Base editing strategies to convert CAG to CAA diminish the disease-causing mutation in Huntington's disease

Doo Eun Choi[1,2], Jun Wan Shin[1,2], Sophia Zeng[1], Eun Pyo Hong[1,2,3], Jae-Hyun Jang[1,2], Jacob M. Loupe[1,2], Vanessa C. Wheeler[1,2], Hannah E. Stutzman[1,4], Benjamin P. Kleinstiver[1,4,5], and Jong-Min Lee[1,2,3,*]

[1] Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

[2] Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

[3] Medical and Population Genetics Program, The Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142, USA

[4] Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA

[5] Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

* Corresponding Author: Jong-Min Lee

185 Cambridge Street, Boston, MA 02114, USA

Phone: 617-726-9724

Email: jlee51@mgh.harvard.edu

Abstract

An expanded CAG repeat in the huntingtin gene (*HTT*) causes Huntington's disease (HD). Since the length of uninterrupted CAG repeat, not polyglutamine, determines the age-at-onset in HD, base editing strategies to convert CAG to CAA are anticipated to delay onset by shortening the uninterrupted CAG repeat. Here, we developed base editing strategies to convert CAG in the repeat to CAA and determined their molecular outcomes and effects on relevant disease phenotypes. Base editing strategies employing combinations of cytosine base editors and gRNAs efficiently converted CAG to CAA at various sites in the CAG repeat without generating significant indels, off-target edits, or transcriptome alterations, demonstrating their feasibility and specificity. Candidate BE strategies converted CAG to CAA on both expanded and non-expanded CAG repeats without altering *HTT* mRNA and protein levels. In addition, somatic CAG repeat expansion, which is the major disease driver in HD, was significantly decreased by a candidate BE strategy treatment in HD knock-in mice carrying canonical CAG repeats. Notably, CAG repeat expansion was abolished entirely in HD knock-in mice carrying CAA-interrupted repeats, supporting the therapeutic potential of CAG-to-CAA conversion base editing strategies in HD and potentially other repeat expansion disorders.

Abbreviations

HD, Huntington's disease; huntingtin, *HTT*; Q, glutamine; CR, canonical repeat; LI, loss of interruption; DI, duplicated interruption; BE, base editing; CBE, cytosine base editor; PAM, protospacer adjacent motif; gRNA, guide RNA; iPSC, induced pluripotent stem cell; EV, empty vector.

Introduction

Huntington's disease (HD; MIM #143100) [1-3] is one of many trinucleotide repeat disorders caused by expansions of CAG repeats [4-8]. Although the underlying causative genes, pathogenic mechanisms, clinical features, and target tissues may be different [7; 9; 10], these disorders share a cardinal feature: an inverse relationship between age-at-onset and respective CAG repeat length [4; 7; 11-19]. To explain this striking genotype-phenotype correlation that is common to many trinucleotide repeat expansion disorders, a universal mechanism in which length-dependent somatic repeat expansion occurs toward a pathological threshold has been proposed [20]. This mechanism provides a good explanation of the relationship between CAG repeat length and age-at-onset in HD very well as 1) the *HTT* CAG repeat shows increased repeat length mosaicism in the target brain region [21-23], 2) somatic instability is repeat length-dependent [23; 24], and 3) the levels of repeat instability shows correlations with cell type-specific vulnerability and age-at-onset [22; 23; 25]. In addition, somatic repeat instability of an expanded *HTT* CAG repeat appears to play a major role in modifying HD since our genome-wide association studies have revealed that the majority of onset modification signals represent instability-related DNA repair genes [26-29]. Together, these data support the critical importance of CAG repeat length and somatic instability in determining the timing of HD onset.

Recent large-scale genetic analyses of HD subjects have revealed that different DNA repeat sequence polymorphisms have an impact on age-at-onset. Most HD subjects carry an uninterrupted glutamine-encoding CAG repeat followed by a glutamine-encoding CAA-CAG codon doublet (referred to as a canonical repeat; CR) [24; 27; 30]. However, expanded CAG repeats lacking the CAA interruption (loss of interruption; LI) or carrying two consecutive CAA-CAG (duplicated interruption; DI) [24; 27; 30] also exist (S. Figure 1). Surprisingly, the age-at-onset of HD subjects carrying LI or DI alleles is best explained by the length of their uninterrupted CAG repeat, not the encoded polyglutamine length [24; 27; 30]. These human genetics data indicate that introducing CAA interruption(s) into the *HTT* CAG repeat to reduce the length of the uninterrupted repeat is a potential therapeutic avenue to delay the onset of HD. Importantly, a genome engineering technology called base editing (BE) was recently developed, permitting the C-to-T conversion (cytosine base editors; CBEs) or A-to-G conversion (adenine base editors; ABEs) [31-34], where CBEs could in principle be applied to convert CAG codons to CAA to shorten the uninterrupted CAG repeat without altering polyglutamine length or introducing different amino acids. In view of the strong human genetic evidence for the role of the uninterrupted CAG

repeat length in determining HD onset, we have conceived BE strategies of converting CAG codons to CAA within the repeat and evaluated their therapeutic potential in HD.

Material and Methods

Study approval

Subject consents and the overall study were approved by the Mass General Brigham IRB and described previously [27]. Experiments involving mice were approved by the Mass General Brigham Institutional Animal Care and Use Committee.

Age-at-onset of HD subjects carrying LI or DI

Detailed experimental procedures for sequencing of the *HTT* CAG repeat region and determination of the CAG repeat length are described previously [27]. We compared age-at-onset of HD subjects carrying LI or DI to that of HD subjects carrying CR. Expected age-at-onset from CAG repeat length of CR was based on the onset-CAG regression model that we reported previously [19; 26; 27]. For expected age-at-onset based on the polyglutamine length, the same regression model was modified by replacing CAG repeat length with CAG repeat length plus 2 because the glutamine length equals CAG + 2 in CR.

Least square approximation to estimate the additional effects of LI and DI on age-at-onset

Carriers of LI and DI alleles showed slightly earlier and later onset age, respectively, compared to those with CR alleles of the same uninterrupted CAG repeat lengths, suggesting that LI and DI alleles confer additional effects. Thus, we attempted to determine the levels of additional effects of LI and DI that were not explained by the uninterrupted CAG size by taking a mathematical approach that is similar to least square approximation. Briefly, we predicted age-at-onset of HD subject carrying LI or DI alleles using our CAG-onset regression model for CR [19], and subsequently calculated the residual by subtracting predicted onset from observed onset age. We then calculated the sum of square (SS) for LI and DI carriers based on the participant's true uninterrupted CAG repeat length using the following formula.

Sum of squares (SS) = $\sum$ (observed age-at-onset − predicted age-at-onset age) ^{2}

Subsequently, we gradually increased and decreased the CAG repeat length for LI and DI allele carriers, respectively, and calculated SS again to identify CAG repeat size that generated the smallest SS. The

differences between true CAG and CAG repeat length that produced the smallest SS were considered as additional effects of LI and DI alleles on age-at-onset.

HEK293 cell culture, gRNA cloning, and transfection

HEK293 cells (https://www.atcc.org/products/crl-1573) were maintained in DMEM containing L-glutamine supplemented with 10% (v/v) FBS and 1% Penicillin-Streptomycin (10,000U/ml). Cells were maintained at 37 °C and 5% CO2. TrypLE Express (Life Technologies) was used to detach cells for sub-culture. PX552 vector (addgene #60958) was digested using SapI (Thermo) and purified by gel purification (QIAquick Gel Extraction Kit). A pair of oligos for each sgRNA were phosphorylated (T4 Polynucleotide Kinase, Thermo) and annealed by incubating at 37 °C for 30 min, 95 °C for 5 min, and ramping to 4 °C. Annealed oligos were diluted (1:50) and ligated into the digested PX552 vector (T7 ligase, Enzymatic) and incubated at room temperature for 15 min. Then, transformation was performed (One Shot Stbl3, Invitrogen). The inserted gRNA sequences were confirmed by Sanger sequencing. For transfection, cells were seeded in 6-well plates at approximately 65% confluence and treated with 1.66ug of CBEs and 0.7ug of gRNA plasmids on the following day using Lipofectamine 3000 (Invitrogen) according to the manufacturer's protocol. Three days after transfection, cells were harvested for molecular analyses. Genomic DNA was extracted using DNeasy Blood & Tissue kit (QIAGEN). AccuPrime GC-Rich DNA Polymerase (Invitrogen) was used to amplify a region containing the CAG repeat (35 cycles). PCR product was purified by PCR QIAquick PCR Purification Kit (QIAGEN) and subjected to MiSeq (Center for Computational and Integrative Biology DNA Core, Massachusetts General Hospital) and/or Sanger sequencing analysis (Center for Genomic Medicine Genomics Core, Massachusetts General Hospital). Primers for MiSeq sequencing were ATGAAGGCCTTCGAGTCCC and GGCTGAGGAAGCTGAGGA; primers for Sanger sequencing analysis were CAAGATGGACGGCCGCTCAG and GCAGCGGGCCCAAACTCA.

MiSeq data analysis to determine indels and conversion types

Sequence data from the MiSeq sequencing were subject to quality control (QC) by removing sequence reads 1) with mean base Phred quality score smaller than 20, 2) showing the difference between forward and reverse read pair, 3) containing fewer than 6 CAGs, or 4) not involving the full primer sequences. QC-passed data

revealed that HEK293 cells carry 16/17 CAG canonical repeats and therefore are expected to produce 18/19 polyglutamine segments. For QC-passed sequence reads, we determined the proportion of sequence reads containing indels, revealing most indels were sequencing errors. Subsequently, we focused on sequence reads without indels to determine the types of conversion. For each sequence read (not including CAA-CAG interruption), we counted sequence reads containing CAA, CAC, CAG, CCG, CGG, CTG, AAG, GAG, and TAG trinucleotide to determine the types and levels of conversion.

MiSeq data analysis of HEK293 cells treated with BE strategies to determine the sites of conversion

Sequence analysis revealed that BE strategies using CBEs produced mostly CAG-to-CAA conversion. CAG-to-TAG conversion was detected in all samples regardless of BE strategies, suggesting that this type of conversion is also due to amplification/sequencing errors. Therefore, we focused on sequence reads of 16/17 CAG repeats containing only CAG or CAA to determine the sites of CAG-to-CAA conversion. Briefly, we recorded the sites of CAG-to-CAA conversion for each sequence read and summed the number of conversions at a given site for a given sample. Therefore, 30% CAG-to-CAA conversion at the second CAG means 70% and 30% of all sequence reads contain CAG and CAA at the second CAG position, respectively.

Quantification of duplicated interruption and multiple conversion

The proportion of duplicated interruptions was determined from HEK293 cells treated with different combinations of CBEs and gRNAs. Briefly, we counted sequence reads containing duplicated interruption and divided them by the number of all QC-passed sequence reads to calculate the proportion of DI alleles. Similarly, we calculated the proportion of sequence reads containing duplicated interruption and CAG-to-CAA conversions at other sites. We also determined the levels of multiple CAG-to-CAA conversion for each BE strategy. For each sequence read in a sample, we counted the number of CAG-to-CAA conversions regardless of their locations to generate a distribution of numbers of multiple conversion for each sample. Since we counted the conversions regardless of their positions, multiple conversions do not necessarily mean consecutive conversions.

Determination of the transfection efficiency

To determine the effects of transfection efficiency on patterns of base editing, we transfected HEK293 cells with gRNA 2 with combinations of different base editors and performed cell staining. Transfected HEK293 cells were fixed with paraformaldehyde (4%) and permeabilized with Triton X-100 (0.5%). Then, cells were stained with DAPI (0.5uM) and incubated for 30 min before being washed with PBS. The eGFP (enhanced green fluorescent protein) and DAPI (4',6-diamidino-2-phenylindole) images from eight areas in each well were captured using a fluorescence inverted microscope (Nikon Eclipse TE2000-U). The ImageJ analysis program was used to measure the size of a single cell expressing eGFP; we randomly selected 20 cells for each image and averaged their sizes to be used as a reference. We counted the number of pixels covered by eGFP-positive signals, and subsequently divided by the average cell size to obtain the number of eGFP positive cells in each image. This was repeated with the DAPI staining images. The percent transfected was calculated by dividing the number of eGFP-positive cells by that of DAPI-positive cells multiplied by 100.

Base editing in HD patient-derived iPSC and differentiated neurons

An iPSC line carrying adult-onset CAG repeats (42 CAG) was derived from a lymphoblastoid cell line in our internal collection by the Harvard Stem Cell Institute iPS Core Facility (http://ipscore.hsci.harvard.edu/) [35; 36]. HD iPSC cells were dissociated into single cells with Accutase (STEMCELL Tech) and plated on matrigel-coated 24-well plate in mTeSR plus media containing CloneR (STEMCELL Tech) to increase cell viability. The following day, cells at 60~70% confluence were transfected with 1.8ug of BE4max and 0.6ug of gRNA plasmids using Lipofectamine STEM (Invitrogen) according to the manufacturer's protocol. Cells were incubated at 37 °C and 5% CO2 for 5 days for sequencing analysis.

