Nucleic Acids Research Advance Access published April 7, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw222

FAM46 proteins are novel eukaryotic non-canonical poly(A) polymerases

Krzysztof Kuchta[1,2,†], Anna Muszewska[1,3,†], Lukasz Knizewski[1], Kamil Steczkiewicz[1], Lucjan S. Wyrwicz[4], Krzysztof Pawlowski[5], Leszek Rychlewski[6] and Krzysztof Ginalski[1,*]

[1]Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Zwirki i Wigury 93, 02–089 Warsaw, Poland, [2]College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Banacha 2C, 02–097 Warsaw, Poland, [3]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02–106 Warsaw, Poland, [4]Laboratory of Bioinformatics and Biostatistics, M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, WK Roentgena 5, 02–781 Warsaw, Poland, [5]Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences, Nowoursynowska 166, 02–787 Warsaw, Poland and [6]BioInfoBank Institute, Limanowskiego 24A, 60–744 Poznan, Poland

Received December 16, 2015; Accepted March 22, 2016

ABSTRACT

FAM46 proteins, encoded in all known animal genomes, belong to the nucleotidyltransferase (NTase) fold superfamily. All four human FAM46 paralogs (FAM46A, FAM46B, FAM46C, FAM46D) are thought to be involved in several diseases, with FAM46C reported as a causal driver of multiple myeloma; however, their exact functions remain unknown. By using a combination of various bioinformatics analyses (e.g. domain architecture, cellular localization) and exhaustive literature and database searches (e.g. expression profiles, protein interactors), we classified FAM46 proteins as active non-canonical poly(A) polymerases, which modify cytosolic and/or nuclear RNA 3′ ends. These proteins may thus regulate gene expression and probably play a critical role during cell differentiation. A detailed analysis of sequence and structure diversity of known NTases possessing PAP/OAS1 SBD domain, combined with state-of-the-art comparative modelling, allowed us to identify potential active site residues responsible for catalysis and substrate binding. We also explored the role of single point mutations found in human cancers and propose that FAM46 genes may be involved in the development of other major malignancies including lung, colorectal, hepatocellular, head and neck, urothelial, endometrial and renal papillary carcinomas and melanoma. Identification of these novel enzymes taking part in RNA metabolism in eukaryotes may guide their further functional studies.

INTRODUCTION

Proteins adopting the nucleotidyltransferase (NTase) fold play crucial roles in various biological processes, such as RNA stabilization and degradation (e.g. RNA polyadenylation), RNA editing, DNA repair, intracellular signal transduction, somatic recombination in B cells, regulation of protein activity, antibiotic resistance and chromatin remodelling (1). Almost all known members of this large and highly diverse superfamily transfer nucleoside monophosphate (NMP) from nucleoside triphosphate (NTP) to an acceptor hydroxyl group belonging to protein, nucleic acid or small molecule. They are characterized by the presence of a common α/β-fold structure composed of a three-stranded, mixed β-sheet flanked by four α-helices. This common core corresponding to the minimal NTase fold is usually decorated by various additional structural elements and additional domains, depending on the family. Sequence analysis of distinct members of this superfamily revealed the following common sequence patterns in NTase fold domain: hG[GS], [DE]h[DE]h and h[DE]h (where h indicates a hydrophobic amino acid) that include conserved active site residues. Three conserved aspartates/glutamates are involved in the coordination of divalent ions and activation of the acceptor hydroxyl group of the substrate. Two of them (from the [DE]h[DE]h motif) are located on the second core β-strand, while the third carboxylate (from the h[DE]h motif) is placed on the structurally adjacent third β-strand. The hG[GS] pattern is placed at the beginning of a short, sec-

*To whom correspondence should be addressed. Tel: +48 22 554 0800; Fax: +48 22 554 0801; Email: kginal@cent.uw.edu.pl
†These authors contributed equally to the paper as first authors.

ond core α-helix and has a crucial role in harbouring the substrate within the active site (1).

Human members of the NTase fold superfamily are encoded by 43 genes (1). Until now, only one group of potentially active human NTases, belonging to the so-called family-with-sequence-similarity-46 (FAM46), has not been characterized for its exact biological function. Little is known about FAM46 proteins apart from their involvement in various diseases. FAM46A, a gene preferentially expressed in the retina, was reported as a positional candidate for human retinal diseases since it maps within the RP25 locus to chromosome 6p12.1-q14.1 interval where several retinal dystrophy loci are located (2). It was also suggested that a variable number of tandem repeat polymorphisms in FAM46A may be associated with non-small cell lung cancer (3). Another FAM46 paralog, FAM46B, was identified to have lower expression in metastatic melanoma cells (United States Patent US 7615349 B2). FAM46B and FAM46C have been recently described as potential markers for refractory lupus nephritis (4) and multiple myeloma (5–10), respectively. Finally, it was reported that FAM46D is overexpressed in lung and glioblastoma tumors (11), as well as together with FAM46C, in the brain of autistic-like behaving transgenic mice (8).

Functional proteomics studies showed that FAM46A might have many potential protein interacting partners (12). One of them, ZFYVE9, detected in the yeast two-hybrid system, is involved in the recruitment of unphosphorylated SMAD2/SMAD3 to the transforming growth factor-beta receptor (13). Another member of FAM46 family, FAM46C, is recognized as a type I interferon stimulated gene, which enhances the replication of certain viruses (14). In addition, FAM46C can be an anti-viral factor in acute infected long-tailed pygmy rice rats by Andes virus (15). It was also suggested that FAM46C is functionally related in some way to the regulation of translation (6).

To date, several studies have indicated that proteins belonging to FAM46 family might play an essential role in the cell; however, their exact functions remain unknown. In our previous work, we performed a comprehensive classification of the NTase fold proteins and assigned FAM46 proteins to this superfamily as potentially active members (1). This classification allowed us only to speculate that like other active NTases, FAM46 members may catalyze template-independent incorporation of NMP from NTP either to nucleic acid, protein or small molecule. In this study we present an in-depth bioinformatics analysis of the FAM46 family combined with an exhaustive literature and database searches and propose that the FAM46 proteins function as non-canonical poly(A) polymerases. Detailed insight into the sequence and structure diversity of NTases and their additional N- and C-terminal domains allowed us to generate a reliable 3D model for one of the family members (FAM46C) and to confidently identify the potential active site residues responsible not only for catalysis but also for substrate binding. In addition, the obtained structural model for human FAM46C sheds some new light on the molecular role of mutations found in cancer patients in the FAM46 genes. Finally, the broad sampling of sequenced genomes made it possible to track the evolutionary history of FAM46 proteins back to the origin and hypothesize that the FAM46 family members are present not only in animals but also in all four sequenced Dictyosteliidae and two Entamoeba (Amebozoa) genomes.

MATERIALS AND METHODS

Sequence searches

Four human FAM46 paralogs (FAM46A, FAM46B, FAM46C, FAM46D) were used as queries in PSI-BLAST (16) searches performed against the NCBI non-redundant (NR) protein sequence database with E-value threshold of 0.001 until profile convergence. Collected sequences were split into organism-specific sets and clustered with CD-HIT (17) in order to obtain unique sequences. All FAM46 family members were aligned using Mafft (18) with some manual adjustments. The alignment used for phylogeny reconstruction was additionally trimmed by TrimAl (19) to eliminate poorly aligned and thus uninformative regions.

Additionally, proteins containing both NTase and C-terminal four-helical up-and-down bundle fold domains were collected with PSI-BLAST searches performed against the NCBI NR database with E-value threshold of 0.001. Sequences (PDB SEQRES) of the following structures possessing this domain context: pdb|2pbe, pdb|3c18, pdb|3jyy, pdb|3jz0, pdb|4ebs, pdb|1v4a, pdb|3k7d and pdb|1kan, were used as queries.

Analysis of gene and protein features

The architecture of the human FAM46 genes was analysed using the UCSC genome browser (20). Protein localization was predicted with BaCelLo (21), CELLO (22), WoLF PSORT (23), Euk-mPLoc 2.0 (24) and MultiLoc (25). NetNES (26) was used to detect the nuclear export signal (NES). Protein phosphorylation motifs were detected with Eukaryotic Linear Motif (ELM) (27). Gene expression patterns were analysed using the BioGPS database (28). Genes with average z-scores higher than 5 (at least in one probe set) in 'Barcode on normal tissues' dataset were considered as expressed in specific tissue/cell. 'Barcode on normal tissues' dataset provides a survey across diverse normal human tissues from the U133plus2 Affymetrix microarray (28). The z-scores in this dataset are generated with the barcode function from the R package 'frma', which bases on barcode algorithm (29). The domain architecture was analysed using Meta-BASIC (30) and SMART (31).

Structural analysis of known NTases

Initially, known NTase fold families and structures were identified from literature (including our previous classification (1)) and various databases of catalogued protein families (PFAM (32), COG (33) and KOG (34)) and structures (PDB (35) and SCOP (36)). This initial set was then used for a comprehensive, transitive searches for all NTase fold superfamily members using our distant homology detection method Meta-BASIC (30) and Gene Relational DataBase (GRDB) system, as described in our previous work (1). Briefly, Meta-BASIC is a highly sensitive meta-profile alignment method capable of finding very distant similarity between proteins through a comparison of sequence profiles

enriched by predicted secondary structures (meta-profiles). The GRDB system includes precalculated Meta-BASIC connections between 16 230 PFAM, 4825 KOG and 4873 COG families and 38 498 representative proteins of known structure (PDB90). Each family and each structure is represented by its sequence (PDB90) or consensus sequence (PFAM, COG and KOG), sequence profile (generated with PSI-BLAST using the NCBI NR70 derivative) and secondary structure profile (predicted with PSIPRED (37)).

The structural diversity was analysed for all collected NTase superfamily structures clustered at 90% sequence identity. Structures were divided into groups based on their structural similarity using DALI (38). Structure-based alignments were generated for all considered domains (including both the conserved NTase fold and additional N- and C-terminal domains) after manually curated superposition of their structures. Secondary structures were assigned with DSSP (39).

3D model building

Potential templates were identified with Meta-BASIC and the consensus of fold recognition 3D-Jury approach (40) using human FAM46 proteins as queries. The sequence-to-structure alignment between FAM46 family members and all representative structures possessing both NTase and PAP/OAS1 SBD domains was built using a consensus alignment approach and 3D assessment (41) based on Meta-BASIC and 3D-Jury results, PSIPRED secondary structure predictions and conservation of critical active site residues and hydrophobic patterns. The 3D model of human FAM46C protein was built with MODELLER (42) using *Trypanosoma brucei* TUTase 4 (pdb|2ikf) (43) as a template. Finally, the side chain rotamers were optimized using SCWRL3 (44). The overall quality of the modelled structure was checked with ProSA (45). Structure visualization was carried out with Pymol (http://www.pymol.org).

Analysis of protein interactors

Human members of the NTase fold superfamily were identified from the UniProt database (46) using our transitive Meta-BASIC search strategy as described above, starting with all collected NTase fold families and structures. Proteins interacting with human NTase superfamily members were identified using the BioGRID database (version 3.4.133) (47). GO annotations (molecular function and biological process) (48) for detected interactors were taken from the UniProt database. FAM46 interacting partners were also identified with manual literature searches.

