Bell et al. BMC Biology (2020) 18:159
https://doi.org/10.1186/s12915-020-00885-2

BMC Biology

RESEARCH ARTICLE Open Access

Modified base-binding EVE and DCD domains: striking diversity of genomic contexts in prokaryotes and predicted involvement in a variety of cellular processes

Ryan T. Bell, Yuri I. Wolf and Eugene V. Koonin*

Abstract

Background: DNA and RNA of all cellular life forms and many viruses contain an expansive repertoire of modified bases. The modified bases play diverse biological roles that include both regulation of transcription and translation, and protection against restriction endonucleases and antibiotics. Modified bases are often recognized by dedicated protein domains. However, the elaborate networks of interactions and processes mediated by modified bases are far from being completely understood.

Results: We present a comprehensive census and classification of EVE domains that belong to the PUA/ASCH domain superfamily and bind various modified bases in DNA and RNA. We employ the "guilt by association" approach to make functional inferences from comparative analysis of bacterial and archaeal genomes, based on the distribution and associations of EVE domains in (predicted) operons and functional networks of genes. Prokaryotes encode two classes of EVE domain proteins, slow-evolving and fast-evolving ones. Slow-evolving EVE domains in α-proteobacteria are embedded in conserved operons, potentially involved in coupling between translation and respiration, cytochrome c biogenesis in particular, via binding 5-methylcytosine in tRNAs. In β- and γ-proteobacteria, the conserved associations implicate the EVE domains in the coordination of cell division, biofilm formation, and global transcriptional regulation by non-coding 6S small RNAs, which are potentially modified and bound by the EVE domains. In eukaryotes, the EVE domain-containing THYN1-like proteins have been reported to inhibit PCD and regulate the cell cycle, potentially, via binding 5-methylcytosine and its derivatives in DNA and/or RNA. We hypothesize that the link between PCD and cytochrome c was inherited from the α-proteobacterial and proto-mitochondrial endosymbiont and, unexpectedly, could involve modified base recognition by EVE domains. Fast-evolving EVE domains are typically embedded in defense contexts, including toxin-antitoxin modules and type IV restriction systems, suggesting roles in the recognition of modified bases in invading DNA molecules and targeting them for restriction. We additionally identified EVE-like prokaryotic Development and Cell Death (DCD) domains that are also implicated in defense functions including PCD. This function was inherited by eukaryotes, but in
(Continued on next page)

* Correspondence: koonin@ncbi.nlm.nih.gov
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

(Continued from previous page)
animals, the DCD proteins apparently were displaced by the extended Tudor family proteins, whose partnership with Piwi-related Argonautes became the centerpiece of the Piwi-interacting RNA (piRNA) system.

Conclusions: Recognition of modified bases in DNA and RNA by EVE-like domains appears to be an important, but until now, under-appreciated, common denominator in a variety of processes including PCD, cell cycle control, antivirus immunity, stress response, and germline development in animals.

Keywords: Modified bases, DNA and RNA-binding domains, Programmed cell death, Cytochrome c, Self versus non-self-discrimination, Antivirus defense, Restriction-modification, Extended Tudor family proteins, piRNA pathway evolution

Background

DNA and different types of RNA of all organisms and diverse viruses contain a variety of modified bases. These derivatives of the canonical purines and pyrimidines perform a broad range of biological functions including regulation of transcription and translation as well as self- versus non-self-discrimination that is required for protection against biological defense and offense systems, such as restriction endonucleases and antibiotics [1–6]. The intricate networks of interaction and complex processes mediated by modified bases are far from being completely understood.

Modified bases are often recognized by dedicated protein domains. One such domain, widespread in eukaryotes and prokaryotes, is known as EVE (named for Protein Data Bank (PDB) structural identifier 2eve) [7]. Sequence and structure analyses have shown that the EVE domain is a member of the PUA (*p*seudo*u*ridine synthase and *a*rchaeosine transglycosylase)/ASCH (*ASC*-1 *h*omology) superfamily, a widely disseminated and apparently ancient assemblage of nucleic acid-binding domains [8–13]. These domains are generally associated with the translation apparatus, often fused to RNA modification enzymes, and bind RNA themselves [11–14]. Some ASCH domains have also been predicted to bind modified bases [15].

The first EVE domain to be characterized is found in mammalian thymocyte nuclear protein 1 (THYN1/Thy28), in which it comprises the highly conserved C-terminal region [7]. THYN1/Thy28 was identified as a reader of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), as well as further oxidized 5mC derivatives 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), in DNA [17]. Most eukaryotes encode orthologs of Thy28/THYN1 in which EVE is the only recognized domain, although fusions with AT-hook and other domains in fungi have been described, further supporting the role of EVE as a DNA-binding domain in these proteins [15]. The PUA-like SRA (*S*ET and *R*ING-associated) domain also binds 5mC and 5hmC DNA [17, 18]. However, a different PUA-like domain, YTH (*YT*521-B *h*omology), shows the closest structural similarity to EVE [10]. The YTH domain also binds modified bases, recognizing N^6-methyladenosine (m^6A) in RNA, in the case of eukaryotic proteins, and m^6A DNA, in the case of archaeal proteins [19, 20]. The conserved core of the PUA/ASCH superfamily consists of a 5-stranded β-barrel (Fig. 1), often with an α-helix between strands 1 and 2, a structural element that is present in EVE domains, which also contain an additional sixth strand in the β-barrel [10] (Fig. 1).

THYN1/Thy28 was originally identified as one of about 300 previously uncharacterized genes that are preferentially expressed in human CD34+ hematopoietic stem/progenitor cells [21]. Shortly afterwards, a cDNA was isolated from apoptotic avian thymocytes encoding a 242 amino acid protein with 88% amino acid similarity to THYN1/Thy28 [22]. Initial cloning and characterization of murine THYN1/Thy28 established nuclear localization and found protein levels to be the highest in testis, with thymus, spleen, liver, and kidney also displaying substantial expression [23]. In a more recent study, nuclear THYN1/Thy28 has been detected in nearly all human tissues [24].

Several studies have explored the role of THYN1/Thy28 in lymphocyte model systems where programmed cell death (PCD), also known as apoptosis, can be induced by antibody treatment. Decreased THYN1/Thy28 protein expression was observed following induction, suggesting that downregulation of this gene is associated with apoptosis initiation [23]. Conversely, overexpression of THYN1/Thy28 was correlated with inhibition of several apoptotic events, such as loss of mitochondrial membrane potential and caspase-3 activation [25]. Furthermore, these experiments have demonstrated accumulation of cells in G1 phase following THYN1/Thy28 overexpression, suggesting that this protein is involved in the regulation of cell cycle progression.

We were interested in the apparently diverse but poorly characterized functions of the EVE domains, and in particular, in the potential roles of modified base recognition in various biological processes. Here, we report a comprehensive bioinformatic analysis of the broad phyletic distribution of EVE-like domains, with an emphasis on the radiation among Proteobacteria, intriguing associations with base modification-dependent

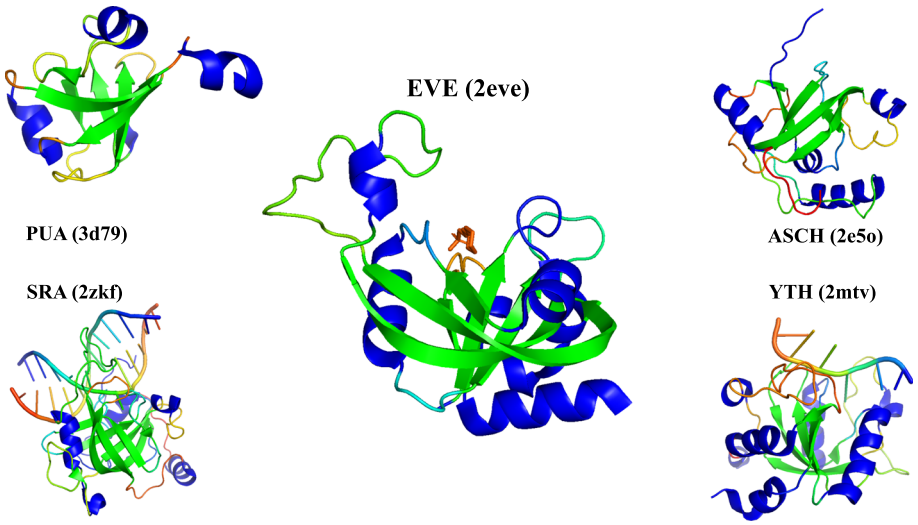

Fig. 1 Structures of EVE and other PUA/ASCH superfamily members. Structures were downloaded from PDB (identifiers in parentheses) and drawn using the PyMOL program [16]. β-strands are colored green, α-helices are colored blue, and loops and ligands are rainbow colored

restriction and toxin-antitoxin systems, and the identification of the Development and Cell Death (DCD) domain as a member of the EVE-like superfamily. We apply the "guilt by association" approach [26–30] to make functional inferences from an extensive comparative analysis of the expanded collection of bacterial and archaeal genomes.

Results

A census of EVE proteins

Our search for EVE proteins using PSI-BLAST and HHpred seeded with profiles derived from multiple alignments of the amino acid sequences of known EVE domains (see "Materials and methods" for details) showed that the EVE domain is most prevalent among Proteobacteria, which harbor the majority of all prokaryotic EVE proteins detected (Additional file 1: Fig. S1) and a plurality of all EVE proteins. CLANS analysis [31] of EVE domains extracted from all EVE proteins in the dataset revealed a diverse cloud of sequences, with four well-defined clusters (Fig. 2). The largest cluster (blue in Fig. 2) consists, mostly, of sequences from β- and γ-proteobacteria, as well as those from the metazoa and fungi. The second largest cluster (red) includes mostly sequences from α-proteobacteria and Bacteroidetes, as well as the majority of plant sequences. Two smaller, almost completely prokaryotic clusters were also identified. The first (green) represents a collection of sequences largely from Proteobacteria, Actinobacteria, and Bacteroidetes. These EVE domains are usually encoded in operonic contexts which imply a role in ligand-activated transcriptional regulation. The second (purple) is mostly made up of sequences from γ-proteobacteria, Firmicutes, and Bacteroidetes and is unique in that the EVE domains in this group are almost always fused to a GNAT-like (*G*CN5-related *N*-*a*cetyl-*t*ransferase) domain.

We chose to focus our initial analysis on the two large clusters which consist, mostly, of proteobacterial EVE domains. α-proteobacteria were the most abundant class in the data, from which the majority of sequences in the second largest cluster (red in Fig. 2) derive.

EVE in α-proteobacteria

The EVE proteins of this class (Fig. 3) are frequently located in a putative operon with the tRNA N^6-adenosine threonylcarbamoyltransferase TsaD, glycerol-3-phosphate dehydrogenase GpsA, and YciI, a small ferredoxin-fold protein homologous to muconolactone isomerases [32]. The sequences of the EVE domains in this group are readily recognizable (RPS-BLAST *E*-values of ~ 1e−42 or better with the pfam01878 query) and form a tight, well-conserved collection with within-group divergence comprising only 35% of the overall divergence between EVE domains (see "Materials and methods" for details).

This highly conserved directional unit (TsaD→GpsA→YciI→EVE) is itself strongly associated with another predicted operon which encodes 3 enzymes of heme biosynthesis, namely, porphobilinogen deaminase (HemC), uroporphyrinogen-III synthase (HemD), and coproporphyrinogen oxidase (HemY/HemG), as well as a diverged homolog of HemX, a putative uroporphyrinogen-III C-methyltransferase that is also homologous to IMMP (*i*nner *m*embrane *m*itochondrial *p*rotein, also known as mitofilin) [33]. In Rhodobacteraceae, HemC is

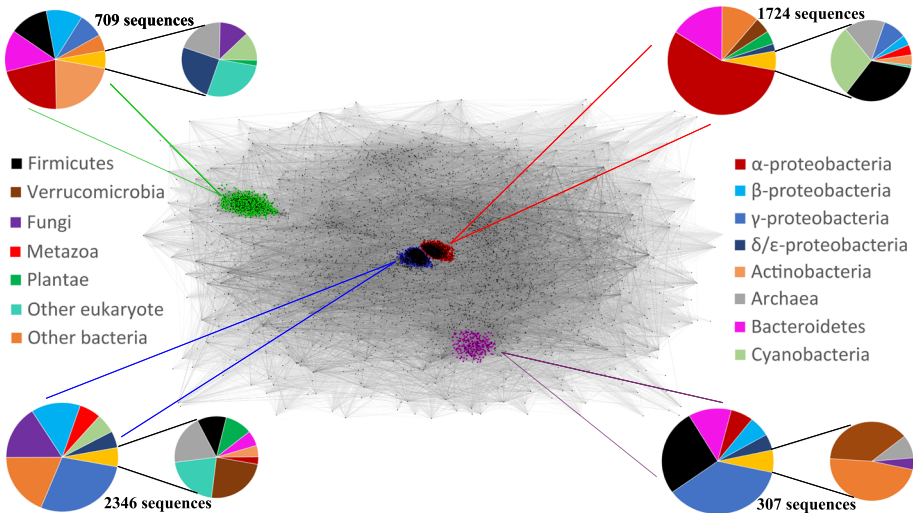

Fig. 2 CLANS analysis of EVE domains. A 2D projection of CLANS clustering analysis of 8403 representative EVE domain sequences. Each sequence is depicted by a dot, and sequence similarity detected by BLAST is indicated by a line, colored in shades of gray according to the BLAST *p* value. Four primary clusters were observed, marked with colors which match colored lines extending toward the pie chart which corresponds to the cluster. The protein sequences used for the CLANS analysis are available as Additional file 2: Dataset 1

missing from this generally well-conserved gene order. Head to head orientation of these putative operons suggests that the promoter regions might overlap, allowing for co-regulation.

The association between the EVE domain and cytochrome c biosynthesis via regulation of heme production in α-proteobacteria is further emphasized by the presence of a cytochrome c biosynthetic cluster (CcmC through CcmI) adjacent to the EVE domain that is conserved in both the Acetobacteraceal branch of Rhodospirillales and Sphingomonadales (Fig. 3) [34]. In Sphingomonadales, a likely operon including the tRNA-modifying enzyme MiaB, which adds a methylthio group to N^6-isopentenyladenosine at position 37 in many tRNAs decoding UNN (the same position modified by TsaD), often occurs between the EVE domain and the cytochrome c biosynthetic operon [35].

A contextual information network graph generated from the pairwise domain associations in prokaryotic EVE protein genomic neighborhoods showed that in α-,

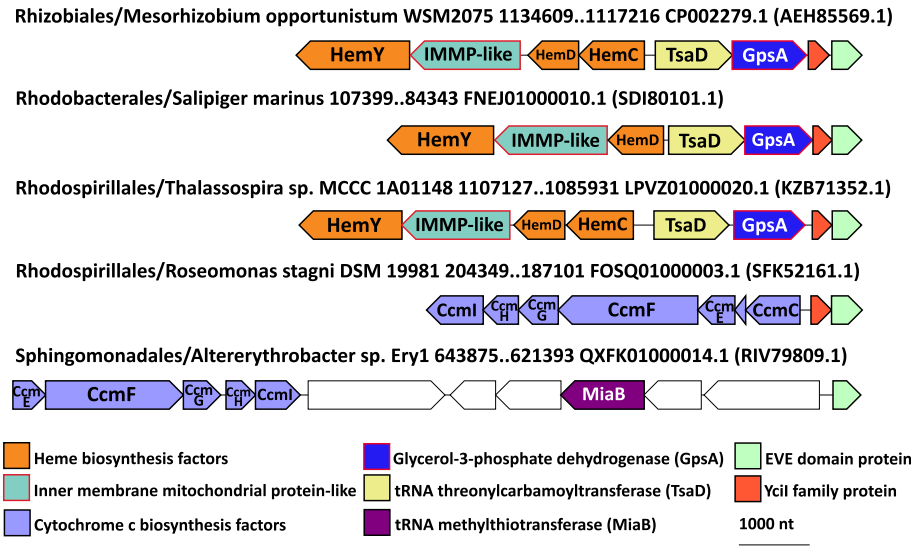

Fig. 3 Conserved genomic context of EVE proteins in α-proteobacteria. Representative EVE protein neighborhoods from α-proteobacteria. Genes are shown as arrows from 5′ to 3′. The order of α-proteobacteria, species, and genomic coordinates for each neighborhood are indicated, as are the GenBank genome accessions and, in parentheses, the GenBank accessions for each EVE protein

β-, and γ-proteobacteria, respectively, the EVE proteins are associated with highly conserved, but largely non-overlapping gene complements (Fig. 4).

EVE in β- and γ-proteobacteria

A prominent exception to the general lack of overlap between the contextual information networks among Proteobacteria is the conservation between β- and γ-proteobacteria of an apparent operonic linkage of EVE proteins and the cell division proteins ZapA and ZapB (Fig. 4). The sequences of the EVE domains in these proteins are also highly recognizable, slightly more so, in fact, than those in α-proteobacteria (RPS-BLAST *E*-values of ~ 5e−55 or better). They likewise form a tight, well-conserved group, with their within-group divergence accounting for only 32% of the overall divergence between EVE domains. The protein-coding gene array ZapB→ZapA→EVE also contains, between ZapA and EVE, a non-coding 6S RNA (*ssrS*) gene. Our analysis of these neighborhoods suggests that the *ssrS* gene is (nearly) always present, based on the positions of the protein-coding genes, leaving a gap sufficient to accommodate the 6S RNA, but are not consistently annotated, conceivably, due to sequence divergence. For this reason, *ssrS* was not included in our calculations that produced the contextual information network graph (Fig. 4).

