6848–6863 *Nucleic Acids Research, 2017, Vol. 45, No. 11* *Published online 11 April 2017*
doi: 10.1093/nar/gkx256

Nonsense-mediated mRNA decay in *Tetrahymena* is EJC independent and requires a protozoa-specific nuclease

Miao Tian[1,2,3], Wentao Yang[1,2], Jing Zhang[1,2], Huai Dang[4], Xingyi Lu[1,2], Chengjie Fu[1] and Wei Miao[1,*]

[1]Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei 430072, China, [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [3]Department of Chromosome Biology, Max F. Perutz Laboratories, University of Vienna, Vienna A-1030, Austria and [4]College of Life Sciences, Northwest Normal University, Lanzhou 730070, China

Received December 10, 2016; Revised March 31, 2017; Editorial Decision April 03, 2017; Accepted April 05, 2017

ABSTRACT

Nonsense-mediated mRNA decay (NMD) is essential for removing premature termination codon-containing transcripts from cells. Studying the NMD pathway in model organisms can help to elucidate the NMD mechanism in humans and improve our understanding of how this biologically important process has evolved. Ciliates are among the earliest branching eukaryotes; their NMD mechanism is poorly understood and may be primordial. We demonstrate that highly conserved Upf proteins (Upf1a, Upf2 and Upf3) are involved in the NMD pathway of the ciliate, *Tetrahymena thermophila*. We further show that a novel protozoa-specific nuclease, Smg6L, is responsible for destroying many NMD-targeted transcripts. Transcriptome-wide identification and characterization of NMD-targeted transcripts in vegetative *Tetrahymena* cells showed that many have exon–exon junctions downstream of the termination codon. However, *Tetrahymena* may lack a functional exon junction complex (EJC), and the *Tetrahymena* ortholog of an EJC core component, Mago nashi (Mag1), is dispensable for NMD. Therefore, NMD is EJC independent in this early branching eukaryote.

INTRODUCTION

Nonsense-mediated mRNA decay (NMD) is an mRNA quality control mechanism that degrades transcripts containing premature termination codons (PTCs) to protect cells from the potential deleterious effects of truncated proteins (1,2). NMD substrates mainly result from nonsense mutation and transcriptional error. However, a major group results from alternative splicing (AS), suggesting that the NMD and AS processes are tightly coupled to maintain a steady-state transcriptome (3–6). Owing to its biological importance, the NMD pathway has been extensively studied in organisms from yeast to humans.

In most species investigated so far, the NMD pathway requires a set of evolutionarily conserved up-frameshift proteins (Upf1, Upf2 and Upf3) to initiate PTC-containing transcript recognition and degradation (1). However, another subset of NMD factors diverge in different eukaryotic lineages (7). Therefore, two major models have been proposed to explain PTC discrimination and the subsequent destruction of PTC-containing transcripts: the faux 3′-UTR model and the exon junction complex (EJC)-enhanced model (7).

The faux 3′-UTR model was initially proposed to describe the NMD mechanism in the unicellular fungus *Saccharomyces cerevisiae* (8). According to this model, the presence of a PTC extends the 3′-UTR region, which then impairs the interaction between cytoplasmic poly(A)-binding protein and eukaryotic release factors (eRFs), enabling Upf1 to interact with eRFs and elicit NMD (8,9). This NMD mechanism was subsequently discovered in many other multicellular organisms (such as humans and *Drosophila*) and is thought to be evolutionarily conserved (10,11). However, some reports appear to contradict this mechanism (2). The EJC-enhanced model was proposed to explain mammalian NMD, as other organisms either lack the EJC (such as *S. cerevisiae*) or their EJC proteins are not required for NMD (such as *Caenorhabditis elegans*) (2,12). The EJC protein complex is recruited 20–24 nt upstream of the spliceosomal intron as a consequence of pre-mRNA splicing, and is displaced by the translating ribosome. It is composed of an inner heterotetrameric core (con-

*To whom correspondence should be addressed. Tel: +86 27 6878 0050; Fax: +86 27 6878 0050; Email: miaowei@ihb.ac.cn
Disclaimer: The funders had no role in the study design, data collection and analysis, the decision to publish or manuscript preparation.

sisting of eIF4A3, Btz, Y14 and Mago nashi) that serves as a binding platform for an outer shell comprising Upf3, Upf2 and other proteins (13,14).The current EJC-enhanced model suggests that transcripts with a termination codon (TC) located more than 50–55 nt upstream of an exon junction are susceptible to NMD (15). An EJC deposited near to the 3′-UTR exon junction could interact with the SURF complex (comprising metazoan-specific Smg1 and the evolutionarily conserved Upf1 and eRF1/3 proteins) to trigger NMD (16,17). However, the EJC-enhanced model has also been challenged (18–20), and alternative models to explain NMD have been recently proposed (2,21). For example, the 'ribosome release model' states that recruitment of Upf1 and additional factors to the termination ribosomes at the PTC could help disassociate downstream post-termination ribosomes as well as mRNPs and resolve RNA secondary structures, which may make the unprotected 3′-UTR region vulnerable to degradation. However, further experiments are required to test these models. In conclusion, a robust molecular basis for the NMD mechanism remains to be elucidated.

In addition to their different PTC recognition mechanisms, eukaryotic lineages have different methods of degrading PTC-bearing transcripts (22). For example, yeast PTC-containing transcripts are deadenylated and/or decapped before exonuclease degradation (23,24). In contrast, mammalian nonsense transcripts are removed via both endo- and exonucleolytic degradation: endonucleolytic degradation of PTC-containing transcripts is initiated by cleavage by the conserved metazoan endoribonuclease Smg6 (25), and exonuclease degradation is initiated by decapping and deadenylation mediated by Smg 5–7 and their interacting partners (22).

Extensive research into the NMD pathways of eukaryotic species ranging from unicellular fungi to mammals has revealed that some components of the PTC recognition and destruction machinery are conserved, while others are divergent. Protozoa are a large group of early branching eukaryotes and, hence, remote relatives of some common model organisms (e.g. yeast, fruit flies, worms, mice and plants). Therefore, investigating protozoan NMD mechanism may improve our understanding of the origins and evolution of this biologically and biomedically important process. The few studies carried out in parasitic protozoa so far suggest that the protozoan NMD mechanism is either primitive or unique (26,27). However, two recent studies in *Paramecium tetraurelia* proved that canonical NMD exists in ciliated protozoa, although the identity of NMD components and their roles in recognizing and destroying PTC-containing transcripts are still poorly understood (28,29).

To investigate the NMD mechanism of ciliated protozoa, we used *Tetrahymena thermophila* (hereafter referred to as *Tetrahymena*) as a model organism because its genome is comparatively well annotated, and genetic manipulation is easier in this species than in other ciliates (30). *Tetrahymena* has two functionally distinct nuclei. During vegetative growth, the germline micronucleus is transcriptionally silent and all protein-coding RNAs are transcribed from the somatic macronuclear genome. In this study, we first identified key factors in the *Tetrahymena* NMD pathway and then evaluated transcriptome variation by deep sequencing after the deletion of individual NMD factors. We found that only one of the two *Tetrahymena* Upf1 homologs, Upf1a, is responsible for directing NMD during vegetative growth, but that the Upf2 and Upf3 homologs are also involved in NMD. Interestingly, the nuclease activity of a metazoan Smg6-like, Nedd4-BP1, bacterial YacP nuclease (NYN) domain-containing protein (named Smg6L) is important for *Tetrahymena* NMD. The discovery of *Tetrahymena* NYN domain-containing Smg6L nuclease homologs in many parasitic and free-living protozoa suggests this protozoa-specific nuclease is involved in destroying PTC-containing transcripts and could be a unique component of protozoan NMD. In contrast, the *Tetrahymena* ortholog of the human EJC core component, Mag1, is not required for NMD in this organism. Further, analysis of *in vivo* and *in vitro* protein interactions indicated that *Tetrahymena* lacks a functional EJC. This finding is consistent with the lack of EJC-binding motifs in Upf3 and Smg6L proteins in *Tetrahymena* and in other ciliated and apicomplexan protozoa. Hence, the NMD pathway is EJC independent in *Tetrahymena* and possibly also in other protozoa. Although EJC is not required for *Tetrahymena* NMD, genome-wide statistical analysis of NMD targets suggests that transcripts with an intron located downstream of the TC are preferentially destroyed via NMD. Therefore, PTC recognition in *Tetrahymena* presumably largely relies on unknown factor(s) related to pre-mRNA splicing.

MATERIALS AND METHODS

Strains, culture conditions and drug treatment

Tetrahymena B2086 wild-type (WT; obtained from the *Tetrahymena* Stock Center at Cornell University) and mutant strains (Supplementary Table S1) were maintained in Super Proteose Peptone (SPP) medium (1% Proteose Peptone, 0.2% glucose, 0.1% yeast extract, 0.003% Sequestrene) at 30°C on a rotary shaker at 135 rpm. The germline micronuclear genome is transcriptionally silent in vegetative growth; therefore, somatic macronuclear gene knockout, mutant and C-terminal hemagglutinin (HA)-tagged strains were generated by homologous recombination as previously described (31). Briefly, the relevant plasmid was introduced into starved B2086 cells via biolistic transformation (32) and transformants were selected by increasing the paromomycin (Sigma-Aldrich, St Louis, MO, USA) concentration in culture medium until all macronuclear loci were replaced via phenotypic assortment (33). Schematic diagrams of all constructs are shown in Supplementary Figures S1 and 2; primer sequences are listed in Supplementary Table S2. Somatic gene knockout and knokdown strains were confirmed by reverse transcription polymerase chain reaction (RT-PCR), transcriptome sequencing and quantitative RT-PCR (qRT-PCR; Supplementary Figure S3). The Smg6L NYN nuclease point mutation (Asp^{820} to Ala; SMG6L-D820A strain) was confirmed by Sanger sequencing. To block protein synthesis, vegetative WT *Tetrahymena* cells were incubated with 100 μg/ml cycloheximide (CHX) (Beyotime, Shanghai, China; 20 mg/ml stock solution in water) for 4 h, and then harvested for downstream analysis.

Gene identification and bioinformatics analysis

NMD factor homologs in *Tetrahymena* and other protists (*Giardia lamblia*, *Ichthyophthirius multifilis*, *Oxytricha trifallax*, *P. tetraurelia*, *Plasmodium falciparum*, *Stylonychia lemnae*, *Toxoplasma gondii*, *Trypanosoma brucei*) were identified using human NMD factor protein sequences as query sequences for BLASTp searching against the respective proteomes (Supplementary Table S3). Orthologs were confirmed by reciprocal BLASTp analysis. Sequences were aligned using ClustalW and colored according to amino acid sequence similarity using ESPript 3.0 (34). Phylogenetic trees were constructed using Mega5 with default settings (neighbor-joining method) (35). Gene expression profiles shown in Figure 1C were retrieved from TetraFGD (http://tfgd.ihb.ac.cn/) (36).

Total RNA isolation and reverse transcription-PCR analysis

Total RNA was extracted from 5 ml samples of vegetatively growing *Tetrahymena* WT and mutant cells (density: 2.5×10^5 cells/ml) using an RNeasy Plus Mini Kit (Qiagen, Valencia, CA, USA) and treated with RQ1 RNase-Free DNase (Promega, Madison, WI, USA) to remove genomic DNA. This preparation (2 μg) was used for cDNA synthesis: first-strand cDNA was synthesized using random hexamers (Promega, Madison, WI, USA) and an M-MLV Reverse Transcriptase kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions with added RNasin ribonuclease inhibitor (Promega, Madison, WI, USA). Titanium Taq DNA Polymerase (Clontech Laboratories, Palo Alto, CA, USA) was used to amplify DNA fragments from both PTC-containing (PTC^+) and normal (FUNC) transcripts within the same reaction using primers targeting a common region. PCR products were resolved by 3% agarose gel electrophoresis (in 0.5 × Tris-borate-EDTA buffer) and stained with ethidium bromide (0.5 μg/ml). The band intensity of PCR products was measured in ImageJ (37) and used to calculate the ratio of PTC^+ and FUNC abundance. Real-time qRT-PCR was carried out using the DNA Engine Opticon 2 System (Bio-Rad/MJ Research, Hercules, CA, USA) in 20 μl reactions containing 10 μl SsoFast EvaGreen PCR mix (Bio-Rad, Hercules, CA, USA) and 0.4 μM primers. All primers used in this study are listed in Supplementary Table S2. cDNA generated from WT *Tetrahymena* total RNA was used as the control and 18S rRNA as the endogenous reference for normalizing gene expression, performed using REST 2009 software (38).