Analysis of single point mutations in cancers

Missense mutations, found in cancer patients, in FAM46 genes were collected from publications and the following databases: cBioPortal (49), ICGC (50) and IntOGen (51). The sequence conservation in FAM46 family was measured based on Jensen–Shannon divergence (JSD) (52) using created FAM46 multiple sequence alignment. The JSD quantifies the similarity between probability distributions with scores ranging from 0 to 1 (53). A background amino acid distribution, estimated from a large sequence set, is used to approximate the distribution of amino acid sites subject to no evolutionary pressure. Positions in an alignment that are found to have amino acid distributions very different from the background distribution are proposed to be functionally important or constrained by evolution. The JSD score was computed using the score_conservation.py program (52) with default parameters, e.g. using BLOSUM62 for the background distribution. Positions in FAM46 multiple sequence alignment with more than 30% gaps were omitted from JSD computations.

Phylogeny

In order to visualize the relationships between FAM46 family members, a phylogenetic analysis was performed with PhyML3.0 (54) using the LG and JTT models, with an estimated gamma parameter and proportion of invariable sites. An approximate branch support was calculated using the aLTR (55) option implemented in PhyML. Branches with supports lower than 0.5 were collapsed. The trees were drawn using iTol (56).

RESULTS AND DISCUSSION

FAM46 family

Firstly, we identified proteins belonging to FAM46 family with an exhaustive PSI-BLAST (16) searches performed against the NCBI NR protein sequence database using all four human FAM46 paralogs (FAM46A, FAM46B, FAM46C and FAM46D) as queries. These searches quickly converged at the third iteration; however, most family members can be easily detected even with a simple BLAST search. This is a feature of compact and very conserved protein families and graphical clustering of all these sequences corroborated this observation. As many of them turned out to be variants or mutants of the same protein (e.g. there are four FAM46 genes in the human genome and 14 proteins in the NR database) we selected 868 protein sequences unique for each organism using sequence clustering at different thresholds followed by manual assessment. It should be noted that some of the detected proteins contain long deletions within conserved regions, what might be due to erroneous gene/exon prediction.

Taxonomic distribution

FAM46 proteins are present in the proteomes of all animals. Supplementary Figure S1 summarizes the taxonomic distribution of all selected 868 FAM46 proteins unique for each organism. Four FAM46 paralogs can be identified in almost all sequenced Vertebrata (with high-quality genomes), but not in Tunicata (*Ciona intestinalis* and *Oikopleura dioica*), Hemichordata (*Saccoglossus kowaleskii*), Echinodermata (*Strongylocentrotus purpuratus*) or Cephalochordata (*Branchiostoma floridae*), which encode only a single FAM46 protein. Specifically, amphibian, bird, reptile and mammal genomes harbour four distinct FAM46 genes. On the other hand, fish proteomes contain six to seven FAM46 paralogs due to lineage specific duplications followed by fast differentiation of the retained paralogs (different in

each of the four analysed fishes). This evolutionary scenario has been already described in teleost fishes (57). An asymmetric acceleration of evolutionary rate in one of the paralogs after the duplication event, manifested by the high protein sequence divergence and usually leading to alignment problems in less conserved regions, was also observed in FAM46 paralogs. FAM46 family members are encoded in all sequenced animal phyla ranging from Arthropoda (Daphnia, Drosophila), Mollusca (*Crassostrea gigas*), Nematoda (*Caenorhabditis elegans*, *Brugia malayi*, *Loa loa*, *Trichinella spiralis*), Platyhelminthes (*Schistosoma mansoni*, *Clonorchis sinensis*), Cnidaria (*Nematostella vectensis*), Placozoa (*Trichoplax adherens*) and Porifera (*Amphimedon queenslandica*). FAM46 genes duplicated and diverged strongly in some Nematoda lineages leading to a variable number of paralogs in the analysed genomes. Moreover, FAM46 proteins are detectable in close metazoan relatives: Choanoflagellida (*Salpingoeca* sp., *Monosiga brevicollis*) and Ichthyosporea (*Sphaeroforma arctica*). Surprisingly, proteins belonging to this family can be also found in Amebozoa (four Entamoeba species, *Polysphondylium pallidum*, *Acytostelium subglobosum* and three Dictyostelium species and *Acanthamoeba castellanii*) and one Diplomonadida (*Guillardia theta*). Choanoflagellida and Ichthyosporea, together with Metazoa, are a sister group of Fungi and Nucleariids. Noteworthy, FAM46 family members are absent in fungal and nucleariid genomes sequenced within the Origins of Multicellularity Project by BROAD. Amebozoa are sometimes grouped with Opisthokonta (Metazoa and Fungi) into a supertaxon Unikonta characterized by a single posterior flagellum in flagellated cells. Summarizing, the presence of FAM46 proteins in proteomes of Metazoa, Choanoflagellida and Amebozoa suggests its origin in the ancestor of Unikonta with further divergence into four distinct conserved representatives in vertebrates.

Phylogeny inference

Phylogenetic relationships were analysed both for a set of 29 representative sequences and for a set of 868 Metazoa, Choanoflagellida, Diplomonadida and Amebozoa FAM46 proteins. Entamoeba, Giardia and Dictostellids form well-separated clades with uncertain branching order (Figure 1). They are, however, clearly separated from the Metazoa-Choanoflagellida clade. *Salpingoeca rosetta* and *M. brevicollis* form a sister clade to Metazoa. Some of invertebrate FAM46 proteins display higher variability at sequence level that can lead to long branches on the phylogenetic tree. The position of basal lineages within the Metazoa is uncertain and possible involvement of long branch attraction phenomenon should be taken into account.

The evolutionary history of FAM46 in the vertebrate genomes is a story of consecutive duplications leading to four highly similar paralogs. All vertebrate genomes analysed in this study retained all four FAM46 paralogs. Surprisingly, we detected the presence of FAM46 proteins in all sequenced Amebozoa genomes.

The divergence time between Choanoflagellida and Metazoa is estimated ∼600MYA (58,59). As FAM46 genes are present in Choanoflagellida, Metazoa and Amebozoa genomes it is possible they were already in the ancestor of all Unikonta, and therefore also in the ancestor of Ophistokonta. The most likely scenario involves an ancient deletion in the ancestor of Fungi. This evolutionary history claims FAM46 would be a very ancient gene, what is not reflected in its encoded amino acid sequence divergence. Provided FAM46 have an ancient origin early in the Unikonts, still the presence of two FAM46 genes in the *G. theta* genome requires clarification. Due to cohabitation, it is plausible that the FAM46 genes in *G. theta* genome appeared via horizontal gene transfer from Choanoflagellida to Diplomonadida. However, the low resolution of these deep branches renders the FAM46 phylogenetic tree (Supplementary Figure S2) uninformative for HGT inference. There is significant evidence for the transfer occurring in the opposite direction (60). The mechanism underlying algae to choanoglafellate transfer is supposed to be based on phagotrophy. We have insufficient data to hypothesize about the possibility of transfer happening from choanoflagellates to algae.

Gene structure

The organization, architecture and regulation of human FAM46 genes and their homologous counterparts in others organisms seem to contribute to their functional diversification. For instance, human FAM46 genes contain 2–3 exons of which 1–2 are coding (Supplementary Table S1) and they encode up to five different transcripts. In addition, antisense transcripts have been also detected, e.g. for human FAM46A (61). Interestingly, we found that the H3K27Ac pattern in the promoter region and along the coding region is completely different for each of human FAM46 genes, what can be related to distinct nucleosome density in these chromatin regions and different expression patterns (62). Additionally, the FAM46D gene is in a repeat dense area as denoted by RepeatMasker, which might be related to the overall chromosome X repeat density.

Domain architecture

Given the high variability of human FAM46 paralogs at the gene organization level, the encoded proteins are surprisingly similar at the sequence level, including the conservation of various motifs. FAM46 proteins seem to have a common two-domain architecture composed of an α/β region (according to secondary structure predictions) followed by an α-helical region. It should be noted that the FAM46 family is a distant outlier in the NTase fold superfamily and cannot be identified with standard sequence comparison methods such as PSI-BLAST. Using a highly sensitive tool for distant homology detection, Meta-BASIC (30), we mapped FAM46 proteins, with the above threshold scores, to several 3D structures including the terminal uridylyl transferase 4 (TUTase 4) from *T. brucei* (pdb|2ikf) (43), 2′-5′-oligoadenylate synthase (OAS) from *S. scrofa* (pdb|1px5) (63), CCA-adding enzyme from *A. fulgidus* (pdb|4x4n) (64), cyclic AMP-GMP synthase from *V. cholerae* (pdb|4u0n) (65), aminoglycoside 6-adenyltransferase from *B. subtilis* (pdb|2pbe) and nuclear factors NF90 and NF45 from *M. musculus* (pdb|4at7) (66). Importantly, FAM46 N-terminal α/β region has weak but

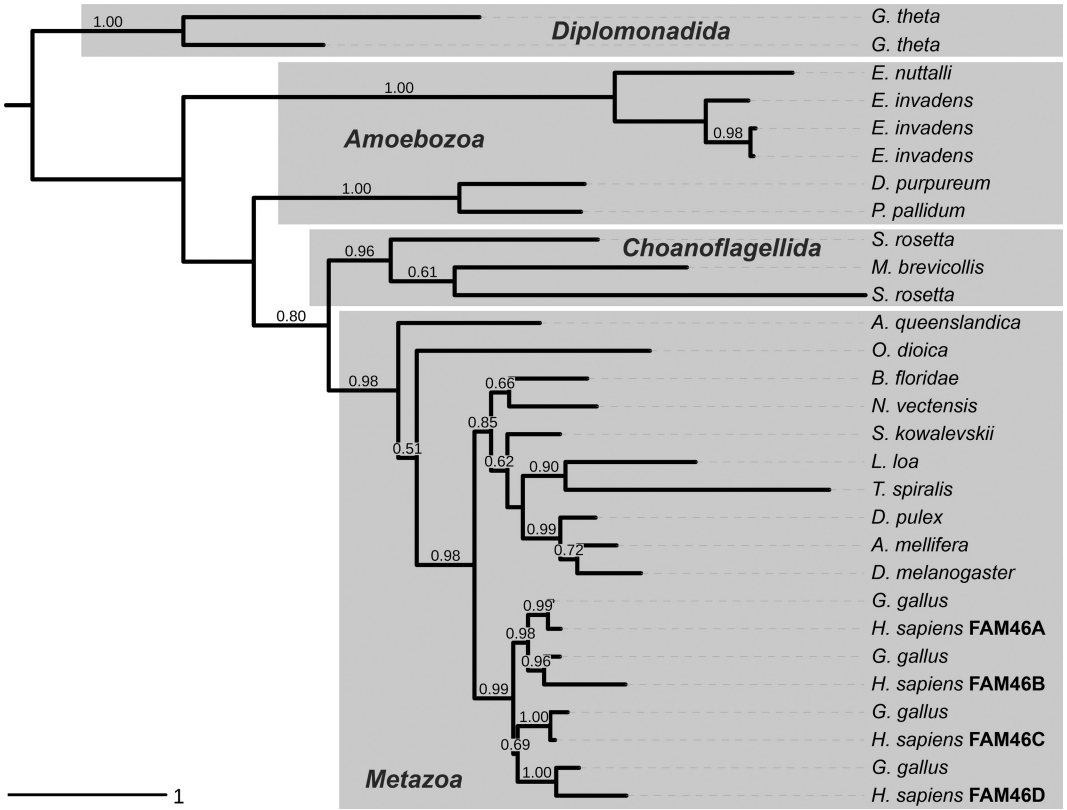

Figure 1. Phylogenetic tree of representative FAM46 protein sequences. Maximum likelihood (ML) analysis for selected 29 family members was carried out using the LG+G model. The approximate likelihood ratio test Shimodaira–Hasegawa-like (SH-like) branch supports above 0.5 are shown. Branches with support lower than 0.5 were collapsed.

evident sequence similarity to the NTase domain, which can be confirmed by several fold recognition servers. Meta-BASIC suggested that the C-terminal α-helical part may be similar either to poly(A) polymerase/2′-5′-oligoadenylate synthetase 1 substrate binding domain (PAP/OAS1 SBD) or the domain of four-helical up-and-down bundle fold (4H), however, it assigned below threshold scores to these predictions. To figure out what protein fold is adopted by the FAM46 C-terminal region, we performed a comprehensive analysis of the structural diversity of all the available NTase superfamily structures, both for their conserved NTase and additional N- and C-terminal domains (Supplementary Figure S3).

