

Two novel PIWI families: roles in inter-genomic conflicts in bacteria and Mediator-dependent modulation of transcription in eukaryotes

*A. Maxwell Burroughs, Lakshminarayan M. Iyer, L. Aravind**

*National Center for Biotechnology Information,
National Library of Medicine,
National Institutes of Health,
Bethesda, MD 20894, USA*

** Correspondence to: aravind@ncbi.nlm.nih.gov*

ABSTRACT

Background: The PIWI module, found in the PIWI/AGO superfamily of proteins, is a critical component of several cellular pathways including germline maintenance, chromatin organization, regulation of splicing, RNA interference, and virus suppression. It binds a guide strand which helps it target complementary nucleic strands.

Results: Here we report the discovery of two divergent, novel families of PIWI modules, the first such to be described since the initial discovery of the PIWI/AGO superfamily over a decade ago. Both families display conservation patterns consistent with the binding of oligonucleotide guide strands. The first family is bacterial in distribution and is typically encoded by a distinctive three-gene operon alongside genes for a restriction endonuclease fold enzyme and a helicase of the DinG family. The second family is found only in eukaryotes. It is the core conserved module of the Med13 protein, a subunit of the CDK8 subcomplex of the transcription regulatory Mediator complex.

Conclusions: Based on the presence of the DinG family helicase, which specifically acts on R-loops, we infer that the first family of PIWI modules is part of a novel RNA-dependent restriction system which could target invasive DNA from phages, plasmids or conjugative transposons. It is predicted to facilitate restriction of actively transcribed invading DNA by utilizing RNA guides. The PIWI family found in the eukaryotic Med13 proteins throws new light on the regulatory switch through which the CDK8 subcomplex modulates transcription at Mediator-bound promoters of highly transcribed genes. We propose that this involves recognition of small RNAs by the PIWI module in Med13 resulting in a conformational switch that propagates through the mediator complex.

1. INTRODUCTION

The PIWI module, found in the PIWI/AGO superfamily of proteins, is a common functional denominator for a wide range of biological processes in eukaryotes. These include, but not are limited to, germline maintenance [1], post-transcriptional gene silencing/RNA interference (RNAi) [2], chromatin dynamics, regulation of transcription [3, 4], regulation of alternative splicing [5], DNA elimination in ciliates [6, 7] and suppression of viral infection [8]. It acts by binding a double-stranded RNA duplex, typically consisting of a targeting RNA strand, referred to as the “guide strand”, and the targeted RNA strand complementary to the guide strand. Binding of the guide strand to the target strand results in either the silencing of specific RNA transcripts, as in the case of transposon silencing during germline maintenance [1, 7] and mRNA silencing during RNAi [2], or is thought to localize crucial factors for regulating processes like transcription [3] and alternative splicing [5]. The PIWI module contains an RNase H fold domain with a conserved triad of residues required for nuclease activity that might participate both in processing the guide strand precursor as well as cleaving target RNAs complementary to the guide strand [9-16]. On several independent occasions the PIWI module has lost the RNase H fold catalytic residues; these inactive versions are still capable of silencing activity by interfering with translation or facilitating degradation of guide strand-bound mRNAs by other nucleases [17].

While the PIWI/AGO superfamily was initially discovered in eukaryotes, orthologs were also identified in a wide range of prokaryotes spanning both the archaeal and bacterial superkingdoms [18, 19]. Despite extensive characterization of these proteins in eukaryotes, the roles of the prokaryotic PIWI (pPIWI) proteins and the nature of their potential double-stranded nucleotide targets have remained murky. Recent analysis detected association with genes encoding several distinct, predicted nucleases, and a general preference for pPIWI genes to be localized in genomic neighborhoods containing genes belonging to known phage-defense systems. This led to a proposal advocating a role for pPIWI proteins as components of novel prokaryotic systems involved in defense against invasive mobile elements [20]. Earlier structural studies observed a tighter binding propensity for single-stranded DNA relative to single-stranded RNA guide strands in pPIWI proteins [21, 22]. They also found, in stark contrast to the eukaryotic PIWI protein, the favored double-stranded substrate for the pPIWI domains to be a DNA-RNA hybrid. These observations suggested that pPIWI proteins might act on DNA-RNA hybrids.

Given recent increase in available genome data, we surveyed the complete scope of eukaryotic and prokaryotic PIWI domains to gain a better understanding of their relationship. Here we report the discovery of two distinctive PIWI families resulting from this survey, the first novel PIWI families to be discovered in well over ten years. One of these is a previously unrecognized bacterial family predicted to be a key component of a RNA-dependent restriction system. The second family is found in the eukaryotic Med13 protein, one of four protein components of the repressive CDK8 subcomplex of the multi-subunit, transcription regulatory Mediator complex. Identification of a PIWI module in Med13 generates a new testable hypothesis regarding the transcription modulatory role of the CDK8 subcomplex.

2. RESULTS AND DISCUSSION

2.1 Discovery of two novel PIWI families

The PIWI module as presently defined in the Pfam database [23] consists of two distinct but functionally tightly coupled domains: an N-terminal three-layered α/β sandwich of the Rossmannoid type, with a four-stranded central β -sheet reminiscent of the TOPRIM domain and the β -sheet crossover occurring after the first β -strand [24] (see Fig. 1A). This domain contributes crucial residues that bind the 5' end of the small RNA guide strand [21, 22, 25-30]. The second domain is the core RNase H domain, which contributes additional, critical residues for guide strand-binding and when preserving the nuclease active site also cleaves the target strand. Prior structural studies on the PIWI module have labeled these two domains as the "MID" and "PIWI" domains, respectively [9, 31]; a convention we adopt henceforth.

We performed profile-profile comparisons using the HHpred program initiated with both single sequences and a HMM derived from a multiple alignment of complete PIWI modules as queries against the complete set of HMMs found in the Pfam and Interpro databases. Interestingly, we observed statistically significant relationships between the PIWI module and two distinct protein families defined by the models "domain of unknown function" DUF3893 and Med13_C (corresponding to a conserved region in the eukaryotic Mediator complex Med13 proteins) from the Pfam database. For instance, a search initiated with a pPIWI module from *Mycobacterium* sp. KMS (gi: 119855142) recovers the DUF3893 profile with p-value= 7×10^{-6} ; 94% probability and the Med13_C profile with p-value= 3.4×10^{-4} ; 90% probability. To further investigate this relationship, we systematically collected all proteins corresponding to the DUF3893 and Med13_C models using

iterative PSI-BLAST searches. The DUF3893-containing proteins were sporadically distributed across a wide range of bacterial lineages including firmicutes, actinobacteria, $\alpha/\beta/\gamma$ -proteobacteria, cyanobacteria, and chloroflexi. The Med13 proteins are widely distributed across eukaryotes including most plants, fungi, animals, slime molds, and stramenopiles as well as basal eukaryotes such as the parabasalid *Trichomonas vaginalis* and the heterolobosean *Naegleria gruberi* (see Additional File 1). In certain lineages additional Med13 paralogs were identified, including those resulting from a duplication event that occurred early in vertebrates [32].

We then constructed multiple sequence alignments of the proteins matching these modules, used them to predict secondary structure, and checked for congruence with existing structures of PIWI modules to determine precisely boundaries of the MID and PIWI domains. This showed that the DUF3893 and Med13_C models currently present in Pfam imprecisely define the domain architectures and boundaries within these proteins, notably excluding regions from both the MID and PIWI domains. Accordingly, we emended the domain boundaries of the DUF3893 and Med13_C models to completely match the predicted structural elements of the two constituent domains (see Fig. 1A). Reciprocal HHpred searches initiated with both single sequences and HMMs derived from the above alignments against a database of HMMs constructed from multiple alignments built using Protein Data Bank (PDB) chains as seeds confirmed relationships with the PIWI domain: an emended representative version of the module matching Pfam DUF3893 (gi: 228927677 from *Bacillus thuringiensis*) recovers the PIWI module from *Archaeoglobus fulgidus*, PDB: 2W42, p-value=6.7x10⁻⁵, probability 90%). Iterative sequence searches with PSI-BLAST further confirmed this relationship: e.g. a search with an emended representative of the module matching Pfam DUF3893 (gi: 269125748 from *Thermomonospora curvatae*) recovers a classical pPIWI domain (gi: 295689105 from *Caulobacter segnis* with e-value=9x10⁻¹⁵, iteration 4). Similarly, a representative of the emended Med13 module (gi: 393215315 from *Fomitiporia mediterranea*) recovers a classical pPIWI module from *Pyrococcus furiosus* (PDB: 1U04, p-value=2.1x10⁻⁴; probability 87%).

