RESEARCH PAPER

RNA Biology 11:4, 360–372; April 2014; © 2014 Landes Bioscience

A highly conserved family of domains related to the DNA-glycosylase fold helps predict multiple novel pathways for RNA modifications

A Maxwell Burroughs* and L Aravind

National Center for Biotechnology Information; National Library of Medicine; National Institutes of Health; Bethesda, MD USA

Keywords: DNA glycosylase, NEMF, Tae2, caliban, FbpA, fibronectin-binding, tRNA 4-thiouridylation, IscS, TusA, base modification

Abbreviations: Fpg, formamidopyrimidine; LUCA, Last Universal Common Ancestor; LECA, Last Eukaryotic Common Ancestor; FMN-DG domain, Formamidopyrimidine, MutM, and Nei/EndoVIII DNA glycosylase; FbpA, fibronectin-binding protein; BER, Base Excision Repair; NFACT domain, NEMF, bacterial FbpA-like proteins, Caliban, and Tae2 domain; NFACT-N domain, NFACT N-terminal domain; NFACT-R domain; NFACT RNA-binding domain; NFACT-C domain-NFACT C-terminal domain; ZnK, zinc knuckle; TGT, transglycosylase; RQC, ribosomal quality control; IRES, Internal Ribosomal Entry Site

A protein family including mammalian NEMF, *Drosophila* caliban, yeast Tae2, and bacterial FpbA-like proteins was first defined over a decade ago and found to be universally distributed across the three domains/superkingdoms of life. Since its initial characterization, this family of proteins has been tantalizingly linked to a wide range of biochemical functions. Tapping the enormous wealth of genome information that has accumulated since the initial characterization of these proteins, we perform a detailed computational analysis of the family, identifying multiple conserved domains. Domains identified include an enzymatic domain related to the formamidopyrimidine (Fpg), MutM, and Nei/EndoVIII family of DNA glycosylases, a novel, predicted RNA-binding domain, and a domain potentially mediating protein–protein interactions. Through this characterization, we predict that the DNA glycosylase-like domain catalytically operates on double-stranded RNA, as part of a hitherto unknown base modification mechanism that probably targets rRNAs. At least in archaea, and possibly eukaryotes, this pathway might additionally include the AMMECR1 family of proteins. The predicted RNA-binding domain associated with this family is also observed in distinct architectural contexts in other proteins across phylogenetically diverse prokaryotes. Here it is predicted to play a key role in a new pathway for tRNA 4-thiouridylation along with TusA-like sulfur transfer proteins.

Introduction

Over 10 y ago, a gene family conserved across all three superkingdoms of life was identified[1] and determined to contain two tandem copies of the nucleic acid-binding helix-hairpin-helix (HhH) domain.[2-4] Based on its phyletic distribution, this family was traced back to the last universal common ancestor (LUCA) of life. The HhH domain pair and its phyletic pattern suggested a general functional role for the family in a nucleic acid-related role in universally conserved pathways: either RNA-metabolism in the context of translation or DNA repair or recombination.[1] Establishment of a relationship between the HhH-domain pair in these proteins to the one found in ribosomal proteins of the S13/S18 family supported the former function in particular.

Since the initial characterization of this family, which includes the Tae2 protein from *Saccharomyces cerevisiae*, the Caliban (Clbn) protein from *Drosophila melanogaster*, the mammalian NEMF proteins, and the so-called fibronectin-binding (FbpA-like) proteins from bacteria, several studies have resulted in attribution of a wide range of functional roles for these proteins. These include fibronectin binding in certain pathogenic bacteria,[5-10] a core component of the ribosome-associating, co-translational degradation complex RQC in yeast,[11,12] regulation of the DNA-damage response in *Drosophila*,[13] and mediation of nuclear export in *Drosophila* and human.[14] Given the disparate nature of these findings, we decided to revisit this gene family using state-of-the-art techniques in sequence analysis and comparative genomics while tapping the wealth of new information that has accumulated in the years since its initial characterization. Here we identify and characterize the distinct globular domains conserved across all members of the gene family in addition to the HhH domain pair. One of these domains is predicted to be an

*Correspondence to: A Maxwell Burroughs; Email: max.burroughs@nih.gov
Submitted: 10/28/2013; Revised: 02/04/2014; Accepted: 02/20/2014; Published Online: 03/05/2014
http://dx.doi.org/10.4161/rna.28302

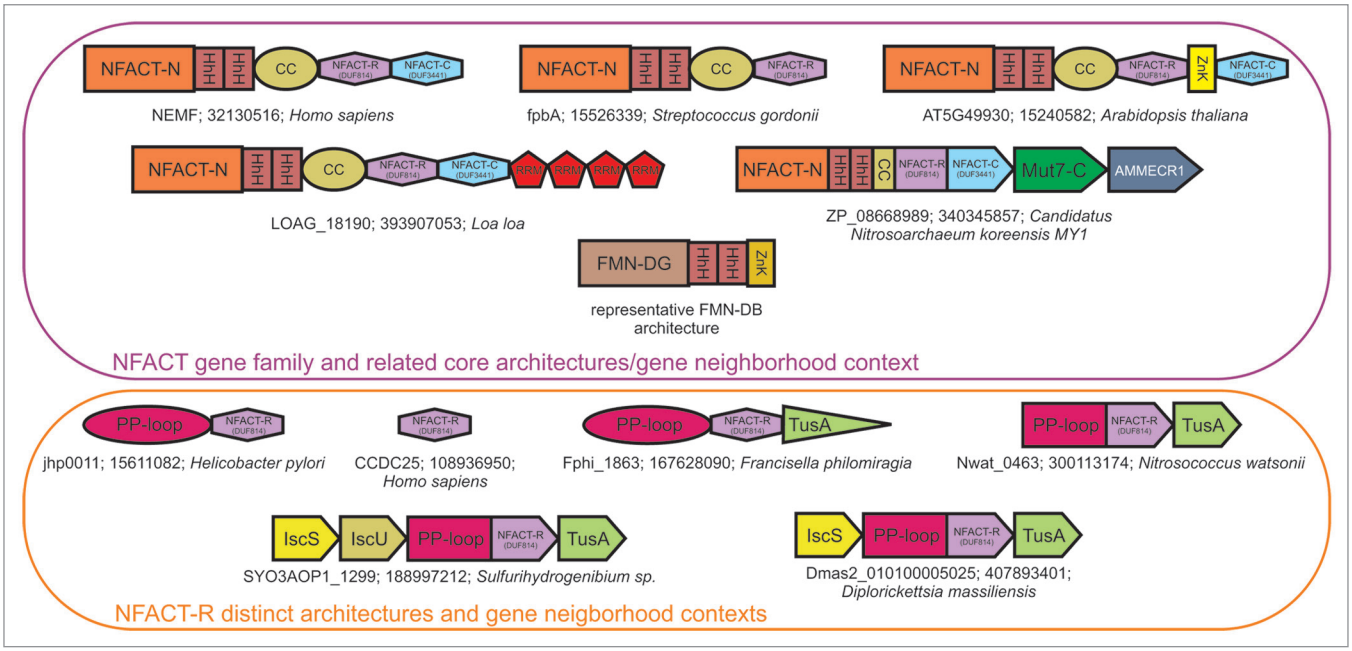

Figure 1. Domain architectures and conserved gene neighborhoods relating to NFACT proteins. Domains architectures are depicted by adjoining polygonal shapes labeled with individual domain names, size of domains are not drawn precisely to scale. Individual genes within conserved gene neighborhoods are depicted with boxed arrows with the arrowhead pointing toward the 3′ end of the gene. Each architecture/neighborhood is labeled with gene name, GenBank gene identifier (gi) number, and organism name separated by semicolons. Architectures/neighborhoods relating to the NFACT gene family are boxed in purple; those relating specifically to independent contexts of the NFACT-R domain are boxed in orange. The coiled-coil region characteristic of the NFACT gene family is depicted as a light green circle in architectures and a light green box in gene neighborhoods. Abbreviations: CC, coiled-coil; HhH, helix-hairpin-helix; ZnK, zinc knuckle; ZnR, zinc ribbon.

enzymatic domain related to the bifunctional DNA glycosylase/endonuclease domain involved in Base Excision Repair (BER), commonly referred to as the Formamidopyrimidine, MutM, and Nei/EndoVIII DNA glycosylase (FMN-DG; also referred to in the literature as Fpg/Nei, Fapy DNA glycosylase, glycosylase/AP-lyase, or Endonuclease VIII) domain.[15-18] We identify shared and distinct features of the active site of these two related domains, implying both similarities and differences in their catalytic mechanisms. Another domain in this gene family is predicted to be a novel RNA-binding domain, with a potential role in a variant of the tRNA 4-thiouridylation pathway present in a subset of prokaryotes. Based on these observations and additional genome contextual evidence, we propose that the fundamental functional role of this ancient gene family is related to processing/modification of double-stranded RNA, perhaps rRNA.

Results

Delineation of the NEMF/FbpA/Caliban/Tae2gene family and its core architectures

To comprehensively characterize this gene family, we collected all related sequences using known members as seeds to initiate sequence profile searches against the non-redundant (nr) protein database at the National Center for Biotechnology Information. Given the presence of a large coiled-coil domain in the gene family, we applied the low complexity seg filter[19] to these searches to avoid inclusion of genes with spurious similarity. Membership of proteins displaying relationships with borderline significance was confirmed by initiating reverse searches. Sequences obtained were then aligned and potential globular regions shared across the gene family were identified by inspection of these alignments after mapping the location of the known HhH domains and the coiled-coil regions onto the alignments (see Methods, **Supplemental Material**).

Orthologs of the gene family across all three superkingdoms of life were identified, including the NEMF, bacterial FbpA-like proteins, Caliban, and Tae2; accordingly, we termed this family NFACT. Representatives of the family were found across all major archaeal lineages including the euryarchaeota, crenarchaeota, korarchaeota, and thaumarchaeota. The NFACT family is also found across most major bacterial lineages, although it is notably absent in the α-, β-, and γ-proteobacterial lineages (despite being present in δ- and ε-proteobacteria) and actinobacteria. In eukaryotes, the NFACT family is again present in all major lineages including the diplomonads, parabasalids, heteroloboseans, kinetoplastids, chromoalveolates, apicomplexa, and the crown-group eukaryotes encompassing the plant, ameobozoan, animal, and fungal lineages (with a notable absence in the basidiomycete fungi). Taken as a whole, despite losses in certain terminal lineages, this phyletic spread unquestionably points to presence of the NFACT family in the LUCA (see **Supplemental Material** for complete sequence and phyletic distribution).