While both the PAP/OAS1 SBD and 4H domains possess four core α-helices C-terminal to NTase domain, only PAP/OAS1 SBD retains the additional (the first core) α-helix located before NTase domain. According to secondary structure predictions, this helix is clearly seen in the FAM46 family members in the conserved region preceding the predicted NTase domain. In addition to a good mapping of predicted and observed core secondary structure elements, FAM46 proteins display also similar conservation of hydrophobic motifs and critical residues for NTP binding (see below) characteristic for the PAP/OAS1 SBD. In our previous studies we showed that such detailed analysis of below threshold Meta-BASIC hits usually enables identification of highly diverged superfamily members which escape detection even with advanced sequence comparison methods. For instance, using this approach we identified restriction endonuclease-like (67) and RNase H-like (68) domains in many uncharacterized and poorly annotated protein families. Finally, we found that proteins embracing both the NTase and 4H domains are mainly encoded in bacteria and rarely found in archeal genomes, with single representatives identified in eukaryotic species, including fungal *Myceliophthora thermophila* (gi|367020986), *Tribulus terrestris* (gi|367039397) and *Rhizophagus irregularis* (gi|552919075), and soil-living amoeba *Dictyostelium discoideum* (gi|66821023). This is consistent with the biological functions played by these proteins, as they participate in antibiotic resistance (e.g. *Staphylococcus aureus* kanamycin nucleotidyltransferase (69), *Enterococcus faecium* lincosamide antibiotic adenylyltransferase (70), *Bacillus subtilis* aminoglycoside 6-adenyltransferase) and nitrogen assimilation (e.g. *Escherichia coli* glutamine synthetase adenyltransferase (71)). In contrast, NTases possessing the PAP/OAS1 SBD can be widely found in eukaryotes. Altogether, results of all these analyses strongly suggest that although displaying little sequence similarity, FAM46 proteins possess PAP/OAS1 SBD consisting of the five right-handed twisted α-helices (with an α1-NTase-α2α3α4α5 topology).

In addition, we found that a few FAM46 proteins possess additional domains inserted inside the NTase domain or located at N- or C-termini (Supplementary Figure S4). It should be noted, however, that the presence of some of these

additional domains may be a result of potentially incorrect gene/exon predictions.

PAP/OAS1 SBD in known NTase structures

To identify conserved PAP/OAS1 SBD residues, critical for binding NTP substrate in an NTase active site, we carried out an exhaustive sequence and structure analysis by generating the structural alignment of all the representative structures possessing both the NTase and PAP/OAS1 SBD domains (Figure 2). The PAP/OAS1 SBD specifically binds a nucleobase of the incoming NTP mainly by amino acids that provide, either directly or indirectly via water molecules, Watson–Crick hydrogen bonds. In addition, a conserved hydrophobic amino acid (e.g. V234 in poly(A) polymerase Pap1 (72) and Y212 in poly(U) polymerase Cid1 (73)), located at the beginning of the third core α-helix of PAP/OAS1 SBD, forms a flat hydrophobic surface for the incoming NTP nucleobase. Proteins containing the PAP/OAS1 SBD also possess another common residue, which is responsible for the recognition of a triphosphate moiety. Conserved lysine/arginine (e.g. K215 in Pap1) located in the second core α-helix of PAP/OAS1 SBD, together with a serine from the NTase domain hG[GS] motif, interact with NTP β- and γ-phosphate groups.

3D model of human FAM46C

Initially, we generated a sequence-to-structure alignment of FAM46 family members with all representative proteins of known structure possessing both the NTase and PAP/OAS1 SBD domains (using their structure-based alignment described above) (Figure 2). Although these structures display very little sequence similarity to the FAM46 proteins, in contrast to our previous work (1) where we focused in general on the most conserved regions of NTase fold common to all NTase superfamily members, here we were able to propose a reliable and complete sequence-to-structure alignment for all conserved regions of both domains. The alignment was guided by secondary structure predictions and conservation of (i) the NTase fold active site motifs, (ii) identified critical PAP/OAS1 SBD residues participating in substrate binding and (iii) hydrophobic patterns responsible for forming the hydrophobic core of the structure.

As a representative of FAM46 family for 3D modelling we selected human FAM46C, which is a potential biomarker for multiple myeloma. FAM46 proteins are very similar in sequence, for instance, four human paralogs share 56–75% amino acid identity within the common region encompassing both domains. In addition, the length of this region in these paralogs differs only by 1–2 residues.

The structure of *T. brucei* TUTase 4 (pdb|2ikf) (43), as assigned the highest Meta-BASIC score among the proteins possessing both NTase and PAP/OAS1 SBD domains, was used as a template to generate the 3D model of human FAM46C, based on the manually derived sequence-to-structure alignment. However, due to the lack of templates with similar insertion between the last core β-strand and the last core α-helix of NTase domain, we were unable to create a reliable model for 70 amino acids of FAM46C in this region. Nevertheless, we can speculate that this insertion in FAM46C should fill the space usually occupied by residues responsible for binding incoming NTP nucleobase and RNA 5′ end. Figure 3 presents a comparison of the FAM46C model and existing structures of non-canonical poly(A) polymerase Trf4p from *Saccharomyces cerevisiae*, which is a part of the Trf4p/Air2p/Mtr4p polyadenylation (TRAMP) complex (74), and the non-catalytic mitochondrial dynamic protein MiD51 from *M. musculus* (75). Importantly, in Trf4p, the region of 53 amino acids between the fourth and fifth core α-helices of the PAP/OAS1 SBD is crucial for binding the RNA 5′ end and a nucleobase of the incoming NTP. The corresponding region in FAM46C is much shorter (only 11 amino acids) and probably is not able to form the interaction interface for a nucleobase. Therefore, it is likely that nucleobase binding residues are located within the 70 amino acids insertion between the last core β-strand and the last core α-helix of the FAM46C NTase domain. In addition, this conserved insertion, composed of the predicted two β-strands and two α-helices (with βαβα order), may also participate in protein–protein interactions similar to the MiD51 receptor which binds the dynamin-related protein 1 (Drp1) via a well-conserved loop located in the NTase domain (75). However, it should be noted that FAM46C, in contrast to MiD51, seems to be an active NTase; therefore, even if the insertion is responsible for protein–protein interactions, it should also play a role in substrate recognition.

Active site

Figure 4 shows a comparison of active sites of human FAM46C (model), poly(A) polymerase Pap1 from *S. cerevisiae* (72), poly(U) polymerase Cid1 from *Schizosaccharomyces pombe* (73), CCA-adding enzyme from *A. fulgidus* (64), 2'-5'-oligoadenylate synthetase OAS1 from *S. scrofa* (76) and the cyclic GMP-AMP synthase (cGAS) from *M. musculus* (77). FAM46 proteins probably function as active NTases because they share all the key motifs in the NTase domain responsible for catalysis and substrate binding, including the [DE]h[DE]h and h[DE]h patterns with three conserved aspartate/glutamate residues (Asp90, Asp92 and Glu166) and hG[GS] motif with Gly73 and Ser74 in human FAM46C. Although we were not able to generate a complete 3D model of human FAM46C, we were able to identify additional residues responsible for NTP binding, which are located in the conserved secondary structure elements. Comparison of active sites of experimentally solved structures showed that proteins encompassing both NTase and PAP/OAS1 SBD usually bind a nucleobase or a ribose-moiety of incoming NTP by a serine or a threonine located just before or in the last core α-helix of NTase domain (Figures 2 and 4). Therefore, it is possible that FAM46C Ser248 may bind, directly or indirectly via a water molecule, 2′-OH or/and 3′-OH hydroxyl group of a ribose-base moiety as it is observed for Thr172 in poly(U) polymerase Cid1 (73) or it can participate in a nucleobase binding similarly to Thr190 in 2′-5′-oligoadenylate synthase OAS1 (76). FAM46C shares also all the conserved residues in PAP/OAS1 SBD responsible for substrate binding. The FAM46C Leu282 probably interacts with a nucleobase of