The same iPSC line was differentiated into neurons using a previously described method [37]. Briefly, the iPSC line was plated on growth factor reduced Matrigel (Corning) in mTeSR Plus media (STEMCELL Technologies). When cells reached ~ 80% confluence, differentiation was initiated by switching to DMEM-F12/Neurobasal media (2:1) supplemented with N2 and retinol-free B27 (N2B27 RA−; Gibco). For the first ten days, cells were supplemented with SB431542 (10 μM; Tocris), LDN-193189 (100 nM; StemGene), and dorsomorphin (200 nM; Tocris). SB431542 was removed from the media on day 5. Cells were maintained in N2B27 RA− supplemented with activin A (25 ng/ml; R&D) on day 9. On day 22, cells were split using Accutase (STEMCELL Technologies) and seeded on a poly-D-lysine/laminin plate with N2B27 media supplemented with

BDNF and GDNF (10 ng/ml each; Peprotech). Media were changed the next day to facilitate neuronal maturation and survival. Cells were fed with new media every two days. For neuronal marker staining, cells were fixed, permeabilized, and blocked using the Image-iT Fix-Perm kit (Invitrogen). Subsequently, cells were stained by Anti-TUBB3 (tubline beta 3; Biolegend Inc, Cat# 801202) in a blocking solution overnight at 4℃. Then, cells were washed with PBS three times for 5 minutes, followed by incubation with Alexa Flour 594 secondary antibodies (Invitrogen) for 1 hour. Finally, cells were washed with PBS three times for 5 minutes and mounted with Vectorshield mounting medium with DAPI (Vector Laboratories). Images were captured by the Leica fluorescence microscope. Differentiated neurons were transfected with 1.8ug of BE4max and 0.6ug of gRNA plasmids using Lipofectamine 3000 according to the manufacturer's protocol. Cells were incubated at 37 °C and 5% CO2 for 7 days for sequencing analysis.

Off-target prediction and experimental validation

Potential off-targets were predicted by the Off-Spotter (https://cm.jefferson.edu/Off-Spotter/) for 8 BE strategies using the gRNA sequences. We allowed a maximum of 4 mismatches to identify potential off-targets that are flanked by the NGG PAM. Given decreased single base specificity at the PAM-proximal sites in the CRISPR-Cas9 genome engineering [38] and the abundance of CAG repeat carrying genes in the human genome, many of our gRNAs (except gRNAs 1 and 2) are predicted to hybridize with many CAG repeat sequences in the genome, generating increased numbers of predicted off-targets. Thus, we performed experimental validation of 1) predicted off-targets for BE strategies 1 and 2 (described here) and 2) genes that cause polyglutamine disorders. For the experimental validation of predicted off-targets, we analyzed HEK293 DNA samples that were used for MiSeq analysis. Briefly, we focused on predicted off-targets in the protein-coding genes for gRNAs 1 and 2 with two mismatches. One and four potential off-targets were predicted for gRNAs 1 (*MINK1*) and 2 (*PINK1, ZNF704, WBP1L, C20orf112*), respectively. We amplified predicted off-target sites of gRNAs 1 and 2 (35 cycles) using the following primers:

MINK1, AGCATGCCTACCTCAAGTCC and CTGGTTTGTCAGCGGGATTC;

PINK1, CTGTACCCTGCGCCAGTA and GGATGTTGTCGGATTTCAGGT;

ZNF704, GGACGGGTTGGACTGGTC and GGGTCCTGGCACTGACTGTG;

WBP1L, CCGACCTCCAACTCCTCCC and GCTGCTCTGTGCCCCCTG; and

C20orf112, GATCTCCGTGGGGCTGAG and CCTACTTCCCTCTCCACAGG.

Amplified DNA samples were analyzed by MiSeq sequencing.

Experimental validation of off-targets in genes causing polyglutamine diseases

Similarly, we amplified genomic regions (35 cycles) containing CAG repeat regions in the genes causing polyglutamine diseases using the following primers:

ATXN1, CCTGCTGAGGTGCTGCTG and CAACATGGGCAGTCTGAGC;

ATXN2, CGGGCTTGCGGACATTGG and GTGCGAGCCGGTGTATGG;

ATXN3, GAATGGTGAGCAGGCCTTAC and TTCAGACAGCAGCAAAAGCA;

CACNA1A, CCTGGGTACCTCCGAGGGC and ACGTGTCCTATTCCCCTGTG;

ATXN7, GAAAGAATGTCGGAGCGGG and CTTCAGGACTGGGCAGAGG;

TBP, AAGAGCAACAAAGGCAGCAG and AGCTGCCACTGCCTGTTG;

ATN1, CCAGTCTCAACACATCACCAT and AGTGGGTGGGGAAATGCTC; and

AR, CTCCCGGCGCCAGTTTGCTG and GAACCATCCTCACCCTGCTG.

Sequencing data analysis was focused on calculating the proportion of sequence reads that contain the CAG-to-CAA conversions.

RNAseq analysis

To determine the molecular consequences of candidate BE strategies, we performed RNAseq analysis. We transfected HEK293 cells with BE4max+empty vector, BE4max+gRNA 1, or BE4max+gRNA2 for 72hours. Subsequently, genomic DNA for MiSeq analysis and cell pellets for RNAseq analysis were generated from replica plates Genome-wide RNAseq analysis (Tru-Seq strand specific large insert RNA sequencing) was performed by the Broad Institute. Sequence data were processed by STAR aligner [39] as part of the Broad Institute's standard RNAseq analysis pipeline. For differential gene expression (DGE) analysis, we used transcripts per million (TPM) data computed by the TPMCalculator (https://github.com/ncbi/TPMCalculator) [40]. Expression levels in approximately 19,000 protein-coding genes based on Ensembl (ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/) were normalized. The DGE analysis was performed by the generalized linear model using a library of “glm” in R package v3.3.1 (https://www.r-project.org/) after

adjustment for two principal components based on RNAseq data, followed by multiple test correction using a FDR method. A multiple test corrected p-value less than 0.05 was considered statistically significant.

Generation and validation of HEK293-51 CAG cells carrying an expanded CAG repeat

HD patient-derived iPSC and neurons showed modest conversion efficiencies, making it technically difficult to characterize molecular consequences of CAG-to-CAA conversion strategies. Thus, we generated HEK293 cells carrying an expanded repeat by replacing one of the non-expanded *HTT* CAG repeats with a 51 CAG repeat. Briefly, we cloned a gRNA (CAGAGCGCAGAGAATGCGCG) into the PX459 vector (Addgene# 62988) to express SpCas9 and gRNA for CRISPR-Cas9 targeting at the *HTT* CAG repeat region. The donor template for homologous recombination was generated by PCR amplification of a human DNA sample carrying 51 CAG repeat into the pCR-Blunt II TOPO plasmid (Invitrogen). Subsequently, HEK293 cells were transfected with PX459 and pCR-Blunt II TOPO plasmids by Lipofectamine 3000 (Invitrogen) for 72 hr. Subsequently, cells were treated with G-418 (Gibco) for 21 days, and surviving cells were re-plated onto 100 cm dishes. After 10 days, visible colonies were picked and maintained separately. Single cell clonal lines were validated by PCR analysis using AccuPrime GC-Rich DNA Polymerase and primer set (ATGAAGGCCTTCGAGTCCC and GGCTGAGGAAGCTGAGGA). The PCR conditions were initial denaturation (95 °C, 3 min), 30 cycles of denaturation (95 °C, 30 sec), annealing (55 °C, 30 sec), extension (72 °C, 40 sec), and final extension (72 °C, 10 min). The PCR products were resolved on a 1.5% agarose gel containing GelRed (Biotium) and visualized under UV light to distinguish expanded from non-expanded CAG repeats. We also performed RT-PCR and immunoblot analysis to confirm the correct integration of the expanded CAG repeat. Briefly, 1 µg of total RNA from the targeted clonal line was subjected to reverse transcription with SuperScript IV Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions followed by PCR analysis using a primer set (ATGAAGGCCTTCGAGTCCC and GGCTGAGGAAGCTGAGGA). For HTT immunoblot analysis, cells were lysed with RIPA Lysis/Extraction Buffer (Thermo) supplemented Halt Protease and Phosphatase Inhibitor Cocktail (Thermo). Whole cell lysate was then separated on NuPAGE 3 to 8%, Tris-Acetate gel (Invitrogen) and transferred to a polyvinylidene fluoride membrane. The membrane was blocked with 5% nonfat dry milk in Tris-buffered saline for 1 h and incubated with primary antibodies for HTT (MAB2166, Sigma-Aldrich) for 12 h at 4 °C. The membrane was washed for 1 h, and blots were incubated with a peroxidase-conjugated

secondary antibody for 1 h then washed for 1 h. The bands were visualized by enhanced chemiluminescence (Thermo). Similar to HEK293 cells, HEK293-51CAG cells were treated with BE4max and candidate gRNAs (i.e., gRNA 1 and gRNA 2) to determine the levels of CAG-to-CAA conversion and the total HTT protein levels. For gRNA 1, we determined the levels of in-frame insertion and deletion right after treatment using methods previously described (Lee 2015).

AAV treatment for a candidate BE strategy and CAG repeat instability in mice

For AAV injection experiments, we used split-intein base editor (v5 AAV) [34]. Forward and reverse oligos (CACCGCTGCTGCTGCTGCTGCTGGA and AAACTCCAGCAGCAGCAGCAGC) (IDT) for gRNA 2 were cloned into the BSmBI site of pCbh_v5 AAV-CBE C-terminal (Addgene, # 137176) and pCbh_v5 AAV-CBE N-terminal (Addgene, # 137175). Cloned vectors were validated by Sanger sequencing, and subsequently packed into AAV9 serotype by UMass Viral Vector Core. HttQ111 HD knock-in mice [41] were maintained on an FVB/N background [42]; AAV9 injections were performed in heterozygous HttQ111/+ mice at 6~11 week. Animal husbandry was performed under controlled temperature and light/dark cycles. After anesthesia was induced using isoflurane, an insulin syringe was inserted into the medial canthus with the bevel of the needle facing down from the eyeball, advanced until the needle tip was at the base of the eye. We injected HD knock-in mice with AAV9 mix (200 µl containing C-terminal and N-terminal split-intein base editor, 1 x 10^{12} vg for each) (experimental group) or PBS (200 µl, control group) by retro-orbital (RO) injection. Ten weeks later, liver and tail samples were collected for instability analysis [43; 44]. Briefly, DNA samples were amplified using primer set (6'FAM-ATGAAGGCC TTCGAGTCCCTCAAGTCCTTC and GGCGGCTGAGGAAGCTGAGGA) and analyzed by ABI3730 to determine the sizes of fragments. Quantification of repeat expansion was based on the expansion index method that we developed previously. The expansion index method robustly quantifies the levels of repeat instability by eliminating potential noise in the fragment analysis results based on the relative peak height threshold [43; 44]. To quantify expansion index in control and mice treated with BE, we applied 10% threshold, and expansion index was calculated based on the highest peak in the tail DNA.

Repeat instability in HD knock-in mice carrying interrupted CAG repeat

To determine the maximal effects of CAG-to-CAA interruption, we analyzed HD knock-in mice carrying interrupted CAG repeat (namely interrupted repeat mice; https://www.jax.org/strain/027418) to HD knock-in mice carrying uninterrupted repeat (namely, pure repeat mice; https://www.jax.org/strain/027417). Repeat in the interrupted repeat mice and pure repeat mice comprises 21 copies of [CAGCAACAGCAACAA] and 105 copies of [CAG], respectively. Both mouse lines were expected to produce huntingtin protein with 105 polyglutamine. Repeat instability in these mice were determined (5 months) by the fragment analysis as described previously [44].

Statistical analysis and software

Statistical analysis of RNAseq data was performed using generalized linear regression analysis. Multiple test correction was performed using false discovery rate using R 3.5.3 [45]. R 3.5.3 was also used to produce plots.