2.2 Characterization of the novel bacterial PIWI family

Structural and architectural features

The above-identified bacterial family which overlaps with the Pfam DUF3893 model displayed two unique, absolutely conserved residues: an arginine and a glutamate (see Fig. 2A). Hence, we refer to this family as the pPIWI-RE family (prokaryotic PIWI with conserved R and E residues).

Secondary structure predictions indicated that the pPIWI-RE family is distinguished from all previously known PIWI domains by the presence of an additional α -helical element following the initial three-stranded beta-meander characteristic of the RNase H fold (see Figs. 1A,2A). We mapped all strongly-conserved residues found in the pPIWI-RE family on to available structures of classical PIWI modules and compared those positions to those required for RNase activity or nucleic acid binding in the latter modules (see Figs. 1B-C, 2A). This showed that the conserved residues in the PIWI and MID domains of the pPIWI-RE family corresponded well to the positions known to be critical for nucleic acid-binding in the cognate domains of classical PIWI modules (see Figs. 1, 2A). In particular, the conserved positions in the MID domain were all clustered in the cleft that specifically binds the 5' end of the guide strand. This suggests that, like classical PIWI domains [33], the pPIWI-RE is likely to recognize small guide strands by anchoring them via the 5' end. The arginine and glutamate characteristic of the pPIWI-RE family mapped to the β -sheet extension, which is unique to the PIWI-like clade (PIWI and Endonuclease V) of the RNase H fold (see Figs. 1A, 2A). We predict that these two residues form a salt bridge across this β -sheet, which probably stabilizes its tertiary structure, and maintains a conformation specific to this family that is required to recognize the guide strand. The RNase catalytic residues are retained only in a subset of the pPIWI-RE family, suggesting that similar to the classical PIWI family they include both active and inactive versions.

The classical PIWI modules are typically fused to several N-terminal RNA-binding domains. In eukaryotic PIWI proteins, in order from the N-terminus, these include the so-called "N-term" domain implicated in unwinding of the double-stranded guide and passenger strands and also guide-target duplexes [34] and the single-stranded RNA-binding PAZ domain which interacts with 3' ends of guide strands. Certain classical PIWI family proteins from kinetoplastids show an OB fold domain instead of the "N-term" domain. Previously studied prokaryotic PIWI proteins display a distinct architecture: in lieu of a PAZ domain they feature the so-called APAZ (Analogous to PAZ) domain suggesting analogous functions for the two domains [20]. Additionally, few pPIWI domains may contain extreme N-terminal fusions to predicted Sir2-domains [20]. The large N-terminal region of the pPIWI-RE family contains a distinct, conserved globular domain that partly overlaps with the Pfam DUF3962 model. Secondary structure predictions indicate that it is likely to adopt a β -strand-rich fold. It neither showed strong congruence with the secondary structural elements of the PAZ or APAZ domain nor did it display the well-conserved sequence motifs characteristic of the PAZ or APAZ domains (see Additional File 1). Furthermore, profile-profile searches did not point to

any relationship between the N-terminal region of the pPIWI-RE family and these domains. Hence, this N-terminal region is likely to contain at least one distinct globular domain, which might nevertheless function analogously to the N-terminal domains in the classical PIWI proteins in mediating additional nucleic acid contacts (see Fig. 2B).

Contextual associations of the pPIWI-RE module

Given the value of contextual information in gleaning insight into the functions of genes [35, 36], we systematically collected conserved gene neighborhoods and domain fusions for the pPIWI-RE domains. Consequently, we observed two distinct genomic contexts for the pPIWI-RE genes with mutually exclusive phyletic patterns (see Fig. 2B): (1) occurrence as a standalone gene (restricted to several *Bacillus* species, proteobacteria *Magnetospirillum gryphiswaldense*, *Pseudomonas putida* and *Azotobacter vinelandii*, and actinobacteria from the genera *Streptomyces* and *Thermomonospora*; Additional file 1). On rare occasions, this version of the pPIWI-RE module might occur fused to an N-terminal Zincin-like metallopeptidase domain. (2) Occurrence as part of a widely distributed three-gene neighborhood. Of the two genes that co-occur with the pPIWI-RE gene we found the first to encode a protein with a conserved restriction endonuclease (REase) fold domain by using profile-profile comparisons with the HHpred program (probability 94% using gi: 158336201 from *Acaryochloris* as a query). These proteins also contain a helical domain with a conserved arginine and Zinc ribbon (ZnR) domain at N-terminus of the REase domain (see Fig. 2B). Moreover, on at least four different occasions these proteins have also acquired further N-terminal HTH domains belonging to the LexA, TetR, MerR and a previously uncharacterized clades [37] (see Fig. 2B). The second gene codes for a Superfamily II (SF-II) DNA helicase. Within SF-II it can be confidently assigned to the DinG-like clade on the basis of two unique structural features that typify them; namely, an iron-binding cysteine-rich region found after strand-2 of the helicase domain [38, 39] and a large helical region inserted between conserved helix-4 and strand-5 which precede the C-terminal P-loop NTPase fold repeat unit characteristic of helicases [40, 41]. The former domain apparently acts as an intracellular sensor of redox potential to regulate activity of the DinG helicase domains [42]. The gene order within this triad is strictly conserved with the REase gene coming first followed by the DinG SF-II helicase and pPIWI-RE genes (see Fig. 2B and Additional file 1). Furthermore, the three genes have either overlapping or very closely spaced termini suggesting they are transcribed as a single polycistronic message.

Functional implications of pPIWI-RE coding systems: A novel RNA-dependent restriction system

The widespread but patchy distribution of the above-described pPIWI-RE containing gene-triads across numerous phylogenetically distant bacteria (Additional file 1) is consistent with this system being disseminated by horizontal gene transfer (HGT). This pattern is reminiscent of bacteriophage restriction systems that confer a selective advantage on recipients due to their role in countering bacteriophage infections [43]. The presence of a gene coding for an REase protein without an associated methylase gene in the pPIWI-RE containing gene-triads is reminiscent of restriction systems such as the Mcr systems that target modified invading DNA [44]. The fact that the REase gene is always the first gene in the operon implies that it would be made before any of the other products and be available to cleave DNA. Hence, like the REases from the Mcr systems, it should have some means of specifically targeting non-self DNA rather than suicidally cleaving the cellular genome upon production. DinG serves as a helicase partner for multiple nuclease domains such as the RNase T-like and RNase D-like nuclease domains (both of which belong the RNase H fold) [45-47]. Hence, it could function as a helicase partner for either the REase or pPIWI-RE or both. Given that these gene triads are parallel to type I and type III restriction-modification (R-M) systems in that they combine REase with helicase genes [48, 49], it is conceivable that the DinG helicase plays a role comparable to the helicases that translocate the target DNA in those R-M systems. However, recent studies on DinG-like helicases, which show that it acts on RNA-DNA duplexes *in vitro* [50] and R-loops (bubble-like structures forming via displacement of one strand of a DNA double helix by a complementary RNA strand [51]) *in vivo* [52], point to further functional complexities. DinG-like helicases are specifically involved in unwinding of R-loops during replication across active transcriptional units [52]. Interestingly, DinG-like helicases have also been found to be components of Type-U CRISPR/Cas systems [53], supporting their action in the context of DNA-RNA hybrid duplexes.

Taken together, these observations allow us to propose a model that can account for the most likely activities of all three products of these gene triads (see Fig. 3A). On the basis of the DinG helicase we posit that the initiating signal recognized by these systems is likely to be a DNA-RNA hybrid structure. These are known to primarily form during transcription and replication of phages [54] or plasmid [55, 56] and relatively infrequently during transcription of the endogenous genome [51]. Therefore, specifically targeting these structures could provide an effective means of restricting transcriptionally active and replicating invasive genomes and their transcripts. In this system the pPIWI-RE module is likely to be deployed as a sensor for the DNA-RNA hybrid, in a manner comparable to the classical pPIWI domain for which there is accumulating evidence for

preferential binding to DNA/RNA hybrids [20, 22, 29]. The catalytically active pPIWI-RE modules might additionally cleave the RNA strand of such hybrid duplexes. Recognition of the DNA-RNA hybrid by the pPIWI-RE module is likely to recruit the DinG helicase for the unwinding and/or the translocation of R-loops, which could further provide a suitable dsDNA substrate for cleavage by the REase domain. Importantly, this hypothesis of DNA-RNA hybrid-directed restriction can explain why the REase protein, which is the first to be transcribed and translated, is unlikely to act on self DNA upon its production. The diverse HTH domains, which are occasionally fused to the N-termini of the REase proteins, could either function as autoregulators of transcription of the gene triad or in providing sequence specificity during restriction.

