Cellular function of the GndA microprotein during heat shock

Jessica J. Mohsen[1,2], Michael G. Mohsen[3,4], Kevin Jiang[1,2], Laura Quinto[3,5], Ane Landajuela[6,7], Erdem Karatekin[6,7,8,9,10], Farren J. Isaacs[3,5], Sarah A. Slavoff[1,2,10,11]

1 Department of Chemistry, Yale University, New Haven, CT 06511
2 Institute for Biomolecular Design and Discovery, Yale University, West Haven, CT 06516
3 Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511
4 Howard Hughes Medical Institute, Yale University, New Haven, CT 06511
5 Systems Biology Institute, Yale University, West Haven, CT 06516
6 Department of Cellular and Molecular Physiology, Yale School of Medicine, New Haven, CT 06510
7 Nanobiology Institute, Yale University, West Haven, CT 06516
8 Wu Tsai Institute, Yale University, New Haven, CT 06511
9 Université de Paris, Saints-Pères Paris Institute for the Neurosciences (SPPIN), Centre National de la Recherche Scientifique (CNRS), 75006 Paris, France
10 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511
11 To whom correspondence should be addressed:

Sarah A. Slavoff
Yale University West Campus
Institute of Biomolecular Design and Discovery
Molecular Innovations Center 232
600 West Campus Drive
PO Box 27392
West Haven, CT 06516
Phone: 203-737-8670
sarah.slavoff@yale.edu

Classification: Biological Sciences; Biochemistry
Keywords: Small open reading frame (sORF); microprotein; 6PGD; GndA; heat shock; electron transport

Abstract

Over the past 15 years, hundreds of previously undiscovered bacterial small open reading frames (sORFs) encoding microproteins of fewer than fifty amino acids have been identified. Biological functions have been ascribed to an increasing number of microproteins from intergenic regions and small RNAs, and many play integral roles in bacterial stress responses. However, despite numbering in the dozens in *Escherichia coli*, and hundreds in humans, same-strand frameshifted sORFs that internally overlap protein coding sequences remain understudied. To provide insight into nested genes, we characterized GndA, a frameshifted 36-amino acid microprotein nested within the 6-phosphogluconate dehydrogenase (6PGD) coding sequence. Using precise genome editing, we demonstrate independent contributions of GndA and 6PGD to cell growth at high temperature. GndA associates with membrane-associated complexes associated with electron transport and ATP generation, and supports ATP homeostasis during heat shock. Functional characterization of GndA thus adds to the catalog of bacterial microproteins that function in stress responses, while providing clear genetic evidence for the importance of an overlapping gene to cellular fitness.

Significance Statement

Same-strand overlapping, or nested, protein coding sequences optimize the information content of size-constrained viral genomes, but were previously omitted from prokaryotic and eukaryotic genome annotations. It was therefore surprising when numerous nested sORFs were recently discovered in bacteria and eukaryotes. Our case study of *E. coli* GndA supports the emerging model that overlapping genes can be expressed and contribute to cellular fitness, as GndA expression supports cell growth at elevated temperature. Characterization of nested sORFs may therefore revise our understanding of the complexity of bacterial and eukaryotic genes.

Main Text

Introduction

Small open reading frames (sORFs) are generally defined as short coding sequences (CDSs) that encode microproteins with biological functions (1–6). Due to an arbitrary 50-amino acid (aa) cutoff for bacterial genome annotation, bacterial sORFs have been overlooked until recently (3). This cutoff was applied because strategies for identification of longer, canonical proteins, which include sequence homology and conservation (7, 8), structural homology (9), and codon usage (10), are not sufficiently sensitive to differentiate coding vs. non-coding sORFs. Microproteins are also difficult to detect using standard proteomic workflows (11), and some microproteins exhibit limited conservation even among related species (12). Despite these challenges, new genomic (13, 14) and proteomic technologies (11) have revealed that microproteins (5) (also referred to as miniproteins (15), micropeptides (16), or small proteins (2, 3)) are numerous in both bacteria and eukaryotes (17). Molecular studies demonstrate that microproteins are functional and, in many cases, membrane associated (18, 19).

A link between bacterial microproteins and stress responses has been established. For example, a foundational study reported upregulation of multiple microproteins in *Escherichia coli* during the heat shock response at 45 °C (20). Additionally, predicted sORFs located in intergenic regions of *Enterobacteriaceae* were found to be expressed under Mg^{2+} starvation (21), and the stressosome cytoplasmic complex is linked to sigma B regulon activation via interaction with a membrane associated microprotein Prli42 (15, 22). Microproteins have also been implicated in mitochondrial ATP generation. In human cells, mitochondrial microproteins,

including BRAWNIN (23), MOCCI (24), SMIM4 and UQCC1 (25), regulate assembly and/or activity of the electron transport chain. Additionally, SLC35A4-MP localizes to the inner mitochondrial membrane where it contributes to cellular respiration (26). These studies position microproteins at the nexus of cellular metabolism and stress responses and suggest that microproteins may broadly play a role in maintaining cellular homeostasis from bacteria to human.

GndA is a 36-aa *E. coli* microprotein with a predicted alpha-helix (27). GndA is encoded by a frameshifted, UUG-initiated sORF nested within the CDS of the *gnd* gene in the +2 reading frame (Fig. 1*A*) (14). In addition to GndA, the *gnd* locus is known to encode 6-phosphogluconate dehydrogenase (6PGD), an enzyme of the pentose phosphate pathway (28). We previously reported that GndA increases in expression during heat shock. Since we detected GndA in fractionated membranes, it is alternatively possible that GndA increases in membrane association or aggregation during heat shock with or without a concomitant change at the protein level. However, despite evidence for its expression and heat shock regulation, the function of the *gndA* sORF has remained unexplored. We therefore sought to determine whether the GndA microprotein functions in the context of the bacterial heat shock response.

Overlapping, frameshifted sORFs are translated in bacteria and eukaryotes, but only a handful have been functionally characterized, and an evolutionary rationale for their existence is lacking. Providing a molecular rationale for fitness benefit of nested sORFs, overlapping microproteins such as MINAS-60 in mammals (29) and *comS* in *B. subtilis* (30–32) positively regulate phenotypes associated with the main CDS that they overlap. We therefore chose to explore the hypothesis that GndA may play a role in cellular metabolism, like the *gnd* gene within which it is nested. Our data support the hypothesis that GndA promotes cell growth and ATP generation at elevated temperature, possibly as a result of its association with protein complexes of the inner membrane, demonstrating that the overlapping ORFs encoded in the *gnd* gene locus both contribute to metabolic homeostasis and cellular fitness.

Results

Given that GndA is nested within 6PGD and increases in co-fractionation with membranes during heat shock (27), we hypothesized that this microprotein might play a role in central metabolism at elevated temperature in *E. coli*. To determine what its role may be, we first examined its cellular localization. Many previously identified microproteins exhibit single transmembrane helices (33–35), and AlphaFold 3 (36) predicts the GndA peptide to form an alpha helix extending from residues 7 to 35 (Fig. 1*B*), consistent with our prior bioinformatic prediction (27). However, our prior study did not determine whether GndA associates with the inner or outer membrane. We therefore carried out subcellular fractionation, revealing association of GndA with the inner membrane when overexpressed from a pET vector in *E. coli* BL21(DE3) cells grown at 37 °C, and with both the inner membrane and the pellet during heat shock (Fig. S1*A*). It is likely that the sedimentation with the pellet during heat shock, as well as the appearance of a second band migrating at higher molecular weight, is a consequence of oligomerization or aggregation (similar to the “sticky” mitochondrial microprotein mitoregulin (37, 38)) and may not reflect its endogenous localization. Nevertheless, these data support membrane association of overexpressed GndA, though we cannot yet determine whether GndA is an integral or peripheral membrane protein. We previously reported that the GndA microprotein increases in expression and/or membrane association during heat shock (27). We therefore examined levels of the 6PGD protein. Endogenous 6PGD remained unchanged at the protein level before and after heat shock (Fig. S1*B*).

While evolutionary conservation is not required for microproteins to contribute to cellular fitness (18), many functional microproteins are conserved (39–41), and conservation is a robust metric to identify microproteins with phenotypic effects (42). We thus hypothesized that if homologs are present in multiple bacterial lineages, GndA is more likely to play a role in cellular fitness. To examine *gndA* conservation, we first performed a translated nucleotide BLAST (tblastn) search using the *E. coli* GndA microprotein sequence. We then retrieved the genomic DNA sequences of species that returned putative homologs in order to manually determine whether syntenic Shine Dalgarno sequence, start and stop codons were present. 23 GndA homologs, all nested within the 6PGD CDS of the respective organism, with > 50% sequence positives were identified in Gram-negative bacteria (Fig. 1*C-D*, Fig. S1*C*). The presence of GndA homologs beyond *E. coli* is consistent with functionality of this microprotein.

