Published online 31 May 2017

Nucleic Acids Research, 2017, Vol. 45, No. 12 ***6995–7020***
doi: 10.1093/nar/gkx494

SURVEY AND SUMMARY

Comprehensive classification of the PIN domain-like superfamily

Dorota Matelska, Kamil Steczkiewicz and Krzysztof Ginalski*

University of Warsaw, CeNT, Laboratory of Bioinformatics and Systems Biology, Zwirki i Wigury 93, 02-089 Warsaw, Poland

Received January 13, 2017; Revised May 18, 2017; Editorial Decision May 23, 2017; Accepted May 24, 2017

ABSTRACT

PIN-like domains constitute a widespread superfamily of nucleases, diverse in terms of the reaction mechanism, substrate specificity, biological function and taxonomic distribution. Proteins with PIN-like domains are involved in central cellular processes, such as DNA replication and repair, mRNA degradation, transcription regulation and ncRNA maturation. In this work, we identify and classify the most complete set of PIN-like domains to provide the first comprehensive analysis of sequence–structure–function relationships within the whole PIN domain-like superfamily. Transitive sequence searches using highly sensitive methods for remote homology detection led to the identification of several new families, including representatives of Pfam (DUF1308, DUF4935) and CDD (COG2454), and 23 other families not classified in the public domain databases. Further sequence clustering revealed relationships between individual sequence clusters and showed heterogeneity within some families, suggesting a possible functional divergence. With five structural groups, 70 defined clusters, over 100,000 proteins, and broad biological functions, the PIN domain-like superfamily constitutes one of the largest and most diverse nuclease superfamilies. Detailed analyses of sequences and structures, domain architectures, and genomic contexts allowed us to predict biological function of several new families, including new toxin-antitoxin components, proteins involved in tRNA/rRNA maturation and transcription/translation regulation.

INTRODUCTION

Metabolism of nucleic acids plays a central role in various cellular processes in all kingdoms of life. Its fundamental component comprises nucleases—highly diverse enzymes that cleave phosphodiester bonds of nucleic acids. Nucleases and their catalytic mechanisms are hugely varied and complex: They can be a protein or RNA; cleave DNA or RNA; may be endonucleases (i.e., cleave a phosphodiester bond more than one nucleotide away from either end of a nucleic acid) or exonucleases (i.e., cleave single nucleotides from an end of a polynucleotide chain); use none, one or two metal ions; and recognize specific substrates based on their structural or sequence features (1). Taking into account their structural folds, they can be classified into evolutionarily related superfamilies, some of which have been broadly described, including GIY-YIG (2), PD-(D/E)XK (3), and RNase H-like (4).

PIN-like domains constitute another major metal-dependent nuclease superfamily with representatives in all kingdoms of life. The name originally refers to the N-terminal domain of an annotated type IV pili twitching motility (PilT) protein (PilT N-terminal domain, PIN) (5). Although this annotation stems from a domain fusion between a PIN domain and a PilT ATPase domain observed in its homologs, a functional link connecting the PIN domains with type IV pili has not been shown yet.

The PIN domain-like superfamily is characterized by a common Rossmanoid fold that consists of a central β-sheet comprising five parallel β-strands, sandwiched with α-helices at both sides (α/β/α sandwich fold, Figure 1). Solved structures and biochemical studies indicate an active site, consisting of well-conserved acidic amino acid residues. The active-site residues are located at the C-terminal region of the core β-strands. According to available crystal structures of holoenzymes and proposed catalytic mechanism, one (6), two (7) or three divalent metal ions, usually Mg^{2+} or Mn^{2+}, are coordinated (8).

*To whom correspondence should be addressed. Tel: +48 22 554 0800; Fax: +48 22 554 0801; Email: kginal@cent.uw.edu.pl

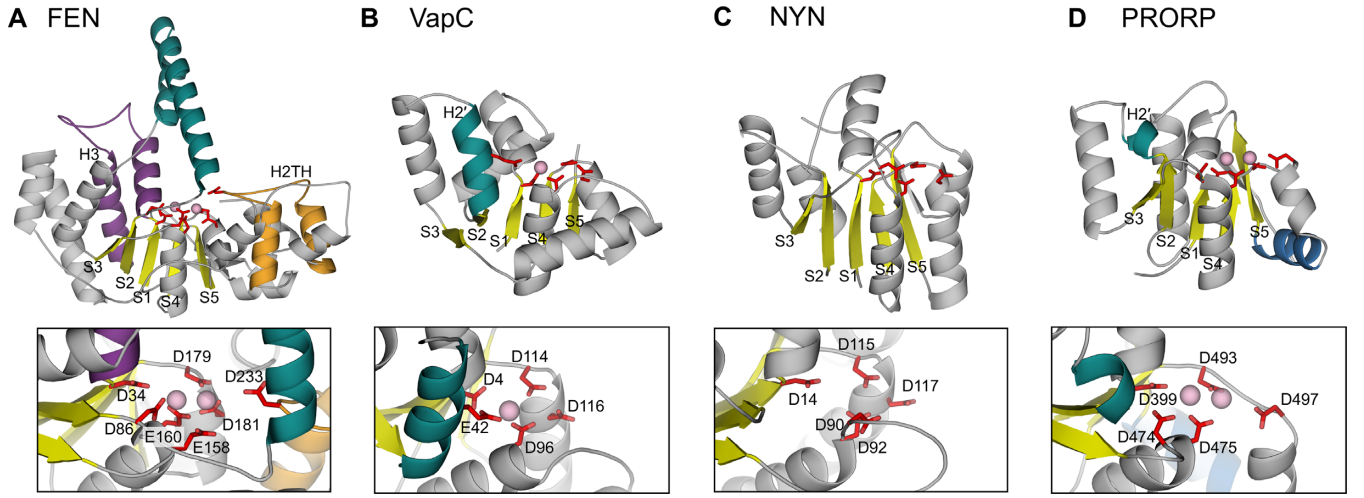

Figure 1. Crystal structures of the major variants of the PIN domain-like fold (top panels) with zoom-in views of their actives sites (bottom panels). (**A**) Structure-specific human FEN-1 nuclease (PDB ID: 3q8k). A 'hydrophobic wedge' between S1 and S2 is shown in dark violet, a 'helical arch' between S2 and H3 is shown in dark green, and a C-terminal helix-2-turn-helix motif (H2TH) in orange. (**B**) A canonical PIN domain of VapC15 from *Mycobacterium tuberculosis* (PDB ID: 4chg). Helix 2′ (H2′) specific for VapC-like domains is shown in dark green. (**C**) VPA0982 from *Vibrio parahaemolyticus* (PDB ID: 2qip). (**D**) PRORP1 from *Arabidopsis thaliana* (PDB ID: 4g24). Short helix 2′ (H2′) is shown in green. β-strands forming the core β-sheet are labeled S1–S5 and shown in yellow, active site residues are shown as red sticks, and metal ions as pink spheres.

The *bona fide* PIN domains characterized with the minimal fold are widely spread in virulence-associated proteins C (VapC). To date, many other nuclease domains have been shown to possess the same core fold, and together they comprise the PIN domain-like superfamily. The use of remote homology detection methods has revealed homology between the canonical PIN nucleases and structure-specific 5′ nucleases (Flap endonucleases, FEN) (9). Later, several other major protein families have been classified as PIN domain-like superfamily members, e.g., Nedd4-BP1/YacP (NYN) (10) and Mut7-C (11). Recently, proteinaceous RNase P (PRORP) responsible for the 5′ tRNA maturation in eukaryotic organelles has turned out to possess a novel variant of the NYN fold (12) (Figure 1). Despite retaining the common core fold and a few conserved active site residues, the PIN-like nucleases are diverse in terms of amino acid sequences, substrate specificities, and catalytic mechanisms.

The structure-specific 5′ nucleases comprise widely spread proteins involved in DNA replication, repair, and recombination. They include flap endonucleases (FENs), 5′–3′ nuclease of DNA polymerase I, the internal domain of Xeroderma Complementation Group G (XPG), and bacteriophage T4 RNase H. They can have 5′–3′ exonucleolytic activity and cleave bifurcated DNA in an endonucleolytic, structure-specific manner (13).

The second major division consists of the PIN domains structurally related to VapC toxins and comprises canonical PIN proteins, which were used to define the core fold. The PIN domains act as endoribonucleases. VapC-like toxins were shown to target various RNAs, including mRNA (14), tRNA (15) and rRNA (16,17), in structure- or sequence-specific manner (18). Eukaryotic homologs of the canonical PIN domains are present in proteins involved in nonsense-mediated mRNA decay (Smg6 (19)), RNA degradation (Rrp44 (20)) and pre-rRNA processing (Nob1 (21), Utp23 (22), Fcf1/Utp24 (23)).

The ubiquitous PIN-domain toxins function in the prokaryotic VapBC-like toxin-antitoxin (TA) systems (24). TA loci were first identified in bacterial plasmids, and they were regarded as involved in stable plasmid maintenance by a so-called 'addiction' mechanism (25). When harbored in the chromosomes of free-living bacteria, the majority of TA systems appear to mediate the general stress response, coordinately enhancing stress survival (26). This is achieved through transcriptional regulation of TA operon expression coupled with posttroperons, in which the antitoxin gene precedes the toxin gene (24). Both genes usually overlap by 1 or 4 nt (in this case, the upstream antitoxin TGA stop codon overlaps with the ATG start codon of the toxin gene in an ATGA sequence) and are co-transcribed from a single promoter located upstream of the antitoxin gene (27). A PIN-domain protein functions as a toxin that can inhibit cell growth or viability by cleavage of the cellular RNA. Antitoxins are small unstable proteins composed of two domains: an N-terminal DNA-binding domain and a C-terminal region involved in toxin binding (28). Formation of the antitoxin-toxin complex results in toxin sequestration and inactivation, and negative autoregulation of the TA operon; otherwise, free antitoxins are degraded by cellular proteases (26). Some genomes harbor many TA operons; for example, the genome of *Mycobacterium tuberculosis* encodes 38 PIN-related TA operons (29).

Despite the central role of the PIN-like domains in the metabolism of nucleic acids, no comprehensive analyses of the relationships between their sequence, structure and function have been performed. Noteworthy, lack of significant sequence similarity between the families makes homology inference a challenging task and hinders new family identification with traditional sequence-based approaches. Here, we combine highly sensitive sequence-based

searches to find distant homologs of known superfamily members and to collect the complete set of proteins possessing PIN-like domain. We discuss predicted and potential function of individual families in the context of their sequences, structures, domain architectures and genomic context. We provide the first systematic classification of the PIN domain-like superfamily, extending our understanding of sequence–structure–function relationships among the nucleases, which may guide further experimental studies.

MATERIALS AND METHODS

Sequence searches, clustering and annotation

Comprehensive identification of PIN-like domains in CDD (v. 3.12) (30), Pfam (v. 30.0) (31) and PDB90 (protein sequences from PDB (32) clustered at 90% sequence identity threshold with CD-HIT (33)) databases was performed using Meta-BASIC (34). Meta-BASIC is a highly sensitive method for distant homology detection based on the comparison of sequence meta-profiles, generated with PSI-BLAST (35) using the NCBI non-redundant protein sequence database derivative (NR70), enriched with secondary structures predicted with PSIPRED (36).

First, known representatives (families and structures) of the PIN domain-like fold were selected from the Pfam (31) (clan CL0280: PIN domain superfamily) and SCOP (v. 1.75, fold c.120) (37) databases. Pfam families were represented as consensus sequences derived from the corresponding seed alignments. Then, they were used as queries in transitive searches with Meta-BASIC (34) against CDD, Pfam and PDB90 databases until no new hits were found. In addition to highly confident hits (i.e., with Meta-BASIC score greater than 40), hits somewhat uncertain (i.e., with Meta-BASIC score between 20 and 40) were also taken into consideration and subjected to further extensive analyses via the Genesilico Metaserver (38) and the HHpred server (39) to identify any potentially correct predictions that may have been placed among the unreliable or incorrect ones. In the case of the Rossmann-fold superfamilies, Meta-BASIC score of 40 corresponds to over 97% of positive predictive value, calculated based on the Pfam clan classifications (data not shown). Correct predictions were selected based on the manual assessment of the conservation of the core secondary structure elements, putative active site residues and hydrophobic positions critical for the PIN domain-like fold, as well as reciprocal Meta-BASIC hits to Pfam, CDD and PDB90 databases. The matrix of pairwise Meta-BASIC scores between all found PIN-like domain families or PDB90 structures can be found in Supplementary Figure S1.

To identify the most complete set of proteins with the PIN-like domains, consensus sequences of the collected families or sequences corresponding to domain structures from PDB90 were used as queries in PSI-BLAST (35) searches against the NCBI NR protein sequence database (as available in April 2015, with six iterations and *E*-value threshold for profile inclusion of 0.001) (40). The collected sequences were clustered at 40% identity threshold with CD-HIT (33) and used as queries in further PSI-BLAST searches (6 iterations, *E*-value < 0.001). The procedure was repeated iteratively until no new hits were found. False positives, which would proliferate in subsequent iterations, were removed. They included new hits with obvious matches to other Pfam or CDD domains not belonging to the PIN domain-like superfamily (with *E*-value $< 10^{-5}$ and sequence coverage $> 50\%$, as obtained with 'hmmscan' (41) against Pfam or RPS-BLAST (35) against CDD).

The final 102,708 NR sequences were annotated with Pfam domains using 'hmmscan' from the HMMER 3.1b package (41) and CDD using RPS-BLAST (35) at the *E*-value cutoff of 10^{-5}. Taxonomic lineages of organisms were assigned according to the NCBI Taxonomy database (40). Transmembrane helices were predicted with TMHMM 2.0 (42). Literature references were searched using PubServer (43). Annotations of the final NR sequences can be found in Supplementary Table S1.

To visualize protein sequence similarities in 2D space, the representatives of sequence clusters at the 80% identity level were subjected to the Fruchterman-Reingold clustering in CLANS (44) (over 70,000 iterations, *E*-value $<$ 0.1). The sequences were assigned into groups, considering highly connected clusters consistent at different *E*-value thresholds (10^{-3}, 10^{-5}, 10^{-10}), and their Pfam and COG/KOG matches. The groups—hereinafter referred to as clusters—were labeled according to the corresponding Pfam families, COG/KOG groups, or after the best-characterized representative. Name with a dot denotes a cluster within a Pfam family ('Pfam'.'subfamily'), PIN_7–PIN_28 correspond to newly defined PIN-like families. Throughout the manuscript, we use a term 'family' interchangeably with 'cluster' if the definition of the cluster is consistent with the current Pfam classification, i.e., it (roughly) corresponds to a Pfam family or it does not have a corresponding entry in Pfam. Relationships between the representative sequences were visualized with Cytoscape 3.2 (45) using the Prefuse Force Directed Layout.

HMM–HMM comparisons

To assess similarities between different PIN-like domain groups defined in the earlier step, their corresponding profile hidden Markov models (HMMs) were compared with HHblits (46). First, sequences corresponding to PIN-like domains were retrieved, clustered at 70% identity with CD-HIT (33), and aligned with 'mafft-linsi' from the MAFFT 7 package (47). For each multiple sequence alignment (MSA), the secondary structure was predicted with HHblits-improved PSIPRED (46) or, in case of groups that include proteins with known tertiary structures, derived from DSSP (48) for respective PDB files. Next, each MSA was filtered with hhfilter ('hhfilter -cov 70 -id 90') and used as an input to build a profile HMM with 'hhmake'. The profiles were compared with 'hhsearch' (49) and the resulting cluster map was visualized with Cytoscape (45).

In addition, HMMs were generated for every cluster with hhbuild from the HHMER 3.1b package (41). They were annotated with computed bit score gathering thresholds and are available as Supplementary Dataset S1. The gathering thresholds correspond to 95% of positive predictive value, calculated based on the results of searching the whole PIN domain-like superfamily sequence dataset.

Structure-based sequence alignment

Structures of PIN-like domains were superimposed within core structural elements using Swiss PDB Viewer (50). The structure-based sequence alignment was derived manually by direct visual comparison of the structures, including respective positions of compared amino acids and their localization in secondary structure elements, to maximize the number of residues aligned between all the analyzed proteins and, where possible, to place insertions and deletions in the loop regions.

Representative sequences of proteins with unknown structures were aligned to the structural alignment in the structurally conserved regions using consensus alignment and 3D assessment approach (51). It was based on secondary structure predictions (36,46), alignments provided by Meta-BASIC (34), and conservation of potential active site residues and hydrophobic profiles within the families.

Structure comparisons

Structures of PIN-like domains superimposed in the previous step were compared to each other using DaliLite v3.3 (52) and GESAMT (53) from the CCP4 v6.5 package (54). As a measure of their similarity, we tested *Z*-score returned by DaliLite and Q-score reported by GESAMT. The scaled scores were represented as a matrix and used as an input for neighbor-joining or UPGMA algorithm implemented in Biopython. Nearly identical (i.e., RMSD < 1 Å for Cα atoms) structures were merged prior to the tree generation. Trees were visualized in Archaeopteryx (55). As the resulting trees had nearly identical topology, we show only the tree based on the Q-score returned by GESAMT (Supplementary Figure S2).

Analysis of genomic neighborhood in prokaryotic genomes

Experimentally confirmed operons were taken from ODB3 (56). Since experimental data on prokaryotic operon architecture is too sparse for comparative genomic analyses, operons were predicted computationally for all fully sequenced genomes available in KEGG GENOME (as of March 2016) (57), using an intergenic distance criterion. Namely, a sequence of genes transcribed from the same strand, with intergenic regions not longer than 100 nt, was considered as a putative operon.

Sequences of PIN-like proteins from the NCBI NR database were mapped to the KEGG GENE database using CD-HIT: if a protein sequence from KEGG GENE was at least 70% identical to any sequence from a given PIN-like cluster, the protein was considered as a representative of that group. Operons containing PIN-like protein-coding genes were further analyzed, i.e., products of their protein-coding genes were mapped to the Pfam and COG families, as described above for the PIN-like proteins.

Prediction of subunits of restriction-modification systems was based on mappings to corresponding Pfam and COG families. Proteins with mappings to the Pfam family HSDR_N (PF04313) or the COG family HsdR (COG4096) were annotated as HsdR subunits. Similarly, HsdM and HsdS subunits were predicted based on mappings to the HsdM_N (PF12161) or HsdM (COG0286), and Methylase_S (PF01420) or HsdS (COG0732) families, respectively.

RESULTS AND DISCUSSION

The first comprehensive catalog of the PIN domain-like superfamily

To identify all protein sequences with PIN-like domains, we conducted an exhaustive approach utilizing both, our state-of-the-art distant homology detection algorithm, Meta-BASIC (as used previously, e.g., in (3,4,58)), and a series of iterative PSI-BLAST searches. The searches were started from the representative sequences of domain families and structures assigned to the PIN domain-like superfamily in Pfam ('PIN' clan, XPG_I and XPG_I_2 families), CDD ('PIN' and 'PIN_SF' superfamilies) and SCOP ('PIN domain-like' superfamily) databases, respectively. As a result of transitive Meta-BASIC searches on Pfam, CDD and PDB90 databases, we identified 105 families and 37 structures, including 41 additional domain families annotated neither in CDD nor Pfam as PIN domain-like superfamily members (Supplementary Table S2, Supplementary Figure S1). Three of them are distant families that have not yet been described as PIN-like domains neither in the literature nor in the databases, i.e., eukaryotic proteins of unknown function DUF1308 (PF07000, KOG4529), bacterial proteins of unknown function DUF4935 (PF16289), and a family of prokaryotic hypothetical proteins COG2454. Among the 38 remaining CCD and Pfam families, previously recognized in the literature as PIN-like domains, yet not covered in the database classifications, several correspond to highly important proteins, e.g., rRNA maturation endonuclease Nob1 (COG1439) (21), potential toxin-antitoxin system component COG4634 (59), and poxvirus G5 proteins (Pox_G5, PF04599) (60–62). In conclusion, we considerably extended both Pfam and CDD definitions of the PIN domain-like superfamily.