In many species of γ-proteobacteria and some β-proteobacteria, the enzyme FAU1/MFTHFS, also known as YgfA, a putative 5-formyltetrahydrofolate cyclo-ligase, is encoded between ZapB→ZapA→SsrS and the EVE protein (Fig. 5). In γ-proteobacteria, another directional gene array is frequently found adjacent to this predicted operon in a head to head orientation, with the potential for the promoter regions to overlap. It encodes an uncharacterized conserved protein (YgfB), an Xaa-Pro aminopeptidase (PepP), and a homolog of 2-octaprenyl-6-methoxyphenol 4-hydroxylase (UbiH), an FAD-dependent oxidoreductase, as well as a homolog of 2-octaprenylphenol 6-hydroxylase (UbiI), both of which are involved in ubiquinone biosynthesis (Fig. 5) [36]. Many of the γ-proteobacterial neighborhoods additionally include genes encoding homologs of ribose-5 phosphate isomerase (RpiA) and L-threonine dehydratase (IlvA).

In β-proteobacteria, the gene coding for the EVE protein is often followed by a gene encoding the ortholog of the TauE sulfite export protein (Fig. 5). In Burkholderiaceae and Neisseriales, a cobalamin (vitamin B-12) biosynthetic cluster is often found immediately adjacent to the ZapAB→SsrS→(FAU1)→EVE unit (Fig. 5). In Burkholderiales, a conserved region encoding a cytochrome c551/c552 family protein, dihydroxy-acid dehydratase (IlvD), a putative transcriptional regulator related to LysR, and prolipoprotein diacylglyceryl transferase (LGT) is adjacent to the ZapAB→SsrS→EVE putative operon.

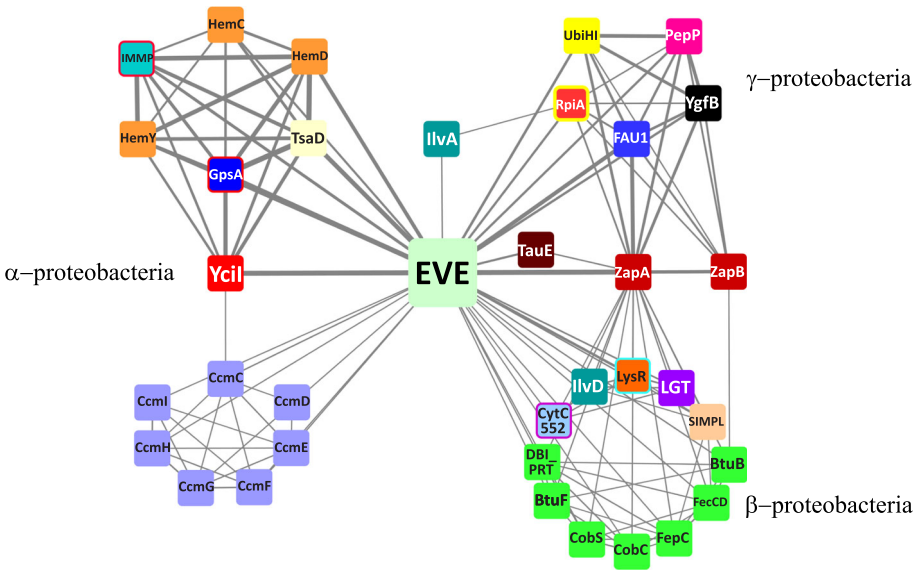

Fig. 4 The most prevalent contextual associations of EVE proteins in the genomes of Proteobacteria. A contextual information network graph of domains detected among genes in the genomic neighborhoods encoding EVE proteins in Proteobacteria. The nodes are clustered by mutual connections, which generally correlate with the distribution of species in which they occur. The thickness of the edges reflects the strength of the association. Nodes representing domains that participate in the same pathway are drawn with the same color. Domain coloring and abbreviations are explained in the text and Figs. 3 and 5. The graph was calculated from the top 400 pairwise associations between domains in EVE protein neighborhoods from 13,388 genomes with unique domain compositions. Singleton associations with EVE and minor networks were removed for clarity

The recently updated phylogeny of the β- and γ-proteobacteria [37] allows some inferences to be made concerning the evolutionary history of the predicted functional systems containing the EVE domain. The taxonomic distribution of the ZapAB→SsrS→(FAU1)→EVE unit covers β-proteobacteria, several early branching members of γ-proteobacteria (Xanthomonadales, Chromatiales, Methylococcales, etc.), and the clade primarily consisting of Pseudomonadales and Oceanospirillales. This broad taxonomic representation implies that the unit was present in the common ancestor of β- and γ-proteobacteria. The VAAP clade (Vibrionales, Alteromonadales, Aeromonadales, and Pasteurellales) have lost this association, and each order, with the exception of Aeromonadales, possesses distinct conserved regions neighboring encoded EVE proteins (Fig. 12, Additional file 1: Figs. S2–4). In *E. coli* K-12, both of the typical EVE-associated γ-proteobacterial operons and their orientations are conserved, but the ZapB and EVE domain proteins have been lost (Fig. 5).

In agreement with the CLANS results, in the phylogenetic tree of the EVE domains, the EVE proteins from most of the higher plants branch from within the α-proteobacterial clade, whereas EVE proteins from the metazoa, fungi, and some plants are more similar to γ-proteobacterial domains, but lie outside of the γ-proteobacterial variation (Additional file 1: Fig. S5). Due to the small size of the EVE domain, phylogenetic analysis cannot confidently identify the prokaryotic ancestry of these domains in eukaryotes, although Proteobacteria are the most likely contributors, with possible multiple acquisitions.

EVE domains in putative ligand-activated antibiotic resistance and other ligand-activated responses

The largest of the almost exclusively prokaryotic clusters from our CLANS analysis (green in Fig. 2) was populated predominantly by domains encoded in the operonic context of a transcription factor and a small molecule ligand-binding domain (Fig. 6). The most frequent putative operons encoded an EVE domain with either a MarR (*m*ultiple *a*ntibiotic *r*esistance) family transcription factor or a YafY family transcription factor. YafY-like factors are a fusion of a putative DNA-binding HTH domain and a WYL domain, a ligand-binding regulator of prokaryotic defense systems [38–40]. The MarR-EVE and YafY-EVE pairs are further associated, most frequently, with a ligand-binding domain of the SPRBCC (*S*TART/*R*HO_alpha_C/PITP/*B*et_v1/*C*oxG/*C*alC) or EhpR (phenazine antibiotic resistance) families. EhpR family proteins contain a vicinal oxygen chelate (VOC) domain, and other VOC domain homologs are also frequently encoded in the neighborhoods of this class of EVE proteins, often replacing SPRBCC domains in

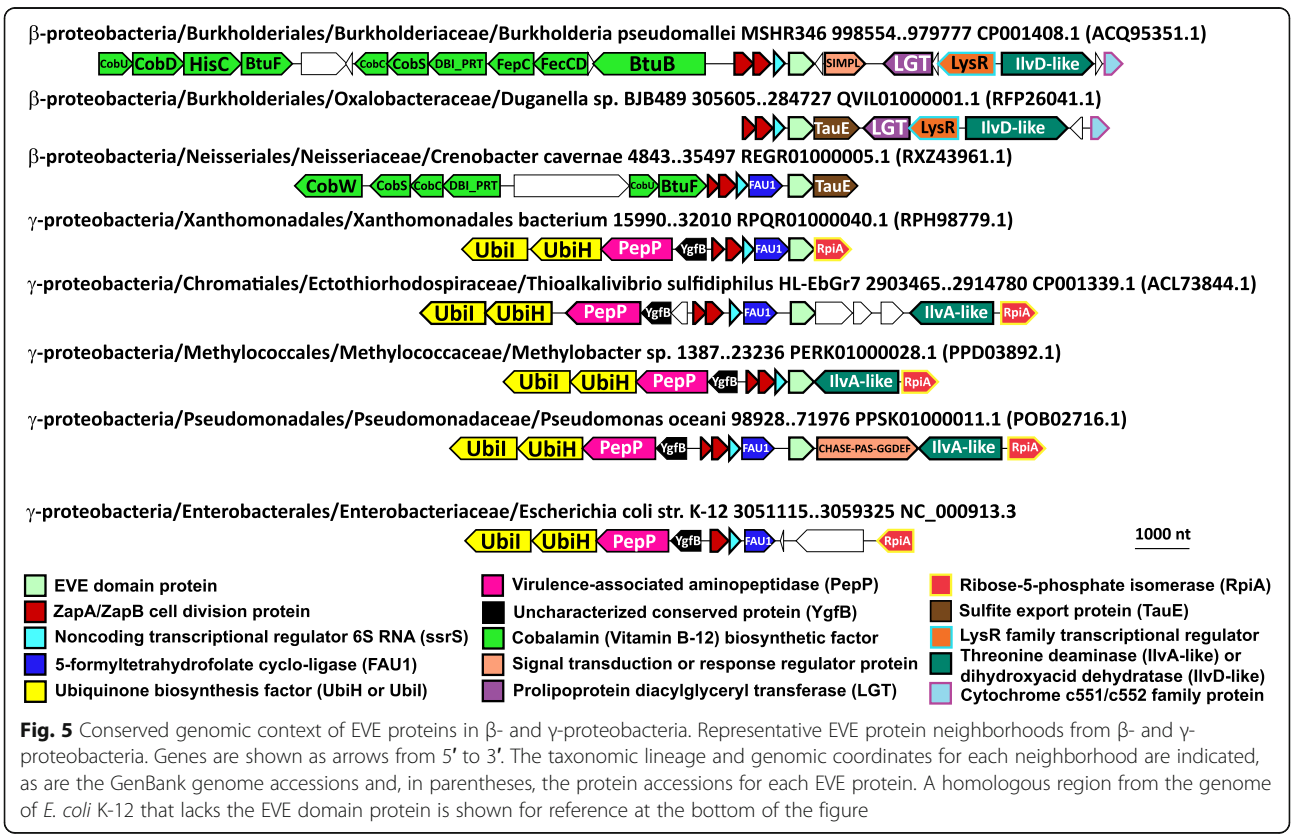

Fig. 5 Conserved genomic context of EVE proteins in β- and γ-proteobacteria. Representative EVE protein neighborhoods from β- and γ-proteobacteria. Genes are shown as arrows from 5′ to 3′. The taxonomic lineage and genomic coordinates for each neighborhood are indicated, as are the GenBank genome accessions and, in parentheses, the protein accessions for each EVE protein. A homologous region from the genome of *E. coli* K-12 that lacks the EVE domain protein is shown for reference at the bottom of the figure

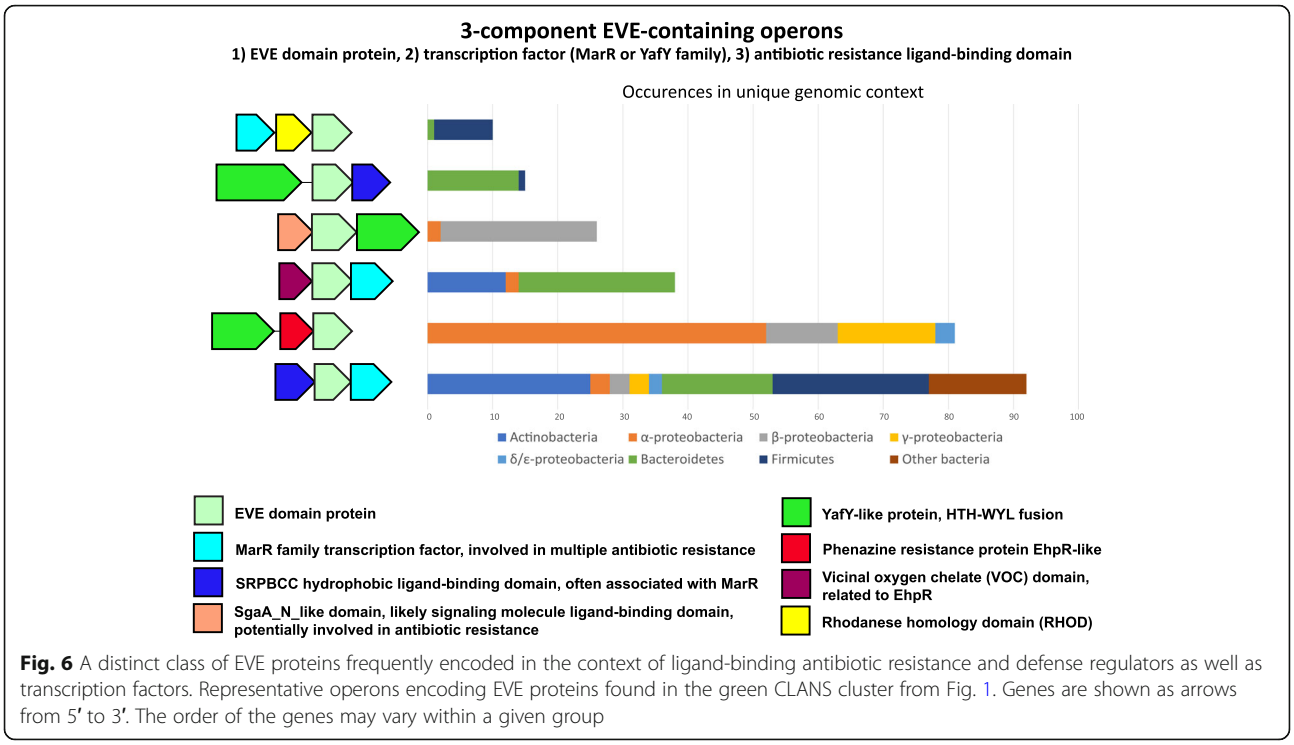

Fig. 6 A distinct class of EVE proteins frequently encoded in the context of ligand-binding antibiotic resistance and defense regulators as well as transcription factors. Representative operons encoding EVE proteins found in the green CLANS cluster from Fig. 1. Genes are shown as arrows from 5′ to 3′. The order of the genes may vary within a given group

association with MarR-EVE pairs (Fig. 6). The sequences of this group of EVE domains formed a distinct clade in our phylogenetic analysis (Additional file 1: Fig. S5). The regions surrounding these apparent 3-component systems are highly diverse. They include putative defense functions in *Nocardia* and related genera, where multiple paralogs of UvrD-like helicase domains fused to Cas4-like PD(D/E)XK phosphodiesterases [41] are present (Additional file 1: Fig. S6). Conversely, in *Azospirillum*, the neighborhoods include translation factor genes and cytochrome c biosynthesis operons, a context that is, surprisingly, closely similar to the distinct classes of EVE proteins in the two largest clusters in our CLANS analysis (Additional file 1: Figs. S7).

EVE as a specificity domain in modification-dependent restriction systems

The EVE proteins in Proteobacteria and eukaryotes found in the two largest clusters we observed with CLANS analysis show high levels of sequence conservation. By contrast, many genome defense systems encompass EVE domains with more pronounced sequence diversity. These domains range from highly significant matches to hits with weaker similarity, and many could be detected only with sensitive methods such as HHpred. A substantial variety of putative modification-dependent (type IV) restriction endonucleases (REs) with core architectures of EVE-PD(D/E)XK phosphodiesterase and EVE-HNH endonuclease were identified in our searches (Fig. 7). Furthermore, we identified numerous proteins containing fusions of the EVE domain with nucleases of the phospholipase D (PLDc) or GIY-YIG superfamilies (Fig. 7). Rare fusions to homologs of the glucosylated 5hmC-dependent RE GmrSD were detected as well.

The EVE domain is also frequently incorporated into homologs of the GTP-dependent DNA translocase McrB. In *E. coli* K-12, McrB, in concert with McrC, a PD(D/E)XK-type nuclease that interacts with McrB hexamers via its N-terminal domain, restricts N4-methylcytosine (4mC)/5mC/5hmC-containing DNA; in this strain, EVE is replaced with a DUF3578 family domain as the specificity module [20, 42–44]. Overall, the EVE-McrB combination is the most common domain architecture among the EVE-containing proteins in defense systems, represented in nearly 300 bacterial and archaeal genera, and is particularly abundant among Firmicutes and Bacteroidetes (Fig. 7).

In diverse archaea, a recurrent partnership was observed between standalone EVE domain proteins and a predicted, uncharacterized restriction system that encodes a SWI2/SNF2 helicase fused to a nuclease (PD(D/E)XK or PLDc family). This gene is expressed in an operon that also encodes a methyltransferase of COG1743 and an uncharacterized DUF499-containing protein (Additional file 1: Fig. S8). Our analysis showed that DUF499 is homologous to CDC6/ORC1 ATPases, which are involved in the recognition of the origin of DNA replication in archaea and eukaryotes [45, 46].

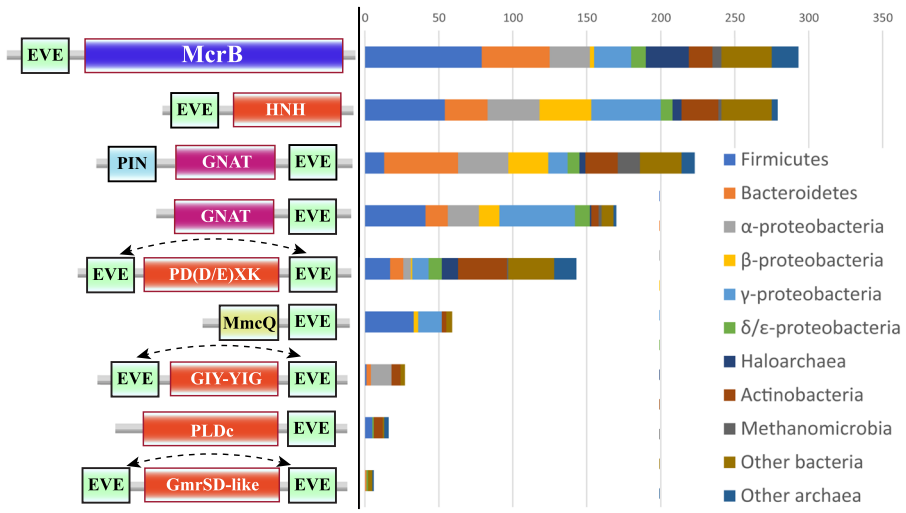

Fig. 7 Phyletic distributions of classes of EVE fusion proteins with predicted roles in prokaryotic modification-dependent restriction and toxin-antitoxin systems. The most common classes of EVE fusion proteins in prokaryotes, with one representative chosen per genus. The representatives contain the core elements depicted (not to scale), with additional domains often present, especially in EVE-McrB proteins. The dotted lines with an arrow at each end indicate that the EVE domain can occur in either position, but not both

EVE domains in toxin-antitoxin systems

A major class of EVE proteins which formed a distinct cluster in our CLANS analysis (purple in Fig. 2) is a fusion of EVE to the C-terminus of a GNAT-like acetyltransferase, often with a PIN RNase domain at the N-terminus (Fig. 7). GNAT and PIN domains both frequently function as toxins [47–49]. This variety of EVE proteins has been described previously in some detail by Iyer et al., who proposed that these proteins acetylate a DNA base, although the frequent presence of a PIN domain suggests that these systems employ RNA as a target or guide [15]. As also addressed in that study, almost all (PIN)-GNAT-EVE operons encode a protein containing a second PUA-like domain, ASCH, and often, also, an AAA+ ATPase of the AAA_17 family. In some cases, mostly in α-proteobacteria, the ASCH domain is fused to a helix-turn-helix (HTH) DNA-binding domain of the xenobiotic response element (XRE) family.