We created a series of strains to examine the function of GndA in heat shock, as well as to separate the phenotypes associated with overlapping GndA and 6PGD (Table S2). *E. coli* K-12 MG1655 cells (WT) were engineered to lack the entire *gnd* gene and thus both CDSs (Δ*gnd*), to delete 6PGD alone (Δ6PGD), or to specifically eliminate GndA expression while introducing minimal or no amino acid changes in the 6PGD reading frame. Three different strategies to abrogate GndA expression were pursued. Two GndA premature stop mutants (ΔGndA^{S116L} and ΔGndA^{A117V}) were generated to truncate GndA after residues 17 and 18. These mutations are however not silent in the 6PGD reading frame, and introduce conservative S116L or A117V mutations in the 6PGD coenzyme domain, respectively, which are reflected in the strain names. We also created a strain with the goal of mutating its translation start site. GndA has previously been reported to initiate at TTG (14), but two additional in-frame ATG codons can be found upstream and adjacent to this site. We therefore generated a 3-start codon ablation mutant (ΔGndA^{SCA}), and all three GndA start site mutations in this strain are silent in the 6PGD +1 reading frame. We note that we have not quantified GndA protein level expression in these strains, due to limited solubility of the microprotein. Because the *gndA* sORF overlaps the portion of the main CDS encoding the coenzyme domain of 6PGD, we additionally generated a 6PGD mutant with a A148V mutation downstream of the sORF but within the same protein domain (6PGD^{A148V}). Finally, to conclusively assign phenotypes to the *gndA* sORF, rescue cells in which GndA was expressed from an arabinose-inducible plasmid were created from each of these mutant lines (ΔGndA^{S116L}-Rescue, ΔGndA^{A117V}-Rescue, ΔGndA^{SCA}-Rescue, 6PGD^{A148V}-Rescue, (Table S2, Fig. S2)).

We then tested the hypothesis that GndA plays a role in cell growth and viability at elevated temperature. ΔGndA^{SCA} cells exhibit a pronounced reduced growth phenotype during stationary phase at 45 °C relative to WT cells (Fig. 2*A*), consistent with a specific requirement for GndA for growth at elevated temperature. Supporting this finding, GndA mutants bearing premature stop codons (ΔGndA^{S116L} and ΔGndA^{A117V}) exhibited comparable, or more pronounced, growth defects, respectively (Fig. 2*B-C*). 6PGD expression and activity also contribute to fitness at elevated temperature; introduction of a point mutation in the coenzyme domain downstream of the *gndA* sORF (6PGD^{A148V}) or deletion of its start codon (Δ6PGD) also caused reduced growth at stationary phase (Fig. 2*D-E*). Interestingly, cells sustaining a deletion of both ORFs (Δ*gnd*) exhibit a lesser growth defect (Fig. 2*F*), possibly indicating a partial rescue of the *gndA* deletion phenotype by loss of 6PGD. The status of *tolC*, which was leveraged to edit the *E. coli* genome, had a very small deleterious effect on growth of 6.2 ± 1.8% at stationary phase at 45 °C relative to WT cells (Fig. S3*A*). In accordance with a specific role for GndA (and 6PGD) during growth at elevated temperature, no strains or controls exhibit changes in growth at 30 °C (Fig. S3*B*). Additionally, lower colony forming units (CFU) were observed in all three GndA mutant strains as well as 6PGD^{A148V} mutant cultures compared to WT, indicating a reduction in viable cells in

the mutants after growth at 45 °C (Fig. 2*G*). These results clearly indicate that both GndA and 6PGD independently contribute to fitness at elevated temperature.

To further demonstrate a direct role for GndA in cellular fitness at high temperature, we expressed GndA from an arabinose-inducible plasmid in selected mutant strains. GndA reintroduction rescued stationary phase growth at 45 °C in both ΔGndA^{S116L} and ΔGndA^{SCA} cells (Fig. 3*A*-*B*). GndA reintroduction also weakly, but statistically significantly, rescued late exponential phase growth of the ΔGndA^{A117V} strain, as well as stationary phase growth of the 6PGD^{A148V} coenzyme domain mutant strain, although no rescue was apparent in late stationary phase for either of these strains (Fig. 3*C*-*D*). GndA expression therefore improves growth at 45 °C to some degree in all *gndA* and 6PGD mutant strains tested.

In order to maintain homeostasis, gene regulatory networks are reshaped in response to stress and genetic perturbations. We hypothesized that the transcriptional state of cells would provide additional insight into the role of GndA in the heat shock response. We performed RNA sequencing (RNA-seq) in ΔGndA^{S116L} and WT cells immediately after heat shock (Fig. 4*A*-*B*). Principal component analysis reveals clustering of individual RNA-seq replicates, as expected (Fig. S4*A*-*D*). ΔGndA^{S116L} cells exhibit upregulation (> 3.5-fold change and p-value < 0.05) of 116 genes upon heat shock relative to WT cells (Fig. 4*A*). 53 of these genes are rescued to WT expression levels in ΔGndA^{S116L}-Rescue cells at 45 °C, demonstrating that their upregulation is a direct consequence of the lack of GndA (Fig. 4*B*). After exclusion of the y-ome (43), gene ontology (GO) analysis (44, 45) of genes rescued by GndA expression during heat shock reveals involvement in the electron transport chain (Fig. 4*C*). Dysregulation of genes encoding proteins of the electron transport chain in heat shocked ΔGndA^{S116L} cells is consistent with the reduced growth phenotype observed in GndA mutants and suggests a cellular compensatory response to defective stress responses in the absence of this microprotein.

Many microproteins bind to and regulate the function of macromolecular complexes (46), so we hypothesized that GndA exerts its effects on cellular fitness through microprotein-protein interactions. In order to provide evidence for GndA interaction partners at endogenous expression levels, we generated GndA^{SPA} cells, in which a sequential peptide affinity (SPA) tag (47, 48) was appended to the N-terminus of the chromosomal copy of GndA. We co-immunoprecipitated endogenously expressed SPA-GndA at both 37 °C and 45 °C and quantified proteins enriched compared to matched cell lysates exposed to anti-FLAG beads pre-incubated with 3xFLAG peptide antigen as negative controls (Fig. 5*A*-*B*). In both samples, we observed enrichment of multiple inner membrane complexes required for electron transport and ATP generation, including members of the *nuo* operon which encodes respiratory complex I (RCI) of the electron transport chain (red dots), as well as multiple subunits of ATP synthase (black dots). We additionally observe enrichment of pyruvate dehydrogenase in both samples (cyan dots). At 45 °C, we observe enrichment of additional proteins associated with the electron transport chain including type II NADH:quinone oxidoreductase (NDH-2) (purple dot) and cytochrome *bd*-I oxidase (salmon dot). The *nuo* operon encodes the protein components of RCI, NADH:quinone oxidoreductase, or NDH-1, an inner membrane complex that catalyzes the first step in the electron transport chain (49–54). We detected enrichment of NuoJ, an RCI subunit, by FLAG-GndA under strong overexpression conditions (Table S3), which we focused on for follow-up validation because it, like GndA, is inner membrane-associated. The enrichment of NuoJ-Myc by FLAG-GndA was validated by co-immunoprecipitation and western blotting when both proteins were strongly overexpressed in *E. coli* (Fig. S5). Taken together, these results are consistent with localization of GndA at the inner membrane, where it directly or indirectly associates with inner membrane complexes involved in metabolism, ATP synthesis, and the electron transport chain.

Given that GndA co-immunoprecipitation enriches RCI and ATP synthase, we hypothesized that GndA could be important for ATP generation at high temperature. We observed a decrease in intracellular ATP at 45 °C in the absence of GndA or 6PGD relative to WT (Fig. 5*C*), indicating that both of these overlapping ORFs contribute to ATP production at elevated temperature. ΔGndA^{S116L}-Rescue cells in which GndA expression was induced for 2 hours show a statistically significant rescue in ATP levels. Δ*gnd* cells exhibit comparable intracellular ATP levels to WT, consistent with the partial rescue in growth at high temperature observed for the same strain (Fig. 2*F*). Δ6PGD cells also exhibit a reduction in ATP levels, consistent with both microprotein and canonical protein influencing cellular fitness and homeostasis at elevated temperature. These results suggest that GndA supports ATP homeostasis at elevated temperature, and that this may be the pathway through which it improves cellular fitness during growth at 45 °C.

Discussion

In this work, we provide evidence for the cellular function of the 36-aa GndA microprotein, which overlaps 6PGD. Our work enriches existing knowledge about frameshifted, internal ORFs as it establishes a role for GndA in the bacterial heat shock response. While dozens of internal, overlapping ORFs have been identified in *E. coli* (14)*, o*nly two were characterized prior to 2019 (*comS* in *B. subtilis* (30–32) and *rpmH* in *T. thermophilus* (55)), and little is known about the molecular functions of nested ORFs. It has been proposed that overlapping gene arrangements may enable co-encoding of proteins with related functions. This model fits the two proteins encoded in the *gnd* locus, GndA and 6PGD, which are both involved in cell growth, viability, and metabolism at elevated temperature.

GndA exhibits conservation in limited species, consistent with *de novo* origination (12, 56). While GndA is present in multiple gram-negative species, it is less widely distributed than 6PGD, which is conserved from bacteria to humans. The *gnd* locus has previously been found to be polymorphic within *E. coli*, possibly due to its proximity to the O-antigen-encoding *rfb* region (57–62). We therefore speculate that mutations in a tolerant region of the 6PGD coenzyme domain resulted in emergence of *gndA* in the +2 reading frame, as has previously been demonstrated for overlapping viral genes within intrinsically disordered regions (63). It will be of interest to test this hypothesis in the context of GndA as well as other nested out-of-frame sORFs in *E. coli* in the future. Nonetheless, the presence of GndA homologs in multiple species after its emergence is consistent with its demonstrated fitness benefit during cellular stress.

