INVESTIGATING THE MOLECULAR MECHANISMS OF SPLICING-PERTURBING SMALL MOLECULES WITH MASSIVELY PARALLEL SEQUENCING IN A MYOTONIC DYSTROPHY 1 MODEL

by

MATTHEW TANNER

A THESIS

Presented to the Department of Chemistry and Biochemistry
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

June, 2014

An Abstract of the Thesis of

Matthew Tanner for the degree of Bachelor of Science
in the Department of Chemistry and Biochemistry to be taken June, 2014

Title: Investigating the Molecular Mechanisms of Splicing-Perturbing Small Molecules with Massively Parallel Sequencing in a Myotonic Dystrophy 1 Model

Approved: _____
J. Andrew Berglund

Myotonic dystrophy is the most common form of adult-onset muscular dystrophy and appears in two forms: myotonic dystrophy 1 (DM1) and 2 (DM2). Both diseases are characterized by progressive muscle degeneration, myotonia, iridescent cataracts, and in severe cases neurodegeneration and cardiac dysfunction. Both forms of myotonic dystrophy are caused by an expansion of repeat DNA in distinct loci in the genome. In DM1, a CTG repeat is expanded from less than 50 repeats in normal individuals to up to several thousand repeats in DM1 patients. The molecular basis of this disease relies on the transcription of these repeats from DNA into RNA. Small molecules that can specifically target the repeats at the DNA level and inhibit their transcription – and thus alleviate the disease symptoms – represent a prime target for the development of therapeutics. This study investigates the capacity of two small molecules, pentamidine and actinomycin D, to reverse the molecular symptoms of DM1 through transcriptional inhibition and provides insight into their DNA target specificity and potential mechanisms of action.

Acknowledgments

I would like to thank Professor Berglund for providing me with the opportunity to involve myself in meaningful research, as well as helping me with countless other opportunities outside of research, including teaching experience, scholarships, and future education. I would also like to thank Ruth Siboni and Stacey Wagner for attentive mentorship in my research as well as the rest of the Berglund lab for help and support along the way.

Table of Contents

Introduction 1

Myotonic dystrophy: physiology and molecular basis 1

Disrupting MBNL1 interactions with CUG repeats through use of small molecules 4

Methods 8

In cellulo reverse transcription-polymerase chain reaction splicing assay 8

Massively parallel sequencing of transcriptomic RNA and analysis 10

Results 13

Actinomycin D and pentamidine have different effects on DM1 mis-splicing events 13

Many splicing factors are differentially expressed following actinomycin D treatment 14

Actinomycin causes widespread changes in cassette exon alternative splicing 19

Specificity of actinomycin D 20

Discussion 26

Actinomycin likely affects alternative splicing *via* differential expression of splicing factors 26

CTG DNA motifs and increased early gene transcripts are associated with decreased expression upon actinomycin treatment 30

Actinomycin activates transcriptional regulatory pathways at higher doses 32

Appendix 34

A primer on the central dogma of molecular biology 34

Appendix figures 39

References 43

List of Figures

Figure 1: CUG repeat RNAs form foci in the nucleus of neurons of DM1 patients and colocalize with splicing regulator MBNL1 1

Figure 2: Different forms and outcomes of alternative splicing 2

Figure 3: Pentamidine rescues mis-splicing and reduces CUG repeat:MBNL1 foci 3

Figure 4: Pentamidine binds DNA and inhibits T7 transcription 4

Figure 5: Actinomycin D binds CTG duplex DNA 5

Figure 6: Schematic of cell culture splicing assay 9

Figure 7: Schematic of ScriptSeq v2 sequencing library preparation 10

Figure 8: Schematic of TopHat alignment and Cufflinks transcript assembly 11

Figure 9: Effects of actinomycin and pentamidine treatment (individual and combinatorial) on DM1 mis-splicing events 13

Figure 10: Clustered heatmap of log2(foldchange) in expression of genes with gene ontology "RNA splicing" in pentamidine-treated samples 15

Figure 11: Clustered heatmap of log2(foldchange) in expression of genes with gene ontology "RNA splicing" in actinomycin-treated samples 16

Figure 12: FPKM of splicing factors implicated in DM1 that are significantly differentially expressed at 25 nM actinomycin 16

Figure 13: Changes in cassette exon alternative splicing following actinomycin treatment 18

Figure 14: Significantly differentially expressed genes ($q < 0.05$) between 0 and 5 nM actinomycin D 20

Figure 15: Actinomycin changes the distribution of aligned reads in early gene sequences of the 74 significantly decreased genes at low actinomycin dosage 22

Figure 16: CTG motifs correlate with decreased expression following actinomycin treatment 23

Figure 17: Average minor groove width may be a predictor of small molecule-gene interactions 25

Figure 18: Enriched motifs in promoter regions 350 bp upstream to 50 bp downstream of transcription start sites of the 3947 differentially expressed genes at 25 nM actinomycin 26

Figure A1: Structure of nucleic acid bases 34

Figure A2: Three DNA nucleotides connected by two phospho-diester bonds 35

Figure A3: dsDNA structure 35

Figure A4: Transcription initiation and elongation 36

Figure A5: Splicing of a primary transcript 37

Figure A6: Polymerization of amino acids into nascent proteins on the ribosome 37

Figure A7: Samples used as replicates in pentamidine titration correlate well with each other 39

Figure A8: Marginal splicing rescue achieved by actinomycin and pentamidine treatment 40

Figure A9: Several of the 74 significantly decreased genes exhibit a shift in the distribution of early gene region reads 41

Figure A10: Correlations of occurrences of all 3-mers and all 4-mers in early gene regions with change in expression of the 74 significantly decreased genes at 5 nM actinomycin 42

INTRODUCTION

Myotonic dystrophy: physiology and molecular basis

Myotonic dystrophy is the most common form of adult-onset muscular dystrophy and is characterized by many multisystemic symptoms: progressive muscle atrophy, myotonia (*i.e.* the inability to relax one's muscles after contracting them), iridescent cataracts, cardiac defects, and in severe cases neurological defects[1]. It is a dominantly inherited genetic disease that affects more than 1 in 8,000 individuals globally[1]. The disease exists in two forms that are genetically distinct but result in nearly identical symptoms: myotonic dystrophy 1 (DM1) and 2 (DM2). Both diseases are caused by different repeat expansions in non-coding regions of different genes. DM1 is caused by a CTG repeat expansion in the 3' untranslated region (3' UTR, the region at the 3' end of a transcript that follows a stop codon) of the *DMPK* gene, ranging from about 30-50 or less repeats in unaffected individuals to up to several thousand in DM1 patients[1,2]. DM2 is caused by a CCTG repeat expansion in the first intron of the *CNBP* gene, and repeat numbers vary from less than 75 in unaffected individuals to a mean of approximately 5000 repeats in DM2 patients[3].

Given the facts that the DM1 and DM2 are autosomal-dominantly inherited (*i.e.* only one, non-sex chromosome containing the expanded repeats is necessary to cause

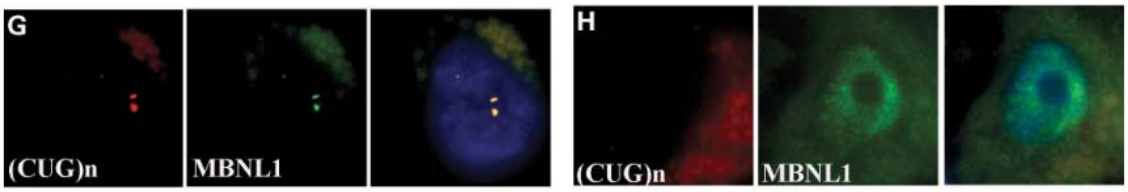

Figure 1. CUG repeat RNAs form foci in the nucleus of neurons of DM1 patients and colocalize with splicing regulator MBNL1. Panel G shows neurons from DM1 patients: CUG repeats in red (visualized *via* fluorescence *in situ* hybridization, FISH), MBNL1 protein in green (visualized *via* immunofluorescence), and DAPI (fluorescent DNA stain) in blue. Panel H shows neurons from non-DM1 individuals: no foci or colocalization observed.[7]

the disease) and disease severity increases and age of onset decreases with increasing repeat length, a small handful of potential mechanisms exist whereby these expanded repeats become pathological. The expanded CTG repeats in DM1 have been shown to reduce expression of *DMPK* and the nearby gene *SIX5*, and indeed mice homozygous (*i.e.* both chromosomal copies) for non-functioning *DMPK* – *DMPK* knockout mice – develop cardiac conduction abnormalities while mice homozygous for non-functioning *SIX5* develop cataracts[4,5]. While these phenomena likely contribute to DM1 pathology, they are independent of the major clinical symptom of DM1 and DM2: muscle myotonia. Instead of acting at the DNA level to disrupt local gene expression, the prevailing model shows that, following transcription, expanded CTG and CCTG repeats become toxic, gain-of-function RNAs. Expanded CUG repeat RNAs form foci in the nuclei of muscle cells and neurons of DM1 patients, and these foci are found to colocalize with the protein MBNL1 (Fig. 1)[7].

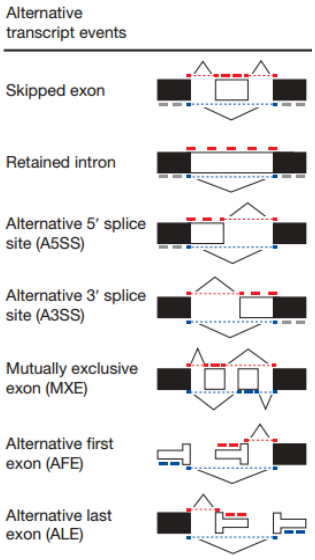

Figure 2. Different forms and outcomes of alternative splicing. Constitutively spliced exons (black boxes) and alternatively spliced exons may be spliced together in many different ways. Each alternative splicing event produces a new transcript (and potentially protein) isoform. MBNL1 regulates many cassette exon (skipped exon) events.[6]

MBNL1 is an evolutionarily conserved protein whose primary cellular function is the regulation of many different pre-mRNA alternative splicing events in the nucleus of a cell and is implicated as an important regulator of tissue differentiation and development in mammals[8]. Many different types of alternative splicing events exist (Fig. 2), but MBNL1 primarily regulates cassette exon (also known as skipped exon) events. MBNL1 colocalization with toxic CUG RNA may then indicate its sequestration

and that free, functional MBNL1 is depleted from the nucleus in DM1 patient cells. Indeed, many MBNL1-regulated alternative splicing events are found to be mis-spliced (*i.e.* the ratio of the two potential splicing isoforms is significantly changed, often shifted towards greater production of a non-functional isoform). Mis-splicing of the insulin receptor (*INSR*) transcript causes the insulin resistance that is observed in DM1 patients, mis-splicing of the voltage-sensitive chloride channel 1 (*CLCN1*) transcript causes the hallmark myotonia, and mis-splicing of cardiac troponin T type 2 and 3 (*TNNT2* and *TNNT3*) transcripts may cause DM1 cardiomyopathy[9,10]. Additionally, *MBNL1* knockout mice develop cataracts and abnormal skeletal muscle histology, and microarray experiments show that the vast majority of mis-splicing events in mice expressing long CUG repeat RNA and in *MBNL1* knockout mice are shared between the two conditions. These data altogether help to cement the primary molecular cause of DM1 and DM2 as the depletion of free MBNL1 *via* sequestration on toxic CUG (DM1) or CCUG (DM2) repeat RNA and subsequent dysregulation of MBNL1-regulated alternative splicing events[10,11].

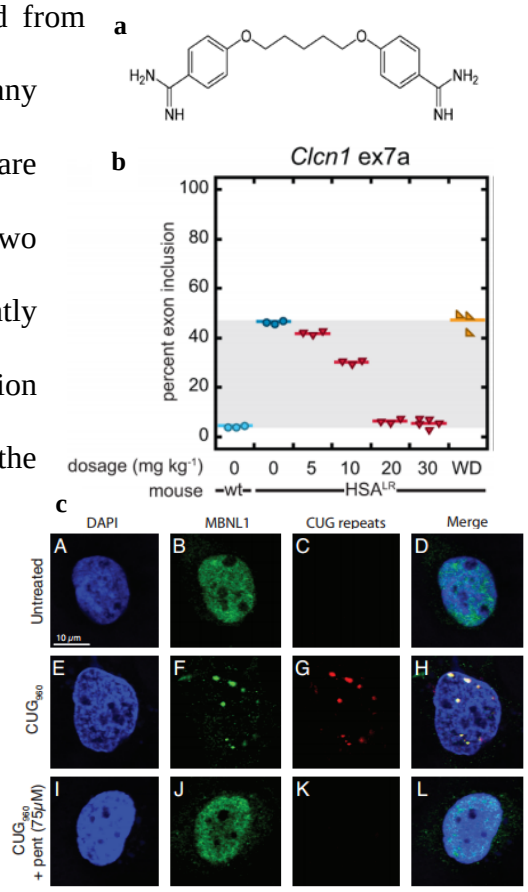

Figure 3. Pentamidine rescues mis-splicing and reduces CUG repeat:MBNL1 foci. **a)** The structure of pentamidine. **b)** The derivatized compound heptamidine reverses *Clcn1* mis-splicing in mice expressing toxic CUG repeat RNA (HSA^{LR} mice)[13]. **c)** In HeLa cells expressing a transcript containing 960 CUG repeats, MBNL1 (green) colocalizes with CUG repeat RNA (red). This colocalization is significantly reduced by 75 μM pentamidine[14].

Disrupting MBNL1 interactions with CUG repeats through use of small molecules

Knowing the fundamental mechanism by which the expanded CTG repeats in the *DMPK* gene cause disease opens the door for the design of therapeutic agents – typically small molecules which may be easily chemically synthesized and derivatized to modify biological functionality – to disrupt any number of biological processes that occur in DM1 pathogenesis. MBNL1 is an RNA-binding protein, and it is through this RNA binding that it normally regulates alternative splicing. *In vitro* RNA binding experiments show that MBNL1 selectively recognizes and binds to the sequence YGCY (where Y is either pyrimidine C or U) in its RNA substrates, explaining its high affinity for toxic CUG repeat RNA[12]. Attempting to identify compounds that may be able to disrupt this MBNL1:CUG repeat RNA interaction, a screen of small molecules revealed that the compound pentamidine is able to reverse some of the mis-splicing effects caused by the toxic CUG repeat RNA as well as reduce the number of CUG repeat:MBNL1 nuclear foci (Fig. 3)[13,14]. It was originally thought that pentamidine may bind directly to the CUG repeat RNA to displace MBNL1, but it is now known that this is likely not the mechanism by which pentamidine is able to reverse DM1 mis-splicing events.

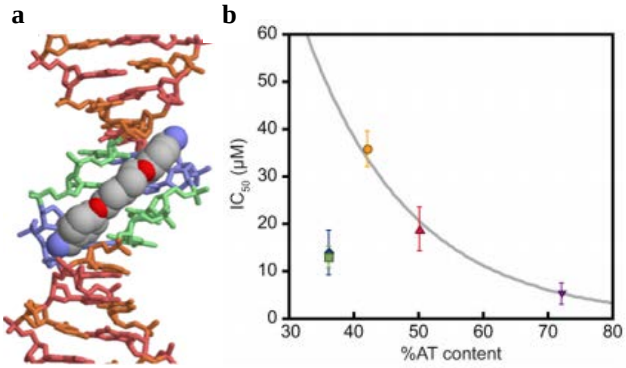

Figure 4. Pentamidine binds DNA, inhibits T7 transcription. **a)** Crystal structure of pentamidine (space-filling model) binding to AATT in the minor groove of dsDNA[15]. **b)** Pentamidine concentration at which 50% of transcription is inhibited (IC_{50}) as a function of template AT content[13].