The distributions of the PIN-GNAT-EVE and the GNAT-EVE fusion proteins among prokaryotes are notably different (Fig. 8). The PIN-GNAT-EVE proteins are frequently found in bacterial and archaeal genomes in a close association with type I restriction-modification (RM) systems (HsdR/M/S operons). A consistent proximity between PIN-GNAT-EVE proteins and other types of defense systems, such as CRISPR-Cas and the COG1743→DUF499→SWI2/SNF2 helicase-nuclease operon described above, was also observed (Fig. 8). By contrast, GNAT-EVE proteins are not associated with type I RM systems but are commonly located within prophages in β- and γ-proteobacteria (Fig. 8).

In addition to the profusion of putative TA systems containing EVE domains, EVE is also regularly found as a standalone protein closely associated with type I RM systems (Fig. 10). These systems often also contain an ASCH domain and are mostly found in archaea. In effect, RM systems exhibit toxin-antitoxin functionality, with the restriction endonuclease playing the role of toxin, whereas the methyltransferase is its antitoxin [47, 50–53]. Accordingly, the EVE domains are likely to play similar roles in these systems, namely, targeting the toxins (including restriction endonucleases) to modified nucleic acids.

MmcQ/YjbR-EVE fusion proteins

Related to the RM and TA system-associated EVE proteins is a class of MmcQ/YjbR-EVE fusions that we found associated with a number of defense gene clusters, as well as signaling, transport, and metabolic factors, mostly, in Firmicutes and γ-proteobacteria (Fig. 7). MmcQ/YjbR (PF04237) has a CyaY-like fold and is also fused to tellurite resistance protein TerB and GNAT-type acetyltransferases in other contexts [54]. MmcQ/YjbR-EVE fusions also frequently contain an N-terminal DUF1831 domain, and in many cases, where this domain is missing, there is a DUF1831-MmcQ/YjbR gene immediately adjacent to MmcQ/YjbR-EVE.

DUF1831-MmcQ/YjbR-EVE fusions, which are the most numerous in our data, are frequently encoded within a genomic context that includes sensor histidine kinases, response regulators, and putative DNA-binding proteins. They are also often associated with ABC-type transport system components. Intriguingly, the large number of currently available *Streptococcus* genomes

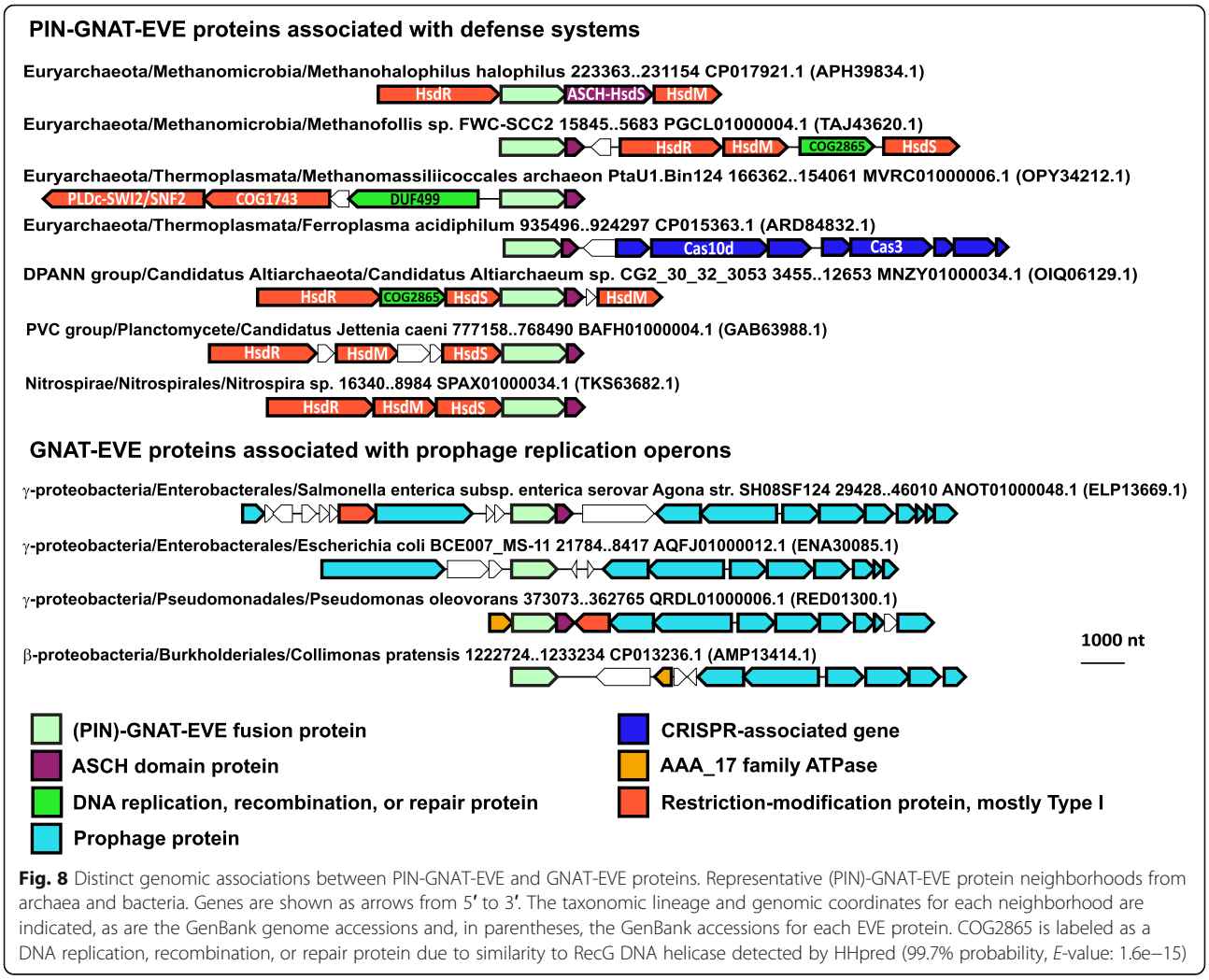

Fig. 8 Distinct genomic associations between PIN-GNAT-EVE and GNAT-EVE proteins. Representative (PIN)-GNAT-EVE protein neighborhoods from archaea and bacteria. Genes are shown as arrows from 5′ to 3′. The taxonomic lineage and genomic coordinates for each neighborhood are indicated, as are the GenBank genome accessions and, in parentheses, the GenBank accessions for each EVE protein. COG2865 is labeled as a DNA replication, recombination, or repair protein due to similarity to RecG DNA helicase detected by HHpred (99.7% probability, *E*-value: 1.6e−15)

enabled the detection of highly variable regions adjacent to the genes encoding DUF1831-MmcQ/YjbR-EVE proteins in conserved positions. These areas often contain mobile genetic elements (MGEs), defense-associated genes (TA modules, CRISPR-Cas systems), as well as uncharacterized, putative defense, transport, secretory, and DNA/protein repair genes (a MsrAB/disulfide interchange factor operon we detected is likely a mobile protein repair system) [55] (Fig. 9). These hotspots for integration (and presumably contraction) adjacent to (DUF1831)-MmcQ/YjbR-EVE genes often include transposases, implying a transposon-type mechanism of mobilization. When these variable gene arrays are large, ancestral, independent mobile modules that were assembled to give rise to them can be predicted by comparison with genomes in which the array is smaller (Fig. 9). Further work will be necessary to establish the relationship in these systems between the mobile genes and those conserved at the borders, including DUF1831-MmcQ/YjbR-EVE. The fusion of MmcQ/YjbR-EVE to a transposase in *Streptococcus lutetiensis* further underscores that this variety of EVE protein might play a role in regulating the acquisition and/or expression of MGEs. We also observed a similar phenomenon in the regions neighboring MmcQ/YjbR-EVE genes in *Actinobacillus* (Additional file 1: Fig. S4).

DCD, an EVE-like domain involved in restriction of modified DNA and PCD in plants

We further identified the Development and Cell Death (DCD) domain as a specificity module comparable in sequence and genomic context to EVE. The DCD domain is rare in prokaryotes and, mostly, is present in archaea and hyperthermophilic bacteria. The DCD domain was originally identified in proteins that are strongly induced during plant development, the hypersensitive response to avirulent pathogens, and reaction to various environmental stresses in plants [56–58]. Although not classified as such previously, we conclude that DCD is a member of the PUA-like superfamily due to the limited

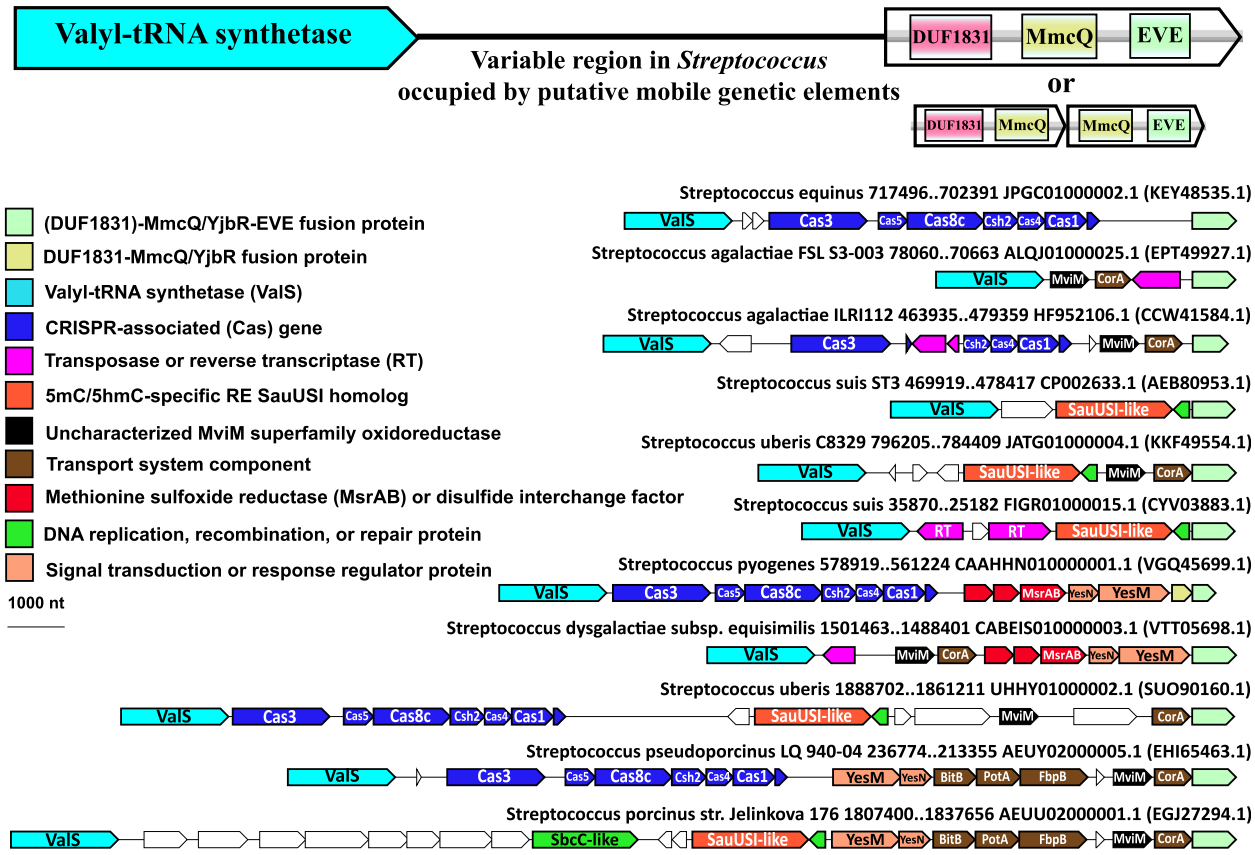

Fig. 9 MmcQ/YjbR-EVE fusion proteins in *Streptococcus*. Representative (DUF1831)-MmcQ/YjbR-EVE protein neighborhoods from *Streptococcus*. Genes are shown as arrows from 5′ to 3′. The genomic coordinates for each neighborhood are indicated, as are the GenBank genome accessions and, in parentheses, the GenBank accessions for each EVE protein. The small "DNA repair" proteins adjacent to the SauUSI homologs are homologs of MutT pyrophosphohydrolase

but significant sequence similarity with EVE detected by profile-profile comparison using HHpred (97.16% probability, *E*-value 0.049). Several of the most highly conserved residues of the EVE domains are present in the DCD domains, and the characteristic secondary structure (βαβαββββ) that forms the EVE β-barrel is also predicted for DCD (Additional file 1: Fig. S9). The DCD domain shows some associations similar to those of defense-related EVE domains, in particular, with type I restriction systems, as well as a fusion to PD(D/E)XK phosphodiesterases and McrB-like domains, and is distinguished by frequent fusion to a Rossmann-fold methyltransferase, which is extremely rare among EVE domains (Fig. 10). These connections imply that, similarly to EVE, DCD domains in prokaryotes recognize methylated bases in DNA and thus contribute to restriction of modified DNA. DCD-methyltransferase fusion protein genes are usually followed by a gene encoding a PD(D/E)XK nuclease, suggesting that they are involved in the additional methylation of modified DNA, recognized by DCD, that could be restricted in the absence of the supplementary methylation.

DCD and YTH: EVE-like domains with roles in modification-dependent DNA restriction systems and eukaryotic modification-based mRNA processing

We performed a comprehensive search for the DCD domain in all available genomes and found that, among eukaryotes, it is not restricted to plants, as originally described, but is also present in many chromist genomes, particularly, in heterokont and haptophyte algal proteins, where it is often fused to another EVE-like domain, YTH (Fig. 11a). The YTH domain is broadly distributed in eukaryotes, has been consistently reported to bind m^6A in eukaryotic mRNAs, and is involved in multiple processes including splicing and polyadenylation, translation/decay balance (notably triaging of mRNA translation during stress), and inhibition of viral RNA replication [19, 60–63]. When fused to the YTH domain in eukaryotes, the DCD domain is also fused to a KH (*K homology*) domain,

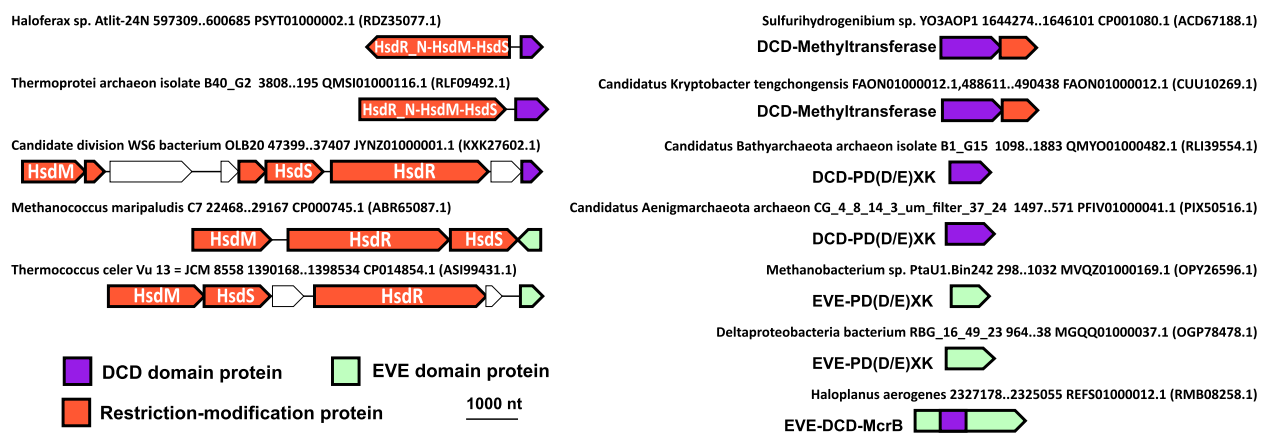

Fig. 10 DCD is an EVE-like domain. Representative DCD and EVE protein neighborhoods from archaea and bacteria. Genes are shown as arrows from 5′ to 3′. The source organism and genomic coordinates for each neighborhood are indicated as are the GenBank genome accessions and, in parentheses, the GenBank accessions for each EVE and DCD protein

and an array of CCCH-type zinc finger (Znf) domains (Fig. 11a). Similar repeated Znfs are conserved in mRNA cleavage and polyadenylation specificity factor 30 (CPSF30) family proteins that are involved in eukaryotic mRNA maturation (Fig. 11a) [64, 65]. CPSF30-like proteins in plants also contain a YTH domain and are orthologous to the Znf-Znf-Znf-YTH-DCD-KH proteins we detected. CPSF30 orthologs in fungi and metazoans only have Znf domains, but YTH domain proteins are integral to the CPSF complexes in vertebrates, where they interact with CPSF6 (Fig. 11a) [66].

