Intron retention induced by microsatellite expansions as a disease biomarker

Łukasz J. Sznajder[a,1,2], James D. Thomas[a,1,3], Ellie M. Carrell[b], Tammy Reid[a], Karen N. McFarland[c], John D. Cleary[a], Ruan Oliveira[a], Curtis A. Nutter[a], Kirti Bhatt[b], Krzysztof Sobczak[d], Tetsuo Ashizawa[e], Charles A. Thornton[b], Laura P. W. Ranum[a], and Maurice S. Swanson[a,2]

[a]Department of Molecular Genetics and Microbiology, Center for NeuroGenetics and the Genetics Institute, College of Medicine, University of Florida, Gainesville, FL 32610; [b]Department of Neurology, University of Rochester, Rochester, NY 14642; [c]McKnight Brain Institute, Department of Neurology and Center for Translational Research in Neurodegenerative Disease, University of Florida, College of Medicine, Gainesville, FL 32610; [d]Department of Gene Expression, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, 61-614 Poznan, Poland; and [e]Neurological Institute, Houston Methodist Hospital, Houston, TX 77030

Edited by Stephen T. Warren, Emory University School of Medicine, Atlanta, GA, and approved March 12, 2018 (received for review September 20, 2017)

Expansions of simple sequence repeats, or microsatellites, have been linked to ∼30 neurological–neuromuscular diseases. While these expansions occur in coding and noncoding regions, microsatellite sequence and repeat length diversity is more prominent in introns with eight different trinucleotide to hexanucleotide repeats, causing hereditary diseases such as myotonic dystrophy type 2 (DM2), Fuchs endothelial corneal dystrophy (FECD), and *C9orf72* amyotrophic lateral sclerosis and frontotemporal dementia (C9-ALS/FTD). Here, we test the hypothesis that these GC-rich intronic microsatellite expansions selectively trigger host intron retention (IR). Using DM2, FECD, and C9-ALS/FTD as examples, we demonstrate that retention is readily detectable in affected tissues and peripheral blood lymphocytes and conclude that IR screening constitutes a rapid and inexpensive biomarker for intronic repeat expansion disease.

amyotrophic lateral sclerosis | intron retention | microsatellite | myotonic dystrophy | RNA splicing

Repetitive elements are a common sequence feature of eukaryotic genomic DNAs and comprise as much as ∼70% of the human genome (1, 2). These repetitive sequences include transposable element families (DNA transposons and LTR and non-LTR retrotransposons) and simple sequence repeats, such as telomeric repeats and a variety of satellites (centromeric, micro, mini, and mega). Microsatellites, which are repeating units of ≤10 base pairs (bp), are a particularly prominent repetitive element class because they are highly polymorphic due to their tendency to form imperfect hairpins, slipped-stranded, quadruplex-like, and other structures resulting in elevated levels of DNA replication and repair errors (3, 4). While these errors result in both repeat contractions and expansions that may provide beneficial gene regulatory activities, expansions cause ∼30 human hereditary diseases (5, 6). Although human introns are significantly longer and denser in repetitive elements compared with exons (7), only eight microsatellite expansion disorders have been linked to intron repeat instability.

In this study, we examined the pathomolecular consequences of both GC- and A/AT-rich intronic microsatellite mutations associated with myotonic dystrophy type 2 (DM2), *C9orf72*-linked amyotrophic lateral sclerosis with frontotemporal dementia (C9-ALS/FTD), Fuchs endothelial corneal dystrophy (FECD), Friedreich's ataxia (FRDA), and spinocerebellar ataxia type 10 (SCA10). We demonstrate the GC-rich CCTG, GGGGCC, and CTG expansions lead to host intron retention (IR) in DM2, C9-ALS/FTD, and FECD, respectively, while A/AT-rich expansions in FRDA and SCA10 do not. Based on these and additional observations, we propose IR as an accessible and inexpensive biomarker for both diagnostic and therapeutic trial purposes.

Results

Sequence Diversity and Positional Bias of Intronic Microsatellite Expansions. The human genome contains ∼80,000 3- to 6-bp microsatellites in introns that could potentially undergo expansion, but only 8 tandem repeats have been documented to expand in hereditary disease (Fig. S1 and Dataset S1). While GC-rich trinucleotide expansions (exp) predominate in exonic regions, intron mutations are composed of 3- to 6-bp repeats that vary considerably in GC content (20–100%) (6, 8). Based on this sequence feature, we divided intronic expansions into GC- and A/AT-rich groups (Fig. 1*A*). In contrast to the majority of A/AU-rich microsatellite RNAs, GC-rich expansions are predicted to form highly stable RNA secondary structures (Fig. S2) (9), increase intron length substantially (Fig. 1*B*), and even multiply intron length several times, such as the SCA36-associated $GGCCTG^{exp}$ mutation in *NOP56* (Fig. S3*A*). SCA10 AUUCU repeats also fold into secondary structures consisting of UCU internal loops closed by AU pairs, but these structures are relatively unstable compared with the hairpins and G-quadruplexes formed by comparable-length GC-rich repeats (10, 11).

Significance

A number of hereditary neurological and neuromuscular diseases are caused by the abnormal expansion of short tandem repeats, or microsatellites, resulting in the expression of repeat expansion RNAs and proteins with pathological properties. Although these microsatellite expansions may occur in either the coding or noncoding regions of the genome, trinucleotide CNG repeats predominate in exonic coding and untranslated regions while intron mutations vary from trinucleotide to hexanucleotide GC-rich, and A/AT-rich, repeats. Here, we use transcriptome analysis combined with complementary experimental approaches to demonstrate that GC-rich intronic expansions are selectively associated with host intron retention. Since these intron retention events are detectable in both affected tissues and peripheral blood, they provide a sensitive and disease-specific diagnostic biomarker.

Author contributions: Ł.J.S. and M.S.S. designed research; Ł.J.S., J.D.T., E.M.C., T.R., K.N.M., J.D.C., R.O., and C.A.N. performed research; E.M.C., T.R., K.N.M., J.D.C., K.B., T.A., C.A.T., and L.P.W.R. contributed new reagents/analytic tools; Ł.J.S., J.D.T., E.M.C., K.S., and M.S.S. analyzed data; Ł.J.S. performed graphics; and Ł.J.S., J.D.T., and M.S.S. wrote the paper.

