 doi:10.20944/preprints202107.0280.v1

Research Article

Comparative Genomic Analysis of the DUF34 Protein Family Suggests Role as a Metal Ion Chaperone or Insertase

Colbie Reed[1], Geoffrey Hutinet[1] and Valérie de Crécy-Lagard[1,2,*]

[1] Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611
[2] Genetics Institute, University of Florida, Gainesville, FL 32611
* Correspondence: vcrecy@ufl.edu

Abstract: Members of the DUF34 (domain of unknown function 34) family, also known as the NIF3 protein superfamily, are ubiquitous across superkingdoms. Proteins of this family have been widely annotated as "GTP cyclohydrolase I type 2" through electronic propagation based on one study. Here, the annotation status of this protein family was examined through comprehensive literature review and integrative bioinformatic analyses that revealed varied pleiotropic associations and phenotypes. This analysis combined with functional complementation studies strongly challenges the current annotation and suggests that DUF34 family members may serve as metal ion insertases, chaperones, or metallocofactor maturases. This general molecular function could explain how DUF34 subgroups participate in highly diversified pathways such as cell differentiation, metal ion homeostasis, pathogen virulence, redox and universal stress responses.

Keywords: comparative genomics; metabolic reconstruction; bioinformatics; conserved unknowns; function prediction; functional annotation; orthology

1. Introduction

Protein families that are both highly conserved across domains of life and poorly characterized are referred to as conserved unknowns [1,2]. Though recent studies that use comparative genomics [3,4], classical genetics [5] and/or biochemistry [6,7] approaches have solved a few of these "orphan" family puzzles, their number remains high [1,8–12]. One of the issues is that, because these conserved proteins often harbour core functional roles, genetic approaches lead to pleiotropic phenotypes, making the elucidation of a precise molecular function quite difficult. For example, the COG0533 and COG0009 proteins involved in the synthesis of the universal tRNA modification threonylcarbamoyladenosine (t^6A) [13–15], were first thought to be in involved in protein degradation [16,17], transcriptional regulation [18], or cell division [14]. Similarly, RidA (reactive intermediate deaminase A), a subgroup within the Rid family of proteins (members also have been referred to as YjgF/YER057c/UK114), was a notable challenge for functional characterization due to the multiple and complex phenotypes associated with mutations in genes of this family in different organisms [19–23].

The DUF34/NIF3 protein family is reportedly ubiquitous, with members found in model organisms such as, *Homo sapiens* (NIF3L1), *Mus musculus* (Nif3l1), *Saccharomyces cerevisiae* (Ngg1-interacting Factor 3/NIF3) [24,25], *Escherichia coli* (YbgI) [26] and *Bacillus cereus* (YqfO) [27]. Despite its conservation, the precise function(s) of members of this family remain undetermined. More than a decade has passed since the family was first formally identified as a target for characterization [24] and even longer since the gene

 doi:10.20944/preprints202107.0280.v1

encoding a homolog of NIF3 in *S. cerevisiae* was first described in *Drosophila melanogaster* [28,29]. Since, it has been linked to a variety of functions across superkingdoms and to several diseases in humans (e.g., juvenile amyotrophic lateral sclerosis, Williams-Beuren Syndrome [30,31], among many others). The role of this protein family remains mysterious, even with recent studies trying to more proximately decifer its function in *E. coli* [32]. Automated annotation databases indicate that the human DUF34 family member, NIF3L1, is highly connected, for example listing 4,178 functional associations for its entry in the Harmonizome database (i.e., 65 datasets, electronically extracted; https://amp.pharm.mssm.edu/Harmonizome/gene/NIF3L1 [33]; accessed June 2021). In addition, an annotation based on a single set of *in vitro* results examining the NIF3 homolog of *Helicobacter pylori* (HP0959) [34] led to the swift percolation of the annotation, "GTP cyclohydrolase I type 2 homolog", throughout many databases, including UniProtKB. This annotation as the first enzyme of tetrahydrofolate biosynthesis is certainly incorrect for the whole protein family, as DUF34 members are found in folate auxotrophs such as *Mycoplasma* [35–37].

A comprehensive analysis of the literature was conducted to catalogue all published knowledge for DUF34 family members, an endeavour that cannot be easily conducted using only simple PubMed searches, as many studies do not mention general family names of genes/proteins for which data has been generated, often only citing species- or system-specific gene names. In parallel, an extensive comparative genomic analysis was performed to investigate the validity of "GTP cyclohydrolase I type 2", a dubious annotation widespread among DUF34 family members, and to ultimately propose a unifying functional role for the family as a metal insertase. With this, it was possible to divide the DUF34 protein family into subgroups by distinctions in structure, complete domain architecture, regulation, occurrence, localization, and functional associations.

2. Materials and Methods

2.1. Capture of literature, structural, and essentiality data.

The strategy used to compile published literature for members of the DUF34 family is detailed in the Supplemental Methods and all websites used, both here and in subsequent analyses, are listed Supplemental Table S1. Most of the public search engines/web crawlers, and searchable libraries/depositories used required text as input while more specialized tools leveraged the use of protein sequences (e.g., PaperBLAST [38]). Protein Data Bank (PDB; RCSB PDB, Research Collaboratory for Structural Bioinformatics PDB)) was used to evaluate and acquire protein crystal structures and respective sequences, related literature, and relevant data files for subsequent search and analysis [38–40]. Crystallization conditions were not considered in these analyses. Structures were edited, aligned using PyMol (Edu PyMol, Educational edition) [41]. MetalPDB was used to survey ions present, indicated or predicted to complex with published protein crystal structures [42,43].

Essentiality data was acquired using multiple different sources listed in Table S1. The BLAST search tool of DEG (Database of Essential Genes) [44] was used, with *H.*

 doi:10.20944/preprints202107.0280.v1

sapiens (NIF3L_HUMAN, Q9GZT8), *Methanocaldococcus jannaschii* (GCH1L_METJA, Q58337), *B. cereus* (Q818H0_BACCR, Q818H0) and *E. coli* (GCH1L_ECOLI, P0AFP6) as inputs. Ogee [45] was used to collect additional essentiality data through the browse function. Predicted essentiality data for *Mycoplasma* species were acquired using pDEG (Database of Predicted Essential Genes) [46].

2.2. Domain analysis.

A first set of sequences of DUF34 family members from model organisms was extracted using OrthoInspector 3.0 (accessed January 30, 2020) [47] using the following input sequences for retrieving sets of sequences per superkingdom: NIF3L_HUMAN (Q9GZT8), GCH1L_METJA (Q58337), and GCH1L_ECOLI (P0AFP6). An additional set of sequences from organisms with published data was extracted from UniProtKB [48] to generate a non-redundant list of 219 sequences to be used in subsequent analyses. The sequences of the corresponding DUF34 proteins were not available for a few organisms with which publications were associated. For *Desulfovibrio desulfuricans*, sequences of the closely related *Desulfovibrio alaskensis G20* were used, and those of *Schistosoma mansoni* were used in the place of *Schistosoma mekongi.* Although described in their respective publications, sequences for DUF34 family members could not be retrieved for three organisms: *Idiosepius paradoxus, Streptomyces sp. SN-1061M, Verrucomicrobium (Termite Associated, TAV) sp. strain 2*. Sequences were aligned using MAFFT (E-INS-i, default settings) [49–51]. Motif and domain logos were generated through use of the WebLogo webserver [52]. Sequence logos were manually aligned using Inkscape [53].

2.4. Absence-Presence, Phyletic Patterns & Homolog/Paralog Co-occurrence.

Species trees were generated with PhyloT (database version 2020.2) and iToL [54] using the organisms collected into a single list that is referenced in Results subsection 3.3. Absence-presence data was acquired, both, through manual curation using advanced searches of common databases (i.e., UniProt, NCBI [55]), subsequent BLAST validation, as well as the use of phyletic patterning tools available through MicrobesOnline (accessed May 2020) [56] and STRING (v11, released January 19, 2019) [57]. Paralogs were identified using EggNOG (EggNOG 5.0) [58] and KEGG Paralog Search (KEGG release 94.1) [59].

2.5. Physical Clustering Analysis.

Physical clustering data was acquired from Gene Context Tool NG (GeConT 3) of the Computational Genomic Group, IBT–UNAM, using the central orthologous group ID known for the DUF34 family, COG0327 (accessed Sept. 2020) [60] and analyzed using a text-mining startegy we developed and termed Physical Clustering Keyword Frequency Analysis (PCKFA). This approach is described in detail in the Supplemental Methods (1.2). Some subsets of families identified using PCKFA were further annotated for use in tables/figures; this was completed using a combination of bioinformatic queries to extract information regarding any key motifs, domains, and/or annotation data linking EC numbers or ligands to the observed COGs, as well as individual proteins (described further in Supplemental Methods).

 doi:10.20944/preprints202107.0280.v1

2.6. Coexpression & gene set enrichment analysis.

Lists of 300 genes coexpressed with DUF34 family members were retrieved for all 10 eukaryotic model organisms available using CoXPresDb (gene sets excluded respective DUF34 homlogs) [61], with the exception of *Caenorhabditus elegans*, as it was discovered that the genome of which does not encode for a DUF34 family member. Organisms for which co-expressed gene sets were retrieved were as follows: *H. sapiens*, *M. musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *D. melanogaster*, *Macaca mulatta*, *Canis lupus familiaris*, *S. cerevisiae*, and *S. pombe*. Protein coregulatory data for *Homo sapiens* was acquired using the ProteomeHD webserver (unsupervised query format) [62]; a threshold of 0.98 was used for co-regulatory data retrieval for NIF3L1 (specific protein reference ID within database: Q9GZT8-2, resulting in 114 total coregulated proteins). Gene set enrichment analyses (GSEA), was performed using two tools: g:GOSt (via g:Profiler webserver) [63], and the functional annotation clustering tool (via DAVID bioinformatic suite) [64–66]. UniProtKB was used to map UniProt IDs to the Entrez Gene IDs of eukaryotic datasets prior to GSEA. If electronic mapping failed for a human identifier, HGNC database was used in manual retrieval (HUGO Gene Nomenclature Committee at the European Bioinformatics Institute [67]). If mapping failed for other organisms for which co-expression or co-regulatory data were retrieved, the "reviewed" entries in UniProtKB were selected over the "unreviewed" duplicates and/or isoforms listed.

2.7. Fusion Analysis.

To analyze fusions present in the family, the protein family as defined by UniProt (e.g., "GTP cyclohydrolase I type 2/NIF3 family") were exported and filtered for all sequences containing InterPro HMM profile signature annotations distinct from those already recognized in Results section 3.5. To optimize coverage of all documented fusions, a second and third approach for curating such homologs were implemented in parallel to the UniProt-dependent approach. For these two complimentary methods, sequences of various domain architectures were directly exported from Pfam (PF01784) and InterPro (IPR036069), independently. Three lists of homologs generated by each method were concatenated and duplicate sequences removed. Fusions identified via the preceding literature review were added, defining the final collection of "noncanonical" homologs. All fusion/arrangement types were further evaluated for legitimacy through manual curation (i.e., comparative annotation review of genome and sequence features) and the assignment of confidence scores: "valid" (highest confidence); "valid, conditional"; "conditional"/"conditional, singleton"; "inconclusive"; "invalid" (lowest confidence, no validity). To ensure results of fusion analyses were comparable to those of other bioinformatics presented, singularly representative COGs and COG descriptions were assigned to the final list of exceptional homologs using CDD Search, subsequently cross-referencing results with EggNOG records for optimal domain descriptions. For more information on data transformation, amendment, and clean-up, see Supplemental Methods (1.3).

 doi:10.20944/preprints202107.0280.v1

2.8. Strain Construction & List.

All strains and oligunecleotides used in this study are listed in Table S2. Two genes of *E. coli*, *ybgI* (encoding for DUF34) and *folE* (encoding for GTP cyclohydrolase I type 1), were cloned independently in pBAD24 between NcoI and SbfI following PCR amplification by Phusion® High-Fidelity DNA Polymerase (New England Biolabs, NEB) using GO285 and GO286 oligonucleotides for *ybgI*, while GO434 and GO435 were used for *folE*. After verification by sequenceing, the plasmids generated were renamed "pGH50" and "pGH101", respectively.

The *ybgI* kanamycin cassette replacement *E. coli* mutants were collected from the Keio Collection [68]. The *folE* kanamycin cassette replacement *E. coli* mutants were collected from a previous study [69]. These mutations were transduced by P1vir into *E. coli K-12 MG1655*. The *ybgI* and *folE* double mutant was obtained by first flipping out the kanamycin cassette from the *ybgI* mutant using pCP20 [70], subsequently tranducing the *folE* mutation using P1vir. Mutation verifications were performed by oneTaq PCR (NEB) using a set of primers internal and external to the gene (GO563 to GO570). Each plasmid, including empty pBAD24, were individually transformed into the control strain and each mutant. Strains were grown at 37 °C using LB supplemented with glucose 0.2%, kanamycin sulfate 50 µg/mL or ampicillin 100 µg/mL when necessary for selection. 2'-deoxythymidine (dT) 0.3 mM was used for *folE* mutants.

2.9. dT Sensitivity Assay.

Strains (WT, single mutants, and double mutants) were grown overnight at 37 °C in LB supplemented with glucose 0.2%, kanamycin sulfate 50 µg/mL (except for WT) and dT 0.3 mM. Each strain was inoculated in various LB with or without dT 0.3 mM at an OD_{600nm} of 0.1, and grown at 37 °C in a bioscreeen (Oy Growth Curves Ab Ltd, Finland) for 40 hours. This experiment was completed in quintuplicate.

2.10. dT Essentiality Complementation Assay.

Strains containing pBAD24 variations were grown overnight at 37 °C in LB supplemented with glucose 0.2%, ampicillin 100 µg/mL and dT 0.3 mM. They were then normalized to an OD_{600nm} of 1.0 in LB, and a 5 µL drop was streaked on LB agar containing ampicillin 100 µg/mL, either glucose or arabinose at 0.2%, and either with or without dT 0.3 mM. These plates were left to grow for 10 hours at 37 °C. This experiment was performed in triplicate.

3. Results and Discussion

3.1. Extensive literature capture and analysis confirms pleiotropic role of DUF34 family members.

While the earliest mention of the family dates back to 1996 when the binding of a yeast homolog to NGG1/ADA3 via a GAL4 fusion domain was noted [71], the first dedicated description of a DUF34 family member was published in 2000 with the isolation and characterization of the human NIF3L1 and its mouse homolog [30]. Only seven papers in PubMed cite the latter study (per June 6, 2021) and 20 mostly unrelated publications cite the former (as of June 6, 2021; studies focused mostly on NGG1/ADA3

 doi:10.20944/preprints202107.0280.v1

or SAGA complex, only 6 demonstrating relevance to DUF34). PaperBLAST, an enginuitive sequence-based literature search tool, searches titles, abstracts, and full publication texts available through Europe PMC [72]. Although it has proven enormously helpful in the quest for published homologs, use of this tool, alone, was found to have some understandable weaknesses leading to oversight or mistakes in the curation of DUF34-relevant publications (e.g., journal paywalls, supplemental data unsearchability, and the reliance upon text-based searches in tandem with sequence-based determination of search terms). We, therefore, expanded our search method by integrating several strategies in tandem with the implementation of homolog keywords and sequences cyclicly with their accumulation (full approach description in Supplemental Methods section 1.1). Ideally, this would allow for the optimized use of each tool, increasing coverage that would be supported and checked through cross-validation. A final collection of sequences and keywords used for sequence-/text-based searches can be found in Data Table 1. The resulting list of curated publications was divided into two groups: "focal" (i.e., homolog mentioned in title or abstract; Table 1) and "non-focal'' (i.e., mention occurs in other publication sections or in supplemental/attached files). The complete collection of focal/non-focal publications is reported in Data Table 2. All individual DUF34 family members with publications are listed in Table S3. Using this integrative search approach, the ultimate total of reference terms reached upwards of 857 and provided DUF34 member-relevant data for ~100 unique organisms. This process increased the total number of DUF34 protein family-relevant papers from <30[1] to 333 distinct publications.

Although the captured data covered all superkingdoms, the distribution of publication counts skewed largely toward bacteria, this domain having the greatest number of "non-focal" publications and, thereby, total publications overall. In contrast, work examining eukaryotic systems contributed the greatest proportion of "focal" publications. Only one "non-focal" publication featured a viral homolog. No publications were found to describe DUF34 family members for any species of plant (*Viridiplantae*), consistent with the absence of DUF34 homologs among annotated plant genomes discussed below.

To discern whether any common functional associations could be extracted from the final DUF34 corpus, word clouds were generated using publication titles of both focal and non-focal publications (Data Table 2, Figure S1). The resulting diagrams predominantly emphasized the systems of study (e.g., "Mycobacterium", "Escherichia", "Bacillus", "yeast") and terms relating to the characterization process (e.g., "reveal", "novel", "analysis", "functional", "identifies", "associated"), both of which observations provided little insight into specific function. However, other less pronounced keywords were indicative of more specific biological contexts, such as "mitochondrial", "DNA repair", "DNA methylation", "[fe]-hydrogenase cofactor

[1] Simple PubMed search using the following query: "DUF34" OR "NIF3" OR "NIF3L1" OR "YbgI" OR "YqfO". Confirmed results again on June 17, 2021 with a total of 25 papers retrieved.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

biosynthesis", "stress", "virulence", "heat", "resistance", and "secreted", for example. Together, these diagrams illustrated that, of the surveyed literature, themes of bacterial pathogen virulence, gene regulation, cell signalling pathways, stress response, as well as metal ion metabolism and related membrane homeostasis, seemed to be emphasized.

Across published data, differences in the localization of DUF34 proteins are reported with no clear consensus. In fungi, for example, family members have been linked to mitochondria (e.g., P53081, *Saccharomyces cerevisiae*), while in model vertebrates (Q9GZT8, *Homo sapiens*; Q9EQ80, *Mus musculus*) and select yeasts (*S. cerevisiae* [73]) homologs have been shown translocating between the nucleus and cytosol, the process of which, in some cases, appeared to be regulated by retinoic acid (Q09GP9, *Bombyx mori* [74]). Although understood as being predominantly cytoplasmic in bacteria, truncated DUF34 homologs are secreted in *Pseudomonas* species as a proposed nematocidal agent [75]. In another case, homologs have been observed to occur at the cellular poles of *E. coli*, co-localizing with PstB (phosphate transporter subunit, ATP-binding) and TktA (transketolase) [32].

Historically, associations of NIF3L1 with human disease have driven much of the impetus for research into this DUF34 homolog [30,31,76,77]. Such links to human disease have been particularly reinforced by many non-focal publications (Table S3; Data Table 2). Indeed, expression of DUF34 in eukaryotes has been associated with several human pathologies, including cancers [78–94], chemotherapeutic drug response [95,96], psychiatric disorders [97,98], cardiovascular disease [99–101], insulin resistance [102], osteoporosis [76,103], inflammation [104], Amyotrophic Lateral Sclerosis (ALS) [30,105], William-Beuren Syndrome [31], as well as several other degenerative and developmental neurological diseases [77,106,107]. The regulation of DUF34 homologs by retinoic acid or biochemical relatives (e.g., all-trans retinoic acid, ATRA; testosterone [Comparative Toxicogenomics Database]) appears to be conserved between humans, mice and select life stages of some insects [74,108–110]. Associations to cell differentiation through gene regulation were also numerous [74,107–109,111–114].

Links to virulence and environmental stress responses dominated the studies of bacterial and fungal DUF34 homologs [32,75,115–128]. In addition, links to regulation of central carbon metabolism were made in *Geobacillus stearothermophilus* [129] and *Bacillus subtilis* [130]. Although ssDNA- and dsDNA-binding properties *in vitro* were observed for at least one archaeal homolog [131], only ssDNA-binding activity has been reported in bacteria [132], observations of which later came under scrutiny in the context of UV-induced DNA damage responses in *E. coli* [32].