YTH also has been reported to bind m^6A in DNA when fused to an McrB homolog in the archaeon *Thermococcus gammatolerans* [20]. Using the sequence of this archaeal YTH domain as a PSI-BLAST query, we detected homologs that, much like DCD, are fused to McrB-like GTPases or PD(D/E)XK nucleases. Most of these YTH-like domains are not clearly distinguishable from EVE domains using HHpred, being modest hits for both types, a pattern that is reminiscent of some prokaryotic DCD domains.

Extended Tudor-DCD fusion proteins in choanoflagellates implicated in the origins of the piRNA pathway

We were unable to identify DCD domains in metazoans. However, when we analyzed the predicted proteins translated from the published transcriptomes of choanoflagellates, the closest unicellular relatives of animals [67, 68], a protein containing a DCD domain fused to an extended Tudor (eTudor) domain was detected in both loricate and non-loricate choanoflagellates, the two main lineages of this phylum (Figs. 11c and 13).

Tudor domains bind post-translationally methylated arginine or lysine residues in eukaryotic proteins [69]. They interact with three main types of modified proteins: histone tails (methylarginine or methyl-lysine), Sm proteins in spliceosomes (methylarginine), and the N-termini of metazoan Piwi-related Argonaute proteins (methylarginine) [69]. The Sm protein-binding Tudor domains present in the splicing factor survival motor neuron (SMN) and related proteins are distinguished from Tudor domains that bind histone tails by an N-terminal α-helix (Fig. 11b) [70]. The Tudor domains that interact with Piwi-related Argonautes are of the eTudor type [59, 69]. The eTudor family is restricted to metazoans, with the exception of Tudor-SN, a highly conserved eukaryotic protein implicated in RNA interference, splicing, microRNA decay, and RNA editing that contains four staphylococcal nuclease (SNase) domains and a single eTudor domain [71–73] (Fig. 11b).

Bioinformatic and structural analyses suggest that the eTudor domain arose when a Tudor domain, related to the Tudor domain in SMN, inserted into the fifth, C-terminal SNase domain of an ancestral multi-SNase protein [59, 70] (Fig. 11b). The resulting domain fusion of Tudor and SNase (hence the name “extended Tudor”) became the ancestor of the eTudor family, in which the catalytic residues from the ancestral SNase domain are mutated, likely rendering it inactive [59, 70, 71, 74] (Fig. 11b). Present in all metazoans, multi-eTudor proteins play crucial roles in the localization of Piwi-related Argonautes and biogenesis of Piwi-interacting RNAs (piRNAs) by interacting with symmetrically dimethylated arginine (SDMA) residues in the Argonaute N-termini, and thus, are essential for repression of transposable elements, modulation of germline mRNA levels, and germ/stem cell immortality [69, 75, 76]. The origins of the complex metazoan multi-eTudor proteins derived from Tudor-SN are fundamental to the understanding of the piRNA pathway and animal germline specification but, currently, remain obscure.

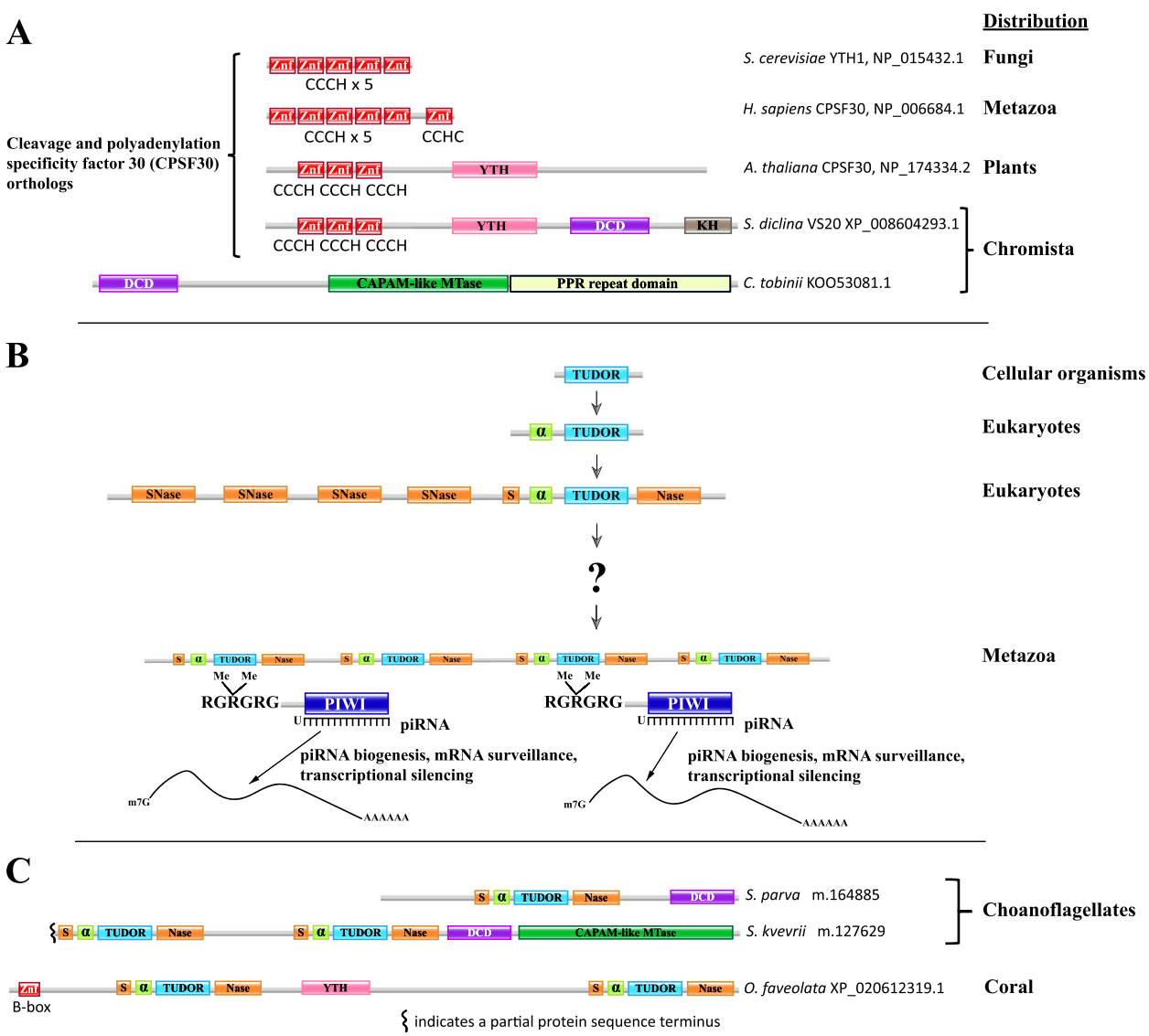

Fig. 11 Evolutionary switch from DCD to eTudor domains. Cleavage and polyadenylation specificity factor 30 (CPSF30) orthologs contain YTH and DCD domains in many heterokont species, and a CAPAM-like putative mRNA methyltransferase containing a DCD domain was detected in a haptophyte species. Multi-eTudor domain proteins evolved from Tudor-SN to produce the family of essential piRNA pathway factors in metazoans, but their ancestry is poorly understood. The eTudor-DCD proteins in choanoflagellates and eTudor-YTH proteins in corals clarify the evolutionary history of the eTudor family. **a** Domain organization of representative orthologs of CPSF30 from budding yeast, human, *Arabidopsis*, and the pathogenic oomycete *S. diclina*, as well as a CAPAM-like putative mRNA methyltransferase from *C. tobinii* which contains a DCD domain and was the only example detected. The RefSeq protein accession numbers are indicated. Zinc finger (Znf) domains are labeled with their type. **b** The core Tudor domain inherited from prokaryotes and present in histone-binding proteins is shown in blue. An N-terminal α-helix conserved in SMN-like proteins and incorporated into the eTudor domain is shown in green. The SNase domain that the SMN-like Tudor domain inserted into is shown in orange, with sections at both termini of the resulting eTudor domain that fold around the inserted Tudor domain to retain the original SNase structure [59]. A multi-eTudor protein is shown interacting with dimethylarginine residues in RGRGRG motifs at the N-termini of two Piwi-related Argonautes. Such motifs are conserved sites of arginine methylation in the Argonautes, as well as Sm proteins bound by the Tudor domains in SMN and Tudor-SN. **c** Domain organization of two choanoflagellate eTudor-DCD proteins and a representative coral eTudor-YTH protein. The species and GenBank or RefSeq protein accession numbers, or identifiers from the predicted protein dataset published by Richter et al. are indicated (Additional file 3: Dataset 2). 5′ partial proteins predicted from transcriptome sequencing are marked at the incomplete terminus. Zinc finger (Znf) domains are labeled with their type

We detected eTudor proteins in choanoflagellates that are not orthologs of Tudor-SN, and these represent the first examples, to our knowledge, to be reported in a non-metazoan organism. These proteins usually also contain a DCD domain, and in some cases, a CAPAM (*cap*-specific *a*denosine *m*ethyltranferase)-like methyltransferase and/or

a second eTudor domain (Figs. 11c and 13). In the process of identifying the CAPAM-like domains, we encountered a misannotation of the Pfam family PCIF1_WW (pfam12237), which, according to our analysis, is not a WW domain, but rather a CAPAM-like methyltransferase. We also observed multi-eTudor proteins fused with a YTH domain in some species of coral, among the earliest branching metazoans (Figs. 11c and 13).

Furthermore, we detected links between the eTudor-DCD/YTH fusion proteins and the ubiquitination pathway and protein degradation. An N-terminal ubiquitin-binding domain (UBA) and B-box Znf domains are present in the choanoflagellate eTudor-DCD and coral eTudor-YTH proteins, respectively (Fig. 13). Similar B-box Znfs are found in TRIM ubiquitin E3 ligases and in the eTudor piRNA pathway factor qin/komo from *Drosophila*, which also contain RING Znfs (Fig. 13) [69, 77]. The eTudor proteins with RING Znfs are conserved throughout the eumetazoa, although in vertebrates, the B-box Znfs appear to have been lost. In the sponge *Amphimedon queenslandica*, a protein with four eTudor domains and an N-terminal MYND-type Znf has been identified, with orthologs present in most metazoans (Fig. 13).

Biochemical functions of EVE-associated proteins

In this section, we present our inferences of the likely biochemical functions of the proteins linked to the EVE domain, both covalently and non-covalently, which we derived from the literature documenting experimental characterization of members of the corresponding protein families. An important caveat is that these inferences, although often direct and likely valid, are inherently less confident than the robust computational results so far described.

EVE in α-proteobacteria

The most prominent contextual association of the EVE domain observed in our study is its inclusion in the putative operon TsaD→GpsA→YciI→EVE in α-proteobacteria. TsaD modifies tRNAs that decode ANN codons (Met, Ile, Thr, Asn, Lys, Arg, Ser) to introduce threonylcarbamoyladenosine (t^6A) at position 37, immediately adjacent to the anticodon [78]. t^6A is a universal modification that is essential for translational fidelity, and mutations in this pathway lead to errors in start codon selection and aberrant frameshifts [78–80]. The α-proteobacterial glycerol-3-phosphate dehydrogenase that is tightly associated with EVE is orthologous to the corresponding mitochondrial enzyme that contributes electrons to the respiratory chain [81]. Thus, the evolutionarily conserved link between the EVE proteins, TsaD and GpsA, implies an unexplored connection between tRNA modification and electron transfer in α-proteobacteria. The YciI protein family has not been thoroughly characterized. One member of this family, TftG from *Burkholderia phenoliruptrix* AC1100, is a dehydrochlorinase requiring a conserved His-Asp dyad for catalysis, a motif that is present in the YciI proteins in the EVE neighborhoods [82]. Fusion of a YciI domain to a σ^{70} factor domain in *Caulobacter vibrioides* and to a BolA transcriptional regulator domain in *Coxiella burnetii* imply that this family may be involved in transcription initiation [83]. Intriguingly, in *E. coli* K-12, the gene encoding TsaD is in a head to head orientation with an operon that encodes σ^{70} factor RpoD, suggesting that a link between tRNA modification and global transcriptional regulation could be ancestral to Proteobacteria. This apparent operon is frequently associated, in a head to head orientation implying possible co-regulation, with a HemC→HemD→IMMP→HemY operon that encodes enzymes of heme biosynthesis. The IMMP ortholog in mitochondria is required for the formation of cristae [84].

EVE in β- and γ-proteobacteria

The cell division proteins strongly associated with EVE in β- and γ-proteobacteria, ZapA and ZapB, interact as a complex with FtsZ, promoting the Z-ring formation during bacterial cytokinesis [85]. Often located between ZapAB and EVE is the enzyme FAU1 (or YgfA), which converts 5-formyltetrahydrofolate (5-formylTHF) to 5, 10-methylenylTHF. 5-formylTHF is a stable storage form of folate that accumulates in dormant cells, such as spores and seeds, whereas 5,10-methylenylTHF is a precursor in purine and methionine biosynthetic pathways [86–88]. Also present in these putative operons, always immediately following ZapA genes, are non-coding 6S RNA (*ssrS*) genes, which express a 184 nucleotide small RNA that functions as a global regulator of transcription in bacteria by binding to the housekeeping σ^{70}-RNA polymerase holoenzyme ($E\sigma^{70}$) [89, 90]. 6S RNAs have been reported to accumulate in *E. coli* cultures during the transition from the exponential to stationary phase of growth, and their effect is to inhibit transcription from most σ^{70}-dependent promoters, which effectively activates the expression of stationary phase-specific genes dependent upon other σ factors, enabling transcriptional adaptation to changing growth conditions [89, 91].

It appears likely that expression of FAU1, in conjunction with ZapAB, SsrS, and the EVE protein, is part of a metabolic switch between proliferative modes. Consistent with this possibility, FAU1 has been implicated in promoting the formation of persister cells and biofilms, and SsrS function is thought to enhance long-term cell survival [92–94]. The SsrS→FAU1 operon, which is broadly conserved in Proteobacteria, including α-

proteobacteria, has been experimentally characterized in *E. coli* K-12, where the dicistronic transcript is processed into mature 6S RNA [95, 96]. Our observations suggest that FAU1 genes are not strictly necessary in these regions and that *ssrS* genes are often flanked, in β- and γ-proteobacteria, but not in α-proteobacteria, by ZapA and EVE genes, which may, like FAU1, be expressed in polycistronic transcripts containing the 6S RNA precursor. No modifications of 6S RNA have been reported, although the consistent, close association with EVE suggests that the 6S RNA might contain modified bases recognized by EVE domains (Fig. 5).

Given its conserved, head to head juxtaposition in γ-proteobacteria with the apparent ZapAB→SsrS→(FAU1)→EVE operon, the operon YgfB→PepP→UbiH→UbiI that has been experimentally characterized in *E. coli* K-12 [97], could be co-expressed and might play a role in cell cycle regulation as well. The PepP ortholog encoded by *Pseudomonas aeruginosa* in this conserved, EVE-containing context has been identified as a critical virulence factor in a *Caenorhabditis elegans* infection model [98]. Homologs of ribose-5 phosphate isomerase (RpiA) and L-threonine dehydratase (IlvA) that are often encoded in these regions likely also participate in the implied, large-scale proliferative regulation.

Many species of Alteromonadales, while lacking the link between EVE and ZapAB→SsrS, encode a cobalamin-dependent radical SAM enzyme in close association with EVE, which might be related to the cobalamin biosynthetic clusters adjacent to ZapAB→SsrS→EVE in β-proteobacteria (Fig. 5, Additional file 1: Fig. S3). Pseudoalteromonadaceae, also lacking ZapAB→SsrS→EVE, encode EVE domains in conserved associations with translation factors and respiration-related enzymes involved in the maturation of cytochrome c and ubiquinone, suggesting that coupling between translation and respiration mediated by EVE domains extends throughout the Proteobacteria and could be ancestral to this phylum (Fig. 12). In these neighborhoods, EVE is tightly linked to factors homologous to acyl-CoA thioesterase TesB and glycerol-3-phosphate O-acyltransferase PlsB, suggesting that the abundance of glycerol-3-phosphate, which can contribute electrons to the respiratory chain via its dehydrogenase [81], as seen in the α-proteobacterial EVE neighborhoods, is modulated by these EVE-associated enzymes.

Our comparative genomic analyses shed light on deeply conserved apparent functions of EVE proteins in α-proteobacteria, where they likely link modulation of cytochrome c maturation with tRNA modification and transcriptional regulation, and in β- and γ-proteobacteria, where they are predominantly implicated in the linkage of cell division, transcriptional, and metabolic regulatory mechanisms, but in some members of these classes, are closely associated with translation and electron transport factors as in α-proteobacteria. As noted above, in *Coxiella burnetii*, a YciI-like protein is fused at the C-terminus to a BolA domain, a transcriptional regulator involved in promoting biofilm formation and repressing motility [32, 83]. Morphological effects of BolA overexpression depend on FtsZ, the interaction partner of ZapAB [99]. This finding suggests that YciI proteins closely tied to EVE in α-proteobacteria might perform a role similar to the YciI homolog fused to BolA, which is likely to involve transcriptional regulation of large-scale biochemical and morphological adaptations to changing conditions. A similar function is conceivably carried out by the ZapAB→SsrS→(FAU1)→EVE regions in β- and γ-proteobacteria, which likely participate in cellular phase shifts between exponential vs. stationary and planktonic vs. biofilm proliferative modes.

