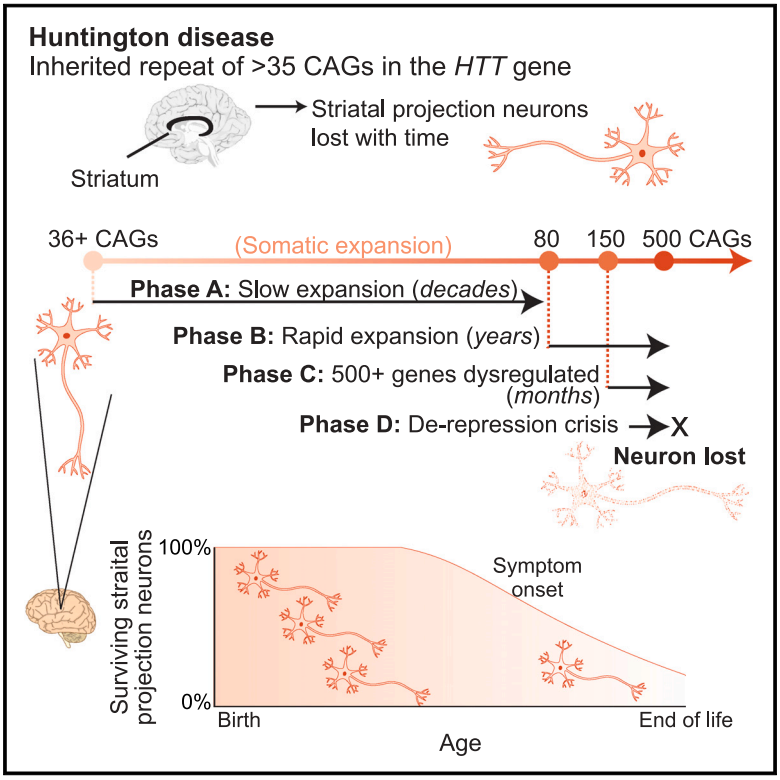
Cell

Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease

Graphical abstract



Highlights

- The HD-causing CAG repeat expands somatically in HD-vulnerable neurons
- Over decades, the unstable alleles expand to more than 10 times their initial length
- The CAG repeat becomes toxic after expanding beyond 150 CAGs
- HD pathogenesis is a DNA-repeat expansion process for almost all of a neuron's life

Authors

Robert E. Handsaker, Seva Kashin, Nora M. Reed, ..., Kiku Ichihara, Sabina Berretta, Steven A. McCarroll

Correspondence

handsake@broadinstitute.org (R.E.H.), skashin@broadinstitute.org (S.K.), sberretta@mclean.harvard.edu (S.B.), smccarro@broadinstitute.org (S.A.M.)

In brief

Single-cell measurement of the Huntington's disease-causing CAG repeat reveals that somatic expansion of this repeat drives pathological changes in neurons, providing insights into disease progression, with implications for therapeutics in HD and potentially other DNA-repeat expansion disorders.

Handsaker et al., 2025, Cell *188*, 1–17
February 6, 2025
https://doi.org/10.1016/j.cell.2024.11.038

CellPress

Cell

CellPress
OPEN ACCESS

Article

Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease

Robert E. Handsaker,[1,2,7,*] Seva Kashin,[1,2,7,*] Nora M. Reed,[1,2,7] Steven Tan,[1,2] Won-Seok Lee,[1,2] Tara M. McDonald,[1,2] Kiely Morris,[3] Nolan Kamitaki,[1,2] Christopher D. Mullally,[1,2] Neda R. Morakabati,[3] Melissa Goldman,[1,2] Gabriel Lind,[1,2] Rhea Kohli,[1,2] Elisabeth Lawton,[3] Marina Hogan,[1,2] Kiku Ichihara,[1,2] Sabina Berretta,[1,3,4,5,8,*] and Steven A. McCarroll[1,2,5,6,8,9,*]

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3]McLean Hospital, Belmont, MA 02478, USA
[4]Department of Psychiatry, Harvard Medical School, Boston, MA 02215, USA
[5]Program in Neuroscience, Harvard Medical School, Boston, MA 02215, USA
[6]Howard Hughes Medical Institute, Boston, MA 02215, USA
[7]These authors contributed equally
[8]Senior author
[9]Lead contact
*Correspondence: handsake@broadinstitute.org (R.E.H.), skashin@broadinstitute.org (S.K.), sberretta@mclean.harvard.edu (S.B.), smccarro@broadinstitute.org (S.A.M.)
https://doi.org/10.1016/j.cell.2024.11.038

SUMMARY

In Huntington's disease (HD), striatal projection neurons (SPNs) degenerate during midlife; the core biological question involves how the disease-causing DNA repeat $(CAG)_n$ in the *huntingtin* (*HTT*) gene leads to neurodegeneration after decades of biological latency. We developed a single-cell method for measuring this repeat's length alongside genome-wide RNA expression. We found that the *HTT* CAG repeat expands somatically from 40–45 to 100–500+ CAGs in SPNs. Somatic expansion from 40 to 150 CAGs had no apparent cell-autonomous effect, but SPNs with 150–500+ CAGs lost positive and then negative features of neuronal identity, de-repressed senescence/apoptosis genes, and were lost. Our results suggest that somatic repeat expansion beyond 150 CAGs causes SPNs to degenerate quickly and asynchronously. We conclude that in HD, at any one time, most neurons have an innocuous but unstable *HTT* gene and that HD pathogenesis is a DNA process for almost all of a neuron's life.

INTRODUCTION

Huntington's disease (HD) is a fatal genetic neurodegenerative disease. Most people who inherit an HD-causing allele have no symptoms for decades, then develop uncontrolled movements (chorea) and cognitive and psychiatric symptoms; the motor symptoms progress to severe impairment, rigidity, and lethality. Persons with HD have severe atrophy of the striatum in which they have lost its principal neurons, striatal projection neurons (SPNs, also called medium spiny neurons or MSNs). No treatments are known to prevent or slow HD.

HD segregates in families in a dominant manner; its genetic cause is an inherited DNA triplet repeat $(CAG)_n$ of variable length, within exon 1 of the *huntingtin* (*HTT*) gene.[1] Most people have inherited alleles with 15–30 consecutive CAGs, but persons with HD have inherited a germline allele with 36 or more consecutive CAGs (36–55 in 98% of cases, 40–49 in 90%).[2] Among persons with HD, longer CAG repeats lead to earlier HD onset, although with substantial inter-individual variation.[3]

Three core aspects of HD are unexplained: its cell-type-specific pathology, its decades-long pre-symptomatic latency, and the series of events by which inherited alleles lead to neurodegeneration.

First, neurodegeneration in HD is highly cell-type specific; in the striatum, most SPNs are lost, while interneurons and glia survive, even though all of these cell types express HTT.

Second, HD symptoms take decades to manifest. Persons who have inherited common HD-causing alleles (40–45 CAGs) reach adulthood with scores on cognitive and motor tests comparable to those of healthy individuals without HD-causing alleles.[4] The average age of clinical motor onset is 40–50 years, preceded by subtle changes in neuroimaging and fluid biomarkers in younger adults.[4,5] The long latency preceding HD symptoms is often attributed to biological processes with slowly cumulative toxicity or to a decades-long lag phase in the development of protein aggregates.

A third mystery involves how inherited *HTT* alleles lead to HD. The encoded protein (HTT), which contains a polyglutamine tract that is encoded by the CAG repeat, has many biological functions; loss, over-expression, and genetic manipulation of *HTT* produce

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

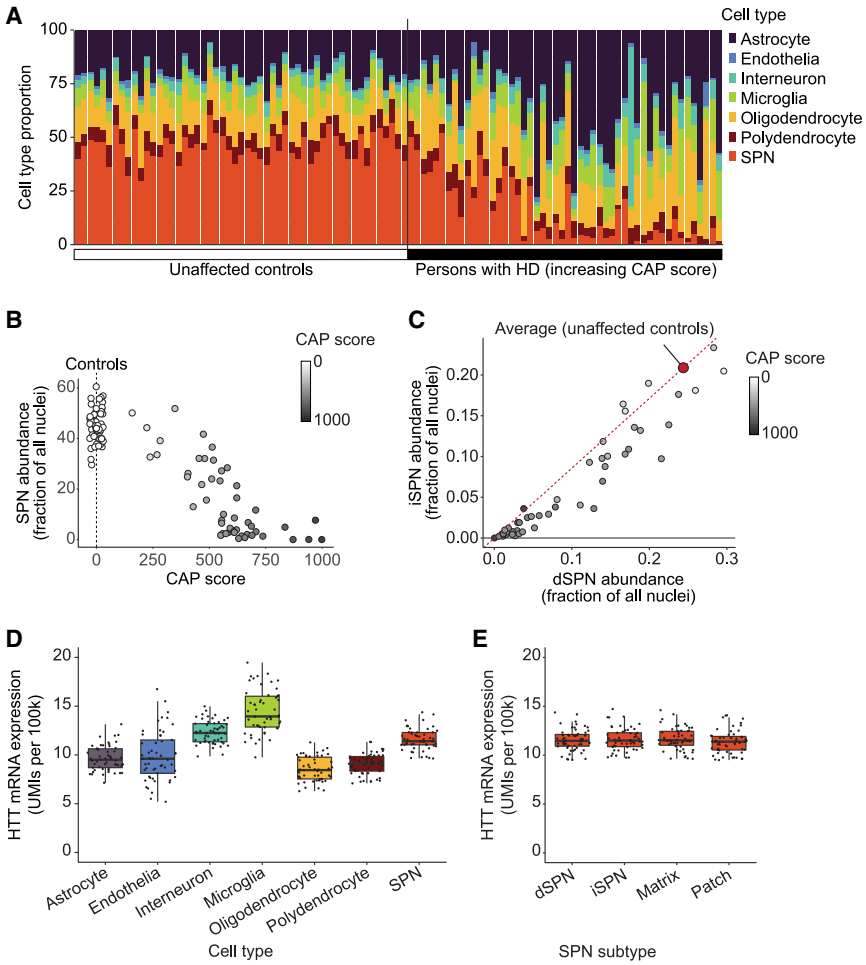


Figure 1. SPN loss in persons with HD

(A) Cell-type proportions in the striatum (anterior caudate) of each donor.