In the case of pPIWI-RE genes occurring independently of the above-described three gene restriction system we found no evidence for the presence of related REase or DinG genes in the same genomes. A simple interpretation would be that these pPIWI-RE modules function similarly to the aforementioned versions, but instead of recruiting restriction machinery they function by themselves. It is possible in these cases they modulate gene expression by cleaving transcripts, physically interfering with transcription (an echo of the action of eukaryotic PIWI proteins), or blocking the release of transcripts from the template DNA [3, 57].

2.3 The PIWI module in eukaryotic Med13

Structural and architectural features of the MedPIWI module

Given the presence of this PIWI module in the Med13 subunit of the Mediator complex, we hereafter refer to it as the MedPIWI module. An inspection of the multiple sequence alignment of the novel eukaryotic family revealed extensive conservation at the positions crucial for nucleic acid-binding in the classical PIWI module including residues interacting with the 5' end of the guide strand in the MID domain (see Figs. 1, 4A). However, this family shows certain distinctive features: 1) absence of the first catalytic aspartate/glutamate found near the C-terminus of strand 1 of the RNase H fold's core β -sheet. 2) The second conserved residue of the catalytic triad, located at the C-terminus of strand-4 of the RNase H fold, is absent with no identifiable compensatory residues. 3) Another charged residue contributing directly to the active site from the C-terminal segment of the final helix of the RNase H fold is also absent (see Fig. 4A). 4) Its RNase H fold shows a reasonably well-conserved aspartate in the loop between strand-1 and strand-2, which is suitably positioned to

contact the bound nucleic acid, based on comparisons to classical PIWI domains [58]. 5) The MedPIWI RNase H fold also shows a near-absolutely conserved aspartate at the C-terminus of strand 2 (see Fig. 4A) that is unlikely to have any role in nucleic acid substrate recognition. Taken together, these observations suggest that none of the MedPIWI modules might be catalytically active. However, they are likely to bind double-stranded nucleic acid substrates, just as the classical PIWI modules.

The MedPIWI modules are distinguished from all other PIWI modules by the presence of extensive disordered regions, often occurring as lineage-specific inserts within both the MID and PIWI domains and also in between the two (indicated by numbers in Figure 4A). This family is also distinguished by a small domain consisting of a likely beta-hairpin followed by a single alpha-helix located immediately N-terminal to the MID domain and might be compared to the small “linker” domains observed in classic PIWI families [20]. Beyond this domain is the Med13-N module corresponding to the Pfam model Med13_N (see Fig. 4B). The conserved core of this region is predicted to adopt an $\alpha+\beta$ structure with a prominent stretch of 6-7 contiguous β strands which could adopt a barrel or sandwich-like fold (Additional file 1). This module is present in all eukaryotic Med13s except those from *Entamoebidae*, where it appears to have been displaced or has degenerated. Thus, the Med13-N module was likely associated with the MedPIWI even in the stem eukaryotes, and is comparable in its location, though not necessarily in function, to the N-terminal domains, such as PAZ, APAZ and that found in the pPIWI-RE family (see above). Some additional lineage-specific globular domains might be present along with an extensive disordered region in the linker connecting the Med13-N module to the rest of the protein. These include a potential Zn-binding domain with two CxC motifs (where “C” is a cysteine residue and “x” is any residue) in animals and other unrelated modules in plants and fungi (see Fig. 4B, Additional File 1). The size and frequency of the lineage-specific inserts and disordered regions roughly corresponds to the total number of units comprising the Mediator complex in a given lineage [32]. Thus, they might represent secondary adaptations for increased inter-subunit contacts within the Mediator complex.

Partners and physical interactions of Med13: functional implications for the MedPIWI module in eukaryotic transcription regulation

The Mediator complex, along with several basal or general transcription factors, is part of the Preinitiation Complex (PIC), which is needed for transcription at promoters of genes transcribed by

RNA polymerase II (pol II) in eukaryotes [59, 60]. The Mediator complex has two basic forms (see Fig. 3B): 1) the core Mediator complex, which is a strong transcriptional coactivator [61] and occupies promoters across the genome [62, 63] and 2) the Mediator-CDK8 complex, which usually has a negative regulatory role and while found to transiently associate across all promoters, associates strongly with only a subset of genes that typically show higher expression levels [62-66]. The latter complex is characterized by the addition of a four subunit subcomplex, CDK8, which, in addition to the MedPIWI-containing Med13, also contains Med12, cyclin C, and the CDK8 kinase. Negative regulation by the CDK8 subcomplex appears to utilize multiple independent, but apparently synergistic, actions of its distinct subunits (see Fig. 3B). The cyclin/kinase pair of the subcomplex phosphorylates the pol II C-terminal tail disrupting the association between pol II and the core Mediator complex [67]. It might also phosphorylate cyclin H in the TFIID complex and inhibit activation of translation by the latter complex [68]. However, previous studies have shown that negative regulation of transcription by the CDK8 subcomplex also occurs independently of the CDK8 kinase activity: the interaction between the CDK8 subcomplex and the core Mediator acts as a modulatory “switch” that allosterically affects the core Mediator-pol II interaction [69, 70] and determines the shift between transient and stable CDK8 subcomplex promoter occupancy. This switch is believed to be dependent on Med12 and Med13 [70, 71], although the exact mode of their action remains murky. In this regard, recent studies utilizing an *in vitro* chromatin-based transcriptional system demonstrated that Med13 is critical for physically linking the CDK8 subcomplex to the core Mediator complex and is specifically required to repress previously activated promoters by barring re-association of a pol II enzyme with the PIC [70].

Given these studies our discovery of a PIWI module in Med13 provides a previously unexplored vista to investigate the mechanism of transcriptional modulation by the CDK8 subcomplex (see Fig. 3B). As the MedPIWI module displays the conserved features related to binding double stranded substrates (see above, Figs. 1B-C, 4A), we posit that this activity is central to the molecular switch that modulates the core Mediator-pol II interactions. We predict two plausible candidates for the substrate oligonucleotide bound by the MedPIWI modules that are consistent with published laboratory studies: 1) it is conceivable that the MedPIWI module retained the ancestral ability to bind DNA-RNA hybrid duplexes, a feature that the ancestral eukaryotic PIWI modules would have presumably possessed when they were acquired from the prokaryotic progenitors. DNA-small RNA hybrids could form close to the transcription start site (TSS) from the small RNA byproducts of polymerase stalling or backtracking [72, 73]. Indeed, such small transcripts have been detected

(commonly referred to as TSSa [74] or tiRNA transcripts [75]) in several global deep-sequencing datasets across a range of animal species [76] and even in association with classical PIWI domains [77]. These could either re-associate with DNA opened as part of the transcriptional bubble formed during re-initiation events or remain associated with open DNA in the wake of repeated pol II passages. This proposal has the attractive feature of explaining the preferential association of Med13 with highly transcribed genes [62-66, 70] because such genes are known to be enriched in small TSS-associated transcripts [75], thereby increasing the chances of formation of DNA-RNA hybrids substrates for the MedPIWI module. The observation that the CDK8 subcomplex association occurs only after initiation of at least a single round of transcription by pol II following PIC assembly [70] also suggests its association might require the availability of previously-transcribed RNA byproducts. Another potential source for small RNAs that could form DNA-RNA hybrids is the small processed antisense transcripts that have been found to be associated with the promoter sites of transcriptionally active genes [3]. 2) Alternatively, like most characterized eukaryotic PIWI modules, the MedPIWI module might bind dsRNA substrates. In this case its action can be compared to the classical eukaryotic PIWI protein AGO2, which has been shown to regulate the positioning of pol II while binding sense-antisense RNA duplexes derived from transcriptionally active genes [3]. Interestingly, these antisense small RNA-AGO2 complexes increase in abundance concomitant with transcriptional activation upon stimuli such as heat shock [3]. It is possible that the MedPIWI module acts in a comparable manner to associate with such promoter-derived small RNAs that could form dsRNA duplexes during active transcription.