Systematic deletions or mutations in the GndA or 6PGD reading frame clearly established that loss of either GndA or 6PGD led to growth defects at 45 °C. More importantly, reintroducing GndA on a plasmid into these mutant strains led to detectable rescue of high-temperature growth, though the extent of rescue varied in different strains. GndA could even rescue the 6PGD-specific mutant strain, suggesting that its fitness benefit is general and that GndA and 6PGD act through different molecular mechanisms or pathways. Puzzlingly, we also reproducibly observe that, unlike mutations in a single ORF, the *gnd* mutant strain lacking both GndA and 6PGD does not exhibit a large growth defect at 45 °C. There are several possible explanations for this. One possibility is that point and truncating mutations in either GndA or 6PGD could hypothetically generate dominant-negative, toxic or aggregation-prone mutant protein fragments; simply deleting both ORFs may be less deleterious because these mutant proteins are not produced. Another hypothetical possibility is that cells lacking either GndA or 6PGD can continue aerobic respiration and ATP synthesis at 45 °C to some extent, while loss of both ORFs may promote complete metabolic reprogramming at high temperature that

paradoxically improves overall cell growth. Nonetheless, the combination of loss-of-function mutations and complementation clearly demonstrate that GndA supports optimal *E. coli* growth at high temperature and that it acts independently of 6PGD.

GndA likely functions in the heat shock response via association with the inner membrane and the protein complexes that reside therein. Endogenously expressed GndA enriches several proteins involved in the electron transport chain, including RCI and cytochrome *bd*-I oxidase, as well as ATP synthase, from cell extracts at both 37 °C and 45 °C. In addition, GndA and complexes of the electron transport chain (52, 64, 65) localize to the inner membrane, supporting their association. It is not yet clear which of these interactions are direct or indirect, though it is interesting to note that cytochrome *bd*-I oxidase, a terminal electron acceptor of the respiratory chain, is already known to associate with two microproteins, CydX and CydH (66, 67). It is therefore reasonable to conclude that GndA exerts its fitness benefit at high temperature via increased association with the inner membrane, where it may encounter respiratory complexes and/or ATP synthase, and future studies will be required to elucidate its direct interactions and molecular mechanism.

The culmination of genetic, phenotypic, interactomic, and molecular evidence presented in this work supports a fitness benefit of increased GndA expression during heat shock by supporting ATP generation. This work expands the emerging paradigm that overlapping, frameshifted ORFs and *de novo* genes may broadly encode microproteins with important cellular functions.

Materials and Methods

Bioinformatics

Next-generation sequencing (NGS): For each sample type consisting of 3 replicates each, 2 µg of extracted RNA per replicate was submitted to the Yale Center for Genome Analysis, where ribosomal RNA depletion and further sample processing was performed. Sequencing was performed with the Illumina NovaSeq system at a depth of approximately 20 million reads per sample. Paired-end reads were sequenced with a read length of 150 base pairs.

Computational analysis of sequencing results: Sequencing data was obtained in gzipped fastq files (.fastq.gz). Read files were trimmed to deplete adapter sequences using bbduk.sh, part of the BBMap array of tools (68). The resulting trimmed read files were then assessed for quality control with fastQC. Subsequently, the reads were mapped to the *E. coli* str. K-12 substr. MG1665 genome using STAR (69). The resulting alignment files (.BAM) were analyzed using featureCounts (70).

The featureCounts program outputs a tab-separated column (.tsv) file. Microsoft Excel was used for downstream analysis. To normalize for gene length, reads per kilobase (RPK) values were calculated for each replicate by dividing the reads for each gene by the length of the gene in kilobases. To normalize for read depth, first the scaling factor for each replicate was calculated by summing up all RPK values and dividing by 10^6. Then, transcripts per million (TPM) values were calculated by dividing the RPK value for each gene by the scaling factor for the replicate. TPM values were used in all subsequent analyses (*i.e.*, volcano plots).

GndA homology studies: tblastn searches were run against the *E. coli* K-12 GndA amino acid sequence (71). Nucleic acid sequences of top hits were aligned with the Clustal Omega multiple sequence alignment program (72).

Protein secondary structure prediction: The AlphaFold server was used to generate a predicted structure for GndA based on the *E. coli* K-12 GndA amino acid sequence (36).

Protein extraction and co-immunoprecipitation (co-IP)

Protein extraction: Equal numbers of cells normalized by OD_{600} were taken from control and heat shocked cultures and pelleted by centrifugation at 21 °C at 4,000 rpm for 5 min. Media was decanted and pellets were resuspended in 1 mL of BPER lysis buffer (Thermo Scientific catalog# 90078) containing DNase I, Lysozyme, and Roche complete protease inhibitor (Sigma catalog #11836170001). Cells were lysed by sonication with a QSonica Misonix Microson Ultrasonic Cell Disruptor XL-2000 for 5 rounds, 5 s on 25 s off, on ice at medium (5) setting. Lysate was clarified by centrifugation at 4 °C at 21,000 rcf for 5 min. Clarified lysate containing solubilized protein was removed without disrupting pelleted material and transferred to a fresh tube. For protein extraction from $GndA^{SPA}$ cells, single colonies were inoculated in LB broth and grown overnight with shaking at 200 rpm at 37 °C. Cultures were diluted 1:100 in 10 mL of fresh LB broth and grown to exponential phase (OD_{600} = 0.3-0.5). Exponential cultures were then grown for 2 h at 37 °C or 45 °C, and were then immediately collected for protein extraction as described.

Co-immunoprecipitation (co-IP): For each clarified lysate, 25 µL of anti-FLAG M2 affinity agarose gel (Sigma catalog #A2220) was washed in 1 mL ice-cold BPER and collected at 1,000 rcf for 1.5 min at 4 °C. Wash supernatant was removed without disrupting agarose gel, and the protein containing clarified lysate was added. The protein lysate was incubated with agarose gel at 4 °C with rotation for 1 h or overnight. Samples were spun down at 1,000 rcf for 1.5 min at 4 °C to collect beads bound to FLAG fusion protein and binding partners. To remove nonspecific binding, supernatant was removed and beads were washed twice with TBS containing 1% Triton X-100 (Sigma catalog #T8787) and 500 mM NaCl at 4 °C with rotation for 3 min. A final wash with 10 inversions by hand was done in TBS containing 1% Triton X-100 and 150 mM NaCl to return proteins to physiological salt concentration. Bound protein was eluted off agarose gel with 30-50 µL 1X 3XFLAG peptide (Sigma catalog #F4799) at 4 °C with rotation for 1 h or boiled. Agarose beads were collected at 4 °C 1,000 rcf for 2 min, and supernatant containing immunoprecipitated protein was removed and stored at -80 °C for downstream western blotting and LC-MS/MS.

pETDuet-1 Co-IP: pETDuet-1 expression vectors encoding 1) FLAG-GndA alone, 2) NuoJ-Myc alone, or 3) both constructs, were transformed into BL21(DE3) cells. Single colonies were inoculated in Luria-Bertani (LB) broth with 100 µg/mL carbenicillin (Sigma catalog #C1389-5G) overnight. After a 1:100 dilution and growth to exponential phase (OD = 0.3-0.5), T7 protein expression was induced at 37 °C with 0.5 mM IPTG. Total protein was extracted by sonication and lysates were clarified twice by centrifugation at 20,000 rcf for 3 min at 4 °C. FLAG-GndA was immunoprecipitated with anti-FLAG M2 affinity agarose gel and samples were eluted with 3XFLAG peptide. Samples were run on both 15% Glycine-SDS-PAGE to resolve NuoJ, and 16% Tricine-SDS-PAGE to resolve GndA. DYKDDDDK tag rabbit monoclonal antibody (Cell Signaling catalog #14793S) and Myc tag rabbit polyclonal antibody (Rockland catalog #600-401-381) were used for Western blotting primary immunostaining. Secondary staining was performed with horseradish peroxidase conjugated rabbit IgG (H&L) (Rockland catalog #611-1302) and visualized with clarity western ECL substrate (BioRad catalog #1705061).

Dual selection for bacterial genome editing

Scarless mutants at the *gnd* and *nuo* loci were generated with an SDS/colicin E1 dual-selectable system (73, 74). WT *E. coli* K-12 MG1655 cells were transformed with a ColE1 kanamycin-resistant plasmid (75) carrying anhydrotetracycline-inducible λ-Red recombineering

genes (Exo, Beta, and Gam). To prepare cell cultures for gene editing, a single colony streaked out on LB agar with kanamycin was inoculated in LB broth with 50 µg/mL kanamycin at 37 °C and shaken at 200 rpm overnight. The culture was diluted 1:100 in LB supplemented with 50 µg/mL kanamycin, and 100 ng/mL of anhydrotetracycline (Sigma catalog #37919) was added after 45 min of incubation to induce the λ-Red recombineering genes. Cultures were grown to OD_{600} = 0.5 and chilled on ice for 20 min. All tips, tubes, cuvettes, and autoclaved milli-Q water were chilled on ice. 1 mL of chilled culture was spun down at 3,000 rpm for 5 min at 4 °C and washed 3 times with chilled milli-Q water. To generate a Δ*tolC* strain from WT MG1655, an oligo mixture was prepared with 1 µL 100 µM oligonucleotide designed to disrupt the *tolC* locus added to 40 µL chilled milli-Q water. Washed cells were gently resuspended in oligo mixture and incubated on ice for 1 min. 50 µL of the cell-oligonucleotide mixture was transferred to a pre-chilled 1 mm electroporation cuvette (BioRad catalog #1652083) and transformed with 1.8 kV, 200 Ω, and 25 µF in a MicroPulser Electroporation Apparatus. Immediately after electroporation, cells were resuspended in 1 mL room temperature SOC medium (Sigma catalog #S1797-10X5ML). 2 mL LB Miller media was additionally added for a total recovery culture volume of 3 mL. Cells were recovered for a minimum of 2 h at 37 °C with 200 rpm shaking to allow turnover of transmembrane TolC before selection. Recovery cultures were diluted 1:100 in LB supplemented with approximately 10 µg/mL colicin E1 (prepared as described previously (76) in strain JC411) and 64 µg/mL vancomycin to select for *tolC* marker displacement, and incubated for 6-12 h at 37 °C with shaking at 200 rpm. Δ*tolC*-selected cultures were plated on antibiotic-free LB agar plates and individual colonies were screened for *tolC* knockout by both negative selection in LB supplemented with 0.005% SDS and colony PCR showing a lack of amplicon from the *tolC* locus.