To independently and systematically classify the PIN domain-like superfamily into families, we first collected sequences homologous to, or representing the above identified families and structures. Consequently, the representative sequences of all identified families and structures were used as starting points in iterative PSI-BLAST searches against the NCBI non-redundant (NR) protein sequence database (see Materials and Methods). As a result, we found over 100,000 proteins that possess PIN-like domains, out of which 9% and 20% could not be captured using Pfam and COG/KOG domain definitions, respectively (Supplementary Table S1). The matching domain sequences were subjected to clustering in CLANS, in which the sequence space of the PIN domain-like superfamily was represented as a sequence similarity network (Figure 2). Since current Pfam and CDD domain definitions are not consistent between the databases (e.g., PF01850 corresponds to 10 different Cluster of Orthologous Groups, COGs) and often comprise more than one densely connected clusters (e.g., DUF4411 splits into two separate clusters if only edges with BLAST *E*-value $< 10^{-5}$ are considered), we independently divided the PIN domain-like superfamily into 70 sequence clusters. The clusters—hereinafter referred to as 'families' if their definition is consistent with the Pfam classification—correspond

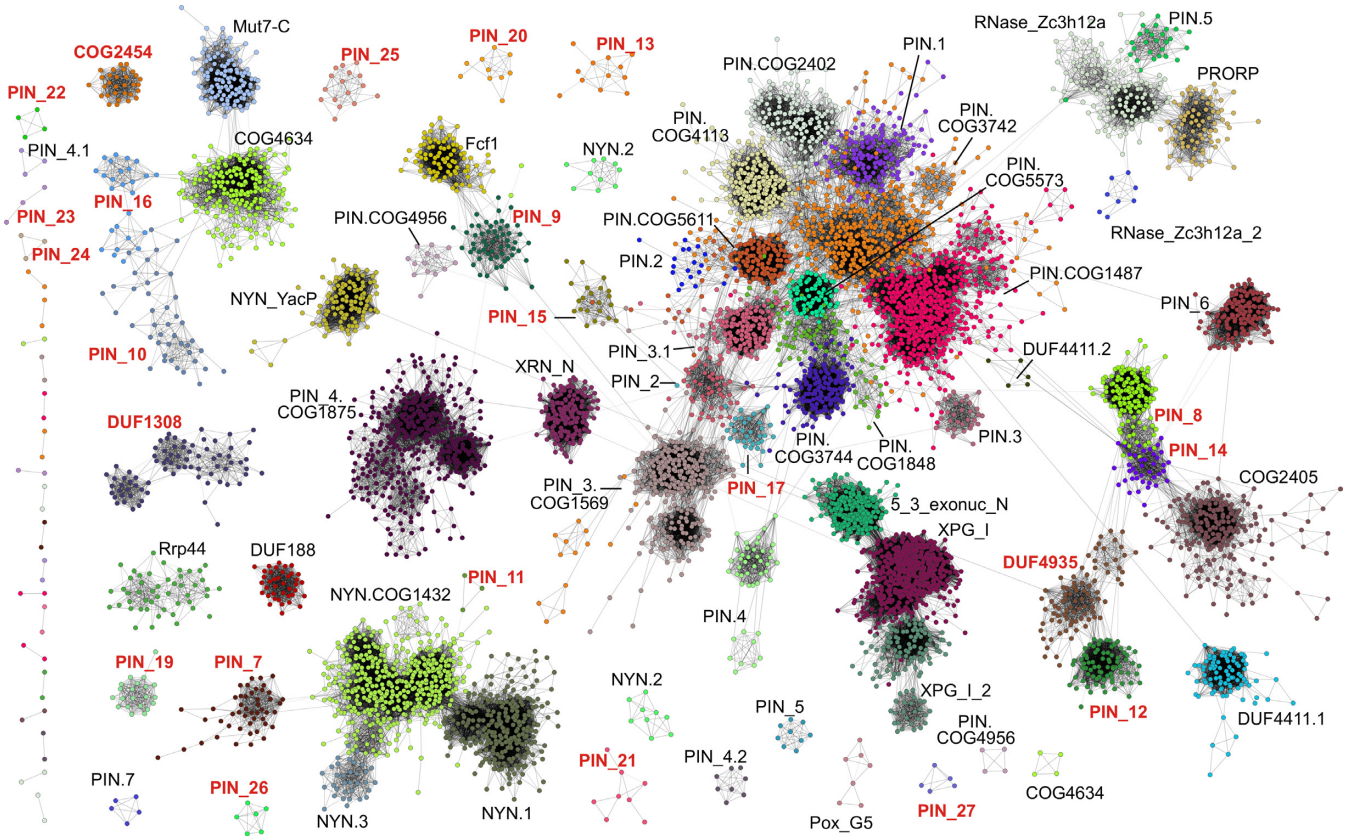

Figure 2. Sequence similarity network of the PIN-like domains. Nodes correspond to the PIN-like domain sequences representing 40% sequence-identity clusters, whereas edges correspond to BLAST *E*-values $< 10^{-5}$. Sequences are colored according to the defined clusters; newly identified PIN-like domains are marked in red. The weighted graph, with weights transformed by –log(*E*-value), was visualized in Cytoscape 3.2 with 10,000 iterations of the Prefuse Force Directed Layout (45).

to groups of highly similar sequences (Figure 2, Supplementary Figure S3). Consequently, we identified 25 clusters that have not been previously included in the Pfam database, out of which 22 cover families previously not recognized as PIN-like (named PIN_7–PIN_22). Biological function, taxonomic distribution, domain architectures, a number of representatives and predicted active site of the defined clusters are summarized in Table 1. Their corresponding profile hidden Markov models are provided in Supplementary Dataset S1.

Structure-based sequence analysis of the PIN-like domains

Due to the high sequence divergence between different PIN domain-like families, the multiple sequence alignment of their representatives could not be generated using standard automatic methods. Instead, to compare sequences of the entire PIN domain-like superfamily, a high-quality sequence alignment was guided by a manually derived structural alignment of known 3D structures. Representative sequences of proteins with unknown structures were aligned to the structural alignment using the consensus alignment and 3D assessment approach (51). Due to the significant divergence of the compared structures in regions outside of the structural core, and uncertainty of predictions for these regions in proteins of unknown structures, a reliable alignment could be generated only for the structural core. As a result, we obtained a multiple sequence alignment of the structurally conserved region common to a majority of the PIN domain-like superfamily proteins, comprising representatives from each PIN-like domain cluster (Figure 3).

The superposition of experimentally solved PIN-like structures showed that the conserved structural core of PIN-like domains consists of five β-strands (S1–S5) and five α-helices (H1–H5), and usually is augmented by peripheral secondary structure elements specific for the individual families (Supplementary Figure S4). The PIN domain-like fold can be classified as Rossmann-like, as it retains a crossover α-helix connecting two parts of the central β-sheet (63). The most conserved elements: S1, S2, crossover helix H4 and S4 define an unambiguously superimposable structural core, while the edge strands S3 and S5, as well as the remaining helices, were more challenging to align (Figure 3). Although on average the PIN-like domains of known structure are ∼130 amino acids (aa) long, the minimal length of the domain bearing all core elements is 103 aa (PDB ID: 2mdt).

The active site is located within the C-terminal region of the β-sheet and is composed of 4–9 negatively charged (Asp/Glu) residues, occasionally supplemented with additional Ser/Thr or Asn/Gln residues. Three of the acidic residues, Asp located downstream to S1, Asp/Glu at the N-terminus of H4, and Asp/Glu downstream to S4 are almost

Table 1. PIN-like domain clusters defined in this study

Cluster name	Structural group	Pfam and COG/KOG domains	Description	Number of active-site residues	PDB IDs	Number of sequences	Assignment to PIN domain-like superfamily	Phyletic distribution
5_3_exonuc_N	FEN	PF02739 (5_3_exonuc_N), COG0258, KOG2519	N-terminal domain of type A DNA polymerases, with 5′–3′ exonuclease and structure-specific endonuclease activities. It functions in DNA replication and repair, cleaves flap structures, including Okazaki fragments.	4–9 (+)	3zd8, 2ihn, 1bgx, 1xo1, 1taq, 1ut5, 1exn, 3h8s, 1tfr	18405 / 238	(164)	Bacteria [17107], Eukaryota [380], dsDNA viruses [243], Archaea [2]
Pox_G5	FEN	PF04599 (Pox_G5)	FEN1-like nucleases conserved in poxviruses, involved in DNA replication and double-strand break repair by homologous recombination (60).	6 (+)	-	98 / 7	(61,62) (fold prediction)	Viruses: Poxviridae [98]
XPG_I	FEN	PF00867 (XPG_I), PF00752 (XPG_N), COG0258, KOG2519, KOG2520, KOG2518, COG5366	Internal domain, which together with the N-terminal domain (XPG_N, PF00752) forms the catalytic domain of the FEN-like structure that contains the active site. In eukaryotes, Rad2/XPG proteins are responsible for a key step of the nucleotide excision DNA repair (NER) pathway, cleaving DNA duplex-containing bubble or loop structures during DNA replication, repair and recombination (165). The UL41 protein (virion host shutoff, vhs) of herpes simplex virus 1 selectively degrades mRNA by endonucleolytic cleavage early in infection (157). The function of archaeal proteins is not known.	7 (+)	1b43, 4wa8, 1ul1, 4q0r, 3ory, 3q8m, 2izo, 1mc8, 3q8k, 3qe9, 1a76, 3qea, 1rxv, 4q0w, 5cnq	6119 / 524	Pfam	Eukaryota [5510], Archaea [363], dsDNA viruses [169], Bacteria [2], unclassified viruses [1]
XPG_I_2	FEN	PF12813 (XPG_I_2), KOG2518	Function unknown. Eukaryotic family of asteroid homologs, in Drosophila possibly functioning in EGFR signaling (166).	6–7 (+)	-	1056 / 123	Pfam	Eukaryota [1056]
XRN_N	FEN	PF03159 (XRN_N), COG5049, KOG2044, KOG2045	In eukaryotes, major 5'–3' exoribonucleases involved in mRNA decay (167), dsRNA accumulation (168), and antiviral responses (169). Function of the viral homologs is not known.	7 (+)	2y35, 3pie, 3fqd	2886 / 167	Pfam	Eukaryota [2835], dsDNA viruses [32], Bacteria [6]
COG2405	VapC	PF11848 (DUF3368), COG2405	Potential toxins from toxin-antitoxin systems, including new three-component systems.	2–5 (+/-, e.g. PDB ID: 2mdt)	2mdt	1675 / 235	Pfam	Bacteria [1248], Archaea [357]
DUF1308	VapC	PF07000 (DUF1308), KOG4529	Function unknown. Present in some mimiviruses, eukaryota and some cyanobacteria, so probably of chloroplast-origin in eukaryota. The Pfam definition comprises two domains: N-terminal, distantly related to the PD-(D/E)XK nucleases, and C-terminal, PIN-like domain, which probably lacks acidic residues from the active site.	2–4 (-/+)	-	762 / 78	This work	Eukaryota [744], Bacteria [10], dsDNA viruses [7]
DUF4411.1	VapC	PF14367 (DUF4411)	Potential toxins from three-component toxin-antitoxin systems, together with HTH DNA-binding domains and DUF955 proteases.	4–5 (+)	-	643 / 91	Pfam	Bacteria [616], Archaea [13]
DUF4411.2	VapC	PF14367 (DUF4411)	Potential toxins from toxin-antitoxin systems.	4–6 (+)	-	14 / 6	Pfam	Bacteria: Bacillales [14]
DUF4935	VapC	PF16289 (DUF4935)	Function unknown. The Pfam definition comprises two regions: the N-terminal one is a PIN-like domain, whereas the α/β C-terminal region does not show homology to any protein of known structure.	4–5 (+)	-	320 / 84	This work	Bacteria [310], Archaea [2]
Fcf1	VapC	PF04900 (Fcf1), COG1412, KOG3164, KOG3165	Maturation of rRNA. In human and yeasts, Utp24 is an essential endoribonuclease processing the 18S rRNA precursor at site A1 and A2 (170). Its homolog, Utp23, appears to be an inactive nuclease, with a general RNA-binding function (22).	2–4 (+, e.g. Utp24 / -, e.g. Utp23)	4mj7	1868 / 90	Pfam	Eukaryota [1859], Bacteria [4], Archaea [2]

Table 1. Continued

Cluster name	Structural group	Pfam and COG/KOG domains	Description	Number of active-site residues	PDB IDs	Number of sequences	Assignment to PIN domain-like superfamily	Phyletic distribution
PIN.COG1487	VapC	PF01850 (PIN), COG1487	VapC-like toxins from toxin-antitoxin systems.	2–4 (+/-)	3zvk, 4chg, 4xgq, 2bsq, 3tnd, 3dbo, 2h1c, 1v96, 3h87, 2h1o, 1y82	10532 / 660	Pfam	Bacteria [9852], Archaea [413], Eukaryota [3]
PIN.COG1848	VapC	PF01850 (PIN), COG1848	Toxins from toxin-antitoxin systems.	3/4 (+/-)	-	938 / 81	Pfam	Bacteria [902]
PIN.COG2402	VapC	PF01850 (PIN), COG2402	Toxins from toxin-antitoxin systems.	3–4 (+/-)	1w8i	1620 / 169	Pfam	Bacteria [1266], Archaea [278], Eukaryota [11]
PIN.COG3742.COG1848.COG4374	VapC	PF01850 (PIN), COG3742, COG4374, COG1848	Toxins from toxin-antitoxin systems.	3–4 (+/-)	-	3265 / 469	Pfam	Bacteria [2421], Archaea [739], Eukaryota [3]
PIN.COG3744	VapC	PF01850 (PIN), COG3744	Toxins from toxin-antitoxin systems.	3–4 (+/-)	-	2506 / 155	Pfam	Bacteria [2429], Archaea [18], Eukaryota [2]
PIN.COG4113	VapC	PF01850 (PIN), COG4113	Toxins from toxin-antitoxin systems.	4 (+)	2fe1, 1v8o, 1v8p	1611 / 243	Pfam	Bacteria [1248], Archaea [301], Eukaryota [1]
PIN.COG4956	VapC	PF01850 (PIN), COG4956	Function unknown. In addition to the PIN domain, the proteins possess TRAM, a putative RNA-binding domain (171) and four predicted transmembrane helices at the N-terminus.	3–4 (+/-)	3ix7	2002 / 20	Pfam	Bacteria [1946], Archaea [6], Eukaryota [2]
PIN.COG5573	VapC	PF01850 (PIN), COG5573	Toxins from toxin-antitoxin systems.	4 (+)	-	695 / 82	Pfam	Bacteria [674], Archaea [6], Eukaryota [1]
PIN.COG5611	VapC	PF01850 (PIN), COG5611	Toxins from toxin-antitoxin systems.	3–4 (+/-)	-	1139 / 140	Pfam	Bacteria [1066], Archaea [36]
PIN.1	VapC	PF01850 (PIN)	Toxins from toxin-antitoxin systems.	3–4 (+/-)	-	1163 / 155	Pfam	Bacteria [1085], Archaea [39]
PIN.2	VapC	PF01850 (PIN)	Toxins from toxin-antitoxin systems.	4–6 (+)	-	671 / 19	Pfam	Bacteria: Actinobacteria [663], Proteobacteria [1]
PIN.3	VapC	PF01850 (PIN)	Toxins from toxin-antitoxin systems.	2–4 (+/-)	-	324 / 32	Pfam	Bacteria [303], Archaea [2]
PIN.4	VapC	PF01850 (PIN)	Function unknown. Transcriptionally coupled to DUF4325-encoding genes.	5–6 (+)	-	119 / 45>	Pfam	Bacteria [108], Archaea [3]
PIN.5	VapC	PF01850 (PIN)	Function unknown.	4 (+)	-	104 / 28	Pfam	Bacteria [104]
PIN.6	VapC	PF01850 (PIN)	Function unknown.	3–5 (+/-)	-	64 / 12	Pfam	Bacteria [58]
PIN.7	VapC	PF01850 (PIN)	Function unknown.	4 (+)	-	26 / 5	Pfam	Bacteria: Cyanobacteria [22], Proteobacteria [2], Bacteroidetes/Chlorobi group [1]
PIN_2	VapC	PF10130 (PIN_2), COG5378	Toxins from toxin-antitoxin systems.	3–4 (+/-)	-	222 / 51	Pfam	Bacteria [144], Archaea [71]
PIN_3.COG1569	VapC	PF13470 (PIN_3), COG1569	Typically, toxins from toxin-antitoxin systems. Some representatives are encoded in operons comprising ATP-grasp ligase, ATPase and HNH nuclease, which were proposed to constitute a novel conflict system, where RNA ligase would neutralize toxic behavior of the nucleases (172).	4–6 (+)	-	3548 / 331	Pfam	Bacteria [3384], Archaea [52], Eukaryota [5]
PIN_3.1	VapC	PF13470 (PIN_3)	Toxins from toxin-antitoxin systems.	4–7 (+)	-	963 / 161	Pfam	Bacteria [906], Archaea [6]
PIN_4.COG1875	VapC	PF13638 (PIN_4), COG1875	In bacteria, toxins from toxin-antitoxin systems. In eukaryota, SMG5 and SMG6 are components of nonsense-mediated mRNA decay (NMD) machinery (161). Swt1 is an endoribonuclease that participates in quality control of nuclear messenger ribonucleoprotein particles and can associate with the nuclear pore complex (173). In bacteria and bacteriophages, fused with PhoH domain to form PhoH2 proteins, which function as sequence-specific RNA helicases and RNases, likely responding to nutrient stress (174).	3–4 (+/-, e.g. SMG5, PDB ID: 2hwy)	2hwx, 2hwy, 2hww, 2dok	6518 / 423	Pfam	Bacteria [4328], Eukaryota [2041], dsDNA viruses [22]