In conclusion, we hypothesize that the modulatory switch mediated by the CDK8 subcomplex probably depends on the ability of the MedPIWI module to recognize small transcripts associated with active promoters that form either DNA-RNA or dsRNA duplexes. This binding induces a conformational change that propagates through the rest of the complex to allosterically impact the interaction of the Mediator with pol II. Binding of duplexes by the MedPIWI module might also influence the deployment of the additional layers of control that depend on the CDK8 subcomplex, such as the activity of the CDK8 kinase [67, 68] and Med12-mediated histone H3K9 SET domain methyltransferase (G9a) recruitment [71] (see Fig. 3B). Intriguingly, in a small number of cases, association of the CDK8 subcomplex with the core Mediator results in the Med13- and Med12-dependent transcriptional activation rather than repression [78, 79]. While this manuscript was under review, a study was published demonstrating the role of enhancer-associated long non-coding RNAs (lncRNAs) in facilitating this process of activation of transcription by the CDK8

subcomplex along with the core mediator [80] (see Fig. 3B). It was demonstrated that in animals these activating lncRNAs interact with the Med12 subunit of the CDK8 complex and cause it to catalyze Histone H3 serine 10 phosphorylation rather than the above-mentioned negative regulatory phosphorylations of Cyclin H and the RNA polymerase C-terminal tail. H3S10 phosphorylation has a positive regulatory role probably by inhibiting the repressive H3K9 methylation among other actions. We suspect that interaction with these enhancer-derived lncRNAs is unlikely to be the primary function of the MedPIWI module because it is conserved across eukaryotes and appears to be required for actions of the CDK8 complex beyond activated transcription. However, we cannot rule out that the lncRNA might interact with processed small RNAs to form duplexes that might be recognized MedPIWI module to regulate transcription in certain conditions.

3. EVOLUTIONARY CONSIDERATIONS

The new PIWI families reported here also offer an opportunity to reassess the natural history of the PIWI/AGO superfamily. The pPIWI-RE family shows a relatively smaller spread across the prokaryotic tree (see Additional File) compared to the classical pPIWI proteins [20]. Hence, it is possible that pPIWI-RE descended from an RNase-active classical pPIWI module in bacteria and was subsequently dispersed to diverse lineages via HGT. The multiple independent losses of the RNase H fold catalytic residues in the pPIWI-RE family are comparable to the classical PIWI modules [20]. Thus, not just active processing of RNA, but also non-catalytic binding of duplexes containing RNA appears to have been widely used across the PIWI/AGO superfamily. Indeed, this function appears to have been the dominant theme in the case of the MedPIWI family. The phyletic patterns of Med13 are closely correlated with the three other subunits of the CDK8 complex. They are present in several basal eukaryotes and are widespread across the eukaryotic tree strongly supporting the presence of a complete CDK8 complex in the last eukaryotic common ancestor (LECA). Thus, the CDK8 subcomplex and an ancestral version of the core Mediator complex appear to have been in place by the LECA, suggesting that antagonistic regulatory interactions of these complexes was a feature of transcription regulation in the common ancestor of extant eukaryotes.

Earlier studies had indicated that at least one member of the classical PIWI family was already present in the LECA [85]. Prior to LECA, in the eukaryotic stem lineage, this PIWI protein appears to have undergone a duplication giving rise to a version with a dedicated role in transcription regulation and a second version primarily involved as a standalone protein in diverse processes

involving small non-coding RNAs. The former version appears to have functionally associated with the other emerging subunits of the CDK8 complex with a corresponding rapid divergence in sequence. At least in the latter version there appears to have been a specificity shift towards dsRNA from the likely ancestral pPIWI preference for binding DNA/RNA hybrid duplexes [20, 22, 29]. The classical PIWI family is also widely conserved across archaea [19], suggesting that the stem eukaryotes could have possibly inherited the ancestral PIWI protein directly from their archaeal progenitor. Given the functional connections now known or inferred across the PIWI/AGO superfamily (each of the two families discussed here and the classical PIWI proteins) to regulation of transcription, it is conceivable that even in archaea (and possibly other prokaryotes) PIWI proteins function in transcription regulation, beyond the proposed role in defense against genomic parasites. If this were the case, then the two primary eukaryotic versions merely reflect partitioning of the ancestral roles into distinct proteins. Thus, our identification of a novel eukaryotic PIWI family could also have implications for the functions of the prokaryotic PIWI domains.

GENERAL CONCLUSIONS

The two novel families of PIWI modules described here are the first such discoveries since the initial characterization of the PIWI/Argonaute family in eukaryotes and their close prokaryotic counterparts over a decade ago [18, 86, 87]. While considerably divergent from these earlier-characterized versions, both families are predicted to bind double-stranded substrates based on the strong conservation residues at positions corresponding to nucleic acid binding sites in the classical PIWI modules in both of the novel families (see Figs. 1, 2, and 4). Moreover, their predicted functions fit within the spectrum of previously observed functional roles for different members of the PIWI superfamily. Thus, despite the considerable divergence from the classical PIWI family at the sequence level the new families appear to have maintained the characteristic ability of this clade of RNase H fold proteins to operate on RNA-containing duplexes. Nevertheless, the predicted functions of the two newly described families present some previously unobserved features. The pPIWI-RE family offers the first example for a potential RNA-dependent restriction system in prokaryotes that is distinct from the previously characterized CRISPR/Cas-type systems [53]. In particular it presents some parallels to the Type-II CRISPR/Cas systems which combine a RNase H fold nuclease with a HNH endoDNase that is also found in several restriction systems [53]. Thus, it emerges as the first clear example of a PIWI family member directing and coordinating a DNA- and RNA- based defensive response against genomic parasites in bacteria. This system could potentially be developed as a reagent to cleave target DNA using a RNA guide. Our prediction implicating the

MedPIWI family in recognition of RNA-containing duplexes offers an entirely new mechanism for the action of the CDK8 subcomplex both in terms of the modulation of transcription at the promoters of highly expressed genes and providing the first delineation of the criterion underlying the transition from transient CDK8 subcomplex co-occupancy at sites of core Mediator occupancy to sustained CDK8 subcomplex association resulting in repressive activity [62] (see Fig. 3B). This research also further fuels the broader emerging theme implicating ncRNAs in modulation of transcription at sites of initiation [3, 80]. This hypothesis could be investigated via a combination of ChIP-seq experiments on CDK8 subcomplex members and MedPIWI module immunoprecipitation-sequencing.

MATERIALS AND METHODS

Iterative profile searches with the PSI-BLAST [88] and JACKHMMER [89] programs were used to retrieve homologous sequences in the protein non-redundant (NR) database at the National Center for Biotechnology Information (NCBI). For most searches a cut-off e-value of 0.01 was used to assess significance. In each iteration, the newly detected sequences that had e-values lower than the cut-off were examined for conserved motifs to detect potential homologs in the twilight zone. Similarity-based clustering was performed using the BLASTCLUST program (<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>) to cluster sequences at different thresholds. Multiple sequence alignments were built using the Kalign [90] and MUSCLE [91] programs, followed by manual adjustments based on profile-profile alignment, secondary structure prediction and structural alignments. Consensus secondary structures were predicted using the JPred program [92]. Remote sequence similarity searches were performed using profile-profile comparisons with the HHpred program [93]. Gene neighborhoods were extracted and analyzed using a custom PERL script that operates on the Genbank genome or whole genome shotgun files. The protein sequences of all neighbors were clustered using the BLASTCLUST program to identify related sequences in gene neighborhoods. Each cluster of homologous proteins was then assigned an annotation based on the domain architecture or shared conserved domain. A complete list of Genbank gene identifiers for proteins investigated in this study is provided in the Additional File 1. Structure similarity searches were conducted using the DALIite program [94] and structural alignments were generated by means of the MUSTANG program [95].

Authors' contributions

AMB collected data; AMB, LMI and LA analyzed the data; AMB and LA wrote the manuscript that was read and approved by all authors.

Acknowledgements

The authors' research is supported by the intramural funds of the US Department of Health and Human Services (National Library of Medicine, NIH).

Competing interests

The authors declare that they have no competing interests.

Additional file

Additional file 1 provides access to: 1) comprehensive list of Genbank identifiers, architectures and operons of modules uncovered in this study. 2) A comprehensive set of alignments of domains reported here in text format.

FIGURE LEGENDS

Figure 1. Spatial conservation of active site and nucleotide-binding residues in the MID and PIWI domains. (A) Topology diagram depicting the structural features and critical binding regions in the domains. MID and PIWI designations are provided at the top of the diagram. The β -sheet extension unique to the PIWI clade of the RNase H fold is labeled and shaded in grey. Locations of key active site residues are marked with green lines. Active site and general regions of nucleotide-binding are shaded and labeled. (B) Cartoon renderings of active site and nucleotide binding regions of a solved PIWI domain structure in complex with double-stranded nucleotide guide/passenger strands (PDB: 3HO1 [30]). Residues in the structure involved in guide strand binding with cognate conserved residues in pPIWI-RE and MedPIWI families are colored in yellow. The guide strand is colored in tan, passenger strand in light blue.

