2017-07-10

Alterations in mRNA 3'UTR Isoform Abundance Accompany Gene Expression Changes in Huntington's Disease

Lindsay S. Romo
University of Massachusetts Medical School

ALTERATIONS IN mRNA 3′UTR ISOFORM ABUNDANCE ACCOMPANY GENE EXPRESSION CHANGES IN HUNTINGTON'S DISEASE

A Dissertation Presented

By

LINDSAY S. ROMO

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences, Worcester
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

July 10, 2017

MD/PhD Program

ALTERATIONS IN mRNA 3′UTR ISOFORM ABUNDANCE ACCOMPANY GENE EXPRESSION CHANGES IN HUNTINGTON'S DISEASE

A Dissertation Presented
By
LINDSAY S. ROMO

The signature of the Thesis Advisor signifies
validation of the Dissertation content

_____
Neil Aronin, M.D., Thesis Advisor

The signatures of the Dissertation Defense Committee signify
completion and approval as to style and content of the Dissertation

_____
Vivian Budnik, Ph.D., Member of the Committee

_____
Allan Jacobson, Ph.D., Member of the Committee

_____
Silvia Corvera, M.D., Member of the Committee

_____
David Housman, Ph.D., External Member of the Committee

The Signature of the Chair of the Committee signifies that the written dissertation meets
the requirements of the Dissertation Committee

_____
Sean Ryder, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Science signifies
that the student has met all graduation requirements of the School.

_____
Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

July 10th, 2017
MD/PhD Program

DEDICATION

To my parents, for inspiring me and sacrificing for me since the day I was born, and to my daughter, for whom I hope to do the same.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Neil Aronin, without whom this work would not have been possible. He has supported me, trusted me, and given me free reign to explore my ideas. Most importantly, he is an example to me of how it is possible to be both brilliant and kind, a leader and a listener, a doctor and a scientist, and a father and a professional. I hope to emulate him in my career and personal life.

I would like to acknowledge my committee members, who have believed in me and given me direction when I was lost. Without your help, I would have struggled to design experiments or interpret results.

Thank you to all of the past and present members of the Aronin lab, in particular Edith Pfister, who has been a sounding board for my ideas, and Kathy Chase and Lori Kennington, who have been my work mothers and helped me with my experiments and personal life when I was overwhelmed. A special thank you as well to Rachael Miller who took over my SNP-linkage to CAG-repeat extended-range project when I became too busy and who is bringing it to completion. Thank you also to Ami Ashar-Patel, for designing the PAS-seq protocol on which my project is heavily dependent, and for commiserating with me through this long journey.

Finally, thank you to my family: to my parents for their support and motivation, to my husband for pulling the extra weight when I could not, and to my daughter for putting up with my occasional physical and mental absences for the past year.

ABSTRACT

Huntington's disease is a neurodegenerative disorder caused by expansion of the CAG repeat in huntingtin exon 1. Early studies demonstrated the huntingtin gene is transcribed into two 3′UTR isoforms in normal human tissue. Decades later, researchers identified a truncated huntingtin mRNA isoform in disease but not control human brain. We speculated the amount of huntingtin 3′UTR isoforms might also vary between control and Huntington's disease brains.

We provide evidence that the abundance of huntingtin 3′UTR isoforms, including a novel mid-3′UTR isoform, differs between patient and control neural stem cells, fibroblasts, motor cortex, and cerebellum. Both alleles of huntingtin contribute to isoform changes. We show huntingtin 3′UTR isoforms are metabolized differently. The long and mid isoforms have shorter half-lives, shorter polyA tails, and more microRNA and RNA binding protein sites than the short isoform.

3′UTR Isoform changes are not limited to huntingtin. Isoforms from 11% of genes change abundance in Huntington's motor cortex. Only 17% of genes with isoform alterations are differentially expressed in disease tissue. However, gene ontology analysis suggests they share common pathways with differentially expressed genes. We demonstrate knockdown of the RNA binding protein CNOT6 in control fibroblasts results in huntingtin isoform changes similar to those in disease fibroblasts. This study further characterizes Huntington's disease molecular pathology and suggests RNA binding protein expression may influence mRNA isoform expression in the Huntington's disease brain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

COPYRIGHT MATERIALS

The work presented in this dissertation is accepted at *Cell Reports* as manuscript "Alterations in mRNA 3′UTR isoform abundance accompany gene expression changes in human Huntington's disease brains" by Romo, Ashar-Patel, Pfister, and Aronin.

CHAPTER I: INTRODUCTION

1.1 Huntington's disease

Huntington's disease (HD) is an inherited invariably fatal neurodegenerative disorder. Individuals usually develop symptoms in their mid-thirties and death occurs within two decades. Early symptoms are psychiatric and cognitive changes. As the disease progresses motor symptoms and dementia develop. Motor symptoms start with aberrant eye motions, progress to involuntary movements (chorea), and end with bradykinesia and rigidity. In late stages, patients experience significant cognitive decline and have difficulty speaking, walking, and swallowing. At that point, long-term institutionalization is necessary. HD prevalence varies. In Asia, prevalence is low at less than 0.5 individuals per 100,000, whereas in Europe, the prevalence is ten times higher. In the United States, 2 of every 100,000 individuals are affected[1]. Currently, there is no treatment that alters the course of the disease. Because patients decline over decades and ultimately require intensive care, HD imposes a great emotional burden on families, and a significant cost on society.

George Huntington first described the autosomal dominant heritability of HD in 1872[2]. Analysis of recombination events in HD families narrowed the region down to a novel gene on chromosome 4[3–6]. Sequencing of the gene revealed a CAG (glutamine) repeat near its 5′ end[7]. Control DNA had a CAG repeat number of 35 or less, while HD patient DNA had a repeat number of 35-100. Repeat length is inversely correlated with age of onset and symptom severity[8,9]. The CAG number often increases between generations, especially when paternally transmitted[10–12]. Patients homozygous for the expansion have similar age of onset and phenotype; however, some studies show they

have more rapid and severe disease symptoms[13,14]. The gene was named huntingtin (*HTT*) in honor of George Huntington.

HTT protein is expressed ubiquitously, though its levels are highest in neurons and testes[15]. HTT knockout mice die prior to gastrulation[16–19]. After gastrulation, HTT is required for development of the central nervous system and homeostasis of cerebrospinal fluid[17,20–23]. HTT's role in development is independent of CAG repeat-length, and humans homozygous for the expansion develop normally[13,21]. One functional *HTT* allele is sufficient for normal development[24,25]. However, reduction of HTT levels below 50% of normal results in impaired neurogenesis[20,21]. Wild-type HTT protein interacts with other proteins to protect cells against toxic stimuli, mediate vesicle transport and endocytosis, and modulate synaptic activity; many of these functions are impaired in HD[26–28]. Mutant HTT protein causes neuronal disease and death by damaging autophagy, vesicle transport, neurotransmitter signaling, and mitochondrial function[27]. HD Pathology is most significant in the striatum and cortex, although by the time of death there is widespread brain atrophy[29,30].

While much is known about HTT protein, fewer studies have explored *HTT* mRNA. The *HTT* gene is transcribed into a sense and anti-sense transcript; the sense transcript makes up the vast majority of *HTT*[31]. Sense *HTT* mRNA is almost fourteen kilobases long, thus its structure is complex. The CAG repeat region forms a hairpin stabilized by the neighboring CCG repeat; the stability of the hairpin increases with CAG expansion[32]. *HTT* mRNA is normally localized diffusely around the cytoplasm, but in HD it forms nuclear foci[32]. Allele specific RT-qPCR of RNA from HD patient and control

brains revealed there is more mutant than wild-type *HTT* in striatum and cortex from HD early-grade brains[33]. Thus the stability, localization, and abundance of *HTT* mRNA change in HD.

Wild-type and mutant *HTT* mRNA are spliced into multiple isoforms, although splice variants are of low abundance compared to the canonical splice isoform[34]. In human and mouse brains, *HTT* can be alternatively spliced to lack exon 29[35]. This isoform is expressed at lower levels in HD knock-in mice than in wild-type mice. Exon 29 contains a putative binding site for an RNA nuclear export protein, suggesting the decrease in the exon 29-skipping isoform may cause *HTT* nuclear retention[35]. Another study identified five new *HTT* splice isoforms that ranged from 0.8% to 13.6% of the expression of the canonical transcript[36]. High throughput sequencing of human brain revealed 18 additional alternative splice isoforms ranging in expression from 1% to 7% to that of the canonical transcript[34]. It has yet to be determined if or how these splice isoforms contribute to HD.

Two alternatively polyadenylated 3′UTR isoforms make up the majority of *HTT* mRNA[37]. The putative polyA signal for the short isoform, AGUAAA, has lower in vitro cleavage and polyadenylation efficiency than that of the long isoform, AUUAAA[38]. The long isoform is 13.7 kilobases whereas the short isoform is 10.3 kilobases, and their relative abundance differs across tissues[37]. The long isoform is more abundant in fetal brain, while the short isoform is more abundant in cell lines and lymphoblasts[37]. The isoforms may play different roles in cells; transfected *HTT* exon 1-3′UTR constructs are

localized differently depending on the 3′UTR length, and the shorter construct forms more aggregates[39].

In HD patients and mouse models but not controls, a small amount *HTT* mRNA is misspliced and polyadenylated at a cryptic polyadenylation signal in intron 1[40]. Translation of this isoform will produce a toxic N-terminal protein[27]. The isoform may be produced via serine/argenine-rich splicing factor 6 (SRSF6), which binds to expanded *HTT* mRNA with much higher affinity than wild-type mRNA. There is a stop codon at the very beginning of intron 1, and immunoprecipitation of N-terminal HTT detected a protein of the appropriate size in HD mouse models but not controls[40]. However, it is unknown whether the protein represents translation of the truncated isoform, or cleavage of full-length HTT. Production of the truncated *HTT* mRNA isoform indicates *HTT* is differently polyadenylated in HD (Fig. 1.1, bottom).

I hypothesized that the 3′UTR alternative polyadenylation of *HTT* and other genes changes in HD. There are two lines of evidence to support this hypothesis: first, as described, *HTT* itself is aberrantly polyadenylated in disease[40]; second, molecular processes linked to polyadenylation including DNA, RNA, and protein expression are aberrant in HD[41–43]. Epigenetic modifications, transcription rate, mRNA expression, and protein abundance affect polyA site selection[44,45]. Alterations in these processes in disease could impact alternative polyadenylation. I sought to characterize the 3′UTR isoform expression of *HTT* and other genes in HD.

A study of alternative polyadenylation could provide recommendations or targets for HD therapy. Epigenetic and mRNA expression changes contribute to pathology in

HD models[27,41,42]. If polyadenylation is also altered in disease, characterization of 3′UTR isoform abundances in HD may reveal novel therapeutic targets. In *HTT*, some single nucleotide polymorphisms (SNPs) that are frequently heterozygous in HD patients are exclusive to the long 3′UTR isoform[46]. These sites are ideal targets for interfering RNAs that discriminate between wild-type and mutant mRNAs; however, the short *HTT* 3′UTR isoform is more pathogenic in cells[39,46,47]. Examination of *HTT* mRNA 3′UTR isoform abundance and processing in HD patients may provide recommendations for the design of mutant *HTT*-lowering therapies.

1.2 Alternative polyadenylation

1.2.1 Mechanism of alternative polyadenylation.

Messenger RNA 3′ endonucleolytic cleavage and polyadenylation occur co-transcriptionally (Fig. 1.1, top). During the initiation of transcription, the 3′ processing machinery assembles on RNA polymerase II[48]. This machinery is a multi-protein complex consisting of the cleavage and polyadenylation specificity factor (CPSF), cleavage factors I and II (CF I_m, CF II_m), and the cleavage and stimulation factor (CstF)[49]. CF I_m and CF II_m bind UGUA motifs[50]. Downstream of these motifs, CPSF recognizes a hexanucleotide sequence, called a polyA signal. The canonical polyA signal is AAUAAA; however, nucleotide variations also elicit cleavage and polyadenylation at lower efficiency[38,51,52]. Downstream of the polyA site, CstF recognizes U- or GU-rich regions and influences the cleavage location and efficiency[53]. After the 3′ end processing machinery binds the nascent transcript, it is cleaved 10-30 nucleotides downstream of the polyA signal; the location of CPSF and CstF binding determines the exact cleavage site[54]. The rate and location of cleavage is dependent on the concentration of these 3′ processing factors[55]. After a transcript is cleaved, the nuclear polyA polymerase adds up to 300 adenosines to the 3′ end of the nascent mRNA. Transcripts that have been spliced and polyadenylated are exported to the cytoplasm, where cytoplasmic adenylases and deadenylases further modulate the polyA tail length[56].

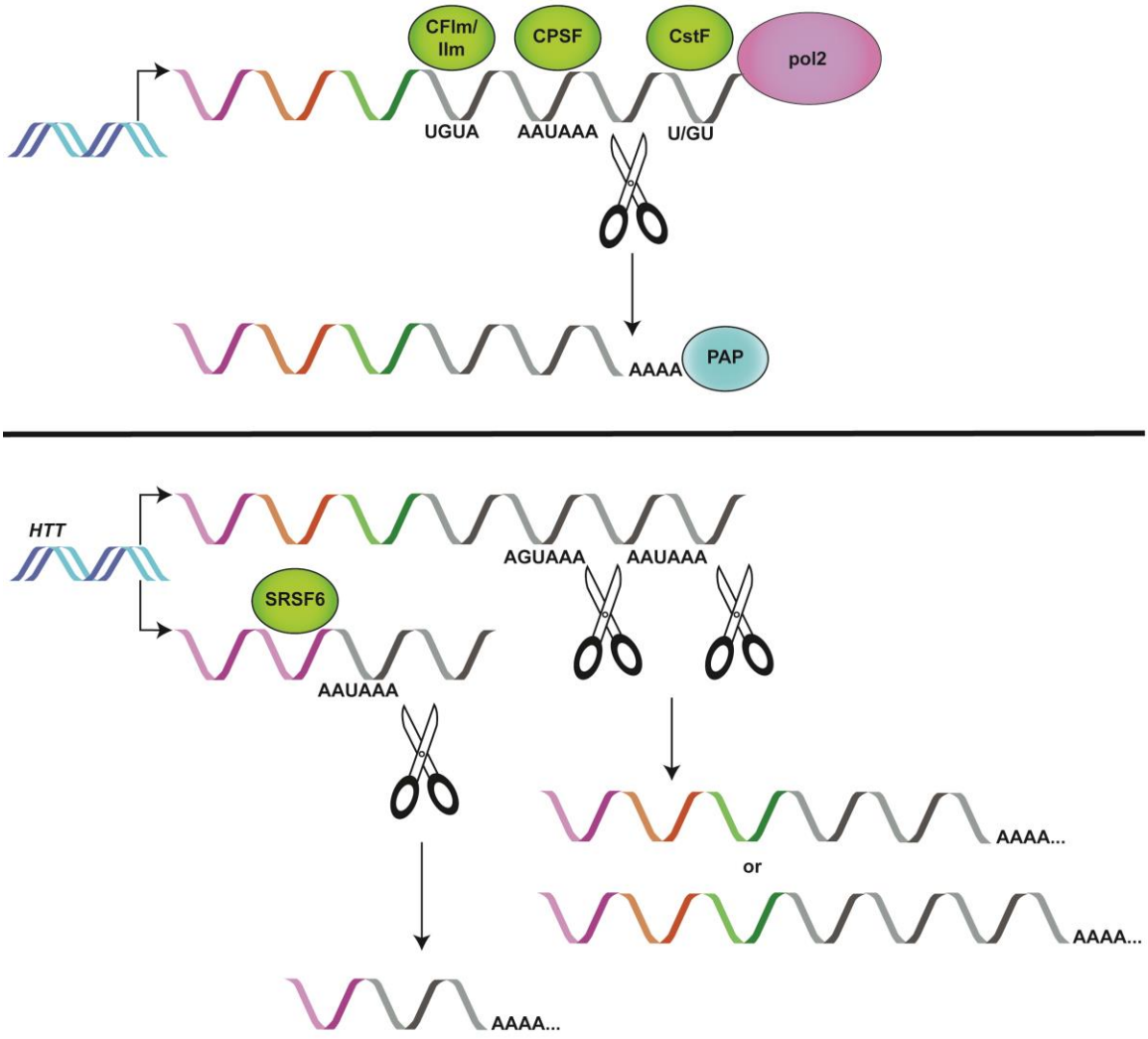


Figure 1.1. *HTT* mRNA is alternatively polyadenylated. (Top) During 3′ UTR (gray) cleavage and polyadenylation, CFIm and CFIIm bind UGUA motifs upstream of the polyA signal. CPSF recognizes and binds the polyA signal (AAUAAA). Downstream of the polyA signal, CstF recognizes U/GU-rich motifs. The relative location of CstF to the polyA signal determines the cleavage site. After cleavage, the transcript is polyadenylated by the polyA polymerase (PAP). **(Bottom)** The *HTT* 3′ UTR has two known polyA signals, a proximal AGUAAA and distal AAUAAA. The 3′ UTR polyA site selected for *HTT* differs depending on the tissue. Mutant *HTT* binds the SRSF6 splicing factor, leading to inappropriate retention of intron 1 (bottom). The misspliced transcript is cleaved and polyadenylated at the cryptic intron 1 AAUAAA polyA signal, resulting in a truncated mRNA that may be translated into a pathogenic protein.

Like *HTT*, most mammalian mRNAs have more than one polyA signal and site, a phenomenon termed alternative polyadenylation[57]. Alternative polyadenylation events fall into four classes. Most commonly, cleavage and polyadenylation occur at two different polyA signals within the same terminal 3′UTR, which does not affect the protein coding sequence. The other three classes involve alternative splicing and alter the coding sequence. Alternative splicing may generate a different terminal exon and 3′UTR. Missplicing can lead to cleavage and polyadenylation after a cryptic polyA signal within an intron, as occurs in the misspliced mutant *HTT* isoform. Finally, cleavage and polyadenylation can initiate prematurely within the coding region. These last two events produce a truncated transcript and protein. PolyA signal strength depends on the sequence of the signal and the surrounding elements. Only 58% of human polyA signals are the canonical signal[58]. In mRNAs with more than one 3′UTR polyA signal, AAUAAA is usually the most distal; however, many transcripts are cleaved and polyadenylated at weaker, more proximal polyA signals[58–62].

Although proximal polyA signals are usually weaker than distal signals, they are transcribed first, giving them more time for recognition by CPSF. In a *Drosophila* polymerase II mutant with slower elongation rates, the polo gene shifts to shorter isoforms[63]. The transcription elongation factor ELL2 is loaded with CstF subunit 64 on RNA polymerase II; knockdown of ELL2 in plasma cells leads the immunoglobulin heavy-chain complex (*Igh*) mRNA to shift to a proximal polyA signal[64]. Pausing of RNA polymerase II at the promoter of genes with extended 3′UTRs in *Drosophila* results in recruitment of the RNA binding protein ELAV, which inhibits cleavage and

polyadenylation at proximal polyA signals[65]. Transcription activators can increase usage of proximal polyA signals by recruiting 3′ processing factors[66]. GAL4-VP16 recruits the RNAP II-associated factor (PAF) elongation complex to genes and stimulates cleavage and polyadenylation of transcripts[67]. Mediator interacts with mRNA processing factor heterogeneous nuclear ribonucleoprotein L (hnRNP L) to affect alternative polyadenylation at transcriptionally active target genes[68]. Epigenetics also impact polyA site usage. The mouse imprinted genes *H13* and *Herc3/Nap1I5* unmethylated alleles undergo internal polyadenylation, while the methylated alleles are polyadenylated at a downstream site[69,70].

Changes in the concentration of 3′ processing proteins can lead cells to shift to more proximal or distal polyA signals. Overexpression of CstF subunit 64 (CstF-64) in B cells causes *IgM* to shift to its proximal polyA site[71]. During T cell activation, increased levels of CstF-64 promote polyadenylation at proximal polyA sites in the nuclear factor of activated T-cells (*NF-ATc*) gene[72]. Knockdown of another 3′ processing factor, CFIm68, causes a widespread shift to proximal polyA signals in human embryonic kidney (HEK293) cells[60]. Higher expression of 3′ processing factors in proliferating cells than in non-dividing cells likely leads to shorter isoforms[45,73].

Other RNA binding proteins affect alternative polyadenylation. Polypyrimidine tract binding protein (PTB) and *Drosophila* sex lethal protein (SXL) compete with CtsF to bind to the pyrimdine-rich region downstream of the polyA signal, reducing efficacy of cleavage and polyadenylation and promoting distal polyA site usage[74,75]. Similarly, the RNA binding protein HuR regulates its own alternative polyadenylation by competing

with CstF[76]. In mouse brains, the RNA binding protein Nova hinders cleavage and polyadenylation at some polyA sites[77]. Epithelium-specific splicing regulatory proteins ESRP1 and 2 bind UG-rich elements downstream of polyA sites and prevent their use, potentially via inhibition of CstF binding[78]. Knockdown of another RNA binding protein, polyA binding protein nuclear 1 (PABPN1), caused a global shift towards proximal polyA signals in cells, likely because PABPN1 competes with CPSF for binding at proximal polyA signals[79]. U1 small nuclear ribonucleoprotein (snRNP) binds mRNAs in introns to prevent premature cleavage and polyadenylation at cryptic poly A sites near the transcript 5′ end[80]. Moderate decreases in U1 snRNP levels shift cleavage and polyadenylation to proximal 3′UTR polyA sites[81]. Knockdown of DICER1 causes many mRNAs to shift towards shorter isoforms, suggesting a role for DICER1 in isoform abundance beyond microRNA-mediated repression[82].

Tissue-specific alternative polyadenylation contributes to tissue-specific protein expression[83]. Many genes that are ubiquitously transcribed are expressed tissue-specifically via alternative polyadenylation, whereas genes that are transcribed in only one tissue often have a single 3′UTR polyA site[84]. During cell transformation and differentiation, a subset of mRNAs alter their 3′UTR length to change their protein expression[84,85]. Tissue-specific 3′UTR lengths allow fine-tuning of the expression of mRNAs targeted by ubiquitously expressed microRNAs[62,84]. Highly proliferative tissues such as testes generally have shorter 3′UTR lengths, with fewer binding sites, while quiescent tissues such as brain have longer 3′UTR lengths[62,86,87]. Several factors contribute to tissue-specific alternative polyadenylation including the distance between

tandem 3′UTR polyA sites, the position of the polyA site within the 3′UTR, and tissue-specific expression of the 3′ processing machinery or other RNA binding proteins that regulate alternative polyadenylation[88,89].

1.2.2 Consequences of alternative polyadenylation.

Trans-acting proteins bind *cis* elements in the 3′UTR to determine mRNA localization and stability[90,91]. In yeast, localization of the protein Ash1 is dependent on interactions between the 3′UTR and the actin cytoskeleton[92]. Similarly, beta and alpha actin mRNA require the 3′UTR for appropriate cellular localization[93]. Alternative polyadenylation can change mRNA localization. Brain-derived neurotrophic factor (*BDNF*) is transcribed into two predominant 3′UTR isoforms. The longer isoform is localized to dendrites, where it plays a role in long term potentiation[94]. The shorter *BDNF* 3′UTR stabilizes mRNA in cell bodies after depolarization[95]. Another neuronal mRNA, calcium/calmodulin dependent protein kinase II (*CaMKII*), is transcribed into two isoforms, the longer of which is localized to dendrites[96]. Isoform expression also mediates protein localization independently of mRNA localization. During B cell activation, *IgM* shifts to its shorter isoform, resulting in secreted IgM protein[71]. The long *CD47* isoform acts as a scaffold to recruit RNA binding proteins to the site of *CD47* translation, where they promote translocation of the protein to the membrane rather than endoplasmic reticulum[97].

Isoforms of the same mRNA may have different stability. Destabilizing adenosine and uridine-rich elements, or AREs, may be exclusive to one isoform[98]. Different polyA

tail lengths can cause different stabilities[56,99]. The polyA tail binds polyA binding protein (Pab1p) during translation[100]. Longer polyA tails usually stabilize mRNAs and enhance their translation[101]. Transcripts with polyA tail lengths less than twenty exhibit markedly reduced translation[102]. MicroRNAs target and destabilize mRNAs. In T cells, mRNAs with extended 3′UTRs have twice as many microRNA target sites as mRNAs with short 3′UTRs and produce less protein[87]. Most microRNA sites are located after the first polyA site[103]. RNA binding proteins bind predominately to the 3′UTR[104]. In yeast, deletions in many RNA binding proteins resulted in decreased mRNA stability[105]. The RNA binding protein UPF1 accumulates on long 3′UTRs, promoting nonsense-mediated decay[106–108]. Modulation of mRNA stability by RNA binding proteins leads to changes in the amount of protein translated[109].

Primary T cells transfected with reporters encoding extended 3′UTR isoforms produce half as much protein as cells transfected with the shorter isoform of the same mRNA[87]. In cancer cells, shorter isoforms of oncogenes produce ten times more protein than longer isoforms[110]. Neuronal activation leads the long but not short *BDNF* isoform to bind polyribosomes and increase *BDNF* translation, whereas the short *BDNF* isoform mediates basal *BDNF* translation at rest[111]. These findings may be mRNA or cell-type specific, as 3′UTR length does not affect the translation efficiency or stability of most mRNAs in mouse fibroblasts[112]. Finally, alternative splicing and polyadenylation can produce isoforms with different coding sequences, changing the structure and function of the translated protein. Polyadenylation in the coding region of the glutamyl-prolyl tRNA synthetase (*EPRS*) results in a truncated protein that increases the translation of its

targets, whereas the full-length protein represses translation[113]. Thus, isoform 3′UTR length affects protein synthesis and function.

1.2.3 Alternative polyadenylation changes in disease.

3′UTR isoform shifts occur in several diseases. Cancer cell lines exhibit shortening of several mRNA 3′UTRs compared to non-transformed cells lines[110]. Shorter oncogene isoforms contribute to transformation in these cells, indicating 3′UTR shortening may contribute to cancer transformation independently of genetic mutations[110]. In patients with non-anterior uveitis, a single nucleotide polymorphism in the interferon regulatory factor 5 (*IRF5*) mRNA creates a polyadenylation site, resulting in a short isoform that is associated with the development of macular edema[114]. In Parkinson's disease, single nucleotide polymorphisms associated with disease promote formation of the longest isoform of alpha synuclein. Increased levels of the alpha synuclein long isoform result in accumulation of the protein and localization to mitochondria[115]. Research on aberrant polyadenylation in disease is still ongoing, and altered expression of alternatively polyadenylated isoforms may be a feature of many more diseases.

1.3 3′UTR Alternative Polyadenylation in HD

Alternative polyadenylation and splicing happen co-transcriptionally, and changes in epigenetics, transcription, and mRNA expression impact polyA site selection[48,116]. Recent studies have found HD cells are characterized by altered epigenetics, gene expression, and mRNA splicing[41,42,117]. Both mutant HTT protein and mutant *HTT* mRNA contribute to these molecular changes. These findings indicate many processes that impact polyA site selection are deregulated in HD. 3′UTR isoforms can have different localization, stability, translation efficiency, and function. If aberrant alternative polyadenylation of *HTT* or other genes occurs in HD, it may impact mRNA and protein metabolism.

1.3.1 Mutant HTT protein disrupts processes linked to alternative polyadenylation.

The epigenome is altered in HD (Fig 1.2).

Epigenetic changes can alter polyA site selection[69,70]. Early studies showed CREB-binding protein (CBP), a histone acetyltransferase (HAT), is sequestered in HTT aggregates[118]. Sequestration of CBP leads to hypo-acetylation and cell toxicity[119,120]. Further studies revealed global HAT expression is decreased in HD cell models[121]. Hypo-acetylation is restricted to a subset of genes that maintain neuronal identity; across several HD mouse and cell models, neuron-specific genes are hypo-acetylated compared to controls, while total acetylation levels are unchanged[122]. The promoters of the hypo-acetylated genes lack methyl groups on histone H3 lysine 4 (H4K4me3), a marker of transcriptional activation[122]. In mice expressing N-terminal *HTT* (N171-82Q), most

acetylation changes are not associated with transcription changes; however, a subset of hypo-acetylated genes is down-regulated[123]. In striatum of another N-terminal transgenic mouse (R6/1), hypo-acetylated and down-regulated DNA regions include areas that drive transcription of nearby genes involved in maintaining neuronal identity[124]. Some pathways enriched among hypo-acetylated genes in N171-82Q mice are also enriched among differentially expressed genes in HD patient brains, suggesting hypo-acetylation and down-regulation of neuronal genes is a feature of HD pathology[123].

Many genes are differently methylated in ST*Hdh*^{Q111} cells and patient brains compared to controls[125,126]. Aberrant methylation alters expression of a subset of genes, including those involved in regulation of gene expression such as transcription factors and chromatin remodelers[125]. In humans, H3K4me3 profiling identified many peaks differently enriched between patients and controls; however, a later study of total methylation found few differences, potentially because the latter study corrected for cell heterogeneity differences between HD and control brains[125,127]. Further studies are needed to elucidate methylation changes in HD.

Correction of aberrant methylation and acetylation improves phenotypes in HD mouse models, suggesting epigenetic changes cause pathology[41]. In transgenic R6/2 mice and in striatal cells derived from *HTT* knock-in mice (ST*Hdh*^{Q111}), there is hypo-acetylation of specific down-regulated genes; treatment with HDAC inhibitors normalized gene expression and reduced motor phenotypes[128,129]. HDAC inhibitors also exert trans-generational benefits; offspring of male N171-82Q mice treated with HDAC inhibitors exhibit a blunted phenotype[130]. Reducing levels of an H3K4me3 demethylase

in mouse primary neurons transduced with mutant *HTT* normalizes expression of down-regulated genes and reduces disease phenotypes[131]. These studies suggest epigenetic changes contribute to HD pathology by decreasing expression of a subset of genes.

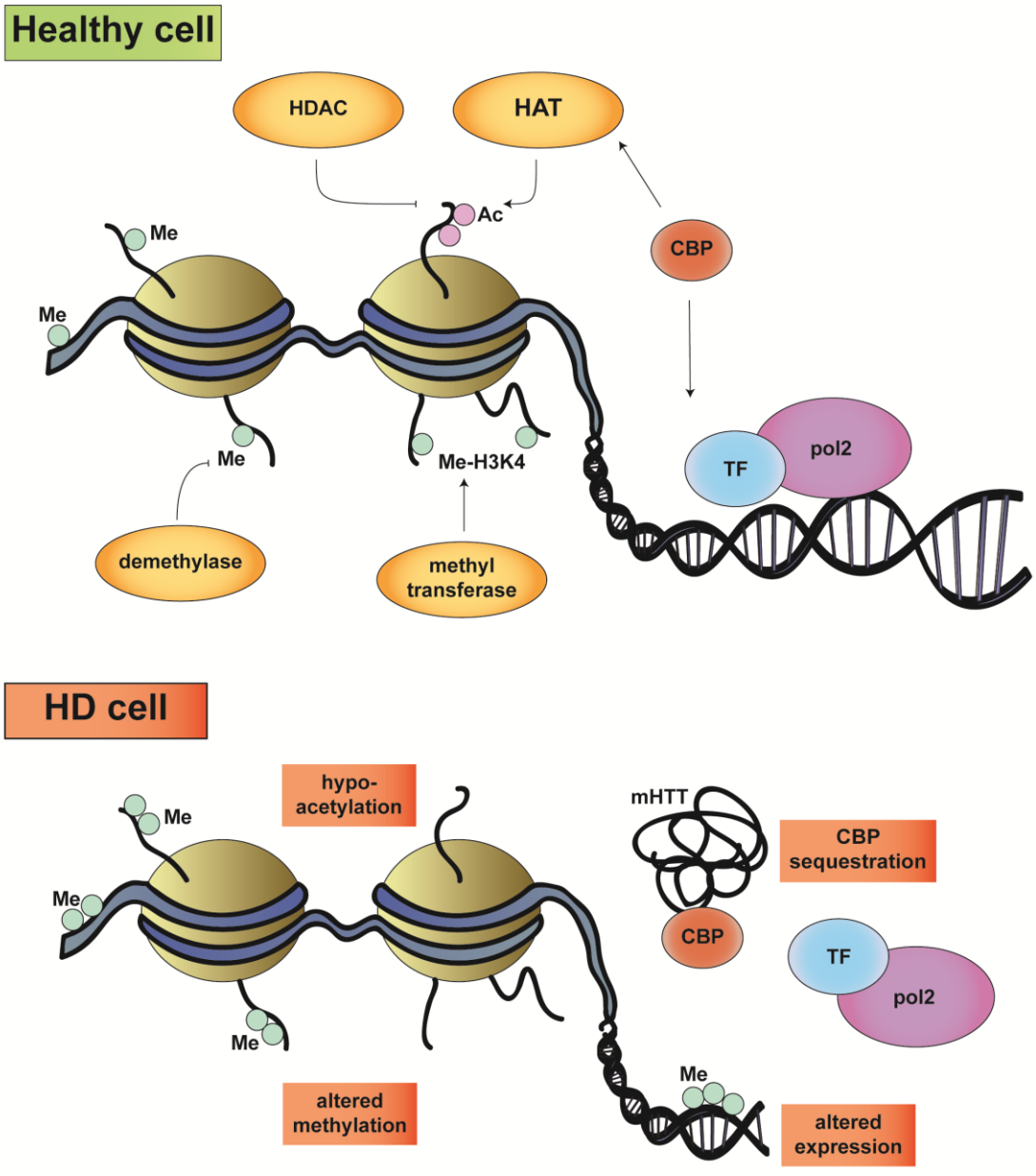


Figure 1.2. Mutant HTT protein exerts epigenetic changes. (Top) In healthy cells, DNA (blue) is wrapped around histones (gold). Repressive methyl groups added to DNA and histones

(Me) by methyltransferases cause DNA to be wrapped tighter, while activating methyl groups such as H3K4 trimethyl cause DNA to be wrapped looser. In contrast, acetylation of histones by histone acetyltransferases (HATs) is always activating, while removal of acetyl groups by histone deacetylases (HDACs) is repressive. Looser DNA is accessible to transcription factors (TFs) and RNA polymerase (pol2), allowing transcription. CBP acetylates histones and recruits transcription factors. **(Bottom)** In HD cells, there is decreased acetylation and H3K4 methylation, and increased repressive methylation of a subset of genes. CBP is sequestered in nuclear mutant HTT inclusions. These changes lead to decreased transcription of a subset of genes in HD.

RNA expression changes in HD.

Transcription and RNA expression impact alternative polyadenylation[116,132]. In HD, epigenetic alterations and aberrant transcription factor binding change mRNA and microRNA expression (Fig. 1.3)[41,123]. Early research found mRNA of the neuropeptides enkephalin and substance P is decreased in HD patient striatum[133]. R6/2 mouse brains also exhibited decreased mRNA for receptors of glutamine, dopamine, acetylcholine, and adenosine compared to controls[134]. Microarrays of striatal mRNAs from pre-phenotypic R6/2 and N171-82Q mice found components of signaling pathways, and inflammation mediators were differentially expressed[135]. Most genes with predominately striatal expression are down-regulated in R6/1 mice and HD patients[136]. Striatal genes differentially expressed in both HD patients and Yac128 mice, which harbor the full-length mutant *HTT* transgene, included pro-inflammatory *Cd4*, interferon-induced *Indo*, and clathrin adaptor complex protein *AP1S1*[137]. However, as neuronal loss is profound in patient striatum, these findings may be due to decreased neurons and increased glia rather than gene expression changes.

Genome-wide studies have revealed widespread mRNA expression changes in HD. Microarrays of mRNA from human HD and control brains found 21%, 3%, 1%, and 0% of genes were differentially expressed in HD striatum, motor cortex, cerebellum, and prefrontal cortex[138]. The magnitude of expression alterations in the striatum correlated with disease grade. Importantly, these changes were not due to neuronal loss, as microarray from laser-captured neurons yielded similar results[138]. The top genes differentially expressed in mouse models with truncated or full-length mutant *HTT* are also differentially expressed in HD patients[139]. This result indicates the expanded polyglutamine repeat changes expression of a subset of mRNAs.

Some changes in mRNA expression in HD are due to aberrant transcription factor binding. Wild-type HTT interacts with the transcriptional repressor REST/NRSF to sequester it to the cytoplasm[140]. That interaction is lost in HD, resulting in increased binding of REST/NRSF to DNA and repression of neuron-specific genes[140–142]. Increased REST activity in HD may also be due to increased REST transcription; R6/2 mice exhibit an Sp1-dependent increase in REST transcription compared to controls[143]. Sp1 is a transcription activator that recruits the transcription initiation factor TFIID to promoters[144]. In contrast to REST, most Sp1 targets are down-regulated in HD. Sp1 and TFIID both interact with full-length HTT[145,146]. This interaction is strengthened with polyglutamine expansion, preventing Sp1 and TFIID from binding target DNA[145]. In HD patient brains, many down-regulated genes have decreased binding to Sp1[147].

Mutant HTT interactions with the TFIID complex prevent cAMP-responsive element-binding protein (CREB) from promoting expression of target genes including

peroxisome proliferator-activated receptor gamma coactivator-1 α (PGC-1α), a transcriptional activator of metabolic genes[148]. Overexpression of TAF4, a subunit of TFIID, ameliorates transcriptional deregulation and cell death in HD cell models[149]. TATA binding protein (TBP), another component of TFIID, accumulates in HD patient brains[150]. Trinucletide expansions within the TBP gene produce a syndrome indistinguishable from Huntington's disease, suggesting impaired TBP contributes to HD pathology[151]. Mutant HTT traps CBP, a protein that associates with CREB to activate transcription[118,120]. The mutant HTT N-terminus binds p53, decreasing transcription of p53 target genes[118,146]. In contrast, binding between N-terminal mutant HTT and nuclear receptor co-repressor (N-CoR) may lead to increased transcription of some genes, as N-CoR is a transcriptional repressor[152].

RNA expression changes in HD extend to microRNAs (Fig. 1.3). HTT colocalizes in P-bodies with argonaute 2 (Ago2), and ST*Hdh*^{Q111} cells have compromised gene silencing[153]. Nuclear REST reduces expression of miR-124a, an inhibitor of non-neuronal mRNA expression[154]. Decreased transcription of miR-124a in HD contributes to loss of neuronal identity[155]. Other microRNA genes targeted by REST are down-regulated in HD[156]. Two of these, miR-9 and 9*, target REST and its cofactor, coREST; decreased expression of miR-9 and 9* in HD may disinhibit the REST complex in a feed-forward loop[155]. Another, miR-146a, targets TBP, a transcription factor implicated in HD pathology[150,157]. Genome-wide studies from human frontal cortex and dorsal caudate found 150 microRNAs are differentially expressed, many of them REST targets and some with target sites in genes known to be differentially expressed in HD[158].

Normalizing the levels of some microRNAs reduces pathology in HD models. *HTT*-targeting microRNAs 150, 146a, and 125b are reduced in ST*Hdh*^{Q111} cells; transfection of these microRNAs decreases aggregates and increases cell viability[159]. MiR-34b is elevated in pre-symptomatic HD patient plasma; knockdown of miR-34b with antisense oligonucleotides alters the toxicity of mutant HTT in cells[160]. Overexpression of miR-22, which is down-regulated in HD patients, led to decreased caspase activity and improved cell viability in primary neurons exposed to mutant HTT[161]. In contrast, increased expression of miR-196a in HD may be neuroprotective; overexpression of miR-196a improves phenotypes in mouse models[162].

These studies measured steady-state mRNA and microRNA levels. However, the correlation between altered mRNA expression and aberrant transcription factor binding indicates some mRNA expression differences are due to transcription changes[42]. As transcription is linked to alternative polyadenylation, these findings suggest alternative polyadenylation may change in HD. In addition, aberrant mRNA levels in HD may change protein abundances. The mRNA expression of 3′-end processing proteins is tightly linked to cell proliferative potential, suggesting mRNA expression and protein expression of these genes are correlated[45]. Many mRNAs are differentially expressed in HD; if the expression of 3′-end processing factors changes, it may cause widespread shifts in alternative polyadenylation.