Transcriptome sequencing and data analysis

Total RNA samples (3 μg) from vegetatively growing *ΔUPF1a*, *ΔMAG1*, *ΔSMG6L*, *SMG6L-NYNmu* and WT B2086 *Tetrahymena* cells were used for constructing pair-end Illumina sequencing libraries and analyzed with a HiSeq 2000 system (Illumina, San Diego, CA, USA). Two biological replicates each of *ΔUPF1a* and WT B2086 were used. After sequencing, filtered reads were mapped to the *Tetrahymena* macronuclear genome (obtained from http://ciliate.org/, release 2014) using the TopHat algorithm (Supplementary Table S4) and novel transcripts were identified using the Cufflinks toolkit (39). Differential gene expression and exon expression (or usage) analyses were carried out using the Cuffdiff and DEXSeq algorithms, respectively (40). The following criteria were used to identify transcripts upregulated in mutants: (i) the transcript FPKM (fragments per kilobase of exon per million fragments mapped) should be 2-fold higher in mutants than in WT cells; (ii) the transcript should be significantly upregulated in a sample with biological replicates; or (iii) the transcript FPKM should be higher and exon expression significantly different in the mutant than in the WT strain (for *ΔUPF1a* only). To reduce the likelihood of ambiguous results due to low read coverage, putative upregulated transcripts with FPKM values of <3 were discounted. Similar criteria were applied for downregulated transcripts. Gene expression levels are listed in Supplementary Table S5. Gene Ontology (GO) enrichment analysis of genes differentially expressed in mutants was done using the Bingo plugin (version 2.44) with default settings in Cytoscape (version 2.8.3) (41,42). Detailed descriptions of all data analysis procedures are provided in Supplementary Figure S4. All transcriptome sequencing data described in this study have been deposited in the NCBI Gene Expression Omnibus (43) and are accessible through GEO Series accession number GSE90899.

Identification of PTC-containing transcripts and analysis of AS events

PTC-containing transcripts, including those with putative upstream open reading frames (uORFs), were identified using the getorf algorithm and custom Perl scripts (44). A PTC-containing transcript is canonically defined as having the same translation start site as a normal transcript annotated in the current *Tetrahymena* Genome Database (TGD) but with a TC prior to that of the normal transcript. uORF-containing transcripts form a subcategory of PTC-containing transcripts, and are defined according to published criteria (45). Briefly, a putative uORF-containing transcript should share an ORF with the normal transcript but should have an additional translation start site within its 5′-UTR region that drives translation of an extra putative protein containing more than 20 residues. In addition, the entire uORF region should be covered by transcriptome sequencing reads.

The AStalavista algorithm was used to classify AS events (46). All AS events were further inspected by manually comparing structural differences between splicing isoforms in Integrative Genomics Viewer (IGV) (47). To generate the sequence logo for splice sites from both normal and PTC-containing transcripts, a custom Perl script was first used to extract sequences from around the splicing donor and acceptor sites. After alignment, these sequences were submitted to the Pictogram online server (http://genes.mit.edu/pictogram.html) for plotting the sequence logo and calculating its information content. Structural features of PTC-containing and normal transcripts were analyzed using a series of custom Perl scripts, and statistical analysis and graph plotting were carried out with R software. Detailed descriptions of data analysis procedures can be found in Supplementary Figure S4.

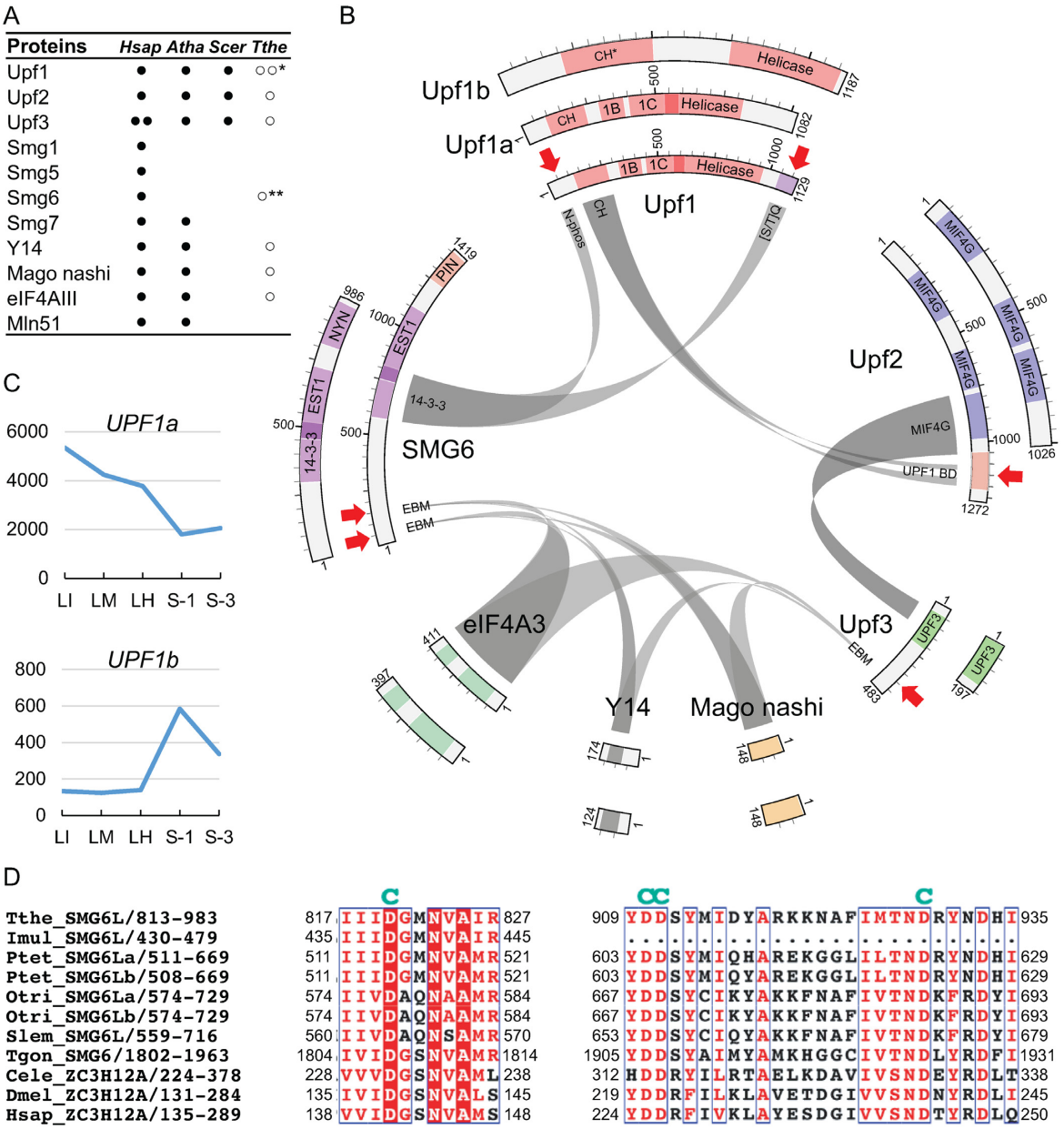

Figure 1. Identification of NMD factor homologs in *Tetrahymena*. (**A**) A list of confirmed NMD factors in *Homo sapiens* (*Hsap*), *Aradidopsis thaliana* (*Atha*) and *Saccharomyces cerevisiae* (*Scer*; solid circles) and NMD factor homologs in *Tetrahymena* (*Tthe*; open circles). *The two Upf1 homologs in *Tetrahymena* were named Upf1a and Upf1b according to their protein sequence similarity with the *H. sapiens* protein. **Tetrahymena* Smg6L contains a NYN ribonuclease domain instead of the PIN domain found in the human Smg6 protein. (**B**) Comparison of protein domains in *Tetrahymena* NMD factor homologs and human NMD factors. Outer circle, *Tetrahymena* NMD factor homologs; inner circle, human NMD factors. Gray ribbons indicate known interactions between human NMD proteins; arrows indicate key protein–protein interaction domains absent or less well conserved in *Tetrahymena*. (**C**) Relative gene expression profiles for the two *Tetrahymena* Upf1 homologs under conditions of vegetative growth (LI, LM and LH correspond to low, medium and high cell density, respectively) and starvation (S-1 and S-3 indicate 1 and 3 h of starvation, respectively). Vertical axes indicate gene expression in arbitrary units. (**D**) Sequence alignment of the NYN ribonuclease domain of protozoan Smg6L proteins and metazoan Zc3h12a proteins showing highly conserved key residues in the catalytic center of NYN ribonuclease domain (labeled 'C'; alignment of the whole region can be found in Supplementary Figure S8E). Cele, *Caenorhabditis elegans*; Dmel, *Drosophila melanogaster*, Imul, *Ichthyophthirius multifiliis*; Otri, *Oxytricha trifallax*; Ptet, *Paramecium tetraurelia*; Slem, *Stylonychia lemnae*; Tgon, *Trypanosoma gondii*.

Immunoprecipitation and mass spectrometry data analysis

To pull down HA-tagged proteins from *Tetrahymena* cell extracts, strains expressing endogenous levels of C-terminally HA-tagged proteins were maintained in SPP medium. The WT strain was used as the control. Cells were harvested during vegetative growth (cell density: 2–3 × 10^5 cells/ml) and resuspended in lysis buffer (30 mM Tris-HCl, 20 mM KCl, 2 mM $MgCl_2$, 1 mM phenylmethylsulfonyl fluoride, 0.1% Triton-X 100, 150 mM NaCl) containing cOmplete proteinase inhibitor (Roche Diagnostics, Indianapolis, IN, USA). Soluble proteins were incubated with EZview anti-HA agarose beads (Sigma-Aldrich, St Louis, MO, USA) in lysis buffer for 2 h at 4°C. After stringent washing in lysis buffer, HA-tagged proteins were eluted using HA peptides (Sigma-Aldrich, St Louis, MO, USA) according to the manufacturer's instructions. Immunoprecipitation (IP) products were analyzed by silver staining and immunoblotted with an anti-HA tag antibody (clone 16B12, Covance, Berkeley, CA, USA). To identify co-immunoprecipitating proteins, an aliquot of each IP product was tryptic digested and analyzed with an LTQ-Orbitrap tandem mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). The MS/MS dataset was analyzed with Mascot software using the *Tetrahymena* protein sequence (downloaded from the TGD; http://ciliate.org/index.php/home/downloads). MS/MS raw data are listed in Supplementary Table S6. Confidence scores for protein interactions identified by mass spectrometry were evaluated with SAINTexpress software: confidence was set at a Bayesian false discovery rate (FDR) value of <0.05 (48).

Recombinant protein purification and GST pull-down analysis

The complete coding sequences for *Tetrahymena* Upf3, Y14, Mag1 and eIF4A3 proteins, the Upf1a cysteine- and histidine-rich (CH) domain (Upf1a-CH, 91–240 aa) and the Upf2 C-terminus (Upf2-Cter, 922–1026 aa) were codon optimized and synthesized for *Escherichia coli* expression (sequences are listed in Supplementary Table S2). To generate an N-terminally GST-tagged or N-terminally 6 × His-tagged recombinant protein, the codon-optimized sequence was cloned into the pGEX-4T-1 (Amersham Biosciences, Piscataway, NJ, USA) or pET-28a (+) (Novagen, Madison, WI, USA) vector, respectively. Expression constructs were transformed into competent *E. coli* BL21 (DE3) cells. GST and GST-tagged proteins were purified using Glutathione Sepharose 4B beads (GE Healthcare, Waukesha, WI, USA) and His-tagged proteins were purified using His-Pur Ni-NTA Resin (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions.