In this comprehensive review of the literature for members of the DUF34 family, observations and functional associations were highly pleiotropic and could be the result of many indirect effects. The only precise molecular function proposed with compelling biochemical evidence is the role as a metal ion insertase in metallocofactor biogenesis described for the homologs of *Methanocaldococcus jannaschii* [133] and *Methanococcus maripaludis* [134].

Table 1. Focal publications featuring members of the DUF34 protein family.

Name	Organisms	Phenotype, Biological Relevance	Reference

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

YqfO/ BC_4286	*Bacillus cereus*	Inserted domain similar to PII-like/CutA1 family proteins; present in select bacterial clades; domain may regulate catalytic activity	[135]
YqfO/ BSU_25170	*Bacillus subtilis subsp. subtilis str. 168*	With YlxR, coregulates *tsaEBD* (t^6A synthesis [62]); disruption impairs *tsaEDB* regulation, loss of glucose-induction of *sigX* via PDHc expression dysregulation	[130]
BmNIF3l	*Bombyx mori*	Translocates to nucleus from cytoplasm upon ATRA tx; higher transcript levels in differentiating tissues; no expression detected in egg stage	[43]
YbgI/b0710	*Escherichia coli*	Structure, homohexameric toroid; monomers possess dinuclear metal ion-binding site; putatively involved in DNA repair	[26]
		No survival impairment upon mutant UV tx; polar localization during cell division (co-localized with PstB, TktA); GlmS putative interaction partner; mutant sensitive to antibiotics affecting cell wall synthesis	[32]
XynX	*Geobacillus stearothermophilus*	Negatively regulates expression of *xynA* (encodes a secreted xylanase); may be negatively regulated by *xylR*	[129]
NIF3L1/ ALS2CR1/ CALS-7/ MDS015/ My018	*Homo sapiens*	Ubiquitously expressed during embryonic development; strong over-expression in spermatogonia-derived, teratocarcinoma cell lines; Isolated, characterized; cytosolic subcellular localization; highly conserved N-, C-terminal regions; shares inserted region of its murine homolog (CutA1-like)	[24]
		NIF3L1 interacts with splice variant, NIF3L1 BP1 (THOC7), cytosolic colocalization; C-terminal leucine zipper-like domain of variant mediates interaction; not indicated in repression in NIH3T3 cells; binding partner, NIF3L1 BP1, demonstrates additional passive presence in the nucleus	[25]
		Retinoic acid-induced binding, cooperative translocation with Trip15/CSN2 from cytosol to nucleus (early neuronal development, silences differentiation suppressor Oct-3/4); ubiquitous expression, important in neuronal development	[108]
		Detected in brain, spinal cord and lymphocytes; observed as two distinct transcripts with similar patterns of expression; highest levels of both transcripts in heart, skeletal muscle, testis; smaller transcript was expressed at higher level than the other; no deletions, polymorphisms linked to ALS patients relative to controls; 1 of 6 candidates eliminated for causative link to ALS2	[30]
		1 of 4 hypermethylated, significant differential expression shared between two cancellous bone specimen groups: osteoarthritis, osteoporosis	[76]
		With 14-3-3, co-regulates transcriptional of Wbscr14 by preventing its nuclear localization via complex formation (Wbscr14 participates in complex-mediated transcription of lipogenic enzymes, promoting fat accumulation)	[31]
		Included in 7.5-Mb interstitial deletion on 2q32.3-33.1 (28 genes) in patient diagnosed with SATB2-Associated 2q32-q33 microdeletion syndrome	[66]
		Significantly associated with triptolide chemosensitivity in lymphoblast cell lines	[136]
		COPS2 point mutations consistent with previously defined NIF3L1-COPS2 co-repression interaction model (limited; pathogenesis associated COPS2 mutations: S120C, N144S, Y159H, R173C)	[137]
HP0959	*Helicobacter pylori*	GTP-binding, hydrolysis *in vitro*, biologically irrelevant pH, temperature	[34]
HcgD/MJ0927	*Methanocaldococcus jannaschii*	Proposed iron chaperone required for FeGP cofactor biosynthesis	[133]
		Homohexameric via 2 interfaced homotrimeric units; binds to ssDNA/dsDNA	[131,138]
Nif3l1/ 1110030G24Rik	*Mus musculus*	Isolated, characterized; ubiquitous expression across tissues; cytosolic localization; highly conserved N-, C-terminal regions; shares inserted region of human homolog	[24]
		Retinoic acid-induced binding, cooperative translocation with Trip15/CSN2 from cytosol to nucleus (early neuronal development, results in silence of differentiation suppressor Oct-3/4); ubiquitous tissue expression, important in neuronal development	[108]
WP_046236688 WP_032702676 PP_1038 VT47_06255 WP_017124074 WP_054077596	*Pseudomonas sp.*	("YqfO03") small, secreted protein; demonstrated high potency as nematicide against *C. elegans*, *M. incognita*; free-standing YqfO domain-containing protein (no NIF3/DUF34 domains) is member of the NIF3 protein family	[75]
Nif3/	*Saccharomyces cerevisiae*	Determined to have dual/multiple localizations (cytosolic, mitochondrial)	[73]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

YGL221C			
SA1388	*Staphylococcus aureus*	Central domain of NIF3 homolog has high structural similarity to CutA1 (family linked to cation tolerance, homeostasis)	[139]
SP1609	*Streptococcus pneumoniae*	Described as a member of same orthologous group (COG2384) as TrmK, RpoD protein families via structural alignment (*incorrect**)	[140]
TTHA1606	*Thermus thermophilus HB8*	Binds to ssDNA (very weakly, *in vitro*)	[132]
NIF3-like protein superfamily	NA	(electronic translation) describes family members of model organisms (Eukaryota, Bacteria), structures published prior to 2007	[141]

3.2. Conservation of metal binding site but variability of metal identity across DUF34 structures.

To complement the literature search, PDB was queried using select DUF34 sequences (YqfO, *B. subtilis*, P54472; NIF3L1, *H. sapiens*, Q9GZT8; YbgI, *E. coli*, P0AFP6; MJ0927, *M. jannaschii*, Q58337) as input. These initial queries returned 15 unique structure entries of DUF34 proteins from six different organisms (5 bacteria, 1 archaea) (Table 2). Text-based queries of PDB were also performed using "NIF3", yielding a total of 27 structures, of which only 16 were discernible members of the DUF34 family. False positives returned as a result of the text-based PDB query were all structures belonging to the Carboxy-terminal domain RNA polymerase II polypeptide A small phosphatase 1 (CTDSP1) homolog of *H. sapiens*, of which was found to have the former alias of "NIF3", an acronym for "NLI-interacting factor 3" (Q9GZU7). In total, after cross-checking this list of 16 PDB-derived structures with those retrievable via simple literature search, 16 individual protein structures were verified and were found to represent two superkingdoms and, within these, seven distinct organisms (eight structures respectively from each, bacteria and archaea).

DUF34 monomers form a homohexameric quaternary structure assembled through the trimerization of homodimers in a "head-to-tail", tessellating fashion. This homohexameric toroid is conserved across published structures with the central opening averaging a diameter of 31 Å (range: 24-38 Å). In some cases, this toroid is modified by the addition of trimeric "lids" to each side of the central opening, creating a cage-like structure; the monomeric structural features constituting these "lids" are the inserted P_{II}-like domains observed in the DUF34 family members belonging to select bacterial clades, fungi, and vertebrates [135]. These inserted domains forming these trimeric "lids" have been described as highly flexible, affecting the resolution of the corresponding architecture [135,139].

Table 2. Published structures of DUF34 protein family members.

Name	Organisms	Ligands	P_{II} domain	PDB	Phenotype	Reference
YbgI	*Escherichia coli*	(2)Fe^{3+}	No	**1NMO**	NA	[26]
		(2)Mg^{2+}	No	**1NMP**		
HcgD/ MJ0927	*Methanocaldococcus jannaschii*	(1)Cl^-, (2)Fe^{3+}	No	**3WSD**	Weaker Fe1 site under oxidized conditions *in vitro*	[133]
		(2)Fe^{2+}, (1)PO_4^{3-}	No	**3WSE**		
		(1)Fe^{3+}, (1)citrate	No	**3WSF**		
		(1)Fe^{2+}, (1)citrate	No	**3WSG**		
		(1)Fe^{3+}, (1)SO_4^{2-}	No	**3WSH**		
		(1)Fe^{2+}, (1)PO_4^{3-}	No	**3WSI**		
		NA	No	**4IWG**	Binds to ssDNA, dsDNA *in vitro*	[131,138]
		NA	No	**4IWM**		
SA1388	*Staphylococcus*	(2)Zn^{2+}, (1)B3P	Yes	**3LNL**	Cavity diameter= 38Å;	[139]

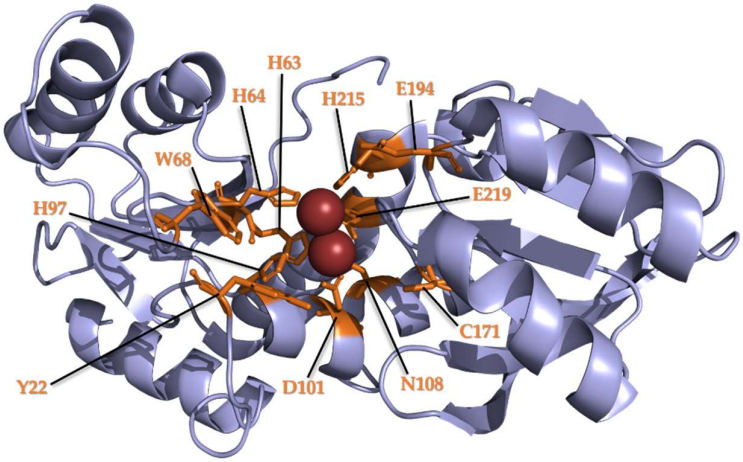
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

	aureus	(2)Zn^{2+}	Yes	**2NYD**	opening edge length= 20Å (triangular opening)	
SP1609	*Streptococcus pneumoniae*	NA	No	**2FYW**	NA	*PDB only*
TTHA1606	*Thermus thermophilus*	NA	No	**2YYB**	Binds ssDNA not dsDNA *in vitro*	[132]
Sthe_0840	*Sphaerobacter thermophilus*	(7)Cl^{-}*, (14)FMT*, (1)ACT*	No	**3RXY**	NA	*PDB only*
YqfO	*Bacillus cereus*	(2)Zn^{2+}, (1)HEPES, (1)TRS	Yes	**2GX8**	NA	[135]

Asterisk indicates that ion count is per the respective asymmetrical unit as opposed to per monomer.

A dinuclear metal-binding active site predicted to be catalytic, not structural [26] is highly conserved across available structures of DUF34 family proteins (Table 2). This active site structure is defined by a central cleft per monomer within which two divalent metal ions bind [26]. The nature of these divalent metal ions varies: from iron found in both bacterial and archaeal homologs [26,133] to zinc found in bacterial homologs containing the additional P_{II}-like domain (i.e., SA1388 of *Staphylococcus aureus*; YqfO of *Bacillus cereus*) [135,139]. This difference in metal ion-binding does not appear to be attributable to the additional domain as the topology of the active site has been described as remaining entirely undisturbed, or "identical", between homologs with and without the distinct domain architecture [135,139].

Figure 1. Dinuclear metal-binding site of E. coli DUF34 homolog, YbgI.



Crystal structure of YbgI (DUF34 homolog, E. coli) illustrating conserved residues of the protein family specific to the monomeric cleft of the active site and its dinuclear metal center. Highly conserved residues noted by Ladner et al. [26] to demonstrate involvement in the structure of the binding pocket are distinctively colorized, annotated (orange; residue identity and location labeled accordingly).

The metal ion-binding sites found in bacterial DUF34 structures contain seven highly conserved residues: five histidines, one glutamate, one aspartate [26,139] (Figure 1). These seven residues are conserved in both YbgI and YqfO forms, the latter possessing the additional, central "YqfO-like" domain [135]. The localization of the active sites within the inside of the toroid's central channel is ubiquitous, however, solvent-accessibility of this space differs between the two types of quaternary structure, the "cage-like" prolate spheroid with trimeric "lids" demonstrating greater restriction of access to active sites [132,135]. It should be noted that one outlier publication regarding the archaeal DUF34 family member, MJ0927 of *M. jannaschii* (4IWG, 4IWM), appears to differ greatly from all other descriptions of quaternary structure for this family [131,138], even contradicting several structures published for the same homolog (3WSD, 3WSE, 3WSF, 3WSG, 3WSH, 3WSI), of which even go as far as to resolve the active site in different states of oxidation [133]. This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

anomalous structure is described as a homohexameric spheroid with three openings (~33Å in diameter), instead of the single, central opening of the toroid conserved in all other published structures of the DUF34 family.

3.3. Family wide and superkingdom-specific signature motifs.

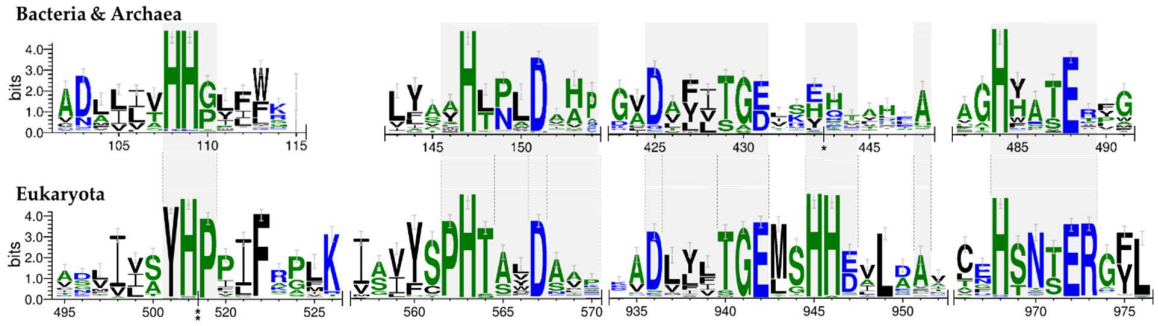
The NIF3/DUF34 family is large, containing 6,804 member sequences in Pfam (Pfam release 32.0), and its members span all kingdoms of life. Previous studies have already shown that proteins of this family can have different domain architectures [26,131,132,135,139] but no systematic, comparative analysis of the architectural distinctions had ever been performed across all superkingdoms. We, therefore, set out to classify the proteins of the DUF34 family into different subtypes based on the domain arrangements and the presence-absence of specific sequence motifs. Because several DUF34 protein structures were available (Table 2), these were used to guide alignment choices and to ultimately map conserved residues.

To resolve subtypes within the DUF34 family, multiple sequence alignments were initially performed inclusive of members across all superkingdoms. Ortholog sequences were extracted from OrthoInspector for each superkingdom (Data Table 3), and structure-based alignments were generated for each group using the MultAlin and ESPript webservers (Figure S2) [142,143]. The motifs were divided into three groups, or "tiers", based on their degree of cross-superkingdom conservation. Four motifs were found to be conserved across all three superkingdoms (logos with distinct tiers for all three superkingdoms are shown in Figure S3). These conserved residues of tier 1 were all integral to the metal binding pocket and are the residues described in Figure 2.

The most notable difference in the more highly conserved motifs was within the dual-histidine motif of the N-terminal region (Figure 2). In eukaryotes, the first histidine residue is replaced by a tyrosine, which may alter the dimensions of the binding pocket (Figure 1). Another notable distinction in eukaryotes is the second histidine pair ((M/L)xHH) located after the C-terminal "Dxxx(T/S)G(E/D)" motif (Figure 2). As no published structures for eukaryotic homologs were available, a model of a representative tertiary structure was generated using the Phyre2 fold prediction webserver (Figure S4). This alignment suggested that the additional histidine pair did not contribute to the binding pocket (Fig. S4, d), and was, instead, positioned exposed on the protein surface, implying a possible role in protein-protein interactions; however, characterizations of this and similar structures have demonstrated a putative involvement in the architecture of the cleft of the active site formed upon dimerization [139]. A final distinguishing feature observed in eukaryotic tier 1 sequences is an additional arginine residue following the C-terminal "HxxxE" motif of the C-terminus, a final motif indicated as a likely contributor to the binding pocket [26,135].

Figure 2. Key motifs of bacteria and archaea compared to those of eukaryota.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1



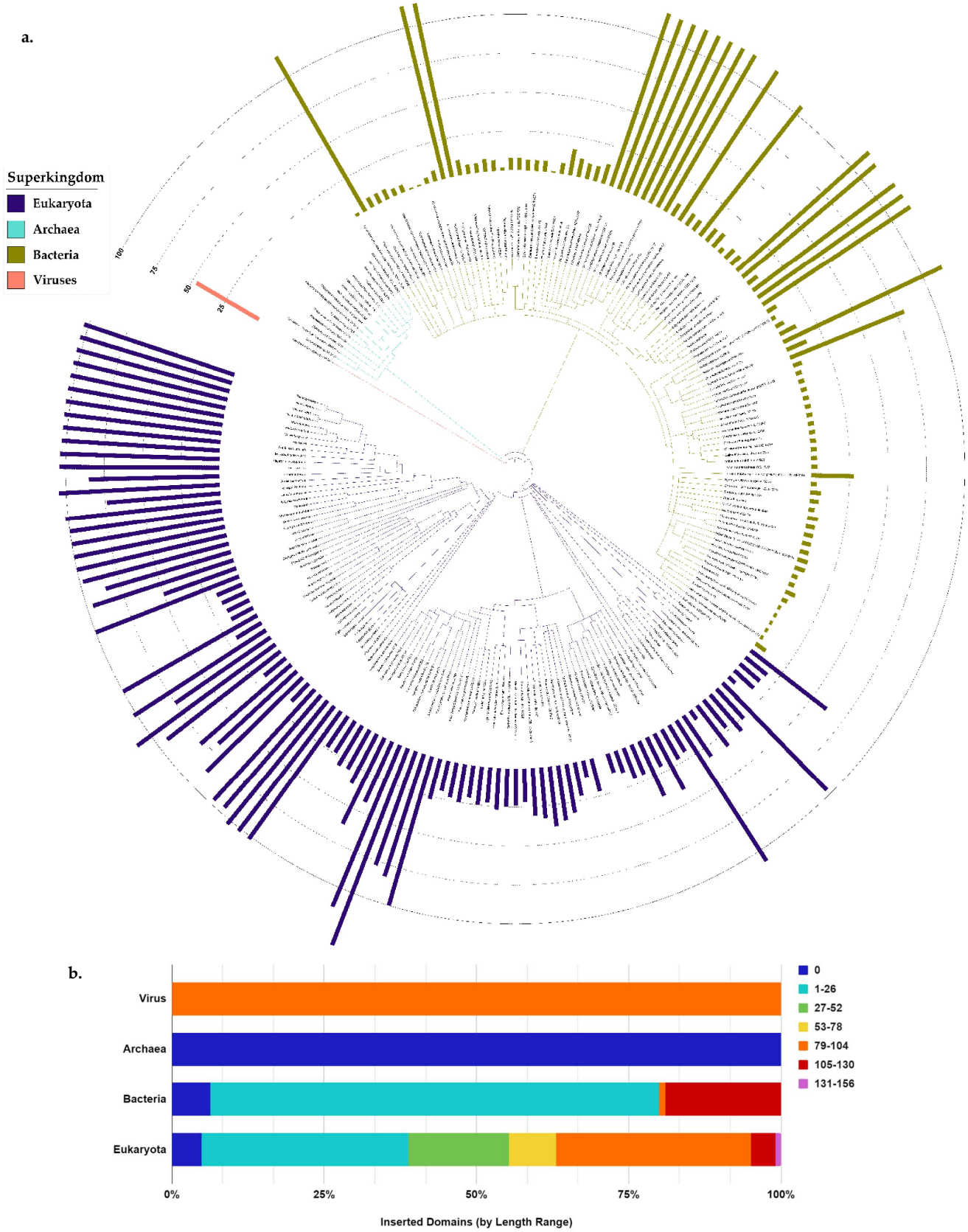
Sequences were aligned for eukaryotic sequences, separately, and, for bacterial and archaeal sequences, combined. A multiple motif method was used to determine and compare family signatures. Full figure illustrating distinct levels of conservation per superkingdom can be examined in Figure S3.