All mutant strains, including Δ*tolC*, Δ*gnd*, Δ6PGD, $\Delta GndA^{S116L}$, $\Delta GndA^{A117V}$, $\Delta GndA^{SCA}$, $6PGD^{A148V}$, and $GndA^{SPA}$, were generated using sequential *tolC* selection/counterselection. At each target locus (*gnd*), the target gene was first knocked out in the MG1655 Δ*tolC* strain by knocking in the *tolC* cassette. This then served as a displaceable marker to select for knock in of tagged and/or mutant genes. First the Δ*gnd* strain was constructed via double-stranded DNA (dsDNA) recombineering, replacing the native locus with *tolC*. Recombineering cassettes were generated by amplifying the *tolC* cassette from MG1655 cells, with primers encoding 50-bp overhangs surrounding the gene targeted for deletion (Table S1). PCR products were purified by PCR purification (Qiagen catalog #28104) and full-length cassettes were subsequently isolated by agarose gel purification (Qiagen catalog #28506) and verified with Sanger sequencing. The Δ*tolC* strain bearing the recombineering plasmid was cultured and induced as described above, then electroporated with 1 µL of dsDNA substrate (100 – 400 ng of dsDNA). For positive *tolC* selection, cells were recovered for 2 h in SOC with LB, then plated on LB agar plates supplemented with 0.005% SDS. Individual knockout strains were confirmed for *tolC* integration by colony PCR and Sanger sequencing.

To generate tagged and mutant strains, the Δ*gnd* strain was subjected to another round of dsDNA recombineering to displace the *tolC* marker with the desired mutant sequence. Tagged and mutant cassettes (Table S1) were generated by colony PCR and overlap extension PCR with Phusion polymerase, or amplification from plasmids purchased from Genscript, and confirmed by Sanger sequencing. The Δ*gnd* strain was cultured and induced as described above. Purified cassettes were electroporated into the strains, and cells were subjected to colicin E1 and vancomycin counterselection as in the initial *tolC* knockout experiment. After counterselection, mutant colonies were plated on LB agar and tested for desired mutations by negative selection in LB supplemented with 0.005% SDS as well as colony PCR and Sanger sequencing.

Proteomics LC-MS/MS

Quantitative proteomics was performed as previously reported (77). Gel slices containing entire lanes were digested with trypsin at 37 °C for 14-16 h. The resulting peptide mixtures were extracted from the gel, dried, subjected to ethyl acetate extraction to remove residual detergent, de-salted with a peptide cleanup C18 spin column (Thermo Scientific catalog #89870), then resuspended in 35 µL 0.1% formic acid (FA), followed by centrifugation at 15,000 rcf at 4 °C for 30 min. A 5-µL aliquot of each sample was injected onto a pre-packed column attached to an Easy-nLC 1200 (Thermo Scientific) in-line with a Thermo Scientific Q Exactive Plus Hybrid Quadrupole-Orbitrap mass spectrometer. A 125-min gradient was used to further separate the peptide mixtures as follows (solvent A: 0.1% FA; solvent B: acetonitrile (ACN) with 0.1% FA): linear gradient from 5% B to 45% B over 90 min, 45% B to 85% B over 1 min, isocratic flow at 85% B was maintained for 10 min, followed by a gradient from 85% B to 5% B over 1 min, and isocratic flow at 5% B over 23 min. The full MS was collected over the mass range of 300-1,700 m/z with a resolution of 70,000 and the automatic gain control (AGC) target was set as 3×10^6. MS/MS data was collected using a top 20 high-collisional energy dissociation method in data-dependent mode with a normalized collision energy of 28.0 eV and a 1.6 m/z isolation window. MS/MS resolution was 17,500 and dynamic exclusion was 90 s. Files were analyzed using MaxQuant, and oxidation of methionine and N-terminal acetylation were set as variable modifications. *E. coli* K-12 Uniprot plus NuoJ-FLAG and GndA was used as the database for searching. Protein quantitation was accomplished with MaxQuant LFQ (version 2.0.2.0) or spectral counting (78). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (79) partner repository with the dataset identifier PXD065696 and 10.6019/PXD065696.

Bacterial heat shock

Single colonies were inoculated in LB broth and grown overnight with shaking at 200 rpm at 30 °C or 37 °C. Cultures were then diluted 1:100 in fresh LB broth and grown to exponential phase (OD_{600} = 0.3-0.5). Each culture was split into two. One half (control) was returned to the original incubator shaker (30-37 °C). The other half (heat shock culture) was warmed for 5 min in a 45 °C water bath with shaking by hand until an internal temperature of 45 °C was reached. Then, the heat shock culture was kept in a 45 °C water bath for 20 min with shaking by hand. Control and heat shock cultures were harvested simultaneously. Control cultures were grown at 30 °C to observe the heat shock effect over a larger temperature change.

RNA-seq

RNA extraction after heat shock was carried out using the RNeasy Protect Bacteria kit (Qiagen catalog #74524). 5×10^8 *E. coli* cells were harvested from each sample using an RNeasy mini spin column. Qiagen protocol 4 was followed for enzymatic lysis with lysozyme (DOT Scientific catalog #DSL38100) and 20 µL proteinase K (NEB catalog #P8107S). Qiagen protocol 7 was followed for RNA purification including DNase I digestion on-column (Qiagen catalog# 79254). Total RNA was eluted in 50 µL RNase free water. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (80) and are accessible through GEO Series accession number GSE269907 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE269907).

Growth curve measurements

WT and GndA mutant *E. coli* cell lines were streaked out on LB agar plates and incubated at 30 °C or 37 °C overnight. Cells containing a pBAD33 plasmid designed to rescue expression of GndA were grown on LB agar with 25 µg/mL chloramphenicol. Single colonies were inoculated in LB broth (with chloramphenicol for cells carrying pBAD33) and grown overnight at 30 °C or 37 °C with shaking at 200 rpm as appropriate for the experimental control conditions. Cultures were

then diluted 1:100 and grown to exponential phase (OD_{600} = 0.3-0.6). For transient heat shock growth curves, exponential cultures were first heat shocked by hand as described, then all samples were normalized to OD_{600} = 0.1, and finally transferred to a sterile round bottom 96-well plate with lid (Corning catalog #3788). Four replicates per cell line were incubated at 30 °C with continuous orbital shaking at 205 cpm while growth was monitored in a BioTek Synergy Neo2 Multimode Reader. To monitor cell growth under prolonged heat shock for greater than 20 min, cultures were brought to 45 °C directly in the plate reader. Cultures were first grown to exponential phase (OD_{600} = 0.3-0.6) and then normalized to OD_{600} = 0.2 and transferred to a sterile round bottom 96-well plate with lid. 0.02% L-(+)-arabinose was added to Rescue cell cultures immediately before transferring to the 96-well plate. Four replicates per cell line were incubated at 45 °C with continuous orbital shaking while growth was monitored in a BioTek Synergy Neo2 Multimode Reader. OD_{600} readings were taken at 15 min intervals.

ATP luminescence assays

BacTiter-Glo Microbial Cell Viability Assay (Promega catalog #G8230) was used to compare total cellular ATP in WT and *gnd* mutant lines after growth at 45 °C. Promega protocols 3.A and 3.B were followed. Single colonies were inoculated in LB (with chloramphenicol for cells carrying pBAD33) and grown overnight at 37 °C at 200 rpm in an incubator shaker. Cultures were diluted 1:100 and grown to exponential phase (OD_{600} = 0.3-0.5), then normalized to OD_{600} = 0.2. 100 µL of culture was aliquoted into a sterile round bottom 96-well plate with lid (Corning catalog #3788) and grown at 45 °C in a plate reader with OD_{600} readings as described above for growth curve measurements. After 2 hours of growth at 45 °C, 50 µL of culture per replicate was aliquoted into a 96-well white Lumitrac polystyrene microplate (Greiner catalog# 655075), and 50 µL of BacTiter-Glo reagent was added to each well with a multichannel pipette. The plate was loaded into a BioTek Synergy Neo2 Multimode Reader set to 23 °C. Orbital shaking at 425 cpm was applied for 10 s. Luminescence was recorded once per min for 20 min with an integration time of 40 ms per well, read height 4.5 mm, and gain of 135. 50 µL of BacTiter-Glo reagent was added to 50 µL LB to measure background luminescence to subtract from experimental readings, which were then normalized to OD_{600} measurements taken at the 2-hour growth timepoint for three replicates.

