Review

Repeat-associated RNA structure and aberrant splicing∗

Melissa A. Halea, Nicholas E. Johnsona, J. Andrew Berglunda,b,*

a Department of Neurology, Virginia Commonwealth University, Richmond, VA 23298, USA
b The RNA Institute, Department of Biological Sciences, University at Albany, State University of New York, Albany, NY 12222, USA

1. Introduction

RNA splicing is a complex and dynamic process by which protein coding sequences (exons), split within the genome by intervening non-coding regions (introns), are reconstituted into a continuous, mature mRNA molecule by the macromolecular complex known as the spliceosome. Alternative splicing (AS), defined as the differential inclusion of coding or non-coding regions within a single mRNA, significantly expands the complexity of the cellular transcriptome and proteome, increasing genomic diversity from a limited population of protein-coding genes [1–4]. Regulation of AS is controlled by two major interacting factors: (1) cis-regulatory sequences within the pre-mRNA that influence splice site selection and (2) their recognition by trans-acting factors, spliceosomal U snRNPs and RNA binding proteins (RBPs), that function to drive specific AS patterns. Precise modulation of AS “choices” is critical for proper gene expression in a developmental, spatio-temporal, and tissue-dependent manner [5,6].

While the process of RNA splicing relies on multiple layers of regulatory elements, many years of research have shown that RNA structure within pre-mRNAs can significantly impact splicing outcomes by both inhibiting and enhancing spliceosome assembly on intron/exon junctions. Specifically, structural elements within a target pre-mRNA often act as the scaffolding for splicing machinery, altering the availability of binding motifs and regulatory sequences. For example, RNA stem loops stabilized by RBPs often block access to key cis-regulatory RNA sequences by the spliceosome, subsequently leading to the exclusion of specific exons from the final mRNA product. Conversely, long-range intronic RNA structures have been shown to enhance splicing by bringing these regulatory sequences closer together within physical space (reviewed in [7,8]). Consequently, genetic mutations on AS that is causative of human disease. One such class of genetic diseases termed microsatellite repeat expansion disorders result in part due to the formation of pathogenic RNA structures that impact global patterns of AS.

Microsatellite repeats, defined as repetitive sequence motifs of one to six nucleotides, make up about 3% of the human genome [9]. While the function of these microsatellites is largely unknown, expansions in the number of repeats within these genomic regions have been linked with over 30 hereditary human disorders, including myotonic dystrophy type 1 and 2 (DM1 and DM2), amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD), and spinocerebellar ataxias (SCAs) (reviewed in [10,11]). From a pathogenic perspective, the effect of these tandem repeats is extensive and often dependent on the location of the repeat. While found in all regions of genetic architecture, including non-coding introns, 5′/3′ untranslated regions (UTRs), and...
coding exons, nearly all are impacted by pathogenic RNA gain-of-function mechanisms that arise from transcription of the repetitive element. RNAs produced from genes containing these expanded repeats form persistent RNA structural elements that aggregate to form intracellular “foci.” These structured RNAs become toxic to the cell often due to their ability to bind and sequester RBPs, including those that regulate AS. More recently, these expanded RNAs have been shown to be translated in all three reading frames in a non-AUG dependent manner through a mechanism known as RAN translation producing toxic peptides that likely contribute to disease pathogenesis (reviewed in [12,13]).

The best examples of this RNA gain-of-function disease mechanism are DM1 and DM2. Transcription of an expanded CTG/CCTG DNA repeat leads to the formation of toxic CUG/CCUG repeat containing RNAs (referred to as CUG<sup>exp</sup>/CCUG<sup>exp</sup>) that become detained within the nucleoplasm in RNA foci. These CUG<sup>exp</sup>/CCUG<sup>exp</sup> RNAs sequester the muscleblind-like (MBNL) family of proteins, key developmental regulators of RNA splicing that are important for driving fetal to adult mRNA isoform transitions [14–18]. Detainment of MBNL proteins by these toxic RNAs leads to global dysregulation of RNA metabolism, and the missplicing that ensues has been linked to a variety of symptoms in patients. In recent years, the model of RNAs forming intracellular inclusions that sequester RBPs has been shown to more broadly apply to additional microsatellite repeat expansion disorders.

In this review we will describe what is currently known about repeat-associated RNA structures, most notably the CUG<sup>exp</sup>/CCUG<sup>exp</sup> RNAs of DM1 and DM2, respectively. In addition, we will also describe how these RNAs impact global AS outcomes as well as survey what is known about toxic, structured RNAs and aberrant splicing in other repeat-associated disorders. Finally, we will discuss how modifiers of RNA structure can alter AS patterns of RBPs impacted in these diseases and serve as potential therapeutic agents to ameliorate disease symptoms.

2. Myotonic dystrophy (DM) – a model for repeat-associated RNA expansion diseases

While over 30 microsatellite repeat expansion disorders have been identified to date, the genomic expansion of CTG trinucleotide repeats linked with myotonic dystrophy type 1 (DM1) was one of the first to be discovered in 1992 [19–21]. Since this microsatellite’s initial description and the additional discovery of the expanded CCTG DM2 locus in 2001 [22,23], years of research have focused on understanding the impacts of these CTG/CCTG expansions on cellular systems and how disruption of downstream molecular processes, specifically AS, leads to disease symptoms. As other expanded microsatellites have been discovered and linked to specific disease presentations, the molecular hallmarks of DM have been used as a guide for characterization of the molecular defects in the contexts of these disorders. In this section we will review and discuss this prevailing, repeat-associated RNA gain-of-function model for disease pathogenesis, beginning with the unique structural properties of the CUG<sup>exp</sup>/CCUG<sup>exp</sup> RNAs and ending with their impacts on AS outcomes in DM.

2.1. Genetics and disease presentation of myotonic dystrophy

DM is the second most common form of adult-onset muscular dystrophy following Duchenne muscular dystrophy with an incidence of approximately 1 in 8000 [24]. While the disease presents in two genetically and clinically distinct forms (DM type 1 and type 2, DM1 and DM2), both are multi-systemic, neurodegenerative disorders that impact the musculoskeletal, gastric, cardiac, and central nervous system (CNS) organs and tissues. Phenotypically, patients with DM present with a core triad of progressive muscle weakness, early-onset cataracts, and myotonia, or the delayed relaxation of skeletal muscle following voluntary contraction. Additionally, patients have symptoms of insulin resistance, reduced fertility, cognitive impairment, and cardiac dysfunction including conduction abnormalities and arrhythmia [25,26]. Overall, the average lifespan for patients with DM1 is 55 years, with death often attributed to cardiac arrhythmias, pulmonary failure, and neoplasms [27].