Conflict of interest statement: M.S.S. is a member of the scientific advisory board of Locana, Inc.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE101824).

[1]Ł.J.S. and J.D.T. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ljsznajder@ufl.edu or mswanson@ufl.edu.

[3]Present address: Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1716617115/-/DCSupplemental.

Published online April 2, 2018.

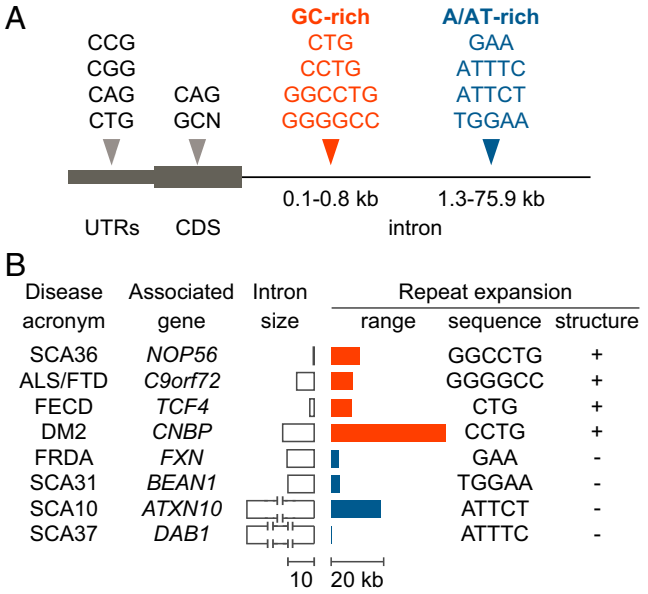

Fig. 1. GC- and A/AT-rich intronic microsatellite expansion mutations. (*A*) Sequences of disease-associated microsatellites located in exons, including untranslated regions (UTRs) and coding sequences (CDS), and introns. Also indicated are the distances to the nearest splice site for intronic GC- and A/AT-rich microsatellites. (*B*) Intronic microsatellite diseases and associated genes, organized according to microsatellite splice site proximity, are shown with relative lengths of host introns (white bars) vs. repeat expansions (color bars), repeat expansion sequences, and their potential to form stable secondary structures.

To identify features that distinguish pathogenic intronic microsatellites from unexpanded repetitive elements, we mapped the locations of repeats in introns and noted a positional bias for disease-relevant microsatellites toward splice sites (ss) with GC-rich microsatellites localized within 0.07–0.8 kb of splice sites, while A/AT-rich repeats were positioned 1.3–75.9 kb, often in downstream introns (Fig. 1*A* and Fig. S3 *A*–*D*). Because RNA structures and microsatellites are both known to influence splicing regulation (12–14), these observations led us to speculate that the GC-rich intronic microsatellite expansions alter RNA structures and/or transacting factor accessibility, resulting in impairment of spliceosome recruitment (Fig. S3*C*). Therefore, we tested if GC-rich microsatellite expansions caused misprocessing of host introns in affected brain and muscle tissues, as well as more accessible cells and tissues, including fibroblasts and blood.

CNBP Intron 1 Retention in DM2. To test our hypothesis that GC-rich expansions disrupt splicing of their host introns, we first selected the DM2 $CCTG^{exp}$ mutation since it is the largest microsatellite expansion reported to date (Fig. S3*A*). *CNBP* is also the most widely and highly expressed intronic expansion disease gene, increasing our ability to confidently measure its RNA processing pattern in a variety of tissues (Fig. S3*E*). While the $CCTG^{exp}$ is located in a large 12-kb intron (i), *CNBP* i1, it is only ∼0.8 kb upstream of the 3′ss (Figs. S1 and S3 *A* and *B*). Furthermore, the effect of the $CCTG^{exp}$ on *CNBP* expression is currently controversial, with some studies reporting no effect, and other studies a decrease, in CNBP RNA and protein levels in DM2 cells and tissues (15–17).

To detect potential CNBP pre-mRNA misprocessing, we queried publicly available strand-specific RNA-sequencing (RNA-seq) datasets obtained from DM2, the related disease DM1, Duchenne muscular dystrophy (DMD), and unaffected skeletal and cardiac muscle (18). As predicted, read coverage across *CNBP* i1 was observed in the variety of DM2 muscles, but not in control samples (Fig. 2*A*). For all RNA-seq experiments, we calculated three distinct metrics: (*i*) relative enrichment in reads spanning intron–exon junctions; (*ii*) average per base pair read coverage across the retained intron; and (*iii*) the fraction of intron-containing molecules (IR ratio) for all four CNBP introns using IRFinder (19, 20). As expected, IR was exclusively elevated for DM2 CNBP i1 with an IR ratio of ∼0.35, while splicing of introns 2, 3, and 4 was unaffected (Fig. 2*B* and Fig. S4 *A*–*C*). To confirm CNBP i1 retention using additional patient samples and an alternative experimental approach, we analyzed microarray datasets obtained from DM2, DM1, facioscapulohumeral muscular dystrophy (FSHD), and unaffected control muscle biopsies (21, 22) and observed statistically significant and specific CNBP i1 retention in DM2 compared with this large control cohort (Fig. 2*C*). This analysis of other muscular dystrophies increased our confidence that CNBP i1 retention was DM2-specific rather than reflecting a general myopathic feature (23). Since bidirectional transcription of *CNBP* i1 occurs (24), we quantified strand-specific RNA-seq read coverage to confirm that our analysis was not confounded by antisense transcription and found >99.9% of reads originated from sense molecules in muscles (Fig. S4*D*).

CNBP i1 retention was validated by using RT-PCR from biopsied skeletal muscle (tibialis anterior; TA) since RNA degradation is minimized in these samples. IR detection from the 3′ ss allowed selective amplification of the retained intron from pre-mRNA intermediates and simultaneous analysis of introns 1, 2, and 3 (Fig. 2*D*). In agreement with whole transcriptomic data, we also detected selective, and up to a sixfold increase in, CNBP i1 retention in DM2 biopsied skeletal muscle (Fig. 2 *E* and *F*), brain frontal cortex (Fig. S4*E*), and lymphoblastoid cell lines (LCLs) (Fig. 2*G*) compared with disease and unaffected controls.

