Targeted 3'-end RNA sequencing uncovers cryptic polyadenylation in Huntington's disease linked to somatic instability and CAG repeat purity

Ane Velasco-Bilbao[1], María Manterola[1], Alvaro Herrero-Reiriz[1], María Carazo-Hidalgo[1], Anna Misiukiewicz[2], Olatz Arnold-Garcia[1,3-4], Esther Perez-Navarro[3,12,13], Martina Hallegger[5,6], Jernej Ule[5], Alberto Rabano[3,7], Adolfo López de Munain[3,8-11], Marta Olejniczak[2], Veronica Brito[3,12,13], Lorea Blazquez[1,3,14, *]

[1] Neurogenetics, RNA biology and therapies group. Department of Neurosciences, Biogipuzkoa Health Research Institute. San Sebastián, Spain.

[2] Department of Genome Engineering, Institute of Bioorganic Chemistry, Polish Academy of Sciences. Poznan, Poland.

[3] Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas (CIBERNED), Instituto de Salud Carlos III. Madrid, Spain.

[4] Department of Cellular and Molecular Medicine. University of California, San Diego. La Jolla, CA, USA.

[5] RNA Networks group. The Francis Crick Institute. London, UK.

[6] Oxford-GSK Institute of Molecular and Computational Medicine, Centre for Human Genetics, University of Oxford, UK.

[7] CIEN Foundation, Reina Sofia Alzheimer Center, ISCIII. Madrid, Spain.

[8] Neuromuscular Diseases group. Department of Neurosciences, Biogipuzkoa Health Research Institute. San Sebastián, Spain.

[9] Department of Neurology. CSUR Rare Neuromuscular Disorders. Donostia University Hospital-OSAKIDETZA. San Sebastián, Spain

[10] Department of Neurosciences. University of the Basque Country. San Sebastián, Spain.

[11] Department of Medicine. School of Medicine. University of Deusto. Bilbao, Spain.

[12] Departament de Biomedicina, Facultat de Medicina, Institut de Neurosciències, Universitat de Barcelona. Barcelona, Spain.

[13] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). Barcelona, Spain.

[14] Ikerbasque, Basque Foundation for Science. Bilbao, Spain.

* Correspondence to Lorea Blázquez

Email: lorea.blazquezgarcia@bio-gipuzkoa.eus

Keywords:
RNA-processing, Huntington´s disease, targeted 3´end sequencing, cryptic polyadenylation, *HTT1a*

Abstract

Huntington's disease (HD) is a progressive neurodegenerative disorder, caused by a CAG trinucleotide expansion in the first exon of the *HTT* gene. The full-penetrance threshold for HD is established at 40 CAG repeats, which are translated into a polyglutamine (polyQ) tract within the huntingtin (HTT) protein. Beyond this threshold, somatic repeat instability in the brain and CAG repeat purity can further modulate disease onset and severity.

HTT can be processed into multiple N-terminal fragments. Among these, the exon 1 HTT fragment containing an expanded polyQ tract is considered the most pathogenic. While some studies propose this fragment is generated by HTT proteolytic cleavage, others indicate it is primarily generated by the choice of a premature, cryptic polyadenylation site within *HTT* intron 1. Here, we apply a targeted RNA sequencing approach to systematically characterize *HTT* cryptic polyadenylation in several HD mouse models, human cell lines, and HD postmortem brain tissue. This method, known as 3´-end targeted RNA sequencing or 3TRS, enables an accurate, quantitative, and cost-effective measurement of alternative polyadenylation. We describe an experimental and computational protocol that can also be adapted to additional transcripts and disease models.

We show that activation of *HTT* cryptic polyadenylation is highly selective for the distal polyadenylation site located 7.3kb downstream intron 1 and requires very long, uninterrupted CAG repeat expansions. In humanized YAC128 mice, cryptic *HTT* expression was detected across all inspected brain regions, whereas in knock-in mice it correlated strongly with tissue-specific somatic repeat instability, being most prominent in the striatum. In human cell lines, pure CAG tracts triggered cryptic polyadenylation, while CAA interruptions within the repeats completely suppressed *HTT1a* expression. In HD patient samples, cryptic *HTT* transcripts were detected only in a juvenile-onset fibroblast line carrying an ultralong CAG expansion and in brain regions exhibiting high somatic instability. Together, these data support a model in which long and unstable CAG repeat expansions drive HD pathogenesis by activating *HTT* cryptic polyadenylation. By enabling quantitative comparison of cryptic and canonical *HTT* isoforms, 3TRS provides a robust framework to investigate *HTT1a* biogenesis and to support the development and evaluation of *HTT*-lowering therapeutic strategies.

Main text

Introduction

HD is a progressive neurodegenerative disorder with an autosomal dominant inheritance pattern caused by a CAG trinucleotide expansion in the first exon of the *HTT* gene (1). The repeat threshold to develop HD with full-penetrance is established in n≥40 CAG repeats, although recent studies have demonstrated that only long (>150) somatic repeat expansions gain toxicity in the brain (2, 3). In *HTT*, the CAG repeat expansion is transcribed and translated to a polyQ tract within the huntingtin protein (HTT). HTT can be processed into multiple N-terminal fragments (4). Amongst them, the exon 1 HTT fragment with an expanded polyQ tract is considered the most pathogenic, and its expression is enough to produce the most aggressive HD mouse models (5, 6). Although several proteases that generate HTT N-terminal fragments have been identified, the enzyme responsible for the biogenesis of the smallest and most toxic exon 1 HTT fragment is still undetermined (7–10).

RNA misprocessing is a molecular hallmark of human disease, which involves alterations in RNA splicing and 3´-end processing (11, 12). While RNA mis-splicing is a well-established feature of neurodegeneration (13), the implication of aberrant 3´-end processing in disease has been investigated in less detail. In the context of HD, previous work from the Bates group has shown that expanded CAG repeats above a pathologic threshold lead to incomplete splicing of *HTT* intron 1 to generate an exon 1-intron 1 mRNA transcript. This isoform, known as *HTT1a*, is translated into the highly pathogenic exon 1 HTT fragment (14–16). Using multiple mouse and human HD models, several cryptic polyadenylation sites (pA-site) were described by rapid amplification of cDNA ends (3´RACE). Moreover, RT-PCR, quantitative PCR (qPCR), digital PCR and probe-based assays have been developed to measure *HTT* intron 1 fragments (14–19). Although informative, these assays do not rely on RNA sequencing, which limits the interpretation of the results, especially when low copy numbers of aberrant transcripts are generated (16, 17).

HTT lowering strategies are a major focus of therapeutic development for HD (20), especially those targeting *HTT1a* isoform with antisense RNA molecules (21, 22). To develop succesful therapies, it is crucial to apply quantitative methods that accurately measure the expression of the targeted transcript, also in the context of all *HTT* isoforms. In this work, we performed targeted 3´-end RNA sequencing to measure all different *HTT* transcripts in a sensitive, quantitative and cost-efficient manner. With this approch, we confirmed *HTT*/*Htt* cryptic polyadenylation in several human and mouse HD models. In humanized mouse models and cell lines, we clearly demonstrate that the distal cryptic pA-site, located 7.3kb within intron 1, is the one activated in the presence of an expanded CAG tract. In mouse *Htt*, however, both cryptic pA-sites within *Htt* intron 1 are used to a similar extent, specially upon somatic repeat expansion in the brain. Our data also demonstrate that CAA interruptions within the repeats block the expression of *HTT1a* transcript. In human HD samples, *HTT* cryptic polyadenylation occurs only in the presence of very long CAG tracts in a cell line from a juvenile-onset HD patient, or upon somatic instability in specific brain regions of postmortem tissue. Our data support a model in which very long, uninterrupted CAG-repeat expansions drive HD pathogenicity activating *HTT* intron 1 cryptic polyadenylation and *HTT1a* expression.

Results

HTT distal cryptic poly-A site in intron 1 is specifically activated in the YAC128 mouse model upon CAG repeat expansion

To quantify 3′-end processing events, an in-house protocol for 3′-end targeted RNA sequencing (3TRS) was developed. This method was designed to overcome limitations associated with conventional RNA-sequencing approaches and to provide a straightforward, cost-effective workflow for library preparation interrogating pre-defined 3'-ends. The protocol involves three main steps: (1) reverse transcription of each sample independently with a barcoded oligo-dT, (2) a targeted second strand synthesis step in a pool of RT samples with an oligo containing unique molecular identifiers (UMIs) for unique cDNA count, and (3) PCR amplification with oligos containing Illumina adapter sequences followed by sequencing. Raw sequencing data were analyzed using a custom bioinformatic pipeline implemented in Nextflow and executed within the EPI2ME open-source analysis platform. The pipeline generated tab-separated value

(TSV) files containing duplicated and deduplicated read counts, quality control reports, and BAM files, which can be easily visualized using genome browsers such as the Integrative Genomics Viewer (IGV) (Fig. 1A). Further methodological details are provided in the Materials and Methods section.