(B) SPN loss with HD progression (increasing CAP score). Unaffected controls in white; among persons with HD, darker shades of gray represent increasing CAP score. These same data are shown on a log scale in Figure S2A.

(C) Decline in iSPNs (D2 SPNs, y axis) and dSPNs (D1 SPNs, x axis) with HD progression. Gray shading as in (B).

(D and E) Expression of *HTT* transcripts (units: UMIs per 100,000) in the nuclei of (D) striatal cell types and (E) SPN subtypes, among 53 control (unaffected) donors. Boxes represent the interquartile range; whiskers extend beyond the hinges by 1.5 times the interquartile range.

See also Figures S1 and S2.

diverse phenotypes in many species and cell types.[6] Diverse biological hypotheses are considered plausible for HD, with recent studies focusing on embryonic development,[7] mitochondria,[8] vascular cells,[9] microglia,[10] and long-range circuitry effects.[11]

An important clue may reside in the long-observed phenomenon of somatic mosaicism in HD. The length of the CAG repeat varies somatically; this somatic mosaicism is pronounced in the brain,[12,13] is greater in neurons than in glia,[14,15] and is greater in persons with earlier-than-expected motor symptom onset.[16] The biological significance of somatic mosaicism in *HTT* has been debated for 30 years, with a dominant view that somatic instability simply modifies the inherent toxicity of HD-causing alleles. However, the recent discovery of common human genetic polymorphisms that modify age at onset[17] suggests that much disease-significant biology may involve somatic instability of the repeat.[18] HD motor onset is delayed by a synonymous CAG-to-CAA variant (within the CAG repeat) that reduces the repeat's instability without shortening the encoded polyglutamine.[17,19] Age of HD onset is also shaped by common genetic variation at many DNA maintenance genes, including *MSH3*, *FAN1*, *MLH1*, *LIG1*, *PMS1*, and *PMS2*.[17,20] Proteins encoded by these genes affect DNA-repeat stability.[21–29]

In this work, to uncover the pathophysiological process in HD and its relationship to the CAG repeat in *HTT*, we developed a molecular approach to measure this repeat at single-cell resolution, concurrent with the same cells' genome-wide RNA expression. This approach allowed us to recognize biological changes that result directly and cell autonomously from the CAG repeat's somatic expansion.

RESULTS

SPN vulnerability, HTT expression, and case-control differences in HD

We first used conventional single-nucleus RNA sequencing (snRNA-seq)[30,31] to analyze RNA expression in 581,273 nuclei sampled from the anterior part of the caudate nucleus—the largest component of the striatum and the region most affected in HD—from 50 persons with HD and 53 unaffected controls (mean 5,643 cell nuclei per donor) (Table S1). Each nucleus was assigned to one of seven major cell types, based on the RNAs it expressed (Figures 1A and S1).

The loss of SPNs during the course of HD was apparent in the declining numbers of SPN nuclei (as a fraction of all cell nuclei sampled) (Figures 1A and 1B). To analyze together persons with many different ages and inherited CAG-repeat lengths, we used the CAG-age-product (CAP) score, a common estimate of onset and progression in HD,[32] which is calculated as *age* * (*inheritedCAGlength* − 33.66). Persons with CAP scores up to 300, generally corresponding to the long latent period prior to clinical motor onset, tended to have SPN proportions just slightly lower than the average unaffected control brain donor (Figures 1B and S2A). In contrast, SPNs appeared to be lost at a substantial rate in donors with a CAP score greater than 350 (generally, donors with manifest HD symptoms), as evidenced by a steep downward slope in the relationship of SPN abundance to CAP score (Figures 1B and S2A). Persons with a CAP

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

CellPress
OPEN ACCESS

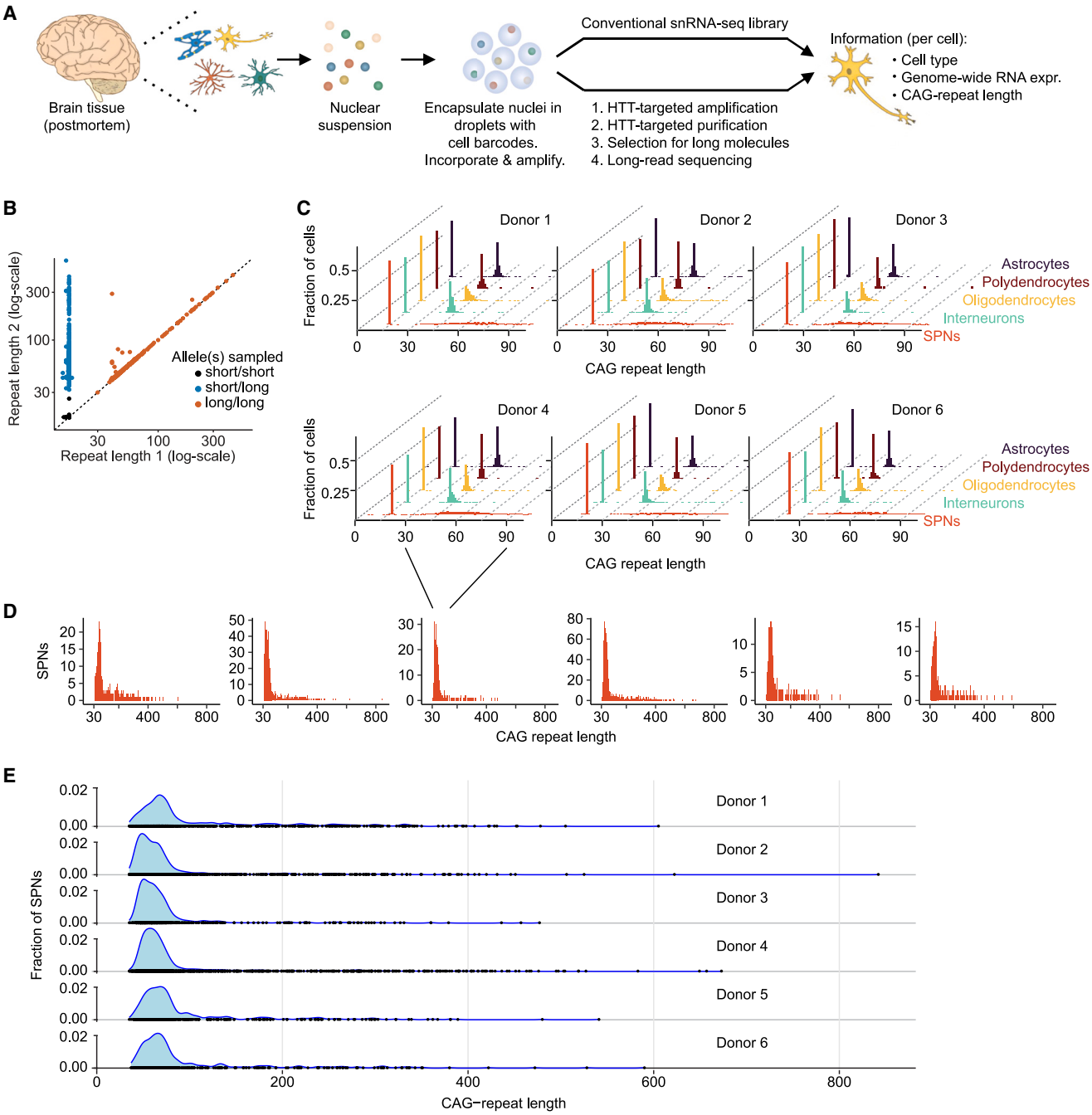


Figure 2. Single-cell analysis of *HTT* CAG-repeat length and genome-wide RNA expression in the same nuclei

(A) Molecular approach. Two sequencing libraries are prepared from the same set of barcoded nuclear cDNAs. The first is a conventional snRNA-seq library. The second library samples the CAG-repeat sequence in *HTT* gene transcripts and is analyzed by long-read sequencing. The presence of shared cell barcodes in the two libraries allows each CAG-repeat sequence to be matched to the RNA-expression profile of the nucleus from which it was sampled.