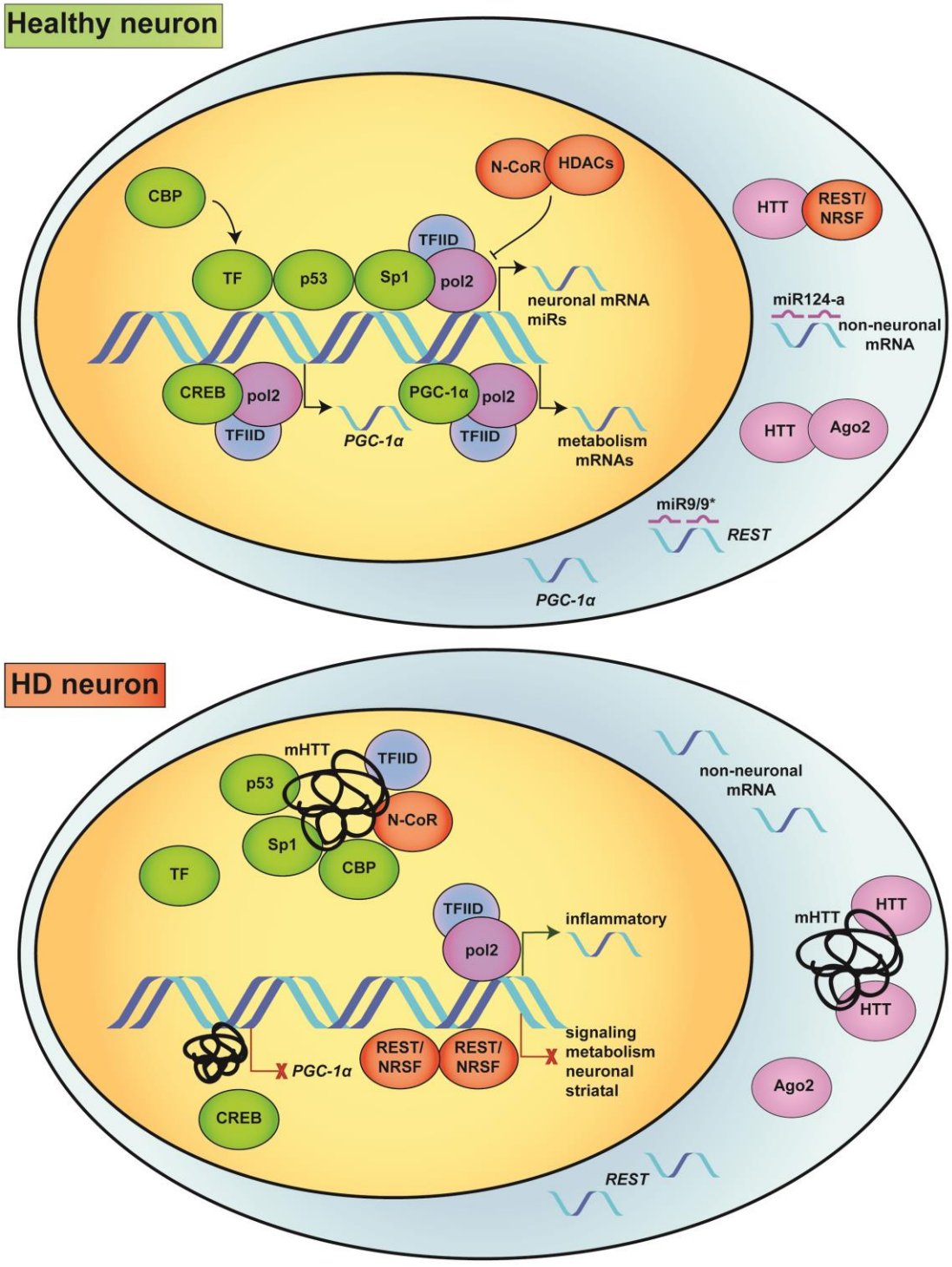


Figure 1.3. Mutant HTT protein impacts RNA expression. (Top) In healthy neurons, several factors interact to modulate transcription. In the cytoplasm, wild-type HTT sequesters the

transcription repressor REST/NRSF, allowing target neuronal mRNAs and microRNAs (miRs) to be transcribed. Wild-type HTT associates with Ago2 in P-bodies to modulate microRNA-mediated mRNA silencing. MicroRNA 124a silences non-neuronal mRNAs, and microRNA 9/9* silence REST mRNA. In the nucleus, activators such as CBP recruit transcription factors (TFs) such as p53 and Sp1 to DNA. Transcription factors recruit TFIID and RNA pol2, resulting in transcription of neuronal mRNAs and microRNAs. The CREB transcription factor promotes transcription of PGC-1α, which promotes transcription of metabolism mRNAs. The nuclear repressor N-CoR interacts with HDACs to inhibit transcription. **(Bottom)** In the cytoplasm of HD neurons, mutant HTT aggregates sequester wild-type HTT, allowing REST/NSRF to enter the nucleus and altering mRNA silencing via Ago2. Reduced levels of microRNAs 124a and 9/9* result in de-repression of REST and non-neuronal mRNAs. In the nucleus, REST/NRSF binds target genes and HTT inclusions sequester transcription factors p53 and Sp1, the transcription activator CBP, and the transcription initiation complex TFIID. This results in decreased transcription of genes involved in signaling and neuronal and striatal genes. Mutant HTT prevents CREB from activating transcription of PGC-1α, reducing transcription of metabolic genes. The repressor N-CoR is sequestered in HTT inclusions, and inflammatory genes are more highly expressed.

RNA processing is aberrant in HD.

Alternative polyadenylation co-occurs with mRNA splicing. Deep sequencing of RNA from the motor cortex of HD patients revealed 593 splicing events that differed between patients and controls[117]. Splicing changes may be due to altered expression of splicing factors (Fig. 1.4). Four splicing factors—PTBP1, SFRS4, RBM4, and SREK1—were differentially expressed in HD. Of these, PTBP1 showed the highest correlation between its expression and splicing events of its target genes[117]. Other splicing changes in HD patient brains may be due to altered levels of serine/arginine-rich splicing factor-6 (SRSF6)[163]. SRSF6 binds the expanded CAG repeat and may promote mis-splicing of

HTT into the disease-associated isoform terminating in intron 1[40]. SRSF6 modulates the splicing of tau protein, and increased levels of SRSF6 in HD lead to increased production of the tau splice isoform associated with neurodegeneration[163–165]. Altered SRSF6 levels also result in missplicing of the primary microtubule associated protein in dendrites, MAP2[166]. These studies show several processes linked to alternative polyadenylation in the cell— chromatin relaxation, RNA expression, and RNA splicing— are aberrant in HD.

1.3.2 Mutant *HTT* mRNA disrupts processes linked to alternative polyadenylation.

Like mutant HTT protein, mutant *HTT* mRNA causes deregulation of mRNA splicing (Fig. 1.5)[43]. Triplet repeat RNA regions form stable hairpin structures[167]. Crystal structures of CAG repeat RNA reveal a helical duplex, and in vitro nuclease analysis of *HTT* mRNA found the CAG repeats form a hairpin stabilized by the neighboring CCG repeats[32,168]. RNA hairpins can bind and sequester proteins, causing toxicity[167].

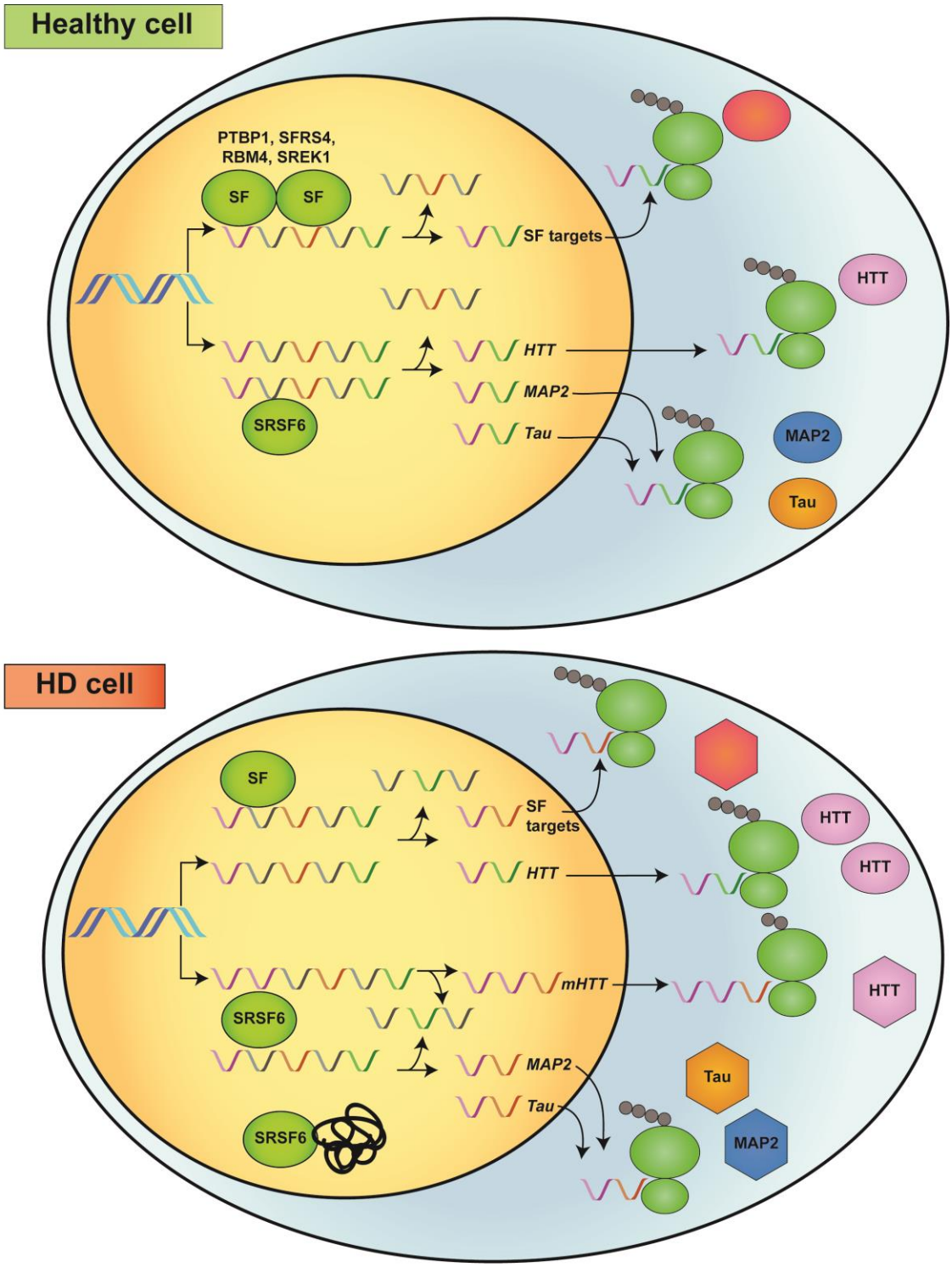


Figure 1.4. Mutant HTT protein impacts RNA processing. (Top) Splicing factors (SF) bind to nascent mRNA transcripts and mediate splicing and removal of introns (gray) and alternative

exons (colored). The splicing factor SRSF6 modulates splicing of *MAP2* and *Tau*. Correctly spliced mRNAs are translated in the cytoplasm into their canonical protein isoforms (ellipses). **(Bottom)** In HD cells, splicing factors including PTBP1, SFRS4, RBM4, and SREK1 have decreased expression, resulting in missplicing of target mRNAs. SRSF6 accumulates in HTT inclusions, and increased SRSF6 levels alter MAP2 and Tau splicing. Misspliced transcripts may be translated into different protein isoforms (hexagons). Most *HTT* mRNA is spliced correctly (ellipses); however, SRSF6 binds to some mutant but not wild-type *HTT (mHTT)*, producing a truncated HTT protein (hexagon).

The atomic structure of CAG repeats and CUG repeats is similar[168]. In type 1 myotonic dystrophy, CUG repeats in the untranslated region of *DMPK* mRNA form nuclear foci that bind and sequester the splicing protein musclebind (MBNL1), leading to extensive splicing changes[169–171]. Most of the pathology in myotonic dystrophy can be explained by sequestration of MBNL1[172]. Although the trinucleotide repeat in *DMPK* is much longer than the repeat in *HTT*, The similarity between the structures of CAG and CUG repeats prompted researchers to study whether mutant *HTT* mRNA sequesters MBNL1. Indeed, expanded *HTT* mRNA forms nuclear foci that colocalize with MBNL1[32]. Expression of long untranslated CAG-repeat RNA can induce MBNL1-rescued splicing changes, and HD patient fibroblasts exhibit splicing changes similar to normal cells transfected with MBNL1 siRNAs[173]. These findings suggest *HTT* mRNA transcripts sequester MBNL1, leading to aberrant splicing.

Interactions between expanded *HTT* mRNA and MBNL1 induce toxicity. Expression of non-translated CAG repeats in *Drosophila* causes neurodegeneration[174]. Transgenic mice that express 200 CAG repeats in the 3′UTR of EGFP exhibit muscle pathology due to MBNL1 sequestration, similar to that in myotonic dystrophy[175].

Neuroblastoma cells transfected with an non-translated N-terminal fragment of *HTT* suffer cytotoxicity comparable to that in cells transfected with the same protein fragment[176]. Transfection of full-length mutant *HTT* into these cells results in nuclear foci that increase with overexpression of MBNL1[176]. Researchers deleted the start codon or inserted repeats in a gene 3′UTR to create these non-translated RNAs. **R**epeat-**a**ssociated **n**on-AUG (RAN) translation occurs in HD, so some toxicity may be attributable to RAN products[177]. However, most experiments found no detectible protein produced by the constructs[175,176]. These studies suggest interactions between mutant *HTT* mRNA and MBNL1 induce toxicity, as occurs in myotonic dystrophy.

Interactions between MBNL1 and Mutant *HTT* mRNA may directly alter alternative polyadenylation. As described, MBNL is sequestered by triplet repeat mRNAs[32,169]. In addition to modulating alternative splicing, MBNL proteins bind the 3′UTR of target mRNAs; depletion of Mbnl in mouse embryo fibroblasts results in widespread aberrant alternative polyadenylation events[178]. In myotonic dystrophy mouse models and patients, misregulation of alternative polyadeylation leads to a persistent neonatal polyadenylation profile[178]. This study indicates triplet repeat expansion mRNAs can cause widespread changes in alternative polyadenylation via interactions with RNA binding proteins[178]. *HTT* sequestration of MBNL1 may result in aberrant polyadenylation.

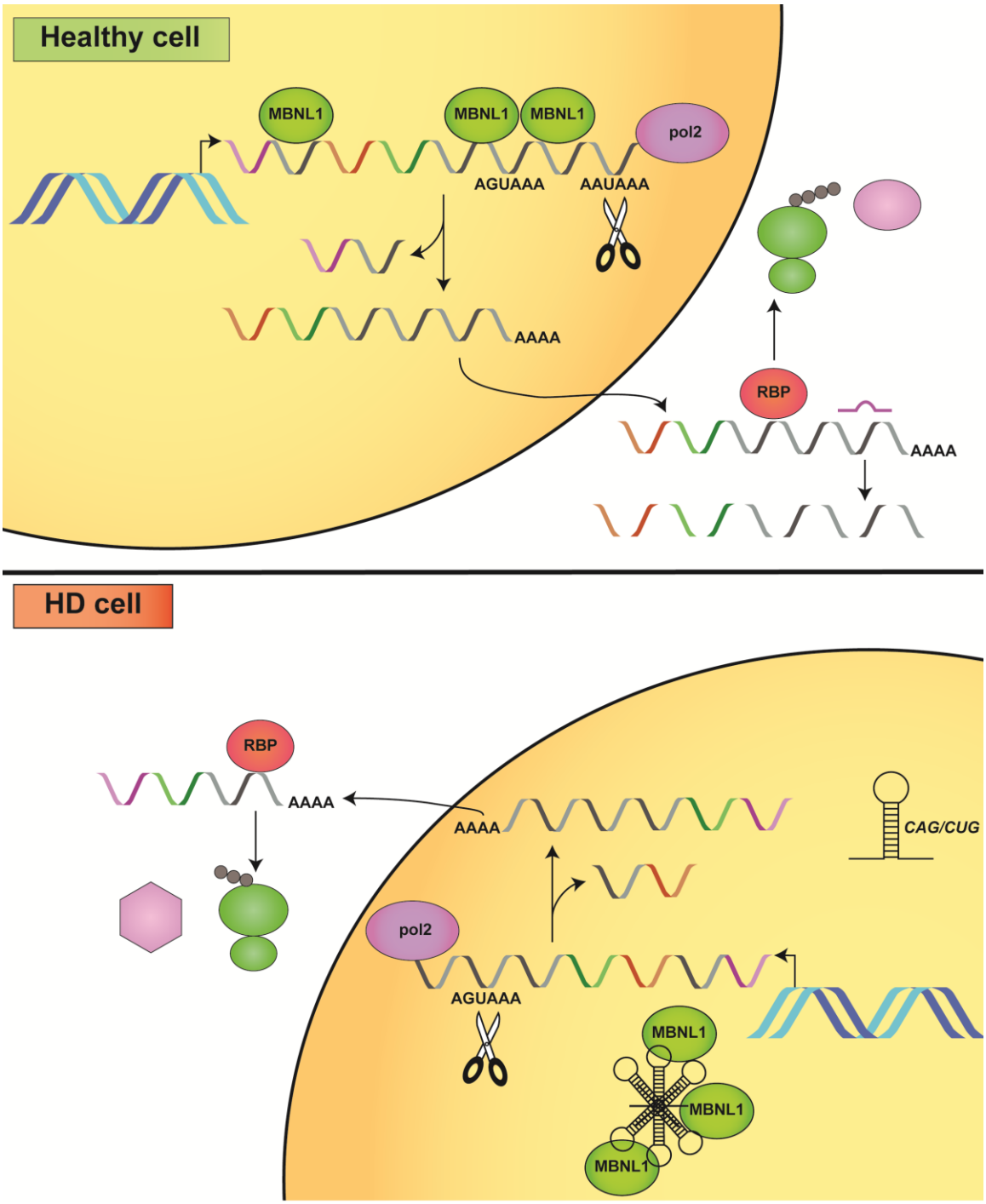


Figure 1.5. Trinucleotide-repeat RNAs impair splicing and alternative polyadenylation. (Top) In healthy cells, MBNL1 associates with mRNA in the ORF and 3′UTR to modulate splicing and alternative polyadenylation. **(Bottom)** In disease cells expressing trinucleotide repeat expansions, such HD and myotonic dystrophy neurons, repeat RNAs form hairpins that aggregate

into nuclear foci and sequester MBNL1. Sequestration of MBNL1 causes missplicing and aberrant polyadenylation of MBNL1 target transcripts. Altered splicing can result in aberrant protein isoforms (hexagon), and altered alternative polyadneylation can impact mRNA stability, localization, and translation via differences in *cis* elements such as destabilizing motifs and RNA binding protein (RBP) and microRNA target sites.

1.4 Conclusions

It has been known since the early 1990s that *HTT* mRNA is processed into two 3′UTR isoforms[179]. Since then, studies have found the 3′UTR directs mRNA localization, stability, translation, and function. During preparation of this dissertation, a study found the long *HTT* 3′UTR localized mutant exon 1 constructs to dendrites, whereas the short *HTT* 3′UTR constructs formed more aggregates, suggesting *HTT* 3′UTR length affects HD pathogenesis[39]. The recent finding that *HTT* is prematurely cleaved and polyadenylated in intron 1 in HD patient brains shows there are changes in *HTT* alternative polyadenylation in HD[40]. Gene expression and splicing, which occur at the same time as alternative polyadenylation, are deregulated in HD[48,117,138,180]. Although these studies show that *HTT* 3′UTR isoforms may differ in their pathogenicity, that *HTT* alternative polyadenylation changes in HD, and that processes linked to alternative polyadenylation are disrupted in HD, no study has explored the abundance or metabolism of 3′UTR isoforms in disease cells.

I hypothesized that alternative polyadenylation is disrupted in HD. Here, I provide evidence that the abundance of *HTT* mRNA 3′UTR isoforms differs between HD patients and non-HD controls, including a novel mid-length 3′UTR isoform. I show that *HTT* mRNA 3′UTR isoforms have different localizations, half-lives, polyA tail lengths, microRNA sites, and RNA binding protein sites. Isoform alterations in HD are not unique to *HTT*. There are widespread changes in mRNA 3′UTR isoform expression in human HD motor cortex. Genes with isoform changes are associated with pathways dysfunctional in HD. I demonstrate that knockdown of the RNA binding protein CNOT6

leads to isoform amounts similar to those in HD motor cortex. My findings indicate that altered 3′UTR isoform expression of *HTT* and a subset of other genes is a feature of molecular pathology in HD motor cortex.

CHAPTER II: THE ABUNDANCE OF *HTT* MESSENGER RNA ISOFORMS CHANGES IN HD

Preface

The work presented in this chapter is accepted at *Cell Reports* as manuscript "Alterations in mRNA 3′UTR isoform abundance accompany gene expression changes in human Huntington's disease brains" by Romo, Ashar-Patel, Pfister, and Aronin.

This work was a collaborative effort. Caleigh Smith and Stevie Yang assisted with optimizing the allele-specific qPCR. Faith Conroy, and Rachael Miller performed mouse genotyping. Ami Ashar-Patel developed the PAS-seq protocol. Yasin Kaymaz provided advice on PAS-seq data analysis. Alicia Bicknell provided the ethinyl-uridine pulse chase protocol. Brian Quattrochi provided the universal primer sequence for isoform-specific qPCR. Julia Alterman provided the siRNA to the *HTT* open reading frame. The DiFiglia lab (Massachusetts General Hospital) provided HD and control neural stem cells.

2.1 Summary

There are three known isoforms of *HTT* mRNA. A truncated (7.9 kb) isoform is generated via termination at a cryptic polyadenylation (polyA) signal in the first intron[40]. This isoform is of low abundance. Its presence in HD patients and mouse models, but not controls, suggests it may be associated with disease[40]. The two predominant *HTT* mRNA isoforms are generated by alternative polyadenylation in the 3′UTR[37]. The short (10.3kb) 3′UTR isoform predominates in dividing cells such as lymphoblasts[37]. The long (13.7kb) isoform predominates in nondividing cells such as neurons[37]. It has not been established whether the abundance or metabolism of *HTT* mRNA 3′UTR isoforms changes in HD. Although many genes exhibit expression and splicing differences in HD, 3′UTR isoform expression changes have not been investigated[34,138,180].

Quantitative PCR of the *HTT* long isoform revealed the abundance of *HTT* mRNA 3′UTR isoforms differs between human HD and control motor cortex, cerebellum, fibroblasts, and neural stem cells. Isoform shifts are tissue-specific, occur in at least two peripheral tissues, and arise from both the mutant and the wild-type allele. PolyA site sequencing identified a novel mid-length *HTT* 3′UTR isoform that is conserved in mice and also changes abundance in HD. I found the short, mid, and long *HTT* isoforms have different localizations, half-lives, polyA tail lengths, RNA binding protein sites, and microRNA sites. These results further our understanding of *HTT* mRNA metabolism and provide recommendations for the design of allele-specific HD therapies.

2.2 Results

2.2.1 The abundance of *HTT* mRNA 3′UTR isoforms changes in Huntington's disease.

I identified two abundant *HTT* mRNA 3′UTR isoforms in public 3′ sequencing data (Fig. 2.1A)[84]. As in previous studies, I found *HTT* mRNA isoform abundance varies across normal tissues: the longer isoform predominates in the brain, neuronal precursor (NTERA2) cells, breast, and ovary, whereas the shorter isoform predominates in testes, B-cells, muscle, and human embryonic kidney (HEK293) cells (Fig. 2.1B)[37].

To determine whether *HTT* mRNA isoform abundance changes in HD, I performed RT-qPCR of the *HTT* long isoform on patient cerebellum, motor cortex, fibroblasts, and neural stem cells normalized to total *HTT* expression. HD cerebellum samples were from grade 2, 3, or 4 brains, whereas motor cortex samples were from grade 1 and 2 brains (Table 2.1). HD brain grades are based on the degree of striatal degeneration and gliosis. Grades range from 1 (50% neuronal loss) to 4 (95% of neuronal loss)[181]. There is limited involvement of the cerebellum in all grades, although cerebellar atrophy and sporadic Purkinje cell loss is seen in grades 3-4[30,181,182]. Cortical neuronal loss and gliosis are variable between samples[181,183]. Typically, 10% of cortical neurons are lost in grade 1 with minimal gliosis, and up to 40% are lost in grade 2 with evident gliosis [181,183]. Neuronal loss can confound gene expression studies. I chose to analyze *HTT* isoform expression in cortex and cerebellum rather than in the more- affected striatum because neuronal loss is more limited in these regions.

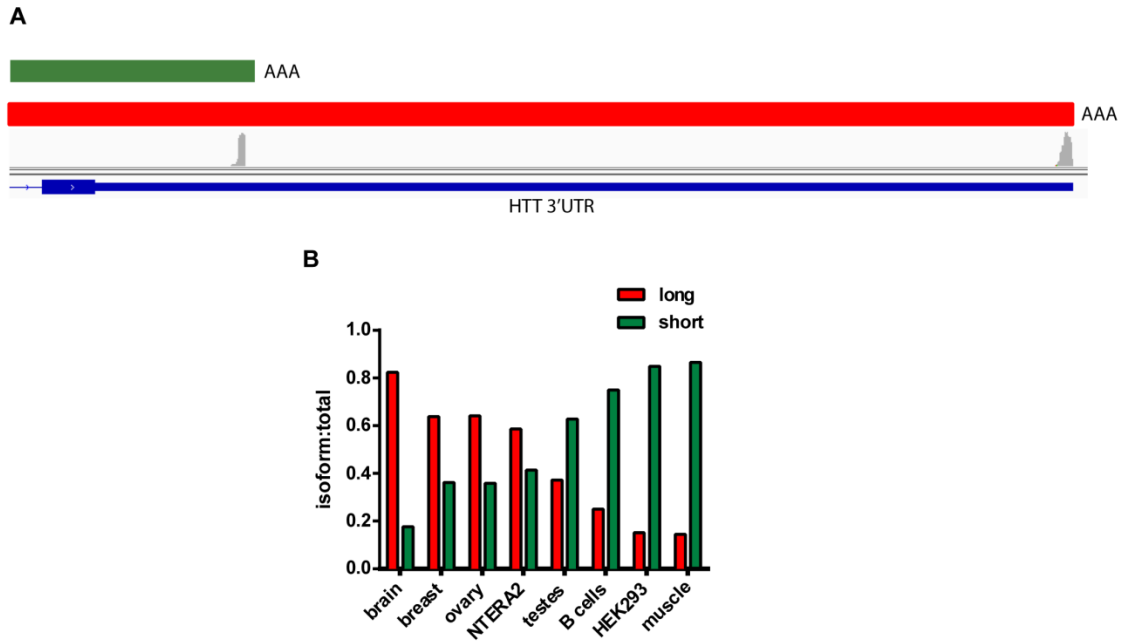


Figure 2.1. *HTT* 3′UTR isoforms have tissue-specific abundance.

(A) NTERA2 cell 3′seq reads mapping to the *HTT* 3′UTR. (B) Quantification of *HTT* 3′UTR isoform 3′seq reads relative to total *HTT* 3′UTR isoform reads across several human tissues and cell lines.

Table 2.1. Human samples used for qPCR and PAS-Seq.

Sample	origin	region	age	sex	CAG #	diagnosis	RIN	PMD (hr:min)
T164	NY	***CB***	41	M			6.8	7:01
T2010	NY	***CB***	92	F			6	14:40
T190	NY	***CB***	92	F			4.7	4:00
H109	NZ	MCx	81	M	15/18		7	7
H122	NZ	MCx	72	F	15/17		7.2	9
H123	NZ	MCx	78	M	17/20		7	7.5
H137	NZ	MCx	77	F	18/20		7	12
H148	NZ	MCx	64	M	17/18		6.3	7
H150	NZ	MCx	78	M	16/22		5.6	12
T343	NY	***MCx***	62	M			6.2	8:16
T144	NY	***MCx, CB***	88	F		ALS	6.2, 5.3	
T139	NY	***MCx***, ***CB***	57	M			6.1, 6	
T551	NY	***MCx, CB***	77	F			6.4, 7.2	13:34
T1488	NY	***MCx, CB***	88	F		AD	6.3, 6.3	12:30
T145	NY	***MCx***, CB	57	F			5, 5.2	9:14
T1327	NY	***MCx***, CB	72	F		AD	4.4, 6.3	25:55
T128	NY	***CB***	50	M		HD 4/4	7.8	41:25
T310	NY	***CB***	49	F	?/42	HD 4/4	7.4	8:30
T550	NY	***CB***	82	F		HD 2/4	3	19:29
T269	NY	CB	61	F	?/47	HD 4/4	6.1	23:00
T150	NY	CB	60	M	?/46	HD 4/4	7.2	12:45
T1991	NY	CB	53	M	?/46	HD 3/4	8	11:17
HC086	NZ	***MCx***	46	M		HD 1/4	2.4	18
HC132	NZ	***MCx***	32	M		HD 1/4	5.2	14
HC074	NZ	MCx	42	F	17/42	HD 1/4	3.1	11
HC051	NZ	MCx	58	M	16/43	HD 1/4	5.6	4.5
HC081	NZ	MCx	70	F	19/41	HD 1/4	7	8
HC083	NZ	MCx	80	M	20/40	HD 1/4	2.7	9
HC092	NZ	***MCx***	72	M	17/41	HD 1/4	6.9	5
HC103	NZ	***MCx***	41	M	19/39	HD 1/4	3.6	11
HC105	NZ	***MCx***	67	F	15/42	HD 1/4	7.5	9
HC132	NZ	MCx	32	M	17/47	HD 1/4	2.5	14
HC137	NZ	***MCx***	83	M	17/42	HD 1/4	4	13
HC061	NZ	MCx	65	M	18/46	HD 2/4	6.4	6
HC062	NZ	*MCx*	78	F	18/43	HD 2/4	5.7	6
HC069	NZ	MCx	50	F	16/46	HD 2/4	2.9	20
HC076	NZ	MCx	71	M	19/42	HD 2/4	3.2	16
HC080	NZ	MCx	45	F	24/43	HD 2/4	6.5	15
HC088	NZ	MCx	57	F	17/44	HD 2/4	2.6	19

HC111	NZ	MCx	91	F	15/40	HD 2/4	2.4	18
HC113	NZ	MCx	58	M	28/44	HD 2/4	6.4	14
HC114	NZ	MCx	53	F	21/47	HD 2/4	6.6	12
HC115	NZ	MCx	56	M	16/46	HD 2/4	4	16
HC120	NZ	MCx	51	M	10/46	HD 2/4	7	15
HC133	NZ	MCx	65	M	17/43	HD 2/4	4.9	14
T273	NY	***MCx, CB***	67	M	?/43	HD 2/4	6.4, 6.1	32:35
T289	NY	***MCx, CB***	54	F		HD 2/4	6.1, 6.2	24:30
T309	NY	MCx, ***CB***	80	F	?/41	HD 2/4	4.9, 6.5	9:25
T461	NY	MCx, ***CB***	77	M	?/42	HD 2/4	5, 6.1	19:28
T3049	NY	***MCx, CB***	72	M		HD 2/4	6, 6	40:50
T1482	NY	***MCx, CB***	76	F		HD 2/4	6.3, 6	61:55

average age	**M/F**
73.50	7/9
60.72	20/16

Samples were obtained from the New York (NY) and the Neurological Foundation of New Zealand (NZ) brain banks. Samples highlighted in red are HD while those with blue are controls. Purple text corresponds to cerebellum while green text corresponds to motor cortex. Samples in ***bold italics*** were also used for PAS-Seq. HD=Huntington's disease, ALS=amyotrophic lateral sclerosis, AD=Alzheimer's disease, M=male, F=female.

I found a significant (p=0.01) 1.6-fold increase in the long 3′UTR isoform relative to total *HTT* mRNA in patient cerebellum. Conversely, there is a significant (p=0.00003) 2.5-fold decrease in the long isoform in patient motor cortex (Fig. 2.2A). Patient fibroblasts and neural stem cells also exhibit a significant decrease in the long isoform (Fig. 2.2B, 2-fold, p=0.0003; and 3.3-fold, p=0.00004 respectively). These results demonstrate the abundance of *HTT* mRNA 3′UTR isoforms changes in HD in a tissue and cell-specific manner.

2.2.2 *HTT* mRNA isoform changes extend to liver and muscle and arise from both alleles.

I used HD model mice to determine if disease-associated isoform changes extend beyond brain. I bred mice that lack murine *Htt*, but harbor full length human *HTT* with either 18 or 128 CAG repeats on a yeast artificial chromosome (Yac18, Yac128)[16,184]. Like patients, Yac128 cerebellum exhibits an increase (1.7-fold, p=0.05) in the long *HTT* mRNA isoform compared to yac18 mouse cerebellum. Isoform differences extend to muscle (quadriceps) and liver, with a 2.2-fold increase (p=0.007) and a 1.6-fold decrease (p=0.03) in the *HTT* long mRNA isoform respectively (Fig. 2.3). These findings indicate that at least two non-brain tissues exhibit tissue-specific *HTT* mRNA isoform changes.

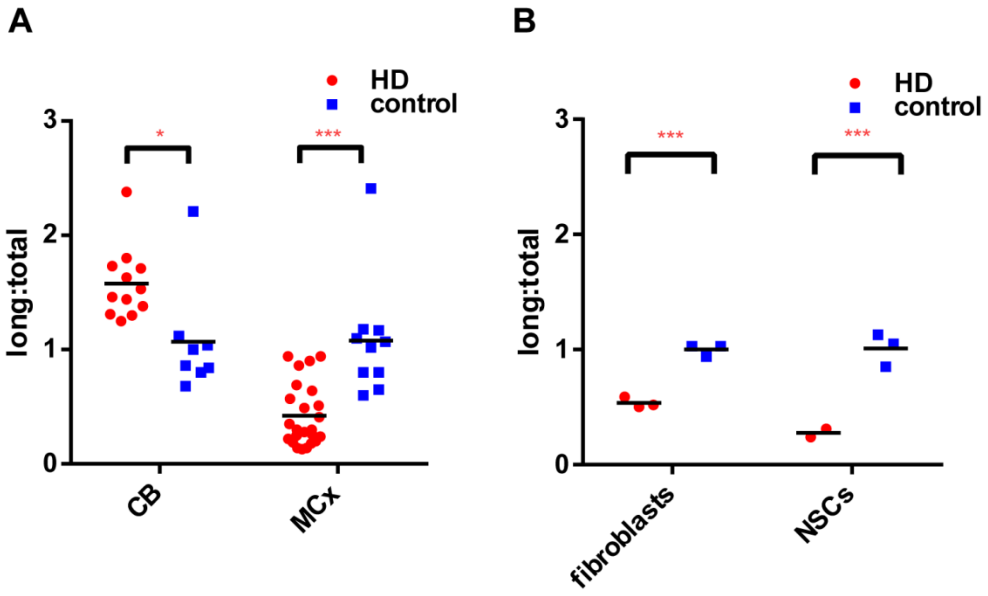


Figure 2.2. *HTT* 3′UTR isoforms change their relative amounts in HD.

(A) qPCR of the *HTT* mRNA long 3′UTR isoform (long) normalized to total *HTT* expression (total) in HD patient and control cerebellum (CB) and motor cortex (MCx). *, **, and *** signify $p<0.05$, 0.005, and 0.0005. (B) qPCR as in C, using patient or control fibroblasts or neural stem cells (NSCs) differentiated from fibroblasts.

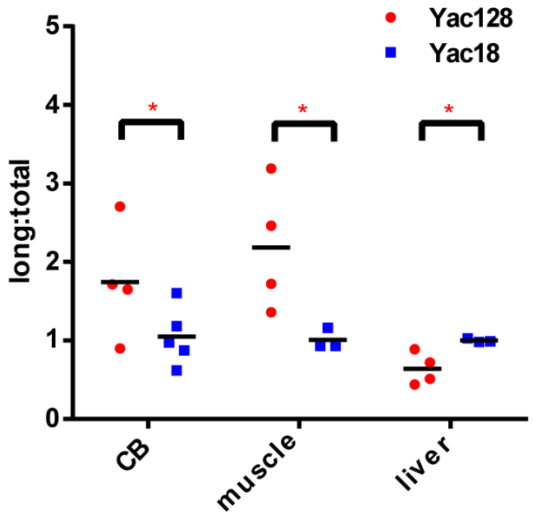


Figure 2.3. *HTT* 3′UTR isoform changes extend to muscle and liver in an HD transgenic mouse model.

Quantitative PCR of the *HTT* long 3′UTR isoform (long) normalized to total *HTT* expression (total) across Yac128 and Yac18 mouse tissues. CB=cerebellum. *, **, and *** signify $p<0.05$, 0.005, and 0.0005.

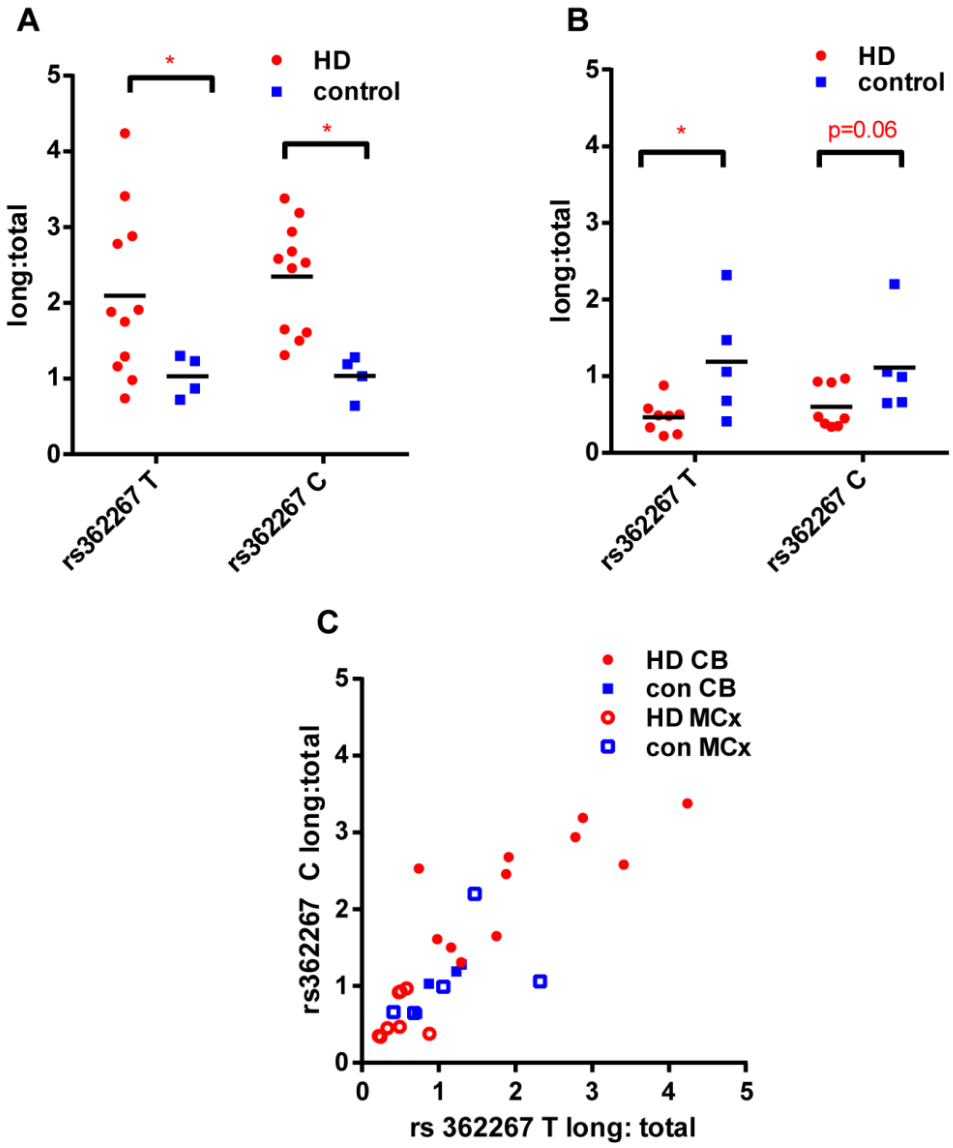


Figure 2.4. Both *HTT* alleles change their 3′UTR isoform amounts in HD.