To investigate the effect of CAG repeat expansion in human *HTT* 3'-end processing, 3TRS was performed in brain tissue from YAC128 transgenic mouse model and compared to wild-type (WT) mice (23). This model was selected because it expresses the full-length human mutant *HTT* transgene, thereby enabling the study of human *HTT* 3′-end processing in a relevant *in vivo* context. Importantly, YAC128 mice also retain expression of the endogenous mouse *Htt* gene, allowing the simultaneous investigation of 3′-end processing in both the human and mouse transcripts. To distinguish and quantify the different *HTT* and *Htt* isoforms, second strand specific oligonucleotides were designed to selectively amplify five transcript variants (Fig 1B): (i) two previously reported *HTT* transcripts generated by the aberrant activation of a proximal (hCPA1) or a distal (hCPA2) cryptic pA-sites in the human *HTT* intron 1 (15), (ii) two canonical human *HTT* isoforms, generated by the activation of either a proximal (hPA1) or a distal (hPA2) alternative pA-site in the *HTT* 3' UTR (24, 25) and (iii) the full-length mouse transcript, which uses the canonical pA-site in the *Htt* 3′ UTR (mPA). Expression of all 3´-ends was measured in the RNA isolated from the striatum (STR; n=4 WT and n=6 YAC), cortex (CTX, n=4 WT and n=7 YAC) and hippocampus (HPC, n=4 WT and n=6 YAC) of YAC128 mice. The endogenous mouse *Htt* transcript, which uses the canonical pA-site (mPA), was detected in all samples, with comparable expression levels between WT and YAC128 mice. Both canonical human *HTT* isoforms, utilising the proximal (hPA1) and distal (hPA2) pA-sites, were detected exclusively in YAC128 mice carrying the human mutant *HTT* gene. Notably, transcripts using intronic cryptic pA-site were also present only in YAC128 samples. Among these, the distal cryptic pA-site (hCPA2) was consistently detected across all three brain regions analysed, whereas the proximal site (hCPA1) showed minimal or no expression (Fig. 1C).

Sequencing results were further visualised using IGV, which showed the expected sequences from the second strand synthesis oligo to the polyadenylation cleavage site (Fig. 1D, Sup. Fig. 1A). Moreover, 3TRS protocol was specific, as no reads corresponding to other transcripts under study were detected in the libraries prepared with each second strand synthesis oligo (Sup. Fig. 1B). Interestingly, there were no significant differences in the usage of distal cryptic pA-site in human *HTT* (hCPA2) between the different brain regions analyzed, whereas in general a higher expression of canonical human and mouse Huntingtin transcripts was detected in the brain cortex (Sup. Fig. 1C).

Targeted 3´-end sequencing allows multiplexing several mouse and human Huntingtin transcripts

After quantifiying that in human *HTT* gene the distal cryptic poly-A site (hCPA2) is used in the presence of CAG expanded repeats in the YAC128 mouse model, the protocol was optimised to sequence all *HTT/Htt* transcripts simultaneously, so that the number of unique cDNA counts for each pA-site can be compared. To achieve this, second strand cDNA synthesis was performed in a multiplexed reaction combining all second-strand oligonucleotides targeting the predefined pA-sites in the *HTT* and *Htt* genes (Fig. 2A). Importantly, there were no significant differences between the unique cDNA read counts for each target using singleplex or multiplex reactions (Sup. Fig. 2A), which demonstrates that multiple oligos can be mixed in the second strand synthesis reaction to simplify the protocol and get comparable reads of each interrogated pA-site.

Across all three brain regions, activation of distal hCPA2 cryptic pA-site, but not proximal hCPA1, was confirmed in YAC128 mice (Fig. 2B). In multiplexed 3TRS aligned sequencing reads were visualised from a single BAM file which contains the 3´-end fragments from all the interrogated human pA-sites (Sup. Fig. 2B). As in the individual reactions, no significant differences in hCPA2 unique cDNA counts were observed between the different brain regions analysed (Fig. 2C). Since a higher level of canonical human *HTT* and mouse *Htt* was present in the cortex, cryptic *HTT* expression level was calculated relative to all canonical *HTT* isoforms. This analysis showed that the ratio of cryptic hCPA2 versus canonical *HTT* was significantly

155 higher in the striatum compared to cortex (Fig. 2D). *HTT* fragment analysis confirmed that the YAC128
156 brain shows the expected repeat length distribution, with no appreciable somatic repeat instability (Fig 2E).

157 **Both *Htt* cryptic poly-A sites are used in *Hdh*$^{Q111/+}$ knock-in mice**

158 In contrast to YAC128 mice, the *Hdh*$^{Q111/+}$ knock-in (KI) mouse model carries a chimeric human/mouse exon
159 1 with a 111 CAG repeat expansion within the endogenous mouse *Htt* gene (26) (Fig. 3A). Therefore, the
160 CAG expansion occurs in the *Htt* genomic context, resembling more closely HD patients. Previous studies
161 in HD KI mice have described two cryptic poly-A sites in *Htt* intron 1 (mCPA1 and mCPA2), which differ
162 from the human *HTT* cryptic poly-A sites. These sites are located at positions 680 bp and 1145 bp in *Htt*
163 intron 1, respectively (Fig. 3A) (15). In order to investigate activation of cryptic polyadenylation in the mouse
164 *Htt* gene upon CAG repeat expansion, RNA samples from the brain cortex (CTX), hippocampus (HPC),
165 and striatum (STR) of WT (N=4) and *Hdh*$^{Q111/+}$ (N=3) mice were analysed using 3TRS. As in the YAC128
166 model, second strand synthesis oligos were designed to selectively amplify three transcript variants (Fig
167 3B): (i) the mouse canonical poly-A site (mPA), arising from the wild-type and the mutant forms, (ii) the
168 mouse proximal cryptic poly-A site (mCPA1) and (iii) the mouse distal cryptic poly-A site (mCPA2). Based
169 on previous results in the YAC128 model, all second strand oligos were multiplexed in a single reaction
170 (Fig. 3B).

171 Compared to the YAC128 mouse model, sequencing results showed that both *Htt* mCPA1 and mCPA2
172 sites were activated in *Hdh*$^{Q111/+}$ mice upon CAG repeat expansion. Moreover, the KI mice exhibited the
173 highest number of cryptically polyadenylated transcripts in the striatum, whereas the differences obtained
174 in the cortex and the hippocampus of *Hdh*$^{Q111/+}$ mice compared to wild-type did not reach statistical
175 significance (Fig. 3C). When the expression level of each cryptic transcript was corrected against canonical
176 *Htt* expression, we confirmed that cryptic polyadenylation is significantly higher in the striatum (Fig. 3D).
177 Since previous studies have reported that the CAG repeat expansion undergoes tissue-specific somatic
178 instability, most prominently in the striatum (27), we investigated whether the highest activation of cryptic
179 polyadenylation observed in the striatum was associated with the level of somatic instability in that tissue.
180 Indeed, *Htt* fragment analysis showed that cryptic polyadenylation increases upon somatic repeat
181 expansion in the striatum (Fig. 3E). Aligned sequencing reads were visualised in IGV using the mouse
182 genome as reference, which confirmed that cryptic polyadenylation in *Htt* intron 1 is activated in the striatum
183 of *Hdh*$^{Q111/+}$ mice, whereas canonical transcripts were present in all tissues from wild-type and *Hdh*$^{Q111/+}$
184 mice (Sup. Fig. 3B).

185 ***HTT* cryptic polyadenylation in HEK293 HD cell lines**

186 To translate the findings from HD mouse models into a human cellular context, we studied HEK293 cell
187 lines carrying homozygous CAG repeat expansions, with or without CAA interruptions, in the endogenous
188 *HTT* locus (18) (Fig. 4A). Following the same approach as before, a multiplexed library targeting all human
189 *HTT* transcripts (hCPA1, hCPA2, hPA1 and hPA2), was prepared and sequenced (Fig. 4B).

190 Regarding canonical *HTT* expression, no significant changes in the proximal (hPA1) or the distal (hPA2)
191 poly-A sites were observed between WT, CAG-expanded and CAA-interrupted HEK293 cell lines (Fig. 4C).
192 Interestingly, while in the YAC128 HD mouse model the distal canonical poly-A site (hPA2) was more
193 frequently used than the proximal one (hPA1), in HEK293 HD cells the proximal site was more active. These
194 findings align with previous reports indicating that the longer isoform (hPA2) is more abundant in brain
195 tissue, whereas the shorter isoform (hPA1) predominates in peripheral tissues (25). Concerning cryptic
196 polyadenylation, the distal site (hCPA2), but not the proximal one (hCPA1) was again only used in HEK293
197 cells carrying homozygous CAG repeat expansions (N=7). Remarkably, CAA interruptions (N=4) avoided
198 *HTT* cryptic poly-A site usage, to a similar extent as wild-type HEK293 cells (N=3), which also showed no
199 cryptic polyadenylation (Fig. 4C). When the expression level of cryptic *HTT* was corrected with canonical
200 *HTT* expression, it confirmed that cryptic *HTT* expression occurs exclusively in the context of pure CAG
201 repeat tracts (Fig. 4D). These findings suggest that the presence of CAA interruptions may prevent the

aberrant activation of cryptic poly-A sites in the context of expanded CAG repeats, potentially contributing to their protective effect in HD phenotype (28).

In CAG-expanded HEK293 cells repeat length did not correlate with the *HTT1a* level (Sup. Fig. 4A). Long-read sequencing of *HTT* CAG-repeat fragment confirmed the expected motif distribution (Fig 4E). Aligned sequencing reads were visualised in IGV, which confirmed that *HTT* cryptic polyadenylation was exclusive from HEK293-CAG HD cells, whereas both canonical sites were recognized across all the three different cell lines (WT, CAA, and CAG) (Sup. Fig. 4B).

***HTT* cryptic polyadenylation in HD patient samples**

To assess *HTT* cryptic polyadenylation in HD patient-derived samples, we performed 3TRS in primary fibroblasts derived from control (WT, n = 8) and HD patients (HD, n = 17). Control samples contained less than 40 CAG repeats in the *HTT* gene. HD patient-derived fibroblasts were divided in four groups according to their CAG-repeat size: CAG 40-50 (n=13), CAG 56 (n=1), CAG 81 (n=1) and CAG 180 (n=1). Library preparation was performed multiplexing second strand oligos targeting all cryptic and canonical proximal and distal poly-A sites (Fig. 5A). In control and HD fibroblasts, reads mapping the canonical proximal poly-A site were significantly more abundant than reads mapping the distal site, although no differences between control and CAG 40-50 groups were obtained. Regarding cryptic site activation, there were no reads mapping the proximal cryptic poly-A site (hCPA1) in none of the samples, and reads mapping the distal site (hCPA2) were only present in the fibroblast cell line with 180 CAG repeats (Fig. 5B), meaning that ultra-long CAG repeats are required to activate the distal cryptic poly-A site in human fibroblasts. With 3TRS, the unique cDNA reads mapping each of the interrogated 3´-ends can be compared against each other. This analysis revealed that cryptic *HTT* transcripts represent around 5% of all canonical *HTT* transcripts, only in the CAG-180 fibroblast cell line (Fig. 5C). *HTT* fragment analysis confirmed the expected repeat length distribution, shown for some representative samples (Sup. Fig. 5A).