(B) Concordance between pairs of measurements of CAG-repeat length from different *HTT* RNA transcripts (with different UMIs) in the same nucleus (same cell barcode). For each such transcript pair, the longer of the two CAG-repeat measurements is shown on the y axis. Nuclei in which both measurements are from the long (HD-causing) allele (orange) make it possible to measure precision and error rate.

(C) Distributions of CAG-repeat length measurements by donor and cell type.

(D) Distributions of CAG-repeat length measurements in SPNs (for each donor in C), showing only the long (HD-causing) allele and the much wider range of CAG-repeat lengths that HD-causing alleles attain in SPNs. Note that the mode/peak in (D) corresponds to the distribution for the long (HD-causing) allele in (C).

(legend continued on next page)

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

score greater than 600 (corresponding to HD stages with greatly advanced caudate atrophy) appeared to have lost 80%–99% of their SPNs. These results are consistent with findings that caudate atrophy commences subtly about 10–25 years before the onset of motor symptoms then escalates later.[4,5]

Two SPN subtypes defined by their connectivity and gene expression are direct-pathway SPNs (dSPNs, D1 SPNs) and indirect-pathway SPNs (iSPNs, D2 SPNs). iSPNs comprised 47% (±6%) of the SPN population in controls but a smaller fraction in persons with HD ($p = 8 \times 10^{-6}$, Wilcoxon test, Figure 1C), indicating that iSPNs become vulnerable earlier on average than dSPNs. Since iSPNs inhibit motor programs while dSPNs initiate such programs, the earlier average loss of iSPNs (which is consistent with stereological measurements[33]) may underlie the prominence of chorea (involuntary movements) as an early motor symptom in HD.[33] (Relative losses of patch [striosomal] and matrix [extra-striosomal] SPNs were somewhat more variable across individuals [Figure S2B].)

A long-standing hypotheses for HD pathology involves continuous lifelong damage from a toxic mutant HTT protein or the slow development of toxic protein aggregates; we thus sought to better understand whether *HTT* expression levels[34,35] could help explain the profound vulnerability of SPNs or the more modest relative vulnerabilities of iSPNs (relative to dSPNs). Expression levels (bi-allelic) of *HTT*, as a fraction of all mRNA transcripts, were slightly lower in SPNs than in interneurons and only modestly higher in SPNs than in most glia (Figure 1D). *HTT* expression levels in dSPNs and iSPNs were indistinguishable ($p = 0.56$, paired t test, Figure 1E). Individuals varied in *HTT* expression levels, but accelerated SPN loss (relative to CAP score) did not associate with having higher *HTT* expression levels (Figures S2C and S2D).

In every caudate cell type, thousands of genes were differentially expressed (on average) between persons with HD and unaffected individuals (Methods S1 section "Case-control analyses"). This broadly altered gene expression potentially reflected the profound consequences of HD, which cause atrophy and de-vascularization of the caudate and thus a greatly changed context for the remaining cells. Indeed, almost all such gene expression changes also associated (in an HD-cases-only analysis) with the extent of a donor's earlier SPN loss (Methods S1 section "Case-control analyses").

Measuring somatic CAG-repeat expansion alongside RNA expression

We turned to investigating whether there were cell-autonomous gene expression changes associated with a cell's own CAG-repeat length. We developed a molecular approach for ascertaining the CAG repeat of *HTT* RNA transcripts (together with cell and molecular barcodes) alongside analysis of genome-wide RNA expression in the same cell nuclei (Figure 2A). In our approach, each CAG-repeat sequence is matched (using the cell barcodes) to the gene expression profile of the cell from which it is derived and thus to the identity and biological state of that cell (Figure 2A). We deeply sampled nuclei from the caudate of six persons with clinically manifest HD (Table S1). We were able to acquire measurements for approximately 10% of SPNs and interneurons (which have the largest nuclei and snRNA libraries) and smaller fractions (2%–5%) of non-neuronal cell types.

We also developed companion analytical approaches. Despite the well-known distorting effects of PCR upon DNA repeats and molecular-size distributions, sets of sequence reads with the same cell barcode and molecular barcode (i.e., from the same *HTT* RNA transcript) exhibited similar repeat lengths (Figure S3A), the consensus of which we used in downstream analyses. When CAG-repeat length could be measured on multiple *HTT* transcripts (with distinct molecular barcodes) in the same nucleus, these measurements also agreed (Figure 2B).

Long somatic CAG-repeat expansions in SPNs

The *HTT* CAG repeat exhibited different lengths in different cells and cell types. Astrocytes, oligodendrocytes, polydendrocytes (OPCs), microglia, and interneurons exhibited modest CAG-repeat instability, with almost all cells having a CAG repeat within a few units of the modal length (Figures 2C and S3). However, SPNs exhibited profound somatic expansion of the HD-causing allele (Figures 2C and 2D). This pattern was present in all (6/6) of the persons with HD whose caudate we deeply sampled by this approach (Figures 2C, 2D, and S3B). The distinction between SPNs and striatal interneurons was particularly notable, since all are inhibitory (GABAergic) neurons that share a developmental lineage. (Among interneurons, cholinergic interneurons exhibited more expansion than other interneurons, although much less expansion than SPNs [STAR Methods; Methods S1 section "somatic expansion in caudate cell types"].)

Somatic expansion was allele specific, strongly affecting the HD-causing allele but not the other inherited allele (Figures 2C and S3B), suggesting that somatic instability is affected in *cis* by an allele's own CAG-repeat length. In an unaffected (control) donor, neither allele exhibited substantial somatic instability (Figure S3C).

The six individuals' distributions of SPN CAG-repeat lengths had a characteristic shape that visually resembled the profile of an armadillo (Figure 2E). The bulk of the distribution (the "body") reflected substantial expansion in almost all SPNs: 95%–98% of each donor's SPNs had expanded beyond the inherited (germline) length, reaching a median CAG-repeat length of 60–73 CAGs (20–31 CAGs longer than the same donors' germline *HTT* alleles of 40–43 CAGs).

The second feature (the "tail") involved a prominent minority of SPNs with far longer expansions (100–500+ CAGs) (Figure 2E). In SPNs from each donor, this long right tail commenced at about 100 CAGs and tapered only slowly across a wide range (100 to 500+ CAGs), suggesting a second, much faster phase of somatic

(E) The full-length distributions for the HD-causing CAG repeat in SPNs for each donor in (C). Blue shaded areas are smoothed density estimates of the repeat length distribution. Overplotted black points show the measurements in individual SPNs. In all six donors, the repeat length distribution exhibits an armadillo-like shape, in which the DNA repeat in most of a donor's SPNs has undergone modest expansion (up to 100 CAGs) but in a small fraction of SPNs has undergone greater expansion (up to 500+ CAGs). Figure S3 and Methods S1, section "repeat expansion dynamics," contain additional visualizations.
See also Figure S3.