Colony forming unit measurements

WT and mutant *gnd* strains were grown in a plate reader at 45 °C as described for growth curve measurements. After 24 h the plate was immediately removed from 45 °C incubation and serial dilutions were performed for each replicate per strain. 5 µL of each diluted culture was plated on LB agar. Plates were incubated overnight at 37 °C. Visible colonies were manually counted and normalized to dilution factor and final OD_{600} reading.

Western blotting

Standard western blotting procedures were followed with modification (81) using a Mini-PROTEAN Tetra Cell, Mini-Trans blot module, and power supply (Bio-Rad catalog# 1658033). Protein samples were resolved on 16% Tricine-SDS-PAGE (82) for proteins with molecular weights less than 20 kDa, and 15% Glycine-SDS-PAGE was used for proteins with molecular weights above 20 kDa. 4X SDS loading buffer was added to protein samples extracted from equivalent *E. coli* samples. PAGE was run at 50-80 V as proteins migrated through the stacking gel, and at 100-120 V through the resolving gel until the protein xtra prestained ladder (Bio-Rad catalog # 1610377) achieved good separation. Gels were equilibrated in pre-chilled transfer buffer with 15% methanol and then protein was transferred to a 0.45 µm nitrocellulose membrane (Sigma catalog# 10600002) via wet electroblotting on ice for 1 h at 300 mV. Membrane was blocked in 3% BSA in TBST for 2 h with rocking at room temperature. Primary antibody was added 1:3,000 in 3% BSA in TBST and incubated at 4 °C overnight with nutation.

The membrane was washed three times in TBST for 5 min at room temperature with rocking. Secondary antibody was added 1:10,000 in 3% BSA in TBST and incubated 1 h at room temperature with rocking. Three washes in TBST were performed. Chemiluminescent imaging was done with horseradish peroxidase conjugated IgG and clarity western ECL substrate (BioRad catalog# 1705061). Images were taken on BioRad ChemiDoc XRS+ Gel Imaging System.

Acknowledgments
We thank all current and former members of the Slavoff and Breaker labs for helpful discussions. This work was funded in part by the National Institutes of Health (1R01GM155404-01), a Mark Foundation for Cancer Research Emerging Leader Award, a Paul G. Allen Frontiers Group Distinguished Investigator Award, and a Sloan Research Fellowship (FG-2022-18417) (to S.A.S.). J.J.M. was supported in part by a Yale University fellowship associated with the NIGMS Chemistry-Biology Interface Training Program (5T32GM067543) and a Roberts Fellowship from the Yale University Department of Chemistry. M.G.M. is a Howard Hughes Medical Institute (HHMI) awardee of the Life Sciences Research Foundation, and received supplementary support from HHMI funds awarded to the Breaker lab.

Figures and Figure Legends

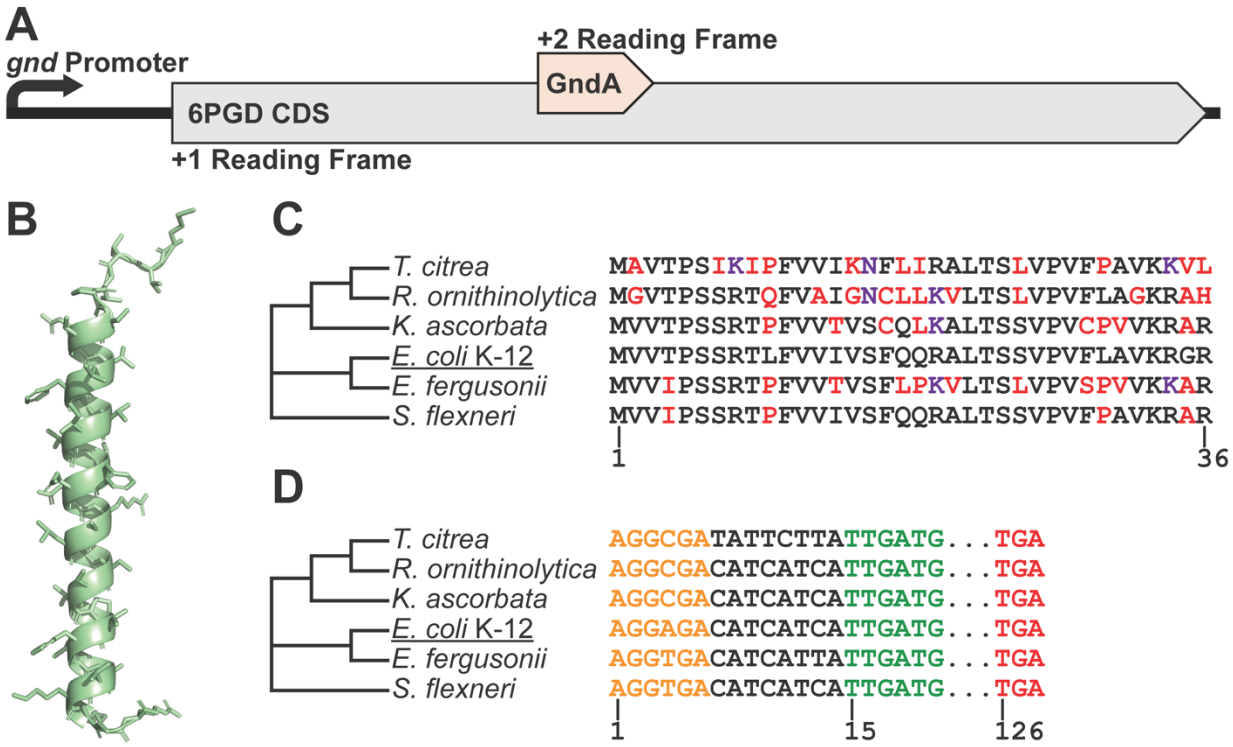

Fig. 1. The *gndA* small open reading frame (sORF) is nested within the 6-phosphogluconate dehydrogenase (6PGD) coding sequence. (*A*) The *E. coli* K-12 MG1655 *gnd* gene locus encodes two proteins with distinct amino acid sequences (27). Schematic of the *gnd* gene with the coding sequence (CDS) for 6PGD defined as the +1 reading frame. The GndA microprotein is encoded by a nested sORF in the +2 reading frame. (*B*) High confidence GndA secondary structure prediction generated with AlphaFold 3 (36). (*C*) GndA conservation in Gram-negative bacteria. Residues in black are conserved, in purple are positives (71), and in red are not conserved. (D) The coding and regulatory sequences for GndA homologs were aligned and Shine Dalgarno sequence (orange), cryptic TTG start codon (green), traditional ATG start codon (green), and stop codon (red) were manually identified.