Next, we tested if CNBP i1 is developmentally regulated in unaffected organs and excluded this possibility by RT-PCR in human fetal and adult tissues (Fig. S4*F*) and using an in silico model of muscle development (Fig. S4*G*) (25). Then, we addressed the question of whether CNBP i1 was a retained or detained intron, since the latter is incompletely spliced RNA that is not exported into the cytoplasm (26). The subcellular localization of $CCUG^{exp}$ RNA was detected by RNA-FISH in patient-derived fibroblasts by using a repeat-specific probe, and, although the level of mutant RNA in foci was higher in the nucleus, $CCUG^{exp}$ RNA was also readily detectable in the cytoplasm of DM2, but not control, cells (Fig. 2*H* and Fig. S4 *H*–*J*). This in situ analysis was confirmed by subcellular fractionation of control and DM2 fibroblasts followed by RT-PCR. Using 3′ss and 5′ss RT-PCR assays, the CNBP i1-containing mRNAs were clearly evident in the cytoplasm of DM2 unlike the controls (Fig. 2*I*). The increased level of i1 inclusive RNA expression in the cytoplasm was also confirmed by RT-PCR with three different primer sets localized in i1 proximal and distal to the 5′ss, as well as just upstream of the $CCUG^{exp}$ (Fig. S4*K*). Finally, we determined if CNBP i1 retention resulted in introduction of a premature termination codon (PTC) and nonsense-mediated decay (NMD) by treating cells with G418 to induce PTC read-through (27). For DM2 fibroblasts, G418 significantly increased the IR ratio, indicating that NMD reduced cytoplasmic levels of CNBP i1-containing mRNA (Fig. S4*L*). Therefore, multiple experimental approaches and patient samples confirmed selective retention of CNBP i1 in DM2 tissues and cells and that CCUG expansions were associated with IR both in vivo and in cell cultures.

Host Intron Retention as an Accessible Biomarker. To test CNBP i1 retention as a potential blood biomarker, peripheral blood lymphocytes (PBLs) were isolated from DM2 patient blood together with disease (DM1 and ALS) and unaffected controls. In agreement with our findings in other tissues, retention of CNBP i1 was enhanced in DM2 PBLs compared with controls, while splicing of introns 2, 3, and 4 in DM2 was unimpaired, and all CNBP introns in the ALS samples were spliced (Fig. 3 *A* and *B* and Fig. S5*A*). As observed in muscle, these reads primarily originated from the sense strand (Fig. S5*B*). To clarify the relationship between IR and $CCTG^{exp}$ length, genomic DNA Southern blot