Finally, postmortem brain tissue from five HD patients was analyzed, including cortex (CTX, N=5), hippocampus (HPC, N=2), and striatum (STR, N=2). Using the 3TRS strategy described above for human fibroblasts, canonical *HTT* transcripts were detected in brain samples (Fig. 5D), where a higher abundance of the distal poly-A site was observed compared to human fibroblasts. Regarding cryptic polyadenylation, our results show that only the distal poly-A is active but not in all HD brain patients and regions. Cryptic polyadenylation was only detected in those brain regions with the highest somatic instability indexes (Fig. 5E and Sup. Fig. 5B).

Discussion

In this work we have established a targeted RNA-sequencing strategy to investigate the expression of *HTT* transcripts generated by cryptic polyadenylation in the context of HD. Our results agree with previous studies that initially described *HTT1a* expression by incomplete *HTT* splicing in the presence of expanded CAG repeats (15–17, 29). Compared to previous studies however, which primarily detect *HTT* intron 1 fragments by RT-PCR (18, 19), quantitative PCR (qPCR) (15, 16), digital PCR (17) or probe based-assays (14) as a proxy for *HTT1a* expression, the strategy presented here involves sequencing the 3´-ends of polyadenylated *HTT* transcripts. Our approach (i) provides quantitative information from 3´-end processed mRNA isoforms and avoids quantification of yet unprocessed *HTT* intron 1 molecules; (ii) precisely differentiates between alternative isoforms generated by the usage of nearby poly-A sites; (iii) allows multiplexing of dozens of samples early in the library preparation protocol, adding a sample barcoding step, and (iv) avoids PCR-based amplification biases introducing Unique Molecular Identifiers (UMIs) during library preparation. Finally, the protocol allows multiplexing several oligos in the second strand synthesis reaction, which allows comparing absolute values from alternative isoforms, also a limitation from previously used strategies (14, 16, 17).

Cryptic polyadenylation in *HTT*/*Htt* was first described by 3'RACE (15), a method that confirmed the usage of a poly-A site, but not in a quantitative-manner. Follow-up studies developed quantitative assays, including qPCR (16), digital PCR (17) and a QuantiGene probe-based assay to measure *HTT1a* expression (14). Nonetheless, these assays detect intron 1 sequences using either amplification primers (qPCR) or

251 fluorescent probes (QuantiGene) but are not specific to detect polyadenylated transcripts. Indeed, primers
252 or probes that bind proximal *HTT* intron 1 sequences cannot differentiate between transcripts that use the
253 proximal or the distal poly-A sites, because all distal transcripts contain the proximal intron 1 sequences.
254 Our results using 3TRS clearly demonstrate that in YAC128 transgenic mice, as in all human samples
255 evaluated in this study, *HTT1a* transcript derives exclusively from the activation of the distal poly-A site
256 (hCPA2). Although a slight, but not significant, increase in the number of reads derived from the proximal
257 cryptic site (hCPA1) was detected when performing singleplex 3TRS in the YAC128 mouse model, this
258 difference was no longer present when all second strand oligos were multiplexed, which allows the
259 comparison of absolute values from alternative isoforms. *HTT1a* expression also shows species-specific
260 differences, since in the *Hdh* $^{Q111/+}$ knock-in mice model, which contains the mouse *Htt* intron 1 sequence
261 downstream the expanded CAG repeat, both proximal and distal poly-A sites are used.

262 Our results also have important implications regarding HD pathogenicity. Using 3TRS we demonstrate that
263 activation of *HTT* cryptic polyadenylation occurs in mouse and human brain, in a mechanism that is
264 exacerbated upon the somatic CAG-repeat instability that occurs in specific regions of the brain. In YAC128
265 mice, *HTT* hCPA2 expression was comparable across the striatum, cortex and hippocampus. This could
266 be due to the mutant allele being in a genetically engineered chromosome, or to the presence of
267 interspersed CAA codons along the exon 1 polyQ tract that stabilize the CAG repeat and prevent germline
268 and somatic instability (30, 31). Compared to cortex, however, the striatum showed a slight but significantly
269 higher ratio of cryptic versus canonical *HTT*, which may somehow contribute to the selective striatal
270 neuronal loss reported in this model (23). In *Hdh*$^{Q111/+}$ mice, the striatum clearly showed the highest level
271 of proximal and distal cryptic *HTT*, a finding that correlated with a higher somatic instability in this brain
272 region and may explain the stronger vulnerability of the striatum in this HD model (32, 33). These results
273 underscore the relevance of using HD models where CAG repeat expansions are introduced within the
274 endogenous locus to investigate HD pathology.

275 The data obtained in HEK293-CAG and CAA cell lines support a model where pure CAG repeats are
276 required to activate *HTT* cryptic polyadenylation and *HTT1a* expression. Previuos studies have shown that
277 the presence of CAA interruptions within the CAG repeat tract of the *HTT* gene are associated with a milder
278 HD phenotype (28, 34). Our results suggest that CAA interruptions may exert a protective effect by
279 modifying the regulatory landscape of *HTT* cryptic polyadenylation, reducing *HTT1a* expression. The results
280 obtained in these cell lines, however, should be interpreted with caution. HEK293-CAG lines are
281 homozygous for CAG repeat expansion which could explain why cryptic polyadenylation is detected with
282 only 41 CAG repeats, and without a positive correlation with increasing CAG repeat lengths. We cannot
283 either disregard that the genetic background, the limited number of samples or the non-neural origin of
284 HEK293 cell line influences these results.

285 In HD patient-derived samples, our results indicate that ultra long CAG repeat expansions are required for
286 the activation of *HTT* cryptic poly-A sites. In HD fibroblasts, *HTT1a* expression was detected only in a cell
287 line derived from a juvenile-onset HD patient with 180 CAG repeats. These findings are consistent with
288 previous studies (16, 17), but the use of 3TRS offers additional advantages. First, it enables a direct
289 comparison of cryptic versus canonical read counts, revealing that approximately 5% of all *HTT* transcripts
290 in this cell line correspond to the toxic *HTT1a* isoform. Second, the incorporation of UMIs avoids and
291 corrects PCR amplification bias, and reveals that *HTT1a* transcripts arise from a very small number of
292 unique RNA molecules that generate a disproportionately high number of duplicated sequencing reads.
293 Finally, analysis of human postmortem brain tissue demonstrates that *HTT* distal cryptic poly-A site is
294 activated in HD brain. Importantly, the highest expression of cryptic *HTT* was not detected in the most
295 affected brain regions, such as the striatum, but in those brain regions with the highest somatic instability
296 indexes. This is likely because the striatum undergoes extreme atrophy in HD, losing ~95% of the neurons
297 by late-stage disease (35). Future studies that apply 3TRS to RNA isolated from neuronal subpopulations
298 rather than bulk tissue will help us understand the complex interplay between somatic repeat expansion
299 and *HTT* cryptic polyadenylation. In the meantime, our results support a model where long-somatic DNA
300 repeat expansions acquired through aging drive neurodegeneration in HD, promoting *HTT* cryptic

301 polyadenylation, *HTT1a* RNA and protein expression and disease pathology, in line with recently published
302 studies (2, 3, 36).

303 This study also has important implications for the development of therapeutic strategies targeting *HTT1a*
304 isoform with antisense RNA molecules (21, 22). The succesful evaluation of ongoing and future therapies
305 requires quantative, sensitive, and cost-effective RNA-sequencing-based methods which can
306 simultaneously evaluate the expression of all *HTT* isoforms using standardized protocols that allow
307 multiplexing several samples.

308 In addition, 3TRS could help elucidate the mechanisms driving *HTT1a* formation, which are yet not fully
309 understood. Several studies point to a complex scenario involving splicing factors, RNA polymerase II
310 transcriptional speed, and epitranscriptomic modifications (37–39). Our approach also captures
311 physiological switches in alternative poly-A site usage. According to our data, the short *HTT* 3´UTR isoform
312 (hPA1; 10.3kb) predominates in dividing cells, as observed in HEK293 cells and fibroblasts, whereas the
313 long isoform (hPA2; 13.7kb) is more abundant in brain, as previously reported (24).

314 The implications of aberrant 3′-end processing in human disease remain poorly understood, partly due to
315 technical challenges to detect alternative and cryptic 3´-ends from standard short-read RNA-sequencing
316 data (40). The 3TRS approach presented here can be readily adapted to quantify multiple transcripts
317 generated by alternative polyadenylation across diverse physiological and/or disease scenarios. We
318 provide comprehensive experimental details and a standardized analysis pipeline that requires minimal
319 coding and computational expertise, enabling a straightforward implementation of this method.

320 Materials and Methods

321 Human postmortem brain tissue

322 All experiments in this study were performed in accordance with relevant guidelines and regulations. Human
323 HD brain samples (Sup. Table 1) were obtained from Fundación CIEN, after the approval of the Clinical
324 Research Ethics Committee of the Basque Country (code LBG-NDD-2023-01/23-30). Informed consent
325 was obtained from all subjects involved in the study.