Table 1. Continued

Cluster name	Structural group	Pfam and COG/KOG domains	Description	Number of active-site residues	PDB IDs	Number of sequences	Assignment to PIN domain-like superfamily	Phyletic distribution
PIN_4.1	VapC	PF13638 (PIN_4)	Function unknown.	4 (+)	-	25 / 5	Pfam	Eukaryota [21], Bacteria [4]
PIN_4.2	VapC	PF13638 (PIN_4)	Function unknown.	3–4 (+/-)	-	27 / 10	Pfam	Bacteria [23], Eukaryota [4]
PIN_5	VapC	PF08745 (PIN_5), COG1458	Function unknown. The gene co-occurrence patterns suggest that it may interact with RNA ligase from TIGR01209 family, tRNA methyltransferase and tRNA-synthetase.	4 (+)	-	197 / 9	Pfam	Archaea [128], Bacteria [65]
PIN_6	VapC	PF17146 (PIN_6), COG1439	In eukaryotes, Nob1 proteins are endoribonucleases involved in 18S rRNA maturation (21,175). Function of archaeal homologs is not known.	3–4 (+/-)	2lcq	1316 / 141	(21) (fold prediction)	Eukaryota [989], Archaea [280], Bacteria [2]
Rrp44	VapC	KOG2102	RNA degradation within exosome. Rrp44 (DIS3) acts as an Mn-dependent endoribonuclease from the exosome core (20,162). DIS3 has a paralog, DIS3L, with a disfunctional PIN-like domain (176).	3–5 (+/-)	2wp8, 4ifd, 4pmw, 5c0w	167 / 60	(177) (fold prediction), (178) (crystal structure)	Eukaryota [167]
PIN_8	VapC	-	Function unknown.	4–5 (+)	-	458 / 139	This work	Bacteria [441], Archaea [5], dsDNA viruses [1]
PIN_9	VapC	COG1412	Function unknown. Archaea-specific Fcf1-like domains, not matching the Fcf1 Pfam model.	4–5 (+)	1o4w	353 / 56	(179)	Archaea [311], Bacteria [1]
PIN_12	VapC	-	Function unknown. Related to DUF4935.	4–5 (+)	-	240 / 79	This work	Bacteria [240]
PIN_13	VapC	-	Potential toxins from toxin-antitoxin systems.	1–3 (-)	-	218 / 13	This work	Bacteria: Actinomycetales [218]
PIN_14	VapC	-	Potential toxins from three-component toxin-antitoxin systems, together with HTH DNA-binding domains and DUF955 proteases.	3–4 (+/-)	-	213 / 41	This work	Bacteria [204], Archaea [3]
PIN_15	VapC	-	Mainly potential toxins from toxin-antitoxin systems. The domains are fused with GCN5-related acetyltransferases or potential RNA-binding (82) PUA-like domains, and transcriptionally coupled with PUA-HTH fusion proteins (Dalk_4501, MPET_RS08500).	3–5 (+/-)	-	182 / 22	This work	Bacteria [161], Archaea [12], Eukaryota [1], dsDNA viruses [1]
PIN_17	VapC	-	Potential toxins from toxin-antitoxin systems. The PIN-like domain is fused with an acetyltransferase domain, and encoded upstream to HTH-ASCH fusion protein genes (COG4933).	3–5 (+/-)	-	117 / 35	This work	Bacteria [117]
PIN_18	VapC	-	Function unknown.	4–6 (+)	-	97 / 1	This work	Archaea: Euryarchaeota [93], unclassified Archaea [1]
PIN_19	VapC	-	Function unknown.	3–4 (+/-)	-	54 / 22	This work	Bacteria [51], Archaea [1]
PIN_20	VapC	-	Potential toxins from three-component toxin-antitoxin systems, together with HTH DNA-binding domains and DUF955 proteases.	4–6 (+)	-	50 / 11	This work	Bacteria: Actinomycetales [50]
PIN_21	VapC	-	Function unknown.	4 (+)	-	42 / 8	This work	Archaea [39], Bacteria [2]
PIN_22	VapC	-	Function unknown.	4–5 (+)	-	29 / 4	This work	Bacteria: Clostridium [29]
PIN_23	VapC	-	Function unknown.	3–4 (+/-)	-	25 / 7	This work	Bacteria [20], Archaea [4]
PIN_24	VapC	-	Function unknown.	3–4 (-/+)	-	23 / 3	This work	Bacteria: Cyanobacteria [23]
PIN_25	VapC	-	Potential toxins from toxin-antitoxin systems.	3–4 (+/-)	-	20 / 17	This work	Bacteria [17]
PIN_26	VapC	-	Potential toxins from toxin-antitoxin systems.	4–5 (+)	-	20 / 6	This work	Bacteria: Firmicutes [20]
PIN_27	VapC	-	Function unknown.	4 (+)	-	8 / 5	This work	Archaea: Euryarchaeota [7], unclassified Archaea [1]

Table 1. Continued

Cluster name	Structural group	Pfam and COG/KOG domains	Description	Number of active-site residues	PDB IDs	Number of sequences	Assignment to PIN domain-like superfamily	Phyletic distribution
PIN_28	VapC	-	Function unknown.	4 (+)	-	8 / 2	This work	Archaea: Sulfolobaceae [8]
COG2454	NYN	COG2454	Function unknown. Fused N-terminally with alpha-helical DUF434. In some archaea, located within ribosomal or tRNA operons.	4–5 (+)	-	291 / 30	This work	Bacteria [200], Archaea [86]
DUF188	NYN	PF02639 (DUF188), COG1671	Function unknown.	5 (+)	-	4252 / 42	Pfam	Bacteria [4125], Eukaryota [3]
NYN.COG1432	NYN	PF01936 (NYN), COG1432	Function unknown. The cluster comprises LabA-like proteins, which in *Synechococcus elongatus* are involved in negative feedback expression regulation of the circadian clock protein KaiC (97,98).	1–5 (+/-)	2qip	8754 / 460	Pfam	Bacteria [7883], Archaea [360], Eukaryota [260], dsDNA viruses [1]
NYN.1	NYN	PF01936 (NYN)	Function unknown.	3–6 (+/-)	-	1813 / 277	Pfam	Eukaryota: Viridiplantae [1151], Opisthokonta [638], other [17]
NYN.2	NYN	PF01936 (NYN)	Function unknown. Majority encoded downstream to the genes encoding putative tRNA methyltransferases TrmB. *M. tuberculosis* Rv0207c was implicated in a unique heme uptake system (105).	2–6 (+/-)	-	682 / 19	Pfam	Bacteria [509], Eukaryota [167]
NYN.3	NYN	PF01936 (NYN)	Function unknown.	2–4 (+/-)	-	225 / 35	Pfam	Eukaryota [214], Bacteria [10]
NYN_YacP	NYN	PF05991 (NYN_YacP), COG3688	Function unknown.	4–7 (+)	-	2959 / 120	Pfam	Bacteria [2588], Eukaryota [168]
PIN_7	NYN	-	Function unknown.	3–4 (+/-)	-	657 / 54	This work	Bacteria [603], Eukaryota [1]
PIN_11	NYN	-	Function unknown. C-terminal domain of bilaterial ZNF451 proteins, comprising 887–1002 region in isoform 1 of human ZNF451 (Uniprot ID: Q9Y4E5–1). In higher eukaryotes, fused with zinc-finger motifs.	3–4 (+/-)	-	284 / 3	This work	Eukaryota: Eumetazoa [282]
PRORP	PRORP	PF16953 (PRORP)	Processing of pre-tRNA at the 5′-end in mitochondria and chloroplasts (12).	4–5 (+)	4g23, 4g24, 4xgl, 5diz	830 / 79	(127) (crystal structure)	Eukaryota [820], dsDNA viruses [1]
RNase_Zc3h12a	PRORP	PF11977 (RNase_Zc3h12a), KOG3777	Two evolutionary separated groups. In higher eukaryotes, MCPIP1 (Zc3h12a) is involved in regulation of mRNA decay (70) and miRNA turnover (132), and cleavage of viral RNA (131). Unknown function in bacteria.	0–5 (+/-)	3v32, 3v33	2122 / 130	Pfam	Eukaryota [1925], Bacteria [154], Archaea [23]
RNase_Zc3h12a_2	PRORP	PF14626 (RNase_Zc3h12a_2)	Function unknown. *C. elegans* eri-9 protein interacts with DICER in endogenous RNAi pathway (180).	2–5 (+/-)	-	26 / 8	Pfam	Eukaryota: Chromadorea [26]
COG4634	Mut7-C	COG4634	In bacteria, potential toxins from toxin-antitoxin systems (59).	3–5 (+/-)	-	1732 / 204	(59)	Bacteria [1536], Archaea [117], Eukaryota [4], dsDNA viruses [1]
Mut7-C	Mut7-C	PF01927 (Mut7-C), COG1656	Function unknown. PIN domain-like fold with an inserted zinc ribbon at the C terminus. In eukaryotes, the Mut7-C domain is fused N-terminally to the 3′–5′ exonuclease RNase D family domain, whereas in archaea, it is a standalone module and in bacteria, it is fused with a ubiquitin member of potential RNA-binding function (11).	2–4 (+/-)	-	1869 / 125	(11)	Bacteria [1055], Eukaryota [556], Archaea [210]
PIN_10	Mut7-C	-	Potential toxins from toxin-antitoxin systems, related to COG4634. Recently, a crystal structure of its DUF433-containing antitoxin VapB45 (Rv2018) from *M. tuberculosis* was solved (PDB ID: 5af3).	3–4 (+/-)	-	297 / 43	This work	Bacteria [297]

Table 1. Continued

Cluster name	Structural group	Pfam and COG/KOG domains	Description	Number of active-site residues	PDB IDs	Number of sequences	Assignment to PIN domain-like superfamily	Phyletic distribution
PIN_16	Mut7-C	-	Potential toxins from toxin-antitoxin systems.	3–4 (+/-)	-	150 / 26	This work	Bacteria [150]

Clusters are named according to the corresponding Pfam families, COG/KOG groups, or after the best-characterized representative. Name with a dot denotes a cluster within a Pfam family ('Pfam'.'subfamily'). Matches to Pfam and CDD were computed with HMMER (41) and RPS-BLAST (35), respectively, at the *E*-value cutoff of 10^{-5}. The number of active site residues was predicted based on the conservation of acidic residues at the positions corresponding to known active sites. '+'/'-' in parentheses denotes the presence of predicted active/inactive nucleases. 'PDB IDs' refer to PDB IDs of solved structures within PDB90 (proteins with known structure clustered at 90% sequence identity). 'Number of sequences' is based on the NCBI NR database (40). Numbers following slash refer to the number of representatives at 40% identity based on clustering of the corresponding sequence sets with CD-HIT (33). In 'Assignment to PIN domain-like superfamily', 'Pfam' refers to the clan CL0280 (PIN) in the Pfam database (31). Taxonomic lineages of organisms were assigned according to the NCBI Taxonomy database (40). Numbers in square brackets in 'Phyletic distribution' refer to numbers of sequences from the NCBI NR database.

always invariant, whereas the position of the remaining one varies across different families, and in most cases it is situated downstream to S2 (Figure 3). All the active site residues contribute to metal ions binding either directly or via a network of water molecules. Mutagenesis studies proved that at least four acidic residues are required to sustain the nuclease activity of the PIN-like domains (23).

The PIN domain-like superfamily could be divided into several major groups, taking into account (a) positions of the active site residues (Figure 3), (b) structural clustering (Supplementary Figure S2), (c) presence of additional secondary structure elements inserted in the structural core (Figure 3, Supplementary Figure S4), and (d) clustering of profile hidden Markov models of the derived clusters (Supplementary Figure S5). Consistently with previous observations (10), we could distinguish five major structural classes, represented accordingly by FEN, VapC (canonical PIN domains), NYN, PRORP and Mut7-C families, out of which only the Mut7-C group lacks known structure and probably represents a deteriorated version of the PIN domain-like fold (Figure 1). It should be noted that differences between VapC, NYN and PRORP groups are subtle, and the above division may be iteratively improved upon release of new structures.

Description of the PIN-like groups

In the following sections, we will characterize the major structural groups and clusters belonging to them, with a focus on the domains of unknown function. For brief descriptions of all the defined clusters, the reader is referred to Table 1.

Group 1: Structure-specific FEN-like nucleases

The FEN-like division encompasses the earliest structurally described PIN-folded protein domains (64) and their catalytic mechanism was studied from the structural perspective in detail (13). These structure-specific nucleases are characterized by the most extensive elaborations to the core fold. They comprise 5′–3′ nuclease domains of polymerase I (5_3_exonuc_N (PF02739)), XRN (XRN_N (PF03159)), FEN and XPG (both belonging to XPG_I (PF00867)), Asteroid (XPG_I_2 (PF12813)), and poxviral G5R (Pox_G5 (PF04599)) proteins.

The structure specificity is thought to be achieved owing to a structural element inserted between S2 and H3, named the 'helical arch' (65) (or 'tower domain' in XRN1 (66)), which promotes the proper orientation of the processed substrate. The arch domain harbors a lysine residue (Lys93 in XRN1 (66) or Lys83 in T5 exonuclease (67)), which is essential for the exonuclease processivity. It binds to the 5′-phosphate of the product DNA and presumably acts as an electrostatic catalyst (68,69). The arch is partly disordered in substrate-free FEN structures and gains structure upon threading of the DNA substrate, according to the proposed 'disorder-thread-order' mechanism (69).

Additional insertion, in FEN-like exonucleases known as 'hydrophobic wedge', is composed of two helices between S1 and S2 and its role is to break the substrate path and to position 3′-flap (13). Other groups, except for Mut7-C, also contain helices between S1 and S2, yet in the 5′–3′ exonucleases, they are usually considerably longer. Interestingly, RNase_Zc3h12a, RNase_Zc3h12a_2, NYN_YacP and COG2454 contain a similarly long predicted insertion, which could be aligned to the FEN-specific extensions of H1 and H2 (Figure 3). Among them, RNase_Zc3h12a was studied experimentally—its representative, MCPIP1, was shown to possess only endonucleolytic activity (70).

At the C-terminus of the PIN-like domain, the FEN-like nucleases are fused with a helix-two-turn-helix (H2TH) or helix-three-turn-helix (H3TH) domain, which binds to dsDNA (71,72). All FEN-like domain structures also feature an expanded β-sheet, i.e., with the additional sixth antiparallel β-strand at the C-terminus (Supplementary Figure S4). XPG_I, XRN_N, and, according to the sequence mappings, also XPG_I_2 and Pox_G5, contain a seventh β-strand, N-terminally adjacent to S1 from the core β-sheet. Unlike 5_3_exonuc_N, all XPG_I and XRN_N structures determined to date have an additional seventh β-strand and a conserved PCNA-binding motif at the C-terminus (71). These features underlie a split of structure-specific PIN-like nucleases into two groups in the structural clustering (Supplementary Figure S2).

In the FEN-like proteins, residues from both the PIN-like and H2TH/H3TH domains contribute to the active site. Within the PIN core, they retain a conserved pattern of six active site residues (conserved in 80% of sequences), i.e., four invariant aspartates located downstream to S1, S2, S4 and upstream to H5, supplemented by two Asp/Glu/Gln residues upstream to H4 (Figure 3).

Group 2: VapC-like nucleases

The 'VapC-like' variant of the PIN domain-like fold, found in toxic endonucleases, involves a stripped-down core of

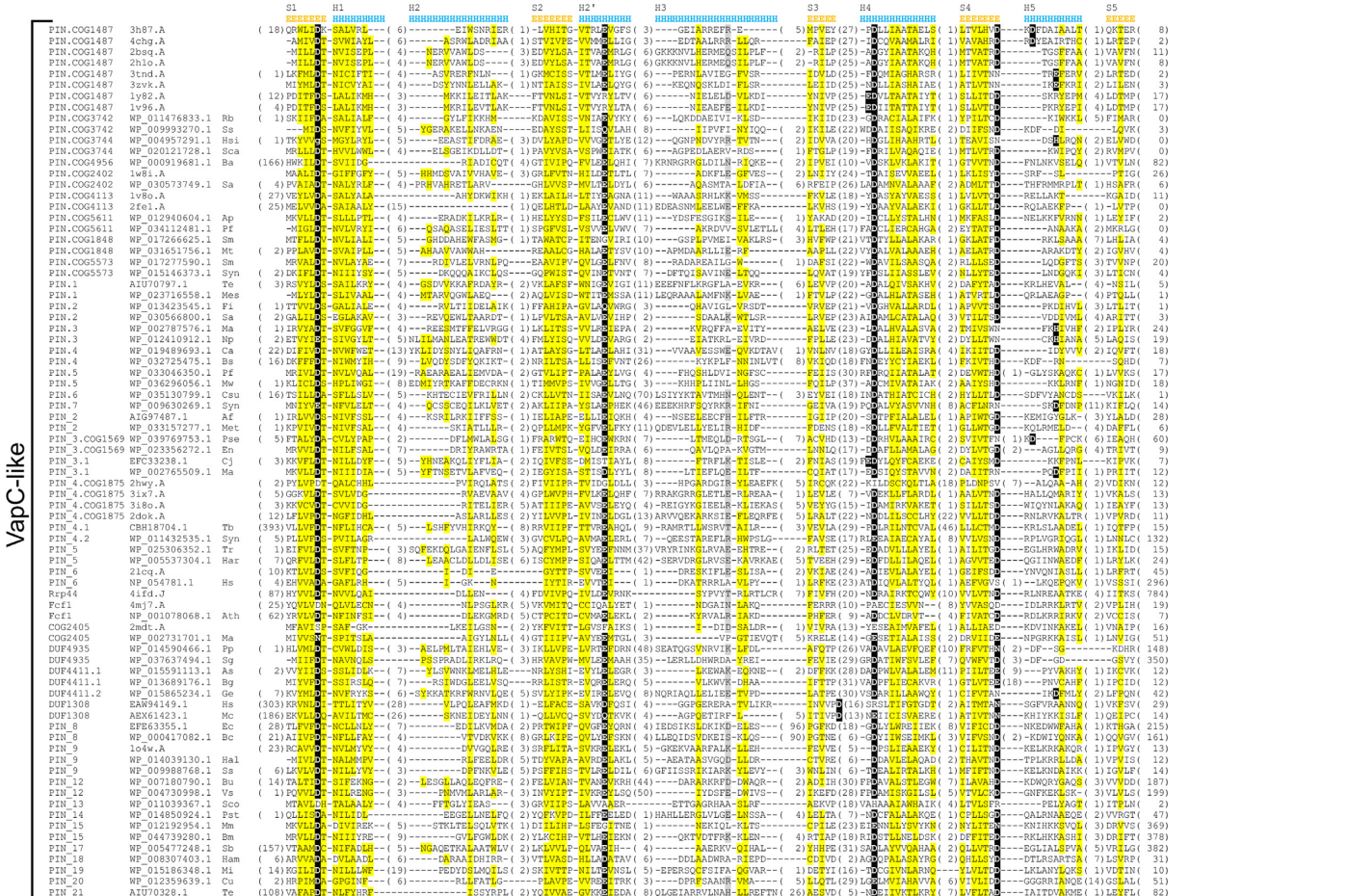

Figure 3. Multiple sequence alignment of the conserved core elements of the PIN domain-like superfamily. The sequence blocks (VapC-like, FEN-like, NYN-like, PRORP and Mut7-C) correspond to the defined structural groups. Each defined cluster is represented by one or more sequences, labeled with NCBI accession numbers or, for proteins of known structure, PDB codes. The numbers of excluded residues are specified in parentheses. Residue conservation is denoted with the following scheme: uncharged, highlighted in yellow; polar, highlighted in gray; known or potential active site residues, highlighted in black. Secondary structure elements (E, β-strand; H, α-helix) are shown above the corresponding alignment blocks. Abbreviated species are defined in Supplementary Dataset S2.

the FEN-like nucleases lacking the helical arch and the H2TH/H3TH element (Figure 1). The arch is replaced by a shorter helical insert H2′, located between S2 and H3 and harboring a catalytic Glu residue. It substitutes Asp from S2, which is present in FEN-like, but absent in VapC-like nucleases. The helical insertion facilitates a cleft, which accepts a C-terminal α-helix of VapB or FitA antitoxins and stabilizes it upon interference with the nuclease catalytic site (73). During toxin-antitoxin assembly, through elaborative multimerization events, a tetramer of heterodimers is formed (74), capable of binding to a promoter region of the toxin-antitoxin operon (75).

Apart from the prototypic VapC proteins involved in TA systems, the above structural features can be attributed to many other PIN-like clusters (altogether 49 identified clusters, Figure 3), including well-described domains involved in tRNA and rRNA maturation (Fcf1, PIN_6, Rrp44 families).

PIN-domain toxin-antitoxin systems. The majority of canonical PIN domain proteins in prokaryotes are toxin-like components encoded in toxin-antitoxin (TA) operons. The largest, central cluster of the sequences (Figure 2) corresponds to VapC (virulence-associated protein C) proteins from prokaryotic VapBC TA systems. VapB is a transcription factor-like protein acting as an inhibitor and VapC is the PIN-domain ribonuclease toxin. Since the proteins are encoded by a *vapBC* operon, often with overlapping open reading frames, it is relatively straightforward to predict analogous systems. Based on the exhaustive genomic neighborhood analysis of the defined PIN-like clusters, we found that 28 of them comprise ribonucleases acting in potential TA systems (Table 1). In total, we predicted that 20% PIN domain-like superfamily proteins encoded in fully sequenced prokaryotic genomes (6,525 out of 32,958) are involved in TA systems. In addition to already annotated PIN (PF01850), PIN_2 (PF10130), PIN_3 (PF13470), PIN_4 (PF13638) and COG4634 families, we found that representatives of COG2405 (DUF3368 (PF11848)), DUF4411 (PF14367) and nine families newly defined in this study (i.e., PIN_10, PIN_13, PIN_14, PIN_15, PIN_16, PIN_17, PIN_20, PIN_25, PIN_26) are encoded in direct proximity to transcription factor-like genes, suggesting their role in TA modules. On the other hand, we noticed that some PIN subfamilies (i.e., PIN.COG4956, PIN.4, PIN.5, PIN.6, PIN.7) are characterized by different genomic contexts, which suggests subfunctionalization within this canonical PIN cluster (Table 1).