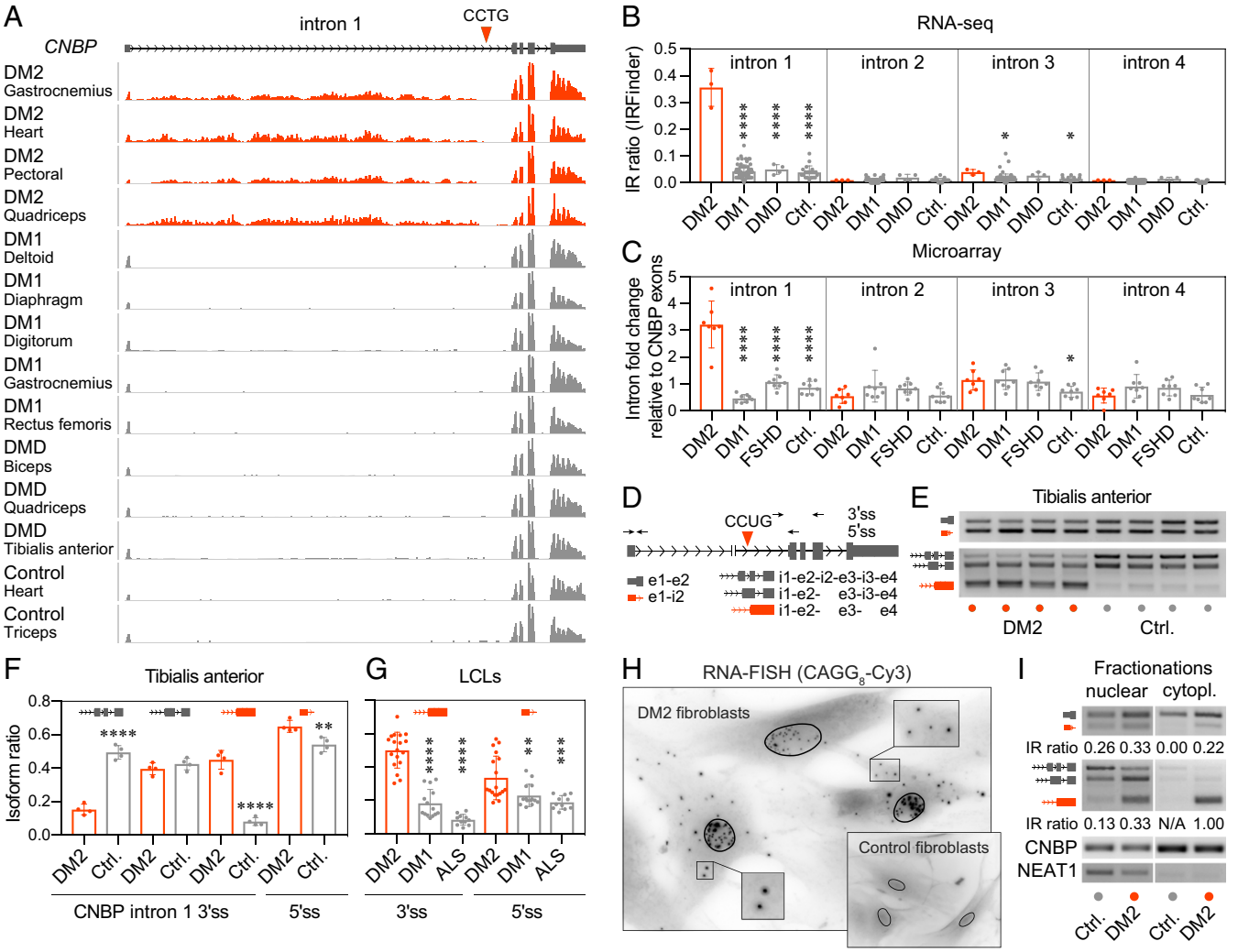

Fig. 2. CNBP intron 1 retention in DM2. (*A*) University of California, Santa Cruz (UCSC) genome browser view of the *CNBP* gene with the intronic CCTG position indicated (triangle). Wiggle plots represent DM2 skeletal muscles (gastrocnemius, pectoral, and quadriceps) and heart RNA-seq data with disease control skeletal muscle (DM1 deltoid, digitorum, gastrocnemius, and rectus; DMD biceps, quadriceps, and TA) and unaffected control heart and triceps. (*B*) CNBP IR ratios calculated by IRFinder. Only samples with a transcript integrity number >75% were analyzed (Fig. S4*A*). Bar graph shows mean ± SD for RNA-seq data from 3 DM2, 74 DM1, 4 DMD, and 19 control (Ctrl.) skeletal or cardiac muscles. (*C*) Human microarray analysis of the fold change of the four CNBP introns relative to the absolute exon signal for seven DM2, eight FSHD, and eight unaffected control vastus lateralis biopsy, and eight DM1 autopsy, muscles. For *B* and *C*, one-way ANOVA with Dunnett's multiple comparison test: $*P < 0.0332$; $****P < 0.0001$. (*D*) Schematic of the CNBP i1 5′ss (3-primers) and 3′ss (2-primers) RT-PCR assay. The IR ratio reflects the relative amount of the isoform with only retained i1 relative to other PCR products. (*E*) RT-PCR analysis of CNBP i1 retention for age-matched biopsied DM2 ($n = 4$) and unaffected control ($n = 4$) TA muscles. (*F*) Isoform ratio calculated based on CNBP i1 5′ss and 3′ss RT-PCR assays. (*G*) CNBP i1 5′ss and 3′ss analysis of DM2 ($n = 18$), DM1 ($n = 14$), and ALS ($n = 11$) LCLs. Bar graph shows mean ± SD for CNBP i1 retention ratio. For *F* and *G*, one-way ANOVA with Dunnett's multiple comparison test: $**P < 0.0068$; $***P < 0.0005$; $****P < 0.0001$. (*H*) RNA-FISH for $CCUG^{exp}$ detection in DM2 fibroblasts using a repeat-specific probe, $CAGG_8$-Cy3. Nuclei are outlined based on DAPI staining (Fig. S4*H*). (*I*) Subcellular fractionation of DM2 fibroblasts confirms the presence of CNBP i1 mRNA in the cytoplasm. DM2 and control fibroblast nuclear and cytoplasmic fractions were analyzed by CNBP i1 5′ss and 3′ss RT-PCR analysis. CNBP and NEAT1 are fractionation controls.

analysis was performed on DM2 patient PBLs, which were categorized as either carrying no (control, DM1), small (100–400 CCTGs; some of these patients were presymptomatic), or large (>1,000 CCTGs) *CNBP* expansions (Fig. 3*C*). RT-PCR analyses of DM2, DM1, and ALS PBLs demonstrated that CNBP i1 retention was dependent on repeat length (Fig. 3*D*) with a fourfold increase in i1 retention between DM2 PBLs with large expansions versus unaffected and disease (DM1 and ALS) controls (Fig. 3*E* and Fig. S5 *C* and *D*). Interestingly, PBLs with predominantly small expansions also showed a twofold increase in CNBP i1 retention versus controls, although these populations also displayed length mosaicism with larger expansions detectable at a reduced level. To examine if intron retention is restricted to the mutant *CNBP* allele, we took advantage of a DM2-linked A>C (rs1871922) *CNBP* i1 SNP previously linked to the mutant DM2 allele (16, 28). Using CNBP i1 5′ss assay primers, we amplified both genomic DNA (gDNA) and cDNA from DM2 PBLs and fibroblasts, and Sanger sequencing revealed overrepresentation of the DM2-linked SNP in cDNA compared with gDNA, indicating preferential i1 inclusion in mutant CNBP RNA (Fig. 3*F*). Because DM2 is a dominant disease and the CNBP i1 retention signal was diluted by transcripts originating from the unexpanded allele, we also confirmed that CNBP i1 retention was twofold higher in homozygous versus heterozygous DM2 patient fibroblasts (Fig. 3*G* and Fig. S5 *E* and *F*). Based on these observations, we concluded that intron retention is a useful DM2 blood biomarker.

While selective retention of CNBP i1 was only observed in DM2 cells and tissues, it was not clear if the $CCUG^{exp}$ mutation