EVE as a specificity domain in modification-dependent restriction systems

The most common function of EVE domains in this capacity likely entails flipping out a modified cytosine derivative from a DNA helix for scrutiny in the EVE's binding pocket and targeting endonuclease activity to the neighboring DNA, given sufficient affinity for the modified sequence. This role can be inferred from the comparison with the SRA domain, which shares the PUA-like fold with EVE. The SRA domain has been characterized in considerable detail, including the base-flipping 5mC DNA-binding mechanism and characterization of its function as a modified DNA specificity module in type IV REs which restrict DNA containing 5mC, 5hmC, and glucosylated 5hmC [18, 100–103]. Therefore, EVE domains deployed in modification-dependent restriction likely use a base-flipping mechanism, similar to that of SRA, to sense modified cytosine in various sequence contexts, although some might bind derivatives of the other pyrimidine bases, thymine or uracil, which are hypermodified in some phage genomes [103]. Yet other EVE domains might preferentially bind modified adenine, given that EVE is also structurally similar to the YTH domain, which binds m^6A in DNA and RNA [10, 20].

We detected a remarkable variety of combinations of EVE and restriction endonuclease domains, implying intense pressure to evolve diverse restriction strategies to provide immunity from a vast and highly varied population of viruses with modified genomes. Some of these modifications bound by EVE domains could be effective in inhibiting defense by CRISPR-Cas adaptive immunity systems, in addition to type I, II, and III REs, as has been reported for glucosylated 5hmC [103, 104]. The EVE domains appear to play a major role in meeting the demand for defenses tailored to this threat.

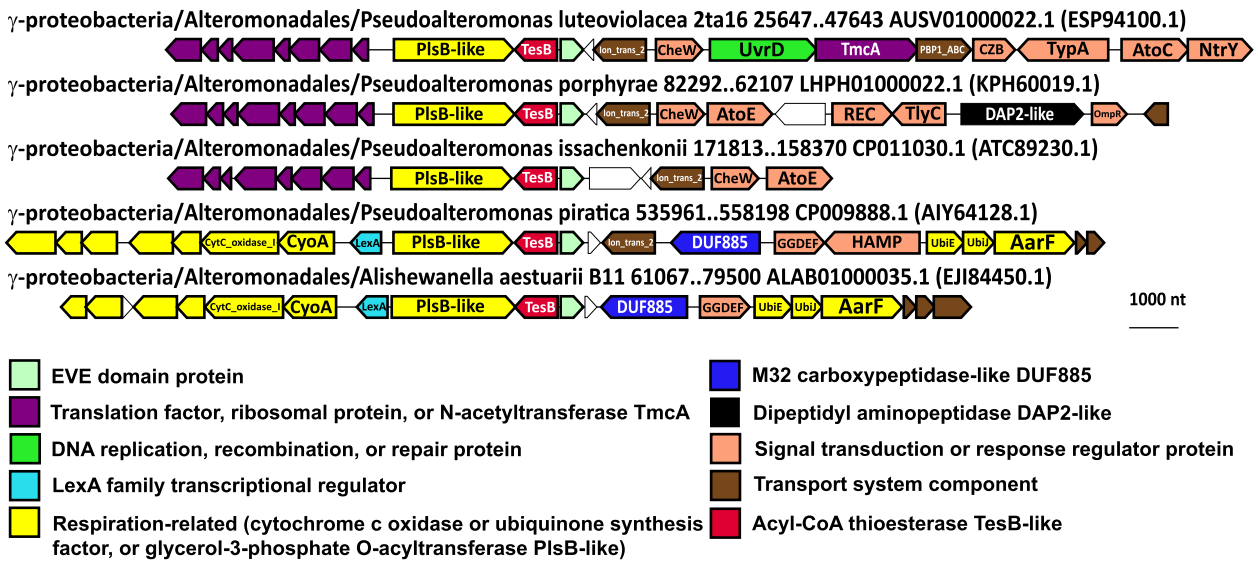

Fig. 12 Conserved genomic context of EVE proteins in Pseudoalteromonadaceae. Representative EVE protein neighborhoods from Pseudoalteromonadaceae. Genes are shown as arrows from 5′ to 3′. The taxonomic lineage and genomic coordinates for each neighborhood are indicated, as are the GenBank genome accessions and, in parentheses, the GenBank accessions for each EVE protein

EVE domains in toxin-antitoxin systems

None of the modules (PIN, GNAT, HTH, EVE, ASCH) in these systems have been experimentally characterized, but the predicted toxin or antitoxin activity of the associated domains strongly suggests that, in this case, the EVE proteins are components of toxin-antitoxin (TA) systems. The PIN domain RNases function as toxins in a broad variety of bacterial, and especially, archaeal TA systems [47, 48]. The GNAT domains also typically function as toxins [49], and HTH domain-containing antitoxins have been described as well [105]. The mechanistic details of these (PIN)-GNAT-EVE protein activities await experimental investigation, but some functional hints emerged from our analysis. The differential distribution of PIN-GNAT-EVE and GNAT-EVE proteins (Fig. 8) is a potentially important clue. Furthermore, in *Methanohalophilus* genomes, the associated ASCH domain is fused to the specificity subunit (HsdS) of a type I RM system, implying that it confers modification specificity to the RM complex (Fig. 8). Combined with the frequent occurrence of HTH-ASCH fusions, these associations suggest that ASCH domains in the (PIN)-GNAT-EVE operons bind modified DNA.

The uncharacterized, putative defense system we detected, COG1743→DUF499→SWI2/SNF2-nuclease→EVE, and the type I RM systems associated with standalone EVE and ASCH proteins, arguably represent TA systems as well (Fig. 10, Additional file 1: Fig. S8). We predict that DUF499 proteins in the former system might interfere with the replication of foreign DNA containing modified bases, which is discriminated by the EVE domain and restricted by the SWI2/SNF2 helicase-nuclease fusion protein, activities that are potentially toxic. The likely role of the methyltransferase is to prevent restriction of the host genome, and thus to serve as an antitoxin, by methylating a base in the sequence recognized by the SWI2/SNF2-nuclease, a modification not recognized by the EVE domain. The presence of this factor also implies that the system can restrict unmodified DNA without requiring recognition by the associated EVE domain.

Similarly, type I RM systems that are associated with EVE domains are likely to target modified DNA, and the presence of the type I methyltransferase (HsdM) suggests restriction of both unmodified and modified DNA can occur (Fig. 10). The methyltransferase in these systems can be predicted to generate modified bases that are not recognized by the associated EVE domain and prevent restriction by the type I endonuclease subunit (HsdR), which is also capable of restricting modified DNA that is discriminated by the EVE domain.

MmcQ/YjbR-EVE fusion proteins

The homology between MmcQ/YjbR and the mitochondrial iron homeostasis protein CyaY suggests that MmcQ/YjbR could be an iron-binding protein as well [106] although the functional residues and electrostatic potential are not conserved between these domains [107]. A more convincing functional prediction for MmcQ/YjbR has been made based on structural and electrostatic surface similarity to the C-terminus of T4 bacteriophage transcription factor MotA, known as MotCF. Although there is only a limited sequence similarity between with MmcQ/YjbR and MotCF, conserved

residues are concentrated in the putative DNA-binding region of MotCF, strongly suggesting that, like MotCF, MmcQ/YjbR interacts with DNA [107]. Furthermore, multiple MmcQ/YjbR homologs have been shown to adopt a "double wing" DNA-binding fold similar to MotCF [107, 108]. In our dataset, one example of a GIY-YIG nuclease-MmcQ/YjbR-EVE fusion and another of a PLDc nuclease-Helicase-MmcQ/YjbR-EVE fusion are present, implicating the EVE domain as a modified DNA base specificity module in these proteins, whereas MmcQ/YjbR might contribute sequence specificity.

DUF1831, often fused at the N-terminus to MmcQ/YjbR-EVE proteins, shows remote structural similarity to TBP-like (TATA-binding) fold proteins, which include S-adenosyl-methionine decarboxylase [109]. Analysis of the genomic neighborhood context of DUF1831 genes supports a role in metabolism of amino acids, particularly, methionine [109]. The DUF1831-MmcQ/YjbR-EVE proteins can be encoded in a putative operon with peptide methionine sulfoxide reductase MsrAB and disulfide interchange factors which likely recycle it. In general, however, the function of these complex EVE proteins is likely to be multifaceted (Fig. 9).

Extended Tudor-DCD fusion proteins in choanoflagellates implicated in the origins of the piRNA pathway

The identification of eTudor-DCD fusion proteins encoded in the transcriptomes of loricate and non-loricate choanoflagellates implies that the common ancestor of the extant choanoflagellates as well as the metazoa, that apparently descended from colonies of non-loricate choanoflagellate-like ancestors [67, 110], already encoded the eTudor-DCD protein. Choanoflagellates, most likely, acquired the genes encoding DCD proteins by horizontal gene transfer (HGT) following the loss of DCD in the ancestor of the opisthokonts. This route of evolution is strongly suggested by the observed absence of DCD from all opisthokonts, other than the choanoflagellates, although the possibility of inheritance from an early eukaryotic ancestor cannot be completely ruled out [56]. Consistent with this scenario, the extant choanoflagellates are thought to have acquired substantial portions of their genomes via HGT, including from algae [111, 112]. In both haptophyte algae (chromists) and non-loricate choanoflagellates, the DCD domain is fused to a CAPAM-like domain, and in the latter case, also to an eTudor domain (Fig. 13) [113]. Furthermore, as noted above, DCD is present in conserved CPSF30-like factors in many heterokont species (also chromists). Therefore, we infer that chromist algae are the putative source of DCD in choanoflagellates that might have acquired it via HGT.

In heterokont DCD proteins that are orthologs of CPSF30 and so can be predicted to participate in mRNA maturation, the DCD domain likely binds modified RNA, either exclusively or in addition to binding modified DNA. Its primary target could be m^5C in mRNA, given the affinity of EVE domains for modified cytosine, and the presence, in the same proteins, of a YTH domain, which has consistently been reported to bind m^6A in eukaryotic mRNA [19, 60–63, 66, 114]. It is probable that the DCD and YTH domains in these proteins recognize distinct modifications. It is this type of DCD domain, which is likely to bind modified RNA, that can be predicted to have fused to eTudor in choanoflagellates.

The nature of the ligand of the choanoflagellate DCD domains is suggested by the other domains to which they are covalently linked. The CAPAM-like methyltransferase fused to DCD in *S. kvevrii* (a choanoflagellate) and *C. tobinii* (a haptophye alga) is homologous to the RNA methyltransferase in the CAPAM protein which methylates m^6A in vertebrate mRNAs (Fig. 13) [113]. In the DCD-CAPAM-like methyltransferase fusion proteins, DCD occupies the same N-terminal position as a helical domain that, in CAPAM, is involved in the recognition of the m^7G mRNA cap and directing methylation to m^6A [113]. Therefore, the DCD domain in these proteins might be involved in targeting additional modifications of modified mRNAs, perhaps, containing m^5C. Furthermore, the presence of N-terminal RanBP2-type Znfs in the eTudor-DCD protein we detected in *A. spectabilis* also implies RNA binding, as well as participation in splicing and/or nuclear export (Fig. 13) [115, 116]. The eTudor domain itself, in the absence of Piwi-related Argonaute proteins, is likely to bind SDMA residues in spliceosomal Sm proteins during mRNA maturation, as shown for the homologous domains in the Tudor-SN and SMN proteins [72, 117].

Similar to the choanoflagellate eTudor-DCD protein we identified in *A. spectabilis*, which contains an N-terminal UBA domain, metazoan eTudor proteins often contain N-terminal Znfs implicated in ubiquitination and protein degradation, and one well-conserved type possesses N-terminal MYND-type Znfs (Fig. 13). The MYND-type Znf in the *Aedes aegypti* eTudor protein Veneno, which contains two eTudor domains, is required for the localization of Veneno to putative piRNA processing germ granules [118]. The consistent presence of N-terminal Znf domains, in many metazoan eTudor proteins and one loricate choanoflagellate eTudor protein, led us to surmise that the incomplete N-termini in the partial eTudor protein sequences we identified in non-loricate choanoflagellates likely harbor a type of N-terminal Znf as well, which has not yet been observed (Figs. 13 and 14).

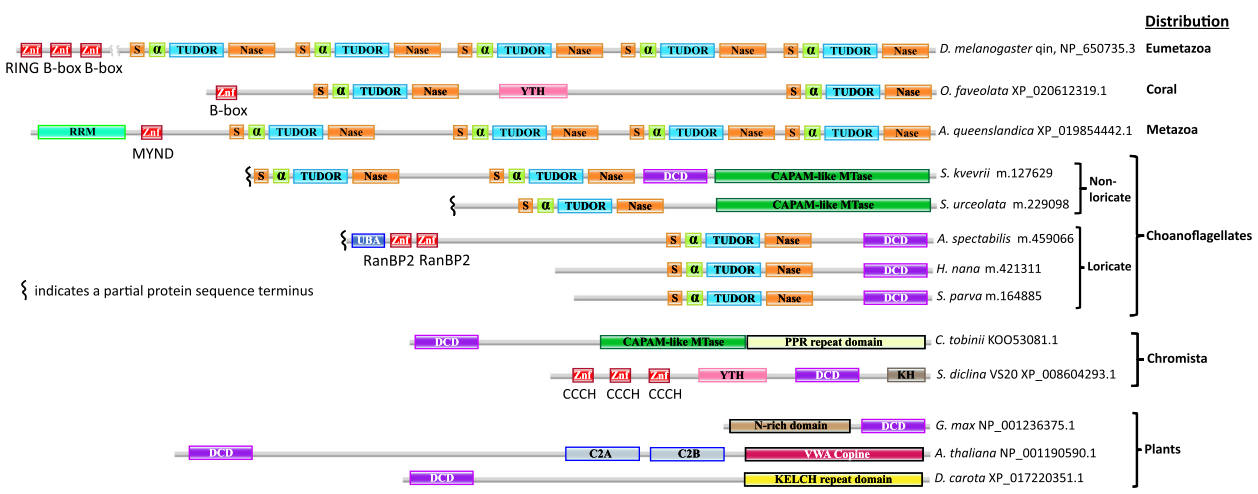

Fig. 13 Representatives of major classes and informative examples of DCD and eTudor proteins in plants, chromists, choanoflagellates, and metazoans. Domain organization of representatives of major classes of DCD proteins in plants and heterokonts, and major classes of coral and eumetazoan eTudor proteins. Especially informative individual examples from haptophyte algae and choanoflagellates are also shown. The species and GenBank or RefSeq protein accession numbers or identifiers from the predicted protein dataset published by Richter et al. are indicated. 5′ partial proteins predicted from transcriptome sequencing are marked at the incomplete terminus. Zinc finger (Znf) domains are labeled with their type. The phyletic distribution of each protein class is denoted as well. Choanoflagellate predicted protein sequences that are not deposited as separate entries in GenBank are available as Additional file 3: Dataset 2

Discussion

The comprehensive analysis of the genomic neighborhoods of prokaryotic EVE proteins described here has a variety of functional and evolutionary implications.

Implications for the evolution of PCD from proteobacterial and defense-related EVEs

PCD in eukaryotes reportedly involves THYN1-like EVE domains, which are broadly distributed and show high sequence similarity to proteobacterial EVEs. The conservation of the EVE genomic context in α-proteobacteria and Pseudoalteromonadaceae suggests the possibility that eukaryotic PCD evolution exploited a proteobacterial mechanism that couples modulation of energy production with translation via cytosine methylation in tRNAs, along with the EVE proteins that recognize these modifications. The sequence of events in the intrinsic PCD pathway in animals is centered around mitochondria that integrate signals of stress or damage and, in response, release proteins from the intermembrane space into the cytosol to initiate PCD [119–122]. Foremost among these proteins is the heme-containing cytochrome c, an essential component of the respiratory electron transport chain [34, 121–123]. Cytosolic cytochrome c binds to apoptotic protease activating factor 1 (Apaf-1), which then recruits pro-caspase-9 to assemble a multi-subunit complex, the apoptosome, starting a complex cascade of proteolytic caspase activity that results in massive protein degradation, internucleosomal DNA cleavage, and global mRNA decay [124, 125]. Intriguingly, roles for tRNA and stress-induced, tRNA-derived tiRNAs in the intrinsic PCD pathway have recently come to light. Multiple studies have demonstrated an interaction between tRNA/tiRNAs and cytochrome c in mammalian cells that inhibits the formation of the apoptosome and promotes cell survival [120, 126]. In the case of tiRNAs, which are generated from tRNA cleavage near the anticodon, modifications of the tRNA, such as 5-methylcytosine (m^5C), which might be recognized by an EVE domain, have been reported to negatively regulate their biogenesis [127]. Furthermore, mitochondrial IMMP, which is orthologous to the protein closely associated with the EVE domain in α-proteobacteria, has been implicated in eukaryotic PCD [128]. In addition, the heme biosynthesis enzymes encoded in the same neighborhoods with EVE proteins in α-proteobacteria are involved in the maturation of cytochrome c, which requires heme as a cofactor, and therefore, are linked to one of the central effectors of eukaryotic PCD [34].