Figure 2. Multiple sequence alignment and contextual information of the pPIWI-RE pPIWI-RE family. (A) An alignment along with representatives of the classical PIWI module is shown. Regions of poor conservation are replaced with numbers representing the length of the excised region. The consensus sequence is provided on the bottom line. Strongly-conserved residues are shaded in red and colored in white. Residues involved in catalytic RNase H activity are shaded in red and colored in yellow. Columns in alignment are color-coded based on conservation of shared chemical properties: yellow, hydrophobic/aliphatic (h/l); green, small/tiny (s/u); purple, charged (c/+/-); blue, polar (p); orange, hydroxyl group-containing (o); grey, large (b). Conserved residues involved in nucleotide binding across both the classical and pPIWI-RE PIWI modules or residues contributing to nuclease activity are denoted above the appropriate column in the alignment by “*” and “^”, respectively. The predicted salt bridge-forming arginine and glutamate residues unique to the pPIWI-RE module are denoted by “&”. Residues which may be conserved in classical PIWI modules but not present in the pPIWI-RE module are denoted by “%”. Boundaries of the MID and PIWI domains are noted above the secondary structure prediction. Sequences in the alignment are labeled to the left of the alignment with gene name, organism abbreviation, and gene identifier number (gi number), demarcated by underscores. (B) Representative domain architectures and conserved gene neighborhoods involving the pPIWI-RE module. Genes within a conserved neighborhood are depicted as arrows with the direction of the arrowhead pointing in the 5' to 3'. Labels below each architecture or neighborhood provide the gene name, organism abbreviation, and gi number for a representative protein. The characteristic C-rich and helical regions of the DinG-type helicase are represented by yellow lettering and purple coils, respectively. Domain abbreviations: ZR, zinc ribbon; X, conserved globular region found N-terminal to MID and pPIWI-RE domains; Y, conserved, largely α -helical domain with conserved arginine residue found N-terminal to ZR and REase domains; Z, largely α -helical domain found N-terminal to DinG-type helicase. Organism abbreviations may be found in Additional File 1.

Figure 3. Schematic representation of predicted functions of the pPIWI-RE and MedPIWI domains. (A) pPIWI-RE domain associates with DNA-RNA hybrid structure present during R-loop formation in an invasive DNA element, resulting in recruitment of the DinG helicase and endoDNase REase domains. (B) Regulation of the core Mediator complex via the CDK8 subcomplex

is depicted, beginning at left with 1) simplified representation of the PIC, poised for initiation of transcription. 2) In absence of CDK8 subcomplex, the core Mediator complex recruits pol II and transcription is initiated. 3) Kinase activity-independent repression of transcription: the CDK8 subcomplex (depicted as yellow oval) transiently associates with core Mediator complexes across the genome [62]; availability of a small RNA binding substrate for the MedPIWI domain in the Med13 component of the CDK8 subcomplex triggers shift from transient association to repressive role of CDK8 subcomplex, triggering conformational switch in the Mediator-CDK8 combined complex which blocks pol II re-association. 4) lncRNA-mediated transcriptional activation: association of Med12 with activating lncRNA transcribed and looping from distal enhancer element (depicted as box colored in green) facilitates CDK8 kinase-mediated phosphorylation of transcriptional-activating histone H3 serine 10, resulting in association of pol II and transcriptional activation [80]. 5) Additional layers of CDK8 subcomplex-mediated transcriptional repression: CDK8 kinase phosphorylates TFIIH [68] or C-terminal domain of pol II [67] and Med12-mediated recruitment of SET domain methyltransferase (G9a) methylates histone H3 lysine 9 [71], all resulting in repression of transcription. Abbreviations: P, phosphorylation event; Me, methylation event; S, switch resulting in conformational change.

Figure 4. Multiple sequence alignment and domain architectures of the MedPIWI family. (A) Multiple sequence alignment with representatives of the classical PIWI module is shown. Organization, numbering, labeling, consensus abbreviations, and coloring of the alignment are as described in the legend to Figure 2. Conserved residues involved in nucleotide binding across both the classical and MedPIWI modules are denoted above the appropriate column in the alignment by “*”. Residues which may be conserved in classical PIWI modules but not in the MedPIWI module are denoted by “%”. (B) Representative domain architectures of the MedPIWI module. The small green box immediately upstream of the MID domain represents the conserved, small “linker” domain. Other unlabeled domains represent potential lineage-specific domains while CxC refers to the animal-specific, potential zinc-binding domain (see Additional File 1). Organism abbreviations may be found in Additional File 1.

References

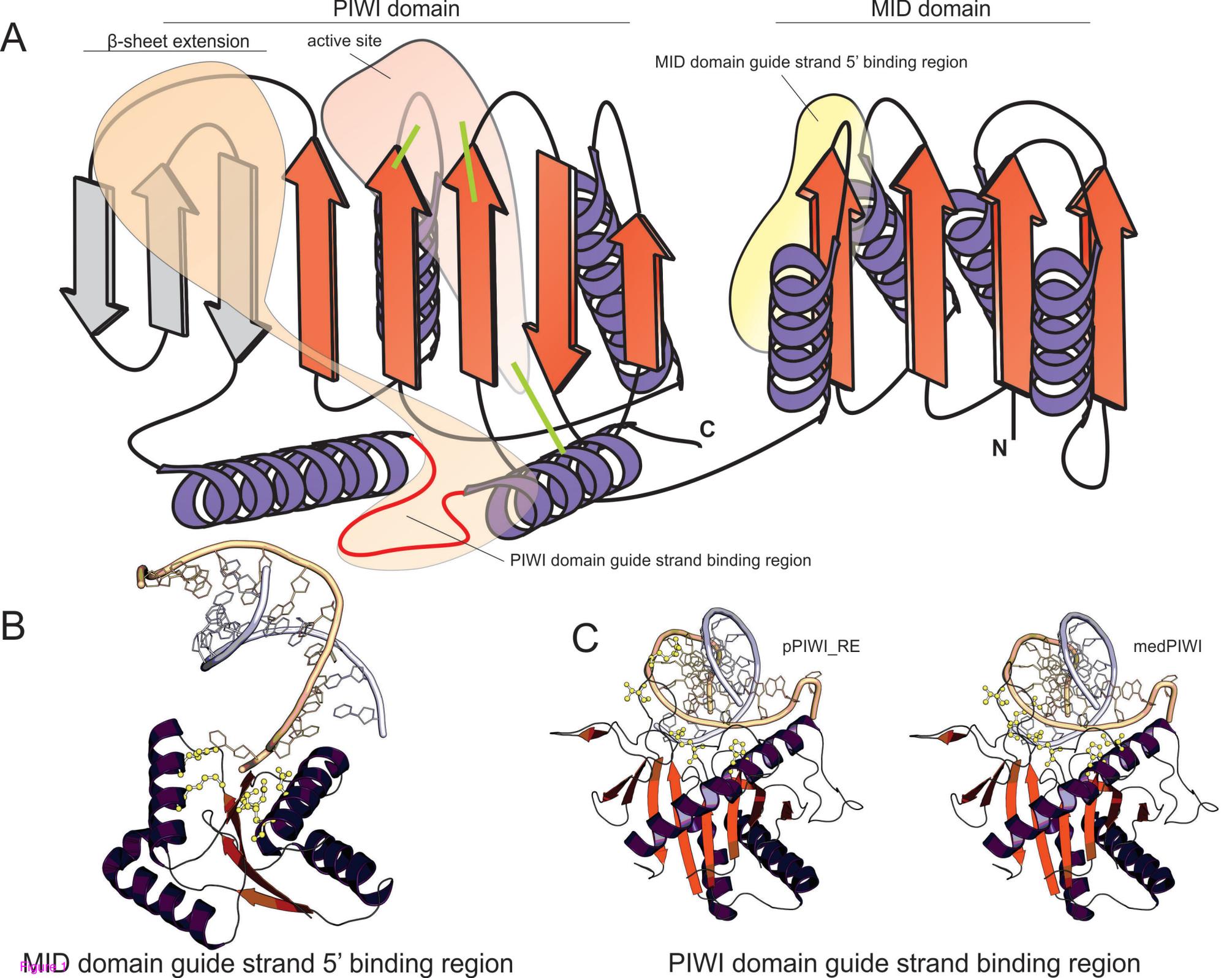
1. Cox DN, Chao A, Baker J, Chang L, Qiao D, Lin H: **A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal.** *Genes & development* 1998, **12**:3715-3727.
2. Murchison EP, Hannon GJ: **miRNAs on the move: miRNA biogenesis and the RNAi machinery.** *Current opinion in cell biology* 2004, **16**:223-229.
3. Cernilogar FM, Onorati MC, Kothe GO, Burroughs AM, Parsi KM, Breiling A, Lo Sardo F, Saxena A, Miyoshi K, Siomi H, et al: **Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila.** *Nature* 2011, **480**:391-395.
4. Halic M, Moazed D: **Dicer-independent primal RNAs trigger RNAi and heterochromatin formation.** *Cell* 2010, **140**:504-516.
5. Ameyar-Zazoua M, Rachez C, Souidi M, Robin P, Fritsch L, Young R, Morozova N, Fenouil R, Descostes N, Andrau JC, et al: **Argonaute proteins couple chromatin silencing to alternative splicing.** *Nature structural & molecular biology* 2012, **19**:998-1004.
6. Mochizuki K: **RNA-directed epigenetic regulation of DNA rearrangements.** *Essays in biochemistry* 2010, **48**:89-100.
7. Chalker DL, Yao MC: **DNA elimination in ciliates: transposon domestication and genome surveillance.** *Annual review of genetics* 2011, **45**:227-246.
8. Aliyari R, Ding SW: **RNA-based viral immunity initiated by the Dicer family of host immune receptors.** *Immunological reviews* 2009, **227**:176-188.
9. Song JJ, Smith SK, Hannon GJ, Joshua-Tor L: **Crystal structure of Argonaute and its implications for RISC slicer activity.** *Science* 2004, **305**:1434-1437.
10. Rand TA, Petersen S, Du F, Wang X: **Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation.** *Cell* 2005, **123**:621-629.
11. Matranga C, Tomari Y, Shin C, Bartel DP, Zamore PD: **Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes.** *Cell* 2005, **123**:607-620.
12. Miyoshi K, Tsukumo H, Nagami T, Siomi H, Siomi MC: **Slicer function of Drosophila Argonautes and its involvement in RISC formation.** *Genes & development* 2005, **19**:2837-2848.
13. Leuschner PJ, Ameres SL, Kueng S, Martinez J: **Cleavage of the siRNA passenger strand during RISC assembly in human cells.** *EMBO reports* 2006, **7**:314-320.
14. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ: **Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila.** *Cell* 2007, **128**:1089-1103.
15. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC: **A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila.** *Science* 2007, **315**:1587-1590.
16. Yang JS, Lai EC: **Dicer-independent, Ago2-mediated microRNA biogenesis in vertebrates.** *Cell Cycle* 2010, **9**:4455-4460.
17. Djuranovic S, Nahvi A, Green R: **A parsimonious model for gene regulation by miRNAs.** *Science* 2011, **331**:550-553.
18. Cerutti L, Mian N, Bateman A: **Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain.** *Trends in biochemical sciences* 2000, **25**:481-482.
19. Aravind L, Koonin EV: **Eukaryote-specific domains in translation initiation factors: implications for translation regulation and evolution of the translation system.** *Genome research* 2000, **10**:1172-1184.