326 Cell culture and RNA extraction

327 RNA samples from wild-type, YAC128 (23) and heterozygous *Hdh*$^{Q111/+}$ knock-in mice (41) were kindly
328 obtained from Dr. Verónica Brito (UB, Universidad de Barcelona). Engineered HEK293T cells, homozygous
329 with different numbers of CAG repeats at the *HTT* locus, were kindly provided by Dr. Marta Olejniczak
330 (Polish Academy of Sciences, Poland) and are listed in Sup. Table 2 (18). Cells were cultured in Dulbecco's
331 modified Eagle's medium (Gibco) supplemented with 10% fetal bovine serum (Teknovas) and 1% penicillin-
332 streptomycin (Gibco) at 37 °C with 5% CO2. Cells were split when they reached 80% confluence and tested
333 for mycoplasma using Venor PCR-based mycoplasma detection kit (MP0025). RNA from cultured cells was
334 isolated using the Maxwell® RSC SimplyRNA Cells Kit (Promega, AS1390) according to the manufacturer's
335 protocol. RNA from human skin fibroblasts was provided by Dr. Veronica Brito (UB; Universidad de
336 Barcelona), obtained as previously described (37). RNA from human brain samples was extracted using
337 the Maxwell® RSC SimplyRNA Tissue Kit (Promega, AS1340) on the Maxwell® RSC instrument, following
338 homogenisation with a pellet pestle in the lysis buffer provided by the kit.

339 *HTT* CAG repeat fragment analysis

340 100 ng RNA was converted to cDNA using the High-Capacity cDNA Reverse Transcription Kit (Applied
341 Biosystems, 4368814) following manufacturer's instructions. CAG repeat sizes were determined using a
342 human-specific PCR assay (42), using the 6-FAM fluorescently labeled forward primer (5'-6FAM-
343 ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC-3') (IDT), and reverse primer (5'-
344 GGCGGCTGAGGAAGCTGAGGA-3'). 50 µl reactions were performed, using KOD Xtreme Hot Start DNA
345 Polymerase (Sigma, 71975-M) according to manufacturer's instructions. Cycling conditions were 2 min at

346 94 ºC, 32 cycles of 10 seconds (s) at 98 ºC, 30 s at 64 ºC and 2 min at 68 ºC, and finished by holding the
347 reaction at 4 ºC. The reaction was then concentrated and purified using Nucleospin Gel and PCR clean-up
348 (Macherey-Nagel, 740609), eluting the samples at 15 µl. Products were resolved using the SeqStudio
349 Genetic Analyzer (Applied Biosystems), and GeneScan 1200 LIZ (Applied Biosystems, 4379950) was used
350 as an internal size standard. CAG repeat size distribution plots were generated using GeneMapper v6
351 software. Somatic CAG expansion indexes were calculated as previously described (43), using a 10%
352 relative peak height threshold cut-off. PCR products containing CAG repeats were sequenced by
353 Plasmidsaurus (Germany, Europe), using the long-read amplicon sequencing service with Oxford
354 Nanopore Sequencing Technology.

355 **Library preparation for 3´-end targeted RNA sequencing (3TRS)**

356 For reverse transcription (RT), 500-1000 ng of each RNA was used as starting material in a total volume of
357 11 µl. For each sample, 1 µl of reverse transcription primer at 250nM containing a sample barcode (Sup.
358 Table 3) and 1µl of dNTPs at 10 mM (NEB, N0447L) were added. After 5 minutes of denaturation at 65 ºC,
359 samples were cooled to 37 ºC and held in a thermocycler. Each sample was reverse transcribed using
360 SuperScript IV reverse transcriptase. A mastermix containing 4 µl of SuperScript IV Buffer, 1 µl of 0.1M
361 DTT, 1 µl of RNAsin® (Promega, N2615) and 1 µl of SuperScript IV reverse transcriptase (Invitrogen
362 18090050) was prepared and pre-warmed at 42 ºC for 2-3 min before being added (7 µl / sample) to the
363 samples held at 37 ºC in the thermocycler. The RT reaction was then performed at 50 ºC for 10 min and
364 the reverse transcriptase enzyme was inactivated at 80 ºC for 10 min and cooled to 25 ºC. After this step,
365 pools containing the RT reaction of six samples were mixed and purified. For purification, 2.5X Purification
366 Beads (Lexogen SKU 022.96) were added, well mixed and incubated for 4-5min. Then, they were allowed
367 to bind to the magnet (for 2-3min until the supernatant is clear), and the supernatant was taken off. The
368 wash step was performed three times with freshly prepared 80% ethanol (500 µl) without taking the tube
369 off the magnet. In the last wash, all the ethanol was removed (adding a spin step) and air-dried for 5 min.
370 Finally, the sample was eluted by adding 50 µl of molecular biology grade water, incubated for 4-5 min, and
371 allowed to bind for 4-5 min in the magnet. Finally, the sample containing the single-stranded cDNA (ss-
372 cDNA) was collected in a new tube.

373 For the second strand synthesis, 15 µl of High-Fidelity Phusion 2X Master Mix (NEB, M0531L); 13 µl of the
374 ss-cDNA and 2 µl of the second strand synthesis oligos specific for our targets of interest (Sup. Table 3)
375 were mixed in a final volume of 30 µl. The initial concentration of the second strand oligo mix was 7.5 µM
376 independently of the number of targets multiplexed in each reaction. Each primer contains a unique
377 molecular identifier (UMI), consisting of a random nucleotide sequence that allows demultiplexing PCR
378 amplified cDNA reads to unique cDNA molecules. Initially, 6-nt long UMIs were employed for the data
379 generated in Figures 1-3. Subsequent experiments were performed with 10-nt long UMIs to increase
380 sequence complexity. The second strand synthesis reaction consists of a 2 min incubation at 98 ºC; 60 s
381 at 55 ºC (expected annealing temperature for the primers) and a 5 min incubation at 72 ºC to complete the
382 second strand synthesis. At this point, samples were held at 10 ºC and purified using 1X purification beads
383 (Lexogen, SKU 022.96), following the manufacturer's instructions.

384 After the second strand synthesis, the double-stranded cDNA (ds-cDNA) was amplified by PCR. A common
385 forward primer containing the Illumina sequencing adapter sequence was used for all the reactions. The
386 reverse primer contained the Illumina sequencing adapter and a 6-nt-long index (i7) that differs for each
387 oligo and allows indexed sequencing of each library (Sup. Table 3). The PCR reaction was set to a 30 µl
388 final volume, containing 15 µl of the Phusion High-Fidelity 2X Master Mix (NEB, M0531L), 1.5 µl of each
389 primer at 10 µM and 12 µl of purified ds-cDNA. The PCR program started at 98 ºC for 30 s, followed by 25
390 cycles of 10 s at 98 ºC, 30 s at 60 ºC and 30 s at 72 ºC. As a final extension step, samples were kept for 5
391 min at 72 ºC and then held at 10 ºC. The size of the PCR product was first assessed by automated
392 electrophoresis with the QIAxcel Advanced System using the QIAxcel DNA Screening kit (QIAGEN). Then
393 it was purified from a 4% E-Gel™ EX Agarose Gel (Invitrogen, G401004) using the E-Gel Power Snap
394 Electrophoresis System (Invitrogen, G8300). In these gels, fragments between 150 bp and 400 bp were
395 selected for band excision. The library was then purified using MiniElute PCR purification columns (Qiagen,
396 28004). Before sequencing, library size and integrity was analysed in the Agilent TapeStation using the

High Sensitivity D1000 ScreenTape Assay (Agilent, 5067- 5584) and library concentration was quantified with the Qubit™ dsDNA High Sensitivity Assay Kit (Invitrogen, Q32851).

3TRS library sequencing and data processing

All 3TRS libraries were sequenced in a NovaSeq X Plus (Illumina) using Novogene pre-made library sequencing service. Sequencing data were analysed using a custom pipeline developed in the laboratory. Quality control of the reads was performed using *FastQC* v0.11.9 (available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc). After that, samples were trimmed and demultiplexed by barcode with *Cutadapt* v4.2 (44), followed by the extraction of the UMI sequences using *UMI-tools* v1.1.2 (45). Reads were then aligned to the human (GRCh38.p13) or the mouse (GRCm39 M30) genome with *STAR* v2.7.10b, incorporating the gene and transcript annotations from GENCODE release v41 or vM30 with the *STAR* v2.7.10b aligner (46) . Only alignments with up to 10 multiple mapping locations were retained, and filtering was applied based on mismatch rate and alignment score relative to read length. Once the BAM files were obtained, duplicated reads were collapsed based on their UMI sequences with *UMI-tools* and indexed with *samtools* v1.16.1 (47). Finally, target regions were quantified using *featureCounts* v2.0.3 (48) from the *Subread* package. A *Nextflow* implementation of this workflow is available on *GitHub*, along with all of its containerized dependencies, with extensive information on how to install and run it (https://github.com/blazquezlab/3TRS_analysis_pipeline).The pipeline was implemented within the EPI2ME user-friendly interface and requires four input files: (1) a FASTQ file containing the raw sequencing data, (2) a threeprime.csv file which contains barcode sequences assigned to each sample, (3) a fiveprime.csv file with the information of the individual libraries pooled in each sequencing experiment and (4) a .saf file containing the chromosome coordinates of the target sequences of interest.

Statistical analysis

Statistical analyses were performed using the GraphPad Prism 10.4.2 software for macOS. Student's t-test was used when analyzing differences between two groups. More than two group comparisons were performed with one-way ANOVA followed by multiple comparisons tests. Dunnett's correction was applied when comparisons were performed against one control group. Tukey´s correction was applied when comparisons were performed between all possible groups. The alpha level for statistical significance was set at < 0.05.

Author Contributions

LB designed and supervised the study in collaboration with AVB, VB, JU, MH, MO and ALM. AVB, MM and OAG performed experiments. AHR, MCH, AVB, MM and LB analyzed data. AM generated the HEK293-CAA modified cell lines and EPN collected animal tissue. AR collected and clinically characterized human postmortem brain tissue. LB oversaw project administration and provided supervision. The initial draft was written by LB and AVB and all authors contributed to review and editing of the manuscript.