3.4. A variable central insertion occurs in some DUF34 family members.

Alignments performed per superkingdom revealed a large diversity in the lengths of aligned sequences (Data Table 4). The spacing between the Tier 1 motifs seemed to vary greatly with the superkingdom. To better understand the occurrence and distribution of lengths for this inserted domain, the regions between the "YxxHxxxxD" and "Dxxx(T/S)G(E/D)" motifs were manually extracted, lengths measured, and their values were then superimposed onto a species tree (Figure 3). With this, it was revealed that the inserted domains were relatively well conserved in select clades of bacteria, a finding reminiscient to an earlier observation made by Godsey *et al.* [135]. Unexpectedly, an inserted region was frequent in proteins from higher-order eukaryota, but was entirely absent from archaeal homologs. Among eukaryotic DUF34 proteins, the insertion sizes followed a pattern of diminishing length from vertebrate to invertebrate homologs (from higher-order to lower-order eukaryota) (Figure 3). In contrast, the length of this domain was relatively stable among bacterial homologs, if occurring at all, with 28.3% harbouring a large form of the insertion (~100 aa), while the remaining sequences lacked the domain entirely. Outside of the regions observed in vertebrates, the sizes of this domain varied greatly, especially in members of invertebrate bilateria and fungi, the latter taxon demonstrating domains of the shortest lengths. Only one viral DUF34 member, MIMI_R836 (Q5UQI9) of *Acanthamoeba polyphaga mimivirus*, was retrieved from published data and its length was notably dominated by the inserted domain.

Figure 3. Inserted domain lengths across model taxa.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1



 doi:10.20944/preprints202107.0280.v1

(a) Lengths of inserted domains were measured for each homolog. Sequences (organisms listed in Data Table 4) were aligned per superkingdom for delimiting domains, which then allowed for the measurement of each inserted region (if present). An evolutionary tree was generated using PhyloT and iToL, and were mapped with the lengths of inserted domains within each respective homolog. For all inserted domain lengths measured (shown in (a)), these data were used to generate (b) a histogram illustrating counts by ranges of domain lengths per superkingdom. The color key denoting each of the 7 different ranges of inserted domain lengths is provided.

3.5. The DUF34 family can be split in eight interconnected subgroups.

To further characterize domain architectures and examine possibilities of functional subclasses, we collected the annotated domains linked to DUF34 family members, specifically leveraging InterPro HMM profile signature identifiers and EggNOG group IDs (Clusters of Orthologous Groups or COGs) (Figure 4; Data Table 5). Various overlapping combinations of COGs and HMM profile signatures were observed, generating a set of specific architectural patterns that were used to delineate alphabetically named subgroups (i.e., A-G). Most DUF34 members fell within one of two keystone COGs. The first, COG0327 (subgroup A; Figure 4a), is predominantly defined by the presence of two specific HMM profile signatures, IPR036069 and IPR002678, and largely defines the shared bases across subgroups. COG0327 is further divided by HMM profile signatures into two subgroups, subgroup B and subgroup C (Figure 4a), the former containing an animal-specific signature (IPR017222) and the latter harbouring a bacteria-specific signature (IPR017221). Although subgroup C was described by InterPro-defined HMM profile signature annotations as being limited to bacteria, nearly all proteins observed within this subgroup belonged to eukaryotes. All members of subgroup B occurred in eukaryota. The second keystone COG of the DUF34 family, COG3323, was defined by the presence of IPR015867 and IPR036069 (subgroup D; Figure 4a), with IPR036069 being shared between COG3323 and COG0327. The addition of a third HMM profile signature, IPR004323, to the pairing of IPR015867 and IPR036069 defined a fifth subgroup, subgroup E. Homologs containing all three keystone COG-definitive signatures (i.e., IPR002678, IPR015867 and IPR036069) was determinate for fusions of COG0327 and COG3323. These fusions were observed to occur in two forms: subgroup F and subgroup G, the latter of which was defined by the additional bacteria-specific signature, IPR017221 (Figure 4a), a signature previously noted in the definition of subgroup C.

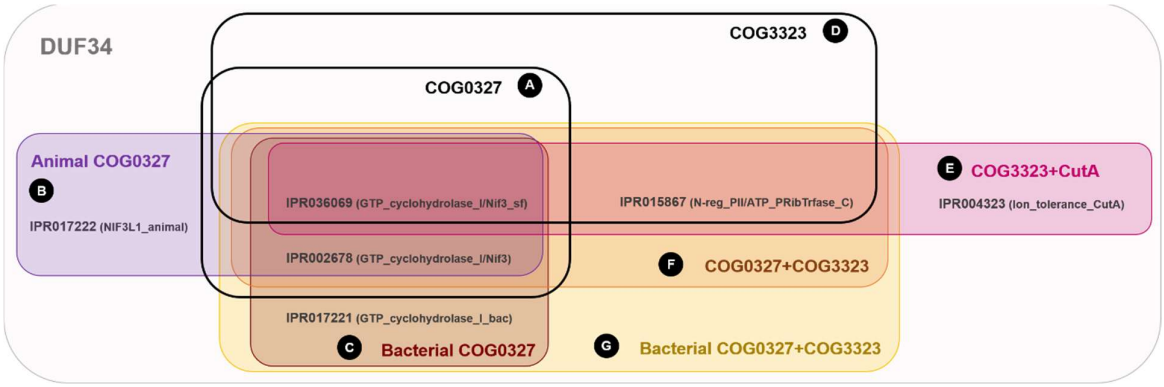
The D-G subgroups can be differentiated from the A-C subgroups by the presence of a "HPYE" motif attributable to the HMM profile signature, IPR015867 (Figure S5, a-b). It can also be noted that subgroups D and E can be viewed as stand-alone forms of the inserted domain found in subgroups F and G. For example, for the DUF34 paralogs of *B. cereus*, BC_2685 (Q81CR2) and BC_4286 (Q818H0), the latter sequence was found to contain an inserted domain bearing high similarity to the former (31.0% identity, 48.0% similarity; EMBOSS Matcher; Figure S6, d) (Figure 4b). This same paralog, BC_2685, was identified as a member of the CutA1 protein family (PF03091). Interestingly, this YqfO-like paralog was also found to have greater identity to the CutA1 homolog of *H. sapiens* (O60888; 29.4% identity, 47.1% similarity) than to that of other bacteria (i.e., *E. coli*; P69488; 25.6% identity, 55.8% similarity). Interestingly, the final glutamate residue of the key motif also distinguishing DUF34 protein family member inserted domains, "HPYE" of the IPR015867 HMM signature profile (Figure S6, g), was replaced by a glutamine in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

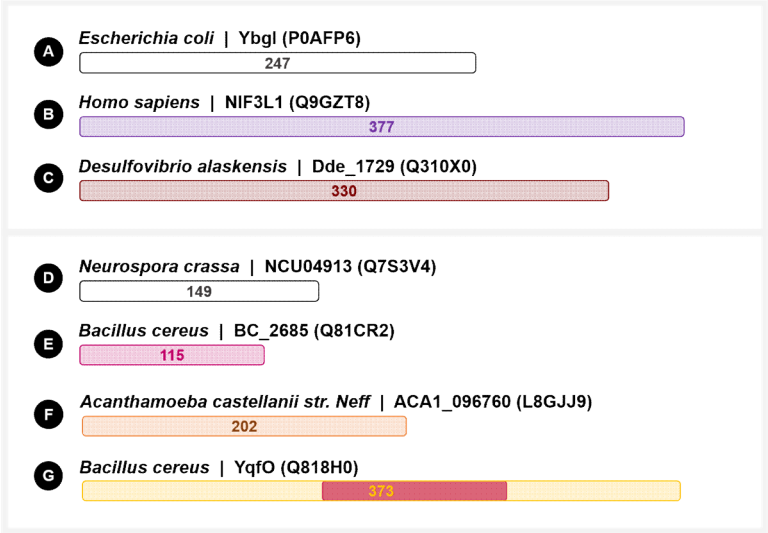
the CutA1 of *E. coli*, a replacement also observed in the inserted domain of NIF3L1, the DUF34 homolog of *H. sapiens*. The CutA1 protein family (formerly known as DUF190) has historically been linked to divalent cation tolerance, copper sensitivity and cytotoxicity (PF03091; IPR004323; COG1324) [144–150]; however, due to characteristics of the quaternary structure (trimers form ferrodoxin-like folds [151]), roles in signal transduction and regulation have also been suggested [152–154]. More recently, refute of the protein's involvement in metal ion tolerance has led to predictions of CutA1 proteins acting in a small molecule carrier or signaling capacity [155,156]. Still, the functions of all three "CutA" proteins remain under-defined with only small attributions put forward for each, in addition to CutA1: CutA2 (DsbD) is thought to have disulfide oxidoreductase activity [157]; and CutA3 (YjdC) has been annotated as an HTH-type transcriptional regulator (TetR/AcrR family), more specifically a negative regulator of nitroreductase NfnB [158].

Figure 4. COG-InterPro HMM signature profile relationships and defined subgroups across DUF34 family members.

a.



b.



Sequences of organisms across the DUF34 protein family, including all fusions and paralogs, were analyzed for co-occurrence relationships of COGs and HMM-determined InterPro family/superfamily/domain annotations. All organism homologs, paralogs & fusions were validated using eggNOG and KEGG Paralog Search. Sequences missing InterPro annotation were analysed by NCBI CDD Search and InterProScan Search. See Data Table 5 for categories and respective COG designations/InterPro signature profiles in tabular format. Sequence source organisms considered were those also observed in Data Table 4. Groups were designated by differential keystone signatures shown in (a) and select representative sequences of subgroups (A-G) are shown (b).

3.6. Taxonomic distribution suggests that the NIF3 (COG0327) and YqfO-like (COG3323) domains have different functions.

Contrary to expectations for the universal conservation established by past publications, particularly in eukaryota, DUF34 appeared absent from the eukaryotic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

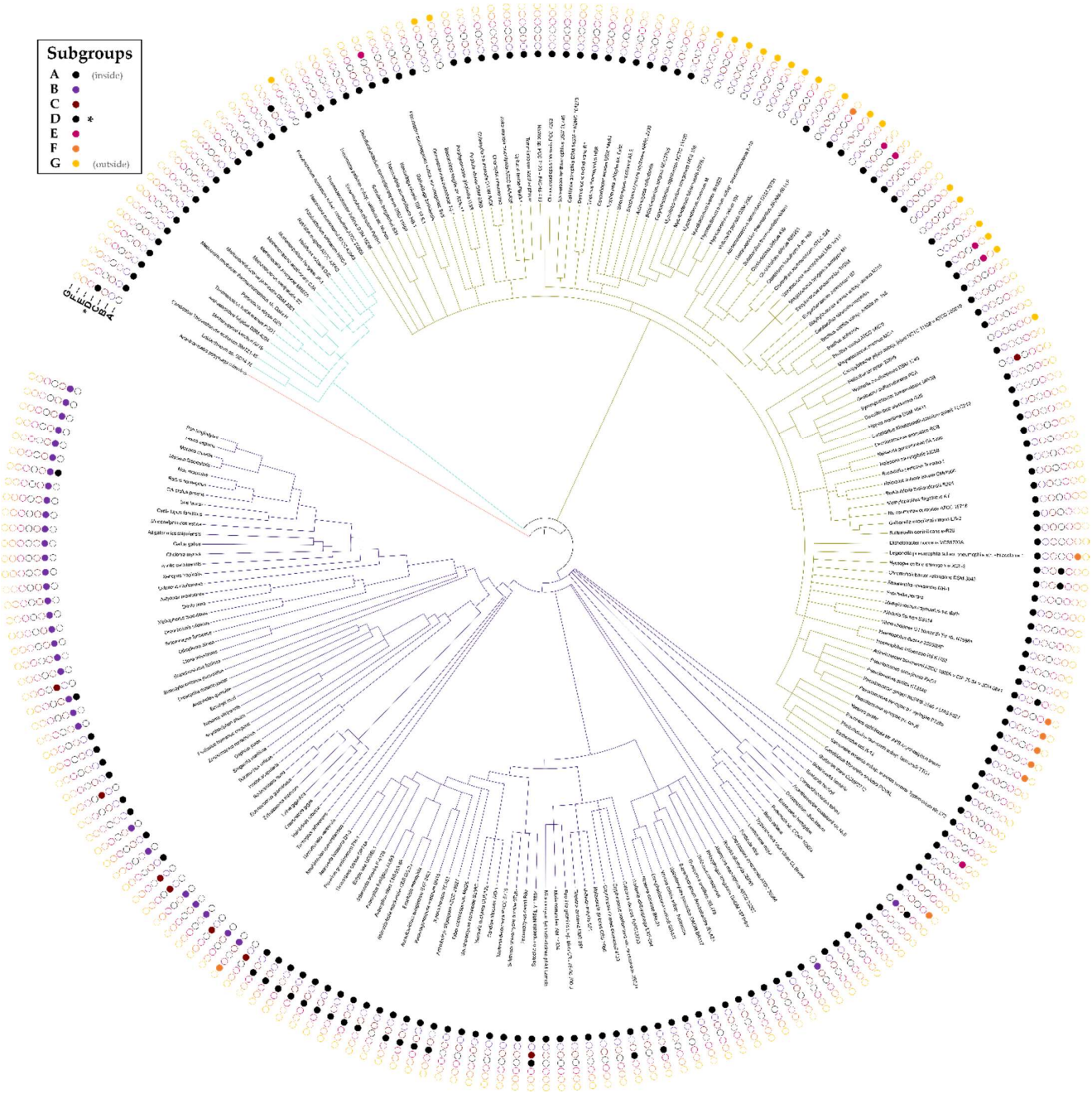
clade of *Viridiplantae* with the closest incidence of homologs occurring in select haptophyta. Although some sequence-based queries of NCBI's databases indicated the existence of a partial homolog belonging to a specific eudicot (i.e., histidinol dehydrogenase chloroplastic isoform X1, GEY60218.1; GFD1148.1; KYP77406.1), these few observations appear largely uncorroborated and were suspected to be products of bacterial contamination. *Caenorhabditis elegans*, a common model organism, was also observed to lack a DUF34 homolog. Among the organisms analyzed, Archaea exclusively harboured DUF34 members of subgroup A (Figure 5). The animal-specific subgroup B, was restricted to *Metazoa*, occurring ubiquitously across *Euteleostomi*. Subgroup A often replaced the animal-specific subgroup B in other lower-order clades of *Metazoa* including, but not limited to: *Arthropoda*, *Annelida* and *Mollusca* (Figure 5). Subgroup A also demonstrated the greatest overall prevalence and broadest taxonomic range, being observed in the majority of organisms across the three major superkingdoms. Almost all bacteria lacking a subgroup A homolog harbored a subgroup G, the bacterial COG0327-COG3323 fusion, in its place. Of all YqfO-like (COG3323) variants of the DUF34 family (subgroups D-G), only subgroup G was ever observed to occur without a subgroup A, B or C form also present. The only exception to this pattern of subgroup absence-presence was *Acanthamoeba polyphaga mimivirus* (tax ID: 212035), which was found to only encode a subgroup D homolog. Interestingly, the DUF34 form annotated as being specific to bacteria, subgroup C, was exclusively observed among select species of non-metazoan bilateria, only occurring in a single bacteria (i.e., *Desulfovibrio alaskensis*).

Approximately three-quarters of the genomes analyzed encoded only one subgroup of the DUF34 family. In organisms with two or more subgroups, the most frequent combination was the co-occurance of either a subgroup A, B or C with any member of subgroups D-G. Although seldom, subgroups A, B and/or C were observed to co-occur together, most often in pairs, in eukaryotic organisms, but never in bacteria, archaea or viruses. Only members of subgroup G ever occured alone more than once without any subgroups A-C. This suggests that this is the only form that can functionally replace any one of the A-C forms, and that the stand-alone versions of the inserted domains definitive of subgroups D or E, relative to subgroups A-C, certainly perform a different function.

In a larger survey of available complete bacterial genomes (JGI-IMG/M; accessed January 2020), DUF34 homologs annotated as belonging to both COGs (subgroups D-G) COG3323 and COG0327, occurred in 18% of complete bacterial genomes, while a much larger fraction of the bacterial family members (66%) were found to encode only the COG0327 designation (Subgroups A-C) (Data Table 6) [159–161].

Figure 5. Absence-presence of DUF34 architectural domain subgroups.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1



Absence-presence data of COGs and HMM-determined InterPro family/superfamily/domain signature profiles added to a species tree, generated using organisms harboring published homologs and those used in alignments acquired via OrthoInspector (Data Table 4) Proteins are designated as categories A-G, as detailed in Figure 4 and Data Table 5. These homologous domains are classified in the map according to their HMM-defined DUF34 domain identities (see Figure 4a).

3.7. Physical clustering and coexpression further links the DUF34 family to metal ion homeostasis and iron sulfur-cluster metabolism.

To determine associations based on physical clustering, gene neighborhoods for members of the DUF34 family were examined using the IBT–UNAM Computational Genomic Group's Gene Context Tool (GCT). The GCT webserver was used to retrieve collections of commonly clustered COGs of DUF34-encoding operons for taxonomic subsets of bacterial and archaeal DUF34 family members (Data Table 7, a). These data were then used to develop a method of text analysis-enabled assessment of COG and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

COG description keyword/phrase frequencies, the methods of which are described further in the Supplemental Methods section (1.2). This approach will be referred to, henceforth, as Physical Clustering Keyword Frequency Analysis (PCKFA). Using PCKFA, COGs and their descriptions were examined for common annotations and trends that could inform on potential functional associations. PCKFA of COG identifiers was used to generate a ranked list of co-occurring COGs. This data was sorted by frequency to generate a final list of the top 20 highest-ranking COGs occurring across all taxonomic ranges (Table 4). Upon closer review of the associated functional annotation, it was determined that 65% (13) of the top 20 most frequently co-occurring COGs of DUF34-containing operons were either predicted or confirmed to be "metal ion-binding/-dependent", an incidence notably greater than the one-third of proteins within PDB predicted to require metal ions [162]. Three of the 13 metal ion-binding/-dependent COGs within those ranking within the top 20 were found to bind Fe-S clusters (Table 4). Despite the diversity of operon compositions that were observed within and between the data's selected taxonomic ranges (Data Table 7), keywords linked to metal ion homeostasis and Fe-S cluster-dependent processes recurred with notable frequency (Figures S7, a).

Representative operons were curated to facilitate more granular, context-driven analyses investigating the observed trends (Data Table 7, d-e). With an initial survey of metal bias based only on COG descriptions, whether or how many of the encoded COGs might be linked to pathways involving metal ions and/or Fe-S clusters remained unclear. This was largely due to the generally poor functional annotation statuses for many of the COGs retrieved. Therefore, the individual sequences constituting these operons were investigated thoroughly using functional annotation and key background literature (as described in Methods) to investigate annotations for any catalytic dependencies or interactions with metals ions. In 13 of the 51 selected bacteria (25.5%), COG0327 was observed to occur alone, and, of those not encoded alone (38 of 51), 31 were found to encode at least one protein with supported annotations of metal-binding/-dependence (81.6% of operons; count inclusive of Fe-S cluster-containing proteins) (Data Table 7 & 8). Similar incidence was observed across archaeal representative operons with 3 of 9 archaeal COG0327 proteins (33.3%) being encoded alone, and, of those not, five were found to encode at least one metal-binding/-dependent protein (5 of 6 operons; ~83%).

