Repeat-associated non-ATG (RAN) translation in neurological disease

John D. Cleary and Laura P.W. Ranum*

Center for NeuroGenetics, Departments of Molecular Genetics & Microbiology and Neurology, College of Medicine, Genetics Institute, University of Florida, Gainesville, FL, USA

Received July 28, 2013; Revised and Accepted July 29, 2013

Well-established rules of translational initiation have been used as a cornerstone in molecular biology to understand gene expression and to frame fundamental questions on what proteins a cell synthesizes, how proteins work and to predict the consequences of mutations. For a group of neurological diseases caused by the abnormal expansion of short segments of DNA (e.g. CAG•CTG repeats), mutations within or outside of predicted coding and non-coding regions are thought to cause disease by protein gain- or loss-of-function or RNA gain-of-function mechanisms. In contrast to these predictions, the recent discovery of repeat-associated non-ATG (RAN) translation showed expansion mutations can express homopolymeric expansion proteins in all three reading frames without an AUG start codon. This unanticipated, non-canonical type of protein translation is length- and hairpin-dependent, takes place without frameshifting or RNA editing and occurs across a variety of repeat motifs. To date, RAN proteins have been reported in spinocerebellar ataxia type 8 (SCA8), myotonic dystrophy type 1 (DM1), fragile X tremor ataxia syndrome (FXTAS) and C9ORF72 amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD). In this article, we review what is currently known about RAN translation and recent progress toward understanding its contribution to disease.

INTRODUCTION

Repeat-expansion disorders are a class of neurological and neuromuscular diseases caused by the expansion of short repetitive elements within the human genome. The genic location of the expansion has been traditionally used to classify these disorders into coding expansions caused by protein gain-of-function effects, and non-coding expansions caused by either a loss-of-function of the affected gene or RNA gain-of-function effects (1–3). For protein gain-of-function diseases, the expansion mutation is translated as part of a larger open-reading frame (ORF), resulting in the expression of a mutant protein that disrupts cellular function and induces toxicity. For example, in Huntington’s disease (HD) the CAG expansion mutation is translated as part of a larger open-reading frame (ORF), resulting in the expression of a mutant protein that disrupts cellular function and induces toxicity. For example, in Huntington’s disease (HD) the CAG expansion mutation is translated as part of the huntingtin protein, which results in protein aggregation and cellular dysfunction (4). For RNA gain-of-function disorders, non-coding expansion RNAs accumulate as nuclear foci that sequester RNA-binding proteins and lead to a loss of their normal function (5,6). For example, in myotonic dystrophy type 1 (DM1) and type 2 (DM2), CUG or CCUG expansion RNAs sequester MBNL proteins from their normal splicing targets, such that the resulting MBNL loss-of-function leads to alternative splicing dysregulation (7–10). The recent discovery of repeat-associated non-ATG (RAN) translation (11) showed that microsatellite expansions do not follow the canonical rules of translation initiation and can generate a series of unexpected repeat proteins. This finding opens the door to new paradigms in disease mechanisms and cell biology. In this review, we discuss the discovery of RAN translation, what is currently known about its molecular biology and progress toward understanding its contribution to disease.

INITIAL DISCOVERY OF RAN TRANSLATION IN SCA8

RAN translation was initially discovered by Zu et al. (11) while investigating the molecular mechanisms of spinocerebellar ataxia type 8 (SCA8). SCA8 is a dominantly inherited, slowly...
progressive neurodegenerative disorder caused by a CTG•CAG repeat expansion (12). Both RNA and protein disease mechanisms likely operate in SCA8 as bidirectional transcription produces both a CUG expansion transcript that forms RNA foci (13) and a CAG expansion transcript with an unusual ATG-initiated ORF encoding a nearly pure polyGln expansion protein (Fig. 1A) (14). The first evidence for RAN translation came when Zu et al., trying to separate the RNA and protein gain-of-function effects, found that removing the only ATG initiation codon within an SCA8 minigene did not prevent expression of the polyGln protein (Fig. 1B) (11). Subsequent experiments with epitope-tagged minigenes showed CAG expansions lacking an ATG initiation codon produce distinct homopolymeric protein products in all the three reading frames, polyGln, polyAla and polySer (Fig. 1C). Because these findings were novel, and completely unexpected multiple approaches were used to characterize the transcripts and to establish the identity of these proteins.

Analysis of polyribosome-bound transcripts showed no evidence of RNA editing that could have introduced a start codon (11). Immunoprecipitation and analysis of C- and N-terminal epitope-tagged constructs demonstrated that RAN translation does not require frameshifting, and that RAN occurs in all the reading frames even in the presence of an ATG-initiated ORF. Additionally, a combination of epitope tags, tritium labeling and mass spectrometry unequivocally proved that these proteins contain expanded polyAla, polyGln or polySer repeat tracts. For polyalanine, mass spectrometry identified a series of N-terminal peptides containing varying numbers of alanines. No peptides containing N-terminal methionine were identified. These data suggest that translation in the polyAla reading frame begins with an alanine, and that start sites occur at various positions throughout the length of the repeat tract.