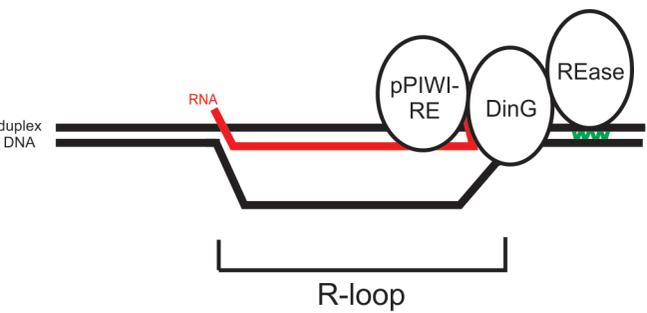
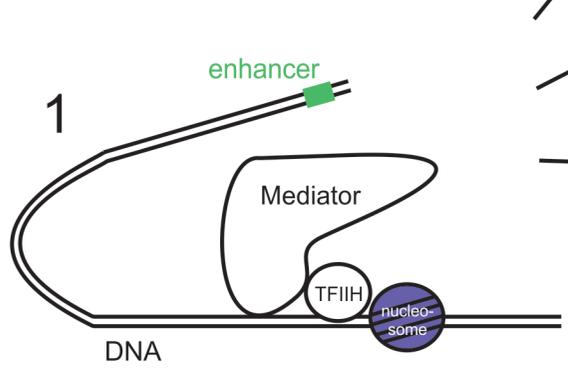
20. Makarova KS, Wolf YI, van der Oost J, Koonin EV: **Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements.** *Biology direct* 2009, **4**:29.
21. Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, Patel DJ: **Structural basis for 5'-end-specific recognition of guide RNA by the *A. fulgidus* Piwi protein.** *Nature* 2005, **434**:666-670.
22. Yuan YR, Pei Y, Ma JB, Kuryavyi V, Zhadina M, Meister G, Chen HY, Dauter Z, Tuschl T, Patel DJ: **Crystal structure of *A. aeolicus* argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage.** *Molecular cell* 2005, **19**:405-419.
23. Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families database.** *Nucleic acids research* 2010, **38**:D211-222.
24. Aravind L, Leipe DD, Koonin EV: **Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins.** *Nucleic acids research* 1998, **26**:4205-4213.
25. Frank F, Sonenberg N, Nagar B: **Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2.** *Nature* 2010, **465**:818-822.
26. Boland A, Tritschler F, Heimstadt S, Izaurralde E, Weichenrieder O: **Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein.** *EMBO reports* 2010, **11**:522-527.
27. Parker JS, Roe SM, Barford D: **Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex.** *Nature* 2005, **434**:663-666.
28. Wang Y, Sheng G, Juraneck S, Tuschl T, Patel DJ: **Structure of the guide-strand-containing argonaute silencing complex.** *Nature* 2008, **456**:209-213.
29. Wang Y, Juraneck S, Li H, Sheng G, Tuschl T, Patel DJ: **Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex.** *Nature* 2008, **456**:921-926.
30. Wang Y, Juraneck S, Li H, Sheng G, Wardle GS, Tuschl T, Patel DJ: **Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes.** *Nature* 2009, **461**:754-761.
31. Parker JS, Roe SM, Barford D: **Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity.** *The EMBO journal* 2004, **23**:4727-4737.
32. Bourbon HM: **Comparative genomics supports a deep evolutionary origin for the large, four-module transcriptional mediator complex.** *Nucleic acids research* 2008, **36**:3993-4008.
33. Boland A, Huntzinger E, Schmidt S, Izaurralde E, Weichenrieder O: **Crystal structure of the MID-PIWI lobe of a eukaryotic Argonaute protein.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:10466-10471.
34. Kwak PB, Tomari Y: **The N domain of Argonaute drives duplex unwinding during RISC assembly.** *Nature structural & molecular biology* 2012, **19**:145-151.
35. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome research* 2000, **10**:1074-1077.
36. Huynen M, Snel B, Lathe W, 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome research* 2000, **10**:1204-1210.
37. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM: **The many faces of the helix-turn-helix domain: transcription regulation and beyond.** *FEMS microbiology reviews* 2005, **29**:231-262.
38. Pugh RA, Honda M, Leesley H, Thomas A, Lin Y, Nilges MJ, Cann IK, Spies M: **The iron-containing domain is essential in Rad3 helicases for coupling of ATP hydrolysis to DNA translocation and for targeting the helicase to the single-stranded DNA-double-stranded DNA junction.** *The Journal of biological chemistry* 2008, **283**:1732-1743.
39. Rudolf J, Makrantonis V, Ingledew WJ, Stark MJ, White MF: **The DNA repair helicases XPD and FancJ have essential iron-sulfur domains.** *Molecular cell* 2006, **23**:801-808.