Results

Effects of CAG-CAG codon doublet on age-at-onset in HD patients

Previously, we and others reported that most HD subjects carry canonical repeats (CR) comprising an uninterrupted expanded CAG repeat followed by CAA-CAG [24; 27; 30]. Although infrequent, uninterrupted CAG repeats followed by 1) no CAA-CAG (LI; 0.23% in our previous GWA data) and 2) two CAA-CAG codon doublets (DI; 0.76% in our previous GWA data) also exist (S. Figure 1). In HD subjects carrying LI alleles, the length of the CAG repeat and polyglutamine segment are identical. However, the polyglutamine length is greater by 2 and 4, respectively, compared to the CAG repeats in CR and DI alleles (S. Figure 1). Since CR, LI, and DI alleles with the same uninterrupted CAG repeat lengths have different polyglutamine sizes, they have provided a powerful tool to investigate the relative importance of the CAG repeat in DNA vs. polyglutamine in protein in determining onset age. For example, if polyglutamine length played an important role in determining age-at-onset, onset of LI and DI allele carriers, who respectively have 2 fewer and 2 more glutamines compared to CR allele carriers, would be significantly later and earlier compared to CR allele carriers with the same uninterrupted CAG repeats (S. Figures 2A and 2B). In stark contrast to these predictions, the onset ages of LI or DI allele carriers are best explained by their respective CAG repeat sizes, not polyglutamine length (S. Figures 2C and 2D). Furthermore, age-at-onset of DI allele carriers is significantly delayed compared to that of LI allele carriers with the same uninterrupted CAG repeat size even though DI alleles encode 4 more glutamines than LI alleles (Student t-test p-value, 1.007E-12) (S. Figure 2D). Together, the data indicate that age-at-onset in HD is determined primarily by the uninterrupted length of the CAG repeat, but there may also be additional effects of different CAA-interruption structures since the CAG repeat length does not fully explain age-at-onset in LI and DI allele carriers (S. Figure 2D) [46]. Therefore, we performed least square approximation to calculate the magnitudes of the additional effects of LI and DI alleles on age-at-onset. Briefly, we varied the individual CAG repeat length to identify the repeat size that best explains the observed age-at-onset of carriers of these LI and DI alleles relative to CR alleles. The age-at-onset of the LI allele carriers is best explained when 3 CAGs are added to the true CAG repeat length (Figures 1A and 1B; S. Figure 3D) while the DI allele carriers behave with respect to age-at-onset as if they have one less CAG than their true CAG repeat length (Figures 1B and 1C; S. Figure 3F). These data suggest that switching a CR allele to a DI allele would delay onset by 1) shortening the uninterrupted CAG repeat by two CAG repeats and 2) conferring an additional effect

comparable to a reduction in length of one CAG. For example, if a DI allele were generated from a CR allele with 43 uninterrupted CAGs by converting the 42nd CAG to CAA using base editing strategies, the age-at-onset is predicted to be delayed by approximately 12 years (Figure 1D), illustrating the robustness of therapeutic base editing strategies.

Cytosine base editors and gRNAs to convert CAG to CAA in the HTT CAG repeat

Recent advancements in genome editing technologies have led to the development of CBEs that are capable of efficient C-to-T conversion (Figure 2A) [31; 32; 47-49]. In principle, CR can be converted to DI if CBEs target the non-coding strand of the *HTT* CAG repeat (Figure 2B). In this study, we tested 4 CBEs comprised of various cytosine deaminases and SpCas9 enzymes with different protospacer-adjacent motif (PAM) specificities to explore the feasibility of CAG-to-CAA conversion as a putative treatment for HD. BE4 is the fourth-generation base editor which was engineered from BE3 to increase the editing efficiencies and decrease the frequency of undesired by-products (Figure 2C) [47]. BE4 exhibited high levels of C-to-T editing activity on the target sites harboring NGG PAMs [47]. The activity window of BE4 is position 4-8, counting from the PAM distal end of the spacer (where the PAM is positions 21-23) (Figure 2B) [48]. We tested the BE4max (Addgene #112093) in this study, which is a codon optimized version of BE4 with improved nuclear localization [48]. Due to the sparsity and lack of NGG PAM sites near and within the CAG repeat, respectively, CAG-to-CAA conversion using BE4 was expected to be somewhat limited. Therefore, we also explored engineered CBEs containing SpCas9 variants that target an expanded range of PAM sequences, including SpCas9-NG [50] and SpG [51] (Figure 2C). Since these variants are capable of targeting sites with NGN PAMs, they might permit higher density targeting near or within the CAG repeat. The nucleotide preceding the target cytosine also affects the C-to-T conversion efficiency in CBEs, especially when a G precedes the C [31; 52; 53]. Thus, engineered deaminase domains have been explored to improve C-to-T conversion in the GC-contexts [32; 49]. For instance, an evolved CDA1-based BE4max variant (evoCDA1) showed substantially higher editing on GC targets [49], which is relevant to the nucleotide context on the non-coding strand of the *HTT* CAG repeat (CTGCTG). Therefore, we explored the use of canonical BE4max-SpCas9, BE4max-SpCas-NG, BE4max-SpG, and evoCDA1-BE4max-SpG (henceforth referred to as BE4max, BE4-NG, BE4-SpG, and evo-SpG, respectively) (Figure 2C).

To achieve CAG-to-CAA conversion in the *HTT* CAG repeat, we designed 3 groups of gRNAs (S. Table 1) based on the sites of predicted hybridization (Figure 2D). Aiming at converting CAGs at the front-end of the repeat, gRNAs 1 and 2 were designed to hybridize with a region involving the upstream of the repeat and conventional NGG PAMs. The gRNAs 1 and 2 contain 10 and 2 non-CAG bases at the PAM-proximal ends, respectively (S. Table 1; S. Figure 4). Considering the activity window of the BE4 (i.e., 13th-17th nucleotide from the PAM) (S. Figure 4, green boxes in the gRNAs), BE4max-gRNAs 1 and BE4max-gRNA 2 were predicted to convert the 1st/2nd and 4th/5th CAG to CAA, respectively (S. Figure 4; sequences with green highlight). The gRNAs 3, 4, and 5 comprised the CAG repeat sequence (S. Table 1) and therefore, were predicted to hybridize throughout the *HTT* CAG repeat and potentially other CAG repeat-containing genes (S. Figure 4). The gRNAs 3, 4, and 5 were predicted to utilize NAA/NTG, NGA/NCT, and NGG/NGC PAMs, respectively (S. Figure 4). Lastly, gRNAs 6, 7, and 8 were designed to convert CAGs at the back-end of the repeat (S. Table 1). Available PAM sites for these gRNAs are NCT, NGC, and NTG (S. Figure 4). Considering the predicted gRNA-target hybridization sites and conversion windows, these three gRNAs might generate the duplicated interruption that is found in HD patients. (S. Figure 4).

Predominant CAG-to-CAA conversion without significant indels by BE strategies for HD

We then characterized 32 BE strategies (i.e., combinations of 4 CBEs and 8 gRNAs). We first determined whether BE strategies for HD produced indels. Since low base editing efficiencies might result in proportionally low levels of indels leading to an underestimation of their frequencies, we used HEK293 cells, which showed high levels of base editing efficiencies [54; 55]. Our MiSeq sequence analyses revealed that HEK293 cells carry two CRs (16 and 17 CAGs) and showed approximately 10% of basal levels of indel ('Cell' in S. Figure 5), which reflects errors due to the difficulty in sequencing the CAG repeat. Nevertheless, transfection of plasmids for BE strategies did not significantly increase the levels of indels compared to cells without any treatment (Cell) or cells treated with empty vector (EV) (S. Figure 5). The lack of significant indel formation was quite expected because the cytosine base editors that we tested use nickases (Figure 2C) [47; 48].

Since most sequence reads containing indels might be sequencing errors, we focused on sequence reads without indels to determine the types of base conversions. HEK293 cells without any treatment (Cell) or cells treated with empty vector (EV) showed low but detectable levels of CAG-to-CAA and CAG-to-TAG

conversions (S. Table 2; S. Figure 6), also reflecting sequencing errors. However, the levels of CAG-to-CAA conversion were significantly increased over baseline sequencing errors in cells treated with some BE strategies (S. Table 2). For example, BE4max in combination with gRNAs 1, 2, 5, and 7 resulted in efficient CAG-to-CAA conversion (Figures 3A). Given the availability of the NGG PAMs (S. Figure 4), robust CAG-to-CAA conversion by gRNAs 1, 2, and 5 was somewhat anticipated for BE4max. However, high levels of CAG-to-CAA conversion by the BE4max-gRNA 7 combination (Figure 3A; S. Figure 6A) were unexpected because the anticipated hybridization site does not provide the NGG PAM (S. Figure 4) that is required for the optimal activity of BE4max. The BE4-NG robustly produced CAG-to-CAA conversions with gRNAs 1, 2, and 3; although not significant, gRNAs 5 and 7 also generated high levels of CAG-to-CAA conversions (Figure 3B; S. Figure 6B). BE4-SpG with the combinations with gRNAs 1 and 2 resulted in significant levels of CAG-to-CAA conversions (Figure 3C; S. Figure 6C). Overall, the CAG-to-CAA conversion was higher in evo-SpG compared to other base editors; gRNAs 1, 2, 4, and 8 produced significant CAG-to-CAA conversions (Figure 3D; S. Figure 6D). These data indicated that our BE strategies primarily generated CAG-to-CAA conversion without significant indel formation. Patterns of conversions also indicated that sites with NGG PAMs (gRNAs 1, 2, and 5) permitted the highest levels of CAG-to-CAA conversion for BE4max and CBEs with relaxed PAM specificities.

Sites of CAG-to-CAA conversion

Subsequently, we determined conversion sites for different BE strategies. The patterns of conversion sites were similar for BE4max, BE4-NG, and BE4-SpG in gRNA 1, showing the most conversion at the second CAG with decreased levels of conversion at the first CAG (Figure 4A; S. Table 3). In contrast, evo-SpG-gRNA 1 combination showed higher editing efficiencies with the maximum conversion at the second CAG with comparable levels of conversions at the first and third CAGs (Figure 4A, cyan; S. Table 3). The gRNA 2 showed similar patterns as gRNA 1 except that conversion sites were shifted to the right; the highest conversion occurred at the 4th CAG by BE4max, BE4-NG, and BE4-SpG (Figure 4B; S. Table 3).

The gRNAs 3 and 4, which were designed to hybridize throughout the CAG repeat, did not generate CAG-to-CAA conversion in combination with BE4max (Figures 4C and 4D, red; S. Table 3) because of the lack of a NGG PAM. Although modest, BE4-NG and BE4-SpG converted the 5th CAG to CAA (Figures 4C and 4D;

S. Table 3), potentially due to the possibility that NAA (gRNA 3) and NGA (gRNA 4) PAMs supported the base editing activity of BE4-NG and BE4-SpG. The gRNAs 3 and 4 produced higher levels of CAG-to-CAA conversion in evo-SpG again (Figures 4C and 4D, cyan), and interestingly, CAG-to-CAA conversions were not limited to the 5th CAG (S. Table 3). The gRNA 5 with BE4max efficiently converted the 6th CAG (Figure 4E, red), which was unexpected; conversions by other base editors were lower but widespread throughout the repeat (Figure 4E).

BE strategies designed to convert CAGs at the back-end of the repeat were tested using gRNA 6, 7, and 8. Although less robust, the patterns of conversion by gRNA 6 (Figure 4F) were similar to those of gRNA 4 (Figure 4D). Since only one nucleotide is different between gRNA 6 and gRNA 4, it appeared that gRNA 6 behaved like gRNA 4 despite one mismatch, favoring the NGA PAM instead of the less optimal NCT PAM (S. Figure 7A). The same explanation might account for the similar patterns of conversion sites for gRNA 7 (Figure 4G) and gRNA 5 (Figure 4E); efficient conversion at the 6th CAG by BE4max-gRNA 7 might be due to the interaction of gRNA 7 at the target site of gRNA 5 (with one mismatch) in favor of the NGG PAM (S. Figure 7B). The gRNA 8 generated CAG-to-CAA conversions only in evo-SpG. Although this group of gRNAs was designed to hybridize with the back-end of the CAG repeat, higher levels of conversion were observed at the front-end CAGs and throughout the repeat (Figures 4F-4H). These results suggest that one PAM-distal mismatch might be tolerated by base editors in favor of targets sites harboring more robust PAMs. Also, our data revealed that as expected, BE4max is highly dependent on the NGG PAM, resulting in CAG-to-CAA conversion at specific CAG sites, while evo-SpG is more efficient in conversion leading to broader targeting due to its relaxed PAM requirement.

Generation of duplicated interruption by BE strategies

Next, we determined the levels of duplicated interruption in the same HEK293 cell MiSeq data. As shown in S. Figure 8, BE4max and evo-SpG did not produce significant amounts of the DI that is found in humans (S. Figure 1). However, BE4-NG (S. Figure 8B) and BE4-SpG (S. Figure 8C) produced modest but significant levels of DI in combinations with gRNAs 5 and 7 (0.5%~1% increase over the basal levels). Modest levels of DI alleles compared to conversions at other sites might be due to the lack of canonical PAMs (e.g., NGG) at the specific site (approximately 18 nucleotides upstream of CAA-CAG interruption). We also observed that gRNA 5

and 7 relatively increased the number of sequence reads containing both DI and CAG-to-CAA conversions at other sites (S. Figures 9B and 9C), indicating that CTG trinucleotides on the non-coding strand of the repeat contributed to modest but widespread CAG-to-CAA conversion throughout the repeat. Similarly, CAG-to-CAA conversion was not confined to specific sites in evo-SpG in combination with gRNAs 3-8 as DI alleles generated by these strategies also contained CAG-to-CAA conversions at other sites (S. Figure 9D). Since increased conversion efficiency in evo-SpG could not be explained by the transfection efficiency (S. Figure 10), these data indicate that evo-SpG has a significantly wider conversion window. In agreement with this, the most frequent number of conversions in a given sequence read by evo-SpG was greater than that of other base editors (S. Figure 11; S. Table 4).

Evaluation of off-target effects

We then evaluated the levels of off-target conversions using Off-Spotter. As summarized in S. Table 5, gRNAs 1 and 2 showed relatively smaller numbers of predicted off-targets due to unique sequences near the PAMs. As expected, gRNAs that were designed to hybridize throughout the CAG repeat showed increased numbers of predicted off-targets. Similarly, gRNAs to convert CAG at the back-end of the repeat showed larger numbers of predicted off-targets, potentially due to the fact that unique sequences are distal to the PAMs. Subsequently, we performed two sets of follow-up off-target validations. For gRNAs 1 and 2, we experimentally evaluated predicted off-targets focusing on protein-encoding genes; one and four genes were predicted off-target sites for gRNA 1 and gRNA 2, respectively, and all showed low levels of conversion compared to on-target (S. Table 6). We also characterized the levels of off-target conversion in other CAG repeat-containing genes focusing on 8 polyglutamine disease genes (S. Tables 7). As predicted, gRNAs 1 and 2 showed low-level conversions in the CAG repeats of other polyglutamine disease genes in general (S. Table 8; S. Figure 12). In contrast, gRNAs 3-8 produced variable but higher levels of conversion in some polyglutamine disease genes depending on the availability of preferred PAMs (S. Figure 12).

Allele specificities and molecular outcomes of candidate BE strategies

Subsequently, we evaluate the levels of allele specificity of candidate BE strategies (BE4max-gRNA 1 and BE4max-gRNA 2) in patient-derived induced pluripotent stem cells (iPSC carrying 41 CAG CR) [35; 36] and

differentiated neurons (S. Figure 13). As shown in S. Table 9, transfection of gRNA 1 and gRNA 2 produced modest CAG-to-CAA conversion on both mutant and normal *HTT* (approximately < 3%). Overall, low conversion efficiencies by transfection and transduction of AAV (adeno-associated virus; data not shown) represent difficulties in delivery in these cell types [35; 56], posing a challenge to determining the levels of allele specificity of BE strategies. To overcome these technical difficulties, we developed a HEK293 clonal line carrying an expanded *HTT* CAG repeat (Figure 5A) by replacing one of normal CAG repeats with a 51 CAG canonical repeat (namely HEK293-51 CAG) (S. Figure 14). A candidate BE strategy (i.e., BE4max-gRNA 1) did not increase the levels of in-frame insertion/deletion in the mutant or normal *HTT* repeat (Figures 5B and 5C). Subsequent analysis revealed that a candidate BE strategy BE4max-gRNA 1 produced high levels of CAG-to-CAA conversions on both expanded and non-expanded canonical repeats (Figure 5D). Although very modest, conversion was significantly higher in the non-expanded repeat (uncorrected p-value, 0.04996), which can be explained by slightly reduced conversion on the mutant *HTT* due to higher GC content in the expanded CAG repeat. However, the candidate BE strategies did not alter huntingtin protein levels (Figures 5E and 5F) at the time of treatment, supporting the safety of candidate BE strategies. We also performed RNAseq analysis to identify genes whose expression levels were altered by candidate BE strategies in HEK293 cells. Candidate strategies such as BE4max-gRNA 1 and BE4max-gRNA 2 produced significant on-target CAG-to-CAA conversions (Figure 6A), but the levels of *HTT* mRNA were not altered by either treatment (filled red circles in S. Figure 15 and Figure 6). In addition, RNAseq data analysis showed that neither BE strategy induced significant gene expression changes in any genes (false discovery rate, 0.05) (S. Figure 5). When comparing all HEK293 samples treated with either BE strategies (n=8) to those treated with EV (n=4), the shape of volcano plot mimicked random sample comparison (Figures 6B and 6C), implying the lack of impacts of candidate BE strategies on transcriptome.

Effects of base conversion on the CAG repeat instability in vivo

The limited cargo capacity of AAV has been circumvented by the intein-split base editor, and the feasibility of BE strategies targeting non-repetitive sequences has been demonstrated in mouse models of human diseases [34; 57; 58]. Taking advantage of the split BE system, we determined whether a candidate HD BE strategy could target the CAG repeat and result in a decrease in somatic repeat expansion, which was hypothesized to be the

major disease driver [20]. Since striatal and liver repeat instability share certain underlying mechanisms [59-66], and *in vivo* delivery might be more efficient in the liver compared to the brain [34], we used AAV9 to evaluate a candidate BE strategy in the liver. As expected, somatic CAG repeat expansion index in the liver of HD knock-in mice carrying around 110 CAGs showed a positive correlation with the inherited CAG repeat length (as represented in tail DNA) and the age of mice (Figures 7A and 7B) [16; 23; 41; 44; 67; 68]. Unfortunately, we could not determine the sequence modification in treated mice by sequencing because of 1) very long CAG repeats in these mice, 2) modest levels of base conversion, and 3) high levels of errors when sequencing the CAG repeat (S. Figure 6). However, when the effects of the tail CAG repeat size and age of mice were corrected, retro-orbital injection of AAV9 for split CBE (v5 AAV-CBE) and gRNA 2 significantly decreased the levels of repeat expansion (Figure 7C; p-value, 1.78E-6). Nevertheless, the expansion index in treated and control mice was largely overlapping (Figure 7A), suggesting that the effects of BE treatment were very modest. We speculate that 1) insufficient dosage due to difficulty in producing high titer viral package for big cargo (i.e., 5KB) [34; 69], 2) limited delivery [70], and/or 3) difficulty in targeting the very long CAG repeat resulted in modest effects. Given those limitations, we also analyzed a mouse model containing interrupted repeat to determine the maximum effects of the interruption on the repeat expansion. HD knock-in mice carrying 105 interrupted CAG repeat (https://www.jax.org/strain/027418) showed complete loss of repeat expansion compared to 105 CAG uninterrupted repeat mice (https://www.jax.org/strain/027417) (Figures 7D and 7E), suggesting that CAA interruption could completely suppress the most important disease modifier (i.e., CAG repeat expansion).

Discussion

Recent advances in genome engineering provide powerful tools to interrogate the relationships among genes, functions, and diseases. For example, CRISPR-Cas9-based editing approaches have revolutionized the investigation of individual genes of interest and also have begun to be applied to humans to treat diseases [71-75]. Base editing (BE), which can convert a single nucleotide to another, represents a newly developed and highly versatile genome engineering technology [31; 33]. BE has advantages over other genome engineering approaches with respect to safety and clinical applicability. BE employing nickase Cas9 does not intentionally create double-stranded DNA breaks (DSBs) [31; 33], minimizing potential adverse effects. Also, BE with low off-targeting is being actively developed, adding an additional layer of safety [76]. The majority of well-characterized disease-causing mutations are point mutations, and therefore many genetic disorders can be addressed by BE strategies [77]. The robustness of BE has been demonstrated in models of genetic disorders caused by point mutations [57; 58; 77; 78], and the first human trial employing base editing has already been started [79; 80]. However, many human disorders are caused by other types of mutations, such as expansions of DNA repeats [7; 16; 81] for which BE may not seem like an ideal tool. In contrast to this commonly held notion, we show that BE strategies could also address diseases that are caused by expanded repeats, broadening their target space and applicability.

In HD, multiple studies have shown that the uninterrupted CAG repeat length in *HTT* gene, not the polyglutamine length in huntingtin protein, determines age-at-onset [24; 27; 30]. Age-at-onset of HD subjects carrying DI alleles not only supports this notion directly, but also points to novel therapeutic strategies. For example, converting CAG to CAA would decrease the length of uninterrupted CAG repeat without changing the length of polyglutamine or altering huntingtin protein. Indeed, our candidate BE strategies could shorten the length of uninterrupted CAG repeat by converting CAG to CAA at various sites in the CAG repeat without causing significant indels or off-target effects. In support, our candidate BE strategy modestly but significantly reduced the levels of CAG repeat expansion in mice, and HD knock-in mice carrying the CAA-interrupted repeats showed virtually zero repeat expansion. Given the role of the uninterrupted CAG repeat length as the most important disease determinant and a pivotal role for repeat instability in the modification of HD [26; 27], our data support the therapeutic potential of CAG-to-CAA conversion BE strategies in HD.

Our data are relevant for a number of reasons. Firstly, genetically supported targets significantly increase the success rate in clinical development [82]; our BE strategies derive directly from human genetic observations in HD individuals [24; 27; 30] that point to the uninterrupted CAG repeat length in *HTT* as the most direct therapeutic target in HD. Based on these human data, CAG-to-CAA conversion even near the 5'-end of the expanded CAG repeat may produce robust onset-delaying effects. In addition, if BE strategies are applied to fully penetrant 40 or 41 CAG canonical repeats, the repeats can become reduced penetrant (i.e., 36-39). Similarly, BE strategies may be able to convert some of the reduced penetrant CAG repeats (e.g., 36 and 37 CAG) to non-pathogenic (CAG < 36), which can prevent the manifestation of the disease. Secondly, lessons from the recent huntingtin-lowering clinical trial [83; 84] implied the importance of allele-specific approaches [35; 36]. Although CAG-to-CAA conversions in our experiments occurred on both mutant and normal *HTT*, our candidate BE strategies are expected to produce mutant allele-specific consequences. The amino acid sequence and levels of huntingtin were not altered, while the length of the uninterrupted CAG repeat was shortened. On the mutant allele, this shortening is expected to reduce the somatic instability of the repeat, reducing its disease-producing potential. On the normal allele, inherited variation in the length of the CAG repeat has not been associated with an abnormal phenotype [19], so the shortening of the CAG repeat is expected to be benign. Importantly, the mutant allele-specific consequences can be achieved without relying on individual genetic variations beyond the CAG repeat. Various SNP-targeting allele-specific approaches have been proposed [35; 36; 85; 86], but most of these can be applied only to a subset of the HD population depending on heterozygosity at the target site. Our BE strategies can achieve allele-specific consequences without targeting SNPs, and, therefore, can be applied to all HD subjects, representing a huge advantage over SNP-targeting allele-specific strategies. Lastly, BE with relaxed PAM requirements [51; 77] has increased the applicability of BE, and our study has further expanded the target space of this powerful technology. BE is appropriately viewed as a tool to correct disease-causing point mutations or to modify gene expression by introducing early stop codons or altering splice sites [77; 87; 88]. However, our study demonstrates that base conversion can address disease-causing repeat expansion mutations without involving DSB. The ramifications apply not only to HD but also to numerous other diseases that are caused by expansions of repeats [4-8], offering alternative therapeutic approaches for the repeat expansion disorders.

Although promising, hurdles must be overcome before CAG-to-CAA conversion BE strategies are applied to humans. The BE strategies that we evaluated did not robustly generate DI alleles that are found in humans, potentially due to the possibility that PAM at specific sites that are required to generate DI did not sufficiently support the activity of CBEs that we tested. Therefore, new CBEs that can efficiently generate DI alleles will greatly facilitate the development of rational treatments for HD. Also, the inability to directly target alternative toxic species such as RAN translation or exon 1A huntingtin fragment [89-92] may represent one of the limitations of our BE strategies for HD. Still, if the levels of those alternative toxic species are dependent of the length of uninterrupted CAG repeat, CAG-to-CAA conversion strategies may be able to ameliorate alternative toxic species-mediated HD pathogenesis. Although BE strategies can address the primary disease driver in principle, they may not produce any significant clinical benefits if they are applied too late in the disease. As we previously speculated, the timing of treatment might have negatively impacted the outcomes of the first ASO *HTT* lowering trial [35; 36; 83; 84]. Considering the evidence for significant levels of neurodegeneration at the onset of characteristic clinical manifestations [93], CAG-to-CAA conversion treatments may not produce any clinical improvements if applied late. With the expectation of mutant-specific consequences, we reason that BE strategies can be applied quite early without involving deficiency-related adverse effects [94-97] because CAG-to-CAA conversion is predicted to neither alter the amino acid sequence nor changes the expression levels of *HTT*. Regardless, these temporal aspects and safety features have to be determined. Finally, like other gene targeting strategies, the development of effective delivery methods is critical for applying BE therapeutically. The expansion-decreasing effects of our initial AAV injection experiments, while significant, may have been limited compared to cell culture systems by inefficient delivery and difficulty in targeting the repeat sequence. For the successful application of BE strategies to human HD, efficient delivery methods will be critical.

Given the lack of effective treatments for HD and the premature terminations of highly anticipated *HTT*-lowering clinical trials such as GENERATION-HD1 [83] and VIBRANT-HD (https://www.hda.org.uk/media/4418/novartis-vibrant-hd-community-letter-final-pdf.pdf), aiming at the most relevant target is becoming increasingly important. Our data reveal relevant strategies for addressing the target most strongly supported by human HD genetic data, the uninterrupted CAG repeat in *HTT*, therefore offer new opportunities for blocking the disorder at its cause. Although both great promise and significant hurdles exist

for the clinical application of BE strategies in HD, our data demonstrate the proof-of-concept of this technology as the basis for developing a rational treatment for HD and, potentially, for other repeat expansion disorders.

Acknowledgements

We thank Drs. Marcy E. MacDonald, James F. Gusella, and David Liu for helpful discussion. This work was supported by grants from Harvard NeuroDiscovery Center, NIH (NS105709, NS119471, NS091161, NS049206), and CHDI Foundation. B.P.K. was also supported by an MGH ECOR Howard M. Goodman Award and a CHDI Research Agreement (14962).

Declaration of Interests

V.C.W. was a founding scientific advisory board member with financial interest in Triplet Therapeutics Inc. Her financial interests were reviewed and are managed by Massachusetts General Hospital and Mass General Brigham in accordance with their conflict of interest policies. V.C.W. is a scientific advisory board member of LoQus23 Therapeutics Ltd. and has provided paid consulting services to Acadia Pharmaceuticals Inc., Alnylam Inc., Biogen Inc. and Passage Bio. V.C.W. has received research support from Pfizer Inc. B.P.K is an inventor on patents and/or patent applications filed by Mass General Brigham that describe genome engineering technologies. B.P.K. is a consultant for EcoR1 capital and is a scientific advisory board member of Acrigen Biosciences, Life Edit Therapeutics, and Prime Medicine. J-ML consults for Life Edit Therapeutics and serves in the advisory board of GenEdit Inc.

Data availability

RNAseq data of control and targeted iPSC clones have been deposited in Dryad (https://doi.org/10.5061/dryad.k3j9kd5cb).

References

1. Huntington, G. (1872). On chorea. Med Surg Rep 26, 320-321.
2. Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72, 971-983.
3. Bates, G.P., Dorsey, R., Gusella, J.F., Hayden, M.R., Kay, C., Leavitt, B.R., Nance, M., Ross, C.A., Scahill, R.I., Wetzel, R., et al. (2015). Huntington disease. Nature reviews Disease primers 1, 15005.
4. Gusella, J.F., and MacDonald, M.E. (2000). Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. Nature reviews Neuroscience 1, 109-115.
5. Ross, C.A. (2002). Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders. Neuron 35, 819-822.
6. Di Prospero, N.A., and Fischbeck, K.H. (2005). Therapeutics development for triplet repeat expansion diseases. Nature reviews Genetics 6, 756-765.
7. Orr, H.T., and Zoghbi, H.Y. (2007). Trinucleotide repeat disorders. Annual review of neuroscience 30, 575-621.
8. Depienne, C., and Mandel, J.L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? American journal of human genetics 108, 764-785.
9. Paulson, H.L., Bonini, N.M., and Roth, K.A. (2000). Polyglutamine disease and neuronal cell death. Proceedings of the National Academy of Sciences of the United States of America 97, 12957-12958.
10. Gatchel, J.R., and Zoghbi, H.Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. Nature reviews Genetics 6, 743-755.
11. Orr, H.T., Chung, M.Y., Banfi, S., Kwiatkowski, T.J., Jr., Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P., and Zoghbi, H.Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nature genetics 4, 221-226.
12. Pulst, S.M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., et al. (1996). Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nature genetics 14, 269-276.
13. Stevanin, G., Durr, A., and Brice, A. (2000). Clinical and molecular advances in autosomal dominant cerebellar ataxias: from genotype to phenotype and physiopathology. European journal of human genetics : EJHG 8, 4-18.
14. Zoghbi, H.Y., and Orr, H.T. (2000). Glutamine repeats and neurodegeneration. Annual review of neuroscience 23, 217-247.
15. Schols, L., Bauer, P., Schmidt, T., Schulte, T., and Riess, O. (2004). Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. The Lancet Neurology 3, 291-304.
16. Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nature reviews Genetics 6, 729-742.
17. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., et al. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nature genetics 4, 398-403.
18. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. Nature genetics 4, 387-392.
19. 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.
20. Kaplan, S., Itzkovitz, S., and Shapiro, E. (2007). A universal mechanism ties genotype to phenotype in trinucleotide diseases. PLoS computational biology 3, e235.
21. Telenius, H., Kremer, B., Goldberg, Y.P., Theilmann, J., Andrew, S.E., Zeisler, J., Adam, S., Greenberg, C., Ives, E.J., Clarke, L.A., et al. (1994). Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nature genetics 6, 409-414.
22. Shelbourne, P.F., Keller-McGandy, C., Bi, W.L., Yoon, S.R., Dubeau, L., Veitch, N.J., Vonsattel, J.P., Wexler, N.S., Group, U.S.-V.C.R., Arnheim, N., et al. (2007). Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Human molecular genetics 16, 1133-1142.
23. 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. Human molecular genetics 29, 2551-2567.
24. 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., team, T.-H., 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.
25. Swami, M., Hendricks, A.E., Gillis, T., Massood, T., Mysore, J., Myers, R.H., and Wheeler, V.C. (2009). Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. Human molecular genetics 18, 3039-3047.

26. GeM-HD Consortium. (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell 162, 516-526.
27. GeM-HD Consortium. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell 178, 887-900 e814.
28. 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., et al. (2021). Huntington's Disease Pathogenesis: Two Sequential Components. Journal of Huntington's disease 10, 35-51.
29. Lee, J.M., Chao, M.J., Harold, D., Abu Elneel, K., Gillis, T., Holmans, P., Jones, L., Orth, M., Myers, R.H., Kwak, S., et al. (2017). A modifier of Huntington's disease onset at the MLH1 locus. Human molecular genetics 26, 3859-3867.
30. Wright, G.E.B., Collins, J.A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., Becanovic, K., Drogemoller, 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. American journal of human genetics 104, 1116-1126.
31. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 533, 420-424.
32. Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Mochizuki, M., Miyabe, A., Araki, M., Hara, K.Y., et al. (2016). Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science 353.
33. Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. Nature 551, 464-471.
34. Levy, J.M., Yeh, W.H., Pendse, N., Davis, J.R., Hennessey, E., Butcher, R., Koblan, L.W., Comander, J., Liu, Q., and Liu, D.R. (2020). Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. Nature biomedical engineering 4, 97-110.
35. Shin, J.W., Hong, E.P., Park, S.S., Choi, D.E., Seong, I.S., Whittaker, M.N., Kleinstiver, B.P., Chen, R.Z., and Lee, J.M. (2022). Allele-specific silencing of the gain-of-function mutation in Huntington's disease using CRISPR/Cas9. JCI insight 7.
36. Shin, J.W., Shin, A., Park, S.S., and Lee, J.M. (2022). Haplotype-specific insertion-deletion variations for allele-specific targeting in Huntington's disease. Molecular therapy Methods & clinical development 25, 84-95.
37. Fjodorova, M., and Li, M. (2018). Robust Induction of DARPP32-Expressing GABAergic Striatal Neurons from Human Pluripotent Stem Cells. Methods in molecular biology 1780, 585-605.
38. Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nature biotechnology 31, 827-832.
39. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.
40. Vera Alvarez, R., Pongor, L.S., Marino-Ramirez, L., and Landsman, D. (2019). TPMCalculator: one-step software to quantify mRNA abundance of genomic features. Bioinformatics 35, 1960-1962.
41. 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. Human molecular genetics 8, 115-122.
42. Lloret, A., Dragileva, E., Teed, A., Espinola, J., Fossale, E., Gillis, T., Lopez, E., Myers, R.H., MacDonald, M.E., and Wheeler, V.C. (2006). Genetic background modifies nuclear mutant huntingtin accumulation and HD CAG repeat instability in Huntington's disease knock-in mice. Human molecular genetics 15, 2015-2024.
43. 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 systems biology 4, 29.
44. Lee, J.M., Pinto, R.M., Gillis, T., St Claire, J.C., and Wheeler, V.C. (2011). Quantification of age-dependent somatic CAG repeat instability in Hdh CAG knock-in mice reveals different expansion dynamics in striatum and liver. PloS one 6, e23647.
45. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. Behavioural brain research 125, 279-284.
46. McAllister, B., Donaldson, J., Binda, C.S., Powell, S., Chughtai, U., Edwards, G., Stone, J., Lobanov, S., Elliston, L., Schuhmacher, L.N., et al. (2022). Exome sequencing of individuals with Huntington's disease implicates FAN1 nuclease activity in slowing CAG expansion and disease onset. Nature neuroscience 25, 446-457.
47. Komor, A.C., Zhao, K.T., Packer, M.S., Gaudelli, N.M., Waterbury, A.L., Koblan, L.W., Kim, Y.B., Badran, A.H., and Liu, D.R. (2017). Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. Science advances 3, eaao4774.
48. Koblan, L.W., Doman, J.L., Wilson, C., Levy, J.M., Tay, T., Newby, G.A., Maianti, J.P., Raguram, A., and Liu, D.R. (2018). Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. Nature biotechnology 36, 843-846.

49. Thuronyi, B.W., Koblan, L.W., Levy, J.M., Yeh, W.H., Zheng, C., Newby, G.A., Wilson, C., Bhaumik, M., Shubina-Oleinik, O., Holt, J.R., et al. (2019). Continuous evolution of base editors with expanded target compatibility and improved activity. Nature biotechnology 37, 1070-1079.
50. Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O.O., Gootenberg, J.S., Mori, H., et al. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science 361, 1259-1262.
51. Walton, R.T., Christie, K.A., Whittaker, M.N., and Kleinstiver, B.P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science 368, 290-296.
52. Kim, Y.B., Komor, A.C., Levy, J.M., Packer, M.S., Zhao, K.T., and Liu, D.R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nature biotechnology 35, 371-376.
53. Gehrke, J.M., Cervantes, O., Clement, M.K., Wu, Y., Zeng, J., Bauer, D.E., Pinello, L., and Joung, J.K. (2018). An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. Nature biotechnology 36, 977-982.
54. Fu, J., Li, Q., Liu, X., Tu, T., Lv, X., Yin, X., Lv, J., Song, Z., Qu, J., Zhang, J., et al. (2021). Human cell based directed evolution of adenine base editors with improved efficiency. Nature communications 12, 5897.
55. Xu, L., Zhang, C., Li, H., Wang, P., Gao, Y., Mokadam, N.A., Ma, J., Arnold, W.D., and Han, R. (2021). Efficient precise in vivo base editing in adult dystrophic mice. Nature communications 12, 3719.
56. Duong, T.T., Lim, J., Vasireddy, V., Papp, T., Nguyen, H., Leo, L., Pan, J., Zhou, S., Chen, H.I., Bennett, J., et al. (2019). Comparative AAV-eGFP Transgene Expression Using Vector Serotypes 1-9, 7m8, and 8b in Human Pluripotent Stem Cells, RPEs, and Human and Rat Cortical Neurons. Stem cells international 2019, 7281912.
57. Villiger, L., Grisch-Chan, H.M., Lindsay, H., Ringnalda, F., Pogliano, C.B., Allegri, G., Fingerhut, R., Haberle, J., Matos, J., Robinson, M.D., et al. (2018). Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. Nature medicine 24, 1519-1525.
58. Koblan, L.W., Erdos, M.R., Wilson, C., Cabral, W.A., Levy, J.M., Xiong, Z.M., Tavarez, U.L., Davison, L.M., Gete, Y.G., Mao, X., et al. (2021). In vivo base editing rescues Hutchinson-Gilford progeria syndrome in mice. Nature 589, 608-614.
59. Mangiarini, L., Sathasivam, K., Mahal, A., Mott, R., Seller, M., and Bates, G.P. (1997). Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. Nature genetics 15, 197-200.
60. Manley, K., Shirley, T.L., Flaherty, L., and Messer, A. (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nature genetics 23, 471-473.
61. Kovtun, I.V., and McMurray, C.T. (2001). Trinucleotide expansion in haploid germ cells by gap repair. Nature genetics 27, 407-411.
62. Kennedy, L., Evans, E., Chen, C.M., Craven, L., Detloff, P.J., Ennis, M., and Shelbourne, P.F. (2003). Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. Human molecular genetics 12, 3359-3367.
63. Kovalenko, M., Dragileva, E., St Claire, J., Gillis, T., Guide, J.R., New, J., Dong, H., Kucherlapati, R., Kucherlapati, M.H., Ehrlich, M.E., et al. (2012). Msh2 acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. PloS one 7, e44273.
64. Pinto, R.M., Dragileva, E., Kirby, A., Lloret, A., Lopez, E., St Claire, J., Panigrahi, G.B., Hou, C., Holloway, K., Gillis, T., et al. (2013). Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. PLoS genetics 9, e1003930.
65. Ament, S.A., Pearl, J.R., Grindeland, A., St Claire, J., Earls, J.C., Kovalenko, M., Gillis, T., Mysore, J., Gusella, J.F., Lee, J.M., et al. (2017). High resolution time-course mapping of early transcriptomic, molecular and cellular phenotypes in Huntington's disease CAG knock-in mice across multiple genetic backgrounds. Human molecular genetics 26, 913-922.
66. Loupe, J.M., Pinto, R.M., Kim, K.H., Gillis, T., Mysore, J.S., Andrew, M.A., Kovalenko, M., Murtha, R., Seong, I., Gusella, J.F., et al. (2020). Promotion of somatic CAG repeat expansion by Fan1 knock-out in Huntington's disease knock-in mice is blocked by Mlh1 knock-out. Human molecular genetics 29, 3044-3053.
67. Kacher, R., Lejeune, F.X., Noel, S., Cazeneuve, C., Brice, A., Humbert, S., and Durr, A. (2021). Propensity for somatic expansion increases over the course of life in Huntington disease. eLife 10.
68. Kennedy, L., and Shelbourne, P.F. (2000). Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? Human molecular genetics 9, 2539-2544.
69. Wu, Z., Yang, H., and Colosi, P. (2010). Effect of genome size on AAV vector packaging. Molecular therapy : the journal of the American Society of Gene Therapy 18, 80-86.
70. Carvalho, L.S., Turunen, H.T., Wassmer, S.J., Luna-Velez, M.V., Xiao, R., Bennett, J., and Vandenberghe, L.H. (2017). Evaluating Efficiencies of Dual AAV Approaches for Retinal Targeting. Frontiers in neuroscience 11, 503.
71. Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346, 1258096.
72. Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell 157, 1262-1278.

73. Stadtmauer, E.A., Fraietta, J.A., Davis, M.M., Cohen, A.D., Weber, K.L., Lancaster, E., Mangan, P.A., Kulikovskaya, I., Gupta, M., Chen, F., et al. (2020). CRISPR-engineered T cells in patients with refractory cancer. Science 367.
74. Gillmore, J.D., Gane, E., Taubel, J., Kao, J., Fontana, M., Maitland, M.L., Seitzer, J., O'Connell, D., Walsh, K.R., Wood, K., et al. (2021). CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. The New England journal of medicine 385, 493-502.
75. Wang, B., Iriguchi, S., Waseda, M., Ueda, N., Ueda, T., Xu, H., Minagawa, A., Ishikawa, A., Yano, H., Ishi, T., et al. (2021). Generation of hypoimmunogenic T cells from genetically engineered allogeneic human induced pluripotent stem cells. Nature biomedical engineering 5, 429-440.
76. Neugebauer, M.E., Hsu, A., Arbab, M., Krasnow, N.A., McElroy, A.N., Pandey, S., Doman, J.L., Huang, T.P., Raguram, A., Banskota, S., et al. (2022). Evolution of an adenine base editor into a small, efficient cytosine base editor with low off-target activity. Nature biotechnology.
77. Rees, H.A., and Liu, D.R. (2018). Base editing: precision chemistry on the genome and transcriptome of living cells. Nature reviews Genetics 19, 770-788.
78. Newby, G.A., Yen, J.S., Woodard, K.J., Mayuranathan, T., Lazzarotto, C.R., Li, Y., Sheppard-Tillman, H., Porter, S.N., Yao, Y., Mayberry, K., et al. (2021). Base editing of haematopoietic stem cells rescues sickle cell disease in mice. Nature 595, 295-302.
79. Kingwell, K. (2022). Base editors hit the clinic. Nature reviews Drug discovery 21, 545-547.
80. Eisenstein, M. (2022). Base editing marches on the clinic. Nature biotechnology 40, 623-625.
81. McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. Nature reviews Genetics 11, 786-799.
82. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nature genetics 47, 856-860.
83. Relations, R.G.M. (2021). Roche provides update on tominersen programme in manifest Huntington's disease. In. (
84. Sheridan, C. (2021). Questions swirl around failures of disease-modifying Huntington's drugs. Nature biotechnology 39, 650-652.
85. Shin, J.W., Kim, K.H., Chao, M.J., Atwal, R.S., Gillis, T., MacDonald, M.E., Gusella, J.F., and Lee, J.M. (2016). Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. Human molecular genetics.
86. Monteys, A.M., Ebanks, S.A., Keiser, M.S., and Davidson, B.L. (2017). CRISPR/Cas9 Editing of the Mutant Huntingtin Allele In Vitro and In Vivo. Molecular therapy : the journal of the American Society of Gene Therapy 25, 12-23.
87. Kuscu, C., Parlak, M., Tufan, T., Yang, J., Szlachta, K., Wei, X., Mammadov, R., and Adli, M. (2017). CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. Nature methods 14, 710-712.
88. Kim, K., Ryu, S.M., Kim, S.T., Baek, G., Kim, D., Lim, K., Chung, E., Kim, S., and Kim, J.S. (2017). Highly efficient RNA-guided base editing in mouse embryos. Nature biotechnology 35, 435-437.
89. Yang, S., Yang, H., Huang, L., Chen, L., Qin, Z., Li, S., and Li, X.J. (2020). Lack of RAN-mediated toxicity in Huntington's disease knock-in mice. Proceedings of the National Academy of Sciences of the United States of America 117, 4411-4417.
90. Neueder, A., Landles, C., Ghosh, R., Howland, D., Myers, R.H., Faull, R.L.M., Tabrizi, S.J., and Bates, G.P. (2017). The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. Scientific reports 7, 1307.
91. Banez-Coronel, M., Ayhan, F., Tarabochia, A.D., Zu, T., Perez, B.A., Tusi, S.K., Pletnikova, O., Borchelt, D.R., Ross, C.A., Margolis, R.L., et al. (2015). RAN Translation in Huntington Disease. Neuron 88, 667-677.
92. Sathasivam, K., Neueder, A., Gipson, T.A., Landles, C., Benjamin, A.C., Bondulich, M.K., Smith, D.L., Faull, R.L., Roos, R.A., Howland, D., et al. (2013). Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. Proceedings of the National Academy of Sciences of the United States of America 110, 2366-2370.
93. Paulsen, J.S., Langbehn, D.R., Stout, J.C., Aylward, E., Ross, C.A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L.J., et al. (2008). Detection of Huntington's disease decades before diagnosis: the Predict-HD study. Journal of neurology, neurosurgery, and psychiatry 79, 874-880.
94. Lopes, F., Barbosa, M., Ameur, A., Soares, G., de Sa, J., Dias, A.I., Oliveira, G., Cabral, P., Temudo, T., Calado, E., et al. (2016). Identification of novel genetic causes of Rett syndrome-like phenotypes. Journal of medical genetics 53, 190-199.
95. Rodan, L.H., Cohen, J., Fatemi, A., Gillis, T., Lucente, D., Gusella, J., and Picker, J.D. (2016). A novel neurodevelopmental disorder associated with compound heterozygous variants in the huntingtin gene. European journal of human genetics : EJHG 24, 1826-1827.
96. Dietrich, P., Johnson, I.M., Alli, S., and Dragatsis, I. (2017). Elimination of huntingtin in the adult mouse leads to progressive behavioral deficits, bilateral thalamic calcification, and altered brain iron homeostasis. PLoS genetics 13, e1006846.
97. Wang, G., Liu, X., Gaertig, M.A., Li, S., and Li, X.J. (2016). Ablation of huntingtin in adult neurons is nondeleterious but its depletion in young mice causes acute pancreatitis. Proceedings of the National Academy of Sciences of the United States of America 113, 3359-3364.

Figure Legend

Figure 1. Effects of CAA interruption on HD age-at-onset.

(A-C) Least square approximation was performed to estimate the additional effects of LI (red circles in panel b and c) and DI on age-at-onset (green circles in panel c and d). We varied the CAG length of HD participants carrying LI or DI, and subsequently calculated sum of square (SS) to identify the CAG repeat that explained the maximum variance in age-at-onset of these allele carriers. Y-axis and X-axis represent age-at-onset and CAG repeat length, respectively. Grey circles and black trend lines respectively represent HD participants with CR alleles and their onset-CAG relationship. SS means sum of square.

(D) To illustrate the magnitude of the impact of a therapeutic base editing strategy of converting a CR allele to a DI allele by changing CAG to CAA, an example of a CR of 43 CAG (45 glutamine) with a mean observed onset of 48 years is displayed. In this example, therapeutic conversion of the 42nd CAG to CAA by BE would produce a DI allele of 41 CAG (45 glutamine). Considering the additional effect of DI alleles in HD patients, a 41 CAG / 45 glutamine DI allele would produce an onset similar to a CR allele of 40 CAG / 42 glutamine, with a mean onset age of 60. Therefore, CAG-to-CAA conversion in HD subjects with 43 CR repeats could delay onset by 12 years.

Figure 2. Cytosine base editors and gRNAs for CAG-to-CAA conversion in HD.

(A) Constituents of base editing are displayed.

(B) Schematic of cytosine base editors (CBEs) that can generate C-to-T edits within a finite edit window at a fixed distance from the PAM.

(C) CBE variants described in the literature and used in this study (lower 4) are shown, including the evoCDA1-based SpG CBE that should function more efficiently in GC nucleotide contexts. Protospacer-adjacent motif (PAM), guide RNA (gRNA), uracil glycosylase inhibitor (UGI), rat APOBEC1 deaminase domain (rAPO1), evolved CDA1 cytosine deaminase domain (evoCDA).

(D) The target region, gRNAs, and expected hybridization sites of the 8 gRNAs are shown.

Figure 3. Levels of CAG-to-CAA conversion by BE strategies.

Only CAG-to-CAA conversion showed significantly increased levels over the baseline sequencing errors. Thus, we calculated the percentage of CAA in the cells that were treated with a combination of cytosine base editors (A, BE4max; B, BE4-NG; C, BE4-SpG; and D, evo-SpG) and gRNAs. HEK293 cells without any treatment (i.e., Cell) were combined (n=8) and plotted for each base editor. EV represents HEK293 cells treated with a base editor and empty vector for gRNA. *, significant by Bonferroni corrected p-value < 0.05 (8 tests for each base editor).

Figure 4. Sites of CAG-to-CAA conversion by BE strategies.

We calculated the percentage of sequence reads containing CAA at specific sites relative to all sequence reads. For example, 27.7% conversion at the 2nd CAG by BE4max-gRNA 1 (top left panel, red) means 27.7 % of all sequence reads from the 16 or 17 CAG alleles have CAA at the 2nd CAG. X-axis and y-axis represent the position of the CAG and percent conversion. Each panel represents a tested gRNA. Plots were based on the mean of 3 independent transfection experiments in HEK293 cells after subtracting corresponding empty vector (EV)-treated cell data. Red, blue, purple, and cyan traces represent BE4max, BE4-NG, BE4-SpG, and evo-SpG.

Figure 5. Allele specificity and molecular outcomes of candidate BE strategies.

(A) To overcome the limitations of patient-derived iPSC and differentiated neurons, we developed HEK293 carrying an adult-onset CAG repeat by replacing one of the normal repeats with 51 canonical CAG (namely HEK293-51 CAG). Red and green bars represent respectively mutant and normal *HTT* in HEK293-51 CAG cells.

(B and C) The HEK293-51 CAG cells were treated with BE4max-gRNA 1 and analyzed to determine the levels of in-frame insertion (B) and in-frame deletion (C) at the time of treatment.

(D) The HEK293-51 CAG cells were treated with the gRNA 1 and analyzed by MiSeq to determine the levels of allele specificity. Conversion efficiency on the Y-axis indicates the percentage of sequence reads containing the CAG-to-CAA conversion at the target site. * represents uncorrected p-value < 0.05 by Student t-test.

(E) Original HEK293 cells and HEK293-51 CAG cells were treated with empty vector (EV), or candidate BE strategies (BE4max-gRNA 1 and BE4max-gRNA 2) and subjected to immunoblot analysis; representative blot is shown in panel E.

(F) Four independent experiments were performed, and we performed one-sample t-test to determine whether BE-treated cells show different total HTT protein levels compared to EV-treated cells. Nothing was significant by p-value < 0.05.

Figure 6. RNAseq analysis of BE strategies confirms the lack of transcriptome alternation.

(A) HEK293 cells were treated with empty vector (EV) or candidate BE strategies such as BE4max-gRNA 1 (gRNA 1), and BE4max-gRNA 2 (gRNA 2) for RNAseq analysis. MiSeq analysis was also performed to judge the levels of CAG-to-CAA conversion. ****, p-value < 0.0001 by Student t-test (n=4).

(B) Confirming the lack of significantly altered genes in BE4max-gRNA 1 or BE4max-gRNA 2, we compared all BE-treated samples (n=8) with all EV-treated samples (n=4) to increase the power in the RNAseq differential gene expression analysis. Each circle in the volcano plot represents a gene analyzed in the RNAseq; *HTT* is indicated by a filled red circle. A red horizontal line represents false discovery rate of 0.05, showing that none was significantly altered by candidate BE strategies.

(C) We also compared two groups of randomly assigned samples (6 samples vs. 6 samples) to understand the shape of the volcano plot when there were no significant genes.

Figure 7. Impacts of CAA interruption on CAG repeat instability.

(A - C) DNA samples (liver and tail) of BE-treated mice were analyzed to quantify somatic repeat expansion. We performed linear regression analysis to model the levels of repeat expansion as a function of treatment, CAG repeat in tail (A), age (B), and with other covariates (i.e., experimental batch, sex, tail CAG and age). Summary of the statistical analysis is summarized in the panel C.

(D and E) To determine the maximal impacts of CAA interruption on the repeat expansion, HD knock-in mice carrying CAA interrupted repeats were analyzed. Liver samples of 105 uninterrupted CAG repeat (D) and interrupted repeat (E) were analyzed at 5 months. Representative fragment analysis is displayed. Red arrows

indicate the modal alleles representing inherited CAG repeats; peaks at the right side of the modal peaks (red arrows) represent expanded repeats.