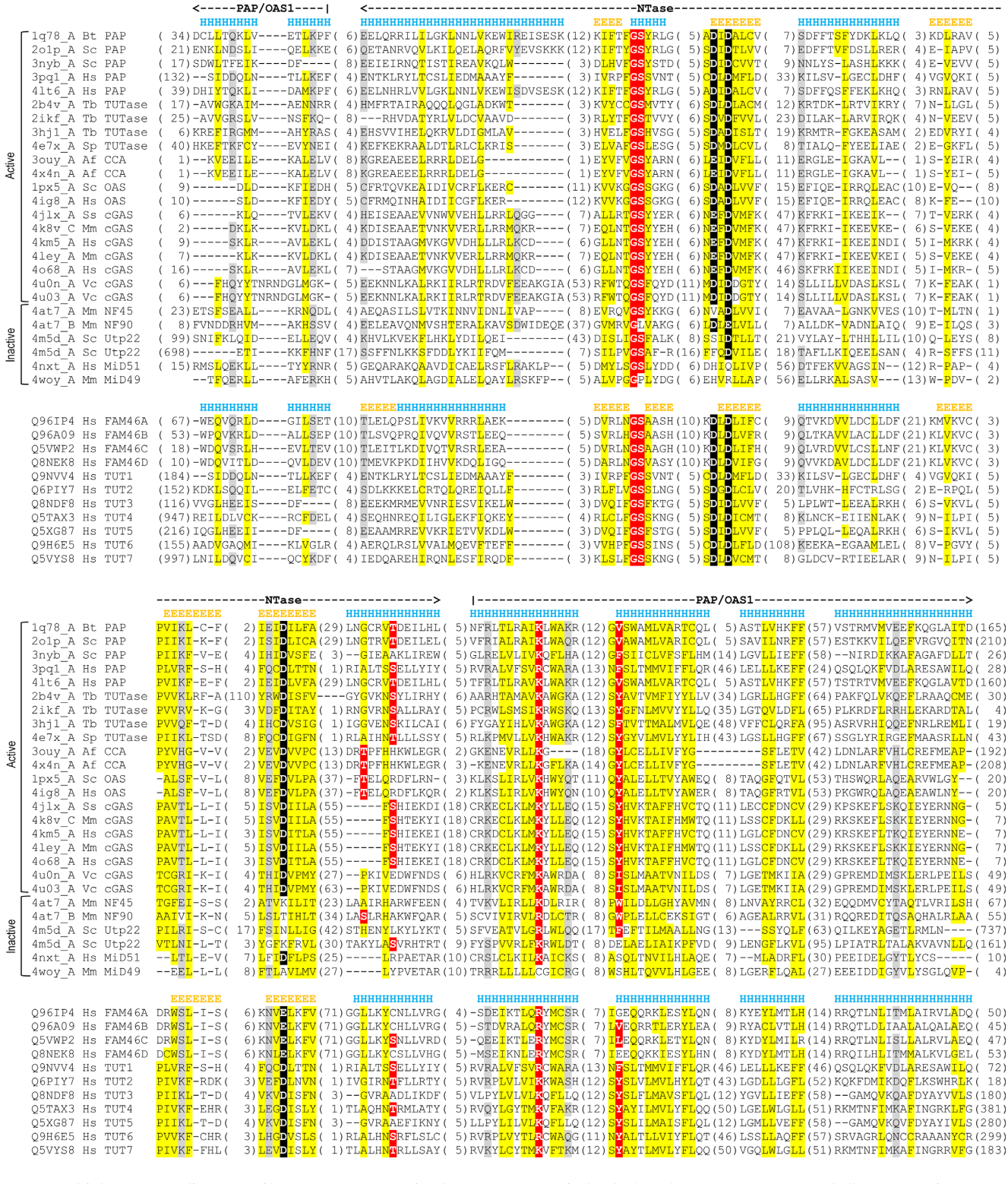

Figure 2. Multiple sequence alignment of human FAM46 proteins, human non-canonical poly(A) polymerases (TUT1-7) and all representative structures possessing both the NTase and PAP/OAS1 SBD domains. Only conserved regions of the domains are shown. Sequences are labelled with PDB code or UniProt ID. The numbers of excluded residues are specified in parentheses. Residue conservation is denoted with the following scheme: uncharged, highlighted in yellow; polar, highlighted in grey; invariant active site residues involved in catalysis, highlighted in black; critical substrate binding residues, highlighted in red. Locations of observed and predicted secondary structure elements are marked above the corresponding alignment blocks. Abbreviations: PAP, poly(A) polymerase; TUTase, terminal uridylyltransferase; CCA, CCA-adding enzyme; OAS, oligoadenylate synthetase; cGAS, cyclic GMP-AMP synthase; NF45 and NF90, nuclear factors NF45 and NF90; Utp22, U3 small nucleolar RNA-associated protein 22; MiD51 and MiD49, mitochondrial dynamics proteins MiD51 and MID49; Ss, *S. scrofa*; Tb, *T. brucei*; Af, *A. fulgidus*; Hs, *H. sapiens*; Mm, *M. musculus*; Sp, *S. pombe*; Sc, *S. cerevisiae*; Vc, *V. cholerae*; Bt, *B. taurus*. Sequence-to-structure alignment for FAM46 proteins can be assigned higher confidence in the NTase domain.

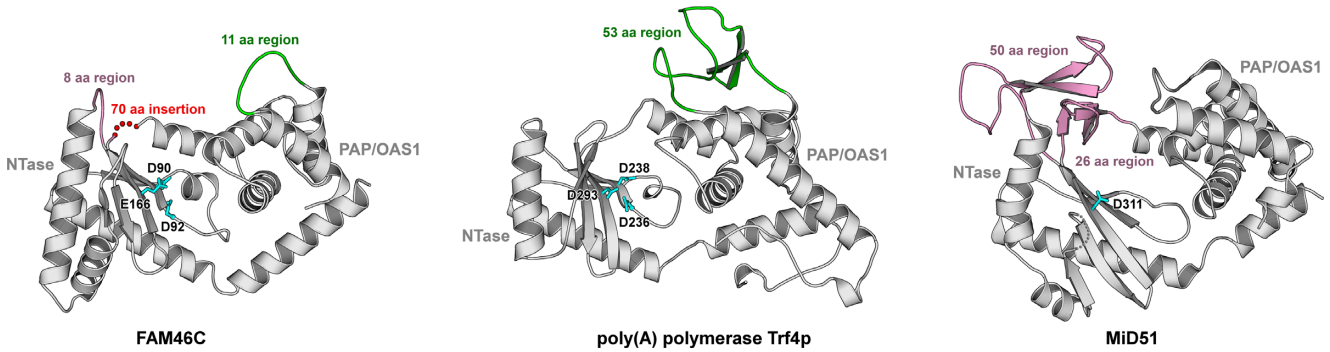

Figure 3. Comparison of 3D model of human FAM46C and available structures of non-canonical poly(A) polymerase Trf4p (pdb|3nyb) and mitochondrial dynamics protein MiD51 (pdb|4oaf). Regions in MiD51 responsible for protein–protein interactions and their potential counterpart in FAM46C are coloured pink. The region between the fourth and fifth core α-helices of PAP/OAS1 SBD in FAM46C and Trf4p (critical for nucleobase binding) is shown in green. The region not modelled in FAM46C (70 amino acids) is denoted by red dots. The conserved active site carboxylates are shown in blue.

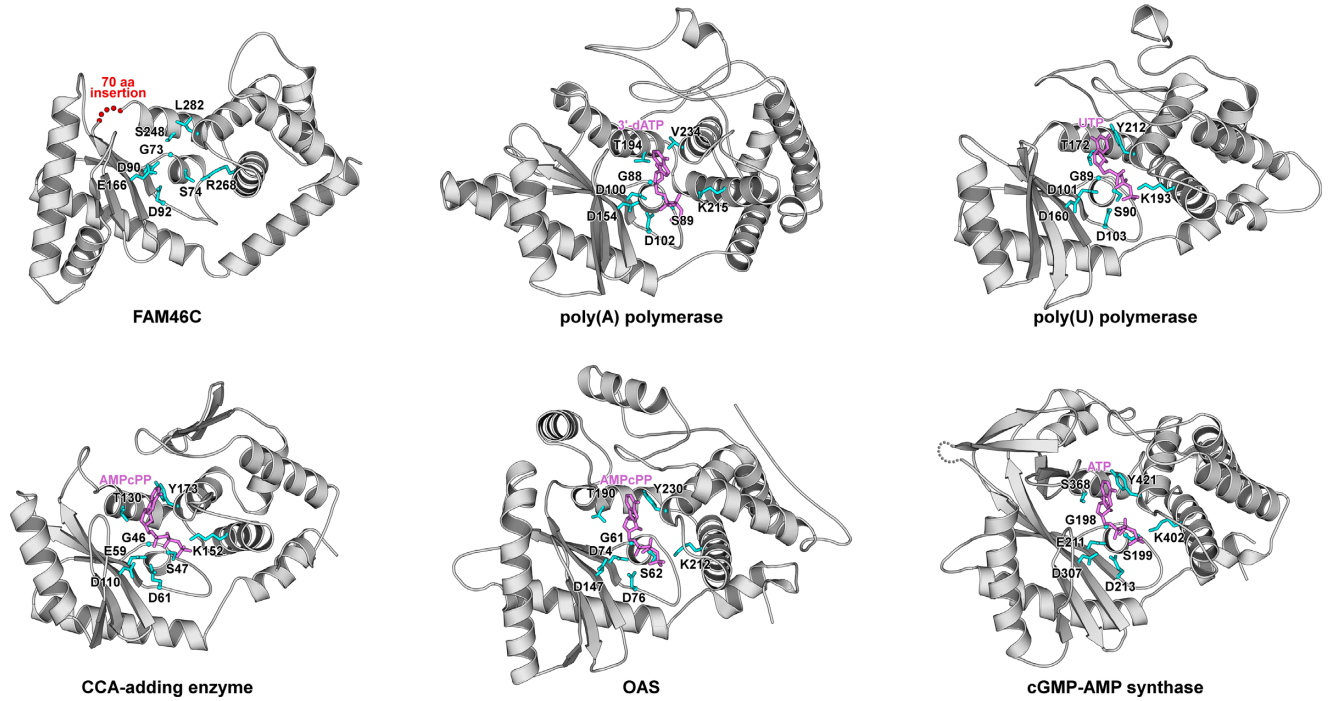

Figure 4. Comparison of the active sites of FAM46C, poly(A) polymerase (Pap1, pdb|1fa0), poly(U) polymerase (Cid1, pdb|4fhp), CCA-adding enzyme (pdb|4x4r), OAS (OAS1, pdb|4rwo) and cyclic GMP-AMP synthase (Mb21d1, pdb|4k97). Only NTase and PAP/OAS1 SBD domains are shown. The region not modelled in FAM46C (70 amino acids) is denoted by red dots. Conserved amino acids critical for catalysis and substrate binding are shown in blue.

the incoming NTP like Tyr212 in poly(U) polymerase Cid1 or makes van der Waals contacts with the ribose-base moiety of NTP similar to Val234 in poly(A) polymerase Pap1 (72). Finally, NTP β- and γ-phosphates most likely interact with the conserved Arg268 in addition to Ser74 from the hG[GS] motif.

Cellular localization and tissue specificity

According to various servers predicting subcellular localization, the FAM46 proteins seem to be localized in both the cytoplasm and nucleus. In addition, three human paralogs (FAM46B, FAM46C and FAM46D) harbour potential leucine-rich NES, located at the end of the C-terminal PAP/OAS1 SBD domain. As a consequence, it is likely that proteins belonging to the FAM46 family shuttle between the nucleus and cytoplasm.

We also analysed the gene expression data available in the BioGPS database (28) and found that each of the human FAM46 paralogs has a different tissue/cell expression pattern (Supplementary Table S1). Different expression patterns probably indicate various biological processes in which FAM46 proteins participate. According to the BioGPS database, FAM46A, FAM46B and FAM46C are potentially expressed in 81, 18 and 66 tissues/cells, respectively, while FAM46D can be found only in sperm (Supplementary Table S1).

Interacting partners

To get more hints at the biological function of FAM46, we compared all human NTase superfamily members according to GO molecular function and biological process of their protein interactors identified in the BioGRID database (47) (Supplementary Figure S5). We found that FAM46 binding partners share a common GO functions and processes mostly with interactors of human NTase fold proteins from the following four groups (described in Supplementary Table S2): interleukin enhancer-binding factors, non-canonical poly(A) polymerases (TUT), poly(A) polymerases (PAP) and zinc finger RNA-binding proteins (ZFR). Similarly to FAM46 family members, proteins belonging to these groups also possess additional PAP/OAS1 SBD domain. In addition, almost all of them retain the same biological function—DNA/RNA binding, including poly(A) RNA binding and participate in the same process—transcription.

