronyan, K., Morozov, Y. I., Graber, J. J., Chernev, A., Schwinghammer, K., Urlaub, H., Anikin, M., Cramer, P., and Temiakov, D. (2017) Mechanism of Transcription Anti-termination in Human Mitochondria. *Cell* **171**, 1082-1093 e1013

39. Doherty, A. J., Serpell, L. C., and Ponting, C. P. (1996) The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Res* **24**, 2488-2497

40. Winterbourn, C. C. (1995) Toxicity of iron and hydrogen peroxide: the Fenton reaction. *Toxicol Lett* **82-83**, 969-974

41. Imlay, J. A. (2003) Pathways of oxidative damage. *Annu Rev Microbiol* **57**, 395-418

42. Hill, T. M., Sharma, B., Valjavec-Gratian, M., and Smith, J. (1997) sfi-independent filamentation in Escherichia coli Is lexA dependent and requires DNA damage for induction. *J Bacteriol* **179**, 1931-1939

43. Dillingham, M. S., and Kowalczykowski, S. C. (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev* **72**, 642-671, Table of Contents

44. Pages, V. (2016) Single-strand gap repair involves both RecF and RecBCD pathways. *Curr Genet* **62**, 519-521

45. Kryshtafovych, A., Moul, J., Basle, A., Burgin, A., Craig, T. K., Edwards, R. A., Fass, D., Hartmann, M. D., Korycinski, M., Lewis, R. J., Lorimer, D., Lupas, A. N., Newman, J., Peat, T. S., Piepenbrink, K. H., Prahlad, J., van Raaij, M. J., Rohwer, F., Segall, A. M., Segurian, V., Sundberg, E. J., Singh, A. K., Wilson, M. A., and Schwede, T. (2016) Some of the most interesting CASP11 targets through the eyes of their authors. *Proteins* **84 Suppl 1**, 34-50

46. Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kubler, J., Lozajic, M., Gabler, F., Soding, J., Lupas, A. N., and Alva, V. (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **430**, 2237-2243

47. Soding, J., Biegert, A., and Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244-248

48. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432

49. Katoh, K., Rozewicki, J., and Yamada, K. D. (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* **20**, 1160-1166

50. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190

51. Geer, L. Y., Domrachev, M., Lipman, D. J., and Bryant, S. H. (2002) CDART: protein homology by domain architecture. *Genome Res* **12**, 1619-1623
52. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-2504

53. Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., and Deutschbauer, A. (2015) Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *mBio* 6, e00306-00315

54. Mendler, K., Chen, H., Parks, D. H., Lobb, B., Hug, L. A., and Doxey, A. C. (2019) AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* 47, 4442-4448

55. Parks, D. H. C.; Chaumeil, P.A.; Rinke, C.; Mussig, A.J.; Hugenholtz, P. (2019) Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. *BioRxiv*

56. Carrio, M. M., and Villaverde, A. (2001) Protein aggregation as bacterial inclusion bodies is reversible. *FEBS Lett* 489, 29-33

57. Lakshminarasimhan, M., Maldonado, M. T., Zhou, W., Fink, A. L., and Wilson, M. A. (2008) Structural impact of three Parkinsonism-associated missense mutations on human DJ-1. *Biochemistry* 47, 1381-1392

58. Liu, Y., Masson, J. Y., Shah, R., O'Regan, P., and West, S. C. (2004) RAD51C is required for Holliday junction processing in mammalian cells. *Science* 303, 243-246

59. Xu, X., and Liu, Y. (2009) Dual DNA unwinding activities of the Rothmund-Thomson syndrome protein, RECQ4. *EMBO J* 28, 568-577

60. Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9, 671-675

61. Otwinowski, Z., and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276, 307-326

62. Karplus, P. A., and Diederichs, K. (2012) Linking crystallographic model and data quality. *Science* 336, 1030-1033

63. Evans, P. R., and Murshudov, G. N. (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* 69, 1204-1214

64. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Struct Biol* 75, 861-877
69. Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., 3rd, Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315

70. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-10041

71. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* **32**, W665-667

72. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612

73. Fenn, T. D. R., D.; Petsko, G. A. (2003) POVScript +: A Program for Model and Data Visualization Using Persistence of Vision Raytracing. *J. Appl. Crystallogr.* **36**, 944-947

74. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., and Ben-Tal, N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350
Table 1. Genes that show strong and conserved co-fitness with DUF328 genes.
Data extracted from the Fitness Browser (http://fit.genomics.lbl.gov/cgi-bin/myFrontPage.cgi)

| Organism | YaaA homolog | Cofitness Hit | Description | Ortholog cofitness |
|----------|---------------|---------------|-------------|-------------------|
| *Shewanella loihica* PV-4 | Shew_1093 | Shew_1611 | sbcB, exonuclease I | 0.66 0.63 |
| *Shewanella* sp. ANA-3 | Shewana3_3143 | Shewana3_1126 | recA, recombinase A | 0.63 0.72 |
|  |  | Shewana3_1104 | recombination and repair protein | 0.62 0.67 |
|  |  | Shewana3_2572 | sbcB, exonuclease I | 0.66 0.66 |
|  |  | Shewana3_0011 | recF, recombination protein F (RefSeq) | 0.57 0.75 |
| *Shewanella oneidensis* MR-1 | SO3540 | SO1328 | transcriptional regulator, LysR family | 0.68 0.65 |
|  |  | SO3430 | recA protein | 0.72 0.63 |
| *Pseudomonas syringae pv. syringae* B728a ΔmexB | Psyrr_1064 | Psyrr_2850 | hypothetical protein | no 0.8 |
|  |  | Psyrr_3174 | uroporphyrinogen-III C-methyltransferase / precorrin-2 dehydrogenase | no 0.78 |
| *Dechlorosoma suillum* PS | Dsui_2322 | Dsui_2345 | hypothetical protein | no 0.88 |
|  |  | Dsui_3354 | adenine specific DNA methylase Mod | no 0.88 |
| *Caulobacter crescentus* NA1000 | CCNA_03496 | CCNA_02062 | DNA repair protein recN | 0.67 0.62 |
| *Acidovorax* sp. GW101-3H11 | Ac3H11_1593 | Ac3H11_784 | Hydrogen peroxide-inducible genes activator | 0.65 0.68 |
|  |  | Ac3H11_2454 | Exodeoxyribonuclease I (EC 3.1.11.1) | 0.66 0.66 |
|  |  | Ac3H11_452 | Polyphosphate kinase (EC 2.7.4.1) | 0.49 0.8 |
|  |  | Ac3H11_135 | Catalase (EC 1.11.1.6) / Peroxidase (EC 1.11.1.7) | no 0.79 |
| Strain name | Genotype / Relevant Characteristics | Reference / Source |
|-------------|------------------------------------|--------------------|
| *E. coli* MG1655 | wild type | (Liu Y. et al, 2011)/Imlay lab, UIUC |
| ΔyaaA | As MG1655 plus Δ(yaaA1::cat)1 | (Liu Y. et al, 2011)/Imlay lab, UIUC |
| ΔqueD | As MG1655 plus ΔqueD::cat | This study |
| Δtgt | As MG1655 plus Δtgt::cat | H. Mori Collection |
| ΔqueD ΔyaaA | As ΔyaaA plus ΔqueD::cat | This study |
| Δtgt ΔyaaA | As ΔyaaA plus Δtgt::cat | This study |
| Hpx⁻ | As MG1655 plus Δ(ahpC-ahpF') kan::′ahpF Δ(katG17::Tn10)1 Δ(katE12::Tn10)1 | (Liu Y. et al, 2011)/Imlay lab, UIUC |
| Hpx⁻ ΔyaaA | As Hpx⁻ plus Δ(yaaA1::cat)1 | (Liu Y. et al, 2011)/Imlay lab, UIUC |
| Hpx⁻ ΔqueD | As Hpx⁻ plus ΔqueD::cat | This study |
| Hpx⁻ ΔyaaA ΔqueD | As Hpx⁻ ΔyaaA plus ΔqueD::cat | This study |
**Figure 1. Phylogenetic distribution of DUF328 proteins and representative genomes in Bacteria.** (A) Results of an AnnoTree query using PFAM family PF03883 (DUF328) in Bacteria with the resolution of genus level. Branches are highlighted in blue for phyla that harbor members of the DUF328 family. The trees were generated in AnnoTree with the e-value cutoff of 0.00001 and adapted for visualization: http://annotree.uwaterloo.ca/app/?qtype=pfam&qstring=PF03883&eval=0.00001. (B) Representative genome neighborhood diagrams of PF03883 (DUF328) family (red arrows). YaaJ: PF01235 (Sodium:alanine symporter family), QueF: PF14489 (QueF-like protein), 2OG: PF13640 (2OG-Fe(II) oxygenase superfamily), DEAD:PF00270 (DEAD/DEAH box helicase), helicase: PF00271 (Helicase conserved C-terminal domain), RecQ-like: PF16124 (RecQ zinc-binding), adh: PF00106 (short chain dehydrogenase), AphC: PF00578 (AhpC/TSA family), Nif: PF01784 (NIF3 (NGG1p interacting factor 3), zf: PF02591 (C4-type zinc ribbon domain).
**Figure 2. Schematics of examples of PF03883/IPR005583/DUF328 fusion proteins.**