HeLa cells expressing long CUG repeat RNA show decreased CUG RNA levels on northern blots with increasing pentamidine concentrations, suggesting that

pentamidine may be either inhibiting transcription of the repeats or destabilizing the repeats and hastening their degradation[13]. Pentamidine is known to bind the minor groove of DNA with a particular preference for binding in A/T-rich regions (Fig. 4a)[15]. *In vitro* transcription reactions using the viral T7 RNA polymerase and doped with pentamidine show that the production of RNA is generally inhibited more strongly when the template sequence has greater A/T content – however, transcription from templates containing 54 CTG or CAG repeats is preferentially inhibited by pentamidine despite its relatively low A/T content (Fig. 4b)[13]. We hypothesize that pentamidine is able to rescue DM1 mis-splicing events by preferentially binding to CTG repeat DNA and inhibiting the elongation of the RNA polymerase through the repeats, thus reducing toxic CUG repeat RNA levels and increasing levels of free MBNL1.

Based on this hypothesis, a drug that has been classically known for its strong DNA binding and potent inhibitory effects on transcription, actinomycin D, was investigated. Actinomycin is a complex small molecule that is synthesized by *Streptomyces spp.* soil bacteria and has been FDA approved as a chemotherapeutic agent since 1964, owing to its potent ability to inhibit transcription in fast-growing tumor cells. Since actinomycin has been previously crystallized bound to a CTG stem-loop (Fig. 5), we hypothesize that actinomycin may

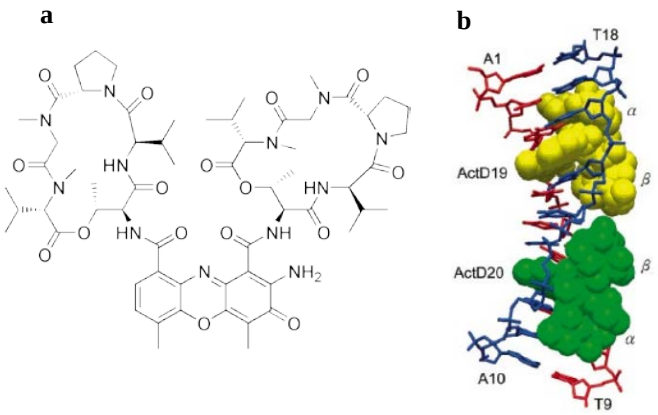

Figure 5. Actinomycin D binds CTG duplex DNA. **a)** The structure of actinomycin D, consisting of two depsipeptide rings (top) connected to a phenoxazone ring (bottom). **b)** Two actinomycin molecules (yellow and green) intercalating at GpC steps in an ATGCTGCAT duplex[16]. Actinomycin is known to bind preferentially to GpC steps[17].

have – at sub-therapeutic doses wherein the drug does not inhibit global transcription – a specificity for binding to CTG repeats and effectively inhibiting their transcription[16].

In this study, we investigate the capacity of actinomycin to reverse DM1 mis-splicing events in an *in cellulo* splicing system, both individually and in combinatorial treatments with pentamidine. We furthermore observe changes in transcriptional profiles of HeLa cells expressing toxic CUG repeat RNA and treated with a titration of either pentamidine or actinomycin using the recently-developed and powerful DNA sequencing technology known as massively parallel or next-generation sequencing. This sequencing technology became available in 2005 and quickly developed into a revolutionary tool for genetic studies. Prior to the advent of this technology, DNA sequencing was typically performed using Sanger sequencing, which allows for the determination of only one DNA sequence at a time with a limit of about 1000 bp for each sequence. For comparison, the haploid human genome comprises about 3.2 billion bp of DNA, requiring about 3.2 million runs of Sanger sequencing to cover the human genome once. The Human Genome Project – which spanned 13 years of research from 1990 to 2003, depended on the work of thousands of researches, and cost $2.7 billion – employed Sanger sequencing to construct its draft of the human genome. Now, a single researcher could use the same Illumina Hi-Seq 2000 massively parallel sequencer used in this project to sequence a full human genome thirty times over in a little more than a week at a cost of less than $10,000. Instead of investigating how any single gene at a time responds to some treatment – as one would have to do using prior technology – researchers can apply this sequencing technology to determine how the entire genome responds to a treatment, or they may use this technology to search for single-nucleotide

mutations (a veritable needle in the haystack that is the 6.4 billion bp diploid human genome) in a population to help determine the genetic determinants of a heretofore unexplained disease. Massively parallel sequencing is a key instrument in the forthcoming era of personalized medicine, wherein some drug treatments and therapeutics may be customized to each patient based upon associations between studied variations within his fully sequenced genome and genetic interactions with certain therapeutic treatments.

This sequencing technology has also become a powerful tool on the other end of the drug development pipeline. Traditionally, useful drugs – the majority of which are small molecules, often extracted from natural sources or chemically synthesized – have been identified *via* large screens of compounds for activities that are relevant to certain disease processes. With the aid of recently available technology, such as computationally-demanding molecular modeling and massively parallel DNA sequencing, rational drug design has sought to investigate and target a molecule's interactions with cellular processes from the ground up, and this movement has become a potent force in drug discovery and development. In our investigation of pentamidine and actinomycin and their respective potentials as starting points in the development of possible therapeutics for myotonic dystrophy 1, the utilization of massively parallel sequencing technology allows for the swift characterization of many aspects of how these drugs interact with the cell, since cellular responses are often carried out *via* regulation of genetic networks, resulting in the production of different amounts of RNA from differentially expressed genes. All the RNA produced in a cell can then be enzymatically copied out into DNA fragments that can then be submitted for massively

parallel sequencing – a method known as RNA-Seq. By identifying *all* the genes that are either activated or deactivated upon drug treatment through the use of next-generation sequencing technology, we are able to better understand how these drugs interact with a multitude of cellular processes, and we may infer whether these interactions are mediated through direct and sequence-specific interactions of each drug with DNA in certain genes or whether it is mediated through cellular activation of transcriptional networks in response to the drug's presence. Since we are effectively sequencing a very large sample of all the RNAs in our cell cultures, we are also able to detect changes in alternative splicing across the transcriptome – the suite of all transcripts that a cell produces – and characterize these changes. This global approach provides a much more complete picture of the efficacy and specificity of these small molecule drugs and provides insight into both their potential mechanisms of action as well as further strategies in the development of small molecule therapeutics for DM1.

METHODS

In cellulo reverse transcription-polymerase chain reaction splicing assay

The cell culture splicing assay comprises the main tool for directly determining whether small molecules are capable of reversing DM1-induced, MBNL1-regulated mis-splicing events (workflow shown in Fig. 6 and described in Coonrod, *et al.*[13]). Briefly, HeLa cells (originally taken from a human cervical adenocarcinoma) are cultured in Dulbecco's Modified Eagle Medium (DMEM high glucose with GlutaMAX™, Life Technologies) supplemented with 5% fetal bovine serum (FBS) to approximately 85-90% confluency. These cells are subsequently trypsinized (TrypLE™, Life Technologies) and plated in 6-well plates at 2.0×10^5 cells per well.

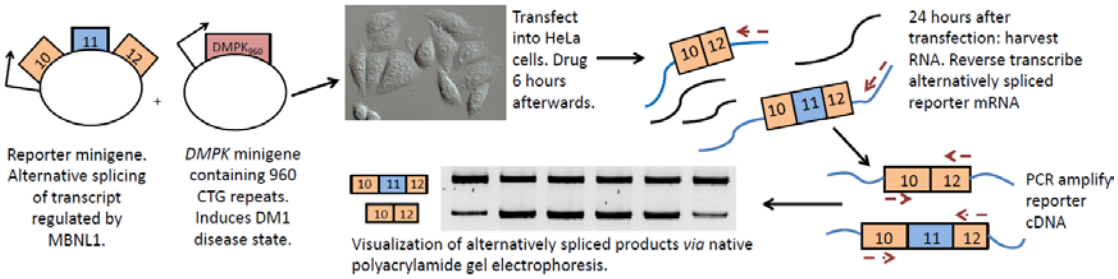

Figure 6. Schematic of cell culture splicing assay. Circles represent plasmid vectors. Colored boxes represent exons (in this case exons 10, 11, and 12 of the INSR minigene). Dashed arrows represent reporter-specific RT and PCR primers.

Plated cells are grown for approximately 24 hrs in DMEM, after which they are washed with 1X PBS (phosphate buffered saline) and placed in Opti-MEM® reduced serum media (Life Technologies) and transfected with 1000 ng total of plasmid DNA using Lipofectamine 2000® (Life Technologies) as described by the manufacturer's protocol. In control samples (wild-type, WT), cells are transfected with 500 ng of reporter minigene plasmid (whose transcript contains a single, MBNL1-regulated alternative splicing event) and 500 ng of empty vector (pcDNA3) plasmid. In experimental samples, cells are transfected with 500 ng of reporter minigene plasmid and 500 ng of $DMPK_{960}$ plasmid. $DMPK_{960}$ plasmid consists of a genomic fragment containing exons 11 through 15 of the human *DMPK* gene with 960 interrupted CTG repeats in the 3' UTR of exon 15, and its expression is driven by a strong cytomegalovirus (CMV) promoter[18].

Six hours following transfection, cells are washed with 1X PBS, placed in DMEM, and experimental samples are drugged with either pentamidine (hydrochloride salt, synthesized on site), actinomycin (Sigma-Aldrich), or a combination of the two. Eighteen hours following drugging, whole-cell RNA is harvested using an RNeasy kit (Qiagen) and quantitated using a NanoDrop 2000 UV spectrophotometer (Thermo Fisher). 500 ng of RNA from each sample is DNased with DNase I (New England

BioLabs) for 1 hr at 37 °C, and 100 ng of DNased-RNA is then subject to reverse transcription (RT) with SuperScript II reverse transcriptase (Life Technologies) using reporter-specific primers for 1 hr at 42 °C[22]. The resulting cDNAs are then amplified *via* polymerase chain reaction (PCR) using reporter-specific primers[18,19]. PCR products are then separated on a 6% 19:1 bis:acrylamide non-denaturing polyacrylamide gel, which is then stained with SYBR Green I dye (Life Technologies) and imaged using an Alpha Imager HP system (Alpha Innotech). Bands representing alternatively spliced transcripts are quantified using Alpha Imager HP software (Alpha Innotech), and percent exon inclusion is calculated as the ratio of the intensity of the included band to the sum of intensities of included and excluded bands (and *vice versa* for percent exon exclusion).

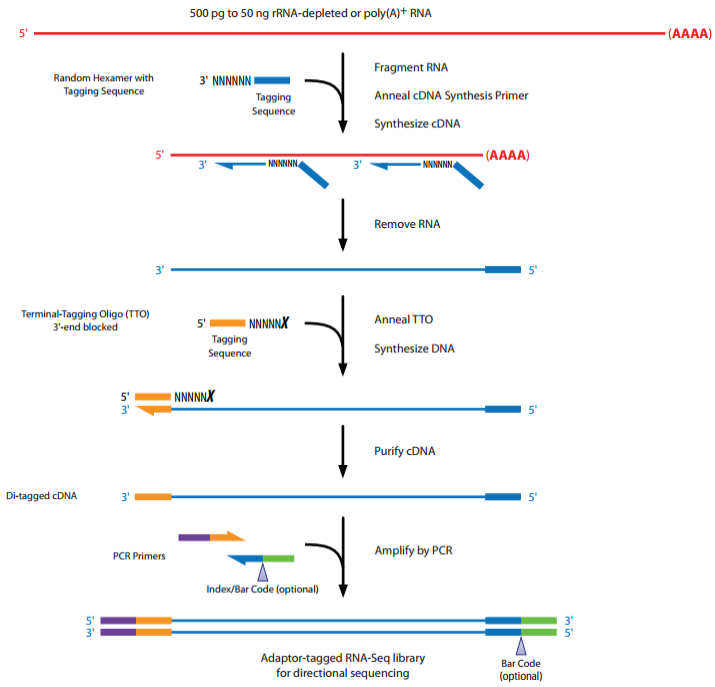

Figure 7. Schematic of ScriptSeq v2 sequencing library preparation[22]. cDNA fragments from each sample were amplified using reverse PCR primers with unique barcode sequences to allow for multiplexing during sequencing.

Massively parallel sequencing of transcriptomic RNA (RNA-Seq) and analysis

RNA-Seq is an increasingly popular method of characterizing cellular RNA expression profiles (*i.e.* the transcriptome) by taking advantage of technological advances in massively parallel sequencing of short DNA fragments[21]. Whole-cell RNA

from two replicates of DM1 HeLa cells expressing an INSR reporter minigene plasmid and treated with none, 5 nM, and 25 nM actinomycin D (as described previously) was used for RNA-Seq. For each sample, 10 µg of RNA was treated with DNase RQ1 (Promega) for 1 hr at 37 °C to remove any contaminating DNA, and the DNased RNA was recovered *via* phenol-chloroform-isoamyl alcohol extraction followed by ethanol precipitation and quantitated by NanoDrop. Subsequently, 1.3 µg of DNased RNA was depleted of abundant ribosomal RNA (rRNA) using a Ribo-Zero™ rRNA removal kit (Epicentre). rRNA-depleted RNA was quantitated using a Qubit Fluorometer (Life Technologies) and approximately 15 ng of rRNA-depleted RNA was used to generate sequencing libraries using the ScriptSeq™ v2 kit (Epicentre) as described by the manufacturer's protocol. The ScriptSeq workflow is outlined in Fig. 7. Thirteen rounds of PCR amplification of the finished sequencing libraries were performed,

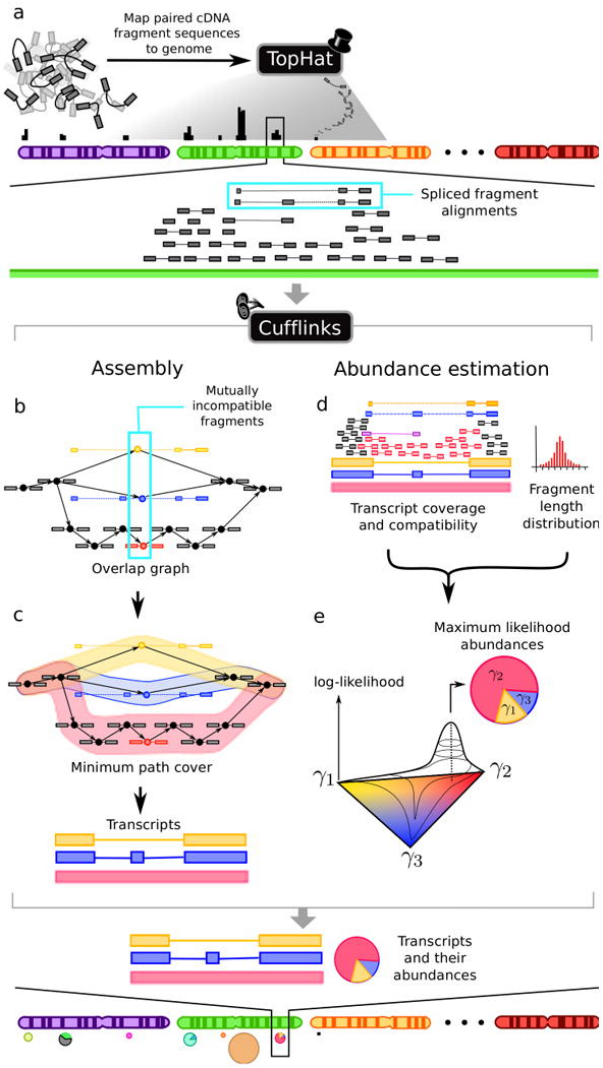

Figure 8. Schematic of TopHat alignment and Cufflinks transcript assembly[24]. **a)** TopHat aligns short reads to reference genome. **b-e)** Cufflinks generates a model of possible transcript isoforms at each locus based on aligned reads and uses minimum path and maximum likelihood algorithms to estimate isoform abundance.

and the amplified libraries were size-selected and purified using Agencourt AMPure XP beads (Beckman-Coulter) and quantitated and characterized *via* Qubit Fluorometer and Fragment Analyzer™ capillary electrophoresis (Advanced Analytical). Each of the six libraries was combined in equimolar amounts to a total of 20 nM and submitted for single-end, 101-base pair sequencing on the Illumina Hi-Seq 2000 massively parallel sequencer at the University of Oregon Genomics Core Facility. An analogous procedure was used with cells treated with none (two replicates), 10 μM, 20 μM, 40 μM, 50 μM, 60 μM, and 80 μM pentamidine. While the pentamidine RNA-Seq experiment did not contain true replicates in the treated samples, gene expression between the closest paired dosages (10 and 20 μM, 40 and 40 μM, and 60 and 80 μM) were found to correlate well with each other and were thus used as replicates for differential analysis (Appendix Fig. A7).