For GST pull-down, 20 μg purified GST-tagged protein or GST (control) was mixed with the same molar ratio of purified His-tagged proteins and incubated with 20 μl Glutathione Sepharose 4B resin at 4°C for 30 min. The resin was then washed with 10 bed volumes of 1 × phosphate buffered saline buffer. Bound proteins were eluted with elution buffer (50 mM Tris-HCl, 10 mM glutathione, pH 8.0) and analyzed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) followed by Coomassie blue staining.

RESULTS

Bioinformatics identification of putative NMD factors in *Tetrahymena*

Human protein sequences for key evolutionarily conserved NMD factors (Upf 1–3), metazoan-specific SMG proteins (Smg1 and Smg 5–7) and EJC core components (Y14, Mago nashi, eIF4A3 and Mln51) were retrieved from the UniProt database and homology searches were performed to identify *Tetrahymena* NMD factor candidates (search results shown in Figure 1A; detailed list in Supplementary Table S3). *Tetrahymena* harbors two mammalian Upf1 homologs: Upf1a has 49% identity and Upf1b has 37% identity to human Upf1. Sequence alignment of the *Tetrahymena* Upf1 homologs suggested that, as in *S. cerevisiae* Upf1, both lack the C-terminal [S/T]Q-rich motif present in human Upf1 (Figure 1B). Upf1a is more likely than Upf1b to be a key player in *Tetrahymena* NMD because its N-terminal cysteine- and histidine-rich (CH) domain and central helicase domain, which are essential for Upf1 function in the NMD pathway, are more highly conserved (for a detailed sequence analysis, see Supplementary Figure S5). Moreover, *UPF1a* mRNA is much more abundant than *UPF1b* mRNA during vegetative growth (Figure 1C) (49).

Protein sequence analysis showed that *Tetrahymena* Upf2 contains three conserved MIF4G domains, the third of which is required to mediate the interaction between Upf2 and Upf3 (Figure 1B; Supplementary Figure S6A and B). However, the *Tetrahymena* Upf2 C-terminal sequence has weak similarity to its counterparts in yeast and humans (Supplementary Figure S6C). Interaction of the Upf2 C-terminal domain (and possibly also the N-terminal domain in humans) with the Upf1 CH domain is essential for NMD (50,51), and this feature is conserved from yeast to humans (52). Therefore, it was important to determine whether *Tetrahymena* Upf1 and Upf2 proteins can interact. Although Upf3 is the least conserved Upf protein in *Tetrahymena*, sequence alignment identified key residues that might mediate its interaction with Upf2 (Supplementary Figure S7B). Similar to yeast Upf3, *Tetrahymena* Upf3 lacks an EJC-binding motif, which reduces the likelihood that an EJC-enhanced NMD pathway operates in *Tetrahymena*.

SMG proteins are a family of metazoan-specific NMD factors. Of these, Smg1 can directly phosphorylate C-terminal [S/T]Q motifs in metazoan Upf1. Neither the *Tetrahymena* nor *S. cerevisiae* genome has an *SMG1* homolog, and the lack of this protein kinase is consistent with the absence of a target phosphorylation site in *Tetrahymena* and yeast Upf1 homologs (Figure 1B and Supplementary Figure S5A). Homologs of Smg5 and Smg7, which bind to the C-terminal Smg1 phosphorylation site of human Upf1, are also absent in *Tetrahymena*. Interestingly, in contrast to yeast, a human Smg6 homolog (named Smg6L, for Smg6 like) with a ribonuclease domain and a conserved Est1 DNA/RNA-binding domain was identified in *Tetrahymena*. The Smg6L nuclease domain resembles the NYN domain in the human ribonuclease Zc3h12a (Mcpip) rather than the PilT N-terminus (PIN) domain in the human Smg6 ribonuclease (Supplementary Figure S8A–C).

Further analysis of Smg6L proteins in several other ciliated and apicomplexan protozoa revealed that a C-terminal NYN domain is common and that key residues in the nuclease catalytic center are highly conserved (Figure 1D). Although previous studies revealed that the poorly characterized NYN ribonuclease domain has structural similarity to the PIN domain, there is no evidence to support a role in degrading PTC-containing transcript (53,54). In addition, *Tetrahymena* Smg6L lacks the EJC-binding motif of human Smg6, and its 14-3-3-like domain lacks key evolutionarily conserved residues required to mediate an interaction with phosphorylated Upf1 (Supplementary Figure S8D).

Interestingly, and in contrast to *S. cerevisiae*, *Tetrahymena* has orthologs for three of the four core components of human EJC (Figure 1A): Mago nashi (*Tetrahymena* Mag1), Y14 and eIF4A3. All have high sequence similarity to their metazoan counterparts, suggesting that they play a conserved role in RNA metabolic processes (Supplementary Figure S7C and E). In addition, BLAST searches identified orthologs of some EJC auxiliary factors in *Tetrahymena* (listed in Supplementary Table S3).

Identification of NMD targets by transcriptome sequencing of *UPF1a*-deleted cells

PTCs can be introduced by nonsense mutation, erroneous transcription or aberrant pre-mRNA splicing. In addition, normal transcripts containing an actively translated ORF in the 5′-UTR (i.e. an uORF) or an aberrant feature (e.g. a spliceosomal intron) in the 3′-UTR are also potential targets of mammalian NMD (55) and as such are expressed at very low levels in normal cells. Therefore, disrupting this mRNA quality control system by depleting NMD core factors can lead to the retention of PTC-containing transcripts, thereby dramatically increasing their levels. Hence, variations in PTC-containing transcript expression in different deletion mutants might reveal genuine NMD factors. Unfortunately, because of their low expression levels and aberrant sequence structure, PTC-containing transcripts (except for potential uORF-containing transcripts) are likely to be discarded during genome annotation and are therefore not recorded in the current *Tetrahymena* genome annotation (released in 2014) (56–58).

Bioinformatics analysis of the two *Tetrahymena* Upf1 proteins showed that Upf1a is more likely to be a functional NMD factor. Hence, we generated a macronuclear *UPF1a* knockout strain to investigate variations in PTC-containing transcript expression by transcriptome sequencing (Supplementary Figure S3). Complete *UPF1a* knockout is not lethal, leading only to a modest extension in generation time (Supplementary Table S7).

Over 60 and 64 million sequencing reads were mapped to the *Tetrahymena* genome by Illumina sequencing and read mapping, respectively, of deep sequencing libraries prepared using total RNA extracted from two biological replicates of *ΔUPF1a* and WT cells (Supplementary Table S4). Subsequent gene expression quantification with the Cuffdiff algorithm identified 875 genes with at least 2-fold higher expression in *ΔUPF1a* than in WT cells: these were defined as upregulated genes in *ΔUPF1a* (Supplementary Table S5). As uORFs in transcripts might also trigger NMD, we investigated whether putative uORFs were present in these upregulated genes. In total, 119 putative uORF-coding regions (with a product length of >20 amino acids) were found in the 23 430 quantifiable genes (genes with expression values; accounting for 86.7% of *Tetrahymena* protein-coding genes), and 14 of these uORF genes were upregulated in the *ΔUPF1a* mutant. Fisher's exact test indicated that uORF-containing genes are significantly enriched in the pool of genes upregulated in *ΔUPF1a* (P-value = 0.0003564). As another major group of NMD targets is generated by AS, we also searched for PTC-bearing transcripts derived from AS isoforms. Using the Cufflinks toolkit, the DEXSeq algorithm, and manual validation with IGV, we identified 274 NMD-sensitive transcripts generated by AS. Among these, 25 isoforms harbor a putative uORF (Supplementary Table S8). Detailed descriptions of the analysis procedures can be found in Supplementary Figure S4.

To verify that structural annotation and quantification of the expression of newly assembled PTC-bearing isoforms were accurate, we performed qRT-PCR and/or RT-PCR analyses of 15 gene loci. Nonsense transcripts derived from these gene loci are formed via different types of AS. As expected, the size of RT-PCR products amplified from PTC-containing and normal isoforms were consistent with annotations for novel assembled isoforms (Figure 2; Supplementary Figures S9 and 10). Moreover, qRT-PCR and/or RT-PCR confirmed that PTC-containing isoform levels were significantly higher in *ΔUPF1a* than in WT cells.

In general, blocking translation with an inhibitor such as CHX leads to the inefficient degradation of NMD targets, although some exceptions have been reported (5,59). Therefore, we investigated whether expression of the newly identified PTC-containing transcripts was changed by CHX treatment. Both qRT-PCR and RT-PCR results showed pronounced retention of those PTC-containing transcripts generated by alternative donor splicing after CHX treatment (Figure 2D; Supplementary Figures S9 and 10). This result suggests that NMD targeting of this class of PTC-containing transcripts may rely on translation. Some, but not all, PTC-containing transcripts generated by other types of AS were also retained in CHX-treated cells. The partial overlap between the groups of PTC-containing transcripts retained in NMD mutants and those retained in CHX-treated cells has also been observed in plants (5).

Experimental confirmation of NMD factors

Bioinformatics analysis of *Tetrahymena* NMD factor homologs showed that some domains responsible for mediating interactions between NMD factors in other organisms are not conserved in *Tetrahymena* (Figure 1B). Sequence divergence among these NMD factor homologs prompted us to investigate whether they have conserved functions in the NMD pathway. The loss of NMD factors dramatically upregulates PTC-bearing transcripts. Hence, we generated macronuclear gene knockout (or knockdown) mutants for five newly identified human NMD factors in *Tetrahymena*: the other three Upf proteins, the EJC core component Mag1, and the conserved protozoan Smg6L protein (Supplementary Figure S3). *Bona fide Tetrahymena* NMD factors were identified by measuring the relative expression