Despite several common symptoms, the clinical presentations of DM1 and DM2 patients are distinct. In general, DM1 patients exhibit progressive distal muscle weakness while DM2 patients display progressive proximal muscle weakness [28,29]. Additionally, the most severe repeat expansions in DM1 are more likely to result in global intellectual impairment, which is not commonly associated with DM2 [30,31]. Most importantly, in approximately 10–20% of the DM1 patient population, symptoms manifest at birth; this specific disease presentation is known as congenital myotonic dystrophy (CDM) [32]. Cases of congenital or juvenile-onset of symptoms are remarkably absent for DM2 patients. Infants with CDM present with different symptoms compared to their adult-onset counterparts, specifically muscle hypotonia, respiratory failure, feeding difficulties, and club-foot deformities [33–35]. Additionally, children with CDM frequently have intellectual disability, slurred speech, and a higher incidence of autism spectrum disorder and attention-deficit-hyperactivity disorder [36,37]. Perhaps most interestingly, the hallmark DM1 symptom of myotonia is absent in CDM patients until adolescence. Due to the substantial disease burden, infants with CDM have a 30% mortality rate in the first year of life when mechanical ventilation for breathing assistance is required for greater than 3 months [38]. Overall, the symptomatic profile of CDM patients is distinct from adult-onset DM1 despite a common genetic cause.

At the molecular level, DM1 is caused by a CTG trinucleotide repeat expansion in the 3’ UTR of the dystrophia myotonia protein kinase gene (DMPK) (Fig. 1A) [19–21]. In healthy individuals, DMPK contains between 5 and 37 CTG repeat units while patients with DM1 possess between 50 and 2000 [39,40]. While there is not a strict repeat length associated with CDM, it is rare to have disease-onset at birth with repeat numbers below 750 [40,41]. Individuals that possess 38 to 50 repeat units are identified as carriers of a “pre-mutation” and often will remain asymptomatic. DM2 is linked to a CCTG repeat expansion within intron 1 of the cellular nucleic acid-binding protein (CNBP) gene [22,23]. Patients with DM2 have between 75 and 11,000 repeats [23]. Both expansions are inherited in an autosomal dominant manner and the number of repeats increases through each successive generation, a phenomenon termed genetic anticipation [42]. In general, disease severity escalates and age-of-onset decreases with increasing repeat size, although these correlations are not as strong for DM2 compared to DM1 [41,43–45]. Interestingly, the onset of symptoms in DM2 patients begins much later in life compared to DM1 despite a generally higher repeat load, indicating that both the nucleotide sequence and its genomic context likely contribute to phenotypic differences between these patient populations.

2.2. CUG/CCUG repeat RNAs form metastable stem loop structures that aggregate in nuclear foci

One of the first molecular markers of myotonic dystrophy to be described post-isolation and identification of the CTG trinucleotide repeat expansion associated with DM1 was the nuclear retention of RNA transcripts from the mutant DMPK allele [46,47]. Termed “foci,” these initial observations of ribonuclear inclusions in DM1 cells and tissues redefined hypotheses regarding the molecular pathogenesis of the expanded repeat to include effects beyond the DMPK genetic locus. As part of understanding the pathogenicity of these RNA aggregates, elucidating the structure of the CUG<sup>exp</sup> RNAs and how these structures may contribute to the formation of these unique RNA inclusions has and continues to be a major question and focus within the field.

Initial studies using chemical probing and other in vitro structural techniques determined that CUG RNA repeats, both inside and outside
of the DMPK 3′ UTR sequence context, form semi-stable hairpins that can fold into a multitude of alternatively aligned stem loop structures (Fig. 1B) [48-52]. Electron microscopy visually confirmed formation of these duplex RNAs in vitro [53,54]. As the length of the tandem CUG repeat increases, the stability of these “slippery” hairpins is enhanced. In fact, in a study using CUG repeat RNAs containing 5, 11, 21, or 49 tandem repeats flanked by 30–35 nucleotides of DMPK 3′ UTR, the CUG5 RNA failed to adopt a secondary structure while CUG11, CUG21, and CUG49 formed increasingly longer and stiffer hairpin structures [48]. The failure for a low number, non-pathogenic repeat to form these structural elements speaks to the potential importance of these RNA structures in impacting cellular processes in disease. Another study confirmed these results using RNAs with CUG5 to CUG197 repeat lengths flanked by 378 nucleotides 5′ to the repeat and 505 nucleotides 3′ to the repeat [55]. Using a combination of SHAPE and DMS probing, it was found that while the CUG5 construct formed no secondary structure, longer repeats formed heterogenous, multi-branched hairpins that are structurally isolated from the remainder of the DMPK transcript [55].

Numerous X-ray crystallography studies have provided more detailed views of the CUG repeat RNA duplexes using short oligonucleotides containing up to six tandem repeats [56-60]. These individual structures all indicate that CUGexp RNAs form double-helical structures similar to A-form RNA helices with Watson-Crick base pairs. Within these helices, standard Watson-Crick C-G and G-C base pairs provide stability to the duplex and are interrupted by non-canonical U≡U pairs (Fig. 1B). These U≡U mismatches appear to be “stretched” such that only a single direct hydrogen bond is formed between the uridine residues. Often referred to as the U≡U wobble, these base-pair mismatches generate bulges within the RNA hairpin that contribute to the unique geometry of the helix and are predicted to facilitate the formation of heterogenous, higher order RNA structures both in vitro and in vivo by allowing for “breathing” of the stem loop (additional detailed structural information is reviewed in [61]). The importance of these higher-order RNA structures and their role in the formation of RNA foci was further highlighted when it was revealed that multivalent, intermolecular RNA base-pairing leads to the aggregation of CUGexp RNAs and gelation into spherical compartments in vitro [62]. In this recent study, the authors proposed that this RNA base-pairing dependent phase-transition drives foci formation in vivo. In fact, treatment of DM1 fibroblasts with doxorubicin, a nucleic acid intercalator, significantly reduced both the number and volume of RNA foci observed, possibly through disrupting these RNA-mediated phase-phase transitions [62].

Like the CUG repeats in DM1, the expanded CCTG repeat associated with DM2 also generates CCUGexp RNA foci from the mutant CNBP
allele [23]. Both chemical probing and X-ray crystallography based methodologies using model CCUG RNA repeats revealed that like CU-G^GG^GG^GG RNAs, the CCUG repeats form hairpin structures that adopt A-form RNA helices with G-C and C-G Watson-Crick base pairs separated by tandem C-U/U-C mismatches (Fig. 1B) [51,63,64]. Molecular dynamics simulations indicate that these non-canonical base pairs are dynamic, contributing to the lower thermodynamic stability of the RNA duplex compared to the CUG^GG^GG RNA [51,63,64]. Despite these insights into the structure of CUG^GG^GG RNA in vitro, no studies to date have been performed to define if this structure persists in the context of the endogenous CNBP sequence. Validation of this structural element in a native context is of particular importance for this microsatellite expansion because of the recent observation that intron 1 is abnormally retained within the mutant CNBP pre-mRNA transcript in DM2 cells and patient samples, presumably through disruption of splice site recognition by secondary structures formed by the CUG^GG^GG RNA [65].