Following sequencing, raw sequencing reads (approximately 25 million for each sample) were aligned to the human genome using the splicing-aware aligner TopHat2, transcript models of aligned reads were assembled using Cufflinks, and differential analysis between samples was performed using Cuffdiff2 (Fig. 8)[23,24,25]. The previously mentioned programs were run using University of Oregon's Applied Computational Instrument for Scientific Synthesis (ACISS) supercomputing cluster. Aligned reads were visualized using Integrative Genomics Viewer (IGV)[26]. Differential analysis data was explored using the CummeRbund package for R and parsed and analyzed using the Bash, R, and Python programming environments. Differential splicing analysis was performed using Mixtures of Isoforms (MISO)[27]. Motif enrichment analysis was performed using Hypergeometric Optimization of Motif Enrichment (HOMER)[28].

RESULTS

Actinomycin D and pentamidine have different effects on DM1 mis-splicing events

In order to determine if actinomycin and pentamidine have additive or synergistic effects on MBNL1-regulated splicing events, actinomycin was titrated into HeLa cells expressing a reporter transcript and 960 CUG repeats (CUG_{960}), as described in Fig. 6, both alone and in combination with 20 μM pentamidine. By comparing marginal splicing rescue of actinomycin alone and 20 μM pentamidine alone with the corresponding combinatory treatment, additive or synergistic effects may be deduced. The effects of actinomycin and pentamidine on alternative splicing of three reporter transcripts (insulin receptor, cardiac troponin T, and MBNL1) are shown in Fig. 9. Both

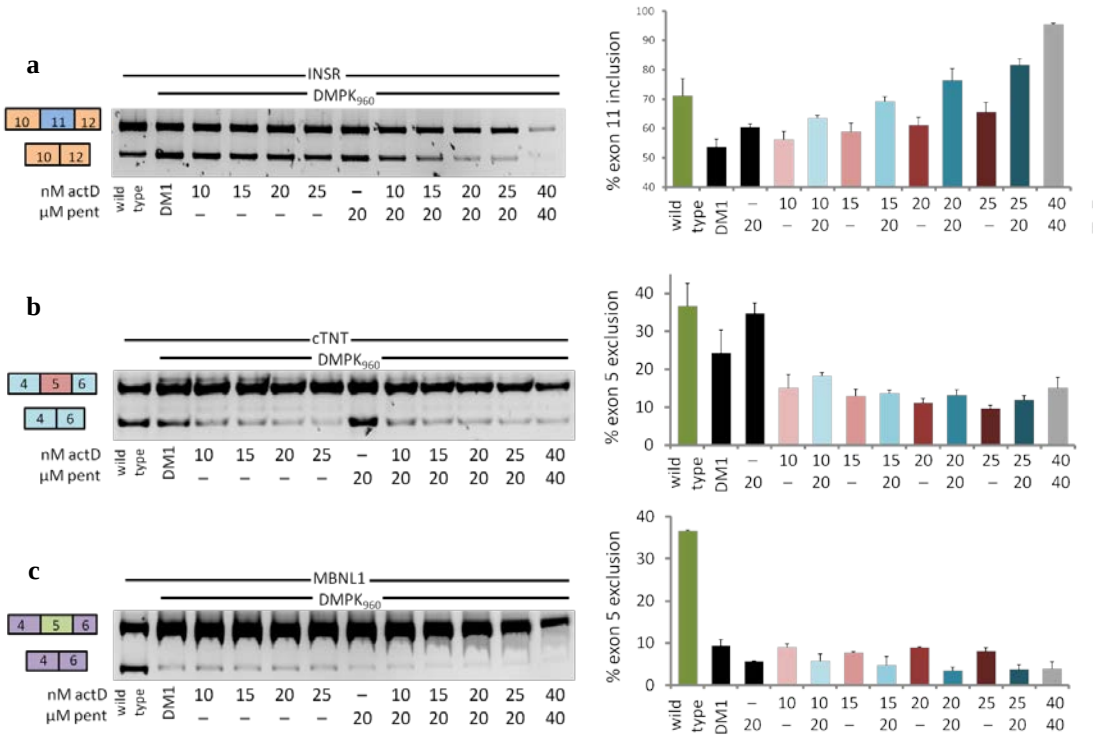

Figure 9. Effects of actinomycin and pentamidine treatment (individual and combinatorial) on DM1 mis-splicing events. Actinomycin (actD) and pentamidine (pent) were titrated into cells (alone or in combination) expressing a reporter transcript and 960 CUG repeats ($DMPK_{960}$). Representative gels with resolved alternatively spliced reporter products shown along with quantification of replicate experiments (**a**, **b**, and **c**). Red bars depict actD-only treatments and blue bars depict actD plus 20 μM pent treatments. Wild-type and DM1 samples received no drug treatment.

actinomycin and pentamidine rescue wild-type splicing in INSR (Fig. 9a), and this splicing rescue has an EC_{50} (effective concentration at which half of full rescue is achieved) of approximately 20-25 nM actinomycin or a combination of 10 nM actinomycin and 20 μM pentamidine. Actinomycin and pentamidine appear to have additive effects at low concentrations, while at high concentrations (25 nM actD and μM pent) more splicing rescue is observed than would be expected by summing the marginal splicing rescues achieved with separate 20 nM actinomycin and 20 μM pentamidine treatments (Appendix Fig. A8). Additionally, at high combinatorial doses significant over-rescue is observed, wherein exon inclusion levels far exceed that of the wild-type state.

Despite the similar effects of actinomycin and pentamidine on the INSR splicing event, the two drugs have opposite effects on cTNT splicing (Fig 9b). While pentamidine rescues strongly on its own at 20 μM, actinomycin drives splicing levels towards a more diseased-like state, decreasing exon 5 exclusion away from wild-type levels and less than DM1 levels. This mis-rescue is modulated by combinatorial treatment with pentamidine, but actinomycin appears to mask much of the rescuing effects of pentamidine (supplementary Fig. A8). In the alternative splicing event found in the MBNL1 transcript, actinomycin D has no apparent effects, neither significantly rescuing nor mis-rescuing splicing, while pentamidine weakly mis-rescues MBNL1 splicing (Fig. 9c).

Many splicing factors are differentially expressed following drug treatment

Although actinomycin D strongly decreases the levels of CUG_{960} RNA by 5 nM as observed in northern blots (Siboni, R., unpublished data), it has inconsistent effects

on MBNL1-regulated splicing events, suggesting that the drug is altering MBNL1-regulated splicing *via* processes other than decreasing CUG_{960} RNA levels and increasing free MBNL1 levels. Pentamidine likewise has inconsistent effects on

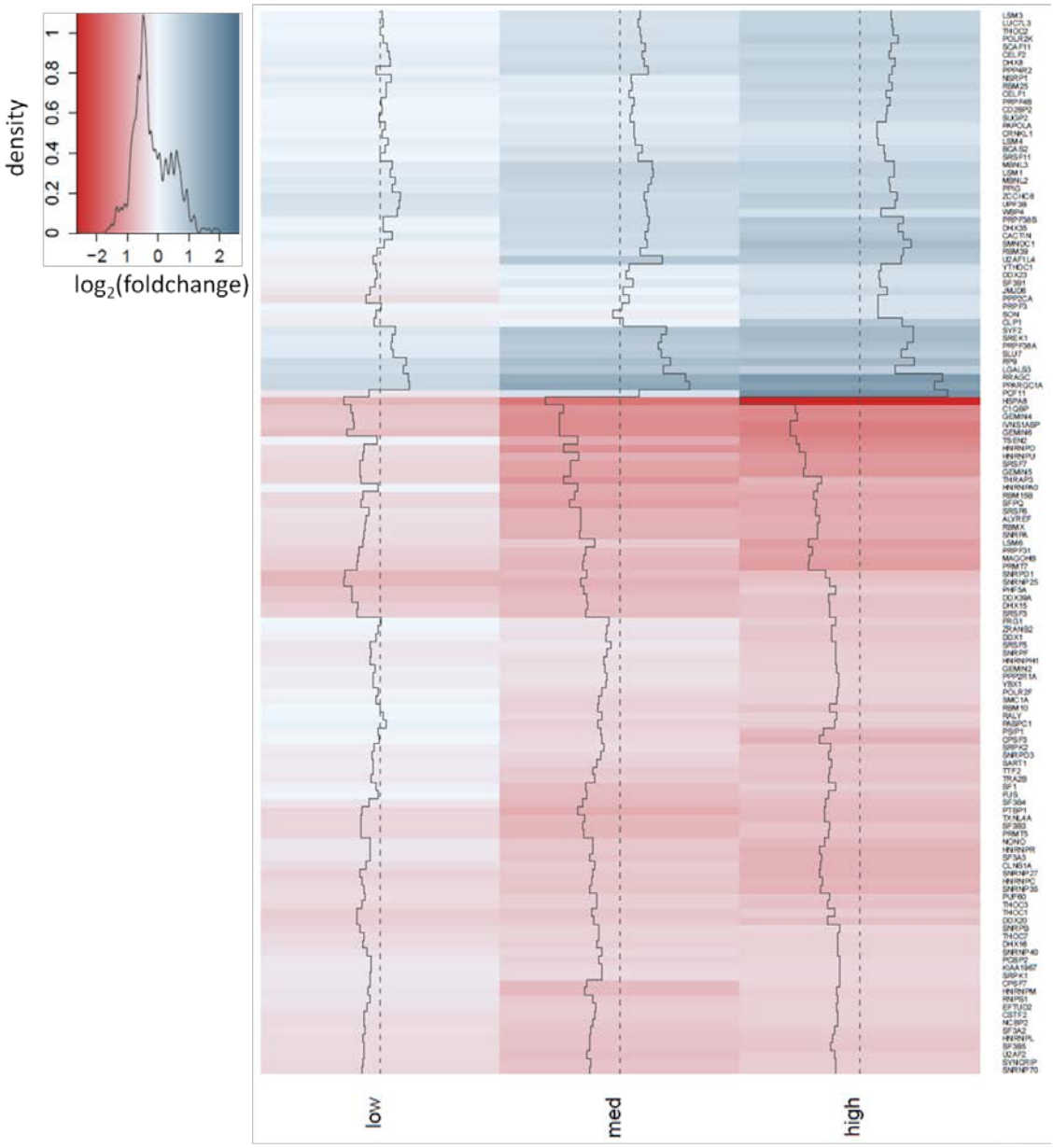

Fig. 10. Clustered heatmap of $\log_2$(foldchange) in expression at low (10, 20 μM), medium (40, 50 μM), and high (60, 80 μM) pentamidine compared to no drug treatment of 135 significantly differentially expressed ($q < 0.05$ between none and high) genes associated with the gene ontology term "RNA splicing" (GO:0008380).

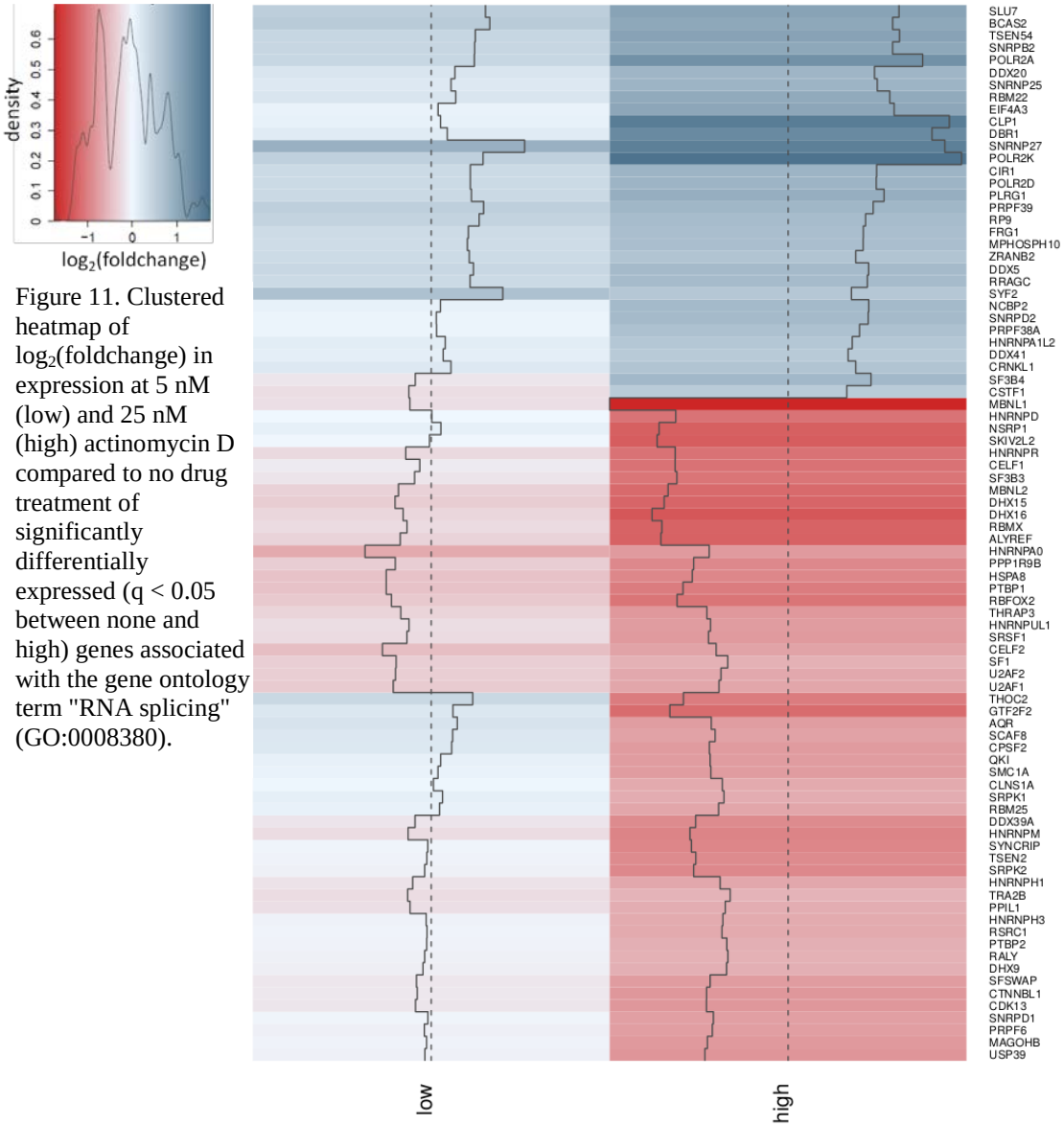

Figure 11. Clustered heatmap of $\log_2$(foldchange) in expression at 5 nM (low) and 25 nM (high) actinomycin D compared to no drug treatment of significantly differentially expressed ($q < 0.05$ between none and high) genes associated with the gene ontology term "RNA splicing" (GO:0008380).