Most commonly, the TA operons consist of two genes, where antitoxin is transcribed upstream of a PIN-domain protein gene and encodes a small single-domain protein. Preferences for neutralizing partners differ among PIN-like clusters (Figure 4, Supplementary Figure S6A). In general, most widespread antitoxins co-occurring with PIN-domain proteins are: RelB/MetJ/Arc-like characterized by a ribbon-helix-helix (RHH) fold (29% of the PIN-like TA-associated domains), AbrB/MraZ/MazE-like with a swapped hairpin fold (double-split beta-barrel in SCOP, 24%) (76), intrinsically disordered YefM/RelE/ParE-like domains with Phd fold (22%), and HigB-like with a helix-turn-helix (HTH) motif (16%). UPF0175 (PF03683,

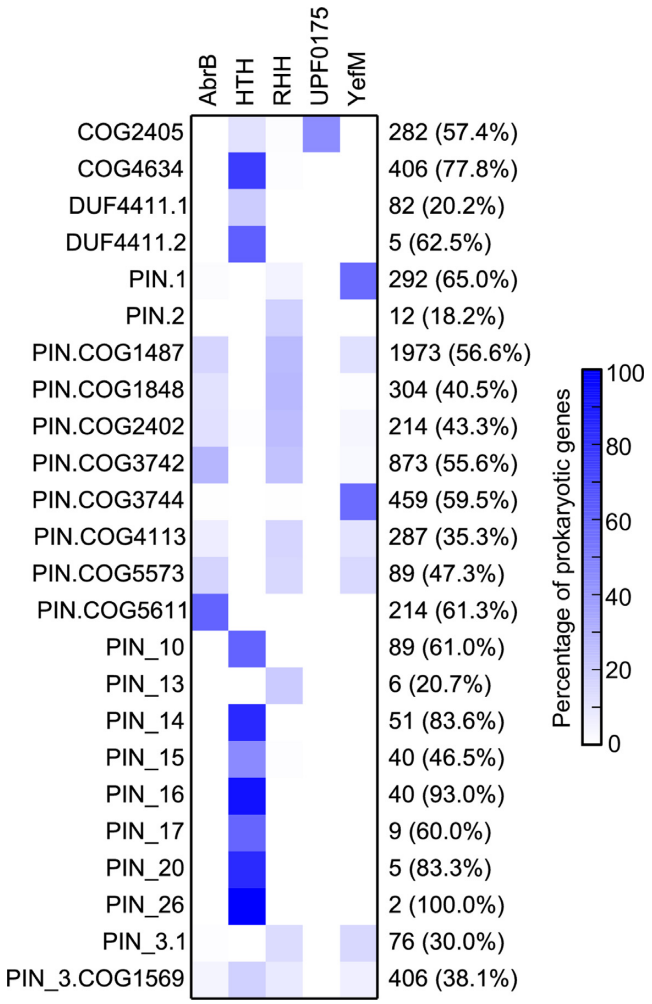

Figure 4. Gene co-occurrence of the PIN-domain toxin and major antitoxin families. For a given PIN-like domain family, percentages correspond to the number of prokaryotic genes located on the same strand and in close proximity (separated by less than 100 nt) to the genes that encode an antitoxin of a given family (AbrB, HTH, RHH, UPF0175 or YefM), in reference to the total number of prokaryotic genes belonging to the family. The calculations are based on the KEGG GENOME database (March 2016) (57). Shown are the PIN-like families that include at least 10% prokaryotic genes encoded in potential toxin-antitoxin operons.

COG2866) is found exclusively in the context of COG2405 and its putative DNA-binding function in TA systems was discussed by Makarova *et al.* (59).

As it was noticed before (77), some PIN-domain toxin genes are potentially encoded within more expanded TA operons. For example, PIN_17 and PIN_15, are associated with GCN5-related N-acetyltransferase (GNAT) and several PUA-like (or ASCH) domains, and are annotated in protein databases as 'acetyltransferase family proteins' (Supplementary Figure S6B). These systems are found sporadically in various bacteria and archaea lineages and are co-transcribed with different genes depending on the species. The GNAT proteins catalyze the transfer of the acetyl group from acetyl coenzyme A (Ac-CoA) to a nitrogen atom in their protein or small molecule substrates (78). In prokaryotes, this reaction is implicated in vari-

ous functions, from antibiotic resistance and xenobiotic metabolism to the regulation of translation (78). The PIN-GNAT-PUA architecture was previously observed in the context of new systems for nucleic acid acetylation, where PUA would act as a recognition module, and RNA would serve as a substrate or as a guide for DNA acetylation (79). The PIN-associated GNAT-like domains exhibit significant sequence similarity to the Pfam acetyltransferase GNAT family (PF00583) and possess the same topology of predicted secondary structure elements (βααββββαβαβ). However, the PIN-associated domain lacks a conserved motif A (Arg/Gln-X-X-Gly-X-Gly/Ala), which in GNATs is responsible for Ac-CoA pyrophosphate binding (80). Therefore, the PIN-like domain proteins contain a degenerated version of the GNAT domain, meaning that—despite the widespread annotation as 'acetyltransferase family proteins'—they probably lack the acetyltransferase activity.

In the discussed operons, a GNAT-like domain is always N-terminally fused with a PIN-like domain and C-terminally fused with a PUA-like domain of RNA-binding function (81,82). A second copy of the PUA domain is always encoded in another operon component, and in α-proteobacteria, it is fused with the DNA-binding HTH domain (Supplementary Figure S6B). Nevertheless, the protein and domain arrangement within the homologous operons is highly variable, even within individual taxonomic phyla. In some genomes of α-proteobacteria and Bacteroidetes, the operons are augmented by additional ATPase protein-coding genes. Furthermore, some proteobacterial and Bacteroidetes operons with the GNAT-PUA combination lack a PIN-like domain, suggesting its dispensable function in these systems.

GNAT proteins located in operons encoding also HTH and RHH DNA-binding proteins were previously hypothesized to be involved in antibiotic resistance (59). However, none of the predicted GNAT proteins associated with nuclease domains has been characterized. A potential GNAT-HTH TA system was discovered and tested experimentally in *Acinetobacter baumannii* (83). Interestingly, its components appeared to function oppositely than expected: the HTH domain proteins act as a toxin arresting bacterial growth, whereas the GNAT domain protein neutralizes this effect. Specific roles of GNAT, PUA and HTH domains in the extended systems utilizing PIN-like domains, as well as a potential upstream signal triggering the response, remain to be elucidated.

In another three-component regulatory system, a PIN-like domain (i.e., COG2405, DUF4411, PIN_14, PIN_20, or PIN_26) is co-transcribed with HTH and protease DUF955 (PF06114, COG2856) domains (Supplementary Figure S6C). In some cases, HTH domains are fused with both, PIN-like domain and DUF955, and also located as a standalone component of the TA systems. DUF955 is a zinc-dependent peptidase-like domain present in IrrE from *Deinococcus deserti* (84), ImmA from *Bacillus subtilis* (85), and transcriptional regulators RamB (86) and PrpC (87). The IrrE central domain is folded in a three-helix bundle, with the second and third helices forming the HTH motif (88). In TA systems, proteases act downstream by degrading free, unstructured antitoxins. Therefore, the Xre family protein of the HTH fold might act as an antitoxin, the activity of which can be both, autoregulated and inhibited by the proteolytic cleavage by the DUF955-HTH fusion protein. This hypothesis is supported by the fact that both *B. subtilis* ImmA and *D. deserti* IrrE were shown to degrade transcriptional repressors of the HTH fold (84,89). As the PIN-associated DUF955 domain retains a conserved HEXXH zinc-binding motif (88), the HTH-DUF955 genes may represent a fused version of these toxin-antitoxin systems. In response to the cellular stress, DUF955 may act as a protease degrading HTH Xre domains encoded within the same gene, and thus activating the expression of the PIN-domain gene. Consequently, this may constitute a novel complex toxin-antitoxin system, in which the upstream signal triggering the protease activity still needs to be determined.

Link of PIN-like toxins to restriction-modification systems. TA systems might preferentially cluster with and stabilize other antivirus defense systems in the so-called defense islands, i.e., discrete DNA segments that include various defense systems (90). Restriction-modification systems (RM) are another type of selfish mobile prokaryotic elements, which sometimes behave as discrete units of life (91). Interestingly, some PIN-related TA systems are organized into more complex clusters related to programmed death, including also subunits of RM systems. A link between TA and RM systems has been recently suggested by Mruk and Kobayashi, who noted that roles of restriction enzymes are similar to those of toxins of TA systems, and modification enzymes correspond to their antitoxins (92). Nevertheless, any direct interaction between TA and RM systems, including involvement of PIN-like nucleases as restriction modules, has not been shown yet.

Based on the genomic neighborhood analysis of PIN-like genes, we found several cases of PIN-like nucleases presumably transcriptionally coupled to type I RM subunits, usually accompanied by other genes (Supplementary Figure S7). In some cases, antitoxin elements are missing, which suggests a different mechanism of regulation of the toxin's activity. Usually, the genomic arrangements are not conserved between closely related species, suggesting recent rearrangements. Interestingly, we did not observe PIN-like nucleases in the context of putative type II RM systems (i.e., without specificity subunit). However, we found three operons encoding a PIN-like nuclease (usually associated with TA systems), HsdM and HsdS subunits, without a clear candidate for a restriction enzyme (Supplementary Figure S7). Among the most complex defense islands, an operon in the genome of *Belliella baltica* was found to encode a newly discovered TA system (DUF433-PIN_10), type I HsdM and HsdS subunits, and a single-chain modification-dependent type IV restriction endonuclease, GmrSD.

DUF4411 (PF14367). DUF4411 remains a functionally uncharacterized domain present in single-domain proteins from bacteria and Archaea, which, according to the sequence clustering, can be separated into two subfamilies, DUF4411.1 and Bacillales-specific DUF4411.2. They are distantly related to the archetypal PIN domain and contain a potential conserved active site that consists of four negatively charged residues. DUF4411.2-coding genes are al-

most always located in the context of DNA-binding HTH domains, suggesting their function within the canonical TA systems. In most Actinobacteria, the genes encoding DUF4411.1 are expressed as mono-cistrons, while in other phyla, they are located downstream to genes encoding fused DUF955 (COG2856, ImmA) and HTH domains (Supplementary Figure S6C). As discussed earlier, this suggests that they may be involved in complex regulatory systems, based on three components: transcriptional regulator, nuclease, and protease. Interestingly, in *Geobacillus* sp. Y412MC61, the gene encoding the HTH-DUF955 fusion protein is located upstream to a gene encoding the COG2405 (another PIN-like domain) protein (GYMC61_2169). As a gene encoding DUF4411.2 (GYMC61_3375) is located downstream to the Xre (HTH) gene, this may represent shuffling of the PIN-like domain proteins within different TA systems.

C-terminal region of DUF1308 (PF07000). According to our predictions, DUF1308 (PF07000) comprises two domains: the N-terminal region distantly related to the PD-(D/E)XK nucleases and the C-terminal part with apparent homology to the known PIN-like domains. The PIN-like domain is widespread among eukaryotes, including animals, plants, and fungi, but also present in some cyanobacteria, Deinococcus, and dsDNA viruses from the Mimiviridae family. DUF1308 covers almost full length of human protein C7orf25 (Uniprot ID: Q9BPX7), with the PIN-like domain encompassing the 245–410 region. Interaction of C7orf25 with threonyl-tRNA synthetase was included in the high-quality human interactome (93), however, its function has not been studied.

The domain retains up to four potential catalytic residues, thus, depending on the protein is predicted to be an active or inactive nuclease. In eukaryotic and viral proteins, it contains a large predicted helical insert after the first canonical strand, whereas in bacterial proteins, the length of this region is comparable to other canonical PIN-domain structures. Since this helical region is located on the side of the potential active site, it may influence substrate discrimination. Interestingly, in a majority of the DUF1308-containing proteins, the PIN-like domain is N-terminally fused with a potentially active PD-(D/E)XK-like domain. However, in some apicomplexa (*Toxoplasma gondii*, *Hammondia hammondi*) and bacteria (*Anabaena sp. 90*, *Nostoc sp. PCC7524*), homologous proteins retain only the PIN-like domain. In cyanobacteria, the DUF1308 proteins are encoded upstream to genes coding for proteins with DUF3349 domains. Although the function of DUF3349 is unknown, analysis of its 3D structure ruled out a possible DNA-binding role (94) and, consequently, the role of the DUF1308 PIN-like domain within TA systems. In a vast majority of proteins, the PD-(D/E)XK-like and PIN-like domains are sole domains; cases of the DUF1308 domain fusions are exceptional and include N-terminal nucleophile (Ntn)-hydrolase domain, DUF1349, and LCCL domains. The Ntn-DUF1308 fusion in some metazoan proteins (*Chelonia mydas*, *Camelus ferus*) provides connection to the 20S proteasome, where the Ntn domains may act as peptidases (95).

N-terminal region of DUF4935 (PF16289). DUF4935 (PF16289) groups hundreds of uncharacterized proteins around 350 residues in length and is found in various bacterial species from Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria. Its N-terminal region (1–200) shows homology to PIN-like domains (PIN, PIN_2, PIN_3, PIN_4) and several crystal structures (e.g., PDB ID: 1w8i, PDB ID: 1y82, PDB ID: 3tnd), and appears centrally in the cluster map of models corresponding to the defined PIN-like families (Supplementary Figure S5). Its less conserved C-terminal part (201–380), comprising several predicted α-helices and β-strands, does not show homology to any protein with solved structure. Therefore, it can be either an addition to the fold core, or a separate domain. Sequence alignment of the N-terminal part to other PIN-like domains with solved structures shows conservation of several acidic residues critical for chelating metal ions (i.e., corresponding to a potential active site in PDB ID: 3tnd). Apart from these conserved residues, the DUF4935 family proteins share a long helical insert following the S2 strand and two α-helices following S3, a feature characteristic for some PIN-like domains (Supplementary Figure S4). Analysis of the genomic context suggests that the DUF4935 genes are usually expressed as mono-cistrons.

Group 3: NYN-like nucleases

The third major group of the PIN domain-like superfamily, NYN, was defined by Anantharaman and Aravind in 2006 (10). Although it spans proteins from all kingdoms of life, none of them has been studied experimentally in the context of nucleolytic function; however, some of the proteins were suggested to act as RNA endonucleases, with specificity achieved through additional domains (96). The original definition of the NYN group comprises two Pfam families, i.e., NYN (PF01936) and NYN_YacP (PF05991), with rather distantly related sequences (Figure 2). We extended the NYN-like group by four other families, two of which were newly defined. Bacterial PIN_7 and eukaryotic PIN_11 cluster closely with NYN, whereas prokaryotic COG2454 and DUF188 seem more related to NYN_YacP (Figure 3).

It was predicted that NYN domains bind one metal ion through four conserved acidic residues (10). According to the predictions and a crystal structure of one representative (PDB ID: 2qip), the NYN domain resembles the canonical PIN fold, however, unlike PIN and FEN variants, it lacks a helical insert after S2, which results in a relatively exposed active site (Figures 1 and 3). The 'cross-over helix', connecting the S3 and S4 strands of the structural core, does not contain other helical inserts, as in some VapC-like structures. This feature distinguishes NYN-like domains from other groups. Notably, in NYN_YacP and COG2454, the predicted H2 helix is considerably longer than in the NYN family.

NYN (PF01936). The NYN family defined by Anantharaman and Aravind was typified by eukaryotic Nedd4-binding protein 1 (N4BP1) and the bacterial YacP proteins (10). However, its current Pfam definition does not cover any of the proteins: N4BP1 belongs to RNase_Zc3h12a (PF11977) and YacP proteins are classified within the

NYN_YacP (PF05991) family. Accordingly, NYN encompasses a large group of moderately similar sequences, i.e., according to our sequence clustering, could be subsequently divided into four groups, here referred to as NYN.COG1432, NYN.1, NYN.2, and NYN.3. Despite the earlier hypothesis, none of their representatives has been studied in the context of the putative nuclease function. Based on the conserved gene neighborhood, bacterial NYN domains were implicated in tRNA or rRNA processing (10), yet, any functional studies are still lacking.

NYN.COG1432 is the most abundant group, present in all domains of life, primarily bacteria, and comprises LabA-like proteins defined under COG1432 (CDD ID: cd06167). In cyanobacterium *Synechococcus elongatus*, LabA and its paralog lalA are involved in negative feedback expression regulation of the circadian clock protein KaiC (97,98). Interestingly, LabA homolog (nicB, PSS_0380) in *Pseudomonas putida* S16 possesses a nicotine-degrading ability, catalyzing the hydroxylation of 6-hydroxy-3-succinoylpyridine (99,100). However, another domain present in the protein, DUF2384, may be responsible for this activity. Some bacterial NYN-coding genes can be found in operons together with genes encoding tRNA/rRNA maturation proteins, such as SpoU (Ferpe_1981 from *Fervidobacterium pennivorans*) and RNase T (PPS_4608 from *Pseudomonas putida* S16), or DNA-modification enzymes, such as Udg4 (Gdia_0379 from *Gluconacetobacter diazotrophicus* PA1 5). In eukaryotes, the LabA-like PIN domains are present mostly in fungi, usually fused with Oskar-TDRD5/TDRD7 HTH (OST-HTH) RNA-binding domains. In symbiotic fungus *Rhizophagus*, the LabA-like PIN domain was previously found in multiple effectors for plant pathogenesis and suggested as a toxin targeting specific host RNAs (101). The PIN-(OST-HTH) domain architecture is commonly found also in bacteria and archaea. Previously, it was predicted as a novel, standalone RNA-degradation system, where the OST-HTH domain would recruit substrates for processing or degradation (102).

NYN.1 is a eukaryote-specific cluster, abundant in opisthokonts and green plants. Some plant species have undergone substantial expansions of the genes encoding this domain, e.g., the *Arabidopsis lyrata* genome bears 91 such paralogs. Human limkain b1 (meiosis arrest female 1 protein, MARF1) was previously shown as a component of mRNA processing bodies through a direct interaction with a central component of the mammalian core-decapping complex, Ge-1 (103). Its role was implicated in the protection against DNA double-strand breaking, as well as in meiosis and retrotransposon surveillance in oocytes (104). MARF1 contains NYN.1, RNA recognition motif (RRM) and OST-HTH domains, however, its specific role in the regulation of mRNA stability still needs to be determined.

NYN.2 groups bacterial and plant LabA-like uncharacterized proteins of high sequence similarity (i.e., nearly 700 sequences cluster to only 19 representatives at 40% sequence identity). In Actinobacteria, the corresponding genes (e.g., Rv0207c from *M. tuberculosis*) are usually found in operons coding for a putative tRNA methyltransferase (TrmB, COG0220; Rv0208c) and membrane efflux pump (YdfJ, COG2409; Rv0206c). Recently, the Rv0202c–Rv0207c genomic region was described to encode a unique mycobacterial heme uptake system implicated in sequestering heme iron from the host (105). However, the role of the NYN protein in this system has not been studied.

NYN.3 comprises mainly fungal (e.g., *Fusarium*, *Aspergillus*, *Penicillum*, *Agaricus*) uncharacterized proteins (usually longer than 400 aa). They do not share a characteristic common domain architecture; however, the NYN.3 domain is typically located at the C-terminus.

NYN_YacP (PF05991). Unlike NYN domains, the NYN_YacP domain is characterized by a conserved 'GYN' motif following the first conserved Asp in strand S1. This feature was described by Anantharaman and Aravind as distinguishing NYN sequences from PIN and FEN groups (10), however, with current Pfam definitions of the families, it is specific only to NYN_YacP. According to our predictions, proteins with NYN_YacP contain two α-helices with conserved basic residues at the C-terminus of the nuclease domain. Conservation of putative catalytic amino acids suggests that the NYN_YacP domain functions as an active ribonuclease, with active site formed by at least four Asp residues, potentially assisted by several Glu and Ser/Thr residues (Figure 3).