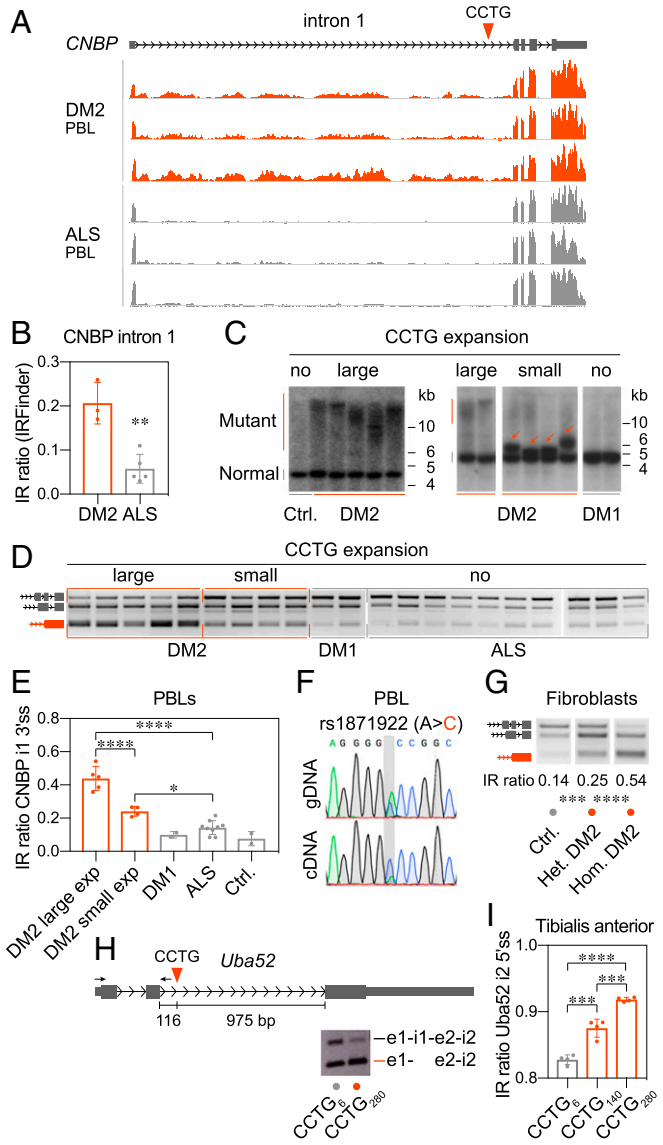

Fig. 3. CNBP intron 1 retention in DM2 as a blood biomarker. (*A*) UCSC browser view of *CNBP* and wiggle plots of PBL RNA-seq data from DM2 (n = 3) and ALS (n = 5) controls. (*B*) CNBP i1 retention ratio calculated by IRFinder (Fig. S5). Two-tailed t test: $^{**}P = 0.0018$. (*C*) Southern blot analysis of genomic DNA derived from DM2 patient PBLs with small (100–400 CCTGs) and large (≥1,000 CCTGs) expansions with DM1 disease and unaffected controls (Ctrl.). (*D*) CNBP i1 3′ss RT-PCR analysis of PBLs from DM2 patients with large (n = 5) and small (n = 4) CCTG expansions, DM1 (n = 2), ALS (n = 9), and unaffected controls (n = 2). (*E*) Bar graph shows mean ± SD for CNBP i1 retention ratio. One-way ANOVA with Tukey's multiple comparison test: $^{*}P = 0.0135$; $^{**}P = 0.0055$; $^{***}P < 0.0008$; $^{****}P < 0.0001$. (*F*) CNBP i1 retained mRNA is specific for the mutant allele. Sanger sequencing trace of gDNA vs. cDNA indicated the DM2-specific SNP predominates in the mRNA/cDNA population. (*G*) CNBP i1 3′ss analysis of heterozygous and homozygous DM2 and control fibroblasts. Two-tailed t test: $^{***}P = 0.0007$; $^{****}P < 0.0001$; three technical replicas. (*H*) Mouse *Uba52* construct with 6, 140, or 280 CCTG repeats inserted downstream of the exon 2 5′ss. (*I*) Mouse TA muscles were electroporated with constructs (n = 4 each), and Uba52 i2 retention was assessed 1 wk later by RT-PCR. Bar graph shows mean ± SD represents Uba52 i2 retention ratio. One-way ANOVA with Tukey's multiple comparison test: $^{***}P < 0.0002$; $^{****}P < 0.0001$.

played a direct role in retention. Thus, we tested whether CCUG repeats induced IR in vivo using a splicing reporter consisting of a modified mouse *Uba52* gene with 6 (control length), 140, or 280 (mutant lengths) CCTG uninterrupted repeats inserted in i2 (Fig. 3*H*). This reporter was selected because *Uba52* is widely expressed, and our studies have indicated that overexpression of this reporter is well tolerated. Plasmids with varying CCTG repeat lengths were electroporated into TA muscles of anesthetized mice, RNAs were isolated 1 wk after electroporation, and Uba52 i2 retention was assessed. In these samples, i2 retention showed a significant increase with CCTG repeat length, indicating that insertion of a $CCTG^{exp}$ downstream of a 5′ss is sufficient to drive IR (Fig. 3*I*). Together, these data support the possibility that CCUG expansions induce IR.

Intron Missplicing in GC-Rich, but Not A/AT-Rich, Microsatellite Expansion Diseases. To test if GC-rich, but not A/AT-rich, microsatellite expansions result in selective IR, we compared IR between two additional GC- and A/AT-rich microsatellite expansion diseases. FECD is caused by a CTG^{exp} in *TCF4* i3, but in contrast to DM2 $CCTG^{exp}$, this mutation is located in the middle of the intron and the CTG^{exp} is considerably smaller (<1.7 kb) (Fig. S1). To detect potential TCF4 i3 retention, we queried publicly available RNA-seq datasets obtained from FECD and control corneal endothelium samples (29, 30). Similar to DM2 *CNBP* i1, we observed an increase in *TCF4* i3 read coverage in FECD, but not in unaffected controls (Fig. 4*A*), with a mean IR ratio of ∼0.18 (Fig. S6*A*). The FECD read distribution across *TCF4* i3 was biased toward the 5′ end and complicated by the presence of an alternative first exon (AFE) with multiple 5′ss in this region (Fig. 4*A*). Therefore, to confirm retention, we analyzed relative enrichment in reads supporting coverage between i3 and flanking exons and average per-nucleotide read coverage across *TCF4* i3 (Fig. S6 *B* and *C*). As expected, both metrics were enriched in FECD samples. Although sense and antisense reads could not be discriminated in FECD RNA-seq datasets, we tested whether antisense transcription occurred at this locus by strand-specific cap analysis of gene expression (CAGE)-seq (31) and found that sense, but not antisense, transcription start sites were detectable (Fig. S6*D*). Analysis of the other strand-specific RNA-seq datasets used in this study also failed to detect antisense transcription across this region.