40. Singleton MR, Dillingham MS, Wigley DB: **Structure and mechanism of helicases and nucleic acid translocases.** *Annual review of biochemistry* 2007, **76**:23-50.
41. Fairman-Williams ME, Guenther UP, Jankowsky E: **SF1 and SF2 helicases: family matters.** *Current opinion in structural biology* 2010, **20**:313-324.
42. Ren B, Duan X, Ding H: **Redox control of the DNA damage-inducible protein DinG helicase activity via its iron-sulfur cluster.** *The Journal of biological chemistry* 2009, **284**:4829-4835.
43. Aravind L, Anantharaman V, Zhang D, de Souza RF, Iyer LM: **Gene flow and biological conflict systems in the origin and evolution of eukaryotes.** *Frontiers in cellular and infection microbiology* 2012, **2**:89.
44. Bickle TA, Kruger DH: **Biology of DNA restriction.** *Microbiological reviews* 1993, **57**:434-450.
45. Bourniquel AA, Bickle TA: **Complex restriction enzymes: NTP-driven molecular motors.** *Biochimie* 2002, **84**:1047-1059.
46. McRobbie AM, Meyer B, Rouillon C, Petrovic-Stojanovska B, Liu H, White MF: **Staphylococcus aureus DinG, a helicase that has evolved into a nuclease.** *The Biochemical journal* 2012, **442**:77-84.
47. Bukowy Z, Harrigan JA, Ramsden DA, Tudek B, Bohr VA, Stevnsner T: **WRN Exonuclease activity is blocked by specific oxidatively induced base lesions positioned in either DNA strand.** *Nucleic acids research* 2008, **36**:4975-4987.
48. Murray NE: **Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle).** *Microbiology and molecular biology reviews : MMBR* 2000, **64**:412-434.
49. Raghavendra NK, Bheemanaik S, Rao DN: **Mechanistic insights into type III restriction enzymes.** *Frontiers in bioscience : a journal and virtual library* 2012, **17**:1094-1107.
50. Voloshin ON, Camerini-Otero RD: **The DinG protein from Escherichia coli is a structure-specific helicase.** *The Journal of biological chemistry* 2007, **282**:18437-18447.
51. Aguilera A, Garcia-Muse T: **R loops: from transcription byproducts to threats to genome stability.** *Molecular cell* 2012, **46**:115-124.
52. Boubakri H, de Septenville AL, Viguera E, Michel B: **The helicases DinG, Rep and UvrD cooperate to promote replication across transcription units in vivo.** *The EMBO journal* 2010, **29**:145-157.
53. Makarova KS, Aravind L, Wolf YI, Koonin EV: **Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems.** *Biology direct* 2011, **6**:38.
54. Kreuzer KN, Brister JR: **Initiation of bacteriophage T4 DNA replication and replication fork dynamics: a review in the Virology Journal series on bacteriophage T4 and its relatives.** *Virology journal* 2010, **7**:358.
55. Itoh T, Tomizawa J: **Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H.** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77**:2450-2454.
56. Kogoma T: **Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription.** *Microbiology and molecular biology reviews : MMBR* 1997, **61**:212-238.
57. Grewal SI, Elgin SC: **Transcription and RNA interference in the formation of heterochromatin.** *Nature* 2007, **447**:399-406.
58. Parker JS, Parizotto EA, Wang M, Roe SM, Barford D: **Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein.** *Molecular cell* 2009, **33**:204-214.
59. Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW: **The mammalian Mediator complex and its role in transcriptional regulation.** *Trends in biochemical sciences* 2005, **30**:250-255.
60. Malik S, Roeder RG: **Dynamic regulation of pol II transcription by the mammalian Mediator complex.** *Trends in biochemical sciences* 2005, **30**:256-263.

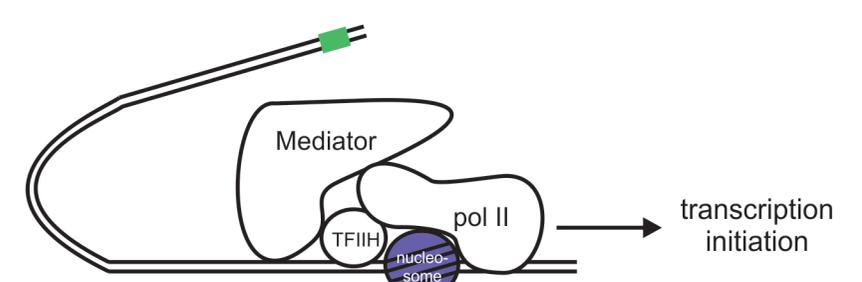
61. Conaway RC, Conaway JW: **Function and regulation of the Mediator complex.** *Current opinion in genetics & development* 2011, **21**:225-230.
62. Andrau JC, van de Pasch L, Lijnzaad P, Bijma T, Koerkamp MG, van de Peppel J, Werner M, Holstege FC: **Genome-wide location of the coactivator mediator: Binding without activation and transient Cdk8 interaction on DNA.** *Molecular cell* 2006, **22**:179-192.
63. Zhu X, Wiren M, Sinha I, Rasmussen NN, Linder T, Holmberg S, Ekwall K, Gustafsson CM: **Genome-wide occupancy profile of mediator and the Srb8-11 module reveals interactions with coding regions.** *Molecular cell* 2006, **22**:169-178.
64. Samuelsen CO, Baraznenok V, Khorosjutina O, Spahr H, Kieselbach T, Holmberg S, Gustafsson CM: **TRAP230/ARC240 and TRAP240/ARC250 Mediator subunits are functionally conserved through evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:6422-6427.
65. Kuchin S, Yeghiayan P, Carlson M: **Cyclin-dependent protein kinase and cyclin homologs SSN3 and SSN8 contribute to transcriptional control in yeast.** *Proceedings of the National Academy of Sciences of the United States of America* 1995, **92**:4006-4010.
66. Gillmor CS, Park MY, Smith MR, Pepitone R, Kerstetter RA, Poethig RS: **The MED12-MED13 module of Mediator regulates the timing of embryo patterning in Arabidopsis.** *Development* 2010, **137**:113-122.
67. Hengartner CJ, Myer VE, Liao SM, Wilson CJ, Koh SS, Young RA: **Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases.** *Molecular cell* 1998, **2**:43-53.
68. Akoulitchev S, Chuikov S, Reinberg D: **TFIIH is negatively regulated by cdk8-containing mediator complexes.** *Nature* 2000, **407**:102-106.
69. Elmlund H, Baraznenok V, Lindahl M, Samuelsen CO, Koeck PJ, Holmberg S, Hebert H, Gustafsson CM: **The cyclin-dependent kinase 8 module sterically blocks Mediator interactions with RNA polymerase II.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:15788-15793.
70. Knuesel MT, Meyer KD, Bernecky C, Taatjes DJ: **The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function.** *Genes & development* 2009, **23**:439-451.
71. Ding N, Zhou H, Esteve PO, Chin HG, Kim S, Xu X, Joseph SM, Friez MJ, Schwartz CE, Pradhan S, Boyer TG: **Mediator links epigenetic silencing of neuronal gene expression with x-linked mental retardation.** *Molecular cell* 2008, **31**:347-359.
72. Taft RJ, Kaplan CD, Simons C, Mattick JS: **Evolution, biogenesis and function of promoter-associated RNAs.** *Cell Cycle* 2009, **8**:2332-2338.
73. Valen E, Preker P, Andersen PR, Zhao X, Chen Y, Ender C, Dueck A, Meister G, Sandelin A, Jensen TH: **Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes.** *Nature structural & molecular biology* 2011, **18**:1075-1082.
74. Seila AC, Core LJ, Lis JT, Sharp PA: **Divergent transcription: a new feature of active promoters.** *Cell Cycle* 2009, **8**:2557-2564.
75. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, et al: **Tiny RNAs associated with transcription start sites in animals.** *Nature genetics* 2009, **41**:572-578.
76. Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ, Rasko JE, et al: **Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans.** *Nature structural & molecular biology* 2010, **17**:1030-1034.
77. Burroughs AM, Ando Y, de Hoon MJ, Tomaru Y, Suzuki H, Hayashizaki Y, Daub CO: **Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin.** *RNA biology* 2011, **8**:158-177.

78. Carrera I, Janody F, Leeds N, Dubeau F, Treisman JE: **Pygopus activates Wingless target gene transcription through the mediator complex subunits Med12 and Med13.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:6644-6649.
79. Gobert V, Osman D, Bras S, Auge B, Boube M, Bourbon HM, Horn T, Boutros M, Haenlin M, Waltzer L: **A genome-wide RNA interference screen identifies a differential role of the mediator CDK8 module subunits for GATA/ RUNX-activated transcription in Drosophila.** *Molecular and cellular biology* 2010, **30**:2837-2848.
80. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhataar R: **Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.** *Nature* 2013, **494**:497-501.
81. Schwartz JC, Younger ST, Nguyen NB, Hardy DB, Monia BP, Corey DR, Janowski BA: **Antisense transcripts are targets for activating small RNAs.** *Nature structural & molecular biology* 2008, **15**:842-848.
82. Han J, Kim D, Morris KV: **Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:12422-12427.
83. Chu Y, Yue X, Younger ST, Janowski BA, Corey DR: **Involvement of argonaute proteins in gene silencing and activation by RNAs complementary to a non-coding transcript at the progesterone receptor promoter.** *Nucleic acids research* 2010, **38**:7736-7748.
84. Janowski BA, Younger ST, Hardy DB, Ram R, Huffman KE, Corey DR: **Activating gene expression in mammalian cells with promoter-targeted duplex RNAs.** *Nature chemical biology* 2007, **3**:166-173.
85. Muljo SA, Kanellopoulou C, Aravind L: **MicroRNA targeting in mammalian genomes: genes and mechanisms.** *Wiley interdisciplinary reviews Systems biology and medicine* 2010, **2**:148-161.
86. Tabara H, Sarkissian M, Kelly WG, Fleenor J, Grishok A, Timmons L, Fire A, Mello CC: **The rde-1 gene, RNA interference, and transposon silencing in C. elegans.** *Cell* 1999, **99**:123-132.
87. Cogoni C, Macino G: **Isolation of quelling-defective (qde) mutants impaired in posttranscriptional transgene-induced gene silencing in Neurospora crassa.** *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**:10233-10238.
88. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25**:3389-3402.
89. Johnson LS, Eddy SR, Portugaly E: **Hidden Markov model speed heuristic and iterative HMM search procedure.** *BMC bioinformatics* 2010, **11**:431.
90. Lassmann T, Frings O, Sonnhammer EL: **Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features.** *Nucleic acids research* 2009, **37**:858-865.
91. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic acids research* 2004, **32**:1792-1797.
92. Cole C, Barber JD, Barton GJ: **The Jpred 3 secondary structure prediction server.** *Nucleic acids research* 2008, **36**:W197-201.
93. Remmert M, Biegert A, Hauser A, Soding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nature methods* 2012, **9**:173-175.
94. Holm L, Rosenstrom P: **Dali server: conservation mapping in 3D.** *Nucleic acids research* 2010, **38**:W545-549.
95. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM: **MUSTANG: a multiple structural alignment algorithm.** *Proteins* 2006, **64**:559-574.