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

CellPress
OPEN ACCESS

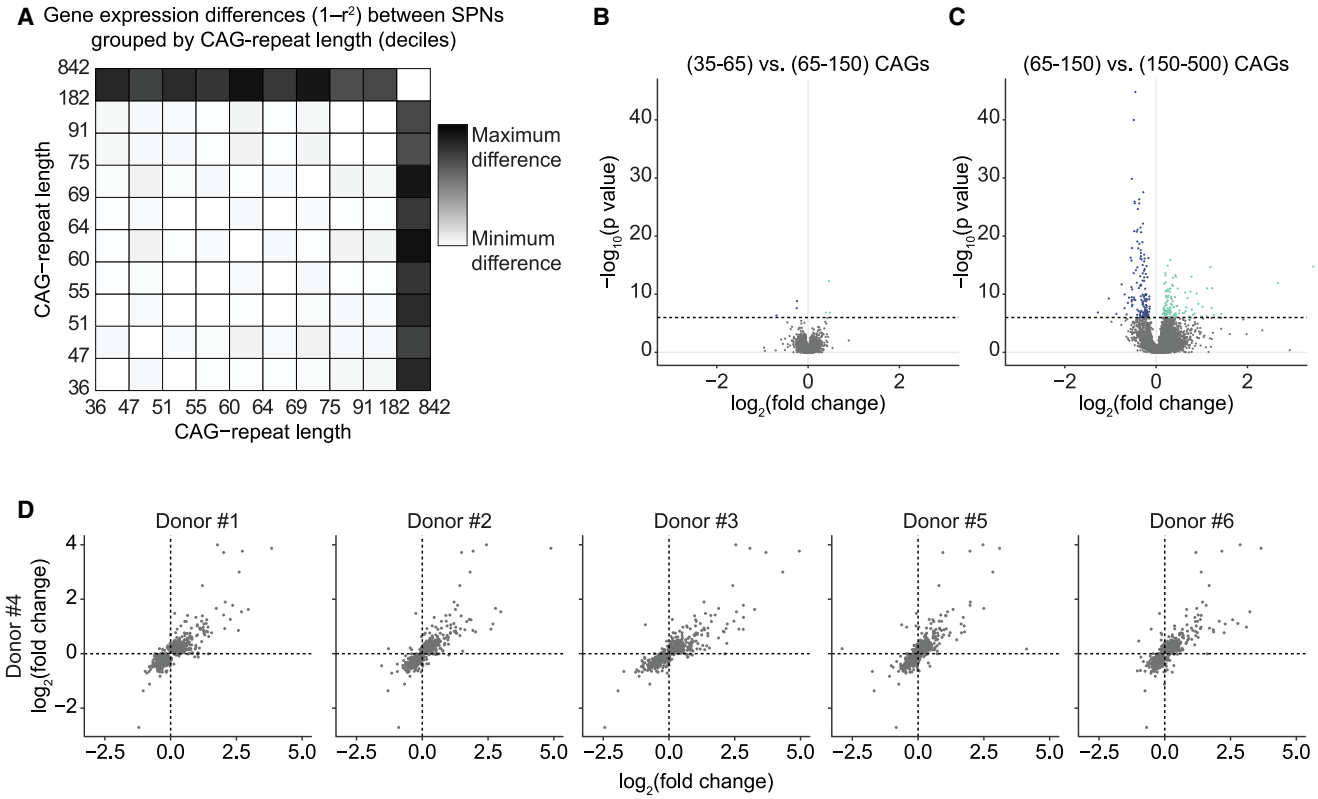


Figure 3. An apparently high length threshold for effects of the *HTT* CAG repeat on SPN biology

(A) Gene-expression comparisons of sets of SPNs (from the same tissue sample) grouped into deciles based on the CAG-repeat length of their HD-causing *HTT* allele. Gray scale: magnitude of gene expression difference (one minus the correlation coefficient); black indicates maximal difference observed in any comparison; white indicates no difference. Figure S4A shows similar analyses of SPNs from six persons with HD.

(B and C) Comparisons (volcano plots) of gene expression between sets of SPNs, sampled from the same person with HD but with CAG-repeat lengths in different ranges. *p* values (y axis) are derived from a Wilcoxon test across the individual SPNs in each group. Fold changes (x axis) are the $\log_2$ ratio of the group averages (positive fold change indicates higher expression in the SPNs with longer repeats). Analyses for all six donors are in Figures S4B–S4E.

(D) Consistency of long(150)-repeat-expansion-associated SPN gene expression changes among individual persons with HD. Each panel is a pairwise comparison of SPN gene expression data involving two persons with HD (x and y axes), in which the values on the two axes are the $\log_2$ fold changes in gene expression when comparing an individual's SPNs with >150 CAGs to the same individual's SPNs with <150 CAGs. Genes whose expression levels change significantly with repeat expansion in at least one of the donors are shown. More analyses are in Figure S4F; see also Methods S1, section "recognizing effects of CAG-repeat length."

See also Figure S4.

expansion that commenced at about 100 CAGs (Figure 2E). We call these phase A and phase B, and further discuss them in a later section.

Our detection of many SPNs with long CAG repeats (100–842 CAGs) (Figures 2D and 2E) contrasted with most earlier human HD studies, many of which detect repeat expansion only in the 35–100 range.[11,15] Critical in recognizing the abundance of these long CAG-repeat expansions was the incorporation of unique molecular identifiers (UMIs) into first-strand cDNAs to address the tendency of subsequent PCR to amplify shorter DNA sequences exponentially more efficiently than longer ones. Notably, a 2003 study, which had utilized small-pool PCR to address this same distorting effect, had also observed many molecules with long (100–1,000 CAG) CAG-repeat tracts in HD brain tissue.[13] The innovation and observation of Kennedy et al.[13] were perhaps insufficiently appreciated at the time.

Somatic repeat expansion to 150 CAGs without consequence

To recognize how a cell is affected by the length of its own *HTT* CAG repeat, we identified allelic series of SPNs naturally arising from the mosaicism within each person with HD. These six allelic series consisted of 467–2,337 SPNs per person, with the CAG-repeat lengths collectively spanning 35–842 CAGs. By comparing SPN gene expression profiles within-person rather than across people, we controlled for the profound non-cell-autonomous effects of each donor's disease state (STAR Methods).

Surprisingly, our analyses detected no apparent cell-autonomous effects of CAG-repeat expansion from 36 to 150 CAGs; however, SPNs with longer expansions (>150 CAGs) had altered expression of hundreds of genes (Figure 3; STAR Methods). This conclusion was supported by many kinds of analyses: simple correlations of gene expression profiles (Figures 3A and S4A);

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

non-parametric statistical tests (as shown in volcano plots and genome-wide distributions of gene-level test statistics; Figures 3B, 3C, and S4B–S4E); and regression of gene expression measurements against CAG-repeat length (Methods S1 section "recognizing effects of CAG-repeat length"). None of these analyses detected any cell-autonomous consequences of repeat expansion to 150 CAGs, but all detected profound effects of CAG-repeat expansion beyond 150 CAGs (Figures S4A–S4E).

Continuously escalating changes (phase C) beyond 150 CAGs

We found that not only the high CAG-repeat-length threshold (~150) but also the ensuing gene expression changes in SPNs with expansions beyond 150 CAGs were highly similar from person to person (Figures 3D and S4F).

Our analysis identified more than 700 genes whose expression levels were affected by CAG-repeat length (Table S2; STAR Methods; Methods S1 section "recognizing effects of CAG-repeat length"). These genes exhibited two kinds of relationships to CAG-repeat length: one set of genes exhibited continuous changes in expression levels as the CAG repeat further expanded beyond 150 CAGs; another set of genes exhibited more discrete and dramatic changes in a specific subset of SPNs with even longer CAG-repeat expansions (generally >250 CAGs).

More than 500 genes exhibited incipient and escalating gene expression changes to the extent the CAG repeat had expanded beyond 150 units (Figures 4A, S5, and S6A). We refer to this as phase C (continuous change) and to the affected genes as C− (downregulated) and C+ (upregulated) genes.

Repeat-length-associated expression changes were almost undetectable at 150–180 repeats, but analyses that drew upon all the phase C genes together indicated that these changes had commenced by about 150 repeats (Figure 4B; STAR Methods). This pattern was shared across persons with HD (Figure 4B) and was present in both direct and indirect SPNs as well as in patch (striosomal) and matrix (extra-striosomal) SPNs (Figure S5).

The genes whose expression declined in those SPNs with a repeat longer than 150 CAGs (the C- genes) were among the most strongly expressed genes in SPNs; almost all were expressed more strongly by healthy SPNs than by other types of inhibitory neurons (Figure 4C). These included *PDE10A*, *PPP2R2B*, *PPP3CA*, *PHACTR1*, and *RYR3* and more than 100 other genes that normal SPNs express more strongly than striatal interneurons do (Figure 4C). This pervasive relationship suggests that a core biological change in phase C involves the erosion of SPN identity features that distinguish SPNs from other kinds of inhibitory neurons. Many of these genes also have important physiological functions; for example, genes encoding the potassium channel subunits KCNA4, KCNAB1, KCND2, KCNH1, KCNK1, KCNQ5, and KCTD1 all declined in expression during phase C, a change that might affect SPN physiology.

Expression levels of *HTT* did not detectably change with CAG-repeat expansion (Figure S6B). (Note that these transcripts might in principle exhibit altered posttranscriptional processing,[36] but 3′ snRNA-seq data are not informative about this.)

Although the relationship of phase C changes to a cell's own CAG-repeat length was strong and clear (Figure 4), such changes appear to have been hard to recognize in bulk-tissue and sorted-SPN studies because they arise asynchronously in individual SPNs and thus are present in only a small fraction of SPNs at any given moment in time. Earlier studies have focused primarily on changes that our own analysis suggested were the result of earlier SPN loss, as they were experienced equally by all remaining SPNs (regardless of CAG-repeat length) and to an extent predicted by a donor's earlier SPN loss (Methods S1 section "Case-control analyses").

We found stronger alignment between our findings and analyses of a specific HD mouse model (Q175), which begins life with a CAG-repeat tract of >170 CAGs in all cells. In such mice, SPNs, interneurons, and glia all exhibited reduced expression of genes that distinguished them from one another.[37] The presence of this long repeat in all cell types may also explain the diverse cell-type-specific pathologies, including vascular and oligodendrocyte pathologies,[9,38] exhibited by Q175 mice.

De-repression crisis (phase D)

A distinct set of more than 100 genes that are normally repressed in SPNs also exhibited repeat-length-dependent change but with a different relationship to CAG-repeat length (Table S2). These genes remained repressed even in most SPNs with long expansions (>150 CAGs) but tended to become de-repressed in those SPNs in which the phase C changes had progressed to the greatest degree (Figures 5A and 5B). In the cells in which this de-repression had occurred, it tended to involve very many genes concurrently. We refer to this state as "de-repression crisis" (phase D).