We also analysed all 61 FAM46 protein interactors identified in the BioGRID database and the literature in more details (Supplementary Table S3). We found that each FAM46 paralog has a different set of interacting protein partners; therefore, it is likely that each paralog participates in a different biological process in the cell. Most of the interacting partners play important roles in development, including cellular proliferation and cell differentiation. We noticed that many FAM46 interactors share a common molecular functions (e.g. nucleic acids binding, including binding of the mRNA poly(A) tail) or biological processes (e.g. protein modification, transcription). Specifically, 26 FAM46 interacting partners may directly bind RNA and/or DNA. This group includes: transcription (co)factors (RHOXF2, TBX4, NR2F2, SOX5, NRF1, TRIP6), transcription activators (Znf322, Cxxc5), RNA stabilization factors (ELAVL1, BCCIP, HDLBP, Pabpc1, Pabpc4), proteins involved in transcription (POLR1A, POLR1E, POLR2J), proteins participating in mRNA translation (EIF4G3, Pabpc1, Pabpc4), proteins taking part in DNA repair (POLE2, Rad23b, WRAP53) and other proteins, which play various roles such as DNA helicase DDX11, putative RNA exonuclease (44M2.3), mitochondrial translation optimization factor 1 (MTO1), SUMO-conjugating enzyme UBC9 (UBE2I), 14-3-3 protein zelta/delta (Ywhaz) involved in signal transduction (78) and ATXN1, which may participate in RNA export (79). Similarly to human FAM46A (80), nine protein interactors (ELAVL1, BCCIP, EIF4G3, MTO1, Ywhaz, TRIP6, UBE2I, Pabpc1, Pabpc4) participate or may participate in RNA poly(A) tail binding. Another seven FAM46 interacting partners (EGLN2, DAZAP2, KEAP1, DCAF6, KLHDC2, ZNHIT6, MVP) were shown or suggested to cooperate with or regulate proteins, which are able to bind directly to nucleic acids. Specifically, the Egl nine homolog 2 (EGLN2) targets the transcriptional complex HIF-α subunit for proteasomal degradation (81). KEAP1 targets transcription factor Nrf2 for ubiquitination and degradation (82). DCAF6 enhances the transcriptional activity of nuclear receptors NR3C1 and AR (83). The physical interaction between KLHDC2 and a bZIP transcription factor (LZIP) leads to the repression of the LZIP-dependent transcription (84). DAZAP2 regulates stress or germ granules–ribonucleoprotein complexes (85,86). MVP is a part of an evolutionary highly conserved ribonucleoprotein particles (vaults) (87,88), while ZNHIT6 may be involved in snoRNP biogenesis (89). Another functional group of FAM46 interactors contains proteins, which take part in protein modification (mostly in protein degradation). This group embraces proteases (serine protease HTRA1, caspase-like protease ESPL1), Kunitz-type protease inhibitor 1 (SPINT1), Kelch-like ECH-associated protein 1 (KEAP1) participating in Nrf2 ubiquitination and degradation (82), EGLN2 targeting HIF-α for proteasomal degradation (81), E3 ubiquitin-protein ligases (RNF14, PARK2), CUL4-associated factor 6 (DCAF6) functioning as substrate-recruiting module for CUL4-DDB1 E3 ubiquitin-protein ligase complex (90), SUMO-conjugating enzyme UBC9 (UBE2I), ubiquitin carboxyl-terminal hydrolase 4 (USP4), a regulatory subunit 6A of the 26S proteasome (Psmc3) and BAG6, which is crucial in ubiquitin-mediated protein degradation of defective or misallocated polypeptides (91). Among all FAM46 interactors, we were able to identify two kinases: Polo-like kinase 4 (PLK4) and non-receptor tyrosine-protein kinase SYK (SYK). PLK regulates cell cycle progression, mitosis and cytokinesis (92), while SYK mediates signal transduction and differentiation, particularly in B-cell development (93,94). Another interacting partner, the FYVE domain-containing protein 9 (ZFYVE9) identified as a SMAD2/3-binding protein, may also regulate the proliferation of hepatic cells during zebrafish embryogenesis (95). The last identified functional group contains proteins responsible for an intra and extracellular cell transport, including DYNC1H1—a heavy chain of cytoplasmic dynein 1 (96,97), RIN3—a small GTPase, which participates in intracellular membrane trafficking (98) and AP2B1, which plays a pivotal role in many vesicle trafficking pathways within the cell (99).

Mutations in cancers

Recent studies have identified numerous somatic mutations in various cancer patients leading to single point mutations in the human FAM46 proteins (Supplementary Table S4). For instance, the human FAM46C gene was reported as a causal driver of multiple myeloma (6). In addition, a single FAM46C mutation (Y247N) was identified in *hem6* mice with hypochromic anemia, which affects terminal spermiogenesis and terminal stages of erythroid differentiation (100). This study showed that male *hem6* mice produce sperm with defects detectable by phase contrast microscopy and fluorescence microscopy. To analyze the role of these mutations, we mapped them onto a 3D model of FAM46C (Figure 5) and found that they can be divided into two groups. The first group includes mutations that are located in a highly conserved area of the active site and its close vicinity, and probably may decrease/increase FAM46 catalytic activity and/or affect substrate binding (e.g. change the preference for the type of incorporated NTP). This group embraces the mouse *hem6* mutation and the majority of mutations found in multiple myeloma patients as well as several mutations from other cancers. The average JSD (52) score for all mutated amino acid positions in multi-

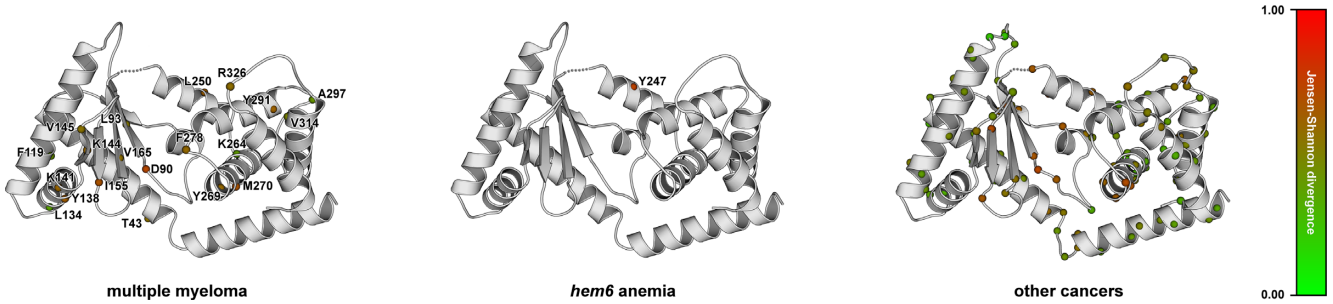

Figure 5. Missense mutations in FAM46 family members found in cancer patients and *hem6* mouse. The positions of the corresponding single point mutations, mapped onto a 3D model of human FAM46C, are shown as spheres. The spheres are coloured according to JSD score, which refers to the amino acid conservation in FAM46 family.

ple myeloma (reported in Supplementary Table S4) and for the mutated residue 247 in *hem6* FAM46C is 0.56 and 0.74, respectively (higher JSD scores correspond to higher sequence conservation). In benchmarks JSD approach, which considers also estimated conservation of sequentially neighbouring sites, performed better than traditional measures (e.g. Shannon entropy or Sum-of-pairs measure) in identifying functionally important residues (52). In comparison, the average JSD score for five FAM46 active site residues: glutamic/aspartic acids, glycine and serine is 0.65. Mutations belonging to the second group are located mostly on the protein surface (usually with JSD scores below 0.4), in evolutionary low conserved regions. Those mutations may affect protein–protein interactions or alternatively might not play any crucial role in the reported cancers.

In addition, we selected all mutations of highly conserved residues with a JSD score higher than 0.65 (the average JSD for five FAM46 active site residues from conserved NTase motifs). It allowed us to identify mutations found in a number of malignancies (highlighted in orange in Supplementary Table S4), which probably have the largest impact on protein activity and may be connected with diagnosed cancers. Consequently, we suggest that, in addition to multiple myeloma, FAM46 genes may be also involved in pathogenesis of various other cancer subtypes including liver hepatocellular carcinoma, bladder urothelial carcinoma, head and neck squamous cell carcinoma, uterine corpus endometrial carcinoma, kidney renal papillary cell carcinoma, lung adenocarcinoma, ductal adenocarcinoma, colorectal adenocarcinoma, primary plasma cell leukemia and skin cutaneous melanoma.

Potential function

The results of our sequence and structure analyses suggest that the FAM46 proteins are active NTases, which have both the NTase fold and PAP/OAS1 SBD domains. Active NTases possessing the PAP/OAS1 SBD are known to participate in tRNA maturation (CCA-adding enzymes), RNA degradation (TUTases, poly(A) polymerase in TRAMP complex), mRNA maturation (poly(A) polymerases) and in a defense response to viruses and bacteria (2′-5′-oligoadenylate synthases and cyclic GMP-AMP synthases). Although it was shown that both FAM46A and FAM46C are induced by interferon I and II (14) and that FAM46C is one of the interferon-stimulated genes (ISGs) which modify viral (YFV and VEEV) replication during infection, it is unlikely that FAM46 proteins are antiviral enzymes like OASes. Unlike replication inhibiting ISGs (such as OASL, Mab-21 and C6orf150), FAM46C slightly enhances the replication of certain viruses (14). FAM46 family members do not possess also an $H(X_5)CC(X_6)C$ motif (conserved among vertebrate cGAS members) located between the NTase and PAP/OAS1 SBD domains. This motif, which resembles most closely HCCC-type zinc-ribbons found in TAZ domains, is required for efficient cytosolic DNA recognition (101). Finally, we investigated the possibility that FAM46 proteins may be novel non-canonical poly(A) polymerases participating in RNA 3′ end modification like TUTases or poly(A) polymerases GLD-2 and GLD-3. This hypothesis is consistent with M. Tian studies (100), where it was shown that mutated FAM46C may modulate the poly(A) tails of specific transcripts during erythroid differentiation. The author identified a single FAM46C mutation (Y247N) in *hem6* mice and showed that it might cause an accelerated, progressive shrinkage of the poly(A) tail in four transcripts (alpha-globin, Alas2, Hbb-b1 and Ftl1) and probably does not have any effect on poly(A) tails in two transcripts (Fth1 and beta-actin). Additionally, a Y247N mutation led to an increase of expression levels of 152 transcripts, resulted in a decrease of expression levels of 29 transcripts, and did not have any effect on 29 erythroid transcripts (100). It should be noted, however, that M. Tian used an indirect approach to analyze the poly(A) tail lengths in the six aforementioned transcripts. His strategy of indirect poly(A) tail length assay assumed that the poly(A) tails do not possess any nucleotides other than adenosines; therefore, he was not able to identify the real length of the modified poly(A) tails if they also contain other nucleotides. Thus, it is likely that he observed shortening of mRNA poly(A) tails for specific transcripts if FAM46C is responsible for the addition of adenosines to the RNA 3′ end. The FAM46C mutation (Y247N) might have weakened the processivity of FAM46C resulting in poly(A) tails shrinkage. On the other hand, FAM46C may be a non-canonical poly(A) polymerase which adds cytidines or uridines to the RNA 3′ end. In this scenario, FAM46C may participate in transcript degradation by modifying the poly(A) tails. This hypothesis is consistent with the fact that up to 152 transcripts increased expression levels in a *hem6* mutant. In this case, the observed shortening of the poly(A) tails in four

transcripts may be a side effect of cell deregulation. In both scenarios, FAM46 proteins may play a very important role in mRNA stability as active non-canonical poly(A) polymerases rather than some other factors, which prevent early mRNA degradation by disrupting interactions between ribonuclease docking complex and RNA as suggested by M. Tian (100). Our functional assignment is also in line with the facts that mouse FAM46C may bind directly or through a complex to RNA CU-rich motifs (100) and FAM46A may bind to poly(A) tails (80). According to Chapman *et al.*, the expression of FAM46C is highly correlated with the expression of ribosomal proteins and initiation and elongation factors involved in protein translation (6). They proposed that FAM46C is functionally related in some way to the regulation of translation (e.g. as a mRNA stability factor), however, they did not assign any exact function to this protein. Recent studies revealed that the poly(A) tail length impacts gene expression in some processes such as inflammation, learning and memory (102), and there is a clear correlation between the poly(A) tail length and translational efficiency in early development stages in zebrafish and African clawed frogs (103). Therefore, it is possible that the correlation observed by Chapman *et al.* is the effect of length change of the poly(A) tails.