Table 4. Top 20 COGs found to occur in operons containing COG0327.

Rank	COG	Name/Description	Metal(s)	References (PMID, BRENDA, EC)
1	**COG0327**	Putative GTP cyclohydrolase 1 type 2, NIF3 family	Fe^{2+}/Fe^{3+}, Zn^{2+}, Mg^{2+}	14519207, 24931373, 17187687, [26.88.147.156], [26.89.148.157]
2	**COG1579**	Predicted nucleic acid-binding protein DR0291, contains C4-type Zn-ribbon domain	Zn^{2+}	22408721
3	**COG0568**	DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32)	Zn^{2+}, Mg^{2+}	29514271, [2.7.7.6]
4	**COG0358**	DNA primase (bacterial type)	Zn^{2+}, Mg^{2+}, Mn^{2+}	1511009, [2.7.7.101]
5	**COG0457**[a]	Tetratricopeptide (TPR) repeat	*NA*	*None listed*
6	**COG2384**	tRNA A22 N1-methylase	*NA*	[2.1.1.217]
7	**COG0079**	Histidinol-phosphate/aromatic aminotransferase or cobyric	*NA*;	32973726, [2.6.1.9]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

		acid decarboxylase	Co (cobalamin)	
8	**COG0240**	Glycerol-3-phosphate dehydrogenase	*NA*	[1.1.1.94]
9	**COG0328**	Ribonuclease HI (RnhA)	Mg^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+}	16601679, [3.1.26.4]
10	**COG0500**[b]	SAM-dependent methyltransferase	*NA*	[2.1.1.242]
11	**COG0513**[c]	Superfamily II DNA and RNA helicase (SrmB/RhlB)	Mg^{2+}, Mn^{2+}	[3.6.4.13]
12	**COG0596**	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase MenH and related esterases, alpha/beta hydrolase fold (MhpC)	*NA*	[3.7.1.14]
13	**COG0655**	Multimeric flavodoxin WrbA, includes NAD(P)H:quinone oxidoreductase	Most req. Fe-S cluster; subtypes without Fe-S clusters	[1.6.5.2], [1.6.5.6]
14	**COG0752**	Glycyl-tRNA synthetase, alpha subunit	Mg^{2+}, Mn^{2+}, Co^{2+}	4295604, [6.1.1.14]
15	**COG0826**	23S rRNA C2501 and tRNA U34 5'-hydroxylation protein RlhA/YrrN/YrrO, U32 peptidase family; ubiquinone biosynthesis protein, UbiU/YhbU	Fe-S cluster/Fe, Ca^{2+}	31289180, 1317840
16	**COG1028**	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	Co^{2+}, Fe/Fe^{2+}, Mg^{2+}, Mn^{2+}, Zn/Zn^{2+}	[1.1.1.2]
17	**COG1897**	Homoserine O-succinyltransferase	*NA*	[2.3.1.31], [2.3.1.46]
18	**COG0177**[d]	Endonuclease III (Nth)	Fe-S cluster, Ca^{2+}, Co^{2+}, Fe/Fe^{2+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, Zn^{2+}	8092678, [4.2.99.18]
19	**COG0477**[d]	MFS family permease (includes anhydromuropeptide permease AmpG, ProP)	*NA*	*None listed*
20	**COG0494**[e]	8-oxo-dGTP pyrophosphatase MutT and related house-cleaning NTP pyrophosphohydrolases, NUDIX family	Co^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+}	[3.6.1.13]

Exceptions representative operons relative to table contents:

[a]Proteins containing TPR repeat domains present in archaeal operons.

[b]SAM-dependent methyltransferase domains present (not designated COG0500).

[c]Though not assigned COG0513, helicase domain-containing proteins are present (e.g., Era/COG1159, YhaM/COG3481).

[d]MutY is present (COG1194), another endonuclease family member.

[e]MutM/NUDIX domain containing proteins are present (COG0266).

Of all COGs encoded by COG0327-containing representative operons, COG1579 co-occurred most frequently. This COG was also determined through PCKFA to be the top-most ranked in, both, singular occurrence and paired occurrence with COG0327 across taxonomic ranges (Figure S7, b-c). COG1579, is a family of unknown function (DUF164) that is conserved primarily among bacterial clades, although homologs are found also in archaea. Members of this group have been linked to functional roles in chemotaxis, flagellin synthesis, type III secretion systems (i.e., *Helicobacter pylori* and *Chlamydia trachomatis* [120,163–165]), and bacteria-induced host cell maturation (i.e., *Mycobacterium avium* [166,167]) but the molecular mechanisms involved remain mysterious. The homolog of *Mycobacterium tuberculosis* has been noted as an essential gene under some circumstances [168]. COG1579 members have an obvious link because of the presence of a domain belonging to the zf-RING_7 Pfam family (PF02591 [169]). A characteristic feature of the zf-RING_7 family is the presence of a C4-type zinc-ribbon domain with two pairs of cysteines in a CxxC-x (18–26)-CxxC (zinc-finger) motif capable of binding zinc ions. Published structures (5Y06/5Y05 of *M. smegmatis* [164]; 4ILO of *Chlamydia trachomatis* [165]) demonstrate an unusual coiled-coil structure that is book-ended by the aforementioned distinctive zinc-finger domain.

Table 5. Metal ion interactions of proteins encoded by representative operons.

	Biological Category	**COG**	**Protein**	**Metal(s)**	**References (PMID, BRENDA, EC)**
Bacteria					
Metal-binding/	DNA and RNA metabolism	COG0328	RnhA	Mg^{2+}, Mn^{2+}, Co^{2+},	16601679, [3.1.26.4]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

-dependent				Ni^{2+}	
		na	RNase P	Mg^{2+}, Mn^{2+}, Zn^{2+}	15867194, [3.1.26.5]
		COG0125	Tmk	Mg^{2+}	28627020, [2.7.4.9]
		COG0752	GlyQ/S	Mg^{2+}, Mn^{2+}, Co^{2+}	4295604, [6.1.1.14]
		COG3935	DnaD	Mg^{2+}	18703019
		COG0358	DnaG	Zn^{2+}, Mg^{2+}, Mn^{2+}	1511009, [2.7.7.101]
		COG0568	RpoD/SigA	Zn^{2+}, Mg^{2+}	29514271, [2.7.7.6]
		COG3481	YhaM	Mn^{2+}, Co^{2+}	12399495, 9868367
		COG0319	YbeY	Zn^{2+}, Ni^{2+}	15632286, 16511207
		COG1381	RecO	Zn^{2+}	25170075, 15719017
		COG0228	RpsP	Mg^{2+}, Mn^{2+}	8730873
		na	Initiator $tRNA^{Met}$	Mg^{2+}	4563246
	A proportion of cases ▶	COG0336	TrmD	Mg^{2+}, Mn^{2+}, Ca^{2+}	25219964, [2.1.1.228]
		COG0266	Fpg/MutM/Nei	Zn^{2+}, Fe^{2+}, Mg^{2+}, Mn^{2+}, Ca^{2+}, Cu^{2+}/Ni^{2+}*	7955043, [4.2.99.18]
		COG0232	Dgt	Mg^{2+}, Mn^{2+}	25694425, [3.1.5.1]
	Membrane biosynthesis; transport; signalling	COG1137	LptB	Mg^{2+}	19234479, 11080142, [3.6.3.-]
		COG1496	YfiH/RL5/PgeF	Zn^{2+}, Cu^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Cd^{2+}	16498617, 16740638, 28612943, [1.7.2.1, 1.10.3.3, 1.10.3.2, 1.16.3.1]
		COG0541	Ffh	Mg^{2+}	14696184, [3.6.5.4]
		COG1127	MlaF/ttg2A	Mg^{2+} (ATP-Mg^{2+})	25916755
	Regulation	COG1692	YmdB	Fe^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Ca^{2+}	24163345, [3.1.4.16]
		COG0265	DegQ	Zn^{2+}, heme, Mn^{2+}, Ca^{2+}, Fe^{2+}, Mg^{2+}	23695557, 23176475, 18723647 [3.4.21.107]
		COG3744	VapC16	Mg^{2+}, Mn^{2+}	28575517
		COG0645	Zeta toxin/P-loop containing NTPase	Mg^{2+}	21445328
		COG1366	SpollAA	Mg^{2+}	15236958
		COG0466	Lon	Mg^{2+}, Mn^{2+}, Ca^{2+}	16511355, [3.4.21.53]
		COG1579	DUF164, zinc ribbon containing	Zn^{2+}	22408721
		COG0642	BaeS/NtrC/AtoS	Fe/Fe^{2+}/Fe^{3+}, Cu^{+}/Cu^{2+}, Ag^{+}, Mg^{2+}	26950881, 21886814, [2.7.13.3]
		COG2204	AtoC-REC/NtrC (PilR-like)	Mg^{2+}, Mn^{2+}, Sr^{2+}, Zn^{2+}/Pb^{2+}, Ag+, Fe/Fe^{2+}/Fe^{3+}, Cu^{+}/Cu^{2+}	REC domain: cd00156; CheY-like receiver: 8257674; Other: [2.7.13.3], 11243806 (BRITE/KEGG)
		na	VanSB-like (MGA_0021)	Ag^{+}, Cu^{+}/Cu^{2+}, Fe/Fe^{2+}/Fe^{3+}, Mg^{2+}	[2.7.13.3]**
	Small molecule metabolism	COG0346	GloA	Ni^{2+}, Co^{2+}, Cd^{2+}, Mn^{2+}, Mg^{2+}, Zn^{2+}, Ca^{2+}, Fe^{2+}	21820381, VOC-like domain: cd07245; [3.4.17.13]
		COG0041	PurE	Mg^{2+}	2464576, [5.4.99.18]
	fusion ▶	COG0328-COG0406	RnhA-CobC	Mg^{2+}	6312898
	fusion ▶	COG1211-COG0245	IspD-IspF	Mn^{2+}, Mg^{2+}	21543842, [2.7.7.60], 10694574, [4.6.1.12]
		COG0699	CrfC	Zn^{2+}, Mg^{2+}	[3.6.5.5]
		COG2049, COG1984, COG1540	PxpB, PxpC, PxpA	Mg^{2+}, Mn^{2+}	28830929, [3.5.2.9]
Heme/metalloenzyme/ metallocofactor biosynthesis, transport (Non-metal-dependent indicated by asterisk)		COG2274	SunT	Mg^{2+}	26201595 (cd02424: Peptidase_C39E)
	Cytochrome c maturation ▶	COG1131, COG1277	CcmA/GldA (CcmABC), GldF/NosY*	Heme (Cu) (GldAFG)	28472044, 11948149; P-loop: cl38936; COG1277 [CDD]; 12618453

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

	Homogentisic acid, pyomelanin production/transport ►	COG0760	SurA	Mg^{2+}	12429090, [5.2.1.8]
		COG0778	RdxA/NfnB-like	Fe/Fe^{2+}	[1.13.11.55]**
		COG1127	MlaF/ttg2A, MlaB (MlaBCDEF/ Ttg2ABCDE)	Fe^{3+}, Mg^{2+} (ATP-Mg^{2+}); Fe^{2+}(assoc. enzymes)	20870774, 23858455, [3.6.3.-]; SulP antagonist domain: cd07042; 15236958
		COG0079	CobC*	Co (cobalamin)	32973726, [2.6.1.9]
Fe-S cluster	Base excision repair; DNA repair, maintenance	COG1194	MutY	$[4Fe\text{-}4S]^{2+}$	25445713
		COG1533	SplB	$[2Fe\text{-}2S]^{2+}$, $[4Fe\text{-}4S]^{2+}$	16829676
		COG0415	PhrB	[4Fe-4S]	23589886
	Fe-S cluster biogenesis	COG0694	NfuA/NifU	[4Fe-4S]	22966982
	Regulation	*na*	ENOG5030DA8 (BC_4767)	Fe/Fe^{2+}, Fe-S, [2Fe-2S], [4Fe-4S], Mg^{2+}, Ca^{2+}	[7.1.1.2]**
		COG0826	PrtC/UbiU/YhbU	Fe-S cluster/Fe, Ca^{2+}	31289180, 1317840
	Regulation of virulence, membrane traits and interactions with environment	COG5007	BolA	[2Fe-2S]	27951647
Archaea					
Metal-binding/ -dependent	DNA and RNA metabolism	COG2016	Tma20	Zn^{2+}, Mg^{2+}	12054814
		COG0024	Map	Co^{2+}, Fe^{2+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, Zn^{2+}	9811545, [3.4.11.18]
		COG1833	Uri	Zn^{2+}	16646971
		COG5431	Shu2-like, SWIM domain containing	Zn^{2+}	29069504
		COG0640	ArsR	Zn^{2+}, Co^{2+}, Ni^{2+}	10995250, 9466913; HTH ArsR: cd00090
		COG0013	AlaS	Zn^{2+}, Mg^{2+}	16374837, [6.1.1.7]
	Small molecule metabolism	COG0388	YafV (YafV-AguA fusion)	Zn^{2+}, Ni^{2+}, Mg^{2+}, Co^{2+}, Fe	pfam00795 [3.5.5.1, 3.5.1.4, 3.5.1.12, 3.5.1.6]
	Metal chelation, homeostasis	COG0679	RhaT	Se	SelP domain: cl04615
	Cofactor biosynthesis, maturation	COG4074	Mth/Hmd	Zn^{2+}	8599536, [1.12.98.2]
		COG4015	HcgE	Mg^{2+}	25882909
		COG0502	HcgA/BioB	$[4Fe\text{-}4S]^{1+}$, $[2Fe\text{-}2S]^{1+}$, Fe/Fe^{2+}	22095926, [2.8.1.6]
		COG4018	FlpA/HmdC/HcgG	[4Fe-4S]	22095926
	Uncharacterized	*na*	Mhun_0036	Zn^{2+}, Mg^{2+}	18433773
		na	Mhun_0038	Zn^{2+}	PF04434: SWIM zinc finger domain

***Indicates EC number acquisition via KEGG GFIT (⩾30% identity); *indicates interaction with metals is indirect*

Despite the high clustering frequencies discernible for several co-occurring COGs, a single link between DUF34 homologs and a distinct metabolic area remained unclear. The diversity of metals associated with proteins encoded by DUF34-containing operons failed to support a preference for a single metal or metal ion-complex, although zinc and iron were found to be common interactors, second to magnesium and manganese. In addition, many of the families listed in Table 5 were found to interact with several metal ions (up to eight) with averages, across the table, of ~2.5 different metals for bacterial proteins and ~1.9 for archaeal proteins. Several metal-dependent/-binding COGs found to frequently cluster within DUF34-containing operons across taxa (Table 4) were also common among representative operons (Data Table 7). When compared to all available PDB structures (PDB 2020), the relative abundance of metal-binding proteins across both

 doi:10.20944/preprints202107.0280.v1

archaeal and bacterial representative operons was observed to be significant (Data Table 8; Figure S8-S10). As stated above and in Table 5, a strong assocation with Fe-S cluster associated proteins was observed (7 of the 40 bacterial and 2 of the 14 archaeal metal-binding proteins analyzed). Examples include HcgA/BioB, and HmdC/HcgG (FlpA homolog) in archaea, and MutY, SplB, NfuA, PhrB and BolA/in bacteria.

Because DUF34 is conserved across bacteria, archaea and most eukaryotes, and as physical clustering was appropriate for only two of three superkingdoms [170], coexpression (top 300 co-expressed, CoXPresDb; Data Table 9, sheets d.1-d.10) and coregulation databases (ProteomeHD; Data Table 10, a) were consulted to identify trends in putative functional associations of eukaryotic DUF34 family members shared with those observed through preceding analyses with bacterial and archaeal family members. Interestingly, a number of genes directly involved in iron homeostasis and Fe-S cluster biogenesis were observed to occur in most eukaryotic organisms surveyed (Data Table 9; Figure S11). BolA or BolA-like family members occurred in *H sapiens*, *M. mulatta*, and *S. cerevisiae*. However, in absence of a BolA-like homolog, *S. pombe* showed coexpression of a Fe-S cluster biogenesis factor, caf17 (IBA57-like; SPAC21E11.07), a member of the GcvT and CAF17 families [171]. Upon further review of the top 100 genes coexpressed in *H. sapiens*, YAE1D1 (57002, Yet Another Essential domain-containing 1), a highly conserved protein essential to cytosolic Fe-S cluster protein assembly (CIA) complex [172], was also observed. Although a Yae1 homolog was not observed in the acquired datasets for either yeast, another essential component of the CIA complex, the Fe-S cluster-binding ATPase, Nbp35 (2543416, *S. pombe*; 852789, *S. cerevisiae*), was found within the top 130 coexpressed genes of each. Genes encoding this protein were found coexpressed with NIF3L1 homologs in three eukaryotes of the 10 total for which data was retrieved. Similar trends associating Fe-S cluster proteins and pathways were observed upon gene functional classification analyses of the same sets of coexpressed genes using the DAVID bioinformatic suite (Data Table 9, e.1-e.10).

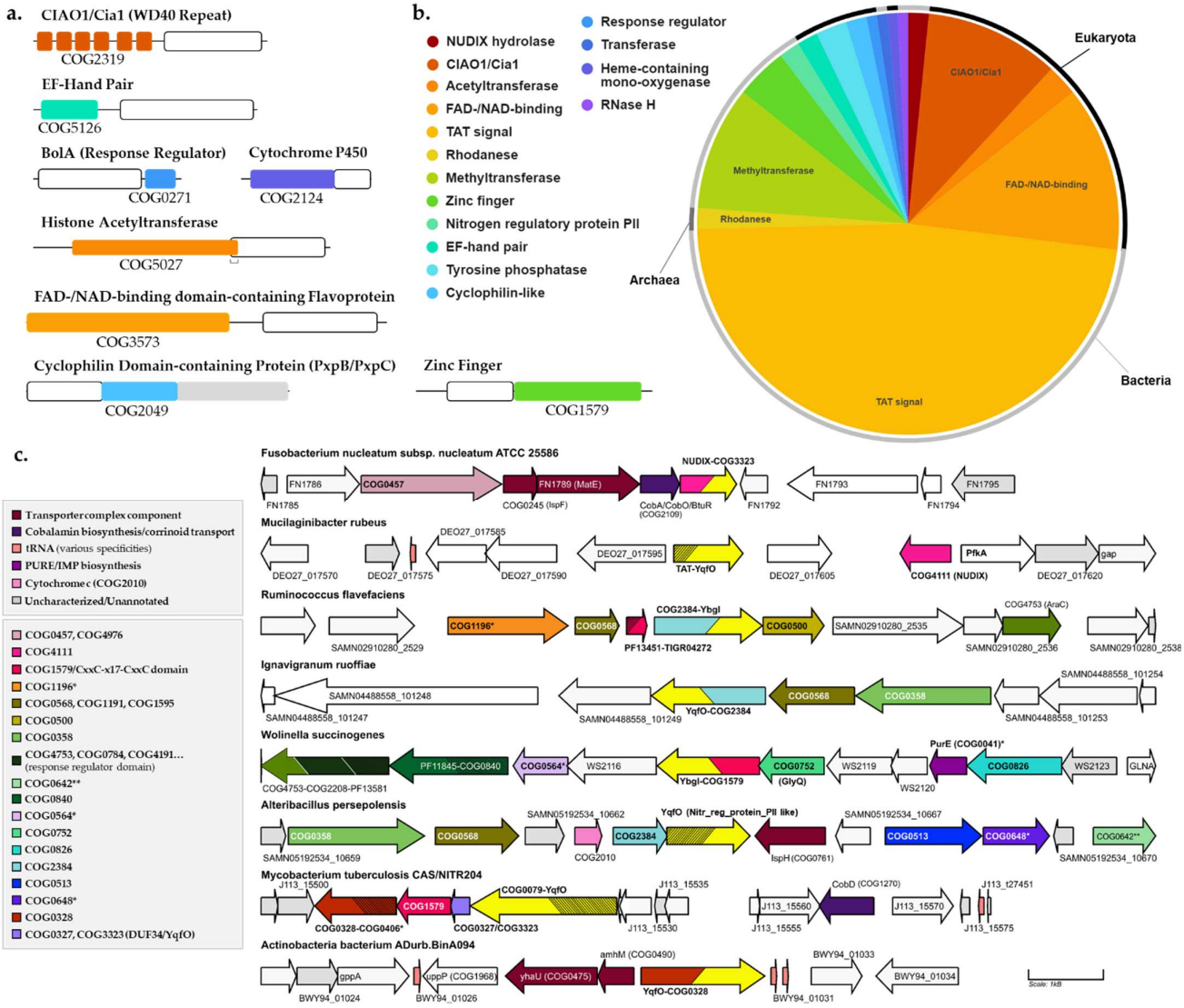
3.8 DUF34 fusions fortify links to metals and metallocofactors, most notably Fe-S clusters.