(A) Allele-specific qPCR using SNP rs362267C/T in the *HTT* long 3′UTR isoform (long) normalized to total *HTT* expression (total) in patient and control cerebellum. (B) Allele-specific qPCR in patient and control motor cortex, as in A. (C) Correlation of *HTT* allele abundance within each HD or control (con) cerebellum (CB) or motor cortex (MCx) sample; r=0.83.

Altered *HTT* mRNA isoform abundance in HD patient brains could arise from the wild-type allele, the mutant allele, or both. To determine which allele is responsible for the changes, I performed allele-specific qPCR on human cerebellum and motor cortex samples heterozygous for single nucleotide polymorphism (SNP) rs362267 in the *HTT* long, but not short, 3′UTR. This SNP heterozygosity is not linked to the CAG repeat; in some patients, the C allele is expanded, whereas in others the T allele is expanded. The long isoform qPCR probe matched either the C or the T SNP. These probes efficiently and specifically amplify the matching but not the nonmatching allele. The amounts of both mutant and wild-type long *HTT* isoform are changed in HD cerebellum (p=0.01, 0.04) and motor cortex (p=0.01, 0.06) compared to controls (Fig. 2.4A,B). The direction of isoform changes is consistent between alleles in each patient (Fig. 2.4C, correlation coefficient 0.83). Thus the disease-associated change in *HTT* mRNA isoforms arises from both alleles.

2.2.3 PolyA site sequencing identified a novel conserved mid-3′UTR isoform of *HTT* mRNA whose abundance changes in disease.

To identify unknown 3′UTR isoforms of *HTT*, I used a new **p**oly**A** **s**ite **seq**uencing method, PAS-seq (Fig. 2.5). PAS-seq requires minimal RNA input to produce high quality sequences, or reads. A high percentage (~95%) of these reads map to the genome (Fig. 2.6). I observe good inter-sample correlation using this method (average Pearson's r=0.88). The PAS-seq reverse transcription primer can anneal to mRNA polyA tails and polyA sequences encoded in the genome. I excluded

genomically-primed reads from my analysis (Fig. 2.5B, see methods)[185].

Using PAS-seq, I identified a novel 12.5kb isoform of *HTT* expressed in both control and disease human cerebellum and motor cortex with abundance similar to the short 3′UTR isoform (Fig. 2.7A). The mid isoform has the features of a true polyA site and is not due to internal polyA priming (see methods, CleanUpdTSeq). I performed PAS-seq on mice to test if the mid 3′UTR isoform is conserved. I used wild-type and Q140 HD mice, which harbor 140 CAG repeats knocked into the mouse *HTT* genomic locus[186]. The mid-3′UTR isoform, and the short and long 3′UTR isoforms, are present in wild-type and Q140 mice (Fig. 2.7A). A polyA signal is required for mRNA cleavage and polyadenylation. There is a putative polyA signal (AAUGAA) located twenty nucleotides upstream of the mid-3′UTR isoform polyA site that has lower in vitro polyA efficiency than the short and long isoform signals[38,187]. This polyA signal and site as well as the surrounding regions are conserved between mice and humans (Fig. 2.7B,C). I re-analyzed public 3′ sequencing data to quantify the mid isoform[188]. Like the short and long isoforms, the mid isoform exhibits tissue-specific expression, although it is less than 3% of total HTT in all tissues (Fig. 2.7D).

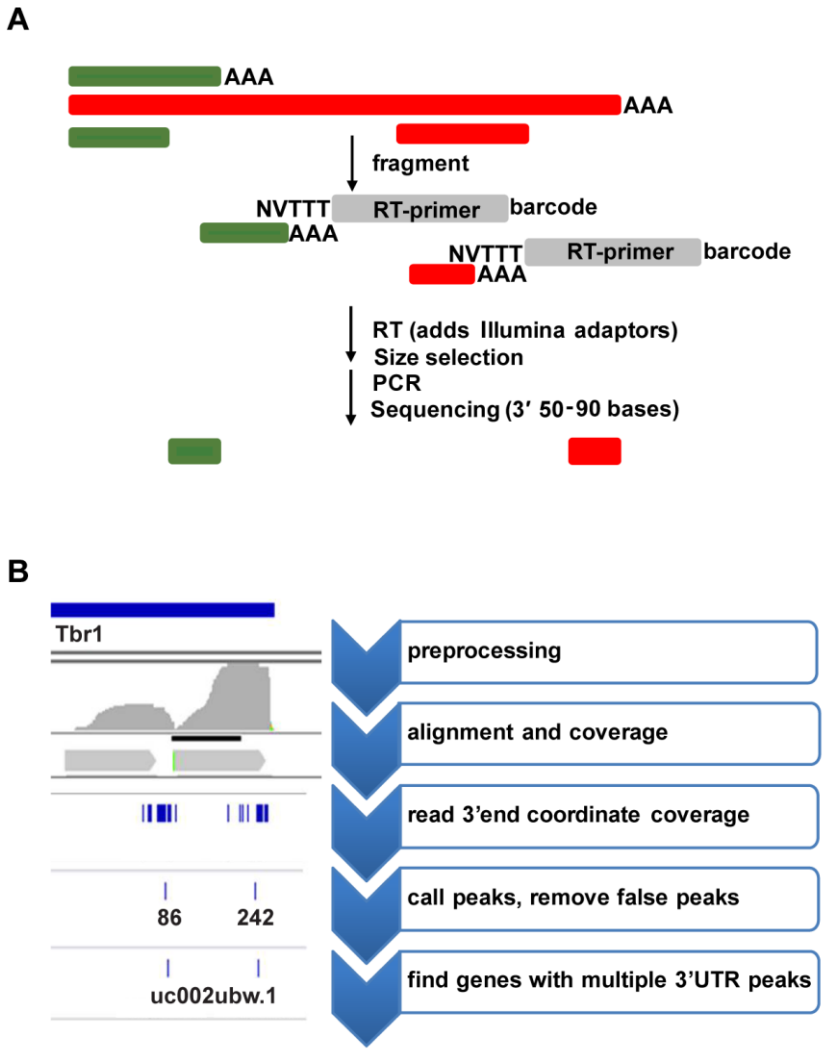


Figure 2.5. PAS-seq assays global isoform abundance. (A) Schematic of PAS-seq library preparation. (B) Schematic of the PAS-seq data analysis pipeline on Tbr1, an exemplary gene with two 3′UTR isoforms.

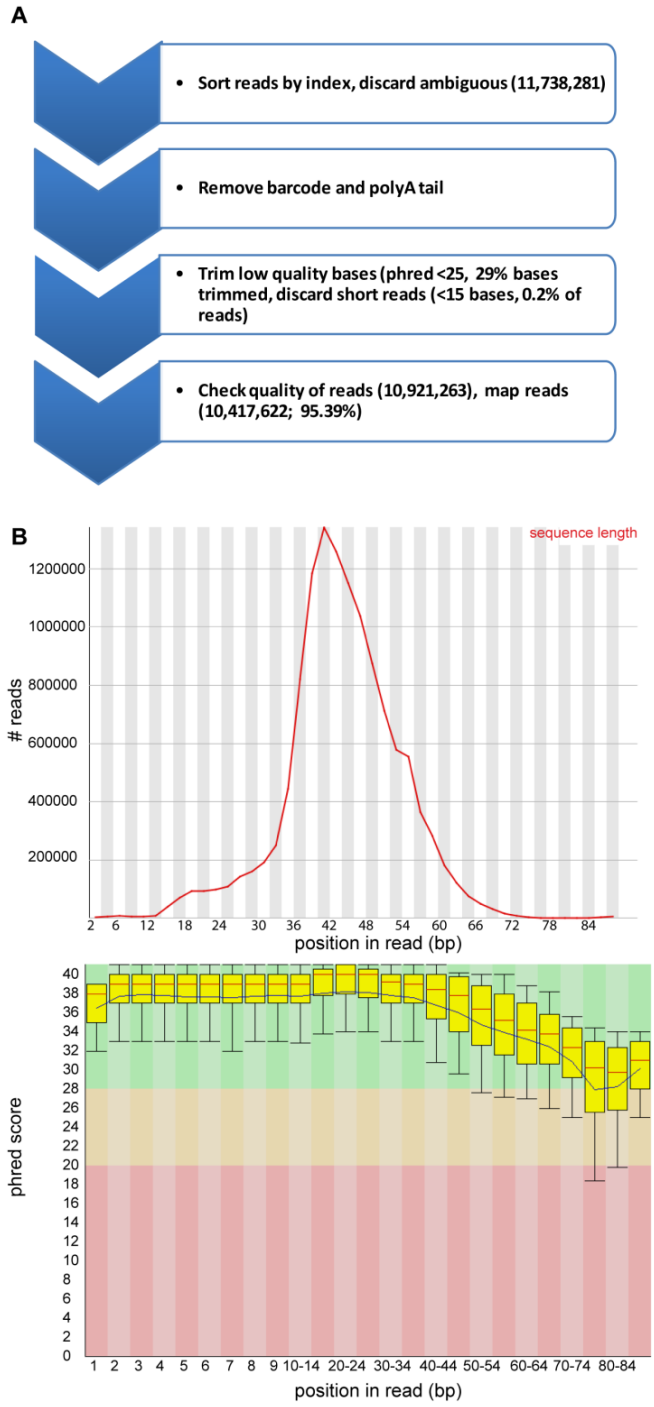


Figure 2.6. PAS-Seq produces high quality reads. (A) Representative library pre-processing workflow showing retention of the majority of reads through quality-control steps. (B) Representative library sequence length distribution (top) and per base phred score (bottom) after pre-processing, generated by the Fastqc package.

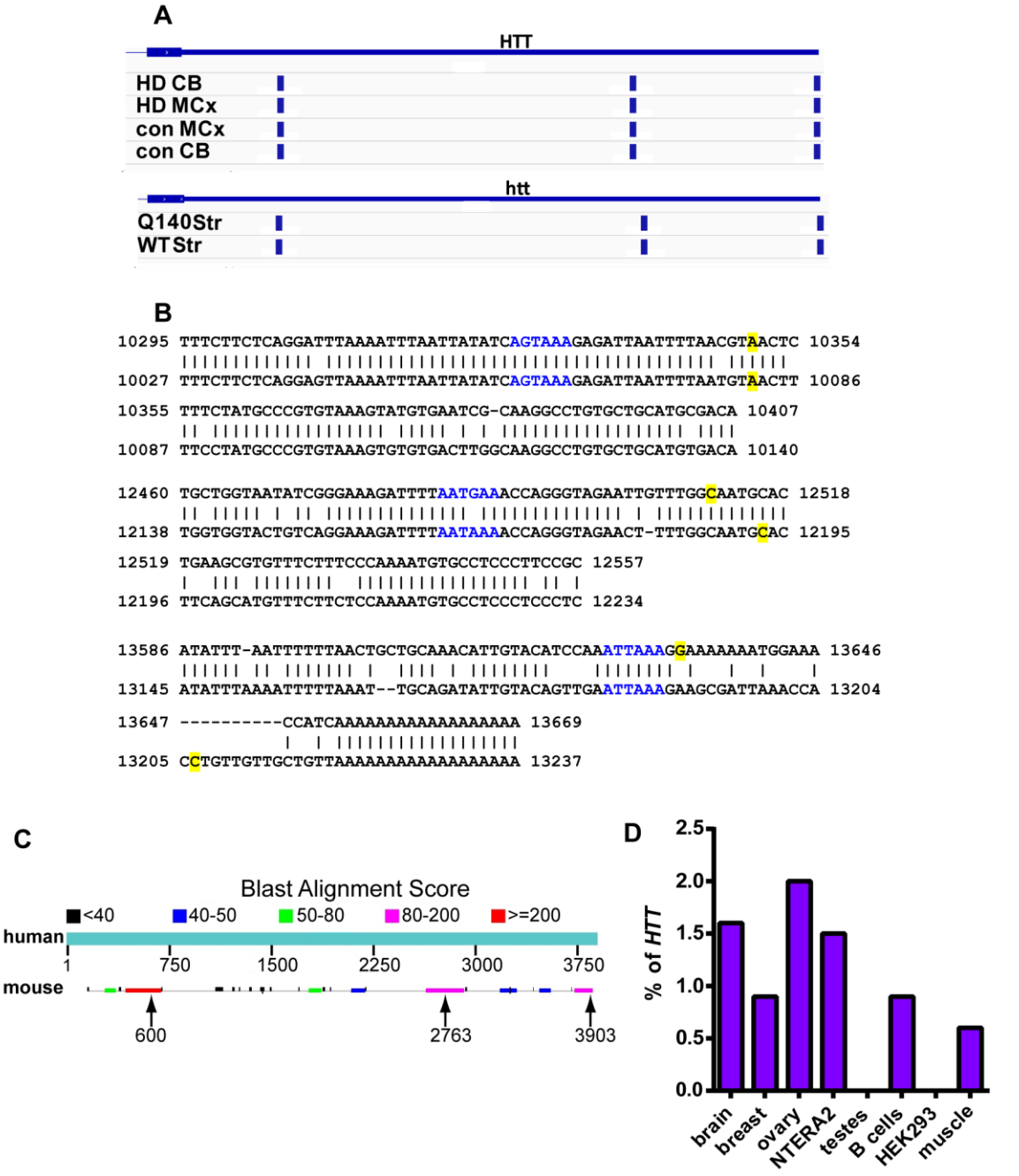
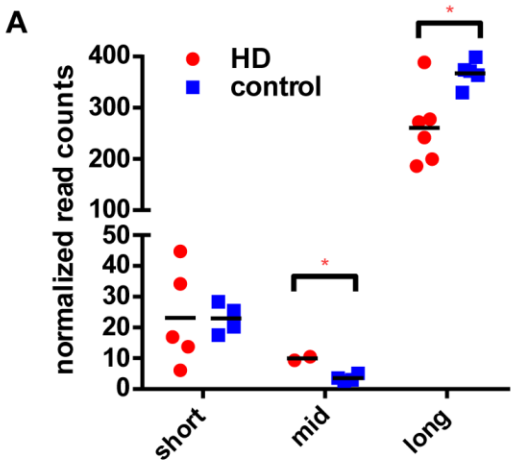
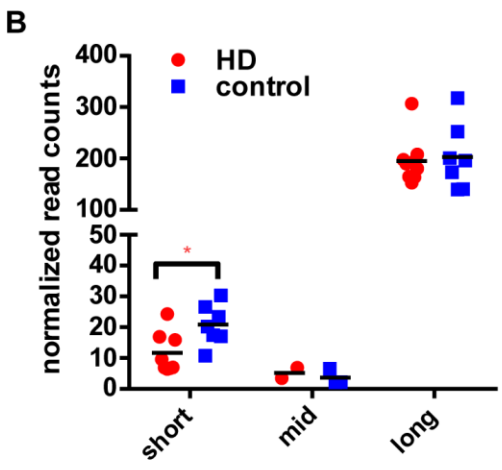


Figure 2.7. PAS-seq revealed a novel *HTT* mid-3′UTR isoform with tissue-specific abundance. (A) PAS-seq peak coordinates mapping to the *HTT* 3′UTR from HD patient and control cerebellum (CB) and motor cortex (MCx), and Q140 and wild-type (WT) mouse striatum (Str). (B) Needleman-Wunsch alignment of the human (top) and murine (bottom) *HTT* 3′UTR around the short, mid, and long 3′UTR isoform polyA signals (blue) and sites (yellow). (C) NCBI

BLAST alignment score of the human and mouse *HTT* 3′UTR; arrows indicate the three polyA sites at 600, 2763, and 3903 bases into the human 3′UTR. (D) *HTT* mid 3′UTR isoform 3′seq reads relative to total *HTT* 3′UTR reads across several human tissues and cell lines.

The abundance of the mid-3′UTR isoform also changes in HD. I compared PAS-seq read counts (normalized to sequencing depth) for each *HTT* isoform. In HD motor cortex, the abundance of the short isoform is unchanged while the abundance of the mid-isoform is increased by 2.8-fold (p=0.005) and the abundance of the long isoform is decreased 1.4-fold (p=0.01). In cerebellum, the abundance of the short isoform decreases by 1.8-fold (p=0.02), and the abundance of the mid and long isoform remains unchanged, indicating the increase in the long isoform relative to total *HTT* detected by qPCR is mostly due to a decrease in the short isoform (Fig. 2.2A). To verify *HTT* isoform shifts are not solely due to changes in gene expression, I calculated the fraction of total *HTT* each isoform contributes. Indeed, isoform changes were still significant after normalizing to gene expression (Fig. 2.8A,B bottom tables). In addition, isoform changes can't be explained by gene expression alterations: total *HTT* expression is decreased by 1.4-fold in HD motor cortex but unchanged in cerebellum compared to controls (Fig. 2.8C).





	HD reads	control reads	fold change
short	27.38	22.88	1.20
mid	9.91	3.57	2.78*
long	260.97	367.41	1.41*

	HD reads	control reads	fold change
short	11.69	20.83	1.78*
mid	5.16	3.6	1.43
long	195.03	202.88	1.04

	HD % total HTT	control % total HTT	% change
short	9.18	5.81	3.37
mid	3.32	0.91	2.41*
long	87.5	93.28	-5.79*

	HD % total HTT	control % total HTT	% change
short	5.52	9.16	-3.64*
mid	2.43	1.59	0.84
long	92.05	89.25	2.8*

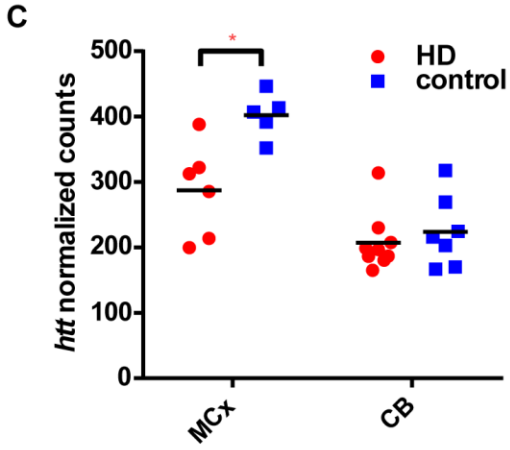


Figure 2.8. The abundance of the *HTT* mid-3′UTR isoform also changes in HD. (A) Read counts of each isoform normalized to depth (graph, top table) as well as contribution of each 3′UTR HTT isoform to total *HTT* expression (bottom table) in control and HD motor cortex. (B) Read counts as in A, in control and HD cerebellum. (C) Total *HTT* expression in HD versus control motor cortex (MCx) and cerebellum (CB). * signifies $p<0.05$.

2.2.4 *HTT* mRNA 3′UTR isoforms have different localizations and half-lives.

3′UTR length affects mRNA stability, translation, and localization[132]. To determine how shifts in *HTT* isoform abundance affect *HTT* mRNA metabolism, I assessed the localization and stability of *HTT* isoforms in SH-SY5Y (human neuroblastoma) cells. These cells have neuronal characteristics and express short, mid, and long HTT 3′UTR isoforms, making them a model of *HTT* mRNA metabolism in human brain[189]. To assay the localization of *HTT* isoforms, I extracted RNA from nuclear and cytoplasmic compartments and performed isoform specific qRT-PCR on RNA from each (Fig. 2.9). I found about three times as much of all *HTT* isoforms are localized in the cytoplasm than in the nucleus in SH-SY5Y (Fig. 2.10A). To determine if isoform localization changes in disease, I assayed RNA from HD and wild-type fibroblasts. All *HTT* isoforms had a lower cytoplasmic to nuclear expression ratio in fibroblasts than in SH-SY5Y cells. As in SH-SY5Y cells, I found little difference between the localization of the HTT long and short isoforms; however, the mid isoform was significantly more abundant in the cytoplasm than the other isoforms in wild-type ($p<0.0001$) and disease ($p=0.02$) fibroblasts. The ratio of cytoplasmic to nuclear transcripts is significantly lower in HD fibroblasts than in wild-type fibroblasts for all three isoforms (short $p=0.01$, mid $p=0.0003$, long $p=0.009$), suggesting more *HTT* mRNA localizes to the nucleus in HD, as has been reported[32].

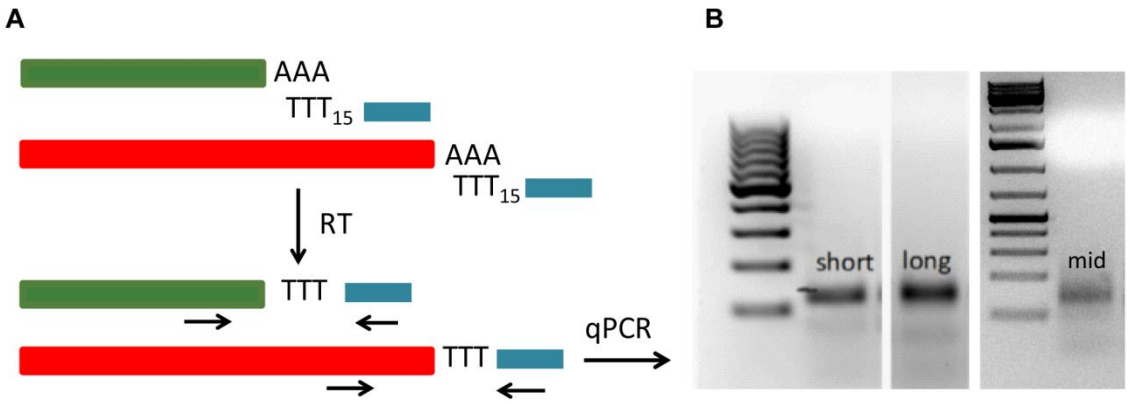


Figure 2.9. Isoform specific qPCR separately measures all *HTT* isoforms. (A) Total RNA is reverse transcribed (RT) with an oligo d (T) primer with a 3′ universal adaptor sequence (blue). cDNA is amplified during qPCR with an *HTT*-specific forward primer and the universal reverse primer. (B) The short and mid isoform forward primers will also anneal to the long isoform; however, the long isoform cannot be amplified during the short extension time, yielding only the short or mid isoform PCR product when reactions are run on a gel.

To study the stability of *HTT* isoforms, I incubated SH-SY5Y cells with labeled uridine, and collected labeled RNA 0, 3, 6, 9, and 12 hours after washing away the modified nucleotide (see methods). I performed isoform specific qRT-PCR on the RNA to determine the relative stability of the *HTT* isoforms. I found the *HTT* short, mid, and long mRNA isoforms have significantly different half-lives; the short isoform half-life is longer than the long isoform half-life, and the mid isoform is the shortest lived (Fig. 2.10B). This finding suggests that the altered relative abundance of *HTT* isoforms in HD brains is accompanied by a change in total *HTT* half-life.

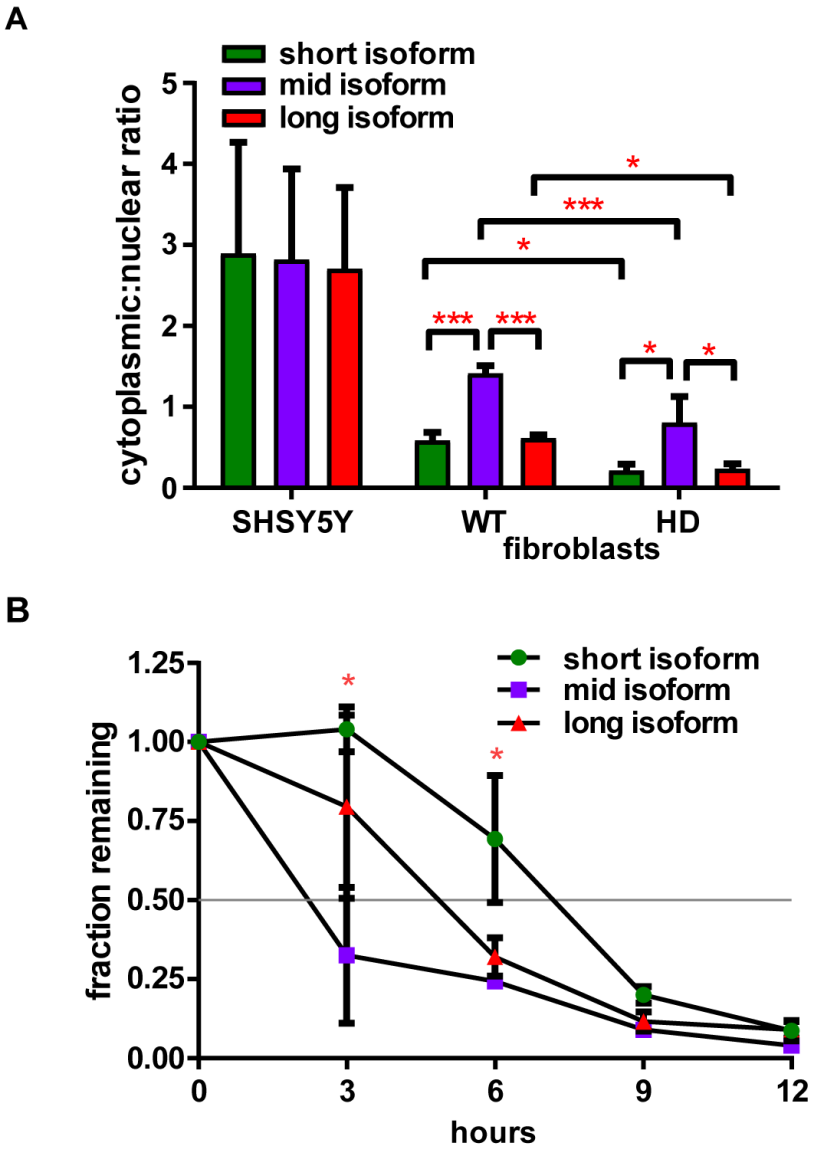


Figure 2.10. *HTT* 3′UTR isoforms have different localization and half-lives.

(A) Cytoplasmic versus nuclear abundance of *HTT* 3′UTR isoforms across cell lines. Data points are the average of triplicates with standard deviation. *, **, and *** signify $p<0.05$, 0.005, and 0.0005. (B) Abundance of labeled *HTT* 3′UTR isoforms in SH-SY5Y cells at various time points after a 12-hour eithinyl-uridine pulse measured by isoform-specific qPCR. Data points are the average of triplicates with standard deviation.

2.2.5 *HTT* mRNA 3′UTR isoforms have different polyA tail lengths, RNA binding protein sites, and microRNA sites.

The different stability of the *HTT* 3′UTR isoforms could be due to different polyA tail lengths, microRNA sites, or RNA binding protein sites[56,132]. To determine the polyA tail length of the *HTT* 3′UTR isoforms, I added a universal sequence to the 3′ end of the mRNA polyA tail that was used as a priming site during isoform-specific PCR. PCR products were resolved on a gel, enabling quantification of poly-A tail length (Fig. 2.11A). I found the *HTT* short isoform has a poly-A tail length of about sixty in both SH-SY5Y cells and in patient and control fibroblasts, whereas the *HTT* long and mid isoforms have very short polyA tail lengths of about ten and four (Fig. 2.11B,C). Recent studies found polyA tail length is correlated with stability, and the rate of mRNA translation drops rapidly as the polyA tail length decreases below twenty nucleotides[102,190]. The short polyA tail length of the mid and long *HTT* isoforms compared to the short isoform may explain their lower half-lives.

To identify microRNA sites that differ between isoforms, I used the bioinformatics tool TargetScan. This program predicts microRNA target sites based on several local factors known to affect microRNA binding strength and specificity[191]. I screened predicted microRNA binding sites for microRNAs expressed in human brain. I found several sites exclusive to the mid and long *HTT* 3′UTR (Fig. 2.12A). These sites may contribute to differences in stability between these isoforms and the short *HTT* isoform[192]. To test this hypothesis, I transfected SH-SY5Y cells with mimics of microRNAs 221, 137, or both (Fig. 2.12B). MicroRNA 137 exclusively binds the long

isoform, while microRNA 221 has an 8-mer binding site in the long isoform and a 6-mer binding site in the mid and long isoforms. I found the microRNA 137 mimic had no effect on isoform abundances, but transfection of the microRNA 221 or microRNA 221 and 137 mimics reduced the abundance of the long and mid but not short isoform relative to total *HTT* (2.12C) or to *GAPDH* (Fig. 2.12D). These results show at least one microRNA exclusively degrades the long and mid *HTT* transcripts, likely contributing to the lower stability of these isoforms compared to the short isoform.

To identify RNA binding protein sites that differ between isoforms, I used the CLIPdb and starBase CLIP-Seq databases[193,194]. I found several brain-expressed RNA binding protein sites exclusive to the long or mid and long *HTT* isoforms that may affect their stability (Fig. 2.12B). MicroRNAs and RNA binding proteins likely interact with *HTT* binding sites in motor cortex and cerebellum; however, brain cell and region-specific expression of these factors is not established. Based on these results, I expect the relative decrease of the *HTT* long isoform in HD motor cortex is accompanied by an increase in *HTT* translation due to increased polyA tail length and increased stability.

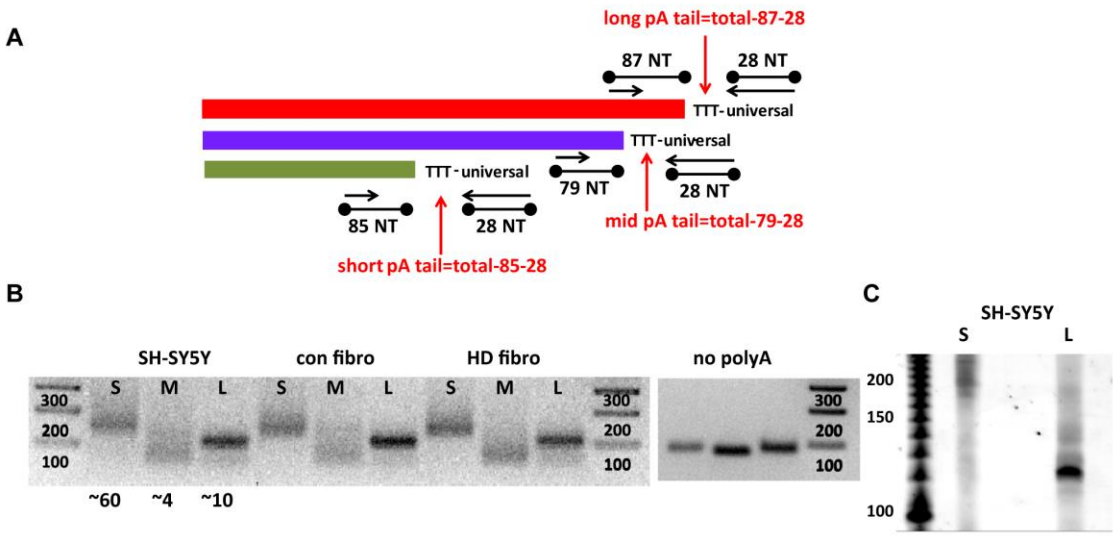


Figure 2.11. *HTT* 3′UTR isoforms have different polyA tail lengths.

(A) A universal sequence is added to the 3′ end of the polyA tail during reverse transcription. *HTT* isoform-specific PCR is then performed with the universal sequence as a reverse priming site and an isoform-specific region as the forward priming site. The polyA tail length is the PCR product length after subtraction of the universal primer length (28 nucleotides, NT) and the distance from the forward primer to the polyA site (87 NT for the long isoform, 79 NT for the mid isoform, 85 NT for the short). (B) Agarose gel of *HTT* 3′UTR isoform PCR products from the polyA tail length assay; S=short, M=mid, L=long. (C) Polyacrylamide gel of *HTT* short and long 3′UTR isoform PCR products from the polyA tail length assay.

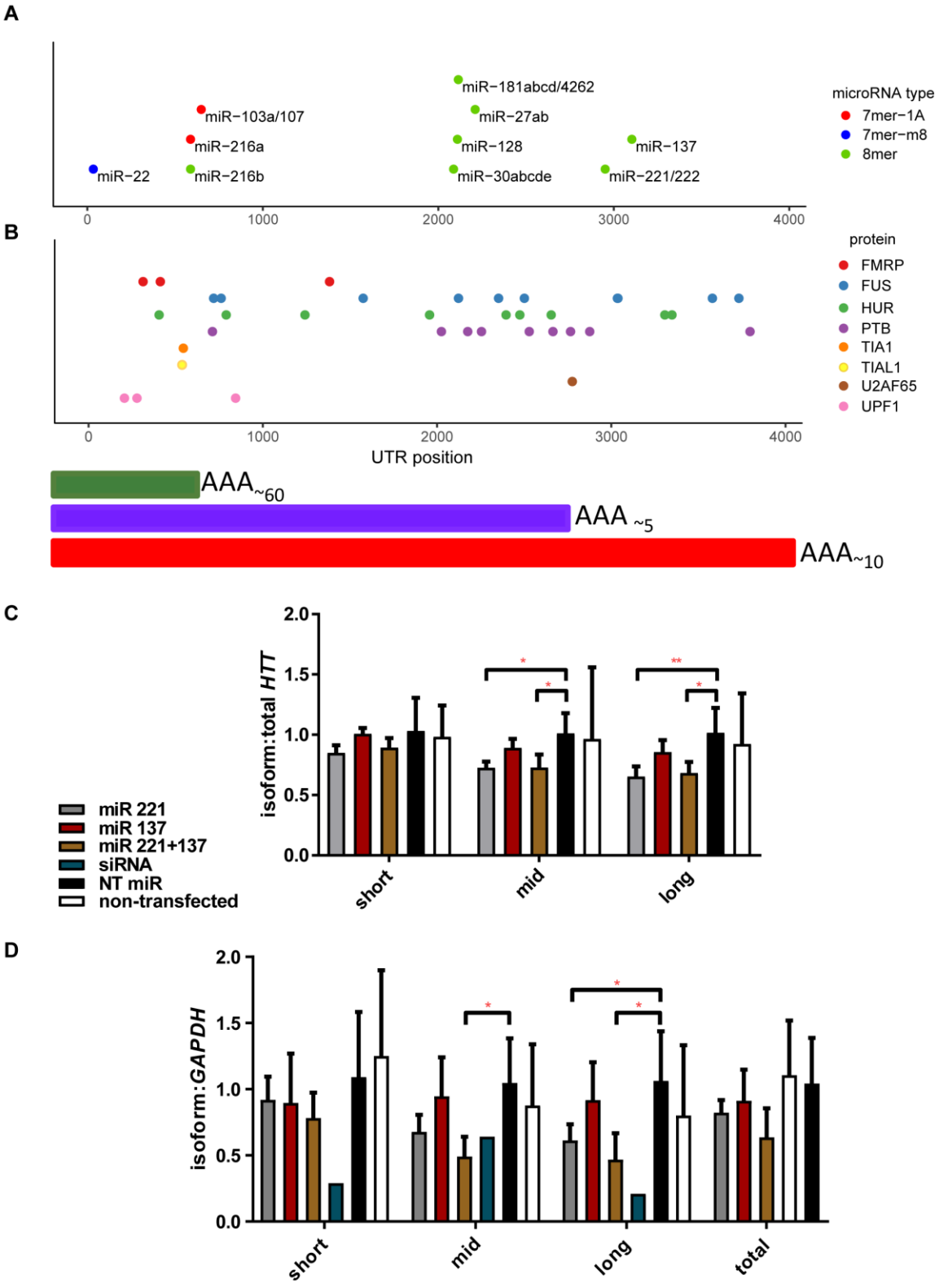

A
miR−181abcd/4262
miR−103a/107
miR−27ab
miR−216a
miR−128
miR−137
miR−22
miR−216b
miR−30abcde
miR−221/222
0
1000
2000
3000
4000
microRNA type
7mer−1A
7mer−m8
8mer
B
protein
FMRP
FUS
HUR
PTB
TIA1
TIAL1
U2AF65
UPF1
UTR position
AAA~60
AAA ~5
AAA~10
C
miR 221
miR 137
miR 221+137
siRNA
NT miR
non-transfected
isoform:total HTT
short
mid
long
D
isoform:GAPDH
short
mid
long
total

Figure 2.12. *HTT* 3′UTR isoforms have different RNA binding protein sites and microRNA target sites. (A) Brain-expressed microRNA target sites in the *HTT* 3′UTR predicted by TargetScan**.** (B) Brain-expressed RNA binding protein sites in the *HTT* 3′UTR identified in public CLIP-Seq datasets. (C) Quantitative PCR of *HTT* 3′UTR isoforms normalized to total *HTT* expression in SH-SY5Y cells after transfection of microRNA mimics with targets exclusive to the *HTT* long or mid and long 3′UTR isoforms. *, **, and *** signify $p<0.05$, 0.005, and 0.0005. (D) Quantitative PCR of *HTT* 3′UTR isoforms normalized to *GAPDH* expression in SH-SY5Y cells after transfection of microRNA mimics with targets exclusive to the *HTT* long and mid 3′UTR isoforms. NT (non-targeting) miR is a *C. Elegans* microRNA mimic.

2.3 Discussion

I sought to characterize the metabolism of *HTT* 3′UTR isoforms and to determine whether altered *HTT* 3′UTR isoform abundance is a feature of HD. I found the abundance of *HTT* 3′UTR isoforms is altered in HD motor cortex, cerebellum, fibroblasts, and neural stem cells. Abundance changes extend to peripheral tissues in Yac128 model mice and arise from the mutant and wild-type allele. PAS-seq identified a novel mid 3′UTR isoform of *HTT* that also changes in disease. I found *HTT* 3′UTR isoforms have different localization, half-lives, polyA tail lengths, microRNA sites, and RNA binding protein sites. Thus, isoform abundance changes in HD likely impact total *HTT* metabolism.

HTT mRNAs form nuclear foci in HD cells that sequester splicing proteins[32]. I found the predominant *HTT* isoforms have similar localization, and the nuclear abundance of all *HTT* isoforms increases in HD; thus, the increase in nuclear *HTT* in HD is likely not due to isoform shifts. While the distribution of the *HTT* long and short 3′UTR isoforms is similar, the mid isoform is significantly more cytoplasmic. The increase in the mid isoform in HD brain may be accompanied by an increase in cytoplasmic *HTT* accessible to the ribosome. However, this isoform is of low abundance and is unlikely to have a great effect on HTT protein expression. My study only assayed nuclear versus cytoplasmic isoform localization. The 3′UTR is known to direct mRNAs to dendrites[96,111]. A recent study of *HTT* isoforms found the long but not short 3′UTR localizes *HTT* mRNA to dendrites in cultured neurons[39]. Further studies are necessary to determine if the dendritic localization of *HTT* isoforms changes in HD.

Isoform changes may contribute to region-specific pathology in HD. The short-isoform has a longer half-life and polyA tail than the other isoforms. Thus, the shift to the short isoform in HD motor cortex is expected to result in an increase in total *HTT* half-life and translation[101]. This is in contrast to findings from a recent study that found there is no difference in *HTT* mRNA abundance between cells transfected with a *HTT* short isoform construct and those transfected with a long isoform construct[39]. However, researchers measured steady-state mRNA levels of an overexpressed construct rather than stability of endogenous *HTT* transcribed from its native gene context. The study also found the short isoform construct produced more HTT protein and aggregates in cells, suggesting the mutant short isoform is more pathogenic than the long isoform[39]. Thus, the shift to the short *HTT* isoform in HD motor cortex likely contributes to HD pathogenesis, while the shift to the long isoform in the cerebellum may be protective. The study did not examine the role of the mid isoform, which also increases in HD. Further research is necessary to determine the impact of *HTT* isoform changes on HD pathogenesis, and whether isoform changes are casual or consequential.