Next, we tested IR in C9-ALS/FTD, where the $GGGGCC^{exp}$ mutation is located in *C9orf72* i1 between AFEs 1a and 1b (Fig. S1). Expansion of the GC-rich repeat alters the activity of promoters upstream of exons 1a and 1b, although transcription from the latter is more severely compromised (32). To determine if this type of expansion mutation also resulted in IR in brain and blood, we assessed C9orf72 i1 retention using RNA-seq strand-specific datasets from C9-ALS/FTD and control samples (33). In agreement with prior results from cell lines (34), we observed increased RNA-seq read coverage across *C9orf72* i1 in C9-ALS/FTD cortex and cerebellum, but not in sporadic (s)ALS and unaffected control, brain samples, although similar to FECD, the read distribution was biased toward the 5′ end of this intron (Fig. 4*B*). Since IR was previously noted in LCLs (34) and *C9orf72* expression is particularly high in myeloid cells (35), we also analyzed RNA-seq for C9-ALS/FTD, sALS ($GGGGCC^{exp}$ negative), and DM2 PBLs (Fig. S7 *A* and *B*). Similar to the brain samples, high read coverage across ∼2.5 kb downstream of *C9orf72* e1b was observed in PBLs, suggesting the existence of an unannotated AFE and/or alternative e1b 5′ss (Fig. 4*B*). In agreement with this possibility, splice junction reads were obtained between C9orf72 i1 and e2, and previously unannotated junctions were validated by RT-PCR and Sanger sequencing by using PBL, LCL, and cortex samples (Fig. S7 *C* and *D*). To test if a novel AFE existed, we analyzed C9orf72 sense-strand CAGE-seq data and identified reads supporting the existence of a novel exon, which we named e1c, and our analysis was confirmed by FANTOM5 consortium annotated transcripts (Fig. S7*C*). To determine if e1c expression was altered in C9-ALS/FTD brain, we used C9-37 and -500 BAC transgenic mouse models for C9-ALS/FTD which express either 37 or 500 GGGGCC repeats, respectively (36). Using human-specific C9orf72 e1c-e2 primers, we detected an elevated signal for C9-500 compared with C9-37 (Fig.

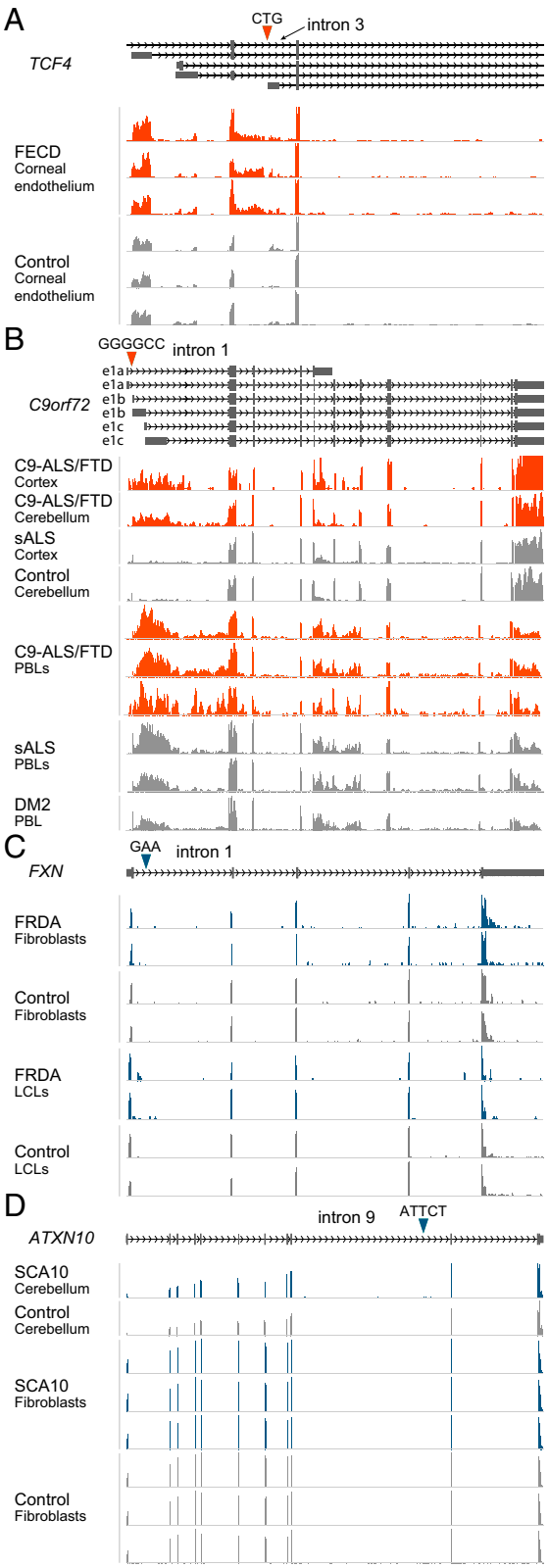

Fig. 4. Intron retention induced by GC-rich, but not A/AT-rich, microsatellite expansion mutations. UCSC genome browser views of *TCF4*, *C9orf72*, *FXN*, and *ATXN10* genes with their respective intronic CTG, GGGGCC, GAA, and ATTCT expansion positions are shown with wiggle plots representing RNA-seq data. (*A*) FECD and control corneal endothelial cells (Fig. S6). (*B*) *C9orf72* C9-ALS/FTD, sALS, and DM2 cortex, cerebellum, and PBLs (Fig. S7). (*C*) FRDA and control fibroblasts and LCLs. (*D*) SCA10 and control cerebella and fibroblasts.

S7*E*), which indicated the presence of GGGGCC^{exp} DNA, possibly due to increased transcription initiation at e1c.

Next, to quantify C9orf72 i1 retention in PBLs, we computed read coverage in the intron downstream of e1c (Fig. S7*F*). Due to e1a and e1b low expression and/or dysregulation (Fig. S7 *G–I*), we quantified relative i1-e2 read junction coverage only (Fig. S7*J*). As for our previous analyses, antisense C9orf72 transcripts did not obscure our findings in PBL samples (Fig. S7*K*). Since detection of C9orf72 i1 retention in human tissues was challenging by RT-PCR (34) and is subject to signal dilution due to the presence of transcripts originating from unexpanded alleles, we used human-specific C9orf72 i1 3′ss primers and confirmed elevated C9orf72 i1 retention for C9-500 compared with C9-37 transgenic models (Fig. S7 *L–O*).

In addition to the GC-rich DM2, FECD, and C9-ALS/FTD mutations, we also examined two A/AT-rich expansions, the FRDA GAA^{exp} in *FXN* i1 and the SCA10 ATTCT^{exp} in *ATXN10* i9. Although the FRDA GAA^{exp} mutation reduces transcription of the mutant *FXN* allele, a previous study reported that GAA^{exp} induced intron missplicing in hybrid and *FXN* minigene splicing reporter assays (37). However, we failed to detect IR for FXN i1 in FRDA fibroblasts and LCLs (Fig. 4*C* and Fig. S8 *A* and *B*). Moreover, IR was not detected for the ATXN10 AUUCU^{exp} i9 mutation in either SCA10 cerebellum or fibroblasts (Fig. 4*D* and Fig. S8*C*). Overall, we concluded that repeat-induced host intron misprocessing is a general feature of GC-rich, but not A/AT-rich, microsatellite expansion diseases.