Fusions can provide substantial insight into putative functional relationships between their constituent protein families. To better understand the full diversity of fusions across the DUF34 family, three different methods were used as described in the methods section to generate a curated set of 226 sequences of varying validity (Data Table 11, b), covering 47 distinct fusion classes and 65 different fusion subclasses (see Supplemental Methods, 1.3). After further curation focusing on fusions of highest confidence, nine fusion classes were in observed in eukaryotes and seven in bacteria. Eukaryotic fusions of note included those with the following domains: WD40 repeat; BolA (BolA-like); FAD-binding flavoprotein; RING- or THAP-type zinc finger; EF-Hand pair; or histone acetyltransferase (Figure 6a). The most common fusion among eukaryotes were those containing the WD40 repeat domain, CIAO1/Cia1 (COG2319), of which is thought to play a role in Fe-S cluster biogenesis. Somewhat consistent with this finding, a fusion with BolA was also observed (COG0271, PF01722; *Fusarium oxysporum Fo47*). It was also remarked that the neighboring of BolA family members, a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

phenomenon shared by at least one bacterial representative operon (Data Table 7, d.1-d.2), was not necessarily uncommon in fungal genomes, as Bol2, for example, is divergently encoded immediately upstream of DUF34 in *S. cerevisiae*.

Figure 6. DUF34 fusions and select gene neighborhoods.



(a) Domain architectures of DUF34 fusions. Domain rendering dimensions and positions are approximate. DUF34 domains are rendered in white with black outlines. Domain colors correspond to key shown in panel b. COGs of fusion domains are listed below each. Fusions deemed "invalid" or "inconclusive" were excluded for panels a and b. (b) Pie chart of DUF34 fusions (126 sequences, total). Outer halo surrounding chart indicates the superkingdoms in which respective fusions were observed (Eukaryota: black; Archaea: dark gray; Bacteria: light gray). (c) Neighborhoods of select bacterial and archaeal fusions are shown (12kb, each), all of at least "conditional" validation confidence (Data Table 11). DUF34 is depicted in bright yellow and fusion domains are indicated by hashing or alternative coloring. For DUF34 sequence labels, "YqfO" denotes a sequence also containing inserted domain, COG3323, while "YbgI" denotes a sequence without the inserted COG3323 domain. Rendered fusion domains do not reflect exact sizes or locations. Color key is divided into two sets of identities (gray boxes): (top) general metabolic theme or specific annotation with bioinformatic precedent; and (bottom) COGs observed in physical clustering analysis (PCA). COGs also observed in PCA (Table 4) are shown in bold. Six minor exceptions to the top-20 rank cut-off are shown in bold with an asterisk (*): COG1196 (top 31st); COG0564 (top 23rd); COG0648 (top 25th); COG0406 (top 48th) in a fusion with COG0328; and COG0041 (top

 doi:10.20944/preprints202107.0280.v1

36th). Others observed in rep. operons but were ranked beyond the "minor exception" threshold (exceeded top-50) in PCA are shown without additional symbols, not bolded: COG0245 (116th) and COG0761 (61st). Finally, one not observed in PCA (not bolded) but was in at least one rep. operon (double asterisk, **): COG0642 (SAMN05192534_10671 of *A. persepolensis*; rep. operon, *Desulfurispirillum indicum S5*) (Data Table 7). *Note: COG4111 (NUDIX hydrolase), present in panel c (neighborhood of M. rubeus), was absent from PCA (any rank) and rep. operons, despite the fusion with COG3323 in F. nucleatum having been resolved in preceding homolog capture and literature review.*

Notable bacterial fusions included domains belonging to COG1579, COG2384, and COG0328, all three COGs having occurred independently in the top-20 ranked COGs determined through PCKFA that were also metal-binding, in addition to being observed among bacterial representative operons (COG1579, *Wolinella succinogenes ATCC 29543*; COG2384, *Ruminococcus flavefaciens Sab67*; COG0328, *Clostridia bacterium 1MN72D_59_214* (taxid: 2044939)). Although without recognizable COGs, the most common gene fusion among bacteria were TAT signals, a sequence feature neglected at the protein annotation level. While the neighborhoods of many bacterial fusions appeared very diverse (Figure 6b), 55% (11) of the top-20 co-occurring COGs of the DUF34 family (Table 4) were represented at least once across all observed neighborhoods. Additionally, genes encoding proteins involved in cofactor biosynthesis, corrinoid/siderophore/metal ion transport, metal- and metal ion stress-dependent processes, as well as DNA/RNA metabolism (e.g., de novo purine biosynthesis), were pronounced among these selected neighborhoods.

3.9. A role of the DUF34 family protein in folate synthesis is precluded by bioinformatic and experimental evidence.

GTP cyclohydrolase I activity was reported using an *in vitro* assay with the *H. pylori* DUF34 family member, HP0959, expressed in *E. coli* [34]. With the roll-out of UniRule, an automated curation and annotation transfer program, by UniProtKB, the annotation of "GTP cyclohydrolase I type 2" was subsequently electronically propagated across thousands of proteins without further substantiation or review outside of this singular publication.

The canonical GTP cyclohydrolase I (GCYHI) enzymes catalyze a complex reaction, the formation of H_2-neopterin-triphosphate (H_2NTP) from GTP, required for the first step of tetrahydrofolate (THF) synthesis in most bacteria [173–175]. H_2NTP is also a precursor to the cofactor BH_4 and to 7-cyano-deazaguanine ($preQ_0$) and intermediate in the synthesis of modified RNA and DNA bases [176,177]. Two non-orthologous protein families have been shown to harbour GCYHI activity [178]. The first, COG0302 (PF01227), was first characterized as FolE in *E. coli* K12 and is called GTP cyclohydrolase I type 1 [35]. The second named FolE2 and part of the COG1469 (PF02649) family was discovered much more recently and is called GTP cyclohydrolase I type 2 [179]. The distribution of the two families in Bacteria and Archaea vary greatly, some have FolE1, some FolE2 and some have both [4,180]. Humans encode FolE as the first step of BH_4 synthesis but no other folate enzyme [176]. A minority of bacteria are auxotrophic for THF, requiring the uptake of a folate source; hence, they do not encode any *de novo* folate biosynthesis enzymes [181]. However, as folate transporters are not present in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

most bacteria that are folate prototrophs, it follows that the *de novo* THF synthesis genes are often found to be essential in these organisms [35,36]. Folate prototrophy is common in most plants (*Viridiplantae*). although minor differences are observed among specific pathway contributors between select clades [182].

Despite the proposed role of the *H. pylori* DUF34 protein (HP0959) in folate synthesis [34], this hypothesis is not supported by the patterns of occurrence of DUF34 family members across folate auxotrophs or prototrophs. Indeed, organisms prototrophic for folate do not encode DUF34 proteins (e.g., plants), whereas folate auxotrophs, such as *M. genitalium*, do. In general, genes encoding DUF34 proteins are not essential with a few exceptions (Table 6). The gene encoding for GTP cyclohydrolase I, *folE*, is essential in *E. coli*, as is expected in most folate prototrophic bacteria [37]. The same essentiality, however, is not observed in mutants of *ybgI* in *E. coli* (Table 6). Moreover, this would imply that YbgI lacks the GTP cyclohydrolase I activity necessary to effectively compensate for the absence of *folE*, an alternative explanation to this compensory failure being that the gene had not been sufficiently expressed in previously tested conditions to do so. An additional observeration of note, however, is that even the YbgI-encoding operon, as a whole, has been reported as being non-essential in *E. coli* [183]. Although DUF34/NIF3 homologs are considered non-essential in an overwhelming majority of bacteria for which data is available (Table 6), one published case of bacterial DUF34 homolog mutant inviability was found, but it occurred in the context of using a specialized method of mutagenesis in *H. pylori* (i.e., *in vitro* mutagenesis using the Tn7 transposon) [184]. Moreover, this case stands out compared to other systems again in that the homolog is essential for *H. pylori*, a rare observation among DUF34 family members (Table 6).

Table 6. Essentiality data of DUF34 homologs.

Organism	Gene/ORF	Essentiality	Source
Acinetobacter baumannii ATCC 17978	A1S_2494	NE	Ogee
Acinetobacter sp. ADP1	ACIAD1347	NE	Ogee
Bacillus cereus ATCC 14579	YqfO/BC_4286	NE	DEG
Bacillus subtilis subsp. subtilis str. 168	BSU25170	NE	Ogee
Burkholderia cenocepacia J2315	BCAL0327	NE	Ogee
Drosophila melanogaster	FBgn0014092	NE	Ogee
Escherichia coli K-12 MG1655	YbgI/b0710	NE	DEG, Ogee
Escherichia coli O23b:H4-ST131	EC958_0819	NE	Ogee
Francisella tularensis subsp. novicida U112	FTN_1077	NE	Ogee
Haemophilus influenzae Rd KW20	HI0105	NE	Ogee
Helicobacter pylori ATCC 700392 / 26695	HP0959	E*	DEG, Ogee
Homo sapiens	NIF3L1	C*	DEG, Ogee
Methanocaldococcus jannaschii DSM 2661	MJ0927	NE	DEG
Mycobacterium tuberculosis H37Rv	Rv2230c	NE	Ogee
Mycoplasma agalactiae PG2	PE_MAG0990	E**	pDEG
Mycoplasma pulmonis UAB CTIP	MYPU_4560	NE	Ogee
Neisseria gonorrhoeae MS11	NGFG_01855	NE	Ogee
Pseudomonas aeruginosa PAO1	PA4445	NE	Ogee
Pseudomonas aeruginosa UCBPP-PA14	PA14_57740	NE	Ogee
Saccharomyces cerevisiae	NIF3	NE	Ogee

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

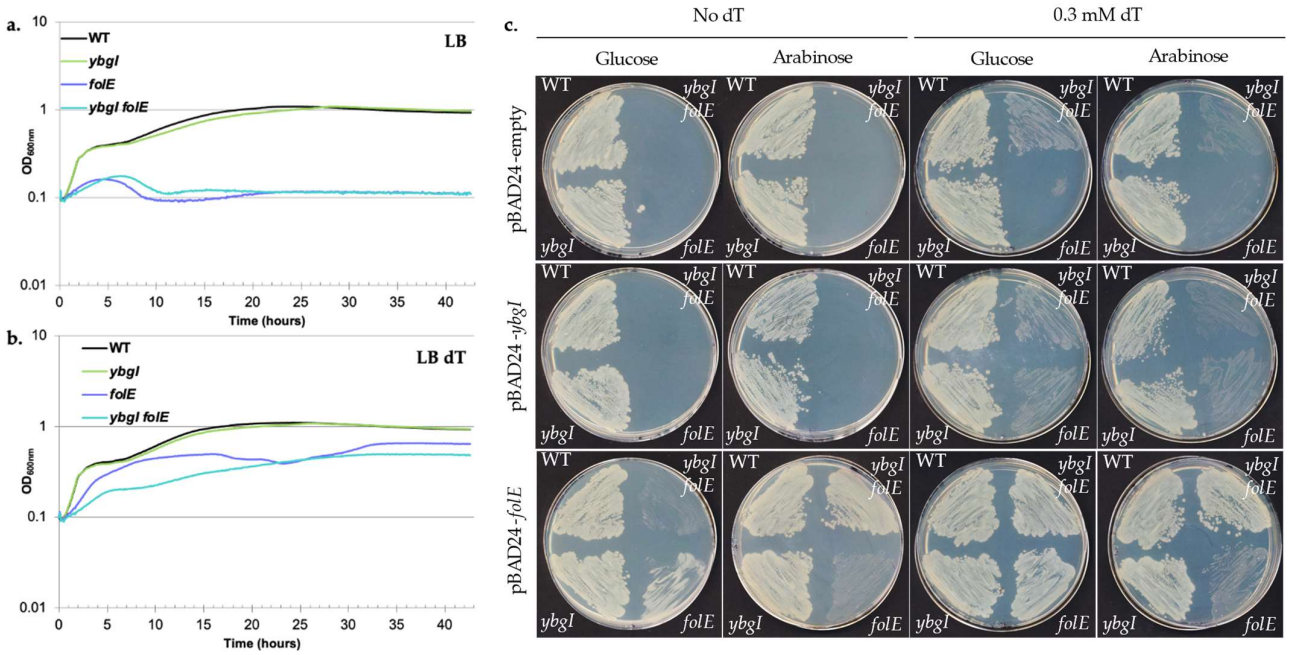
Salmonella enterica subsp. enterica serovar Typhi str. CT18	STY0751	NE	Ogee
Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344	SL1344_0692	NE	Ogee
Schizosaccharomyces pombe	SPCC126.12	NE	Ogee
Shewanella oneidensis MR-1	SO_2621	NE	Ogee
Staphylococcus aureus subsp. aureus NCTC 8325	SAOUHSC_01660	NE	Ogee
Vibrio cholerae O1 str. C6706	VC_2093	NE	Ogee
Yersinia pestis KIM	y1272	NE	Ogee

C = "conditional", E = "essential", NE = "non-essential"
**** denote that a single study directly related to the consensus determined for essentiality may not indicate evidence of essentiality with clarity or specificity***
***** indicates singular entry from pDEG, which contains predicted essential genes of sequenced Mycoplasma species; this particular gene was indicated to be an unique fusion with other genes that may be alternative sources of implicated essentiality***

With differences in essentiality considered, a series of completementation assays were performed to better illustrate the relationship of *ybgI* to *folE* and the folate biosynthetic pathway. The esssentiality of folate in *E. coli* is partially linked to the *de novo* synthesis of thymidine, as the thymidilate synthase (ThyA, [185]), that catalyzes the formation of dTMP from dUTP, uses THF as a cofactor. It was previously reported hat complementing the growth media with dT allowed a *folE* mutant of *E. coli* to grow at a low rate [177]. The *ybgI* mutant of *E. coli* had a similar growth compared to a WT in presence and in absence of dT, while *folE* mutant could only grow in presence of dT (Figure 7). Interestingly, the double mutant also required dT to grow but grew at a lower rate than the *folE* single mutant, eventually reaching the same final OD as the *folE* single mutant (Figure 7a-b). Expression of *E. coli folE in trans* complemented the essentiality of dT upon plating for, both, the single and double mutants (Figure 7c), whereas the expression of *E. coli ybgI in trans* did not complement this phenotype. It can be noted that the overexpression of *folE* in the single mutant did not fully complement the growth phenotype, while successfully doing so in the double mutant (Figure 7c, + arabinose). The WT was not impacted by the overexpression of *folE*, eliminating the hypothesis for a toxicity of high FolE levels but revelaed a genetic interation between *ybgI* and *folE* that is also observed with the better growth of the double mutant on dT compared to the single *folE* mutant. Futher studies will have to be performed to dissect this inteaction but it can be noted that FolE is a metal dependent zinc -requiring enzyme [186].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 July 2021 doi:10.20944/preprints202107.0280.v1

Figure 7. DUF34 of *E. coli*, *ybgI*, fails complementation in the absence of *folE*.



Plates were imaged after 20 hours of growth at 37 °C. (a-b) dT essentiality assay. WT, single mutants and double mutant (*folE*, *ybgI*) strains have been grown at 37 °C in LB supplemented in absence (a) or presence (b) or dT 0.3 mM. Each curve shown is averaged of across 5 replicates. (c) dT essentiality complementation assay. WT, single mutants and double mutant (*folE*, *ybgI*) strains, containing various derivatives of pBAD24 encoding for either *E. coli* YbgI or FolE, have been streaked on LB plates supplemented with Ampicillin 100 µg/mL in the presence of either 0.2% glucose for repression of the gene expression, or 0.2% arabinose for overexpression of the gene of interest, and in presence or absence of dT 0.3 mM.

4. Conclusion

In this comprehensive comparative genomic analysis of the DUF34 family, we presented a collection of arguments refuting a role in folate synthesis as a GTP cyclohydrolase I type 2 in most organisms, including the gram-negative model, *E. coli*. While we concede that it is possible the *in vitro* GTP cyclohydrolase activity described for the DUF34 member of *H. pylori*, *HP0959*, may still accurately reflect the enzyme's ability, further controls—such as site-directed mutagenesis of essential residues or *in vivo* complementation data—would be necessary to ensure that the observed activity was not related to a contaminating endogenous enzyme. Additionally, without experimental localization of the homolog's activity, the conditions required for *HP0959* to act as a GTP cyclohydrolase remain extreme for most biological contexts (i.e., very low optimal pH). In light of our analyses, the propagation of this annotation should therefore be limited until further experimental work is conducted.

The published quorum emphasizes a pleiotropic role of the DUF34 that is typical of a core molecular function. We propose that members of this family have a general metal ion insertase function that may vary in substrate and target with individual members and clades. The only member of this family with notable biochemical and structural

 doi:10.20944/preprints202107.0280.v1

characterization is the archaeal HcgD, which has been proposed to act as iron chaperone in the maturation of the iron-guanylylpyridinol (FeGP) cofactor required by [Fe]-hydrogenase [133]. The structural data presented here strongly link the DUF34 family to metal homeostasis, while the physical clustering, fusion, and co-expression data also suggest a metal link, most notably to Fe-S clusters. Proving metal insertion activity *in vivo* can be a very difficult task. For example, our group predicted that members of the COG0523 family were involved in metal insertion over 15 years ago and the experimental validation of this prediction are only now being published [187–189]. We believe that the thorough analysis presented here should guide future experimental efforts to solve this long-standing functional enigma for one of the most conserved unknowns remaining to be confidently characterized.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1. Figure S1: Word clouds generated from titles of focal and non-focal publications listed in Data Table 2; Figure S2: Secondary structural annotation by superkingdom using MultAlign-based ESPRIPT analyses; Figure S3: Complete DUF34/NIF3 homolog sequence logos across and for each superkingdom (eukaryota, archaea, bacteria) with three tiers of relative conservation; Figure S4: Phyre2 generated model of NIF3L1 (*H. sapiens*) structurally aligned with YqfO to illustrate binding pockets, residues differences within and adjacent to the active site; Figure S5: Motif differences in sequences of the D-G subgroups with and without the IPR015867 HMM profile signature annotation; Figure S6: Pairwise alignments of *B. cereus* DUF34 paralogs; Figure S7: PCKFA of COGs and COG descriptions; Figure S8: Abundances of metal ion ligand annotations across published protein structures; Figure S9: Relative abundances of metal-binding proteins per distinct ion across representative operons comparing those of bacteria and archaea to those observed in PDB; Figure S10: Relative abundances of metal-binding proteins per distinct ion as fractions of all encoded proteins across representative operons; Figure S11: Distributions of GO terms retrieved for each set of top 300 co-expressed genes of eukaryotic DUF34 family members; Figure S12: STRING network of GSEA output of DUF34 co-regulated genes of *H. sapiens*; Table S1: All resources used in systematic literature review and subsequent analyses; Table S2: Lists of strains and oligos used in growth assays; Table S3: Formatted table of all organisms, genes/proteins with published data (both focal and non-focal publications); Data Table 1: Table of search terms used and generated in the literature review/data capture process; Data Table 2: Catalog of all focal and non-focal publications collected through comprehensive literature review and data capture process of the DUF34 protein family; Data Table 3: Model organism sequences used in initial sequence alignments across and for each superkingdom exported from OrthoInspector (FASTA format); Data Table 4: Collating lists of sequences from model organisms (exported from OrthoInspector) and those acquired from comprehensive data capture and literature review (Table S3); Data Table 5: All COGs and InterPro signature profiles of the DUF34 family including paralogs and some fusions; Data Table 6: "IMG-occurrence" data sheet; Data Table 7: Physical clustering keyword frequency analysis (PCKFA) and representative operons; Data Table 8: Representative operon metal-binding protein abundance; Data Table 9: CoXPresDb (eukaryota) exports of the top 300 co-expressed genes of DUF34; Data Table 10: Co-regulated genes of *Homo sapiens* DUF34 homolog; Data Table 11: Concatenated list of sequences indicated to be possible non-canonical fusions of the DUF34 family; Data Table 12: STRING network export generated following the results of Data Table 10.