The conserved core of the NFACT family found across all members is formed by four domains interrupted by the coiled-coil region (**Fig. 1**; **Supplemental Material**): from N terminus to C terminus

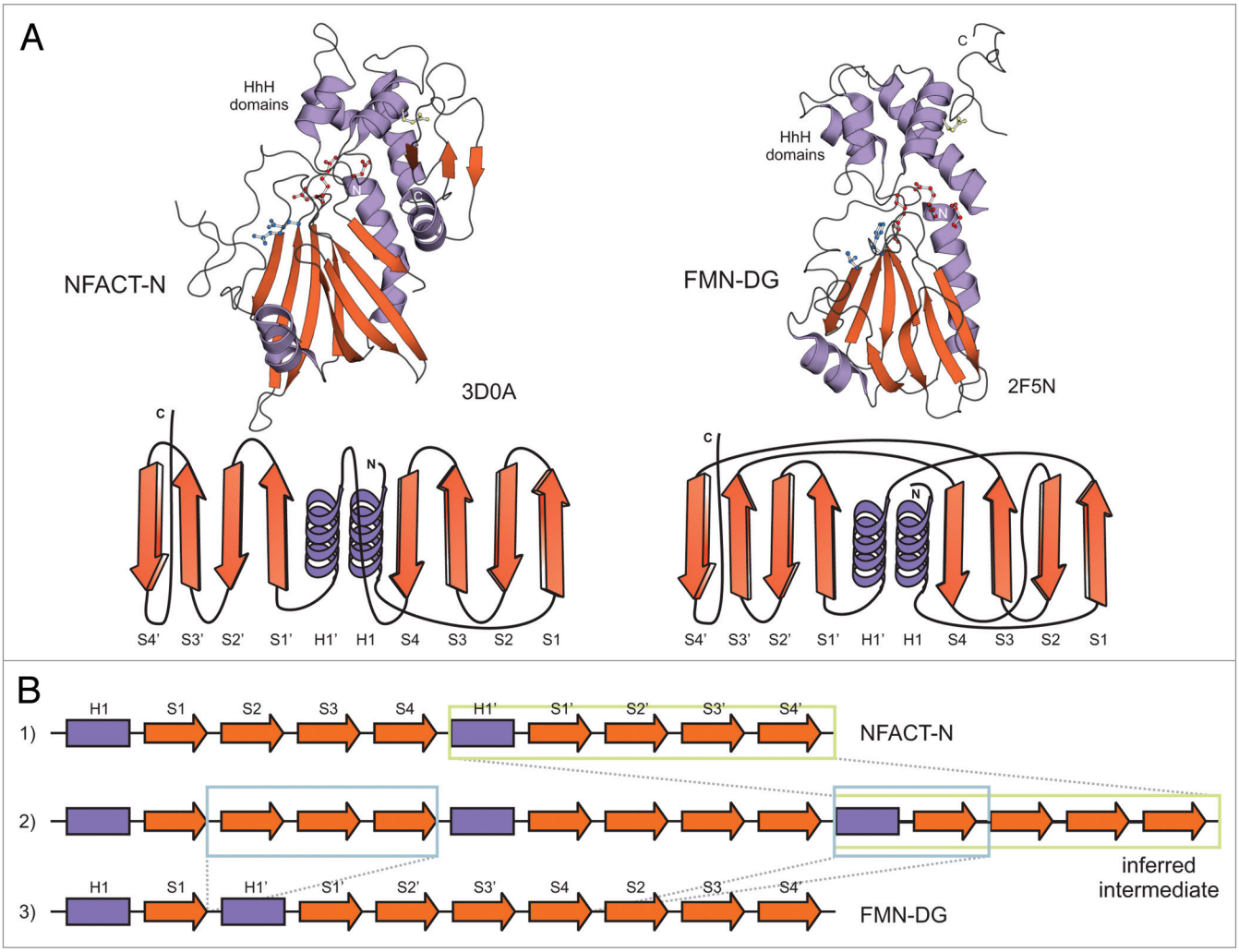

Figure 2. Structural relationship between NFACT-N and FMN-DG domains. (**A**) Cartoon renderings of solved crystal structures (top) for the NFACT-N (left) and FMN-DG domains (right) accompanied by corresponding topology diagrams (bottom). Residues conserved across the domains are rendered as ball-and-stick in the cartoons; active site/enzymatic residues are colored in red, residues with a likely direct or indirect role in substrate recognition are colored in blue, and the conserved asparagine/histidine found in the HhH domains is colored in yellow. The PDB identifier is provided to the right of both cartoons. The labeling scheme provided below each diagram reflects the spatial/evolutionary conservation of each element as evident from the solved crystal structures and as referred to in the text. (**B**) Stepwise scenario for the emergence of the FMN-DG domain from the ancestral NFACT-N domain. Domains are depicted as arrays of secondary structure elements to show the wiring between elements.

these entail an uncharacterized N-terminal domain, the two HhH domains, the coiled-coil region, and a domain currently annotated as DUF814 (Domain of Unknown Function 814) in Pfam.[20] The first three domains from the N-terminus are currently incorrectly annotated as a single domain in Pfam: the FbpA domain. We propose renaming the N-terminal domain the NFACT-N (for NEMF, FbpA, Caliban, Tae2, N-terminal) domain and separating it from the downstream HhH domains (**Fig. 1**). In archaea and eukaryotes, an additional C-terminal domain annotated in Pfam as DUF3441 is present, clearly establishing the archaeal version of the family as the one inherited by eukaryotes. This core (NFACT-N+HhH+HhH+coiled-coil+DUF814[+DUF3441]) has proven resistant to domain accretion during evolution, although a few limited elaborations are observed in eukaryotes. In plants, *Entamoeba*, and *Giardia*, a zinc knuckle domain (ZnK) insertion is present between the DUF814 and DUF3441 domains. The most parsimonious explanation for this unusual phyletic distribution is independent, secondary acquisition of the ZnK in these three distant lineages, a scenario supported by the lack of specific sequence similarity across the different ZnKs. While the ZnK is not universally present in chlorophyte alga, its presence in *Ostreococcus* suggests that it was acquired early in the evolution of Viridiplantae. One additional potential domain fusion of note is the C-terminal fusion to four copies of the RNA-binding RRM domain in the roundworm *Loa loa* (**Fig. 1**).

To better understand the roles of the distinct domains in the NFACT proteins, we investigated in detail the previously uncharacterized domains in NFACT proteins using sensitive sequence-profile searches.

NFACT-N domain

Iterative profile searches initiated with NFACT-N domain sequences and their downstream HhH domains against the

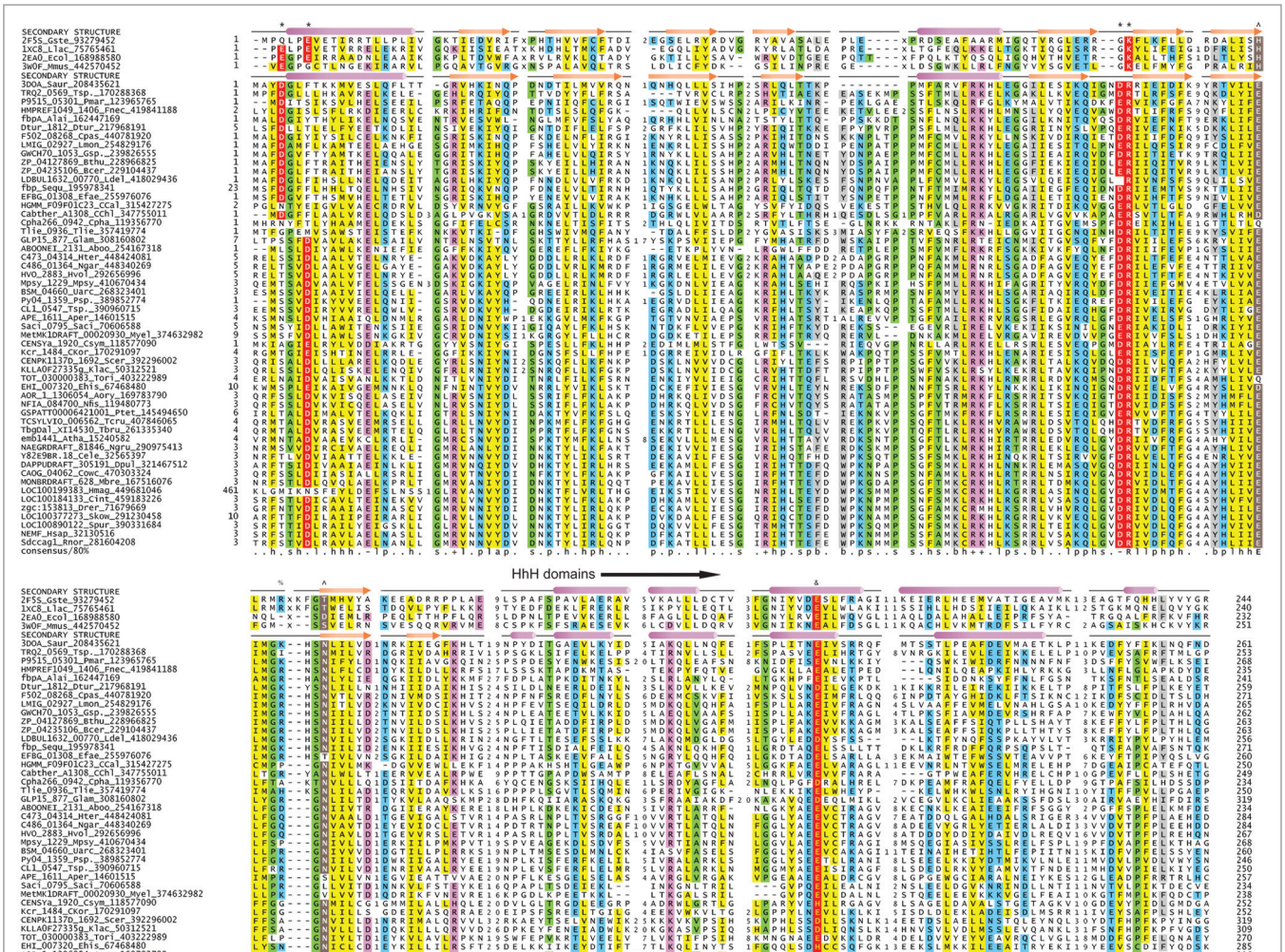

Figure 3. See page 364 for figure legend.

nr database recovered significant matches extending along the length of the DNA glycosylase (FMN-DG) domain and its characteristic C-terminal HhH domain pair. For example, a sequence from the *H. sapiens* NEMF protein recovered the FMN-DG and HhH domains from the fungus *Togninia minima* (gi: 500258929, iteration: 4, e-value: $2e^{-4}$) and the acidobacterium *Terriglobus roseus* (gi: 390956237, iteration: 4, e-value: $7e^{-4}$) in PSI-BLAST. Reciprocal searches confirmed these relationships: a search initiated with the same sequence from *Terriglobus* yielded the NFACT-N from *Methanocella arvoryzae* (gi: 147920849, PSI-BLAST iteration: 1, e-value: $4e^{-3}$). As an independent means of confirmation, a HMM profile constructed from the multiple sequence alignment of the NFACT-N and HhH domains was searched against a database of HMMs constructed for individual pdb entries using the HHpred program. In addition to detecting PDB: 3doa (structure of a N-terminal NFACT fragment), this search again revealed a significant relationship between the NFACT-N and HhH domains and cognate domains in FMN-DG; for example, significant matches are retrieved for FMN-DGs from *Arabidopsis thaliana* (PDB: 3twl, p-value: $1e^{-5}$, probability: 95.9%) and *Geobacillus stearothermophilus* (PDB: 3u6p, p-value: $6.2e^{-5}$, probability: 94.0%).