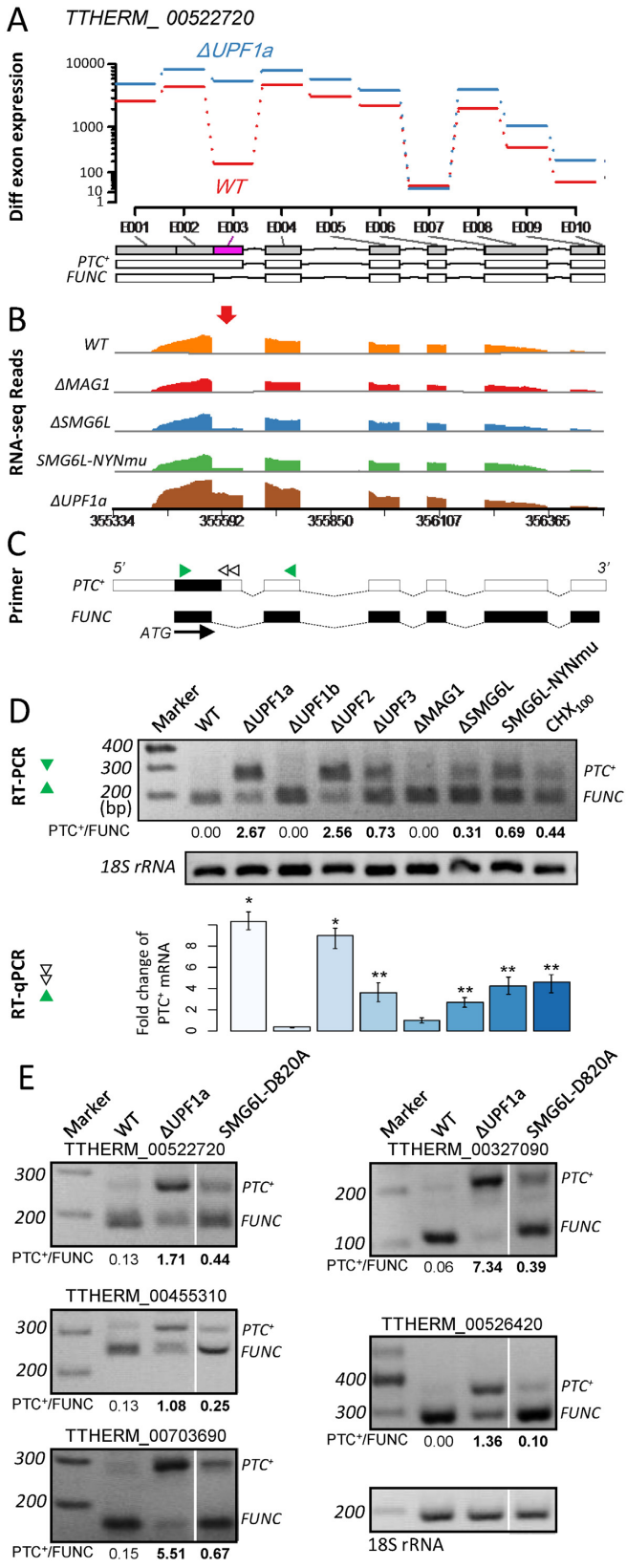

Figure 2. Analysis of the structure and expression of PTC-containing transcripts. (**A**) The PTC-introducing alternatively spliced first exon of *UBC9* is significantly upregulated in *ΔUPF1a* compared with WT cells (indicated by the counting bin E003). The short horizontal lines indicate the relative expression of each exon region in different samples; the purple box indicates an alternative exon with significant differential expression. (**B**) Accumulation of sequencing reads derived from the PTC-introducing alternative exon (red arrow) of *UBC9* can be observed in *ΔUPF1a* and both *SMG6L* mutants, but not in WT and *ΔMAG1* strains. All *y*-axes are set to the same scale. (**C**) Model of the gene encoding full-length Ubc9 protein, showing PTC-containing (PTC^+) and normal (FUNC) transcripts. The open reading frame of each transcript is shown in black; triangles indicate the locations of primers used for RT-PCR and qRT-PCR analyses. (**D**) RT-PCR and qRT-PCR analyses of PTC-containing transcripts in different cells. 'PTC^+/FUNC' indicates the band intensity of the PTC-containing transcript relative to the normal transcript; CHX_{100}, cycloheximide (100 μg/ml) treated cells (see 'Materials and Methods'section). **P*-value < 0.05, ***P*-value < 0.01. (**E**) Mutagenesis of a key catalytic residue (Asp^{820}) in the Smg6L NYN domain leads to significant accumulation of a subset of PTC-bearing transcripts (uncropped agarose gel images are shown in Supplementary Figure S11).

of PTC-bearing isoforms in mutants of several NMD factor by RT-PCR.

Results of qRT-PCR and/or RT-PCR showed that all PTC-bearing isoforms tested were significantly upregulated in *ΔUPF2* and *ΔUPF3* cells compared with the WT control, suggesting that Upf2 and Upf3 are involved in the *Tetrahymena* NMD pathway (Figure 2D; Supplementary Figures S9 and 10). In contrast, loss of the less highly conserved *Tetrahymena UPF1b* gene did not perturb PTC-containing transcript degradation, revealing that its protein product is not required for NMD. By carefully comparing the extent of PTC-containing transcript upregulation in different mutants, we found that *UPF1a* and *UPF2* gene locus disruption causes the greatest retention of nonsense transcripts, while *UPF3* knockout has a reduced effect.

We next investigated whether the protozoa-specific Smg6L nuclease is involved in NMD. For this, we first investigated changes in nonsense transcript levels in *SMG6L* knockout cells by qRT-PCR. Levels of the vast majority of NMD targets tested were significantly perturbed in *SMG6L* knockout cells (Figure 2; Supplementary Figures S9 and 10), suggesting that the protozoa-specific Smg6L protein has an evolutionarily conserved role in NMD. Next, to investigate whether Smg6L nuclease activity is required for NMD, we produced a *Tetrahymena* strain expressing endogenous levels of a Smg6L mutant with a truncated NYN ribonuclease domain (termed *SMG6L-NYNmu*). Complete deletion of the NYN domain-coding sequence was confirmed by both RT-PCR and transcriptome sequencing of the SMG6L-NYNmu-expressing strain (Supplementary Figure S3). As for whole gene deletion, qRT-PCR indicated substantial upregulation of NMD targets in *SMG6L-NYNmu* compared with WT cells (Figure 2; Supplementary Figures S9 and 10). In addition, *in silico* analysis of the sequence and structural features of the *Tetrahymena* Smg6L NYN domain revealed conservation of the residues essential for ribonuclease activity (Figure 1D). We therefore substituted Smg6L Asp^{820} (positively charged; corresponding to human Zc3h12a/Mcpip Asp^{141}) with a neutral alanine to form SMG6L-D820A. Sanger sequencing analysis of the mutated region confirmed complete substitution of the target residue (Supplementary Figure S3). Unsurprisingly, NMD targets were also upregulated in *SMG6L-D820A* cells (Figure 2E). However, in contrast to *ΔSMG6L*

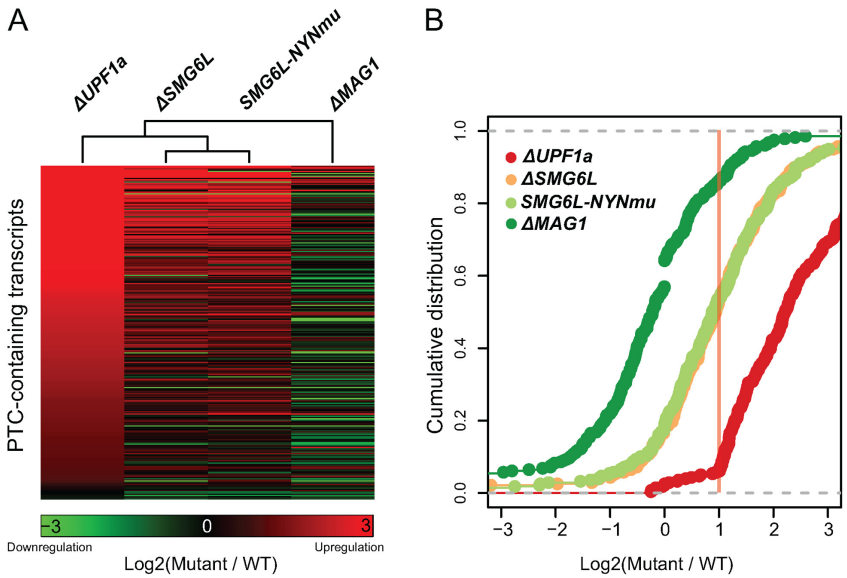

Figure 3. Fold change in expression of PTC-containing transcripts in the *ΔUPF1a*, *ΔSMG6L*, *SMG6L-NYNmu* and *ΔMAG1* strains. (**A**) Heatmap representation of the fold change in expression for each PTC-containing transcript in *ΔUPF1a*, *ΔSMG6L*, *SMG6L-NYNmu* and *ΔMAG1* mutants compared with WT cells. Pearson correlation was used for hierarchical clustering analysis. (**B**) Cumulative distribution by fold change in expression of PTC-containing transcripts in each mutant.

and *SMG6L-NYNmu* cells, substitution of Asp^{820} led to significant upregulation of only a subset of NMD targets. This result suggests that mutation of a critical residue in the catalytic center of Smg6L only partially impairs its function in the NMD pathway. To summarize, our data demonstrate that the protozoa-specific Smg6L protein has an evolutionarily conserved role in directing NMD target degradation.

Because several distinct mechanisms (and their corresponding factors) govern mRNA turnover in eukaryotes, we next investigated whether *Tetrahymena* NMD depends solely on Smg6L-mediated mRNA decay. When comparing the relative expression of PTC-containing transcripts in *ΔUPF1a*, *ΔSMG6L* and *SMG6L-NYNmu* cells, only about 50% of targets showed twofold upregulation in both the *ΔSMG6L* and *SMG6L-NYNmu*s strains, far fewer than in the *UPF1a* deletion strain (Figure 3).

EJC-binding motifs were not identified in *Tetrahymena* Upf3 and Smg6L proteins, prompting us to investigate whether EJC is involved in the *Tetrahymena* NMD pathway. For this, we generated a macronuclear *MAG1* knockout strain: complete absence of this locus was confirmed by RT-PCR and RNA-seq analysis of vegetative *ΔMAG1* cells (Supplementary Figure S3). qRT-PCR analysis of NMD target expression in *ΔMAG1* versus WT cells suggested that Mag1 protein does not contribute to the *Tetrahymena* NMD pathway (Figure 2; Supplementary Figures S9 and 10). This finding was supported by transcriptome analysis showing that very few NMD targets were upregulated in *MAG1* knockout cells (Figure 3). Moreover, hierarchical clustering (Pearson correlation) of NMD revealed a common pattern of target expression in *ΔUPF1a*, *ΔSMG6L*, and *SMG6L-NYNmu* cells, but not in *ΔMAG1* cells (Figure 3A).

In vivo and *in vitro* interactions of *Tetrahymena* NMD factors with EJC core component orthologs

A crucial step in delineating the molecular mechanisms of *Tetrahymena* NMD is defining how NMD factors interact. To identify *in vivo* interacting partners of the core *Tetrahymena* NMD factor, Upf1a, we generated a strain expressing endogenous levels of HA-tagged Upf1a. First, we investigated the *in vivo* Upf1a interactome by IP of HA-tagged proteins, followed by silver staining and immunoblotting. A preliminary analysis of the silver staining pattern identified two distinct bands specific to the Upf1a-HA sample (Figure 4A); immunoblotting revealed that the stronger band is HA-tagged Upf1a.

To identify Upf1a co-eluting protein(s), IP products were tryptic digested and analyzed by tandem mass spectrometry. Interestingly, Smg6L (117.9 kDa) was the top hit (Supplementary Table S9). This is consistent with the silver staining result: a band just below Upf1a (124.6 kDa) was specific to the Upf1a IP sample (Figure 4A). Upf2 also co-eluted with Upf1a, although this finding was only supported by two MS/MS spectra (Supplementary Table S9). We did not observe the Upf2 band (122.5 kDa) in the silver stained gel, probably because its similar molecular weight to Upf1a prevented resolution of the proteins by SDS-PAGE. Other proteins with putative functions in RNA processing and protein synthesis also co-eluted exclusively with Upf1a, suggesting that Upf1a functions in RNA metabolism. However, the confirmed NMD factor, Upf3, and the three EJC core component orthologs, Mag1, Y14, and eIF4A3, failed to co-purify with Upf1a.

To rule out the possibility that our IP method was too stringent to identify weak protein interactions, we cross-linked *Tetrahymena* cells with 0.1% paraformaldehyde (PFA) before performing IP and tandem mass spectrometry

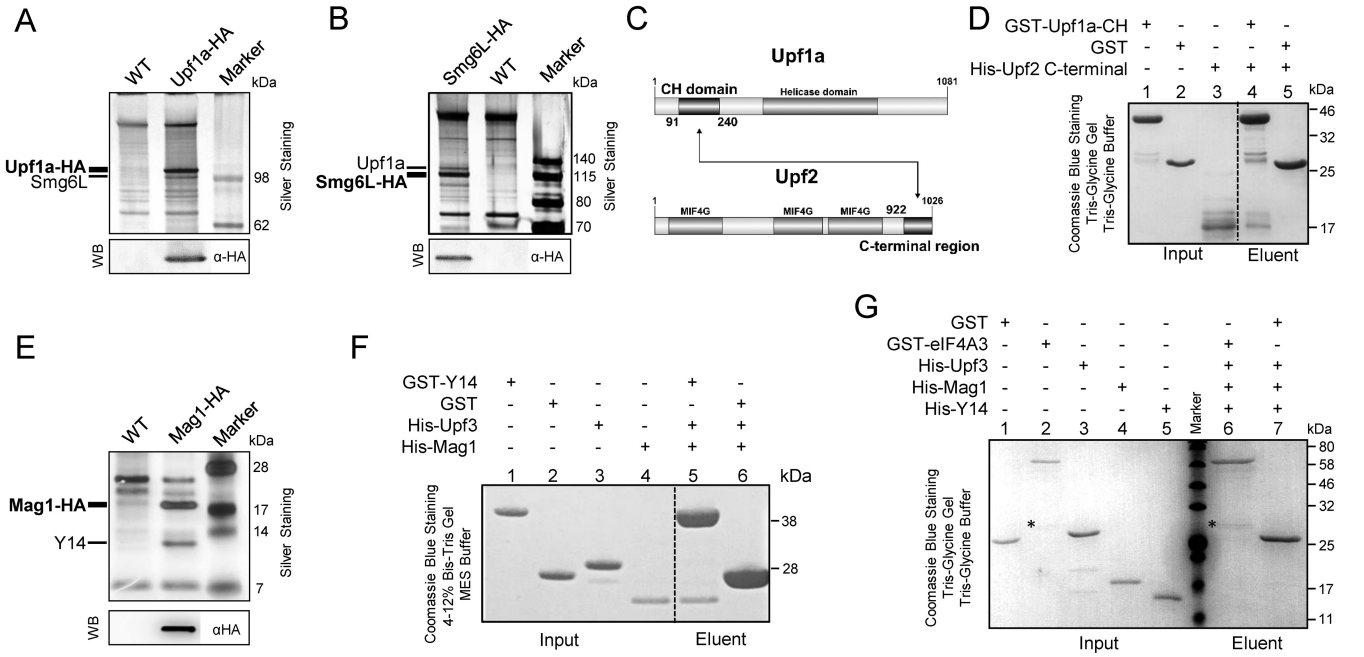

Figure 4. Protein interactions of *Tetrahymena* NMD factors and EJC homologs. (**A** and **B**) Silver-stained Upf1a and Smg6L IP products. Reciprocal IP-coupled mass spectrometry shows co-purification of Smg6L with Upf1a proteins (Supplementary Table S9). (**C** and **D**) GST pull-down shows that the Upf1a CH domain physically interacts with the Upf2 C-terminus. The His-tagged Upf2 C-terminal sequence co-purifies with the GST-tagged Upf1a CH domain but not with GST. (**E**) Silver-stained Mag1 IP products. (**F**) GST pull-down confirms that Mag1 and Y14 interact *invitro* and that neither interacts with Upf3. Lanes 5 and 6 show His-tagged Mag1 co-purification with GST-tagged Y14, but not with GST. (**G**) GST pull-down shows no protein interaction between GST-tagged eIF4A3 and other EJC homologs (Mag1-Y14 and Upf3). *A contaminant that could not be completely removed during eIF4A3 purification. All interacting proteins identified by IP-coupled mass spectrometry are listed in Supplementary Table S9.