While many studies have been performed to define the structures of the CUG^GG^GG/CCUG^GG^GG RNAs in vitro, little work has been performed to elucidate the structures of these expanded RNA repeats in vivo. Characterizing the nature of these RNA structures within the cellular environment will be critical for both understanding how these mutant RNAs perturb cellular processes and how to therapeutically target these pathogenic molecules. With the continued development of increasingly sensitive, specific, and high-throughput technologies for determination and visualization of RNA structure in vivo, such as SHAPE-MaP and LASER-seq, elucidation of these RNA structures in vivo may soon be possible [66–69].

2.3. Muscleblind-like (MBNL) proteins are sequestered by the expanded CUG/CCUG RNA repeats

Support for an RNA gain-of-function mechanism of disease pathogenesis in the DM field accelerated with the observation that human muscleblind-like (MBNL) proteins co-localize with CUG^GG^GG/CCUG^GG^GG RNA foci in DM cells (modeled in Fig. 1A) [70–72]. Initially discovered as a key developmental regulator of muscle differentiation in Drosophila [73,74], the family of MBNL proteins are a highly conserved group of RBPs found in nearly all metazoans that regulate RNA metabolism, most notably AS, during tissue-specific development [75]. Expressed across a variety of tissue types, the three human MBNL paralogs (MBNL1, MBNL2, and MBNL3) are specifically implicated in regulating fetal to adult mRNA isoform transitions in heart, skeletal muscle, and brain [14–18]. Beyond AS, MBNL proteins have been linked to the regulation of other RNA metabolic processes including RNA localization [76], turnover [77], gene expression [78,79], alternative polyadenylation [80], and micro-RNA processing [81].

Several studies using a wide variety of methods have found that YGCV (Y = C or U) is the consensus MBNL binding motif within its RNA targets. Both SELEX (systematic evolution of ligands by exponential enrichment) and computational analysis of motif enrichment in events derived from splicing sensitive microarrays identified YGCV, and in particular UGGC, as the most common MBNL1 binding motif, especially when repeated several times over a short sequence region [79,82]. Ultraviolet cross-linking and immunoprecipitation combined with deep sequencing (CLIP-seq) of MBNL/RNA complexes confirmed that YGCV tetra-nucleotide sequences were indeed a major binding motif for all three MBNL paralogs in vivo [18,76,77,83]. The CUG^GG^GG/CCUG^GG^GG RNAs expressed in DM1/DM2 contain many repeating YGCV motifs, and MBNL1 has been shown to bind these repeats in vitro [84]. This compact and concentrated grouping of ideal MBNL binding elements attract and sequester MBNL proteins within nuclear foci, reducing their functional concentration within the cell. In accordance with this model, MBNL binding increases proportionally with the CUG repeat expansion size as the number of binding sites multiplies (Fig. 2A) [71].

The binding of MBNL proteins to their target RNA sequence elements is mediated by four highly-conserved CCCH-type zinc finger (ZF) motifs (ZF1–ZF4) that in all three MBNL paralogs form two tandem RNA binding domains (ZF1-2 and ZF3-4) that each bind YGCV RNA motifs. NMR studies have shown that each ZF pair interacts with single-stranded RNA molecules in a 1:1 stoichiometric ratio [85]. An NMR derived structure of MBNL1 ZF1-2 in complex with a 15-nucleotide sequence of the human cardiac troponin T (TNNT2) pre-mRNA containing a single CCGU binding motif revealed that the single-stranded RNA molecule wraps around the ZF-2 domain to form the binding interface [85]. This mode of RNA binding is supported by an X-ray crystal structure of the MBNL1 ZF3-4 domain in complex with a short 5′-CCGU-GU-3′ RNA molecule. Analysis of this structure indicates that the MBNL1-RNA interaction is based on the stacking of aromatic and arginine residues with the RNA guanine and cytosine bases [86]. Overall, these independent structures support a model in which the domains of MBNL interact with the Watson-Crick face of the GC dinucleotide within a target YGCV RNA motif, a mode of RNA binding whereby the binding of MBNL destabilizes RNA secondary structure [87]. In fact, stabilization of CUG/CCUG RNA duplexes significantly decreases MBNL binding affinity in vitro, further illustrating that MBNL proteins interface with the CUG^GG^GG/CCUG^GG^GG RNAs by disrupting the RNA duplex to expose a single-stranded RNA binding-surface (Fig. 1A) [87,88].

Importantly, while intermolecular RNA base-pairing is suggested to contribute to nuclear foci formation and stability, the binding of MBNL to the CUG^GG^GG/CCUG^GG^GG RNAs has also been implicated. RNA silencing of MBNL expression significantly reduces foci number in DM1 cells, indicating that the interaction of MBNL with these toxic RNAs aids in stabilizing these aberrant nuclear inclusions [89–91]. Interestingly, microscopy-based studies with CUG^GG^GG RNAs demonstrated that these RNA-protein aggregates are dynamic structures in which MBNL proteins rapidly assemble and disassemble on the RNA molecules, presumably by disrupting local RNA duplex formation [91,92]. These experiments reveal that MBNL proteins regularly diffuse between the nucleoplasm and the CUG^GG^GG foci, demonstrating that the foci are not static aggregates, but are stochastic in nature.

2.4. MBNL sequestration by the expanded CUG/CCUG RNAs leads to aberrant splicing in DM

During cell development and differentiation, groups of RBPs work together to regulate complex networks of AS events to facilitate control of complicated cellular process [5]. Global analysis of AS during development has shown that splicing transitions occur for groups of genes simultaneously and specific RBPs contribute to this splicing coordination [93,94]. In the case of DM, years of research have revealed that depletion of MBNL from the nucleoplasm via sequestration by the "toxic RNAs" leads to multi-systemic dysregulation of AS. While other molecular perturbations including altered gene expression, upregulation of CELF1, and RAN translation (reviewed in [95]) are observed in DM, mis-splicing of MBNL targets remains the prevailing model of disease pathology [95,96].

In brief, MBNL proteins act as positional-dependent regulators of AS in a transcript-dependent manner whereby they act as enhancers or repressors of exon inclusion. In general, if MBNL binds to YGCV regulatory elements upstream or within a regulated exon it represses inclusion while if it binds downstream of a regulated exon it enhances inclusion [76,79,82]. In general, MBNL mRNA and protein levels rise during tissue differentiation. This is especially dramatic in heart, skeletal muscle, and brain where elevated levels of MBNL1 and MBNL2 have been shown to play a major role in promoting transitions from fetal to adult AS patterns [97]. Analysis of key splicing events in MBNL1 knockout (KO) mice revealed that MBNL proteins control the coordinated transitions of AS for a large subset of events that are regulated in a temporal manner during development within these tissues [15,98].