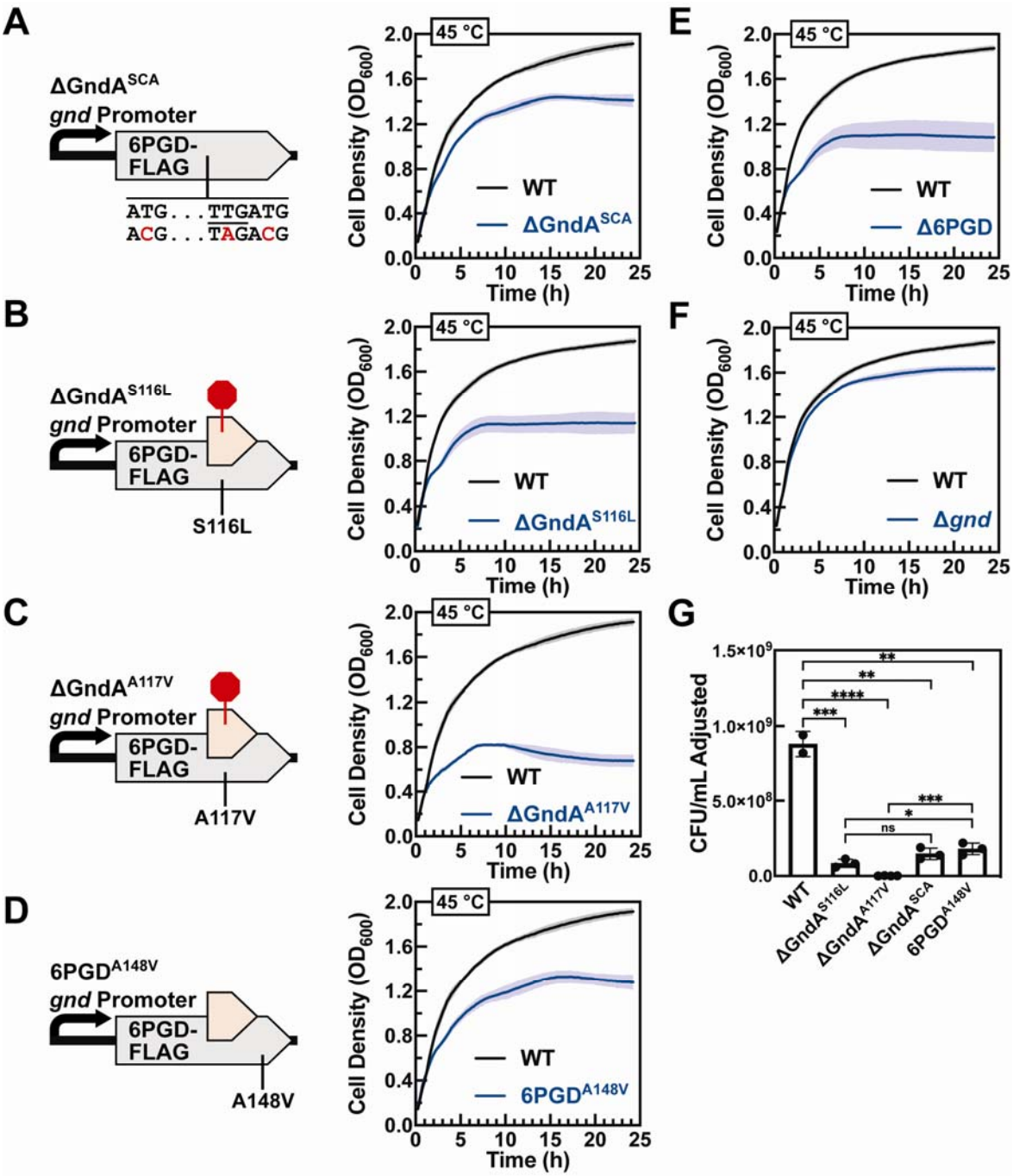

Fig. 2. GndA mutant *E. coli* exhibit a reduced growth phenotype and reduced colony forming units at elevated temperature. (*A*) Schematic of ΔGndA^{SCA} cell line genotype and growth curves (OD_{600}) for WT and ΔGndA^{SCA} cells at 45 °C. (*B*) Schematic of ΔGndA^{S116L} cell line genotype and growth curves (OD_{600}) for WT and ΔGndA^{S116L} cells at 45 °C. (*C*) Schematic of ΔGndA^{A117V} cell line genotype and growth curves (OD_{600}) for WT and ΔGndA^{A117V} cells at 45 °C. (*D*) Schematic of 6PGD^{A148V} cell line genotype and growth curves (OD_{600}) for WT and 6PGD^{A148V} cells at 45 °C. (*E*) Growth curves (OD_{600}) for WT and Δ6PGD cells at 45 °C. (*F*) Growth curves (OD_{600}) for WT and Δ*gnd* cells at 45 °C. Error (shading above and below curves) calculated from 4 replicates. (*G*) Colony forming units (CFU) for WT and *gnd* mutants after growth at 45 °C adjusted to OD_{600}. **** *p*-value < 0.00005.

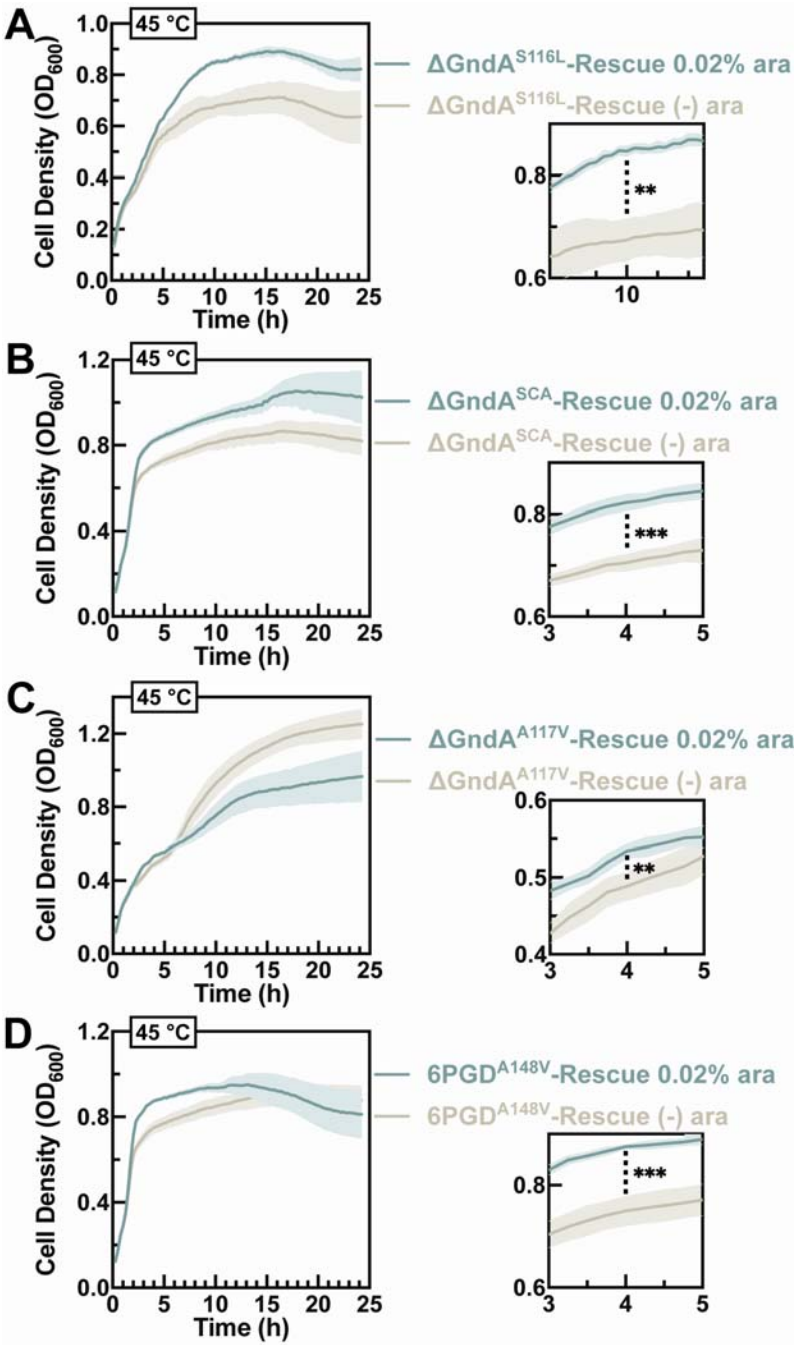

Fig. 3. GndA rescues growth at elevated temperature. (*A*) Growth curves comparing ΔGndA^{S116L} cells transformed with a pBAD33 expression vector encoding GndA (ΔGndA^{S116L}-Rescue) rescued with 0.02% arabinose at 45 °C. (*B*) Growth curves comparing ΔGndA^{SCA} cells transformed with a pBAD33 expression vector encoding GndA (ΔGndA^{SCA}-Rescue) rescued with 0.02% arabinose at 45 °C. (*C*) Growth curves comparing ΔGndA^{A117V} cells transformed with a pBAD33 expression vector encoding GndA (ΔGndA^{A117V}-Rescue) rescued with 0.02% arabinose at 45 °C. (*D*) Growth curves comparing 6PGD^{A148V} cells transformed with a pBAD33 expression vector encoding GndA (6PGD^{A148V}-Rescue) rescued with 0.02% arabinose at 45 °C. Error (shading above and below curves) calculated from 3-4 replicates. *** *p*-value < 0.0005.