Phase D was associated with still longer CAG repeats (Figure 5A). De-repression was rare (<4%) even among SPNs with 150–250 CAGs, but it became very common (>50%) in SPNs with 350 or more CAGs (Figure 5C). Notably, in nuclei in which phase D changes were detected, the number of de-repressed genes exhibited little relationship to CAG-repeat length (Figure S7A), a pattern distinct from the phase C changes, which were well predicted by an SPN's CAG-repeat length at the time of analysis (Figure 4B). We interpret this to mean that while phase C changes proceed on a timescale similar to that of fast CAG-repeat expansion, phase D changes progress with far faster kinetics once initiated.

The 100+ genes we found to be de-repressed in phase D had distinct biological features in common. They included most of the genes at the HOXA, HOXB, HOXC, and HOXD loci, as well as noncoding RNAs (*HOTAIR*, *HOTTIP*, *HOTAIRM1*) at these same loci (Figure 5D). These genes are involved in cell specification in the brain and other organs[39] and are normally expressed during early embryonic development but not in adult neurons. The de-repressed genes at other loci included dozens of transcription factor genes (including *FOXD1*, *LHX9*, *ONECUT1*, *POU4F2*, *SHOX2*, *SIX1*, *TBX5*, *TLX2*, *ZIC4*) that are normally expressed in other neural cell types but not in SPNs.

The de-repression of so many transcription factor genes could in principle lead to the expression of genes normally expressed in other neural cell types. Indeed, phase D SPNs expressed many genes that are normally expressed in

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

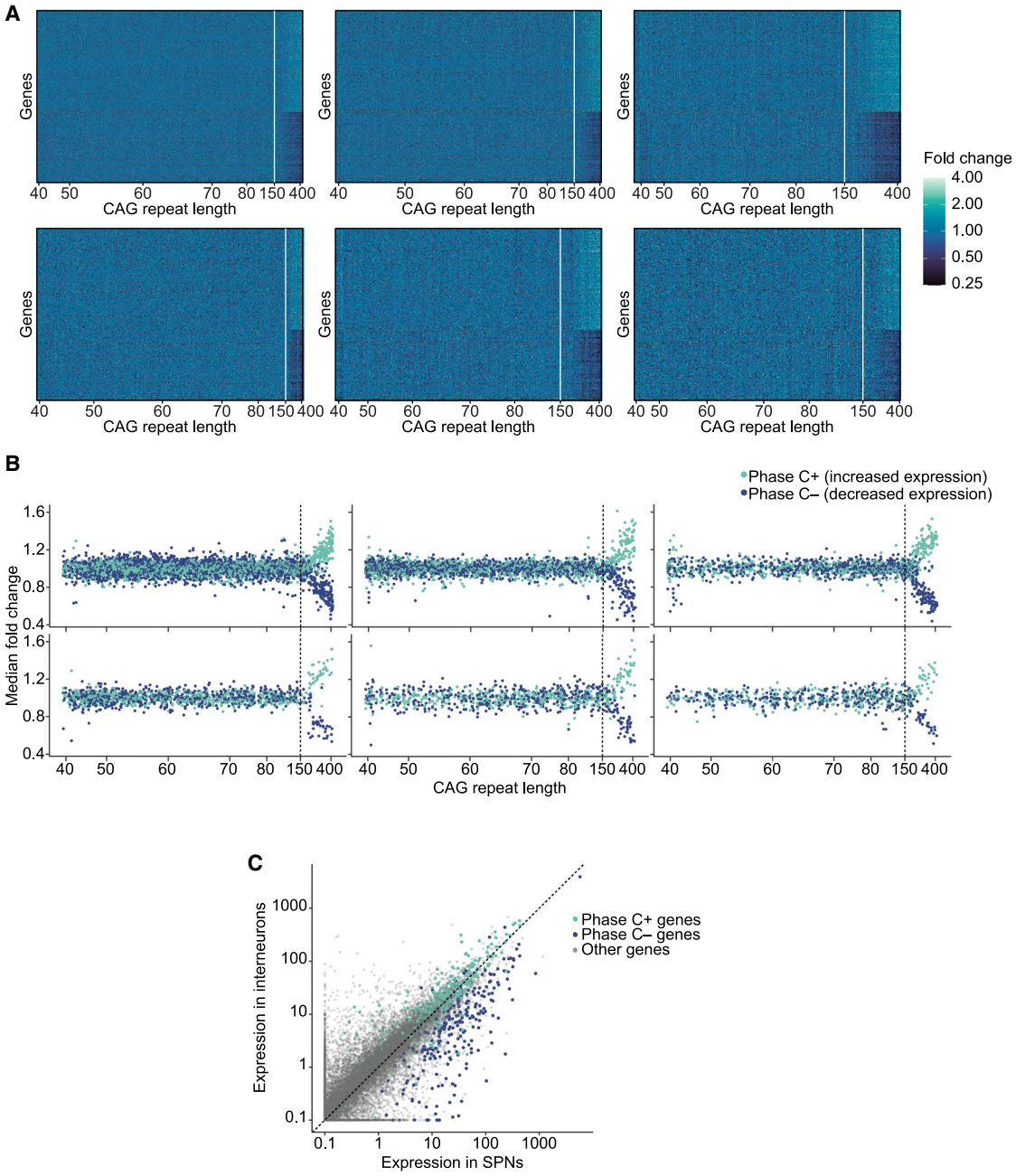


Figure 4. Gene-expression changes in SPNs with somatic CAG-repeat expansion beyond 150 CAGs

(A) On each plot, one donor's individual SPNs are ordered from left to right by the length of their *HTT* CAG repeat (the columns of the heatmap). Each row shows expression data for a specific gene in these SPNs. (The genes shown are genes found to change in expression with repeat expansion.) Shades of each facet show the level of expression of that gene in that SPN, relative to the average SPN with repeat length <150 in that donor. Example trajectories for individual genes are in Figure S6A.

(B) As in (A), on each plot, individual SPNs are ordered from left to right by their CAG-repeat length. Each SPN is represented by both a blue and a green point. Blue points show the median fold change of a set of 192 genes, which decreases in expression with CAG-repeat expansion (C– genes). Green points show the median fold change of a set of 274 genes, which increases in expression with CAG-repeat expansion (C+ genes). See also Table S2.

(C) Genes that decline in expression during phase C (C– genes) are genes that are more strongly expressed in healthy SPNs than in striatal interneurons. Gray points: all genes (cell-type-specific expression levels in unaffected individuals). Colored circles: genes whose expression levels decline (blue) or increase (green) in SPNs with *HTT* CAG-repeat expansion beyond 150 units (in phase C).

See also Figures S5 and S6.

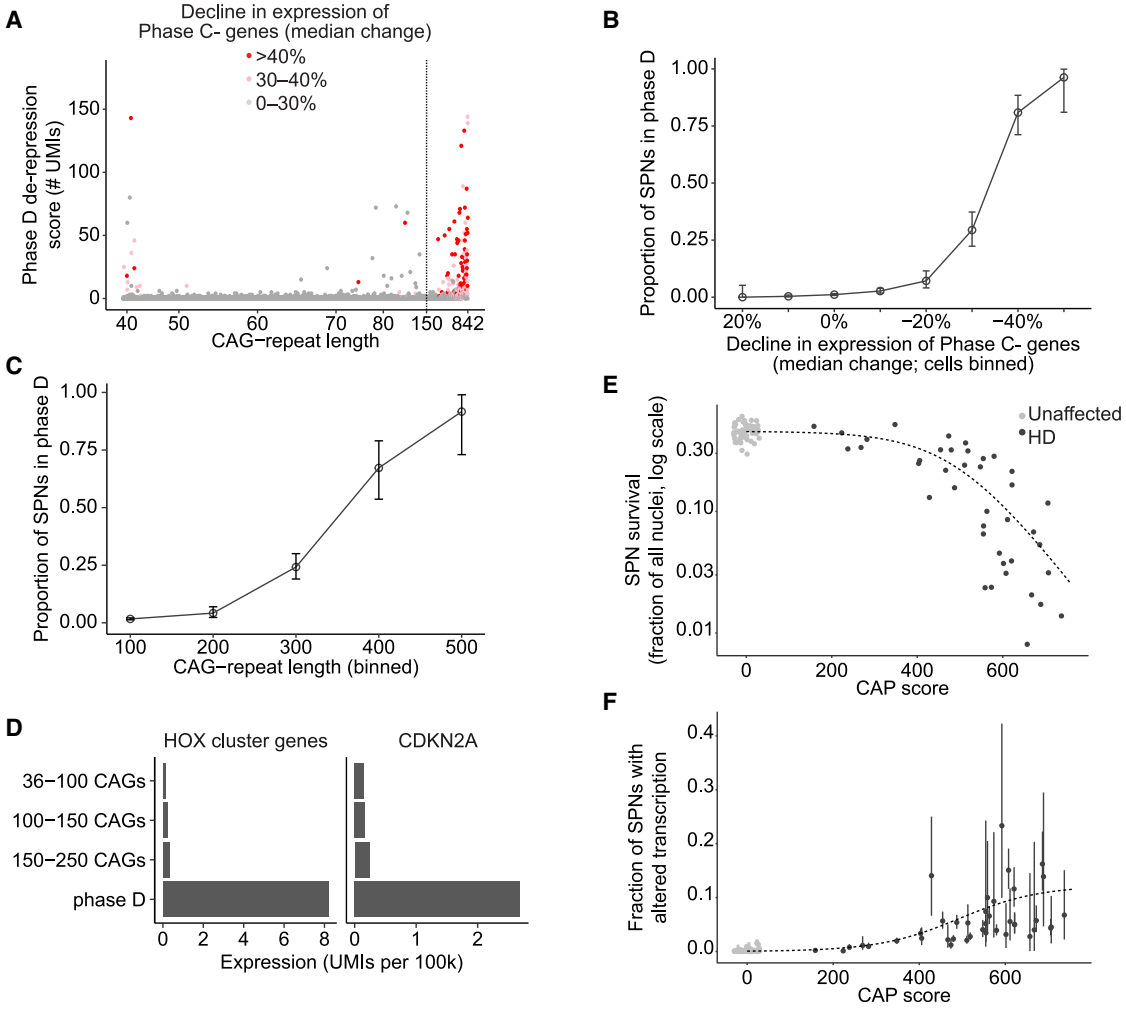


Figure 5. De-repression crisis (phase D) and subsequent SPN elimination (phase E)