using exactly the same procedure as before. In the crosslink IP, Smg6L was again identified as the top hit (spectrum number: 8), but this time Upf2 was the second top hit (spectrum number 6, Supplementary Table S9). However, neither Upf3 nor other EJC homologs were identified as Upf1a-interacting proteins after PFA cross-linking. Lack of Upf1b involvement in the *Tetrahymena* NMD pathway was confirmed by IP-coupled mass spectrometric analysis of the Upf1b interactome: no interaction between Upf1b and any known NMD factor was observed in normal and PFA-crosslink IPs (Supplementary Table S9).

We next performed reciprocal IPs using a strain expressing endogenous levels of HA-tagged Smg6L. As expected, Upf1a was the top hit for Smg6L co-purifying proteins, confirming the interaction between Upf1a and Smg6L (Figure 4B and Supplementary Table S9). However, IP experiments were not performed in the presence of RNase, so further experiments are required to demonstrate whether the Upf1a–Smg6L interaction is direct or mediated by RNA. Upf2, Upf3, and *Tetrahymena* EJC core component orthologs all failed to co-purify with Smg6L. The latter finding is consistent with the lack of an EJC-binding motif in Smg6L (Figure 1B).

To characterize the Upf1a–Upf2 interaction mechanism, we performed a GST pull-down assay to determine whether the Upf1a CH domain interacts with the Upf2 C-terminal region. Interestingly, although *in silico* analysis indicated that these two regions lack the key conserved residues required for protein binding and have limited amino acid similarity in relevant binding areas (Supplementary Figures S5B and 6C), GST pull-down suggested that they do interact *in vitro* (Figure 4C and D). Human Upf2 is reported to bind to Upf1 in a bipartite manner, in which both the C-terminal α-helical and β-hairpin motifs of Upf2 contribute to the interaction with Upf1 (51). The secondary structure of the *Tetrahymena* Upf2 C-terminus was therefore analyzed using the JPred algorithm (60), revealing a α-helical motif and a β-hairpin motif with structural similarity to the corresponding motifs in human Upf2 (Supplementary Figure S6C). These structures may therefore mediate the Upf1a–Upf2 interaction.

Loss-of-function analysis of the EJC core component, Mag1, demonstrated that this protein is not involved in the NMD pathway. Therefore, we doubt whether *Tetrahymena* EJC components can interact, as they do in mammals. To delineate the *in vivo* Mag1 interactome, an HA tag coding sequence was added to the 3′ terminus of the endogenous *MAG1* ORF region by homologous recombination. IP-coupled tandem mass spectrometry of the protein product identified an EJC component core component ortholog, Y14 (13.9 kDa), as the top hit for Mag1 co-eluting proteins (Supplementary Table S9). Consistent with the MS/MS result, silver staining analysis of Mag1-HA IP products identified a specific band of ∼14 kDa (Figure 4E). However, no other EJC component orthologs were identified in Mag1 IPs (Supplementary Table S9). As expected, Smg6L and Upf3 (which have no EJC-binding motifs), and Upf1a and Upf2 did not co-purify with Mag1.

GST pull-down confirmed that Mag1 and Y14 interact *in vivo* (Figure 4F). This result, along with the Mag1 IP data,

suggests that Mag1 and Y14 interact directly. When His-tagged Upf3 was also included in the GST pull-down assay, it did not interact with Y14 (Figure 4F). Thus, *Tetrahymena* Upf3 does not interact with the Mag1–Y14 heterodimer. We next performed GST pull-down to test whether Upf3 and the Mag1–Y14 heterodimer interact with the *Tetrahymena* EJC homolog, eIF4A3. Unsurprisingly, none of these proteins co-eluted with GST-tagged eIF4A3 (Figure 4G). In conclusion, *in vivo* and *in vitro* analyses of interactions among *Tetrahymena* EJC core component orthologs demonstrated that Y14–Mag1, Upf3 and eIF4A3 do not interact. Therefore, the evolutionarily conserved EJC core complex is probably absent in this organism.

Bioinformatics analysis of the structural features of NMD-associated transcripts in *Tetrahymena*

To investigate the features of PTC-introducing AS events in *Tetrahymena*, we first categorized AS subtypes using the AStalavista algorithm (46), followed by manual inspection in the IGV genome browser. Figure 5A depicts typical AS subtypes. Our analysis showed that nearly 50% of PTCs are introduced by alternative donor splice sites and fewer by alternative acceptor splice sites (Figure 5B, outer circle). Interestingly, 21 NMD targets were generated via the recently defined exitron splicing process (61). Moreover, putative uORF-containing transcripts were mainly produced by alternative 5′-UTR (Figure 5B, inner circle). By comparing the flanking sequences of PTC-introducing AS sites with those of normal splice sites, we showed that PTC-introducing AS sites have a canonical 'GT–AG' boundary. However, the composition of intronic nucleic acids proximal to the exon–intron boundary is more divergent in PTC-introducing splice sites than in canonical splice sites (Supplementary Figure S12). Therefore, aberrant splicing at these 'weak' splice sites might be the main cause of PTC-containing transcripts.

According to the canonical EJC-dependent NMD model in mammals, the main feature of NMD targets is the presence of a PTC located >50 nt upstream of the last spliceosomal intron. In contrast, NMD in *S. cerevisiae*, the fruit fly and the nematode generally functions in a '3′-UTR-dependent' manner and NMD targets are transcripts with longer 3′-UTRs. We therefore attempted to identify the PTC recognition mechanism by investigating the structural features of *Tetrahymena* NMD targets.

First, we used custom Perl scripts to investigate whether the 3′-UTRs of PTC-bearing transcript are significantly longer than those of normal transcripts (i.e. without PTCs). The 3′-UTR regions of *Tetrahymena* transcripts are poorly characterized: according to the current release of the annotated *Tetrahymena* genome, only 1235 genes (4.5% of the total) have annotated 3′-UTR regions. Using RNA-seq data, we re-annotated the 3′-UTRs of normal transcripts. As a result, the number of genes with known 3′-UTR regions increased to 4781. Using this expanded dataset, the median 3′-UTR length was compared in normal and NMD-sensitive transcripts (consisting of both PTC-containing transcripts and putative uORF-containing transcripts) using the Wilcoxon rank sum test. Statistical analysis showed that NMD-sensitive transcripts have significantly longer 3′-UTR regions compared with normal transcripts (*P*-value = 1.8e-135; Figure 5C). To investigate the relationship between 3′-UTR length and the levels of transcript expression in *ΔUPF1a* and WT cells, we first classified normal and NMD-sensitive transcripts into four groups according to their 3′-UTR length distribution (Figure 5D). The median fold change difference in transcript expression was compared among the different groups using the Wilcoxon rank sum test. This analysis revealed that the median fold change in expression was significantly higher for transcripts with longer 3′-UTRs (Figure 5D, groups 3 and 4) than for those with shorter 3′-UTRs (Figure 5D, groups 1 and 2). These results suggest that long 3′-UTR sequences may represent a marker for *Tetrahymena* NMD targets. However, the modest fold change in expression of transcripts with long 3′-UTRs in *ΔUPF1a* cells persuaded us to investigate other representative features of NMD targets.

Previous research into the *Schizosaccharomyces pombe* NMD mechanism suggested that an exon junction near to the TC triggers NMD (62). Therefore, we performed a statistical analysis of TC-proximal exon junction localization in 18 387 intron-containing normal transcripts and 285 intron-containing PTC-bearing transcripts. Exon junctions were enriched around PTCs compared with normal TCs (Figure 5E and Supplementary Figure S13). Therefore, to investigate whether transcripts with an exon junction near to a TC are more susceptible to NMD, we classified both normal and PTC-bearing transcripts into six groups according to the relative distance from the TC to the nearest exon junction, and analyzed their susceptibility to NMD by comparing the median fold change in transcript expression among groups (Figure 5F). Interestingly, NMD susceptibility seems to be modestly enhanced if an exon junction is located near to a TC (Figure 5F, group 4), but is greatly enhanced if the TC-proximal exon junction is downstream (rather than upstream) of the TC, even if it's further away (Figure 5F, groups 5 and 6). This result suggests that the *Tetrahymena* PTC recognition mechanism is not similar to that of *S. pombe*, but instead may resemble that of the mammalian EJC-enhanced model. Indeed, only 14.3% of transcripts within group 4 have exon junctions located over 55 nt downstream of the TCs, compared with 76.5 and 90.8% in groups 5 and 6, respectively (Supplementary Figure S14).

To determine whether an exon junction within the 3′-UTR (or 3′-UTR intron) is a characteristic of *Tetrahymena* NMD targets, we first compared NMD susceptibility by measuring the fold change in expression of all transcripts with different 3′-UTR lengths and either containing or lacking a 3′-UTR intron. As depicted in Figure 5G, transcripts with a 3′-UTR intron are more susceptible to NMD than those without. In addition, for transcripts with a 3′-UTR intron, those with a long 3′-UTR had a significantly higher fold change in expression compared with those with a relatively short 3′-UTR (Figure 5G, group 4 versus groups 5 and 6). Besides, as depicted in Figure 5H (group 1 and 2), transcripts without a 3′-UTR intron are generally not susceptible to NMD, even if there is an exon junction close to the TC. In contrast, transcripts containing a 3′-UTR intron are highly susceptible to NMD, even if their TCs are not proximal to an exon junction (Figure 5H, group 3). Therefore, the presence of a 3′-UTR-located spliceosomal intron