My study of *HTT* isoform abundance provides specific recommendations for allele-specific *HTT*-lowering therapies. Many frequently-heterozygous SNPs are located in the *HTT* 3′UTR[46]. I found the abundance of the long isoform decreases 2.5-fold in HD motor cortex compared to controls. Thus SNPs exclusive the long *HTT* isoform are not ideal therapeutic targets. However, further characterization of the role of *HTT* isoforms in HD is necessary. If the short isoform is more associated with pathology in HD patients, it will be the ideal target for RNA interference in HD.

2.4 Materials and methods

2.4.1 Samples. The New York Brain Bank and The Neurological Foundation of New Zealand Brain Bank supplied fresh frozen human brain samples. Three HD and three control fibroblast lines were obtained from the Coriell Cell Repository. The M. DiFiglia lab kindly supplied neural stem cell pellets (Massachusetts General Hospital). SH-SY5Y cells were obtained from ATCC. The Cure Huntington's Disease Initiative (CHDI) supplied Q140 mice. Mice heterozygous for null murine *Htt* were supplied by X. W. Yang (University of California Los Angeles) and bred with Yac18 mice supplied by M. Hayden (University of British Columbia) and Yac128 mice (Jackson Labs). Wild-type mice were obtained from Jackson Labs. For Q140, wild-type, Yac18 null-null, and Yac128 null-null mouse experiments, 7-9 month mice were sacrificed, and brain regions were dissected and stored in RNA-later prior to RNA extraction.

2.4.2 Mouse genotyping. Mouse ear punches were digested in 50mM sodium hydroxide at 95°C for twenty minutes. Tris-HCl (pH 8.0) was added to a final concentration of 0.1M. To determine genotype, DNA in 1μl of the supernatant was amplified and resolved as follows. For Yac18 and Yac128 mice, PCR was performed with PrimeStar GXL polymerase (Clontech) and 0.5μM forward and reverse primers 5′-GCCTCCGGGGACT GCCGTGC-3′ and 5′-CGGCTGAGGCAGCAGCGGCT-3′. The reaction was incubated at 98°C for 1 minute; 30 cycles of 98°C for 10 seconds and 68°C for 30 seconds; and 68°C for 10 minutes. For null murine *Htt* mice, PCR was performed with FlashTaq Master Mix (Empirical Bioscience) and 0.4μM primers as follows: for murine *Htt*, 5′-

ACGCATCCGCCTGTCAATTCTG-3′ and 5′-CTGAAACGACTTGAGCGACTC-3′; and for the null cassette, 5′-AACACCGAGCCGACCCTGCAG-3′ and 5′-CCACCATGATATTCGGCAAGCAG-3′. The reaction was incubated at 94°C for 16.5 minutes; 35 cycles of 94°C for 20 seconds, 65 °C for 30 seconds, and 72°C for 30 seconds; and 72°C for 5 minutes. For Q140 mice, PCR was performed with FlashTaq Master Mix (Empirical Bioscience) and 0.4μM forward and reverse primers 5′-CTGCACCGACCGTGAGTCC-3′ and 5′-GAAGGCAC TGGAGTCGTGAC-3′. The reaction was incubated at 95°C for 5 minutes; 40 cycles of 94°C for 20 seconds, 67°C for 30 seconds, and 72°C for 15 seconds; and 72°C for 5 minutes. All PCR products were resolved on a 1.5% agarose, 1x TAE gel and compared to expected product size for each genotype.

2.4.3 Quantitative RT-PCR. Total RNA was extracted using Trizol (Ambion), and RNA quality was assessed by Bioanalyzer (Agilent). RNA was treated with TurboDNAse (Ambion), and cDNA was synthesized from 2μg total RNA using SuperScript IV (Invitrogen) with oligo $d(T)_{20}$ priming per manufacturer's instructions. Quantitative PCR was performed with 1.8μl cDNA, 1μM forward and reverse primers, and QuantiFast SYBR green master mix (Qiagen). The reaction was incubated at 95°C for 5 minutes, followed by 40 cycles of 95°C for 10 seconds and 60°C for 30 seconds. The isoform:total mRNA or total mRNA:*GAPDH* mRNA was calculated using the ΔΔCt method. All primers were validated by relative standard curve, and primer pairs amplified targets with 85-115% efficiency in the linear range. I verified primer specificity by running qPCR

products on a 2% agarose, 1x TAE gel, extracting the bands using the QIAquick Gel Extraction Kit (Qiagen), and Sangar sequencing the bands with the forward and reverse PCR primers. I compared HD and control sample averages with unpaired two-tailed t-tests using GraphPad Prism version 6.00 for Windows, GraphPad Software, La Jolle California USA, www.graphpad.com. I considered $p<0.05$ significant.

2.4.4 Allele-specific quantitative RT-PCR. RNA was extracted and cDNA was synthesized as above. Quantitative PCR was performed using 1.8µl cDNA, 2uM forward and reverse primer, 150nM probe, and 1x Type-it Fast SNP PCR master mix (Qiagen). The reaction mix was incubated at 95°C for 5 minutes, followed by 40 cycles of 95°C for 15 seconds, 67.5°C for 15 seconds, and 72°C for 30 seconds. The long *HTT* isoform allele:total mRNA was calculated using the ΔΔCt method. All primer-probe sets were validated as above. I compared HD and control sample averages as above ("Quantitative RT-PCR").

2.4.5 Ethinyl-uridine (EU) pulse chase: in vitro transcription of spike-in. To amplify firefly luciferase DNA with the T7 promoter, 10ng template DNA (plasmid with firefly luciferase) was combined with 0.3µM forward and reverse primers and 1x Phusion High Fidelity PCR Master Mix (NEB). The reaction was incubated at 98°C for 30 seconds; 35 cycles of 98°C for 15 seconds, 52°C for 10 seconds, and 72°C for 20 seconds; and 72°C for 1 minute. The reaction was run on a 1% agarose gel to ensure there was only one product and then cleaned with the Promega PCR Purification Kit. T7 RNA polymerase

(Promega) performed in vitro transcription. Briefly, 1μg of the firefly PCR product was added to a reaction with 1x transcription buffer; 1μl RNAseOut (Thermo Fisher Scientific); 0.5mM each of rCTP, rATP, and rGTP; 0.48mM rUTP; 0.02mM E-UTP (Jena Bioscience); and 1μl T7 RNA polymerase. The reaction was incubated for 2 hours at 37°C, and then treated with RQ1 DNase (Promega). In vitro transcripts were purified on Centrispin 20 columns (VWR). To precipitate the purified transcripts, one-tenth volume 3M sodium acetate and one volume of phenol chloroform were added, and the reaction was vortexed for one minute followed by centrifugation at 12,000g for 10 minutes. One volume of chloroform was added to the top layer, and the mixture was vortexed and centrifuged again. The top layer was kept, 2.5 volumes of ethanol and 4μg glycogen were added, and the RNA was precipitated overnight at -20°C. The mixture was then centrifuged at 12,000g for 40 minutes at 4°C. The pellet was washed with 70% ethanol, spun 10 minutes at 12,000g and 4°C, and resuspended in water. In vitro transcripts were Bioanalyzed (Agilent) to ensure a single product.

2.4.6 Ethinyl-uridine (EU) pulse chase. I modified the Click-iT Nascent RNA Capture Kit (Life Technologies) as follows. SH-SY5Y cells were pulsed in 200μM EU for 14 hours and then chased by two washes with PBS before addition of unmodified media. Cytoplasmic RNA fractions were collected by lysing cells 10 minutes in hypotonic lysis buffer (20mM Tris-HCl, 15mM NaCl, 10mM EDTA, 0.5% NP-40, 0.1% Triton-X-100) followed by centrifugation at 1,200g for 10 minutes at 4°C. To extract total cytoplasmic RNA, the supernatant was incubated at 42°C for one hour with 1% SDS and 200μg/ml

Proteinase K (Ambion) and then ethanol-precipitated as above. The RNA was treated with TurboDNAse (Ambion) and cleaned on Clean and Concentrator columns (Zymo). Five micrograms RNA was biotinylated per kit instructions with the addition of 0.02ng EU-containing firefly spike-in RNA per reaction, and RNA was ethanol-precipitated as above. Beads pulled down 4µg RNA as instructed with the exceptions 10% trimethylamine was used instead of Click-iT RNA binding buffer in the RNA binding reaction, and beads were ultimately resuspended in 25µl 50mM Tris-HCl instead of wash buffer 2. SuperScript IV (Invitrogen) reverse transcribed captured RNA in a 50µl reaction with 800rpm shaking and the universal reverse transcription primer. Quantitative PCR was then performed as above ("Quantitative RT-PCR").

2.4.7 Cell fractionation. Cytoplasmic RNA fractions were collected by lysing cells 10 minutes in hypotonic lysis buffer (20mM Tris-HCl, 15mM NaCl, 10mM EDTA, 0.5% NP-40, 0.1% Triton-X-100) followed by centrifugation at 1,200g for 10 minutes at 4°C. To extract RNA, the supernatant (cytoplasmic) or pellet (nuclear) was incubated at 42°C for one hour with 1% SDS and 200µg/ml Proteinase K (Ambion) and then ethanol-precipitated as above. The RNA was treated with TurboDNAse (Ambion) and submitted to quantitative RT-PCR as above ("Quantitative RT-PCR"). The ratio of cytoplasmic to nuclear expression of each isoform was compared by Anova with Tukey's multiple comparison test.

2.4.8 PolyA tail length assay. RNA was extracted with Trizol (Ambion), treated with

TurboDNAse (Ambion), and cleaned with Clean and Concentrator columns (Zymo). I then used the PolyA Tail Length Assay (Affymetrix) to add GI tails to mRNA. Tailed mRNAs were reverse transcribed by SuperScript IV (Invitrogen) from the polyA tail universal reverse primer (see Table 2.2). The short, mid, and long isoform forward primers were used with the isoform specific universal reverse primer to amplify the polyA tail. PCR products were resolved on a 2.5% agarose, 1x TAE gel. The bands were extracted using the QIAquick Gel Extraction Kit (Qiagen) and Sangar sequenced with the forward and reverse primers to ensure isoform PCR specificity.

2.4.9 MicroRNA and RNA binding protein target site prediction. I used the CLIPdb and starBase v2.0 browsers to identify RNA binding proteins that interact with the *HTT* 3′UTR [193,194]. I included brain-expressed RNA binding proteins binding at sites represented by more than five reads. Brain expression was confirmed with the human protein atlas, www.proteinatlas.org [195]. To identify predicted microRNA binding sites in the *HTT* 3′UTR, I used TargetScanHuman version 7.0[191]. I identified brain-expressed microRNAs with predicted high-affinity binding sites (7mer-1A, 7mer-m8, 8mer). Brain expression was confirmed with miRmine Human miRNA Expression Database [196].

2.4.10 MicroRNA transfection. SH-SY5Y cells were transfected with microRNA mimics 137, 221, or both (Dharmacon) using Lipofectamine RNAimax transfection reagent (Invitrogen) per manufacturer's instructions. After 72 hours, RNA was collected and submitted to quantitative RT-PCR of *HTT* isoforms as above ("Quantitative RT-

PCR").

Table 2.2 Primers used in chapter 2
(AS=allele-specific, NAS=non allele-specific, IS=isoform specific)

Primers (5′-3′)

Target	Forward	Reverse
(NAS) *HTT* long 3′UTR	ATGGATGCATGCCCTAAGAG	CAGTCTCCGATGAGCACAGA
(NAS) total *HTT*	CATGGTGGGAGAGACTGTGA	CAAAGAGCACTTCTGCCACA
(AS) *HTT* long 3′UTR	AAGTGGATTCTGGATGGCCG	ACATGACAGTCGCCAACCTT
(AS) total *HTT*	GTGACCAGGTCCTTTCTCCTG	TGTTCCCAAAGCCTGCTCAC
(IS) *HTT* long 3′UTR	GGAAGGACTGACGAGAGATG	ACGCATCTATGCGCATATCG
(IS) *HTT* short 3′UTR	AGCAGGCTTTGGGAACACTG	ACGCATCTATGCGCATATCG
(IS) *HTT* mid 3′UTR	GTGGCAAGCACCCATCGTAT	ACGCATCTATGCGCATATCG
short *HTT* polyA tail	TAGACACCCGGCACCATTCT	ACGCATCTATGCGCATATCG
long *HTT* polyA tail	GGAAGGACTGACGAGAGATG	ACGCATCTATGCGCATATCG
mid *HTT* polyA tail	GGCTGTGGGGAGATTGCTTT	ACGCATCTATGCGCATATCG
short *HTT* -polyA tail	TAGACACCCGGCACCATTCT	CGTTAAAATTAATCTCTTTACTG
mid *HTT* -polyA tail	GGCTGTGGGGAGATTGCTTT	CCAAACAATTCTACCCTGGT
long *HTT* -polyA tail	GGAAGGACTGACGAGAGATG	GATGGTTTCCATTTTTTTCCTTT
firefly luciferase qRT-PCR	TGATCAAGTACAAGGGCTACCA	GCTGCAGCAGGATAGACTCC

Probes (5′-3′)

(AS) *HTT* long 3′UTR	TCTGCTTGC(C/T)GACTGGCT
(AS) total *HTT*	CCTGCTGGTTGTTGCCAGGTTG

Reverse transcription primers (5′-3′)

Isoform-specific	ACGCATCTATGCGCATATCGTTTTTTTTTTTTTTT
PolyA tail length assay	ACGCATCTATGCGCATATCGCCCCCCCCTTTT

CHAPTER III: WIDESPREAD CHANGES IN 3′ ISOFORM ABUNDANCE ARE A FEATURE OF HD PATHOLOGY

Preface

The work presented in this chapter is accepted at *Cell Reports* as manuscript "Alterations in mRNA 3′UTR isoform abundance accompany gene expression changes in human Huntington's disease brains" by Romo, Ashar-Patel, Pfister, and Aronin.

This work was a collaborative effort. Ami Ashar-Patel developed the PAS-seq protocol. Yasin Kaymaz and Jeff Bailey provided advice on PAS-seq data analysis.

3.1 Summary

Although many genes exhibit expression and splicing differences in HD, 3′UTR isoform expression changes have not been investigated[117,138,173,180]. 3′UTR length is important for mRNA localization, stability, and translation[132]. Many mRNAs use 3′UTR alternative polyadenylation to achieve tissue-specific expression and function[62,84,89]. Long mRNA isoforms contain binding sites for trans-acting factors such as microRNAs that exert dynamic changes in steady-state mRNA levels[87]. Ubiquitously expressed genes, like *HTT*, are the most likely to use alternative polyadenylation to alter mRNA expression in the brain[37,84]. Changes in mRNA 3′UTR length occur in Parkinson's disease, cancer, and myotonic dystrophy, another repeat expansion disease[110,115,178]. In chapter II, I found the abundance of *HTT* 3′UTR isoforms changes in disease. The change arises from both the wild-type and mutant alleles, suggesting the cause is a *trans* factor. If so, other mRNAs besides *HTT* may change their isoform abundances in HD. I sought to determine if alterations in 3′UTR length are a feature of HD.

3.2 Results

3.2.1 Many other genes exhibit changes in isoform abundance in HD motor cortex.

I studied whether alterations in isoform expression are unique to *HTT* mRNA, or features of many mRNAs in HD. To assay transcriptome-wide isoform expression, I performed PAS-seq on motor cortex from grade 1 HD brains (n=6) versus controls (n=5), and cerebellum from grade 2-4 patient brains (n=9) versus controls (n=7). I identified genes with multiple 3′UTR isoforms, and normalized isoform expression to gene expression (see methods). I compared the normalized isoform expression between HD and control samples.

PAS-seq uses oligo(dT) to capture polyA tails during reverse transcription. Oligo(dT) also anneals to genomic polyA stretches. These genomic sites are artefacts. Artefactual polyA sites can lead to overestimation of the number of alternatively polyadenylated genes[185]. I remove these genomically-primed reads from my data. To validate my analysis, I compared the number of alternatively polyadenylated genes identified by PAS-seq to that that identified by the 3Pseq 3′ sequencing method, which avoids oligo(dT) priming[61]. I found 33% of genes in the motor cortex and 30% in the cerebellum display alternative polyadenylation, with over half exhibiting two 3′UTR isoforms (Fig. 3.1). This percentage is similar to that obtained via 3Pseq. Consistency with 3Pseq data indicates PAS-seq accurately captured reads primed from mRNA polyA tails while excluding reads primed from genomic polyA regions. As previously reported in brain tissue, I found long isoforms are more abundant than short isoforms in cerebellum and motor cortex[86].

In the motor cortex, 11% of 2,164 alternatively polyadenylated genes showed a significant (false discovery rate<10%) change in abundance of at least one isoform in HD ($p<0.01$) (Fig. 3.2A). Many isoform abundances changed over 25% in HD motor cortex versus controls, indicating a major shift in isoform expression for the corresponding gene (Fig. 3.2B). In the cerebellum, none of the alternatively polyadenylated genes showed a significant (false discovery rate <10%) change in abundance of at least one isoform (Fig. 3.2C). The false discovery rate threshold corrects for false-positives due to multiple testing, but may reject true positives. To allow comparison of isoform changes between motor cortex and cerebellum, I also included a more sensitive p-value threshold ($p<0.01$). Using this threshold, thirteen percent of alternatively polyadenylated genes exhibit isoform changes in HD motor cortex, whereas five percent of genes exhibit changes in cerebellum. Isoform abundance changes in the cerebellum were not as large as in the motor cortex (Fig. 3.2D). These results indicate that there are widespread isoform changes in HD motor cortex that alter the 3'UTR isoform abundances for many genes, whereas changes in the cerebellum are negligible.

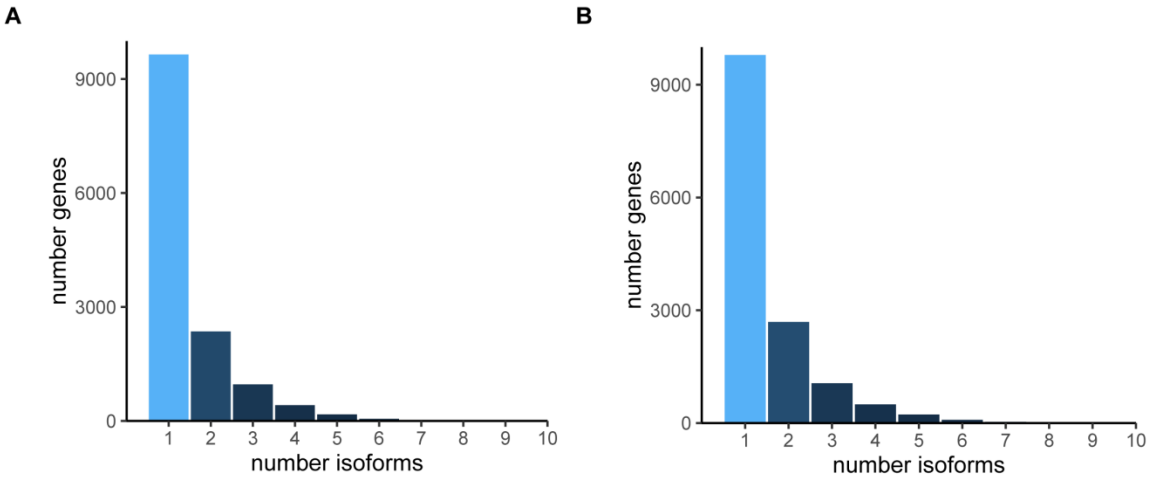


Figure 3.1. Isoform number distributions match previous studies. (A) Histogram of number of 3′UTR isoforms per gene in cerebellum. I included isoforms represented in at least two samples (B) Histogram of number of isoforms per gene in motor cortex.

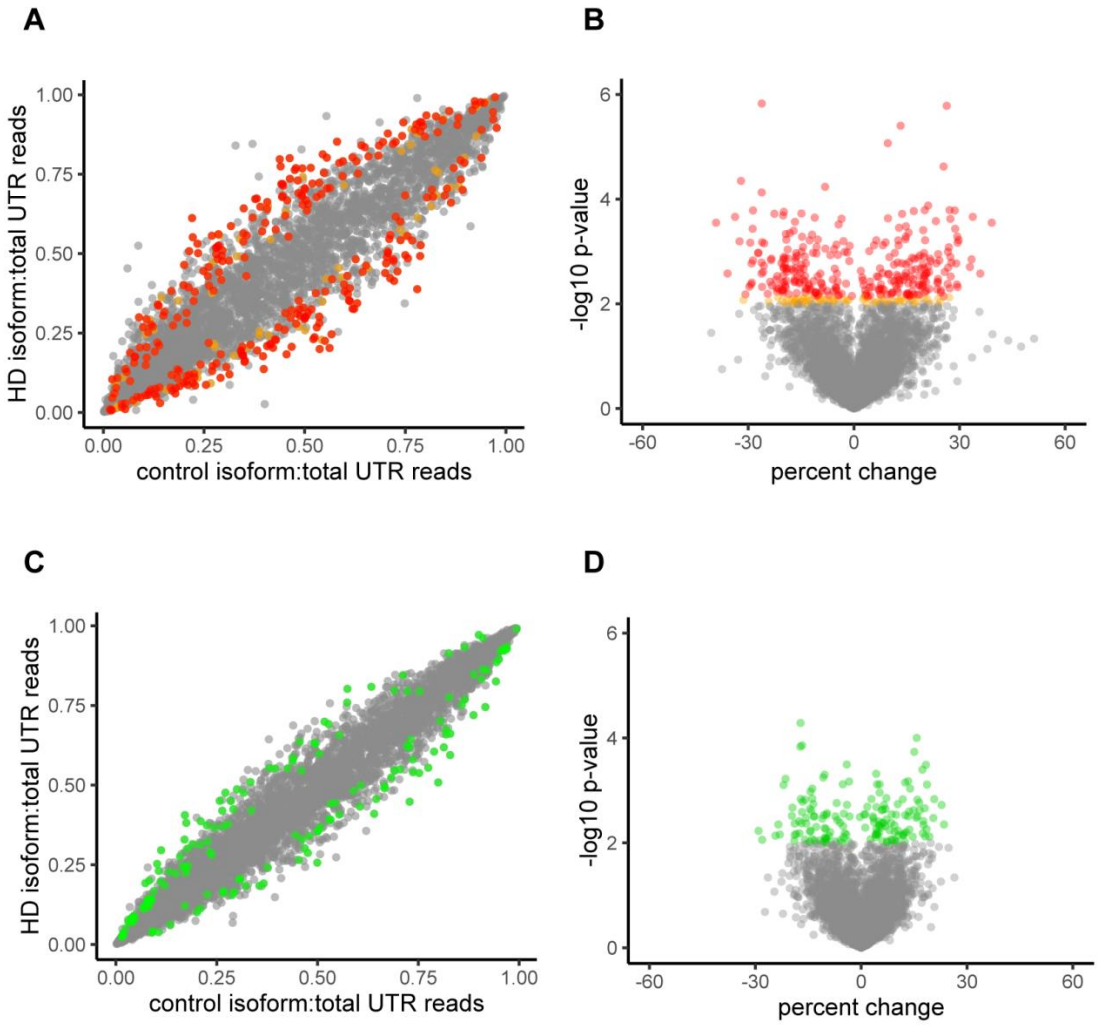


Figure 3.2. Transcriptome-wide PAS-seq analysis identifies a large subset of genes with 3′UTR isoform changes in HD motor cortex.

(A) Scatter plot of PAS-seq reads from control (n=5) and HD (n=6) motor cortex. Each dot represents a unique 3′UTR isoform expressed in at least five HD and control samples. Shown is the number of reads mapping to the isoform divided by the total reads mapping to the gene 3′UTR (isoform fraction). Orange dots indicate isoforms with $p<0.01$, whereas red dots indicate isoforms with adjusted $p<0.1$ (false discovery rate<10%). (B) Volcano plot of reads from control and HD motor cortex, colors as in A. The percent change is calculated as the difference in the isoform fraction between HD and control samples multiplied by 100. (C) Scatter plot of PAS-seq reads from control (n=7) and HD

(n=9) cerebellum, as in A. Green dots indicate isoforms with p<0.01. (D) Volcano plot of reads from control and HD cerebellum, colors as in C.

Neurons die and glia proliferate in the HD brain[29]. Although neuronal loss is limited to around 10% in grade 1 motor cortex, it is possible some or all isoform changes I see are due to changes in cell populations[183]. I found isoform abundance is highly correlated (r=0.87) between the motor cortex and cerebellum despite the different cell milieu and limited neuronal loss in the cerebellum (Fig 3.3). The correlation of isoform abundance between brain regions suggests isoform abundances do not greatly differ between cell types, and a 10% neuronal loss is unlikely to cause substantial shifts in isoform abundance. Previous studies that used HD caudate, which suffers extensive neuronal loss (50-95%), have demonstrated mRNA expression is similar in HD caudate tissue and HD caudate neurons[138]. These results suggest the isoform changes I observe in HD motor cortex are not solely due to neuronal loss.

To determine whether genes with significant (p<0.01) isoform changes shifted towards longer or shorter mRNA 3′UTR isoforms, I calculated a weighted change (see methods). If the weighted change is positive, the gene shifts towards longer mRNA isoforms in HD patients, whereas if it is negative, the gene shifts towards shorter isoforms. I found 129 genes shift towards longer isoforms in HD motor cortex, whereas 110 genes shift towards shorter isoforms (Fig. 3.4A). In the cerebellum, I found 31 genes shifted to longer isoforms, whereas 83 shifted to shorter isoforms (Fig. 3.4B). Thus isoform shifts are gene-specific, and there is no generalized shift towards longer or shorter isoforms in HD.

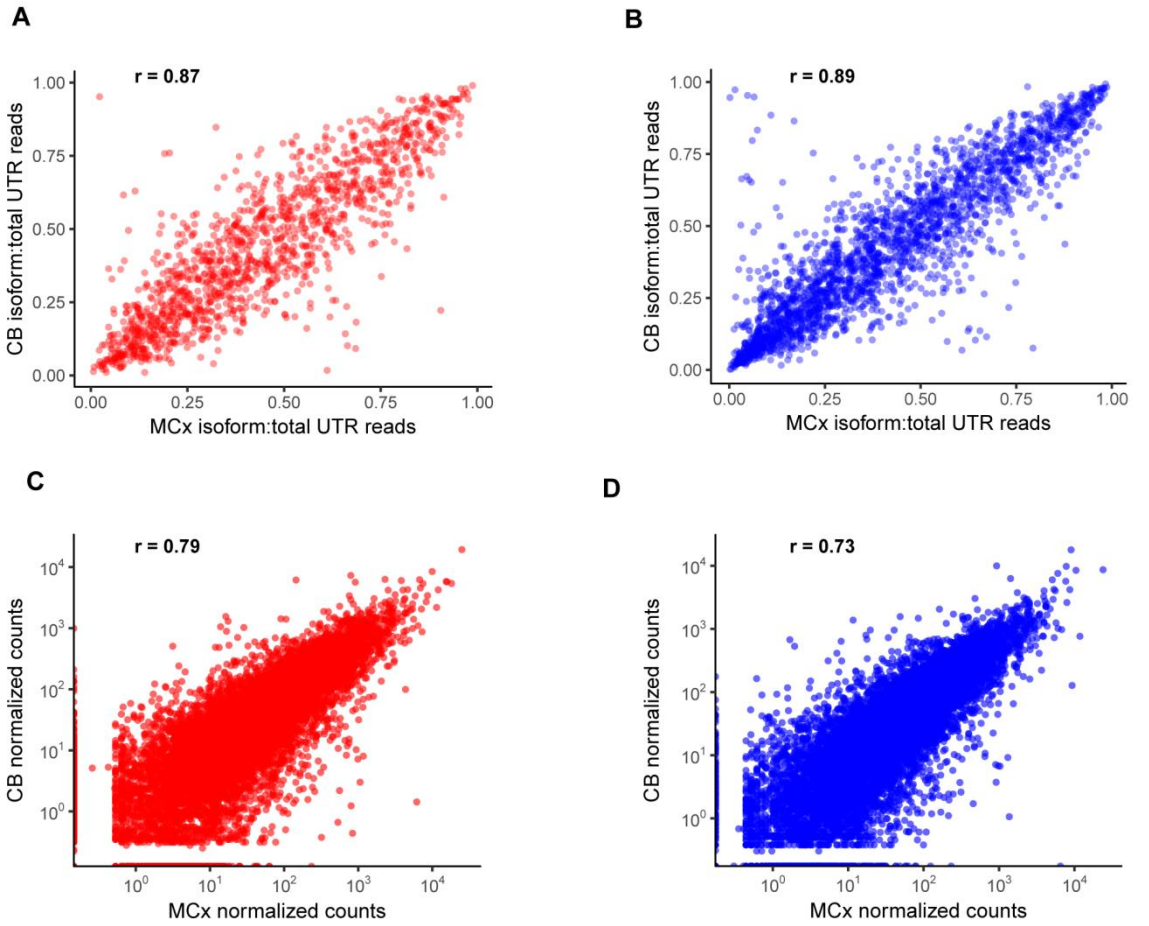


Figure 3.3. The expression of most isoforms and genes is highly correlated between cerebellum and motor cortex.

(A) Scatter plot of PAS-Seq reads from HD motor cortex (MCx, n=6) and cerebellum (CB, n=9). Each dot represents a 3'UTR isoform. Shown is the number of reads mapping to the isoform divided by the total reads mapping to the gene 3′UTR, and the Pearson correlation coefficient (r) of isoform expression between the brain regions. (B) Scatter plot of PAS-Seq reads as in A, from control motor cortex (n=5) and cerebellum (n=7). (C) Scatter plot of PAS-Seq reads from HD motor cortex (MCx, n=6) and cerebellum (CB, n=9). Each dot represents a gene. Shown is the total number of reads mapping to the gene normalized to sequencing depth (normalized counts), and the Pearson correlation coefficient (r) of gene expression between the brain regions. (D) Scatter plot of PAS-Seq reads as in C, from HD motor cortex (n=5) and cerebellum (n=7).

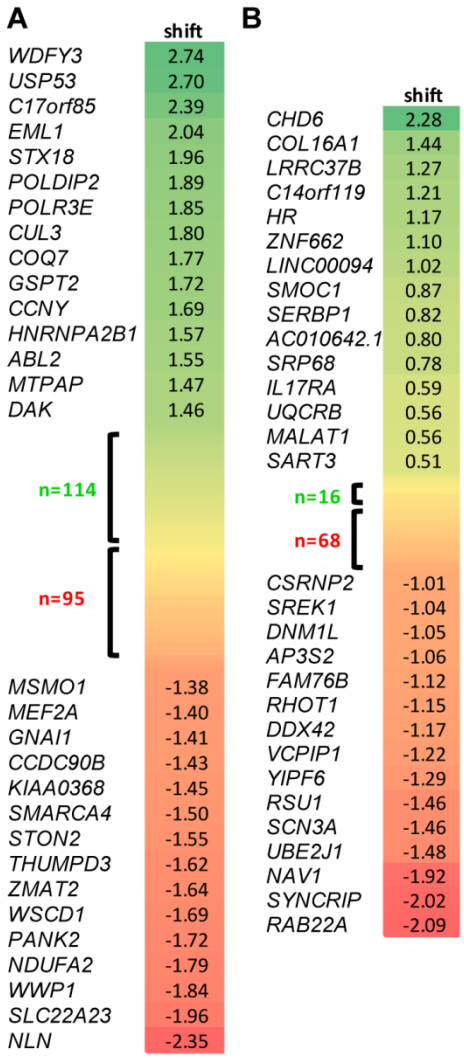


Figure 3.4. There is no global shift to longer or shorter 3′UTR isoforms in HD patient brains.

(A) Heat map of the weighted change (see methods) of genes with one or more isoform differentially expressed in patient motor cortex ($p<0.01$). A positive number indicates the gene shifts towards longer isoforms in HD, whereas a negative number indicates the gene shifts towards shorter isoforms in HD. (B) Heat map of genes with one or more isoform significantly differentially expressed in patient cerebellum, as in A.

I compared significant ($p<0.01$) isoform shifts in the motor cortex and cerebellum to look for common changes. Only 11 genes exhibit mRNA 3′UTR isoform shifts in both the motor cortex and cerebellum (Fig. 3.5A, purple). Of these, 9 showed opposite changes in the motor cortex, with 7 of the 9 shifting towards shorter isoforms in the cerebellum and longer isoforms in the motor cortex. This finding indicates that isoform changes in HD are region-specific.

3.2.2 Genes with isoform shifts are involved in pathways associated with HD pathogenesis.

To identify affected pathways, I performed gene ontology analysis on genes with significant ($p<0.01$) isoform shifts. I found genes that shift to longer 3′UTR isoforms are enriched for different pathways compared to genes that shift to shorter isoforms (Fig. 3.5B). Several of these pathways have been reported to be enriched among genes differentially expressed in HD, including cytokine signaling and production, RNA pol2 transcription, translation, calcium signaling, vesicle-mediated transport, microtubule organization, and DNA binding[138,180]. Some enriched pathways are thought to be involved in HD pathogenesis, including calcium signaling, cytokine signaling, histone acetylation, transcription factor binding, axonal transport, synaptic vesicle localization, dendritic spine morphogenesis, microtubule organizing, and phagocytic vesicles[197–201]. These results suggest aberrant 3′UTR isoform expression may influence pathogenesis in these disease pathways.

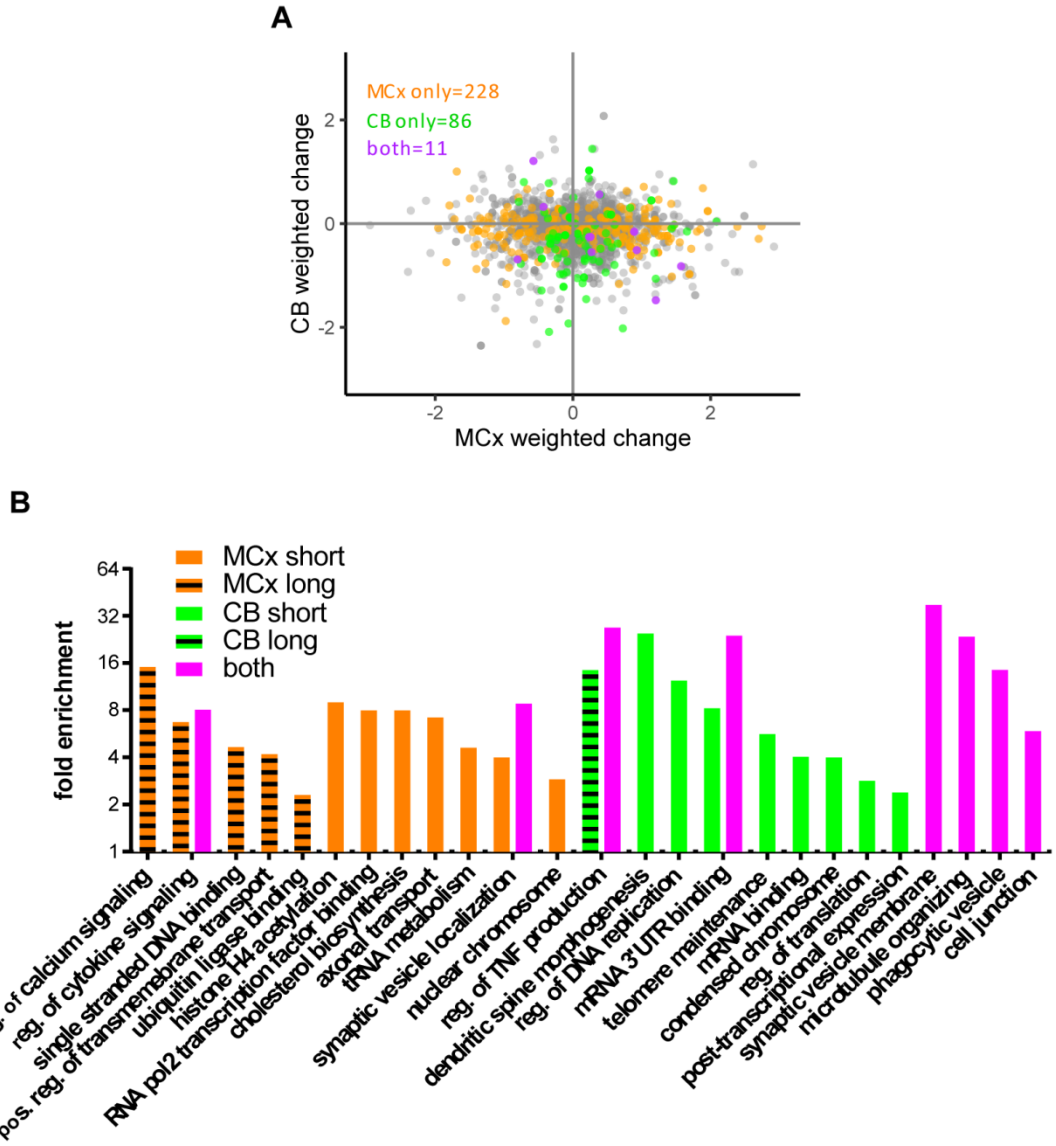


Figure 3.5. HD-associated pathways are enriched among genes with isoform changes.

(A) Comparison of gene weighted changes between patient versus control motor cortex (MCx) and cerebellum (CB). Green dots have significant changes in HD cerebellum, orange dots have significant changes in HD motor cortex, and purple dots have both ($p<0.01$). (B) Gene ontology analysis of genes exhibiting significant ($p<0.01$) shifts to longer (long) or shorter (short) isoforms in HD motor cortex (MCx) and cerebellum (CB).

3.2.3 Many genes are differentially expressed in HD motor cortex.

Widespread gene expression changes are reported in the cortex of late-stage Huntington's disease patients[180]. I sought to determine if the extensive 3′UTR isoform changes in HD motor cortex co-occur with gene expression changes. PAS-seq captures polyadenylated mRNA isoforms. To measure steady-state gene expression, I summed all PAS-seq reads, or isoforms, from each gene (see methods). Consistent with previous studies, many genes were differentially expressed in HD patient motor cortex relative to control motor cortex (false discovery rate<10%, 1.5-fold change) (Fig. 3.6A)[138,180]. Of differentially expressed genes, 700 exhibited increased expression in HD motor cortex, whereas 1008 exhibited decreased expression (Fig. 3.7A). The percent of detected genes exhibiting expression changes is identical to the percent of alternatively polyadenylated genes exhibiting isoform changes (11%), indicating isoform changes are proportional to gene expression changes in HD motor cortex. In contrast, only 1% of genes are differentially expressed (false discovery rate<10%, 1.5-fold change) in patient cerebellum, with 2.5% making the unadjusted p-value cutoff (Fig. 3.6B). Of differentially expressed genes, 89 increased expression in patient cerebellum whereas 64 decreased (Fig. 3.7B). This percent is also consistent with PAS-seq, which found 5% of genes made the unadjusted p-value cutoff for isoform changes. Thus, isoform changes accompany gene expression changes in the HD motor cortex, whereas isoform and gene expression changes are minimal in the cerebellum.