Both the M. Tian and Chapman *et al.* studies are consistent with the results of our analysis of FAM46 interactors and interacting partners of all remaining human NTase fold proteins. We found that FAM46 binding partners share a common GO functions and processes mostly with interactors of those active NTase fold superfamily members which belong to non-canonical poly(A) polymerases and poly(A) polymerases. We showed that over half of the 61 identified FAM46 interactors participate in DNA and/or RNA binding, including nine proteins which can bind mRNA poly(A) tails. Many of FAM46 interacting partners are involved in transcription or translation, like transcription (co)factors (RHOXF2, TBX4, NR2F2, SOX5, NRF1, TRIP6), transcription activators (Znf322, Cxxc5), RNA stabilization factors (ELAVL1, BCCIP, HDLBP, Pabpc1, Pabpc4), proteins involved in transcription (POLR1A, POLR1E, POLR2J), proteins participating in mRNA translation (EIF4G3, Pabpc1, Pabpc4), and proteins taking part in transcription regulation (EGLN2, KEAP1, DCAF6, KLHDC2) and mitochondrial translation optimization (MTO1). Finally, some FAM46 protein interactors regulate or are a part of ribonucleoprotein complexes.

Our domain architecture analysis revealed that proteins belonging to FAM46 family possess only two domains: NTase and PAP/OAS1 SBD, with single exceptions of some additional domains present in a few proteins. Importantly, we were not able to detect any additional conserved domains such as ferredoxin-like, which plays a critical role in processivity of canonical poly(A) polymerases or TUTases (Supplementary Figure S3). The ferredoxin-like domain provides additional interactions with RNA and may enhance its binding, allowing the NTase enzyme to add up to several hundred nucleotides. Therefore, FAM46 proteins acting as non-canonical poly(A) polymerases probably can add only a few nucleotides to the RNA 3′ end.

FAM46 family members seem to be localized both in the cytoplasm and nucleus, like two other human non-canonical poly(A) polymerases, PAPD4 and PAPD5 (104). Considering the physiological functions of FAM46 interactors, we can speculate about the biological processes, in which FAM46 proteins may participate. FAM46A probably cooperates with a subunit RPB11-a of DNA-directed RNA polymerase II, eukaryotic translation initiation factor 4 gamma (eIF4G), high-density lipoprotein-binding protein (HDLBP, Vigilin), while FAM46C may bind to polyadenylate-binding proteins (Pabpc1, Pabpc4) and (together with FAM46A) to ELAV-like protein 1. As a consequence, proteins belonging to FAM46 family can be involved in mRNA (de)stabilization either in the nucleus or cytoplasm. DNA-directed RNA polymerase II transcribes all protein-coding genes and synthesizes many functional non-coding RNAs. The eIF4G3 subunit is a scaffold protein in eIF4F complex, which participates in the recruitment of eukaryotic mRNAs to the ribosome (105). Pabpc1 and Pabpc4 belong to cytoplasmic poly(A) binding proteins (PABPC), which bind specifically to the poly(A) tail of mRNA and are required for poly(A) shortening, ribosome recruitment and translation initiation (106). Another protein interactor, Xenopus Vigilin, can selectively protect *in vitro* vitellogenin mRNA from cleavage by endonuclease PMR-1 (107), while ELAVL1 is described in the literature usually as a stabilization factor, which prevents the degradation of mRNAs possessing short tails (108–110). FAM46 proteins can be also involved in a ribosome biogenesis (like POLR1A, POLR1E interactors (111,112)) or they can (de)stabilize a nuclear pool of extra-ribosomal RPL23 and the pre-60S trans-acting factor eIF6 (like BCCIP interactor (113)). By interacting with telomerase Cajal body protein 1 (WRAP53), FAM46 family members may change 3′ ends of small Cajal body RNAs, which are involved in modifying splicing RNAs (114). Together with the Box C/D snoRNA protein 1 (ZNHIT6), they may also participate in snoRNP biogenesis, which is essential for the processing and modification of rRNA (89). Finally, FAM46 proteins (together with DAZAP2 and major vault proteins (MVP)) may modify RNAs which build ribonucleoproteins complexes like stress granules (85) and vaults composed of MVP, vault poly(adenosine diphosphate-ribose) polymerases (VPARP), telomerase-associated proteins (TEP1) and small untranslated RNAs (vRNAs) (87,88).

The FAM46 family members seem to be highly regulated proteins. The process, in which these new non-canonical poly(A) polymerases participate, is probably determined by their tissue-specific expression and gene organization. As reported in the BioGPS database, tissue expression levels are different for each human FAM46 paralog. Moreover, the human FAM46 proteins are likely to be regulated by phosphorylation. Each human paralog has many phosphorylation patterns detectable with ELM predictor (27) with high probability scores (data not shown). For instance, FAM46A, FAM46B and FAM46C have two potential phosphoserine sites (a LIG_PLK pattern) recognized by the Polo-like kinase, which is a known human FAM46C interacting partner.

CONCLUSION

A comprehensive analysis of various biological information available in literature and databases combined with numerous sequence and structure analyses (including a state-of-the-art distant homology detection, fold recognition and 3D modelling) allowed us to propose that FAM46 members function as cytoplasmic and/or nuclear non-canonical poly(A) polymerases. Four human FAM46 paralogs thus complement the group of already known non-canonical poly(A) polymerases in humans embracing seven proteins: RBM21 (U6 TUTase, Star-PAP, TUT6), hGLD2 (PAPD4, TUT2), hmtPAP (PAPD1, TUT1), POLS (TUT5), PAPD5 (TUT3), ZCCHC6 (TUT7) and ZCCHC11 (TUT4). ZCCHC6 and ZCCHC11 mono-uridylate the 3′ end of specific miRNAs involved in cell differentiation and Homeobox (Hox) gene control (115). The hmtPAP produces poly(A) tails in mitochondria (116). The RBM21 catalyzes the uridylation of U6 snRNA involved in pre-mRNA splicing (117). The hGLD2 generates poly(A) tails of selected cytoplasmic mRNAs (118). The PAPD5 participates in the polyadenylation-mediated degradation of aberrant pre-rRNA and in replication-dependent histone mRNA degradation (119). Unfortunately, we are not able to predict the exact type of RNA that can be modified by FAM46 proteins. However, taking into account all the identified FAM46 interacting partners, we can speculate that FAM46 proteins could modify the 3′ end of mRNAs, small Cajal body RNAs and vRNAs. In addition, they may also participate in snoRNP and ribosome biogenesis, and (de)stabilize a nuclear pool of extra-ribosomal RPL23 and the pre-60S trans-acting factor eIF6.

The FAM46 family members as well as all the known non-canonical poly(A) polymerases share the two following domains: a PAP/OAS1 SBD with an inserted NTase domain right after the first core α-helix. In this work, we showed that proteins with such domain architecture, in addition to highly conserved NTase domain patterns ([DE]h[DE]h, h[DE]h and hG[GS]), possess also three additional, conserved amino acids critical for NTP binding. These residues embrace serine or threonine in the last α-helix of the NTase domain, and lysine/arginine and a hydrophobic amino acid located in the second and third PAP/OAS1 SBD core α-helix, respectively. Although the FAM46 proteins retain serine or cysteine in the last α-helix of the NTase domain, it is possible that the conserved insertion between the last core β-strand and α-helix in FAM46 NTase domain may substitute the role of the conserved Ser/Thr at least for some family members, enabling them to catalyze the modification of selected RNA 3′ ends.

We also performed a systematic search for missense mutations in human FAM46 genes, found in cancer patients. Collected mutation data from various databases and literature, combined with sequence/structure analyses suggest that, in addition to multiple myeloma, FAM46 genes may be also involved in the development of other major malignancies including lung, colorectal, hepatocellular, head and neck, urothelial, endometrial and renal papillary carcinomas and melanoma. We identified several single point mutations of highly conserved FAM46 amino acids that may affect the enzyme catalytic activity, processivity and substrate binding (e.g. by changing the preference for the type of incorporated NTP). Consequently, these mutations can lead to deregulation of specific RNAs as an oncogenic mechanism in multiple myeloma and other cancers. This is consistent with previous studies which showed a correlation between RNA deregulation (e.g. mRNA (120), microRNA (121,122), long non-coding RNA (123), small non-coding RNA (124)) and various diseases including cancers.

Summarizing, this work provides functional and structural annotation for novel and highly important enzymes involved in RNA metabolism in eukaryotes and thus may guide functional studies of these previously uncharacterized proteins. Further experimental investigations should address the predicted activity and clarify potential substrates to provide more insight into the detailed biological roles of these newly detected non-canonical poly(A) polymerases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Wayne Dawson for proofreading the manuscript.