(A) De-repression of genes that are normally silent in SPNs. Points represent individual SPNs; red and pink points are those SPNs whose phase C expression changes (Figure 4B) have progressed beyond the threshold values shown in the legend. y axis: de-repression score, the number of transcripts (UMIs) detected from 107 "phase D" genes that are normally silent in SPNs. Additional visualizations of this relationship are in Figures S7C and S7D.

(B) Fraction of SPNs exhibiting this de-repression phenotype, in relationship to the increasing dysregulation (reduced expression) of the phase C− genes. Error bars represent 95% binomial confidence intervals.

(C) Fraction of SPNs exhibiting this de-repression phenotype, in relation to CAG-repeat length. Error bars represent 95% binomial confidence intervals.

(D) Expression of HOX cluster genes (left) and *CDKN2A* (right) in SPNs of persons with HD.

(E and F) SPN loss and transcriptionopathy as HD progresses.

(E) Relationship of SPN survival (shown on a logarithmic scale) to HD progression as indexed by CAP score. The dashed curve shows a logistic function fit to the SPN survival data; its slope (derivative) estimates average rates of SPN loss as HD progresses. Error bars represent 95% binomial confidence intervals.

(F) Relationship of the frequency of SPN phase C transcriptionopathy (fraction of SPNs, y axis) to HD progression as indexed by CAP score (x axis) for the same donors in (E). The dashed curve is the negative of the derivative of the SPN survival curve from (E) (i.e., is proportional to the estimated rate of SPN loss). In (E) and (F), donors were excluded if they had fewer than 25 surviving SPNs (n = 8) or were beyond the plot range (CAP score >800, n = 2).

See also Figure S7.

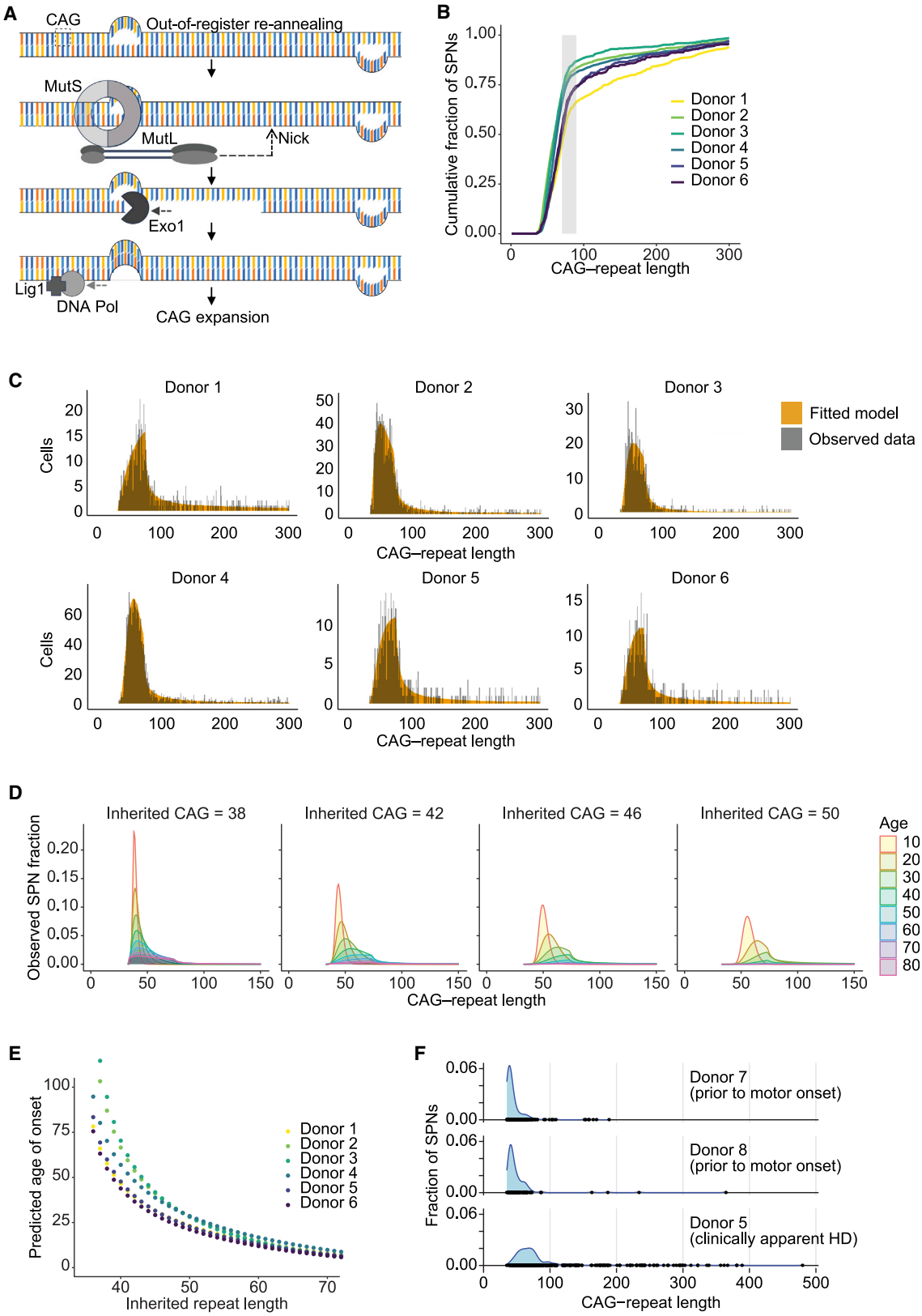
interneurons (*CALB2*, *KCNC2*), in glutamatergic (excitatory) neurons (*SLC17A6*, *SLC17A7*, *SLC6A5*), in astrocytes (*SLC1A2*), in OPCs (*VCAN*), or in oligodendrocytes (*MBP*). These changes suggested that SPNs in phase D were losing negative as well as positive features of SPN cell identity.

Two of the most strongly induced genes in phase D were *CDKN2A* (Figure 5D) and *CDKN2B*, which encode proteins (p16(INK4a) and p15(INK4b)) that promote senescence and apoptosis.[40–43] Ectopic expression of *Cdkn2a* is toxic to neurons.[44] De-repression of *CDKN2A* and *CDKN2B* in phase D SPNs may be an imminent cause of their death.

Interestingly, inactivation of the polycomb repressor complex 2 (PRC2) in adult mice causes a similar set of gene expression changes, including de-repression of Hox genes, other transcription factors, and *Cdkn2a* and *Cdkn2b*, leading within months to SPN loss, motor function decline, and lethality.[45]

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

CellPress
OPEN ACCESS



(legend on next page)

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

Elimination phase (phase E)

The above results suggested that the transcriptional changes in long-repeat (phases C and D) SPNs might lead to their death. Although we cannot observe the same SPNs at multiple points in time, we sought to learn from comparisons across donors who passed away at different stages of caudate atrophy and SPN loss. To do this, we used CAP score as a measure of HD progression (as in Figure 1) in order to bring persons with a variety of ages and inherited CAG-repeat lengths into a single analysis (Figures 5E and 5F).

Across HD progression, the rate of SPN loss can be estimated from the slope of the decline in SPN abundance (on a logarithmic scale) (Figure 5E). This inferred rate of SPN loss increased in tandem with the fraction of donors' SPNs whose RNA expression indicated the presence of phase C transcriptional changes (Figures 5F and S7B). In addition, donors in whom larger fractions of SPNs exhibited this transcriptionopathy tended to be donors with precocious SPN loss (Figures S7C and S7D).

Insights from computational modeling of DNA-repeat expansion dynamics

Our experimental results were hard to reconcile with conventional biological models of HD in which most or all SPNs endure a toxic mutant HTT *simultaneously*. Could a model of *sequential* SPN toxicity—in which, at any one time, most SPNs have a biologically innocuous *HTT* whose repeat length is far below a high toxicity threshold—plausibly explain the relentless loss of SPNs in HD (Figure 1)? Could the decades-long latent period before symptom onset be reconciled with the subsequent, fast loss of SPNs (Figure 1)?