The FMN-DG family is well-distributed across bacteria, absent in archaea, and only present in scattered eukaryotic lineages. The family can be divided into two distinct subfamilies, the Fpg/MutM-like subfamily found in bacteria, plants, and fungi and the eukaryotic Nei subfamily primarily observed in animals.[17,21] Previous analysis indicated that eukaryotic versions of the Fpg/MutM-like subfamily likely emerged via horizontal gene transfer (HGT) from bacteria relatively early in eukaryotic evolution, while the Nei subfamily emerged later following HGT from a bacterial source to the stem of the animal lineage.[17] Our

Figure 3 (See previous page). Multiple sequence alignment of NFACT-N+HhH domains with selected FMN-DG+HhH domain sequences. FMN-DG sequences from solved crystal structures are at the top of the alignment followed by the NFACT-N sequences. Secondary structure elements are depicted as follows: extended loop regions are represented by black lines, β-strands represented by orange arrows, and α-helices represented by purple cylinders. The transition from the core enzymatic domain to the HhH domains is labeled with a black arrow above the alignment. Individual sequences are labeled to the left with gene name, organism name, and gi number separated by an underscore. Gene names are replaced by PDB identifiers where appropriate. Numbers to the left and right of the alignment correspond to amino acid position within the protein encoding the domain. Insert regions are excised and replaced with numbers indicating the length in amino acids of the insert. Due to the "re-wiring" between the core enzymatic domains (**Fig. 2**), FMN-DG sequences are not presented in linear order; "breaks" in this order are marked at appropriate positions with "x." Coloring is based on the consensus line at the bottom of the alignment: h, hydrophobic (shaded in yellow); s, small (shaded in green); l, aliphatic (shaded in yellow); -, negatively charged (shaded in purple); p, polar (shaded in blue); +, positively charged (shaded in purple); a, aromatic (shaded in yellow); b, big (shaded in gray); u, tiny (shaded in green); c, charged (shaded in purple). Columns corresponding to active site residue positions are shaded in red, colored in white, and marked at the top with "*." Columns corresponding to positions involved in either direct or indirect substrate recognition are shaded in brown, colored in white, and marked with "^." The column corresponding to the conserved glutamate/histidine residue in the HhH domains is shaded in red, colored in yellow, and marked with "&." The column corresponding to the conserved lysine/arginine residue specific to NFACT-N is marked with a "%." Organism abbreviations as follows: Aboo, *Aciduliprofundum boonei*; Alai, *Acholeplasma laidlawii*; Aory, *Aspergillus oryzae*; Aper, *Aeropyrum pernix*; Atha, *Arabidopsis thaliana*; Bcer, *Bacillus cereus*; Bthu, *Bacillus thuringiensis*; CCal, *Candidatus Caldiarchaeum*; CChl, *Candidatus Chloracidobacterium*; CKor, *Candidatus Korarchaeum*; Cele, *Caenorhabditis elegans*; Cint, *Ciona intestinalis*; Cowc, *Capsaspora owczarzaki*; Cpas, *Clostridium pasteurianum*; Cpha, *Chlorobium phaeobacteroides*; Csym, *Cenarchaeum symbiosum*; Dpul, *Daphnia pulex*; Drer, *Danio rerio*; Dtur, *Dictyoglomus turgidum*; Ecol, *Escherichia coli*; Efae, *Enterococcus faecalis*; Ehis, *Entamoeba histolytica*; Fnec, *Fusobacterium necrophorum*; Glam, *Giardia lamblia*; Gsp., *Geobacillus* sp.; Gste, *Geobacillus stearothermophilus*; Hmag, *Hydra magnipapillata*; Hsap, *Homo sapiens*; Hter, *Halorubrum terrestre*; Hvol, *Haloferax volcanii*; Klac, *Kluyveromyces lactis*; Ldel, *Lactobacillus delbrueckii*; Llac, *Lactococcus lactis*; Lmon, *Listeria monocytogenes*; Mbre, *Monosiga brevicollis*; Mmus, *Mus musculus*; Mpsy, *Methanolobus psychrophilus*; Myel, *Metallosphaera yellowstonensis*; Nfis, *Neosartorya fischeri*; Ngar, *Natrinema gari*; Ngru, *Naegleria gruberi*; Pmar, *Prochlorococcus marinus*; Psp., *Pyrococcus* sp.; Ptet, *Paramecium tetraurelia*; Rnor, *Rattus norvegicus*; Saci, *Sulfolobus acidocaldarius*; Saur, *Staphylococcus aureus*; Scer, *Saccharomyces cerevisiae*; Sequ, *Streptococcus equi*; Skow, *Saccoglossus kowalevskii*; Spur, *Strongylocentrotus purpuratus*; Tbru, *Trypanosoma brucei*; Tcru, *Trypanosoma cruzi*; Tlie, *Thermovirga lienii*; Tori, *Theileria orientalis*; Tsp., *Thermococcus* sp.; Tsp., *Thermotoga* sp.; Uarc, uncultured archaeon.

analysis identified Nei homologs in the early-branching eukaryote diplomonad *Giardia*, suggesting the possibility that Nei also might have been acquired earlier in eukaryotic evolution (AMB, LA, personal observations).

Thus, the phyletic patterns of NFACT-N point to an origin in the LUCA, whereas the FMN-DGs appear to have emerged first in bacteria. This suggests that the latter are likely to have been derived from the former early in bacterial evolution. To better understand the relationship between NFACT-N and FMN-DNA glycosylases, we constructed a structure-guided super-alignment, first aligning known FMN-DG structures with the 3DOA structure and then adding further NFACT-N/HhH sequences (see Materials and Methods). At this point it became evident that despite the clear homology between the domains, the shared core scaffold had undergone a multi-step structural reorganization via duplication during divergence (**Fig. 2**). Both domains feature a core containing eight β-strands and two α-helices leading into the dyad of HhH motifs. The NFACT-N domain, which was inferred to represent the ancestral condition, has two repeats of a basic structural element, each containing an α-helix leading into a 4-stranded β-meander, yielding a sandwich-like fold with the two stacking β-sheets from each repeat oriented at a roughly 30 degree angle to the other (**Fig. 2A**). In FMN-DG, this structure, including the orientation of the stacking β-sheets, is retained but the connectivity between the helix/meander units has been substantially altered (**Fig. 2**). The most parsimonious explanation for the "re-wiring" of the connectivity in FMN-DG entails the following steps, in some ways reminiscent of the recently elucidated steps underlying the derivation of the FYVE domain from the canonical binuclear treble clef domain core[22] (**Fig. 2B**): (1) duplication of one of the repeats in the original 2-repeat structure yielding a 3-repeat intermediate. (2) Given the packing of the sheets against each other to form a sandwich, the 3-repeat state leads to an inherently unstable condition with competition between alternative repeats to reconstitute the original sandwich. (3) This instability was resolved by natural selection through partial loss of a subset of the elements to reconstitute the original two-sheet sandwich. This reconstitution, rather than proceeding via the loss of a complete superfluous repeat, resulted from complementary, partial loss of elements from repeat 1 and 3, while retaining the overall sandwich structure intact. Using this understanding as a guide, an alignment of the spatially equivalent strands and helices across both domains was constructed. This revealed extensive sequence conservation between the aligned individual elements of both domains, strongly supporting the above proposed scenario (**Fig. 3**).

The FMN-DG enzymes have been studied in the context of removal of DNA bases damaged through oxidation, e.g., 8-oxoguanine (8OG). FMN-DG is distinct from other types of DNA glycosylases in that it also catalyzes the next step in DNA repair, the introduction of an endonucleolytic break at the site of base removal. The catalytic process of the FMN-DG domain includes, in order of steps: (1) recognition of oxidized bases, (2) flipping of the oxidized "base" from the double helix to position the base in the active site pocket, (3) removal of the base, and (4) induction of an endo-nucleolytic break in the DNA backbone following opening of the deoxyribose ring associated with the excised base.[23] Despite extensive research on these activities, the basis for steps 1 and 2 remain relatively poorly understood as the implicated residues appear to vary considerably even among closely related members of the same subfamily of FMN-DGs, a phenomenon probably related to the observed promiscuity in oxidized base substrates. Steps 3 and 4 draw on a set of core residues clustered in the active site pocket, including an absolutely conserved proline residue, one or two conserved glutamate residues located in distinct spatial positions near the N terminus of H1, and a lysine residue found in the loop between strand-1' and strand-2' (**Figs. 2A, 3, and 4A**). Consensus based on experimental studies