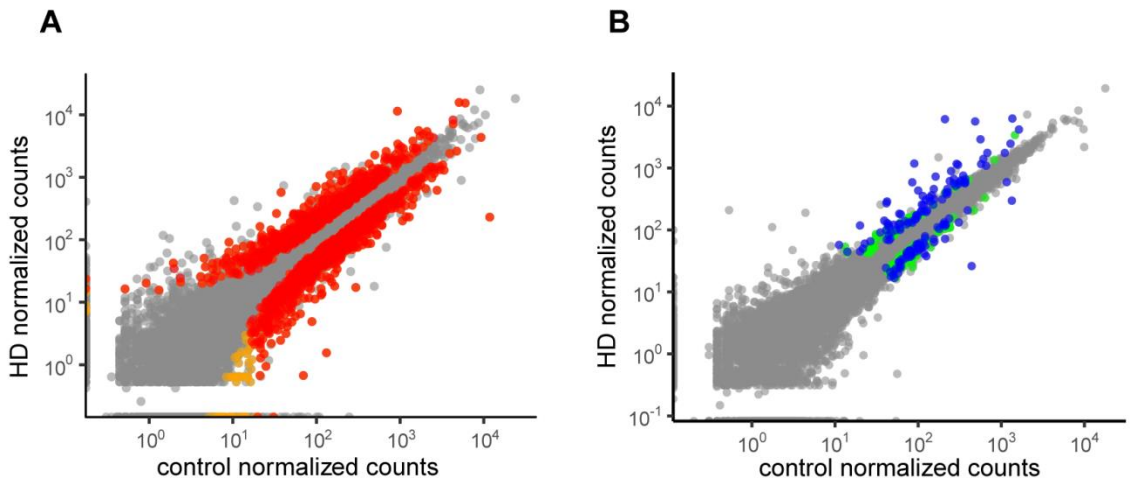


Figure 3.6. Gene expression analysis identifies a large subset of genes with expression changes in HD motor cortex.

(A) Scatter plot of PAS-seq reads from control (n=5) and HD (n=6) motor cortex. Each dot represents the normalized expression of a gene. Orange dots indicate isoforms with $p<0.01$, whereas red dots indicate isoforms with adjusted $p<0.1$ (false discovery rate<10%). (B) Scatter plot of PAS-seq reads from control (n=7) and HD (n=9) cerebellum, as in A. Green dots indicate isoforms with $p<0.01$, whereas blue dots indicate isoforms with adjusted $p<0.1$ (false discovery rate<10%).

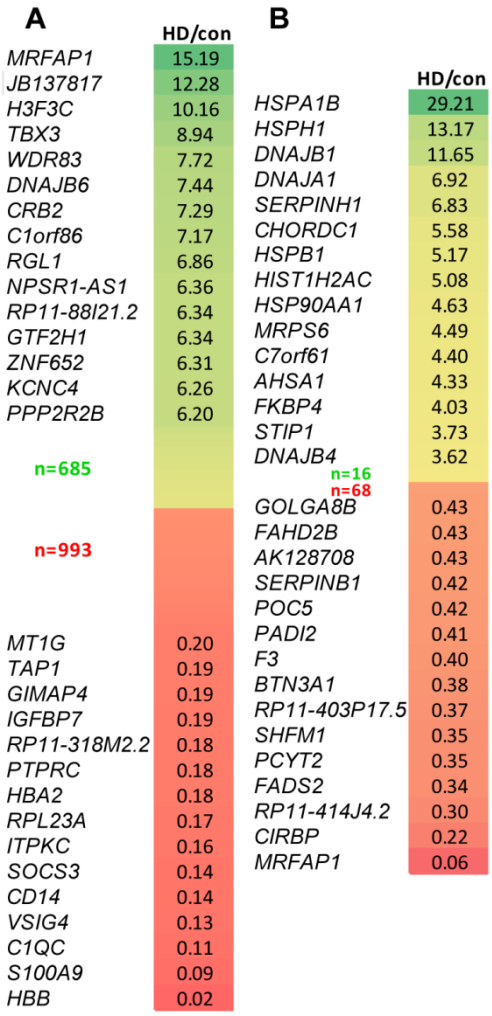


Figure 3.7. Most differentially expressed genes are down-regulated in HD brains.

(A) Heat map of genes significantly differentially expressed in patient motor cortex (false discovery rate<10%). Shown are HD normalized read counts divided by control normalized read counts. (B) Heat map of genes significantly differentially expressed in HD cerebellum (false discovery rate<10%), as in A.

3.2.4 Many isoforms and genes switch expression between HD grade 1 and grade 2.

I compared my gene expression results to those of previous HD patient studies that used microarrays of motor cortex and cerebellum or RNA sequencing from prefrontal cortex[138,180]. I found many genes reported differentially expressed in HD patient brains were also differentially expressed in my data. However, some genes previously reported to increase in HD motor cortex compared to controls were decreased in my samples, and some genes previously reported to decrease were increased in my samples. Quantitative PCR of three of these genes validated my PAS-seq results (Fig. 3.8). In contrast, my cerebellum data largely agreed with previous studies. I speculated that these discrepancies arose because previous studies used motor cortex from a mixture of HD grades, whereas I used motor cortex from grade 1 brains exclusively.

To investigate whether gene expression changes with disease progression, I performed PAS-seq on grade 2 HD motor cortex (n=4) (Fig. 3.9). Although my statistical power was limited by small sample size, I found the magnitude of significant 3′UTR isoform changes was greater than in grade 1, with most genes (68%) shifting to longer isoforms (Fig. 3.9A and D, 3.10A). PAS-seq revealed three outlier genes with extreme 3′UTR isoform shifts in grade 2 motor cortex: *NDFIP1*, *AP1G1*, and *CTSB* (Fig. 3.9A,B). NDFIP1 protein plays a vital role in healing cortical injury, while AP1G1 is essential for formation of clathrin-coated vesicles[202,203]. Axonal transport of these vesicles is impaired in HD neurons[204]. Interestingly, CTSB protein was recently found to mediate the beneficial effect of exercising on hippocampal neurogenesis, an effect that is lost in mouse models of HD[205,206].

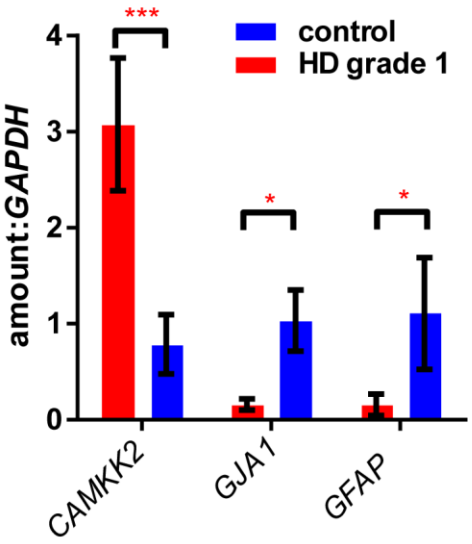


Figure 3.8. qPCR validates select PAS-Seq significant gene expression changes. Grade 1 motor cortex qPCR validation of three differentially expressed genes previously reported down-regulated (CAMKK2) or up-regulated (GJA1, GFAP) in mixed-grade HD motor cortex. Shown is average expression across samples (n=3-4 per group) with standard deviation. *, **, and *** signify $p<0.05$, 0.005, and 0.0005.

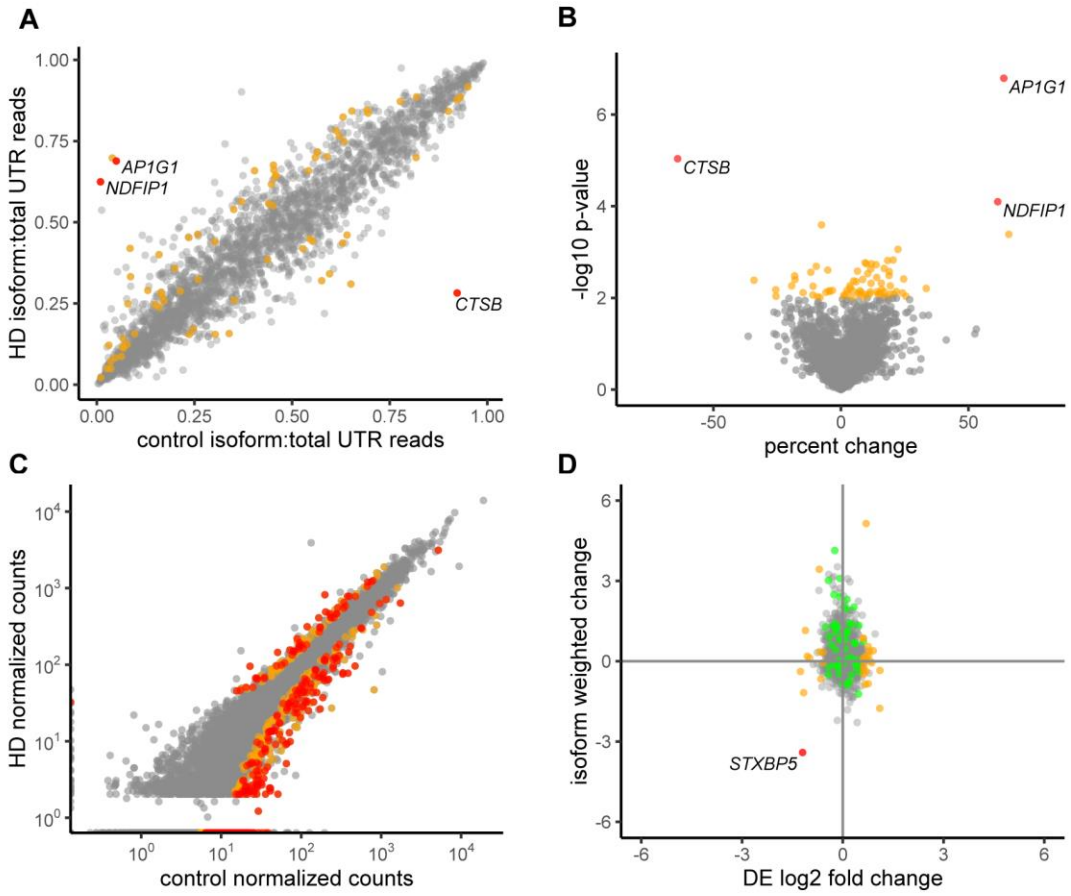


Figure 3.9. Many isoforms and genes are differentially expressed in motor cortex from HD patient grade 2 brains.

(A) Scatter plot of PAS-Seq reads from control (n=5) and HD (n=4) grade 2 motor cortex. Each dot represents a unique 3'UTR isoform expressed in at least four HD and control samples. Shown is the number of reads mapping to the isoform divided by the total reads mapping to the gene 3′UTR (isoform fraction). Orange dots indicate isoforms with $p<0.01$, while red dots indicate isoforms with adjusted $p<0.1$ (false discovery rate<10%). (B) Volcano plot of reads from control and HD grade 2 motor cortex, colors as in A. The percent change is calculated as the difference in the isoform fraction between HD and control samples. (C) Scatter plot of PAS-Seq reads from control and HD grade 2 motor cortex. Each dot represents a gene. Shown is the total number of reads mapping to the gene normalized to sequencing depth (normalized counts). Orange dots

indicate isoforms with $p<0.01$, while red dots indicate isoforms with adjusted $p<0.1$ (false discovery rate<10%).

In contrast to motor cortex from grade 1 samples, the majority of genes (86%) exhibited decreased expression in motor cortex from grade 2 samples (Fig. 3.9C). I identified eight genes with increased expression in previous HD studies that had significantly decreased expression in my grade 1 samples, and six genes with decreased expression in previous HD studies that had significantly increased expression in my grade 1 samples[138]. Eleven of these genes switched gene expression between grade 1 and grade 2 HD motor cortex (i.e., from increased expression relative to controls to decreased expression), and the other three moved from greater than 1.5-fold change to less than 1.5-fold change (Table 3.1; Fig. 3.10B, gold dots). Surprisingly, many other genes with significantly increased expression in grade 1 cortex exhibit significantly decreased expression in grade 2 cortex, and vice versa (Fig. 3.10C, light blue in quadrants II and IV). These results suggest isoform and gene expression changes in HD motor cortex are grade-specific, and gene expression studies should not combine several grades of HD cortex.

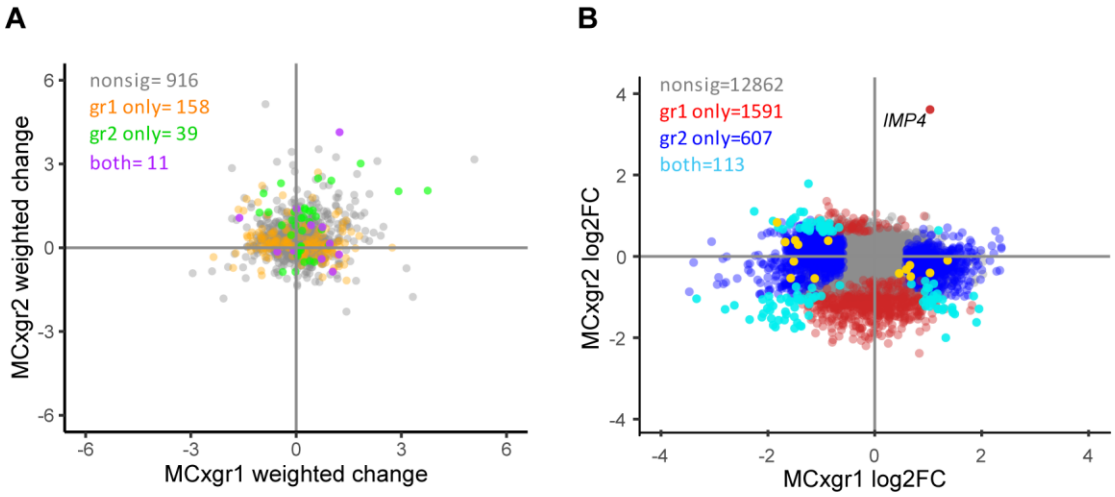


Figure 3.10. Many isoforms and genes switch expression between HD grade 1 and grade 2.

(A) Comparison of gene isoform weighted changes and differential expression (DE) log 2 fold changes in HD versus control grade 2 motor cortex. The weighted change is the difference in the average weighted isoform ratio between HD and control samples (see methods). Green dots have significant ($p<0.01$) isoform changes, orange dots have significant expression changes ($p<0.01$), and red dots have both. (B) Comparison of gene weighted changes in patient versus control grade 1 motor cortex (MCxgr1) and grade 2 motor cortex (MCxgr2), calculated as in A. (C) Comparison of gene expression log 2 fold changes in motor cortex grade 1 and grade 2 motor cortex. Orange dots were previously reported to have opposite expression changes than in my grade 1 data.

Table 3.1. Top genes with opposite direction expression changes in my data compared to previous studies.

	MCxgr1 L2FC	MCxgr2L2FC
ATP2B2	1.36	-0.10
EGR4	1.03	-0.41
CAMKK2	0.67	-0.50
SORT1	0.66	-0.22
PSEN1	0.59	-0.32
LAPTM4B	0.46	-0.42
SOX9	-0.87	0.39
FGF2	-1.12	-0.54
SLC14A1	-1.44	0.29
MT1H	-1.49	0.40
SLC1A3	-1.51	-0.12
CD44	-1.57	-0.53
MT1G	-1.68	0.35
GJA1	-1.83	0.84

Shown are the log2 fold changes (L2FC) in motor cortex (MCx) from grade 1 (gr1) and grade 2 (gr2) HD patient brains for fourteen genes.

3.2.4 Most genes with isoform changes in HD are not differentially expressed.

To determine if genes with isoform changes also exhibit expression changes, I compared gene expression to isoform expression for all alternatively polyadenylated genes in the motor cortex and cerebellum using a significance cutoff of $p<0.01$ (Fig. 3.11A,B). In the motor cortex, only 17% of genes with isoform shifts also exhibit gene expression changes (Fig. 3.11A, red). Of genes exhibiting isoform and gene expression changes, I found no association between isoform shift and gene expression direction, with

about equal numbers of genes in all four quadrants. This dissociation of isoform and gene expression is consistent with studies showing the association between 3′UTR length and isoform stability is gene-specific[112]. In the cerebellum, only 4% of genes with significant isoform changes are also differentially expressed (Fig. 3.11B, red). Of these, all are shifted towards shorter isoforms with two exhibiting increased expression and two exhibiting decreased expression. These results indicate that most genes with isoform changes are not differentially expressed in HD motor cortex and cerebellum.

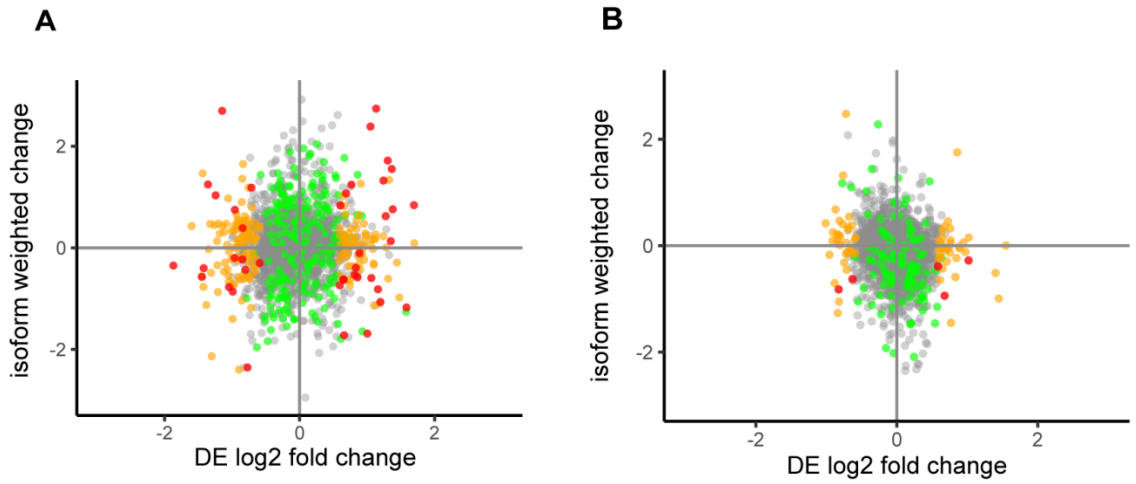


Figure 3.11. Most genes with 3′UTR isoform changes do not exhibit expression changes in HD.
(A) Comparison of gene isoform weighted changes (see methods) and differential expression (DE) log 2 fold changes in HD versus control motor cortex. Green dots have significant isoform changes in HD, orange dots have significant expression changes in HD, and red dots have both ($p<0.01$). (B) Comparison of gene isoform weighted changes and expression log 2 fold changes in HD versus control cerebellum, as in A.

I compared gene expression between the motor cortex and cerebellum to see if there is overlap in differentially expressed genes. I found sixteen (11%) genes differentially expressed in the cerebellum are also differentially expressed in motor cortex (Fig. 3.12A, light blue). Of these, eleven are inversely expressed in motor cortex. I performed gene ontology analysis to identify pathways enriched ($p<0.01$) among genes with increased or decrease expression in the motor cortex, cerebellum, or both (Fig. 3.12B).

My analysis identified several pathways previously reported to be enriched among genes differentially expressed in HD motor cortex: CNS development, protein localization, extracellular exosome, fatty acid biosynthesis, and regulation of TNF signaling[138,180]. I provide a more detailed gene ontology analysis than previous studies, separating genes into categories based on region and direction of change. My data reveals many pathways enriched in both genes with decreased expression exclusively in motor cortex, and genes with increased expression exclusively in cerebellum. Some of these pathways, including protein folding and ubiquitination, are vital for clearance of the mutant HTT protein. Region-specific regulation of these processes may be related to region-specific pathology. In contrast, genes with expression changes in both motor cortex and cerebellum are enriched for immune and stress responses, suggesting dysfunction of these pathways is not region-specific.

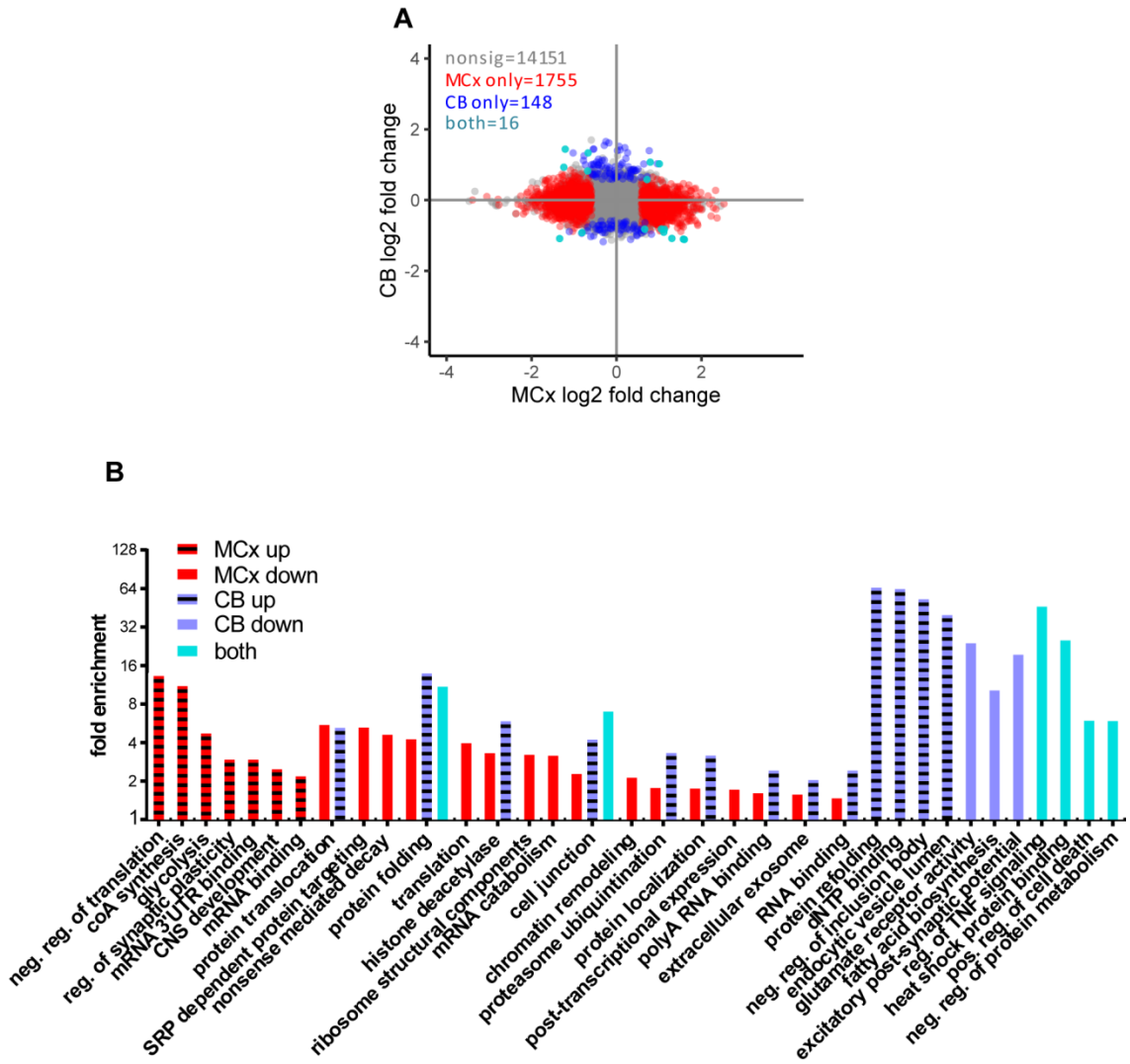


Figure 3.12. Pathways enriched among differentially expressed genes are region-specific.

(A) Comparison of gene expression log-2 fold changes (in motor cortex (MCx) and cerebellum (CB). Red dots have significant expression changes in HD motor cortex, blue dots have significant changes in HD cerebellum, and light blue have both (false discovery rate<10%). (B) Gene ontology analysis of genes exhibiting significant (false discovery rate<10%) expression changes in HD motor cortex (MCx) and cerebellum (CB).

I sought to determine if isoform and gene expression changes are associated with common cell pathways. I found several ontology categories enriched among genes with isoform changes and genes with expression changes including mRNA 3′UTR binding, cell junction, ubiquitination, mRNA binding, and protein transport (Fig. 3.5B, 3.12B). Within each category, genes with isoform differences and genes with expression changes do not overlap; i.e., the same gene does not exhibit aberrant isoform and gene expression. These results indicate that isoform alterations and gene expression changes are separate, but they may affect common processes in HD brains.

3.2.5 Most genes do not show an increase in non-UTR isoforms in HD. Many mRNAs change their splicing patterns in HD[117,173]. These changes may lead to increases in non-3′UTR isoforms, including isoforms that are misspliced or terminated early, such as the truncated *HTT* isoform[40]. To identify non-UTR isoforms, I separated and summed reads mapping to the 3′UTR or to the open reading frame of every gene. I then compared the fraction of non-3′UTR reads in disease and control samples. I found very few genes (2%) with a significant ($p<0.01$) increase in non-3′UTR reads in both patient motor cortex and cerebellum compared to controls, suggesting misspliced transcripts retain the canonical 3′UTR rather than shifting to non-3′UTR polyA signals. *HTT* itself was unchanged, as expected given the low abundance of the truncated intron one isoform[40]. I did not detect the truncated isoform in my data, potentially because PCR of boundary of *HTT* exon 1 and intron 1 is difficult due to its GC-rich content[207].

3.3 Discussion

HD brains exhibit extensive alterations in mRNA expression and splicing[117,138]. I studied whether widespread mRNA 3′UTR isoform changes occur in HD. Using PAS-seq, I identified a large subset of genes with mRNA 3′UTR isoform shifts in HD motor cortex. Genes with isoform changes were involved in pathways known to be aberrant in HD. In contrast, the cerebellum, which is less affected in HD, exhibited minimal changes. As expected, many genes were differentially expressed in HD motor cortex. Differentially expressed genes were mostly separate from genes with isoform changes, although they were involved in similar pathways.

This is the first study identifying widespread 3′UTR isoform alterations in HD. I expect my results are not due to neuronal loss, as motor cortex from grade 1 and cerebellum from grade 2-4 HD brains exhibit minimal cell loss, and isoform abundances for most genes are well-correlated between motor cortex and cerebellum. One caveat of PAS-seq is that it only detects polyadenylated isoforms. Neuron cell bodies contain deadenylated pools of mRNA[208]. Thus some isoform differences I detected may be due to altered deadenylation in HD. I only investigated changes in terminal 3′UTR isoforms. Alternative polyadenylation can also occur in non-canonical 3′UTRs or in introns. Finally, some of my post-mortem samples had low RNA integrity. I found as the RNA integrity number (RIN) fell below a score of 3 (see methods), PAS-seq reads shifted away from the 3′UTR to the open reading frame, indicating mRNA 3′ end degradation. However, I selected only high quality samples for PAS-seq, and RINs were similar between HD and control samples.

Surprisingly, although my HD grade 1 motor cortex differentially expressed genes overlap extensively with previous studies on mixed-grade samples, many show opposite changes in my samples[138,180]. I found almost all of these discrepant genes change expression between grade 1 and grade 2. Although many genes have increased expression in grade 1 motor cortex, few genes show increased expression in grade 2, suggesting loss of transcriptional activators, gain of transcriptional repressors, or changes in mRNA stability with disease progression. I found HDAC transcription repressors are enriched amongst genes down-regulated in motor cortex from grade 1 but not grade 2 HD brains. Decreased expression of these genes may lead to increased transcription in motor cortex from grade 1 brains that is lost by grade 2, when HDAC expression normalizes. Interestingly, HDAC inhibitors increase neuronal survival in HD models[197].

Widespread changes in isoform abundance may lead to aberrant mRNA metabolism in HD. Some tandem 3′UTR isoforms exhibit unique localization and function. The long but not short 3′UTR isoform of brain-derived neurotrophic factor mRNA is translated in dendrites, where the protein is vital for pruning and spine enlargement[94]. The long isoform of *CD47* mRNA acts as a scaffold for proteins that translocate the CD47 protein to the cell membrane, whereas the short isoform localizes the protein to the endoplasmic reticulum[97]. Loss of normal mRNA isoform abundance can contribute to disease. The alpha synuclein long 3′UTR isoform is preferentially localized to mitochondria; increased abundance of the extended alpha synuclein mRNA may contribute to mitochondrial dysfunction in Parkinson's disease[115]. Several genes change their 3′UTR isoform amounts in cancer and myotonic dystrophy[110,178]. In cancer

cell lines, oncogenes shift to shorter isoforms, resulting in increased protein production[110].

I demonstrate that genes with isoform changes in HD motor cortex are associated with pathways disrupted in HD including calcium signaling, cytokine signaling, histone acetylation, transcription factor binding, axonal transport, synaptic vesicle localization, dendritic spine morphogenesis, microtubule organizing, and phagocytic vesicles. Altered metabolism of mRNAs in disease pathways could result in aberrant localization or function of proteins implicated in HD pathogenesis. I didn't see a concerted shift to shorter or longer isoforms in HD motor cortex, but studies have shown stability is isoform-specific[112]. Proteomic work is necessary to determine if genes with isoform changes exhibit alterations in protein abundance, and whether the proteins are involved in HD pathology.

3.4 Materials and methods

3.4.1 PAS-seq library preparation. For PAS-seq method including adaptor, primer, and barcode sequences, see Ashar-Patel 2017, manuscript in preparation. Briefly, 5-10μg total RNA was collected using Trizol (Ambion), and RNA quality was assessed by Bioanalyzer (Agilent). RNA was treated with TurboDNAse (Ambion) and cleaned with Clean and Concentrator columns (Zymo). Cleaned RNA was fragmented in 1x SuperScript III first strand buffer (Invitrogen) at 94°C for 5 minutes 30 seconds and immediately placed on ice. SuperScript III (Invitrogen) reverse transcribed fragmented RNA from the custom PAS-seq oligo d(T) primer, and 20 minute incubation with RNAse I hydrolyzed remaining RNA (Life Technologies). Samples were run on a denaturing 8M urea, 10% polyacrylamide gel, and 160-200 nucleotide cDNA fragments were eluted overnight in elution buffer (300mM NaCl, 10mM EDTA). Spin-X columns (Corning) removed remaining gel pieces, and eluted cDNA was precipitated by incubation with 50% isopropanol at -20°C for 30 minutes followed by centrifugation at maximum speed for 45 minutes at 4°C. The pellet was washed with 70% ethanol and collected by spinning at maximum speed for 10 minutes at 4°C. Dried pellets were resuspended in water. cDNA was circularized by incubation with CircLigase II (Epicentre) at 60°C for 4 hours. Phusion polymerase (NEB) amplified the insert and adaptors with PE 1.0 and PE 2.0 primers (Illumina) by incubation at 98°C for 30 seconds; 14 cycles of 98°C for 5 seconds, 62°C for 10 seconds, and 72°C for 10 seconds; and a final extension of 72°C for 2 minutes. The PCR products were run on a non-denaturing 10% polyacrylamide gel. The 215-254 nucleotide library was extracted and eluted/precipitated as above. Libraries were

Bioanalyzed (Agilent) for quality control, and sequenced with single-end 100 base-pair sequencing on Illumina HiSeq 2000 or MiSeq instruments.

3.4.2 PAS-seq analysis: isoform and gene expression quantification. I sorted indexed reads into samples and removed the barcode and randomer (see Ashar-Patel 2017, in preparation). Cutadapt trimmed the 3′ polyA tail and low quality bases (phred<25), excluding reads less than 15 nucleotides[209]. I checked the quality of the remaining reads using Fastqc (Andrews, S., 2010, http://www.bioinformatics.babraham. ac.uk/projects/fastqc) before mapping them to the hg19 or mm10 genome with Bowtie [210]. Bedtools converted the bam files to bed files and computed the read 3′ end coverage of each coordinate (PAS-seq read 3′ ends represent polyA sites) [211]. Because the site of cleavage and polyA is heterogeneous, I wrote a peak-calling program to combine reads within 24 nucleotides of the coordinate with the most 3′ end coverage for each polyA site. I used 24 nucleotides because the majority of cleavage events from the same polyA signal occur within 24 nucleotides of the most 5' cleavage site [187]. My peak-calling program generated slightly different coordinates between samples for the same isoform due to small inter-sample variations in isoform 3′ end coverage. I considered isoform peaks between samples to be the same if they were within 40 nucleotides of each other, as 40 nucleotides is the maximum difference between cleavage sites of the same isoform [187]. I mapped peaks with more than five reads to gene ORF or 3′UTRs, excluding peaks with non-unique mapping. For each gene, I defined the 3′UTR as the terminal (most 3′) coordinates from the UCSC table browser. I used CleanUpdTSeq to remove genomically-

primed peaks, a Bayesian classifier that performs better than heuristic methods at removing false positives while retaining true positives [185]. I identified alternatively polyadenylated genes as genes with more than one 3′UTR peak. For differential expression analysis, I combined all reads mapping to a gene, either the ORF or the 3′UTR. Raw and processed data can be accessed at http://www.ncbi.nlm.nih.gov /geo/query/acc.cgi?acc=GSE96099. My full bioinformatics pipeline is available in Appendix 2.

3.4.3 PAS-seq analysis: HD versus control isoform comparison. I normalized the expression of each isoform in alternatively polyadenylated genes by dividing the isoform abundance by the total reads mapping to the gene 3′UTR. I compared the normalized isoform abundance between HD and control samples using t-tests with or without multiple-hypothesis testing correction using the Benjamini-Hochburg procedure. I considered an adjusted p-value <0.1 to be significant unless otherwise stated (false discovery rate<10%). I did not place a fold or percent change cutoff because small changes in isoform abundances of some genes may affect cellular functions. However, most (>75%) significantly changing isoforms changed by 5% or more, and those with smaller changes tended to be low-abundance isoforms. Only isoforms represented in at least five disease and control samples were considered. These and further PAS-seq analyses were performed using R 3.3.2 (R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/).

3.4.4 PAS-seq analysis: isoform weighted change. For each gene with a significant change in the abundance of at least one isoform, I determined a weighted change. I weighed each isoform ratio by the isoform number multiplied by ten. For instance, for a gene transcribed into two isoforms of equal abundance (isoform ratios of 0.5), the shorter isoform weighted ratio is 5 (0.5 x 10 x 1) whereas the longer isoform weighted ratio is 10 (0.5 x 10 x 2). I then calculated the difference in the average weighted isoform ratios between HD and control samples for the gene. A positive weighted change indicates a shift towards longer isoforms in HD samples, whereas a negative weighted change indicates a shift towards shorter isoforms in HD samples.

3.4.5 PAS-seq analysis: HD versus control gene expression comparison. The total reads mapping to each gene were compared between patients and controls using the DESeq2 differential expression package[212]. I considered genes with a fold change greater than 50% and false discovery rate <10% (Benjamini Hochburg correction) significantly differentially expressed.

3.4.6 Gene ontology analysis. For isoform analysis, the list of genes with significant ($p<0.01$) changes in at least one 3′UTR isoform in HD motor cortex or cerebellum were compared to the background total expressed APA genes in each region using Panther gene ontology analysis[213]. I considered categories significantly enriched if they had more than five genes, enrichment greater than 1.5 fold compared to background, and p value less than 0.01. For gene expression analysis, the list of genes with significant (false

discovery rate<10%) expression changes in HD motor cortex or cerebellum were compared to the background total expressed genes in each region using Panther gene ontology analysis as above.

3.4.7 PAS-seq validation RT-qPCR. Total RNA was extracted using Trizol (Ambion), and RNA quality was assessed by Bioanalyzer (Agilent). RNA was treated with TurboDNAse (Ambion), and cDNA was synthesized from 2µg total RNA using SuperScript IV (Invitrogen) with oligo $d(T)_{20}$ priming per manufacturer's instructions. Quantitative PCR was performed with 1.8µl cDNA, and QuantiFast SYBR green master mix (Qiagen). The CAMKK2, GJA1, and GFAP PrimePCR Assay primer mixes (Biorad) were used at 1x in the PCR reaction. Reactions were incubated at 95°C for 5 minutes, followed by 40 cycles of 95°C for 10 seconds and 60°C for 30 seconds on a StepOnePlus Real-Time PCR System (Thermo Fisher Scientific). All primers were validated by relative standard curve, and primer pairs amplified targets with 85-115% efficiency in the linear range. I verified primer specificity by running qPCR products on a 2% agarose, 1x TAE gel, extracting the bands using the QIAquick Gel Extraction Kit (Qiagen), and Sangar sequencing the bands with the forward and reverse PCR primers.

Table 3.2 Primers used in chapter 3 (AS=allele-specific, NAS=non allele-specific, IS=isoform specific

Primers (5′-3′)		
Target	**Forward**	**Reverse**
CAMKK2	PrimePCR Assay primer mix (Biorad)	
GJA1	PrimePCR Assay primer mix (Biorad)	
GFAP	PrimePCR Assay primer mix (Biorad)	
GAPDH	GAGTCAACGGATTTGGTCGT	TTGATTTTGGAGGGATCTCG
Reverse transcription primers (5′-3′)		
Isoform-specific	ACGCATCTATGCGCATATCGTTTTTTTTTTTTTT	

CHAPTER IV: RNA BINDING PROTEINS MAY AFFECT 3′UTR ISOFORM ABUNDANCE

Preface

The work presented in this chapter is accepted at *Cell Reports* as manuscript "Alterations in mRNA 3′UTR isoform abundance accompany gene expression changes in human Huntington's disease brains" by Romo, Ashar-Patel, Pfister, and Aronin.

4.1 Summary

RNA binding proteins mediate alternative polyadenylation, and altered expression of RNA binding proteins can lead to shifts in 3′UTR length[55]. There are many examples of altered expression of RNA binding proteins modulating alternative polyadenylation site selection. Increased levels of 3′ processing factor CstF-64 promote proximal polyA sites, whereas knockdown of CFIm68 also shifts mRNAs towards proximal polyA sites[60,71,72]. RNA binding proteins PTB, SXL, and PABPN1 promote distal polyA site usage[74,75,79]. Similarly, knockdown of U1 snRNP and ELL2 results in shifts to proximal polyA sites[64,81].

As many genes are differentially expressed in HD, I reasoned altered expression of an RNA binding protein could cause alterations in 3′UTR isoforms. I identified CNOT6 as an RNA binding protein differentially expressed in HD motor cortex. Knockdown of CNOT6 in wild-type fibroblasts results in *HTT* 3′UTR isoform shifts similar to those seen in HD fibroblasts and motor cortex. In addition, the 3′UTR isoforms of *SECISBP2L*, another gene with isoform shifts in HD, exhibit shifts similar to those in HD motor cortex.

4.2 Results

4.2.1 Decreasing expression of the RNA binding protein CNOT6 leads some genes to change isoform abundance. A change in the expression of an RNA binding protein may lead to changes in isoform abundance by altering isoform production rates during alternative polyadenylation, or by affecting isoform decay rates post-transcriptionally. To identify candidate RNA binding proteins, I searched for genes differentially expressed in HD motor cortex with the RNA binding protein GO-term (Fig. 4.1). Of those, twenty were involved in mRNA alternative polyadenylation or stability. I reasoned an RNA binding protein responsible for *HTT* isoform shifts would show opposite expression in motor cortex and cerebellum. Of the nine fitting that description, two were subunits of the Ccr4-not complex that plays a widepread role in mRNA metabolism[214,215]. Of the two, CNOT6 is more differentially expressed in HD motor cortex and cerebellum. CNOT6 is the subunit of the Ccr4-not complex that catalyzes deadenylation of mRNAs. Because CNOT6 is increased in HD cerebellum but decreased in motor cortex, I expected mRNAs targeted by CNOT6 to exhibit opposite changes in the cerebellum and motor cortex. For instance, if a CNOT6 target shifts to shorter 3′UTR isoforms in HD motor cortex, I expect it to change to longer isoforms in HD cerebellum. Indeed, 82% of genes with isoform alterations in motor cortex and cerebellum, including *HTT*, exhibit opposite isoform deviations.

	FC MCx	FC CB
YBX1	-3.09	1.43
CNOT6	-2.10	1.26
CNOT7	-2.00	1.21
CASC3	-1.60	1.12
CPEB4	1.72	-1.16
THOC5	2.13	-1.19
CPEB1	2.16	-1.26
CELF2	2.58	-1.43
PCBP2	2.61	-1.33

Figure 4.1. RNA binding proteins differentially expressed in HD motor cortex include Ccr4-not complex components CNOT6 and CNOT7. Algorithm to identify proteins that might cause 3′UTR isoform changes in HD patient brains identified *CNOT6* and *CNOT7*, members of the multi-functional Ccr4-not complex that plays a role in many aspects of RNA metabolism. . Numbers are fold changes (FC) in mRNA expression between HD and control motor cortex (MCx) or cerebellum (CB).