Acknowledgments

We are grateful to HD patients, controls and their families for donating biological samples for research. We would like to acknowledge Prof. Gill Bates and Dr. Sandra Fienko (Institute of Neurology, UCL, London) for their initial involvement providing RNA samples that were not included in this work, and Dr. Sonia Vazquez-Sanchez, Dr. Carlos Marinas and Dr. Jone López-Erauskin, from Prof. Cleveland group (UCSD, USA) for critical discussion of this work. We want to acknowledge all members of the 'Neurogenetics, RNA biology and therapies' group at Biogipuzkoa Health Research Institute for shared discussion. We finally want to thank the DIPC Atlas Supercomputing Center and their support team for data storage and technical assistance with computational data analysis.

Competing Interest Statement

The authors declare that they have no conflict of interest.

Funding

This study has been funded by Instituto de Salud Carlos III (ISCIII) through the projects PI19/00468 and PI22/00598 awarded to LB, and co-funded by the European Union and Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas (CIBERNED). This research was also supported by the Spanish Ministry of Science and Innovation, through the project CNS2023-144441 awarded to LB and Ramon y Cajal (RYC2018-024397-I) fellowship. The Education Department of the Basque Government supported this study through the IKUR strategy (NEURODEGENPROT) awarded to ALM and IKERBASQUE fellowship (RF/2019/001), awarded to LB. Funding to AVB and MM came from the Basque Government Doctoral Training Program (PRE_2021_1_0171 and PRE_2023_1_0126).

References

1. MacDonald,M.E., Ambrose,C.M., Duyao,M.P., Myers,R.H., Lin,C., Srinidhi,L., Barnes,G., Taylor,S.A., James,M., Groot,N., *et al.* (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.
2. Aldous,S.G., Smith,E.J., Landles,C., Osborne,G.F., Cañibano-Pico,M., Nita,I.M., Phillips,J., Zhang,Y., Jin,B., Hirst,M.B., *et al.* (2024) A CAG repeat threshold for therapeutics targeting somatic instability in Huntington's disease. *Brain*, **147**, 1784–1798.
3. Handsaker,R.E., Kashin,S., Reed,N.M., Tan,S., Lee,W.-S., McDonald,T.M., Morris,K., Kamitaki,N., Mullally,C.D., Morakabati,N.R., *et al.* (2025) Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease. *Cell*, **188**, 623-639.e19.
4. Landles,C., Sathasivam,K., Weiss,A., Woodman,B., Moffitt,H., Finkbeiner,S., Sun,B., Gafni,J., Ellerby,L.M., Trottier,Y., *et al.* (2010) Proteolysis of Mutant Huntingtin Produces an Exon 1 Fragment That Accumulates as an Aggregated Protein in Neuronal Nuclei in Huntington Disease. *Journal of Biological Chemistry*, **285**, 8808–8823.
5. Pouladi,M.A., Morton,A.J. and Hayden,M.R. (2013) Choosing an animal model for the study of Huntington's disease. *Nat Rev Neurosci*, **14**, 708–721.
6. Davies,S.W., Turmaine,M., Cozens,B.A., DiFiglia,M., Sharp,A.H., Ross,C.A., Scherzinger,E., Wanker,E.E., Mangiarini,L. and Bates,G.P. (1997) Formation of Neuronal Intranuclear Inclusions Underlies the Neurological Dysfunction in Mice Transgenic for the HD Mutation. *Cell*, **90**, 537–548.
7. Lunkes,A., Lindenberg,K.S., Ben-Haïem,L., Weber,C., Devys,D., Landwehrmeyer,G.B., Mandel,J.-L. and Trottier,Y. (2002) Proteases Acting on Mutant Huntingtin Generate Cleaved Products that Differentially Build Up Cytoplasmic and Nuclear Inclusions. *Mol. Cell*, **10**, 259–269.
8. Ratovitski,T., Nakamura,M., D'Ambola,J., Chighladze,E., Liang,Y., Wang,W., Graham,R., Hayden,M.R., Borchelt,D.R., Hirschhorn,R.R., *et al.* (2007) N-Terminal Proteolysis of Full-Length Mutant Huntingtin in an Inducible PC12 Cell Model of Huntington's Disease. *Cell Cycle*, **6**, 2970–2981.
9. Tebbenkamp,A.T.N., Crosby,K.W., Siemienski,Z.B., Brown,H.H., Golde,T.E. and Borchelt,D.R. (2012) Analysis of Proteolytic Processes and Enzymatic Activities in the Generation of Huntingtin N-Terminal Fragments in an HEK293 Cell Model. *PLoS ONE*, **7**, e50750.
10. Bent,M.L. van der, Evers,M.M. and Vallès,A. (2022) Emerging Therapies for Huntington's Disease – Focus on N-Terminal Huntingtin and Huntingtin Exon 1. *Biol.: Targets Ther.*, **16**, 141–160.
11. Scotti,M.M. and Swanson,M.S. (2016) RNA mis-splicing in disease. *Nat Rev Genet*, **17**, 19–32.
12. Boreikaitė,V. and Passmore,L.A. (2023) 3′-End Processing of Eukaryotic mRNA: Machinery, Regulation, and Impact on Gene Expression. *Annu. Rev. Biochem.*, **92**, 199–225.
13. Mehta,P.R. (2023) The era of cryptic exons: implications for ALS-FTD.
14. Fienko,S., Landles,C., Sathasivam,K., McAteer,S.J., Milton,R.E., Osborne,G.F., Smith,E.J., Jones,S.T., Bondulich,M.K., Danby,E.C.E., *et al.* (2022) Alternative processing of human *HTT* mRNA with implications for Huntington's disease therapeutics. *Brain*, **145**, 4409–4424.
15. Sathasivam,K., Neueder,A., Gipson,T.A., Landles,C., Benjamin,A.C., Bondulich,M.K., Smith,D.L., Faull,R.L.M., Roos,R.A.C., Howland,D., *et al.* (2013) Aberrant splicing of *HTT* generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 2366–2370.
16. Neueder,A., Landles,C., Ghosh,R., Howland,D., Myers,R.H., Faull,R.L.M., Tabrizi,S.J. and Bates,G.P. (2017) The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. *Sci Rep*, **7**, 1307.
17. Hoschek,F., Natan,J., Wagner,M., Sathasivam,K., Abdelmoez,A., Einem,B.V., Bates,G.P., Landwehrmeyer,G.B. and Neueder,A. (2024) Huntingtin HTT1a is generated in a CAG repeat-length-dependent manner in human tissues. *Mol Med*, **30**, 36.
18. Dabrowska,M., Ciolak,A., Kozlowska,E., Fiszer,A. and Olejniczak,M. (2020) Generation of New Isogenic Models of Huntington's Disease Using CRISPR-Cas9 Technology. *Int J Mol Sci*, **21**.
19. Tong,H., Yang,T., Liu,L., Li,C., Sun,Y., Jia,Q., Qin,Y., Chen,L., Zhao,X., Zhou,G., *et al.* (2023) Aberrant splicing of mutant huntingtin in Huntington's disease knock-in pigs. *Neurobiol. Dis.*, **187**, 106291.
20. Farag,M., Tabrizi,S.J. and Wild,E.J. (2025) Huntington's disease clinical trials update: October 2025. *J. Huntingt.'s Dis.*, 10.1177/18796397251399751.
21. Papadopoulou,A.S., Alterman,J., Landles,C., Smith,E.J., Conroy,F., Phillips,J., Canibano-Pico,M., Nita,I.M., Osborne,G.F., Iqbal,A., *et al.* (2025) Lowering the HTT1a transcript as an effective therapy for Huntington's disease. *bioRxiv*, 10.1101/2025.06.10.658804.