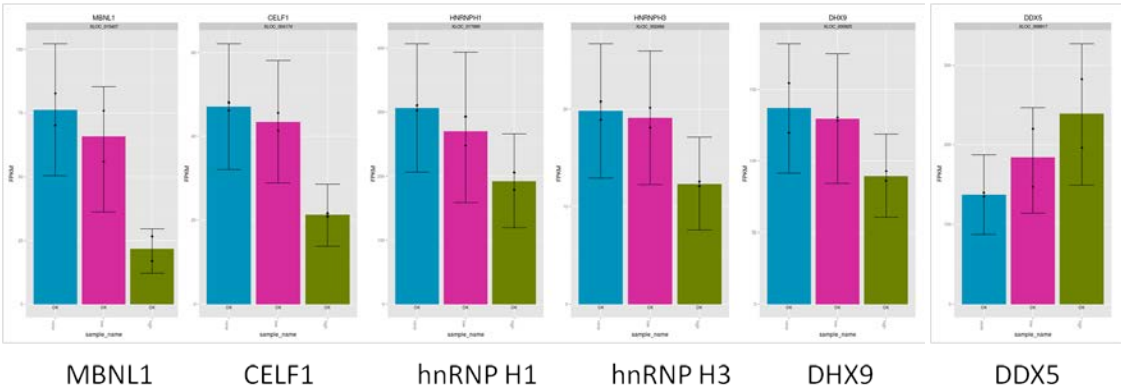

Figure 12. FPKM (fragments mapped per kilobase of gene exon model per million mapped reads) of splicing factors that are implicated in DM1 and are significantly differentially expressed at 25 nM actinomycin D. Blue bars depict no actinomycin, magenta bars depict 5 nM (low) actinomycin, and green bars depict 25 nM (high) actinomycin. Each dot represents the value of a replicate experiment.

MBNL1-regulated splicing events, rescuing INSR and cTNT but slightly mis-rescuing MBNL1. An RNA-Seq experiment was performed to examine pentamidine- and actinomycin-induced global changes in transcriptomic profiles of CUG_{960}-expressing cells that may help to explain the observed changes in alternative splicing. At 60-80 μM pentamidine, 135 genes with the associated Gene Ontology term "RNA splicing" (GO:0008380) are significantly differentially expressed ($q < 0.05$) (Fig. 10), while at 25 nM actinomycin, 86 genes such genes are differentially expressed (Fig. 11). Pentamidine induces an approximately two-fold increase in MBNL2 and MBNL3, which are thought to be able to act as alternative splicing regulators redundant with MBNL1 and may account for some of the observed splicing rescue[29]. Most striking in actinomycin-treated cells is the observation that the gene that decreases the greatest in expression is MBNL1 at an approximately four-fold decrease (Fig. 12), which is accompanied by an approximately 1.5-fold decrease in MBNL2. Many proteins involved in the basal splicing complex are also changed in expression, with U2AF1, U2AF2, SF1, and SF3B3 decreasing and SF3B4 increasing in expression. U2AF1 and U2AF2 are together responsible for binding the intronic polypyrimidine tract, while SF1 is responsible for binding the intronic branch-point sequence, and all are required for canonical assembly of the spliceosome. SF3B3 and SF3B4 are subunits of the U2 snRNP, a component of the major spliceosome, and SNRNP25 is a component of the U12 minor spliceosome complex.

Of the 86 differentially expressed genes with the associated Gene Ontology "RNA splicing", several have been previously implicated in DM1-misplicing: CELF1, hnRNP H1, hnRNP H3, DHX9, and DDX5[30]. CELF1 and hnRNP H1 are thought to

generally antagonize MBNL1-promoted splicing outcomes (*i.e.* if MBNL1 promotes exon inclusion in a certain splicing event, these factors will promote exclusion)[30]. These factors were previously identified as increasing two- to three-fold in DM1 myoblasts expressing expanded CUG repeats compared to those expressing normal CUG repeat tracts, and upon treatment with 25 nM actinomycin all but one significantly decreases in expression (Fig. 12)[30]. PTBP1 has been previously identified as co-regulating several alternative splicing events with MBNL1, and both PTBP1 and PTBP2 are decreased in expression following actinomycin treatment[31].

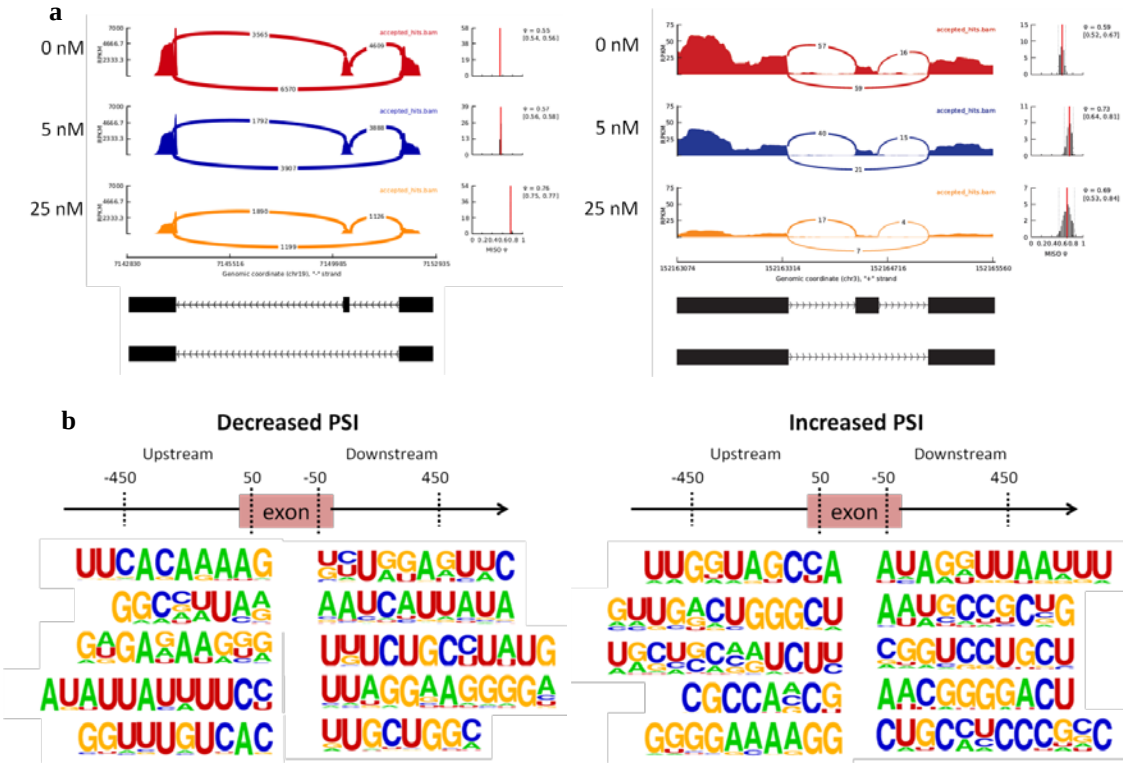

Figure 13. Changes in cassette exon alternative splicing following actinomycin D treatment. **a)** Sashimi plots of previously examined splicing events. The INSR event represents the spliced transcript from the transfected reporter, while the MBNL1 event is endogenous. Connecting arcs depict splice junction-spanning reads, and estimated PSI (percent spliced in) distributions are shown at the right. **b)** Cassette exon splicing events that experienced significant change (Bayes factor greater than 20 and change in PSI of greater than 0.05) between 0 and 25 nM actinomycin were separated into groups by either decreasing (424 events) or increasing (309 events) PSI. Upstream and downstream sequences flanking these exons in a window of 50 bp into the exon and 450 bp into the intron were subjected to motif enrichment analysis using HOMER with a background set of analogous sequences from 3000 randomly chosen cassette exon splicing events. P-values for motifs shown range from 10^{-6} to 10^{-14}.

Actinomycin D causes widespread changes in cassette exon alternative splicing

Alternative splicing of the transfected INSR reporter, as detected in the RNA-Seq data by the program MISO, is observed to increase in exon inclusion at 25 nM actinomycin in concordance with previously performed splicing experiments with this reporter, while alternative splicing of the endogenous MBNL1 transcript likewise increases slightly in exon inclusion at 25 nM actinomycin (Fig 13a). However, with a large number of proteins involved in RNA splicing changed in expression at 25 nM actinomycin, one may expect pronounced changes in alternative splicing across the transcriptome that are not limited to MBNL1-regulated events. Indeed, at 25 nM actinomycin 424 annotated cassette exon alternative splicing events were significantly decreased in percent exon inclusion (PSI, percent spliced in) and 309 events were significantly increased in PSI as identified by MISO. Since many regulators of alternative splicing are RNA-binding proteins (including MBNL1, CELF1, PTBP1, and hnRNPs) whose specificity is at least in part governed by sequence-specific RNA binding, it may be expected that the RNA sequences flanking these differentially spliced exons are enriched in regulatory sequences that correspond to the binding sites of differentially expressed splicing factors. As such, the sequences extending from 50 bp inside these cassette exons to 450 bp into the upstream or downstream introns were subjected to motif enrichment analysis with analogous sequences from 3000 randomly chosen annotated cassette exon splicing events as a background model using HOMER. Selected enriched motifs from each group of sequences are shown in Fig. 13b.

Of note, each group of sequences is enriched for motifs containing G/A runs, and all but one contains the sequence GGGA. This GGGA motif forms the core of the

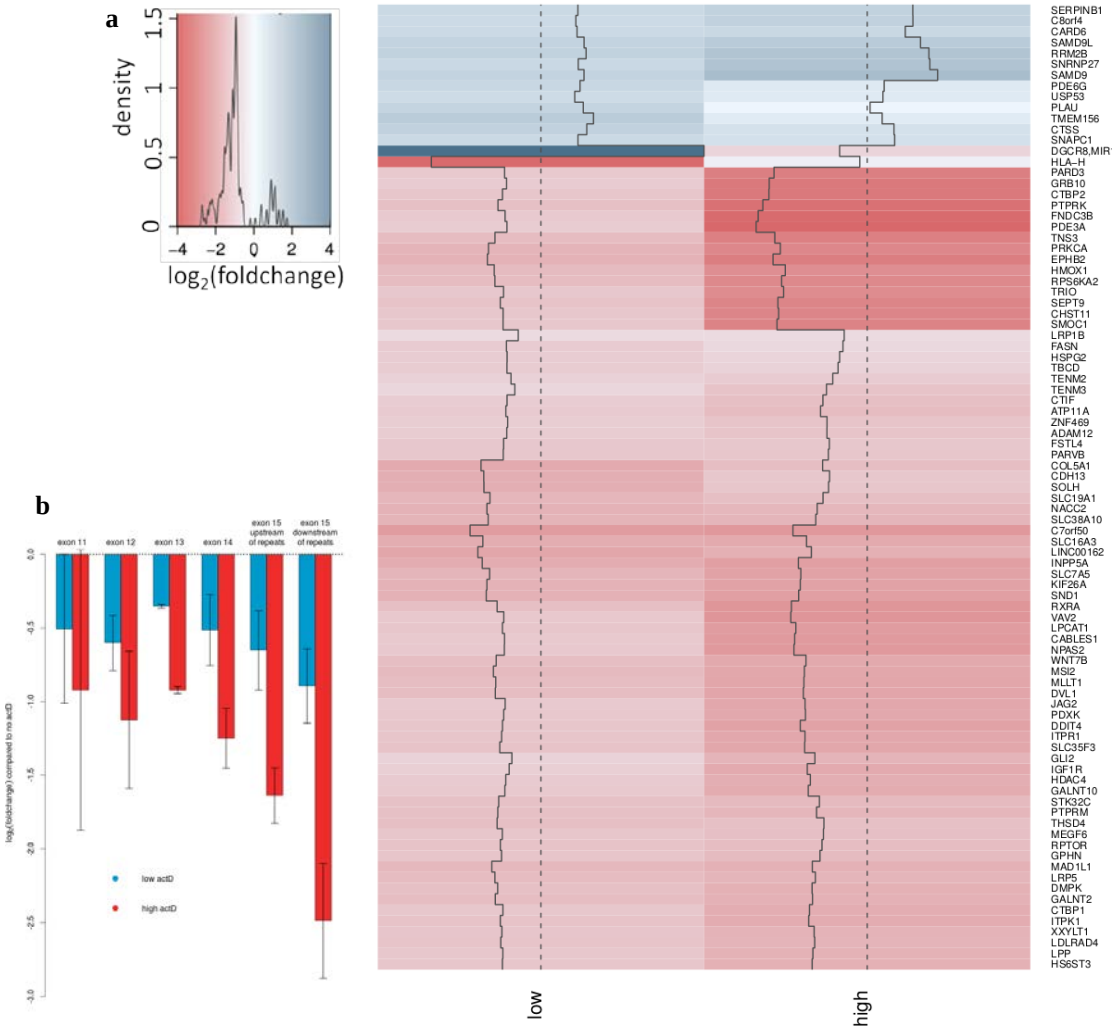

Figure 14. Significantly differentially expressed genes ($q < 0.05$) between 0 and 5 nM actinomycin. **a)** Clustered heatmap depicting $\log_2$(foldchange) in expression of 89 genes between none and either 5 nM (low) or 25 nM (high) actinomycin. **b)** $\log_2$(foldchange) in normalized aligned read counts between none and low (blue) or high (red) actinomycin across the transfected exons of DMPK.

consensus binding sequence of hnRNP H1, which is decreased two-fold at 25 nM actinomycin[32]. MBNL1's consensus sequence YGCY appears frequently in downstream sequences of exons that decreased in PSI and upstream sequences of exons that increased in PSI. Other common motifs are A/U-rich sequences and U/G-rich sequences.