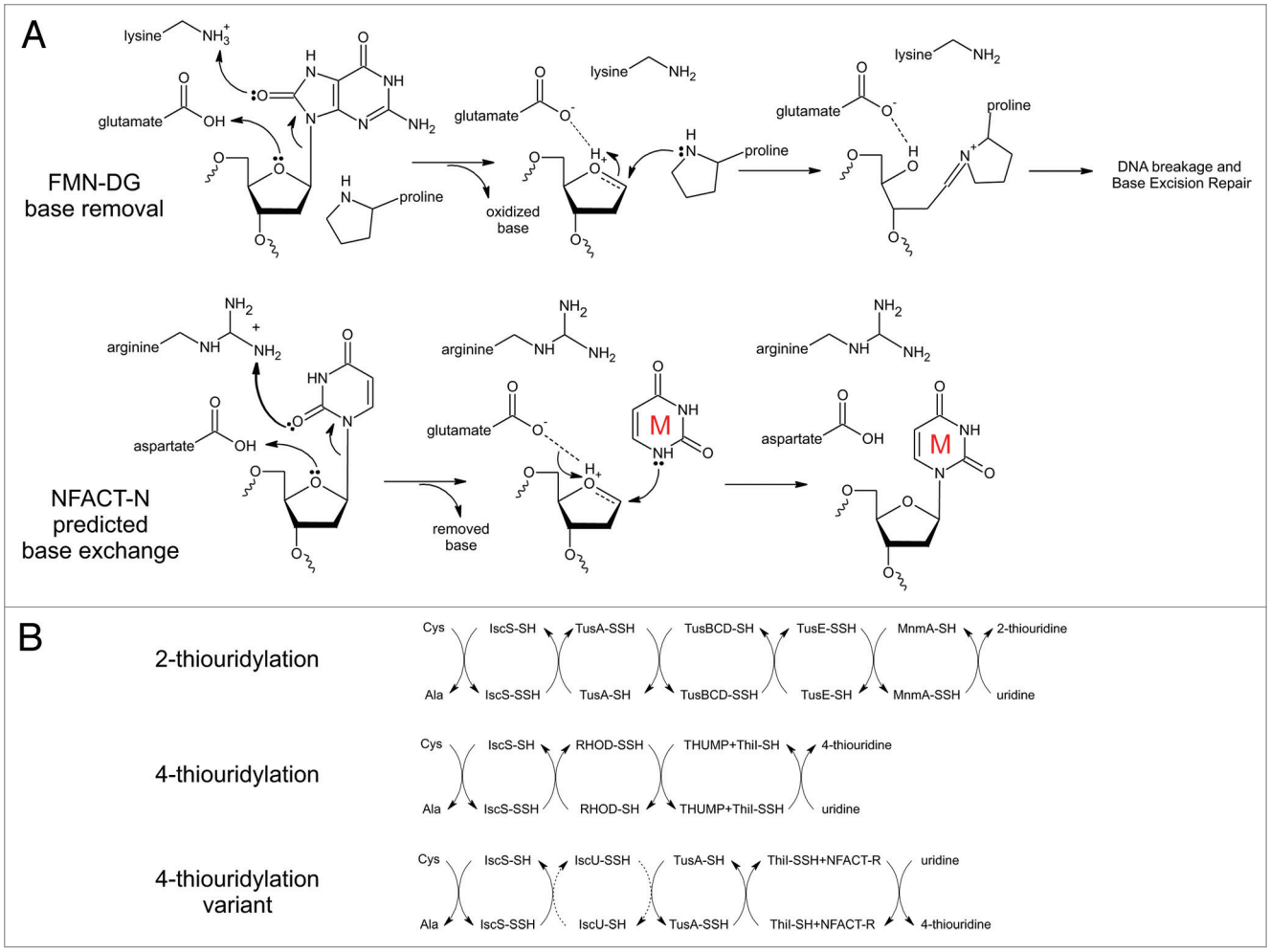

Figure 4. Known and predicted reactions. (**A**) Base removal and ring-opening steps catalyzed by FMN-DNA glycosylases (top) and predicted analogous steps catalyzed by NFACT-N during a potential base-exchange reaction (bottom). The introduced free base in the NFACT-N reaction is labeled with a red "M," indicating the base is potentially modified in some way despite being shown here as a uridine. (**B**) 2-thiouridylation, canonical 4-thiouridylation, and the novel, predicted 4-thiouridylation sulfur relay pathways. Potential intermediate step in the novel 4-thiouridylation pathway involving transfer to a conserved cysteine on IscU is shown in dotted lines, reflecting uncertainty in whether this step is present in all organisms containing the ThiI+NFACT-R fusion or whether it is restricted to a subset of them.

currently holds that the proline residue forms a Schiff-base intermediate with the C1' atom of the damaged base and that at least one of the two conserved glutamate residues is involved in initiating the base removal step.[24-32] The role of the lysine has been debated but appears to also be involved in the base removal step while possibly also being a key player in the backbone cleavage reaction.[27,31] In comparison, the NFACT-N domain harbors an absolutely conserved arginine positionally equivalent to the lysine. Helix-1 in NFACT-N also contains a well-conserved aspartate found in one or the other of the two positions where conserved glutamates are found in the FMN-DG domain (**Figs. 2A and 3**). Strikingly, NFACT-N does not contain any absolutely conserved prolines, either at the N terminus or elsewhere in the domain (**Fig. 3**). NFACT-N also features an additional absolutely conserved aspartate in the predicted active site pocket, one residue downstream of the aforementioned arginine, which could compensate for the second aspartate seen in the N terminus of H1 in the cognate FMN-DG domain (**Figs. 2A and 3**). Two additional well-conserved residues outside of the active site pocket with conserved cognates in FMN-DG include a glutamate found near the C terminus of strand-3′ and a well-conserved asparagine found at the N terminus of strand-4'. The glutamate residue appears equivalent to a highly conserved histidine residue in FMN-DG, which plays a role in substrate nucleic acid binding;[28] however, in NFACT-N, this residue appears to form a conserved salt bridge with a positively charged residue in the linker region found between the NFACT-N domain and the HhH domains. The conserved NFACT-N asparagine appears positionally equivalent to a polar residue typically taking the form of a serine or threonine in FMN-DG domains that contributes to nucleic acid recognition (**Figs. 2A and 3**); in NFACT-N, this residue might help position a well-conserved NFACT-N-specific arginine/lysine, which points out from the predicted catalytic core and could be involved in nucleic acid recognition (**Fig. 3**).

Crosslinking experiments capturing FMN-DG active site intermediates clearly indicate that proline-mediated Schiff base

formation occurs following base excision and deoxyribose ring-opening but prior to nicking of the phosphodiester backbone (**Fig. 4A**).[27] One of the two conserved H1 aspartates is consistently implicated in base removal upstream of the Schiff base formation, although which aspartate is involved may depend on the substrate. Conflicting views on the lysine indicate a role in either base-removal upstream of Schiff base formation[31] or in initiating the DNA cleavage reaction.[27] The lack of a conserved proline or any compensatory amine-bearing residue in the cognate NFACT-N active site unequivocally establishes NFACT-N as incapable of forming the Schiff base intermediate and thus is unlikely to be involved in DNA cleavage as catalyzed by FMN-DG. However, like FMN-DG, NFACT-N displays an arginine equivalent to lysine in the former and two acidic active site residues. Hence, based on the above-outlined spatial position- and residue- conservation between the domains, we predict NFACT-N potentially catalyzes base-removal as observed in FMN-DG (**Fig. 4A**, see below). These observations, combined with the knowledge that the FMN-DG domains were derived from the ancestral NFACT-N domain, also potentially assists in distinguishing between proposed roles for the lysine in the FMN-DG reaction mechanism. As the conserved proline, Schiff base formation, and nuclease activity are unique to FMN-DG, they must necessarily be catalytic innovations secondary to the roles of the ancestral aspartate/glutamate and arginine/lysine. Thus, a role for the lysine/arginine in base removal during glycosylase activity is likely to be the ancestral role (**Fig. 4A**), although we cannot rule out that the lysine has secondarily acquired an additional role in DNA cleavage in FMN-DG. This role for the lysine is also consistent with mutational studies finding base removal, as opposed to DNA cleavage, to be most affected by lysine substitution.[27]

One additional residue conserved in both NFACT-N- and FMN-DG-fused HhH domains is noteworthy: a well-conserved glutamate residue (sometimes replaced by histidine in NFACT-N) found in the first helix of the second HhH domain, which mediates a backbone contact with a distinctive, conserved loop structure immediately C-terminal to the HhH domains in NFACT-N and C-terminal to an inserted zinc ribbon domain found downstream of the HhH domains in FMN-DG. This glutamate-backbone contact positions the HhH domains for interaction with the nucleic acid substrate on the side opposite to the NFACT-N and FMN-DG active sites (**Figs. 2A and 3**).Thus, this structural constraint appears to have maintained similar active site clefts for the two families with the HhH motifs forming a nucleic acid binding "cap" in both cases to accommodate a double-stranded substrate.

DUF814 domain

Searches with individual DUF814 sequences as seeds failed to recover any remote relationships with other known domains. A multiple sequence alignment constructed for the domain indicates the DUF814 domain consists of an α/β structure with at least seven β-strands and three α-helices. DUF814 domains of the NFACT gene family contain a well-conserved DxxxH motif with the two conserved residues at either end of the third predicted strand in addition to a conserved downstream serine residue found in the second helix of this domain (**Supplemental Material**). These domains are present in two additional contexts: (1) N-terminal fusion to a PP-loop domain in a diverse range of bacteria including firmicutes, nitrosporae, fusobacteria, synergistetes, spirochetes, delta-, epsilon-, and some gamma-proteobacteria, aquificae, dictyoglomi, planctomycetes as well as few thaumarchaeota and (2) as a solo domain C-terminally fused to a small coiled-coil region present across all eukaryotic lineages, typified by the CCDC25 protein in humans and the Jlp2 protein in *Saccharomyces cerevisiae* (**Supplemental Material**). An alignment constructed of only members of the PP-loop-fused DUF814 family revealed the same core secondary structure but lacked the conserved residues found in the NFACT family, instead featuring two nearly-absolutely conserved arginines N-terminal to the first strand and the first helix, respectively (**Supplemental Material**). The PP-loop domain belongs to the ThiI-like PP-loop family, which catalyzes 4-thiouridylation at nucleotide 8 of bacterial and archaeal tRNA.[33] ThiI-like domains were previously thought to be universally fused to an N-terminal RNA-binding THUMP domain[34] with versions from many bacteria and a few archaea additionally fused C-terminally to a Rhodanese (RHOD) domain containing an absolutely conserved cysteine residue. Bacteria and archaea lacking the C-terminal RHOD domain likely interact with a stand-alone RHOD domain during thiouridylation.[35] Canonical ThiI-mediated 4-thiouridylation of tRNA begins with mobilization of sulfur from free cysteine via the IscS desulfurase, followed by transfer of the sulfur to the conserved cysteine residue found in the RHOD domain (**Fig. 4B**). In parallel, the THUMP domain binds and positions the tRNA and the PP-loop domain activates the oxo-group of the target uridine using ATP to form an adenylated intermediate.[36,37] Sulfotransfer to the uridine then proceeds through either direct attack by the RHOD-bound persulfide sulfur on the adenylated intermediate or by a free sulfide generated after attack by an additional sulfur provided by the conserved cysteine residue internal to the core PP-loop domain (**Fig. 4B**).[38]

Two striking observations regarding the ThiI+DUF814 proteins were immediately apparent: (1) the ThiI+DUF814 protein shows an almost entirely mutually exclusive phyletic distribution pattern with respect to the THUMP-containing ThiI-like enzymes (**Supplemental Material**) and (2) while the ThiI+DUF814 proteins lack both THUMP and RHOD fusions, they retain the conserved residues required for ATP utilization/adenylation found across all PP-loop domains in addition to conserving the internal cysteine residue and CxxC motif characteristic and specific to the ThiI-like PP-loop domains[39] (**Supplemental Material**). It is also worth noting the variable flexible loop region covering the ThiI active site[35] appears to be more elaborate in the ThiI-DUF814 fusion proteins than canonical ThiI-like domains and houses several unique, strongly conserved motifs (GRxRxxQ and TxxE and a glutamine) (**Supplemental Material**).