FUNDING

Foundation for Polish Science [TEAM/2010-6 to K.G.]; Polish National Science Centre [2011/02/A/NZ2/00014 and 2014/15/B/NZ1/03357 to K.G.]; National Centre for Research and Development [INNOTECH-K2/HI2/19/184217/NCBR/13 to L.R.]; European Commission [FP7-KBBE-2011-289646 to L.R.]. A.M. and K.S. were the recipients of the fellowship from the Ministry of Science and Higher Education. Funding for open access charge: Polish National Science Centre [2014/15/B/NZ1/03357].
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Barragan,I., Borrego,S., Abd El-Aziz,M.M., El-Ashry,M.F., Abu-Safieh,L., Bhattacharya,S.S. and Antinolo,G. (2008) Genetic analysis of FAM46A in Spanish families with autosomal recessive retinitis pigmentosa: characterisation of novel VNTRs. *Ann. Hum. Genet.*, **72**, 26–34.
3. Etokebe,G.E., Bulat-Kardum,L., Munthe,L.A., Balen,S. and Dembic,Z. (2014) Association of variable number of tandem repeats in the coding region of the FAM46A gene, FAM46A rs11040 SNP and BAG6 rs3117582 SNP with susceptibility to tuberculosis. *PLoS One*, **9**, e91385.
4. Benjachat,T., Tongyoo,P., Tantivitayakul,P., Somparn,P., Hirankarn,N., Prom-On,S., Pisitkun,P., Leelahavanichkul,A., Avihingsanon,Y. and Townamchai,N. (2015) Biomarkers for refractory Lupus nephritis: a microarray study of kidney tissue. *Int. J. Mol. Sci.*, **16**, 14276–14290.
5. Boyd,K.D., Ross,F.M., Walker,B.A., Wardell,C.P., Tapper,W.J., Chiecchio,L., Dagrada,G., Konn,Z.J., Gregory,W.M., Jackson,G.H. *et al.* (2011) Mapping of chromosome 1p deletions in myeloma

identifies FAM46C at 1p12 and CDKN2C at 1p32.3 as being genes in regions associated with adverse survival. *Clin. Cancer Res.*, **17**, 7776–7784.
6. Chapman,M.A., Lawrence,M.S., Keats,J.J., Cibulskis,K., Sougnez,C., Schinzel,A.C., Harview,C.L., Brunet,J.P., Ahmann,G.J., Adli,M. *et al.* (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature*, **471**, 467–472.
7. Kortum,K.M., Langer,C., Monge,J., Bruins,L., Zhu,Y.X., Shi,C.X., Jedlowski,P., Egan,J.B., Ojha,J., Bullinger,L. *et al.* (2015) Longitudinal analysis of 25 sequential sample-pairs using a custom multiple myeloma mutation sequencing panel (M(3)P). *Ann. Hematol.*, **94**, 1205–1211.
8. Hamilton,S.M., Spencer,C.M., Harrison,W.R., Yuva-Paylor,L.A., Graham,D.F., Daza,R.A., Hevner,R.F., Overbeek,P.A. and Paylor,R. (2011) Multiple autism-like behaviors in a novel transgenic mouse model. *Behav. Brain Res.*, **218**, 29–41.
9. Kuehl,W.M. and Bergsagel,P.L. (2012) Molecular pathogenesis of multiple myeloma and its premalignant precursor. *J. Clin. Invest.*, **122**, 3456–3463.
10. Barbieri,M., Manzoni,M., Fabris,S., Ciceri,G., Todoerti,K., Simeon,V., Musto,P., Cortelezzi,A., Baldini,L., Neri,A. *et al.* (2015) Compendium of FAM46C gene mutations in plasma cell dyscrasias. *Br. J. Haematol.*, doi:10.1111/bjh.13793.
11. Bettoni,F., Filho,F.C., Grosso,D.M., Galante,P.A., Parmigiani,R.B., Geraldo,M.V., Henrique-Silva,F., Oba-Shinjo,S.M., Marie,S.K., Soares,F.A. *et al.* (2009) Identification of FAM46D as a novel cancer/testis antigen using EST data and serological analysis. *Genomics*, **94**, 153–160.
12. Colland,F., Jacq,X., Trouplin,V., Mougin,C., Groizeleau,C., Hamburger,A., Meil,A., Wojcik,J., Legrain,P. and Gauthier,J.M. (2004) Functional proteomics mapping of a human signaling pathway. *Genome Res.*, **14**, 1324–1332.
13. Etokebe,G.E., Kuchler,A.M., Haraldsen,G., Landin,M., Osmundsen,H. and Dembic,Z. (2009) Family-with-sequence-similarity-46, member A (Fam46a) gene is expressed in developing tooth buds. *Arch. Oral Biol.*, **54**, 1002–1007.
14. Schoggins,J.W., Wilson,S.J., Panis,M., Murphy,M.Y., Jones,C.T., Bieniasz,P. and Rice,C.M. (2011) A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, **472**, 481–485.
15. Campbell,C.L., Torres-Perez,F., Acuna-Retamar,M. and Schountz,T. (2015) Transcriptome markers of viral persistence in naturally-infected andes virus (bunyaviridae) seropositive long-tailed pygmy rice rats. *PLoS One*, **10**, e0122935.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
18. 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.
19. Capella-Gutierrez,S., Silla-Martinez,J.M. and Gabaldon,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
20. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G., Rhead,B. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
21. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
22. Yu,C.S., Chen,Y.C., Lu,C.H. and Hwang,J.K. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.
23. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
24. Chou,K.C. and Shen,H.B. (2010) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS One*, **5**, e9931.
25. Hoglund,A., Donnes,P., Blum,T., Adolph,H.W. and Kohlbacher,O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.
26. la Cour,T., Kiemer,L., Molgaard,A., Gupta,R., Skriver,K. and Brunak,S. (2004) Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng. Des. Sel.*, **17**, 527–536.
27. Dinkel,H., Van Roey,K., Michael,S., Davey,N.E., Weatheritt,R.J., Born,D., Speck,T., Kruger,D., Grebnev,G., Kuban,M. *et al.* (2014) The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.*, **42**, D259–D266.
28. Wu,C., Macleod,I. and Su,A.I. (2013) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.*, **41**, D561–D565.
29. McCall,M.N., Uppal,K., Jaffee,H.A., Zilliox,M.J. and Irizarry,R.A. (2011) The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.*, **39**, D1011–D1015.
30. Ginalski,K., von Grotthuss,M., Grishin,N.V. and Rychlewski,L. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.*, **32**, W576–W581.
31. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. U. S. A.*, **95**, 5857–5864.
32. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
33. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
34. Koonin,E.V., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Krylov,D.M., Makarova,K.S., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.
35. 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.
36. 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.
37. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
38. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
39. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
40. Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
41. 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.
42. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
43. Stagno,J., Aphasizheva,I., Rosengarth,A., Luecke,H. and Aphasizhev,R. (2007) UTP-bound and Apo structures of a minimal RNA uridylyltransferase. *J. Mol. Biol.*, **366**, 882–899.
44. Wang,Q., Canutescu,A.A. and Dunbrack,R.L. Jr (2008) SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.*, **3**, 1832–1847.
45. Wiederstein,M. and Sippl,M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, **35**, W407–W410.
46. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
47. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
48. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

49. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.
50. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
51. Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Tamborero,D., Schroeder,M.P., Jene-Sanz,A., Santos,A. and Lopez-Bigas,N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.
52. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
53. Lin,J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, **37**, 145–151.
54. Guindon,S., Dufayard,J.F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
55. Anisimova,M. and Gascuel,O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.*, **55**, 539–552.
56. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.
57. Brunet,F.G., Roest Crollius,H., Paris,M., Aury,J.M., Gibert,P., Jaillon,O., Laudet,V. and Robinson-Rechavi,M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.*, **23**, 1808–1816.
58. Peterson,K.J. and Butterfield,N.J. (2005) Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 9547–9552.
59. King,N., Westbrook,M.J., Young,S.L., Kuo,A., Abedin,M., Chapman,J., Fairclough,S., Hellsten,U., Isogai,Y., Letunic,I. *et al.* (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*, **451**, 783–788.
60. Sun,G., Yang,Z., Ishwar,A. and Huang,J. (2010) Algal genes in the closest relatives of animals. *Mol. Biol. Evol.*, **27**, 2879–2889.
61. Matsuzaka,Y., Tounai,K., Denda,A., Tomizawa,M., Makino,S., Okamoto,K., Keicho,N., Oka,A., Kulski,J.K., Tamiya,G. *et al.* (2002) Identification of novel candidate genes in the diffuse panbronchiolitis critical region of the class I human MHC. *Immunogenetics*, **54**, 301–309.
62. Teif,V.B., Vainshtein,Y., Caudron-Herger,M., Mallm,J.P., Marth,C., Hofer,T. and Rippe,K. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.*, **19**, 1185–1192.
63. Hartmann,R., Justesen,J., Sarkar,S.N., Sen,G.C. and Yee,V.C. (2003) Crystal structure of the 2′-specific and double-stranded RNA-activated interferon-induced antiviral protein 2′-5′-oligoadenylate synthetase. *Mol. Cell*, **12**, 1173–1185.
64. Kuhn,C.D., Wilusz,J.E., Zheng,Y., Beal,P.A. and Joshua-Tor,L. (2015) On-enzyme refolding permits small RNA and tRNA surveillance by the CCA-adding enzyme. *Cell*, **160**, 644–658.
65. Zhu,D., Wang,L., Shang,G., Liu,X., Zhu,J., Lu,D., Wang,L., Kan,B., Zhang,J.R. and Xiang,Y. (2014) Structural biochemistry of a Vibrio cholerae dinucleotide cyclase reveals cyclase activity regulation by folates. *Mol. Cell*, **55**, 931–937.
66. Wolkowicz,U.M. and Cook,A.G. (2012) NF45 dimerizes with NF90, Zfr and SPNR via a conserved domain that has a nucleotidyltransferase fold. *Nucleic Acids Res.*, **40**, 9356–9368.
67. 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.
68. 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.
69. Sakon,J., Liao,H.H., Kanikula,A.M., Benning,M.M., Rayment,I. and Holden,H.M. (1993) Molecular structure of kanamycin nucleotidyltransferase determined to 3.0-A resolution. *Biochemistry*, **32**, 11977–11984.
70. Morar,M., Bhullar,K., Hughes,D.W., Junop,M. and Wright,G.D. (2009) Structure and mechanism of the lincosamide antibiotic adenylyltransferase LinB. *Structure*, **17**, 1649–1659.
71. Xu,Y., Zhang,R., Joachimiak,A., Carr,P.D., Huber,T., Vasudevan,S.G. and Ollis,D.L. (2004) Structure of the N-terminal domain of Escherichia coli glutamine synthetase adenylyltransferase. *Structure*, **12**, 861–869.
72. Bard,J., Zhelkovsky,A.M., Helmling,S., Earnest,T.N., Moore,C.L. and Bohm,A. (2000) Structure of yeast poly(A) polymerase alone and in complex with 3′-dATP. *Science*, **289**, 1346–1349.
73. Lunde,B.M., Magler,I. and Meinhart,A. (2012) Crystal structures of the Cid1 poly (U) polymerase reveal the mechanism for UTP selectivity. *Nucleic Acids Res.*, **40**, 9815–9824.
74. Hamill,S., Wolin,S.L. and Reinisch,K.M. (2010) Structure and function of the polymerase core of TRAMP, a RNA surveillance complex. *Proc. Natl Acad. Sci. U. S. A.*, **107**, 15045–15050.
75. Losón,O.C., Liu,R., Rome,M.E., Meng,S., Kaiser,J.T., Shan,S.O. and Chan,D.C. (2014) The mitochondrial fission receptor MiD51 requires ADP as a cofactor. *Structure*, **22**, 367–377.
76. Lohofener,J., Steinke,N., Kay-Fedorov,P., Baruch,P., Nikulin,A., Tishchenko,S., Manstein,D.J. and Fedorov,R. (2015) The activation mechanism of 2′-5′-oligoadenylate synthetase gives new insights into OAS/cGAS triggers of innate immunity. *Structure*, **23**, 851–862.
77. Gao,P., Ascano,M., Wu,Y., Barchet,W., Gaffney,B.L., Zillinger,T., Serganov,A.A., Liu,Y., Jones,R.A., Hartmann,G. *et al.* (2013) Cyclic [G(2′,5′)pA(3′,5′)p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. *Cell*, **153**, 1094–1107.
78. Fu,H., Subramanian,R.R. and Masters,S.C. (2000) 14-3-3 proteins: structure, function, and regulation. *Annu. Rev. Pharmacol. Toxicol.*, **40**, 617–647.
79. Irwin,S., Vandelft,M., Pinchev,D., Howell,J.L., Graczyk,J., Orr,H.T. and Truant,R. (2005) RNA association and nucleocytoplasmic shuttling by ataxin-1. *J. Cell Sci.*, **118**, 233–242.
80. Castello,A., Fischer,B., Eichelbaum,K., Horos,R., Beckmann,B.M., Strein,C., Davey,N.E., Humphreys,D.T., Preiss,T., Steinmetz,L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.
81. Jaakkola,P., Mole,D.R., Tian,Y.M., Wilson,M.I., Gielbert,J., Gaskell,S.J., von Kriegsheim,A., Hebestreit,H.F., Mukherji,M., Schofield,C.J. *et al.* (2001) Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. *Science*, **292**, 468–472.
82. Zhang,D.D. and Hannink,M. (2003) Distinct cysteine residues in Keap1 are required for Keap1-dependent ubiquitination of Nrf2 and for stabilization of Nrf2 by chemopreventive agents and oxidative stress. *Mol. Cell. Biol.*, **23**, 8137–8151.
83. Tsai,T.C., Lee,Y.L., Hsiao,W.C., Tsao,Y.P. and Chen,S.L. (2005) NRIP, a novel nuclear receptor interaction protein, enhances the transcriptional activity of nuclear receptors. *J. Biol. Chem.*, **280**, 20000–20009.
84. Zhou,H.J., Wong,C.M., Chen,J.H., Qiang,B.Q., Yuan,J.G. and Jin,D.Y. (2001) Inhibition of LZIP-mediated transcription through direct interaction with a novel host cell factor-like protein. *J. Biol. Chem.*, **276**, 28933–28938.
85. Kim,J.E., Ryu,I., Kim,W.J., Song,O.K., Ryu,J., Kwon,M.Y., Kim,J.H. and Jang,S.K. (2008) Proline-rich transcript in brain protein induces stress granule formation. *Mol. Cell. Biol.*, **28**, 803–813.
86. Forbes,M.M., Rothhamel,S., Jenny,A. and Marlow,F.L. (2015) Maternal dazap2 regulates germ granules by counteracting dynein in zebrafish primordial germ cells. *Cell Rep.*, **12**, 49–57.
87. Kedersha,N.L. and Rome,L.H. (1986) Isolation and characterization of a novel ribonucleoprotein particle: large structures contain a single species of small RNA. *J. Cell Biol.*, **103**, 699–709.
88. Kedersha,N.L., Heuser,J.E., Chugani,D.C. and Rome,L.H. (1991) Vaults. III. Vault ribonucleoprotein particles open into flower-like structures with octagonal symmetry. *J. Cell Biol.*, **112**, 225–235.
89. McKeegan,K.S., Debieux,C.M., Boulon,S., Bertrand,E. and Watkins,N.J. (2007) A dynamic scaffold of pre-snoRNP factors facilitates human box C/D snoRNP assembly. *Mol. Cell. Biol.*, **27**, 6782–6793.