22. Bragg,R.M., Landles,C., Smith,E.J., Osborne,G.F., Cantle,J.P., Bates,G.P. and Carroll,J.B. (2025) Selective targeting of mutant huntingtin intron-1 improves rescue provided by antisense oligonucleotides. *bioRxiv*, 10.1101/2025.07.29.665998.
23. Slow,E.J., Raamsdonk,J. van, Rogers,D., Coleman,S.H., Graham,R.K., Deng,Y., Oh,R., Bissada,N., Hossain,S.M., Yang,Y.-Z., *et al.* (2003) Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease. *Hum. Mol. Genet.*, **12**, 1555–1567.
24. Lin,B., Rommens,J.M., Graham,R.K., Kalchman,M., MacDonald,H., Nasir,J., Delaney,A., Goldberg,Y.P. and Hayden,M.R. (1993) Differential 3′ polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. *Hum. Mol. Genet.*, **2**, 1541–1545.
25. Romo,L., Ashar-Patel,A., Pfister,E. and Aronin,N. (2017) Alterations in mRNA 3′ UTR Isoform Abundance Accompany Gene Expression Changes in Human Huntington's Disease Brains. *Cell Reports*, **20**, 3057–3070.
26. Wheeler,V.C., White,J.K., Gutekunst,C.-A., Vrbanac,V., Weaver,M., Li,X.-J., Li,S.-H., Yi,H., Vonsattel,J.-P., Gusella,J.F., *et al.* (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Hum. Mol. Genet.*, **9**, 503–513.
27. Lee,J.-M., Pinto,R.M., Gillis,T., Claire,J.C.St. and Wheeler,V.C. (2011) Quantification of Age-Dependent Somatic CAG Repeat Instability in Hdh CAG Knock-In Mice Reveals Different Expansion Dynamics in Striatum and Liver. *PLoS ONE*, **6**, e23647.
28. Wright,G.E.B., Collins,J.A., Kay,C., McDonald,C., Dolzhenko,E., Xia,Q., Bečanović,K., Drögemöller,B.I., Semaka,A., Nguyen,C.M., *et al.* (2019) Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. *Am. J. Hum. Genet.*, **104**, 1116–1126.
29. Gipson,T.A., Neueder,A., Wexler,N.S., Bates,G.P. and Housman,D. (2013) Aberrantly spliced HTT, a new player in Huntington's disease pathogenesis. *RNA Biol.*, **10**, 1647–1652.
30. Pouladi,M.A., Stanek,L.M., Xie,Y., Franciosi,S., Southwell,A.L., Deng,Y., Butland,S., Zhang,W., Cheng,S.H., Shihabuddin,L.S., *et al.* (2012) Marked differences in neurochemistry and aggregates despite similar behavioural and neuropathological features of Huntington disease in the full-length BACHD and YAC128 mice. *Hum. Mol. Genet.*, **21**, 2219–2232.
31. Farshim,P.P. and Bates,G.P. (2018) Mouse Models of Huntington's Disease. *Methods Mol. Biol.*, **1780**, 97–120.
32. Wheeler,V.C., White,J.K., Gutekunst,C.-A., Vrbanac,V., Weaver,M., Li,X.-J., Li,S.-H., Yi,H., Vonsattel,J.-P., Gusella,J.F., *et al.* (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Hum. Mol. Genet.*, **9**, 503–513.
33. Carroll,J.B., Deik,A., Fossale,E., Weston,R.M., Guide,J.R., Arjomand,J., Kwak,S., Clish,C.B. and MacDonald,M.E. (2015) HdhQ111 Mice Exhibit Tissue Specific Metabolite Profiles that Include Striatal Lipid Accumulation. *PLoS ONE*, **10**, e0134465.
34. Consortium,G.M. of H.D. (GeM-H., Lee,J.-M., Correia,K., Loupe,J., Kim,K.-H., Barker,D., Hong,E.P., Chao,M.J., Long,J.D., Lucente,D., *et al.* (2019) CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*, **178**, 887-900.e14.
35. Vonsattel,J.-P., Myers,R.H., Stevens,T.J., Ferrante,R.J., Bird,E.D. and Richardson,E.P. (1985) Neuropathological Classification of Huntington's Disease. *J. Neuropathol. Exp. Neurol.*, **44**, 559–577.
36. Belgrad,J. and Khvorova,A. (2024) More than 185 CAG repeats: a point of no return in Huntington's disease biology. *Brain*, **147**, 1601–1603.
37. Pupak,A., Rodríguez-Navarro,I., Sathasivam,K., Singh,A., Essmann,A., Toro,D.D., Ginés,S., Pinto,R.M., Bates,G.P., Ørom,U.A.V., *et al.* (2024) m6A modification of mutant huntingtin RNA promotes the biogenesis of pathogenic huntingtin transcripts. *EMBO Rep*, 10.1038/s44319-024-00283-7.
38. Neueder,A., Dumas,A.A., Benjamin,A.C. and Bates,G.P. (2018) Regulatory mechanisms of incomplete huntingtin mRNA splicing. *Nat Commun*, **9**, 3955.
39. Mason,M.A., Gomez-Paredes,C., Sathasivam,K., Neueder,A., Papadopoulou,A.-S. and Bates,G.P. (2020) Silencing Srsf6 does not modulate incomplete splicing of the huntingtin gene in Huntington's disease models. *Sci. Rep.*, **10**, 14057.
40. Shah,A., Mittleman,B.E., Gilad,Y. and Li,Y.I. (2021) Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biol.*, **22**, 291.

561 41. Wheeler,V.C., Auerbach,W., White,J.K., Srinidhi,J., Auerbach,A., Ryan,A., Duyao,M.P., Vrbanac,V.,
562 Weaver,M., Gusella,J.F., *et al.* (1999) Length-Dependent Gametic CAG Repeat Instability in the
563 Huntington's Disease Knock-in Mouse. *Hum. Mol. Genet.*, **8**, 115–122.
564 42. Mangiarini,L., Sathasivam,K., Mahal,A., Mott,R., Seller,M. and Bates,G.P. (1997) Instability of highly
565 expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nat. Genet.*, **15**, 197–
566 200.
567 43. Lee,J.-M., Ramos,E.M., Lee,J.-H., Gillis,T., Mysore,J.S., Hayden,M.R., Warby,S.C., Morrison,P.,
568 Nance,M., Ross,C.A., *et al.* (2012) CAG repeat expansion in Huntington disease determines age at
569 onset in a fully dominant fashion. *Neurology*, **78**, 690–695.
570 44. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads.
571 *EMBnetJ.*, **17**, 10–12.
572 45. Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular
573 Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
574 46. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and
575 Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
576 47. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and
577 Subgroup,1000 Genome Project Data Processing (2009) The Sequence Alignment/Map format and
578 SAMtools. *Bioinformatics*, **25**, 2078–2079.
579 48. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for
580 assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

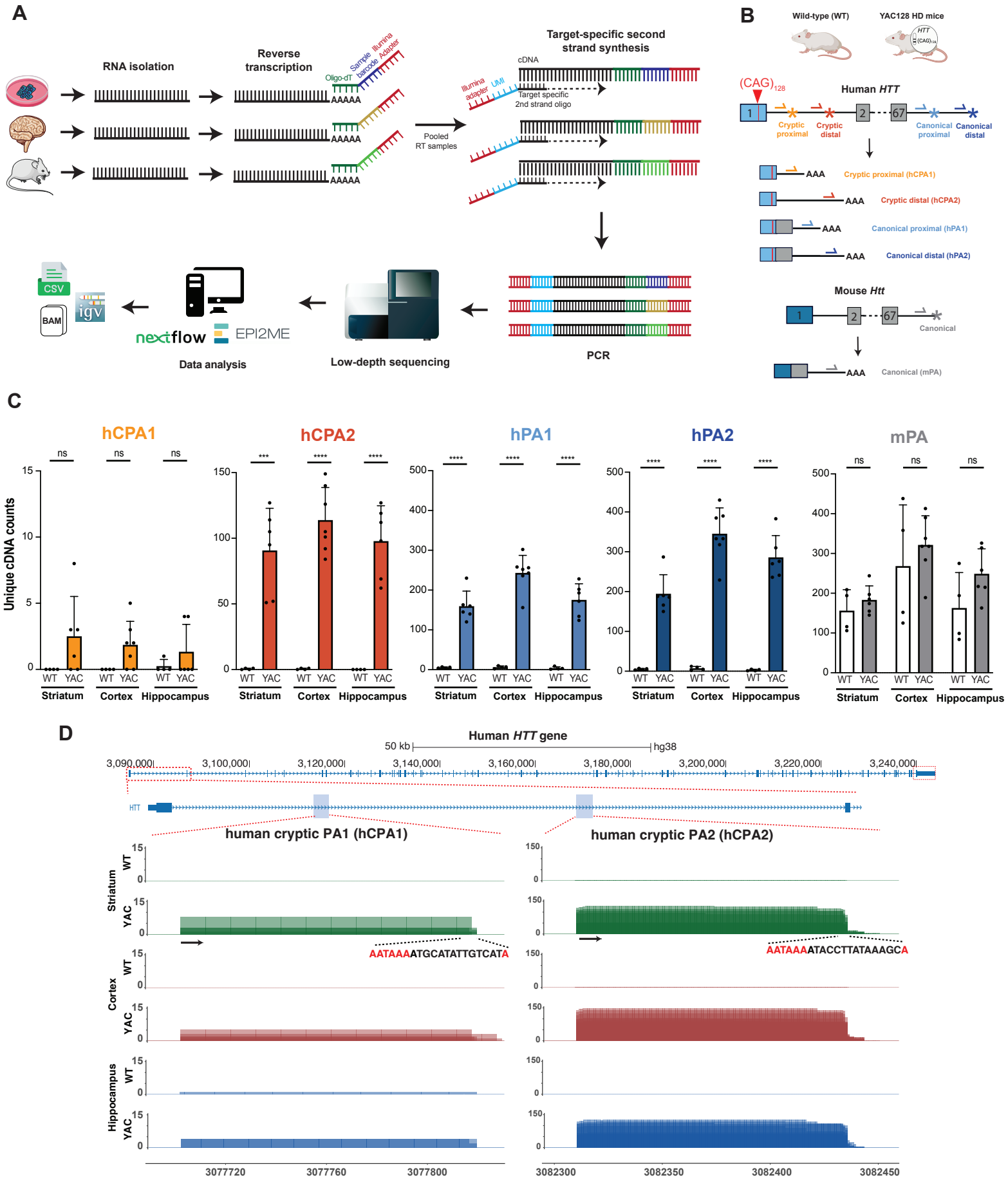

Figure 1: *HTT* RNA processing in the YAC128 mouse model. A. Schematic representation of the 3TRS method developed in this study. **B.** Representation of the second strand oligo design strategy used to selectively amplify the 3´-end of five individual *HTT/Htt* transcripts. **C.** Unique cDNA read counts of cryptic and canonical human *HTT* and mouse *Htt* transcripts in the three different brain regions analyzed in the WT and the YAC128 HD mouse model. Data are presented as mean + SD. Statistical significance was calculated using an unpaired Student´s t-test: ns (non-significant; P ≥ 0.05), *** (P < 0.001), and **** (P < 0.0001). **D.** Sashimi plots of WT and YAC128 mouse brain regions showing the reads mapping the expected *HTT* 3´-ends. The arrow indicates the location of the oligonucleotide used for the second strand synthesis. The sequence at the end of the read indicates the polyadenylation signal (PAS) and the cleavage sites in red.

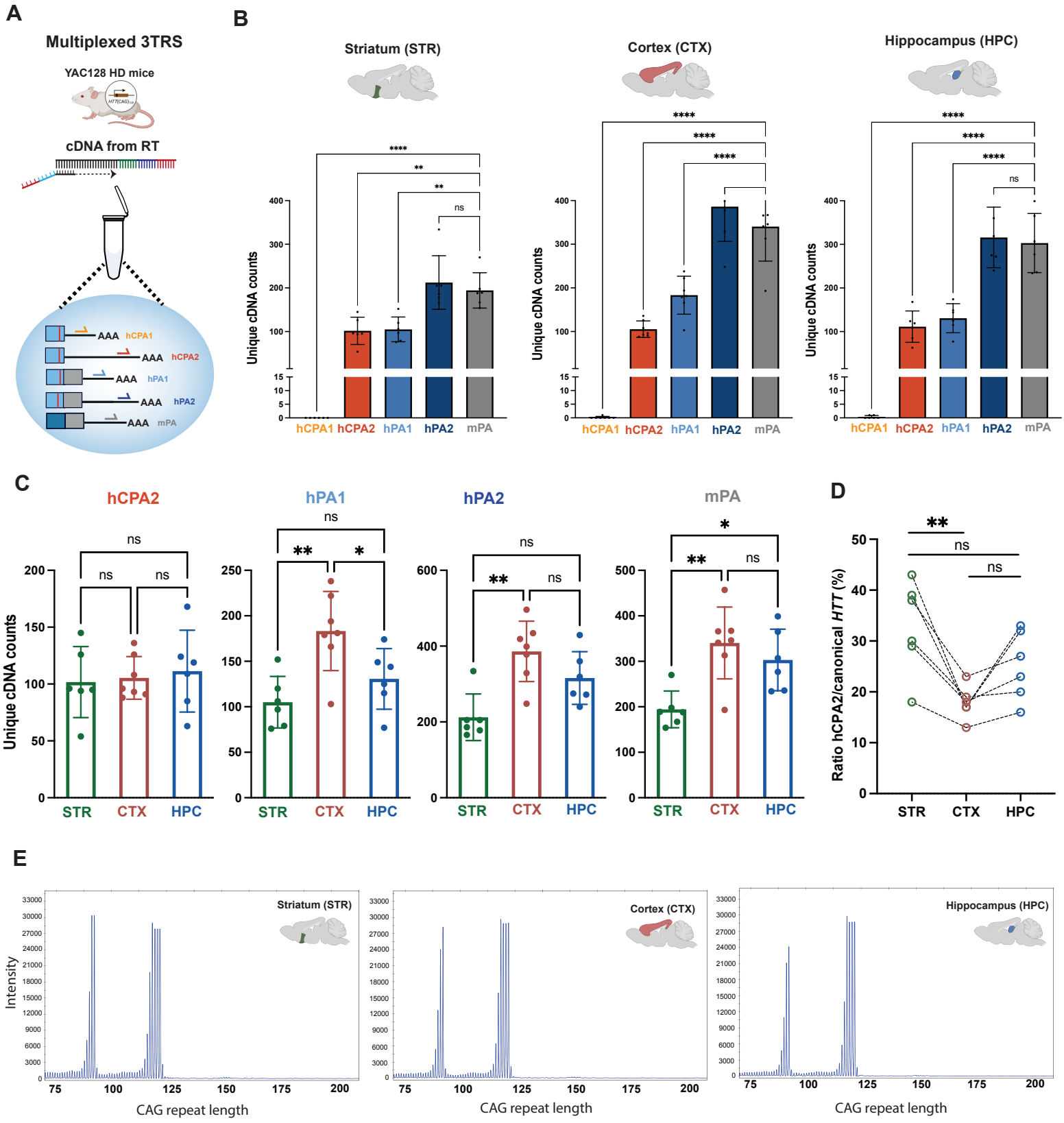

Figure 2: Multiplexed 3TRS in the YAC128 mouse model. A. Representation of the second strand synthesis reaction multiplexing all oligonucleotides targeting the predefined poly-A sites in the *HTT/Htt* transcripts. **B.** Unique cDNA read counts for all interrogated *HTT/Htt* 3´-ends across the three brain regions analyzed in the YAC128 mouse model. Data are presented as mean + SD. Statistical significance was calculated with a one-way ANOVA test followed by Dunnett´s multiple comparisons analysis: ns (non-significant; P ≥ 0.05), ** (P ≤ 0.01), *** (P ≤ 0.001) and **** (P ≤ 0.0001). **C.** Unique cDNA read counts of human *HTT* and mouse *Htt* transcripts in the three different brain regions analyzed in the YAC128 HD mouse model. Data are presented as mean + SD. Statistical significance, indicated as ns (non-significant; P ≥ 0.05), * (P < 0.05) and ** (P ≤ 0.01), was calculated with a one-way ANOVA test followed by Tukey´s multiple comparisons analysis. **D.** Ratio of cryptic hCPA2 versus all canonical *HTT* isoforms in the striatum (STR), cortex (CTX) and hippocampus (HPC) of YAC128 transgenic mice. Data are presented as individual values. Statistical significance, indicated as ns (non-significant; P ≥ 0.05) and ** (P ≤ 0.01), was calculated using a one-way ANOVA test followed by Tukey´s multiple comparisons analysis. **E.** Length distribution of *HTT* CAG tract in the striatum, cortex or hippocampus of one representative YAC128 mouse.

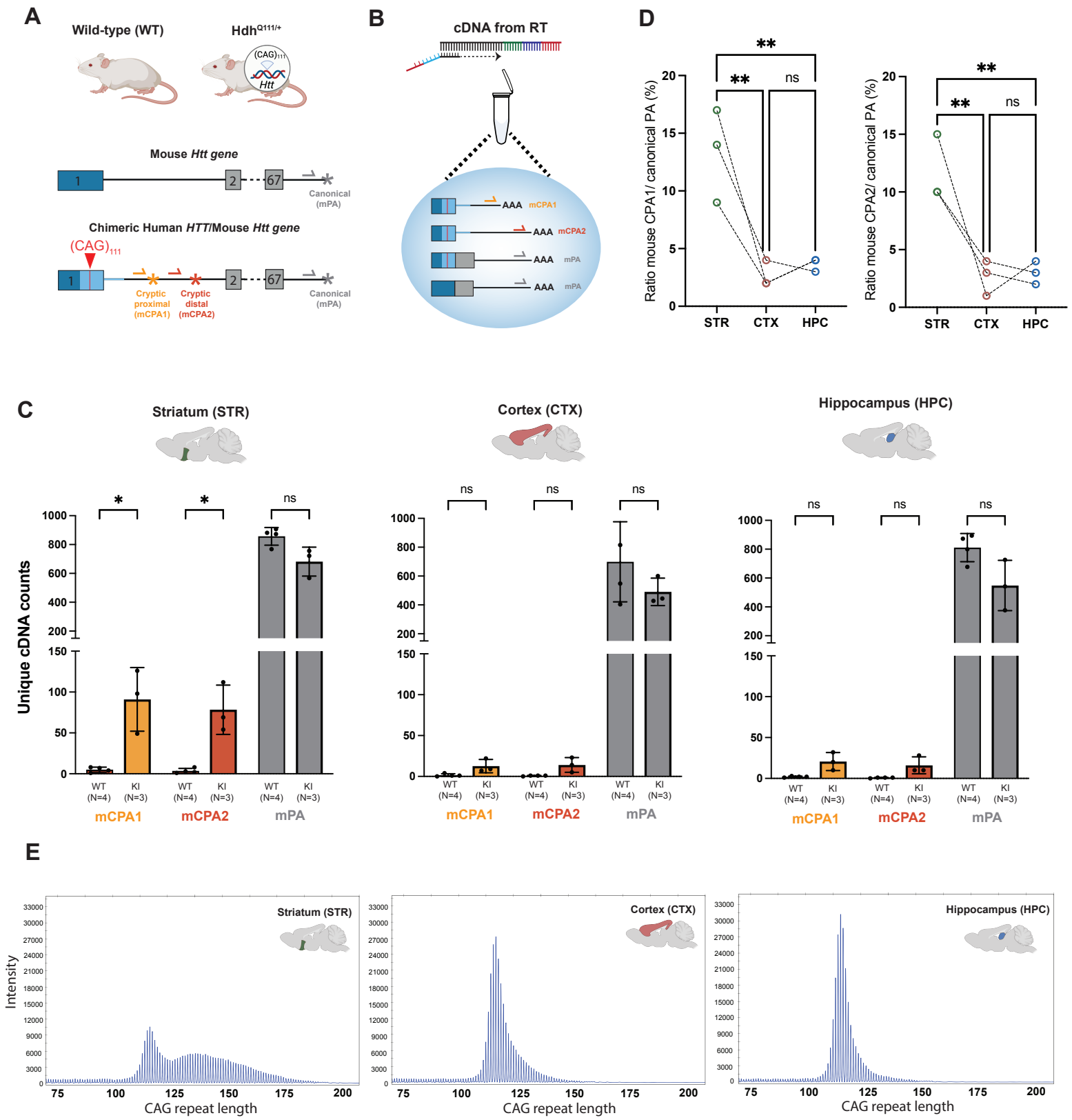

Figure 3: *Htt* RNA processing in the *HdhQ111/+* mouse model. A. Representation of the second strand oligo design strategy used to selectively amplify the 3´-end of three mouse *Htt* transcripts. **B.** Representation of the second strand synthesis reaction multiplexing all oligonucleotides targeting the predefined poly-A sites in the *Htt* transcripts. **C.** Unique cDNA read counts of cryptic and canonical mouse *Htt* transcripts in the three different brain regions analyzed in the WT and the *HdhQ111/+* HD mouse model. Data are presented as mean + SD. Statistical significance was calculated using an unpaired Student´s t-test: ns (non-significant; P ≥ 0.05) and * (P < 0.05). **D.** Ratio of cryptic versus canonical *Htt* isoforms in the striatum (STR), cortex (CTX) and hippocampus (HPC) of *HdhQ111/+* mice. Data are presented as individual values. Statistical significance is indicated as ns (non-significant; P ≥ 0.05) and ** (P ≤ 0.01), after calculation with a one-way ANOVA test followed by Tukey´s multiple comparisons analysis. **E.** Length distribution of *Htt* CAG tract in the striatum, cortex or hippocampus of one representative *HdhQ111/+* mouse.