Author Contributions: Conceptualization, Valérie de Crécy-Lagard and Colbie Reed; Data curation, Colbie Reed; Formal analysis, Colbie Reed; Investigation, Geoffrey Hutinet and Valérie de Crécy-Lagard; Methodology, Geoffrey Hutinet; Project administration, Valérie de Crécy-Lagard; Visualization, Colbie Reed; Writing – original draft, Geoffrey Hutinet and Colbie Reed; Writing – review & editing, Geoffrey Hutinet, Valérie de Crécy-Lagard and Colbie Reed

Funding: "This research was funded by the National Institutes of Health grant number GM70641 to V. dC.-L and by funds from University of Florida Dept of Microbiology and Cell Sciences

Institutional Review Board Statement: Not applicable.

Acknowledgments: Early preliminary bioinformatics and initial complementation assays (not discussed, not shown) were performed by undergraduate student, Rouyi Zhang. Institutional support provided by the Department of Microbiology and Cell Science of the University of Florida. Additional appreciation is noted for the developers of UniProt for their helpful feedback and correspondence relating to current annotation statuses of proteins relevant to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Danchin, A.; Fang, G. Unknown unknowns: essential genes in quest for function. *Microb. Biotechnol.* **2016**, *9*, 530–540, doi:10.1111/1751-7915.12384.
2. Niehaus, T.D.; Thamm, A.M.; de Crécy-Lagard, V.; Hanson, A.D. Proteins of unknown biochemical function - A persistent problem and a roadmap to help overcome it. *Plant Physiol.* **2015**, *169*, pp.00959.2015, doi:10.1104/pp.15.00959.
3. de Crécy-Lagard, V.; Haas, D.; Hanson, A.D. Newly-discovered enzymes that function in metabolite damage-control. *Curr. Opin. Chem. Biol.* **2018**, *47*, 101–108, doi:10.1016/j.cbpa.2018.09.014.
4. De Crécy-Lagard, V.; Phillips, G.; Grochowski, L.L.; Yacoubi, B. El; Jenney, F.; Adams, M.W.W.; Murzin, A.G.; White, R.H. Comparative genomics guided discovery of two missing archaeal enzyme families involved in the biosynthesis of the pterin moiety of tetrahydromethanopterin and tetrahydrofolate. *ACS Chem. Biol.* **2012**, *7*, 1807–1816, doi:10.1021/cb300342u.
5. 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.; et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **2018**, *557*, 503–509, doi:10.1038/s41586-018-0124-0.
6. Kolker, E. Identification and functional analysis of "hypothetical" genes expressed in *Haemophilus influenzae*. *Nucleic Acids Res.* **2004**, *32*, 2353–2361, doi:10.1093/nar/gkh555.
7. Ghodge, S.V. Mechanistic Characterization and Function Discovery of Phosphohydrolase Enzymes from the Amidohydrolase Superfamily, Texas A&M University, 2015.
8. Tan, C.L. The absence of universally-conserved protein-coding genes. *bioRxiv* **2019**, 842633, doi:10.1101/842633.
9. Rödelsperger, C.; Prabh, N.; Sommer, R.J. New Gene Origin and Deep Taxon Phylogenomics: Opportunities and Challenges. *Trends Genet.* **2019**, *35*, 914–922, doi:10.1016/j.tig.2019.08.007.
10. Alam, M.T.; Takano, E.; Breitling, R. Prioritizing orphan proteins for further study using phylogenomics and gene expression profiles in *Streptomyces coelicolor*. *BMC Res. Notes* **2011**, *4*, 325, doi:10.1186/1756-0500-4-325.
11. Wood, V.; Lock, A.; Harris, M.A.; Rutherford, K.; Bähler, J.; Oliver, S.G. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol.* **2019**, *9*, 180241, doi:10.1098/rsob.180241.
12. Nagy, L.G.; Merényi, Z.; Hegedüs, B.; Bálint, B. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Res.* **2020**, *48*, 2209–2219, doi:10.1093/nar/gkz1241.
13. Thiaville, P.C.; Iwata-Reuyl, D.; DeCrécy-Lagard, V. Diversity of the biosynthesis pathway for threonylcarbamoyladenosine (t^6A), a universal modification of tRNA. *RNA Biol.* **2014**, *11*, 1529–1539, doi:10.4161/15476286.2014.992277.
14. El Yacoubi, B.; Hatin, I.; Deutsch, C.; Kahveci, T.; Rousset, J.-P.; Iwata-Reuyl, D.; G Murzin, A.; de Crécy-Lagard, V. A role for the universal Kae1/Qri7/YgjD (COG0533) family in tRNA modification. *EMBO J.* **2011**, *30*, 882–893, doi:10.1038/emboj.2010.363.
15. El Yacoubi, B.; Lyons, B.; Cruz, Y.; Reddy, R.; Nordin, B.; Agnelli, F.; Williamson, J.R.; Schimmel, P.; Swairjo, M.A.; De Crécy-Lagard, V. The universal YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA. *Nucleic Acids Res.* **2009**, *37*, 2894–2909, doi:10.1093/nar/gkp152.
16. Sutherland, D.R.; Abdullah, K.M.; Cyopick, P.; Mellors, A. Cleavage of the cell-surface O-sialoglycoproteins CD34, CD43,

CD44, and CD45 by a novel glycoprotease from *Pasteurella haemolytica*. *J. Immunol.* **1992**, *148*, 1458–64.

17. Nichols, C.E.; Lamb, H.K.; Thompson, P.; El Omari, K.; Lockyer, M.; Charles, I.; Hawkins, A.R.; Stammers, D.K. Crystal structure of the dimer of two essential *Salmonella typhimurium* proteins, YgjD & YeaZ and calorimetric evidence for the formation of a ternary YgjD-YeaZ-YjeE complex. *Protein Sci.* **2013**, *22*, 628–40, doi:10.1002/pro.2247.
18. Edvardson, S.; Prunetti, L.; Arraf, A.; Haas, D.; Bacusmo, J.M.; Hu, J.F.; Ta-Shma, A.; Dedon, P.C.; de Crécy-Lagard, V.; Elpeleg, O. tRNA N6-adenosine threonylcarbamoyltransferase defect due to KAE1/TCS3 (OSGEP) mutation manifest by neurodegeneration and renal tubulopathy. *Eur. J. Hum. Genet.* **2017**, *25*, 545–551, doi:10.1038/ejhg.2017.30.
19. Niehaus, T.D.; Gerdes, S.; Hodge-Hanson, K.; Zhukov, A.; Cooper, A.J.L.; ElBadawi-Sidhu, M.; Fiehn, O.; Downs, D.M.; Hanson, A.D. Genomic and experimental evidence for multiple metabolic functions in the RidA/YjgF/YER057c/UK114 (Rid) protein family. *BMC Genomics* **2015**, *16*, 382, doi:10.1186/s12864-015-1584-3.
20. Downs, D.M.; Ernst, D.C. From microbiology to cancer biology: the Rid protein family prevents cellular damage caused by endogenously generated reactive nitrogen species. *Mol. Microbiol.* **2015**, *96*, 211–9, doi:10.1111/mmi.12945.
21. Irons, J.L.; Hodge-Hanson, K.; Downs, D.M. RidA Proteins Protect against Metabolic Damage by Reactive Intermediates. *Microbiol. Mol. Biol. Rev.* **2020**, *84*, 1–28, doi:10.1128/MMBR.00024-20.
22. Lambrecht, J.A.; Schmitz, G.E.; Downs, D.M. RidA proteins prevent metabolic damage inflicted by PLP-dependent dehydratases in all domains of life. *MBio* **2013**, *4*, e00033-13, doi:10.1128/mBio.00033-13.
23. Borchert, A.J.; Ernst, D.C.; Downs, D.M. Reactive enamines and imines *in vivo*: lessons from the RidA paradigm. *Trends Biochem. Sci.* **2019**, *44*, 849–860, doi:10.1016/j.tibs.2019.04.011.
24. Tascou, S.; Uedelhoven, J.; Dixkens, C.; Nayernia, K.; Engel, W.; Burfeind, P. Isolation and characterization of a novel human gene, *NIF3L1*, and its mouse ortholog, *Nif3l1*, highly conserved from bacteria to mammals. *Cytogenet. Genome Res.* **2000**, *90*, 330–336, doi:10.1159/000056799.
25. Tascou, S.; Kang, T.W.; Trappe, R.; Engel, W.; Burfeind, P. Identification and characterization of NIF3L1 BP1, a novel cytoplasmic interaction partner of the NIF3L1 protein. *Biochem. Biophys. Res. Commun.* **2003**, *309*, 440–448, doi:10.1016/j.bbrc.2003.07.008.
26. Ladner, J.E.; Obmolova, G.; Teplyakov, A.; Howard, A.J.; Khil, P.P.; Camerini-Otero, R.D.; Gilliland, G.L. Crystal structure of *Escherichia coli* protein YbgI, a toroidal structure with a dinuclear metal site. *BMC Struct. Biol.* **2003**, *3*, 7, doi:10.1186/1472-6807-3-7.
27. Baysal, Ö.; Lai, D.; Xu, H.-H.; Siragusa, M.; Çalışkan, M.; Carimi, F.; da Silva, J.A.T.; Tör, M. A Proteomic Approach Provides New Insights into the Control of Soil-Borne Plant Pathogens by *Bacillus* Species. *PLoS One* **2013**, *8*, e53182, doi:10.1371/journal.pone.0053182.
28. Ashburner, M.; Misra, S.; Roote, J.; Lewis, S.E.; Blazej, R.; Davis, T.; Doyle, C.; Galle, R.; George, R.; Harris, N.; et al. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. *Genetics* **1999**, *153*, 179–219.
29. Geisler, R.; Bergmann, A.; Hiromi, Y.; Nüsslein-Volhard, C. cactus, a gene involved in dorsoventral pattern formation of *Drosophila*, is related to the IκB gene family of vertebrates. *Cell* **1992**, *71*, 613–621, doi:10.1016/0092-8674(92)90595-4.
30. Hadano, S.; Yanagisawa, Y.; Skaug, J.; Fichter, K.; Nasir, J.; Martindale, D.; Koop, B.F.; Scherer, S.W.; Nicholson, D.W.; Rouleau, G.A.; et al. Cloning and characterization of three novel genes, ALS2CR1, ALS2CR2, and ALS2CR3, in the juvenile amyotrophic lateral sclerosis (ALS2) critical region at chromosome 2q33-q34: Candidate genes for ALS2. *Genomics* **2001**, *71*, 200–213, doi:10.1006/geno.2000.6392.
31. Merla, G.; Howald, C.; Antonarakis, S.E.; Reymond, A. The subcellular localization of the ChoRE-binding protein, encoded by the Williams–Beuren syndrome critical region gene 14, is regulated by 14-3-3. *Hum. Mol. Genet.* **2004**, *13*, 1505–1514, doi:10.1093/hmg/ddh163.
32. Sergeeva, O. V.; Bredikhin, D.O.; Nesterchuk, M. V.; Serebryakova, M. V.; Sergiev, P. V.; Dontsova, O.A. Possible Role of

Escherichia coli Protein YbgI. *Biochem.* **2018**, *83*, 270–280, doi:10.1134/S0006297918030070.

33. Rouillard, A.D.; Gundersen, G.W.; Fernandez, N.F.; Wang, Z.; Monteiro, C.D.; McDermott, M.G.; Ma'ayan, A. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, *2016*, baw100, doi:10.1093/database/baw100.
34. Choi, H.-P.; Juarez, S.; Ciordia, S.; Fernandez, M.; Bargiela, R.; Albar, J.P.; Mazumdar, V.; Anton, B.P.; Kasif, S.; Ferrer, M.; et al. Biochemical Characterization of Hypothetical Proteins from *Helicobacter pylori*. *PLoS One* **2013**, *8*, e66605, doi:10.1371/journal.pone.0066605.
35. Adams, N.E.; Thiaville, J.J.; Proestos, J.; Juárez-Vázquez, A.L.; McCoy, A.J.; Barona-Gómez, F.; Iwata-Reuyl, D.; de Crécy-Lagard, V.; Maurelli, A.T. Promiscuous and adaptable enzymes fill "holes" in the tetrahydrofolate pathway in *Chlamydia* species. *MBio* **2014**, *5*, 1–14, doi:10.1128/mBio.01378-14.
36. De Crécy-Lagard, V. Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. *Comput. Struct. Biotechnol. J.* **2014**, *10*, 41–50, doi:10.1016/j.csbj.2014.05.008.
37. Hutchison, C.A.; Peterson, S.N.; Gill, S.R.; Cline, R.T.; White, O.; Fraser, C.M.; Smith, H.O.; Venter, J.C. Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science (80-.).* **1999**, *286*, 2165–2169, doi:10.1126/science.286.5447.2165.
38. Berman, H.M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242, doi:10.1093/nar/28.1.235.
39. Burley, S.K.; Berman, H.M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J.M.; Dutta, S.; et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **2019**, *47*, D464–D474, doi:10.1093/nar/gky1004.
40. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank. A Computer-Based Archival File for Macromolecular Structures. *Eur. J. Biochem.* **1977**, *80*, 319–324, doi:10.1111/j.1432-1033.1977.tb11885.x.
41. Schrodinger The PyMOL Molecular Graphics System, Version 2.0, Schrodinger, LLC. 2015.
42. Andreini, C.; Cavallaro, G.; Lorenzini, S.; Rosato, A. MetalPDB: A database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2013**, *41*, 312–319, doi:10.1093/nar/gks1063.
43. Putignano, V.; Rosato, A.; Banci, L.; Andreini, C. MetalPDB in 2018 : a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2018**, *46*, D459–D464, doi:10.1093/nar/gkx989.
44. Luo, H.; Lin, Y.; Gao, F.; Zhang, C.-T.; Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements: Table 1. *Nucleic Acids Res.* **2014**, *42*, D574–D580, doi:10.1093/nar/gkt1131.
45. Chen, W.-H.; Lu, G.; Chen, X.; Zhao, X.-M.; Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.* **2017**, *45*, D940–D944, doi:10.1093/nar/gkw1013.
46. Lin, Y.; Zhang, R.R. Putative essential and core-essential genes in *Mycoplasma* genomes. *Sci. Rep.* **2011**, *1*, 53, doi:10.1038/srep00053.
47. Nevers, Y.; Kress, A.; Defosset, A.; Ripp, R.; Linard, B.; Thompson, J.D.; Poch, O.; Lecompte, O. OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Res.* **2019**, *47*, D411–D418, doi:10.1093/nar/gky1068.
48. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515, doi:10.1093/nar/gky1049.
49. Landan, G.; Graur, D. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pacific Symp. Biocomput. 2008, PSB 2008* **2008**, *24*, 15–24, doi:10.1142/9789812776136_0003.
50. Penn, O.; Privman, E.; Ashkenazy, H.; Landan, G.; Graur, D.; Pupko, T. GUIDANCE: A web server for assessing alignment confidence scores. *Nucleic Acids Res.* **2010**, *38*, 23–28, doi:10.1093/nar/gkq443.

51. Sela, I.; Ashkenazy, H.; Katoh, K.; Pupko, T. GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **2015**, *43*, W7–W14, doi:10.1093/nar/gkv318.
52. Crooks, G.; Hon, G.; Chandonia, J.; Brenner, S. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190, doi:10.1101/gr.849004.1.
53. Minatani, K. Proposal for SVG2DOT: - An Interoperable Tactile Graphics Creation System Using SVG outputs from Inkscape. *Stud. Health Technol. Inform.* **2015**, *217*, 506—511.
54. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, 256–259, doi:10.1093/nar/gkz239.
55. Bethesda (MD): National Library of Medicine (US), N.C. for B.I. National Center for Biotechnology Information (NCBI)[Internet] Available online: https://www.ncbi.nlm.nih.gov/.
56. Dehal, P.S.; Joachimiak, M.P.; Price, M.N.; Bates, J.T.; Baumohl, J.K.; Chivian, D.; Friedland, G.D.; Huang, K.H.; Keller, K.; Novichkov, P.S.; et al. MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucleic Acids Res.* **2009**, *38*, 396–400, doi:10.1093/nar/gkp919.
57. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613, doi:10.1093/nar/gky1131.
58. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309–D314, doi:10.1093/nar/gky1085.
59. Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **2019**, *47*, D590–D595, doi:10.1093/nar/gky962.
60. Martinez-Guerrero, C.E.; Ciria, R.; Abreu-Goodger, C.; Moreno-Hagelsieb, G.; Merino, E. GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Res.* **2008**, *36*, 176–180, doi:10.1093/nar/gkn330.
61. Obayashi, T.; Kagaya, Y.; Aoki, Y.; Tadaka, S.; Kinoshita, K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.* **2019**, *47*, D55–D62, doi:10.1093/nar/gky1155.
62. Kustatscher, G.; Grabowski, P.; Schrader, T.A.; Passmore, J.B.; Schrader, M.; Rappsilber, J. Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **2019**, *37*, 1361–1371, doi:10.1038/s41587-019-0298-5.
63. Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, H.; Vilo, J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **2019**, *47*, W191–W198, doi:10.1093/nar/gkz369.
64. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1–13, doi:10.1093/nar/gkn923.
65. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44–57, doi:10.1038/nprot.2008.211.
66. Jiao, X.; Sherman, B.T.; Huang, D.W.; Stephens, R.; Baseler, M.W.; Lane, H.C.; Lempicki, R.A. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **2012**, *28*, 1805–1806, doi:10.1093/bioinformatics/bts251.
67. Bruford, E.A.; Braschi, B.; Denny, P.; Jones, T.E.M.; Seal, R.L.; Tweedie, S. Guidelines for human gene nomenclature. *Nat. Genet.* **2020**, *52*, 754–758, doi:10.1038/s41588-020-0669-3.
68. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K.A.; Tomita, M.; Wanner, B.L.; Mori, H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2006**, *2*, 2006.0008, doi:10.1038/msb4100050.