In *Francisella philomiragia*, the ThiI+DUF814 protein is additionally fused to a TusA/SirA-like (TusA) domain at the C terminus (**Fig. 1**). Inspection of the gene neighborhoods surrounding the ThiI+DUF814 fusion proteins revealed further association with a TusA domain, the only conserved gene neighborhood association found across phylogenetically diverse bacteria. Within these neighborhoods, a subset of bacteria including several

aquificae and a single γ-proteobacterium additionally associates with an IscS-like desulfurase with aquificae further containing an IscU/NifU protein (**Fig. 1**; **Supplemental Material**). While this is the first reported linkage between ThiI-like PP-loop domains and TusA domains, TusA domains combine with a distinct PP-loop family, the IscS desulfurases, and several additional domains to catalyze tRNA 2-thiouridylation (**Fig. 4B**). In this pathway, IscS-abstracted sulfur is first transferred to a conserved cysteine in TusA, initiating a complex sulfur transfer pathway, which continues until the PP-loop ATPase domain incorporates the sulfur into the base via adenylation (**Fig. 4B**).[40] Alignment of the TusA proteins associating with ThiI+DUF814 revealed retention of the canonical conserved cysteine essential for sulfur transfer (**Supplemental Material**).

The above web of contextual information presents several reasons strongly supporting the ThiI+DUF814 proteins being part of a distinct pathway catalyzing 4-thiouridylation in a subset of prokaryotes: (1) mutual exclusivity in phyletic distributions between ThiI+DUF814 and other ThiI-like enzymes indicates functional equivalence of the two (**Supplemental Material**).[41] (2) The high degree of sequence similarity between the ThiI enzymes regardless of domain fusions and the specific conservation of critical residues necessary for ThiI's role in 4-thiouridylation. (3) The presence of the TusA domain, which can compensate for the absence of the RHOD domain and act as a sulfur acceptor prior to transfer to the tRNA. This proposed 4-thiouridylation pathway is predicted to proceed as follows (**Fig. 4B**): (1) Analogous to its role in 2-thiouridylation of tRNA, the TusA domain likely accepts sulfur abstracted from the free cysteine pool in the cell by the IscS-like desulfurase domain;[40,42] the sulfur relay could also include transfer through IscU/NifU based on its scattered presence in gene neighborhoods (**Fig. 1**), potentially representing the first known involvement of IscU/NifU in tRNA thiolation as opposed to its standard role in FeS cluster biogenesis.[42] (2) TusA then functions similar to the RHOD domain in standard 4-thiouridylation pathways by interacting with ThiI and positioning the donor sulfur near the active site of ThiI. (3) The DUF814 domain binds the tRNA substrate in lieu of the THUMP domain, positioning it for ThiI-catalyzed adenylation of the target uridine and ultimately sulfur transfer. Given these observations and previous connections between the NFACT gene family and nucleic acid-binding, we can also predict DUF814 acts as an RNA-binding domain in the NFACT gene family; henceforth, we refer to this domain as the NFACT RNA-binding (NFACT-R) domain.

Solo NFACT-R domains prototyped by eukaryotic CCDC25/Jlp2-like proteins have not been previously experimentally characterized; however, functional coupling networks detect strong associations with several genes encoding diverse RNA-binding domains in humans,[43] consistent with a role for the domain in RNA-binding. CCDC25 shares interactions with proteins which also interact with NFACT proteins, although the two have yet to be directly linked. The CCDC25/Jlp2-like family of NFACT-R domains exhibits strong conservation in eukaryotes suggesting purifying selection, a feature of RNA-associated domains functioning in core biological processes like base modification and translation. The close relationship between the solo NFACT-R domain and the version in the NFACT proteins along with its pan-eukaryotic distribution suggests that it likely emerged through partial duplication from a NFACT precursor prior to the Last Eukaryotic Common Ancestor (LECA). Hence, we predict that it is likely to function in an RNA-binding role in a core cellular function, perhaps even in the same complex as the NFACT proteins (see below).

DUF3441 domain

In the Pfam database, DUF3441 is presented as an eukaryote-specific domain. However, we recovered divergent yet clearly homologous versions at the C-terminus of archaeal NFACT proteins. For example, a PSI-BLAST search with the C-terminal domain from the archaeon *Aeropyrum pernix* recovered the entirety of the eukaryotic NFACT DUF3441 domain from the annelid *Capitella teleta* (gi: 443707183, iteration: 4, e-value: $2e^{-03}$). Given its presence as a core NFACT domain, we rename this domain the NFACT-C domain (NFACT C-terminal domain). Secondary structure predictions based on a multiple sequence alignment suggest that it adopts an α/β fold. In contrast to other domains in NFACT proteins, there is little absolute conservation of residues outside of a strongly-conserved PG motif (**Supplemental Material**); however, several positions in the domain are retained as different polar or charged residues. This pattern suggests NFACT-C could mediate protein–protein contacts within a larger complex rather than playing a catalytic role.

Contextual analysis of the NFACT gene family

We then investigated the NFACT family itself for potential conserved gene neighborhoods and functional interaction networks, as this information can provide insight into function of uncharacterized domains.[44,45] Gene neighborhoods extracted for bacterial NFACT genes did not recover any conserved associations; however, archaeal NFACT genes formed a conserved gene neighborhood across all archaeal lineages, including the euryarchaeota, crenarchaeota, nanoarchaeota, thaumarchaeota, and the caldiarchaeota[46] with two genes coding for: (1) An active Mut7-C RNase domain of the PIN nuclease fold with its accompanying C-terminal Zn-ribbon domain[47,48] and (2) the AMMECR1 domain containing the RAGNYA fold[49] (**Supplemental Material**). The conserved connection to the Mut7-C RNase is again strongly suggestive of an RNA-related role for NFACT. Additionally, AMMECR1 has previously been linked to involvement in an as-yet-uncharacterized RNA base modification, potentially entailing the transfer of a modifying group onto an RNA base via a conserved cysteine.[49] Notably, most AMMECR1 domains found in this conserved neighborhood retains all previously predicted enzymatic residues (except the crenarchaeon *Vulcanisaeta* where the predicted catalytic cysteine residue is replaced by a serine; see **Supplemental Material**).

Functional interaction networks constructed primarily by co-expression patterns and protein–protein interaction data were also suggestive of an RNA-related role for the NFACT family.[43] The yeast Tae2 protein is most strongly linked to a cluster of proteins involved in ribosomal subunit biogenesis. Consistent with this, recent studies identified it as a part of the large (60S) ribosomal subunit interacting Ribosomal Quality Control Complex (RQC).[11,12]

Cognate NEMF proteins from mammals were predominantly linked to proteins harboring diverse RNA-binding domains, many with demonstrated roles in splicing and rRNA biogenesis.

Discussion

Emerging picture of NFACT as a potential RNA-modifying enzyme

Multiple independent lines of evidence support a RNA-related role for NFACT proteins: (1) the previously determined evolutionary relationship between the NFACT HhH domains and those observed in the S13/S18 ribosomal proteins, (2) the evolutionary relationship between the NFACT-N domain and the FMN-DG catalytic domain suggesting NFACT might function as an enzyme operating on bases in double stranded nucleic acids, (3) conserved operonic associations in archaea with the AMMECR1 and the Mut7-C RNase domain pointing in the direction of RNA processing and modification, and (4) experimental evidence from yeast[11,12] and functional network associations suggesting one or more rRNAs as probable substrates.

In terms of RNA-modification reactions, parallels can be drawn between the glycosylase reaction catalyzed by FMN-DG and base-modification mechanisms catalyzed by structurally unrelated RNA-modifying enzymes: the pseudouridine synthases catalyzing formation of pseudouridine and the tRNA-guanine transglycosylases (TGTs) catalyzing the "base-swapping" mechanism, which inserts the preQ0 base precursor of archaeosine in archaea[50] and queuosine in bacteria and eukaryotes.[51-53] Members of the pseudouridine synthase fold, which includes the TruB, RluA, RsuA, and TruA-like families,[54,55] similar to FMN-DG require an initial base-flipping step to position a base in the active site prior to modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine "base rotation" reaction resulting in formation of pseudouridine in tRNA substrates.[56,57] Several families of pseudouridine synthases additionally contain a well-conserved lysine or arginine residue in the active site, which forms a salt bridge with the catalytic aspartate. Despite similarities, the aspartate in FMN-DG plays only a transient role in the catalytic mechanism, sharply contrasting the centrality of the aspartate in the pseudouridine synthase mechanism, which is thought to directly form a Michael adduct with a carbon in the pyrimidine ring of the uridine as a reaction intermediate.[56,58] Thus, it seems unlikely NFACT would follow the mechanism template established in pseudouridine synthases. TGTs, members of the TIM-barrel fold, also require an absolutely conserved aspartate residue when catalyzing the swapping of a tRNA guanine base for the preQ0 precursor base. In contrast to the pseudouridine synthase mechanism and closer to the FMN-DNA glycosylase mechanism, this aspartate is involved in the initial step of the reaction facilitating the removal of the base from the sugar backbone.[51] A further parallel with FMN-DG is observed in the base-exchange step wherein N9 of the imidazole ring of the preQ0 base directly attacks the C1' of the "base-less" sugar, resulting in attachment of the preQ0 to the tRNA sugar backbone (**Fig. 4A**). This step strikingly resembles the attack by the nitrogen atom belonging to the conserved FMN-DG proline residue on the C1' of the DNA substrate backbone following base removal. It is possible that the secondary emergence of the proline and the Schiff base observed in FMN-DG activity displaced a nitrogen-based attack from an incoming base similar to the elucidated TGT mechanism in the ancestral NFACT-N domain (**Fig. 4A**). This reasoning leads to a proposal that the NFACT proteins could possibly catalyze a base-exchange reaction in which a regular RNA base is initially removed and replaced by a new base, similar to the action of the TGTs.