To determine if changes in *CNOT6* expression exert changes in isoform expression, I transfected control human fibroblasts with siRNAs targeting *CNOT6* or an off-target control (*MAP4K4*). The *CNOT6* siRNAs reduced *CNOT6* mRNA expression over 80% in these cells (Fig. 4.2, $p<0.0001$). I chose *HTT* and *SECISBP2L* (SECIS Binding Protein 2 Like) as candidate *CNOT6* targets. *HTT* and *SECISBP2L* shift to longer 3′UTR isoforms in HD cerebellum and shorter isoforms in HD motor cortex. I collected RNA from transfected cells for qPCR analysis of the *HTT* long 3′UTR isoform versus total *HTT*, and of the *SECISBP2L* long 3′UTR isoform versus total *SECISBP2L*. I found knock down of CNOT6 resulted in a decrease in the *HTT* and *SECISBP2L* long isoforms similar to that seen in HD patient motor cortex and in HD fibroblasts (Fig. 4.3A, $p=0.03$, 0.001). *HTT* isoform-specific qPCR found that, as in HD patient motor cortex, the abundance of the short isoform didn't change while abundance of the long isoform decreased by almost two-fold (Fig.4.3B, $p=0.03$). These results suggest changes in the expression of *CNOT6* may influence the abundance of some 3′UTR isoforms in patient motor cortex and cerebellum.

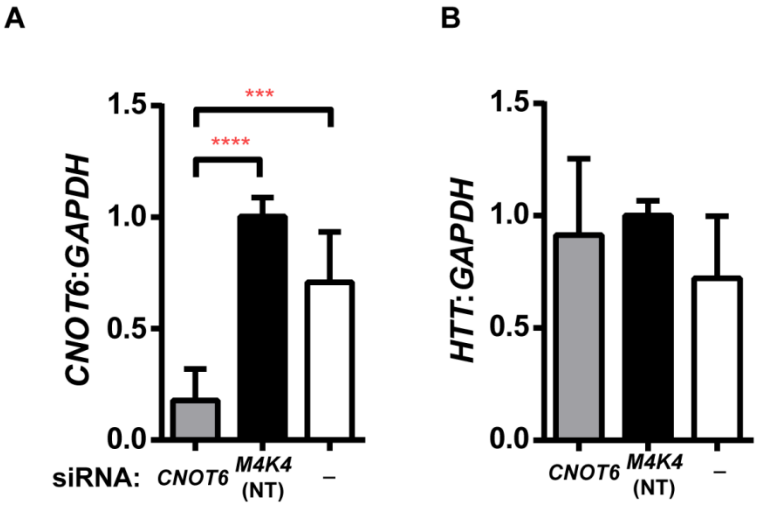


Figure 4.2. CNOT6 siRNAs reduce *CNOT6* but not *HTT* mRNA expression in wild-type fibroblasts. (A) qRT-PCR quantification *CNOT6* mRNA compared to *GAPDH* mRNA after transfection of *CNOT6* siRNA (gray), no siRNA (white), or non-targeting (NT) *MAP4K4* control siRNA (black) into wild-type fibroblasts. *, **, and *** signify $p<0.05$, 0.005, and 0.0005. (B) qRT-PCR quantification HTT (*HTT*) mRNA compared to *GAPDH* mRNA after transfection of *CNOT6* siRNA (gray), no siRNA (white), or non-targeting (NT) *MAP4K4* control siRNA (black) into wild-type fibroblasts.

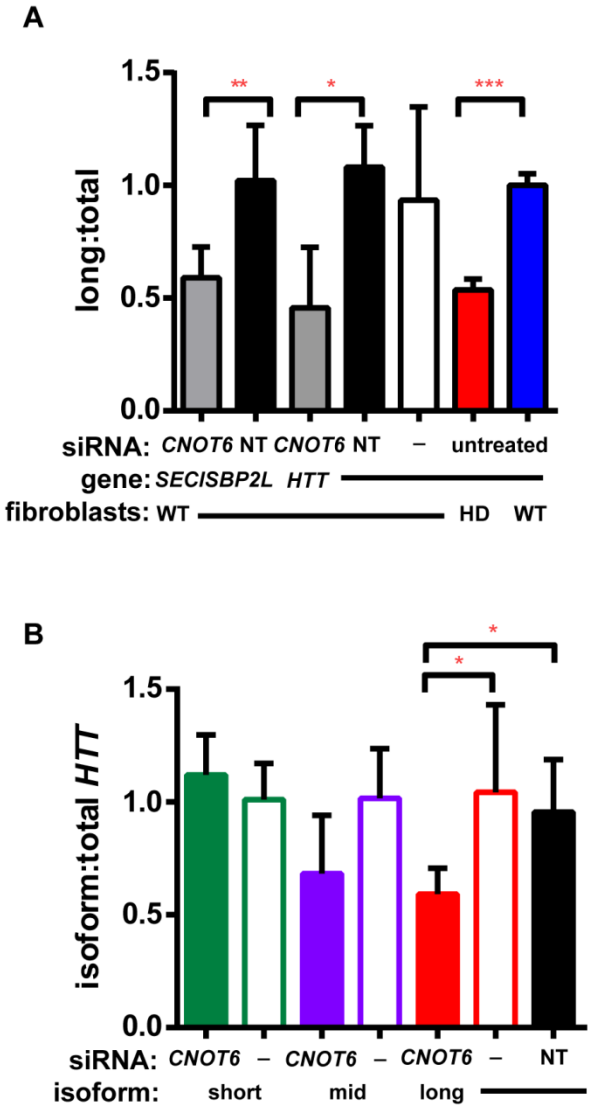


Figure 4.3. Changes in the expression of CNOT6 influence isoform abundances. (A) qRT-PCR quantification of the *HTT* long isoform (long) compared to total *HTT* expression (total) or of the *SECISBP2L* long isoform (long) compared to total *SECISBP2L* expression (total) after transfection of *CNOT6* siRNA (gray) or non-targeting (NT) *MAP4K4* control siRNA (black) into wild-type fibroblasts. Shown is the average of at least three transfections with standard deviation. Red and blue bars are untreated fibroblasts, as in Fig. 1D. *, **, and *** signify $p<0.05$, 0.005, and 0.0005. (B) Isoform-specific qRT-PCR quantification of each *HTT* long isoform (long) compared to

total *HTT* expression after transfection of *CNOT6* siRNA (colored), no siRNA (white), or non-targeting (NT) *MAP4K4* control siRNA (black) into wild-type fibroblasts.

4.3 Discussion

I sought to determine a possible mechanism for widespread 3′UTR isoform changes in disease. I found knockdown of CNOT6, an RNA binding protein involved in mRNA metabolism differentially expressed in HD brain, produced *HTT* mRNA isoform alterations similar to those seen in HD motor cortex. My results point to a model in which changes in the expression of RNA binding proteins in the HD brain lead to changes in the expression of mRNA 3′UTR isoforms (Fig. 4.4). RNA binding proteins may interact with nascent mRNA during transcription and alternative polyadenylation, leading to changes in polyA site selection and isoform production. For instance, aberrant expression of 3′ end modulatory factors known to correlate to polyA site selection may lead to isoform abundance changes[45]. I found three 3′ end processing factors are differentially expressed in HD grade 1 motor cortex: *CPSF2*, *PCBP2*, and *THOC5*. In addition to interacting with transcripts during alternative polyadenylation, RNA binding proteins may adhere to isoforms post-transcriptionally, affecting the stability and decay of isoforms and leading to changes in isoform abundance. I show a decrease in the CNOT6 deadenylase is associated with a decrease in the *HTT* long isoform, indicating the long isoform is more sensitive to changes in CNOT6 expression, perhaps due to its short poly-A tail.

Further studies are necessary to determine the mechanism of 3′UTR isoform changes in HD. I found *HTT* and *SECISBP2L* isoform amounts are responsive to changes in *CNOT6* mRNA levels. However, I only tested two mRNAs for response to *CNOT6* levels. Altered *CNOT6* expression may not explain all isoform changes I identified. Other RNA binding proteins, such as 3′ end processing factors in Fig. 4.1, likely affect isoform

abundance in HD. Exploration of the causes of isoform changes in HD motor cortex may result in novel therapeutic targets.

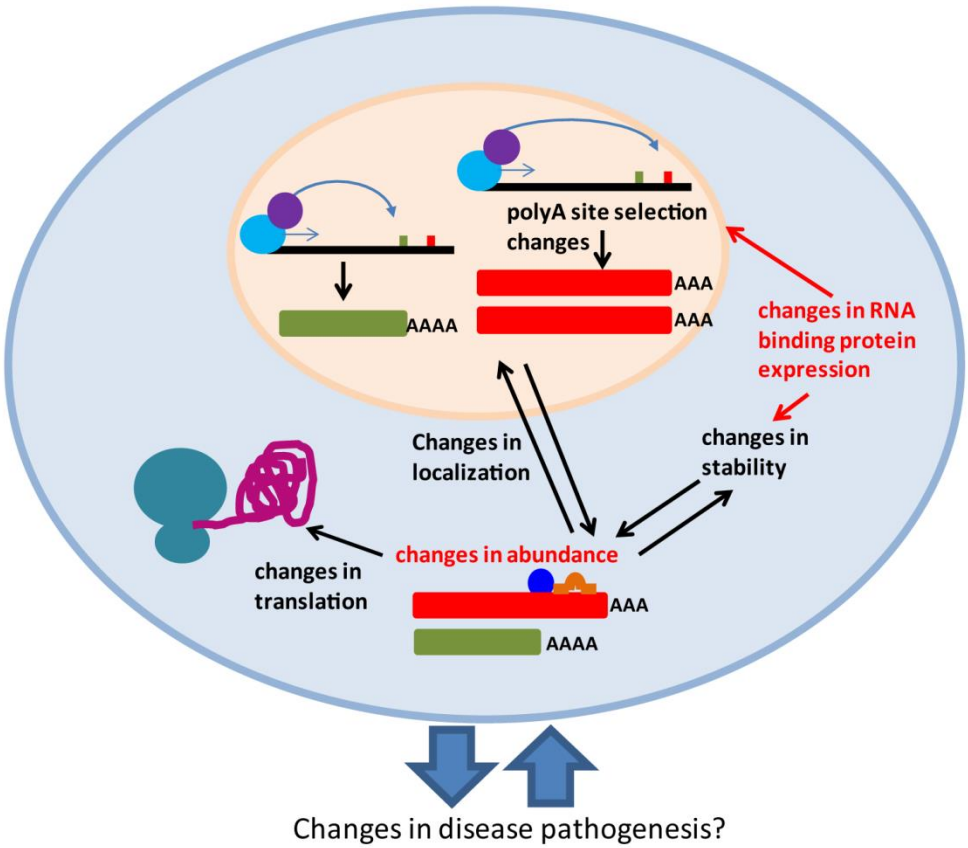


Figure 4.4. Changes in the expression of RNA binding proteins may influence alterations in isoform abundances in HD. Model of mRNA 3′UTR isoform changes in HD. Changes in the expression of RNA binding proteins lead to changes in mRNA 3′UTR isoform (red, green) abundance either by altering the production of some isoforms during transcription and alternative polyadenylation, or by affecting the stability of some isoforms post-transcriptionally. Widespread changes in isoform abundance likely lead to changes in stability, localization, and translation via different interactions with RNA binding proteins (dark blue) or microRNAs (orange), or differences in polyA tail length.

4.4 Materials and methods

4.4.1 CNOT6 siRNA transfection. I transfected wild-type fibroblasts with three CNOT6 siRNAs (Qiagen FlexiTube) using Lipofectamine RNAimax transfection reagent (Invitrogen) per manufacturer's instructions. After 72 hours, RNA was collected, DNAsed, reverse transcribed, and submitted to quantitive RT-PCR.

4.4.2 Quantitative RT-PCR. Total RNA was extracted using Trizol (Ambion), and RNA quality was assessed by Bioanalyzer (Agilent). RNA was treated with TurboDNAse (Ambion), and cDNA was synthesized from 2µg total RNA using SuperScript IV (Invitrogen) with oligo $d(T)_{20}$ priming per manufacturer's instructions. Quantitative PCR was performed with 1.8µl cDNA, 1µM forward and reverse primers, and QuantiFast SYBR green master mix (Qiagen). Reactions were incubated at 95°C for 5 minutes, followed by 40 cycles of 95°C for 10 seconds and 60°C for 30 seconds. on a StepOnePlus Real-Time PCR System (Thermo Fisher Scientific). The CAMKK2, GJA1, and GFAP PrimePCR Assay primer mixes (Biorad) were used at 1x in the PCR reaction. All primers were validated by relative standard curve, and primer pairs amplified targets with 85-115% efficiency in the linear range. I verified primer specificity by running qPCR products on a 2% agarose, 1x TAE gel, extracting the bands using the QIAquick Gel Extraction Kit (Qiagen), and Sangar sequencing the bands with the forward and reverse PCR primers.

Table 4.1 Primers used in chapter 4 (AS=allele-specific, NAS=non allele-specific, IS=isoform specific

Primers (5′-3′)

Target	Forward	Reverse
HTT long 3′UTR	ATGGATGCATGCCCTAAGAG	CAGTCTCCGATGAGCACAGA
total *HTT*	CATGGTGGGAGAGACTGTGA	CAAAGAGCACTTCTGCCACA
SECISBP2L long 3′UTR	TGAAGGGATTTTCATCTTTTGCTGT	AGTACAGAAGAATCCCCACTAGTAT
total *SECISBP2L*	CCCTCATCCCTTCCAACCTTT	ACTTTCAAGGTGTGATGTGCCA
CNOT6	ACAGAGTGCCTATGAGAGTGGC	CAGGATGCCTAAGGTGTTCAGC
GAPDH	GAGTCAACGGATTTGGTCGT	TTGATTTTGGAGGGATCTCG
(IS) *HTT* long 3′UTR	GGAAGGACTGACGAGAGATG	ACGCATCTATGCGCATATCG
(IS) *HTT* short 3′UTR	AGCAGGCTTTGGGAACACTG	ACGCATCTATGCGCATATCG
(IS) *HTT* mid 3′UTR	GTGGCAAGCACCCATCGTAT	ACGCATCTATGCGCATATCG

Reverse transcription primers (5′-3′)

Isoform-specific	ACGCATCTATGCGCATATCGTTTTTTTTTTTTTT

CHAPTER V: CONCLUSIONS

I show alternatively polyadenylated isoforms of full-length *HTT* mRNA, including a novel mid-3′UTR isoform, change their steady-state abundance in HD cells and tissues. Isoform changes are tissue-specific and arise from both *HTT* alleles. *HTT* isoform changes in HD likely impact total *HTT* mRNA metabolism: *HTT* isoforms have different localizations, stabilities, polyA tail lengths, microRNA sites, and RNA binding protein sites. Isoform changes extend beyond *HTT*. Eleven percent of mRNAs exhibit altered 3′UTR isoform abundance in HD motor cortex. Most genes with isoform changes are not differently expressed, but they are enriched in pathways involved in HD pathogenesis. These results suggest isoform changes may contribute to HD pathogenesis independently of gene expression changes. Altered expression of RNA binding proteins may affect isoform abundance in HD. I found knockdown of the multifunctional RNA binding protein CNOT6 results in altered isoform abundances in cells. This is the first study to characterize *HTT* 3′UTR isoform metabolism and show extensive 3′UTR changes are a feature of HD.

Isoform alterations are not surprising, given DNA and RNA expression and processing are deregulated in HD[41,123]. Further studies are necessary to elucidate the role of 3′UTR isoform changes in HD. Several questions require further investigation. The mechanism of isoform changes remains to be established, and it is unclear if or how changes in isoform expression cause pathology. The etiology of HD is multifactorial, and it is likely some molecular and cellular changes are causes, some are results, and some are epiphenomenon of pathology. Here, I discuss these unresolved questions and the directions future research could take to answer them.

5.1 Isoform changes may be due to differential expression of RNA binding proteins in HD.

I hypothesized isoform changes are due to differential expression of RNA binding proteins in HD. Isoform changes may arise due to altered synthesis of 3′UTR isoforms during alternative polyadenylation, or due to altered decay of 3′UTR isoforms after mRNA processing. I expect differential expression of RNA binding proteins involved in mRNA 3′ processing will affect isoform production, whereas differential expression of proteins involved in mRNA stability will affect decay. Changes in 3′UTR isoform synthesis or decay could alter isoform steady state abundance.

Knockdown of the deadenylase CNOT6, a subunit of the multifunctional Ccr4-Not complex that regulates mRNA metabolism, alters the expression of *HTT* and *SECISBP2L* 3′UTR isoforms. CNOT6 is more differentially expressed in HD motor cortex than in HD cerebellum, which is consistent with greater isoform changes in motor cortex than in cerebellum (Fig. 4.1). CNOT6 and the Ccr4-Not complex impact the stability and deadenylation of specific mRNAs[214]. Differential expression of *CNOT6* in HD could result in widespread changes in stability, polyA tail length, or both. Global trends in mRNA polyA tail lengths can be determined by polyA tail sequencing (PAT-seq) with or without CNOT6 knockdown in control fibroblasts[101]. If differential CNOT6 expression in HD impacts global polyA tail lengths, I would expect a transcriptome-wide shift in polyA tail length after CNOT6 knockdown. If polyA tail length changes occur, they may be a cause of isoform abundance alterations, as a transcript's polyA tail length can impact its stability[101,102]. CNOT6 may also modulate isoform stability independently

of deadenylation, and changes in polyA tail length maybe an epiphenomena to isoform abundance alterations.

The mRNA of several proteins involved in mRNA 3′ processing is differentially expressed in HD brains including Cleavage and Polyadenylation Specificity Factor 2 (CPSF2), Poly(C) Binding Protein 2 (PCBP2), and THO Complex Subunit 5 Homolog (THOC5). CPSF2 recognizes and binds the polyA signal during cleavage and polyadenylation[45]. PCBP2 binds poly-cytosine regions near polyA signals in the 3′UTR and enhances cleavage and polyadenylation[216]. THOC5 recruits cleavage factor 1 (CF1) to nascent mRNAs, promoting cleavage and polyadenylation[217]. Aberrant expression of these 3′ processing proteins in HD may affect synthesis of 3′UTR isoforms to impact steady-state 3′UTR isoform abundance. However, altered expression of 3′ processing proteins generally leads to global shifts in 3′UTR length to predominately shorter or longer isoforms, which I did not observe in my data[45].

The impact of each of these RNA binding proteins on 3′UTR isoform abundance can be tested in vitro. Each protein can be knocked down in control fibroblasts, and then PAS-seq can be performed on the fibroblasts' mRNA. If isoform changes occur in control fibroblasts, RNA binding protein expression can then be supplemented to determine if they rescue isoform changes. Isoform changes may also be due to aberrant RNA binding protein activity rather than expression. Many proteins are sequestered into HTT aggregates in HD[198]. Sequestration of proteins involved in isoform synthesis or decay into aggregates may alter or ablate their function. Here, I only measured differential expression of RNA binding proteins, not altered activity. To identify RNA binding

proteins that may be sequestered by mutant HTT, mutant HTT could be immunoprecipitated from HD fibroblasts followed by mass spectrometry or co-immunoprecipitation for RNA binding proteins. Knockdown of these RNA binding proteins then be performed as described followed by *HTT* isoform-specific qPCR or PAS-seq. If sequestration of RNA binding proteins into aggregates impairs their function, I would expect knockdown of the RNA binding proteins in control fibroblasts would result in isoform changes similar to those in HD fibroblasts.

RNA binding proteins may interact directly or indirectly with target transcripts to impact isoform abundance. To identify proteins that directly interact with *HTT* and other mRNAs with isoform changes, the candidate RNA binding protein could be immunoprecipitated and attached mRNAs could be submitted to PAS-seq. Alternatively, *HTT* transcripts could be pulled down and attached proteins submitted to mass spectrometry. In addition, genes with isoform changes in HD motor cortex could be queried for common motifs that may be sites for RNA binding proteins.

These experiments will help determine RNA binding proteins that may be responsible for isoform changes in HD. If the causative RNA binding proteins are involved in mRNA stability, I expect isoform alterations arise at steady-state; if the causative RNA binding proteins are involved in mRNA 3′ processing, I expect changes arise during alternative polyadenylation. Whether isoform changes arise during or after transcription and alternative polyadenylation can be assessed by ethinyl-uridine pulse chase in cell models. HD fibroblasts exhibit changes in *HTT* isoform expression compared to controls and may exhibit transcriptome-wide changes. HD or control

fibroblasts can be pulsed with ethinyl-uridine for one hour, during which *HTT* mRNA is transcribed but does not decay, or twelve hours, after which isoform abundances reach steady-state levels (see Figure 2.10B). Cells can then be washed, and labeled mRNA can be collected and *HTT* isoform-specific qPCR or PAS-seq performed. If isoform changes arise during alternative polyadenylation, they will be evident after one-hour incubation with modified uridine. If they arise only at steady state, they will only be present after the twelve-hour incubation.

If isoform changes contribute to HD pathogenesis, RNA binding proteins or other factors involved could be novel therapeutic targets. In addition, if *HTT* isoforms differentially impact HD pathogenesis, specific *HTT* isoforms may be effective targets of *HTT*-lowering therapies.

5.2 *HTT* 3′UTR isoform changes may contribute to HD pathogenesis.

The relative increase in the *HTT* short isoform in HD patient motor cortex may contribute to pathogenesis. During preparation of this dissertation, a study was published by Xu, An, and Xu exploring the impact of *HTT* 3′UTR length on cellular HD pathogenesis[39]. Researchers created constructs expressing mutant *HTT* exon 1 or full-length GFP followed by the short or long *HTT* 3′UTR. They found transfection of the long *HTT* 3′UTR directed GFP expression to the dendrites of cultured rat neurons, whereas transfection of the short 3′UTR restricted expression to the soma. *HTT* exon 1 constructs with the short 3′UTR produced more aggregates in cells, likely due to increased translational efficiency of the short 3′UTR construct compared to the long construct. This is consistent with my finding that the short *HTT* 3′UTR isoform has increased stability and polyA tail length compared to the long isoform, both of which are correlated to translation efficiency[102]. Given these findings, I would expect that the relative increase in the *HTT* short 3′UTR isoform in HD motor cortex results in increased HTT aggregates compared to the cerebellum, where there is a decrease in the *HTT* short 3′UTR isoform.

Further experiments are necessary to determine if *HTT* 3′UTR isoforms have different pathogenicity. The Xu study used artificial, overexpressed isoform constructs out of the normal *HTT* genomic context in embryonic cultured neurons. The constructs did not express full-length *HTT* cDNA, but only *HTT* exon 1, or the *GFP* open reading frame, followed by the short or long *HTT* 3′UTR. These constructs are not spliced or transcribed normally and likely have different secondary structure and associated

proteins, which may impact isoform stability, localization, and translation. Studies on endogenous *HTT* mRNA are ideal to assess isoform pathogenicity. HD or control patient induced pluripotent stem cells with varying CAG numbers could be differentiated into medium spiny neurons[218]. Prior to differentiation, CRISPR/Cas9 technology could be applied to mutate the polyA sites of the short, mid, or long isoform, and clonal populations could be generated that express only one *HTT* 3′UTR isoform. HD medium spiny neurons expressing the short, mid, and long *HTT* 3′UTR isoform could be compared. Ribosome profiling could be used to compare isoform translation rates, ethinyl-uridine pulse chase could be used to confirm isoform stabilities, and RNA fluorescent in-situ hybridization (FISH) could be used to compare isoform localization. To test isoform pathogenicity, cell stresses or BDNF withdrawal could be applied to HD medium spiny neurons expressing each *HTT* isoform, and cell toxicity could be compared to controls[218]. These experiments will compare the metabolism and pathogenicity of each endogenous *HTT* 3′UTR isoform, and will test the hypothesis that the short *HTT* 3′UTR isoform is more pathogenic in neuronal cells.

There is variation in *HTT* 3′UTR isoform abundance across patient brain samples (see Fig. 2.8). I found isoform abundance is not correlated to CAG repeat number ($p>0.5$ for all isoforms). If *HTT* 3′UTR isoforms have different pathogenicity, variations in isoform abundance may contribute to variable age of onset in patients with the same CAG repeat number. I do not know the age of onset for my patient samples. Once the pathogenicity of isoforms has been determined, the contribution of isoform abundance to age of onset can be determined. The abundance of each *HTT* isoform in multiple patients

with similar CAG repeat numbers could be determined by PAS-seq, and the power of isoform abundance to predict age of onset could be determined by regression analysis.

If the short *HTT* isoform is more pathogenic, it will be an effective target for HD RNA and DNA therapies. The polyA site of the short isoform could be mutated using CRISPR/Cas9, leading to decreased production of the pathogenic isoform without inactivating the gene entirely, which could be potentially deleterious given its pro-survival functions. Targeting the *HTT* short isoform mRNA would be more challenging, as it shares its entire sequence with the mid and long isoforms. Interfering RNAs or antisense oligonucleotides could be designed to target the interface between the isoform and the polyA tail. In contrast, if the long isoform is more pathogenic, therapies could target sequences unique to the long 3′UTR. Isoform-specific therapies may enhance targeting of pathogenic *HTT* mRNA transcripts while sparing protective transcripts.

5.3 Changes in the abundance of 3′UTR isoforms of other genes may contribute to HD pathogenesis.

Genes with 3′UTR isoform changes in HD patient brains are implicated in disease pathogenesis. Two proteins that modulate mutant HTT aggregation are among the top five genes with isoform changes in motor cortex from grade 1 HD brains: WD repeat and FYVE domain-containing protein 3 (*WDFY3*), and WW Domain Containing E3 Ubiquitin Protein Ligase 1 (*WWP1*). *WDFY3*, which exhibits the greatest shift towards longer mRNA 3′UTR isoforms, is involved in autophagy of ubiquitinated bodies and promotes turnover of mutant HTT protein[219,220]. In contrast, *WWP1*, which exhibits the third greatest shift towards shorter isoforms, ubiquitinates mutant HTT to prevent its turnover and promote aggregate formation[221]. WDFY3 and WWP1 are likely involved in neuropathology, as HTT aggregation is thought to play a central role in HD pathogenesis[27]. Mutations in another gene that shifts to shorter isoforms in HD motor cortex, pantothenate kinase 2 (*PANK2*), are associated with Hallervorden-Spatz syndrome, an HD-like syndrome characterized by neurodegeneration and severe dystonia, suggesting changes in *PANK2* expression might contribute to neurodegeneration in HD[222].

In the cerebellum, Collagen Type XVI Alpha 1 Chain (*COL16A1*) exhibits the second greatest shift to longer 3′UTR isoforms. *COL16A1* encodes a collagen protein that maintains integrity of the extracellular matrix. In neurons null for the *HTT* gene, *COL16A1* and other extracellular matrix proteins are most down-regulated, suggesting wild-type HTT promotes their expression[223]. The Sodium Voltage-Gated Channel Alpha

Subunit 3 (*SCN3A*) shifts to shorter 3′UTR isoforms in HD cerebellum. Decreased expression of the beta subunits of these voltage-gated sodium channels contributes to excitotoxicity in HD[224]. Ubiquitin-conjugating enzyme E2 J1 (*UBE2J1*), which shifts towards shorter isoforms in HD, catalyzes attachment of ubiquitin to proteins. *UBE2J1* is a target of REST and CoREST, transcriptional repressors with increased activity in HD[140,225].

In motor cortex from grade 2 HD brains, there are three outlier genes with extreme isoform shifts: Nedd4 Family Interacting Protein 1 (*NDFIP1*), Adaptor Related Protein Complex Gamma 1 Subunit (*AP1G1*), and Cathepsin B (*CTSB*). NDFIP1 protein plays a vital role in healing cortical injury, wheras AP1G1 is essential for formation of clathrin-coated vesicles[202,203]. Wild-type HTT interacts with other proteins to regulate clathrin-mediated endocytosis; this process is disrupted in HD[28,226]. CTSB protein was recently found to mediate the beneficial effect of exercising on hippocampal neurogenesis, an effect that is lost in mouse models of HD, suggesting CTSB is impaired in disease[205,206].

Other genes with large isoform shifts in HD patient brains have not been implicated in pathogenesis, but they may still play a role in disease. In motor cortex, Ubiquitin Specific Peptidase 53 (*USP53*), Nuclear Cap Binding Subunit 3 (*NCBP3*), Echinoderm Microtubule Associated Protein Like 1 (*EML1*), NADH:Ubiquinone Oxidoreductase Subunit A2 (*NDUFA2*), Solute Carrier Family 22 Member 23 (*SLC22A23*), and Neurolysin (*NLN*) shift to longer or shorter 3′UTR isoforms in HD. These genes are involved in cytokine signaling (*USP53*), mRNA nuclear export

(*NCBP3*), assembly of microtubules (*EML1*), mitochondrial metabolism (*NDUFA2*), anion transport (*SLC22A23*), and oligopeptide hydrolysis (*NLN*). Several of these pathways, including cytokine signaling, mitochondrial metabolism, and ion transport, are involved in HD pathogenesis[27]. Genes with isoform changes are also enriched for gene ontology pathways involved in HD pathogenesis such as calcium signaling, cytokine signaling, histone acetylation, transcription factor binding, axonal transport, synaptic vesicle localization, dendritic spine morphogenesis, microtubule organizing, and phagocytic vesicles[197–201].

Changes in the steady-state abundance of these 3′UTR isoforms of may contribute to HD pathogenesis. The impact of 3′UTR length on these mRNAs is unknown, and the fold change in isoform expression required for biological changes has not been established. However, as these genes are involved in HD pathogenesis, 3′UTR isoform changes may contribute to altered protein function and toxicity. For instance, WWP1 and WDFY3 modulate HTT protein turnover. If different mRNA 3′UTR isoforms of these proteins are localized differently or have different stability, isoform changes may affect the amount and localization of the resulting proteins. Altered expression or localization of WWP1 and WDFY3 may result in decreased turnover of HTT aggregates and increased pathology. Alternatively, isoform changes may be protective, resulting in increased turnover of HTT. In either scenario, isoform changes may impact HD pathogenesis. Further studies are necessary to determine if isoform changes in HD change the metabolism and translation of mRNAs such as *WWP1* and *WDFY3*.

Cell models can be used to determine the effect of isoform changes on total mRNA metabolism. HD fibroblasts exhibit *HTT* 3′UTR isoform changes, and may also exhibit transcriptome-wide changes like HD patient brains. HD and control fibroblast mRNA can be submitted to PAS-seq to determine if HD fibroblasts exhibit widespread isoform changes compared to controls. HD mouse model primary neurons or striatal cell lines can also be tested. In addition, patient or control fibroblasts differentiated into neural stem cells could be tested[218]. All of these disease cells display a phenotype, although the fibroblast phenotype is subtle[218,227,228].

Once a cell model is identified, mRNA metabolism can be assayed in disease and control cells. Global mRNA stability can be measured via ethinyl uridine pulse chase followed by PAS-seq. Transcript localization can be tested using polyA tail FISH. Isoform translation rates can be determined with ribosome profiling. These experiments will measure differential metabolism of mRNA isoforms of the same gene in normal cells, and identify alterations in mRNA metabolism in disease cells. Because 3′UTR length impacts mRNA localization, stability, and translation, I expect genes with isoform changes will also exhibit changes in metabolism of their mRNAs[132]. Altered mRNA metabolism may impact the abundance and localization of proteins involved in HD pathogenesis, potentially contributing to neurotoxicity.

Once the mechanism of isoform alterations is identified, the impact of isoform changes on HD pathogenesis can be determined. For instance, if decreased CNOT6 levels affect global isoform abundance in HD, CNOT6 can be exogenously supplied in control cells to normalize protein levels. PAS-seq can be performed to confirm exogenous

CNOT6 rescues some or all isoform changes. Cells could then be assayed for HD phenotypes to determine if normalization of isoform abundances had any impact on disease pathology. Results would have to be interpreted with care, since CNOT6 or other RNA binding proteins may reduce HD pathogenesis independently of isoform abundances, even if exogenous expression shifts isoform amounts towards normal.

If 3′UTR isoform changes contribute to HD pathogenesis, RNA binding proteins that impact 3′UTR isoform expression may be novel therapeutic targets. The ideal HD therapy will correct the mutant *HTT* gene or mRNA. However, it is difficult to target mutant *HTT* without targeting the neuroprotective wild-type *HTT*. In addition, gene and RNA therapies using short oligonucleotides have many off-target effects[229]. Down-regulated RNA binding proteins could be rescued by exogenous protein expression. As there are few treatments for HD, any novel therapeutic targets may be beneficial to patients.

5.4 Conclusions: Treating HD.

In conclusion, in this dissertation I detail a novel feature of HD pathology: changes in the abundance of mRNA 3′UTR isoforms, including *HTT*. The mechanism and effect of these changes remains to be established. HD pathogenesis is multifactorial. Isoform changes may be causes or consequences of HD pathology. However, expression of the mutant *HTT* mRNA and protein is central to all pathology in HD, and *HTT*-lowering treatments may alter the course of HD and improve the lives of patients. This dissertation characterizes the abundance, stability, and localization of the predominant *HTT* mRNA isoforms in normal and disease cells. These findings deepen our understanding of HD, and will aid the design of therapeutics.

Many putative treatments aim to reduce expression of *HTT* mRNA or DNA before it can cause toxicity[229]. Antisense oligonucleotides, small interfering RNAs (siRNAs), and artificial microRNAs have been designed to lower *HTT* mRNA levels in pre-clinical studies[229]. Antisense oligonucleotides are single-stranded DNA molecules that anneal to *HTT* mRNA. The RNA-DNA duplex is cleaved by RNAse H. siRNAs are double-stranded RNAs. The strand complementary to the *HTT* mRNA is loaded into the RNA induced silencing complex (RISC), which cleaves target *HTT* mRNA. Artificial microRNAs are designed with backbones and structure identical to an endogenous microRNA, but the region that gets loaded into RISC anneals to *HTT* mRNA, resulting in cleavage or translation suppression. Enzymes that cut DNA including Zinc finger nucleases and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR/Cas9) have been used to therapeutically target the *HTT* gene[230,231]. All of these

therapies have to be modified or packaged into viruses or lipid vesicles for delivery to the brain[229]. Delivery is difficult, as direct injection in the brain via neurosurgery is invasive and could potentially cause infection, whereas intrathecal delivery saturates the brain surface while missing deeper structures such as the striatum, and peripheral delivery is largely cleared by the liver and may not cross the blood brain barrier[229].

An additional challenge of all *HTT*-lowering strategies is allele specificity. Wild-type HTT performs many important functions in cells, and reduction below 50% of normal levels is deleterious[26]. The ideal HD therapy would target mutant but not wild-type *HTT*. However, mutant and wild-type *HTT* DNA and mRNA are identical except for the CAG repeat length and heterozygous SNPs. Putative *HTT* therapies aim to target the CAG expansion, or SNPs in the mutant but not wild-type allele[46,47,232,233].

As described, if one mutant *HTT* 3′UTR isoform is more pathogenic, it could be targeted at the DNA or RNA level, whereas if one isoform is protective, it could be supplemented. I show *HTT* 3′UTR isoforms have different localizations. *HTT* mRNA-lowering therapies could be directed to cellular compartments enriched for *HTT* isoforms. The region unique to the long mRNA isoform the ideal target for RNA-lowering therapies, as I show there is a decrease in the long *HTT* 3′UTR isoform in patient motor cortex. Thus, my findings will aid in design of therapies targeting *HTT* mRNA and DNA.

HD is a complicated disease with changes in many cellular processes: epigenetics, transcription, splicing, translation, metabolism, signaling, vesicle transport, and survival[198]. Here, I describe another cell process deregulated in HD: 3′UTR isoform abundance. How isoform changes fit into the complex landscape of HD pathogenesis

remains to be established. However, *HTT* mRNA and protein are at the core of all changes in HD. Many *HTT*-lowering therapies are nearing clinical trials. As there is currently no disease-altering treatment for HD, these therapies hold great promise for the families devastated by HD. The findings described in this dissertation reveal a new aspect of HD pathology and provide some recommendations for HD therapy. It is my hope that soon there will be a disease-altering treatment for this terrible disease.

APPENDIX I: LINKING SNPs TO THE CAG REPEAT EXTENDED RANGE ("SLIC-er")

Preface

This work was a collaborative effort. Jonathon Watts designed the modified oligonucleotides. Rachael Miller and I performed the screens.

A.1 Summary

Mutant HTT protein causes pathology at the DNA, RNA, and protein level and disrupts the function of wild-type HTT. Wild-type HTT protein interacts with several other proteins to play a role in neuronal survival, synaptic activity, and vesicle transport. Putative HD therapies aim to decrease *HTT* mRNA before it is translated into the pathogenic protein[229]. However, it may be detrimental to decrease mRNA levels from both *HTT* alleles, given the important functions of wild-type HTT. Thus, several therapeutic approaches aim to target mutant but not wild-type *HTT* mRNA.

Antisense oligonucleotides, small interfering RNAs (siRNAs), and artificial microRNAs can lower *HTT* mRNA levels in vitro and in vivo[229]. Making these therapies allele-specific is challenging. The mutant and wild-type *HTT* mRNAs are nearly identical except for the CAG repeat expansion; designing antisense oligos or siRNAs targeting the CAG repeat is virtually impossible given the short length of these therapeutics. One strategy for allele-specific silencing is via microRNA-like RNA duplexes targeting the CAG repeat, which more effectively silence a target when more duplexes are bound. Since there are more CAG repeats on the mutant allele, these duplexes cause more silencing of the mutant than wild-type allele[233]. However, there is still substantial silencing of the wild-type allele using this method. Another strategy targets siRNAs to the Δ2642 deletion in *HTT* exon 58, which is linked to the mutant allele[232]. This treatment will only work on the subset of patients that have this mutation on the mutant but not wild-type allele. A third strategy for allele-specific silencing is targeting frequently-heterozygous SNPs in *HTT* with siRNAs or antisense oligonucleotides, which can be

engineered to effectively silence the mutant but not wild-type allele in vitro[46,47]. To be effective, these therapies require knowledge of which SNP corresponds to the mutant allele.

Our lab has published a method for **S**NP-**li**nkage to the **C**AG repeat in HD patients (SLiC)[234]. In the method, *HTT* mRNA is reverse transcribed, and a PCR is performed from exon 1 to the SNP using primers with restriction enzyme sites. The PCR products are cleave with restriction enzymes and then ligated into circular cDNA. An inverse PCR is then performed spanning the SNP and exon 1, and the products are sequenced to determine the linkage of the SNP to the CAG repeat[234]. This method is effective for SNPs proximal to exon1; however its efficacy drops for more distal SNPs due to decreased efficiency of the long-range PCR. The ideal SNP-linkage method would be effective regardless of the SNP position, as several frequently heterozygous SNPs are located in the HTT 3′UTR, up to 12kb away from exon 1[46].

I developed a method for **S**NP-**li**nkage to the **C**AG repeat at **e**xtended **r**anges, SLiC-er (Fig. A.1). Total RNA is treated with RNAsH and a SNP-specific DNA oligonucleotide, resulting in cleavage of one allele but not the other. *HTT* mRNA is reverse transcribed with an oligo d(T) primer, and exon 1 PCR is performed on full-length cDNA and resolved via agaorse gel electrophoresis. Little full-length mRNA is transcribed from the allele matching the oligonucleotide, as most of it is cleaved during the RNAse H step. Thus, the oligo-matching allele is faint or absent on the gel, enabling linkage of the SNP to the allele.