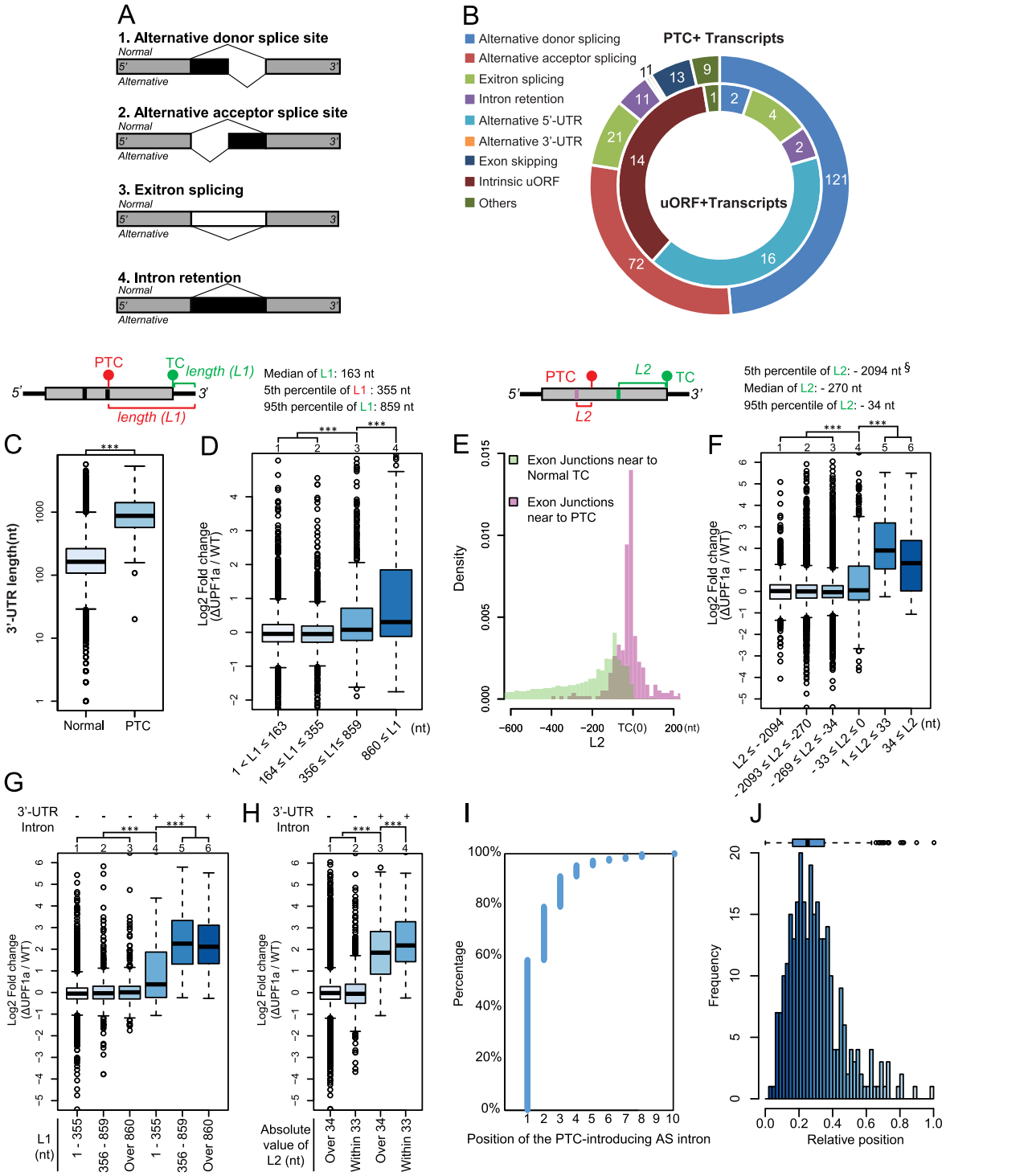

Figure 5. Bioinformatics analysis of PTC-bearing transcripts. (**A**) Schematic representation of several typical alternative splicing (AS) types. (**B**) Pie charts showing the relative proportions of PTC-introducing AS events (outer circle), uORF-introducing AS events and uORFs within normal transcripts (inner circle). (**C**) Comparison of 3′-UTR length showing that 3′-UTRs of PTC-containing transcripts are significantly longer than those of normal transcripts. (**D**) Comparison of 3′-UTR length and relative fold change in expression of transcripts in *ΔUPF1a* and WT cells. Transcripts were divided into four groups according to 3′-UTR length (based on the length distribution of 3′-UTRs in normal and PTC-containing transcripts). L1, length of the 3′-UTR region. (**E**) Localization and distribution of exon junctions near to a termination codon (TC) in intron-containing normal and PTC-bearing transcripts. L2, relative distance from the TC to its nearest exon junction. §Negative value indicates an exon junction located upstream of the TC. (**F**) Comparison of the positions of exon junctions near to a TC and the relative fold change in expression of transcripts in *ΔUPF1a* and WT cells.(**G**) Comparison of 3′-UTR length and relative fold change in expression of transcripts with or without a 3′-UTR intron in *ΔUPF1a* and WT cells. (**H**) Comparisons of the position of the exon junction near to a TC and the relative fold change in expression of transcripts with or without 3′-UTR intron in *ΔUPF1a* and WT cells. ***P-value < 0.001 (Wilcoxon rank sum test). (**I**) Analysis of PTC-introducing AS events suggests they are highly likely to occur within the first two 5′ introns. (**J**) Analysis of PTC localization within transcripts shows they are highly likely to occur within the 5′-proximal region of transcripts.

(or exon junction) is more likely to be a marker for NMD targets. Moreover, for transcripts with a 3′-UTR intron, those with TC-proximal exon junctions (Figure 5H, group 4) have a slightly (but significantly) higher median fold change in gene expression in NMD-deficient cells, which suggests that, in the presence of a 3′-UTR intron, a TC-proximal exon junction could enhance NMD.

To determine how 3′-UTR introns are formed, we analyzed the distribution of PTC-introducing AS sites (Figure 5I). Interestingly, nearly 80% of PTC-introducing AS events occurred within the first two 5′ introns. *Tetrahymena* intron-containing genes have an average of 5.1 introns. Consequently, we found that >80% of PTCs generated by AS are located in the 5′ half of PTC-containing transcripts (Figure 5J). Therefore, introduction of a PTC into the 5' proximal region redefines or 'extends' the 3′-UTR region, thus raising the possibility that a spliceosomal intron (or exon junction) is located within the redefined 3′-UTR region (Figure 5J).

Bioinformatics analysis of the function of PTC-bearing genes

We performed GO enrichment analysis to reveal the biological processes and molecular functions of PTC-bearing genes (i.e. NMD targets) and thus the biological role of NMD in vegetative cells. First, PTC-bearing genes were not associated with a particular biological process, suggesting that NMD controls genes are involved in a range of biological processes. This is consistent with the notion that the NMD pathway functions in general mRNA quality control. Second, analysis of the molecular functions of PTC-bearing genes showed significant enrichment of those encoding proteins with nucleic acid binding, GTPase and methyltransferase activities (Supplementary Table S10). These data suggest that the NMD pathway is important for maintaining these molecular functions at a steady state. Moreover, because eukaryotic homologs are lacking, 135 of PTC-bearing genes have no GO terms at all. Since ancient genes shared by many species are more likely to be functionally characterized, these 135 PTC-bearing genes are probably ciliate specific (63).

We also checked the description of each PTC-containing gene from the TGD and used the Kyoto Encyclopedia of Genes and Genomes pathway database to assign these genes to particular biological pathways. Similar to observations in human cells (64), some PTC-containing genes are involved in spliceosome formation, for example, the *Tetrahymena* ortholog of human splicing factor SF1, and *Tetrahymena* U1 and U2 snRNPs (Supplementary Table S8). Expression of some human core spliceosomal genes is reported to be critically autoregulated by a coupled AS–NMD mechanism (64,65). Therefore, our identification of PTC-containing transcripts derived from genes encoding core splicing components provides a starting point to investigate whether such a post-transcriptional gene regulatory mechanism exists in this early branching eukaryote.

DISCUSSION

Identification and characterization of conserved and lineage-specific *Tetrahymena* NMD factors

Tetrahymena has two Upf1 homologs, but only one (Upf1a) is required for NMD in vegetative growth, while the other (Upf1b) is dispensable (Figure 2; Supplementary Figures S9 and 10). Compared with Upf1b, Upf1a has a slightly higher sequence similarity to human Upf1a (Supplementary Figure S5). Moreover, Upf1a serves as a binding platform for other NMD factors (e.g. Smg6L and Upf2, see Figure 4 A–D and Supplementary Table S9), whereas Upf1b does not interact with any known NMD factor (Supplementary Table S9). As Upf1 is a multifaceted protein required for NMD as well as mediating DNA replication, telomere maintenance and several mRNA degradation processes (66), we cannot exclude the possibility that Upf1b may also function in processes other than the NMD.

Initially, the weak sequence similarity between the C-terminal domains of *Tetrahymena* Upf2 and its metazoan counterpart made us doubt the existence of a physical interaction between Upf1a and Upf2 (Supplementary Figure S6C). However, *in vivo* and *in vitro* analyses confirmed that Upf1a and Upf2 interact. Comparison of the secondary structures of *Tetrahymena* and human Upf2 C-terminal regions suggested that the *Tetrahymena* Upf2 C-terminus may adopt a structure resembling the Upf1-binding domain of human Upf2, thus enabling it to bind Upf1a (Supplementary Figure S6C). Moreover, conservation of this structural determinant in human and *Tetrahymena* Upf2 proteins suggests that the interaction between the Upf1 CH domain and the Upf2 C-terminus may be conserved throughout eukaryotes.

Tetrahymena Upf3 contains the conserved residues required to interact with the third MIF4G domain of Upf2 (Supplementary Figure S7B). Although Upf3 did not co-elute with Upf2 in Upf1 IP experiments, the possibility that Upf3 binds to Upf2 cannot be ruled out. We clearly showed that Upf3 is an important mediator of *Tetrahymena* NMD (Figure 2; Supplementary Figures S9 and 10). However, the extent of PTC-containing transcript accumulation was lower in *ΔUPF3* cells than in *ΔUPF1* and *ΔUPF2* cells. A Upf3-independent NMD pathway has been described in human cell lines (67); therefore, the presence of a similar pathway in *Tetrahymena* may explain the partial retention of PTC-containing transcripts in *ΔUPF3* cells.

Interestingly, a homolog of the metazoan Smg6 endoribonuclease is also involved in *Tetrahymena* NMD. The metazoan-specific PIN domain-containing Smg6 protein was first identified in *C. elegans*. Its endoribonuclease activity has since been reported to be involved in PTC-bearing transcript degradation in *Drosophila* and many other higher eukaryotes (25,68–70). Metazoan Smg6 is recruited to PTC-containing transcripts via direct interaction with Upf1 (25,71,72). Similarly, we observed an interaction between *Tetrahymena* Smg6L and Upf1a (Figure 4A and B; Supplementary Table S9). However, as our co-IP experiments were performed without RNase treatment, we cannot rule out the possibility that this interaction is indirect and RNA mediated. Unlike the metazoan homolog,

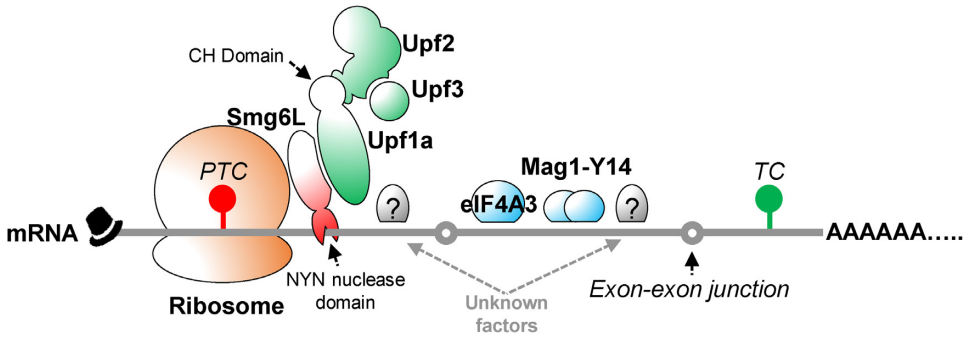

Figure 6. Schematic diagram showing how *Tetrahymena* NMD factors are involved in PTC-containing transcript degradation. Analyses of gene function, protein interactions and transcript structures have shown that the *Tetrahymena* NMD pathway functions in an EJC-independent manner. The evolutionary conserved NMD factor Upf1a plays a central role in the NMD pathway: it serves as a binding platform for Upf2 (and possibly Upf3) and recruits the protozoa-specific nuclease Smg6L to degrade PTC-containing transcripts. Although NMD-targeted transcripts are enriched with exon–exon junctions downstream of the TC, the EJC core component Mag1 is not required for NMD and not all EJC homologs can interact with one another. Therefore, further investigations are needed to identify possible novel factor(s) involved in PTC identification.