Representatives were retrieved from pfam and Conserved Domain Architecture Retrieval Tool (CDART). H2O2_YaaD: PF03883; GIY-YIG: PF01541; SpoVK: COG0464; Ser_Rec: Ser_Recombinase Superfamily, cd00338; Rec: recombinase, pfam07508; Zn: Recombinase zinc beta ribbon domain, pfam13408; AlbA_2: Putative DNA-binding domain, PF04326; HTH: Helix-turn-helix domains.
Figure 3: YaaAEc binds to diverse DNA constructs with nanomolar affinity. (A-C) show EMSA of recombinant YaaAEc with various DNA constructs illustrated below each panel. The concentration of YaaAEc is shown at the top of each panel and the fraction of the total DNA signal that is shifted from the free position is quantified beneath each lane. In (D), the binding of YaaAEc to the 12 nt bubble duplex DNA from EMSA is fitted to a single-site, positively cooperative binding model. Data were measured in triplicate and verified for multiple preparations of YaaAEc. The Hill coefficient was fitted as 3.1.
Figure 4: Three-dimensional structure of YaaAEc defines a new fold. (A) and (B) show two views of the ribbon diagram of YaaAEc with α-helices lettered and β-strands numbered. The views in (A) and (B) are related by the 90° rotation indicated by the arrow. YaaAEc defines a new protein fold class. (C) shows the electrostatic potential surface for YaaAEc looking down into the apical cleft as calculated by APBS (see Experimental Procedures). Electrostatic potential is colored from blue (+10 kT/e) to red (-10 kT/e), where T is 298 K. (D) shows the same view into the apical cleft as (C) with key residues labeled. The green and gold residues compose two groups related by an approximate dyad axis that passes between residue pairs E130, Y175 and K9, K209.
Figure 5. YaaA_Ec possesses long stretches of non-standard structures extending from the core to the periphery of the protein. (A) shows a ribbon diagram of YaaA_Ec with the conserved SGXYG motif in yellow and the surrounding stretches of peptides adopting structures that are neither α-helix or β-strand in magenta. (B) is a close-up view of this region, where the dotted lines show a network of hydrogen bonds that mediate interactions between the conserved SGXYG motif and the surrounding residues. Tyrosine residues are highly represented in this region.
Figure 6. YaaA<sub>Ec</sub> has a helix-hairpin-helix (HhH) DNA binding motif. (A) shows a ribbon diagram of YaaA<sub>Ec</sub> with the region that is structurally conserved in other DNA binding proteins colored magenta. The classical HhH motif is defined by αB and αC (labeled). (B) shows a superposition of YaaA<sub>Ec</sub> (grey, blue) with the HhH motif (yellow) and bound DNA from XPF-ERCC1 endonuclease (PDB 6SXB). This illustrates one potential DNA binding mode for YaaA<sub>Ec</sub> and other DUF328 proteins.
The DUF328 family member YaaA is a DNA-binding protein with a novel fold
Janani Prahlad, Yifeng Yuan, Jiusheng Lin, Chou-Wei Chang, Dirk Iwata-Reuyl, Yilun Liu,
Valérie de Crécy-Lagard and Mark A. Wilson

J. Biol. Chem. published online August 12, 2020

Access the most updated version of this article at doi: 10.1074/jbc.RA120.015055

Alerts:
  * When this article is cited
  * When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts