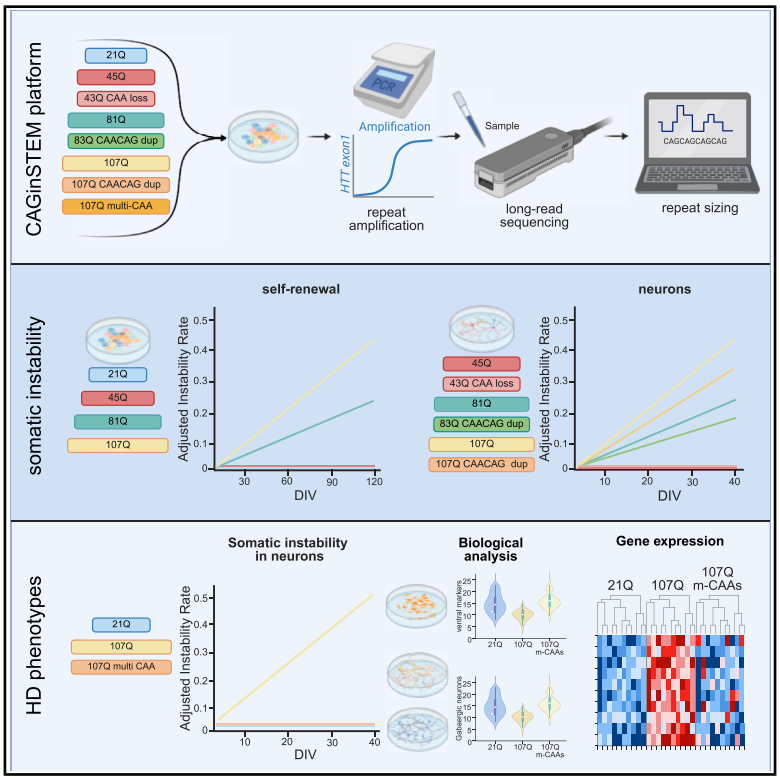
Cell Reports

Resource

A human CAGinSTEM platform for decoding HTT repeats' somatic instability links CAG interruption to HD pathology in neurons

Graphical abstract



Highlights

- CAGinSTEM, a CRISPR-engineered stem cell platform to model HTT repeat instability
- Long-read sequencing enables accurate, sequence-resolved HTT repeat sizing
- The CAGinSTEM platform recapitulates human HD haplotype instability *in vitro*
- Multiple CAA interruptions abolish CAG instability and rescue neuronal HD pathology

Authors

Martina Zobel, Gianluca Damaggio, Maria Lidia Mignogna, ..., Andrea Scolz, Raffale Iennaco, Elena Cattaneo

Correspondence

dario.besusso@unimi.it (D.B.),
elena.cattaneo@unimi.it (E.C.)

In brief

Zobel et al. developed a CRISPR-engineered human stem cell platform to study Huntington's disease repeat instability. Using long-read sequencing, they demonstrated that CAG repeat expansion occurs in both proliferating cells and neurons, mirroring clinical observation. Introducing multiple CAA interruptions completely stabilized the repeat and reversed disease-related cellular phenotypes *in vitro*.

Zobel et al., 2025, Cell Reports *44*, 116685
December 23, 2025
https://doi.org/10.1016/j.celrep.2025.116685

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

A human CAGinSTEM platform for decoding HTT repeats' somatic instability links CAG interruption to HD pathology in neurons

Martina Zobel,[1,2,15] Gianluca Damaggio,[1,2,3,15] Maria Lidia Mignogna,[1,2,10,15] Dario Besusso,[1,2,15,*] Davide Scalzo,[1,2] Andrea Cossu,[1,2,11] Camilla Trovesi,[1,2,12] Mariacristina Crosti,[2] Francesco Cortina,[1,2] Ilaria Campus,[1,2,13] Giulio Formenti,[1,2,14] Saveria Mazzara,[2,4] Francesco Gregoretti,[5] Laura Antonelli,[5] Gennaro Oliva,[5,6] Chiara Zuccato,[1,2] Vincenza Colonna,[7,8] Paola Conforti,[1,2] Matteo Cereda,[1,2] Riccardo Lorenzo Rossi,[2,9] Simone Maestri,[1,2] Andrea Scolz,[1,2] Raffale Iennaco,[1,2] and Elena Cattaneo[1,2,16,*]

[1]Laboratory of Stem Cell Biology and Pharmacology of Neurodegenerative Diseases, Department of Biosciences, University of Milan, Milan 20122, Italy
[2]Istituto Nazionale Genetica Molecolare, Romeo ed Enrica Invernizzi, Milan 20122, Italy
[3]University of Naples Federico II, Naples, Italy
[4]Department of Computing Sciences and Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, Milan, Italy
[5]National Research Council, Institute for High Performance Computing and Networking, Naples 80131, Italy
[6]CSCS - Swiss National Supercomputing Centre, Via Trevano 131, Lugano 6900, Switzerland
[7]National Research Council, Institute of Genetics and Biophysics, Naples 80131, Italy
[8]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
[9]IFOM ETS - The AIRC Institute of Molecular Oncology, Milan 20139, Italy
[10]Present address: Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan 20132, Italy
[11]Present address: Evotec International GmbH, Göttingen 37079, Germany
[12]Present address: Axxam, Bresso-Milan 20091, Italy
[13]Present address: *In Vivo* Laboratory, CompLife Italia Srl, Garbagnate Milanese, Italy
[14]Present address: The Rockefeller University, New York, NY 10065, USA
[15]These authors contributed equally
[16]Lead contact
*Correspondence: dario.besusso@unimi.it (D.B.), elena.cattaneo@unimi.it (E.C.)
https://doi.org/10.1016/j.celrep.2025.116685

SUMMARY

Somatic CAG instability in the mutant Huntingtin (*HTT*) gene is increasingly recognized as a key hallmark of Huntington's disease (HD). Using our novel human CAGinSTEM platform, we manipulated *cis* genetic elements influencing instability in human HD neurons, monitoring repeat length. Quality-controlled CRISPR-engineered stem cells with increasing CAG lengths and clinical haplotypes were analyzed using third-generation sequencing. Our findings link interruptions in the CAG repeat, especially the loss or duplication of the penultimate CAA of canonical alleles, to significant instability modulation. Notably, four internal CAA interruptions completely abolish CAG instability, reversing HD phenotypes such as altered striatal fate acquisition and nuclear disorganization. This platform highlights the role of *cis* modifiers, emphasizing the direct influence of *HTT* DNA repeat composition on CAG instability and providing a robust framework for modeling *HTT* repeat instability *in vitro*.

INTRODUCTION

Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder caused by the expansion of the CAG trinucleotide repeat tract within the exon 1 of the *Huntingtin* (*HTT*) gene.[1] This expansion results in the production of a mutant HTT protein with an elongated N-terminal poly-glutamine (polyQ) stretch. Ultimately, this leads to cell-type-specific neuronal dysfunction, primarily affecting striatal medium spiny projection neurons (MSNs) and cortical neurons, causing their demise.[2] Although disease-modifying therapies are not yet available, gene silencing[3] and cell replacement approaches[4–6] are currently under investigation.

The length of the CAG repeat tract plays a crucial role in determining the onset and severity of the disease, with longer repeats associated with earlier onset and more rapid disease progression.[7,8] However, the CAG size alone accounts for only about 60% of the variability in age of onset (AOO) among patients.[9] More recent investigations have demonstrated that not only the number but also the purity of the CAG repeat influences HD onset.[10–13]

In most human alleles (>95%), the CAG repeat tract carries a characteristic CAA interruption in the second-to-last repeat [(CAG)n-CAA-CAG]. The loss of this CAA is linked to earlier AOO [single loss of interruptions, CAA-loss, producing an uninterrupted (CAG)n], while the duplication of the terminal

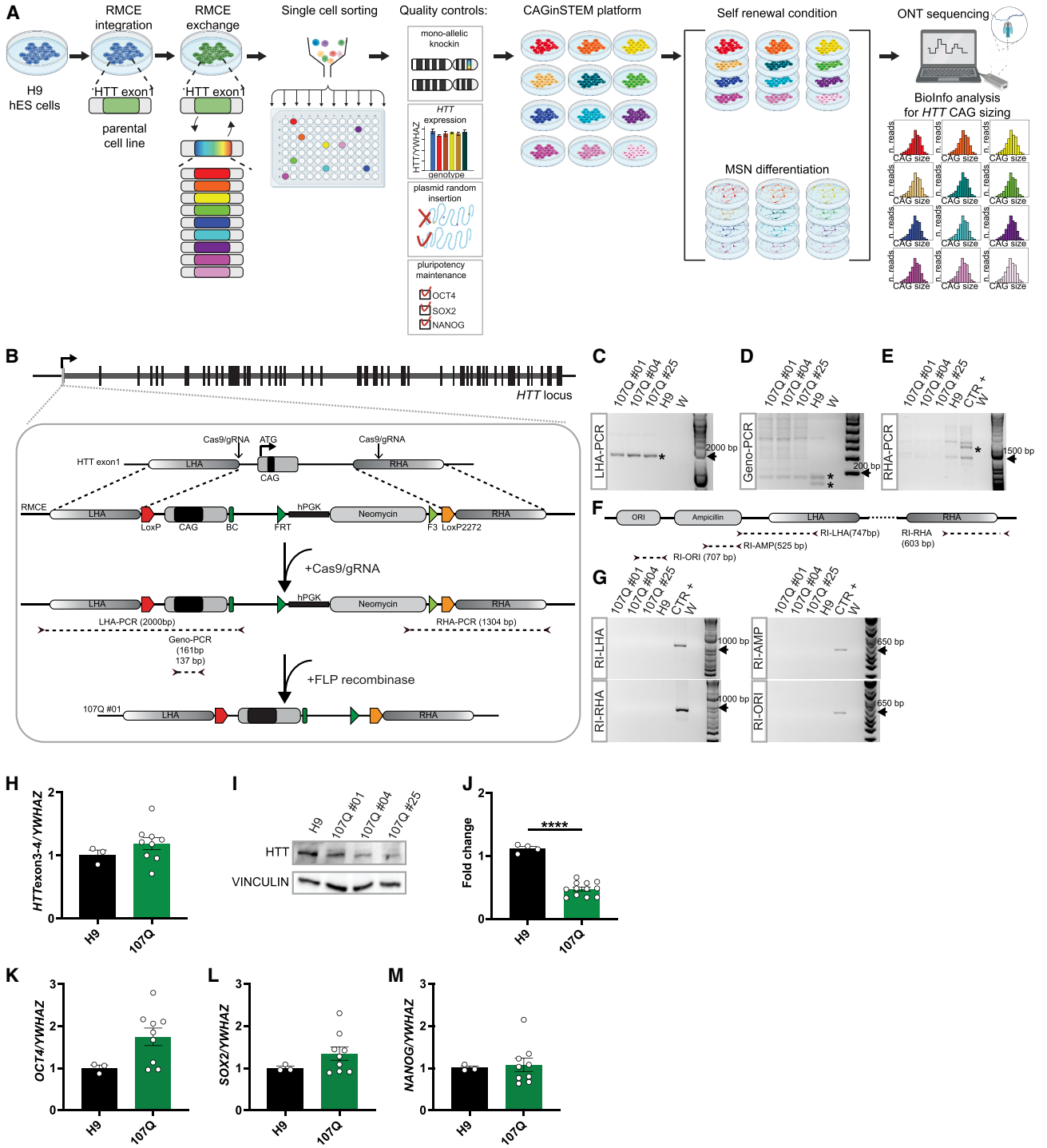


Figure 1. Generation and characterization of RMCE-parental cell line

(A) Schematic representation of the experimental strategy to generate the CAGinSTEM platform. H9 hESCs are edited to generate an RMCE parental cell line that is used to then exchange multiple HTT-exon 1 variants. Single clones carrying desired HTT-exon 1 modification are isolated following FACS sorting and subjected to quality controls to verify monoallelic knockin editing, the absence of plasmids' random insertions into the whole genome, comparable HTT expression to H9 parental cell lines, and pluripotency retention. The isolated clones are maintained in self-renewal and differentiated into striatal neurons, and the CAG instability of HTT-exon 1 is analyzed by Oxford Nanopore Technologies (ONT). A bioinformatic pipeline is developed to account for HTT CAG size.

(B) Editing strategy to generate RMCE-H9 cell lines corresponding to clones: 107Q #01, 107Q #04, and 107Q #25 carrying (CAG)105-111-CAACAG. BC, barcode. LHA and RHA are left and right homology arms, respectively. Segmented lines surrounded by black arrows indicate PCR amplicons.

(legend continued on next page)

CAA-CAG tract [CAACAG-dup, (CAG)n-CAA-CAG-CAA-CAG] is associated with delayed AOO when adjusted for uninterrupted CAG count.[10–13] However, using MiSeq sequencing, the recent GeM-HD consortium study revealed a different and more complex scenario, also depending on whether somatic instability is measured in blood or brain.[13]

Furthermore, postmortem brain tissues of individuals with HD exhibit mosaicism in the size of CAG repeats, with a bias toward expansions, particularly evident in post-mitotic neurons of the striatum and cerebral cortex.[14] This phenomenon has also been observed in animal models of HD.[15] Additionally, a new transgenic mouse model with 120 uninterrupted CAG repeats shows striatum-specific HD pathology and transcriptional dysregulation, which are strongly mitigated in the CAA-interrupted bacterial artificial chromosome (BAC)-HD model with a matched polyQ stretch, further confirming the crucial role of DNA repeat purity beyond glutamine content.[16,17]

Building on these observations, the current hypothesis on HD pathogenesis suggests that the germline-inherited CAG repeat undergoes somatic expansion, also in post-mitotic neurons, persisting throughout an individual's lifetime until it exceeds a cell type-specific toxicity threshold. Specifically, recent works reported cell-type-specific somatic mosaicism in brain tissues postmortem from patients with HD, with MSNs and cholinergic interneurons showing the highest levels of somatic expansions. Notably, extreme somatic expansions were identified in a small fraction of MSNs only, and these events were directly linked to transcriptional dysregulation.[18,19] Such extreme expansions, above 150 CAG repeats, are thought to be responsible for differential vulnerability, initiating neuronal dysfunction and ultimately leading to cell death.[18–21] However, the exact link between the CAG nucleotide sequence and its expansion dynamics, as well as the temporal pattern of CAG repeat changes in neurons and their role in disease pathology, remains elusive.

To address this knowledge gap, we have developed a novel CRISPR-engineered human stem cell platform comprising multiple cell lines harboring varying sizes and repeat compositions, named "CAGinSTEM platform." Paired with third-generation long-read sequencing by Oxford Nanopore Technologies (ONT), this platform enables longitudinal monitoring of CAG repeat instability at single-molecule resolution in differentiated neurons. Our approach utilizes recombinase-mediated cassette exchange (RMCE) to install a modified exon 1 into the endogenous *HTT* locus, leveraging the properties of human embryonic stem cells (hESCs) to study CAG dynamics in disease-relevant post-mitotic neurons. This strategy allows for the quantification of the CAG repeat length changes over time, providing a unique opportunity to study the dynamics of CAG repeat instability in a controlled cellular environment. Our findings further support a direct link between somatic instability and CAG tract purity, with instability increasing as the tract lengthens. Introducing multiple central CAA interruptions—never observed in patients—completely abolishes somatic instability in both proliferating cells and post-mitotic neurons, while also reversing HD *in vitro* phenotypes, including MSN density and nuclear epigenetic organization. This validates the model's suitability for studying CAG instability modifiers and their mechanisms. We conclude that CAG repeat purity directly affects CAG instability over time, and that inserting multiple CAAs and preventing the formation of a pathological pure CAG stretch mitigates HD-associated neuronal pathology.

RESULTS

Generation of the human CAGinSTEM platform carrying increasing HTT CAG lengths

To investigate the impact of CAG length on repeat instability, we employed CRISPR-assisted gene targeting to generate CAGinSTEM, a human stem cell-based platform for *in vitro* repeat instability modeling. This platform includes a diverse array of validated hESC lines with an identical genetic background, differing only by the specific *cis* modifier under examination (Figure 1A). Initially, we established a parental cell line carrying a monoallelic RMCE flanking *HTT* exon 1 (Figures 1A–1C), creating an efficient model for introducing any *HTT* exon 1 variants. Using this system, we successfully established a $(CAG)_{105}$-CAACAG/107Q RMCE parental cell line by integrating the cassette into a single allele of the endogenous *HTT* locus (Figure 1B). The RMCE contains the modified exon 1, followed by a unique genetic barcode and a neomycin resistance cassette for downstream isolation of recombinant clones. Our workflow (Figures 1A and 1B) included editing confirmation by PCR and removal of the antibiotic-resistance cassette to avoid confounding effects (Figures 1B and S1A).

Due to the intrinsic instability of the CAG tract, upon exon 1 exchange, we isolated three clones with repetitions of 105, 107, and 111 uninterrupted CAGs [$(CAG)_{105\text{-}111}$-CAACAG], referred to as 107Q clones (Figure 1B; Table 1). The monoallelic knockin of the expected CAG length in 107Q clones was validated using two PCR methods (left homology arm [LHA]-PCR, Figures 1B and 1C; and genotype PCR [geno-PCR], Figures 1B and 1D).

(C) LHA-PCR to assess monoallelic knockin. * PCR product of 2,000 bp.
(D) Geno-PCR to assess monoallelic knockin. * PCR product of 161 bp (H9 with CAG25CAACAG/27Q) and of 137 bp (H9 with CAG17CAACAG/19Q).
(E) PCRs to assess the excision of the antibiotic selection cassette. * PCR product at 1,304 bp.
(F) Schematic representation of donor plasmid elements outside the homology arms. Segmented lines surrounded by black arrows indicate PCR amplicons.
(G) PCRs to assess plasmid random insertions (RI) into the genome (RI-LHA 747 bp, RI-RHA 603 bp, RI-AMP 525 bp, and RI-ORI 707 bp).
(H) RT-qPCR analysis of *HTT*. *YWHAZ* is used as the housekeeping gene.
(I) Western blot of HTT. Vinculin is used as the housekeeping gene.
(J) Quantification of HTT total protein expression in relative pixel intensity normalized on the intensity of the vinculin band.
(K–M) RT-qPCR analysis of *OCT4* (K), *SOX2* (L), and *NANOG* (M). *YWHAZ* is used as the housekeeping gene.
For all agarose gels, 1 kb plus ladder is used as the molecular weight marker, and arrows indicate molecular weight. W (water) and CTR+ are the negative and the positive controls, respectively.
Data are presented as the mean ± SEM, dots represent different culturing passages of a clone, unpaired Student's *t* test is applied to compare 107Q clones to H9 cells, and statistical analysis is reported in Table S1. ****$p < 0.0001$.

Table 1. Allele structure of isolated clones.

Expected CAG length and/or haplotype	No. of screened clones	No. of validated clones	Original 107 Q parental cell line	Resulted CAG length and/or haplotype	Nomenclature
$(CAG)_{105}$CAACAGCCGCCA (CCG)n	48	3	N/A	Clone #25 $(CAG)_{111}$-CAACAG Clone #01 $(CAG)_{105}$-CAACAG Clone #04 $(CAG)_{107}$-CAACAG	107Q
$(CAG)_{19}$CAACAGCCGCCA (CCG)n	164	3	107Q #1 107Q #4 107Q #1	Clone #06 $(CAG)_{19}$-CAACAG Clone #14 $(CAG)_{19}$-CAACAG Clone #20 $(CAG)_{19}$-CAACAG	21Q
$(CAG)_{43}$CAACAGCCGCCA (CCG)n	129	3	107Q #25 107Q #25 107Q #25	Clone #12 $(CAG)_{43}$-CAACAG Clone #51 $(CAG)_{43}$-CAACAG Clone #60 $(CAG)_{43}$-CAACAG	45Q
$(CAG)_{79}$CAACAGCCGCCA (CCG)n	102	3	107Q #25 107Q #1 107Q #4	Clone #06 $(CAG)_{79}$-CAACAG Clone #08 $(CAG)_{78}$-CAACAG Clone #47 $(CAG)_{79}$-CAACAG	81Q
$(CAG)_{107}$CCGCCA(CCG)n	32	2	107Q #1 107Q #1	Clone #08 $(CAG)_{108}$ Clone #12 $(CAG)_{113}$	107Q CAA-loss
$(CAG)_{105}(CAACAG)_2$CCG CCA(CCG)n	31	3	107Q #1 107Q #1 107Q #1	Clone #12 $(CAG)_{105}$-$(CAACAG)_2$ Clone #14 $(CAG)_{105}$-$(CAACAG)_2$ Clone #26 $(CAG)_{105}$-$(CAACAG)_2$	107Q CAACAG-dup
$[(CAG)_{21}CAA]_4(CAG)_{17}$CAA CAGCCGCCA(CCG)n	31	3	107Q #1 107Q #1 107Q #1	Clone #01 $[(CAG)_{21}CAA]_4(CAG)_{17}$CAACAG Clone #14 $[(CAG)_{21}CAA]_4(CAG)_{17}$CAACAG Clone #16 $[(CAG)_{21}CAA]_4(CAG)_{17}$CAACAG	107Q multi-CAAs
$(CAG)_{79}(CAACAG)_2$CCGCCA (CCG)n	61	3	107Q #25 107Q #4 107Q #4	Clone #05 $(CAG)_{78}$-$(CAACAG)_2$ Clone #32 $(CAG)_{78}$-$(CAACAG)_2$ Clone #33 $(CAG)_{79}$-$(CAACAG)_2$	83Q CAACAG-dup
$(CAG)_{43}$CCGCCA(CCG)n	40	2	107Q #4 107Q #4	Clone #04 $(CAG)_{43}$ Clone #14 $(CAG)_{43}$	43Q CAA-loss

Since H9 cell lines are heterozygous for CAG repeats (17 and 25 CAG) in the *HTT* exon 1 locus, geno-PCR was used to confirm the targeting of the shorter allele for all the clones (Figures 1B and 1D). The excision of the antibiotic-resistance cassette was verified by right homology arm (RHA)-PCR (Figures 1B and 1E), showing the absence of the expected PCR product compared to the positive control (CTR+), and random insertions of the targeting plasmids were ruled out using four distinct PCRs mapping various plasmid portions (Figure 1F) showing that all the clones appear negative compared to the positive control. *HTT* locus transcriptional activity was tested by quantitative reverse-transcription PCR (RT-qPCR), showing equivalent transcriptional level of *HTT* in the different clones and the H9 cell line (Figure 1H). Total HTT protein levels, measured by western blot, showed a mild reduction in the expression in all clones compared to the H9 cell line, consistent with findings in other isogeneic stem cell allelic series[22] (Figures 1I and 1J; Table S1). A normal 46,XX karyotype was confirmed for all clones by G-banding analysis (Figure S1B). Pluripotency genes were expressed at similar levels across clones, as shown by RT-qPCR (Figures 1K–1M and S1H; Table S1) and immunocytochemistry for OCT4 and SOX2 (Figure S1C).

Cre-mediated exchange of the RMCE cassette was used to integrate *HTT* exon 1 variants with different CAG lengths into the parental cell line, generating the first components of the CAGinSTEM human platform (Figures 1A, 2A, and 2B). We integrated $(CAG)_{19}$-CAACAG, $(CAG)_{43}$-CAACAG, and $(CAG)_{78}$-CAACAG, referred to as 21Q, 45Q, and 81Q, respectively. All 21Q and 45Q clones showed the expected repeat length, while the 81Q clones included one with 80Q and two with 81Q (Table 1). After antibiotic selection, recombined cells underwent nucleofection with Flp-recombinase to remove the neomycin selection marker. Clones were isolated by single-cell sorting and verified as previously described for the 107Q line (Figures 1A and 2C–2G). The isolated clones and their respective CAG length are listed in Table 1 and were rigorously verified for locus integrity (Figures 2C–2E), absence of plasmids' random insertions (Figures 2F and 2G), *HTT* mRNA expression level (Figure 2H; Table S1), HTT protein level (Figures 2I and 2J; Table S1), and pluripotency markers (Figures S1D–S1H; Table S1).

CAG sizing by third-generation sequencing

Precise CAG sizing retaining DNA sequence information is inherently complex.[23,24] To address this, we developed a quantitative method to determine the CAG length within the engineered *HTT* exon 1 locus through genomic PCR amplification and amplicon long-read sequencing with ONT. Analysis of ONT sequence data was performed using a novel computational Nextflow pipeline, optimized for assessing CAG repeat instability (Figures 3A

CellPress
OPEN ACCESS

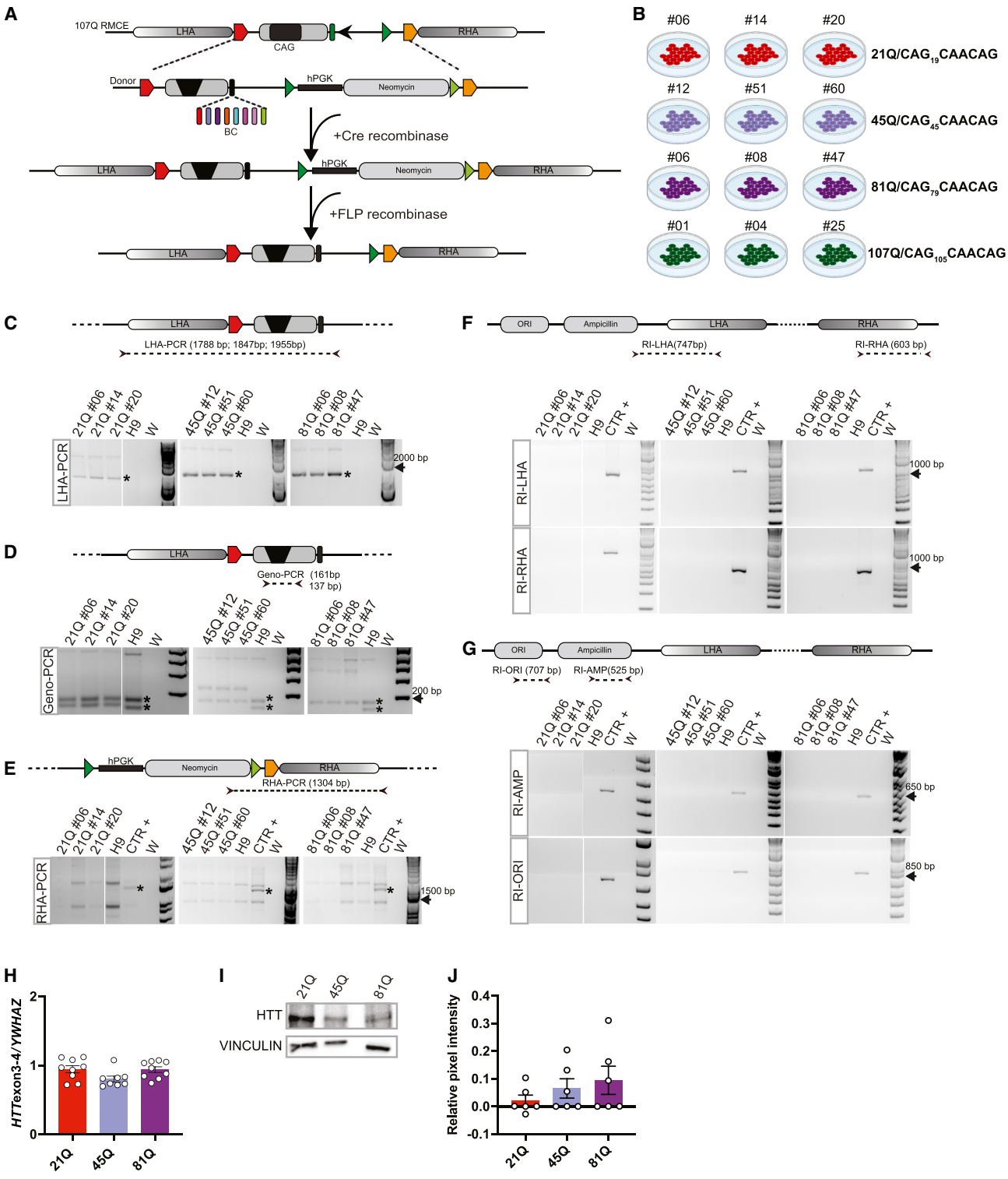


Figure 2. Generation and characterization of CAGinSTEM platform carrying incrementally longer CAG length in HTT exon 1

(A) Exchange strategy to generate hESC carrying an increasing CAG length. BC, barcode; colors identify the specific barcode for each donor. LHA and RHA are left and right homology arms, respectively.

(B) Schematic representation of clones isolated and characterized for each genotype.

(legend continued on next page)

and S2A). This process included base-calling with per-sample demultiplexing, alignment to the human reference genome, and CAG repeat size determination using Straglr.[25] The instability for trinucleotide repeat expansions was evaluated to assess the likelihood or propensity of CAG repeats to undergo expansion. This index examines the distribution of CAG sizes, often represented by peaks of different lengths.[26] In addition to the ONT barcode, our reads incorporate a cell line and repeat-specific barcode. This allows for an additional demultiplexing step and enables the assignment of each read to the originating cell line in the case of a mixed culture (21Q, 45Q, 81Q, and 107Q) with multiple CAG lengths. This feature of the CAGinSTEM platform was demonstrated at different time points during the *in vitro* striatal differentiation process (Figure 3B).

To exclude PCR-induced sequence artifacts or bias, we compared CAG size distributions obtained from direct sequencing of linearized plasmids containing 19 and 79 pure CAG repetitions (21 and 81Q) (Figures 3C and 3E) with those from the same plasmids after PCR amplification or from PCR on genomic DNA extracted from the corresponding clones (Figures 3D and 3E). Comparable normalized peak heights of CAG size distributions, centered around the same main peak, were observed between direct sequencing and PCR amplification of plasmid and genomic DNA, confirming the robustness of the procedure (Figure 3E; Table S1). Finally, to validate the equivalence between different PCR designs, such as ONT-PCR and LHA-PCR (the latter tailored for specific amplification of the edited allele only), we conducted a comparative analysis of CAG size distributions in gDNA amplified using both methods, after selecting reads from the recombinant allele and leveraging the cell line barcode. No significant differences were found in the normalized peak heights of CAG size distributions, centered around the main peak, between the two methods (Figure 3E; Table S1). This confirms the suitability of a PCR-based third-generation sequencing approach for accurately quantifying *HTT* CAG size, while retaining allele-specific genomic sequences.

CAG length influences repeat instability in mitotically active stem cells and human striatal neurons

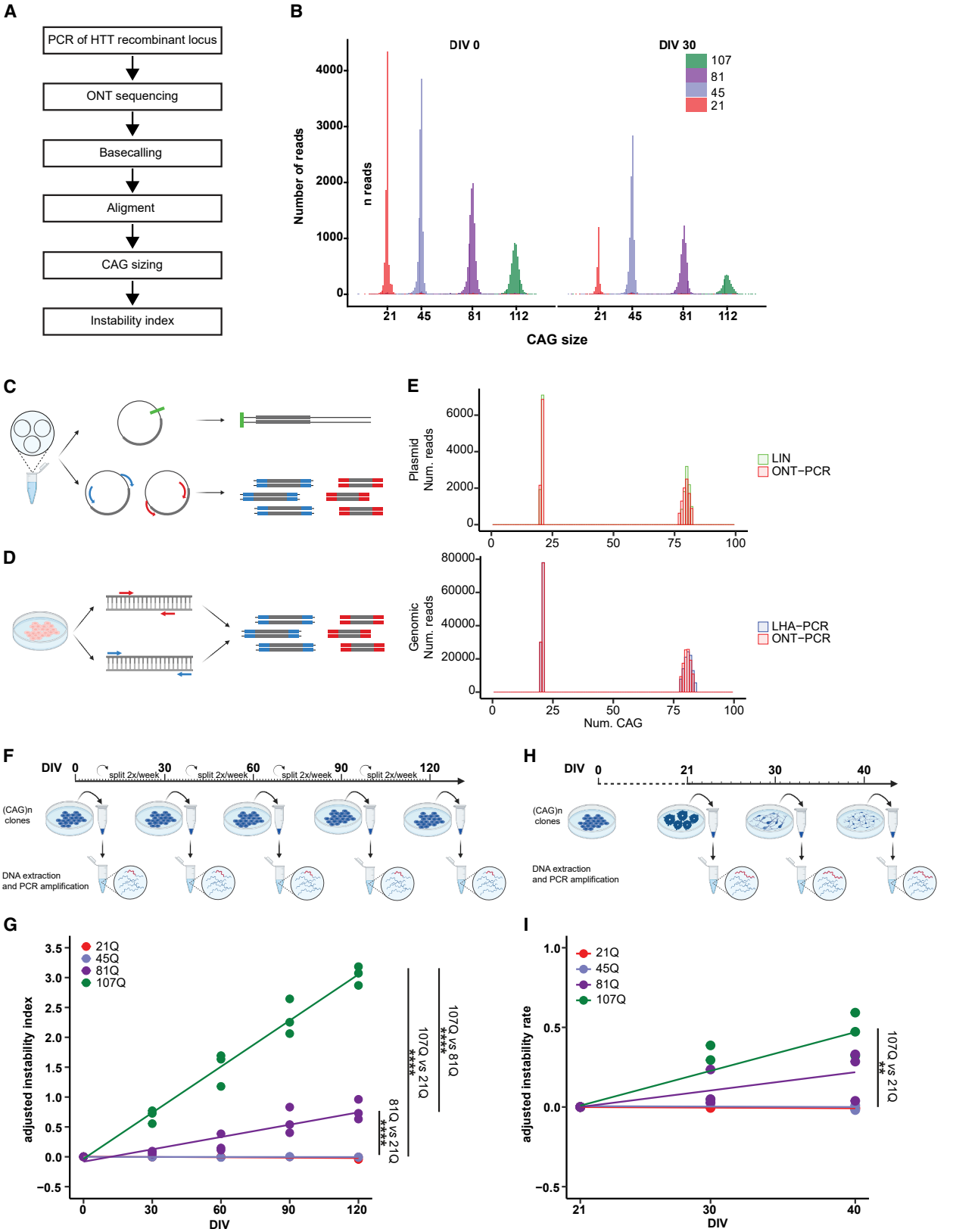
Given the high propensity of the *HTT*-CAG tract to expand in both mitotically dividing and post-mitotic cells, a mosaic of cells with varying repeat lengths develops over the lifetime of individuals with HD.[27] This is particularly evident in striatal neurons, the most affected cell type in HD.[28–31] Accordingly, we aimed to study the dynamics of CAG instability as a function of CAG length in both mitotically dividing cells and in post-mitotic striatal neurons.

To assess the trend of CAG instability during self-renewal, we cultured cell lines with increasing CAG lengths for 120 days *in vitro* (DIV), passaging them twice a week and collecting cell pellets every 30 DIV (Figure 3F).

As previously described,[32] the repeat instability index is a static measurement of repeat expansions and/or contractions observed in a population of cells, tissues, or organisms, regardless of the main repeat length frequency of the distribution. To better capture changes in repeat distribution over time from the beginning of the experiment (DIV0), we adopted a modified version by computing the instability index at each time point, referred to as the DIV0 main peak and called "adjusted instability rate" (AIR; see STAR Methods). This value directly represents the average number of triplet repeat changes occurring over time from DIV0 (Figure S2D). Using this approach to track the dynamics of CAG instability, we first showed that clones of 107Q and 81Q, despite small differences in uninterrupted CAG length (see Table 1), exhibit the same AIR trend within their respective cell lines (Figures S2E and S2F; Table S1). A significant increase in AIR over time was observed in 81Q and 107Q clones, but not in the 45Q clones, compared to the 21Q control line (Figure 3G; Table S1). Of note, we also detected a statistically significant difference in AIR levels between 107Q and 81Q clones, suggesting that AIR increases with uninterrupted CAG repeat length within the expanded range. Similar to a recent mouse study,[16] both 81Q and 107Q clones displayed a linear increase in CAG instability over time, with a modal CAG repeat expansion of 0.00275 and 0.025 repeat units/day, respectively (Figure 3G; Table S1).

Somatic expansion of the *HTT* CAG tract occurs stochastically throughout the lifespan of individuals with HD and in HD animal models. This process leads to a mosaic of cells with varying repeat lengths,[15,27,33] with MSNs being the most affected cell type.[29–31] Therefore, we investigated the dynamics of *HTT* CAG repeat by subjecting the CAGinSTEM platform cell lines to striatal neuronal differentiation and repeat sizing.[34] Cell pellets

(C) Upper: schematic representation of donor plasmid. Segmented lines surrounded by black arrows indicate PCR amplicons. Lower: LHA-PCR to assess monoallelic knockin. * PCR product of 1,788 bp for 21Q, 1,847 bp for 45Q and 1,955 bp for 81Q.
(D) Upper: schematic representation of donor plasmid. Segmented lines surrounded by black arrows indicate PCR amplicons. Lower: geno-PCR is used to assess monoallelic knockin. * PCR product of 161 bp (H9 with CAG25CAACAG/27Q) and of 137 bp (H9 with CAG17CAACAG/19Q).
(E) Upper, schematic representation of donor plasmid. Segmented lines surrounded by black arrows indicate PCR amplicons. Lower panel, RHA-PCR to assess the excision of the antibiotic selection cassette. * PCR product at 1,304 bp.
(F) Upper: schematic representation of donor plasmid elements outside the homology arms. Segmented lines surrounded by black arrows indicate PCR amplicons. Lower, RI-LHA (747 bp) and RI-RHA (603 bp) PCRs to assess plasmid random insertions (RI) into the genome.
(G) Upper: schematic representation of donor plasmid elements outside the homology arms. Segmented lines surrounded by black arrows indicate PCR amplicons. Lower panel, RI-AMP (525 bp), RI-ORI (707 bp), and RI-ORI PCRs to assess plasmid random insertions into the genome.
(H) RT-qPCR analysis of *HTT* in 21Q, 45Q, and 81Q cell lines. *YWHAZ* is used as the housekeeping gene.
(I) Western blot of HTT. Vinculin is used as the housekeeping gene.
(J) Quantification of HTT total protein expression in relative pixel intensity normalized to the intensity of the vinculin band.
For all agarose gels, 1 kb plus ladder is used as the molecular weight marker, and arrows indicate molecular weight. W (water) and CTR+ are the negative and the positive controls, respectively, of the PCR reaction.
Data are presented as the mean ± SEM. Dots represent different culturing passages of the clone. Kruskal–Wallis test followed by Dunn's post hoc test (H) or one-way ANOVA followed by Tukey's post hoc test (J) are performed, and statistical analysis is reported in Table S1.



(legend on next page)

were collected from precursor cells during neuronal specification (DIV21 and 30), as well as from terminally differentiated neurons at DIV40, to measure CAG instability (Figure 3H). The propensity of the different cell lines and clones for neuronal differentiation was qualitatively assessed through the expression of ßIII-tubulin and MAP2, two markers associated with neuronal differentiation and maturation (Figure S2G). At the end of differentiation, all clones exhibited approximately 60% post-mitotic neurons and about 20% of cycling cells (Figures S2H and S2I; Table S1), both contributing to AIR dynamics. Tracking the dynamics of CAG instability during striatal differentiation, we found that 107Q clones showed a statistically significant increase in CAG instability over time compared to 21Q control clones. In contrast, 45Q and 81Q clones exhibited relatively stable CAG repeats, with no significant differences observed in current culture conditions (Figure 3I; Table S1). Furthermore, CAG instability accumulated linearly in both 81Q and 107Q from DIV21 to DIV40 (Figure 3I; Table S1). We conclude that a linear increase in CAG instability can be detected during both neuronal differentiation and self-renewal using the CAGinSTEM platform.

CAG instability is modulated by CAG stretch composition

We expanded our CAGinSTEM platform to model the recently identified variants of the CAG repeat terminal codons CAA-CAG.[10,11,13] First, we introduced a duplication of the CAACAG in the 107Q and 81Q lines (Figure 4A), generating cell lines with an additional CAACAG interruption (109Q CAACAG-dup; 83Q CAACAG-dup; Figures 4A, S3A–S3J, and S4A–S4L; Tables 1 and S1). Second, we generated cell lines in which the final CAACAG was replaced with an uninterrupted CAGCAG (107Q CAA-loss; 43Q CAA-loss; Figures 4A, S3A–S3J, and S4M–S4X; Tables 1 and S1). Finally, to investigate the role of CAA interruptions in the *HTT* repeat, we created additional lines carrying a synthetic allele containing four CAA interruptions at positions 22, 44, 67, and 89 within the CAG stretch. This allele, denoted 107Q multiple central interruptions (multi-CAAs), follows the structure $(CAG)_{21}CAA]_4(CAG)_{17}CAACAG$, encoding 107 glutamines (Figures 4A and S4A–S4J; Tables 1 and S1). Notably, this latter composition lacks any pure CAG stretch of a diagnostically pathological length. All genotypes retain the intervening sequence $CCGCCA(CCG)_n$ found in the corresponding proline-rich region of the canonical human allele.[35]

To assess the impact of these modifications on CAG instability, we examined the mitotically dividing cells, which exhibit a higher instability rate than neurons. As shown in Figure 3F, cell lines were passaged twice a week, and pellets were collected every 30 DIV (Figure 3F), up to DIV120. By measuring CAG instability over time, we found that duplicating the terminal CAA interruption (109Q CAACAG-dup) significantly reduced the AIR compared to 107Q clones with a typical allele (Figure 4B; Table S1). Similarly, 83Q CAACAG-dup clones exhibited a significant reduction in CAG instability compared to 81Q clones over the 120-day period (Figure 4C; Table S1). These results confirm a direct link between non-canonical CAA interruptions and repeat instability, validating the CAGinSTEM platform as an *in vitro* model for studying *HTT* CAG instability.

Recent human data indicate that the CAA-loss allele advances disease onset by 5.0 years on average, whereas the CAA/CCA-loss leads to 10.0 years earlier onset. Additionally, a reduction in somatic expansion was observed in blood samples.[13] However, in 107Q CAA-loss clones, we detected no significant change in instability, compared to canonical 107Q clones with a similar number of uninterrupted CAG repeats (Figure S3K; Table S1). To further investigate this, we examined the impact of CAACAG loss in a $(CAG)_{43}$ pure repeat, matching the allele described in the families with HD CAACAG-loss.[10,11,13] By comparing the instability of control canonical 45Q and modified 43Q CAA-loss cell lines, we observed a significant increase in CAG instability over time in the 43Q CAACAG-loss cells, despite their identical number of uninterrupted CAG repeats (Figure 4D; Table S1). Remarkably, introducing multiple CAA interruptions within the CAG stretch (107Q multi-CAAs lines) completely abolished repeat instability over time compared to the canonical 107Q line (Figure 4E; Table S1). This finding aligns with the observations from transgenic animal models carrying multiple interruptions in the repeat stretch.[17]

Overall, our findings highlight the significant impact of CAA terminal interruptions on the stability of the CAG tract, with multiple interruptions effectively preventing CAG instability over time.

Figure 3. CAG sizing to evaluate instability in hESC lines carrying HTT-CAG, increasing length in self-renewal and neuronal differentiation

(A) Schematic representation of all the steps needed to go from the enrichment of DNA carrying HTT exon 1 to the instability evaluation.

(B) Exploiting cell line-specific barcode for mixed culture experiments.

(C) Schematic overview of the HTT exon 1 PCR enrichment validation workflow using plasmid DNA. Plasmids were linearized for direct sequencing (green bar), and HTT exon 1 was enriched via non-allele-specific amplification (ONT-PCR, red arrows) for amplicon sequencing.

(D) Schematic overview of the HTT exon 1 PCR enrichment validation workflow using genomic DNA. Genomic DNA was extracted, and HTT exon 1 was amplified using both non-allele-specific (ONT-PCR, red arrows) and allele-specific (LHA-PCR, blue arrows) approaches for amplicon sequencing.

(E) Validation of the PCR enrichment protocol. Top: CAG sizing from the plasmid through direct sequencing after linearization (LIN) and sequencing after non-allele-specific amplification (ONT-PCR). Bottom: CAG sizing from genomic DNA after non-allele-specific (ONT-PCR) and allele-specific (LHA-PCR) amplification. Two CAG lengths (21Q and 81Q) were considered. Again, only CAG sizes with a relative frequency higher than 20% of the most frequent CAG size are reported. An exact two-sample Kolmogorov-Smirnov test is applied, and statistical analysis is reported in Table S1.

(F) Schematic representation of self-renewal maintenance of clones.

(G) Graph represents the adjusted CAG instability rate in self-renewal.

(H) Schematic representation of neuronal differentiation protocol.

(I) Graph represents the adjusted CAG instability rate in terminally differentiated neurons. Each dot represents a clone. For 21Q #6, 107Q #1, 107Q #4, and 107Q #25 clones, each dot represents the average of at least three experimental replicates; two-way ANOVA followed by Tukey's multiple comparisons test is applied, and statistical analysis is reported in Table S1. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and $****p < 0.0001$.

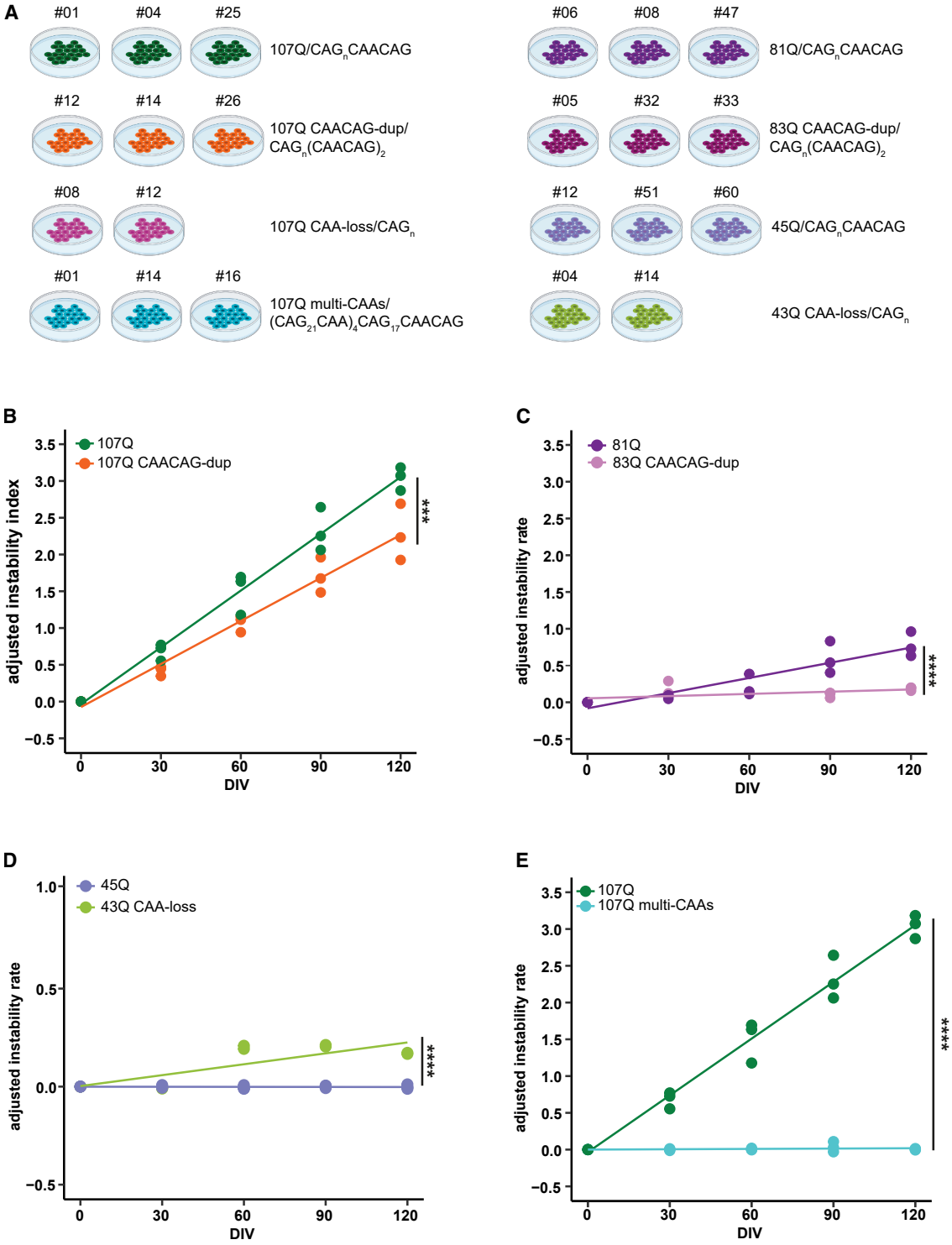
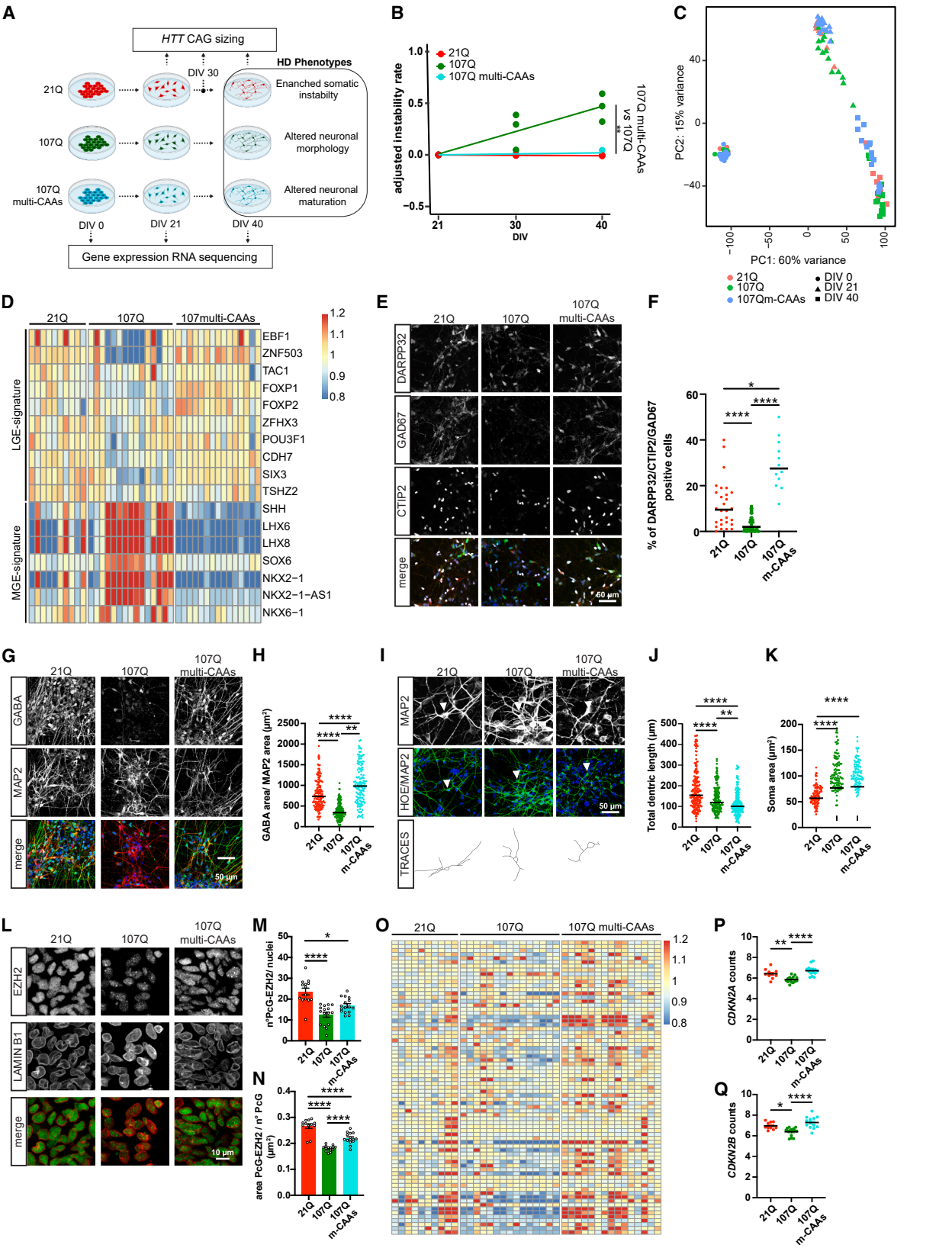


Figure 4. CAG instability is modulated by CAG stretch composition

(A) Schematic representation of clones subjected to instability evaluation in self-renewal conditions.

(B–E) Graph represents adjusted CAG instability rate over 120 DIV of 107Q (data as in Figure 3G), 107Q CAACAG-dup, 81Q (data as in Figure 3G), 83Q CAACAG-dup, 45Q (data as in Figure 3G), 43Q CAA-loss, and 107Q multi-CAA clones in self-renewal. Each dot represents a clone. For 107Q#1, 107Q#4, and 107Q#25 clones, each dot represents the average of at least three experimental replicates; two-way ANOVA followed by Tukey's multiple comparisons test is applied, and statistical analysis is reported in Table S1. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$.



(legend on next page)

Uninterrupted CAG repeat length as the key determinant of intracellular pathology in HD neurons

The complete recovery of CAG instability over time observed in the proliferating 107Q multi-CAAs cell line raises the question of whether central interruptions could also mitigate CAG instability in post-mitotic striatal neurons—the earliest affected cell type in HD. To address this, we subjected 107Q multi-CAAs, 107Q, and 21Q control cell lines to striatal differentiation as described in Figure 3H. At the end of differentiation, although 107Q multi-CAAs exhibited slightly fewer p27-positive cells compared to 107Q, this did not correspond to differences in Ki67-positive cells. Additionally, transcript levels for both markers were comparable across all genotypes, suggesting no significant differences in cell cycle distribution (Figures S5A–S5C; Table S1). However, longitudinal CAG sizing confirmed the complete suppression of repeat instability in 107Q multi-CAA neurons compared to canonical 107Q cells (Figures 5A and 5B; Table S1), indicating that an uninterrupted CAG repeat contributes to some extent to the rate of expansion, even in post-mitotic cells.

Building on previous findings that multiple repeat interruptions mitigate striatum-specific transcriptional dysregulation in mice,[17] we next examined transcriptional signatures through genome-wide RNA sequencing of 21Q, 107Q, and 107Q multi-CAA cell lines during striatal differentiation (at DIV0, DIV21, and DIV40, Figure 5A; Table S1). Notably, the 107Q multi-CAA cell lines carry an extreme CAG length encoding for a pathologically long polyQ protein, which is expressed at comparable levels to 107Q (Figure S5D; Table S1). While all genotypes clustered together at DIV0, genotype-specific transcriptional differences emerged at DIV21 and DIV40 (Figures 5C and S5E). To explore variability between HD and control cultures, we analyzed differentially expressed genes (DEGs) at all time points (Figures S5F–S5G; Table S3). Notably, among all DEGs, a subset of 379 genes was downregulated and 261 upregulated at terminal differentiation in both 107Q vs. 21Q and 107Q vs. 107Q multi-CAA comparisons. This indicates that there are some HD-associated transcriptional alterations that are also found in the same direction in the 107Q vs. 107Q multi-CAA comparison, suggesting that for these genes, the 107Q multi-CAA mimics the 21Q line. In other words, such alterations can be reversed—without altering the number of glutamines in the HTT protein—by introducing multiple synonymous CAA interruptions. Notably, this transcriptional comparison found no differential expression of mismatch repair genes, providing no evidence that differences in expression levels of these genes contribute to the observed phenotype.

Neurons obtained from human pluripotent stem cells carrying the HD mutation have been reported to exhibit characteristic pathological features, including impaired striatal neuron differentiation, abnormal neuronal morphology, and altered HTT1a production, some of which are also observed in HD mouse models.[17,36–43] Consistently, bulk RNA sequencing showed a reduction in lateral ganglionic eminence (LGE)-specific transcripts, like EBF1 and SIX3, in canonical 107Q lines relative to control 21Q lines, accompanied by an increase in medial ganglionic eminence (MGE)-like signature, such as SHH and NKX2.1 transcripts, suggesting a shift in fate determination in the presence of the HD gene, at both DIV21 and terminal differentiation (Figures 5D, S5H, and S5I; Table S1). Notably, the presence of multiple CAA interruptions in the 107Q multi-CAA clones restored the differentiation marker profile to a state similar to the control (Figures 5D, S5H, and S5I; Table S1). This was corroborated by a significant reduction in the yield of MSNs and GABAergic neurons in 107Q cells compared to 21Q controls, which was fully recovered in the 107Q multi-CAA (Figures 5E–5H; Table S1). Collectively, this analysis confirms that an

Figure 5. Uninterrupted CAG repeat length is the key determinant of intracellular pathology in HD neurons

(A) Schematic representation of the experimental paradigm.

(B) Graph shows the adjusted CAG instability rate during neuronal differentiation of 21Q, 107Q (data as in Figure 3I), and 107Q multi-CAA clones. Each dot represents a clone. For 21Q#6, 107Q#1, 107Q#4, and 107Q#25 clones, each dot represents the average of at least three experimental replicates; two-way ANOVA followed by Tukey's multiple comparisons test is applied, and statistical analysis is reported in Table S1.

(C) Principal-component analysis.

(D) Heatmap of VST-transformed expression values normalized by row for LGE transcriptional signature and MGE transcriptional signature.

(E) Representative immunofluorescence images of 21Q, 107Q, and 107Q CI cells stained with DARPP32 (green), GAD67 (red), and CTIP2 (gray) to identify MSNs at DIV40. Hoechst is used to counterstain nuclei. Scale bars, 50 μm.

(F) Quantification of the percentage of triple-labeled positive cells.

(G) Representative immunofluorescence images of 21Q, 107Q, and 107Q multi-CAA cells stained with GABA (green) and MAP2 (red) to identify GABAergic neurons at DIV40. Hoechst (blue) is used to counterstain nuclei. The scale bar is 50 μm.

(H) Quantification of the area covered by GABA signal over MAP2 area, which is normalized to the number of nuclei.

(I) Representative immunofluorescence images of 21Q, 107Q, and 107Q multi-CAA cells stained with MAP2 (green) and Hoechst (blue) to identify neurons at DIV40. Representative traces of neuronal morphology of MAP2+ neurons are shown. Scale bars, 50 μm.

(J) Quantification of total dendritic length based on MAP2+ neurons.

(K) Quantification of soma area based on MAP2+ neurons.

(L) Representative immunofluorescence images of 21Q, 107Q, and 107Q multi-CAA cells stained with EZH2 (green) and lamin B1 (red). Scale bars, 10 μm.

(M) Quantification of EZH2 PcG bodies number.

(N) Quantification of EZH2 PcG bodies area (μm^2).

(O) Heatmap of VST-transformed expression values normalized by row for PRC2-regulated genes. The complete list of genes is in Table S3.

(P) Graph represents *CDKN2A* counts at DIV40 of neuronal differentiation within each replicate of 21Q, 107Q, and 107Q multi-CAAs clones.

(Q) Graph represents *CDKN2A* counts at DIV40 of neuronal differentiation within each replicate of 21Q, 107Q, and 107Q multi-CAAs clones.

Data are presented as mean ± SEM. For (F, H, M, and N), dots represent images from at least two clones/genotype from at least three independent biological replicates. For (J) and (K), each dot represents a single neuron. For (P) and (Q), each dot represents a clone from at least two clones/genotype from at five independent biological replicates; one-way ANOVA followed by Tukey's post hoc test (N and P) or Kruskal-Wallis test followed by Dunn's post hoc test (F, H, J, K, M, and Q) are performed and statistical analysis is reported in Table S1. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$.

uninterrupted CAG repeat promotes suboptimal LGE differentiation while favoring an increased proportion of MGE-like neuronal identities. These findings strongly suggest that the impaired striatal differentiation observed in 107Q neurons is driven by CAG purity rather than polyQ tract length. As further support of this hypothesis, comparison of the fraction of reads mapping to HTT intron 1 as a proxy for HTT1a transcript levels exhibits a significantly higher abundance of HTT1a in 107Q lines, but not in 107Q multi-CAAs, compared to the 21Q line (Figure S5J; Table S1). Altered dendritic development has been reported in cortical neurons derived from HD-induced pluripotent stem cells.[39] To examine neuronal morphology, we assessed microtubule-associated protein 2 (MAP2)-positive neuronal outgrowths at DIV40 of differentiation. Interestingly, 21Q neurons exhibited greater mean neurite lengths and reduced soma areas compared to both 107Q and 107Q multi-CAA cells (Figures 5I–5K), indicating that this phenotype is polyQ-dependent and not rescued by the multiple CAA interruptions.[17]

These findings emphasize that CAG repeat purity, independent of polyQ length, governs repeat instability and striatal differentiation abnormalities in HD neurons.

CAA interruptions restore nuclear morphology and chromatin organization in HD neurons

HD neurons have also been associated with nuclear morphology alterations and increased lamin B1 levels in patient brains and adult mouse models.[44–46] To investigate this, we immunostained 21Q, 107Q, and 107Q multi-CAA cell lines at DIV40 for lamin B1. Compared to 21Q controls, canonical 107Q neurons displayed a significant reduction in nuclear area and increased lamin B1 intensity (Figures S6A, S6B, and S6D; Table S1), both of which were fully recovered in 107Q multi-CAA cells (Figures S6A, S6B, and S6D; Table S1). Nuclear circularity, however, remained unaffected across all lines (Figures S6A and S6C; Table S1). The reduced nuclear area observed in 107Q neurons raises the possibility of increased H3K9me3-marked constitutive heterochromatin, which is closely associated with nuclear envelope morphology.[47,48] Consistent with this, we found that H3K9me3 puncta were significantly larger in 107Q neurons compared to 21Q, suggesting more compact heterochromatin in the presence of mutant HTT (Figures S6E–S6G; Table S1). Importantly, puncta size was fully recovered in 107Q multi-CAAs cells, suggesting that CAA interruption influences chromatin organization (Figures S6E–S6G; Table S1). Given PRC2's role in regulating facultative heterochromatin and cell-type-specific silencing of developmentally regulated genes,[49–52] we next investigated both the protein levels of key PRC2 components and the organization of nuclear bodies in our CAGinSTEM platform. Specifically, we performed western blot analyses and assessed EZH2 puncta formation at DIV40. We found no significant differences in the total protein levels of PRC2 components or in global H3K27me3 levels across the different genotypes (Figures S6H and S6I; Table S1), consistent with previous findings in knockin mouse ESC lines.[53] Notably, canonical 107Q neurons displayed a significant reduction in the number and size of polycomb group (PcG) foci per nucleus compared to 21Q controls (Figures 5L–5N, S6J, and S7K; Table S1), consistent with prior reports of mutant HTT disrupting PRC2 organization.[13,52–54] Notably, all phenotypes were partially restored in 107Q multi-CAA cells, further supporting a role for uninterrupted CAG repeats in HD pathology. Moreover, transcriptional dysregulation in HD striatal cells has been reported in postmortem human tissues and in the KI-Q175 HD mouse model,[53–55] implicating partial PRC2 loss of function. In addition, PRC2 facilitates neuron specification during differentiation and contributes to the repression of harmful transcriptional programs in adult neurons. We therefore explored the effect of CAA interruptions on PRC2-related gene signature derived from neuron-specific Ezh2 conditional knockout mice[52] in the CAGinSTEM lines during differentiation. We detected a significant reduction of PRC2-regulated transcripts in canonical 107Q cells compared to unexpanded 21Q controls, with partial recovery in multi-CAA clones (Figures 5O–5Q and S6L–S6N; Tables S1 and S3).

Taken together, our findings suggest that uninterrupted CAG repeat drives HD phenotypes in neurons, leading to transcriptional dysregulation and disrupted cell identity, which may contribute to long-term neuronal loss. Furthermore, these results highlight the primary role of *HTT* DNA repeat composition, rather than the HTT protein, in governing *HTT* CAG instability and HD-related phenotypes, further validating the CAGinSTEM platform as a robust *in vitro* model for studying progressive genetic and phenotypic variations in HD.

DISCUSSION

In this study, we have established a genetically engineered platform using hESCs that harbor barcoded variants of *HTT* exon 1, featuring diverse CAG sizes and nucleotide compositions integrated into the native *HTT* locus. This innovative CAGinSTEM platform, distinguished by its ability to monoallelically modify *HTT* exon 1, facilitates the generation of quality-checked cell lines and specialized cell types within a relatively short time frame. Crucially, we introduced an ONT third-generation long-read sequencing method for precise CAG sizing, ensuring an accurate assessment of repeat length and composition over time. By employing the AIR for tracking CAG expansion over time, we could average out possible biases arising from the ONT error rate, as potential biases are shared between samples across all time points. Moreover, by comparing CAG size distributions between linearized plasmids and amplified genomic DNA, we showed that our PCR amplification protocol did not introduce any relevant bias. While providing very high target enrichment, the PCR amplification came at the cost of erasing DNA methylation and potential genomic variants outside of the amplified locus, which may potentially play a role in somatic instability.

Our findings elucidate the dynamics of CAG instability across different lengths during both stem cell propagation and differentiation according to a striatal protocol. We observed that clones harboring longer canonical CAG repeats, specifically 81Q and 107Q, exhibit a linear increase in instability over time, in contrast to clones with 45Q or fewer repeats, which remain stable under similar conditions. This correlation between CAG length and instability rate supports and recapitulates *in vitro* the model whereby longer CAG tracts are increasingly prone to somatic

expansion, progressively leading to transcriptional dysregulation and pathogenic consequences.

Our study aligns with the two-stage hypothesis of HD pathogenesis, which proposes that while inherited CAG repeats are not initially pathogenic, they are susceptible to somatic expansion over time, eventually triggering disease onset.[16,17] We show that the rate of CAG expansion correlates with repeat length, following a linear relationship between expansion rate and CAG size. This relationship influences both the timing of onset and the trajectory of disease progression. Recent studies on mismatch repair genes in mouse *Htt* models and human cells have reported a similar linear correlation between repeat expansion and time.[16,56–58] Importantly, our platform models clinically relevant haplotypes, offering valuable insights into how these variations impact CAG instability and associated cellular phenotypes.

The two-stage hypothesis also provides an explanation for why rodent models of HD require the introduction of very large CAG tracts (>100) to display disease-like features, as their limited lifespan does not provide enough time for typical human pathogenic ranges (39–44 CAG) to expand somatically beyond the critical threshold of toxicity.[59] Therefore, a fully HTT-humanized cell system, such as our CAGinSTEM platform, provides an optimal context to study the dynamics and mechanisms of somatic instability.

In this study, we demonstrate that the single loss of a CAA interruption increases instability in 43Q CAA-only lines compared to the canonical 45Q allele. Conversely, duplication of the CAACAG motif reduces instability in both 81Q and 107Q lines. Notably, multi-CAA 107Q lines, engineered to contain multiple central CAA interruptions, completely abolish the typical pattern of somatic expansion observed in HD models. In addition, these interruptions reverse several HD-associated *in vitro* phenotypes. These include impaired acquisition of striatal identity, altered transcriptional profiles in the LGE lineage, increased HTT1a levels, and disorganized nuclear and chromatin architecture that includes a partial disruption of PRC2-related gene signature. These findings align with a model where pure CAG repeats form stable hairpin structures during DNA replication that are able to promote slippage events and therefore repeat expansion.[60] The introduction of CAA interruptions supposedly disrupts these secondary structures, preventing the DNA polymerase slippage that drives somatic instability. By preventing the formation of these pure CAG-dependent secondary structures, CAA interruptions enhance repeat stability during cell division and aging of post-mitotic neurons, potentially blocking the primary driver of HD progression.

A growing body of work in both mouse and human systems implicates reduced polycomb complex activity, particularly PRC2, in HD-associated toxicity. Transcriptional parallels between PRC2-deficient mouse models and cells derived from patients with HD support the hypothesis that partial loss of PRC2 function contributes to aberrant gene expression in a cell-type-specific manner. This epigenetic misregulation may impair neural fate specification during development and/or promote the inappropriate reactivation of developmental genes, ultimately leading to neuronal vulnerability and degeneration.[16,19,55] Our CAGinSTEM data support this model and show that PRC2-related phenotypes are dependent on CAG instability, as the introduction of multiple CAA interruptions into the *HTT* repeat restores several of these transcriptional and chromatin abnormalities to near-normal levels. These results raise the possibility that introducing CAA interruptions into the repeat could have therapeutic value in HD. Although a CRISPR-based approach has been recently reported in a proof-of-principle study with some positive outcomes,[56,61,62] the translation of such an approach remains highly speculative as gene-editing technologies face significant challenges in the efficiency of delivery and precision of modification in post-mitotic human neurons *in vivo*.

Moreover, our findings point to the intriguing possibility that some individuals may carry pathogenic-range CAG alleles yet remain asymptomatic due to the naturally occurring central CAA interruptions, suggesting a natural protective mechanism at play.

In conclusion, our study contributes to the evolving understanding of HD pathogenesis, shifting focus from a purely polyglutamine-centric view to one that emphasizes the role of somatic CAG expansion. Through the use of a genetically precise, humanized CAGinSTEM platform, we provide compelling support for the two-stage model. By investigating clinically relevant haplotypes, we highlight potential therapeutic avenues for modifying CAG tract purity to mitigate disease phenotypes. Despite its *in vitro* nature, our platform offers unique opportunities for high-resolution screening of modifiers of CAG instability and a deeper understanding of the molecular underpinnings of HD. Collectively, these insights advance the goal of developing effective interventions to halt or delay HD progression.

Limitations of the study

Despite the strengths of our system, its *in vitro* nature imposes certain limitations. In particular, the absence of detectable repeat instability in some cell lines (e.g., 45Q) may be influenced by factors such as culture duration and detection sensitivity. Thus, apparent mitotic stability does not rule out the emergence of instability over longer time frames. Moreover, while providing the advantages of an isogenic system, our hESC-based model may not fully recapitulate the aspects of age-dependent neurodegenerative phenotypes observed in patient neurons and cannot represent individual cell polymorphisms, as these are masked by bulk analysis.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Elena Cattaneo (elena.cattaneo@unimi.it).

Materials availability

Plasmids and cell lines generated in this study will be made available upon request, but we may require a completed materials transfer agreement.

All reagents generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

- The complete codebase, including the Nextflow workflow and the R scripts, is available at the following GitHub repository: https://github.com/GianlucaDamaggio/T-Rex.

- Raw genomic and transcriptomic sequencing data have been submitted to NCBI SRA (BioProject PRJNA1074134; https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1074134).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank C. Cordiglieri and A. Fasciani of the Imaging Facility from INGM for scientific and technical assistance. We thank M.L. Sarnicola of the Flow Cytometry Facility from INGM for technical assistance. We acknowledge A. Via (Department of Biochemical Sciences "Alessandro Rossi Fanelli," Sapienza University, Rome) for the support in the establishment of bioinformatic infrastructure for the storage and analysis of sequencing data. This work received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no 742436). This study was also partially funded by Telethon GMR23T1059.

AUTHOR CONTRIBUTIONS

Conceptualization, D.B., C.T., C.Z., E.C., G.F., S. Maestri, and R.L.R.; methodology, M. Zobel, M.L.M., D.B., P.C., and G.D.; investigation, M. Zobel, M.L.M., D.S., A.C., I.C., F.C., P.C., A.S., R.I., M. Crosti, F.G, G.O, and L.A.; bioinformatic analysis, G.D., S. Maestri, and V.C.; writing – original draft, M. Zobel, D.B., M.L.M., and G.D.; writing – review & editing, M. Zobel, D.B., and E.C.; funding acquisition, E.C.; and supervision, E.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - hES cell culture
- METHOD DETAILS
  - Plasmid donor library and sgRNA design
  - Generation of H9-RMCE parental cell line
  - Generation of the CAGinSTEM platform with increasing CAG length and/or modified CAG stretch composition within HTT exon1
  - gDNA extraction and PCR
  - PCR clean-up and ONT library preparation
  - Striatal differentiation
  - Immunofluorescence and confocal imaging
  - Western blotting
  - RNA isolation and qRT-PCR
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ONT sequence data analysis
  - Bulk RNA-seq

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2025.116685.

Received: May 13, 2025
Revised: October 10, 2025
Accepted: November 17, 2025

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. https://doi.org/10.1016/0092-8674(93)90585-E.
2. Ross, C.A., and Tabrizi, S.J. (2011). Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurol. *10*, 83–98. https://doi.org/10.1016/S1474-4422(10)70245-3.
3. Tabrizi, S.J., Ghosh, R., and Leavitt, B.R. (2019). Huntingtin Lowering Strategies for Disease Modification in Huntington's Disease. Neuron *101*, 801–819. https://doi.org/10.1016/j.neuron.2019.01.039.
4. Besusso, D., Schellino, R., Boido, M., Belloli, S., Parolisi, R., Conforti, P., Faedo, A., Cernigoj, M., Campus, I., Laporta, A., et al. (2020). Stem Cell-Derived Human Striatal Progenitors Innervate Striatal Targets and Alleviate Sensorimotor Deficit in a Rat Model of Huntington Disease. Stem Cell Rep. *14*, 876–891. https://doi.org/10.1016/j.stemcr.2020.03.018.
5. Schellino, R., Besusso, D., Parolisi, R., Gómez-González, G.B., Dallere, S., Scaramuzza, L., Ribodino, M., Campus, I., Conforti, P., Parmar, M., et al. (2023). hESC-derived striatal progenitors grafted into a Huntington's disease rat model support long-term functional motor recovery by differentiating, self-organizing and connecting into the lesioned striatum. Stem Cell Res. Ther. *14*, 189. https://doi.org/10.1186/s13287-023-03422-4.
6. Holley, S.M., Reidling, J.C., Cepeda, C., Wu, J., Lim, R.G., Lau, A., Moore, C., Miramontes, R., Fury, B., Orellana, I., et al. (2023). Transplanted human neural stem cells rescue phenotypes in zQ175 Huntington's disease mice and innervate the striatum. Mol. Ther. *31*, 3545–3563. https://doi.org/10.1016/j.ymthe.2023.10.003.
7. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., and Kalchman, M.A. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat. Genet. *4*, 398–403. https://doi.org/10.1038/ng0893-398.
8. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., and Abbott, M. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat. Genet. *4*, 387–392. https://doi.org/10.1038/ng0893-387.
9. Lee, J.-M., Ramos, E.M., Lee, J.-H., Gillis, T., Mysore, J.S., Hayden, M.R., Warby, S.C., Morrison, P., Nance, M., Ross, C.A., et al. (2012). CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology *78*, 690–695. https://doi.org/10.1212/WNL.0b013e318249f683.
10. 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. https://doi.org/10.1016/j.ajhg.2019.04.007.
11. Lee, J.-M., Correia, K., Loupe, J., Kim, K.-H., Barker, D., Hong, E.P., Chao, M.J., Long, J.D., Lucente, D., Vonsattel, J.P.G., et al. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell *178*, 887–900.e14. https://doi.org/10.1016/j.cell.2019.06.036.
12. Ciosi, M., Maxwell, A., Cumming, S.A., Hensman Moss, D.J., Alshammari, A.M., Flower, M.D., Durr, A., Leavitt, B.R., Roos, R.A.C., and TRACK-HD team; and et al.. (2019). A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. EBioMedicine *48*, 568–580. https://doi.org/10.1016/j.ebiom.2019.09.020.
13. Lee, J.-M., McLean, Z.L., Correia, K., Shin, J.W., Lee, S., Jang, J.-H., Lee, Y., Kim, K.-H., Choi, D.E., Long, J.D., et al. (2025). Genetic modifiers of somatic expansion and clinical phenotypes in Huntington's disease highlight shared and tissue-specific effects. Nat. Genet. *57*, 1426–1436. https://doi.org/10.1038/s41588-025-02191-5.

14. Wheeler, V.C., Auerbach, W., White, J.K., Srinidhi, J., Auerbach, A., Ryan, A., Duyao, M.P., Vrbanac, V., Weaver, M., Gusella, J.F., et al. (1999). Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. Hum. Mol. Genet. *8*, 115–122. https://doi.org/10.1093/hmg/8.1.115.

15. Møllersen, L., Rowe, A.D., Larsen, E., Rognes, T., and Klungland, A. (2010). Continuous and Periodic Expansion of CAG Repeats in Huntington's Disease R6/1 Mice. PLoS Genet. *6*, e1001242. https://doi.org/10.1371/journal.pgen.1001242.

16. Wang, N., Zhang, S., Langfelder, P., Ramanathan, L., Gao, F., Plascencia, M., Vaca, R., Gu, X., Deng, L., Dionisio, L.E., et al. (2025). Distinct mismatch-repair complex genes set neuronal CAG-repeat expansion rate to drive selective pathogenesis in HD mice. Cell *188*, 1524–1544.e22. https://doi.org/10.1016/j.cell.2025.01.031.

17. Gu, X., Richman, J., Langfelder, P., Wang, N., Zhang, S., Bañez-Coronel, M., Wang, H.-B., Yang, L., Ramanathan, L., Deng, L., et al. (2022). Uninterrupted CAG repeat drives striatum-selective transcriptionopathy and nuclear pathogenesis in human Huntingtin BAC mice. Neuron *110*, 1173–1192.e7. https://doi.org/10.1016/j.neuron.2022.01.006.

18. Mätlik, K., Baffuto, M., Kus, L., Deshmukh, A.L., Davis, D.A., Paul, M.R., Carroll, T.S., Caron, M.-C., Masson, J.-Y., Pearson, C.E., and Heintz, N. (2024). Cell-type-specific CAG repeat expansions and toxicity of mutant Huntingtin in human striatum and cerebellum. Nat. Genet. *56*, 383–394. https://doi.org/10.1038/s41588-024-01653-6.

19. 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. https://doi.org/10.1016/j.cell.2024.11.038.

20. Donaldson, J., Powell, S., Rickards, N., Holmans, P., and Jones, L. (2021). What is the Pathogenic CAG Expansion Length in Huntington's Disease? J. Huntingtons Dis. *10*, 175–202. https://doi.org/10.3233/JHD-200445.

21. Hong, E.P., MacDonald, M.E., Wheeler, V.C., Jones, L., Holmans, P., Orth, M., Monckton, D.G., Long, J.D., Kwak, S., Gusella, J.F., and Lee, J.M. (2021). Huntington's Disease Pathogenesis: Two Sequential Components. J. Huntingtons Dis. *10*, 35–51. https://doi.org/10.3233/JHD-200427.

22. Ooi, J., Langley, S.R., Xu, X., Utami, K.H., Sim, B., Huang, Y., Harmston, N.P., Tay, Y.L., Ziaei, A., Zeng, R., et al. (2019). Unbiased Profiling of Isogenic Huntington Disease hPSC-Derived CNS and Peripheral Cells Reveals Strong Cell-Type Specificity of CAG Length Effects. Cell Rep. *26*, 2494–2508.e7. https://doi.org/10.1016/j.celrep.2019.02.008.

23. Maestri, S., Scalzo, D., Damaggio, G., Zobel, M., Besusso, D., and Cattaneo, E. (2025). Navigating triplet repeats sequencing: concepts, methodological challenges and perspective for Huntington's disease. Nucleic Acids Res. *53*, gkae1155. https://doi.org/10.1093/nar/gkae1155.

24. Cattaneo, E., Scalzo, D., Zobel, M., Iennaco, R., Maffezzini, C., Besusso, D., and Maestri, S. (2025). When repetita no-longer iuvant: somatic instability of the CAG triplet in Huntington's disease. Nucleic Acids Res. *53*, gkae1204. https://doi.org/10.1093/nar/gkae1204.

25. Chiu, R., Rajan-Babu, I.-S., Friedman, J.M., and Birol, I. (2021). Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. *22*, 224. https://doi.org/10.1186/s13059-021-02447-3.

26. Lee, J.-M., Zhang, J., Su, A.I., Walker, J.R., Wiltshire, T., Kang, K., Dragileva, E., Gillis, T., Lopez, E.T., Boily, M.-J., et al. (2010). A novel approach to investigate tissue-specific trinucleotide repeat instability. BMC Syst. Biol. *4*, 29. https://doi.org/10.1186/1752-0509-4-29.

27. McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. Nat. Rev. Genet. *11*, 786–799. https://doi.org/10.1038/nrg2828.

28. La Spada, A.R. (1997). Trinucleotide repeat instability: genetic features and molecular mechanisms. Brain Pathol. *7*, 943–963. https://doi.org/10.1111/j.1750-3639.1997.tb00895.x.

29. Shelbourne, P.F., Keller-McGandy, C., Bi, W.L., Yoon, S.-R., Dubeau, L., Veitch, N.J., Vonsattel, J.P., Wexler, N.S., US-Venezuela Collaborative Research Group; Arnheim, N., and Augood, S.J. (2007). Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Hum. Mol. Genet. *16*, 1133–1142. https://doi.org/10.1093/hmg/ddm054.

30. Telenius, H., Kremer, H.P., Theilmann, J., Andrew, S.E., Almqvist, E., Anvret, M., Greenberg, C., Greenberg, J., Lucotte, G., and Squitieri, F. (1993). 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. https://doi.org/10.1093/hmg/2.10.1535.

31. Aronin, N., Chase, K., Young, C., Sapp, E., Schwarz, C., Matta, N., Kornreich, R., Landwehrmeyer, B., Bird, E., and Beal, M.F. (1995). CAG expansion affects the expression of mutant Huntingtin in the Huntington's disease brain. Neuron *15*, 1193–1201. https://doi.org/10.1016/0896-6273(95)90106-x.

32. Nakamori, M., Panigrahi, G.B., Lanni, S., Gall-Duncan, T., Hayakawa, H., Tanaka, H., Luo, J., Otabe, T., Li, J., Sakata, A., et al. (2020). A slipped-CAG DNA-binding small molecule induces trinucleotide-repeat contractions in vivo. Nat. Genet. *52*, 146–159. https://doi.org/10.1038/s41588-019-0575-8.

33. Mouro Pinto, R., Arning, L., Giordano, J.V., Razghandi, P., Andrew, M.A., Gillis, T., Correia, K., Mysore, J.S., Grote Urtubey, D.-M., Parwez, C.R., et al. (2020). Patterns of CAG repeat instability in the central nervous system and periphery in Huntington's disease and in spinocerebellar ataxia type 1. Hum. Mol. Genet. *29*, 2551–2567. https://doi.org/10.1093/hmg/ddaa139.

34. Conforti, P., Bocchi, V.D., Campus, I., Scaramuzza, L., Galimberti, M., Lischetti, T., Talpo, F., Pedrazzoli, M., Murgia, A., Ferrari, I., et al. (2022). In vitro-derived medium spiny neurons recapitulate human striatal development and complexity at single-cell resolution. Cell Rep. Methods *2*, 100367. https://doi.org/10.1016/j.crmeth.2022.100367.

35. Dawson, J., Baine-Savanhu, F.K., Ciosi, M., Maxwell, A., Monckton, D.G., and Krause, A. (2022). A probable cis-acting genetic modifier of Huntington disease frequent in individuals with African ancestry. HGG Adv. *3*, 100130. https://doi.org/10.1016/j.xhgg.2022.100130.

36. Conforti, P., Besusso, D., Bocchi, V.D., Faedo, A., Cesana, E., Rossetti, G., Ranzani, V., Svendsen, C.N., Thompson, L.M., Toselli, M., et al. (2018). Faulty neuronal determination and cell polarization are reverted by modulating HD early phenotypes. Proc. Natl. Acad. Sci. *115*, E762–E771. https://doi.org/10.1073/pnas.1715865115.

37. Lim, R.G., Al-Dalahmah, O., Wu, J., Gold, M.P., Reidling, J.C., Tang, G., Adam, M., Dansu, D.K., Park, H.-J., Casaccia, P., et al. (2022). Huntington disease oligodendrocyte maturation deficits revealed by single-nucleus RNAseq are rescued by thiamine-biotin supplementation. Nat. Commun. *13*, 7791. https://doi.org/10.1038/s41467-022-35388-x.

38. Ring, K.L., An, M.C., Zhang, N., O'Brien, R.N., Ramos, E.M., Gao, F., Atwood, R., Bailus, B.J., Melov, S., Mooney, S.D., et al. (2015). Genomic Analysis Reveals Disruption of Striatal Neuronal Development and Therapeutic Targets in Human Huntington's Disease Neural Stem Cells. Stem Cell Rep. *5*, 1023–1038. https://doi.org/10.1016/j.stemcr.2015.11.005.

39. Mehta, S.R., Tom, C.M., Wang, Y., Bresee, C., Rushton, D., Mathkar, P.P., Tang, J., and Mattis, V.B. (2018). Human Huntington's Disease iPSC-Derived Cortical Neurons Display Altered Transcriptomics, Morphology, and Maturation. Cell Rep. *25*, 1081–1096.e6. https://doi.org/10.1016/j.celrep.2018.09.076.

40. 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. USA *110*, 2366–2370. https://doi.org/10.1073/pnas.1221891110.

41. Landles, C., Osborne, G.F., Phillips, J., Canibano-Pico, M., Nita, I.M., Ali, N., Bobkov, K., Greene, J.R., Sathasivam, K., and Bates, G.P. (2024). Mutant huntingtin protein decreases with CAG repeat expansion:

implications for therapeutics and bioassays. Brain Commun. *6*, fcae410. https://doi.org/10.1093/braincomms/fcae410.

42. Hoschek, F., Natan, J., Wagner, M., Sathasivam, K., Abdelmoez, A., von Einem, B., 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. https://doi.org/10.1186/s10020-024-00801-2.
43. Conforti, P., Besusso, D., Brocchetti, S., Campus, I., Cappadona, C., Galimberti, M., Laporta, A., Iennaco, R., Rossi, R.L., Dickinson, V.B., and Cattaneo, E. (2020). RUES2 hESCs exhibit MGE-biased neuronal differentiation and muHTT-dependent defective specification hinting at SP1. Neurobiol. Dis. *146*, 105140. https://doi.org/10.1016/j.nbd.2020.105140.
44. Alcalá-Vida, R., Garcia-Forn, M., Castany-Pladevall, C., Creus-Muncunill, J., Ito, Y., Blanco, E., Golbano, A., Crespí-Vázquez, K., Parry, A., Slater, G., et al. (2021). Neuron type-specific increase in lamin B1 contributes to nuclear dysfunction in Huntington's disease. EMBO Mol. Med. *13*, e12105. https://doi.org/10.15252/emmm.202012105.
45. Coffinier, C., Jung, H.-J., Nobumori, C., Chang, S., Tu, Y., Barnes, R.H., Yoshinaga, Y., De Jong, P.J., Vergnes, L., Reue, K., et al. (2011). Deficiencies in lamin B1 and lamin B2 cause neurodevelopmental defects and distinct nuclear shape abnormalities in neurons. Mol. Biol. Cell *22*, 4683–4693. https://doi.org/10.1091/mbc.e11-06-0504.
46. Gasset-Rosa, F., Chillon-Marinas, C., Goginashvili, A., Atwal, R.S., Artates, J.W., Tabet, R., Wheeler, V.C., Bang, A.G., Cleveland, D.W., and Lagier-Tourenne, C. (2017). Polyglutamine-Expanded Huntingtin Exacerbates Age-Related Disruption of Nuclear Integrity and Nucleocytoplasmic Transport. Neuron *94*, 48–57.e4. https://doi.org/10.1016/j.neuron.2017.03.027.
47. Irmak, D., Fatima, A., Gutiérrez-Garcia, R., Rinschen, M.M., Wagle, P., Altmüller, J., Arrigoni, L., Hummel, B., Klein, C., Frese, C.K., et al. (2018). Mechanism suppressing H3K9 trimethylation in pluripotent stem cells and its demise by polyQ-expanded huntingtin mutations. Hum. Mol. Genet. *27*, 4117–4134. https://doi.org/10.1093/hmg/ddy304.
48. Padeken, J., Methot, S.P., and Gasser, S.M. (2022). Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. Nat. Rev. Mol. Cell Biol. *23*, 623–640. https://doi.org/10.1038/s41580-022-00483-w.
49. Keenan, C.R. (2021). Heterochromatin and Polycomb as regulators of haematopoiesis. Biochem. Soc. Trans. *49*, 805–814. https://doi.org/10.1042/BST20200737.
50. Becker, J.S., Nicetto, D., and Zaret, K.S. (2016). H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet. *32*, 29–41. https://doi.org/10.1016/j.tig.2015.11.001.
51. Becker, J.S., McCarthy, R.L., Sidoli, S., Donahue, G., Kaeding, K.E., He, Z., Lin, S., Garcia, B.A., and Zaret, K.S. (2017). Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. Mol. Cell *68*, 1023–1037.e15. https://doi.org/10.1016/j.molcel.2017.11.030.
52. von Schimmelmann, M., Feinberg, P.A., Sullivan, J.M., Ku, S.M., Badimon, A., Duff, M.K., Wang, Z., Lachmann, A., Dewell, S., Ma'ayan, A., et al. (2016). Polycomb repressive complex 2 (PRC2) silences genes responsible for neurodegeneration. Nat. Neurosci. *19*, 1321–1330. https://doi.org/10.1038/nn.4360.
53. Biagioli, M., Ferrari, F., Mendenhall, E.M., Zhang, Y., Erdin, S., Vijayvargia, R., Vallabh, S.M., Solomos, N., Manavalan, P., Ragavendran, A., et al. (2015). Htt CAG repeat expansion confers pleiotropic gains of mutant huntingtin function in chromatin regulation. Hum. Mol. Genet. *24*, 2442–2457. https://doi.org/10.1093/hmg/ddv006.
54. Pearl, J.R., Shetty, A.C., Cantle, J.P., Bergey, D.E., Bragg, R.M., Coffey, S.R., Kordasiewicz, H.B., Hood, L.E., Price, N.D., Ament, S.A., and Carroll, J.B. (2025). Altered huntingtin–chromatin interactions predict transcriptional and epigenetic changes in Huntington's disease. Dis. Model. Mech. *18*, dmm052282. https://doi.org/10.1242/dmm.052282.
55. Malaiya, S., Cortes-Gutierrez, M., Herb, B.R., Coffey, S.R., Legg, S.R.W., Cantle, J.P., Colantuoni, C., Carroll, J.B., and Ament, S.A. (2021). Single-Nucleus RNA-Seq Reveals Dysregulation of Striatal Cell Identity Due to Huntington's Disease Mutations. J. Neurosci. *41*, 5534–5552. https://doi.org/10.1523/JNEUROSCI.2074-20.2021.
56. Mouro Pinto, R., Murtha, R., Azevedo, A., Douglas, C., Kovalenko, M., Ulloa, J., Crescenti, S., Burch, Z., Oliver, E., Kesavan, M., et al. (2025). In vivo CRISPR–Cas9 genome editing in mice identifies genetic modifiers of somatic CAG repeat instability in Huntington's disease. Nat. Genet. *57*, 314–322. https://doi.org/10.1038/s41588-024-02054-5.
57. Ferguson, R., Goold, R., Coupland, L., Flower, M., and Tabrizi, S.J. (2024). Therapeutic validation of MMR-associated genetic modifiers in a human *ex vivo* model of Huntington disease. Am. J. Hum. Genet. *111*, 1165–1183. https://doi.org/10.1016/j.ajhg.2024.04.015.
58. Bunting, E.L., Donaldson, J., Cumming, S.A., Olive, J., Broom, E., Micläuș, M., Hamilton, J., Tegtmeyer, M., Zhao, H.T., Brenton, J., et al. (2025). Antisense oligonucleotide–mediated MSH3 suppression reduces somatic CAG repeat expansion in Huntington's disease iPSC–derived striatal neurons. Sci. Transl. Med. *17*, eadn4600. https://doi.org/10.1126/scitranslmed.adn4600.
59. Crook, Z.R., and Housman, D. (2011). Huntington's Disease: Can Mice Lead the Way to Treatment? Neuron *69*, 423–435. https://doi.org/10.1016/j.neuron.2010.12.035.
60. Khristich, A.N., and Mirkin, S.M. (2020). On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J. Biol. Chem. *295*, 4134–4170. https://doi.org/10.1074/jbc.REV119.007678.
61. Choi, D.E., Shin, J.W., Zeng, S., Hong, E.P., Jang, J.-H., Loupe, J.M., Wheeler, V.C., Stutzman, H.E., Kleinstiver, B., and Lee, J.-M. (2024). Base editing strategies to convert CAG to CAA diminish the disease-causing mutation in Huntington's disease. eLife *12*, RP89782. https://doi.org/10.7554/eLife.89782.
62. Matuszek, Z., Arbab, M., Kesavan, M., Hsu, A., Roy, J.C.L., Zhao, J., Yu, T., Weisburd, B., Newby, G.A., Doherty, N.J., et al. (2025). Base editing of trinucleotide repeats that cause Huntington's disease and Friedreich's ataxia reduces somatic repeat expansions in patient cells and in mice. Nat. Genet. *57*, 1437–1451. https://doi.org/10.1038/s41588-025-02172-8.
63. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319. https://doi.org/10.1038/nbt.3820.
64. Patel, H., Ewels, P., Manning, J., Garcia, M.U., Peltzer, A., Hammarén, R., Botvinnik, O., Talbot, A., Sturm, G., bot, nf-core, et al. (2024). nf-core/rnaseq: nf-core/rnaseq v3.17.0 - Neon Newt. (Zenodo). https://doi.org/10.5281/zenodo.13986791.
65. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
66. Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2023). Biostrings: Efficient manipulation of biological strings. Version 2.66.0 (Bioconductor version: Release (3.16)). https://doi.org/10.18129/B9.bioc.Biostrings.
67. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. *7*, 203–214. https://doi.org/10.1089/10665270050081478.
68. Sebestyén, E., Marullo, F., Lucini, F., Petrini, C., Bianchi, A., Valsoni, S., Olivieri, I., Antonelli, L., Gregoretti, F., Oliva, G., et al. (2020). SAMMY-seq reveals early alteration of heterochromatin and deregulation of bivalent genes in Hutchinson-Gilford Progeria Syndrome. Nat. Commun. *11*, 6274. https://doi.org/10.1038/s41467-020-20048-9.
69. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
70. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

71. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

72. Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

73. Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics *28*, 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

74. Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. *20*, 631–656. https://doi.org/10.1038/s41576-019-0150-2.

75. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

76. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

77. Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. *4*, 1521. https://doi.org/10.12688/f1000research.7563.2.

78. Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., van den Brand T, et al. (2023). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. Version 3.4.2. https://CRAN.R-project.org/package=ggplot2.

79. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. *20*, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248.

80. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

81. Kolde, R. (2019). pheatmap: Pretty Heatmaps. Version 1.0.12.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal anti-OCT3/4 (C-10)	Santa Cruz Biotechnology	Cat#sc-5279; RRID:AB_628051
Rabbit anti-SOX2	Millipore	Cat# AB5603; RRID:AB_2286686
Rabbit βIII-TUBULIN	BioLegend	Cat# 802001; AB_2564645
Mouse monoclonal anti-MAP2 Clone Ap20	BD Biosciences	Cat# 556320; RRID:AB_396359
Rabbit polyclonal anti-Ki67	Abcam	Cat#ab15580; RRID:AB_443209
Mouse anti-p27	Cell Signaling	Cat#3698; RRID:AB_2077832
Rabbit anti GABA	Sigma	Cat#A2052; RRID:AB_477652
Rat monoclonal anti-CTIP2 [25B6]	Abcam	Cat#ab18465; RRID:AB_2064130
Rabbit monoclonal anti-DARPP32 [EP720Y]	Abcam	Cat# ab40801; RRID:AB_731843
Mouse monoclonal anti-GAD67	Millipore	Cat# MAB5406; RRID:AB_2278725
Rabbit monoclonal EZH2 (D2C9)	Cell Signaling	Cat# 5246;RRID:AB_10694683
Rabbit H3K9me3	Abcam	Cat# ab8898; RRID:AB_306848
Mouse monoclonal Lamin b1	Merk	Cat# amab91251; N/A
Rabbit monoclonal Huntingtin (D7F7)	Cell Signaling	Cat# 5656;RRID:AB_10827977
Rabbit EZH2	Cell Signaling	Cat# 4905; RRID:AB_2278249
Rabbit Polyclonal SUZ12	Active Motif	Cat# 39357; RRID: AB_2614929
EED	Abcam	Cat# ab240650; RRID: AB_2922803
Rabbit polyclonal H3K27Me3	Millipore	Cat# 07–449; RRID: AB_310624
Rabbit polyclonal H3	Abcam	Cat# ab1791; N/A
Mouse Monoclonal Vinculin	Sigma	Cat# V9131; RRID: AB_477629
AlexaFluor Goat Anti-Rabbit 488	Life Technologies	Cat#A11008; RRID:AB_143165
AlexaFluor Goat Anti-Mouse 568	Life Technologies	Cat#A11004; RRID:AB_2534072
AlexaFluor Goat Anti-Rat 647	Life Technologies	Cat#A21247; RRID:AB_141778
HRP-conjugated Goat anti-Rabbit IgG	Bio-Rad	Cat#170–6515; RRID:AB_11125142
HRP-conjugated Goat anti-Mouse IgG	Bio-Rad	Cat#170–6516; RRID:AB_11125547
Chemicals, peptides, and recombinant proteins		
B27 Supplement	Life Technologies	17504–044
B27 w/o Vit A supplement	Life Technologies	12587–010
brain-derived neurotrophic factor (BNDF)	PeproTech	450–02
Cell Culture water	Sigma Aldrich	W4502-1L
Cultrex	Bio-Techne	343201001
DKK-1	PeproTech	120–30
DMEM/F12	Life Technologies	21331–020
Dulbecco's PBS w/o Calcium w/o Magnesium (PBS)	Euroclone	ECB40041
EDTA 0.5M pH 8.0	Millipore	324506
ESGRO Complete Accutase	Millipore	SF006
GlutaMAX (100×)	Life Technologies	35050–38
HaltTM Protease and Phosphatase Inhibitor Cocktail 1 mM	Thermo Fisher Scientific	78440
Hoechst 33342	Invitrogen	H3570
LDN	CHDI Foundation	00396388-0001-006
mTeSRTM1 basal medium	STEMCELL Technologies	85851
mTeSRTM1 Supplement (5×)	STEMCELL Technologies	85851

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Normal Goat Serum (NGS)	Vector Laboratories	S-1000
N2 Supplement	Life Technologies	17502–048
Penicillin/Streptomycin solution (100×)	Euroclone	ECB3001D
PMSF 1mM	Sigma Aldrich	P7626
ProLong Diamond	ThermoFisher Scientific	P36961
Y-27632 (ROCKi)	CHDI Foundation	00197406-0001-007
SB431542	CHDI Foundation	00447536-0000-002
SHH C-25 II	R&D System	464-SH-2MG
Trypan Blue Solution	Life Technologies	15250–06
TRIzolTM Reagent	Life Technologies	15596018
SCR7	TOCRIS	14892-97-8
G418	ThermoFisher	10131027
SPRIselect beads	Beckman Coulter	B23318
Paraformaldehyde (PFA)	Sigma-Aldrich	P6148
Critical commercial assays		
Clarity Western ECL Substrate	Bio-Rad	1705061
DNA-freeTM DNase Treatment and Removal	Thermo Fisher Scientific	AM1906
iScript cDNA Synthesis Kit Bio-Rad 1708891	Bio-Rad	1708891
NucleoSpin Tissue® kit	Machery-Nagel	740952
Human Stem Cell Nucleofector® Kit	Lonza	VPH-5012
PrimeSTAR® HS DNA Polymerase with GC Buffer	Takara	R044A
Phusion High–Fidelity DNA Polymerase	Thermo Fisher	F530S
PrimeStar Max Polymerase	Takara	R045A
Q5® High-Fidelity DNA Polymerase	NEB	M0491
Ligation sequencing kit	ONT	SQK-LSK110
Native barcoding kit	ONT	EXP-NBD104 EXP-NBD114
Pierce BCA Protein Assay Kit	Thermo Fisher Scientific	23225
SsoFast EvaGreen" Supermix	Bio-Rad	172–5202
Deposited data		
Raw genomic sequencing data	NCBI SRA	BioProject PRJNA1074134 https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1074134
Raw transcriptomic sequencing data	NCBI SRA	BioProject PRJNA1074134 https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1074134
Experimental models: Cell lines		
Human: H9 (WA-09) hESC line	WiCell Research Institute	NIHhESC-10-0062
H9-RMCE parental 107Q/$(CAG)_{105}$-CAACAG	This paper	N/A
H9-21Q/$(CAG)_{19}$-CAACAG	This paper	N/A
H9-45Q/$(CAG)_{43}$-CAACAG	This paper	N/A
H9-81Q/$(CAG)_{78}$-CAACAG	This paper	N/A
H9-107Q CAACAG-dup	This paper	N/A
H9-107Q multi-CAAs	This paper	N/A
H9-83Q CAACAG-dup	This paper	N/A
H9-43Q CAA-loss	This paper	N/A
H9-107Q CAA-loss	This paper	N/A
Experimental models: Organisms/strains		
Primers sequence for PCR and qRT-PCR see Table S2	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Recombinant DNA		
pCAG-Flpe:GFP	Addgene	13788
pCAG-Cre:GFP	Addgene	13776
pUC57_HDR_RMCE_$(CAG)_{19}$-CAACAG/21Q_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$(CAG)_{43}$-CAACAG/45Q_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$(CAG)_{79}$-CAACAG/81Q_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$(CAG)_{105}$-CAACAG/107Q_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_CAG_{107}/107Q_LOI_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$(CAG)_{105}(CAACAG)_2$/ 107Q_DUP_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$([(CAG)_{21}CAA]_4(CAG)_{17}$ CAACAG)/107Q_multi CAAs_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$(CAG)_{78}(CAACAG)_2$ CAACAG/81Q_DUP_PuroR	Genscript	custom DNA sinthesis
pUC57_HDR_RMCE_$(CAG)_{45}$/43Q_LOI_PuroR	Genscript	custom DNA sinthesis
Software and algorithms		
nf-core rnaseq v3.14.0-gb89fac3 Nextflow pipeline	Di Tommaso et al.[63] Patel et al.[64]	https://doi.org/10.5281/zenodo.13986791
DESeq2 package v1.12.3 and Dependencies	Love et al.[65]	t http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html.
R 4.3.1	N/A	https://cran.r-project.org/bin/windows/base/old/4.3.1/
GenomicAlignments v1.36.0	Pagès et al.[66]	https://doi.org/10.18129/B9.bioc.Biostrings
Guppy v5.0.7	ONT proprietary software	N/A
Straglr v1.2.0	Chiu et al.[25]	https://github.com/bcgsc/straglr
BLASTN v2.12	Zhang et al.[67]	N/A
FIJI - ImageJ	N/A	https://fiji.sc/
CFX Manager Software	N/A	Bio-Rad
GraphPad Prism	N/A	https://www.graphpad.com/
NIS-elements Analysis softwere	Nikon	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

hES cell culture

Human embryonic stem (hES) H9 female cell line (WiCell) and derived edited clones were cultured on cultrex (80–120 μg/mL, Bio-Techne)-coated dishes in complete mTeSR1 medium (STEMCELL Technologies) plus 1× Penicillin/Streptomycin solution (Euroclone). The medium was changed daily, and cells were dissociated twice a week with PBS (Euroclone) plus 0.5 mM EDTA (Millipore) for passaging. hES H9 and derived edited cells were checked by Q-banding analyses (BioKryo, Italia) to evaluate the karyotype. The lines were regularly tested and maintained mycoplasma-free (Eurofins Genomics).

METHOD DETAILS

Plasmid donor library and sgRNA design

The donor plasmids pUC57_HDR_RMCE_(CAG)19-CAACAG/21Q_PuroR, pUC57_HDR_RMCE_(CAG)43-CAACAG/45Q_PuroR, pUC57_HDR_RMCE_(CAG)79-CAACAG/81Q_PuroR and pUC57_HDR_RMCE_(CAG)105-CAACAG/107Q_PuroR were synthesized by GenScript. PuroR was substituted by NeoR with cut and paste cloning strategy using SpeI/NotI restriction enzyme (NEB). Similarly, the donor plasmids pUC57_HDR_RMCE_CAG107/107Q_LOI_PuroR, pUC57_HDR_RMCE_(CAG)105(CAACAG)2/107Q_DUP_PuroR, pUC57_HDR_RMCE_([(CAG)21CAA]4(CAG)17CAACAG)/107Q_multiCAAs_PuroR, pUC57_HDR_RMCE_(CAG)78(CAACAG)2CAACAG/81Q_DUP_PuroR and pUC57_HDR_RMCE_(CAG)45/43Q_LOI_PuroR were synthesized by GenScript. All pUC57 plasmids contains the HTT exon1 variant followed by a construct-specific barcode and the antibiotic resistance cassette (neomycin or puromycin resistance gene under the control of hPGK promotor) enclosed by FRT sites. All the genetic elements are enclosed between WT *LoxP* and the *LoxP*2272 sites placed in the same orientation. Outside *LoxP* sites, homology arms were designed to drive Cas9-assisted targeting by homologous directed repair (HDR). gRNA2 (GTGTGAGGCAGAACCTGCGG) and gRNA1i (GGCAC

TTAAAACAGCCT) used for the integration of pUC57_HDR_107Q_PuroR donor plasmid were synthetized by IDT. Alt-R CRISPR-Cas9 tracrRNA (IDT 1072534) and the two designed Alt-R CRISPR-Cas9 crRNAs were assembled individually in duplex buffer (IDT #11050112) for 5 min at 95°C followed by slow cool down to RT.

Generation of H9-RMCE parental cell line

H9 cells were detached in AccutaseTM (Millipore) following the manufacturer's instruction. 2 × 106 cells were nucleofected with 1 μg of pUC57_HDR_RMCE_107Q_PuroR donor plasmid and 20 pmol of equimolar Alt-R S.p. HiFi Cas9 Nuclease V3 (IDT # 1081060) and sgRNA RNP complex using Nucleofector kit (Lonza), in Nucleofector II (Amaxa biosystems) using B-016 program. After nucleofection, cells were plated in mTeSR media with 10 μM RI (Y-27632, provided by CHDI foundation) and 1 μM of SCR7 (TOCRIS) for 24 h. Two days after nucleofection, cells were daily treated with 0.1 mg/mL G418 (ThermoFisher) for 5 days to select and isolate recombinant cells. The presence of RMCE cassette was validated by LHA-PCR, Geno-PCR and RHA-PCR (see Table S2 and "Genomic DNA extraction and PCR" paragraph).

To eliminate potential confounding effects of the antibiotic selection cassette, 2 × 106 targeted H9 cells were nucleofected with 6 μg of pCAG-Flpe:GFP (Addgene) by using Nucleofector kit (Lonza) using B-016 program. Two days after nucleofection, cells were single sorted in five 96-well plates to isolate individual clones using a FACSaria III SORP cell sorter (BD Biosciences).

The following quality controls were performed: (i) the integration of the RMCE cassette was assessed by LHA-PCR; (ii) the allele-specificity of the targeting was assessed using the genotype-PCR; (iii) the absence of the antibiotic-resistance cassette was confirmed by RHA-PCR; (iv) the absence of plasmid random insertions was assessed by RI-LHA, RI-RHA, RI-AMP, RI-ORI PCRs mapping on distinct portions of donor and general vector sequences not involved in the recombination process (see Table S2 and "Genomic DNA extraction and PCR" paragraph); (v) the functionality of the HTT locus was tested by qRT-PCR; (vi) maintenance of cell pluripotency was evaluated by analyzing OCT4, SOX2, NANOG transcript and protein expression by qRT-PCR (see Table S2 and "RNA isolation and qRT-PCR" paragraph) and immunofluorescence staining (see "Immunofluorescence and confocal imaging" paragraph). Three RMCE-H9 master cell lines were isolated upon targeting with the (CAG)105-CAACAG HTT exon1 modification. The resulting clones carry an unmodified wild type allele is 17 CAG.

Generation of the CAGinSTEM platform with increasing CAG length and/or modified CAG stretch composition within HTT exon1

To generate the entire platform, 2 × 106 cells of the RMCE-H9 parental clones (#01, #04, #25) were nucleofected with 1 μg of pUC57_HDR_RMCE plasmid carrying a specific exon1 variant and 1 μg of pCAG-Cre:GFP (Addgene) by using Nucleofector kit (Lonza), in Nucleofector II (Amaxa biosystems) using the B-016 program. Antibiotic selection was used to enrich for successful transformants by daily treating cells with 0.1 mg/mL G418 or 1 μg/mL puromycin for 5 days. To eliminate potential confounding effects of the antibiotic selection cassettes, 2 × 106 targeted cells were nucleofected with 6 μg of pCAG-Flpe:GFP (Addgene) by using Nucleofector kit (Lonza) in Nucleofector II by using B-016 program, for antibiotic cassette removal. Two days after nucleofection, cells were single sorted into five 96-well plates to isolate individual clones using a FACSaria III SORP cell sorter. Surviving cells (on average 60 clones) were amplified and screened based on the same quality controls described for the isolation of the RMCE-master cell lines. The described workflow was performed for all three previously established H9 -RMCE master cell lines to generate each HTT variant.

gDNA extraction and PCR

gDNA extraction was performed using a NucleoSpin Tissue kit (Machery-Nagel #740952) according to the manufacturer's instructions. The quality and concentration of all extracted DNA samples were verified by spectroscopic analysis (NanoDrop 1000—ThermoScientific).

The LHA-PCR reaction was performed in a total volume of 25 μL with 100 ng of genomic DNA using PrimeSTAR HS DNA Polymerase with GC Buffer (Takara #R044A) following manufacturer instructions. The amplification consisted of 30 cycles of 10 s at 98°C, 5 s at 55°C, and 17 s at 72°C.

The Geno-PCR reaction was performed in a total volume of 20 μL with 50 ng of genomic DNA using Phusion High–Fidelity DNA Polymerase (Thermo Fisher #F530S) following manufacturer's instructions with GC buffer. The amplification consisted of 3′ minutes at 98°C; 30 cycles of 10 s at 98°C, 30 s at 70°C, 5 s at 72°C; 10 min at 72°C.

In the case of pUC57_HDR_RMCE plasmids carrying NeoR, the RHA-PCR was performed with F3polyA_FW and HTT3'_RV2 primers in a total volume of 25 μL with 50 ng of genomic DNA using PrimeStar Max Polymerase (Takara #R045A) following manufacturer's instructions. The amplification consisted of 30 cycles of 10 s at 98°C, 5 s at 55°C, 15 s at 72°C.

In the case of pUC57_HDR_RMCE plasmids carrying PuroR cassette resistance, the RHA PCR was performed with HTT3'_RV2 and PURO_RV2 primers in a total volume of 25 μL with 50 ng of genomic DNA using PrimeStar Max Polymerase (Takara #R045A) following manufacturer's instructions. The amplification consisted of 28 cycles of 10 s at 98°C, 5 s at 55°C, 17 s at 72°C.

The ONT-PCR was performed in a total volume of 25 μL with 50 ng of genomic DNA using PrimeStar Max Polymerase (Takara #R045A) following manufacturer instructions with the addition of 1M Betaine (Sigma #14300). The amplification consisted of 30 cycles of 10 s at 98°C, 5 s at 55°C, 17 s at 72°C.

The RI-LHA PCR and the RI-RHA PCR were performed in a total volume of 25 μL with 50 ng of genomic DNA using Q5 High-Fidelity DNA Polymerase (NEB #M0491) following manufacturer's instructions. The amplification consisted of 30 s at 98°C; 30 cycles of 10 s at 98°C, 30 s at 61°C, 15 s at 72°C; 2 min at 72°C.

The RI-AMP PCR and the RI-ORI PCR were performed in a total volume of 15 μL with 50 ng of genomic DNA using Q5 High-Fidelity DNA Polymerase (NEB #M0491) following manufacturer's instructions. The amplification consisted of 30s seconds at 98°C; 30 cycles of 10 s at 98°C, 30 s at 67°C, 30 s at 72°C; 2 min at 72°C.

PCR products quality and length were assessed by gel electrophoresis. PCR products were sequenced using Sanger technology (Eurofins Genomics) when required. For all the primers sequences used, see Table S2.

PCR clean-up and ONT library preparation

For sequencing using ONT, each PCR amplicon was cleaned-up using a 0.7× concentration of SPRIselect beads (Beckman Coulter, #B23318) to size select the amplicons of interest and get rid of amplicons shorter than ~2000bp. Amplicons were then transferred with a 5× concentration of QIAGEN Buffer PB (QIAGEN, #19066) in a silica membrane column (ZYMO RESEARCH, #C1004) and centrifugated for a minute at 11.000g. After two washes of DNA using Wash solution (ZYMO RESEARCH, #D4003-2), final eluates in nuclease-free water were obtained. Size selected amplicons were then quantified using the Qubit 4 Fluorometer (Invitrogen, #Q33238) coupled with the Qubit1X dsDNA high sensitivity assay kit (Invitrogen, #Q33230). 200 fmol of size selected amplicons were processed for ONT sequencing according to manufacturer's instructions. Briefly, the amplicons library was prepared using the Ligation sequencing kit (ONT #SQK-LSK110) combined with the Native barcoding kit (ONT #EXP-NBD104, #EXP-NBD114) to multiplex up to 12 amplicons. ONT library was loaded on R9.4.1 flow cells following the manufacturer's instructions, and sequencing was performed for 72h. For each R9.4.1 flow cell, the availability of sufficient active pores was confirmed on arrival and directly before the sequencing run.

Striatal differentiation

The striatal differentiation protocol was performed according to Conforti et al. 2022[34] with minor modifications. Briefly, 80% confluent cells were dissociated using Accutase (Millipore #SF006), counted and plated on cultrex-coated dishes at low confluency (10′000 cells/cm^2) in complete mTeSR1 medium supplemented with 10μM ROCK inhibitor (Y-27632, provided by CHDI foundation).

After two days of expansion, cells were exposed to Dual-SMAD inhibition for neuronal induction for 12 days (10μM SB431542, 500nM LDN193189 (provided by CHDI foundation), N2 Supplement (Life Technologies # 17502-048), B27 Supplement without (Life Technologies), 1× Penicillin/Streptomycin solution (Euroclone ECB3001D), 1× GlutaMAX solution (Life Technologies #35050-38) in DMEM/F12 (Life Technologies # 21331-020). On DIV 7, cells were dissociated with PBS plus 0.5 mM EDTA (Millipore #324506) and re-plated in a 1:2 dilution on cultrex (120–180 μg/mL)-coated dishes by supplementing the medium with 10μM ROCK inhibitor. Starting on DIV12 until DIV25 medium was supplemented with 200 ng/mL recombinant human SHH C-25 II (R&D System # 464-SH-2MG) and 100 ng/mL DKK-1 (PrepoTech # 120-30). On DIV21 cells were detached upon Accutase dissociation and re-plated at a cell density of 2 × 10^4 cells/cm^2 cell density on cultrex 160–240 μg/mL -coated dishes. Cells were maintained, in DMEM/F12 plus N2 Supplement, B27 Supplement with retinoic acid (Life Technologies # 17504-044), 20 ng/mL BDNF (PrepoTech #450-02), recombinant human SHH C-25 II (R&D System # 464-SH-2MG) and 100 ng/mL DKK-1 (PrepoTech # 120-30) by partially change medium every three days until day 25. Human SHH C-25 II and DKK-1 were then removed starting from day 26 until the end of the differentiation protocol.

Immunofluorescence and confocal imaging

For ICC cells were fixed for 15 min in ice-cold 4% paraformaldehyde (PFA) (Sigma-Aldrich, #P6148) in sodium phosphate (PBS) buffer, pH 7.4 followed by three washes in PBS. Coverslips were then permeabilized with 0.5%Triton X-100 in PBS for 10 min and incubated in blocking solution (0.25% Triton X-100, 2.5% Normal Goat Serum +0.5mg/mL BSA +2.5%FBS in PBS) for 1 h at room temperature (RT). Cells were incubated with primary antibodies against OCT4 (Santa cruz #5279, 1:100), SOX2 (Millipore #ab5603; 1:200), βIII-TUBULIN (BioLegend, cat. n. 802001, 1:1,000), MAP2 (Becton Dickinson, cat. n. 556320, 1:500), Ki67 (Abcam #15580; 1:5000); p27 (Cell signaling #3698; 1:300); GABA (Sigma, cat. n. A2052, 1:500), CTIP2 (Abcam #ab18465, 1:1000), GAD67 (Millipore #MAB5406, 1:1000), DARPP32 (Abcam #ab40801, 1:250), EZH2 (Cell Signaling #5246, 1:500) and H3K9me3 (Abcam #ab8898, 1:500) overnight at +4°C, or primary antibody against lamininB1 (Merk #amab91251, 1:500) 3 h RT. Then, cells were incubated for 1 h at RT with the appropriate fluorophore-conjugated secondary antibody (Alexa Fluor, Invitrogen, 1:500) and 30 min with 0.1 μg/mL Hoechst (Invitrogen, cod. 33342) to counterstain nuclei. All the antibodies were diluted in blocking solution. Finally, coverslips were mounted with ProLong Diamond (ThermoFisher Scientific #P36961) after washing three times for 30 min with PBS. Pictures of hESCs and differentiated neurons were captured using a Leica TCS SP5 Confocal Laser Scanning Microscope (Leica Microsystems) equipped with a 40× (NA 1.4) oil immersion. The percentage of cells positive for striatal markers DARPP32, CTIP2, and GAD67 was computed manually using the cell counter feature of ImageJ.[34] The quantification of the PcG foci and LaminB was performed with an automated pipeline described in.[68] After loading the Hoechst- or LaminB- stained nuclei and PcGs immunofluorescence, images were converted into grayscale. The measurement of the PcGs intensity, size, and shape was performed after nuclei segmentation by the IdentifyPrimaryObject algorithm. Partial nuclei at the image borders were discarded.

Quantification of Ki67 and p27 positive cells was performed with NIS-element. Confocal images acquired at 40× magnification with 2× digital zoom were analyzed through the General Analysis tool, using a custom pipeline optimized for parameters such as background and object size. The same pipeline was applied uniformly across all samples. The number of Ki67 and p27 positive cells was then normalized over the total number of Hoechst 33342 positive nuclei.

Quantification of the GABA area, based on GABA/MAP2 staining, was performed with NIS-element. Confocal images taken at 40× were automatically quantified by a general analysis tool employing a custom pipeline set on parameters such as background and dimension. The same pipeline was used for all the sub-clones. The area covered by GABA positive cells was then normalized to the area covered by MAP2 positive cells normalized over the total number of Hoechst 33342 positive nuclei.

Quantification of total dendritic length and soma area, based on MAP2 staining, was performed manually with Fiji in blind for genotypes.

For the presentation of confocal images, multiple focal planes with z-spacing of 0.2–0.5 μm were flattened by ImageJ maximum projection, and contrast was enhanced by linear methods using ImageJ (NIH, Bethesda, MD, USA).

Western blotting

hES cells were collected and homogenized in RIPA buffer (Tris-HCl pH8 50 mM, NaCl 150 mM, SDS 0.1%, NP40 1%) supplemented with PMSF 1mM (Sigma-Aldrich) and Halt Protease & Phosphatase Inhibitor Cocktail 1mM (Thermo Fisher Scientific). Homogenized cells were incubated for 10 min at +4°C and then centrifuged at 13,500 g for 20 min at +4°C. Total amount of protein extracts was quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific). 50 μg of proteins were loaded per track onto a 5% polyacrylamide gel for Huntingtin. 30 μg of proteins were loaded for PRC2 subunits and H3K27Me3 onto a 7.5% and 12% polyacrylamide gel respectively. Proteins were transferred onto a nitrocellulose membrane using the Trans-Blot TurboSystem (Bio-Rad) and blocked in 5% BSA (Merck Life Science A3059; dissolved in TBS-0.1% Tween-10). Nitrocellulose membranes were immunoprobed with primary antibody against: Huntingtin (Cell Signaling #5656; 1:5000), EZH2 (Cell Signaling #4905; 1:1000), SUZ12 (Active Motif 39357; 1:1000), EED (Abcam ab240650; 1:1000), H3K27Me3 (Millipore #07–449; 1:1000), H3 (Abcam ab1791; 1:5000) overnight at 4°C. Appropriate HRP-conjugated secondary antibody was detected using Clarity Western ECL Substrate (Bio-Rad). Vinculin (Sigma #V9131; 1: 5000) was used as housekeeping gene. Protein bands were detected by the Chemidoc MP imaging system (Bio-Rad) and densitometric analysis were performed using 'Analyze gels' plugin of ImageJ analysis software (NIH, Bethesda, MD, USA).

RNA isolation and qRT-PCR

Total RNA from hES cells and terminal differentiated neurons was isolated with TRIzol reagent (Life Technologies #15596018) according to the manufacturer's instructions. The integrity of the purified RNA and the absence of genomic DNA contamination were assessed by non-denaturing agarose gel electrophoresis. In the presence of genomic DNA contamination, RNA extracts were treated with the DNA-free DNase Treatment and Removal kit (Invitrogen AM1906). 500ng of total RNA was reverse transcribed with iScript cDNA Synthesis Kit (Bio-Rad #1708891) following the manufacturer's instructions to produce cDNA. Quantitative real time (qRT)-PCR was performed using a CFX96TM Real-Time System (Bio-Rad) and analyzed with the CFX Manager Software (Bio-Rad). All reactions were performed in 15 μL containing 50 ng cDNA and SsoFastTM EvaGreen Supermix (Bio-Rad 1725204). Primer pairs used for *HTT*, *OCT4*, *SOX2* and *NANOG* amplification are reported in Table S2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Graphs and statistical analysis were performed by GraphPad Prism and R v4.3.1 software. First, outlier identification was performed and outlier values were excluded when present. Then, the Shapiro–Wilk normality test was performed to assess the normal distribution of data. When normally distributed, the data were analyzed using an unpaired Student's *t* test or one-way ANOVA followed by Tukey's post hoc *t* test as appropriate. If the data violated the normality test, unpaired Mann-Whitney's *t* test or Kruskal–Wallis test, followed by Dunn's post hoc test, was performed as appropriate. For the CAG instability index, two-way ANOVA followed by Tukey's post hoc test and linear regression were performed. p values < 0.05 were considered statistically significant and are detailed in Table S1.

ONT sequence data analysis

We developed a pipeline that combines existing software with custom scripts. As a preprocessing step, we base-called sequencing reads with Guppy v5.0.7 in "sup" mode, applying a Q-score cutoff of 10 to filter out lower quality reads. While base-calling, Guppy also trims ONT barcodes and compresses fastq files, optimizing storage space. Preprocessed reads are then analyzed using our novel T-Rex Nextflow[63] pipeline. The Nextflow framework exploits containerized technology (Docker, Singularity) for increased reproducibility and increased portability across platforms. Starting from a set of fastq files and a sample sheet reporting metadata for each sample (e.g., clone ID, replicate ID, DIV, and location of the fastq and barcode files), the pipeline performs genome alignment, barcode filtering, CAG sizing and adjusted instability evaluation. In particular, alignment to the GRCh38 reference genome is performed by Guppy, which incorporates Minimap2,[69] generating BAM files. For efficient downstream CAG sizing, we combine groups of 100 BAM files into merged BAM files, using Samtools v1.18[70] for merging and indexing. These merged files are then analyzed with Straglr v1.2.0[25] using a custom BED file (chr4 3074868 3074948 CAG) as –loci parameter, to accurately account for

flanking regions. In parallel, our pipeline scans sequencing reads, looking for the expected barcode within the RMCE cassette. The process begins with converting fastq files to fasta using awk, followed by barcode detection using BLASTN v2.12,[67] with the arguments "-task blastn-short" and "-outfmt 6″ for short sequence tasks and customized output formatting.

Next, we evaluate the adjusted instability rate with a custom R script; this analysis requires three main inputs: a tsv file reporting CAG sizing information by Straglr, a tsv file reporting read-to-genome alignment information by Guppy, and a tsv file reporting read-to-barcode alignment information by BLASTN. The R script first filters sequencing reads based on alignment coverage, genomic coordinate, and BLAST score. To mitigate the effects of reads unbalancing, a resampling method is applied. Specifically, reads across cell clones for each time point are downsampled to the number of the smallest sample at the specific time point, ensuring an even mix of reads generated by the sequencing of forward and reverse strands, to account for context-specific sequencing errors of the Nanopore platform. This normalization process is repeated 100 times, and the adjusted instability rate is then reported as the median of these iterations (Figure S2A). We also tested a different subsampling strategy, by considering the smallest number of reads across samples from all time points, and a variable number of iterations – ranging from 10 to 500. This analysis confirmed that the downsampling strategy and the number of iterations had little effect on AIR values, as shown by the lack of statistically significant differences in a two-way ANOVA analysis, that compared the time course of AIR across variable number of iterations (Figures S2B and S2C). The script evaluates the CAG size distribution for each DIV, clone and replicate, and identifies the most frequent CAG size as the main peak for each condition. The adjusted instability rate is then calculated according to the formula by Lee et al., 2010,[26] with modifications to peak steps as per Nakamori et al., 2020[32] (Figure S2A), using 20% of the highest peak as the minimum frequency threshold and using DIV0 (Self-renewal) or DIV21 (Differentiation) for adjusting the instability index.

The complete codebase, including the Nextflow workflow configurations and the R script, is available at the following GitHub repository: https://github.com/GianlucaDamaggio/T-Rex.

Bulk RNA-seq

RNA was isolated and quality checked as described above. 150 ng of total RNA for each sample was dissolved in RNAse-, DNAse-water. RNA sequencing was carried out by Eurofins Genomics on an Illumina NovaSeq 6000 platforms in two batches, to generate 150 bp paired-end reads. Sequencing reads were preprocessed with nf-core rnaseq v3.14.0-gb89fac3 Nextflow pipeline.[63,64] Specifically, sequencing reads were trimmed and quality filtered with fastp v0.23.4[71] and their sequencing quality was checked with FastQC v0.12.1.[72] Reads generated from residual ribosomal RNA were then identified and removed with SortMeRNA v4.3.4.[73] Filtered reads were aligned to GRCh38 human reference genome with STAR v2.7.10a[74] and the resulting alignments were compressed and saved in sorted BAM files with Samtools v1.17.[75] RSEM v1.3.1[76] was then used to estimate gene expression levels from BAM files, using annotations from Ensembl v.110 GTF file for GRCh38 reference genome.

Gene counts were imported in R environment with *tximport* function[77] and a DESeq object was created with *DESeqDataSetFromTximport* function from DESeq2 package v1.12.3.[65] Genes with less than 5 counts in at least four samples were discarded. A differential expression analysis was then performed using *DESeq* function, using GT and Batch as covariates. Genes were considered as differentially expressed (DEGs) in case padj <0.05 and |log2FC| > 0.25.

The number of DEGs in each comparison (107Q vs. 21Q, 107Q vs. 107Q-multiCAAs, 107Q-multiCAAs vs. 21Q) was plotted as a barplot with ggplot2 v3.5.0[78] and their intersection was plotted using *upset* function from UpSetR v1.4.0 package.[79]

Read counts for DEGs were normalized and transformed with *vst* function from DESeq2 package and corrected for batch effect using *removeBatchEffect* function from Limma v3.56.2 package.[80] Transformed and batch-corrected counts were scaled by the mean expression value of the gene across samples and saturated between 0.9 and 1.1. Such expression values were then plotted as a heatmap with *pheatmap* function from pheatmap v1.0.12 package.[81]

BAM files for reads mapping to *HTT* gene were imported in R using readGAlignmentPairs function from GenomicAlignments v1.36.0,[66] and the number of reads mapping to intron 1 of *HTT* transcript was calculated with findOverlaps function from GenomicAlignments package. The number of *HTT* intron 1 reads, expressed as Transcripts Per Million (TPM), was evaluated for each sample and considered as a proxy for HTT1a levels.

The list of the PRC2 regulated genes was derived from Von Schimmelmann et al.,[52] Nature Neuroscience Table S3 reporting genes found altered upon PCR2-delition in WT mice.