Specificity of actinomycin D

Although actinomycin does reduce the levels of CUG_{960} RNA, it is clear that it also affects the expression of many other genes at the same time. At 5 nM actinomycin,

only 89 genes are significantly ($q < 0.05$) differentially expressed, and of these 74 genes are decreased in expression, suggesting that many of these decreases in expression are directly caused by actinomycin-mediated transcriptional inhibition (Fig. 14a). Of these 89 genes, the most enriched Gene Ontology is "regulation of cell proliferation" (GO:0042127) with 17/89 genes and p-value of 2.3×10^{-6}, and the most enriched Swiss-Prot Protein Information Resource keyword is "phosphoprotein" with 53/89 genes and p-value of 2.1×10^{-6} as calculated by the DAVID Functional Annotation tool, and as such at least some of the differential expression observed in this set is likely due to cellular signaling and subsequent activation of regulatory networks in response to the drug rather than the direct binding of actinomycin to DNA in these genes and subsequent inhibition of transcription[33].

Yet these 74 genes that decreased in expression at 5 nM actinomycin are most sensitive to the drug, and they offer insight to the sequence specificity of actinomycin binding and inhibiting transcription. Viewing aligned reads in the IGV genome browser reveals a shift in the distribution of reads in early gene regions of several genes, wherein increasing actinomycin trends with increasing reads in the first approximately 10 kb of a gene, including intronic regions (Appendix Fig. A9). To quantify this trend, the average $\log_2$(foldchange) in read counts at each base position from the annotated gene start to 50 kb downstream between none and either low or high actinomycin samples was calculated (Fig. 15a). On average, high actinomycin causes increased reads in early gene regions, between approximately 1 and 10 kb downstream from the gene start site. The inset in Fig. 15a shows the pseudo-periodic nature of these changes in read counts, and early peaks are spaced apart by approximately 150 bp. A single-sample t-test shows

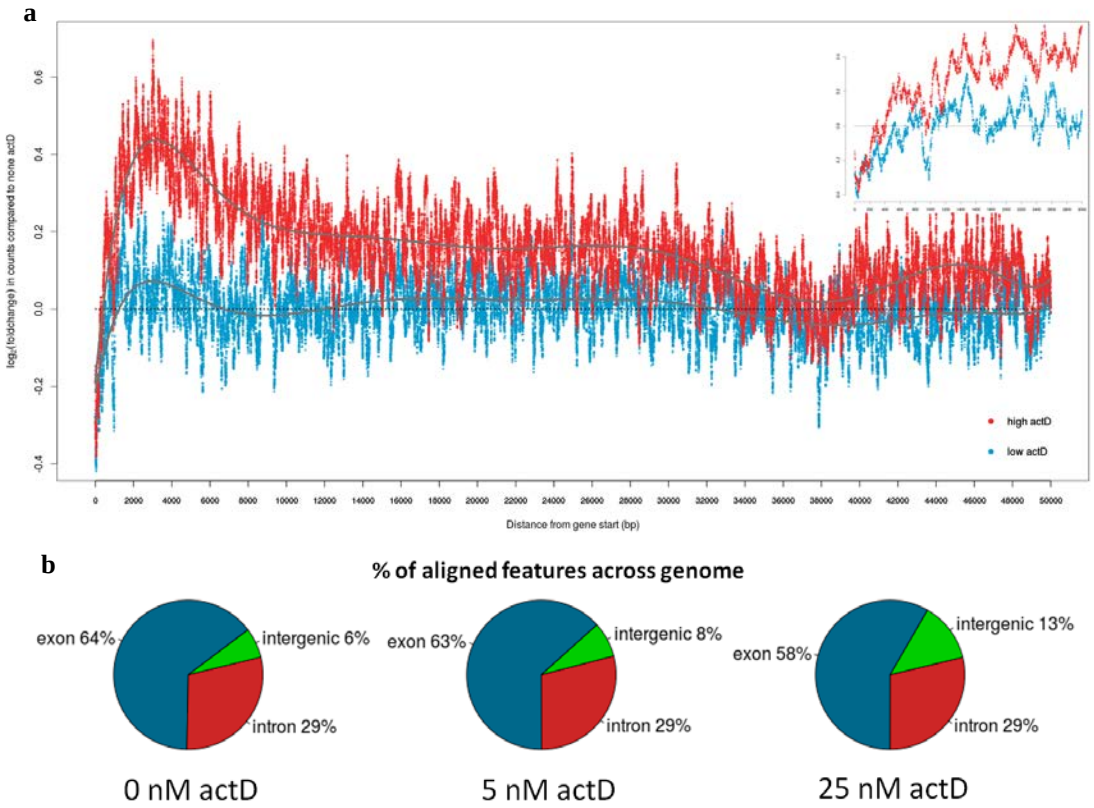

Figure 15. Actinomycin changes the distribution of aligned reads in early gene sequences of the 74 significantly decreased genes at low actinomycin dosage. **a)** The $\log_2$(foldchange) in average aligned read counts at each base position from the annotated gene start to 50 kb downstream shows that at 25 nM actinomycin (red) causes an increase in reads in early gene regions. The inset shows a magnification of the region from 0 to 3 kb, depicting the periodic nature of the change in read counts. Dotted line: zero change in read counts; gray curves: lines of best fit using a 9th-degree polynomial to aid visualization of trajectories. **b)** Actinomycin slightly decreases the number of exonic features and increases the number of intergenic features in aligned reads across the genome. Intronic features remain constant with respect to total aligned features but increase slightly relative to exonic features.

that the change in read counts between 2 and 5 kb are significantly different from 0 with p-values ranging from 5.2×10^{-9} to 0.038. Since this region is intronic for many of the genes in this set, it could be proposed that these increases in aligned reads are a result of increased intron retention due to spliceosomal deficiency. This is an attractive concept, considering the number of components of the spliceosome that are differentially expressed at 25 nM actinomycin, yet the portion of aligned intronic features across the genome only increases 1.1-fold relative to the portion of aligned exonic features (Fig.

15b) and as such this is likely not the source of the observed accumulation of nascent transcripts in early gene regions.

Increased read counts in these early gene regions suggest that there may be certain sequence motifs in these regions to which actinomycin binds strongly, causing elongating polymerases to pause and – along with their nascent transcripts – accumulate along this region. The *DMPK* transcript decreases about two-fold in expression at 5 nM actinomycin (Fig. 14a) and the number of detected transcripts decreases most dramatically downstream of the CTG repeat tract in exon 15 compared to the other exonic portions of the transfected construct (Fig. 14b), suggesting that the expanded CTG repeat tract may be particularly sensitive to actinomycin, and the CTG motif may be such a determinant of decreased expression in early gene regions. Sequences flanking the gene start site for these 74 genes were probed for occurrences of simple motifs, which were then correlated with changes in expression. As shown in Fig. 16, counts of

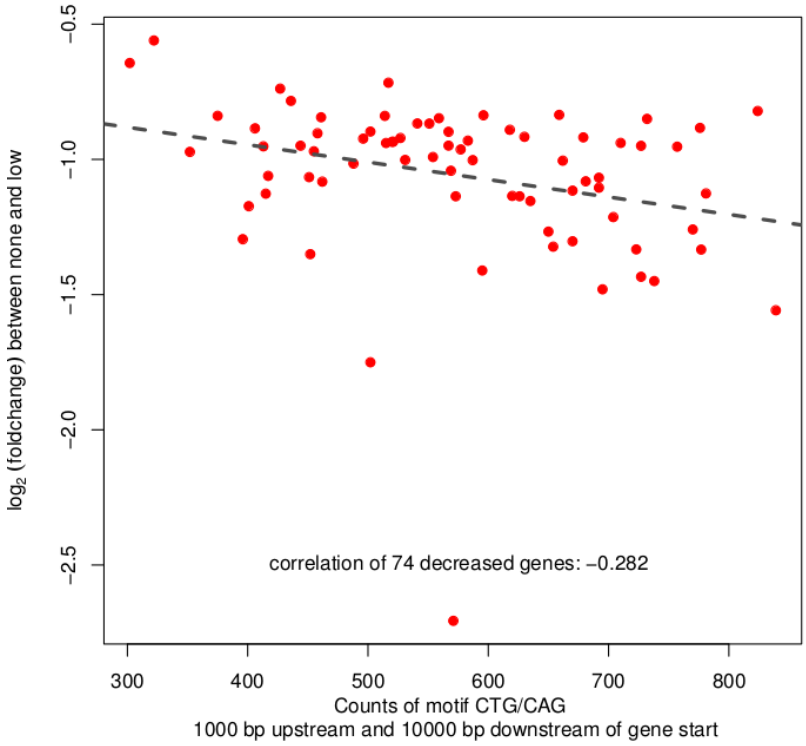

Figure 16. CTG motifs correlate with decreased expression following actinomycin treatment. Each red dot represents a gene that decreased significantly between none and 5 nM (low) actinomycin.

CTG and CAG (representing CTG on the opposite strand) in a region extending from 1 kb upstream to 10 kb downstream of gene start sites correlate negatively (r = -0.282) with $\log_2$(foldchange), suggesting some specificity for sequences with high CTG content. This correlation has greater magnitude than the average absolute values of correlation in this region across all 3-mers or 4-mers, which is 0.178 and 0.166, respectively (Appendix Fig. A10).

Actinomycin and pentamidine both bind DNA in the minor groove (actinomycin furthermore intercalates DNA when bound), preferentially binding GpC dinucleotides and A/T-rich tracts, respectively. G/C-rich sequences are generally associated with wide minor grooves while A/T-rich sequences are generally associated with narrow minor grooves, and for this reason tetranucleotide sequences with approximated average minor groove widths were counted in a region 1 kb upstream to 10 kb downstream of annotated gene start sites of all genes that decreased significantly at low drug dosage, and these counts were correlated with $\log_2$(foldchange) and plotted in Fig. 17[34]. These correlations were then compared with respective average minor groove widths, producing an overall correlation of r = 0.650 for pentamidine and r = -0.439 for actinomycin. These observed trends suggest that pentamidine and actinomycin may recognize DNA structure in addition to DNA sequence: pentamidine may bind strongly to narrow minor grooves, more severely inhibiting transcription of genes with many of these binding sites (and producing a positive overall correlation of expression with minor groove width) while actinomycin may bind more strongly to wide minor grooves to more severely inhibit transcription (producing a negative overall correlation).

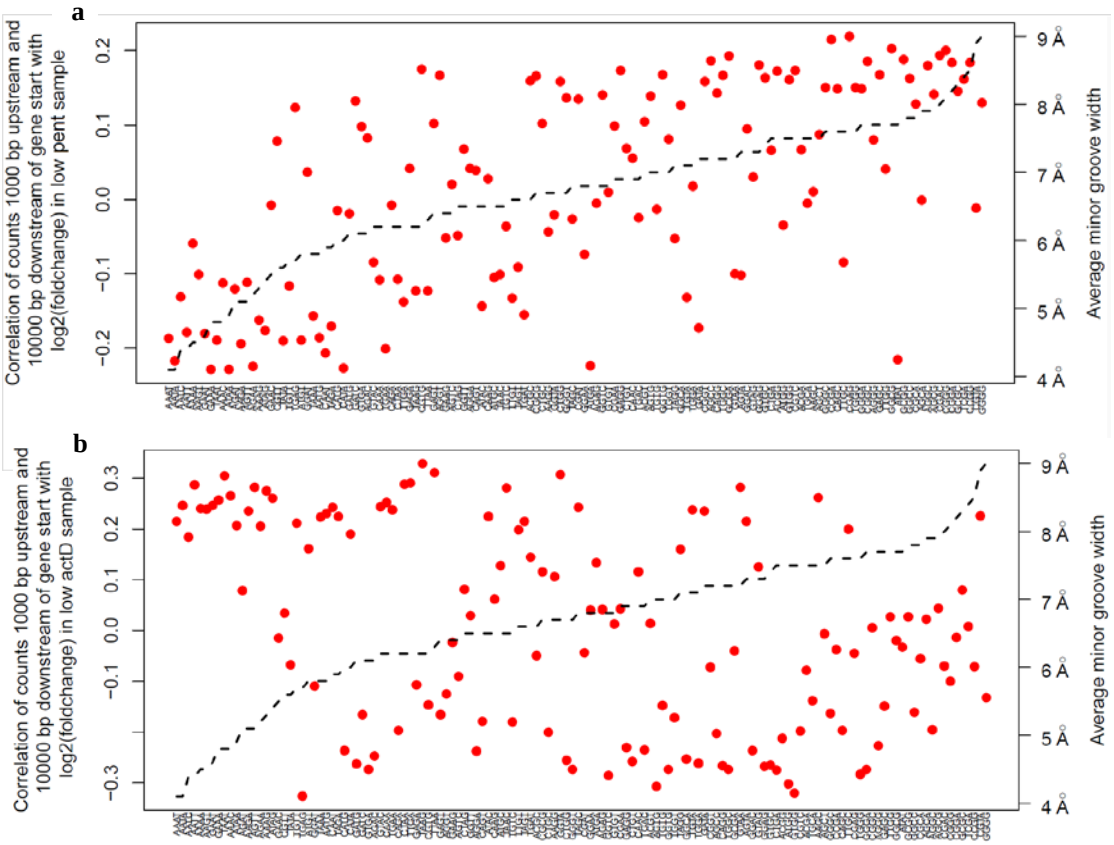

Figure 17. Average minor groove width may be a predictor of small molecule-gene interactions. **a)** Correlations of counts of tetranucleotide sequences in early gene regions with respective average minor groove widths with $\log_2$(foldchange) in expression of 641 genes that decreased significantly ($q < 0.05$) in expression at low pentamidine dosage. This trend has a correlation of $r = 0.650$ with average minor groove widths. **b)** Analogous correlations with 74 genes that decreased significantly at low actinomycin dosage, with $r = -0.439$ for this trend of correlations with average minor groove widths. Dashed line represents average minor groove width as calculated by Rohs, *et al.*[34].

At higher dosages, actinomycin appears to have very widespread effects on differential expression across the genome, as 3947 genes are significantly ($q < 0.05$) differentially expressed at 25 nM actinomycin. The most enriched gene ontology as identified by DAVID is "regulation of transcription" (GO:0006355) with p-value less than 10^{-24}, and as such a large amount of the differential expression observed at this dosage is likely due to activation of signaling pathways and regulatory networks. Indeed, promoter regions of these genes in a region extending 350 bp upstream to 50 bp downstream of the transcription start site are enriched in sequences that correspond well with previously established DNA binding motifs of several ubiquitous transcription

factors, including p53 (TP53) and members of the SMAD, STAT, and IRF (interferon-regulatory factor) families (Fig. 18).