Additional experiments on RAN translation (11) demonstrated a number of features. First, immunofluorescence showed RAN proteins expressed from all three reading frames can accumulate in a single cell, although more frequently only one or two RAN proteins were found. Second, RAN proteins expressed across CAG expansions increase apoptosis, suggesting a potential contribution to disease. Third, RAN proteins are expressed across hairpin-forming CAG but not across non-hairpin-forming CAA repeats in cell culture. These data suggest that structured RNAs may be required for RAN translation. Fourth, RAN translation also occurs across CUG expansion transcripts. Fifth, longer CAG repeat tracts are associated with the simultaneous expression of multiple protein products with a different length threshold required for translation in each frame (Table 1). Taken together, these data demonstrated that CAG and CUG expansion transcripts undergo a novel type of protein translation in which homopolymorphic proteins are expressed in all the three reading frames without an ATG-initiation codon.

IN VIVO EVIDENCE FOR RAN TRANSLATION IN SCA8

After establishing that RAN translation occurs in transfected cells, Zu et al. looked for evidence that SCA8 RAN proteins are expressed in vivo (11). SCA8 is characterized by severe cerebellar atrophy with Purkinje cell degeneration and loss of granule cells (14). Zu et al. (11) developed antibodies against the unique C-terminal region of the predicted CAG-encoded polyAla RAN protein and showed polyAla-positive immunostaining in cerebellar Purkinje cells from human SCA8 but not control autopsy tissue. These antibodies also detected polyAla RAN proteins in Purkinje cells from an established mouse model of SCA8 (14). SCA8 Purkinje cells are also known to accumulate CUG RNA foci (13) and poly-Gln inclusions (14). Although additional studies are needed to understand the effects of RAN proteins in SCA8, the accumulation of the SCA8 polyAla protein in Purkinje cells suggests that RAN proteins may contribute to disease.

RAN TRANSLATION IN MYOTONIC DYSTROPHY TYPE 1

Additional in vivo evidence for RAN translation was demonstrated in myotonic dystrophy (11). DM1, one of the best examples of an RNA gain-of-function disease (5), is caused by a CUG expansion in the 3’ UTR of the DMPK gene (15–17). Antisense transcripts in the CAG direction have also been reported (11,18). To determine whether RAN translation also occurs for DM1,
Zu et al. (11) performed immunostaining with two types of antibodies: (i) a well-established monoclonal antibody that detects expanded polyGln tracts (19,20) and (ii) a novel antibody developed to detect the unique C-terminal region of the predicted CAG encoded poly-Gln RAN protein (11). Positive immunostaining was observed in DM1 myoblasts, skeletal muscle and blood. Similar staining was found in an established DM1 mouse model (21,22), which showed staining of cardiomyocytes and leukocytes (11). Additionally, in both humans and mice polyGln aggregates co-localized with caspase-8, an early indicator of polyGln-induced apoptosis (23). Although RNA gain-of-function effects in DM1 are known to cause specific alternative splicing changes, these results suggest the possibility that RAN translation may also contribute to this disorder.

The discovery of RAN translation combined with growing evidence that many microsatellite expansion mutations are transcribed in both directions (2) suggests that in addition to previously considered gene products, expansion mutations may also express up to six additional RAN proteins (Fig. 2)—each of which may contribute to disease (Fig. 3). Consistent with this prediction, RAN translation has recently been reported in two additional disorders: fragile X-associated tremor ataxia syndrome (FXTAS) (24) and C9ORF72 amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD) (25–27).

Table 1. In vitro characteristics and in vivo detection of RAN translation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SCA8</td>
<td>CAG</td>
<td>polyGln</td>
<td>&gt;42 repeats</td>
<td>ATG-polyGln, cerebellum and brain stem</td>
<td>(11)</td>
</tr>
<tr>
<td>DM1</td>
<td>CAG</td>
<td>polyAla</td>
<td>&gt;73 repeats</td>
<td>Cerebellum</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>polySer</td>
<td>&gt;58 repeats</td>
<td>ND</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>polyGln</td>
<td>ND</td>
<td>Myoblasts, skeletal muscle, peripheral blood leukocytes</td>
<td>(11)</td>
</tr>
<tr>
<td>FXTAS</td>
<td>CGG</td>
<td>polyGly</td>
<td>&gt;30 repeats</td>
<td>Frontal cortex, cerebellum, hippocampus</td>
<td>(24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>polyAla</td>
<td>&gt;88 repeats</td>
<td>ND</td>
<td>(24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>polyArg</td>
<td>UD</td>
<td>ND</td>
<td>(24)</td>
</tr>
<tr>
<td>ALS/FTD</td>
<td>GGGGCC</td>
<td>polyGlyPro</td>
<td>&gt;145 repeats</td>
<td>Cerebellum, hippocampus, iPSC-derived neurons, neocortex, medial and lateral geniculate nuclei, testes</td>
<td>(25–27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>polyGlyAla</td>
<td>&gt;38 repeats</td>
<td>Cerebellum, hippocampus</td>
<td>(26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>polyGlyArg</td>
<td>UD</td>
<td>Cerebellum, hippocampus</td>
<td>(26)</td>
</tr>
</tbody>
</table>