While more speculative in nature, it is possible to obtain certain clues regarding the nature of the potential target of the predicted NFACT-N catalytic domain: if NFACT-N were to act similar to FMN-DG, then among the endogenous bases in RNA uridine or cytidine contain spatially comparable carbonyl groups to FMN-DG substrates.[18] In addition to formation of pseudouridine and several pseudouridine derivatives, a range of rRNA base modifications have been previously characterized and include various forms of methylated, hydroxylated, and acetylated bases.[59,60] However, while pseudouridine and deazaguanine (e.g., preQ0) formation shares some kinship with base-exchange reactions and at least some modified rRNA bases do exist as free bases in the cell,[61] most well-studied base modifications typically proceed via direct enzymatic attachment of a chemical group on an existing base and not through a base-exchange reaction. Additionally, it is not clear that NFACT catalyzes a terminal reaction on a base: a potential exchanged base could, similar to preQ0 attachment via TGT, be followed by further modifications catalyzed by distinct enzymes. Another possibility is that NFACT could act as an rRNA repair enzyme, replacing damaged bases with normal bases. Oxidation of RNA bases, particularly in conjunction with cell death, is currently an area of emerging research interest;[62,63] given the connections between NFACT-N and FMN-DG, this possibility could warrant additional investigation.

This proposed role for NFACT in rRNA base-modification potentially unifies some of the disparate experimental findings on these proteins. The yeast NFACT protein Tae2 appears to activate the transcription factor Hsf1 in response to translational stress detected by the RQC,[11] ultimately leading to degradation of peptides derived from mRNAs lacking a stop codon.[12] It is conceivable that this proceeds via an effect of the proposed modification on translation fidelity consistent with the observed role of certain rRNA modifications, such as pseudouridylation, on ribosomal stability and translation fidelity.[64] Translation of Internal Ribosomal Entry Sites (IRES)-bearing mRNAs, which play a role in responding to stress conditions across eukaryotes,[65] also depends on rRNA base modifications like pseudouridylation.[66,67] Hence, it is conceivable that the proposed modification mediated by NFACT proteins might also intersect with IRES-mediated translation under stress. The *Drosophila* NFACT protein caliban has been proposed to be part of a network including p53, caspase, and Hid proteins during DNA damage-induced apoptosis.[13,14] Here again the modification could play a role as part of a stress response ensuing from DNA damage, as many key oncogenes and apoptosis factors are translated via IRES.

Certain studies have suggested that members of the NFACT family are virulence factors of pathogenic bacteria involved in

mediating adhesion and invasion in light of its capacity to bind fibronectin/fibrinogen.[6-10,68-72] These studies have also sought to demonstrate its presence in the extracellular space around some of these bacteria.[7,73,74] From an evolutionary perspective, however, the nearly universal phyletic distribution pattern of NFACT is indicative of a broadly conserved fundamental functional role. Hence, any role restricted to a subset of pathogenic species in the bacterial superkingdom is unlikely to represent an ancestral role. Building on this, the strong evolutionary conservation observed across the NFACT gene family is inconsistent with typical bacterial virulence factors involved in invasion, which are under strong selective pressure to diverge rapidly in response to evolving host factors.[75] NFACT's notable presence in the extracellular space is puzzling, particularly since groups studying it as FbpA and our own computational analyses failed to identify any means of transport to the cell surface either coded as a signal within NFACT (i.e., a signal anchor) or via other transport factors. One possible explanation for this, not explored in the current literature, is that the extracellular NFACT proteins are byproducts of cell lysis during biofilm formation;[76] consistent with this, these proteins have been identified in extracellular space occupied by biofilm with no apparent direct attachment to the cell membrane.[7,73,74] This would also be consistent with recent results questioning various aspects of potential direct NFACT roles in mediating adhesion and invasion during infection.[69,77-80] Thus, the proposed extracellular fibronectin/fibrinogen-binding role is unlikely to be a general one for this family and might merely reflect promiscuous interactions mediated by the extensive coiled-coil regions and facilitated by biofilm formation. However, we cannot rule out the possibility that in certain pathogenic bacteria NFACT might contribute specifically to biofilm formation or adhesion.

General conclusions

Through synthesis of sequence, genome, functional interaction, and structural data, we propose a potential role for the NFACT proteins conserved across the three superkingdoms of life in RNA-base modification in the context of translation. We present evidence that this function is mediated by the catalytic activity of the NFACT-N domain, shown here to be related to DNA glycosylases. We also predict an RNA-binding role for the NFACT-R (formerly DUF814) domain. In a diverse subset of bacteria lacking the canonical 4-thiouridylation pathway,[81] NFACT-R is also predicted to contribute to a novel variant of the tRNA 4-thiouridylation pathway. This pathway appears to combine the ThiI domain observed in the canonical tRNA 4-thiouridylation pathway, the TusA domain hitherto known only from the tRNA 2-thiouridylation pathway, and the NFACT-R domain. This observation points to a novel variation in the ancient thiolated base biosynthesis pathways in bacteria and shows how they have repeatedly drawn components from a common pool of players involved in synthesis of sulfur-containing metabolites.[48,82]

Materials and Methods

Iterative sequence profile searches were performed using the web implementation of the PSI-BLAST program[83] (with the following listed parameters different from default: -num_descriptions 20000, -evalue 20, -comp_based_stats 1, -pseudocount 30, -psi-blast_threshold 0.01) and web version 1.5 of the JACKHMMER (http://hmmer.janelia.org/search/jackhmmer) program run with default parameters against the non-redundant (NR) protein database at the National Center for Biotechnology Information (NCBI). Multiple sequence alignments were built by the Kalign2[84] and MUSCLE[85] programs with default parameters, followed by manual adjustments on the basis of profile–profile and structural alignments. Similarity-based clustering for both classification and culling of nearly identical sequences was performed using the BLASTCLUST program with empirically determined length and score threshold parameters (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The web-based implementation of the HHpred program[86] based on the HHsuite-2.0.15 software package with default parameters was used for profile–profile comparisons, searching against the pdb70 and pfamA_27.0 pre-configured databases. Structure similarity searches were performed using the DaliLite v. 3 program.[87] Secondary structures were predicted using the JPred 3 program with default parameters.[88] For previously known domains, the Pfam database release 27.0[20] was used as a guide and augmented by addition of newly detected divergent members. Structural visualization and manipulations were performed using the Open-Source PyMOL 1.5.0.3 (http://www.pymol.org) program. Funcoup3.0 was used to analyze contextual information based in interaction and expression data.[89]

Gene neighborhoods were determined using either the PTT file (downloadable from the NCBI ftp site) or the GenBank file in the case of whole genome shot gun sequences. The protein sequences of all neighbors were clustered using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) to identify related sequences in gene neighborhoods. Each cluster of homologous proteins were then assigned an annotation based on the domain architecture or conserved shared domain. Neighborhoods were further refined by ensuring that genes are unidirectional on the same strand of DNA and shared a putative common promoter to be counted as a single operon. If they were on opposite strands they were examined for potential bidirectional promoter sharing patterns. In-house Perl scripts were used to automate this analysis of genome context.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Burroughs AM and Aravind L's research is supported by the funds of the Intramural Research Program of the NIH, National Library of Medicine.

Supplemental Material

Supplemental material may be found here:
www.landesbioscience.com/journals/rnabiology/article/28302