Tetrahymena Smg6L has an NYN ribonuclease domain (instead of a PIN domain) in its C-terminus. In eukaryotic proteins, the NYN ribonuclease domain is usually found along with other RNA-binding domains and may therefore function in RNA processing, e.g. microRNA biogenesis, virus RNA degradation, tRNA processing and small nucleolar RNA maturation (53,73–75). However, to our knowledge, a requirement for an NYN domain-containing protein in PTC-containing transcript destruction has not been reported. Our observation that many PTC-containing transcripts are dramatically retained in *ΔSMG6L* cells, as well as in cells expressing Smg6L protein with a truncated NYN domain or mutated NYN catalytic site, provides evidence that this NYN domain-containing protein is involved in *Tetrahymena* NMD (Figures 2 and 3; Supplementary Figures S9 and 10). It will be interesting to discover whether Smg6L endoribonuclease function is necessary for *Tetrahymena* NMD. Our identification of Smg6L homologs in other protozoa (including both free-living and parasitic protists; Figure 1D, Supplementary Figure S8 and Table S3) suggests that an NYN domain-containing Smg6L protein may be specifically required for the NMD pathway in ciliated and apicomplexan protozoa. It is thus possible that this conserved NYN domain-containing protein gained a new function in PTC-bearing transcript degradation by fusing to an Est1 DNA/RNA-binding domain. This finding improves our understanding of the functional divergence of NYN domain-containing proteins during evolution. Although Smg6L interacts with Upf1a, transcriptome analysis of *ΔUPF1a* and *ΔSMG6L* mutant cells showed that around 50% of PTC-containing transcripts are specifically retained in *ΔUPF1a* cells (and not in Smg6L mutants; Figure 3B). A similar observation that substantial NMD occurs in humans and *Drosophila* with SMG6 deficiency (76–78) is explained by partial compensation for the loss of SMG6 by an Smg6-independent pathway in metazoan cells (16,70,78). Therefore, additional factor(s) may act redundantly with Smg6L in the degradation of NMD targets in *Tetrahymena*.

Tetrahymena NMD is EJC independent

Unlike in *S. cerevisiae*, three orthologs of mammalian EJC core components (Mago nashi, Y14 and eIF4A3) and homologs of many EJC auxiliary components are present in *Tetrahymena* (Figure 1A and Supplementary Table S3). However, *in vivo* and *in vitro* protein interaction analyses showed that not all of *Tetrahymena* EJC orthologs interact with one another (Figure 4E–G). For example, Mag1 was found to directly interact with Y14 proteins, but not with other EJC components (Figure 4E–G and Supplementary Table S9), supporting the notion that Mago nashi and Y14 protein co-evolved as a heterodimer (79). Our results are similar to those of Choudhury *et al*., who recently reported that the *Drosophila* Y14–Mago heterodimer is not likely form a complex with eIF4A3 (80); moreover, *Drosophila* NMD is mainly EJC independent (81,82). Comparative protein sequence analysis suggested that the lack of interaction among *Tetrahymena* EJC component orthologs might be due to the absence of residues required to mediate eIF4A3–Y14/Mag1 interactions (Supplementary Figure S7C–E). Consistent with this, knockout of *the Tetrahymena* EJC core component, Mag1, did not affect NMD pathway function (Figures 2 and 3; Supplementary Figures S9 and 10). Moreover, the EJC-binding motif present in the metazoan NMD factors Smg6 and Upf3 is absent in *Tetrahymena* NMD factors (Supplementary Figures S7A and 8A). Inspection of the amino acid sequences of EJC orthologs and NMD factor homologs showed that a lack of residues and motifs required to mediate these protein–protein interaction is a common feature of other protists (data not shown). Overall, these data indicate that the NMD pathway of *Tetrahymena*, and probably of all ciliated protozoa, functions in an EJC-independent manner.

NMD is generally believed to facilitate intron gain events, and the extent of intron proliferation correlates with the complexity (or robustness) of NMD mechanism (83–85). For instance, the EJC-enhanced NMD pathway seems to be exclusive to lineages that may have undergone intron gain and thus have a high intron density, such as vertebrates and plants (Supplementary Figure S15) (86). In contrast, EJC components are mostly dispensable for NMD in lineages

that have undergone intron loss (such as *Drosophila* and nematodes) (81,82,87–89). For example, recent reports indicate that the *Drosophila* EJC proteins associate with nascent transcripts in an intron-independent manner and that the Y14–Mago nashi heterodimer is unlikely to form a complex with eIF4A3 (80). Our experimental confirmation of an EJC-independent NMD mechanism in *Tetrahymena* provides supportive evidence that NMD is largely independent of EJC in lineages that have undergone intron loss.

We identified 274 PTCs induced by different types of AS in *Tetrahymena* (Figure 5B, outer circle; Supplementary Table S8), despite its relatively intron-poor genome. Therefore, aberrant AS is still an important mechanism for inducing PTC-containing transcripts in this organism. We found that a large proportion (around 80%; Figure 5I) of PTC-introducing aberrant AS events took place within the first two 5′ introns and, correspondingly, that >85% of PTCs are located in the 5′ half of transcripts (Figure 5J). This process extends the 3′-UTR and thus increases the likelihood of spliceosomal introns being located downstream of the TC. Interestingly, transcripts that have an intron within the 3′-UTR region are more likely to be NMD targets in *Tetrahymena* (Figure 5G and H). Although this finding suggests an EJC-enhanced NMD mechanism, our experiments ruled out this possibility in *Tetrahymena* (Figures 2–4). Therefore, preferential NMD targeting of transcripts with an intron located downstream of the TC in *Tetrahymena* presumably relies on unknown factors related to pre-mRNA splicing (Figure 6). Nevertheless, we noticed that 16.5% of PTCs are located within the last exon and are still targeted by NMD (Supplementary Figure S13, right panel), suggesting 3′-UTR intron is not required for eliciting NMD degradation of these transcripts. Therefore, as found for all organisms investigated to date, some NMD events cannot be explained by current NMD models (2). Additional NMD mechanism(s) may be necessary to target such transcripts in *Tetrahymena*; alternatively, NMD may be simply a passive consequence of premature termination of translation, as suggested by Brogna *et al*. (2). For example, retention of PTC-containing transcripts by the *Tetrahymena ΔUPF1a* mutant seems to fit the recently proposed 'ribosome release model' (2). According to this model, in WT cells, Upf1a and its interacting proteins may disassociate proteins (including post-termination ribosomes and mRNPs) bound to the coding sequence downstream of the PTC, thus making the unprotected 3′-UTR region vulnerable to degradation. Therefore, the preferential localization of PTCs in the 5′ half of PTC-containing transcripts in *Tetrahymena* may cause exposure of a large fraction of the coding region downstream of the PTC to nucleases, which may facilitate its degradation (Figure 5J). In contrast, in *ΔUPF1a* cells, the corresponding regions of PTC-containing transcripts remain coated with proteins, enabling the transcript to evade degradation. However, further experiments are needed confirm that this model explains *Tetrahymena* NMD. Overall, our identification of NMD factors and their interacting partners, along with sequence characterization of NMD-targeted transcripts, provides a starting point for further investigations into the NMD mechanism in this early branching eukaryote.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Prof. Josef Loidl (University of Vienna) for his critical comments on the manuscript. We thank Dr Yifan Liu (University of Michigan) for sharing valuable experiences and many reagents for generating knockout mutants, expressing epitope-tagged proteins and performing IP experiments. We also thank Dr Jie Xiong, Ms Dongxia Yuan (Institute of Hydrobiology, CAS) and Mr Kangping Zhao (Northwest Normal University) for their help in analyzing transcriptome sequencing data; and Dr Anura Shodhan (University of Vienna) for her suggestions in English writing.

FUNDING

Natural Science Foundation of China [31525021, 91631303 to W.M.]; Projects of International Cooperation and Exchanges Ministry of Science and Technology of China [2013DFG32390 to W.M.]; Austrian Science Fund (FWF) [P27313-B20 to M.T.]. Funding for open access charge: Natural Science Foundation of China [31525021, 91631303 to W.M.].
Conflict of interest statement. None declared.

REFERENCES

1. Chang,Y.F., Imam,J.S. and Wilkinson,M.E. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.*, **76**, 51–74.
2. Brogna,S., McLeod,T. and Petric,M. (2016) The Meaning of NMD: Translate or Perish. *Trends Genet.*, **32**, 395–407.
3. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 189–192.
4. Kawashima,T., Douglass,S., Gabunilas,J., Pellegrini,M. and Chanfreau,G.F. (2014) Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae. *PLoS Genet.*, **10**, e1004249.
5. Drechsel,G., Kahles,A., Kesarwani,A.K., Stauffer,E., Behr,J., Drewe,P., Ratsch,G. and Wachtera,A. (2013) Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell*, **25**, 3726–3742.
6. Ni,J.Z., Grate,L., Donohue,J.P., Preston,C., Nobida,N., O'Brien,G., Shiue,L., Clark,T.A., Blume,J.E. and Ares,M. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, **21**, 708–718.
7. Brogna,S. and Wen,J.K. (2009) Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.*, **16**, 107–113.
8. Amrani,N., Ganesan,R., Kervestin,S., Mangus,D.A., Ghosh,S. and Jacobson,A. (2004) A faux 3′-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature*, **432**, 112–118.
9. Muhlrad,D. and Parker,R. (1999) Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *RNA*, **5**, 1299–1307.
10. Behm-Ansmant,I., Gatfield,D., Rehwinkel,J., Hilgers,V. and Izaurralde,E. (2007) A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *EMBO J.*, **26**, 1591–1601.
11. Eberle,A.B., Stalder,L., Mathys,H., Orozco,R.Z. and Muhlemann,O. (2008) Posttranscriptional gene regulation by spatial rearrangement of the 3′ untranslated region. *PLoS Biol.*, **6**, e92.