DISCUSSION

Actinomycin likely affects alternative splicing *via* differential expression of splicing factors

As observed in the *in cellulo* splicing results (Fig. 9), actinomycin has varying effects on MBNL1-regulated alternative splicing events – rescuing wild-type INSR splicing, but mis-rescuing cTNT splicing – despite its capacity to strongly decrease the abundance of CUG_{960} RNA at as little as 5 nM. As such, actinomycin must be affecting alternative splicing through cellular processes other than MBNL1 sequestration on long CUG repeats, and indeed RNA-Seq shows that 86 genes with the gene ontology "RNA splicing" are differentially expressed at 25 nM actinomycin. Of these genes, MBNL1 decreases the most severely, and MBNL2, which is able to regulate alternative splicing events similarly to MBNL1, decreases in expression as well. Conversely, four splicing regulators that have been found to be upregulated two- to three-fold in DM1 myoblasts were decreased in expression

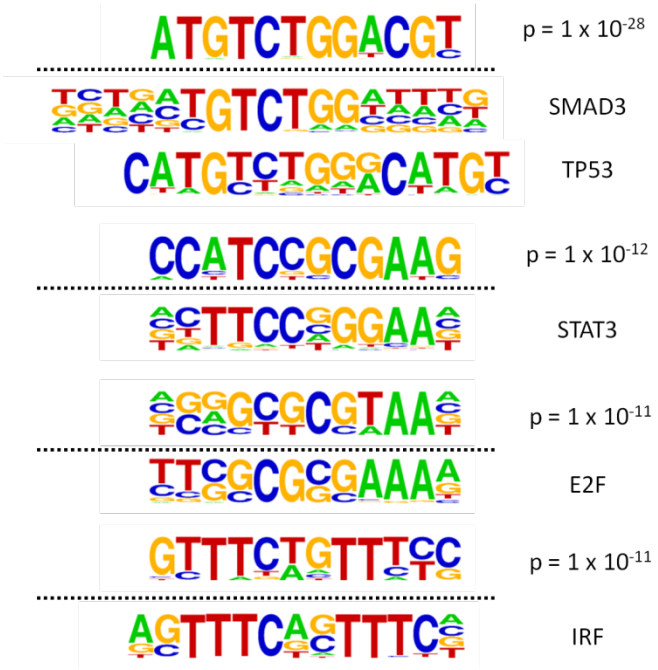

Figure 18. Enriched motifs (above dashed lines) and p-values of enrichment in promoter regions 350 bp upstream to 50 bp downstream of transcription start sites of the 3947 differentially expressed genes at 25 nM actinomycin. Previously determined binding sequences of transcription factors are shown below the dashed lines of their respective enriched motifs.

approximately two-fold at 25 nM actinomycin. Together, these results help to explain actinomycin's inconsistent ability to rescue wild-type alternative splicing: the INSR, cTNT, and MBNL1 splicing events are all regulated by MBNL1, yet each is also likely regulated by a suite of other factors that may not regulate the other events. CELF1, for example, promotes exclusion of INSR exon 11[35]. If inclusion of this exon is strongly negatively regulated by CELF1 and only weakly positively regulated by MBNL1, then decreases in both CELF1 and MBNL1 may result in greater exon inclusion as is observed. In addition, several MBNL1 splicing events, including INSR and cTNT, show differing sensitivity to MBNL1 levels in an MBNL1-inducible cell culture system (Wagner, S., unpublished data). These experiments reveal that the INSR splicing event is marginally sensitive to MBNL1 levels, requiring only a relatively small amount of MBNL1 in order to achieve maximum exon inclusion, whereas cTNT is more sensitive to MBNL1 levels and requires much more induced MBNL1 in order to achieve maximum exon exclusion. These trends support the observed rescue of INSR splicing and mis-rescue of cTNT splicing following actinomycin-mediated MBNL1 depletion, assuming that INSR exon inclusion is additionally promoted by changing levels of another splicing regulator such as the loss of CELF1. In this fashion, decreases in MBNL1 levels in combination with differential expression of many other splicing regulators will affect many different splicing events in a myriad of different – and difficult to predict – ways.

In addition to revealing the large number of splicing factors that are differentially expressed at 25 nM actinomycin, RNA-Seq also exposed transcriptome-wide changes in alternative splicing, including 424 cassette exon splicing events that

decreased in PSI by at least 0.05 and 309 such events that increased in PSI. Differential splicing observed in these events is likely a result of differential expression of splicing factors, and as a result the RNA sequences flanking these cassette exons contain some RNA targets of enriched or depleted splicing factors. MBNL1's consensus binding sequence, YGCY, is found in downstream sequences of exons that decreased in PSI and in upstream sequences of exons that increased in PSI (Fig. 13b). Additionally, the motif UUGC, which occurs in downstream sequences of exons that decreased in PSI, closely matches that previously found as MBNL1's target sequence in an *in vitro* high-throughput binding and sequencing experiment[36]. This corroborates the general trend that MBNL1 promotes inclusion when bound downstream of an exon and exclusion when bound upstream of an exon, since following MBNL1 depletion at 25 nM actinomycin the opposite trends in PSI are observed, assuming that these enriched MBNL1-binding motifs represent true, regulatory MBNL1 binding sites. Likewise, the motifs UGU and UGUGU are found throughout all groups of sequences and may represent targets of CELF1, which has been reported to bind UGU motifs[37,38]. CELF1 antagonizes MBNL1-regulated splicing in a number of transcripts, and the approximately three-fold decrease in CELF1 at 25 nM actinomycin may help to diminish some of the splicing dysregulation caused by a depletion of MBNL1 at this dosage[39].

At 25 nM actinomycin, ten hnRNPs – including PTBP1 (hnRNP I) and RALY (an hnRNP C homolog) are differentially expressed, and all but one are decreased in expression[40]. hnRNPs are generally expressed ubiquitously and at high levels in human tissues, binding to nascent transcripts as part of an hnRNP complex that mediates a

diverse range of processes, including RNA splicing, export, and stability[40,41]. hnRNP D binds to A/U-rich sequences and is decreased approximately two-fold at 25 nM actinomycin (Fig. 11), and A/U-rich motifs appear in all groups of sequences except for upstream sequences of events that increased in PSI[41]. hnRNP H, which is decreased approximately two-fold at 25 nM actinomycin, is known to bind a GGGA core, and CLIP-Seq studies have identified G/A runs as strong binding sequences as well[40,42]. These G/A runs are found in all groups of enriched motifs, suggesting that hnRNP H may regulate exon inclusion in a complex fashion, independent of whether it binds up- or downstream of the exon. Another splicing factor that decreases approximately two-fold in expression, hnRNP M, binds U/G-rich sequences, and these may be found in upstream sequences of exons that decreased in PSI (GGUUUGU) as well as scattered throughout the other groups of sequences[42]. Since hnRNPs are involved in many different aspects of RNA processing other than splicing, it may be expected that these processes, such as polyadenylation, RNA stability, RNA export, and RNA localization are likewise affected by actinomycin treatment. In the case of polyadenylation, the number of reads containing poly(A) runs of 100 or greater consecutive As increases approximately 1.5-fold from 0 to 25 nM actinomycin, although it is not possible to determine if this observed increase truly reflects retained competency of the polyadenylation complex or if it is due to other effects, such as a regulatory shift towards polyadenylated transcripts. It is worthwhile to note, however, that two components of the polyadenylation complex, CSTF1 and CPSF2, increase and decrease respectively in expression at 25 nM actinomycin.

CTG motifs, wide minor groove DNA motifs, and increased early gene transcripts are associated with decreased expression upon actinomycin treatment

While 733 cassette exon alternative splicing events are differentially spliced at 25 nM actinomycin, only 237 such events are differentially spliced at 5 nM actinomycin. Likewise, at 25 nM actinomycin 3947 genes are differentially expressed while only 89 genes are differentially expressed at 5 nM actinomycin. The DMPK transcript containing 960 CUG repeats is particularly sensitive to 5 nM actinomycin, decreasing in expression approximately two-fold and decreasingly mostly severely in the region just downstream of the CTG repeat tract (Fig. 14b). Additionally, occurrences of CTG motifs in early gene regions correlate with greater decreases in expression at 5 nM (Fig. 16). Together, these results suggest that actinomycin may bind CTGs particularly strongly to preferentially inhibit transcription of a small number of genes at low concentrations. CTGs and other G/C-rich sequences generally confer greater width to the DNA minor groove, and actinomycin is classically known to bind preferentially at GpC dinucleotides. Indeed, sequences with narrow minor grooves typically correlate positively with expression of the 74 decreased genes at low actinomycin while sequences with wide minor grooves correlate negatively (Fig. 17b). In effect, genes containing greater numbers of narrow minor groove sequences were associated with less decrease in expression while genes containing greater numbers of wide minor groove sequences were associated with greater decreases in expression, which may be caused by actinomycin preferentially binding to sequences with wide minor grooves. In contrast, the opposite trend is observed regarding the 642 genes that decreased significantly at low pentamidine dosage, suggesting that pentamidine preferentially binds and inhibits the transcription of narrow minor grooves (Fig. 17a).

The 74 genes that significantly decreased in expression at 5 nM actinomycin exhibit a distinctive pattern when examining base-by-base changes in expression in early gene regions: at 25 nM actinomycin, the number of mapped reads decreases from approximately 0 to 100 bp downstream of the annotated gene start site and then increases significantly from 1 to 6 kb with a peak of about 1.5-fold increase in mapped reads, eventually returning to near zero change in expression at about 35 kb downstream of the gene start (Fig. 15a). An attenuated parallel of this trend is also observed at 5 nM actinomycin. This increase in mapped reads in early – and often intronic – gene regions implies increased polymerase pausing, causing a back-up of polymerases and increase of aligned reads in early gene regions despite an overall decrease in gene expression. This pausing would preclude splicing of the nascent transcript if the polymerases pause after passing a 5' splice site but before reaching a 3' splice site, and as such many of these regions of increased aligned reads fall in the gene's first intron (Appendix Fig. A9). Actinomycin has been shown to pause DNA polymerases *in vitro* by binding single-stranded DNA with sequence specificity and forming heat-stable complexes, and high-affinity binding to unwound DNA in transcription bubbles and inhibiting elongation is hypothesized to be actinomycin's biological mechanism of action[43,44].

The inset of Fig. 15a shows a magnification of the region from 0 to 3 kb downstream of the gene start, exposing the pseudo-periodic nature of changes in aligned reads: in the first 600 bp, each trough in the curve representing 5 nM actinomycin is separated by about 150 bp. While this may simply be an artifact in the data, this trend could potentially represent chromatin-dependent interactions of actinomycin with DNA, since about 146 bp of DNA are wound around each histone octamer and each

nucleosome is connected by about 20 bp of linker DNA in classical 10 nm euchromatin[45]. Since both histones and actinomycin are tight minor groove binders, it may be that actinomycin does not bind easily to DNA packaged into histones but rather binds readily to the histone-free linker DNA to cause a periodic pattern of polymerase pausing. A ChIP-Seq experiment with RNA polymerase II would reveal in much better detail any conspicuous patterns of polymerase pausing during elongation in the presence of actinomycin.

Increases in aligned reads in these early gene regions may be caused by increased intron retention across the transcriptome (as could be expected considering the observed differential expression of many components of the core spliceosome), but the portion of all intronic features in genome-wide aligned reads stays constant with respect to all aligned features and decreases only 1.1-fold with respect to aligned exonic features (Fig. 15b). Of note, the portion of intergenic features increases two-fold from 0 nM to 25 nM actinomycin. Further analysis will be necessary to characterize the nature of these actinomycin-inducible intergenic transcripts: they may be unannotated protein-coding genes, unannotated long non-coding RNAs (lncRNAs), or potentially products of bidirectional transcription initiation (as is observed in PRKCA in Appendix Fig. A9).

Actinomycin activates transcriptional regulatory pathways at higher doses

While 5 nM actinomycin does not appear to induce any widespread regulatory cascades, by 25 nM it is clear that signal transduction is driving a substantial portion of the differentially expressed genes. Actinomycin is a strong inducer of p53 signaling at as low as 10 nM, and p53's DNA target sequence is found enriched in promoters of genes that are differentially expressed at 25 nM actinomycin (Fig. 18)[46]. p53 is a

ubiquitous tumor-suppressor protein whose disruption is observed in a large number of cancers, and as such its activation may be beneficial in actinomycin's use as a chemotherapeutic. IRF transcription factors mediate many cellular immune responses to pathogens, the STAT protein family regulates cell growth and apoptosis in response to growth factors and cytokines, E2F transcription factors regulate the cell cycle and apoptosis, and SMAD proteins are involved in signal transduction across multiple signaling pathways[47]. DNA binding sequences of each of these transcription factor families are enriched in promoters of differentially expressed genes at 25 nM actinomycin, depicting the wide scope of signaling pathways that actinomycin activates. DNA-binding motifs for IRF, STAT, and E2F transcription factors are found in the promoter regions of *MBNL1* and *MBNL2*, offering potential regulatory mechanisms by which MBNL transcript levels are strongly downregulated upon treatment with 25 nM actinomycin.

APPENDIX

A Primer on the Central Dogma of Molecular Biology

As first stated by Francis Crick (of Watson and Crick fame) in 1958, the central dogma of molecular biology describes the flow of genetic information that underlies cellular processes: genetic information is stored in deoxyribonucleic acid (DNA) and copied out as messenger ribonucleic acid (mRNA) which then serves as a code for the production of proteins, which are the major functional units of organisms. The term dogma is, as admitted by Crick himself, an unfortunate misnomer given the endless amount of scientific evidence that affirms these biological processes. Furthermore, the central dogma does not consider many processes that have only in the past few decades become well-understood, such as alternative splicing of mRNA. In order to understand the molecular – and thus ultimately physiological – effects and implications of myotonic dystrophy and its potential small molecule therapeutics, we must briefly review nucleic acid structure and synthesis, RNA transcription, and mRNA splicing.

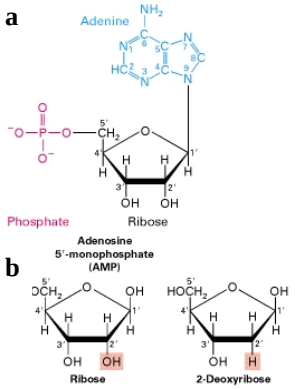

Figure A1. Structure of nucleic acid bases. **a)** Adenosine, a nucleotide base of RNA. **b)** RNA and DNA ribose rings differ by one oxygen at the 2' position.

Each type of nucleic acid (DNA and RNA) is composed of four major nucleotide bases that form a polymer of certain length and sequence (Fig. A1). DNA may contain adenosine (A), cytidine (C), guanine (G), and/or thymidine (T) bases, while RNA may contain A, C, G, and/or uridine (U) bases. The bases of nucleic acids are complementary, owing to the specific hydrogen bonds that may be formed between bases and similar width of complementary base pairs: A in one strand of nucleic acid

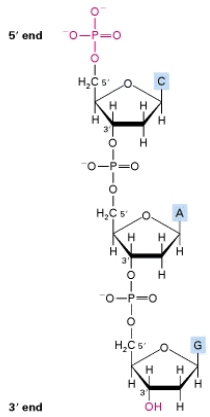

Figure A2. Three DNA nucleotides connected by two phospho-diester bonds.[1]

pairs with T or U on another strand (and *vice versa*) and similarly C pairs with G. Nucleic acids also have directionality. A single strand of nucleic acid has a 5' end and a 3' end (as designated by either a free 5' phosphate group or a free 3' -OH group as seen in Fig. A1). Between each free end of a nucleic acid strand, each nucleotide is connected to its neighbors through these aforementioned groups, forming phosphodiester bonds (Fig. A2). While RNA is typically single-stranded, DNA is almost always double-stranded (dsDNA), with each strand running antiparallel in an opposite direction (5' to 3' *vs* 3' to 5') to the other (Fig. A3).