References

1. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol 2000; 1:H0009; PMID:11178258; http://dx.doi.org/10.1186/gb-2000-1-5-research0009
2. Shao X, Grishin NV. Common fold in helix-hairpin-helix proteins. Nucleic Acids Res 2000; 28:2643-50; PMID:10908318; http://dx.doi.org/10.1093/nar/28.14.2643
3. Doherty AJ, Serpell LC, Ponting CP. The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. Nucleic Acids Res 1996; 24:2488-97; PMID:8692686; http://dx.doi.org/10.1093/nar/24.13.2488
4. Aravind L, Walker DR, Koonin EV. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res 1999; 27:1223-42; PMID:9973609; http://dx.doi.org/10.1093/nar/27.5.1223
5. Dramsi S, Bourdichon F, Cabanes D, Lecuit M, Fsihi H, Cossart P. FbpA, a novel multifunctional Listeria monocytogenes virulence factor. Mol Microbiol 2004; 53:639-49; PMID:15228540; http://dx.doi.org/10.1111/j.1365-2958.2004.04138.x
6. Christie J, McNab R, Jenkinson HF. Expression of fibronectin-binding protein FbpA modulates adhesion in Streptococcus gordonii. Microbiology 2002; 148:1615-25; PMID:12055283
7. Holmes AR, McNab R, Millsap KW, Rohde M, Hammerschmidt S, Mawdsley JL, Jenkinson HF. The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence. Mol Microbiol 2001; 41:1395-408; PMID:11580843; http://dx.doi.org/10.1046/j.1365-2958.2001.02610.x
8. Yamasaki T, Hitsumoto Y, Katayama S, Nogami Y. Fibronectin-binding proteins of Clostridium perfringens recognize the III1-C fragment of fibronectin. Microbiol Immunol 2010; 54:221-7; PMID:20377750; http://dx.doi.org/10.1111/j.1348-0421.2010.00201.x
9. Hennequin C, Janoir C, Barc MC, Collignon A, Karjalainen T. Identification and characterization of a fibronectin-binding protein from Clostridium difficile. Microbiology 2003; 149:2779-87; PMID:14523111; http://dx.doi.org/10.1099/mic.0.26145-0
10. Henderson B, Nair S, Pallas J, Williams MA. Fibronectin: a multidomain host adhesin targeted by bacterial fibronectin-binding proteins. FEMS Microbiol Rev 2011; 35:147-200; PMID:20695902; http://dx.doi.org/10.1111/j.1574-6976.2010.00243.x
11. Brandman O, Stewart-Ornstein J, Wong D, Larson A, Williams CC, Li GW, Zhou S, King D, Shen PS, Weibezahn J, et al. A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. Cell 2012; 151:1042-54; PMID:23178123; http://dx.doi.org/10.1016/j.cell.2012.10.044
12. Defenouillère Q, Yao Y, Mouaikel J, Namane A, Galopier A, Decourty L, Doyen A, Malabat C, Saveanu C, Jacquier A, et al. Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products. Proc Natl Acad Sci U S A 2013; 110:5046-51; PMID:23479637; http://dx.doi.org/10.1073/pnas.1221724110
13. Wang Y, Wang Z, Joshi BH, Puri RK, Stultz B, Yuan Q, Bai Y, Zhou P, Yuan Z, Hursh DA, et al. The tumor suppressor Caliban regulates DNA damage-induced apoptosis through p53-dependent and -independent activity. Oncogene 2013; 32:3857-66; PMID:22964637; http://dx.doi.org/10.1038/onc.2012.395
14. Bi X, Jones T, Abbasi F, Lee H, Stultz B, Hursh DA, Mortin MA. Drosophila caliban, a nuclear export mediator, can function as a tumor suppressor in human lung cancer cells. Oncogene 2005; 24:8229-39; PMID:16103875; http://dx.doi.org/10.1038/sj.onc.1208962
15. Lu AL, Li X, Gu Y, Wright PM, Chang DY. Repair of oxidative DNA damage: mechanisms and functions. Cell Biochem Biophys 2001; 35:141-70; PMID:11892789; http://dx.doi.org/10.1385/CBB:35:2:141
16. Yang W. Structure and mechanism for DNA lesion recognition. Cell Res 2008; 18:184-97; PMID:18157156; http://dx.doi.org/10.1038/cr.2007.116
17. Grin IR, Zharkov DO. Eukaryotic endonuclease VIII-like proteins: new components of the base excision DNA repair system. Biochemistry (Mosc) 2011; 76:80-93; PMID:21568842; http://dx.doi.org/10.1134/S000629791101010X
18. Prakash A, Doublié S, Wallace SS. The Fpg/Nei family of DNA glycosylases: substrates, structures, and search for damage. Prog Mol Biol Transl Sci 2012; 110:71-91; PMID:22749143; http://dx.doi.org/10.1016/B978-0-12-387665-2.00004-3
19. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996; 266:554-71; PMID:8743706; http://dx.doi.org/10.1016/S0076-6879(96)66035-2
20. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. The Pfam protein families database. Nucleic Acids Res 2012; 40:D290-301; PMID:22127870; http://dx.doi.org/10.1093/nar/gkr1065
21. Wallace SS, Bandaru V, Kathe SD, Bond JP. The enigma of endonuclease VIII. DNA Repair (Amst) 2003; 2:441-53; PMID:12713806; http://dx.doi.org/10.1016/S1568-7864(02)00182-9
22. Burroughs AM, Iyer LM, Aravind L. Functional diversification of the RING finger and other binuclear treble clef domains in prokaryotes and the early evolution of the ubiquitin system. Mol Biosyst 2011; 7:2261-77; PMID:21547297; http://dx.doi.org/10.1039/c1mb05061c
23. Zharkov DO, Shoham G, Grollman AP. Structural characterization of the Fpg family of DNA glycosylases. DNA Repair (Amst) 2003; 2:839-62; PMID:12893082; http://dx.doi.org/10.1016/S1568-7864(03)00084-3
24. Zharkov DO, Rieger RA, Iden CR, Grollman AP. NH2-terminal proline acts as a nucleophile in the glycosylase/AP-lyase reaction catalyzed by Escherichia coli formamidopyrimidine-DNA glycosylase (Fpg) protein. J Biol Chem 1997; 272:5335-41; PMID:9030608; http://dx.doi.org/10.1074/jbc.272.8.5335
25. Sun B, Latham KA, Dodson ML, Lloyd RS. Studies on the catalytic mechanism of five DNA glycosylases. Probing for enzyme-DNA imino intermediates. J Biol Chem 1995; 270:19501-8; PMID:7642635; http://dx.doi.org/10.1074/jbc.270.33.19501
26. Tchou J, Grollman AP. The catalytic mechanism of Fpg protein. Evidence for a Schiff base intermediate and amino terminus localization of the catalytic site. J Biol Chem 1995; 270:11671-7; PMID:7744806; http://dx.doi.org/10.1074/jbc.270.19.11671
27. Zharkov DO, Golan G, Gilboa R, Fernandes AS, Gerchman SE, Kycia JH, Rieger RA, Grollman AP, Shoham G. Structural analysis of an Escherichia coli endonuclease VIII covalent reaction intermediate. EMBO J 2002; 21:789-800; PMID:11847126; http://dx.doi.org/10.1093/emboj/21.4.789
28. Gilboa R, Zharkov DO, Golan G, Fernandes AS, Gerchman SE, Matz E, Kycia JH, Grollman AP, Shoham G. Structure of formamidopyrimidine-DNA glycosylase covalently complexed to DNA. J Biol Chem 2002; 277:19811-6; PMID:11912217; http://dx.doi.org/10.1074/jbc.M202058200
29. Jiang D, Hatahet Z, Melamede RJ, Kow YW, Wallace SS. Characterization of Escherichia coli endonuclease VIII. J Biol Chem 1997; 272:32230-9; PMID:9405426; http://dx.doi.org/10.1074/jbc.272.51.32230
30. Castaing B, Fourrey JL, Hervouet N, Thomas M, Boiteux S, Zelwer C. AP site structural determinants for Fpg specific recognition. Nucleic Acids Res 1999; 27:608-15; PMID:9862987; http://dx.doi.org/10.1093/nar/27.2.608
31. Sugahara M, Mikawa T, Kumasaka T, Yamamoto M, Kato R, Fukuyama K, Inoue Y, Kuramitsu S. Crystal structure of a repair enzyme of oxidatively damaged DNA, MutM (Fpg), from an extreme thermophile, Thermus thermophilus HB8. EMBO J 2000; 19:3857-69; PMID:10921868; http://dx.doi.org/10.1093/emboj/19.15.3857
32. Fromme JC, Verdine GL. DNA lesion recognition by the bacterial repair enzyme MutM. J Biol Chem 2003; 278:51543-8; PMID:14525999; http://dx.doi.org/10.1074/jbc.M307768200
33. Aravind L, Anantharaman V, Koonin EV. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins 2002; 48:1-14; PMID:12012333; http://dx.doi.org/10.1002/prot.10064
34. Aravind L, Koonin EV. THUMP--a predicted RNA-binding domain shared by 4-thiouridine, pseudouridine synthases and RNA methylases. Trends Biochem Sci 2001; 26:215-7; PMID:11295541; http://dx.doi.org/10.1016/S0968-0004(01)01826-6
35. Waterman DG, Ortiz-Lombardía M, Fogg MJ, Koonin EV, Antson AA. Crystal structure of Bacillus anthracis ThiI, a tRNA-modifying enzyme containing the predicted RNA-binding THUMP domain. J Mol Biol 2006; 356:97-110; PMID:16343540; http://dx.doi.org/10.1016/j.jmb.2005.11.013
36. Mueller EG, Palenchar PM. Using genomic information to investigate the function of ThiI, an enzyme shared between thiamin and 4-thiouridine biosynthesis. Protein Sci 1999; 8:2424-7; PMID:10595545; http://dx.doi.org/10.1110/ps.8.11.2424
37. Tesmer JJ, Klem TJ, Deras ML, Davisson VJ, Smith JL. The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. Nat Struct Biol 1996; 3:74-86; PMID:8548458; http://dx.doi.org/10.1038/nsb0196-74
38. Iwata-Reuyl D. An embarrassment of riches: the enzymology of RNA modification. Curr Opin Chem Biol 2008; 12:126-33; PMID:18294973; http://dx.doi.org/10.1016/j.cbpa.2008.01.041
39. McRobbie AM, Meyer B, Rouillon C, Petrovic-Stojanovska B, Liu H, White MF. Staphylococcus aureus DinG, a helicase that has evolved into a nuclease. Biochem J 2012; 442:77-84; PMID:22166102; http://dx.doi.org/10.1042/BJ20111903
40. Ikeuchi Y, Shigi N, Kato J, Nishimura A, Suzuki T. Mechanistic insights into sulfur relay by multiple sulfur mediators involved in thiouridine biosynthesis at tRNA wobble positions. Mol Cell 2006; 21:97-108; PMID:16387657; http://dx.doi.org/10.1016/j.molcel.2005.11.001
41. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 2001; 55:709-42; PMID:11544372; http://dx.doi.org/10.1146/annurev.micro.55.1.709
42. Shi R, Proteau A, Villarroya M, Moukadiri I, Zhang L, Trempe JF, Matte A, Armengod ME, Cygler M. Structural basis for Fe-S cluster assembly and tRNA thiolation mediated by IscS protein-protein interactions. PLoS Biol 2010; 8:e1000354; PMID:20404999; http://dx.doi.org/10.1371/journal.pbio.1000354