90. Angers,S., Li,T., Yi,X., MacCoss,M.J., Moon,R.T. and Zheng,N. (2006) Molecular architecture and assembly of the DDB1-CUL4A ubiquitin ligase machinery. *Nature*, **443**, 590–593.
91. Rodrigo-Brenni,M.C., Gutierrez,E. and Hegde,R.S. (2014) Cytosolic quality control of mislocalized proteins requires RNF126 recruitment to Bag6. *Mol. Cell*, **55**, 227–237.
92. Arquint,C., Gabryjonczyk,A.M., Imseng,S., Bohm,R., Sauer,E., Hiller,S., Nigg,E.A. and Maier,T. (2015) STIL binding to Polo-box 3 of PLK4 regulates centriole duplication. *Elife*, **4**, e07888.
93. Cheng,A.M., Rowley,B., Pao,W., Hayday,A., Bolen,J.B. and Pawson,T. (1995) Syk tyrosine kinase required for mouse viability and B-cell development. *Nature*, **378**, 303–306.
94. Carnevale,J., Ross,L., Puissant,A., Banerji,V., Stone,R.M., DeAngelo,D.J., Ross,K.N. and Stegmaier,K. (2013) SYK regulates mTOR signaling in AML. *Leukemia*, **27**, 2118–2128.
95. Liu,N., Li,Z., Pei,D. and Shu,X. (2013) Zfyve9a regulates the proliferation of hepatic cells during zebrafish embryogenesis. *Int. J. Dev. Biol.*, **57**, 773–778.
96. Fiorillo,C., Moro,F., Yi,J., Weil,S., Brisca,G., Astrea,G., Severino,M., Romano,A., Battini,R., Rossi,A. *et al.* (2014) Novel dynein DYNC1H1 neck and motor domain mutations link distal spinal muscular atrophy and abnormal cortical development. *Hum. Mutat.*, **35**, 298–302.
97. Vallee,R.B., McKenney,R.J. and Ori-McKenney,K.M. (2012) Multiple modes of cytoplasmic dynein regulation. *Nat. Cell Biol.*, **14**, 224–230.
98. Hutagalung,A.H. and Novick,P.J. (2011) Role of Rab GTPases in membrane traffic and cell physiology. *Physiol. Rev.*, **91**, 119–149.
99. Collins,B.M., McCoy,A.J., Kent,H.M., Evans,P.R. and Owen,D.J. (2002) Molecular architecture and functional model of the endocytic AP2 complex. *Cell*, **109**, 523–535.
100. Tian,M. (2010) *The molecular cloning and characterization of Fam46c RNA stability factor*. PhD dissertation, Harvard University.
101. Kranzusch,P.J., Lee,A.S., Berger,J.M. and Doudna,J.A. (2013) Structure of human cGAS reveals a conserved family of second-messenger enzymes in innate immunity. *Cell Rep.*, **3**, 1362–1368.
102. Weill,L., Belloc,E., Bava,F.A. and Mendez,R. (2012) Translational control by changes in poly(A) tail length: recycling mRNAs. *Nat. Struct. Mol. Biol.*, **19**, 577–585.
103. Subtelny,A.O., Eichhorn,S.W., Chen,G.R., Sive,H. and Bartel,D.P. (2014) Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, **508**, 66–71.
104. Schmidt,M.J. and Norbury,C.J. (2010) Polyadenylation and beyond: emerging roles for noncanonical poly(A) polymerases. *Wiley Interdiscip. Rev. RNA*, **1**, 142–151.
105. Villa,N., Do,A., Hershey,J.W. and Fraser,C.S. (2013) Human eukaryotic initiation factor 4G (eIF4G) protein binds to eIF3c, -d, and -e to promote mRNA recruitment to the ribosome. *J. Biol. Chem.*, **288**, 32932–32940.
106. Kuhn,U. and Wahle,E. (2004) Structure and function of poly(A) binding proteins. *Biochim. Biophys. Acta*, **1678**, 67–84.
107. Cunningham,K.S., Dodson,R.E., Nagel,M.A., Shapiro,D.J. and Schoenberg,D.R. (2000) Vigilin binding selectively inhibits cleavage of the vitellogenin mRNA 3′-untranslated region by the mRNA endonuclease polysomal ribonuclease 1. *Proc. Natl Acad. Sci. U. S. A.*, **97**, 12498–12502.
108. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Ascano,M. Jr, Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339.
109. Fan,X.C. and Steitz,J.A. (1998) Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *EMBO J.*, **17**, 3448–3460.
110. Peng,S.S., Chen,C.Y., Xu,N. and Shyu,A.B. (1998) RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. *EMBO J.*, **17**, 3461–3470.
111. Engel,C., Sainsbury,S., Cheung,A.C., Kostrewa,D. and Cramer,P. (2013) RNA polymerase I structure and transcription regulation. *Nature*, **502**, 650–655.
112. Laferte,A., Favry,E., Sentenac,A., Riva,M., Carles,C. and Chedin,S. (2006) The transcriptional activity of RNA polymerase I is a key determinant for the level of all ribosome components. *Genes Dev.*, **20**, 2030–2040.
113. Wyler,E., Wandrey,F., Badertscher,L., Montellese,C., Alper,D. and Kutay,U. (2014) The beta-isoform of the BRCA2 and CDKN1A(p21)-interacting protein (BCCIP) stabilizes nuclear RPL23/uL14. *FEBS Lett.*, **588**, 3685–3691.
114. Venteicher,A.S., Abreu,E.B., Meng,Z., McCann,K.E., Terns,R.M., Veenstra,T.D., Terns,M.P. and Artandi,S.E. (2009) A human telomerase holoenzyme protein required for Cajal body localization and telomere synthesis. *Science*, **323**, 644–648.
115. Thornton,J.E., Du,P., Jing,L., Sjekloca,L., Lin,S., Grossi,E., Sliz,P., Zon,L.I. and Gregory,R.I. (2014) Selective microRNA uridylation by Zcchc6 (TUT7) and Zcchc11 (TUT4). *Nucleic Acids Res.*, **42**, 11777–11791.
116. Chang,J.H. and Tong,L. (2012) Mitochondrial poly(A) polymerase and polyadenylation. *Biochim. Biophys. Acta*, **1819**, 992–997.
117. Trippe,R., Guschina,E., Hossbach,M., Urlaub,H., Luhrmann,R. and Benecke,B.J. (2006) Identification, cloning, and functional analysis of the human U6 snRNA-specific terminal uridylyl transferase. *RNA*, **12**, 1494–1504.
118. Kwak,J.E., Wang,L., Ballantyne,S., Kimble,J. and Wickens,M. (2004) Mammalian GLD-2 homologs are poly(A) polymerases. *Proc. Natl Acad. Sci. U.S.A.*, **101**, 4407–4412.
119. Mullen,T.E. and Marzluff,W.F. (2008) Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5′ to 3′ and 3′ to 5′. *Genes Dev.*, **22**, 50–65.
120. Lokody,I. (2014) RNA dynamics: destabilizing mRNAs promotes metastasis. *Nat. Rev. Cancer*, **14**, 578.
121. Palmero,E.I., de Campos,S.G., Campos,M., de Souza,N.C., Guerreiro,I.D., Carvalho,A.L. and Marques,M.M. (2011) Mechanisms and role of microRNA deregulation in cancer onset and progression. *Genet. Mol. Biol.*, **34**, 363–370.
122. Jansson,M.D. and Lund,A.H. (2012) MicroRNA and cancer. *Mol. Oncol.*, **6**, 590–610.
123. Gao,W., Chan,J.Y. and Wong,T.S. (2014) Long non-coding RNA deregulation in tongue squamous cell carcinoma. *Biomed. Res. Int.*, **2014**, 405860.
124. Ravo,M., Cordella,A., Rinaldi,A., Bruno,G., Alexandrova,E., Saggese,P., Nassa,G., Giurato,G., Tarallo,R., Marchese,G. *et al.* (2015) Small non-coding RNA deregulation in endometrial carcinogenesis. *Oncotarget*, **6**, 4677–4691.