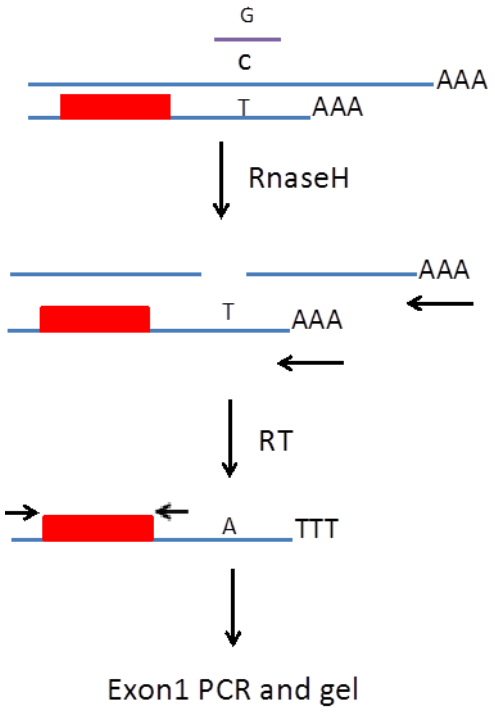


Figure A.1. Schematic of SNP-linkage to CAG-repeat extended range (SLIC-er) method. The DNA oligo is shown in purple; red represents the expanded CAG repeat; arrows represent primers.

A.2 Results

We designed oligonucleotides complementary to the region surrounding five SNPs frequently heterozygous in HD patients: rs363099 in exon 29, rs362336 in exon 48, rs362331 in exon 50, rs362273 in exon 57, and rs362306 in the 3′UTR. For each SNP, we designed three unmodified DNA oligonucleotides with the nucleotide complementary to the SNP at or around the center position (Table A.1). Total RNA from Yac128 mice was DNAse treated and subjected to incubation with RNAse H and each candidate oligonucleotide or an off-target oligonucleotide. RNAse H degraded the RNA strand in the DNA-RNA hybrids. Treated RNA was then reverse transcribed from an oligo(dT) primer with conditions optimized to produce full-length HTT cDNA (see methods). Transcripts reverse transcribed from mRNA cleaved by RNAse H terminate at the cleavage site and do not reach HTT exon 1. To assay the efficacy of RNAse H cleavage, the CAG repeat in exon 1 or a region distal to the SNP in the 3′UTR was PCR amplified. Products were resolved on a 2% agarose gel. There was reduced PCR product for all three oligonucleotides for each SNP compared to the off-target control and the distal 3′UTR (Fig. A.2). SNP rs362306A/G located in the HTT 3′UTR was most efficiently silenced.

Table A.1 Unmodified DNA oligonucleotides.

rs363099 (T/C) Exon 29

1 5′- CTGAGCGAAGAAACC -3′
2 5′- TGAGCGAAGAAACCC -3′
3 5′- GCTGAGCGAAGAAAC -3′

rs362336 (A/G) Exon 48

1 5′- CCGCCGGTTCTGCAG -3′
2 5′- CGCCGGTTCTGCAGG -3′
3 5′- GCCGCCGGTTCTGCA -3′

rs362331 (C/T) Exon 50

1 5′- CACAGTGGATGAGGG -3′
2 5′- ACAGTGGATGAGGGA -3′
3 5′- ACACAGTGGATGAGG -3′

rs362273 (G/A) Exon 57

1 5′- TGATCTGCAGCAGCA -3′
2 5′- GATCTGCAGCAGCAG -3′
3 5′- TTGATCTGCAGCAGC -3′

rs362306 (A/G) 3' UTR

1 5′- GCAGCTGTAACCTGG -3′
2 5′- CAGCTGTAACCTGGC -3′
3 5′- AGCAGCTGTAACCTG -3′

Unmodified DNA oligonucleotides. The SNP is shown in red. Oligonucleotides were designed to match the Yac but not Bac transgene.

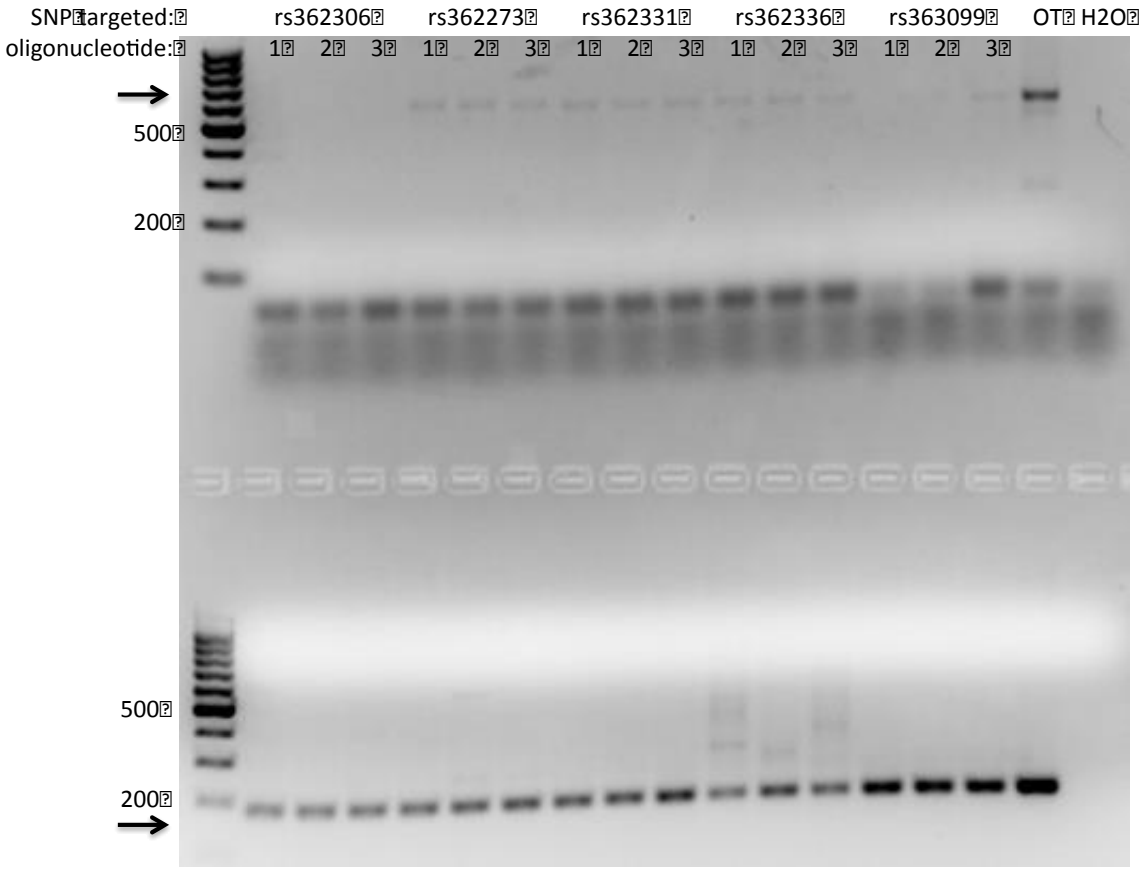


Figure A.2. Screen of RNAse H cleavage of five frequently-heterozygous SNPs via unmodified DNA oligonucleotides. RNA was extracted from Yac128 mouse tissue, treated with RNAseH and a matching or off-target DNA oligonucleotide, and reverse transcribed. PCR of HTT exon 1 (top) or 3′UTR distal to the SNP (bottom) was performed on the cDNA. OT signifies off-target (oligonucleotide complementary to human beta-actin).

We designed two sets of modified oligonucleotides to discriminate between alleles heterozygous at rs362306. Our designs were based on studies that found oligonucleotides with central DNA regions surrounded by modified RNA bases are best at discriminating between *HTT* alleles in cells[47]. The first set of modified oligonucleotides we designed contained a 4-base DNA gap flanked on each end with

2′-O-methyl RNA, with the SNP site in two different positions (Table A.2). The second set of modified oligonucleotides contained a 6-base DNA gap flanked on each end with 2′-O-methyl, with the SNP site in varying positions as well as an additional mismatch. Studies have shown adding an additional mismatch to siRNAs heightens alelle specificity[46]. To test allele discrimination by our oligonucleotides, we treated total RNA from Bac21 or Yac128 mice as described. The oligonucleotides were complementary to the Yac128 but mismatched to the Bac21 rs362306. We found all oligonucleotides discriminated between the Yac and Bac HTT alleles (Fig. A.3). Oligonucleotides similar to these could be used to link the SNP to the CAG repeat in patients prior to allele-specific therapy.

Table A.2. Modified oligonucleotides targeting rs362306.

1	5′- GCAGCTGTAACCUG-3′
2	5′- AGCUGTAACCUGGC-3′
3	5′- GCAGCAGTAACCUG-3′
4	5′- AGCUGTATCCUGGC-3′
5	5′- AGCUGTAAACUGGC-3′
6	5′- AGCUGTATCCUGGC-3′

Black, red, or green bases are unmodified DNA; blue are 2′-O-methyl RNA. Red indicates the SNP. Green bases are an additional mismatch. Oligonucleotides were designed to match the Yac but not Bac transgene.

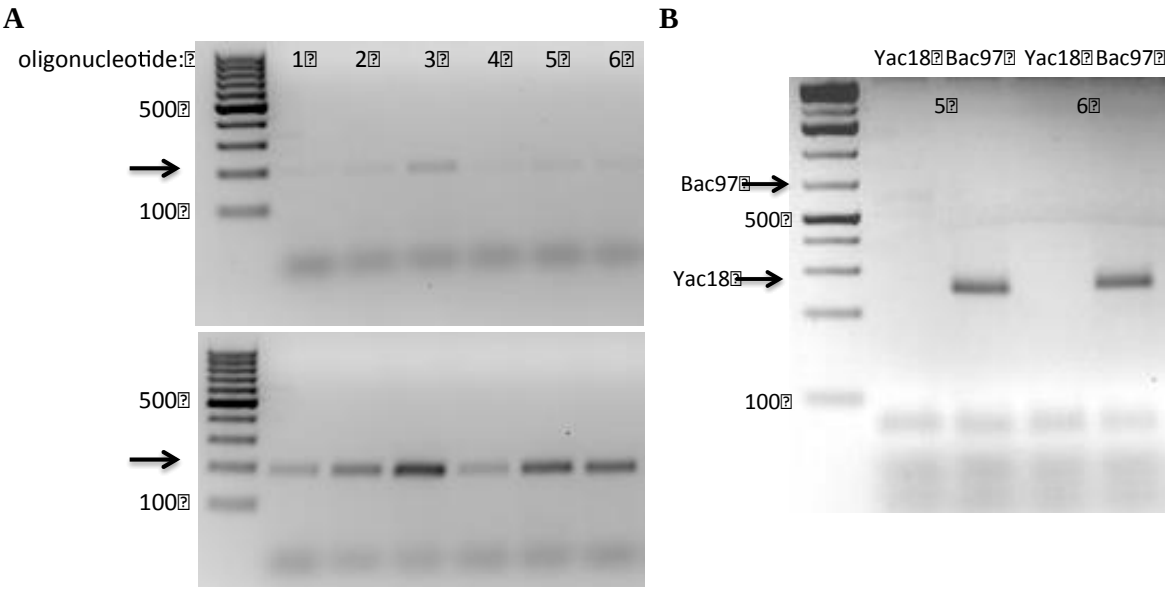


Figure A.3. Screen of SNP rs362306 discrimination by modified oligonucleotides and RNAseH. (A) RNA was extracted from Yac18 or Bac97 mouse tissue, treated with RNAseH and a modified oligonucleotide matching Yac but not Bac rs362306, and reverse transcribed. PCR of HTT exon67 was performed on the cDNA. (B) PCR of HTT exon 1 was performed on cDNA of RNA treated with modified oligonucleotide 5 or 6 (see table A.2).

A.3 Discussion

Our improved method to link a heterozygous SNP to the CAG repeat allows robust linkage of SNPs as far as the 3′UTR. Several further studies are necessary to validate the method. We have shown allele discrimination by oligonucleotides on RNA from homozygous mice. However, the method will be used clinically on heterozygous RNA from HD patients. The next step is to test the method on RNA from Yac/Bac mice, which are heterozygous for rs362306. The allele-SNP linkage is known for these mice, making them ideal for SLIC-er validation. Once the method is working well on RNA from Yac/Bac mice, it must be tested on RNA from heterozygous patient cells. Efficacy should also be shown for other SNPs in addition to rs362306. Once this method is optimized and validated, it will be easier and more reliable than existing methods and can be used clinically to direct therapeutics to the mutant but not wild-type allele.

A.4 Materials and methods

A.4.1 RNAse H treatment. Total RNA was extracted from mouse tissues using Trizol (Ambion), and RNA was treated with TurboDNAse (Ambion). One microgram of RNA was heated at 95° for 2 minutes with 1μg of the complementary oligonucleotide (IDT) in 1x RNAseH buffer (NEB). The reaction was slowly cooled to room temperature, and 4 units of E. coli derived RNase H enzyme (NEB) and RNase OUT (Invitrogen) were added. The reaction was incubated at 37° for 4 hours, then 65° for 20 minutes.

A.4.2 Reverse transcription. Treated RNA was reverse transcribed from an oligo(dT) primer using SuperScript IV (Invitrogen) per manufacturer protocol with the following modifications: 2M betaine and 0.6M trehalose were added to optimize production of full-length HTT, and the reaction was incubated at 42° for 2.5 hours[235]. Reactions were cleaned on a DNA clean and concentrate column (Zymo).

A.4.3 *HTT* exon 67 PCR. At least 20ng cDNA was PCR amplified by GoTaq Green Master Mix (Promega) with 2μM forward and reverse primers 5′-GGAGCCTCTCCTGCTTCTTT-3′ and 5′-CACCTCAAGCACAGACTGGA-3′. The reaction was incubated at 94°C for 5 minutes; 30 cycles of 94°C for 30 seconds, 60°C for 30 seconds, and 72°C for 30 seconds; and 72°C for 5 minutes. Products were resolved on a 2% agarose, 1x TAE gel.

A.4.4 *HTT* 3′UTR PCR. At least 20ng cDNA was PCR amplified by Phusion Master Mix (ThermoFisher Scientific) with 2μM forward and reverse primers 5′-ATGGATGCATGCCCTAAGAG-3′ and 5′-CAGTCTCCGATGAGCACAGA-3′. The reaction was incubated for 35 cycles of 98°C for 30 seconds, 64°C for 20 seconds, and 72°C for 20 seconds followed by 72°C for 5 minutes. Products were resolved on a 2% agarose, 1x TAE gel.

A.4.5 *HTT* Exon 1 PCR. At least 100ng cDNA was PCR amplified by PrimeStar GXL polymerase (Clontech) with 1μM forward and reverse primers 5′-GCCTCCGGGGACT GCCGTGC-3′ and 5′-CGGCTGAGGCAGCAGCGGCT-3′. The reaction was incubated at 98°C for 1 minute; 30 cycles of 98°C for 10 seconds and 68°C for 30 seconds; and 68°C for 10 minutes. Products were resolved on a 2% agarose, 1x TAE gel.

APPENDIX II: PAS-SEQ ANALYSIS CODE

```
#Purple text corresponds to instructions for input to Unix terminal
#Red text corresponds to names of shell files or names of R/python files run within shells
#Green text corresponds to commenting within shell or R/python files

#first, gunzip and combine all fastq into one file called read1.fastq (cat *.fastq > read1.fastq)
#then, run: bsub -q long -n 24 -W 8:00 -R "rusage[mem=120000]" -o output.log -e error.log ./index_reads.sh

#index_reads.sh:
#!/bin/bash
python index_sort.py

#files for index_reads.sh:
#index_sort.py:

#input fastq contains all R1 reads
inputfq=open('read1.fastq')
inputfq=inputfq.readlines()
#define indices (samples)
indices=['GTGAT','CATCG','AGCTA','GGAGC','ACTGT', 'ATCTG',
        'CGGGA','TAGCC','TGACT','ACAAG','ATTCA','TCTAC']
#loop over reads and search for index, write line to output
for index in indices:
        outputfq=open(str(index)+'.fastq','w')
        for row in range(1,len(inputfq),4):
                if index in inputfq[row][0:5]:
                        outputfq.write(inputfq[row-1])
                        outputfq.write(inputfq[row])
                        outputfq.write(inputfq[row+1])
                        outputfq.write(inputfq[row+2])

#move all indexed files to their own folders, name by sample so 'sample_index.fastq'
#in folders, run bsub -q long -n 24 -W 4:00 -R "rusage[mem=12000]" -o output.log -e error.log ../preprocess_shell.sh index

#preprocess_shell.sh
#!/bin/bash

#need clean directory with this sh file, fastq file, bowtie pre-indexed hg19 genome for alignment (or symbolic link to the files), cleanUpdTSeq.R, cleanUpdTSeq.sh
#input ($1)='filename' from filename.fastq
```

```
module load cutadapt/1.9
module load openssl/1.0.1g

#remove barcode+randomer and polyA from reads
cutadapt -u 12 $1.fastq > noadapt.fastq
cutadapt -a AAAAAAA noadapt > pAtrimmed.fastq

#trim low quality bases, throw out reads <15nt after trimming
cutadapt –q 25 –m 15 pAtrimmed.fastq > preprocessed.fastq

#check quality of remaining reads
module load fastqc/0.10.1
fastqc preprocessed.fastq

#run bsub -q long -n 24 -W 8:00 -R "rusage[mem=50000]" -o output.log -e error.log
../PAS-Seq_all.sh
sample_name

#PAS-Seq_all.sh
#!/bin/bash
#align reads to genome, first download pre-indexed hg19 for bowtie (find online), need
human UTR
#coordinates from UCSC table browser in file named "hg19_3primeUTRcoordinates(+/-
).bed" with no
#header and gene coordinates in “hg19genes(+/-).bed” with no header

#$1 input is sample name
module load bowtie/1.0.0
bowtie -S hg19 preprocessed.fastq $1_aligned.sam

#convert SAM to BAM
module load samtools/0.0.19
samtools view -S -b $1_aligned.sam > $1_aligned.bam

#sort and index BAM file for viewing with IGV
samtools sort $1_aligned.bam $1_aligned_sorted
samtools index $1_aligned_sorted.bam $1_aligned_sorted.bai

#convert BAM to BED file to identify peaks and falsely primed reads
module load bedtools/2.25.0
bamToBed -i $1_aligned_sorted.bam > $1_aligned_sorted.bed

#compute read 3′ end NT coverage for + and – strand reads
module load ucsc_cmdline_util/12_16_2013
```

```
fetchChromSizes hg19 > hg19.chrom.sizes
genomeCoverageBed -strand + -bg -3 -i $1_aligned_sorted.bed -g hg19.chrom.sizes >
$1_coverage+.bed
genomeCoverageBed -strand - -bg -3 -i $1_aligned_sorted.bed -g hg19.chrom.sizes >
$1_coverage-.bed

#call and collapse reads onto peaks
module load R/3.0.2
Rscript ../peak_call_pw.R $1_coverage+.bed $1_peaks+.bed
Rscript ../peak_call_pw.R $1_coverage-.bed $1_peaks-.bed

#find gene names 3'UTR peaks are in
intersectBed -a $1_peaks+.bed -b ../hg19_3primeUTRcoordinates+.bed -wa -wb >
$1_UTRpeaks+.bed
intersectBed -a $1_peaks-.bed -b ../hg19_3primeUTRcoordinates-.bed -wa -wb >
$1_UTRpeaks-.bed

#find gene names all peaks are in
intersectBed -a $1_peaks+.bed -b ../hg19genes+.bed -wa -wb > $1_peaks_genes+.bed
intersectBed -a $1_peaks-.bed -b ../hg19genes-.bed -wa -wb > $1_peaks_genes-.bed

#files for PAS-Seq_all.sh
#peak_call_pw.R

#input 3′ end coverage file, define output filename
args=commandArgs(T)
input_coverage=args[1]
output_peaks=args[2]

peaks_all<-read.table(input_coverage)
colnames(peaks_all)=c("chr","start","stop","coverage")

#split reads by chromosome
peaks_split = split(peaks_all, peaks_all$chr)
all_data<-list()

#loop over chromosomes to call peaks
for (m in 1:length(peaks_split)) {
        #make dataframe of reads from the chromosome, count number of reads in the
dataframe
        peaks<-as.data.frame(peaks_split[m])
        colnames(peaks)=c("chr","start","stop","coverage")
        n_reads<-nrow(peaks)
        #make the weights for a histogram with each peak position weighted by its
```

```
coverage
        hist_weight<-rep(peaks[1,3],peaks[1,4])
        for (n in 2:nrow(peaks)) {
                weighted1<-rep(peaks[n,3],peaks[n,4])
                hist_weight<-c(hist_weight,weighted1)
        }
        #make the histogram with bins 48 nucleotides wide
        range<-seq(from = (peaks[1,3]-48), to=(peaks[n_reads,3]+48), by=48)
        histo_data<-hist(hist_weight,breaks=range)
        bin_counts<-as.matrix(histo_data$counts)
        breaks<-as.matrix(histo_data$breaks)
        bins_breaks<-as.data.frame(cbind(breaks[1:nrow(breaks)-
1],breaks[2:nrow(breaks)],bin_counts))
        colnames(bins_breaks)=c("bin_start", "bin_end","count")
        bins_breaks=bins_breaks[bins_breaks$count>4,]
        nbins<-nrow(bins_breaks)
        #find maximum coordinate of peaks in bins with >4 reads; look upstream and
        downstream 24nt to offset histogram bins. Change coverage at that position to the
        sum of all surrounding area (24nt up and downstream)
        window_peaks<-matrix(ncol=4)
        colnames(window_peaks)=colnames(peaks)
        for (k in 1:nrow(bins_breaks)){
                above<-peaks[peaks$stop>(bins_breaks[k,1]-24),]
                window<-above[above$stop<=(bins_breaks[k,2]+24),]
                max_position<-window[which.max(window[,4]),]
                t<-peaks[peaks$stop>as.numeric((max_position[3]-24)),]
                v<-t[t$stop<as.numeric((max_position[3]+24)),]
                max_position[4]=sum(v[,4])
                window_peaks=rbind(window_peaks,max_position)
        }
        window_peaks=window_peaks[2:nrow(window_peaks),]
        window_peaks=unique(window_peaks)
        #remove any peaks that are on the border of bins
        for (n in 1:(nrow(window_peaks)-1)) {
                if (!is.na(window_peaks[n,3]) & !is.na(window_peaks[n+1,3])&
as.numeric(window_peaks[n+1,3])-as.numeric(window_peaks[n,3])<25) {
                        mat<-rbind(window_peaks[n,],window_peaks[n+1,])
                        index_remove<-which.min(mat[,4])
                        index<-which(window_peaks[,3]==mat[index_remove,3])
                        window_peaks[index,3]<-NA
                }
        }
        window_peaks=na.omit(window_peaks)
        all_data[[m]]=window_peaks
```

```
}

all_cov=all_data[[1]]
for (t in 2:length(all_data)){
        all_cov=rbind(all_cov,all_data[[t]])
}
all_cov=all_cov[order(as.numeric(row.names(all_cov))),]

write.table(all_cov, file=output_peaks, sep='\t', row.names=FALSE, col.names=FALSE,
quote=FALSE)

#clean up + and – strand 3′UTR and total gene peaks with cleanUpdTSeq to remove false
polyA sites
#run bsub -q long -n 24 -W 18:00 -R "rusage[mem=120000]" -o output.log -e error.log
../cleanUpdTSeq.sh sample_name

#cleanUpdTSeq.sh
#!/bin/bash
module load R/3.2.0
Rscript ../cleanUpdTSeq.R $1_UTRpeaks+.bed $1_UTRpeaks-.bed
$1_UTRpredictions.tsv $1_UTRpredictions.bed
Rscript ../cleanUpdTSeq.R $1_peaks_genes+.bed $1_peaks_genes-.bed
$1_all_predictions.tsv $1_all_predictions.bed

#files for cleanUpdTSeq.sh
#cleanUpdTSeq.R
#takes 4 command line arguments: input+ peaks, input- peaks, outputfilename (tsv file of
+ and – strand true peaks), outputfilename (bed file of + and - strand true peaks)

#load required modules
#source("http://bioconductor.org/biocLite.R")
#biocLite("cleanUpdTSeq")
library(cleanUpdTSeq)
library(BSgenome.Hsapiens.UCSC.hg19)

#read in data, convert to GRanges (peaks)
args=commandArgs(T)
input_plus_peaks=args[1]
input_minus_peaks=args[2]
output_filename=args[3]
bed_peaks=args[4]

testSet_plus <- read.table(input_plus_peaks, header=FALSE)
testSet_plus = data.frame(testSet_plus[,1:4],testSet_plus[,8])
```

```
colnames(testSet_plus)=c('chr','start','stop','score','geneID')
testSet_minus <- read.table(input_minus_peaks, header=FALSE)
testSet_minus = data.frame(testSet_minus[,1:4],testSet_minus[,8])
colnames(testSet_minus)=c('chr','start','stop','score','geneID')

testSetall<-rbind(testSet_plus, testSet_minus)
peak_names<-cbind(as.matrix(1:nrow(testSetall)),as.matrix(testSetall[,5]))

#make testSet 6 columns with strand info in 6th column. column5=coverage
strand_plus<-rep('+',nrow(testSet_plus))
strand_minus<-rep('-',nrow(testSet_minus))
strand<-c(strand_plus,strand_minus)
testSet<-cbind.data.frame(testSetall[,1:3],peak_names[,1],testSetall[,4],strand)
colnames(testSet)=c('chr','start','stop','name','score','strand')
peaks <- BED2GRangesSeq(testSet, withSeq=FALSE)

#build feature vectors
testSet.NaiveBayes <- buildFeatureVector(peaks, BSgenomeName=Hsapiens,
upstream=40, downstream=30, wordSize=6, alphabet=c("ACGT"),
sampleType="unknown", replaceNAdistance=30, method="NaiveBayes",
ZeroBasedIndex=1, fetchSeq=TRUE)

#test against datasets (3PSeq)
data(data.NaiveBayes)
predictTestSet(data.NaiveBayes$Negative, data.NaiveBayes$Positive,
        testSet.NaiveBayes=testSet.NaiveBayes,outputFile=
        output_filename, assignmentCutoff=0.95)

#find gene names of peaks, remove internally primed peaks
predicted<-read.table(output_filename, header=TRUE)
predicted_peaks<-predicted[predicted[,4]>0,]
rows_left<-predicted_peaks[,1]

#make file of true peaks with gene name and coverage information
testSet[,4]=peak_names[,2]
testSet=testSet[rows_left,]
write.table(testSet, file=bed_peaks, sep='\t', row.names=FALSE, col.names=FALSE,
quote=FALSE)

#Run bsub -q long -n 24 -W 24:00 -R "rusage[mem=16000]" -o output.log -e error.log
../ratios_peaks.sh sample_name

#ratios_peaks.sh
#these programs will normalize the peak read to total reads from the 3′UTR and identify
```

genes with more than one 3′UTR peak. It will also add up all the peaks from the same gene to get the total gene coverage.

```
#!/bin/bash
#load stuff required
module load R/3.1.0

Rscript ../all_pA_ratios.R $1_UTRpredictions.bed $1_mpApeaks.bed $1_mpAratios.bed
Rscript ../all_gene_coverages.R $1_UTRpredictions.bed $1_all_predictions.bed
$1_combined_peaks.bed $1_combined_coverage.bed

#files for ratios_peaks.sh
#all_pA_ratios.R
#takes 4 arguments from command line: 1. input bed file of true peaks in the UTR
(output from cleanUpdTSeq script) with + and - strand genes 2. output name for file of
genes/coverage/strand/coordinate information for genes with multiple pA peaks 3. output
name for file of isoform:total coverage of tandem pA peak genes with gene name, pA
site, and coverage.

args=commandArgs(T)
input_UTR_peaks=args[1]
output_mpA_peaks=args[2]
output_mpA_ratios=args[3]

all_real_peaks <- read.table(input_UTR_peaks, header=FALSE,
col.names=c('chr','start','stop','name','coverage','strand'), stringsAsFactors=FALSE)

#remove peaks that map to more than one gene
p<-all_real_peaks[,1:3]
q<-p[!(duplicated(p) | duplicated(p, fromLast = TRUE)),]
all_real_peaks<-all_real_peaks[rownames(q),]

#make first rows to add to in the loop, I will delete these later. mpA_peaks contains info
for all genes with multiple 3'UTR peaks.
#Ratio contains name and ratio of isoform: total 3'UTR coverage for each peak from a
gene with multiple 3'UTR peaks.
mpA_peaks<-all_real_peaks[1,]
these_peaks<-all_real_peaks[1,]
peak_norm<-these_peaks[,5]/sum(these_peaks[,5])
ratio<-cbind(these_peaks[,4],peak_norm,these_peaks[,5],these_peaks[,1:3])
colnames(ratio)=c('geneID','peak/total','coverage','chr','start','stop')
indices<-0

for (n in 1:(nrow(all_real_peaks))){
```

```
        #only analyze this gene if it hasn't been analyzed yet
        if (length(which(indices==n))==0){
                #find number of 3′UTR peaks that are in this gene
                number_peaks<-
        t(as.matrix(which(all_real_peaks[,4]==all_real_peaks[n,4])))
                these_peaks<-all_real_peaks[number_peaks,]
                #if the number of peaks is greater than one, add peaks from that gene to
                file of mpA peaks, divide peak coverage by the total reads mapping to the
                gene 3′UTR and add peaks to ratio file
                if (length(number_peaks)>1) {
                        mpA_peaks=rbind(mpA_peaks,these_peaks)
                        peak_norm<-these_peaks[,5]/sum(these_peaks[,5])
                        this_gene<-
cbind(these_peaks[,4],peak_norm,these_peaks[,5],these_peaks[,1:3])
                        colnames(this_gene)<-colnames(ratio)
                        ratio=rbind(ratio,this_gene)
                        rownames(ratio)<-NULL
                        indices=data.frame(indices, number_peaks)
                 }
        }
}

#remove the first dummy rows
mpA_peaks=mpA_peaks[2:nrow(mpA_peaks),]
colnames(mpA_peaks)=c('chr','start','stop','geneID','coverage','strand')

ratio=as.data.frame(ratio[2:nrow(ratio),])
ratio <- as.data.frame(lapply(ratio, unlist))

#write outputs
write.table(mpA_peaks, file=output_mpA_peaks, sep='\t', row.names=FALSE,
col.names=FALSE, quote=FALSE)
write.table(ratio, file=output_mpA_ratios, sep='\t', row.names=FALSE,
col.names=FALSE, quote=FALSE)

#all_gene_coverages.R
#takes 2 arguments from command line: 1. input bed file of true peaks (output from
cleanupdt seq script) with + and - strand genes 2. output name for file of
genes/coverage/strand/coordinate/UTRreads combined information.
#this script will remove duplicate gene names or peak coordinates made by intersectbed
in all gene file.

#Define inputs and outputs. Output file has name, total reads for gene, strand, and UTR
reads for gene.
```

```
args=commandArgs(T)
input_UTRpeaks=args[1]
input_allpeaks=args[2]
output_combined_coverage=args[3]

UTR_peaks <- read.table(input_UTRpeaks, header=FALSE,
col.names=c('chr','start','stop','geneID','coverage','strand'))

all_peaks <- read.table(input_allpeaks, header=FALSE,
col.names=c('chr','start','stop','geneID','coverage','strand'))

#remove peaks that fit in more than one gene
p<-UTR_peaks[,1:3]
q<-p[!(duplicated(p) | duplicated(p, fromLast = TRUE)),]
UTR_peaks<-UTR_peaks[rownames(q),]

p<-all_peaks[,1:3]
q<-p[!(duplicated(p) | duplicated(p, fromLast = TRUE)),]
all_peaks<-all_peaks[rownames(q),]

#combine all the peaks from one gene
indices<-0
combined_peaks<-data.frame(matrix(nrow=nrow(all_peaks),ncol=4))
for (j in 1:nrow(all_peaks)) {
        #only analyze this gene if it hasn't been analyzed before
        if (length(which(indices==j))==0){
                #Continue summing reads until they are no longer in the same gene.
                sum=all_peaks[j,5]
                if (j<nrow(all_peaks)){
                        start<-j+1
                        end<-nrow(all_peaks)
                        for (m in start:end){
                                if (all_peaks[j,4]!=all_peaks[m,4]) {
                                        break
                                }
                                else {
                                        indices=data.frame(indices,j)
                                        indices=data.frame(indices,m)
                                        sum=sum+all_peaks[m,5]
                                }
                        }
                }
                #Add all the UTR reads that map to the gene
```

```
                UTR_reads<-which(UTR_peaks[,4]==as.character(all_peaks[j,4]))
                if (length(UTR_reads)>0) {
                        reads<-sum(UTR_peaks[UTR_reads,5])
                }
                else {
                        reads<-0
                }
                combined_peaks[j,1]=toString(all_peaks[j,4])
                combined_peaks[j,2]=sum
                combined_peaks[j,3]=toString(all_peaks[j,6])
                combined_peaks[j,4]=reads
        }
}

combined_peaks=na.omit(combined_peaks)


#write outputs
write.table(combined_peaks, file=output_combined_coverage, sep='\t',
row.names=FALSE, col.names=FALSE, quote=FALSE)

#once you have done all of the above on all samples and moved all *mpAratios* and
*_combined_peaks* to their own folder:
#Run bsub -q long -n 24 -W 18:00 -R "rusage[mem=16000]" -o output.log -e error.log
../combine_samples.sh tissue_name

#combine_samples.sh
#this program combines the gene peak info (isoform ratios or total coverage) for all
samples.
#!/bin/bash
#input is the tissue name for all of the samples (ie cortex)

#load stuff required
module load R/3.1.0

Rscript ../combine_samples_newest.R $1_all_ratios.tsv $1_all_gene_coverage.tsv
$1_fracUTRgenecoverage.tsv

#files for combine_samples.sh
#combine_samples_newest.R

#define outputs, which include the output ratios filename and the output coverage
filename.
args=commandArgs(T)
```

```
output_ratios=args[1]
output_gene_coverage=args[2]
output_fracUTR=args[3]

#import relevant files in the working directory
filenames_ratios<-list.files(getwd(),pattern='*mpAratios*')
filenames_gene_coverage<-list.files(getwd(),pattern='*combined_peaks*')

just_name_filenames<-substr(filenames_ratios,1,8)
list_of_ratios<-lapply(filenames_ratios, read.table)
list_of_gene_cov<-lapply(filenames_gene_coverage, read.table)

names(list_of_ratios) <- filenames_ratios
names(list_of_gene_cov) <- filenames_gene_coverage


#make a list of all unique isoforms represented in the sample datasets for isoform
coverage
UTR_iso_names<-cbind(list_of_ratios[[1]][,1],list_of_ratios[[1]][,5:6])
colnames(UTR_iso_names)<-c('gene','start','stop')
for (s in 2:length(list_of_ratios)){
        aname<-cbind(list_of_ratios[[s]][,1],list_of_ratios[[s]][,5:6])
        colnames(aname)<-c('gene','start','stop')
        UTR_iso_names=rbind(UTR_iso_names,aname)
}
UTR_iso_names=unique(UTR_iso_names)

#make a list of all unique genes represented in the sample datasets for total coverage
gene_names<-as.matrix(list_of_gene_cov[[1]][,1])
for (k in 2:length(list_of_gene_cov)){
        gene_names=rbind(gene_names,as.matrix(list_of_gene_cov[[k]][,1]))
}
gene_names=unique(gene_names)

#make an output dataframe with the first column for gene name and the other columns
for isoform ratios or gene coverage for that isoform/gene in each sample
all_samples_ratios<-
as.data.frame(matrix(ncol=(length(list_of_ratios)+3),nrow=nrow(UTR_iso_names)))
all_samples_gene_coverage<-as.data.frame(matrix(ncol=(length(list_of_gene_cov)+1),
nrow=nrow(gene_names)))
all_samples_fracUTR<-as.data.frame(matrix(ncol=(length(list_of_gene_cov)+1),
nrow=nrow(gene_names)))

colnames(all_samples_ratios)=c('geneID','start','stop',just_name_filenames)
```

```
colnames(all_samples_gene_coverage)=c('geneID',just_name_filenames)
colnames(all_samples_fracUTR)=c('geneID',just_name_filenames)

#fill the dataframe with the ratio/coverage values of the sample if they cover that isoform, if not with an NA
index=0
for (n in 1:nrow(UTR_iso_names)){

	all_samples_ratios[n,1:3]=t(as.matrix(c(toString(UTR_iso_names[n,1]),UTR_iso_names[n,2:3])))
	for (j in 1:length(list_of_ratios)){
		index=which(list_of_ratios[[j]][,1]==toString(UTR_iso_names[n,1]) &
	list_of_ratios[[j]][,5]==UTR_iso_names[n,2] &
	list_of_ratios[[j]][,6]==UTR_iso_names[n,3])
		if (length(index)>0){
			all_samples_ratios[n,j+3]=list_of_ratios[[j]][index,2]
		}
		else {
			all_samples_ratios[n,j+3]=NA
		}
	}
}


#do the same for all gene coverage
index=0
for (n in 1:nrow(gene_names)){
	all_samples_gene_coverage[n,1]=toString(gene_names[n,1])
	for (j in 1:length(list_of_gene_cov)){
		index=which(list_of_gene_cov[[j]][,1]==gene_names[n,1])
		if (length(index)==1){
			all_samples_gene_coverage[n,j+1]=list_of_gene_cov[[j]][index,2]
		}
		else {
			all_samples_gene_coverage[n,j+1]=NA
		}
	}
}

#make a matrix of fraction of total reads corresponding to the 3′UTR for each sample
index=0
for (n in 1:nrow(gene_names)){
	all_samples_fracUTR[n,1]=toString(gene_names[n,1])
	for (j in 1:length(list_of_gene_cov)){
```

```
                index=which(list_of_gene_cov[[j]][,1]==gene_names[n,1])
                if (length(index)==1){
                        fraction<-
list_of_gene_cov[[j]][index,4]/list_of_gene_cov[[j]][index,2]
                        all_samples_fracUTR[n,j+1]=fraction
                 }
                 else {
                        all_samples_fracUTR[n,j+1]=NA
                }
        }
}

#write output
#all_samples_ratios= all mpA peak ratios (iso:totalUTRisos)
#all_samples_gene_coverage= total coverage of entire gene for all genes
#all_samples_fracUTR= UTRreads/totalreads for each gene

write.table(all_samples_ratios, file=output_ratios, sep='\t', row.names=FALSE,
col.names=TRUE, quote=FALSE)
write.table(all_samples_gene_coverage, file=output_gene_coverage, sep='\t',
row.names=FALSE, col.names=TRUE, quote=FALSE)
write.table(all_samples_fracUTR, file=output_fracUTR, sep='\t', row.names=FALSE,
col.names=TRUE, quote=FALSE)

#then, download the combined samples files, and in local R combine close UTR peaks
(<40NT apart) created by differences in peak calling amongst samples, shown here for
MCxGr1 samples:

#input coverage and ratios files
all_ratios<-read.table('HumMCxGr1_all_ratios.tsv',header=TRUE,fill=TRUE,
stringsAsFactors=FALSE)
number_mpA_genes<-nrow(as.matrix(unique(all_ratios[,1])))
number_non_mpA_genes<-number_genes-number_mpA_genes

#combine any peaks that are within 40 nucleotides
ordered <- all_ratios[order(all_ratios$geneID, all_ratios$start),]
rownames(ordered)<-seq(1,nrow(ordered))
#make a first row for the dataframe that I'll delete later
this_peak<-ordered[1,]
indices<-0
for (k in 1:nrow(ordered)) {
        #only analyze this row if it hasn't been analyzed before
        if (length(which(indices==k))==0) {
                this_gene<-ordered[which(ordered[,1]==ordered[k,1]),]
```

```
rownums<-seq(1,nrow(this_gene))
#only analyze if there is more than one row corresponding to this isoform
if (nrow(this_gene)>1) {
	#add this gene to the indices so I can skip the rest of the rows
	indices<-c(indices, rownames(this_gene))
	sums=this_gene[2,2]-this_gene[1,2]
	#find the distance between the coordinates for isoforms from this gene
	if (nrow(this_gene)>2) {
		for (j in 2:(nrow(this_gene)-1)) {
			sum=this_gene[j+1,2]-this_gene[j,2]
			sums=rbind(sums, sum)
		}
	}
	#make a matrix of coordinates that are within 40 nucleotides of each other
	close<-as.matrix(which(sums<40))
	close_indices<-c(close,close+1)
	all<-c(rownums,close_indices)
	 #make a matrix of coordinates that are not within 40 nucleotides of any other peak, add to output. Close=matrix of indices of rows that are within 40 nucleotides of the next row
	not_close<-which(!(all %in% all[duplicated(all)]))
	for (l in 1:length(not_close)) {
		this_peak<-rbind(this_peak, this_gene[not_close[l],])
	 }
	#do the following if there are more than two peaks within 40 nucleotides for the gene
	if (nrow(close)>1) {
		 ind=0
		#combine peaks within 40 nucleotides of each other
		for (r in 1:nrow(close)) {
			#only do this if this row hasn't been analyzed
			if (length(which(ind==r))==0) {
				p=r
				rows<-vector(mode="numeric", length=0)
				#ID all rows within 40 nucleotides of this row
				while ((p<nrow(close) && close[p+1]-close[p]==1) || (p>1 && p<=nrow(close) && close[p]-close[p-1]==1)) {
					rows=c(rows, close[p],close[p]+1)
					ind=c(ind,p)
					 p=p+1
				}
```

```
                                        #combine all close peaks into one row, keep coordinates of 1st
                                        if (length(rows)>0) {
                                                data<-this_gene[unique(rows),]
                                                first<-data[1,]
                                                for (j in 1:(nrow(data)-1)) {
                                                        real<-which(!is.na(data[j+1,4:ncol(data)]))
                                                        first[1,(real+3)]<-data[j+1,(real+3)]
                                                 }
                                                #add to output
                                                this_peak<-rbind(this_peak, first)
                                        }
                                        else {
                                                data<-this_gene[close[r]:(close[r]+1),]
                                                first<-data[1,]
                                                real<-which(!is.na(data[2,4:ncol(data)]))
                                                first[1,(real+3)]<-data[2,(real+3)]
                                                this_peak<-rbind(this_peak, first)
                                        }
                                }
                        }
                }
                #do the following if there are only two peaks within 40 nucleotides of each other
                else if(nrow(close)==1) {
                        real<-which(!is.na(this_gene[(close+1),4:ncol(this_gene)]))
                        this<-this_gene[close,]
                        this[1,(real+3)]<-this_gene[(close+1),(real+3)]
                        this_peak<-rbind(this_peak, this)
                }
        }
        #if there is only one row corresponding to this gene, just add it to the output.
        else {
                this_peak=rbind(this_peak, ordered[k,])
        }
    }
}

this_peak<-this_peak[2:nrow(this_peak),]
```

```
this_peak<-this_peak[order(this_peak$geneID, this_peak$start),]
all_ratios<-this_peak
```

REFERENCES

1. Pringsheim, T. *et al.* The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Mov. Disord.* **27,** 1083–1091 (2012).
2. Huntington, G. On Chorea. *J. Neuropsychiatry Clin. Neurosci.* **15,** 317–321 (2003).
3. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–238 (1983).
4. MacDonald, M. E. *et al.* The Huntington's disease candidate region exhibits many different haplotypes. *Nat. Genet.* **1,** 99–103 (1992).
5. Taylor, S. *et al.* Cloning of the alpha-adducin gene from the huntingtons-disease candidate region of chromosome-4 by exon amplification. *Nat. Genet.* **2,** 223–227 (1992).
6. Ambrose, C. *et al.* A novel G protein-coupled receptor kinase gene cloned from 4p16.3. *Hum. Mol. Genet.* **1,** 697–703 (1992).
7. Macdonald, M. E. *et al.* A Novel Gene Containing a Trinucleotide That Is Expanded and Unstable on Huntington's Disease Chromosomes. *Cell* **72,** 971–983 (1993).
8. Rubinsztein, D. C., Barton, D. E., Davison, D. E. & Ferguson-Smith, M. A. Analysis of the huntingtin gene reveals a trinucleotide-length polymophism in the region of the gene that contains two CCG-rich streches and a corelation between decreased age of onset of Huntington's desease and CAG repeat number. *Hum. Mol. Genet.* **2,** 1713–1715 (1993).
9. Snell, R. G. *et al.* Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat. Genet.* **4,** 393–397 (1993).
10. Ambrose, C. *et al.* Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* **4,** 387–392 (1993).
11. Telenius, H. *et al.* Molecular analysis of juvenile huntington disease: The major influence on (CAG)n repeat length is the sex of the affected parent. *Hum. Mol. Genet.* **2,** 1535–1540 (1993).
12. Goldberg, Y. P. *et al.* Molecular analysis of new mutations for Huntington's disease: intermediate alleles and sex of origin effects. *Nat. Genet.* **5,** 174–9 (1993).
13. Wexler, N. S. *et al.* Homozygotes for Huntington's disease. *Nature* **326,** 194–7 (1987).
14. Squitieri, F. *et al.* Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course. *Brain* **126,** 946–955 (2003).
15. Trottier, Y. *et al.* Cellular localization of the Huntington's disease protein and discrimination of the normal and mutatted form. *Nat. Genet.* **10,** 196–201 (1995).
16. Zeitlin, S., Liu, J.-P., Chapman, D. L., Papaioannou, V. E. & Efstratiadis, A. Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homogue. *Nat. Genet.* **10,** 155–163 (1995).
17. Nasir, J., Floresco, S. & O'Kusky, J. Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell* **81,** 811–823 (1995).