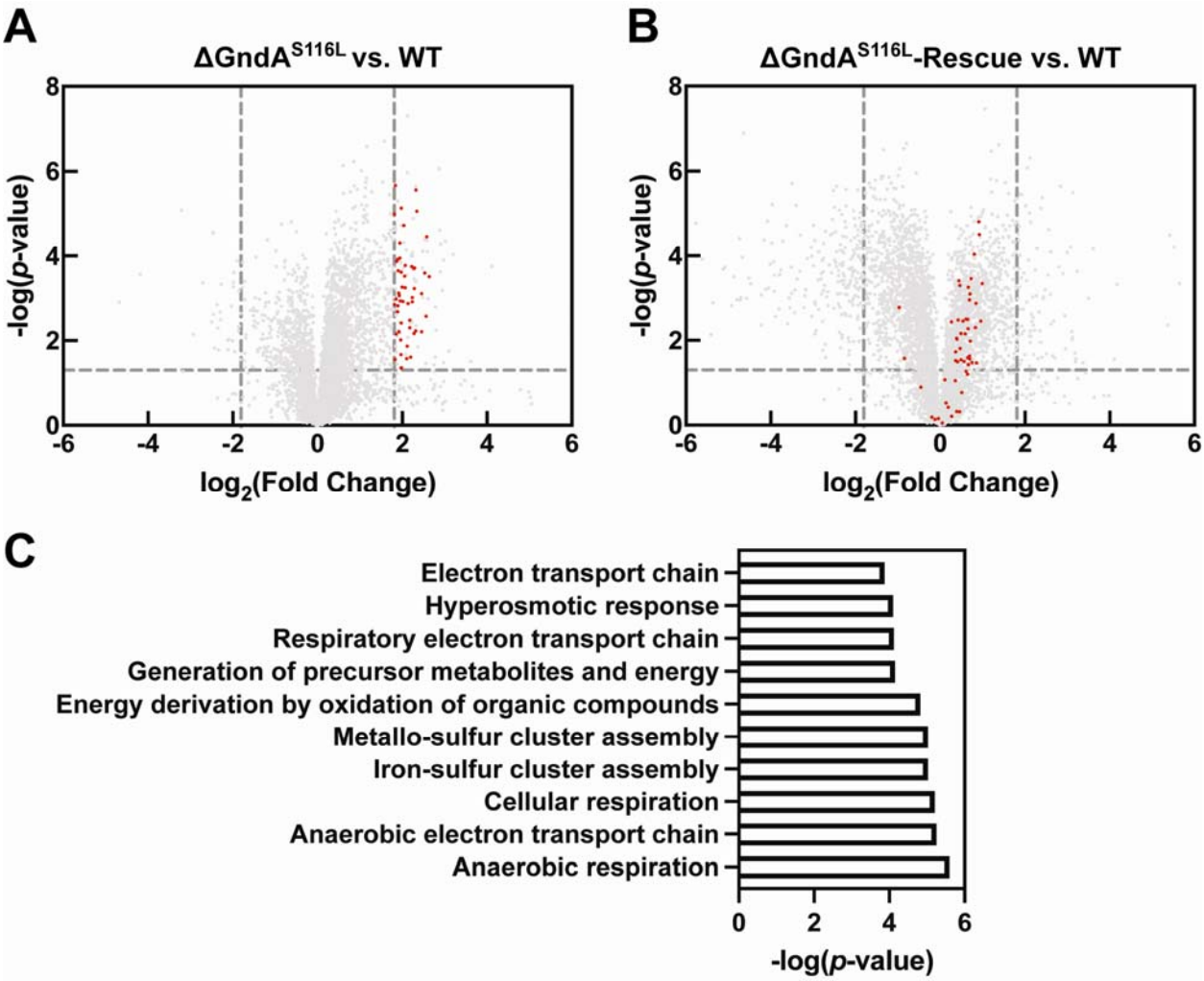

Fig. 4. GndA translation correlates with expression of electron transport chain genes. (*A*) RNA-seq analysis of gene expression changes in ΔGndA^{S116L} vs. WT and (*B*) ΔGndA^{S116L}-Rescue vs. WT. *P*-values were calculated by performing a *t*-test assuming a two-tailed distribution and homoscedasticity (n = 3). Horizontal dotted line signifies *p*-value of 0.05, and vertical dotted lines signifies a fold change of 3.5. One and twelve datapoints (including *ara* genes) are outside the axis limits of volcano plots *A* and *B* respectively. (*C*) GO analysis of genes of known function which are upregulated to transcripts per million (TPM) fold change greater than 3.5 for ΔGndA^{S116L} vs. WT after heat shock and are rescued to WT TPM levels in ΔGndA^{S116L}-Rescue cells.

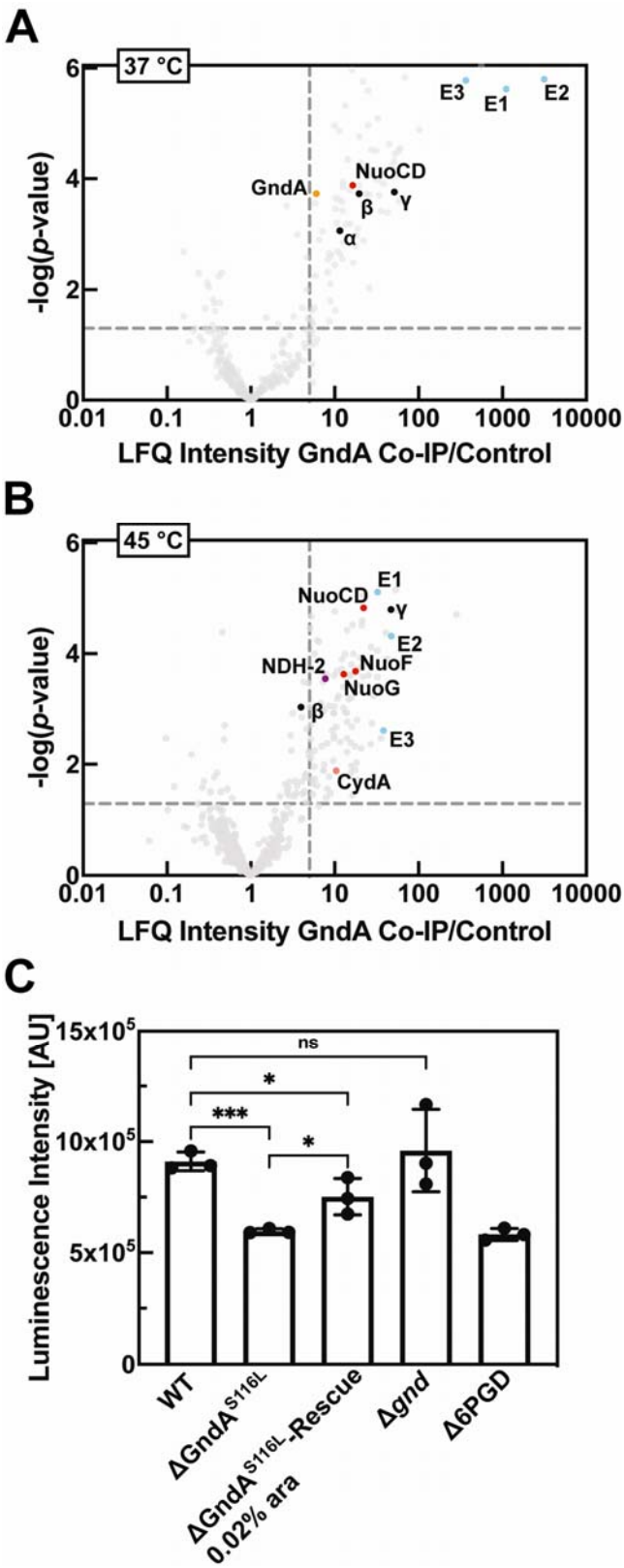

Fig. 5. GndA interaction with the electron transport chain and ATP synthase. (*A*-*B*) Volcano plot of label free quantitative proteomic analysis of endogenous SPA-GndA co-immunoprecipitation from GndA^{SPA} cell lysates with reference to a preincubated 3XFLAG peptide control grown at 37

°C and 45 °C respectively. *P*-values were calculated by performing a *t*-test assuming a two-tailed distribution and homoscedasticity (n = 4). Subunits of complexes associated with endogenous GndA color coded according to complex identity. Orange indicates endogenous GndA, red respiratory complex I (NDH-1), purple type II NADH:quinone oxidoreductase (NDH-2), salmon cytochrome *bd*-I oxidase, black ATP synthase, and cyan pyruvate dehydrogenase. (C) Relative intracellular ATP levels normalized to OD_{600} for WT and *gnd* mutants. *** *p*-value < 0.0005.

References

1. G. Storz, Y. I. Wolf, K. S. Ramamurthi, Small proteins can no longer be ignored. *Annu. Rev. Biochem.* **83**, 753–777 (2014).
2. M. W. Orr, Y. Mao, G. Storz, S. B. Qian, Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.* **48**, 1029–1042 (2020).
3. M. R. Hemm, J. Weaver, G. Storz, Escherichia coli Small Proteome . *EcoSal Plus* **9** (2020).
4. P. Garai, A. Blanc-Potard, Uncovering small membrane proteins in pathogenic bacteria: Regulatory functions and therapeutic potential. *Mol. Microbiol.* **114**, 710–720 (2020).
5. J. J. Mohsen, A. A. Martel, S. A. Slavoff, Microproteins—Discovery, structure, and function. *Proteomics* **23** (2023).
6. A. Khitun, T. J. Ness, S. A. Slavoff, Small open reading frames and cellular stress responses. *Mol. Omi.* **15**, 108–116 (2019).
7. B. Alix, D. A. Boubacar, M. Vladimir, T-REX: A web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* **40**, 573–579 (2012).
8. W. R. Pearson, An introduction to sequence similarity (“homology”) searching. *Curr. Protoc. Bioinforma.*, 1–8 (2013).
9. A. Waterhouse, *et al.*, SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
10. P. Puigbò, I. G. Bravo, S. Garcia-Vallve, CAIcal: A combined set of tools to assess codon usage adaptation. *Biol. Direct* **3**, 1–8 (2008).
11. S. A. Slavoff, *et al.*, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64 (2013).
12. A.-R. Carvunis, *et al.*, Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
13. J. Weaver, F. Mohammad, A. R. Buskirk, G. Storz, Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *MBio* **10**, 1–21 (2019).
14. S. Meydan, *et al.*, Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. *Mol. Cell* **74**, 481-493.e6 (2019).
15. F. Impens, *et al.*, N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes. *Nat. Microbiol.* **2**, 17005 (2017).
16. D. M. Anderson, *et al.*, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595–606 (2015).
17. M. R. Hemm, B. J. Paul, T. D. Schneider, G. Storz, K. E. Rudd, Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol. Microbiol.* **70**, 1487–1501 (2008).
18. A. Wacholder, *et al.*, A vast evolutionarily transient translatome contributes to phenotype and fitness. *Cell Syst.* **14**, 363-381.e8 (2023).
19. S. B. Parikh, C. Houghton, S. B. Van Oss, A. Wacholder, A. Carvunis, Origins, evolution, and physiological implications of de novo genes in yeast. *Yeast* **39**, 471–481 (2022).
20. M. R. Hemm, *et al.*, Small stress response proteins in Escherichia coli: Proteins missed by classical proteomic studies. *J. Bacteriol.* **192**, 46–58 (2010).
21. I. Fesenko, *et al.*, The hidden bacterial microproteome. *Mol. Cell* **85**, 1024-1041.e6 (2025).
22. A. H. Williams, *et al.*, The cryo-electron microscopy supramolecular structure of the bacterial stressosome unveils its mechanism of activation. *Nat. Commun.* **10**, 3005 (2019).
23. S. Zhang, *et al.*, Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat. Commun.* **11**, 1312 (2020).
24. C. Q. E. Lee, *et al.*, Coding and non-coding roles of MOCCI (C15ORF48) coordinate to