12. Hug,N., Longman,D. and Caceres,J.F. (2016) Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.*, **44**, 1483–1495.
13. Le Hir,H., Izaurralde,E., Maquat,L.E. and Moore,M.J. (2000) The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J.*, **19**, 6860–6869.
14. Le Hir,H., Gatfield,D., Izaurralde,E. and Moore,M.J. (2001) The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.*, **20**, 4987–4997.
15. Nagy,E. and Maquat,L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.*, **23**, 198–199.
16. Popp,M.W. and Maquat,L.E. (2013) Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.*, **47**, 139–165.
17. Kashima,I., Yamashita,A., Izumi,N., Kataoka,N., Morishita,R., Hoshino,S., Ohno,M., Dreyfuss,G. and Ohno,S. (2006) Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes Dev.*, **20**, 355–367.
18. Buhler,M., Steiner,S., Mohn,F., Paillusson,A. and Muhlemann,O. (2006) EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3 ' UTR length. *Nat. Struct. Mol. Biol.*, **13**, 462–464.
19. Singh,G., Rebbapragada,I. and Lykke-Andersen,J. (2008) A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol.*, **6**, e111.
20. Huang,L., Lou,C.H., Chan,W., Shum,E.Y., Shao,A., Stone,E., Karam,R., Song,H.W. and Wilkinson,M.F. (2011) RNA homeostasis governed by cell type-specific and branched feedback loops acting on NMD. *Mol. Cell*, **43**, 950–961.
21. He,F. and Jacobson,A. (2015) Nonsense-mediated mRNA decay: degradation of defective transcripts is only part of the story. *Annu. Rev. Genet.*, **49**, 339–366.
22. Nicholson,P. and Muhlemann,O. (2010) Cutting the nonsense: the degradation of PTC-containing mRNAs. *Biochem. Soc. Trans.*, **38**, 1615–1620.
23. Muhlrad,D. and Parker,R. (1994) Premature translational termination triggers mRNA decapping. *Nature*, **370**, 578–581.
24. He,F., Li,X.R., Spatrick,P., Casillo,R., Dong,S.Y. and Jacobson,A. (2003) Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast. *Mol. Cell*, **12**, 1439–1452.
25. Eberle,A.B., Lykke-Andersen,S., Muhlemann,O. and Jensen,T.H. (2009) SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nat. Struct. Mol. Biol.*, **16**, 49–55.
26. Chen,Y.H., Su,L.H. and Sun,C.H. (2008) Incomplete nonsense-mediated mRNA decay in Giardia lamblia. *Int. J. Parasitol.*, **38**, 1305–1317.
27. Delhi,P., Queiroz,R., Inchaustegui,D., Carrington,M. and Clayton,C. (2011) Is there a classical nonsense-mediated decay pathway in trypanosomes? *PLoS One*, **6**, e25112.
28. Contreras,J., Begley,V., Macias,S. and Villalobo,E. (2014) An UPF3-based nonsense-mediated decay in Paramecium. *Res. Microbiol.*, **165**, 841–846.
29. Jaillon,O., Bouhouche,K., Gout,J.F., Aury,J.M., Noel,B., Saudemont,B., Nowacki,M., Serrano,V., Porcel,B.M., Segurens,B. *et al.* (2008) Translational control of intron splicing in eukaryotes. *Nature*, **451**, 359–362.
30. Ruehle,M.D., Orias,E. and Pearson,C.G. (2016) Tetrahymena as a unicellular model eukaryote: genetic and genomic tools. *Genetics*, **203**, 649–665.
31. Gao,S., Xiong,J., Zhang,C.C., Berquist,B.R., Yang,R.D., Zhao,M., Molascon,A.J., Kwiatkowski,S.Y., Yuan,D.X., Qin,Z.H. *et al.* (2013) Impaired replication elongation in Tetrahymena mutants deficient in histone H3 Lys 27 monomethylation. *Genes Dev.*, **27**, 1662–1679.
32. CassidyHanley,D., Bowen,J., Lee,J.H., Cole,E., VerPlank,L.A., Gaertig,J., Gorovsky,M.A. and Bruns,P.J. (1997) Germline and somatic transformation of mating Tetrahymena thermophila by particle bombardment. *Genetics*, **146**, 135–147.
33. Sonneborn,T.M. (1974) Tetrahymena pyriformis. In: King,RC (ed). *Handbook of Genetics*. Springer, NY, pp. 433–467.
34. Robert,X. and Gouet,P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.*, **42**, W320–W324.
35. Tamura,K., Peterson,D., Peterson,N., Stecher,G., Nei,M. and Kumar,S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.
36. Xiong,J., Lu,Y., Feng,J., Yuan,D., Tian,M., Chang,Y., Fu,C., Wang,G., Zeng,H. and Miao,W. (2013) Tetrahymena functional genomics database (TetraFGD): an integrated resource for Tetrahymena functional genomics. *Database*, **2013**, bat008.
37. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
38. Pfaffl,M.W., Horgan,G.W. and Dempfle,L. (2002) Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res.*, **30**, e36.
39. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
40. Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.
41. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
42. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
43. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
44. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
45. Hayden,C.A. and Jorgensen,R.A. (2007) Identification of novel conserved peptide uORF homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC Biol.*, **5**, 32.
46. Foissac,S. and Sammeth,M. (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.*, **35**, W297–W299.
47. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
48. Teo,G.C., Liu,G.M., Zhang,J.P., Nesvizhskii,A.I., Gingras,A.C. and Choi,H. (2014) SAINTexpress: improvements and additional features in significance analysis of INTeractome software. *J. Proteomics*, **100**, 37–43.
49. Miao,W., Xiong,J., Bowen,J., Wang,W., Liu,Y.F., Braguinets,O., Grigull,J., Pearlman,R.E., Orias,E. and Gorovsky,M.A. (2009) Microarray analyses of gene expression during the Tetrahymena thermophila life cycle. *PLoS One*, **4**, e4429.
50. Kadlec,J., Izaurralde,E. and Cusack,S. (2004) The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nat. Struct. Mol. Biol.*, **11**, 330–337.
51. Clerici,M., Mourao,A., Gutsche,I., Gehring,N.H., Hentze,M.W., Kulozik,A., Kadlec,J., Sattler,M. and Cusack,S. (2009) Unusual bipartite mode of interaction between the nonsense-mediated decay factors, UPF1 and UPF2. *EMBO J.*, **28**, 2293–2306.
52. Chakrabarti,S., Jayachandran,U., Bonneau,F., Fiorini,F., Basquin,C., Domcke,S., Le Hir,H. and Conti,E. (2011) Molecular mechanisms for the RNA-dependent ATPase activity of Upf1 and its regulation by Upf2. *Mol. Cell*, **41**, 693–703.
53. Anantharaman,V. and Aravind,L. (2006) The NYN domains: novel predicted RNAses with a PIN domain-like fold. *RNA Biol.*, **3**, 18–27.
54. Xu,J.W., Peng,W., Sun,Y., Wang,X.X., Xu,Y.H., Li,X.M., Gao,G.X. and Rao,Z.H. (2012) Structural study of MCPIP1 N-terminal conserved domain reveals a PIN-like RNase. *Nucleic Acids Res*, **40**, 6957–6965.
55. Celik,A., Kervestin,S. and Jacobson,A. (2015) NMD: At the crossroads between translation termination and ribosome recycling. *Biochimie*, **114**, 2–9.

56. Stover,N.A., Krieger,C.J., Binkley,G., Dong,Q., Fisk,D.G., Nash,R., Sethuraman,A., Weng,S. and Cherry,J.M. (2006) Tetrahymena Genome Database (TGD): a new genomic resource for Tetrahymena thermophila research. *Nucleic Acids Res.*, **34**, D500–D503.
57. Stover,N.A., Punia,R.S., Bowen,M.S., Dolins,S.B. and Clark,T.G. (2012) Tetrahymena genome database Wiki: a community-maintained model organism database. *Database (Oxford)* , **2012**, bas007.
58. Coyne,R.S., Thiagarajan,M., Jones,K.M., Wortman,J.R., Tallon,L.J., Haas,B.J., Cassidy-Hanley,D.M., Wiley,E.A., Smith,J.J., Collins,K. *et al.* (2008) Refined annotation and assembly of the Tetrahymena thermophila genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. *BMC Genomics*, **9**, 562.
59. Carter,M.S., Doskow,J., Morris,P., Li,S., Nhim,R.P., Sandstedt,S. and Wilkinson,M.F. (1995) A regulatory mechanism that detects premature nonsense codons in T-cell receptor transcripts in vivo is reversed by protein synthesis inhibitors in vitro. *J. Biol. Chem.*, **270**, 28995–29003.
60. Drozdetskiy,A., Cole,C., Procter,J. and Barton,G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.
61. Marquez,Y., Hopfler,M., Ayatollahi,Z., Barta,A. and Kalyna,M. (2015) Unmasking alternative splicing inside Protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.*, **25**, 995–1007.
62. Wen,J. and Brogna,S. (2010) Splicing-dependent NMD does not require the EJC in Schizosaccharomyces pombe. *EMBO J.*, **29**, 1537–1551.
63. Zhang,Y.E., Landback,P., Vibranovski,M. and Long,M. (2012) New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays*, **34**, 982–991.
64. Saltzman,A.L., Kim,Y.K., Pan,Q., Fagnani,M.M., Maquat,L.E. and Blencowe,B.J. (2008) Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol. Cell. Biol.*, **28**, 4320–4330.
65. Saltzman,A.L., Pan,Q. and Blencowe,B.J. (2011) Regulation of alternative splicing by the core spliceosomal machinery. *Genes Dev.*, **25**, 373–384.
66. Isken,O. and Maquat,L.E. (2008) The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat. Rev. Genet.*, **9**, 699–712.
67. Chan,W.K., Huang,L., Gudikote,J.P., Chang,Y.F., Imam,J.S., MacLean,J.A. and Wilkinson,M.F. (2007) An alternative branch of the nonsense-mediated decay pathway. *EMBO J.*, **26**, 1820–1830.
68. Pulak,R. and Anderson,P. (1993) mRNA surveillance by the Caenorhabditis elegans smg genes. *Genes Dev.*, **7**, 1885–1897.
69. Gatfield,D. and Izaurralde,E. (2004) Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila. *Nature*, **429**, 575–578.
70. Huntzinger,E., Kashima,I., Fauser,M., Sauliere,J. and Izaurralde,E. (2008) SMG6 is the catalytic endonuclease that cleaves mRNAs containing nonsense codons in metazoan. *RNA*, **14**, 2609–2617.
71. Okada-Katsuhata,Y., Yamashita,A., Kutsuzawa,K., Izumi,N., Hirahara,F. and Ohno,S. (2012) N- and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD. *Nucleic Acids Res.*, **40**, 1251–1266.
72. Chakrabarti,S., Bonneau,F., Schussler,S., Eppinger,E. and Conti,E. (2014) Phospho-dependent and phospho-independent interactions of the helicase UPF1 with the NMD factors SMG5-SMG7 and SMG6. *Nucleic Acids Res.*, **42**, 9447–9460.
73. Suzuki,H.I., Arase,M., Matsuyama,H., Choi,Y.L., Ueno,T., Mano,H., Sugimoto,K. and Miyazono,K. (2011) MCPIP1 ribonuclease antagonizes dicer and terminates microRNA biogenesis through precursor microRNA degradation. *Mol. Cell*, **44**, 424–436.
74. Lin,R.J., Chien,H.L., Lin,S.Y., Chang,B.L., Yu,H.P., Tang,W.C. and Lin,Y.L. (2013) MCPIP1 ribonuclease exhibits broad-spectrum antiviral effects through viral RNA binding and degradation. *Nucleic Acids Res.*, **41**, 3314–3326.
75. Gutmann,B., Gobert,A. and Giege,P. (2012) PRORP proteins support RNase P activity in both organelles and the nucleus in Arabidopsis. *Genes Dev.*, **26**, 1022–1027.
76. Jonas,S., Weichenrieder,O. and Izaurralde,E. (2013) An unusual arrangement of two 14-3-3-like domains in the SMG5-SMG7 heterodimer is required for efficient nonsense-mediated mRNA decay. *Genes Dev.*, **27**, 211–225.
77. Luke,B., Azzalin,C.M., Hug,N., Deplazes,A., Peter,M. and Lingner,J. (2007) Saccharomyces cerevisiae Ebs1p is a putative ortholog of human Smg7 and promotes nonsense-mediated mRNA decay. *Nucleic Acids Res.*, **35**, 7688–7697.
78. Frizzell,K.A., Rynearson,S.G. and Metzstein,M.M. (2012) Drosophila mutants show NMD pathway activity is reduced, but not eliminated, in the absence of Smg6. *RNA*, **18**, 1475–1486.
79. Gong,P.C., Zhao,M. and He,C.Y. (2014) Slow co-evolution of the MAGO and Y14 protein families is required for the maintenance of their obligate heterodimerization mode. *PLoS One*, **9**, e84842.
80. Choudhury,S.R., Singh,A.K., McLeod,T., Blanchette,M., Jang,B., Badenhorst,P., Kanhere,A. and Brogna,S. (2016) Exon junction complex proteins bind nascent transcripts independently of pre-mRNA splicing in Drosophila melanogaster. *Elife*, **5**, e19881.
81. Sauliere,J., Haque,N., Harms,S., Barbosa,I., Blanchette,M. and Le Hir,H. (2010) The exon junction complex differentially marks spliced junctions. *Nat. Structu. Mol. Biol.*, **17**, 1269–1271.
82. Gatfield,D., Unterholzner,L., Ciccarelli,F.D., Bork,P. and Izaurralde,E. (2003) Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. *EMBO J.*, **22**, 3960–3970.
83. Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.
84. Farlow,A., Meduri,E., Dolezal,M., Hua,L.S. and Schlotterer,C. (2010) Nonsense-mediated decay enables intron gain Drosophila. *PLoS Genet*, **6**, e1000819.
85. Tange,T.O., Nott,A. and Moore,M.J. (2004) The ever-increasing complexities of the exon junction complex. *Curr. Opin. Cell Biol.*, **16**, 279–284.
86. Csuros,M., Rogozin,I.B. and Koonin,E.V. (2011) A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput. Biol.*, **7**, e1002150.
87. Longman,D., Plasterk,R.H.A., Johnstone,I.L. and Caceres,J.F. (2007) Mechanistic insights and identification of two novel factors in the C-elegans NMD pathway. *Genes Dev.*, **21**, 1075–1085.
88. Chapin,A., Hu,H., Rynearson,S.G., Hollien,J., Yandell,M. and Metzstein,M.M. (2014) In vivo determination of direct targets of the nonsense-mediated decay pathway in Drosophila. *G3*, **4**, 485–496.
89. Gatfield,D. and Izaurralde,E. (2004) Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila. *Nature*, **429**, 575–578.