In the context of DM, functional concentrations of MBNL proteins
are reduced due to the sequestration of all three MBNL paralogs within CUG<sup>exp</sup>/CCUG<sup>exp</sup> foci. This disruption of MBNL levels within DM cells leads to a variety of mis-splicing events that have been directly linked to patient symptoms. Direct comparison of transcriptome changes in CUG-repeat expressing and Mbnl1 knockout mouse models revealed that loss of MBNL1 accounts for greater than 80% of the splicing pathology observed in DM1 [79]. Within many affected tissues the fetal isoforms of MBNL1 regulated transcripts have been observed. For example, cardiac troponin T type 2 (TNNT2) and cardiac sodium voltage-gated channel alpha subunit 5 (SCN5A) are mis-spliced in the heart [17,99]. These genes contain MBNL1 regulated exons that are included in fetal tissue that lacks MBNL protein expression and excluded in adult tissue due to the high expression of MBNL [17,84,99,100]. Inclusion of these fetal exons in DM1 patients due to MBNL depletion has been linked with cardiac symptoms including conduction defects and arrhythmia [17,99].

While there are many other examples of reversion to fetal splicing isoforms as a consequence of MBNL sequestration resulting in disease symptoms, one of the best characterized is that of the insulin receptor (INSR) transcript. Alternative splicing of a 36 nucleotide exon 11 (e11) within the INSR pre-mRNA leads to the production of the INSR-A (−e11) and INSR-B (+e11) receptor products [101]. While both protein isoforms are responsive to insulin signaling, INSR-B has enhanced signaling capacity and is predominately expressed in tissues responsible for glucose metabolism, including skeletal muscle [102,103]. MBNL proteins bind an evolutionarily conserved intronic enhancer element downstream of e11 to drive exon inclusion and drive production of the INSR-B isoform [104]. As a repercussion of MBNL capture within the...
The hallmark symptom of DM, myotonia, has also been directly linked with mis-splicing of the chloride voltage gated channel 1 (CLCN1) transcript. Loss of MBNL leads to inclusion of exon 7a which possesses a pre-mature stop codon and causes degradation of the CLCN1 message via nonsense mediated decay (Fig. 2B) [110,111]. This reduction in functional CLCN1 protein levels results in reduced conductance of chloride ions in the sarcolemma of muscle tissue and subsequent myotonia [110]. The connection of this mis-splicing event to the loss of MBNL was uncovered in Mbnl1 knockout mice which possess both Clcn1 transcript mis-splicing and the myotonia phenotype independent of CUG<sup>exp</sup> RNA expression [112]. Importantly, MBNL1 overexpression in a DM1 (CUG-repeat expressing) mouse model restores Clcn1 splicing to the adult pattern and rescues myotonia [113]. Characterization of these and other events indicates that the sequestration of MBNL by the CUG<sup>exp</sup>/CCUG<sup>exp</sup> RNAs reverts a vast collection of MBNL pre-mRNA targets to the embryonic splice isoform, producing a disease signature, immature transcriptome.

While extensive evidence has linked MBNL-dependent splicing regulation of specific events with disease phenotypes, RNA sequencing of DM patient tissue samples, most notably DM1 skeletal muscle, has expanded the field's understanding of both the complexity and degree of AS dysregulation within the patient population. In accordance with the general pattern that increasing repeat expansion size leads to an earlier age-of-onset and exacerbated phenotypic presentation, patients with higher CUG repeat loads, in general, present with more extensive splicing dysregulation compared to individuals with repeat numbers in the pre-mutation range [114]. This observation correlates with the RNA gain-of-function model in which increasing CUG/CCUG repeat expression titrates free MBNL proteins from the nucleoplasm, proportionally disrupting adult AS patterns (Fig. 2A). Wagner and colleagues verified this pathogenic disease mechanism via use of a tunable Mbnl1 expressing HEK-293 cell line in which Mbnl1 concentrations can be precisely controlled [114]. Splicing of MBNL-dependent exons in this system displayed dose-dependent responses to changes in MBNL1 concentration with individual events requiring varying amounts of protein to achieve half-maximal exon inclusion, presumably as a consequence of differential cis-motif enrichment and/or expression of other trans-acting factors (Fig. 2C) [114]. Analysis of an additional 120 RNAseq transcriptomes from DM1 skeletal muscle and heart tissue confirmed that AS events often display differential patterns of dysregulation that correlate with their sensitivity to MBNL levels [115]. Interestingly, within this study Wang and colleagues also deduced that different tissues often possess varied splicing severity of the same event, even within the same individual patient, highlighting the complexity of this multi-systemic disease [115].

While next-generation sequencing based techniques have provided new insights into transcriptome-wide alterations in adult-onset DM1 patients, to date little work has been performed to characterize global splicing dysregulation in CDM. RNA-seq analysis of three biceps brachii samples from three young CDM patients confirmed that CDM is a spliceopathy [116]. Comparison of the transcriptomic changes in these samples to those from adult-onset DM1 patients revealed that many of the same AS events are dysregulated in both patient populations, albeit more severely in CDM [116]. Additionally, transcriptomic alterations in a muscle-specific Mbnl1/Mbn2/Mbn3 knockout mouse recapitulate many of the same AS changes identified in CDM patient samples, verifying that disruption of MBNL-dependent AS contributes in part to CDM spliceopathy [116].

Finally, while disruption of AS has been observed in DM2 cells and tissues, large-scale, global transcriptomic analysis of AS perturbations has only been performed for a limited number of DM2 patient samples [115]. Targeted analyses of specific MBNL-dependent events indicate that, in general, while many of the same AS events impacted in DM1 are also affected in DM2 [16,106], mis-splicing is often less severe despite the higher binding affinity and enhanced recruitment of MBNL proteins to the CCUG<sup>exp</sup> RNAs [16,84]. These observations suggest that genetic modifiers, potentially of CCUG<sup>exp</sup> RNA structure, may alter patterns of AS in this patient population. A recent report proposed that RBFOX proteins, another family of RBPs that regulate AS in a developmental manner, bind to and interact with the CCUG<sup>exp</sup> RNAs, limiting MBNL sequestration by this toxic RNA species and alleviating subsequent MBNL-dependent mis-splicing [117]. Further studies are needed to fully characterize global perturbations of AS outcomes in DM2 tissues and their relationship to both CCG repeat load and disease severity.

### 3. Structured repeat RNAs and aberrant splicing in other repeat-expansion disorders

The RNA gain-of-function model of molecular pathogenesis proposed and experimentally validated in the context of DM has served as the foundation for hypotheses of disease pathology for several other genetic disorders linked to microsatellite repeat expansions. However, in many cases the contributions of RNA structural elements formed following expression of the repeat and its relationship to dysregulation of AS splicing are not as clearly defined. Within this section we will review and summarize what is currently known about how RNA structure and AS contribute to disease pathogenesis in other microsatellite repeat expansion disorders with variant disease presentations (summarized in Fig. 3). We will also identify and discuss unexplored or unanswered questions in relationship to these topics.