Moreover, the neighborhoods of EVE proteins in β- and γ-proteobacteria implicate the EVE domain in deeply conserved coordination between proteins that promote cytokinesis (ZapAB), a small RNA that promotes transcriptional adaptation to growth conditions (SsrS, which may be modified and bound by an EVE domain), and a metabolic enzyme (FAU1) involved in persister cell and biofilm formation under environmental stress. In γ-proteobacteria, an associated operon encoding a virulence-related aminopeptidase (PepP) and respiration-related factors of ubiquinone biosynthesis (UbiHI) is also likely to contribute to the overall function of these conserved regions. We cannot yet

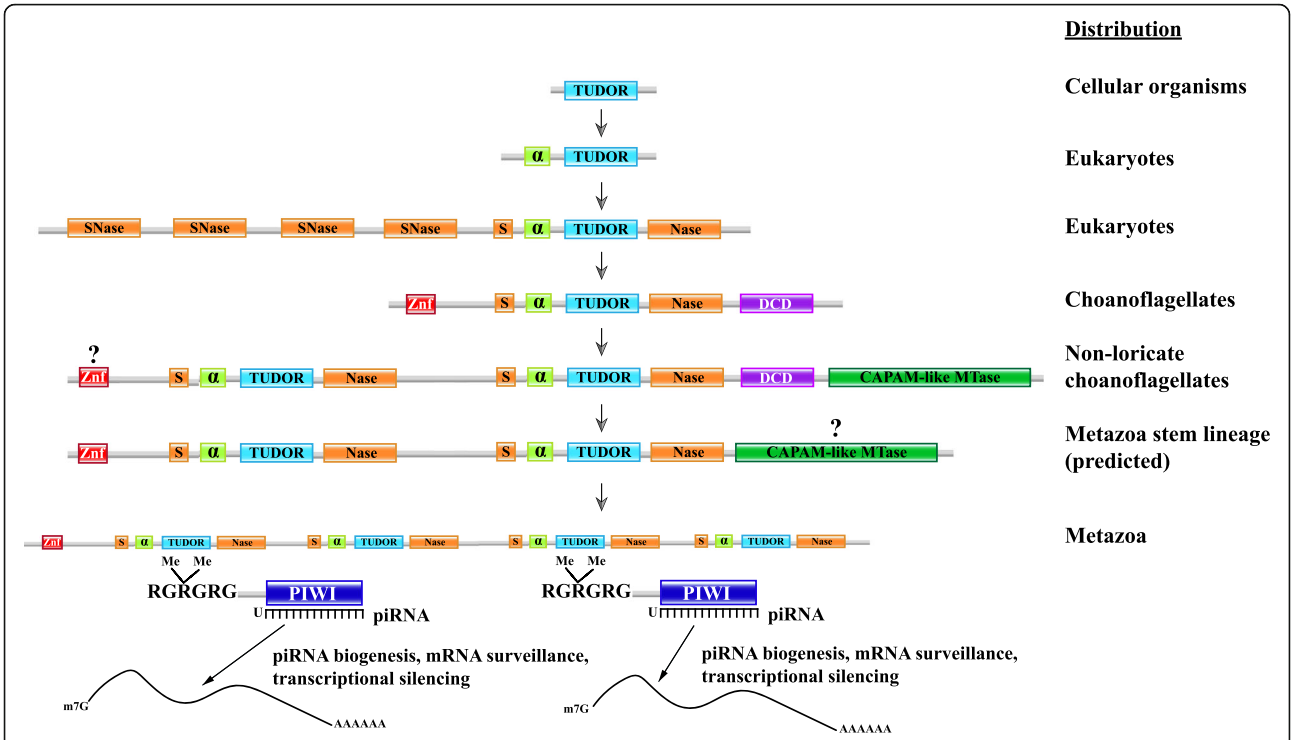

Fig. 14 A hypothetical evolutionary scenario for the metazoan eTudor domain proteins essential to the piRNA pathway. The predicted evolution of Tudor protein function and domain composition is depicted. In the first stage, the core Tudor domain is involved in recognition of methylarginine in histone tails, with an unknown prokaryotic role. Next, this domain is combined with an N-terminal α-helix in an ancestral SMN-like splicing factor in the eukaryotic stem lineage, binding SDMA in spliceosomal Sm proteins. Subsequently, but also during the eukaryotic stem phase, this type of Tudor domain was inserted into an SNase domain in the ancestor of the Tudor-SN-like family, forming the eTudor domain. The Tudor-SN protein class became involved in RNAi and splicing, primarily binding SDMA in spliceosomal Sm proteins and potentially in other proteins as well. During the diversification of eukaryotes, in ancestral choanoflagellates, the eTudor domain was fused to the DCD domain, forming a putative mRNA maturation factor binding SDMA in spliceosomal Sm proteins with eTudor and m^5C RNA with DCD. This type of protein may play a role in mRNA nuclear export, self- vs. non-self-discrimination, and protein degradation, with N-terminal Znf domains that might bind ubiquitin or RNA. At the next step, after the divergence of the loricate and non-loricate choanoflagellates, duplication of the eTudor domain in the non-loricate choanoflagellates allowed elaboration of the function of the eTudor-DCD protein, freeing one eTudor domain from the conserved function of preferential Sm protein binding and allowing it to interact preferentially with other proteins containing SDMA, eventually including Piwi-related Argonautes, but perhaps not yet at this stage. The involvement of the CAPAM-like methyltransferase domain is not necessarily part of the piRNA pathway origins, but the possibility is suggested by the data. In the final link to the extant piRNA systems, we propose a hypothetical factor, potentially still binding methylated Sm proteins, that evolved to target mRNA via interaction with piRNA/Argonaute complexes rather than binding m^5C. These complexes now discriminate spliced and modified self from non-self/aberrant mRNA, regulating nuclear export, as well as transcription via Argonaute-directed chromatin modification, a conserved feature of eukaryotic cells also extensively performed by plants. N-terminal Znfs contribute to eTudor localization to perinuclear germ granules and possibly link them to the ubiquitination and protein degradation system to provide even more control over expression

determine the ancestry of the eukaryotic EVE domains, but these roles of EVE proteins in bacteria tying together energetic, transcriptional, and translational responses could presage the involvement of this domain in eukaryotic PCD.

Defense-related EVE domains, which do not generally fall into the two largest clusters from our CLANS analysis, nevertheless, are likely to be involved in PCD. Modification-dependent restriction systems are generally toxic to cells which express enzymes that catalyze the formation of modified bases they recognize and, thus, are implicated in a form of prokaryotic PCD [103]. Therefore, EVE domains associated with nucleases are potentially involved in both innate immunity and PCD, in cases when a cognate methyltransferase is expressed in the same cell. The connection between TA systems containing EVE domains and its role in prokaryotic PCD is readily apparent. The available evidence concerning MmcQ/YjbR-EVE fusion proteins suggests coordination of environmental sensing and response, metabolism, and defense/PCD that is modulated by that class, which represents yet another way in which EVE domains participate in the complex chains of events involved in cell fate decisions. The DCD domain, a defense-related EVE-like domain, likely shares the innate immunity/PCD role of EVE in prokaryotes, whereas in eukaryotes, it has taken

on more complex functions that could involve modified RNA binding and has resulted in clear involvement of this domain in PCD in plants, as well as a likely role in mRNA maturation in chromists and choanoflagellates, which was apparently important during the evolution of the eTudor proteins and the piRNA pathway in metazoans.

Involvement of (PIN)-GNAT-EVE proteins in virus-host conflicts

In light of the observations described above concerning (PIN)-GNAT-EVE proteins, targeting GNAT to modified DNA or RNA via EVE, conceivably, protects phages against host RM systems, perhaps, via toxicity to the host, whereas the addition of PIN is associated with host defense and likely counteracts the effect of GNAT-EVE. The GNAT-EVE proteins potentially target aminoacyl-tRNAs (aa-tRNAs) that harbor modifications recognized by EVE, given that GNAT toxins have been reported to acetylate aa-tRNAs [49, 129]. Under this scenario, PIN-GNAT-EVE proteins would likely degrade toxic, acetylated aa-tRNAs generated by GNAT-EVE. The role of the putative DNA-binding ASCH domain that is nearly always present in (PIN)-GNAT-EVE operons in this process remains unclear, but it might be involved in regulating the expression of the (PIN)-GNAT-EVE protein. Other accessory proteins, such as AAA_17 family ATPases that are frequently, but not invariably, encoded near these factors can be expected to contribute in non-essential, regulatory capacities. The interplay between phage and host cell proteins with GNAT-EVE architectures appears to be a widespread phenomenon that clearly warrants further inquiry.

Extended Tudor-DCD and the origins of the piRNA pathway

We predicted that DCD proteins in choanoflagellates and chromists bind modified RNA, or less likely, also DNA. It is unclear whether DCD domain proteins in plants, which have been the subject of considerable inquiry, bind modified DNA, RNA, or both, but investigation of their affinities toward modified nucleic acids can be expected to elucidate their roles in plant PCD. Conceivably, given the distributions we observed and the domains fused to DCD and YTH in various phyla, these domains originate from fast-evolving EVE domains involved in restriction of modified DNA that were recruited, early in eukaryogenesis, to recognize modified eukaryotic mRNA.

Mechanistically, the eTudor-DCD proteins in choanoflagellates could be involved in mRNA maturation, with the DCD domain potentially recognizing m^5C in mRNA, whereas the eTudor domain interacts with methylated Sm proteins, and might, in choanoflagellate species yet to be identified that encode an Argonaute protein(s) with N-terminal SDMA residues, interact with those as well. Consistent with this possibility, the m^5C modification of mRNA appears to promote export from the nucleus [130], implicating eTudor-DCD proteins in trafficking of spliced, mature mRNA into the cytoplasm.

This putative function could be an evolutionary foundation for the piRNA pathway, given that the eTudor proteins involved in this pathway generally localize to perinuclear germ granules, which are associated with clusters of nuclear pore complexes (NPCs), and have been proposed to be extensions of the nuclear pore environment [131–133]. These granules are foci of RNA and protein accumulation that appear to determine which mRNAs are permitted to enter the germ cell cytoplasm for translation, primarily, via silencing of unlicensed transcripts by Piwi-related Argonautes, and so are final arbiters of nuclear export [130–134]. Thus, choanoflagellate eTudor-DCD proteins, that can be predicted to interact with m^5C and methylated Sm proteins during mRNA maturation and exit from the nucleus, could have been fundamental contributors to the function of the perinuclear granules that ultimately arose in animal germ cells and which also govern nuclear mRNA export by associating with NPCs [131–133]. Given the role of piRNA pathway eTudor proteins in self- vs. non-self-discrimination, it is possible that eTudor-DCD proteins already play a role in this process in choanoflagellates, ensuring that self mRNAs are correctly spliced, modified, and licensed to exit the nucleus.

The "sudden" appearance of a sizeable group of eTudor proteins and a fully fledged piRNA pathway in the basal metazoans, the sponges, is now illuminated by the identification of a potential transitional form in choanoflagellates [135, 136]. We propose that the choanoflagellate eTudor-DCD protein was an evolutionary bridge from Tudor-SN to the multi-eTudor proteins in early metazoans that initiated the first piRNA pathway (Fig. 14) [70, 137]. The multi-eTudor proteins fused with a YTH domain we detected in several coral species provide additional support for this hypothesis because YTH and DCD are fused to each other in many species of chromist algae, the presumed donor of DCD to choanoflagellates. YTH domain proteins are known to localize to stress granules, which have considerable similarity to germ granules, where they concentrate mRNAs with the m^6A modification to promote their translation [114, 134]. YTH proteins additionally localize to nuclear speckles, also known as interchromatin granule clusters, that are enriched in mRNA maturation factors, where they facilitate processing and nuclear export of m^6A RNA [138]. It is plausible that DCD domain proteins would similarly promote accumulation of modified mRNAs and did so at the dawn of the piRNA pathway in the first metazoans. Such ancestral

activity might underlie the germ granule localization of eTudor proteins and Piwi-related Argonautes complexed with piRNAs, which effectively concentrate RNA, via base pairing between mRNA and piRNA, as well as through binding of SDMA in the Argonaute N-termini by multiple, fused eTudor domains. The activity of Piwi-related Argonaute and eTudor piRNA pathway factors has not been reported to involve mRNA modifications that might once have been recognized by DCD domains, or additional modifications by CAPAM-like methyltransferases, although that possibility merits further inquiry (Fig. 14). The metazoan CAPAM-like methyltransferases could have originated from an ancestor present in a choanoflagellate eTudor-DCD protein, and mRNA modifications might still play a role in the piRNA pathway.

The predicted eTudor proteins from non-loricate choanoflagellates we analyzed are incomplete at their N-termini (Fig. 13), and comparison with full-length eTudor proteins from metazoans led us to speculate that the full-length choanoflagellate proteins (that still remain to be identified) contain N-terminal Znfs. Furthermore, we hypothesize that an evolutionary intermediate eTudor protein existed in the first metazoans which contained multiple eTudor domains and an N-terminal Znf, but lacked the DCD domain, its RNA-binding function replaced by interaction with Piwi-related Argonautes and their associated small RNAs (Fig. 14). Although the variety of the Znf in this founding member is difficult to determine with the available data, a protein with this basic architecture could have served as a foundation for the evolution of the piRNA pathway, ultimately, derived from an ancestral choanoflagellate eTudor-DCD protein. Our survey of metazoan eTudor proteins also revealed a divergent evolutionary trajectory that apparently occurred in nematodes, which lack eTudor proteins with N-terminal Znfs, or with more than two copies of eTudor. A fundamental contributing factor to this outcome was the deletion of the conserved α-helix in one of the eTudor domains in the ancestor of the tandem eTudor proteins that are essential components of the nematode-specific RNA-dependent RNA polymerase (RdRP) complexes (Additional file 1: Fig. S10).

Conclusions

Our comprehensive search for EVE domains demonstrated their wide presence in bacteria and archaea, as well as most eukaryotes. Mechanistically, the common denominator of the EVE family is the binding of methylated bases in DNA and RNA that turns out to be important in a broad variety of functional contexts. The (predicted) biological roles of EVE-like domains are many and, depending on the context, might seem to be in opposition. However, the overarching theme is involvement in immunity, self- vs non-self-discrimination, and stress response/PCD, whether that be through targeting of modified DNA for restriction in diverse prokaryotes, regulation of the proliferation-PCD-dormancy balance in Proteobacteria and eukaryotes (especially during the differentiation of immune cell populations in vertebrates), or export and translation of spliced and modified mRNAs in choanoflagellates.

The linkage between EVE proteins, tRNA modification, and cytochrome c maturation that is suggested by the conserved operonic organization in α-proteobacteria could shed light on the mechanism of the reported anti-apoptotic action of eukaryotic THYN1/Thy28-like EVE proteins and, beyond that, the origin of cytochrome c involvement in eukaryotic PCD. We hypothesize that the proteins encoded in the α-proteobacterial TsaD→GpsA→YciI→EVE operon promote translation coupled with respiration, and downregulation of this operon could induce dormancy. It is unclear if cytochrome c efflux occurs in free-living α-proteobacteria undergoing dormancy or PCD as it does in mitochondria, or this phenomenon evolved during eukaryogenesis, but it would likely be an effective mechanism for rapidly decreasing ATP production in the event of a runaway viral infection.

The EVE domains split into two functionally distinct classes that evolve under different regimes, slowly in the case of those that seem to be involved in basic cellular functions in Proteobacteria and eukaryotes, and fast, in the case of those involved in defense functions and virus-host arms races in diverse bacteria and archaea. The incorporation of EVE domains into numerous RM and TA modules is a remarkable, previously unnoticed pattern in microbial evolution, which emphasizes the various, still under-appreciated, roles that different base modifications play in the intricate virus-host interactions.

The DCD domains present another facet of the PUA/EVE story. The genomic context in prokaryotes implies that they perform functions similar to those of the second class of EVE domains, namely, are involved in antivirus defense via recognition of modified bases in DNA. However, in eukaryotes, the DCD domains become part of a more complicated evolutionary scenario that seems to involve an important aspect of the origins of plants and animals. Plants have retained the DCD domain in multiple proteins that contribute to PCD during plant development as well as stress and pathogen response. In animals, DCD domains apparently have been lost. However, we identified a “smoking gun” in choanoflagellates, the unicellular direct ancestors of animals, where the DCD domains are fused to eTudor proteins. The roles played by eTudor proteins in germline immortality, gametogenesis, and early embryonic development in animals, where regulation of PCD is pivotal [139], are

intriguingly similar to those of DCD proteins in plants. It appears likely, therefore, that DCD domains were important at the earliest stages of the evolution of multicellularity in both plant and animal ancestors, but then, were supplanted by the expanding eTudor family in the animal lineage.

The extent to which the regulation of PCD might still be relevant to eTudor and piRNA function, despite the loss of the DCD domain, remains to be elucidated, but there is evidence suggesting that it could be considerable. In the light of the putative evolutionary connection between the ancient form of PCD and the evolution of the eTudors and the piRNA network, it seems more explicable that conserved piRNA biogenesis factors Tudor-KH and MitoPLD/Zucchini are mitochondrially localized, perinuclear piRNA processing germ granules are also closely associated with mitochondria in addition to nuclear pores, and that reported phenotypes of many piRNA factor mutants include induction of PCD and/or germline mortality [75, 140–144]. PCD is an important part of tissue differentiation, and piRNAs can prevent ectopic expression of somatic genes that might contribute to PCD or senescence of the germline [145, 146]. In animals, populations of piRNAs, with their biogenesis orchestrated by eTudor proteins, Piwi-related Argonautes, and other factors, could have taken the place of the DCD domain in regulating PCD and transposable element mobilization by licensing germline mRNA translation. Whereas in ancestral eukaryotes, m^5C modifications of RNA possibly regulated nuclear export in the absence of a piRNA pathway, when they began to evolve a germline and differentiation into tissues from an embryo in the metazoan lineage, Argonaute-piRNA complexes may have taken over aspects of this screening process, ultimately linking it to Argonaute-directed repressive chromatin modification, a conserved feature of eukaryotic cells, in order to regulate transcription as well as trafficking from the nucleus [147–149]. The consistent presence of ubiquitin-associated Znf domains in eTudor factors suggests an additional role in protein degradation, interconnecting multiple layers of control over expression.