To address these and other questions, and to better appreciate the dynamic processes that might give rise to clinical observations and end-of-life biological measurements, we computationally modeled repeat expansion dynamics over the human lifespan, seeking to understand whether simple models based on an emerging understanding of DNA-repeat expansion mechanisms[46,47] (Figure 6A) would generate repeat length distributions and cell loss trajectories consistent with our experimental results.

In post-mitotic cells such as neurons, DNA-repeat length-change mutations are thought to result from occasional strand misalignment (mispaired repeats) after transcription or transient helix destabilization.[48] Mispaired repeats create extrahelical extrusions ("slip-out" structures) (Figure 6A). Small extrahelical extrusions are recognized by DNA mismatch repair (MMR) complexes, which initiate repair pathways[46] that involve nicking, excision, and resynthesis of one of the two strands. If the two slip-out structures are farther apart than this excision distance, then resynthesis results in a length-change mutation (Figure 6A)—an expansion or contraction, depending on which strand has been nicked and excised. (Some MMR complexes have a strand bias that leads to expansions more frequently than contractions[47].) Experimental observations indicate that repeat expansion tends to occur in small increments.[21,28]

Our simulations adhered as closely as possible to this emerging understanding. We assumed that all SPNs initially had the same (germline) *HTT* allele, that length-change mutations were stochastic expansions or contractions of a small number of CAG units, that the likelihood of mutation increased with repeat length, and that SPN loss occurred among SPNs with >150 repeats. We found mutation-rate and expansion-contraction-bias parameters that optimized the likelihood of the observed data from each person with HD, including the distribution of SPN CAG-repeat lengths and SPN loss at the age of death and brain donation. These simulations are described in detail in Methods S1, section "repeat expansion dynamics"; an animated rendering is in Video S1.

The most challenging aspect of the repeat-length data to explain was its armadillo shape (Figure 2E)—the simultaneous presence of a large majority of SPNs with 40–100 CAGs and a small minority of SPNs with far more (100–800+) CAGs. All the donors we analyzed exhibited this transition across about 70–90 CAGs (Figure 6B). Models in which the increase in the mutation rate was a linear, quadratic, higher-order polynomial, or log-normal function of repeat length did not generate this shape. However, models with two phases of expansion—a slow phase (phase A) that transitioned into a much faster phase (phase B)—generated data that closely matched the experimental data (Figure 6C; Video S1). Our models estimated this transition as occurring over a similar repeat-length interval (70–90 CAGs) in each donor, with the mutation rate increasing at least 6-fold over this range (beyond its general pattern of continuous increase with repeat length). We note that at this length scale (70+ CAGs, 210+ bp), otherwise-mobile

Figure 6. Computational modeling of somatic CAG-repeat expansion dynamics

(A) This schematic illustrates mechanisms (established in earlier work[46]) for non-replicative DNA-repeat expansion in post-mitotic cells. Extrahelical DNA extrusions ("slip-out" structures) form from mispairing within the CAG-repeat tract after strand separation (e.g., due to transcription). The MutSβ complex can bind to these transient structures, initiating DNA excision and resynthesis, which (when initiated on the strand opposite the slip-out) can result in the incorporation of an extra repeat unit. Occurring many times across a human lifetime, this mutational process results in a progressive expansion of the DNA repeat.

(B) Cumulative distributions of CAG-repeat length measurements in SPNs from six deeply sampled donors. The gray shaded region highlights the range (70–90 CAGs) over which somatic expansion appears to greatly accelerate.

(C) Distributions of CAG-repeat length measurements in SPNs from these same donors (black) overlaid with the results of stochastic models (orange) for which a few key parameters (such as mutation rate) have been fitted to each donor's repeat-length and SPN-loss data.

(D) Effect of changing a single variable (inherited/germline CAG-repeat length) in the model for a typical donor, keeping the other fitted parameters fixed. Each curve indicates the predicted CAG-repeat length distribution for surviving SPNs at each decade (ages 10–80).

(E) Model-estimated relationship between inherited germline CAG-repeat length and age at clinical motor onset. As a proxy for age of onset, we used the predicted time at which 25% of a donor's SPNs have been lost. We estimated this age-of-onset proxy at different hypothetical inherited repeat lengths. The shapes of the resulting curves approximate the known relationship between inherited repeat length and age of HD onset.

(F) Observed CAG-repeat length distributions for two donors (donors 7 and 8, Table S1) with HD-causing alleles, who passed away prior to onset of clinical motor symptoms, based on review of their medical records. Data from a typical symptomatic donor (donor 5) are shown at the bottom for comparison (distributions for several other symptomatic donors are in Figure 2E).

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

slip-out structures (Figure 6A) may with increasing likelihood be separated by an intervening nucleosome, greatly reducing the likelihood that they resolve on their own before they are surveilled by MMR complexes.

Fitting the experimental data did not require assuming single-cell heterogeneity in mutation rates: we found that asynchronous SPN toxicity could be explained simply by the asynchronous passage of SPNs from phase A to the subsequent, faster phase B. This asynchronicity arose from the following relationships: (1) length-change mutations were initially rare events (occurring less than once per year per cell across 36–55 CAGs), and (2) such mutations, upon occurring, increased the probability of subsequent mutations.

A fundamental relationship in HD is the association of longer inherited alleles with symptom onset earlier in life—a relationship that is steep in the 36–45 CAG range (across which each extra inherited CAG accelerates symptom onset by more than a year) and has long been thought to reflect increasing HTT toxicity in this range. Our simulations also produced this relationship, but for a different reason: slightly longer inherited alleles bypassed the CAG-repeat lengths at which somatic expansion is most infrequent (occurring less than once per year) (Figures 6D and 6E; Video S2). We note that this was previously predicted on theoretical grounds by Kaplan et al.[49]

Simulation results suggested that the earlier loss of iSPNs relative to dSPNs (Figure 1C)—which had not been explained by *HTT* expression levels (Figure 1E)—may instead be explained by a modestly higher (~15%) rate of somatic expansion in iSPNs (Video S3).

A long-standing mystery about HD involves the long latent period (generally decades) in which persons have no apparent symptoms (HD Integrated Staging System stages 0 and 1, see Tabrizi et al.[5]). Our simulations predicted that persons in this stage might in fact have substantial somatic expansion but with almost all SPNs still in phase A. To test this, we analyzed caudate tissue from two persons with HD who had passed away and contributed their brains for research prior to clinical motor diagnosis and/or without apparent neuropathology upon autopsy. Distributions of CAG-repeat lengths in these donors' SPNs indeed exhibited substantial somatic expansion but few cells with long (>100) expansions (Figure 6F).

We also found that explaining a long-puzzling feature of HD—the transition from slow to rapid atrophy of the caudate—did not require common assumptions of a non-cell-autonomous disease-escalating process (such as inflammation or spreading prions). Rather, the period of more rapid decline corresponded to the period in which the bulk of a person's SPNs reached the end of phase A and more quickly traversed the subsequent pathological phases (whereas only a small number of precociously expanding SPNs did this during the earlier, latent stage).

Our simulations suggest that the average SPN in a person with the most common HD-causing inherited allele (42 repeats) spends 96.4% (SD 2.0%) of its life with 42–150 CAG repeats, i.e., with what our experimental results suggest is an innocuous *HTT* gene.

A pathogenesis model: ELongATE

Our results suggest that an SPN's own CAG repeat becomes toxic only when quite long (>150 CAGs) and that this long repeat is necessary and sufficient for pathology. We propose a model of HD pathogenesis involving a series of phases driven cell autonomously by a neuron's own expanding *HTT* allele (Figure 7). We call this dynamic ELongATE (extra-long repeats acquire toxic effect).

In the first phase (phase A, when an SPN has 36–80 CAGs), an SPN undergoes decades of slow but accelerating repeat expansion. We estimate that an SPN takes 50 years (on average) to expand from 40 to 60 CAGs and then another 12 years to expand from 60 to 80, but with variability both cell to cell and person to person (Methods S1, section "repeat expansion dynamics"). Phase A could be compared with a slowly and capriciously "ticking DNA clock."

As an SPN enters the second phase (phase B, 80–150 CAGs), the rate of expansion greatly accelerates, and the tract may now expand to 150 CAGs in just a few years. Still, as in phase A, the SPN's *HTT* CAG repeat does not appear to affect its own gene expression. Phase B could be compared with a more rapidly and predictably ticking DNA clock.

As an SPN enters the third phase (phase C, 150+ repeat units), hundreds of genes begin to change in expression levels. These changes are initially tiny, but they escalate alongside further repeat expansion (Figure 4), eroding gene expression features of SPN identity (Figure 4C).

In its fourth phase (phase D), an SPN de-represses scores of genes that are typically expressed in other neural cell types or in embryonic development. Phase D SPNs also de-repress *CDKN2A* and *CDKN2B*, which encode proteins that promote senescence and apoptosis.

In the final phase, an SPN is eliminated (phase E). Lost cells do not appear in CAG length and gene expression data, but the effect of their earlier loss is apparent in effects on gene expression in remaining cells of all types (including SPNs) (Methods S1 "Case-control analyses").