#### 3.1. Fuch’s endothelial corneal dystrophy (FECD)

Fuch’s endothelial corneal dystrophy (FECD) is an age-related, degenerative disorder of the endothelium that specifically impacts the cornea. Progressive loss of corneal endothelial cells in affected individuals ultimately results in corneal edema, scarring, and vision loss [118]. FECD impacts approximately one in 20 Americans over the age of 40 years and is currently the leading cause of corneal transplants worldwide [119,120]. While more than 15 genes harboring mutations or SNPs have been associated with this disorder, an expanded CUG repeat tract within intron 3 of TCF4 is the most commonly identified genetic cause of FECD [121]. In general, repeat lengths over 50 confer a significant risk for development of disease symptoms, but as in the DM, the range of observed repeat size varies; cases of over 1000 repeats have been reported [122]. The CUG repeat is inherited in an autosomal dominant pattern and inter-generational instability has also been identified in cases of parent-to-child transmission [123]. Increasing lengths of this trinucleotide repeat are generally correlated with a more severe phenotype and a higher risk of requiring corneal transplant at a younger age [124,125].

While the CUG repeats linked independently with FECD and DM1 are located on separate genes and in distinct regions of genomic architecture (intron vs 3’ UTR, respectively), the similar patterns of genetic inheritance, observations of repeat instability, and comparable impacts to corneal tissue indicated that a similar RNA gain-of-function model for disease pathology could explain the phenotypic presentation in FECD patients. Using the molecular hallmarks of DM1 as a guide for their investigative studies, several groups have recently confirmed that the expanded CUG-repeat containing RNAs transcribed from the TCF4 locus form ribonuclear foci in both patient-derived cell lines and corneal tissue [126,127]. Additional reports confirmed that like in DM1, MBNL proteins co-localize with these toxic RNAs in the nucleus. MBNL sequestration within these nuclear aggregates leads to mis-splicing of many MBNL-regulated events that overlap significantly with those dysregulated in DM1, including Mbnl1, Claspl1, Numal1, Add3, and...
VEGFA [126,128]. Interestingly, a significant proportion of the genes found to be mis-spliced in FECD patient corneal endothelium are associated with cytoskeletal protein binding and cell adhesion, potentially contributing to degradation of this tissue [128]. However, further experimental validation is needed to elucidate causal links between these perturbations in AS and disease symptoms. Overall, these studies define a conserved RNA gain-of-function molecular pathology between FECD and DM1. In accordance with this statement, a recent report found that DM1 patients are susceptible to FECD symptoms despite the lack of an expanded CTG repeat-containing TCF4 allele, indicating that the expression of CUGexp RNAs from the DMPK locus can also drive the observed corneal pathology [129].

Due to the similarities in the molecular presentation of both FECD and DM1, the CUGexp RNAs in FECD are predicted to form similar heterogenous, “slippery” hairpin structures. However, the TCF4-derived CUGexp RNAs could form distinctive secondary structures compared to the same CUGexp – containing RNAs expressed in DM1 tissues due to both variant sequence context and genomic position within their respective pre-mRNA transcripts. These variations may potentially impact foci formation, MBNL binding affinity and association dynamics within nuclear foci, and the severity of mis-splicing. As an example of the differential impacts these repeats can elicit based on their sequence context, the CUGexp in TCF4 intron 3 increases this intron’s retention within the population of processed mRNA transcripts; this same phenomenon is not observed in DM1 due to this microsatellite’s position within the 3’ UTR region [65]. While the CUG repeat itself and RNA gain-of-function mechanism are comparable between these two disorders, potential differences in RNA structure may influence molecular markers of disease.

3.2. Spinocerebellar ataxias

Spinocerebellar ataxias (SCAs) encompass a large group of autosomal dominant disorders characterized by cerebellar atrophy leading to progressive loss of muscle coordination and balance resulting in altered gait, abnormal speech, involuntary eye movements, and poor hand-eye coordination [130]. To date, more than 40 SCAs have been identified, 13 of which are due to expansions of various microsatellite repeats in both coding and non-coding regions of independent genomic loci (reviewed in [130,131]). Many of this group of SCAs are proposed to be caused by an RNA gain-of-function molecular pathology, several of which are highlighted here.

SCA10 is a prevalent ataxia in Latin American populations linked to a pentanucleotide ATTCT expansion within intron 9 of the ATXN10 gene [132,133]. While unaffected individuals possess between 10 and 29 repeat units, the pathogenic allele associated with the onset of disease symptoms harbors between 800 and 4500 ATTCT repeats [133]. The AUUCUexp RNA formed from the mutant ATXN10 allele forms both nuclear and cytoplasmic RNA aggregates in SCA10 patient-derived fibroblasts and transgenic mouse brains ectopically expressing approximately 500 ATTCT repeats [134]. The AUUCUexp RNA was found to co-localize with hnRNP K [134], an RNA binding protein shown to have various roles in RNA metabolism, including RNA splicing (reviewed in [135,136]). More specifically, hnRNP K preferentially binds
pyrimidine-rich RNA sequences in a variety of RNA targets to post-transcriptionally control gene expression [136,137]. Interestingly, hnRNP K is required for effective axonogenesis [138]. As such, dysregulation of functional levels of hnRNP K via sequestration by the pyrimidine-rich AUUCU\textsuperscript{exp} RNA may contribute to the cerebellar degeneration observed in SCA10. While the impacts of hnRNP K depletion on AS and other aspects of RNA metabolism have not been assessed globally, the individual hnRNP K target β-tropomyosin was found to be mis-spliced in SCA10 cells [134]. Crystallographic studies using a tetraloop receptor to facilitate RNA crystallization of two AUUCU RNA repeats revealed that these pentanucleotide RNA units adopt an A-form helical structure in vitro in which UCU mis-matches form internal loops closed by AU pairs (modeled in Fig. 3) [139]. Computational modeling of these RNA repeats indicate that the presence of these large internal loops allow for the RNA helix to be “unzipped” to form single-stranded conformations that hnRNP K could sample to achieve effective binding [139,140]. This relative lack of stability compared to other repeat RNAs shown to fold into A-form helices, such as CUG\textsuperscript{exp} RNAs of DM1, may contribute to why such large expansions are necessary to achieve the molecular manifestations of disease. More specifically, a larger repeat threshold may be required for RNA foci to assemble and trap a significant concentration of RBPs to induce defects in RNA metabolic processes.