The DCD and YTH domains have strikingly similar evolutionary histories and functional associations. Both are found only rarely in prokaryotes, mostly in archaea, where they are associated with restriction of modified DNA. PSI-BLAST searches for both DCD and YTH domains in prokaryotes readily recover restriction-associated EVE domain proteins, implying that they are both essentially varieties of the much more numerous EVE domains, produced from the diversification driven by virus-host conflict. Subsequently, it would appear, during eukaryogenesis, both were plucked from relative obscurity and conscripted into conserved roles in eukaryotes where they are (possibly, in the case of DCD) involved with concentration and processing of modified RNA, especially during stress response/PCD. The piRNA system, born of the partnership of eTudor proteins and Piwi-related Argonautes in the first animals and showing parallels with DCD function in plants, is conceivably related to this ancient RNA concentration and processing mechanism and, possibly, still involves RNA modifications. Consequently, characterization of the roles of choanoflagellate eTudor-DCD proteins as well as metazoan eTudor proteins in PCD and stress response could be important for understanding the origins of animal germline specification. Moreover, the study of restriction systems with EVE-like domains is likely to shed light on the origins of eukaryotic mRNA regulation.

Taken together, the findings reported here suggest multiple connections between PCD, antivirus defense, and various forms of stress response via diverse families of EVE-like domains that recognize modified bases in DNA and RNA. The role of such bases in the coordination of antivirus defense, PCD, cell proliferation, and development remains under-appreciated, perhaps, substantially. These observations open up many experimental directions that can be expected to advance the understanding of the complexity of all these processes.

Materials and methods

Identification and phylogenetic analysis of EVE and DCD domain proteins

A comprehensive search for EVE proteins was initiated with the available multiple sequence alignment pfam01878. This alignment was clustered and each sub-alignment used to produce a position-specific scoring matrix (PSSM) for use as a PSI-BLAST query against the non-redundant (nr) NCBI database (*E*-value ≤ 10) [150]. Manual filtering of these results aided by BLAST and HHpred [151] was followed by extraction of the EVE domains from each protein and similarity clustering to a threshold of 0.85 with MMeqs2 [152]. Selection of one representative per cluster yielded 8403 sequences for CLANS analysis [31] and phylogenetic tree construction. The metrics of within-group similarity in Proteobacteria were calculated using a tree built with the FastTree program [153] from an alignment of these sequences made using PROMALS3D [154]. This tree was rooted at the midpoint, after which all root-to-tip distances were calculated, giving a median tree height of 2.82 with and interquartile range of 2.27–4.28. The subtrees for α-proteobacteria and β/γ-proteobacteria were extracted and the same values were calculated, yielding heights of 0.99 [0.85–1.18] and 0.9 [0.66–1.03], respectively, which represent 35% and 32% of the full tree median height. During the generation of the schematic tree included in the Supplementary Data (Additional file 1,

Fig. S5), the sequences were further clustered to a similarity threshold of 0.5 with MMseqs2 [152], then the sequences in each cluster were aligned with MUSCLE [155]. Next, profile-to-profile similarity scores between all clusters were calculated with HHsearch [156], and a UPGMA dendrogram was generated using the pairwise similarity scores. Clusters with high similarity, defined as a pairwise score to self-score ratio > 0.1, were aligned to each other with HHalign [157]. This procedure was performed for a total of 5 iterations. Finally, each cluster alignment was used to build trees using the FastTree program [153] that were rooted at the midpoint and grafted onto the tips of the UPGMA dendrogram that was generated from the cluster similarity scores.

Searches for DCD and eTudor domains were performed similarly to the EVE domains, using the pfam10539 and pfam00567 alignments. For the choanoflagellate eTudor-DCD proteins, a BLAST database was constructed using the predicted protein sequences published by Richter et al. [68] prior to searching with PSI-BLAST. Multiple sequence alignments of EVE, DCD, and eTudor domains were constructed with MUSCLE and PROMALS3D [154, 155].

Genome neighborhood analysis

Domains in genes neighboring prokaryotic EVE and DCD genes (10 on each side) were identified using PSI-BLAST against alignments of domains in the NCBI Conserved Domain Database (CDD) and Pfam (*E*-value 0.001). Some genes were additionally analyzed manually using HHpred. Gene sequence, coordinate, and directional information was downloaded from GenBank using custom Perl scripts. The contextual information network graph was generated using the Cytoscape program [158]. The thickness of edges between nodes represents the strength of association between domains. For any pair of domains in a given genomic neighborhood, the association was calculated as $1/(n+1)$ where n is the number of intervening domains encoded in the neighborhood by distinct genes between the members of the pair. The association values were first averaged across the neighborhoods within each genus and then averaged between the genera to produce the overall weighted average.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12915-020-00885-2.

Additional file 1: Supplementary Figure 1. Phylogenetic distribution of prokaryotic EVE proteins. **Supplementary Figure 2.** Conserved genomic context of EVE proteins in Vibrionales. **Supplementary Figure 3.** Conserved genomic context of EVE proteins in Alteromonadales. **Supplementary Figure 4.** Conserved genomic context of EVE proteins in Pasteurellales. **Supplementary Figure 5.** Schematic phylogenetic tree of EVE domain cluster representatives. **Supplementary Figure 6.** Conserved genomic context of EVE proteins in *Nocardia* and related genera that are found in a distinct CLANS analysis cluster (green in Fig. 2) . **Supplementary Figure 7.** Conserved genomic context of EVE proteins in *Azospirillum* that are found in a distinct CLANS analysis cluster (green in Fig. 2). **Supplementary Figure 8.** COG1743->DUF499->SWI2/SNF2 helicase-nuclease->EVE defense systems in archaea. **Supplementary Figure 9.** Alignment of EVE and DCD domain representatives. **Supplementary Figure 10.** Hypothesis for the evolution of the RdRP complex eTudor proteins in *C. elegans*. **Supplementary Figure 11.** Alignment of the eTudor domain from *Drosophila* SND1 (Tudor-SN) with eTudor domains in choanoflagellates that are fused to DCD domains.

Additional file 2. EVE domain sequences used for CLANS analysis.

Additional file 3. Predicted protein sequences from a choanoflagellate transcriptome sequencing dataset published by Richter et al. that contain extended Tudor domains and/or DCD domains.

Acknowledgements

The authors thank Dr. Andrew Z. Fire (Stanford University) for critical reading of the manuscript and insightful comments, and Koonin group members for helpful discussions.

Authors' contributions

RB and EVK contributed to the conception and design of the study. RB contributed to the data collection. RB, YIW, and EVK contributed to the data analysis. RB and EVK contributed to the manuscript drafting. RB and EVK contributed to the manuscript revision for critical intellectual content. All authors read and approved the final manuscript.

Funding

This work was supported by intramural funds of the US Department of Health and Human Services (the National Library of Medicine, to EVK).

Availability of data and materials

The datasets supporting the analysis presented in this article are included within the additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 14 July 2020 Accepted: 1 October 2020
Published online: 04 November 2020

References

1. Traube FR, Carell T. The chemistries and consequences of DNA and RNA methylation and demethylation. RNA Biol. 2017;14(9):1099–107.
2. Yanas A, Liu KF. Chapter Seven - RNA modifications and the link to human disease. In: Garcia BA, editor. Methods in Enzymology. 626. Cambridge: Academic Press; 2019. p. 133–46.
3. Harcourt EM, Kietrys AM, Kool ET. Chemical and structural effects of base modifications in messenger RNA. Nature. 2017;541(7637):339–46.
4. Carell T, Brandmayr C, Hienzsch A, Müller M, Pearson D, Reiter V, et al. Structure and function of noncanonical nucleobases. Angew Chem Int Ed. 2012;51(29):7110–31.
5. Seelam PP, Sharma P, Mitra A. Structural landscape of base pairs containing post-transcriptional modifications in RNA. RNA. 2017;23(6):847–59.
6. Hofer A, Liu ZJ, Balasubramanian S. Detection, structure and function of modified DNA bases. J Am Chem Soc. 2019;141(16):6420–9.
7. Miyaji H, Yoshimoto T, Asakura H, Komachi A, Kamiya S, Takasaki M, et al. Molecular cloning and characterization of the mouse thymocyte protein gene. Gene. 2002;297(1):189–96.

8. Song A-X, Chang Y-G, Gao Y-G, Lin X-J, Shi Y-H, Lin D-H, et al. Identification, expression, and purification of a unique stable domain from human HSPC144 protein. Protein Expr Purif. 2005;42(1):146–52.
9. Yu F, Song A, Xu C, Sun L, Li J, Tang L, et al. Determining the DUF55-domain structure of human thymocyte nuclear protein 1 from crystals partially twinned by tetartohedry. Acta Crystallogr Sect D. 2009;65(3):212–9.
10. Bertonati C, Punta M, Fischer M, Yachdav G, Forouhar F, Zhou W, et al. Structural genomics reveals EVE as a new ASCH/PUA-related domain. Proteins. 2009;75(3):760–73.
11. Iyer LM, Burroughs AM, Aravind L. The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism. Bioinformatics. 2005;22(3):257–63.
12. Aravind L, Koonin EV. Novel predicted RNA-binding domains associated with the translation machinery. J Mol Evol. 1999;48(3):291–302.
13. Pérez-Arellano I, Gallego J, Cervera J. The PUA domain–a structural and functional overview. FEBS J. 2007;274(19):4972–84.
14. Kim B-N, Shin M, Ha SC, Park S-Y, Seo P-W, Hofmann A, et al. Crystal structure of an ASCH protein from Zymomonas mobilis and its ribonuclease activity specific for single-stranded RNA. Sci Rep. 2017;7(1):12303.
15. Iyer LM, Zhang D, Maxwell Burroughs A, 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(16):7635–55.
16. Schrodinger LLC. The PyMOL Molecular Graphics System, Version 1.8; 2015.
17. Spruijt Cornelia G, Gnerlich F, Smits Arne H, Pfaffeneder T, Jansen Pascal WTC, Bauer C, et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell. 2013;152(5):1146–59.
18. Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. Nature. 2008;455(7214):826–9.
19. Patil DP, Pickering BF, Jaffrey SR. Reading m6A in the transcriptome: m6A-binding proteins. Trends Cell Biol. 2018;28(2):113–27.
20. Hosford CJ, Bui AQ, Chappie JS. The structure of the Thermococcus gammatolerans McrB N-terminal domain reveals a new mode of substrate recognition and specificity among McrB homologs. J Biol Chem. 2020; 295(3):743–56.
21. Zhang Q-H, Ye M, Wu X-Y, Ren S-X, Zhao M, Zhao C-J, et al. Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem/progenitor cells. Genome Res. 2000;10(10):1546–60.
22. Compton MM, Thomson JM, Icard AH. The analysis of cThy28 expression in avian lymphocytes. Apoptosis. 2001;6(4):299–314.
23. Jiang XZ, Toyota H, Yoshimoto T, Takada E, Asakura H, Mizuguchi J. Anti-IgM-induced down-regulation of nuclear Thy28 protein expression in Ramos B lymphoma cells. Apoptosis. 2003;8(5):509–19.
24. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science. 2017;356(6340):eaal3321.
25. Toyota H, Jiang X-Z, Asakura H, Mizuguchi J. Thy28 partially prevents apoptosis induction following engagement of membrane immunoglobulin in WEHI-231 B lymphoma cells. Cell Mol Biol Lett. 2012;17(1):36–48.
26. Aravind L. Guilt by association: contextual information in genome analysis. Genome Res. 2000;10(8):1074–7.
27. Vey G. Metagenomic guilt by association: an operonic perspective. PLoS One. 2013;8(8):e71484.
28. Rogozin IB, Makarova KS, Wolf YI, Koonin EV. Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Brief Bioinform. 2004;5(2):131–49.
29. Doerks T, von Mering C, Bork P. Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. Nucleic Acids Res. 2004; 32(21):6321–6.
30. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol. 2000;18(6):609–13.
31. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 2004;20(18):3702–4.
32. Willis MA, Song F, Zhuang Z, Krajewski W, Chalamasetty VR, Reddy P, et al. Structure of YciI from Haemophilus influenzae (HI0828) reveals a ferredoxin-like α/β-fold with a histidine/aspartate centered catalytic site. Proteins. 2005; 59(3):648–52.
33. Huynen MA, Mühlmeister M, Gotthardt K, Guerrero-Castillo S, Brandt U. Evolution and structural organization of the mitochondrial contact site (MICOS) complex and the mitochondrial intermembrane space bridging (MIB) complex. Biochim Biophys Acta. 2016;1863(1):91–101.
34. Kranz RG, Richard-Fogal C, Taylor J-S, Frawley ER. Cytochrome c biogenesis: mechanisms for covalent modifications and trafficking of heme and for heme-iron redox control. Microbiol Mol Biol Rev. 2009;73:510–28.
35. Pierrel F, Douki T, Fontecave M, Atta M. MiaB protein is a bifunctional radical-S-adenosylmethionine enzyme involved in thiolation and methylation of tRNA. J Biol Chem. 2004;279(46):47555–63.
36. Meganathan R. Ubiquinone biosynthesis in microorganisms. FEMS Microbiol Lett. 2001;203(2):131–9.
37. Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, et al. Phylogeny of Gammaproteobacteria. J Bacteriol. 2010;192(9):2305.
38. Makarova KS, Anantharaman V, Grishin NV, Koonin EV, Aravind L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. Front Genet. 2014;5:102.
39. Yan WX, Chong S, Zhang H, Makarova KS, Koonin EV, Cheng DR, et al. Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. Mol Cell. 2018; 70(2):327–39.e5.
40. Müller AU, Leibundgut M, Ban N, Weber-Ban E. Structure and functional implications of WYL domain-containing bacterial DNA damage response regulator PafBC. Nat Commun. 2019;10(1):4653.
41. Hudaiberdiev S, Shmakov S, Wolf YI, Terns MP, Makarova KS, Koonin EV. Phylogenomics of Cas4 family nucleases. BMC Evol Biol. 2017;17(1):232.
42. Dila D, Sutherland E, Moran L, Slatko B, Raleigh EA. Genetic and sequence organization of the mcrBC locus of Escherichia coli K-12. J Bacteriol. 1990; 172(9):4888.
43. Sukackaite R, Grazulis S, Tamulaitis G, Siksnys V. The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. Nucleic Acids Res. 2012;40(15):7552–62.
44. Nirwan N, Itoh Y, Singh P, Bandyopadhyay S, Vinothkumar KR, Amunts A, et al. Structure-based mechanism for activation of the AAA+ GTPase McrB by the endonuclease McrC. Nat Commun. 2019;10(1):3058.
45. Akita M, Adachi A, Takemura K, Yamagami T, Matsunaga F, Ishino Y. Cdc6/Orc1 from Pyrococcus furiosus may act as the origin recognition protein and Mcm helicase recruiter. Genes Cells. 2010;15(5):537–52.
46. Makarova KS, Koonin EV. Archaeology of eukaryotic DNA replication. Cold Spring Harbor Perspect Biol. 2013;5(11):a012963-a.
47. Makarova KS, Wolf YI, Koonin EV. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. Biol Direct. 2009;4(1):19.
48. Matelska D, Steczkiewicz K, Ginalski K. Comprehensive classification of the PIN domain-like superfamily. Nucleic Acids Res. 2017;45(12):6995–7020.
49. Yeo CC. GNAT toxins of bacterial toxin–antitoxin systems: acetylation of charged tRNAs to inhibit translation. Mol Microbiol. 2018;108(4):331–5.
50. Mruk I, Kobayashi I. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. Nucleic Acids Res. 2014;42(1):70–86.
51. Vasu K, Nagaraja V. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev. 2013;77(1):53–72.
52. Naito T, Kusano K, Kobayashi I. Selfish behavior of restriction-modification systems. Science. 1995;267(5199):897.
53. Fozo EM, Makarova KS, Shabalina SA, Yutin N, Koonin EV, Storz G. Abundance of type I toxin-antitoxin systems in bacteria: searches for new candidates and discovery of novel families. Nucleic Acids Res. 2010;38(11):3743–59.
54. Anantharaman V, Iyer LM, Aravind L. Ter-dependent stress response systems: novel pathways related to metal sensing, production of a nucleoside-like metabolite, and DNA-processing. Mol BioSyst. 2012;8(12): 3142–65.
55. Lourenço Dos Santos S, Petropoulos I, Friguet B. The oxidized protein repair enzymes methionine sulfoxide reductases and their roles in protecting against oxidative stress, in ageing and in regulating protein function. Antioxidants (Basel). 2018;7(12):191.
56. Tenhaken R, Doerks T, Bork P. DCD – a novel plant specific domain in proteins involved in development and programmed cell death. BMC Bioinformatics. 2005;6(1):169.
57. de Camargos LF, Fraga OT, Oliveira CC, da Silva JCF, Fontes EPB, Reis PAB. Development and cell death domain-containing asparagine-rich protein (DCD/NRP): an essential protein in plant development and stress responses. Theor Exp Plant Physiol. 2019;31(1):59–70.