18. Duyano, M. P. *et al.* Inactivation of the mouse Huntington's disease gene homolog Hdh. *Science (80-.).* **269,** 407–410 (1995).
19. Dragatsis, I., Efstratiadis, A. & Zeitlin, S. Mouse mutant embryos lacking huntingtin are rescued from lethality by wild-type extraembryonic tissues. *Development* **125,** 1529–1539 (1998).
20. Auerbach, W. *et al.* The HD mutation causes progressive lethal neurological disease in mice expressing reduced levels of huntingtin. *Hum. Mol. Genet.* **10,** 2515–2523 (2001).
21. White, J. K. *et al.* Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nat. Genet.* **17,** 404–10 (1997).
22. Dietrich, P., Shanmugasundaram, R., Shuyu, E. & Dragatsis, I. Congenital hydrocephalus associated with abnormal subcommissural organ in mice lacking huntingtin in Wnt1 cell lineages. *Hum. Mol. Genet.* **18,** 142–150 (2009).
23. Reiner, A. *et al.* Neurons lacking huntingtin differentially colonize brain and survive in chimeric mice. *J. Neurosci.* **21,** 7608–19 (2001).
24. Histories, C. & Findings, A. Neuropathologieal Findings in Wolf-Hirsehhorn (4p-) Syndrome. *Acta Neuropathol.* 163–165 (1981).
25. Ambrose, C. M. *et al.* Structure and expression of the Huntington's disease gene: Evidence against simple inactivation due to an expanded CAG repeat. *Somat. Cell Mol. Genet.* **20,** 27–38 (1994).
26. Cattaneo, E., Zuccato, C. & Tartari, M. Normal huntingtin function: an alternative approach to Huntington's disease. *Nat. Rev. Neurosci.* **6,** 919–30 (2005).
27. Zuccato, C., Valenza, M. & Cattaneo, E. Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. *Physiol Rev* **90,** 905–981 (2010).
28. Harjes, P. & Wanker, E. E. The hunt for huntingtin function: Interaction partners tell many different stories. *Trends Biochem. Sci.* **28,** 425–433 (2003).
29. Vonsattel, J. P. G. & DiFiglia, M. Huntington disease. *J. Neuropathol. Exp. Neurol.* **57,** 369–384 (1998).
30. Rüb, U. *et al.* Degeneration of the cerebellum in huntingtons disease (HD): Possible relevance for the clinical picture and potential gateway to pathological mechanisms of the disease process. *Brain Pathol.* **23,** 165–177 (2013).
31. Chung, D. W., Rudnicki, D. D., Yu, L. & Margolis, R. L. A natural antisense transcript at the Huntington's disease repeat locus regulates HTT expression. *Hum. Mol. Genet.* **20,** 3467–3477 (2011).
32. de Mezer, M., Wojciechowska, M., Napierala, M., Sobczak, K. & Krzyzosiak, W. J. Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. *Nucleic Acids Res.* **39,** 3852–3863 (2011).
33. Liu, W. *et al.* Increased Steady-State Mutant Huntingtin mRNA in Huntington's Disease Brain. **2,** 491–500 (2013).
34. Labadorf, A. T. & Myers, R. H. Evidence of extensive alternative splicing in post mortem human brain HTT transcription by mRNA sequencing. *PLoS One* **10,** 1–7 (2015).
35. Hughes, A. C. *et al.* Identification of novel alternative splicing events in the

huntingtin gene and assessment of the functional consequences using structural protein homology modelling. *J. Mol. Biol.* **426,** 1428–1438 (2014).

36. Ruzo, A. *et al.* Discovery of novel isoforms of Huntingtin reveals a new hominid-specific exon. *PLoS One* **10,** (2015).
37. Lin, B. *et al.* Differential 3’ polyadenylation of the huntington disease gene results in two mRNA species with variable tissue expression. *Hum. Mol. Genet.* **2,** 1541–1545 (1993).
38. Sheets, M. D., Ogg, S. C. & Wickens, M. P. Point mutations in AAUAAA and the poly(A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation {Iin vitro}. *Nucleic Acids Res* **18,** 5799–5805 (1990).
39. Xu, H., An, J. J. & Xu, B. Distinct cellular toxicity of two mutant huntingtin mRNA variants due to translation regulation. *PLoS One* 1–19 (2017).
40. Sathasivam, K. *et al.* Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2366–70 (2013).
41. Francelle, L., Lotz, C., Outeiro, T., Brouillet, E. & Merienne, K. Contribution of Neuroepigenetics to Huntington’s Disease. *Front. Hum. Neurosci.* **11,** 17 (2017).
42. Cha, J. H. J. Transcriptional signatures in Huntington’s disease. *Prog. Neurobiol.* **83,** 228–248 (2007).
43. Martí, E. RNA toxicity induced by expanded CAG repeats in Huntington’s disease. *Brain Pathol.* **26,** 779–786 (2016).
44. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14,** 496–506 (2013).
45. Gruber, A. R., Martin, G., Keller, W. & Zavolan, M. Means to an end: Mechanisms of alternative polyadenylation of messenger RNA precursors. *Wiley Interdiscip. Rev. RNA* **5,** 183–196 (2014).
46. Pfister, E. L. *et al.* Five siRNAs targeting three SNPs may provide therapy for three-quarters of Huntington’s disease patients. *Curr. Biol.* **19,** 774–8 (2009).
47. Østergaard, M. E. *et al.* Rational design of antisense oligonucleotides targeting single nucleotide polymorphisms for potent and allele selective suppression of mutant Huntingtin in the CNS. *Nucleic Acids Res.* **41,** 9634–9650 (2013).
48. Glover-Cutter, K., Kim, S., Espinosa, J. & Bentley, D. L. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat. Struct. Mol. Biol.* **15,** 71–8 (2008).
49. Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes. Dev.* **11,** 2755–2766 (1997).
50. Takagaki, Y., Ryner, L. C. & Manley, J. L. Four factors are required for 3’ -end cleavage of pre-mRNAs. *Genes Dev.* **3,** 1711–1724 (1989).
51. Proudfoot, N. & Brownlee, G. Sequence at the 3’end of globin mRNA shows homology with immunoglobulin light chain mRNA. *Nature* **252,** 359–62 (1974).
52. Proudfoot, N. J. Ending the message : poly (A) signals then and now. *Genes Dev.* **25,** 1770–1782 (2011).
53. MacDonald, C. C., Wilusz, J. & Shenk, T. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol. Cell. Biol.* **14,** 6647–6654 (1994).

54. Chen, F., MacDonald, C. C. & Wilusz, J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res.* **23,** 2614–20 (1995).
55. Shi, Y. Alternative polyadenylation : New insights from global analyses. 2105–2117 (2012). doi:10.1261/rna.035899.112.cleavage/polyadenylation
56. Norbury, C. J. Cytoplasmic RNA: a case of the tail wagging the dog. *Nat Rev Cancer* **13,** 643–653 (2013).
57. Derti, a. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22,** 1173–1183 (2012).
58. Beaudoing, E. Patterns of Variant Polyadenylation Signal Usage in Human Genes. **10,** 1001–1010 (2000).
59. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17,** 761–772 (2011).
60. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide Analysis of Pre-mRNA 3' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3' UTR Length. *Cell Rep.* **1,** 753–763 (2012).
61. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469,** 97–101 (2011).
62. Smibert, P. *et al.* Global Patterns of Tissue-Specific Alternative Polyadenylation in Drosophila. *Cell Rep.* **1,** 277–289 (2012).
63. Pinto, P. A. B. *et al.* RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J.* **30,** 2431–2444 (2011).
64. Martincic, K., Alkan, S. a, Cheatle, A., Borghesi, L. & Milcarek, C. Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing. *Nat. Immunol.* **10,** 1102–1109 (2009).
65. Oktaba, K. *et al.* ELAV Links Paused Pol II to Alternative Polyadenylation in the Drosophila Nervous System. *Mol. Cell* **57,** 341–348 (2014).
66. Ji, Z. *et al.* Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol Syst Biol* **7,** 534 (2011).
67. Nagaike, T. *et al.* Transcriptional Activators Enhance Polyadenylation of mRNA Precursors. *Mol. Cell* **41,** 409–418 (2011).
68. Huang, Y. *et al.* Mediator Complex Regulates Alternative mRNA Processing via the MED23 Subunit. *Mol. Cell* **45,** 459–469 (2012).
69. Wood, A. J. *et al.* Regulation of alternative polyadenylation by genomic imprinting. 1141–1146 (2008). doi:10.1101/gad.473408.products
70. Cowley, M., Wood, A. J., Böhm, S., Schulz, R. & Oakey, R. J. Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. *Nucleic Acids Res.* **40,** 8917–8926 (2012).
71. Takagaki, Y., Seipelt, R. L., Peterson, M. L. & Manley, J. L. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87,** 941–952 (1996).
72. Chuvpilo, S. *et al.* Alternative polyadenylation events contribute to the induction of NF-ATc in effector T cells. *Immunity* **10,** 261–269 (1999).
73. Elkon, R. *et al.* E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol.* **13,** R59 (2012).

74. Castelo-Branco, P. *et al.* Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol. Cell. Biol.* **24,** 4174–83 (2004).
75. Gawande, B., Robida, M. D., Rahn, A. & Singh, R. Drosophila Sex-lethal protein mediates polyadenylation switching in the female germline. *EMBO J.* **25,** 1263–1272 (2006).
76. Dai, W., Zhang, G. & Makeyev, E. V. RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res.* **40,** 787–800 (2012).
77. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456,** 464–9 (2008).
78. Dittmar, K. A. *et al.* Genome-Wide Determination of a Broad ESRP-Regulated Posttranscriptional Network by High-Throughput Sequencing. *Mol. Cell. Biol.* **32,** 1468–1482 (2012).
79. Jenal, M. *et al.* The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* **149,** 538–553 (2012).
80. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468,** 664–668 (2010).
81. Berg, M. G. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150,** 53–64 (2012).
82. Neve, J. *et al.* Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Res.* **26,** 24–35 (2016).
83. Ni, T. *et al.* Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics* **14,** 615 (2013).
84. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27,** 2380–96 (2013).
85. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods* **10,** 133–9 (2013).
86. Miura, P. *et al.* Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res.* **23,** 812–825 (2013).
87. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science (80-.).* **320,** 1643–1647 (2008).
88. Weng, L., Li, Y. I., Xie, X. & Shi, Y. Poly (A) code analyses reveal key determinants for tissue-specific mRNA alternative polyadenylation. 813–821 (2016). doi:10.1261/rna.055681.115.4
89. Zhang, H., Lee, J. Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol.* **6,** R100 (2005).
90. Kislauskis, E. H. & Singer, R. H. Determinants of mRNA localization. *Curr. Opin. Cell Biol.* **4,** 975–978 (1992).
91. Andreassi, C. & Riccio, A. To localize or not to localize: mRNA fate is in 3′UTR ends. *Trends Cell Biol.* **19,** 465–474 (2009).
92. Takizawa, P. a, Sil, a, Swedlow, J. R., Herskowitz, I. & Vale, R. D. Actin-dependent localization of an RNA encoding a cell-fate determinant in yeast.

Nature **389,** 90–93 (1997).
93. Kislauskis, E. H., Li, Z., Singer, R. H. & Taneja, K. L. Isoform-specific 3′-untranslated sequences sort α-cardiac and β-cytoplasmic actin messenger RNAs to different cytoplasmic compartments. *J. Cell Biol.* **123,** 165–172 (1993).
94. An, J. J. *et al.* Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* **134,** 175–87 (2008).
95. Fukuchi, M. & Tsuda, M. Involvement of the 3'-untranslated region of the brain-derived neurotrophic factor gene in activity-dependent mRNA stabilization. *J. Neurochem.* **115,** 1222–1233 (2010).
96. Blichenberg, A. *et al.* Identification of a *cis*-acting dendritic targeting element in the mRNA encoding the alpha subunit of Ca^{2+}/calmodulin-dependent protein kinase II. *Eur J Neurosci* **13,** 1881–1888 (2001).
97. Berkovits, B. D. & Mayr, C. Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522,** 363–367 (2015).
98. Barreau, C., Paillard, L. & Osborne, H. B. AU-rich elements and associated factors: Are there unifying principles? *Nucleic Acids Res.* **33,** 7138–7150 (2005).
99. Weill, L., Belloc, E., Bava, F. A. & Mendez, R. Translational control by changes in poly(A) tail length: recycling mRNAs. *Nat Struct Mol Biol* **19,** 577–585 (2012).
100. Coller, J. M., Gray, N. K. & Wickens, M. P. mRNA stabilization by poly (A) binding protein is independent of poly (A) and requires translation. *Genes Dev.* **12,** 3226–3235 (1998).
101. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508,** 66–71 (2014).
102. Park, J. E., Yi, H., Kim, Y., Chang, H. & Kim, V. N. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Mol. Cell* **62,** 462–471 (2016).
103. Legendre, M., Ritchie, W., Lopez, F. & Gautheret, D. Differential repression of alternative transcripts: A screen for miRNA targets. *PLoS Comput. Biol.* **2,** 333–342 (2006).
104. Gebauer, F., Preiss, T. & Hentze, M. W. From Cis -Regulatory Elements to Complex. *Cold Spring Harb. Perspect. Biol.* **4,** 1–14 (2012).
105. Hasan, A., Cotobal, C., Duncan, C. D. S. & Mata, J. Systematic Analysis of the Role of RNA-Binding Proteins in the Regulation of RNA Stability. *PLoS Genet.* **10,** (2014).
106. Nicholson, P. *et al.* Nonsense-mediated mRNA decay in human cells: Mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell. Mol. Life Sci.* **67,** 677–700 (2010).
107. Muhlrad, D. & Parker, R. Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *RNA* **5,** 1299–1307 (1999).
108. Hogg, J. R. & Goff, S. P. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell* **143,** 379–389 (2010).
109. Szostak, E. & Gebauer, F. Translational control by 3'-UTR-binding proteins. *Brief. Funct. Genomics* **12,** 58–65 (2013).
110. Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative

Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138,** 673–684 (2009).
111. Lau, A. G. *et al.* Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). *Proc. Natl. Acad. Sci.* **107,** 15945–15950 (2010).
112. Spies, N., Burge, C. B. & Bartel, D. P. 3′ UTR-Isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23,** 2078–2090 (2013).
113. Yao, P. *et al.* Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell* **149,** 88–100 (2012).
114. Márquez, A. *et al.* Two Functional Variants of IRF5 Influence the Development of Macular Edema in Patients with Non-Anterior Uveitis. *PLoS One* **8,** 6–11 (2013).
115. Rhinn, H. *et al.* Alternative α-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. *Nat. Commun.* **3,** 1084 (2012).
116. Tian, B. & Manley, J. L. Alternative cleavage and polyadenylation: The long and short of it. *Trends Biochem. Sci.* **38,** 312–320 (2013).
117. Lin, L. *et al.* Transcriptome sequencing reveals aberrant alternative splicing in Huntington's disease. *Hum. Mol. Genet.* **25,** 1–13 (2016).
118. Steffan, J. S. *et al.* The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. *Proc Natl Acad Sci U S A* **97,** 6763–6768 (2000).
119. Jiang, H. *et al.* Depletion of CBP is directly linked with cellular toxicity caused by mutant huntingtin. *Neurobiol. Dis.* **23,** 543–551 (2006).
120. Nucifora, F. C. *et al.* Interference by Huntingtin and Atrophin -1 with CBP-Mediated Transcription Leading to Cellular Toxicity. *Science (80-.).* **291,** 2423–2428 (2001).
121. Igarashi, S. *et al.* Inducible PC12 cell model of Huntington's disease shows toxicity and decreased histone acetylation. *Neuroreport* **14,** 565–568 (2003).
122. Guiretti, D. *et al.* Specific promoter deacetylation of histone H3 is conserved across mouse models of Huntington's disease in the absence of bulk changes. *Neurobiol. Dis.* **89,** 190–201 (2016).
123. Valor, L. M., Guiretti, D., Lopez-Atalaya, J. P. & Barco, A. Genomic landscape of transcriptional and epigenetic dysregulation in early onset polyglutamine disease. *J Neurosci* **33,** 10471–10482 (2013).
124. Achour, M. *et al.* Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice. *Hum. Mol. Genet.* **24,** 3481–3496 (2015).
125. Dong, X. *et al.* The role of H3K4me3 in transcriptional regulation is altered in Huntington's disease. *PLoS One* **10,** 1–23 (2015).
126. Ng, C. W. *et al.* Extensive changes in DNA methylation are associated with expression of mutant huntingtin. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2354–9 (2013).
127. De Souza, R. A. G. *et al.* DNA methylation profiling in human Huntington's disease brain. *Hum. Mol. Genet.* **25,** 2013–2030 (2016).

128. Sadri-Vakili, G. *et al.* Histones associated with downregulated genes are hypo-acetylated in Huntington's disease models. *Hum. Mol. Genet.* **16,** 1293–1306 (2007).
129. Thomas, E. A. *et al.* The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 15564–9 (2008).
130. Jia, H., Morris, C. D., Williams, R. M., Loring, J. F. & Thomas, E. a. HDAC inhibition imparts beneficial transgenerational effects in Huntington's disease mice via altered DNA and histone methylation. *Proc. Natl. Acad. Sci.* **112,** E56–E64 (2015).
131. Vashishtha, M. *et al.* Targeting H3K4 trimethylation in Huntington disease. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E3027-36 (2013).
132. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and Consequences of Alternative Polyadenylation. *Mol. Cell* **43,** 853–866 (2011).
133. Augood, S. J., Faull, R. L. M., Love, D. R. & Emson, P. C. Messenger RNA in the striatum of early grade Huntington's disease: a detailed cellular in situ hybridization study. *Neuroscience* **72,** 1023–1036 (1996).
134. Cha, J. H. *et al.* Altered neurotransmitter receptor expression in transgenic mouse models of Huntington's disease. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **354,** 981–9 (1999).
135. Luthi-Carter, R. *et al.* Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. *Hum Mol Genet* **9,** 1259–71. (2000).
136. Desplats, P. A. *et al.* Selective deficits in the expression of striatal-enriched mRNAs in Huntington's disease. *J. Neurochem.* **96,** 743–757 (2006).
137. Mazarei, G. *et al.* Expression analysis of novel striatal-enriched genes in Huntington disease. *Hum. Mol. Genet.* **19,** 609–622 (2009).
138. Hodges, A. *et al.* Regional and cellular gene expression changes in human Huntington's disease brain. *Hum. Mol. Genet.* **15,** 965–77 (2006).
139. Kuhn, A. *et al.* Mutant huntingtin's effects on striatal gene expression in mice recapitulate changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage. *Hum. Mol. Genet.* **16,** 1845–1861 (2007).
140. Zuccato, C. *et al.* Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.* **35,** 76–83 (2003).
141. Zuccato, C. *et al.* Widespread Disruption of Repressor Element- 1 Silencing Transcription Factor / Neuron- Restrictive Silencer Factor Occupancy at Its Target Genes in Huntington's Disease. **27,** 6972–6983 (2016).
142. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science (80-.).* **267,** 1360–3 (1995).
143. Ravache, M., Weber, C., Mérienne, K. & Trottier, Y. Transcriptional activation of REST by Sp1 in huntington's disease models. *PLoS One* **5,** (2010).
144. Pugh, B. F. & Tjian, R. Mechanism of transcriptional activation by Sp1: Evidence for coactivators. *Cell* **61,** 1187–1197 (1990).

145. Dunah, A. W. *et al.* Sp1 and TAFII130 Transcriptional Activity Disrupted in Early Huntington's Disease. *Science (80-.).* **296,** 2238–2243 (2002).
146. Suhr, S. T. *et al.* Identities of sequestered proteins in aggregates from cells with induced polyglutamine expression. *J. Cell Biol.* **153,** 283–294 (2001).
147. Chen-Plotkin, A. S. *et al.* Decreased association of the transcription factor Sp1 with genes downregulated in Huntington's disease. *Neurobiol. Dis.* **22,** 233–241 (2006).
148. Cui, L. *et al.* Transcriptional Repression of PGC-1α by Mutant Huntingtin Leads to Mitochondrial Dysfunction and Neurodegeneration. *Cell* **127,** 59–69 (2006).
149. Shimohata, T. *et al.* Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. *Nat. Genet.* **26,** 29–36 (2000).
150. van Roon-mom, W. M. C. Van *et al.* Insoluble TATA-binding protein accumulation in Huntington ' s disease cortex. **109,** 1–10 (2002).
151. Stevanin, G. *et al.* Huntington's disease-like phenotype due to trinucleotide repeat expansions in the TBP and JPH3 genes. *Brain* **126,** 1599–1603 (2003).
152. Boutell, J. M. *et al.* Aberrant interactions of transcriptional repressor proteins with the Huntington's disease gene product, huntingtin. *Hum. Mol. Genet.* **8,** 1647–1655 (1999).
153. Savas, J. N. *et al.* Huntington's disease protein contributes to RNA-mediated gene silencing through association with Argonaute and P bodies. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 10820–5 (2008).
154. Conaco, C., Otto, S., Han, J.-J. & Mandel, G. Reciprocal actions of REST and a microRNA promote neuronal identity. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2422–7 (2006).
155. Packer, A. N., Xing, Y., Harper, S. Q., Jones, L. & Davidson, B. L. The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease. *J. Neurosci.* **28,** 14341–6 (2008).
156. Johnson, R. *et al.* A microRNA-based gene dysregulation pathway in Huntington's disease. *Neurobiol. Dis.* **29,** 438–445 (2008).
157. Sinha, M., Ghose, J., Das, E. & Bhattarcharyya, N. P. Altered microRNAs in STHdhQ111/HdhQ111 cells: miR-146a targets TBP. *Biochem. Biophys. Res. Commun.* **396,** 742–747 (2010).
158. Martí, E. *et al.* A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res.* **38,** 7219–7235 (2010).
159. Sinha, M., Ghose, J. & Bhattarcharyya, N. P. Micro RNA-214, -150, -146a and -125b target Huntingtin gene. *RNA Biol.* **8,** 1005–1021 (2011).
160. Gaughwin, P. M. *et al.* Hsa-miR-34b is a plasma-stable microRNA that is elevated in pre-manifest Huntington's disease. *Hum. Mol. Genet.* **20,** 2225–2237 (2011).
161. Jovicic, A., Zaldivar Jolissaint, J. F., Moser, R., Silva Santos, M. de F. & Luthi-Carter, R. MicroRNA-22 (miR-22) Overexpression Is Neuroprotective via General Anti-Apoptotic Effects and May also Target Specific Huntington's Disease-Related Mechanisms. *PLoS One* **8,** 2–9 (2013).
162. Cheng, P. H. *et al.* MiR-196a ameliorates phenotypes of huntington disease in cell,

transgenic mouse, and induced pluripotent stem cell models. *Am. J. Hum. Genet.* **93,** 306–312 (2013).
163. Fernandez-Nogales, M. *et al.* Huntington's disease is a four-repeat tauopathy with tau nuclear rods. *Nat Med* **20,** 881–885 (2014).
164. Gu, J. *et al.* Cyclic AMP-dependent protein kinase regulates 9G8-mediated alternative splicing of tau exon 10. *FEBS Lett.* **586,** 2239–2244 (2012).
165. Vuono, R. *et al.* The role of tau in the pathological process and clinical expression of Huntington's disease. *Brain* **138,** 1907–18 (2015).
166. Cabrera, J. R. & Lucas, J. J. MAP2 Splicing is Altered in Huntington's Disease. *Brain Pathol.* 1–9 (2016). doi:10.1111/bpa.12387
167. Krzyzosiak, W. J. *et al.* Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.* **40,** 11–26 (2012).
168. Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. Atomic resolution structure of CAG RNA repeats: Structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res.* **38,** 8370–8376 (2010).
169. Miller, J. W. *et al.* Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J.* **19,** 4439–4448 (2000).
170. Warf, M. B., Diegel, J. V, von Hippel, P. H. & Berglund, J. A. The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9203–9208 (2009).
171. Mankodi, A. *et al.* Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. *Hum. Mol. Genet.* **10,** 2165–2170 (2001).
172. Osborne, R. J. *et al.* Transcriptional and post-transcriptional impact of toxic RNA in myotonic dystrophy. *Hum. Mol. Genet.* **18,** 1471–1481 (2009).
173. Mykowska, A., Sobczak, K., Wojciechowska, M., Kozlowski, P. & Krzyzosiak, W. J. CAG repeats mimic CUG repeats in the misregulation of alternative splicing. *Nucleic Acids Res.* **39,** 8938–8951 (2011).
174. Li, L.-B., Yu, Z., Teng, X. & Bonini, N. M. RNA toxicity is a component of ataxin-3 degeneration in Drosophila. *Nature* **453,** 1107–11 (2008).
175. Hsu, R. J. *et al.* Long Tract of untranslated CAG repeats is deleterious in transgenic mice. *PLoS One* **6,** (2011).
176. Sun, X. *et al.* Nuclear retention of full-length HTT RNA is mediated by splicing factors MBNL1 and U2AF65. *Sci. Rep.* **5,** 12521 (2015).
177. Bañez-Coronel, M. *et al.* RAN Translation in Huntington Disease. *Neuron* **88,** 667–677 (2015).
178. Batra, R. *et al.* Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Mol. Cell* **56,** 311–322 (2014).
179. Lin, B. *et al.* Differential 3' polyadenylation of the huntington disease gene results in two mRNA species with variable tissue expression. *Hum. Mol. Genet.* **2,** 1541–1545 (1993).
180. Labadorf, A. *et al.* RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression. *PLoS One* **10,** 1–21 (2015).

181. Vonsattel, J.-P. *et al.* Neuropathological Classification of Huntinqton's Disease. *J. Neuropathol. Exp. Neurol.* **44,** 559–577 (1985).
182. Vonsattel, J. P. G. Huntington disease models and human neuropathology: Similarities and differences. *Acta Neuropathol.* **115,** 55–69 (2008).
183. Thu, D. C. V *et al.* Cell loss in the motor and cingulate cortex correlates with symptomatology in Huntington's disease. *Brain* **133,** 1094–1110 (2010).
184. Slow, E. J. *et al.* Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease. *Hum. Mol. Genet.* **12,** 1555–1567 (2003).
185. Sheppard, S., Lawson, N. D. & Zhu, L. J. Accurate identification of polyadenylation sites from 3' end deep sequencing using a naive Bayes classifier. *Bioinformatics* **29,** 2564–71 (2013).
186. Menalled, L. B., Sison, J. D., Dragatsis, I., Zeitlin, S. & Chesselet, M.-F. Time course of early motor and neuropathological anomalies in a knock-in mouse model of Huntington's disease with 140 CAG repeats. *J. Comp. Neurol.* **465,** 11–26 (2003).
187. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33,** 201–212 (2005).
188. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27,** 2380–2396 (2013).
189. Datta, P. K., Bonner, J. F., Haas, C. J. & Fischer, I. Considerations for the Use of SH-SY5Y Neuroblastoma Cells in Neurobiology. *Neuronal Cell Cult. Methods Protoc.* **1078,** 35–44 (2013).
190. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications. *Mol. Cell* **53,** 1044–1052 (2014).
191. Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of Mammalian MicroRNA Targets. *Cell* **115,** 787–798 (2003).
192. Broderick, J. A., Salomon, W. E., Ryder, S. P., Aronin, N. & Zamore, P. D. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA* **17,** 1858–1869 (2011).
193. Yang, Y.-C. T. *et al.* CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* **16,** doi: 10.1186/s12864-015-1273-2 (2015).
194. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42,** 92–97 (2014).
195. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347,** 1260419–1260419 (2015).
196. Panwar, B., Omenn, G. S. & Guan, Y. miRmine: A Database of Human miRNA Expression Profiles. *Bioinformatics* **2017,** doi: 10.1093/bioinformatics/btx019 (2017).
197. Moumné, L., Betuing, S. & Caboche, J. Multiple Aspects of Gene Dysregulation in Huntington's Disease. *Front. Neurol.* **4,** 127 (2013).
198. Zuccato, C., Valenza, M. & Cattaneo, E. Molecular Mechanisms and Potential

Therapeutical Targets in Huntington's Disease. *Physiol. Rev.* **90,** 905–981 (2010).
199. Martinez-Vicente, M. *et al.* Cargo recognition failure is responsible for inefficient autophagy in Huntington's disease. *Nat. Neurosci.* **13,** 567–576 (2010).
200. Ellrichmann, G., Reick, C., Saft, C. & Linker, R. A. The Role of the Immune System in Huntington's Disease. *Clin. Dev. Immunol.* **2013,** doi:10.1155/2013/541259 (2013).
201. Trushina, E. *et al.* Microtubule destabilization and nuclear entry are sequential steps leading to toxicity in Huntington's disease. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 12171–12176 (2003).
202. Sang, Q. *et al.* Nedd4-WW Domain-Binding Protein 5 (Ndfip1) Is Associated with Neuronal Survival after Acute Cortical Brain Injury. *J. Neurosci.* **26,** 7234–7244 (2006).
203. Seaman, M. N. J., Sowerby, P. J. & Robinson, M. S. Cytosolic and membrane-associated proteins involved in the recruitment of AP-1 adaptors onto the trans-Golgi network. *J. Biol. Chem.* **271,** 25446–25451 (1996).
204. Trushina, E. *et al.* Mutant Huntingtin Impairs Axonal Trafficking in Mammalian Neurons In Vivo and In Vitro. *Mol. Cell. Biol.* **24,** 8195–8209 (2004).
205. Moon, H. Y. *et al.* Running-Induced Systemic Cathepsin B Secretion Is Associated with Memory Function. *Cell Metab.* **24,** 332–340 (2016).
206. Kohl, Z. *et al.* Physical activity fails to rescue hippocampal neurogenesis deficits in the R6/2 mouse model of Huntington's disease. *Brain Res.* **1155,** 24–33 (2007).
207. Gipson, T. A., Neueder, A., Wexler, N. S. & Bates, G. P. Aberrantly spliced HTT, a new player in Huntington's disease pathogenesis. **6286,** (2017).
208. Udagawa, T. *et al.* Bidirectional control of mRNA translation and synaptic plasticity by the cytoplasmic polyadenylation complex. *Mol. Cell* **47,** 253–66 (2012).
209. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
210. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25.1-R25.10 (2009).
211. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
212. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
213. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41,** 377–386 (2013).
214. Shirai, Y. T., Suzuki, T., Morita, M., Takahashi, A. & Yamamoto, T. Multifunctional roles of the mammalian CCR4-NOT complex in physiological phenomena. *Front. Genet.* **5,** 1–11 (2014).
215. Mittal, S., Aslam, A., Doidge, R., Medica, R. & Winkler, G. S. The Ccr4a (CNOT6) and Ccr4b (CNOT6L) deadenylase subunits of the human Ccr4-Not complex contribute to the prevention of cell death and senescence. *Mol. Biol. Cell*

22, 748–58 (2011).
216. Ji, X., Wan, J., Vishnu, M., Xing, Y. & Liebhaber, S. A. αCP Poly (C) Binding Proteins Act as Global Regulators of Alternative. *Mol. Cell. Biol.* **33,** 2560–2573 (2013).
217. Katahira, J. *et al.* Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. *Nucleic Acids Res.* **41,** 7060–7072 (2013).
218. The HD iPSC Consortium. Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. *Cell Stem Cell* **11,** 264–78 (2012).
219. Isakson, P., Holland, P. & Simonsen, A. The role of ALFY in selective autophagy. *Cell Death Differ.* **20,** 12–20 (2012).
220. Fox, L. M. Autophagy-linked FYVE protein mediates the turnover of mutant huntingtin and modifies pathogenesis in mouse models of Huntington's disease (Doctoral dissertation). (Columbia, 2016).
221. Lin, L. *et al.* Atypical ubiquitination by E3 ligase WWP1 inhibits the proteasome-mediated degradation of mutant huntingtin. **1643,** 103–112 (2016).
222. Martino, D., Stamelou, M. & Bhatia, K. P. The differential diagnosis of Huntington's disease- like syndromes : 'red flags' for the clinician. *J. Neurol. Nuerosurgery, Psychiatry* **84,** 650–656 (2013).
223. Strehlow, A. N. T., Li, J. Z. & Myers, R. M. Wild-type huntingtin participates in protein trafficking between the Golgi and the extracellular space. **16,** 391–409 (2007).
224. Oyama, F., Miyazaki, H., Sakamoto, N., Becquet, C. & Machida, Y. Sodium channel b4 subunit : down-regulation and possible involvement in neuritic degeneration in Huntington's disease transgenic mice. 518–529 (2006). doi:10.1111/j.1471-4159.2006.03893.x
225. Qureshi, I. A., Gokhan, S. & Mehler, M. F. REST and CoREST are transcriptional and epigenetic regulators of seminal neural fate decisions. 4477–4486 (2010). doi:10.4161/cc.9.22.13973
226. Velier, J. *et al.* Wild-type and mutant huntingtins function in vesicle trafficking in the secretory and endocytic pathways. *Exp. Neurol.* **152,** 34–40 (1998).
227. Squitieri, F. *et al.* Abnormal morphology of peripheral cell tissues from patients with Huntington disease. *J. Neural Transm.* **117,** 77–83 (2010).
228. Trettel, F. *et al.* Dominant phenotypes produced by the HD mutation in STHdh(Q111) striatal cells. *Hum. Mol. Genet.* **9,** 2799–2809 (2000).
229. Aronin, N. & Difiglia, M. Huntingtin-lowering strategies in Huntington's disease: Antisense oligonucleotides, small RNAs, and gene editing. *Mov. Disord.* **29,** 1455–1461 (2014).
230. Garriga-canut, M., Agustín-pavón, C., Herrmann, F., Sánchez, A. & Dierssen, M. Synthetic zinc finger repressors reduce mutant huntingtin expression in the brain of R6/2 mice. *PNAS* **109,** E3136–E3145 (2012).
231. Yang, S. *et al.* CRISPR/Cas9-mediated gene editing ameliorates neurotoxicity in mouse model of Huntington's disease. **127,** 2719–2724 (2017).

232. Zhang, Y., Engelman, J. & Friedlander, R. M. Allele-specific silencing of mutant Huntington's disease gene. *J. Neurochem.* **108,** 82–90 (2009).
233. Hu, J., Liu, J. & Corey, D. R. Allele-selective inhibition of huntingtin expression by switching to an miRNA-like RNAi mechanism. *Chem. Biol.* **17,** 1183–1188 (2010).
234. Liu, W., Kennington, L., Rosas, H. & Hersch, S. Linking SNPs to CAG repeat length in Huntington's disease patients. *Nat. Methods* **5,** 951–953 (2008).
235. Spiess, A.-N. & Ivell, R. A Highly Efficient Method for Long-Chain cDNA Synthesis Using Trehalose and Betaine. *Anal. Biochem.* **301,** 168–174 (2002).