SCA36 is associated with an expansion of a GGGCGTG hexanucleotide repeat within intron 1 of NOP56 [141]. In contrast to many of the other SCAs, SCA36 patients present with lower motor neuropathy resulting in a plethora of symptoms including muscle atrophy [141,142]. Consistent with an RNA gain-of-function disease mechanism, the GGCCUCG\textsuperscript{exp} RNAs produced from this genetic locus form nuclear foci in both patient-derived lymphoblastoid cell lines and brain tissue [141,143]. SRSF2 has been shown to co-localize with GGCCUCG\textsuperscript{exp} RNA in these ribonuclear aggregates [141]. SRSF2 (also named SC35) is a member of the serine and arginine-rich (SR) protein family that play a role in the regulation of both constitutive and alternative RNA splicing (reviewed in [144]). To date, no studies have been performed to determine if SRSF2 sequestration leads to widespread AS dysregulation and its potential connection to disease presentation. However, the reported overlap in SRSF1 and SRSF2 mRNA binding targets may present challenges in elucidating if AS is disrupted by capture of SRSF2 within toxic GGCCUCG\textsuperscript{exp} RNA foci [145]. More importantly, no experimental studies have defined specifically how SRSF2 may interact with this hexanucleotide repeat, especially as molecular dynamics simulations predict that these RNA species may fold into G-quadruplexes (modeled in Fig. 3) while structural studies of the SRSF2 RNA-recognition motif indicate that this RBP binds single-stranded RNA [146,147]. Further research is required to address many of these questions regarding the relationship between the GGCCUCG\textsuperscript{exp} RNA structures formed in vitro, sequestration of SRSF2, and AS.

The expansion of a CTG trinucleotide repeat in the 3' UTR of the ATXN8 gene is the genetic element responsible for SCAS [148]. The toxic CUG\textsuperscript{exp} RNA transcripts produced from the pathogenic ATXN8 allele form ribonuclear foci that co-localize with MBNL proteins in both transgenic mice and patient cerebellar tissue [149]. Additionally, these CUG\textsuperscript{exp} transcripts trigger splicing changes of the MBNL1 regulated GABA-A transporter 4 (GAT4) in SCAS human autopsy tissue [149]. Persistence of the GAT4 fetal isoform is predicted to lead to loss of GABAergic inhibition in the granular layer of the cerebellum. This disease-associated reversion to a fetal splicing pattern is consistent with the global pattern of mis-splicing in DM1 and likely indicates other MBNL target RNA transcripts are also mis-spliced in SCAS, although transcriptome-wide changes in AS have yet to be characterized and reported for this specific repeat expansion disorder. Despite sharing the same nucleotide content and 3' UTR location of the CUG repeat with DM1, albeit in a different gene, there are significant differences in the molecular hallmarks of disease presentation. For example, the number of RNA foci observed in impacted tissues varies greatly between these two diseases. In general, many RNA foci are observed in DM1 tissues and the number of nuclear foci increases with longer CUG expansions (reviewed in [150]). Interestingly, this correlation between repeat size and foci number is not found in SCA8. For example, the presence of a single RNA foci was detected in the molecular layer interneurons and Bergmann glia of two distinct SCA8 patient samples possessing extremely disparate CTG repeat expansions (400 or 1000 CTG repeats) [149]. Moving forward in it will be interesting to determine if the sequence context in which these repeat expansions reside and its potential impacts on RNA secondary structures formed may contribute to the different levels of RNA foci observed in SCA8 compared to DM1.

3.3. C9ORF72 amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD)

Amyotrophic lateral sclerosis (ALS) is a devastating, progressive, adult-onset neurogenerative disorder characterized by degeneration of both the upper and lower motor neurons that ultimately results in muscle paralysis and death [151]. Frontotemporal dementia (FTD) is a complex brain disorder defined by degeneration of the frontal and temporal brain regions resulting in disruption of speech, cognitive function, and behavioral skills [152]. Although clinically distinct disorders, patients often present with symptoms of both diseases [153]. The most common mutation associated with both familial ALS and FTD is the expansion of a GGGGCC hexanucleotide repeat (G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp}) within the first intron of the chromosome 9 open reading frame 72 (C9orf72) gene, termed C9-ALS/FTD [154,155]. Like many of the other microsatellite repeat expansion disorders described in this review, the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} displays an autosomal dominant pattern of inheritance [156]. The size of the repeat in C9-ALS/FTD cases ranges from several hundred to several thousand as compared to 2-23 in controls [155,157,158].

While the molecular mechanisms that define disease pathogenesis in C9-ALS/FTD are still under investigation (reviewed in [159]), RNA toxicity caused by the transcribed G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} is thought to play an important role in promoting neuronal degeneration. This repeat is bi-directionally transcribed producing both G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} and G\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs that aggregate into sense and antisense RNA foci, respectively, in C9-ALS/FTD patient brain and spinal cord [155,160-162]. Interestingly, these foci are generally found within separate neuronal cells, although instances of both co-occurrence and co-localization within single nuclei have been observed [161,162].

Several groups have characterized the structure of the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp}/ G\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs in an effort to understand the RNA toxicity of these hexanucleotide repeats. Initial NMR and CD spectroscopy studies utilizing a minimal G\textsubscript{4}C\textsubscript{2} RNA oligomer showed that these RNA repeats form G-quadruplexes in vitro consisting of four parallel stacks of G-tetrad planes connected by single-stranded loops consisting of two cytosines each (modeled in Fig. 3) [163]. Additional reports confirmed that model G\textsubscript{4}C\textsubscript{2} RNAs form both uni- and multimolecular G-quadruplexes in vitro in a repeat length and RNA concentration-dependent manner [164,165]. Using an antibody that recognizes G-quadruplexes (BG-4), aggregates containing these structures were found in C9-ALS/FTD patient-derived fibroblasts and astrocytes [166]. While an increase of these G-quadruplex-containing inclusions was observed in C9-ALS/FTD cells compared to controls, it is important to note that these aggregates were not confirmed to be the result of G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} RNAs, especially given that the BG-4 antibody recognizes G-quadruplexes independent of sequence content. Interestingly, the use of genome-wide RNA structural probing techniques indicates that G-quadruplexes are globally unfolded within eukaryotic cells [167]. These disparate results suggest that 1) the cellular machinery that unfolds RNA regions capable of forming G-quadruplexes are mis-regulated in C9-ALS/FTD or 2) the sheer volume of G-quadruplexes derived from G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} RNAs overwhelms this machinery. Interestingly, the sense G\textsubscript{4}C\textsubscript{2} RNA repeats, but not antisense G\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs, promote phase transitions and RNA gelation in vitro and when ectopically expressed in cells [62,168]. Inhibition of general RNA
base-pairing interactions or targeted disruption of the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} G-quadruplex structure limit these RNA-mediated phase transitions [62,168]. These results indicate that RNA foci formation in C9-ALS/FTD depends upon the specific secondary structures adopted by these G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} RNAs.