Discussion

Intron retention during RNA processing, with potential effects on nuclear retention, nucleocytoplasmic transport, and cytoplasmic turnover, is a conserved regulatory mechanism that impacts a wide range of cellular events, including tissue development, neuronal activity-dependent gene expression, and tumor suppressor inactivation (38–40). In this study, we demonstrate that disease-associated GC-rich intronic microsatellite expansions induce IR in a variety of affected patient cells and tissues, and these retention events are not developmentally regulated. Interestingly, IR does not occur in A/AT-rich intronic expansion diseases, where the mutation is located more distally from the nearest splice site, including FRDA (41) and SCA10 (42). These results are consistent with our hypothesis that GC-rich microsatellite expansions exert an inhibitory effect on splicing by altering RNA structure and/or access of splicing factors to intronic regulatory regions. We also show that these IR events can be readily assayed by RT-PCR assays using peripheral blood, greatly increasing the scope and RNA quality of samples available for analysis. Using DM2 as a model, we show that CNBP i1 retention occurs even with expansion sizes in the low pathogenic range, which suggests that this assay may be informative for presymptomatic patients.

Our results of CNBP i1 retention appear to contradict some earlier studies, based on RNA-FISH and protein analyses that CNBP pre-mRNA splicing and protein levels are unaffected in DM2 tissues and cell lines (17, 43), although other reports have concluded that CNBP levels are reduced in DM2 (15, 16). In contrast to this study, these earlier reports did not directly examine CNBP i1 retention in affected tissues. Nevertheless, our results clearly demonstrate that CNBP i1 retention occurs in DM2 tissues and blood cells. Since CCUG^{exp} RNA foci are a distinguishing feature of DM2 cells, it is likely that only a portion of CNBP pre-mRNAs are aberrantly spliced although the level of misspliced CNBP transcripts is sufficiently high to allow RT-PCR–based detection of i1 retention in blood.

RNA misprocessing has been reported for a number of microsatellite expansion diseases. In DM1 and DM2, the expression of C(C)UG^{exp} RNAs results in foci formation and sequestration of the MBNL family of alternative splicing factors and expression of developmentally inappropriate isoforms for a wide variety of genes (44). In contrast, the Huntington disease CAG^{exp} mutation is associated with HTT i1 misprocessing and cryptic polyadenylation site use (45). For FRDA, the *FXN* GAA^{exp} mutation causes gene silencing, possibly due to impairment of transcriptional elongation (46, 47). While FXN i1 missplicing has been

documented for a *FXN* minigene splicing reporter (37), another group failed to detect misspliced FXN RNAs in FRDA patient-derived fibroblasts and LCLs (47), in agreement with the results of this study. Moreover, our findings have implications for DM2 pathogenesis. IR is a characteristic feature of DM2, and the export of CNBP mRNAs with a selectively retained i1 could facilitate RAN translation of CCUG repeats in the cytoplasm. Indeed, we have recently shown that RAN translation occurs in DM2 (24).

Conventional genetic strategies to map hereditary microsatellite expansion mutations are both time- and labor-intensive and are confounded by penetrance, expressivity, and pedigree ascertainment issues. Given the prevalence of RNA misprocessing in transcripts harboring expanded intronic microsatellites, we speculate that unbiased screening of patient samples with unknown disease etiologies will uncover additional expansions in novel disease-associated genes. The transcriptomic approach that we have developed using DM2 as a model could be used to screen a large cohort of blood samples from patients affected with neurological diseases for specific RNA misprocessing events followed by Southern blot analysis and DNA sequencing to identify novel microsatellite expansion mutations.

Materials and Methods

Patient muscle (autopsy and biopsy), brain (autopsy), and blood samples (DM1, DM2, ALS, and SCA10) were collected following written informed consent as approved by the Universities of Florida and Rochester Institutional Review Boards. PBLs were isolated from the buffy coat of freshly collected whole blood, and red blood cells were preferentially lysed and removed by using the RBC Lysis Buffer (Roche). PBLs were centrifuged, washed once with PBS, and used for either gDNA isolation (Flexigene kit; Qiagen), LCL generation, or total RNA isolation (TRIzol; Thermo Fisher Scientific) per the manufacturer's protocols. All procedures were approved by the Institutional Animal Care and Use Committee (University of Rochester). Additional materials and methods details are described in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank participating patients and A. Berglund and E. Wang for comments on the manuscript. This work was supported NIH Grants NS058901 and NS098819 (to M.S.S. and L.P.W.R.), NS040389 (to L.P.W.R.), NS048843 (to C.A.T.), and NS083564 (to T.A.); a Target ALS grant (to L.P.W.R.); and Muscular Dystrophy Association grants (to M.S.S. and L.P.W.R.). During this study, Ł.J.S. was a postdoctoral fellow of the Myotonic Dystrophy and Wyck Foundations.

1. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 7:e1002384.
2. Padeken J, Zeller P, Gasser SM (2015) Repeat DNA in genome organization and stability. *Curr Opin Genet Dev* 31:12–19.
3. López Castel A, Cleary JD, Pearson CE (2010) Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol* 11:165–170.
4. Iyer RR, Pluciennik A, Napierala M, Wells RD (2015) DNA triplet repeat expansion and mismatch repair. *Annu Rev Biochem* 84:199–226.
5. Mirkin SM (2007) Expandable DNA repeats and human disease. *Nature* 447:932–940.
6. Zhang N, Ashizawa T (2017) RNA toxicity and foci formation in microsatellite expansion diseases. *Curr Opin Genet Dev* 44:17–29.
7. Vanichkina DP, Schmitz U, Wong JJ, Rasko JEJ (2017) Challenges in defining the role of intron retention in normal biology and disease. *Semin Cell Dev Biol* 75:40–49.
8. Pearson CE, Nichol Edamura K, Cleary JD (2005) Repeat instability: Mechanisms of dynamic mutations. *Nat Rev Genet* 6:729–742.
9. Ciesiolka A, Jazurek M, Drazkowska K, Krzyzosiak WJ (2017) Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. *Front Cell Neurosci* 11:97.
10. Handa V, Yeh HJ, McPhie P, Usdin K (2005) The AUUCU repeats responsible for spinocerebellar ataxia type 10 form unusual RNA hairpins. *J Biol Chem* 280:29340–29345.
11. Park H, et al. (2015) Crystallographic and computational analyses of AUUCU repeating RNA that causes spinocerebellar ataxia type 10 (SCA10). *Biochemistry* 54:3851–3859.
12. Hefferon TW, Groman JD, Yurk CE, Cutting GR (2004) A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc Natl Acad Sci USA* 101:3504–3509.
13. Zuccato E, Buratti E, Stuani C, Baralle FE, Pagani F (2004) An intronic polypyrimidine-rich element downstream of the donor site modulates cystic fibrosis transmembrane conductance regulator exon 9 alternative splicing. *J Biol Chem* 279:16980–16988.
14. Wolfe MS (2009) Tau mutations in neurodegenerative diseases. *J Biol Chem* 284: 6021–6025.
15. Huichalaf C, et al. (2009) Reduction of the rate of protein translation in patients with myotonic dystrophy 2. *J Neurosci* 29:9042–9049.
16. Raheem O, et al. (2010) Mutant (CCTG)n expansion causes abnormal expression of zinc finger protein 9 (ZNF9) in myotonic dystrophy type 2. *Am J Pathol* 177: 3025–3036.
17. Margolis JM, Schoser BG, Moseley ML, Day JW, Ranum LP (2006) DM2 intronic expansions: Evidence for CCUG accumulation without flanking sequence or effects on ZNF9 mRNA processing or protein expression. *Hum Mol Genet* 15:1808–1815.
18. Wagner SD, et al. (2016) Dose-dependent regulation of alternative splicing by MBNL proteins reveals biomarkers for myotonic dystrophy. *PLoS Genet* 12:e1006316.
19. Wong JJ, et al. (2013) Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* 154:583–595.
20. Middleton R, et al. (2017) IRFinder: Assessing the impact of intron retention on mammalian gene expression. *Genome Biol* 18:51.
21. Nakamori M, et al. (2013) Splicing biomarkers of disease severity in myotonic dystrophy. *Ann Neurol* 74:862–872.
22. Batra R, et al. (2014) Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Mol Cell* 56:311–322.
23. Bachinski LL, et al. (2014) Most expression and splicing changes in myotonic dystrophy type 1 and type 2 skeletal muscle are shared with other muscular dystrophies. *Neuromuscul Disord* 24:227–240.
24. Zu T, et al. (2017) RAN translation regulated by muscleblind proteins in myotonic dystrophy type 2. *Neuron* 95:1292–1305.
25. Thomas JD, et al. (2017) Disrupted prenatal RNA processing and myogenesis in congenital myotonic dystrophy. *Genes Dev* 31:1122–1133.
26. Boutz PL, Bhutkar A, Sharp PA (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* 29:63–80.
27. Floquet C, Rousset JP, Bidou L (2011) Rescue of nonsense mutated p53 by read-through leads to apoptosis in cancers cells. *Med Sci (Paris)* 27:585–586.
28. Bachinski LL, et al. (2003) Confirmation of the type 2 myotonic dystrophy (CCTG)n expansion mutation in patients with proximal myotonic myopathy/proximal myotonic dystrophy of different European origins: A single shared haplotype indicates an ancestral founder effect. *Am J Hum Genet* 73:835–848.
29. Du J, et al. (2015) RNA toxicity and missplicing in the common eye disease Fuchs endothelial corneal dystrophy. *J Biol Chem* 290:5979–5990.
30. Chen Y, et al. (2013) Identification of novel molecular markers through transcriptomic analysis in human fetal and adult corneal endothelial cells. *Hum Mol Genet* 22: 1271–1279.
31. Forrest AR, et al.; FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* 507:462–470.
32. Haeusler AR, Donnelly CJ, Rothstein JD (2016) The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. *Nat Rev Neurosci* 17:383–395.
33. Prudencio M, et al. (2015) Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. *Nat Neurosci* 18:1175–1182.
34. Niblock M, et al. (2016) Retention of hexanucleotide repeat-containing intron in C9orf72 mRNA: Implications for the pathogenesis of ALS/FTD. *Acta Neuropathol Commun* 4:18.
35. O'Rourke JG, et al. (2016) C9orf72 is required for proper macrophage and microglial function in mice. *Science* 351:1324–1329.
36. Liu Y, et al. (2016) C9orf72 BAC mouse model with motor deficits and neurodegenerative features of ALS/FTD. *Neuron* 90:521–534.
37. Baralle M, Pastor T, Bussani E, Pagani F (2008) Influence of Friedreich ataxia GAA noncoding repeat expansions on pre-mRNA processing. *Am J Hum Genet* 83:77–88.
38. Braunschweig U, et al. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* 24:1774–1786.
39. Mauger O, Lemoine F, Scheiffele P (2016) Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. *Neuron* 92:1266–1278.
40. Jung H, et al. (2015) Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* 47:1242–1248.
41. Sanchez N, et al. (2016) Characterization of frataxin gene network in Friedreich's ataxia fibroblasts using the RNA-Seq technique. *Mitochondrion* 30:59–66.
42. Matsuura T, et al. (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat Genet* 26:191–194.
43. Botta A, et al. (2006) Effect of the [CCTG]n repeat expansion on ZNF9 expression in myotonic dystrophy type II (DM2). *Biochim Biophys Acta* 1762:329–334.
44. Sznajder LJ, et al. (2016) Mechanistic determinants of MBNL activity. *Nucleic Acids Res* 44:10326–10342.
45. Neueder A, et al. (2017) The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. *Sci Rep* 7:1307.
46. Kumari D, Usdin K (2012) Is Friedreich ataxia an epigenetic disorder? *Clin Epigenetics* 4:2.
47. Punga T, Bühler M (2010) Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *EMBO Mol Med* 2:120–129.