The complementary nature and directionality of nucleic acids provide a basis for the transcription of genes from DNA into RNA copies. Genes may be loosely defined as stretches of DNA that are transcribed to ultimately produce a functional product, typically protein although this ultimate product may be RNA. Gene expression is exquisitely regulated in eukaryotes (*i.e.* in a very general sense organisms whose cells have nuclei and various membrane-bound organelles), perhaps most so at the level of transcription. Genes contain many different regulatory elements that interact with protein factors to regulate the initiation of transcription by the RNA-producing enzyme, RNA polymerase (Fig. A4a). RNA polymerase then catalyzes the

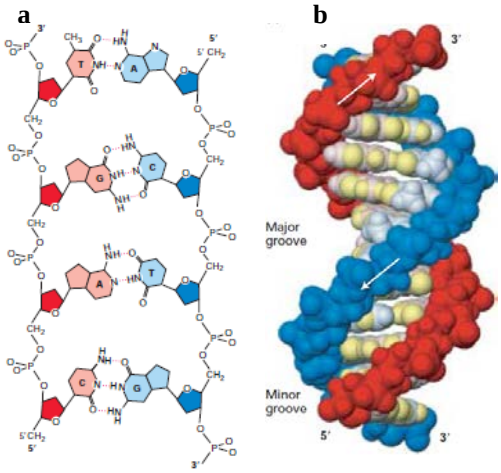

Figure A3. dsDNA structure. **a)** Schematic of antiparallel DNA strands, showing hydrogen bonds between bases. **b)** Structural representation of dsDNA, showing major and minor grooves.[2]

synthesis of an RNA that is complementary to the template strand of DNA and identical to the coding (or non-template) strand of DNA (replacing Ts with Us) in the 5' to 3' direction (Fig. A4b).

Following the transcription of an RNA copy from a gene, the RNA often undergoes extensive processing in the nucleus, of which we will primarily focus on splicing. The primary, unprocessed transcript consists of units called exons and introns (Fig. A5). Exons are stretches of the primary transcript that are included in the processed, mature RNA by the nuclear splicing machinery while introns are stretches of the primary transcript that are ultimately excised during this process. Each splicing reaction is facilitated by a complex of proteins and RNA known as the spliceosome. Exons may be considered constitutive if they are always included in the final RNA product (such as exons 1, 2, and 4 in Fig. A5a), or they may be alternatively spliced if they are not always included (such as exon 3 in Fig A5a). About 94% of human genes are estimated to produce alternatively spliced products[4]. The decision of whether to include or exclude an alternatively spliced exon depends on a multitude of factors, but a substantial contribution comes from the presence and abundance of specific proteins that regulate specific alternative splicing events. These proteins often bind directly to the primary transcript RNA and interact with the spliceosome in such a way as to

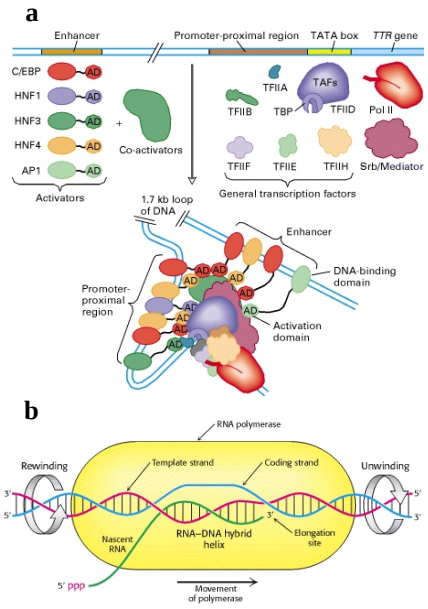

Figure A4. Transcription initiation and elongation. **a)** Many regulatory DNA elements and protein factors assemble and determine the frequency of transcription initiation[1]. **b)** RNA polymerase catalyzes the 5'-3' synthesis of RNA complementary to the template DNA strand[3].

enhance or repress the inclusion of a certain exon. One such regulator of alternative splicing that is central in the pathology of myotonic dystrophy is the protein MBNL1.

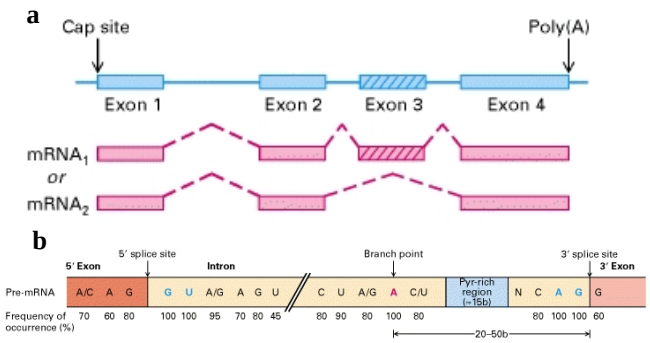

Figure A5. Splicing of a primary transcript. **a)** A canonical eukaryotic gene (in blue) contains exons (thick boxes) and introns (thin lines). The resulting transcript may be alternatively spliced in various ways to include or exclude certain exons from the mature RNA product. **b)** A canonical intron, showing conserved 5' splice site (GU), 3' splice site (AG), branch-point A, and the polypyrimidine tract, all of which are involved in exon or intron recognition by the spliceosome.[1]

Following processing, the mature RNA may then be exported from the nucleus and into the cytoplasm of the cell, where – if the RNA is a protein-encoding mRNA – it undergoes translation to produce a protein product. Each three-letter sequence in a mature mRNA is called a codon and corresponds to a specific amino acid. The ribosome – a large complex of proteins and RNA – loads onto the 5' end of the mRNA and decodes each codon, incorporating the corresponding amino acid into the nascent protein polymer as it reads forward through each codon (Fig. A6). The reading frame of a mature mRNA is then determined by a start codon – typically AUG – which determines where protein synthesis begins and ends at a stop codon – UGA, UAA, or UAG – which signals the ribosome to terminate translation and release the completed protein product.

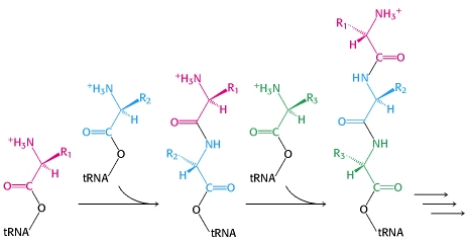

Figure A6. Polymerization of amino acids (colored units) into nascent proteins on the ribosome. Each amino acid is carried to the ribosome by an attached transfer RNA (tRNA), which specifically pairs with the presented mRNA codon[3].

APPENDIX REFERENCES

1. Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular Cell Biology*, 4th edition. New York: W.H. Freeman.

2. Lodish, H. , Berk, A., Kaiser, C.A., Krieger, M., Scott, M.P., Bretscher, A., Ploegh, H., and Matsudaira, P. (2007). *Molecular Cell Biology*, 6th edition. New York: W.H. Freeman.

3. Berg, J.M., Tymoczko, J.L., and Stryer, L. (2002). *Biochemistry*, 5th edition. New York: W.H. Freeman.

4. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470-6.

APPENDIX FIGURES

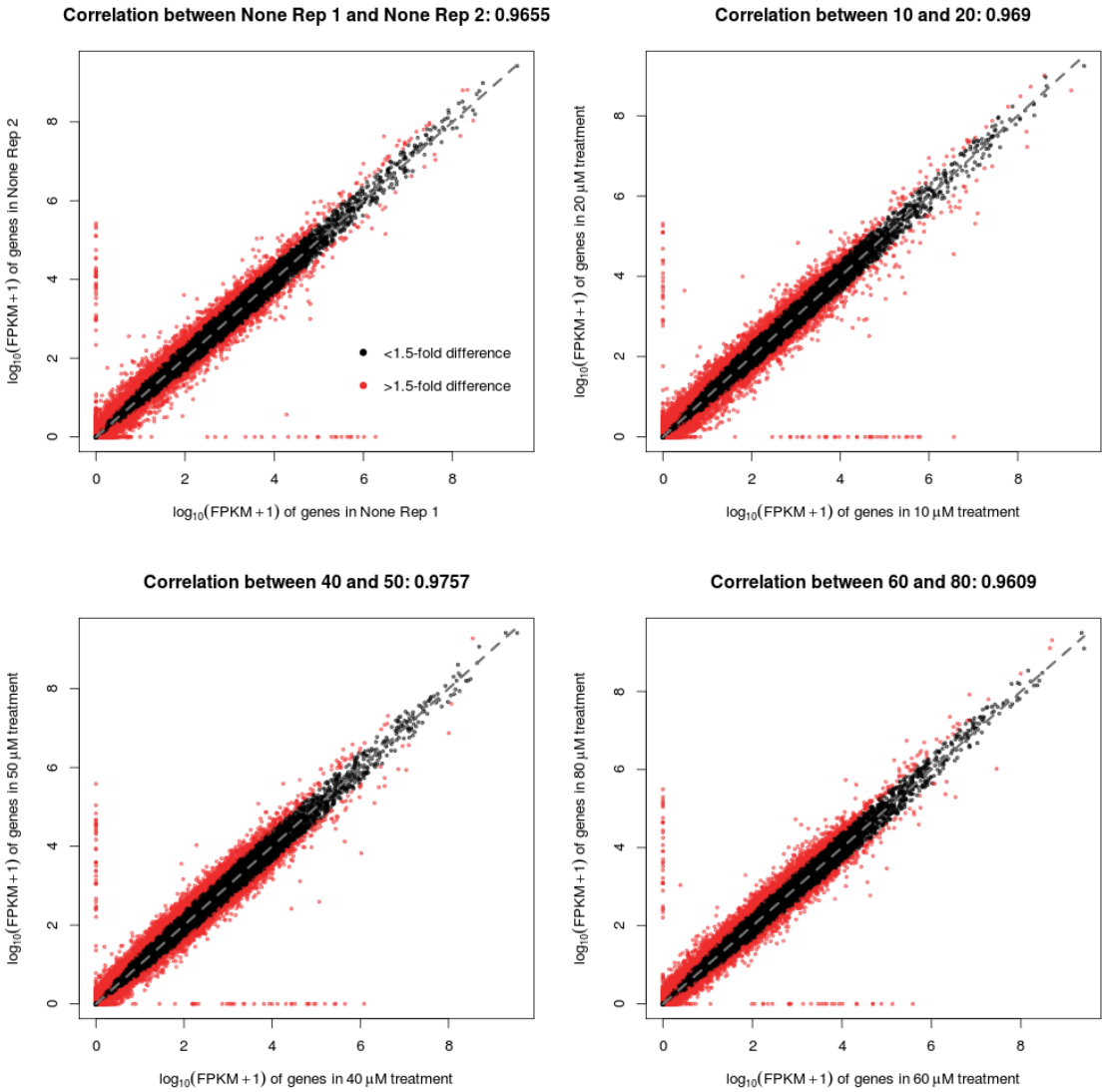

Figure A7. Samples used as replicates in pentamidine titration correlate well with each other. Sequencing data from different pentamidine dosages were paired and used as replicates for differential analysis, and the $\log_{10}$(FPKM+1) of all genes in each sample trends well with its replicate pair. Dashed line represents one-to-one correspondence between samples. Red dots represent genes where greater than a 1.5-fold difference in (FPKM+1) occurred between replicates.

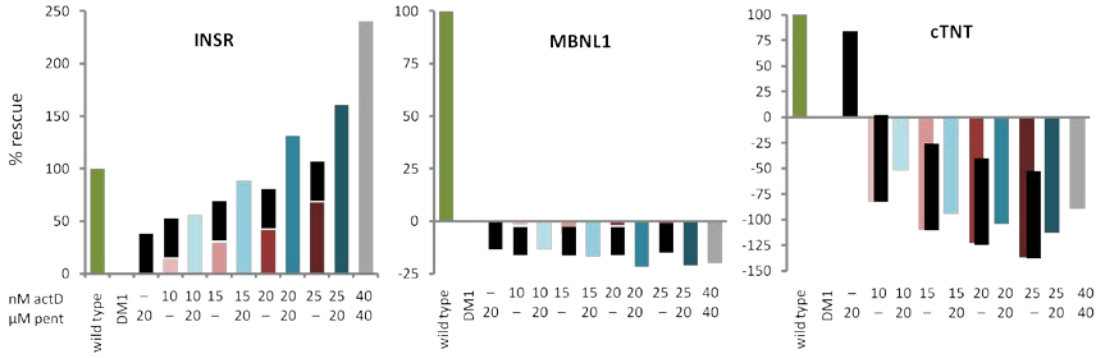

Figure A8. Marginal splicing rescue (as a percentage of rescue between wild-type and DM1 splicing levels) achieved by actinomycin and pentamidine treatment (data transformed from Fig. 8). Black bars superimposed on actinomycin-only treatments represent expected splicing rescue if actinomycin and 20 µM pentamidine act additively, while blue bars represent the observed splicing rescue at these respective combinatorial dosages. Actinomycin and pentamidine appear to have additive effects on INSR splicing at 10 nM actD and 20 µM pent, while past this dosage synergistic effects are observed. Actinomycin appears to mask much more of the rescuing effects of pentamidine than would be expected if the two drugs acted additively in cTNT splicing. While each drug appears to have additive effects in MBNL1 splicing, this is confounded by actinomycin's marginal effects on splicing in this event.