NYN_YacP is ubiquitously found in bacteria and some eukaryotes, mainly plants. The function of any of its representatives is not known, but the gene neighborhood analysis suggests that it can be involved in the central RNA metabolism. In most gram-positive bacteria and eukaryotes, NYN_YacP family members are single-domain proteins. The gene encoding an NYN_YacP-domain protein in *B. subtilis* (BSU00970) is a part of a large gltX-cysE-cysS-mrnC-yacO-yacP operon, which encodes glutamyl-tRNA synthetase, serine acetyltransferase, cysteinyl-tRNA synthetase, mini-ribonuclease 3 processesing 23S rRNA, and putative tRNA/rRNA methyltransferase YacO (106) (Supplementary Figure S6G). The conserved gene neighborhood persists in almost 300 genomes from Firmicutes (mostly Bacilli and Clostridia). A homologous operon in *Listeria monocytogenes* contains also sigma-70 RNA polymerase factor rpoE (107). Altogether, the data suggest that in the gram-positive bacteria NYN_YacP may be involved in the rRNA/tRNA processome complex.

Interestingly, in some Clostridia, NYN_YacP is fused with a TetM-like domain (CDD ID: cd04168). The TetM-like tetracyclin-resistance proteins are paralogs of the translational GTPase EF-G, thus can mimic their structure and function, and actively remove tetracycline from the ribosome in a GTP-hydrolysis-dependent manner (108). Whether the role of the TetM_like domain in TetM-NYN_YacP fusion proteins is related to tetracycline resistance or whether a nuclease can be involved in this process is unclear. Intriguingly, some of the corresponding genes are found to be potentially co-expressed with genes encoding HTH-domain proteins, suggesting a potential functional diversification within this highly conserved nuclease family (Supplementary Figure S6G).

PIN_11. PIN_11 is a newly identified family that comprises a C-terminal domain of human ZNF451 (COASTER) and its homologs. It is specific to Eumetazoa

and distantly related to the bacterial family PIN_7. According to our sequence-to-structure mappings, these two families represent a minimized NYN variant of the PIN domain-like fold (Figure 3). PIN_11 is a potential active nuclease due to the presence of at least four conserved Asp residues in the predicted active site.

PIN_11 corresponds to the 887–1002 region in isoform 1 of human ZNF451 (Uniprot ID: Q9Y4E5–1), while isoform 2 (Uniprot ID: Q9Y4E5–2) is missing the N-terminal beta-strand of the PIN domain-like fold. Human ZNF451, similarly to other higher eukaryotes, contains 11 predicted C2H2-type zinc fingers. It functions in promyelocytic leukemia bodies in the nucleus as a transcriptional cofactor, as a coactivator or corepressor, depending on the factors with which it interacts (109). Moreover, ZNF451 was shown to interact with p300 by the PIN-like domain and to negatively regulate TGF-beta signaling in a p300-dependent and sumoylation-independent manner (109). Interestingly, PIN_11 sequences contain several conserved Cys and His residues within the PIN-like domain, which may suggest stabilization of the domain structure with an embedded short zinc-binding loop. The zinc finger may enhance binding of the PIN-like protein to the target RNA, as observed in MCPIP1 (110).

DUF188 (PF02639). DUF188 is a family of well conserved, uncharacterized bacterial proteins from a wide range of taxonomic phyla, comprising YqxD from *B. subtilis* (BSU25230) and YaiI from *Escherichia coli* (b0387). Multiple sequence alignment of the family shows at least five conserved positions of acidic residues following the predicted core S1, S3 and S4 strands. The predicted secondary structure profile resembles that of the NYN-like variant of the PIN domain-like fold, with a short α-helix after S1 and a short loop that connects S2 and S3. Moreover, similarly to NYN_YacP, DUF188 retains two long predicted α-helices following the PIN-like domain, with several conserved positively charged residues that might be involved in substrate binding.

In *E. coli*, *yaiL* is transcribed as a mono-cistron (111), whereas in Firmicutes, e.g., *B. subtilis* and *Listeria monocytogenes*, YqxD proteins are encoded within RNA polymerase major sigma43 operons. The sigma43 operon comprises *dnaG*, encoding the DNA primase involved in the initiation of chromosome replication, and *rpoD*, which encodes the principal sigma subunit of RNA polymerase (112,113) (Supplementary Figure S6D). Expression of this operon is tightly regulated by at least seven promoters, and activated upon stress exposure (114). *YqxD* was shown to be dispensable for vegetative growth in *B. subtilis*, yet influenced timing of bacterial sporulation (115).

On the other hand, in some closely related Bacilli lineages (*B. cereus*, *B. anthracis*, *B. thuringiensis*), the corresponding gene can be found in the genomic context of GntR-family, MocR-type regulator (ARO8, COG1167; BC3039 in *B. cereus*). The MocR-like protein consists of an N-terminal winged HTH DNA-binding domain and a large C-terminal type I aminotransferase domain (116). The aminotransferase catalyzes the pyridoxal 5′-phosphate (PLP)-dependent reversible transfer of an amino group from the amino acid substrate to an acceptor α-keto acid (117). In analogy to another GntR-family regulator, *B. subtilis* GabR (118), binding of the aminotransferase effector may result in a conformational change that regulates DNA-binding affinity of the HTH domain. However, the functional coupling of DUF188 and COG1167, its similarity to TA systems and the effector of COG1167 still need to be determined.

COG2454. COG2454 encompasses two domains, i.e., an N-terminal α-helical DUF434 (PF04256) and, according to our predictions, a highly conserved C-terminal PIN-like domain. The consensus of secondary structure predictions obtained with Genesilico Metaserver (38) shows that the N-terminal region comprises at least four α-helices with several clusters of positively charged residues, suggesting involvement in substrate binding. Although the structure of the PIN-like domain is not known, both Meta-BASIC and HHpred provided reliable mappings to several FEN-like structures, e.g., FEN1 from *Sulfolobus solfataricus* (PDB ID: 2izo). Similarly to the FEN-like structures, COG2454 seems to contain a long helical insertion following S1, but as an NYN-type domain contains a short helix after S2 (Figure 3). The predicted active site includes at least five highly conserved Asp residues.

The phylogenetic distribution of COG2454 is wide, however, it indicates significant gaps in the classical tree of life. The representatives are rather scattered in many bacterial phyla (mostly in organisms from extreme environments, including Clostridia, Bacteroidetes, Spirochaetes), yet abundant in Crenarchaeota and Euryarchaeota branches of Archaea. While mostly monocistronic in bacterial genomes, in Archaea, typified by *S. solfataricus* P2 SSO0340, the COG2454 genes are located within polycistronic operons encoding proteins involved in translation, phosphate transport and metabolism, tRNA maturation, and rRNA/tRNA themselves (Supplementary Figure S6E). Regarding its genome neighborhood, this family is a strong candidate for an exonuclease involved in tRNA or rRNA maturation.

In Archaea, as in other domains of life, tRNA maturation requires removal of extra sequences at both, 5′- and 3′-ends. While 5′ cleavage is catalyzed by quasi-universal RNase P, the 3′ cleavage is more complex and can proceed via two alternative pathways: endonucleolytic or exonucleolytic. In endonucleolytic pathway, endonuclease RNase Z cleaves pre-tRNA directly downstream of the discriminator nucleotide. However, its activity strongly depends on the presence of the CCA motif, and this relationship depends on the tRNA 3′-trailers present in the organism (119). While in the majority of studied organisms, the activity of RNase Z is inhibited by the CCA motif, RNase Z from *T. maritima* can cleave CCA-containing pre-tRNAs after CCA (120). RNase Z from *P. aerophilus*, whose genome contains both CCA-less and CCA-containing tRNAs, processes exclusively CCA-containing pre-tRNAs *in vitro*. This inherency suggests additional enzymatic activities or cofactors for the tRNA 3′ maturation of CCA-less pre-tRNAs.

Given the strong substrate selectivity of RNase T involved in the exonucleolytic pathway, it has been suggested that another pathway of tRNA maturation must exist. In bacteria, the exonucleolytic pathway is performed via the action of six exoribonucleases discovered so far, among which RNase PH plays a significant role. Archaea encode

only one homolog, which mostly is RNase PH. In *B. subtilis*, it is difficult for RNase PH to remove the nucleotide immediately downstream of the CCA motif (121). Similarly to bacterial genomes, archaeal genomes with COG2454 contain a mixture of CCA-less and CCA-containing pre-tRNAs. Thus, one would expect these species also to possess two 3′-processing pathways. However, the presence of the exonucleolytic pathway is still an open question. It is unlikely that archaea encode only a single exonuclease; other exonucleases need to be discovered (122).

In *S. solfataricus* P2, SSO0340 is co-expressed with a gene encoding alanyl-tRNA synthetase and tRNA-Arg (123), similarly to *Pyrococcus horikoshii* Nob1, which is involved in rRNA maturation (124). Interestingly, the RNA-seq data indicates that the precursor tRNA-Arg transcript contains a 3′-trailer of ∼100 nt in length (123). According to the RNAfold (125) prediction, the precursor folds into a complex secondary structure that includes a stable hairpin right downstream of the 3′-end of the mature transcript. The only enzymes known to be involved in the removal of the tRNA 3′-trailer in Archaea are endonuclease RNase Z and exonuclease RNase PH; however, their substrate specificity in *S. solfataricus* remains unknown (122). Therefore, given the strong conservation of the potential nuclease active site (Table 1) and persistent appearance of the COG2454 genes in close proximity to the tRNA-associated operons in the genomes of Crenarchaeota, it is tempting to speculate that COG2454 may be involved in a new nucleolytic pathway of tRNA maturation.

Group 4: PRORP

Initially, the C-terminal domain of *Arabidopsis thaliana* At2g32230 was listed as a member of the NYN family (10). Later studies have shown that it is a representative of a large family of proteinaceous RNase P (PRORP) present in eukaryotic organelles and, as an RNA endonuclease, it removes the 5′-leader from precursor tRNA (126). However, analysis of known and predicted structures suggests that PRORP domains are distinct from NYN and, together with RNase_Zc3h12a (PF11977) and RNase_Zc3h12a_2 (PF14626), should be considered as a separate group (Figure 3, Supplementary Figure S2). Among those three families, crystal structures were solved for the PRORP (PDB ID: 4g23) and RNase_Zc3h12a (PDB ID: 3v33) families. Unlike NYN-like structures, they possess a short H2′ helix, corresponding to the helical arch of FEN-like nucleases (Figure 1). Moreover, both PRORP1 and ZC3H12A structures revealed unique bent helices between S4 and S5 and an extended core β-sheet (Supplementary Figure S4).

Both PRORP and ZC3H12A nuclease domains act as RNA endonucleases with substrate specificity achieved through additional RNA-binding domains (127,128). Interestingly, they are both fused with zinc-binding motifs. The crystal structure and sequence analysis revealed that PRORP nucleases contain an extended active site composed of 5–7 acidic residues, which can coordinate two metal ions (127,129). Similarly to the RNA-based RNase P, they were suggested to employ a two-metal-dependent catalytic mechanism (127). On the other hand, the crystal structure of ZC3H12A features only one bound metal ion, yet embedded in an expanded water network (128). The smallest family belonging to this group, RNase_Zc3h12a_2, comprises close homologs of *Ceanorhabditis elegans* eri-9 protein, yet, it is characterized by a variable number of predicted active site residues. The catalytic mechanism for the Zc3h12a-like nucleases needs to be studied.

RNase_Zc3h12a (PF11977). Nedd4-binding protein (N4BP1), which was used to define the NYN family by Anantharaman and Aravind (10), currently falls into the RNase_Zc3h12a definition. Together with six other proteins (Zc3h12a/MCPIP1, Zc3h12b/MCPIP2, Zc3h12c/MCPIP3, Zc3h12d/MCPIP4, NYNRIN/CGIN1, and KHNYN/KIAA0323), they constitute an abundant repertoire of human RNase_Zc3h12a domain-containing proteins. Out of them, MCPIP1 has been studied most extensively; it was shown to function in various cellular processes, including intracellular mRNA turnover (130), antiviral defense (131) and miRNA biogenesis (132), and its PIN-like domain acts as an endoribonuclease (70,128). Interestingly, N4BP1, CGIN1, and KHNYN proteins are probably of retroviral origin (133), however, their role in human cells still needs to be discovered.

Group 5: Mut7-C-like domains

Mut7-C (PF01927) clusters closely with three other families of unknown function, i.e., COG4634, PIN_10 and PIN_16. The structure of Mut7-C-like domains is unknown, however, according to our predictions, they are characterized by the most diverged and reduced structure. The alignment to other PIN-like sequences suggests that the Mut7-C-like structure lacks core elements H3 and S3, which are located at the borderline of the central β-sheet (Figure 3), thus, they might be classified as a distinct protein fold. However, the remaining secondary structure elements are apparently homologous to other PIN-like proteins in terms of sequence similarity, as well as localization of the active site residues. Likewise, the Mut7-C-like group should be considered as a shortened version of the PIN fold.

Mut7-C (PF01927). Distant homology of Mut7-C to the characterized PIN domains was first discovered using iterative sequence profile searches by Iyer *et al.* (11), where it was linked to the ubiquitin signaling system due to fusion to ubiquitin-like domains. The PIN-like domain contains a zinc-ribbon inserted at the C-terminus, characterized by four conserved Cys residues. Sequence alignment shows that only the first active site residue following S1 is conserved; the presence of the other residues and, therefore, nuclease activity, depends on the species.

In most eukaryotes, the Mut7-C domain is N-terminally fused with the 3′–5′ exonuclease RNase D family domain (DNA_pol_A_exo1, PF01612); in archaea and most bacteria, it acts as a standalone module, whereas in other bacteria, it is fused with ubiquitin-like domain of potential RNA-binding function (11). In some lower eukaryotes (e.g., Saprolegnia, Aphanomyces), besides the two domains, Mut7 proteins contain an aminoacyl-tRNA editing domain, which is involved in the tRNA editing of mischarged tRNAs. Mut7 protein in *C. elegans* is involved in

transposon silencing and RNA interference through its N-terminal 3′–5′ exonuclease domain (134). Its human homolog EXD3 (Uniprot ID: Q8N9H8) has not been studied; however, it was shown to be involved in cullin-RING ubiquitin ligase network via interaction with Cullin-1 (135). Interestingly, in some Euryarchaeota (e.g., *Haloferax volcanii*, HVO_074), the Mut7-C genes are located downstream to DNA-directed polymerase X, involved in DNA repair (Supplementary Figure S6F).

Reaction mechanism of the PIN-like nucleases

Catalyzed reactions (exo vs. endo). PIN-like nucleases are versatile nucleolytic catalysts, being exploited in central cellular processes. Despite possessing the conserved Rossmann-like fold, their molecular function varies in regard to several mechanical aspects, summarized in Table 2. At a general level, they can act as 5′–3′ exonucleases, cleaving single nucleotides from the 5′ end of a polynucleotide chain, or endonucleases, cleaving an internal phosphodiester bond. PIN-like nucleases typically display only one of the above activities. Members of two FEN-like families, classified in Pfam as internal domain of Xeroderma Pigmentosum Complementation Group G (XPG_I) and N-terminal domain of 5′–3′ exonucleases (5_3_exonuc_N), act both as exo- and endonucleases (136); however, as key players in DNA replication, they also have been most extensively studied, involving decades of research. It should be noted that in some studies the reported 5′–3′ exonuclease activity refers to the structure-specific endonuclease activity, with a requirement of free 5′ end (flap) (137). Although the N-terminal domain of 5′–3′ exonucleases (XRN_N) displays similar structural architecture (Supplementary Figure S4), its representatives have been shown to act only as 5′–3′ exonucleases (138). On the other hand, VapC- and NYN-like nucleases display only endoribonucleolytic activity, recognizing specific RNA secondary structures, and in some cases sequence motifs (139).

Two metal ion catalysis and active site. Despite many solved structures and biochemical studies, the precise catalytic mechanism of the PIN-like nucleases is still a subject of debate (140). Nevertheless, they are generally thought to utilize two metal ion cleavage mechanism, supported by Mg^{2+} or Mn^{2+} ions. In this model, the first ion ('ion B', according to the nomenclature proposed by Yang (1)) is buried deeply in the catalytic site and assists the 3′-O leaving group. The second ion ('ion A'), located on the nucleophile side, is bound more weakly (141). It is generally agreed that ion A plays a more important role (nucleophile formation) than ion B (transition state stabilization) (142). According to crystal structures and theoretical simulations, during the reaction the two metal ions are separated by 3.5–5.2 Å and positioned in line with the phospho-sugar backbone of the substrate (142–144).

Although several acidic residues were listed as critical for catalytic activity of PIN-like nucleases, the carboxylate located in the N-terminal part of the crossover helix H4 seems to play an exceptional role as it spans both ion binding pockets. Moreover, in many nucleases, it coordinates both ions directly, e.g., in VapC (145), EXO1 (68), PRORP (127), or participates in shaping the water network involved in the ion binding (7,28,141,146,147). Also, an invariant Asp located at the strand S1 is essential for catalysis. It is involved rather in ion A binding and barely coordinates it directly, as it is buried much deeper within the active site. The above two residues, together with carboxylate located at S2 (or at the additional helix immediately after S2 in VapC-like proteins, Figure 3), line ion A binding pocket, and in concert with Asp at S4 and adjacent structural elements shape the ion B binding site.

Variations in the active site composition within the PIN domain-like superfamily hinder establishment of a mechanism common to all PIN-like nucleases. The FEN-like nucleases are usually characterized by the most expanded active site of seven acidic residues (Table 2). They form two metal-binding sites (sites A and B), each coordinating one divalent metal ion (13). Much effort has been put to explain varied numbers of metal ions observed in the crystal structures and their interatomic distances exceeding that for a typical two metal ion mechanism (140). Mutagenesis studies have shown that one divalent metal ion is both necessary and sufficient for structure-specific endonuclease activity, whereas two divalent metal ions are required to support exonucleolytic cleavage (141). Additional kinetic analyses have shown that site B is involved in substrate binding rather than chemical catalysis (148) and triggers conformational changes (149). Moreover, as the position of ion B varies across different crystal structures and three metal ions were observed in *E. coli* PolA (PDB ID: 1taq), a mechanism involving three metal ions has been suggested (150).

In contrast, the active site of the canonical PIN-like nucleases is usually formed by 4–5 acidic residues corresponding to site A of the FEN-like enzymes, supplemented by Ser/Thr residues (9). As most structures of VapC-like nucleases have been solved with only one metal ion bound to the active site, their catalytic mechanism is still an open debate. Only recently, a mechanism in which only one metal ion is necessary has been supported by different experimental strategies (151). On the other hand, despite having a stripped-down active site, consisting of five acidic residues, VapC15 and PRORP1 were shown to coordinate two divalent metal ions in the crystal structures (127,145). The two metal ion catalysis in PRORP1 was further confirmed by examination of metal and pH dependence of the substrate cleavage (152).