43. Alexeyenko A, Schmitt T, Tjärnberg A, Guala D, Frings O, Sonnhammer EL. Comparative interactomics with Funcoup 2.0. Nucleic Acids Res 2012; 40:D821-8; PMID:22110034; http://dx.doi.org/10.1093/nar/gkr1062
44. Aravind L. Guilt by association: contextual information in genome analysis. Genome Res 2000; 10:1074-7; PMID:10958625; http://dx.doi.org/10.1101/gr.10.8.1074
45. Huynen M, Snel B, Lathe W 3rd, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000; 10:1204-10; PMID:10958638; http://dx.doi.org/10.1101/gr.10.8.1204
46. Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, Chee GJ, Hattori M, Kanai A, Atomi H, et al. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res 2011; 39:3204-23; PMID:21169198; http://dx.doi.org/10.1093/nar/gkq1228
47. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 2002; 30:1427-64; PMID:11917006; http://dx.doi.org/10.1093/nar/30.7.1427
48. Iyer LM, Burroughs AM, Aravind L. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. Genome Biol 2006; 7:R60; PMID:16859499; http://dx.doi.org/10.1186/gb-2006-7-7-r60
49. Balaji S, Aravind L. The RAGNYA fold: a novel fold with multiple topological variants found in functionally diverse nucleic acid, nucleotide and peptide-binding proteins. Nucleic Acids Res 2007; 35:5658-71; PMID:17715145; http://dx.doi.org/10.1093/nar/gkm558
50. Watanabe M, Matsuo M, Tanaka S, Akimoto H, Asahi S, Nishimura S, Katze JR, Hashizume T, Crain PF, McCloskey JA, et al. Biosynthesis of archaeosine, a novel derivative of 7-deazaguanosine specific to archaeal tRNA, proceeds via a pathway involving base replacement on the tRNA polynucleotide chain. J Biol Chem 1997; 272:20146-51; PMID:9242689; http://dx.doi.org/10.1074/jbc.272.32.20146
51. Kittendorf JD, Barcomb LM, Nonekowski ST, Garcia GA. tRNA-guanine transglycosylase from Escherichia coli: molecular mechanism and role of aspartate 89. Biochemistry 2001; 40:14123-33; PMID:11714265; http://dx.doi.org/10.1021/bi0110589
52. Okada N, Nishimura S. Isolation and characterization of a guanine insertion enzyme, a specific tRNA transglycosylase, from Escherichia coli. J Biol Chem 1979; 254:3061-6; PMID:107167
53. Shindo-Okada N, Okada N, Ohgi T, Goto T, Nishimura S. Transfer ribonucleic acid guanine transglycosylase isolated from rat liver. Biochemistry 1980; 19:395-400; PMID:6986171; http://dx.doi.org/10.1021/bi00543a023
54. Koonin EV. Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res 1996; 24:2411-5; PMID:8710514; http://dx.doi.org/10.1093/nar/24.12.2411
55. Ramamurthy V, Swann SL, Paulson JL, Spedaliere CJ, Mueller EG. Critical aspartic acid residues in pseudouridine synthases. J Biol Chem 1999; 274:22225-30; PMID:10428788; http://dx.doi.org/10.1074/jbc.274.32.22225
56. Huang L, Pookanjanatavip M, Gu X, Santi DV. A conserved aspartate of tRNA pseudouridine synthase is essential for activity and a probable nucleophilic catalyst. Biochemistry 1998; 37:344-51; PMID:9425056; http://dx.doi.org/10.1021/bi971874+
57. Spedaliere CJ, Ginter JM, Johnston MV, Mueller EG. The pseudouridine synthases: revisiting a mechanism that seemed settled. J Am Chem Soc 2004; 126:12758-9; PMID:15469254; http://dx.doi.org/10.1021/ja046375s
58. Gu X, Liu Y, Santi DV. The mechanism of pseudouridine synthase I as deduced from its interaction with 5-fluorouracil-tRNA. Proc Natl Acad Sci U S A 1999; 96:14270-5; PMID:10588695; http://dx.doi.org/10.1073/pnas.96.25.14270
59. Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF. The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res 2011; 39:D195-201; PMID:21071406; http://dx.doi.org/10.1093/nar/gkq1028
60. Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowik A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother KM, et al. MODOMICS: a database of RNA modification pathways--2013 update. Nucleic Acids Res 2013; 41:D262-7; PMID:23118484; http://dx.doi.org/10.1093/nar/gks1007
61. Hall RH. Isolation of 3-Methyluridine and 3-Methylcytidine from Solubleribonucleic Acid. Biochem Biophys Res Commun 1963; 12:361-4; PMID:14070345; http://dx.doi.org/10.1016/0006-291X(63)90105-0
62. Kong Q, Lin CL. Oxidative damage to RNA: mechanisms, consequences, and diseases. Cell Mol Life Sci 2010; 67:1817-29; PMID:20148281; http://dx.doi.org/10.1007/s00018-010-0277-y
63. Li Z, Wu J, Deleo CJ. RNA damage and surveillance under oxidative stress. IUBMB Life 2006; 58:581-8; PMID:17050375; http://dx.doi.org/10.1080/15216540600946456
64. Sumita M, Desaulniers JP, Chang YC, Chui HM, Clos L 2nd, Chow CS. Effects of nucleotide substitution and modification on the stability and structure of helix 69 from 28S rRNA. RNA 2005; 11:1420-9; PMID:16120833; http://dx.doi.org/10.1261/rna.2320605
65. Spriggs KA, Stoneley M, Bushell M, Willis AE. Re-programming of translation following cell stress allows IRES-mediated translation to predominate. Biol Cell 2008; 100:27-38; PMID:18072942; http://dx.doi.org/10.1042/BC20070098
66. Yoon A, Peng G, Brandenburger Y, Zollo O, Xu W, Rego E, Ruggero D. Impaired control of IRES-mediated translation in X-linked dyskeratosis congenita. Science 2006; 312:902-6; PMID:16690864; http://dx.doi.org/10.1126/science.1123835
67. Jack K, Bellodi C, Landry DM, Niederer RO, Meskauskas A, Musalgaonkar S, Kopmar N, Krasnykh O, Dean AM, Thompson SR, et al. rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells. Mol Cell 2011; 44:660-6; PMID:22099312; http://dx.doi.org/10.1016/j.molcel.2011.09.017
68. Osanai A, Li SJ, Asano K, Sashinami H, Hu DL, Nakane A. Fibronectin-binding protein, FbpA, is the adhesin responsible for pathogenesis of Listeria monocytogenes infection. Microbiol Immunol 2013; 57:253-62; PMID:23586629; http://dx.doi.org/10.1111/1348-0421.12030
69. Barketi-Klai A, Hoys S, Lambert-Bordes S, Collignon A, Kansau I. Role of fibronectin-binding protein A in Clostridium difficile intestinal colonization. J Med Microbiol 2011; 60:1155-61; PMID:21349990; http://dx.doi.org/10.1099/jmm.0.029553-0
70. Ohara N, Ohara-Wada N, Kitaura H, Nishiyama T, Matsumoto S, Yamada T. Analysis of the genes encoding the antigen 85 complex and MPT51 from Mycobacterium avium. Infect Immun 1997; 65:3680-5; PMID:9284137
71. Armitige LY, Jagannath C, Wanger AR, Norris SJ. Disruption of the genes encoding antigen 85A and antigen 85B of Mycobacterium tuberculosis H37Rv: effect on growth in culture and in macrophages. Infect Immun 2000; 68:767-78; PMID:10639445; http://dx.doi.org/10.1128/IAI.68.2.767-778.2000
72. Gaultney RA, Gonzalez T, Floden AM, Brissette CA. BB0347, from the lyme disease spirochete Borrelia burgdorferi, is surface exposed and interacts with the CS1 heparin-binding domain of human fibronectin. PLoS One 2013; 8:e75643; PMID:24086600; http://dx.doi.org/10.1371/journal.pone.0075643
73. Torelli R, Serror P, Bugli F, Paroni Sterbini F, Florio AR, Stringaro A, Colone M, De Carolis E, Martini C, Giard JC, et al. The PavA-like fibronectin-binding protein of Enterococcus faecalis, EfbA, is important for virulence in a mouse model of ascending urinary tract infection. J Infect Dis 2012; 206:952-60; PMID:22782954; http://dx.doi.org/10.1093/infdis/jis440
74. Lin YP, Kuo CJ, Koleci X, McDonough SP, Chang YF. Manganese binds to Clostridium difficile Fbp68 and is essential for fibronectin binding. J Biol Chem 2011; 286:3957-69; PMID:21062746; http://dx.doi.org/10.1074/jbc.M110.184523
75. Aravind L, Anantharaman V, Zhang D, de Souza RF, Iyer LM. Gene flow and biological conflict systems in the origin and evolution of eukaryotes. Front Cell Infect Microbiol 2012; 2:89; PMID:22919680; http://dx.doi.org/10.3389/fcimb.2012.00089
76. Rice KC, Bayles KW. Molecular control of bacterial death and lysis. [table of contents.]. Microbiol Mol Biol Rev 2008; 72:85-109; PMID:18322035; http://dx.doi.org/10.1128/MMBR.00030-07
77. Innocentin S, Guimarães V, Miyoshi A, Azevedo V, Langella P, Chatel JM, Lefèvre F. Lactococcus lactis expressing either Staphylococcus aureus fibronectin-binding protein A or Listeria monocytogenes internalin A can efficiently internalize and deliver DNA in human epithelial cells. Appl Environ Microbiol 2009; 75:4870-8; PMID:19482952; http://dx.doi.org/10.1128/AEM.00825-09
78. Pracht D, Elm C, Gerber J, Bergmann S, Rohde M, Seiler M, Kim KS, Jenkinson HF, Nau R, Hammerschmidt S. PavA of Streptococcus pneumoniae modulates adherence, invasion, and meningeal inflammation. Infect Immun 2005; 73:2680-9; PMID:15845469; http://dx.doi.org/10.1128/IAI.73.5.2680-2689.2005
79. Noske N, Kämmerer U, Rohde M, Hammerschmidt S. Pneumococcal interaction with human dendritic cells: phagocytosis, survival, and induced adaptive immune response are manipulated by PavA. J Immunol 2009; 183:1952-63; PMID:19570831; http://dx.doi.org/10.4049/jimmunol.0804383
80. Delvecchio A, Currie BJ, McArthur JD, Walker MJ, Sriprakash KS. Streptococcus pyogenes prtFII, but not sfbI, sfbII or fbp54, is represented more frequently among invasive-disease isolates of tropical Australia. Epidemiol Infect 2002; 128:391-6; PMID:12113482; http://dx.doi.org/10.1017/S0950268802006787
81. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012; 40:D109-14; PMID:22080510; http://dx.doi.org/10.1093/nar/gkr988
82. Iyer LM, Zhang D, Burroughs AM, Aravind L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. Nucleic Acids Res 2013; 41:7635-55; PMID:23814188; http://dx.doi.org/10.1093/nar/gkt573
83. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res 2008; 36:W5-9; PMID:18440982; http://dx.doi.org/10.1093/nar/gkn201

84. Lassmann T, Frings O, Sonnhammer EL. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 2009; 37:858-65; PMID:19103665; http://dx.doi.org/10.1093/nar/gkn1006
85. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004; 5:113; PMID:15318951; http://dx.doi.org/10.1186/1471-2105-5-113
86. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009; 77(Suppl 9):128-32; PMID:19626712; http://dx.doi.org/10.1002/prot.22499
87. Holm L, Kääriäinen S, Rosenström P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics 2008; 24:2780-1; PMID:18818215; http://dx.doi.org/10.1093/bioinformatics/btn507
88. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 2008; 36:W197-201; PMID:18463136; http://dx.doi.org/10.1093/nar/gkn238
89. Schmitt T, Ogris C, Sonnhammer EL. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res 2014; 42:D380-8; PMID:24185702; http://dx.doi.org/10.1093/nar/gkt984