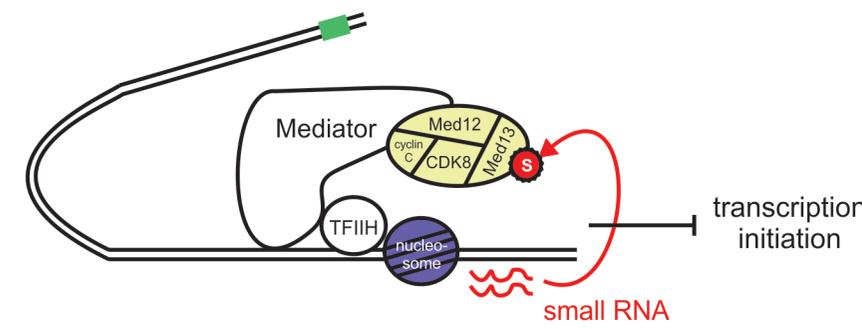


A**B**

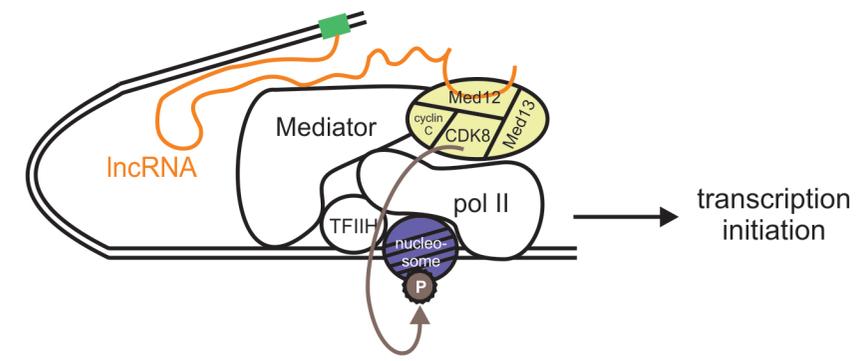
2



3



4



5

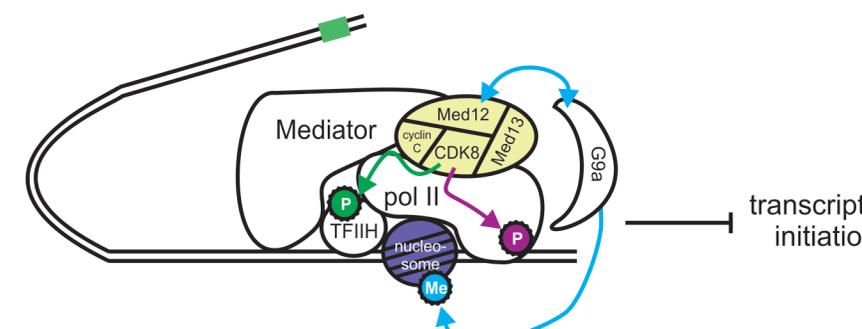
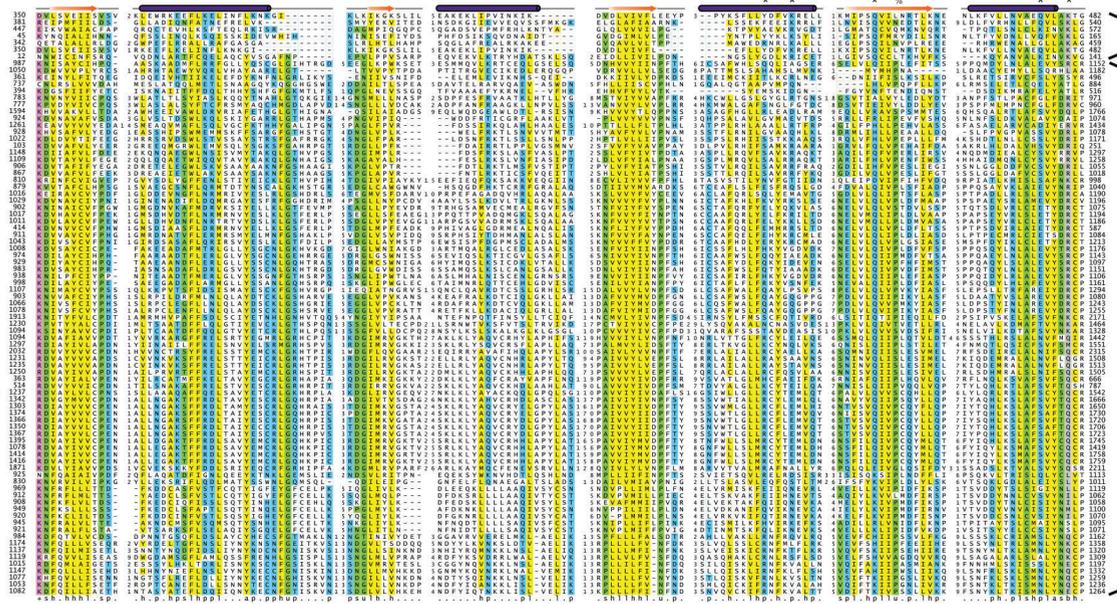


Figure 3

A

MID domain



classical PIWI family

MedPIWI family

B



MED13_Hsap_102468717

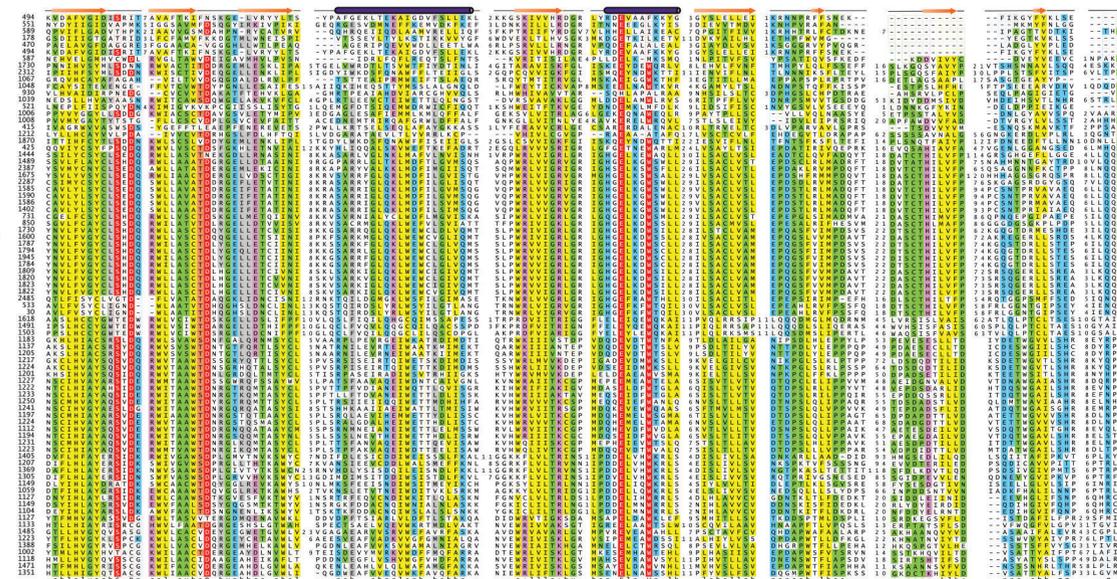


MYCGRDRAFT_109073_Ztri_398398996



GCT_Atha_334183337

PIWI domain



classical PIWI family

MedPIWI family

classical PIWI family

MedPIWI family