Figure 1

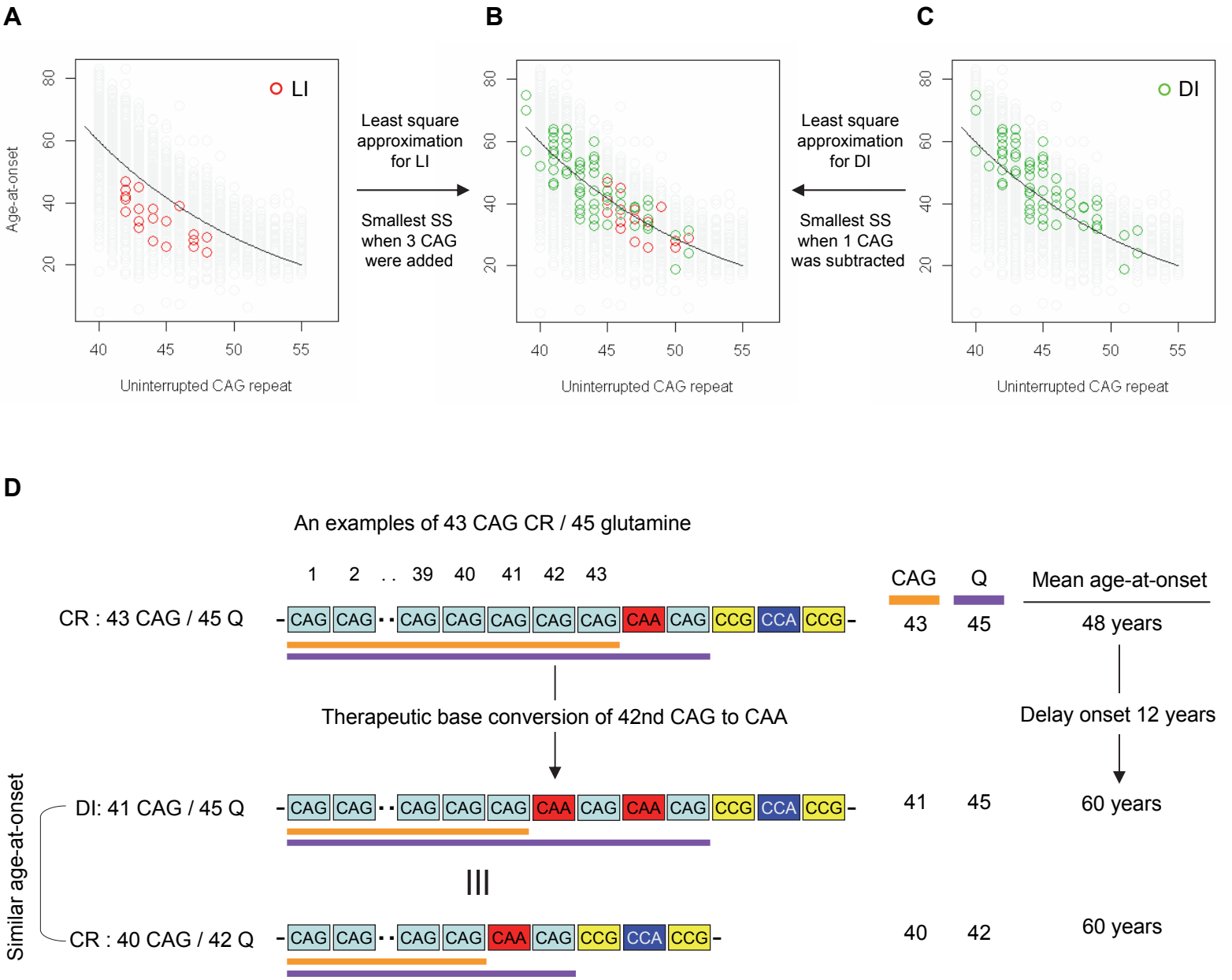


Figure 2

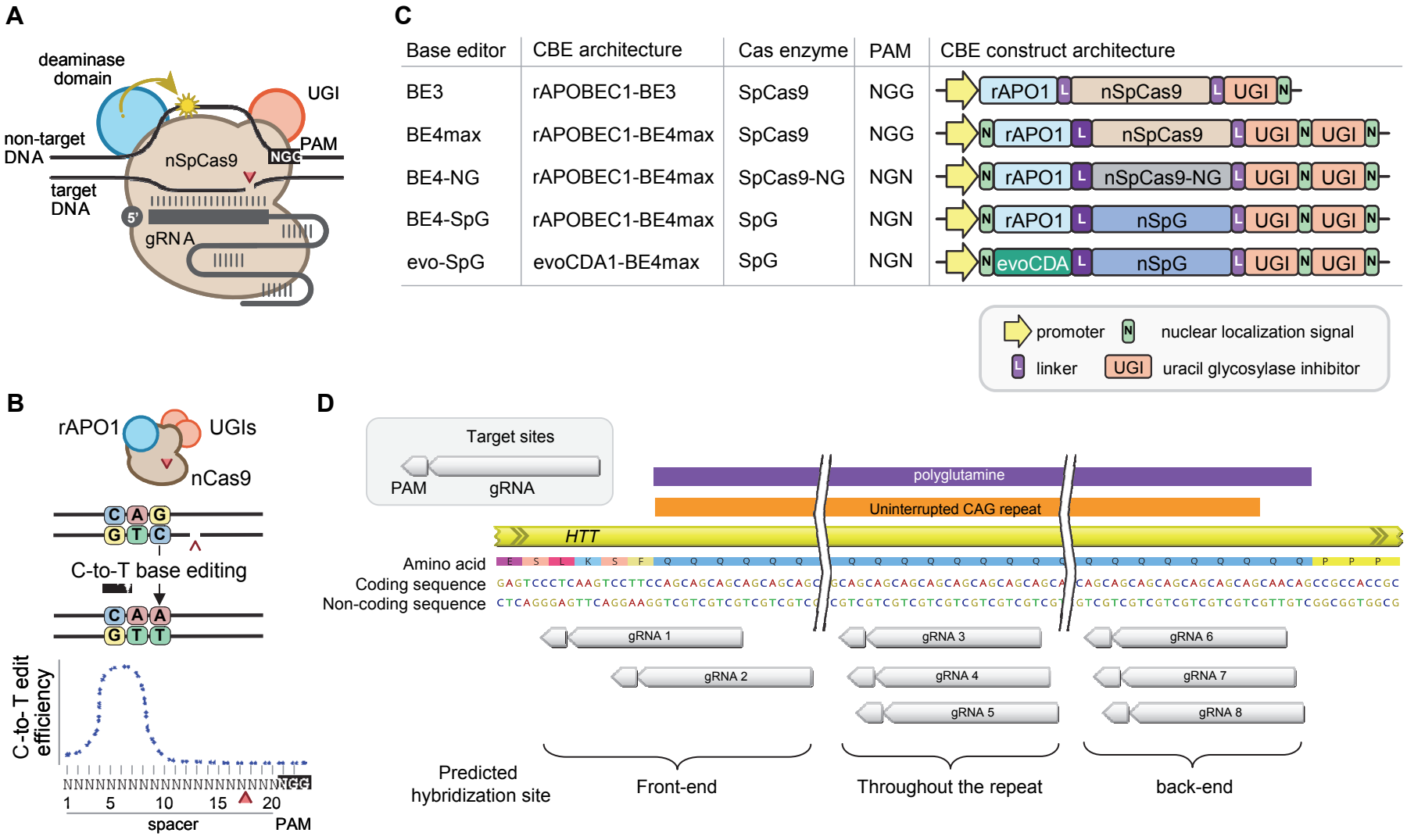


Figure 3

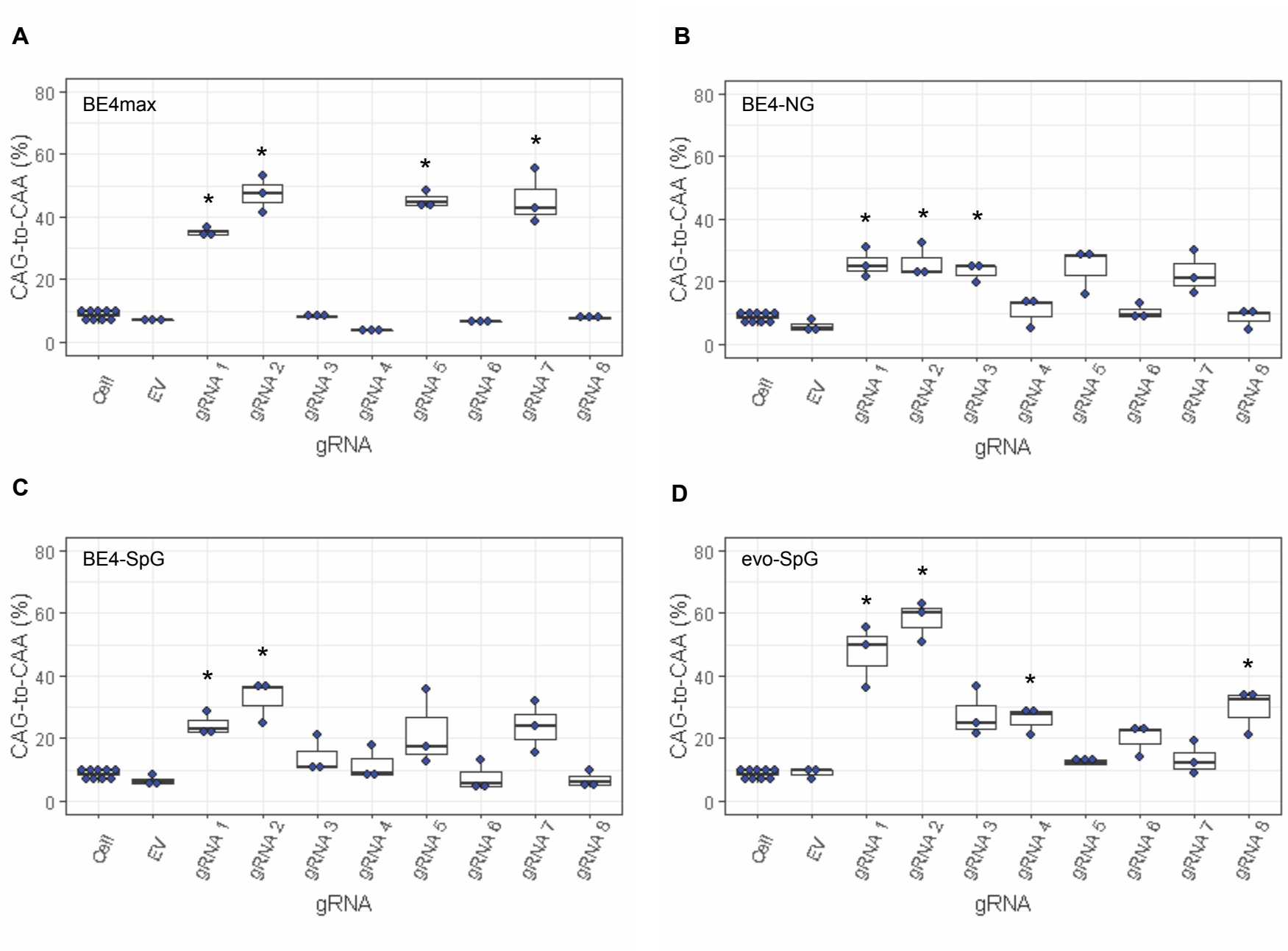


Figure 4

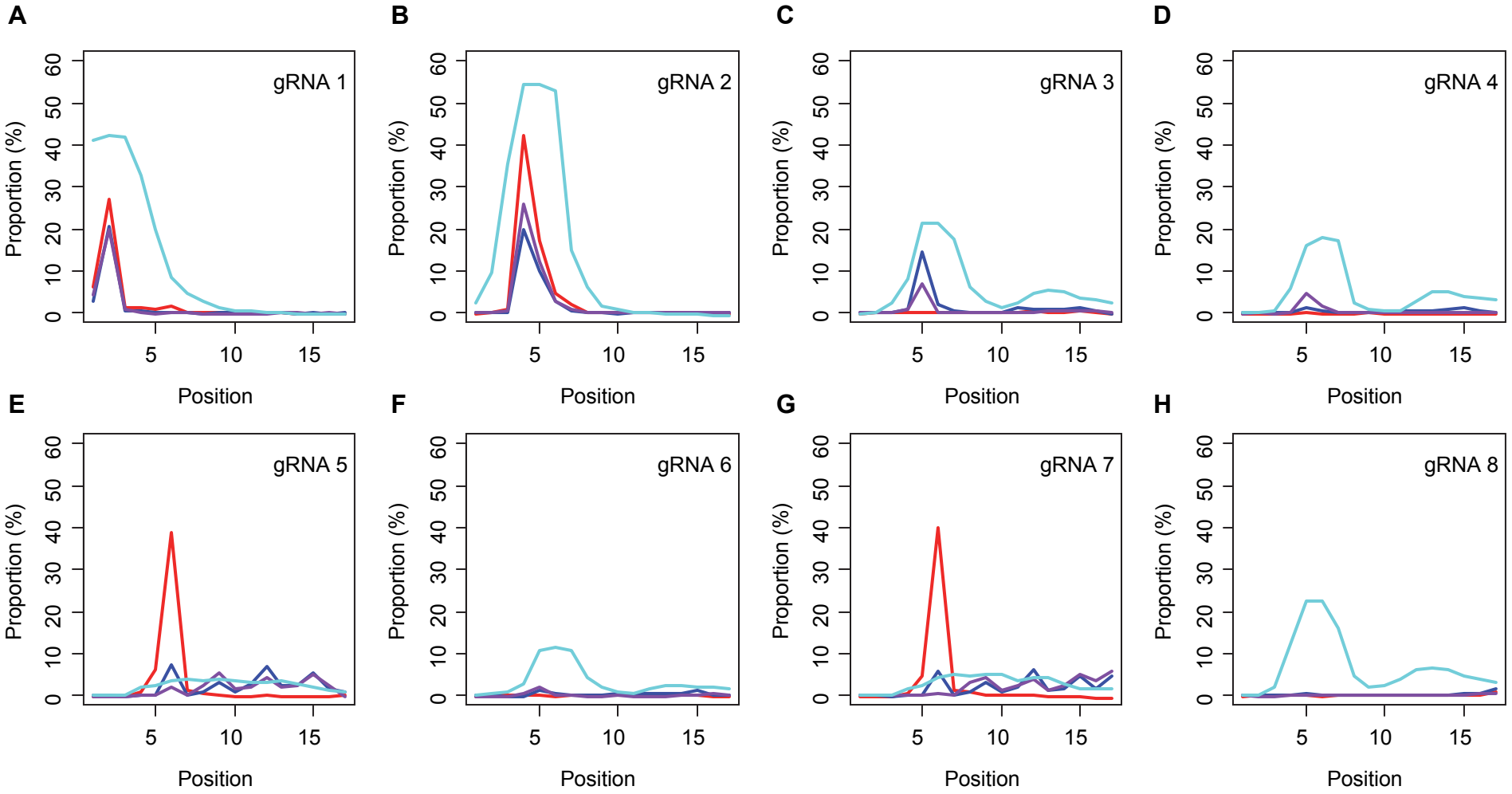


Figure 5

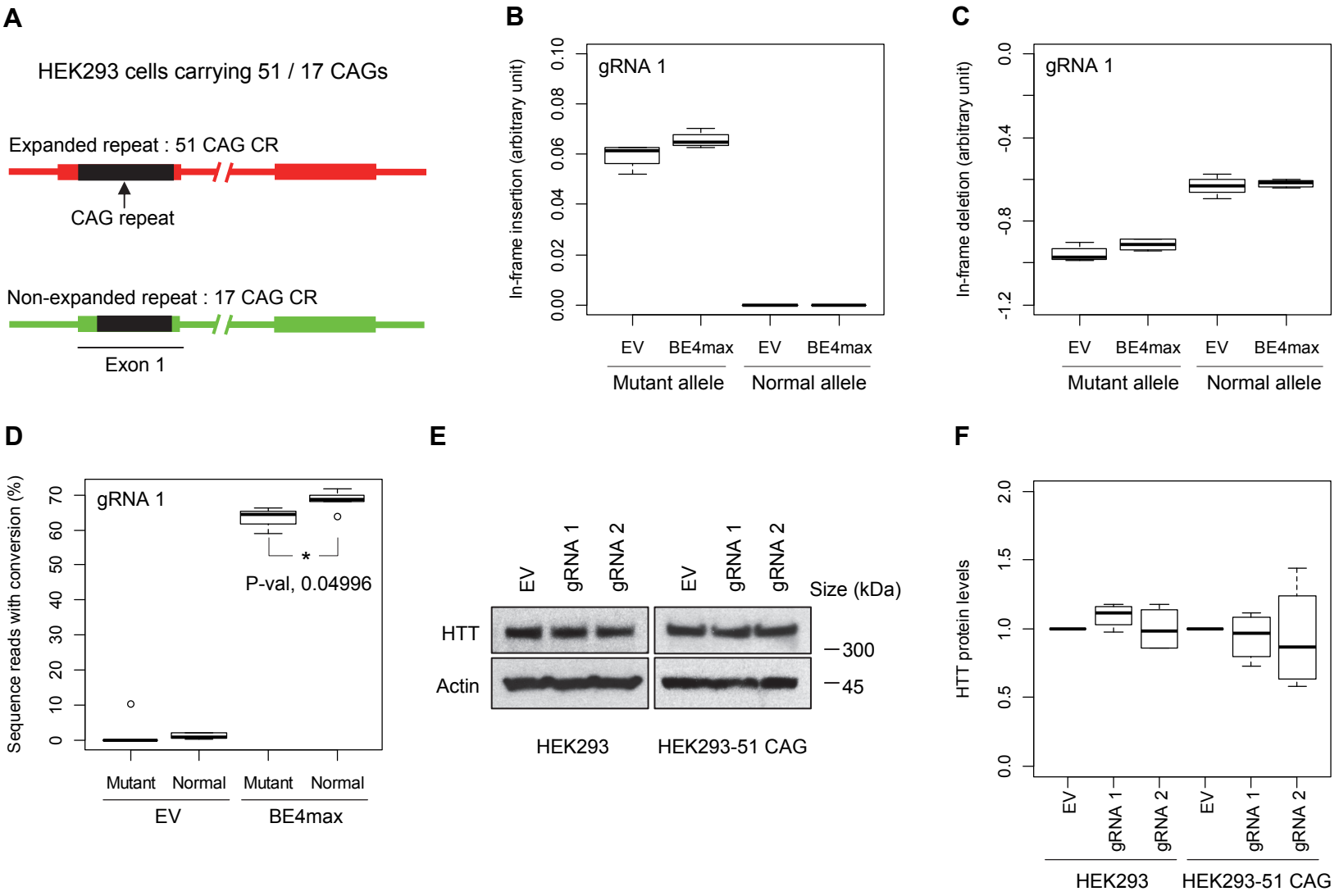


Figure 6

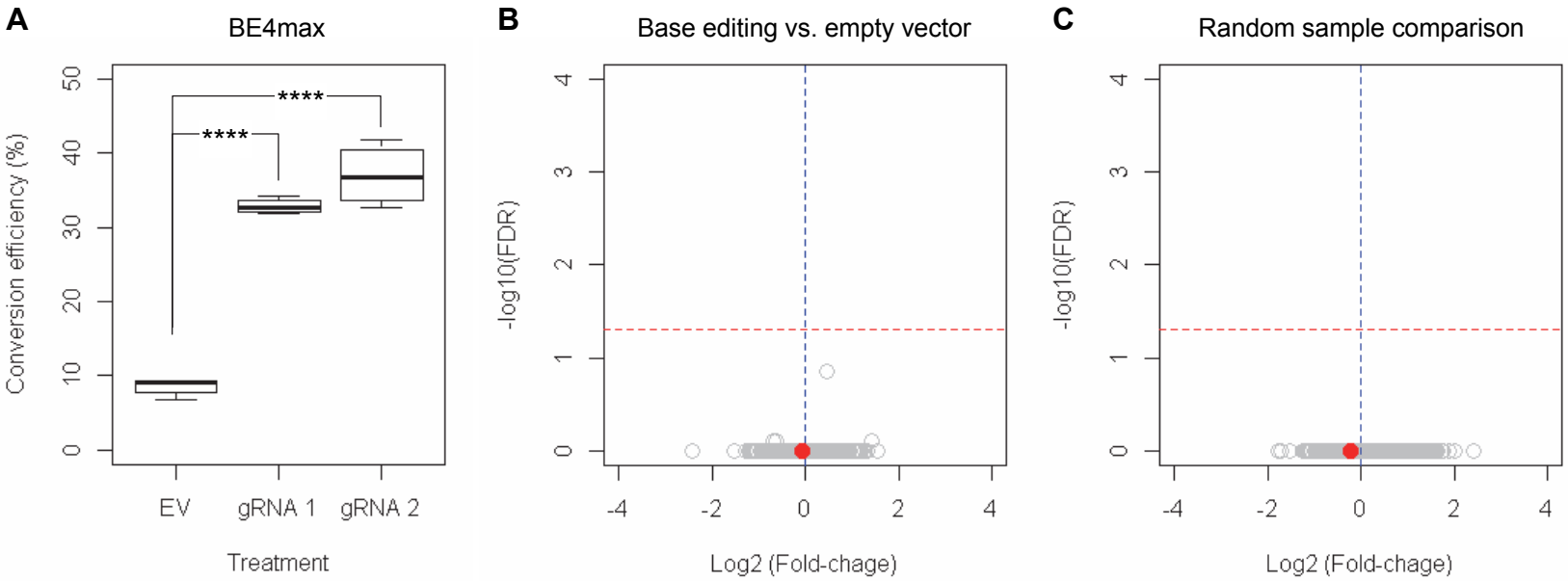
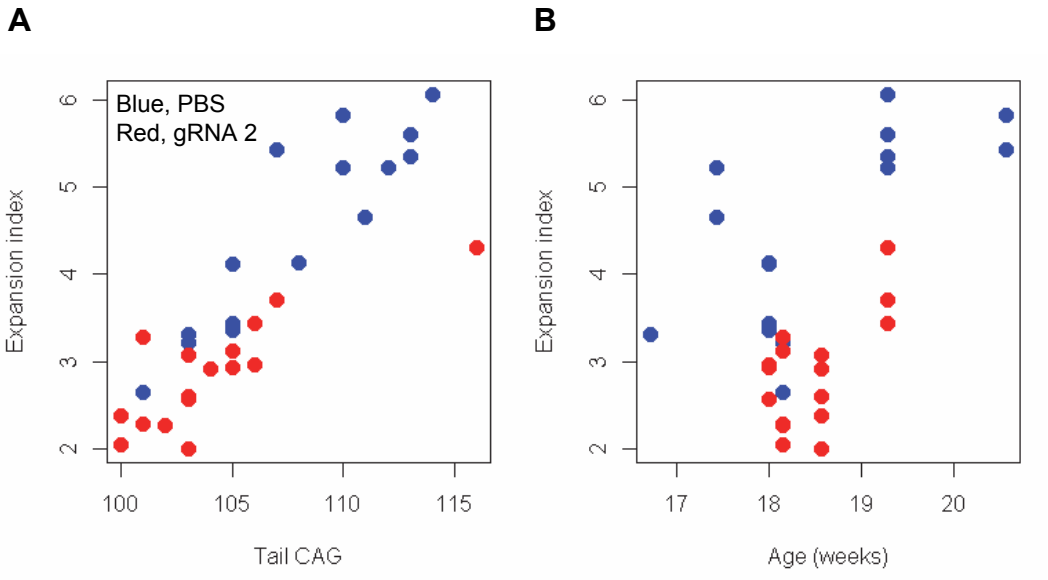


Figure 7



C

Predictor	Effect size	P-value
Batch	-0.23845	0.121678
Sex	-0.19829	0.163209
Tail CAG	0.16994	9.20E-09
Age (week)	0.39639	0.000211
CBE + gRNA 2	-0.91531	1.78E-06