ND, not examined and not determined; UD, examined but undetermined.

*aAntisense transcript.

Figure 2. Model of RAN translation across repeats in coding and non-coding gene regions. Schematic diagram showing mutations located in intronic or exonic regions with expression of distinct RAN proteins in three frames from sense and antisense directions. For expansions in introns, sense and antisense transcripts (not shown) produce RAN proteins with different repeat motifs and distinct C-terminal regions not corresponding to any endogenous proteins. For repeat-expansion mutations located in ORFs, up to six distinct RAN proteins may be produced from sense and antisense transcripts (see upper inset for antisense RAN proteins). The RAN protein expressed in the ORF is predicted to start at or close to the repeat and contain the same C-terminal region as the protein expressed from the canonical ATG-initiated ORF. Variability of RAN proteins will occur when expressed from: sense or antisense transcripts; different repeat motifs and with variations in C-terminal sequences.
RAN TRANSLATION AND THE CGG REPEATS OF FRAGILE X TREMOR ATAXIA SYNDROME (FXTAS)

Fragile X-associated tremor ataxia syndrome (FXTAS) is a late-onset disorder that primarily affects the cerebellum and causes coordination deficits and cognitive decline (28–30). This is caused by a specific range of expanded CGG repeats (55–200 repeats) within the 5′ UTR of the FMR1 gene (28), whereas longer repeats (>200 CGGs) are associated clinically distinct Fragile X syndrome (31). In contrast to the transcriptional silencing and loss of protein expression in Fragile X syndrome (32), FXTAS is associated with increased CGG transcripts that accumulate as RNA foci in human autopsy tissue (33). The associated increased mRNA expression, neurodegeneration and CGG-repeat containing neuronal inclusions (33–35) suggested an RNA gain-of-function mechanism. However, not all aspects of disease pathology, such as inclusion size and associated proteins (34,36), are readily explained by this mechanism. Recent work by Todd et al. (24) has shown that RAN translation may explain some of these incongruous aspects of FXTAS pathology.

Initially, Todd et al. (24), noticed aggregates in a fly model designed to express a non-coding CGG<sub>EXP</sub> mutation upstream of a GFP reporter. This group performed a series of experiments to understand the molecular basis of these aggregates and to test the hypothesis that FXTAS CGG expansion mutations undergo RAN translation. First, they showed evidence from Drosophila, including mass spectrometry, that a high-molecular weight fusion protein is expressed that contains a homopolymeric glycine expansion. Second, in transfected mammalian cells they showed CGG expansions trigger RAN translation in at least two out of three reading frames producing polyGly-GFP and polyAla–GFP fusion proteins. Third, in the polyAla frame, RAN translation is length dependent with polyAla detected using constructs with 88 but not 30 CGG repeats. In contrast,polyGly expression occurred with 88, 50 and 30 CGGs (Table 1). While the poly-Gly protein was produced from constructs containing only 30 repeats, aggregation was only associated with longer repeats tracts. Fourth, these authors performed a number of experiments that indicate translation initiation can begin upstream of the CGG repeat in the polyGly reading frame. Fifth, these authors show evidence that the polyGly RAN protein accumulates as aggregates in several model systems and in human FXTAS brains using several custom C-terminal antibodies. In summary, Todd et al. (24)
RAN TRANSLATION AND C9ORF72 ALS

A large G4C2 hexanucleotide repeat expansion in intron 1 of the C9orf72 gene was recently identified as the most common cause of ALS/FTD (37,38). Repeat tracts in unaffected controls typically contain fewer than 23 G4C2 repeats, while expansions in ALS/FTD patients range from hundreds to more than 1000 repeats (37–39). Initially, haploinsufficiency and RNA gain-of-function were suggested as possible disease mechanisms because the expansion mutation decreases C9orf72 transcript levels and G4C2 expansion transcripts form RNA foci (37). Two recent studies suggest RAN translation as a third possible mechanism (25,26).