Substrate specificity. The XPG-like (belonging to the XPG_I family) and PolA-like (5_3_exonuc_N) nucleases are generally recognized as structure-specific nucleases (Table 2). The structure-specific activity occurs at double strand-single strand junctions in bifurcated nucleic acid substrates such as flap, pseudo-Y and 5′-overhanging hairpin (65,153,154). The structure-specific 5′ nuclease activity of FEN-like domains is used commonly during DNA replication. For example, FEN1 removes RNA from Okazaki fragments, which are formed on the lagging strand (140), whereas DNA polymerase I cleaves primers or damaged nucleotides. Similar activity, specific towards DNA containing duplexes, bubbles or loops, is performed by XPG proteins and plays a central role in nucleotide excision repair (155). XPG proteins with severely impaired endonuclease activ-

Table 2. Functions and catalytic strategies of selected PIN-like domains

Enzyme	PIN-like cluster	Endonucleolytic activity	5′–3′ exonucleolytic activity	Biological function	Active site	Metal ions in tertiary structures
E. coli DNA polymerase I (polA)	5_3_exonuc_N	DNA, RNA and RNA-DNA (preferentially cleaves on the junction between a 5′ single-strand and duplex, i.e., 5′ flap) (65)	DNA (single-stranded (181) and double-stranded (182))	DNA replication: removal of the RNA primers from lagging strand fragments. DNA repair: mediation of the nick translation.	Two metal binding sites, A and B (7 x Asp/Glu: D13, D63, E113, D115, D116, D138, D140)	One Zn^{2+} bound in the crystal structure of *Thermus aquaticus* homolog (PDB ID: 1taq). Nuclease active with Mg^{2+} and Mn^{2+}, but not with Zn^{2+} or Ca^{2+} (65).
E. coli flap endonuclease Xni (ExoIX, ygdG)	5_3_exonuc_N	DNA (5′ flap and pseudo flap-like structures) (72)	— (183)	Unknown biological function. As mutations of this gene are synthetically lethal with those in polymerase I, the protein has been implied in Okazaki fragment maturation (184).	Metal binding site A (5 x Asp/Glu: D9, D50, E102, D104, D127)	Two Mg^{2+} 2.5 Å apart bound in the crystal structure of the complex with DNA (PDB ID: 3zd8). One K^{+} bound at an interface between the H3TH domain and DNA. Ca^{2+} has inhibitory effect (72).
Bacteriophage T4 RNase H (rnh)	5_3_exonuc_N	DNA (junction between a single-strand and duplex) (7,185)	DNA (dsDNA), RNA-DNA (7,185)	DNA replication: removal of the RNA primers from lagging strand fragments (186).	Sites A and B (7 x Asp/Glu: D19, D71, E130, D132, D155, D157, D200)	Two Mg^{2+} 7 Å apart bound in the crystal structure (PDB ID: 1tfr). No metal ions in the complex with fork DNA (PDB ID: 2ihn).
Human exonuclease 1 (EXO1)	XPG_I	DNA (5′ flap and pseudo flap-like structure-specific), RNA-DNA (RNA primer removal from Okazaki fragments) (187,188)	DNA (dsDNA, low activity on ssDNA) (187,188)	DNA replication: removal of the RNA primers from lagging strand fragments (187,188).	Sites A and B (7 x Asp/Glu: D30, D78, E150, D152, D171, D173, D225)	Two Mn^{2+} 4.1 Å apart bound in the crystal structure of the complex with DNA (PDB ID: 3qeb).
Human flap endonuclease 1 (FEN1)	XPG_I	DNA (5′ flap and pseudo flap-like structure-specific, gapped DNA duplex, not ssDNA and dsDNA) (153,189)	DNA (nicked or gapped dsDNA (190), but not ssDNA (191))	DNA replication: removal of the RNA primers from lagging strand fragments, resolution of stalled DNA replication forks. DNA repair: long-patch base excision repair (192).	Sites A and B (7 x Asp/Glu: D34, D86, E158, E160, D179, D181, D233)	Two Mg^{2+} 3.4 Å apart bound in the crystal structure of the complex with PCNA (PDB ID: 1ul1).
Human gap endonuclease 1 (GEN1)	XPG_I	DNA (5′ flap, replication fork, Holliday junction) (193)	— (194)	Homologous recombination: Holliday junction resolution (195).	Sites A and B (7 x Asp/Glu: D30, E75, E134, E136, D155, D157, D208)	One Mg^{2+} bound in the crystal structure of the complex with DNA (PDB ID: 5t9j).
Human DNA repair protein complementing XP-G cells (ERCC5)	XPG_I	DNA (single-stranded structure-specific, including bubble and splayed arm substrates) (189)	?	DNA repair: nucleotide excision repair (NER) (196).	Sites A and B (7 x Asp/Glu: D30, D77, E789, E791, D810, D812, D861)	—
Virion host shutoff protein (UL41)	XPG_I	RNA (mRNA) (157)	?	Decay of host mRNAs (197).	Sites A and B (7 x Asp/Glu: D34, D82, E192, D194, D213, D125, D261) (198)	—
Human exoribonuclease 1 (XRN1)	XRN_N	?	RNA (5′ monophosphorylated single-stranded or duplex substrates (158), mouse (199) and yeast (200) homologs postulated to cleave G4-tetraplex substrates), DNA (yeast homolog cleaves single-stranded DNA (201))	RNA decay: major 5'–3' exoribonuclease in mRNA decay (158). rRNA maturation in yeast (159).	Sites A and B (7 x Asp/Glu: D35, D86, E176, E178, D206, D208, D292)	One Mg^{2+} bound in the crystal structure of *Drosophila melanogaster* homolog (PDB ID: 2y35). One Mn^{2+} bound in the crystal structure of *Kluyveromyces lactis* homolog (PDB ID: 3pif).
S. pombe Rat1/Xrn2	XRN_N	?	RNA (5′ monophosphorylated single-stranded substrates (202))	Transcription termination (203).	Sites A and B (7 x Asp/Glu: D55, D104, E205, E207, D235, D237, D336)	One Mg^{2+} bound in the crystal structure of the complex with Rai1 (PDB ID: 3fqd).
A. thaliana PRORP1	PRORP	RNA (tRNA or tRNA-like structures) (12)	?	tRNA maturation: 5′ maturation of tRNA precursors (126).	Site A (5 x Asp: D399, D493, D497, D474, D475) (204)	Two Mn^{2+} bound in the crystal structure (PDB ID: 4g24).
Human endoribonuclease ZC3H12A (MCPIP1)	RNase_Zc3h12a	RNA (preferentially cleaves a stem-loop structure) (132)	?	mRNA decay (130), miRNA biogenesis regulation (132), viral infection.	Site A (5 x Asp: D141, D225, D226, D244, D248) (128)	One Mg^{2+} bound in the crystal structure (PDB ID: 3v33). Nuclease active with Mg^{2+} and Mn^{2+}, but not with Fe^{2+}, Zn^{2+} or Ca^{2+} (128).
S. cerevisiae Nob1	PIN_6	RNA (single-stranded region of a hairpin, i.e. site D of 18S rRNA) (175)	?	rRNA maturation (175).	Site A (4 x Asp/Glu: D15, E43, D92, D110; but D110 is not essential for function) (124)	NMR structure of *P. horikoshii* homolog (PDB ID: 2lcq). Nuclease active with Mn^{2+}, but not with Mg^{2+} (in *P. horikoshii* homolog).

Table 2. Continued

Enzyme	PIN-like cluster	Endonucleolytic activity	5′–3′ exonucleolytic activity	Biological function	Active site	Metal ions in tertiary structures
Human telomerase-binding protein EST1A (SMG6)	PIN_4.COG1875	RNA (single-stranded, not double-stranded, preferentially cleaves within a degenerate pentameric motif (163)), not single-stranded DNA (161)	?	Nonsense-mediated mRNA decay (205).	Site A (4 x Asp/Glu: D1251, E1282, D1353, D1392)	No metal ions in the crystal structure (PDB ID: 2hww). Nuclease active with Mn^{2+} and, to a much lesser extent, Mg^{2+} (161).
Human exosome complex exonuclease RRP44 (DIS3)	Rrp44	RNA (single-stranded, preferentially 5′ monophosphorylated, as shown for yeast homolog) (162,206)	?	Exosome-mediated mRNA decay (206).	Site A (4 x Asp/Glu: D69, E97, D146, D177)	No metal ions in the crystal structure of the exosome complex (PDB ID: 4ifd). Nuclease active with Mn^{2+}, Mg^{2+} and Zn^{2+}.
Yeast rRNA-processing protein UTP24	Fcf1	RNA (sequence-specific, cleaves sites A1 and A2 of 18S pre-rRNA) (170)	?	rRNA maturation (23).	Site A (4 x Asp/Glu: D68, E105, D139, D157)	No metal ions in the PIN-like domain in the crystal structure (PDB ID: 4mj7). Nuclease active with Mn^{2+} and Mg^{2+}.
M. tuberculosis VapC15	PIN.COG1487	RNA (cleaves $tRNA3^{Leu\text{-}CAG}$) (139)	?	Toxin-antitoxin (with VapB15 as an antitoxin) (145).	Site A (5 x Asp/Glu: D4, E42, D96, D114, D116).	Mn^{2+}-Mg^{2+} pair bound in the crystal structure of the heterotrimeric complex (VapBC2) with antitoxin (PDB ID: 4chg). Both metals are shared by the toxin-antitoxin pair.
M. tuberculosis VapC5	PIN.COG1487	RNA (low activity on double-stranded RNA), no activity on dsDNA (28)	?	Potential toxin-antitoxin (with VapB5 as an antitoxin).	Site A (4 x Asp/Glu: D26, E57, D115, D135)	No metal ions bound in the crystal structure of the complex with antitoxin (PDB ID: 3dbo). Nuclease active with Mg^{2+}.
H. influenzae ribonuclease VapC1	PIN.COG1487	RNA (structure- and sequence-specific, cleaves initiator tRNA between the anticodon stem and loop, but does not cleave mRNA, rRNA or tmRNA), no activity on ssDNA or dsDNA (207,208)	?	Toxin-antitoxin: inhibition of translation initiation and translation activation at elongated codons (15).	Site A (4 x Asp/Glu/Asn: D6, E43, D99, E120 or N117 — polar residue required at this position (209))	Crystal structure not available, only a 3D model (209).
R. felis VapC2	PIN.COG1487	RNA (cleaves single-stranded RNA) (210)	?	Toxin-antitoxin: pathogenesis (210).	Site A (4 x Asp/Glu: D6, E43, D99, E120)	No metal ions in the crystal structure of the complex with VapB2 and its promoter DNA (PDB ID: 3zvk).
P. aerophilum PAE2754	PIN.COG4113	RNA (cleaves single-stranded, G-rich RNA) (160)	?	Potential toxin-antitoxin (with PAE2755 as an antitoxin).	Site A (4 x Asp/Glu: D8, E39, D92, D1180) (211)	No metal ions in the crystal structure of the dimer (PDB ID: 1v8p). Nuclease active with Mn^{2+} and Mg^{2+}.

Further description of the terms used in the table can be found in the main text.

ity are implicated as a background for xeroderma pigmentosum (156). A close XPG homolog from herpes simplex virus 1, UL41 protein (virion host shutoff, vhs), selectively degrades mRNA by endonucleolytic cleavage early in infection (157).

Despite sharing similar structural topology with the structure-specific endonucleases (Supplementary Figure S4), the XRN_N family members were shown to exhibit only 5′–3′ exonucleolytic activity. Both XRN1 and XRN2 proteins, founders of the XRN_N family, preferentially cleave 5′ monophosphorylated RNA. Xrn1p is the main enzyme degrading decapped mRNA in multiple decay pathways in yeast (158), and together with Xrn2p (Rat1) plays role in rRNA maturation (159). The 5′ phosphate is thought to be recognized by a basic pocket in the PIN-like domain formed by four highly conserved residues, from which larger 5′ groups are sterically excluded (PDB ID: 2y35) (66).

Enzymes characterized by a canonical version of the PIN domain-like fold usually act as endoribonucleases, recognizing and cleaving specific structures and/or sequences. Toxic behaviour of bacterial VapC nucleases usually relies on cleavage of RNAs essential for translation, i.e., tRNA or rRNA. Recent large-scale study on cellular targets for 12 VapC toxins in *M. tuberculosis* showed that VapCs are highly target-specific: 11 cleave specific tRNAs, and one recognizes sarcin–ricin loop of 23S rRNA (139). Moreover, the substrate specificity is reflected by the phylogenetic relatedness of the proteins (139). Another systematic study on four VapCs from *P. aerophilum* and *M. tuberculosis* revealed preference of the toxins for G- and GC-rich 4-mer RNA sequences (160). Various eukaryotic homologs of VapC toxins are involved in rRNA maturation (Nob1, Utp23, Fcf1/Utp24), recognizing specific sites in pre-rRNA. Mutagenesis studies for an archaeal Nob1 homolog revealed two residues responsible for specific degradation of the RNA substrate (21). D100 from the active site and R115 were proposed to play role in correct positioning of the substrate with respect to the catalytic center. A question remains whether similar mechanism is employed

by other VapC-like nucleases and how it relates to the proposed two metal ion catalysis.

Besides XRN_N representatives, also proteins characterized by a stripped-down PIN domain-like fold are involved in mRNA decay. ZC3H12A, DIS3 and SMG6 function in large protein complexes and cut single-stranded RNA with various specificities (Table 2). ZC3H12A preferentially cleaves a stem-loop structure within mRNA 3′ UTRs and miRNAs (161), DIS3 recognizes 5′ monophosphorylated ssRNA (162), whereas SMG6 favorably degrades a degenerate pentameric sequence motif (163). However, the mechanism underlying their substrate specificity, including preference for a hydroxyl group at the 2′ position of ribose in the nucleic acid substrate, awaits further studies.

CONCLUSIONS

In this first comprehensive study of the PIN domain-like superfamily, we systematically identified all its sequence representatives, and clustered them according to similarity of the corresponding domains. The Pfam release 30.0 comprised 18 families classified into the PIN (CL0280) clan. Transitive sequence searches led to identification of several new families, including representants of Pfam (DUF1308 (PF07000), DUF4935 (PF16289)) and CDD (COG2454), and 23 other families not classified in these databases. The systematic sequence clustering revealed relationships between individual sequence groups and showed heterogeneity within some families, suggesting the possible function divergence. With the 70 defined clusters, over 100,000 identified proteins, and broad biological functions, the PIN domain-like domain superfamily constitutes one of the largest and most diverse nuclease superfamilies. Based on the high-quality structure-based multiple sequence alignment of their representatives, we predicted nuclease active sites as well as insertions to the structural core, and grouped the clusters into five major structural classes. Detailed analyses of the protein domain architecture, genome context and structure modeling allowed us to predict biological functions of several new families, including new toxin-antitoxin components, proteins potentially involved in tRNA/rRNA maturation and transcription/translation regulation.

Up to date, considerable effort has been made to characterize structures and catalytic mechanisms of the PIN-like nucleases. However, our knowledge about structure and functions of PIN-like domains is largely biased towards VapC and FEN-like domains, which—being considerably distinct—do not provide insights into the roles of individual structural elements and subtle local sequence differences in the substrate specificity. It would be of great interest to study in detail how the structural insertions influence substrate recognition and catalytic mechanism of the nucleases, in particular in the least studied groups, i.e., NYN, PRORP, and Mut7-C. Also, an unsolved question remains: What is the catalytic mechanism of the nucleases (mainly from the VapC-like group), whose structures were solved in the presence of only one metal ion?

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ivan Shabalin and Michal Lazniewski for careful reading of the manuscript. The authors would also like to thank the anonymous reviewers for their insightful suggestions.

FUNDING

Foundation for Polish Science [TEAM to K.G.]; Polish National Science Centre [2011/02/A/NZ2/00014 and 2014/15/B/NZ1/03357 to K.G.]. Funding for open access charge: Polish National Science Centre [2014/15/B/NZ1/03357].

Conflict of interest statement. None declared.

REFERENCES

1. Yang,W. (2011) Nucleases: diversity of structure, function and mechanism. *Q. Rev. Biophys.*, **44**, 1–93.
2. Dunin-Horkawicz,S., Feder,M. and Bujnicki,J.M. (2006) Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC Genomics*, **7**, 98.
3. Steczkiewicz,K., Muszewska,A., Knizewski,L., Rychlewski,L. and Ginalski,K. (2012) Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res.*, **40**, 7016–7045.
4. Majorek,K.a., Dunin-Horkawicz,S., Steczkiewicz,K., Muszewska,A., Nowotny,M., Ginalski,K. and Bujnicki,J.M. (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.*, **42**, 4160–4179.
5. Wall,D. and Kaiser,D. (1999) Type IV pili and cell motility. *Mol. Microbiol.*, **32**, 1–10.
6. Bunker,R.D., McKenzie,J.L., Baker,E.N. and Arcus,V.L. (2008) Crystal structure of PAE0151 from Pyrobaculum aerophilum, a PIN-domain (VapC) protein from a toxin-antitoxin operon. *Proteins*, **72**, 510–518.
7. Mueser,T.C., Nossal,N.G. and Hyde,C.C. (1996) Structure of bacteriophage T4 RNase H, a 5' to 3' RNA-DNA and DNA-DNA exonuclease with sequence similarity to the RAD2 family of eukaryotic proteins. *Cell*, **85**, 1101–1112.
8. Garforth,S.J., Patel,D., Feng,M. and Sayers,J.R. (2001) Unusually wide co-factor tolerance in a metalloenzyme; divalent metal ions modulate endo-exonuclease activity in T5 exonuclease. *Nucleic Acids Res.*, **29**, 2772–2779.
9. Clissold,P.M. and Ponting,C.P. (2000) PIN domains in nonsense-mediated mRNA decay and RNAi. *Curr. Biol.*, **10**, 888–890.
10. Anantharaman,V. and Aravind,L. (2006) The NYN domains. *RNA Biol.*, **3**, 18–27.
11. Iyer,L.M., Burroughs,a.M. and Aravind,L. (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biol.*, **7**, R60.
12. Gobert,A., Gutmann,B., Taschner,A., Gössringer,M., Holzmann,J., Hartmann,R.K., Rossmanith,W. and Giegé,P. (2010) A single Arabidopsis organellar protein has RNase P activity. *Nat. Struct. Mol. Biol.*, **17**, 740–744.
13. Tsutakawa,S.E., Classen,S., Chapados,B.R., Arvai,A.S., Finger,L.D., Guenther,G., Tomlinson,C.G., Thompson,P., Sarker,A.H., Shen,B. *et al.* (2011) Human flap endonuclease structures, DNA double-base flipping, and a unified understanding of the FEN1 superfamily. *Cell*, **145**, 198–211.
14. McKenzie,J.L., Robson,J., Berney,M., Smith,T.C., Ruthe,A., Gardner,P.P., Arcus,V.L. and Cook,G.M. (2012) A VapBC toxin-antitoxin module is a posttranscriptional regulator of metabolic flux in mycobacteria. *J. Bacteriol.*, **194**, 2189–2204.
15. Winther,K.S. and Gerdes,K. (2011) Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 7403–7407.