69. Hutinet, G.; Kot, W.; Cui, L.; Hillebrand, R.; Balamkundu, S.; Gnanakalai, S.; Neelakandan, R.; Carstens, A.B.; Fa Lui, C.; Tremblay, D.; et al. 7-Deazaguanine modifications protect phage DNA from host restriction systems. *Nat. Commun.* **2019**, *10*, 5442, doi:10.1038/s41467-019-13384-y.
70. Datsenko, K.A.; Wanner, B.L. One-step inactivation of chromosomal genes in *Escherichia coli K-12* using PCR products. *Proc. Natl. Acad. Sci.* **2000**, *97*, 6640–6645, doi:10.1073/pnas.120163297.
71. Martens, J.A.; Genereaux, J.; Saleh, A.; Brandl, C.J. Transcriptional Activation by Yeast PDR1p Is Inhibited by Its Association with NGG1p/ADA3p. *J. Biol. Chem.* **1996**, *271*, 15884–15890, doi:10.1074/jbc.271.27.15884.
72. Gou, Y.; Graff, F.; Kilian, O.; Kafkas, S.; Katuri, J.; Kim, J.H.; Marinos, N.; McEntyre, J.; Morrison, A.; Pi, X.; et al. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.* **2015**, *43*, D1042–D1048, doi:10.1093/nar/gku1061.
73. Karniely, S.; Rayzner, A.; Sass, E.; Pines, O. α-Complementation as a probe for dual localization of mitochondrial proteins. *Exp. Cell Res.* **2006**, *312*, 3835–3846, doi:10.1016/j.yexcr.2006.08.021.
74. Chen, J.; Gai, Q.; Lv, Z.; Chen, J.; Nie, Z.; Wu, X.; Zhang, Y. All-trans retinoic acid affects subcellular localization of a novel BmNIF3l protein: functional deduce and tissue distribution of *NIF3l* gene from silkworm (Bombyx mori). *Arch. Insect Biochem. Physiol.* **2010**, *74*, 217–231, doi:10.1002/arch.20364.
75. Manan, A.; Bazai, Z.; Fan, J.; Yu, H.; Li, L. The Nif3-family protein YqfO03 from *Pseudomonas syringae MB03* has multiple nematicidal activities against *Caenorhabditis elegans* and *Meloidogyne incognita*. *Int. J. Mol. Sci.* **2018**, *19*, 3915, doi:10.3390/ijms19123915.
76. Li, Y.; Xie, B.; Jiang, Z.; Yuan, B. Relationship between osteoporosis and osteoarthritis based on DNA methylation. *Int. J. Clin. Exp. Pathol.* **2019**, *12*, 3399–3407.
77. Yu, N.; Shin, S.; Lee, K.-A. First Korean Case of SATB2 -Associated 2q32-q33 Microdeletion Syndrome. *Ann. Lab. Med.* **2015**, *35*, 275, doi:10.3343/alm.2015.35.2.275.
78. Huang, S.; Li, Y.; Chen, Y.; Podsypanina, K.; Chamorro, M.; Olshen, A.B.; Desai, K. V.; Tann, A.; Petersen, D.; Green, J.E.; et al. Changes in gene expression during the development of mammary tumors in MMTV-Wnt-1 transgenic mice. *Genome Biol.* **2005**, *6*, R84, doi:10.1186/gb-2005-6-10-r84.
79. Jostes, S.V. The bromodomain inhibitor JQ1 as novel therapeutic option for type II testicular germ cell tumours: The role of SOX2 and SOX17 in regulating germ cell tumour pluripotency, Rheinischen Friedrich-Wilhelms-Universität, 2019.
80. Wu, J.; Liu, S.; Xiang, Y.; Qu, X.; Xie, Y.; Zhang, X. Bioinformatic Analysis of Circular RNA-Associated ceRNA Network Associated with Hepatocellular Carcinoma. *Biomed Res. Int.* **2019**, *2019*, 1–14, doi:10.1155/2019/8308694.
81. Quigley, D.A.; Fiorito, E.; Nord, S.; Van Loo, P.; Alnaes, G.G.; Fleischer, T.; Tost, J.; Moen Vollan, H.K.; Tramm, T.; Overgaard, J.; et al. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Mol. Oncol.* **2014**, *8*, 273–284, doi:10.1016/j.molonc.2013.11.008.
82. Kusonmano, K.; Halle, M.K.; Wik, E.; Hoivik, E.A.; Krakstad, C.; Mauland, K.K.; Tangen, I.L.; Berg, A.; Werner, H.M.J.; Trovik, J.; et al. Identification of highly connected and differentially expressed gene subnetworks in metastasizing endometrial cancer. *PLoS One* **2018**, *13*, e0206665, doi:10.1371/journal.pone.0206665.
83. Wang, M.; Li, L.; Liu, J.; Wang, J. A gene interaction network-based method to measure the common and heterogeneous mechanisms of gynecological cancer. *Mol. Med. Rep.* **2018**, *18*, 230–242, doi:10.3892/mmr.2018.8961.
84. Antoniali, G.; Serra, F.; Lirussi, L.; Tanaka, M.; D'Ambrosio, C.; Zhang, S.; Radovic, S.; Dalla, E.; Ciani, Y.; Scaloni, A.; et al. Mammalian APE1 controls miRNA processing and its interactome is linked to cancer RNA metabolism. *Nat. Commun.* **2017**, *8*, 797, doi:10.1038/s41467-017-00842-8.
85. Schneeweiss, A.; Hartkopf, A.D.; Müller, V.; Wöckel, A.; Lux, M.P.; Janni, W.; Ettl, J.; Belleville, E.; Huober, J.; Thill, M.; et al. Update Breast Cancer 2020 Part 1 – Early Breast Cancer: Consolidation of Knowledge About Known Therapies. *Geburtshilfe Frauenheilkd.* **2020**, *80*, 277–287, doi:10.1055/a-1111-2431.

86. Codrich, M.; Comelli, M.; Malfatti, M.C.; Mio, C.; Ayyildiz, D.; Zhang, C.; Kelley, M.R.; Terrosu, G.; Pucillo, C.E.M.; Tell, G. Inhibition of APE1-endonuclease activity affects cell metabolism in colon cancer cells via a p53-dependent pathway. *DNA Repair (Amst).* **2019**, *82*, 102675, doi:10.1016/j.dnarep.2019.102675.
87. Wang, L.-J.; Hsu, C.-W.; Chen, C.-C.; Liang, Y.; Chen, L.-C.; Ojcius, D.M.; Tsang, N.-M.; Hsueh, C.; Wu, C.-C.; Chang, Y.-S. Interactome-wide Analysis Identifies End-binding Protein 1 as a Crucial Component for the Speck-like Particle Formation of Activated Absence in Melanoma 2 (AIM2) Inflammasomes. *Mol. Cell. Proteomics* **2012**, *11*, 1230–1244, doi:10.1074/mcp.M112.020594.
88. Lin, C.-Y.; Ström, A.; Vega, V.B.; Kong, S.L.; Yeo, A.L.; Thomsen, J.S.; Chan, W.C.; Doray, B.; Bangarusamy, D.K.; Ramasamy, A.; et al. Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. *Genome Biol.* **2004**, *5*, R66, doi:10.1186/gb-2004-5-9-r66.
89. Xi, Y.; Riker, A.; Shevde-Samant, L.; Samant, R.; Morris, C.; Gavin, E.; Fodstad, O.; Ju, J. Global comparative gene expression analysis of melanoma patient samples, derived cell lines and corresponding tumor xenografts. *Cancer Genomics Proteomics* **2011**, *5*, 1–35, doi:10.1016/j.cortex.2009.08.003.Predictive.
90. Schrader, A.; Meyer, K.; Walther, N.; Stolz, A.; Feist, M.; Hand, E.; von Bonin, F.; Evers, M.; Kohler, C.; Shirneshan, K.; et al. Identification of a new gene regulatory circuit involving B cell receptor activated signaling using a combined analysis of experimental, clinical and global gene expression data. *Oncotarget* **2016**, *7*, 47061–47081, doi:10.18632/oncotarget.9219.
91. Uxa, S.; Bernhart, S.H.; Mages, C.F.S.; Fischer, M.; Kohler, R.; Hoffmann, S.; Stadler, P.F.; Engeland, K.; Müller, G.A. DREAM and RB cooperate to induce gene repression and cell-cycle arrest in response to p53 activation. *Nucleic Acids Res.* **2019**, *47*, 9087–9103, doi:10.1093/nar/gkz635.
92. Xiang, Y.; Zhang, C.-Q.; Huang, K. Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. *BMC Bioinformatics* **2012**, *13*, S12, doi:10.1186/1471-2105-13-S2-S12.
93. Cury, S.S.; Lapa, R.M.L.; de Mello, J.B.H.; Marchi, F.A.; Domingues, M.A.C.; Pinto, C.A.L.; Carvalho, R.F.; de Carvalho, G.B.; Kowalski, L.P.; Rogatto, S.R. Increased DSG2 plasmatic levels identified by transcriptomic-based secretome analysis is a potential prognostic biomarker in laryngeal carcinoma. *Oral Oncol.* **2020**, *103*, 104592, doi:10.1016/j.oraloncology.2020.104592.
94. Qu, S.; Shi, Q.; Xu, J.; Yi, W.; Fan, H. Weighted Gene Coexpression Network Analysis Reveals the Dynamic Transcriptome Regulation and Prognostic Biomarkers of Hepatocellular Carcinoma. *Evol. Bioinforma.* **2020**, *16*, 117693432092056, doi:10.1177/1176934320920562.
95. Chauhan, L.; Jenkins, G.D.; Bhise, N.; Feldberg, T.; Mitra-Ghosh, T.; Fridley, B.L.; Lamba, J.K. Genome-wide association analysis identified splicing single nucleotide polymorphism in CFLAR predictive of triptolide chemo-sensitivity. *BMC Genomics* **2015**, *16*, 483, doi:10.1186/s12864-015-1614-1.
96. Kalari, K.R.; Necela, B.M.; Tang, X.; Thompson, K.J.; Lau, M.; Eckel-Passow, J.E.; Kachergus, J.M.; Anderson, S.K.; Sun, Z.; Baheti, S.; et al. An Integrated Model of the Transcriptome of HER2-Positive Breast Cancer. *PLoS One* **2013**, *8*, e79298, doi:10.1371/journal.pone.0079298.
97. Ahmed, S.S.S.J.; Ahameethunisa, A.R.; Santosh, W.; Chakravarthy, S.; Kumar, S. Systems biological approach on neurological disorders: a novel molecular connectivity to aging and psychiatric diseases. *BMC Syst. Biol.* **2011**, *5*, 6, doi:10.1186/1752-0509-5-6.
98. Malan-Müller, S.; de Souza, V.B.C.; Daniels, W.M.U.; Seedat, S.; Robinson, M.D.; Hemmings, S.M.J. Shedding Light on the Transcriptomic Dark Matter in Biological Psychiatry: Role of Long Noncoding RNAs in D-cycloserine-Induced Fear Extinction in Posttraumatic Stress Disorder. *Omi. A J. Integr. Biol.* **2020**, *24*, 352–369, doi:10.1089/omi.2020.0031.
99. Qiu, L.; Liu, X. Identification of key genes involved in myocardial infarction. *Eur. J. Med. Res.* **2019**, *24*, 22, doi:10.1186/s40001-019-0381-x.
100. Lin, H. Identification of Potential coregenes in Sevoflurance induced Myocardial Energy Metabolismin Patients

Undergoing Off-pump Coronary Artery Bypass Graft Surgery using Bioinformatics analysis. *Res. Sq.* **2019**, 1–16, doi:10.21203/rs.2.17434/v1.

101. Chekouo, T.; Safo, S.E. Bayesian Integrative Analysis and Prediction with Application to Atherosclerosis Cardiovascular Disease. *arXiv* **2020**, 1–48.
102. Winer, D.A.; Winer, S.; Shen, L.; Wadia, P.P.; Yantha, J.; Paltser, G.; Tsui, H.; Wu, P.; Davidson, M.G.; Alonso, M.N.; et al. B cells promote insulin resistance through modulation of T cells and production of pathogenic IgG antibodies. *Nat. Med.* **2011**, *17*, 610–617, doi:10.1038/nm.2353.
103. Xia, B.; Li, Y.; Zhou, J.; Tian, B.; Feng, L. Identification of potential pathogenic genes associated with osteoporosis. *Bone Joint Res.* **2017**, *6*, 640–648, doi:10.1302/2046-3758.612.BJR-2017-0102.R1.
104. Thankam, F.G.; Boosani, C.S.; Dilisio, M.F.; Agrawal, D.K. MicroRNAs associated with inflammation in shoulder tendinopathy and glenohumeral arthritis. *Mol. Cell. Biochem.* **2018**, *437*, 81–97, doi:10.1007/s11010-017-3097-7.
105. Wang, J.C.; Ramaswami, G.; Geschwind, D.H. Gene co-expression network analysis in human spinal cord highlights mechanisms underlying amyotrophic lateral sclerosis susceptibility. *bioRxiv* **2020**.
106. Lv, L.; Zhang, D.; Hua, P.; Yang, S. The glial-specific hypermethylated 3′ untranslated region of histone deacetylase 1 may modulates several signal pathways in Alzheimer's disease. *Life Sci.* **2021**, *265*, 118760, doi:10.1016/j.lfs.2020.118760.
107. Tian, Y.; Voineagu, I.; Paşca, S.P.; Won, H.; Chandran, V.; Horvath, S.; Dolmetsch, R.E.; Geschwind, D.H. Alteration in basal and depolarization induced transcriptional network in iPSC derived neurons from Timothy syndrome. *Genome Med.* **2014**, *6*, 75, doi:10.1186/s13073-014-0075-5.
108. Akiyama, H.; Fujisawa, N.; Tashiro, Y.; Takanabe, N.; Sugiyama, A.; Tashiro, F. The Role of Transcriptional Corepressor Nif3l1 in Early Stage of Neural Differentiation via Cooperation with Trip15/CSN2. *J. Biol. Chem.* **2003**, *278*, 10752–10762, doi:10.1074/jbc.M209856200.
109. Duzyj, C.M.; Paidas, M.J.; Jebailey, L.; Huang, J.; Barnea, E.R. PreImplantation factor (PIF*) promotes embryotrophic and neuroprotective decidual genes: effect negated by epidermal growth factor. *J. Neurodev. Disord.* **2014**, *6*, 36, doi:10.1186/1866-1955-6-36.
110. Akiyama, H. Implication of Trip15/CSN2 in early stage of neuronal differentiation of P19 embryonal carcinoma cells. *Dev. Brain Res.* **2003**, *140*, 45–56, doi:10.1016/S0165-3806(02)00574-6.
111. Boswell, W.T.; Boswell, M.; Walter, D.J.; Navarro, K.L.; Chang, J.; Lu, Y.; Savage, M.G.; Shen, J.; Walter, R.B. Exposure to 4100 K fluorescent light elicits sex specific transcriptional responses in *Xiphophorus maculatus* skin. *Comp. Biochem. Physiol. Part C Toxicol. Pharmacol.* **2018**, *208*, 96–104, doi:10.1016/j.cbpc.2017.09.008.
112. Zuccotti, M.; Merico, V.; Sacchi, L.; Bellone, M.; Brink, T.C.; Bellazzi, R.; Stefanelli, M.; Redi, C.; Garagna, S.; Adjaye, J. Maternal Oct-4 is a potential key regulator of the developmental competence of mouse oocytes. *BMC Dev. Biol.* **2008**, *8*, 97, doi:10.1186/1471-213X-8-97.
113. Skottman, H.; Mikkola, M.; Lundin, K.; Olsson, C.; Strömberg, A.-M.; Tuuri, T.; Otonkoski, T.; Hovatta, O.; Lahesmaa, R. Gene Expression Signatures of Seven Individual Human Embryonic Stem Cell Lines. *Stem Cells* **2005**, *23*, 1343–1356, doi:10.1634/stemcells.2004-0341.
114. Yan, L.; Yao, X.; Bachvarov, D.; Saifudeen, Z.; El-Dahr, S.S. Genome-wide analysis of gestational gene-environment interactions in the developing kidney. *Physiol. Genomics* **2014**, *46*, 655–670, doi:10.1152/physiolgenomics.00035.2014.
115. Gangaiah, D.; Labandeira-Rey, M.; Zhang, X.; Fortney, K.R.; Ellinger, S.; Zwickl, B.; Baker, B.; Liu, Y.; Janowicz, D.M.; Katz, B.P.; et al. *Haemophilus ducreyi* Hfq Contributes to Virulence Gene Regulation as Cells Enter Stationary Phase. *MBio* **2014**, *5*, 1–13, doi:10.1128/mBio.01081-13.
116. Labandeira-Rey, M.; Mock, J.R.; Hansen, E.J. Regulation of Expression of the *Haemophilus ducreyi* LspB and LspA2 Proteins by CpxR. *Infect. Immun.* **2009**, *77*, 3402–3411, doi:10.1128/IAI.00292-09.
117. Spinola, S.M.; Fortney, K.R.; Baker, B.; Janowicz, D.M.; Zwickl, B.; Katz, B.P.; Blick, R.J.; Munson, R.S. Activation of the

CpxRA System by Deletion of *cpxA* Impairs the Ability of *Haemophilus ducreyi* To Infect Humans. *Infect. Immun.* **2010**, *78*, 3898–3904, doi:10.1128/IAI.00432-10.

118. Rahmani-Badi, A.; Sepehr, S.; Fallahi, H.; Heidari-Keshel, S. Erratum: Exposure of *E. coli* to DNA-Methylating Agents Impairs Biofilm Formation and Invasion of Eukaryotic Cells via Down Regulation of the N-Acetylneuraminate Lyase NanA. *Front. Microbiol.* **2016**, *7*, 1–13, doi:10.3389/fmicb.2015.00383.
119. Dunman, P.M.; Murphy, E.; Haney, S.; Palacios, D.; Tucker-Kellogg, G.; Wu, S.; Brown, E.L.; Zagursky, R.J.; Shlaes, D.; Projan, S.J. Transcription Profiling-Based Identification of *Staphylococcus aureus* Genes Regulated by the *agr* and/or *sarA* Loci. *J. Bacteriol.* **2001**, *183*, 7341–7353, doi:10.1128/JB.183.24.7341-7353.2001.
120. Pereira, L.E.; Tsang, J.; Mrázek, J.; Hoover, T.R. The zinc-ribbon domain of *Helicobacter pylori HP0958*: requirement for RpoN accumulation and possible roles of homologs in other bacteria. *Microb. Inform. Exp.* **2011**, *1*, 8, doi:10.1186/2042-5783-1-8.
121. Pomposiello, P.J.; Bennik, M.H.J.; Demple, B. Genome-Wide Transcriptional Profiling of the *Escherichia coli* Responses to Superoxide Stress and Sodium Salicylate. *J. Bacteriol.* **2001**, *183*, 3890–3902, doi:10.1128/JB.183.13.3890-3902.2001.
122. Peng, C.; Andersen, B.; Arshid, S.; Larsen, M.R.; Albergaria, H.; Lametsch, R.; Arneborg, N. Proteomics insights into the responses of *Saccharomyces cerevisiae* during mixed-culture alcoholic fermentation with *Lachancea thermotolerans*. *FEMS Microbiol. Ecol.* **2019**, *95*, 1–16, doi:10.1093/femsec/fiz126.
123. Liang, W.; Bi, Y.; Wang, H.; Dong, S.; Li, K.; Li, J. Gene Expression Profiling of *Clostridium botulinum* under Heat Shock Stress. *Biomed Res. Int.* **2013**, *2013*, 1–8, doi:10.1155/2013/760904.
124. Selby, K.; Mascher, G.; Somervuo, P.; Lindström, M.; Korkeala, H. Heat shock and prolonged heat stress attenuate neurotoxin and sporulation gene expression in group I *Clostridium botulinum strain ATCC 3502*. *PLoS One* **2017**, *12*, e0176944, doi:10.1371/journal.pone.0176944.
125. Anderson, K.L.; Roux, C.M.; Olson, M.W.; Luong, T.T.; Lee, C.Y.; Olson, R.; Dunman, P.M. Characterizing the effects of inorganic acid and alkaline shock on the *Staphylococcus aureus* transcriptome and messenger RNA turnover. *FEMS Immunol. Med. Microbiol.* **2010**, *60*, 208–250, doi:10.1111/j.1574-695X.2010.00736.x.
126. Belvin, B.R.; Gui, Q.; Hutcherson, J.A.; Lewis, J.P. The *Porphyromonas gingivalis* hybrid cluster protein Hcp is required for growth with nitrite and survival with host cells. *Infect. Immun.* **2019**, *87*, doi:10.1128/IAI.00572-18.
127. Aurass, P.; Pless, B.; Rydzewski, K.; Holland, G.; Bannert, N.; Flieger, A. *bdhA-patD* Operon as a Virulence Determinant, Revealed by a Novel Large-Scale Approach for Identification of *Legionella pneumophila* Mutants Defective for Amoeba Infection. *Appl. Environ. Microbiol.* **2009**, *75*, 4506–4515, doi:10.1128/AEM.00187-09.
128. Zhao, W.; Caro, F.; Robins, W.; Mekalanos, J.J. Antagonism toward the intestinal microbiota and its effect on *Vibrio cholerae* virulence. *Science (80-.).* **2018**, *359*, 210–213, doi:10.1126/science.aap8775.
129. Shulami, S.; Shenker, O.; Langut, Y.; Lavid, N.; Gat, O.; Zaide, G.; Zehavi, A.; Sonenshein, A.L.; Shoham, Y. Multiple Regulatory Mechanisms Control the Expression of the *Geobacillus stearothermophilus* Gene for Extracellular Xylanase. *J. Biol. Chem.* **2014**, *289*, 25957–25975, doi:10.1074/jbc.M114.592873.
130. Ogura, M.; Sato, T.; Abe, K. *Bacillus subtilis YlxR*, Which Is Involved in Glucose-Responsive Metabolic Changes, Regulates Expression of *tsaD* for Protein Quality Control of Pyruvate Dehydrogenase. *Front. Microbiol.* **2019**, *10*, 1–15, doi:10.3389/fmicb.2019.00923.
131. Chen, S.-C.; Huang, C.-H.; Yang, C.S.; Kuan, S.-M.; Lin, C.-T.; Chou, S.-H.; Chen, Y. Crystal Structure of a Conserved Hypothetical Protein MJ0927 from *Methanocaldococcus jannaschii* Reveals a Novel Quaternary Assembly in the Nif3 Family. *Biomed Res. Int.* **2014**, *2014*, 1–8, doi:10.1155/2014/171263.
132. Tomoike, F.; Wakamatsu, T.; Nakagawa, N.; Kuramitsu, S.; Masui, R. Crystal structure of the conserved hypothetical protein TTHA1606 from *Thermus thermophilus HB8*. *Proteins Struct. Funct. Bioinforma.* **2009**, *76*, 244–248, doi:10.1002/prot.22397.