RAN translation of the C9orf72 G4C2 hexanucleotide expansion mutation is predicted to result in the expression of dipeptide proteins: GlyPro (GP), GlyArg (GR) and GlyAla (GA). Two groups developed antibodies to these predicted dipeptide motifs and used them to examine patient tissues to look for in vivo evidence of RAN translation (25,26). Mori et al. (26) used antibodies to all three predicted dipeptide products, while Ash et al. (25) focused on the GP frame. Both the groups performed a detailed examination of patient tissues and showed that these antibodies recognize inclusions in C9orf72 ALS/FTD autopsy tissue. In the Mori et al. study (26), the GA antibody, and to a much lesser extent the GP and GR antibodies, detected inclusions in the cerebellum, hippocampus and other brain regions. These inclusions were similar in shape and abundance to typical ALS/FTD inclusions (40) and colocalized with p62 but not phospho-TDP-43 (26). Inclusions that are p62-positive/phospho-TDP43 negative are classic features of ALS/FTD pathology (40–43). In the Ash et al. study (25), the GP antibodies detected widespread neuronal cytoplasmic and intranuclear inclusions throughout the central nervous system. These inclusions were also morphologically similar to the classic ALS inclusions (25). In both the studies, these antibodies did not detect aggregates in C9orf72-negative disease controls (25,26). More recently, Almeida et al. (27) showed that neurons derived from C9orf72-positive iPS cells have GP-positive aggregates, elevated p62 levels and an increased sensitivity to cellular stress induced by autophagy inhibitors. Taken together, data from these studies suggest that dipeptide repeat proteins, expressed by RAN translation, contribute to ALS/FTD.

COMMON THEMES IN RAN TRANSLATION

RAN translation has now been reported in four diseases and has been shown to occur across four different types of repeat motifs: CAG, CUG, CGG and GGGGCC. Among this diversity, several common themes are emerging. First, RAN translation is repeat length-dependent with translation more likely with longer expansion mutations. Second, RAN translation in different reading frames have different length thresholds, such that longer repeats are more likely to result in the accumulation of a cocktail of RAN proteins expressed from different reading frames. It is possible that the simultaneous expression of RAN proteins across long repeats may play a role in anticipation, the earlier onset and increased disease severity associated with longer repeats. Third, all RAN-competent repeat motifs described to date form unusual secondary structures (44–48). Fourth, all disorders in which RAN translation has been reported to date have neurological features.

NEXT STEPS IN RAN TRANSLATION

What are the critical next steps in RAN translation research? From the analysis so far, it is clear that research needs to move beyond the observational and into the mechanistic. For example, what are the precise RNA structural, sequence and protein factor requirements for RAN translation? Answering these questions will yield important clues to the breadth and scope of RAN translation. Future analysis also needs to be extended beyond immunological approaches to more detailed structural and biochemical analyses of RAN translation proteins in disease. Antibody-based techniques are often subject to artifacts and technical problems, which may be particularly problematic for antibodies directed against repeat motifs themselves. Additionally multiple approaches will be necessary to validate results, especially given the possibility of overlap between RAN translation and other cellular processes. For example, the products of RAN translation and frameshifting may appear to be identical when looking at regions only downstream of the repeat motif. Given the discovery of RAN translation, previous reports of frameshifting for disorders such as SCA3 and HD (49–51) warrant re-examination. A more general question is does RAN translation occur across all microsatellite expansion diseases and if so when, where and why? Additional studies will be required to sort out which RAN proteins are toxic and their potential contribution to disease.

CONCLUSIONS

In summary, RAN translation is a novel mechanism that impacts our basic understanding of gene expression, cell biology and disease. Because more than 30 diseases are caused by microsatellite expansion mutations RAN translation may produce an abundant, yet previously unrecognized set of mutant proteins that contribute to a large category of neurological diseases. Additionally, recent evidence from ribosome profiling studies (24,52–58) suggests that translation is more widespread than previously appreciated. Furthermore, because >50% of the human genome consists of repetitive DNA and repetitive, hairpin-forming sequences undergo RAN translation, the discovery of RAN translation could reveal an abundant, yet previously unrecognized category of repeat-containing proteins.

ACKNOWLEDGEMENT

The authors wish to thank Dr Tao Zu and Ms. Yuanjing Liu for helpful comments and suggestions.

Conflict of Interest statement. L.P.W.R. is named as an inventor on patents for a gene test for SCA8 and on pending patents on RAN translation.
FUNDING
This work was supported by the National Institutes of Health to (R01NS058901 and R01NS040389), Muscular Dystrophy Association, Keck Foundation, CHDI and Target ALS to L.P.W.R., and the Myotonic Dystrophy Foundation to J.D.C. Funding to pay the Open Access publication charges for this article was provided by the Center for NeuroGenetics, College of Medicine, University of Florida.

REFERENCES
et al. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron, 72, 245–256.