In contrast to G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp}, the antisense C\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs fold into A-form RNA helices in vitro. The crystal structure of a (CCCCGGG)\textsubscript{CC}CCC RNA oligomer that models the expanded repeat in C9ORF72 antisense RNA transcripts found that two RNAs oligomers packed into an A-form helical structure consisting of four Watson-Crick G-C base pairs separated by two tandem C-C mismatches (modeled in Fig. 3) [169]. Additionally, CD spectra of various C\textsubscript{4}G\textsubscript{2} RNAs containing increasing repeat lengths found that as the number of repeats increased, the propensity to form intramolecular A-form RNA helices was also enhanced [169]. Overall, while investigations to define the secondary structures of the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} and C\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs have been completed in vitro, the secondary structures these RNAs adopt in vivo have not been explored. Additionally, no experimental studies have been performed to determine if the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} and C\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs interact with each other to form differential RNA structures. This line of investigation is especially pertinent given the observation of co-localization of sense and antisense RNA foci [161,162].

Several RBPs have been proposed to bind to the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} RNAs, although identification of binding partners of the antisense C\textsubscript{4}G\textsubscript{2}\textsuperscript{exp} RNAs has yet to be addressed. Through immunofluorescence-based co-localization with RNA foci and RNA-pulldown combined with mass spectrometry a wide range of RBPs have been shown to associate with the G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} RNAs [165,170,171]. These RBPs include SRSF1 [164,170,172], SRSF2 [170,171], hnRNP H [166,170,171], hnRNP A1 [171,172], hnRNP A3 [173], and nucleolin [165], among others [174,175]. While nearly all proteins listed here have roles in the regulation of both constitutive and alternative splicing, for the purposes of this review we will discuss the RBP for which sequestration leads to documented perturbations in AS outcomes, namely hnRNP H. HnRNP H is a member of the heterogenous nuclear ribonucleoproteins (hnRNPs) family, a large group of RBPs that have multiple roles in RNA metabolism, including AS (reviewed in [135]). HnRNP H functions as both an enhancer and repressor of AS by binding to G-rich RNA motifs [176]. In general, exon inclusion is enhanced when intronic regions either upstream or downstream of the target exon are bound while exon exclusion occurs when hnRNP H binds within the regulated exon [177]. AS regulation by this RBP has been specifically implicated for a vast collection of target transcripts critical for a variety of cellular processes including neuronal development and differentiation [178,179].

HnRNP H was shown in several RNA pull-downs to bind G\textsubscript{4}C\textsubscript{2} hexanucleotide-repeat-containing RNAs and co-localize with RNA foci in C9-ALS/FTD cell models and patient brain tissue [170,171]. Additional RNA-crosslinking and immunoprecipitation studies found that hnRNP H represents the major RBP bound to these toxic RNA species and co-localizes with G-quadruplexes as identified by the BG-4 antibody in C9-ALS/FTD patient derived cells [166]. Interestingly, hnRNP H has been shown to interact with G-quadruplex motifs within select mRNAs [180], indicating that the particular RNA structures formed by these expanded RNAs in vitro and potentially in vivo provides an ideal platform for binding and sequestration of this RBP, although the specific nature of this interaction remains unknown.

Consistent with an RNA gain-of-function disease mechanism, depletion of hnRNP H from the nucleoplasm results in spliceopathy; comparison of a panel of hnRNP H regulated target exons between C9-ALS/FTD cerebellum and U87 cells in which hnRNP H expression was knocked down via siRNA treatment were found to be similarity mis-spliced [166]. In addition, RNAseq of C9-ALS/FTD postmortem cerebellum revealed widespread changes in AS for which hnRNP H binding motifs were enriched in the sequences surrounding mis-spliced cassette exons [181]. Further characterization of hnRNP H-dependent spliceopathy in a large cohort of C9-ALS/FTD patient tissue identified a spectrum of mis-splicing across several known hnRNP H target RNAs that generally correlated with relative protein concentration as measured by RBP solubility (i.e. the fraction of hnRNP H not trapped within toxic RNA aggregates) [182]. RNAseq validated the correlation between hnRNP H solubility and transcriptome-wide RNA mis-splicing across multiple C9-ALS/FTD samples [182], a pattern comparable to the relationship between free MBNL proteins levels and mis-splicing in DM1 [114]. Importantly, hnRNP H regulated AS events shown to be mis-spliced in these studies are related to proper neuronal function. One such example is ACHEx which encodes acetylcholinesterase, an enzyme that hydrolyzes the neurotransmitter acetylcholine to terminate synaptic transmission in neuromuscular junctions, a region particularly sensitive to denervation in ALS [183]. HnRNP H also regulates splicing of many other RBP-encoding transcripts with roles in splicing regulation themselves, including SRSF4, FUS, and TDP-43 [184]. This suggests that the range of mis-splicing observed in C9-ALS/FTD patient tissue may extend beyond mRNAs directly regulated by hnRNP H.

Overall, these results highlight a critical link between sequestration of hnRNP H by G\textsubscript{4}C\textsubscript{2}\textsuperscript{exp} RNAs, reduced functional concentrations of free hnRNP H within the nucleoplasm, and subsequent transcriptome-wide splicing dysregulation. Additional characterization of both how alterations in hnRNP H “dosage” alter patterns of mis-splicing and the impacts of sequestration of other RBPs by both the sense and antisense RNA transcripts will aid in further defining how dysregulation of AS contributes to disease symptoms in the context of C9-ALS/FTD.

4. Modifiers of RNA structure in repeat-expansion disorders alter patterns of mis-splicing

The shared features between all the microsatellite repeat expansion diseases described in this review is an RNA gain-of-function pathogenic mechanism centered upon the binding and sequestration of RBPs, specifically AS factors, by the repeat-containing RNA transcripts predicted to fold into secondary structures that promote and drive foci formation. Insights into how variations in RNA sequence and structural content contribute to AS regulation by these same RBPs in a non-disease context can provide valuable information about how to perturb RNA – repeat RNA interactions. In this final section we will highlight how RNA sequence and structure can modify AS regulation by MBNL proteins, an RBP implicated in DM1, DM2, FEDC, and SCA8. We will also briefly describe how observations from basic biological studies are being leveraged to design therapeutics that disrupt repeat-RNA structures as a means to reverse disease symptoms.