58. Hoepflinger MC, Pieslinger AM, Tenhaken R. Investigations on N-rich protein (NRP) of Arabidopsis thaliana under different stress conditions. Plant Physiol Biochem. 2011;49(3):293–302.
59. Liu H, Wang J-YS, Huang Y, Li Z, Gong W, Lehmann R, et al. Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. Genes Dev. 2010;24(17):1876–81.
60. Hazra D, Chapat C, Graille M. m^6A mRNA Destiny: chained to the rhYTHm by the YTH-containing proteins. Genes (Basel). 2019;10(1):49.
61. Xiao W, Adhikari S, Dahal U, Chen Y-S, Hao Y-J, Sun B-F, et al. Nuclear m6A reader YTHDC1 regulates mRNA splicing. Mol Cell. 2016;61(4):507–19.
62. Liao S, Sun H, Xu C. YTH domain: a family of N6-methyladenosine (m6A) readers. Genomics Proteomics Bioinformatics. 2018;16(2):99–107.
63. Zhao YL, Liu YH, Wu RF, Bi Z, Yao YX, Liu Q, et al. Understanding m6A function through uncovering the diversity roles of YTH domain-containing proteins. Mol Biotechnol. 2019;61(5):355–64.
64. Valverde R, Edwards L, Regan L. Structure and function of KH domains. FEBS J. 2008;275(11):2712–26.
65. Shimberg GD, Michalek JL, Oluyadi AA, Rodrigues AV, Zucconi BE, Neu HM, et al. Cleavage and polyadenylation specificity factor 30: an RNA-binding zinc-finger protein with an unexpected 2Fe-2S cluster. Proc Natl Acad Sci U S A. 2016;113(17):4700–5.
66. Kasowitz SD, Ma J, Anderson SJ, Leu NA, Xu Y, Gregory BD, et al. Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. PLoS Genet. 2018;14(5):e1007412.
67. Hoffmeyer TT, Burkhardt P. Choanoflagellate models — Monosiga brevicollis and Salpingoeca rosetta. Curr Opin Genet Dev. 2016;39:42–7.
68. Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. eLife. 2018;7:e34226.
69. Chen C, Nott TJ, Jin J, Pawson T. Deciphering arginine methylation: Tudor tells the tale. Nat Rev Mol Cell Biol. 2011;12(10):629–42.
70. Jin J, Xie X, Chen C, Park JG, Stark C, James DA, et al. Eukaryotic protein domains as functional units of cellular evolution. Sci Signal. 2009;2(98):ra76.
71. Li C-L, Yang W-Z, Chen Y-P, Yuan HS. Structural and functional insights into human Tudor-SN, a key component linking RNA interference and editing. Nucleic Acids Res. 2008;36(11):3579–89.
72. Gao X, Zhao X, Zhu Y, He J, Shao J, Su C, et al. Tudor staphylococcal nuclease (Tudor-SN) participates in snRNP assembly via interacting with symmetrically dimethylated Sm proteins. J Biol Chem. 2012;287(22):18130–41.
73. Elbarbary RA, Miyoshi K, Myers JR, Du P, Ashton JM, Tian B, et al. Tudor-SN-mediated endonucleolytic decay of human cell microRNAs promotes G(1)/S phase transition. Science. 2017;356(6340):859–62.
74. Li C-L, Yang W-Z, Shi Z, Yuan HS. Tudor staphylococcal nuclease is a structure-specific ribonuclease that degrades RNA at unstructured regions during microRNA decay. RNA. 2018;24(5):739–48.
75. Tóth KF, Pezic D, Stuwe E, Webster A. The piRNA pathway guards the germline genome against transposable elements. Adv Exp Med Biol. 2016;886:51–77.
76. Sturm Á, Perczel A, Ivics Z, Vellai T. The Piwi-piRNA pathway: road to immortality. Aging Cell. 2017;16(5):906–11.
77. Tomar D, Singh R. TRIM family proteins: emerging class of RING E3 ligases as regulator of NF-κB pathway. Biol Cell. 2015;107(1):22–40.
78. Swinehart WE, Deutsch CW, Sarachan KL, Luthra A, Bacusmo JM, de Crécy-Lagard V, et al. Specificity in the biosynthesis of the universal tRNA nucleoside N6-threonylcarbamoyl adenosine (t6A) - TsaD is the gatekeeper. RNA. 2020;26(9):1094–103.
79. Missoury S, Plancqueel S, Li de la Sierra-Gallay I, Zhang W, Liger D, Durand D, et al. The structure of the TsaB/TsaD/TsaE complex reveals an unexpected mechanism for the bacterial t6A tRNA-modification. Nucleic Acids Res 2018;46(11):5850–5860.
80. Deutsch C, El Yacoubi B, de Crécy-Lagard V, Iwata-Reuyl D. Biosynthesis of threonylcarbamoyl adenosine (t6A), a universal tRNA nucleoside. J Biol Chem. 2012;287(17):13666–73.
81. Mráček T, Drahota Z, Houštěk J. The function and the role of the mitochondrial glycerol-3-phosphate dehydrogenase in mammalian tissues. Biochim Biophys Acta. 2013;1827(3):401–10.
82. Hayes RP, Lewis KM, Xun L, Kang C. Catalytic mechanism of 5-chlorohydroxyhydroquinone dehydrochlorinase from the YCII superfamily of largely unknown function. J Biol Chem. 2013;288(40):28447–56.
83. Dressaire C, Moreira RN, Barahona S, Alves de Matos AP, Arraiano CM. BolA is a transcriptional switch that turns off motility and turns on biofilm development. mBio. 2015;6(1):e02352.
84. von der Malsburg K, Müller Judith M, Bohnert M, Oeljeklaus S, Kwiatkowska P, Becker T, et al. Dual role of mitofilin in mitochondrial membrane organization and protein biogenesis. Dev Cell. 2011;21(4):694–707.
85. Galli E, Gerdes K. Spatial resolution of two bacterial cell division proteins: ZapA recruits ZapB to the inner face of the Z-ring. Mol Microbiol. 2010;76(6):1514–26.
86. Field MS, Szebenyi DME, Stover PJ. Regulation of de novo purine biosynthesis by methenyltetrahydrofolate synthetase in neuroblastoma. J Biol Chem. 2006;281(7):4215–21.
87. Kruschwitz HL, McDonald D, Cossins EA, Schirch V. 5-Formyltetrahydropteroylpolyglutamates are the major folate derivatives in Neurospora crassa conidiospores. J Biol Chem. 1994;269(46):28757–63.
88. Stover P, Schirch V. The metabolic role of leucovorin. Trends Biochem Sci. 1993;18(3):102–6.
89. Wassarman KM, Storz G. 6S RNA regulates E. coli RNA polymerase activity. Cell. 2000;101(6):613–23.
90. Wassarman KM. 6S RNA, a global regulator of transcription. Microbiol Spectrum. 2018;6(3):RWR-0019-2018.
91. Steuten B, Hoch PG, Damm K, Schneider S, Köhler K, Wagner R, et al. Regulation of transcription by 6S RNAs: insights from the Escherichia coli and Bacillus subtilis model systems. RNA Biol. 2014;11(5):508–21.
92. Trotochaud AE, Wassarman KM. 6S RNA function enhances long-term cell survival. J Bacteriol. 2004;186(15):4978–85.
93. Hansen S, Lewis K, Vulić M. Role of global regulators and nucleotide metabolism in antibiotic tolerance in Escherichia coli. Antimicrob Agents Chemother. 2008;52(8):2718–26.
94. Ren D, Bedzyk LA, Thomas SM, Ye RW, Wood TK. Gene expression in Escherichia coli biofilms. Appl Microbiol Biotechnol. 2004;64(4):515–24.
95. Chae H, Han K, Kim K-S, Park H, Lee J, Lee Y. Rho-dependent termination of ssrS (6S RNA) transcription in Escherichia coli: implication for 3′ processing of 6S RNA and expression of downstream ygfA (putative 5-formyl-tetrahydrofolate cyclo-ligase). J Biol Chem. 2011;286(1):114–22.
96. K-s K, Lee Y. Regulation of 6S RNA biogenesis by switching utilization of both sigma factors and endoribonucleases. Nucleic Acids Res. 2004;32(20):6057–68.
97. Nakahigashi K, Miyamoto K, Nishimura K, Inokuchi H. Isolation and characterization of a light-sensitive mutant of Escherichia coli K-12 with a mutation in a gene that is required for the biosynthesis of ubiquinone. J Bacteriol. 1992;174(22):7352–9.
98. Feinbaum RL, Urbach JM, Liberati NT, Djonovic S, Adonizio A, Carvunis A-R, et al. Genome-wide identification of Pseudomonas aeruginosa virulence-related genes using a Caenorhabditis elegans infection model. PLoS Pathog. 2012;8(7):e1002813.
99. Aldea M, Hernández-Chico C, de la Campa AG, Kushner SR, Vicente M. Identification, cloning, and expression of bolA, an ftsZ-dependent morphogene of Escherichia coli. J Bacteriol. 1988;170(11):5169–76.
100. Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature. 2008;455(7214):818–21.
101. Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, et al. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature. 2008;455(7214):822–5.
102. Liu G, Fu W, Zhang Z, He Y, Yu H, Wang Y, et al. Structural basis for the recognition of sulfur in phosphorothioated DNA. Nat Commun. 2018;9(1):4689.
103. Weigele P, Raleigh E. Biosynthesis and function of modified bases in bacteria and their viruses. Chem Rev. 2016;116(20):12655–87.
104. Vlot M, Houkes J, Lochs SJA, Swarts DC, Zheng P, Kunne T, et al. Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR-Cas effector complexes. Nucleic Acids Res. 2018;46(2):873–85.
105. Chan WT, Espinosa M, Yeo CC. Keeping the wolves at bay: antitoxins of prokaryotic type II toxin-antitoxin systems. Front Mol Biosci. 2016;3:9.
106. Layer G, Ollagnier-de Choudens S, Sanakis Y, Fontecave M. Iron-sulfur cluster biosynthesis: characterization of Escherichia coli CyaY as an iron donor for the assembly of [2Fe-2S] clusters in the scaffold IscU. J Biol Chem. 2006;281(24):16256–63.

107. Singarapu KK, Liu G, Xiao R, Bertonati C, Honig B, Montelione GT, et al. NMR structure of protein yjbR from Escherichia coli reveals 'double-wing' DNA binding motif. Proteins. 2007;67(2):501–4.
108. Feldmann EA, Seetharaman J, Ramelot TA, Lew S, Zhao L, Hamilton K, et al. Solution NMR and X-ray crystal structures of Pseudomonas syringae Pspto_3016 from protein domain family PF04237 (DUF419) adopt a "double wing" DNA binding motif. J Struct Funct Genom. 2012;13(3):155–62.
109. Bakolitsa C, Kumar A, Carlton D, Miller MD, Krishna SS, Abdubek P, et al. Structure of LP2179, the first representative of Pfam family PF08866, suggests a new fold with a role in amino-acid metabolism. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2010;66(Pt 10):1205–10.
110. Cavalier-Smith T. Origin of animal multicellularity: precursors, causes, consequences—the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. Philos Transact Royal Soc B Biol Sci. 2017;372(1713):20150476.
111. Tucker RP. Horizontal gene transfer in Choanoflagellates. J Exp Zool B Mol Dev Evol. 2013;320(1):1–9.
112. Yue J, Sun G, Hu X, Huang J. The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate Monosiga brevicollis. BMC Genomics. 2013;14(1):729.
113. Akichika S, Hirano S, Shichino Y, Suzuki T, Nishimasu H, Ishitani R, et al. Cap-specific terminal N6-methylation of RNA by an RNA polymerase II–associated methyltransferase. Science. 2019;363(6423):eaav0080.
114. Anders M, Chelysheva I, Goebel I, Trenkner T, Zhou J, Mao Y, et al. Dynamic m6A methylation facilitates mRNA triaging to stress granules. Life Sci Alliance. 2018;1(4):e201800113.
115. Nguyen CD, Mansfield RE, Leung W, Vaz PM, Loughlin FE, Grant RP, et al. Characterization of a family of RanBP2-type zinc fingers that can recognize single-stranded RNA. J Mol Biol. 2011;407(2):273–83.
116. Ritterhoff T, Das H, Hofhaus G, Schröder RR, Flotho A, Melchior F. The RanBP2/RanGAP1*SUMO1/Ubc9 SUMO E3 ligase is a disassembly machine for Crm1-dependent nuclear export complexes. Nat Commun. 2016;7:11482.
117. Selenko P, Sprangers R, Stier G, Bühler D, Fischer U, Sattler M. SMN Tudor domain structure and its interaction with the Sm proteins. Nat Struct Biol. 2001;8(1):27–31.
118. Joosten J, Miesen P, Taşköprü E, Pennings B, Jansen PWTC, Huynen MA, et al. The Tudor protein Veneno assembles the ping-pong amplification complex that produces viral piRNAs in Aedes mosquitoes. Nucleic Acids Res. 2018;47(5):2546–59.
119. Elmore S. Apoptosis: a review of programmed cell death. Toxicol Pathol. 2007;35(4):495–516.
120. Saikia M, Jobava R, Parisien M, Putnam A, Krokowski D, Gao X-H, et al. Angiogenin-cleaved tRNA halves interact with cytochrome c, protecting cells from apoptosis during osmotic stress. Mol Cell Biol. 2014;34(13):2450–63.
121. Martínez-Fábregas J, Díaz-Moreno I, González-Arzola K, Díaz-Quintana A, De la Rosa MA. A common signalosome for programmed cell death in humans and plants. Cell Death Dis. 2014;5(7):e1314-e.
122. Hüttemann M, Pecina P, Rainbolt M, Sanderson TH, Kagan VE, Samavati L, et al. The multiple functions of cytochrome c and their regulation in life and death decisions of the mammalian cell: from respiration to apoptosis. Mitochondrion. 2011;11(3):369–81.
123. Ow Y-LP, Green DR, Hao Z, Mak TW. Cytochrome c: functions beyond respiration. Nat Rev Mol Cell Biol. 2008;9(7):532–42.
124. He B, Lu N, Zhou Z. Cellular and nuclear degradation during apoptosis. Curr Opin Cell Biol. 2009;21(6):900–12.
125. Thomas MP, Liu X, Whangbo J, McCrossan G, Sanborn KB, Basar E, et al. Apoptosis triggers specific, rapid, and global mRNA decay with 3′ uridylated intermediates degraded by DIS3L2. Cell Rep. 2015;11(7):1079–89.
126. Mei Y, Yong J, Liu H, Shi Y, Meinkoth J, Dreyfuss G, et al. tRNA binds to cytochrome c and inhibits caspase activation. Mol Cell. 2010;37(5):668–78.
127. Saikia M, Hatzoglou M. The many virtues of tRNA-derived stress-induced RNAs (tiRNAs): discovering novel mechanisms of stress response and effect on human health. J Biol Chem. 2015;290(50):29761–8.
128. Madungwe NB, Feng Y, Lie M, Tombo N, Liu L, Kaya F, et al. Mitochondrial inner membrane protein (mitofilin) knockdown induces cell death by apoptosis via an AIF-PARP-dependent mechanism and cell cycle arrest. Am J Physiol Cell Physiol. 2018;315(1):C28–43.
129. Wilcox B, Osterman I, Serebryakova M, Lukyanov D, Komarova E, Gollan B, et al. Escherichia coli ItaT is a type II toxin that inhibits translation by acetylating isoleucyl-tRNAIle. Nucleic Acids Res. 2018;46(15):7873–85.
130. Yang X, Yang Y, Sun B-F, Chen Y-S, Xu J-W, Lai W-Y, et al. 5-methylcytosine promotes mRNA export — NSUN2 as the methyltransferase and ALYREF as an m5C reader. Cell Res. 2017;27(5):606–25.
131. Updike DL, Hachey SJ, Kreher J, Strome S. P granules extend the nuclear pore complex environment in the C. elegans germ line. J Cell Biol. 2011;192(6):939–48.
132. Sengupta MS, Boag PR. Germ granules and the control of mRNA translation. IUBMB Life. 2012;64(7):586–94.
133. Sheth U, Pitt J, Dennis S, Priess JR. Perinuclear P granules are the principal sites of mRNA export in adult *C. elegans* germ cells. Development. 2010;137(8):1305–14.
134. Voronina E, Seydoux G, Sassone-Corsi P, Nagamori I. RNA Granules in Germ Cells. Cold Spring Harbor Perspect Biol. 2011;3(12):a002774.
135. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature. 2008;455(7217):1193–7.
136. Fierro-Constaín L, Schenkelaars Q, Gazave E, Haguenauer A, Rocher C, Ereskovsky A, et al. The conservation of the germline multipotency program, from sponges to vertebrates: a stepping stone to understanding the somatic and germline origins. Genome Biol Evol. 2017;9(3):474–88.
137. Caudy AA, Ketting RF, Hammond SM, Denli AM, Bathoorn AMP, Tops BBJ, et al. A micrococcal nuclease homologue in RNAi effector complexes. Nature. 2003;425(6956):411–4.
138. Galganski L, Urbanek MO, Krzyzosiak WJ. Nuclear speckles: molecular organization, biological function and role in disease. Nucleic Acids Res. 2017;45(18):10350–68.
139. Schrader S, Kaldenhoff R, Richter G. Expression of novel genes during somatic embryogenesis of suspension-cultured carrot cells (Daucus carota). J Plant Physiol. 1997;150(1):63–8.
140. Bronkhorst AW, Ketting RF. Trimming it short: PNLDC1 is required for piRNA maturation during mouse spermatogenesis. EMBO Rep. 2018;19(3):e45824.
141. Weick E-M, Miska EA. piRNAs: from biogenesis to function. Development. 2014;141(18):3458.
142. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell. 2007;129(1):69–82.
143. Aravin Alexei A, Chan DC. piRNAs meet mitochondria. Dev Cell. 2011;20(3):287–8.
144. Manage KI, Rogers AK, Wallis DC, Uebel CJ, Anderson DC, Nguyen DAH, et al. A tudor domain protein, SIMR-1, promotes siRNA production at piRNA-targeted mRNAs in *C. elegans*. eLife. 2020;9:e56731.
145. Strome S, Updike D. Specifying and protecting germ cell fate. Nat Rev Mol Cell Biol. 2015;16(7):406–16.
146. Meier P, Finch A, Evan G. Apoptosis in development. Nature. 2000;407(6805):796–801.
147. Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet. 2014;15(6):394–408.
148. Wälti MA, Villalba C, Buser RM, Grünler A, Aebi M, Künzler M. Targeted gene silencing in the model mushroom Coprinopsis cinerea (Coprinus cinereus) by expression of homologous hairpin RNAs. Eukaryot Cell. 2006;5(4):732–44.
149. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11(3):204–20.
150. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
151. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J Mol Biol. 2018;430(15):2237–43.
152. Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics. 2016;32(9):1323–30.
153. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
154. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 2007;23(7):802–8.
155. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
156. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2004;21(7):951–60.

157. Söding J, Remmert M, Biegert A, Lupas AN. HHsenser: exhaustive transitive profile search using HMM–HMM comparison. Nucleic Acids Res. 2006; 34(suppl_2):W374–W8.
158. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.

Publisher's Note