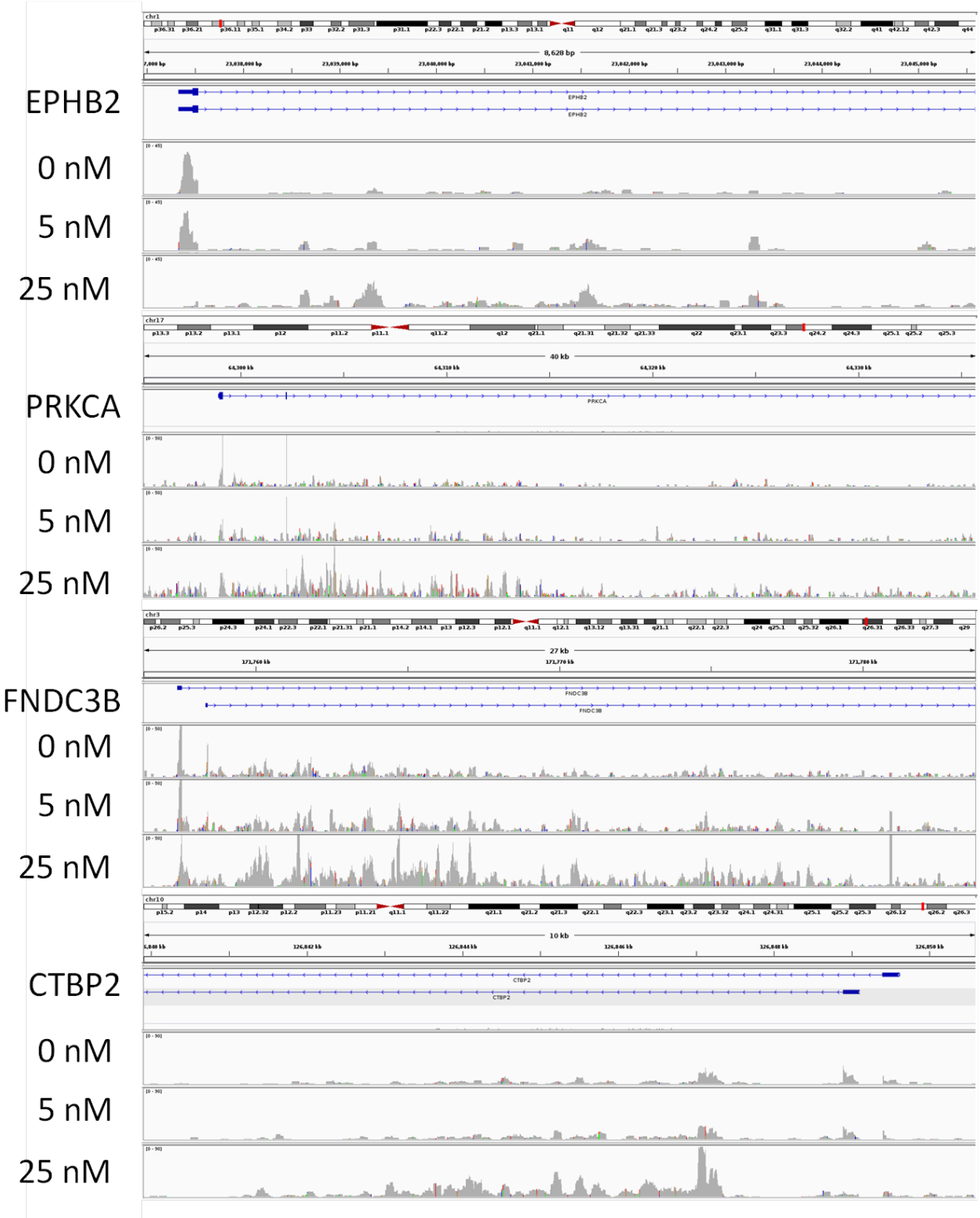

Figure A9. Several of the 74 significantly decreased genes exhibit a shift in the distribution of early gene region reads. EPHB2 (positive strand) and CTBP2 (negative strand) decrease in aligned reads in the first exon, but increase in aligned reads in the first several kb of their respective first introns at 25 nM actinomycin. PRKCA and FNDC3B increase in intronic reads in the first 10-20 kb of their first introns at 25 nM actinomycin.

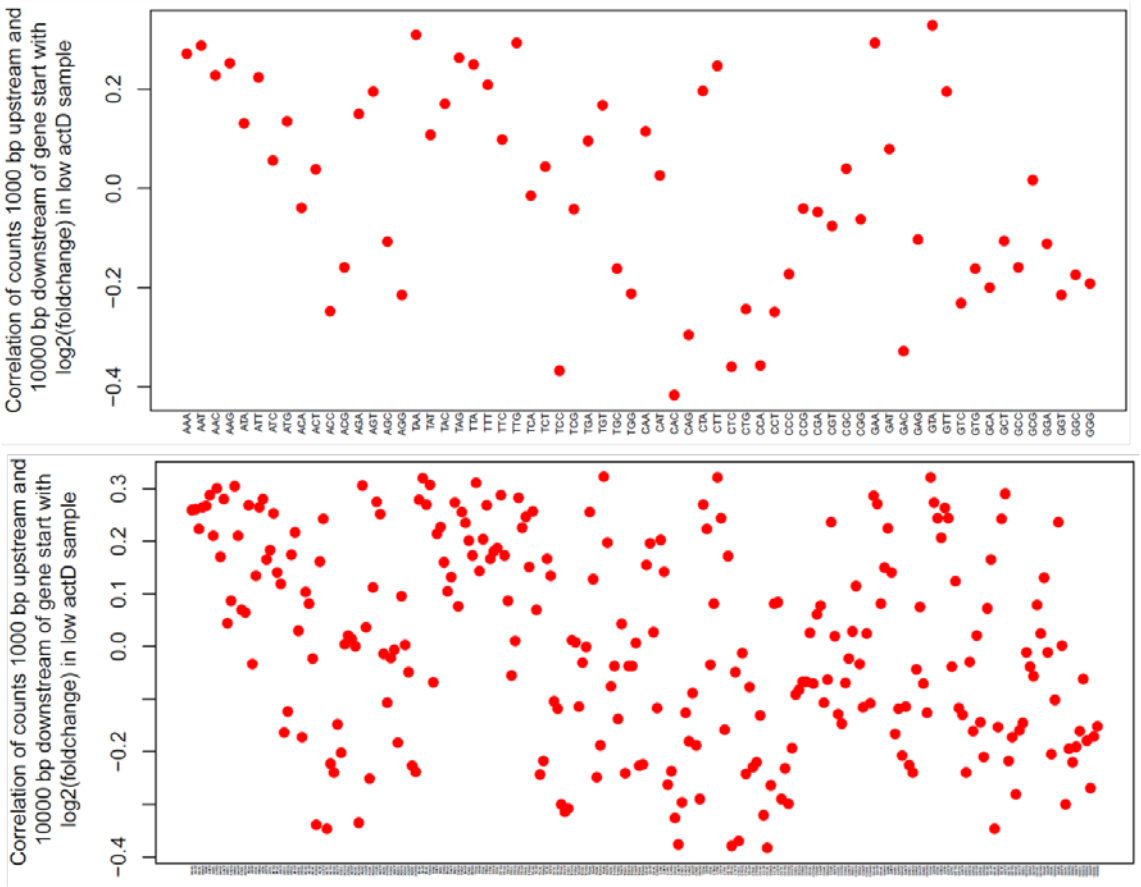

Figure A10. Correlations of occurrences of all 3-mers (top) and all 4-mers (bottom) in early gene regions (1 kb upstream of gene start to 10 kb downstream of gene start) with change in expression of the 74 significantly decreased genes at 5 nM actinomycin. The average magnitude of correlation |r| is 0.178 for all 3-mers and 0.166 for all 4-mers. Average correlations (without taking absolute values) are 0 for both cases.

REFERENCES

1. U.S. National Library of Medicine. "Myotonic dystrophy". Jan. 13th, 2014. http://ghr.nlm.nih.gov/condition/myotonic-dystrophy

2. Bird, T.D. (2013). Myotonic dystrophy type 1. *GeneReviews*. http://www.ncbi.nlm.nih.gov/books/NBK1165/

3. Dalton, J.C., Ranum, L.P.W., and Day, J.W. (2013). Myotonic dystrophy type 2. *GeneReviews*. http://www.ncbi.nlm.nih.gov/books/NBK1466/

4. Berul, C.I., Maguire, C.T., Aronovitz, M.J., Greenwood, J., Miller, C., Gehrmann, J., Housman, D., Mendelsohn, M.E., and Reddy, S. (1999). *DMPK* dosage alterations result in atrioventricular conduction abnormalities in a mouse myotonic dystrophy model. *J Clin Invest* 103(4):R1-7.

5. Klesert, T.R., Cho, D.H., Clark, J.I., Maylie, J., Adelman, J., Snider, L., Yuen, E.C., Soriano, P., and Tapscott, S.J. (2000). Mice deficient in *Six5* develop cataracts: implications for myotonic dystrophy. *Nat Genet* 25:105-9.

6. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470-6.

7. Jiang, H., Mankodia, A., Swanson, M.S., Moxley, R.T., and Thornton, C.A. (2004). Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum Molec Genet* 13(24):3079-88.

8. Han. H., Irimia, M., Ross, J.P., Sung, H.-K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., Wang, E., Trcka, D., Thompson, T., O'Hanlon, D., Slobodenuic, V., Barbosa-Morais, N.L., Burge, C.B., Moffat, J., Frey, B.J., Nagy, A., Ellis, J., Wrana, J.L., and Blencowe, B.J. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*. Published online ahead of print.

9. Dansithong, W., Paul, S., Comai, L., and Reddy, S. (2005). MBNL1 is the primary determinant of focus formation and aberrant insulin receptor splicing in DM1. *J Biol Chem* 280(7):5773-80.

10. Kanadia, R., Johnstone, K.A., Mankodi, A., Lungu, C., Thornton, C.A., Douglas, E., Timmers, A.M., Hauswirth, W.W., and Swanson, M.S. (2003). A muscleblind knockout model for myotonic dystrophy. *Science* 302:1978-80.

11. Du, H., Cline, M.S., Osborne, R.J., Tuttle, D.L., Clark, T.A., Donohue, J.P., Hall, M.P., Shiue, L., Swanson, M.S., Thornton, C.A., and Ares Jr., M. (2010). Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. *Nat Struct Mol Biol* 17(2):187-93.

12. Goes, E.S., Purcell, J., Voelker, R.B., Gates, D.P., and Berglund, J.A. (2010). MBNL1 binds GC motifs embedded in pyrimidines to regulate alternative splicing. *Nucleic Acids Res* 38(7):2467-84.

13. Coonrod, L.A., Nakamori, M., Wang, W., Carrel, S., Hilton, C.L., Bodner, M.J., Siboni, R.B., Docter, A.G., Haley, M.M., Thornton, C.A., and Berglund, J.A. (2013). Reducing levels of toxic RNA with small molecules. *ACS Chem Biol* 8(11):2528-37.

14. Warf, M.B, Nakamori, M., Matthys, C.A., Thornton, C.A., and Berglund, J.A. (2009). Pentamidine reverses the splicing defects associated with myotonic dystrophy. *Proc Natl Acad USA* 106(44):18551-6.

15. Niedle, S., Jenkins, T.C., and Edwards, K.J. (1992). Crystal structure of a pentamidine-oligonucleotide complex: implications for DNA-binding properties. *Biochemistry* 31(31):7104-9.

16. Hou, M.H., Robinson, H., Gao, Y.-G., and Wang, A.H.-J. (2002). Crystal structure of actinomycin D bound to the CTG triplet repeat sequences linked to neurological disease. *Nucleic Acids Res* 30(22):4910-7.

17. Lo, Y.S., Tseng, W.H., Chuang, C.Y., and Hou, M.H. (2013). The structural basis of actinomycin-D binding induces nucleotide flipping out, a sharp bend and a left-handed twist in CGG triplet repeats. *Nucleic Acids Res* 41(7):4284-94.

18. Philips, A.V., Timchenko, L.T., and Cooper, T.A. (1998). Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science* 280:737-741.

19. Purcell, J., Oddo, J.C., Wang, E.T., and Berglund, J.A. (2012). Combinatorial mutagenesis of MBNL1 zinc fingers elucidates distinct classes of regulatory events. *Mol Cell Biol* 32(20):4155-67.

20. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harb Symop Quant Biol* LI:263-73.

21. Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57-63.

22. Epicentre (2012). ScriptSeq v2 RNA-Seq Library Preparation Kit, user protocol.

23. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions, and gene fusions. *Genome Biol* 14:R36-48.

24. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511-5.

25. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachtor, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nat Biotechnol* 31(1):46-53.

26. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14(2):178-92.

27. Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 7:1009-15.

28. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38:576-89.

29. Charizanis, K., Lee, K.Y., Batra, R., Goodwin, M., Zhang, C., Yuan, Y., Shiue, L., Cline, M., Scotti, M.M., Xia, G., Kumar, A., Ashizawa, T., Clark, H.B., Kimura, T., Takahashi, M.P., Fujimura, H., Jinnai, K., Yoshikawa, H., Gomes-Pereira, M., Gourdon, G., Sakai, N., Nishino, S., Foster, T.C., Ares, M. Jr., Darnell, R.B., and Swanson, M.S. (2012). Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy. *Neruon* 75(3):437-50.

30. Paul, S., Dansithong, W., Jog, S.P., Holt, I., Mittal, S., Brook, D.J., Morris, G.E., Comai, L., and Reddy, S. (2011). Expanded CUG repeats dysregulate RNA splicing by altering the stoichiometry of the Muscleblind 1 complex. *J Biol Chem* 286(44):38427-38.

31. Gooding, C., Edge, C., Lorenz, M., Coelho, M.B., Winters, M., Kaminski, C.F., Cherny, D., Eperon, I.C., and Smith, C.W.J. (2013). MBNL1 and PTB cooperate to repress splicing of *Tpm1* exon 3. *Nucleic Acids Res* 41(9):4765-82.

32. Caputi, M., and Zahler, A.M. (2001). Determination of the RNA binding specifity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. *J Biol Chem* 276(47):43850-9.

33. Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* 4(1):44-57.

34. Rohs, R. West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature* 461:1248-53.

35. Sen, S., Talukdar, I., and Webster, N.J.G. (2009). SRp20 and CUG-BP1 modulate insulin receptor exon 11 alternative splicing. *Mol Cell Biol* 29(3):871-80.

36. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L.H., Dale, R.K., Smith, S.A., Yarosh, C.A., Kelly, S.M., Nabet, B., Mecenas, D., Li, W., Laishram, R.S., Qiao, M., Lipshitz, H.D., Piano, F., Corbett, A.H., Carstens, R.P., Frey, B.J., Anderson, R.A., Lynch, K.W., Penalva, L.O.F., Lei, E.P., Fraser, A.G., Blencowe, B.J., Morris, Q.D., and Hughes, T.R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499:172-7.

37. Marquis, J., Paillard, L., Audic, Y., Cosson, B., Danos, O., Le Bec, C., Osborne, H.B. (2006). CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochem J* 400:291-301.

38. Masuda, A., Anderson, H.S., Doktor, T.K., Okamoto, T., Ito, M., Andresen, B.S., and Ohno, K. (2012). CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. *Sci Rep* 2(209):1-10.

39. Kalsotra, A., Xiao, X., Ward, A.J., Castle, J.C., Johnson, J.M., Burge, C.B., and Cooper, T.A. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad USA* 105(51):20333-8.

40. Han, S.P., Tang, Y.H., and Smith, R. (2010). Functional diversity of the hnRNPs: past, present, and perspectives. *Biochem J* 430:379-92.

41. Chaudhury, A., Chander, P., and Howe, P.H. (2010). Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: focus on hnRNP E1's multifunctional regulatory roles. *RNA* 16:1449-62.

42. Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., Ares Jr., M., and Yeo, G.W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Reports* 1:167-78.

43. Rill, R.L., and Hecker, K.H. (1996). Sequence-specific actinomycin D binding to single-stranded DNA inhibits HIV reverse transcriptase and other polymerases. *Biochemistry* 35:3525-33.

44. Paramanathan, T., Vladescu, I., McCauley, M.J., Rouzina, I., and Williams, M.C. (2012). Force spectroscopy reveals the DNA structural dynamics that govern the slow binding of actinomycin D. *Nucleic Acids Res* 40(11):4925-32.

45. Annunziato, A. (2008). DNA packaging: nucleosomes and chromatin. *Nature Education* 1(1):26.

46. Choong, M.L., Yang, H., Lee, M.A., and Lane, D.P. (2009). Specific activation of the p53 pathway by lose dose actinomycin D. *Cell Cycle* 8(17):2810-18.

47. Honda, K., and Taniguchi, T. (2006). IRFS: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. *Nat Rev Immunol* 6:644-58.