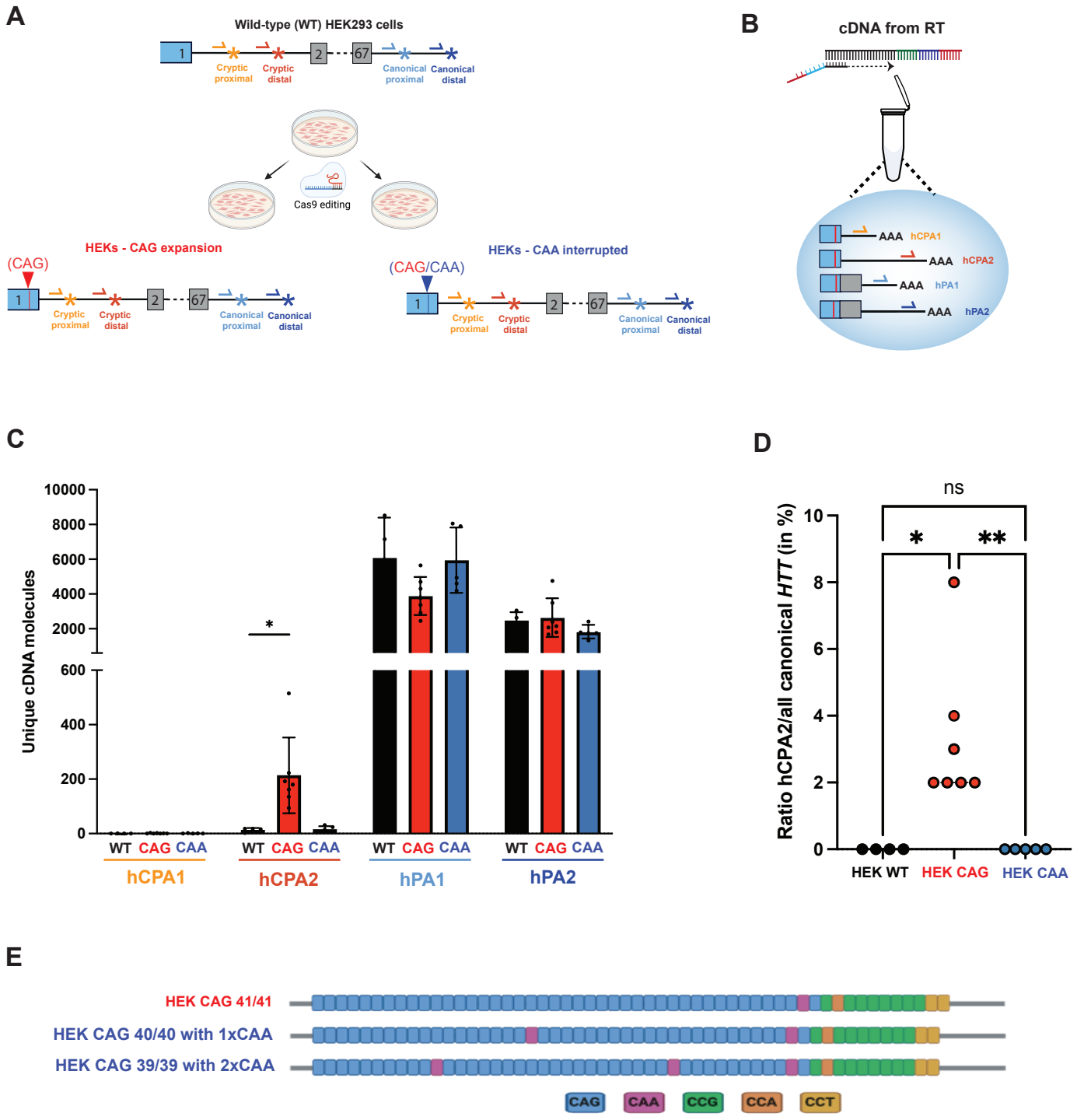

Figure 4: *HTT* cryptic polyadenylation in HEK293 HD cell lines. A. Image of the *HTT* gene in the three HEK293 cell lines analyzed in this study, highlighting the second strand oligos used to amplify the 3´-end of proximal and distal cryptic and canonical *HTT* transcripts. **B.** Representation of the second strand synthesis reaction multiplexing all second strand oligonucleotides targeting all *HTT* transcripts. **C.** Unique cDNA read counts of cryptic and canonical *HTT* transcripts in the three HEK293 cell lines under study. Data are presented as mean + SD. For each *HTT* transcript, statistical significance was calculated using a one-way ANOVA test followed by Dunnett´s multiple comparisons analysis. Only statistically significant results are indicated as * ($P \leq 0.05$). **D.** Ratio of cryptic versus canonical *HTT* isoforms in the three investigated HEK293 cell lines. Data are presented as individual values. Statistical significance is indicated as ns (non-significant; $P \geq 0.05$), * ($P < 0.05$) and ** ($P \leq 0.01$), after calculation with a one-way ANOVA test followed by Tukey´s multiple comparisons analysis. **E.** Motif and length configuration of *HTT* CAG tract in one representative HEK293-CAG cell line and two HEK293-CAA cell lines.

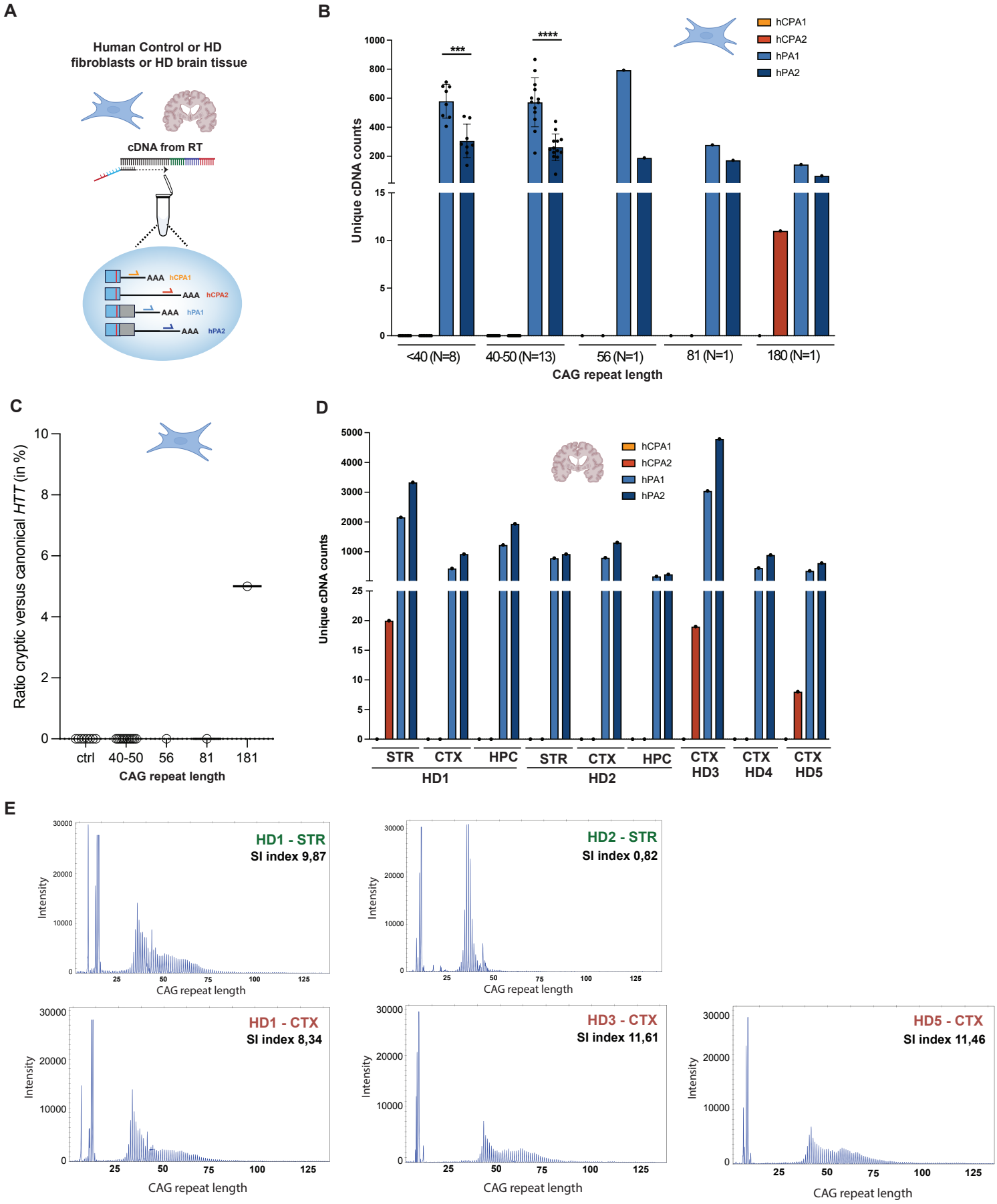

Figure 5: *HTT* cryptic polyadenylation in human HD samples. A. Representation of the second strand synthesis reaction multiplexing all second strand oligonucleotides targeting all *HTT* transcripts, starting from RNA isolated from fibroblasts or postmortem brain tissue. **B.** Unique cDNA read counts of cryptic and canonical *HTT* transcripts in the human fibroblasts under study. Data are presented as mean + SD, whenever more than one sample was available in the group. Statistical analysis for the read counts of canonical *HTT* transcripts was calculated using a paired Student´s t-test: *** (P<0.001), and **** (P<0.0001). **C.** Ratio of cryptic versus canonical *HTT* isoforms in control and HD fibroblasts. **D.** Unique cDNA read counts of cryptic and canonical *HTT* transcripts in the striatum (STR; N=2), cortex (CTX; N=5) and hippocampus (HPC; N=2) of human HD post-mortem brain tissue. Data are represented as individual values. **E.** Length distribution of *HTT* CAG tract in five representative brain regions with or without cryptic *HTT* expression. The somatic instability (SI) index is indicated for each sample.