133. Fujishiro, T.; Ermler, U.; Shima, S. A possible iron delivery function of the dinuclear iron center of HcgD in [Fe]-hydrogenase cofactor biosynthesis. *FEBS Lett.* **2014**, *588*, 2789–2793, doi:10.1016/j.febslet.2014.05.059.
134. Lie, T.J.; Costa, K.C.; Pak, D.; Sakesan, V.; Leigh, J.A. Phenotypic evidence that the function of the [Fe]-hydrogenase Hmd in Methanococcus maripaludis requires seven hcg (hmd co-occurring genes) but not hmdII. *FEMS Microbiol. Lett.* **2013**, *343*, 156–160, doi:10.1111/1574-6968.12141.
135. Godsey, M.H.; Minasov, G.; Shuvalova, L.; Brunzelle, J.S.; Vorontsov, I.I.; Collart, F.R.; Anderson, W.F. The 2.2 Å resolution crystal structure of *Bacillus cereus* Nif3-family protein YqfO reveals a conserved dimetal-binding motif and a regulatory domain. *Protein Sci.* **2007**, *16*, 1285–1293, doi:10.1110/ps.062674007.
136. Lamba, J.K.; Feldberg, T.; Ghosh, T.M.; Bhise, N.; Fridley, B. Abstract 2214: Genome-wide association analysis identified genetic markers associated with triptolide cellular sensitivity using HapMap LCLs as model system. In Proceedings of the Experimental and Molecular Therapeutics; American Association for Cancer Research, 2013; Vol. 73, pp. 2214–2214.
137. Malik, A.; Pande, K.; Kumar, A.; Vemula, A.; Chandramohan, M.R.V. Finding Pathogenic nsSNP's and their structural effect on COPS2 using Molecular Dynamic Approach. *bioRxiv* **2020**, doi:10.1101/2020.10.12.333252.
138. Kuan, S.-M.; Chen, H.-C.; Huang, C.-H.; Chang, C.-H.; Chen, S.-C.; Yang, C.S.; Chen, Y. Crystallization and preliminary X-ray diffraction analysis of the Nif3-family protein MJ0927 from *Methanocaldococcus jannaschii*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2013**, *69*, 80–82, doi:10.1107/S1744309112049408.
139. Saikatendu, K.S.; Zhang, X.; Kinch, L.; Leybourne, M.; Grishin, N. V.; Zhang, H. Structure of a conserved hypothetical protein SA1388 from *S. aureus* reveals a capped hexameric toroid with two PII domain lids and a dinuclear metal center. *BMC Struct. Biol.* **2006**, *6*, 27, doi:10.1186/1472-6807-6-27.
140. Constantine, K.L.; Krystek, S.R.; Healy, M.D.; Doyle, M.L.; Siemers, N.O.; Thanassi, J.; Yan, N.; Xie, D.; Goldfarb, V.; Yanchunas, J.; et al. Structural and functional characterization of CFE88: Evidence that a conserved and essential bacterial protein is a methyltransferase. *Protein Sci.* **2009**, *14*, 1472–1484, doi:10.1110/ps.051389605.
141. Qijing, G.; Zhang, Y. NIF3 类超家族蛋白. *Chinese J. Cell Biol.* **2007**, *29*, 816–820.
142. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **1988**, *16*, 10881–10890, doi:10.1093/nar/16.22.10881.
143. Robert, X.; Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **2014**, *42*, 320–324, doi:10.1093/nar/gku316.
144. Yang, J.; Li, Q.; Yang, H.; Yan, L.; Yang, L.; Yu, L. Overexpression of human CUTA isoform 2 enhances the cytotoxicity of copper to HeLa cells. *Acta Biochim. Pol.* **2008**, *55*, 411–415, doi:10.18388/abp.2008_3089.
145. Gupta, S.D.; Lee, B.T.O.; Camakaris, J.; Wu, H.C. Identification of *cutC* and *cutF* (*nlpE*) genes involved in copper tolerance in *Escherichia coli*. *J. Bacteriol.* **1995**, *177*, 4207–4215, doi:10.1128/jb.177.15.4207-4215.1995.
146. Fong, S.T.; Camakaris, J.; Lee, B.T. Molecular genetics of a chromosomal locus involved in copper tolerance in *Escherichia coli K-12*. *Mol. Microbiol.* **1995**, *15*, 1127–37, doi:10.1111/j.1365-2958.1995.tb02286.x.
147. Tanaka, Y.; Tsumoto, K.; Nakanishi, T.; Yasutake, Y.; Sakai, N.; Yao, M.; Tanaka, I.; Kumagai, I. Structural implications for heavy metal-induced reversible assembly and aggregation of a protein: the case of *Pyrococcus horikoshii* CutA. *FEBS Lett.* **2004**, *556*, 167–74, doi:10.1016/s0014-5793(03)01402-9.
148. Odermatt, A.; Solioz, M. Two trans-acting metalloregulatory proteins controlling expression of the copper-ATPases of *Enterococcus hirae*. *J. Biol. Chem.* **1995**, *270*, 4349–54, doi:10.1074/jbc.270.9.4349.
149. Rensing, C.; Franke, S. Copper Homeostasis in *Escherichia coli* and Other *Enterobacteriaceae*. *EcoSal Plus* **2007**, *2*, ecosalplus.5.4.4.1, doi:10.1128/ecosalplus.5.4.4.1.
150. Bagautdinov, B. The structures of the CutA1 proteins from *Thermus thermophilus* and *Pyrococcus horikoshii*: characterization of metal-binding sites and metal-induced assembly. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2014**, *70*, 404–413, doi:10.1107/S2053230X14003422.

151. Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60*, 2256–2268, doi:10.1107/S0907444904026460.

152. Siltberg-Liberles, J.; Martinez, A. Searching distant homologs of the regulatory ACT domain in phenylalanine hydroxylase. *Amino Acids* **2009**, *36*, 235–249, doi:10.1007/s00726-008-0057-2.

153. Arnesano, F.; Banci, L.; Benvenuti, M.; Bertini, I.; Calderone, V.; Mangani, S.; Viezzoli, M.S. The Evolutionarily Conserved Trimeric Structure of CutA1 Proteins Suggests a Role in Signal Transduction. *J. Biol. Chem.* **2003**, *278*, 45999–46006, doi:10.1074/jbc.M304398200.

154. Forchhammer, K.; Lüddecke, J. Sensory properties of the P_{II} signalling protein family. *FEBS J.* **2016**, *283*, 425–437, doi:10.1111/febs.13584.

155. Selim, K.A.; Tremiño, L.; Marco-Marín, C.; Alva, V.; Espinosa, J.; Contreras, A.; Hartmann, M.D.; Forchhammer, K.; Rubio, V. Functional and structural characterization of PII-like protein CutA does not support involvement in heavy metal tolerance and hints at a small-molecule carrying/signaling role. *FEBS J.* **2021**, *288*, 1142–1162, doi:10.1111/febs.15464.

156. Selim, K.A.; Haffner, M. Heavy Metal Stress Alters the Response of the Unicellular Cyanobacterium *Synechococcus elongatus PCC 7942* to Nitrogen Starvation. *Life* **2020**, *10*, 275, doi:10.3390/life10110275.

157. Koga, R.; Matsumoto, A.; Kouzuma, A.; Watanabe, K. Identification of an extracytoplasmic function sigma factor that facilitates c-type cytochrome maturation and current generation under electrolyte-flow conditions in *Shewanella oneidensis MR-1*. *Environ. Microbiol.* **2020**, *22*, 3671–3684, doi:10.1111/1462-2920.15131.

158. Manina, G.; Bellinzoni, M.; Pasca, M.R.; Neres, J.; Milano, A.; De Jesus Lopes Ribeiro, A.L.; Buroni, S.; Škovierová, H.; Dianišková, P.; Mikušová, K.; et al. Biological and structural characterization of the *Mycobacterium smegmatis* nitroreductase NfnB, and its role in benzothiazinone resistance. *Mol. Microbiol.* **2010**, *77*, 1172–1185, doi:10.1111/j.1365-2958.2010.07277.x.

159. Markowitz, V.M.; Chen, I.M.A.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Anderson, I.; Lykidis, A.; Mavromatis, K.; et al. The integrated microbial genomes system: An expanding comparative analysis resource. *Nucleic Acids Res.* **2009**, *38*, 382–390, doi:10.1093/nar/gkp887.

160. Grigoriev, I. V.; Nordberg, H.; Shabalov, I.; Aerts, A.; Cantor, M.; Goodstein, D.; Kuo, A.; Minovitsky, S.; Nikitin, R.; Ohm, R.A.; et al. The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* **2012**, *40*, 26–32, doi:10.1093/nar/gkr947.

161. Nordberg, H.; Cantor, M.; Dusheyko, S.; Hua, S.; Poliakov, A.; Shabalov, I.; Smirnova, T.; Grigoriev, I. V.; Dubchak, I. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **2014**, *42*, 26–31, doi:10.1093/nar/gkt1069.

162. Waldron, K.J.; Rutherford, J.C.; Ford, D.; Robinson, N.J. Metalloproteins and metal sensing. *Nature* **2009**, *460*, 823–830, doi:10.1038/nature08300.

163. Ryan, K.A.; Karim, N.; Worku, M.; Moore, S.A.; Penn, C.W.; O'Toole, P.W. HP0958 is an essential motility gene in *Helicobacter pylori*. *FEMS Microbiol. Lett.* **2005**, *248*, 47–55, doi:10.1016/j.femsle.2005.05.022.

164. Kumar, A.; Karthikeyan, S. Crystal structure of the MSMEG_4306 gene product from *Mycobacterium smegmatis*. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2018**, *74*, 166–173, doi:10.1107/S2053230X18002236.

165. Barta, M.L.; Battaile, K.P.; Lovell, S.; Hefty, P.S. Hypothetical protein CT398 (CdsZ) interacts with σ54 (RpoN)-holoenzyme and the type III secretion export apparatus in *Chlamydia trachomatis*. *Protein Sci.* **2015**, *24*, 1617–1632, doi:10.1002/pro.2746.

166. Rees, W.D.; Lorenzo-Leal, A.C.; Steiner, T.S.; Bach, H. *Mycobacterium avium Subspecies paratuberculosis* Infects and Replicates within Human Monocyte-Derived Dendritic Cells. *Microorganisms* **2020**, *8*, 994, doi:10.3390/microorganisms8070994.

167. Kim, W.S.; Shin, M.-K.; Shin, S.J. MAP1981c, a Putative Nucleic Acid-Binding Protein, Produced by *Mycobacterium avium subsp. paratuberculosis*, Induces Maturation of Dendritic Cells and Th1-Polarization. *Front. Cell. Infect. Microbiol.* **2018**, *8*, doi:10.3389/fcimb.2018.00206.

168. Sassetti, C.M.; Boyd, D.H.; Rubin, E.J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol.*

Microbiol. **2003**, *48*, 77–84, doi:10.1046/j.1365-2958.2003.03425.x.

169. Lu, S.; Wang, J.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz, D.I.; Marchler, G.H.; Song, J.S.; et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **2020**, *48*, D265–D268, doi:10.1093/nar/gkz991.
170. Yanai, I.; Hunter, C.P. Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. *Genome Res.* **2009**, *19*, 2214–20, doi:10.1101/gr.093815.109.
171. Sheftel, A.D.; Wilbrecht, C.; Stehling, O.; Niggemeyer, B.; Elsässer, H.P.; Mühlenhoff, U.; Lill, R. The human mitochondrial ISCA1, ISCA2, and IBA57 proteins are required for [4Fe-4S] protein maturation. *Mol. Biol. Cell* **2012**, *23*, 1157–1166, doi:10.1091/mbc.E11-09-0772.
172. Cai, K.; Markley, J. NMR as a Tool to Investigate the Processes of Mitochondrial and Cytosolic Iron-Sulfur Cluster Biosynthesis. *Molecules* **2018**, *23*, 2213, doi:10.3390/molecules23092213.
173. Katzemeier, G.; Schmid, C.; Kellermann, J.; Lottspeich, F.; Bacher, A. Biosynthesis of Tetrahydrofolate. Sequence of GTP Cyclohydrolase I from *Escherichia coli*. *Biol. Chem. Hoppe. Seyler.* **1991**, *372*, 991–998, doi:10.1515/bchm3.1991.372.2.991.
174. Cossins, E.A.; Chen, L. Folates and one-carbon metabolism in plants and fungi. *Phytochemistry* **1997**, *45*, 437–452, doi:10.1016/S0031-9422(96)00833-3.
175. Burg, A.W.; Brown, G.M. The biosynthesis of folic acid. 8. Purification and properties of the enzyme that catalyzes the production of formate from carbon atom 8 of guanosine triphosphate. *J. Biol. Chem.* **1968**, *243*, 2349–58.
176. Thöny, B.; Auerbach, G.; Blau, N. Tetrahydrobiopterin biosynthesis, regeneration and functions. *Biochem. J.* **2000**, *347*, 1–16, doi:10.1042/0264-6021:3470001.
177. Phillips, G.; El Yacoubi, B.; Lyons, B.; Alvarez, S.; Iwata-Reuyl, D.; De Crécy-Lagard, V. Biosynthesis of 7-deazaguanosine-modified tRNA nucleosides: A new role for GTP cyclohydrolase I. *J. Bacteriol.* **2008**, *190*, 7876–7884, doi:10.1128/JB.00874-08.
178. El Yacoubi, B.; Bonnett, S.; Anderson, J.N.; Swairjo, M.A.; Iwata-Reuyl, D.; De Crécy-Lagard, V. Discovery of a new prokaryotic type I GTP cyclohydrolase family. *J. Biol. Chem.* **2006**, *281*, 37586–37593, doi:10.1074/jbc.M607114200.
179. Paranagama, N.; Bonnett, S.A.; Alvarez, J.; Luthra, A.; Stec, B.; Gustafson, A.; Iwata-Reuyl, D.; Swairjo, M.A. Mechanism and catalytic strategy of the prokaryotic-specific GTP cyclohydrolase-IB. *Biochem. J.* **2017**, *474*, 1017–1039, doi:10.1042/BCJ20161025.
180. Sankaran, B.; Bonnett, S.A.; Shah, K.; Gabriel, S.; Reddy, R.; Schimmel, P.; Rodionov, D.A.; De Crécy-Lagard, V.; Helmann, J.D.; Iwata-Reuyl, D.; et al. Zinc-independent folate biosynthesis: Genetic, biochemical, and structural investigations reveal new metal dependence for GTP cyclohydrolase IB. *J. Bacteriol.* **2009**, *191*, 6936–6949, doi:10.1128/JB.00287-09.
181. de Crécy-Lagard, V.; El Yacoubi, B.; de la Garza, R.D.; Noiriel, A.; Hanson, A.D. Comparative genomics of bacterial and plant folate synthesis and salvage: Predictions and validations. *BMC Genomics* **2007**, *8*, 1–15, doi:10.1186/1471-2164-8-245.
182. Gorelova, V.; Bastien, O.; De Clerck, O.; Lespinats, S.; Rébeillé, F.; Van Der Straeten, D. Evolution of folate biosynthesis and metabolism across algae and land plant lineages. *Sci. Rep.* **2019**, *9*, 5731, doi:10.1038/s41598-019-42146-5.
183. Gerdes, S.Y.; Scholle, M.D.; Campbell, J.W.; Balázsi, G.; Ravasz, E.; Daugherty, M.D.; Somera, A.L.; Kyrpides, N.C.; Anderson, I.; Gelfand, M.S.; et al. Experimental determination and system level analysis of essential genes in *Escherichia coli MG1655*. *J. Bacteriol.* **2003**, *185*, 5673–84, doi:10.1128/JB.185.19.5673-5684.2003.
184. Salama, N.R.; Shepherd, B.; Falkow, S. Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.* **2004**, *186*, 7926–35, doi:10.1128/JB.186.23.7926-7935.2004.
185. Wahba, A.J.; Friedkin, M. The Enzymatic Synthesis of Thymidylate. *J. Biol. Chem.* **1962**, *237*, 3794–3801, doi:10.1016/S0021-9258(19)84524-6.
186. Rebelo, J.; Auerbach, G.; Bader, G.; Bracher, A.; Nar, H.; Hösl, C.; Schramek, N.; Kaiser, J.; Bacher, A.; Huber, R.; et al. Biosynthesis of Pteridines. Reaction Mechanism of GTP Cyclohydrolase I. *J. Mol. Biol.* **2003**, *326*, 503–516, doi:10.1016/S0022-2836(02)01303-7.

187. Chandrangsu, P.; Huang, X.; Gaballa, A.; Helmann, J.D. *Bacillus subtilis* FolE is sustained by the ZagA zinc metallochaperone and the alarmone ZTP under conditions of zinc deficiency. *Mol. Microbiol.* **2019**, *112*, 751–765, doi:10.1111/mmi.14314.

188. Blaby-Haas, C.E.; Flood, J.A.; Crécy-Lagard, V. de; Zamble, D.B. YeiR: a metal-binding GTPase from *Escherichia coli* involved in metal homeostasis. *Metallomics* **2012**, *4*, 488, doi:10.1039/c2mt20012k.

189. Sydor, A.M.; Jost, M.; Ryan, K.S.; Turo, K.E.; Douglas, C.D.; Drennan, C.L.; Zamble, D.B. Metal Binding Properties of *Escherichia coli* YjiA, a Member of the Metal Homeostasis-Associated COG0523 Family of GTPases. *Biochemistry* **2013**, *52*, 1788–1801, doi:10.1021/bi301600z.