regulate host inflammation and immunity. *Nat. Commun.* **12**, 1–22 (2021).
25. C. Liang, *et al.*, Mitochondrial microproteins link metabolic cues to respiratory chain biogenesis. *Cell Rep.* **40**, 111204 (2022).
26. A. L. Rocha, *et al.*, An Inner Mitochondrial Membrane Microprotein from the SLC35A4 Upstream ORF Regulates Cellular Metabolism. *J. Mol. Biol.* **436**, 168559 (2024).
27. P. Yuan, N. G. D'Lima, S. A. Slavoff, Comparative Membrane Proteomics Reveals a Nonannotated E. coli Heat Shock Protein. *Biochemistry* **57**, 56–60 (2018).
28. Y. Y. Chen, *et al.*, Conformational changes associated with cofactor/substrate binding of 6-phosphogluconate dehydrogenase from Escherichia coli and Klebsiella pneumoniae: Implications for enzyme mechanism. *J. Struct. Biol.* **169**, 25–35 (2010).
29. X. Cao, *et al.*, Nascent alt-protein chemoproteomics reveals a pre-60S assembly checkpoint inhibitor. *Nat. Chem. Biol.* **18**, 643–651 (2022).
30. C. D'Souza, M. M. Nakano, P. Zuber, Identification of comS, a gene of the srfA operon that regulates the establishment of genetic competence in Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 9397–401 (1994).
31. L. W. Hamoen, H. Eshuis, J. Jongbloed, G. Venema, D. van Sinderen, A small gene, designated comS, located within the coding region of the fourth amino acid-activation domain of srfA, is required for competence development in Bacillus subtilis. *Mol. Microbiol.* **15**, 55–63 (1995).
32. S. Meydan, N. Vázquez-Laslop, A. S. Mankin, Genes within Genes in Bacterial Genomes. *Microbiol. Spectr.* **6** (2018).
33. N. Vakirlis, *et al.*, De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
34. E. C. Hobbs, X. Yin, B. J. Paul, J. L. Astarita, G. Storz, Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16696–16701 (2012).
35. D. Du, *et al.*, Structure of the AcrAB–TolC multidrug efflux pump. *Nature* **509**, 512–515 (2014).
36. J. Abramson, *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
37. A. Chugunova, *et al.*, LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. *Proc. Natl. Acad. Sci.* **116**, 4940–4945 (2019).
38. S. Zhang, *et al.*, LINC00116-encoded microprotein mitoregulin regulates fatty acid metabolism at the mitochondrial outer membrane. *iScience* **26**, 107558 (2023).
39. X. Cao, *et al.*, Alt-RPL36 downregulates the PI3K-AKT-mTOR signaling pathway by interacting with TMEM24. *Nat. Commun.* **12** (2021).
40. Z. Na, *et al.*, Mapping subcellular localizations of unannotated microproteins and alternative proteins with MicroID. *Mol. Cell* **82**, 2900-2911.e7 (2022).
41. Y. Chen, *et al.*, Unannotated microprotein EMBOW regulates the interactome and chromatin and mitotic functions of WDR5. *Cell Rep.* **42**, 113145 (2023).
42. T. F. Martinez, *et al.*, Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
43. S. Ghatak, Z. A. King, A. Sastry, B. O. Palsson, The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res.* **47**, 2446–2454 (2019).
44. M. Ashburner, *et al.*, Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
45. S. A. Aleksander, *et al.*, The Gene Ontology knowledgebase in 2023. *Genetics* **224**, 1–14 (2023).
46. C.-L. Sandmann, *et al.*, Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol. Cell* **83**,

994-1011.e18 (2023).
47. M. Zeghouf, *et al.*, Sequential Peptide Affinity (SPA) System for the Identification of Mammalian and Bacterial Protein Complexes. *J. Proteome Res.* **3**, 463–468 (2004).
48. F. Fontaine, R. T. Fuchs, G. Storz, Membrane localization of small proteins in Escherichia coli. *J. Biol. Chem.* **286**, 32464–32474 (2011).
49. U. Weidner, *et al.*, The Gene Locus of the Proton-translocating NADH□: Ubiquinone Oxidoreductase in Escherichia coli. *J. Mol. Biol.* **233**, 109–122 (1993).
50. T. Friedrich, The NADH:ubiquinone oxidoreductase (complex I) from Escherichia coli. *Biochim. Biophys. Acta - Bioenerg.* **1364**, 134–146 (1998).
51. J. M. Berrisford, C. J. Thompson, L. A. Sazanov, Chemical and NADH-induced, ROS-dependent, cross-linking between sublimits of complex I from Escherichia coli and Thermus thermophilus. *Biochemistry* **47**, 10262–10270 (2008).
52. R. G. Efremov, L. A. Sazanov, Structure of the membrane domain of respiratory complex I. *Nature* **476**, 414–421 (2011).
53. H. Erhardt, *et al.*, Organization of the Escherichia coli aerobic enzyme complexes of oxidative phosphorylation in dynamic domains within the cytoplasmic membrane. *Microbiologyopen* **3**, 316–326 (2014).
54. T. Friedrich, D. K. Dekovic, S. Burschel, Assembly of the Escherichia coli NADH:ubiquinone oxidoreductase (respiratory complex I). *Biochim. Biophys. Acta - Bioenerg.* **1857**, 214–223 (2016).
55. R. Feltens, M. Gossringer, D. K. Willkomm, H. Urlaub, R. K. Hartmann, An unusual mechanism of bacterial gene expression revealed for the RNase P protein of Thermus strains. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5724–9 (2003).
56. A. Pavesi, Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic Review. *Genes (Basel).* **12**, 809 (2021).
57. K. Nelson, R. K. Selander, Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (gnd) in enteric bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10227–10231 (1994).
58. M. Biseric̀, J. Y. Feutrier, P. R. Reeves, Nucleotide sequences of the gnd genes from nine natural isolates of Escherichia coli: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic gnd locus. *J. Bacteriol.* **173**, 3894–3900 (1991).
59. D. E. Dykhuizen, L. Green, Recombination in Escherichia coli and the definition of biological species. *J. Bacteriol.* **173**, 7257–7268 (1991).
60. P. I. Tarr, *et al.*, Acquisition of the rfb-gnd Cluster in Evolution of Escherichia coli O55 and O157. *J. Bacteriol.* **182**, 6183–6191 (2000).
61. J. R. Elder, M. Bugarel, H. C. den Bakker, G. H. Loneragan, K. K. Nightingale, Interrogation of single nucleotide polymorphisms in gnd provides a novel method for molecular serogrouping of clinically important Shiga toxin producing Escherichia coli (STEC) targeted by regulation in the United States, including the “big six” non-O157 . *J. Microbiol. Methods* **129**, 85–93 (2016).
62. A. L. Cookson, *et al.*, Culture independent analysis using gnd as a target gene to assess Escherichia coli diversity and community structure. *Sci. Rep.* **7**, 1–11 (2017).
63. S. Willis, J. Masel, Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes. *Genetics* **210**, 303–313 (2018).
64. P. Kolata, R. G. Efremov, Structure of Escherichia coli respiratory complex I reconstituted into lipid nanodiscs reveals an uncoupled conformation. *Elife* **10**, 1–32 (2021).
65. L. A. Sazanov, A giant molecular proton pump: structure and mechanism of respiratory complex I. *Nat. Rev. Mol. Cell Biol.* **16**, 375–388 (2015).
66. C. E. VanOrsdel, *et al.*, The Escherichia coli CydX Protein Is a Member of the CydAB Cytochrome bd Oxidase Complex and Is Required for Cytochrome bd Oxidase Activity. *J.*

Bacteriol. **195**, 3640–3650 (2013).
67. S. Safarian, *et al.*, Active site rearrangement and structural divergence in prokaryotic respiratory oxidases. *Science (80-.).* **366**, 100–104 (2019).
68. B. Bushnell, J. Rood, E. Singer, BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
69. A. Dobin, *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
70. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
71. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
72. F. Madeira, *et al.*, Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
73. J. A. Devito, Recombineering with tolC as a Selectable/Counter-selectable Marker: Remodeling the rRNA Operons of Escherichia coli. *Nucleic Acids Res.* **36**, 1–12 (2008).
74. C. J. Gregg, *et al.*, Rational optimization of tolC as a powerful dual selectable marker for genome engineering. *Nucleic Acids Res.* **42**, 4779–4790 (2014).
75. R. Lutz, Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
76. S. A. Schwartz, D. R. Helinski, Purification and Characterization of Colicin E1. *J. Biol. Chem.* **246**, 6318–6327 (1971).
77. A. Khitun, S. A. Slavoff, Proteomic Detection and Validation of Translated Small Open Reading Frames. *Curr. Protoc. Chem. Biol.* **11**, 1–27 (2019).
78. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A Human Short Open Reading Frame (sORF)-encoded Polypeptide That Stimulates DNA End Joining. *J. Biol. Chem.* **289**, 10950–10957 (2014).
79. Y. Perez-Riverol, *et al.*, The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* **53**, D543–D553 (2025).
80. R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–10 (2002).
81. P.-C. Yang, T. Mahmood, Western blot: Technique, theory, and trouble shooting. *N. Am. J. Med. Sci.* **4**, 429 (2012).
82. H. Schägger, Tricine-SDS-PAGE. *Nat. Protoc.* **1**, 16–22 (2006).