4.1. Alternative splicing regulation by MBNL is fine-tuned by RNA secondary structure

Several independent studies have demonstrated that MBNL proteins bind to YGGY regulatory elements within a vast collection of target pre-mRNAs to drive AS outcomes. However, the contributions of RNA motif number and their structural organization within a MBNL – regulated transcript have remained elusive. A collection of recent studies have aimed to more clearly define how variations in RNA structure can modulate AS by this RBP. Using large-scale combinatorial mutagenesis of RNA sequences surrounding a single UGCU motif in conjunction with high throughput RNA sequencing analysis, it was discovered that MBNL binding was enhanced not only by the creation of additional YGGY motifs but by integration of nucleotides predicted to perturb RNA structure [185]. These observations are consistent with the reported reduction of MBNL1 binding to CUG-repeat RNA duplexes stabilized via pseudouridine substitution of uridines [87]. Additionally, the mutations identified using this approach also enhanced MBNL splicing activity when placed into a minigene system [185], indicating that the sequence context surrounding a defined MBNL regulatory element can influence splicing outcomes by modulating the structural neighborhood and limiting the availability of a target RNA motif in vivo.
Using a hybrid MBNL-regulated minigene system, another study confirmed that enhancement of RNA structure reduces MBNL1 AS activity [186]. Specifically, the authors found that increasing the number of UGCU binding elements generally correlated with increased regulatory activity and binding affinity when the motifs were predicted to be unstructured or arranged on a singular side of an RNA hairpin [186]. On the other hand, increased distance between UGCU motifs in a single-stranded context, re-distribution to the opposite face of a RNA helix, or structural stabilization of flanking RNA sequence space all reduced MBNL1’s splicing activity independent of the motif number and, in some contexts, MBNL binding affinity [186]. Overall, these studies are consistent with a model of MBNL splicing regulation whereby local modulation of RNA secondary structure can have profound impacts on splicing regulation and represents yet another complex and sensitive regulatory component that can augment MBNL-mediated AS. These splicing regulation and represents yet another complex and sensitive modulation of RNA secondary structure can have profound impacts on either the DM1 CUGexp or C9-ALS/FTD G4C2 repeat expansion disorder, disruption of these pathogenic RNA structures is a prevailing strategy for therapeutic design and development. Both small molecules and antisense oligonucleotides (ASOs) have been characterized as potential therapeutic molecules via their action to bind and interfere with the formation of repeat RNA foci and prevent the association of RBPs with these toxic RNAs (reviewed in [187–190]).

ASOs are short, synthetic, single-stranded oligonucleotides containing modifications for stability that are designed to bind target RNA sequences via complementary base-pairing, providing a high level of specificity for the pathogenic repeat RNAs. While some ASOs are engineered to target repeat RNAs for degradation as a means to reduce the toxic repeat RNA load [191], others are designed to disrupt repeat RNA aggregation and prevent RBPs binding. In fact, ASOs designed to bind the expanded CUG repeats in F5DC and DM1 have been shown to reduce the total amount of RNA foci leading to the cellular re-distribution of MBNL and correction of mis-splicing in both patient-derived cells and mouse tissue [192–194]. ASOs targeting the C9-ALS/FTD G_{C>G}^\text{exp} also block foci formation and nuclear accumulation of ADARB2 [195]. In agreement with these observations, ASOs engineered to interact with either the DM1 CUG^exp or C9-ALS/FTD G_{C>G}^\text{exp} RNAs disrupt RNA phase-transitions in vitro and in cells by interfering with intermolecular base-pairing [52].

Small molecules that target toxic repeat RNA structures have also been identified as potential compounds for therapeutic intervention. By taking advantage of the unique intramolecular RNA structures formed by these toxic RNAs as defined by in vitro structural studies, several groups have screened for and rationally designed compounds that bind these RNA structures with high specificity. In the context of DM1, a plethora of small molecules have been characterized that bind CUG^exp RNAs structures (reviewed in [196]). These compounds displace MBNL from the toxic RNAs and rescue mis-splicing [196]. Small molecules reported to bind to the specific RNA structures formed by the DM2 CUG^exp [63,197,198], SCA10 AUUCU^exp [199], and C9-ALS/FTD G_{C>G}^\text{exp} RNAs [200] in vitro have also been characterized; in each case improvements in disease-associated molecular defects were observed. Overall, leveraging the unique structural properties of these toxic RNAs presents a promising pathway for the development of effective therapeutic compounds across a range of repeat expansion diseases [189,190].

5. Concluding remarks and future directions

As the number of genetic diseases associated with expanded microsatellite repeats continues to expand and the molecular mechanisms of each disease are unraveled, it has become clear that common themes of disease pathology persist independent of the repeat’s genomic location or nucleotide content. These conserved themes include (1) the RNAs transcribed from the pathogenic DNA repeat associate via inter- and intramolecular base-pairing to form anomalous RNA aggregates, (2) these expanded RNAs possess the capacity to adopt RNA structures with variant stability depending on the sequence content and context of the repeat as evaluated by in vitro structural studies, (3) the toxic RNAs provide a compact and condensed platform of RBP binding motifs, (4) these toxic RNAs capture and sequester RBPs reducing their functional concentration in vivo, and (5) impacts on RBP(s) “dosage” within the cell have profound effects on RNA metabolism, most notably RNA splicing, that is related to the regulatory function of the specific RBP(s) sequestered by the disease-causative RNAs.

While significant progress has been made to define the secondary structures of these expanded repeat RNAs in vitro, experimental limitations have made determining the structures and dynamics of these RNAs in vivo challenging. These gaps in our understanding of the structural behavior of these repeat RNAs in vivo limits insights into both how RBPs are sequestered within RNA foci and why the association of particular RNAs to RBPs has such deleterious impacts on AS transcriptome-wide. Additionally, even though the binding of MBNL proteins to the CUG^exp/CCUG^exp RNAs in DM1/DMD has been well described, the specific RBPs that are captured by other expanded RNA repeats still need to be refined and the molecular impacts of these observations analyzed; this is especially important in the case of C9-ALS/FTD, where many RBPs have been shown to co-localize with the G_{C>G}^\text{exp} RNAs. Also, further characterization of the relationship between levels of individual RBP sequestration and potential dose-dependent alterations in the regulation of target transcripts may provide valuable insights into the variability and complexity in phenotypic presentation of this group of disorders. Finally, recent studies have revealed that RNA modifications can alter local RNA secondary structure to limit the availability of RNA binding motifs for AS factors like hnRNP C and hnRNP G, resulting in altered splicing outcomes of target mRNAs [201,202]. Determining if nucleotides within the toxic RNAs are modified and how these modifications may influence in vivo secondary structures, foci formation, and RBP association is a new and important course of study in the field to explore how this layer of post-transcriptional gene regulation may contribute to the molecular mechanism of disease. Uncovering the answers to many of these questions will be vital to both increasing our basic knowledge of the shared RNA gain-of-function disease mechanism in repeat expansion disorders and for the generation of effective therapeutics that target this component of disease pathology.

Transparency document

The Transparency document associated with this article can be found, in online version.

Acknowledgements

We appreciate helpful discussions with members of the Berglund and Johnson labs. Research in the Berglund and Johnson labs is supported by grants R01GM121862 from the National Institute of General Medical Sciences (USA) R01NS104010 and P50NS048843 from National Institute of Neurological Disorders and Stroke (USA).

References


...
M. A. Hale, et al.

BBA - Gene Regulatory Mechanisms xxx (xxxx) xxx-xxx