16. Winther,K.S., Brodersen,D.E., Brown,A.K. and Gerdes,K. (2013) VapC20 of Mycobacterium tuberculosis cleaves the sarcin-ricin loop of 23S rRNA. *Nat. Commun.*, **4**, 2796.
17. Cruz,J.W. and Woychik,N.A. (2016) tRNAs taking charge. *Pathog. Dis.*, **74**, ftv117.
18. Cruz,J.W., Sharp,J.D., Hoffer,E.D., Maehigashi,T., Vvedenskaya,I.O., Konkimalla,A., Husson,R.N., Nickels,B.E., Dunham,C.M. and Woychik,N.A. (2015) Growth-regulating Mycobacterium tuberculosis VapC-mt4 toxin is an isoacceptor-specific tRNase. *Nat. Commun.*, **6**, 7480.
19. Huntzinger,E., Kashima,I., Fauser,M., Sauliere,J. and Izaurralde,E. (2008) SMG6 is the catalytic endonuclease that cleaves mRNAs containing nonsense codons in metazoan. *RNA (N. Y.)*, **14**, 2609–2617.
20. Schneider,C., Leung,E., Brown,J. and Tollervey,D. (2009) The N-terminal PIN domain of the exosome subunit Rrp44 harbors endonuclease activity and tethers Rrp44 to the yeast core exosome. *Nucleic Acids Res.*, 1–14.
21. Fatica,A., Tollervey,D. and Dlakić,M. (2004) PIN domain of Nob1p is required for D-site cleavage in 20S pre-rRNA. *RNA (N. Y.)*, **10**, 1698–1701.
22. Lu,J., Sun,M. and Ye,K. (2013) Structural and functional analysis of Utp23, a yeast ribosome synthesis factor with degenerate PIN domain. *RNA (N. Y.)*, **19**, 1815–1824.
23. Bleichert,F., Granneman,S., Osheim,Y.N., Beyer,A.L. and Baserga,S.J. (2006) The PINc domain protein Utp24, a putative nuclease, is required for the early cleavage steps in 18S rRNA maturation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 9464–9469.
24. Arcus,V.L., McKenzie,J.L., Robson,J. and Cook,G.M. (2011) The PIN-domain ribonucleases and the prokaryotic VapBC toxin-antitoxin array. *Protein Eng. Des. Sel.*, **24**, 33–40.
25. Jensen,R.B. and Gerdes,K. (1995) Programmed cell death in bacteria: proteic plasmid stabilization systems. *Mol. Microbiol.*, **17**, 205–210.
26. Gerdes,K., Christensen,S.K. and Løbner-Olesen,A. (2005) Prokaryotic toxin-antitoxin stress response loci. *Nat. Rev. Microbiol.*, **3**, 371–382.
27. Pandey,D.P. and Gerdes,K. (2005) Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res.*, **33**, 966–976.
28. Miallau,L., Faller,M., Chiang,J., Arbing,M., Guo,F., Cascio,D. and Eisenberg,D. (2009) Structure and proposed activity of a member of the VapBC family of toxin-antitoxin systems. VapBC-5 from Mycobacterium tuberculosis. *J. Biol. Chem.*, **284**, 276–283.
29. Arcus,V.L., Rainey,P.B. and Turner,S.J. (2005) The PIN-domain toxin-antitoxin array in mycobacteria. *Trends Microbiol.*, **13**, 360–365.
30. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
31. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
32. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
33. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
34. Ginalski,K., von Grotthuss,M., Grishin,N.V. and Rychlewski,L. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.*, **32**, W576–W581.
35. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
36. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
37. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
38. Kurowski,M.A. and Bujnicki,J.M. (2003) GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.*, **31**, 3305–3307.
39. Soding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
40. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
41. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
42. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
43. Jaroszewski,L., Koska,L., Sedova,M. and Godzik,A. (2014) PubServer: literature searches by homology. *Nucleic Acids Res.*, **42**, W430–W435.
44. Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
45. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
46. Remmert,M., Biegert,A., Hauser,A. and Söding,J. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
47. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
48. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
49. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
50. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
51. Ginalski,K. and Rychlewski,L. (2003) Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins*, **53** (Suppl. 6), 410–417.
52. Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.
53. Krissinel,E. (2012) Enhanced fold recognition using efficient short fragment clustering. *J. Mol. Biochem.*, **1**, 76–85.
54. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G., McCoy,A. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.
55. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
56. Okuda,S. and Yoshizawa,A.C. (2011) ODB: a database for operon organizations, 2011 update. *Nucleic Acids Res.*, **39**, D552–D555.
57. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
58. Kuchta,K., Knizewski,L., Wyrwicz,L.S., Rychlewski,L. and Ginalski,K. (2009) Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. *Nucleic Acids Res.*, **37**, 7701–7714.
59. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2009) Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol. Direct*, **4**, 19.
60. Senkevich,T.G., Koonin,E.V. and Moss,B. (2009) Predicted poxvirus FEN1-like nuclease required for homologous recombination, double-strand break repair and full-size genome formation. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 17921–17926.
61. Iyer,L.M., Balaji,S., Koonin,E.V. and Aravind,L. (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.*, **117**, 156–184.

62. Da Silva,M., Shen,L., Tcherepanov,V., Watson,C. and Upton,C. (2006) Predicted function of the vaccinia virus G5R protein. *Bioinformatics*, **22**, 2846–2850.
63. Rao,S.T. and Rossmann,M.G. (1973) Comparison of super-secondary structures in proteins. *J. Mol. Biol.*, **76**, 241–256.
64. Youngsoo,K. and Eom,S.H. (1995) Crystal structure of Thermus aquaticus DNA polymerase. *Nature*, **376**, 612–616.
65. Lyamichev,V., Brow,M.A. and Dahlberg,J.E. (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science*, **260**, 778–783.
66. Jinek,M., Coyle,S.M. and Doudna,J.A. (2011) Coupled 5' nucleotide recognition and processivity in Xrn1-mediated mRNA decay. *Mol. Cell*, **41**, 600–608.
67. Garforth,S.J., Ceska,T.A., Suck,D. and Sayers,J.R. (1999) Mutagenesis of conserved lysine residues in bacteriophage T5 5'-3' exonuclease suggests separate mechanisms of endo-and exonucleolytic cleavage. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 38–43.
68. Orans,J., McSweeney,E.A., Iyer,R.R., Hast,M.a., Hellinga,H.W., Modrich,P. and Beese,L.S. (2011) Structures of human exonuclease 1 DNA complexes suggest a unified mechanism for nuclease family. *Cell*, **145**, 212–223.
69. Patel,N., Atack,J.M., Finger,L.D., Exell,J.C., Thompson,P., Tsutakawa,S., Tainer,J.A., Williams,D.M. and Grasby,J.A. (2012) Flap endonucleases pass 5'-flaps through a flexible arch using a disorder-thread-order mechanism to confer specificity for free 5'-ends. *Nucleic Acids Res.*, **40**, 4507–4519.
70. Matsushita,K., Takeuchi,O., Standley,D.M., Kumagai,Y., Kawagoe,T., Miyake,T., Satoh,T., Kato,H., Tsujimura,T., Nakamura,H. *et al.* (2009) Zc3h12a is an RNase essential for controlling immune responses by regulating mRNA decay. *Nature*, **458**, 1185–1190.
71. Hosfield,D.J., Mol,C.D., Shen,B. and Tainer,J.A. (1998) Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity. *Cell*, **95**, 135–146.
72. Anstey-Gilbert,C.S., Hemsworth,G.R., Flemming,C.S., Hodskinson,M.R.G., Zhang,J., Sedelnikova,S.E., Stillman,T.J., Sayers,J.R. and Artymiuk,P.J. (2013) The structure of Escherichia coli ExoIX–implications for DNA binding and catalysis in flap endonucleases. *Nucleic Acids Res.*, **41**, 8357–8367.
73. Mattison,K., Wilbur,J.S., So,M. and Brennan,R.G. (2006) Structure of FitAB from Neisseria gonorrhoeae bound to DNA reveals a tetramer of toxin-antitoxin heterodimers containing pin domains and ribbon-helix-helix motifs. *J. Biol. Chem.*, **281**, 37942–37951.
74. Dienemann,C., Bøggild,A., Winther,K.S., Gerdes,K. and Brodersen,D.E. (2011) Crystal structure of the VapBC toxin-antitoxin complex from Shigella flexneri reveals a hetero-octameric DNA-binding assembly. *J. Mol. Biol.*, **414**, 713–722.
75. Wilbur,J.S., Chivers,P.T., Mattison,K., Potter,L., Brennan,R.G. and So,M. (2005) Neisseria gonorrhoeae FitA interacts with FitB to bind DNA through its ribbon-helix-helix motif. *Biochemistry*, **44**, 12515–12524.
76. Coles,M., Djuranovic,S., Söding,J., Frickey,T., Koretke,K., Truffault,V., Martin,J. and Lupas,A.N. (2005) AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. *Structure*, **13**, 919–928.
77. Anantharaman,V. and Aravind,L. (2003) New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. *Genome Biol.*, **12**, R81.
78. Vetting,M.W., Luiz,L.P., Yu,M., Hegde,S.S., Magnet,S., Roderick,S.L. and Blanchard,J.S. (2005) Structure and functions of the GNAT superfamily of acetyltransferases. *Arch. Biochem. Biophys.*, **433**, 212–226.
79. Iyer,L.M., Zhang,D., Burroughs,A.M. and Aravind,L. (2013) Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res.*, **41**, 7635–7655.
80. Neuwald,A.F. and Landsman,D. (1997) GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. *Trends Biochem. Sci.*, **22**, 154–155.
81. Pérez-Arellano,I., Gallego,J. and Cervera,J. (2007) The PUA domain - a structural and functional overview. *FEBS Journal*, **274**, 4972–4984.
82. Iyer,L.M., Burroughs,a.M. and Aravind,L. (2006) The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism. *Bioinformatics*, **22**, 257–263.
83. Jurenaite,M., Markuckas,A. and Sužiedeliene,E. (2013) Identification and characterization of type II toxin-antitoxin systems in the opportunistic pathogen acinetobacter baumannii. *J. Bacteriol.*, **195**, 3165–3172.
84. Ludanyi,M., Blanchard,L., Dulermo,R., Brandelet,G., Bellanger,L., Pignol,D., Lemaire,D. and de Groot,A. (2014) Radiation response in Deinococcus deserti: IrrE is a metalloprotease that cleaves repressor protein DdrO. *Mol. Microbiol.*, **94**, 434–449.
85. Bose,B., Auchtung,J.M., Lee,C.A. and Grossman,A.D. (2008) A conserved anti-repressor controls horizontal gene transfer by proteolysis. *Mol. Microbiol.*, **70**, 570–582.
86. Gerstmeir,R., Cramer,A., Dangel,P., Schaffer,S. and Eikmanns,B.J. (2004) RamB, a novel transcriptional regulator of genes involved in acetate metabolism of Corynebacterium glutamicum. *J. Bacteriol.*, **186**, 2798–2809.
87. Masiewicz,P., Brzostek,A., Wolanski,M., Dziadek,J. and Zakrzewska-Czerwinska,J. (2012) A novel role of the PrpR as a transcription factor involved in the regulation of methylcitrate pathway in Mycobacterium tuberculosis. *PloS One*, **7**, e43651.
88. Vujicic-Zagar,A., Dulermo,R., Le Gorrec,M., Vannier,F., Servant,P., Sommer,S., de Groot,A. and Serre,L. (2009) Crystal structure of the IrrE protein, a central regulator of DNA damage repair in deinococcaceae. *J. Mol. Biol.*, **386**, 704–716.
89. Bose,B., Auchtung,J.M., Lee,C.A. and Grossman,A.D. (2008) A conserved anti-repressor controls horizontal gene transfer by proteolysis. *Mol. Microbiol.*, **70**, 570–582.
90. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
91. Kobayashi,I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.*, **29**, 3742–3756.
92. Mruk,I. and Kobayashi,I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.*, **42**, 70–86.
93. Rolland,T., Tasan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
94. Buchko,G.W., Phan,I., Myler,P.J., Terwilliger,T.C. and Kim,C.Y. (2011) Inaugural structure from the DUF3349 superfamily of proteins, Mycobacterium tuberculosis Rv0543c. *Arch. Biochem. Biophys.*, **506**, 150–156.
95. Brannigan,J.A., Dodson,G., Duggleby,H.J., Moody,P.C., Smith,J.L., Tomchick,D.R. and Murzin,A.G. (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature*, **378**, 416–419.
96. Stoll,B. and Binder,S. (2016) Two NYN domain containing putative nucleases are involved in transcript maturation in Arabidopsis mitochondria. *Plant J.*, **85**, 278–288.
97. Taniguchi,Y., Katayama,M., Ito,R., Takai,N., Kondo,T. and Oyama,T. (2007) labA: a novel gene required for negative feedback regulation of the cyanobacterial circadian clock protein KaiC. *Genes Dev.*, **21**, 60–70.
98. Taniguchi,Y., Nishikawa,T., Kondo,T. and Oyama,T. (2012) Overexpression of lalA, a paralog of labA, is capable of affecting both circadian gene expression and cell growth in the cyanobacterium Synechococcus elongatus PCC 7942. *FEBS Lett.*, **586**, 753–759.
99. Tang,H., Wang,S., Ma,L., Meng,X., Deng,Z., Zhang,D., Ma,C. and Xu,P. (2008) A novel gene, encoding 6-hydroxy-3-succinoylpyridine hydroxylase, involved in nicotine degradation by Pseudomonas putida strain S16. *Appl. Environ. Microbiol.*, **74**, 1567–1574.
100. Wang,M., Yang,G., Min,H. and Lv,Z. (2009) A novel nicotine catabolic plasmid pMH1 in Pseudomonas sp. strain HF-1. *Can. J. Microbiol.*, **55**, 228–233.

101. Zhang,D., Burroughs,A.M., Vidal,N.D., Iyer,L.M. and Aravind,L. (2016) Transposons to toxins: the provenance, architecture and diversification of a widespread class of eukaryotic effectors. *Nucleic Acids Res.*, **44**, 3513–3533.
102. Anantharaman,V., Zhang,D. and Aravind,L. (2010) OST-HTH: a novel predicted RNA-binding domain. *Biol. Direct*, **5**, 13.
103. Bloch,D.B., Li,P., Bloch,E.G., Berenson,D.F., Galdos,R.L., Arora,P., Malhotra,R., Wu,C. and Yang,W. (2014) LMKB/MARF1 localizes to mRNA processing bodies, interacts with Ge-1, and regulates IFI44L gene expression. *PLoS One*, **9**, e94784.
104. Su,Y.-Q., Sun,F., Handel,M.a., Schimenti,J.C. and Eppig,J.J. (2012) Meiosis arrest female 1 (MARF1) has nuage-like function in mammalian oocytes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 18653–18660.
105. Tullius,M.V., Harmston,C.A., Owens,C.P., Chim,N., Morse,R.P., McMath,L.M., Iniguez,A., Kimmey,J.M., Sawaya,M.R., Whitelegge,J.P. *et al.* (2011) Discovery and characterization of a unique mycobacterial heme acquisition system. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5051–5056.
106. Redko,Y., Bechhofer,D.H. and Condon,C. (2008) Mini-III, an unusual member of the RNase III family of enzymes, catalyses 23S ribosomal RNA maturation in B. subtilis. *Mol. Microbiol.*, **68**, 1096–1106.
107. Toledo-Arana,A., Dussurget,O., Nikitas,G., Sesto,N., Guet-Revillet,H., Balestrino,D., Loh,E., Gripenland,J., Tiensuu,T., Vaitkevicius,K. *et al.* (2009) The Listeria transcriptional landscape from saprophytism to virulence. *Nature*, **459**, 950–956.
108. Dönhöfer,A., Franckenberg,S., Wickles,S., Berninghausen,O., Beckmann,R. and Wilson,D.N. (2012) Structural basis for TetM-mediated tetracycline resistance. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 16900–16905.
109. Feng,Y., Wu,H., Xu,Y., Zhang,Z., Liu,T., Lin,X. and Feng,X.-H. (2014) Zinc finger protein 451 is a novel Smad corepressor in transforming growth factor-β signaling. *J. Biol. Chem.*, **289**, 2072–2083.
110. Yokogawa,M., Tsushima,T., Noda,N.N., Kumeta,H., Enokizono,Y., Yamashita,K., Standley,D.M., Takeuchi,O., Akira,S. and Inagaki,F. (2016) Structural basis for the regulation of enzymatic activity of Regnase-1 by domain-domain interactions. *Sci. Rep.*, **6**, 22324.
111. Cho,B.K., Zengler,K., Qiu,Y., Park,Y.S., Knight,E.M., Barrett,C.L., Gao,Y. and Palsson,B.O. (2009) The transcription unit architecture of the Escherichia coli genome. *Nat. Biotechnol.*, **27**, 1043–1049.
112. Wang,L.F. and Doi,R.H. (1986) Nucleotide sequence and organization of Bacillus subtilis RNA polymerase major sigma (sigma 43) operon. *Nucleic Acids Res.*, **14**, 4293–4307.
113. Metzger,R., Brown,D.P., Grealish,P., Staver,M.J., Versalovic,J., Lupski,J.R. and Katz,L. (1994) Characterization of the macromolecular synthesis (MMS) operon from Listeria monocytogenes. *Gene*, **151**, 161–166.
114. Liao,C.T., Wen,Y.D., Wang,W.H. and Chang,B.Y. (1999) Identification and characterization of a stress-responsive promoter in the macromolecular synthesis operon of Bacillus subtilis. *Mol. Microbiol.*, **33**, 377–388.
115. Zuberi,A.R. and Doi,R.H. (1990) A mutation in P23, the first gene in the RNA polymerase sigma A (sigma 43) operon, affects sporulation in Bacillus subtilis. *J. Bacteriol.*, **172**, 2175–2177.
116. Rigali,S., Derouaux,A., Giannotta,F. and Dusart,J. (2002) Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J. Biol. Chem.*, **277**, 12507–12515.
117. Sung,M.H., Tanizawa,K., Tanaka,H., Kuramitsu,S., Kagamiyama,H., Hirotsu,K., Okamoto,A., Higuchi,T. and Soda,K. (1991) Thermostable aspartate aminotransferase from a thermophilic Bacillus species. Gene cloning, sequence determination, and preliminary x-ray characterization. *J. Biol. Chem.*, **266**, 2567–2572.
118. Okuda,K., Ito,T., Goto,M., Takenaka,T., Hemmi,H. and Yoshimura,T. (2015) Domain characterization of Bacillus subtilis GabR, a pyridoxal 5'-phosphate-dependent transcriptional regulator. *J. Biochem.*, **158**, 225–234.
119. Nashimoto,M. (1997) Distribution of both lengths and 5' terminal nucleotides of mammalian pre-tRNA 3' trailers reflects properties of 3' processing endoribonuclease. *Nucleic Acids Res.*, **25**, 1148–1154.
120. Minagawa,A., Takaku,H., Takagi,M. and Nashimoto,M. (2004) A novel endonucleolytic mechanism to generate the CCA 3' termini of tRNA Molecules in Thermotoga maritima. *J. Biol. Chem.*, **279**, 15688–15697.
121. Wen,T., Oussenko,I.A., Pellegrini,O., Bechhofer,D.H. and Condon,C. (2005) Ribonuclease PH plays a major role in the exonucleolytic maturation of CCA-containing tRNA precursors in Bacillus subtilis. *Nucleic Acids Res.*, **33**, 3636–3643.
122. Hartmann,R.K., Gossringer,M., Spath,B., Fischer,S. and Marchfelder,A. (2009) The making of tRNAs and more - RNase P and tRNase Z. *Prog. Mol. Biol. Transl. Sci.*, **85**, 319–368.
123. Wurtzel,O., Sapra,R., Chen,F., Zhu,Y., Simmons,B.A. and Sorek,R. (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res.*, **20**, 133–141.
124. Veith,T., Martin,R., Wurm,J.P., Weis,B.L., Duchardt-Ferner,E., Safferthal,C., Hennig,R., Mirus,O., Bohnsack,M.T., Wöhnert,J. *et al.* (2012) Structural and functional analysis of the archaeal endonuclease Nob1. *Nucleic Acids Res.*, **40**, 3259–3274.
125. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.
126. Holzmann,J., Frank,P., Löffler,E., Bennett,K.L., Gerner,C. and Rossmanith,W. (2008) RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. *Cell*, **135**, 462–474.
127. Howard,M.J., Lim,W.H., Fierke,C.a. and Koutmos,M. (2012) Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 16149–16154.
128. Xu,J., Peng,W., Sun,Y., Wang,X., Xu,Y., Li,X., Gao,G. and Rao,Z. (2012) Structural study of MCPIP1 N-terminal conserved domain reveals a PIN-like RNase. *Nucleic Acids Res.*, **40**, 6957–6965.
129. Lechner,M., Rossmanith,W., Hartmann,R.K., Thölken,C., Gutmann,B., Giegé,P. and Gobert,A. (2015) Distribution of Ribonucleoprotein and Protein-Only RNase P in Eukarya. *Mol. Biol. Evol.*, **32**, msv187.
130. Mizgalska,D., Wegrzyn,P., Murzyn,K., Kasza,A., Koj,A., Jura,J., Jarzab,B. and Jura,J. (2009) Interleukin-1-inducible MCPIP protein has structural and functional properties of RNase and participates in degradation of IL-1beta mRNA. *FEBS J.*, **276**, 7386–7399.
131. Lin,R.J., Chien,H.L., Lin,S.Y., Chang,B.L., Yu,H.P., Tang,W.C. and Lin,Y.L. (2013) MCPIP1 ribonuclease exhibits broad-spectrum antiviral effects through viral RNA binding and degradation. *Nucleic Acids Res.*, **41**, 3314–3326.
132. Suzuki,H.I., Arase,M., Matsuyama,H., Choi,Y.L., Ueno,T., Mano,H., Sugimoto,K. and Miyazono,K. (2011) MCPIP1 ribonuclease antagonizes dicer and terminates microRNA biogenesis through precursor microRNA degradation. *Mol. Cell*, **44**, 424–436.
133. Marco,A. and Marín,I. (2009) CGIN1: a retroviral contribution to mammalian genomes. *Mol. Biol. Evol.*, **26**, 2167–2170.
134. Ketting,R.F., Haverkamp,T.H.A., van Luenen,H.G.A.M. and Plasterk,R.H.A. (1999) Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell*, **99**, 133–141.
135. Bennett,E.J., Rush,J., Gygi,S.P. and Harper,J.W. (2010) Dynamics of cullin-RING ubiquitin ligase network revealed by systematic quantitative proteomics. *Cell*, **143**, 951–965.
136. Grasby,J.A., Finger,L.D., Tsutakawa,S.E., Atack,J.M. and Tainer,J.A. (2012) Unpairing and gating: sequence-independent substrate recognition by FEN superfamily nucleases. *Trends Biochem. Sci.*, **37**, 74–84.
137. Xu,Y., Derbyshire,V., Ng,K., Sun,X.C., Grindley,N.D. and Joyce,C.M. (1997) Biochemical and mutational studies of the 5'-3' exonuclease of DNA polymerase I of Escherichia coli. *J. Mol. Biol.*, **268**, 284–302.
138. Nagarajan,V.K., Jones,C.I., Newbury,S.F. and Green,P.J. (2013) XRN 5'→3' exoribonucleases: structure, mechanisms and functions. *Biochim. Biophys. Acta*, **1829**, 590–603.
139. Winther,K., Tree,J.J., Tollervey,D. and Gerdes,K. (2016) VapCs of Mycobacterium tuberculosis cleave RNAs essential for translation. *Nucleic Acids Res.*, **44**, 9860–9871.
140. Tomlinson,C.G., Atack,J.M., Chapados,B., Tainer,J.A. and Grasby,J.A. (2010) Substrate recognition and catalysis by flap

endonucleases and related enzymes. *Biochem. Soc. Trans.*, **38**, 433–437.
141. Feng,M., Patel,D., Dervan,J.J., Ceska,T., Suck,D., Haq,I. and Sayers,J.R. (2004) Roles of divalent metal ions in flap endonuclease-substrate interactions. *Nat. Struct. Mol. Biol.*, **11**, 450–456.
142. Yang,W., Lee,J.Y. and Nowotny,M. (2006) Making and breaking nucleic acids: two-Mg2+-Ion catalysis and substrate specificity. *Mol. Cell*, **22**, 5–13.
143. Rosta,E., Nowotny,M., Yang,W. and Hummer,G. (2011) Catalytic mechanism of RNA backbone cleavage by ribonuclease H from quantum mechanics/molecular mechanics simulations. *J. Am. Chem. Soc.*, **133**, 8934–8941.
144. Dupureur,C.M. (2008) Roles of metal ions in nucleases. *Curr. Opin. Chem. Biol.*, **12**, 250–255.
145. Das,U., Pogenberg,V., Subhramanyam,U.K.T., Wilmanns,M., Gourinath,S. and Srinivasan,A. (2014) Crystal structure of the VapBC-15 complex from Mycobacterium tuberculosis reveals a two-metal ion dependent PIN-domain ribonuclease and a variable mode of toxin-antitoxin assembly. *J. Struct. Biol.*, **188**, 249–258.
146. Hwang,K.Y., Baek,K., Kim,H.Y. and Cho,Y. (1998) The crystal structure of flap endonuclease-1 from Methanococcus jannaschii. *Nat. Struct. Biol.*, **5**, 707–713.
147. Devos,J.M., Tomanicek,S.J., Jones,C.E., Nossal,N.G. and Mueser,T.C. (2007) Crystal structure of bacteriophage T4 5' nuclease in complex with a branched DNA reveals how flap endonuclease-1 family nucleases bind their substrates. *J. Biol. Chem.*, **282**, 31713–31724.
148. Tomlinson,C.G., Syson,K., Sengerova,B., Atack,J.M., Sayers,J.R., Swanson,L., Tainer,J.A., Williams,N.H. and Grasby,J.A. (2011) Neutralizing mutations of carboxylates that bind metal 2 in T5 flap endonuclease result in an enzyme that still requires two metal ions. *J. Biol. Chem.*, **286**, 30878–30887.
149. Zheng,L., Li,M., Shan,J., Krishnamoorthi,R. and Shen,B. (2002) Distinct roles of two Mg2+ binding sites in regulation of murine flap endonuclease-1 activities. *Biochemistry*, **41**, 10323–10331.
150. Syson,K., Tomlinson,C., Chapados,B.R., Sayers,J.R., Tainer,J.A., Williams,N.H. and Grasby,J.A. (2008) Three metal ions participate in the reaction catalyzed by T5 flap endonuclease. *J. Biol. Chem.*, **283**, 28741–28746.
151. Dupureur,C.M. (2010) One is enough: insights into the two-metal ion nuclease mechanism from global analysis and computational studies. *Metallomics*, **2**, 609–620.
152. Howard,M.J., Klemm,B.P. and Fierke,C.A. (2015) Mechanistic studies reveal similar catalytic strategies for phosphodiester bond hydrolysis by protein-only and RNA-dependent ribonuclease P. *J. Biol. Chem.*, **290**, 13454–13464.
153. Harrington,J.J. and Lieber,M.R. (1995) DNA structural elements required for FEN-1 binding. *J. Biol. Chem.*, **270**, 4503–4508.
154. Garforth,S.J. and Sayers,J.R. (1997) Structure-specific DNA binding by bacteriophage T5 5'-3' exonuclease. *Nucleic Acids Res.*, **25**, 3801–3807.
155. Habraken,Y., Sung,P., Prakash,L. and Prakash,S. (1993) Yeast excision repair gene RAD2 encodes a single-stranded DNA endonuclease. *Nature*, **366**, 365–368.
156. Lalle,P., Nouspikel,T., Constantinou,A., Thorel,F. and Clarkson,S.G. (2002) The founding members of xeroderma pigmentosum group G produce XPG protein with severely impaired endonuclease activity. *J. Invest. Dermatol.*, **118**, 344–351.
157. Taddeo,B., Zhang,W. and Roizman,B. (2006) The U(L)41 protein of herpes simplex virus 1 degrades RNA by endonucleolytic cleavage in absence of other cellular or viral proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 2827–2832.
158. Jones,C.I., Zabolotskaya,M.V. and Newbury,S.F. (2012) The 5' → 3' exoribonuclease XRN1/Pacman and its functions in cellular processes and development. *Wiley Interdiscip. Rev. RNA*, **3**, 455–468.
159. Henry,Y., Wood,H., Morrissey,J.P., Petfalski,E., Kearsey,S. and Tollervey,D. (1994) The 5' end of yeast 5.8S rRNA is generated by exonucleases from an upstream cleavage site. *EMBO J.*, **13**, 2452–2463.
160. McKenzie,J.L., Duyvestyn,J.M., Smith,T., Bendak,K., MacKay,J., Cursons,R., Cook,G.M. and Arcus,V.L. (2012) Determination of ribonuclease sequence-specificity using Pentaprobes and mass spectrometry. *RNA (N. Y.)*, **18**, 1267–1278.
161. Glavan,F., Behm-Ansmant,I., Izaurralde,E. and Conti,E. (2006) Structures of the PIN domains of SMG6 and SMG5 reveal a nuclease within the mRNA surveillance complex. *EMBO J.*, **25**, 5117–5125.
162. Schaeffer,D., Tsanova,B., Barbas,A., Reis,F.P., Dastidar,E.G., Sanchez-Rotunno,M., Arraiano,C.M. and van Hoof,A. (2009) The exosome contains domains with specific endoribonuclease, exoribonuclease and cytoplasmic mRNA decay activities. *Nat. Struct. Mol. Biol.*, **16**, 56–62.
163. Schmidt,S.A., Foley,P.L., Jeong,D.H., Rymarquis,L.A., Doyle,F., Tenenbaum,S.A., Belasco,J.G. and Green,P.J. (2015) Identification of SMG6 cleavage sites and a preferred RNA cleavage motif by global analysis of endogenous NMD targets in human cells. *Nucleic Acids Res.*, **43**, 309–323.
164. Eom,S., Wang,J. and Steitz,T. (1996) Structure of Taq polymerase with DNA at the polymerase active site. *Nature*, **382**, 278–281.
165. Harrington,J.J. and Lieber,M.R. (1994) Functional domains within FEN-1 and RAD2 define a family of structure-specific endonucleases: implications for nucleotide excision repair. *Genes Dev.*, **8**, 1344–1355.
166. Kotarski,M.a., Leonard,D.a., Bennett,S.a., Bishop,C.P., Wahn,S.D., Sedore,S.a. and Shrader,M. (1998) The Drosophila gene asteroid encodes a novel protein and displays dosage-sensitive interactions with Star and Egfr. *Genome*, **41**, 295–302.
167. Sun,M., Schwalb,B., Pirkl,N., Maier,K.C., Schenk,A., Failmezger,H., Tresch,A. and Cramer,P. (2013) Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Mol. Cell*, **52**, 52–62.
168. Burgess,H.M. and Mohr,I. (2015) Cellular 5'-3' mRNA exonuclease Xrn1 controls double-stranded RNA accumulation and anti-viral responses. *Cell Host Microbe*, **17**, 332–344.
169. Li,Y., Yamane,D. and Lemon,S.M. (2015) Dissecting the roles of the 5' exoribonucleases Xrn1 and Xrn2 in restricting hepatitis C virus replication. *J. Virol.*, **89**, 4857–4865.
170. Wells,G.R., Weichmann,F., Colvin,D., Sloan,K.E., Kudla,G., Tollervey,D., Watkins,N.J. and Schneider,C. (2016) The PIN domain endonuclease Utp24 cleaves pre-ribosomal RNA at two coupled sites in yeast and humans. *Nucleic Acids Res.*, **44**, 9016.
171. Anantharaman,V., Koonin,E.V. and Aravind,L. (2001) TRAM, a predicted RNA-binding domain, common to tRNA uracil methylation and adenine thiolation enzymes. *FEMS Microbiol. Lett.*, **197**, 215–221.
172. Burroughs,A.M. and Aravind,L. (2016) RNA damage in biological conflicts and the diversity of responding RNA repair systems. *Nucleic Acids Res.*, **44**, 8525–8555.
173. Skruzny,M., Schneider,C., Racz,A., Weng,J., Tollervey,D. and Hurt,E. (2009) An endoribonuclease functionally linked to perinuclear mRNP quality control associates with the nuclear pore complexes. *PLoS Biol.*, **7**, e8.
174. Andrews,E.S.V. and Arcus,V.L. (2015) The mycobacterial PhoH2 proteins are type II toxin antitoxins coupled to RNA helicase domains. *Tuberculosis*, **95**, 385–394.
175. Lamanna,A.C. and Karbstein,K. (2009) Nob1 binds the single-stranded cleavage site D at the 3'-end of 18S rRNA with its PIN domain. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 14259–14264.
176. Tomecki,R., Kristiansen,M.S., Lykke-Andersen,S., Chlebowski,A., Larsen,K.M., Szczesny,R.J., Drazkowska,K., Pastula,A., Andersen,J.S., Stepien,P.P. *et al.* (2010) The human core exosome interacts with differentially localized processive RNases: hDIS3 and hDIS3L. *EMBO J.*, **29**, 2342–2357.
177. Makarova,K.S., Aravind,L., Galperin,M.Y., Grishin,N.V., Tatusov,R.L., Wolf,Y.I. and Koonin,E.V. (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.*, **9**, 608–628.
178. Makino,D.L., Baumgärtner,M. and Conti,E. (2013) Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex. *Nature*, **495**, 70–75.
179. Levin,I., Schwarzenbacher,R., Page,R., Abdubek,P., Ambing,E., Biorac,T., Brinen,L.S., Campbell,J., Canaves,J.M., Chiu,H.-J. *et al.* (2004) Crystal structure of a PIN (PilT N-terminus) domain (AF0591) from Archaeoglobus fulgidus at 1.90 A resolution. *Proteins*, **56**, 404–408.

180. Pavelec,D.M., Lachowiec,J., Duchaine,T.F., Smith,H.E. and Kennedy,S. (2009) Requirement for the ERI/DICER complex in endogenous RNA interference and sperm development in Caenorhabditis elegans. *Genetics*, **183**, 1283–1295.
181. Amblar,M., Sagner,G. and Lopez,P. (1998) Purification and properties of the 5'-3' exonuclease D10A mutant of DNA polymerase I from Streptococcus pneumoniae: a new tool for DNA sequencing. *J. Biotechnol.*, **63**, 17–27.
182. Kelly,R.B., Atkinson,M.R., Huberman,J.A. and Kornberg,A. (1969) Excision of thymine dimers and other mismatched sequences by DNA polymerase of Escherichia coli. *Nature*, **224**, 495–501.
183. Allen,L.M., Hodskinson,M.R.G. and Sayers,J.R. (2009) Active site substitutions delineate distinct classes of eubacterial flap endonuclease. *Biochem. J.*, **418**, 285–292.
184. Fukushima,S., Itaya,M., Kato,H., Ogasawara,N. and Yoshikawa,H. (2007) Reassessment of the in vivo functions of DNA polymerase I and RNase H in bacterial cell growth. *J. Bacteriol.*, **189**, 8575–8583.
185. Bhagwat,M., Hobbs,L.J. and Nossal,N.G. (1997) The 5'-exonuclease activity of bacteriophage T4 RNase H is stimulated by the T4 gene 32 single-stranded DNA-binding protein, but its flap endonuclease is inhibited. *J. Biol. Chem.*, **272**, 28523–28530.
186. Hollingsworth,H.C. and Nossal,N.G. (1991) Bacteriophage T4 encodes an RNase H which removes RNA primers made by the T4 DNA replication system in vitro. *J. Biol. Chem.*, **266**, 1888–1897.
187. Qiu,J., Qian,Y., Chen,V., Guan,M.X. and Shen,B. (1999) Human exonuclease 1 functionally complements its yeast homologues in DNA recombination, RNA primer removal, and mutation avoidance. *J. Biol. Chem.*, **274**, 17893–17900.
188. Lee,B.I. and Wilson,D.M. 3rd (1999) The RAD2 domain of human exonuclease 1 exhibits 5' to 3' exonuclease and flap structure-specific endonuclease activities. *J. Biol. Chem.*, **274**, 37763–37769.
189. O'Donovan,A., Davies,A.A., Moggs,J.G., West,S.C. and Wood,R.D. (1994) XPG endonuclease makes the 3' incision in human DNA nucleotide excision repair. *Nature*, **371**, 432–435.
190. Liu,R., Qiu,J., Finger,L.D., Zheng,L. and Shen,B. (2006) The DNA-protein interaction modes of FEN-1 with gap substrates and their implication in preventing duplication mutations. *Nucleic Acids Res.*, **34**, 1772–1784.
191. Lindahl,T., Gally,J.A. and Edelman,G.M. (1969) Deoxyribonuclease IV: a new exonuclease from mammalian tissues. *Proc. Natl. Acad. Sci. U.S.A.*, **62**, 597–603.
192. Shen,B., Singh,P., Liu,R., Qiu,J., Zheng,L., Finger,L.D. and Alas,S. (2005) Multiple but dissectible functions of FEN-1 nucleases in nucleic acid processing, genome stability and diseases. *BioEssays*, **27**, 717–729.
193. Ip,S.C., Rass,U., Blanco,M.G., Flynn,H.R., Skehel,J.M. and West,S.C. (2008) Identification of Holliday junction resolvases from humans and yeast. *Nature*, **456**, 357–361.
194. Rass,U., Compton,S.A., Matos,J., Singleton,M.R., Ip,S.C., Blanco,M.G., Griffith,J.D. and West,S.C. (2010) Mechanism of Holliday junction resolution by the human GEN1 protein. *Genes Dev.*, **24**, 1559–1569.
195. Garner,E., Kim,Y., Lach,F.P., Kottemann,M.C. and Smogorzewska,A. (2013) Human GEN1 and the SLX4-associated nucleases MUS81 and SLX1 are essential for the resolution of replication-induced Holliday junctions. *Cell Rep.*, **5**, 207–215.
196. Scherly,D., Nouspikel,T., Corlet,J., Ucla,C., Bairoch,A. and Clarkson,S.G. (1993) Complementation of the DNA repair defect in xeroderma pigmentosum group G cells by a human cDNA related to yeast RAD2. *Nature*, **363**, 182–185.
197. Smiley,J.R. (2004) Herpes simplex virus virion host shutoff protein: immune evasion mediated by a viral RNase? *J. Virol.*, **78**, 1063–1068.
198. Liu,Y.F., Tsai,P.Y., Lin,F.Y., Lin,K.H., Chang,T.J., Lin,H.W., Chulakasian,S. and Hsu,W.L. (2015) Roles of nucleic acid substrates and cofactors in the vhs protein activity of pseudorabies virus. *Vet. Res.*, **46**, 141.
199. Bashkirov,V.I., Scherthan,H., Solinger,J.A., Buerstedde,J.M. and Heyer,W.D. (1997) A mouse cytoplasmic exoribonuclease (mXRN1p) with preference for G4 tetraplex substrates. *J. Cell. Biol.*, **136**, 761–773.
200. Liu,Z. and Gilbert,W. (1994) The yeast KEM1 gene encodes a nuclease specific for G4 tetraplex DNA: implication of in vivo functions for this novel DNA structure. *Cell*, **77**, 1083–1092.
201. Page,A.M., Davis,K., Molineux,C., Kolodner,R.D. and Johnson,A.W. (1998) Mutational analysis of exoribonuclease I from Saccharomyces cerevisiae. *Nucleic Acids Res.*, **26**, 3707–3716.
202. Xue,Y., Bai,X., Lee,I., Kallstrom,G., Ho,J., Brown,J., Stevens,A. and Johnson,A.W. (2000) Saccharomyces cerevisiae RAI1 (YGL246c) is homologous to human DOM3Z and encodes a protein that binds the nuclear exoribonuclease Rat1p. *Mol. Cel. Biol.*, **20**, 4006–4015.
203. West,S., Gromak,N. and Proudfoot,N.J. (2004) Human 5' → 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, **432**, 522–525.
204. Howard,M.J., Karasik,A., Klemm,B.P., Mei,C., Shanmuganathan,A., Fierke,C.A. and Koutmos,M. (2016) Differential substrate recognition by isozymes of plant protein-only Ribonuclease P. *RNA (N. Y.)*, **22**, 782–792.
205. Eberle,A.B., Lykke-Andersen,S., Muhlemann,O. and Jensen,T.H. (2009) SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nat. Struct. Mol. Biol.*, **16**, 49–55.
206. Lebreton,A., Tomecki,R., Dziembowski,A. and Séraphin,B. (2008) Endonucleolytic RNA cleavage by a eukaryotic exosome. *Nature*, **456**, 993–996.
207. Daines,D.A., Wu,M.H. and Yuan,S.Y. (2007) VapC-1 of nontypeable Haemophilus influenzae is a ribonuclease. *J. Bacteriol.*, **189**, 5041–5048.
208. Winther,K.S. and Gerdes,K. (2009) Ectopic production of VapCs from Enterobacteria inhibits translation and trans-activates YoeB mRNA interferase. *Mol. Microbiol.*, **72**, 918–930.
209. Hamilton,B., Manzella,A., Schmidt,K., DiMarco,V. and Butler,J.S. (2014) Analysis of non-typeable Haemophilous influenzae VapC1 mutations reveals structural features required for toxicity and flexibility in the active site. *PLoS One*, **9**, e112921.
210. Audoly,G., Vincentelli,R., Edouard,S., Georgiades,K., Mediannikov,O., Gimenez,G., Socolovschi,C., Mege,J.L., Cambillau,C. and Raoult,D. (2011) Effect of rickettsial toxin VapC on its eukaryotic host. *PloS One*, **6**, e26528.
211. Arcus,V.L., Bäckbro,K., Roos,A., Daniel,E.L. and Baker,E.N. (2004) Distant structural homology leads to the functional characterization of an archaeal PIN domain as an exonuclease. *J. Biol. Chem.*, **279**, 16471–16478.