Importantly, individual SPNs enter the fast phases (B, C, D, E) at different times, an asynchrony that our modeling suggests can be explained largely by the variable amounts of time that individual neurons take to traverse phase A. Phase A introduces this asynchrony because each neuron's expansion results from low-frequency stochastic length-change mutations (initially occurring less than once per year), with each expansion event increasing the likelihood of subsequent further expansion.

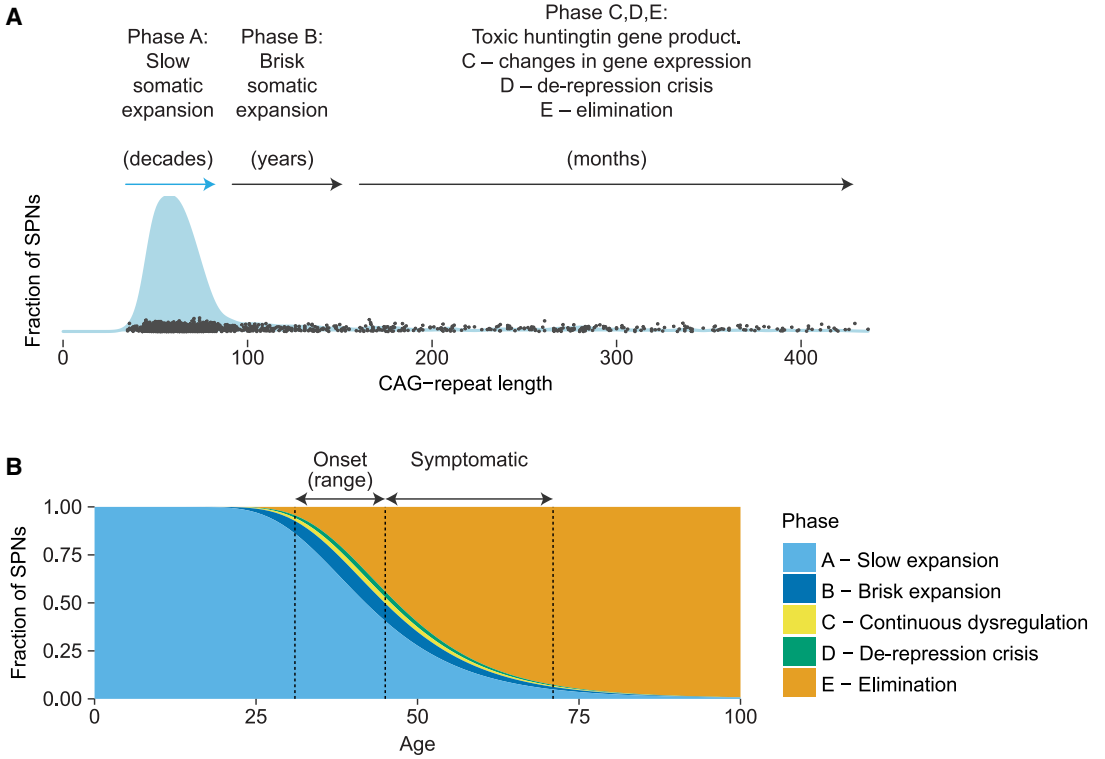
DISCUSSION

Three biological questions

Biological research on HD has long been animated by three puzzles. What is toxic to cells about the inherited alleles that cause HD? Why is this toxicity so cell-type specific? And why are HD symptoms preceded by decades of biological latency? Our experiments and analyses suggest surprising answers to all three puzzles.

The surprising answer to the first puzzle—the biological nature of the toxicity of inherited HD-causing alleles—is that such alleles are in fact innocuous and remain so even after decades of somatic expansion (phase A in Figure 7). Among SPNs sampled from the same donor and tissue, we found no cell-autonomous gene expression consequence of CAG-repeat

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038



Phase		CAG-repeat length	Somatic expansion dynamic	Time in phase (est.)	Gene expression changes associated with neuron's own *HTT* CAG repeat length (relative to nearby neurons of the same type)		
					Genes affected	Dynamic	Biological theme
A	Slow DNA-repeat Expansion	36 to 80	Slow (initially <1/yr) but increasing	Decades	n/a	n/a	n/a
B	Brisk DNA-repeat Expansion	80 to 150	Fast (10% to 50% per year)	Years	n/a	n/a	n/a
C	Continuous gene-expression change	> 150	Fast	Months	Hundreds	Gradually escalating	Cell-identity erosion
D	De-repression crisis	>> 150	Fast	Months	Hundreds	Quickly escalating	De-repression crisis
E	Eliminated				(Atrophy-associated gene-expression changes in surviving cells)		

Figure 7. ELongATE: A model for SPN pathology in HD

(A) Individual neurons pass asynchronously through five key pathological phases, spending >95% of their lives in a long period of DNA-repeat expansion (a ticking DNA clock, phases A and B) with a biologically harmless (but unstable) *HTT* gene. Individual neurons asynchronously exit phase A and proceed through the subsequent, faster phases.

(B) Prediction of the fraction of SPNs in each of the five phases, across the latent, peri-onset and then progressive stages of HD. The estimated trajectories are based on the data from a representative donor. The indicated ranges for clinical motor onset and escalating symptoms are approximate. The illustrated onset range, representing loss of 20% to 50% of a donor's SPNs, was inferred from available age-of-onset data for the brain donors whose tissue we analyzed. The time estimates listed for each phase are for persons who inherit the more common HD-causing alleles (40–45 CAGs).

length across a wide range of repeat lengths (36–150), including those inherited by almost all patients—but we found profound, and likely quite toxic, gene expression changes in SPNs with longer (>150) repeats. A potential interpretation is that the apparent threshold for an inherited *HTT* allele to be disease causing (36–40 CAGs) reflects not that such alleles encode toxic RNAs or proteins but that such alleles are sufficiently unstable as to be likely to expand beyond 150 repeats within a human lifetime. We propose that the key question is not what is toxic about inherited *HTT* alleles but rather what toxicity is acquired with expansion beyond about 150 repeats. Future molecular research on *HTT* RNA and protein might optimally focus on phenomena

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

that change at long (~150) repeat lengths as opposed to the modest repeat lengths (35–100) observed in most cells.

We propose that the answer to the second question—the cell-type specificity of cell death in HD—is that the *HTT* CAG repeat reaches the high toxicity length threshold only in certain cell types (Figure 2C). The key question is perhaps not why SPNs are more vulnerable to a toxic *HTT* but rather why somatic instability varies by cell type.

The surprising answer to the third puzzle—the apparently slow toxicity caused by mutant *HTT*—may be that once a neuron develops a harmful *HTT*, that neuron's decline is not actually slow. Our results suggest that once the toxicity threshold is crossed and cell-autonomous biological changes start, these changes progress to cell death over months rather than decades—one to two orders of magnitude faster than previously thought. Individual neurons thus tend to experience their own *HTT* toxicity asynchronously.

Scores of phenomena have been described in animal and cellular models of HD and proposed to explain or contribute to HD pathogenesis. A long-standing challenge has been to identify which of these changes are disease-driving mechanisms, which are reactive mechanisms, and which arise only in models but are not features of HD in humans. The potential centrality of the ELongATE dynamic (Figure 7) in HD is supported by human genetic findings that HD "modifiers"—common alleles that affect the age at which HD motor symptoms commence—arise from *MSH3*, *FAN1*, *MLH1*, *LIG1*, *PMS1*, and *PMS2*,[17] genes that are functionally united not only by roles in DNA repair but by more specific effects on the stability of DNA repeats.[21–29,47]

Therapeutic implications

The most important implication of our findings may be for developing therapies for HD and perhaps other DNA-repeat disorders.

The focus of almost all therapies in advanced clinical development for HD is on lowering HTT expression; these candidate therapies utilize diverse approaches including antisense oligonucleotides, small interfering RNAs, splicing modulation, and gene editing.[50] Under conventional models for HD pathology, HTT lowering has a compelling rationale: if inherited HD-causing alleles encode a toxic protein (or become toxic after just modest somatic expansion), and if the cell-biological process by which such alleles lead to neuronal death is decades long, then even a partial reduction in HTT production might greatly delay disease. However, HTT-lowering treatments have so far been unsuccessful in HD clinical trials,[51–53] a disappointment which may result from adverse biological effects of therapy[51] and/or lack of efficacy, including the possibility that the toxic *HTT* entity is an alternative *HTT* isoform[36] that is not targeted by these treatments.

Our model for HD pathogenesis (Figure 7) suggests two surprising biological challenges for HTT-lowering therapies. First, at any time, very few SPNs may actually have a toxic HTT protein from whose lowering they might benefit (Figure 7). (At the same time, most neurons may be deriving positive biological function from HTT.[54]) Second, even once an SPN arrives at cell-biological toxicity (phases C and D in Figure 7) and may benefit from HTT lowering, its expected lifetime may be months rather than decades. In short, HTT-directed therapeutic efforts will need to address the possibility that HTT toxicity is brief, asynchronous, and intense rather than long, synchronous, and indolent.

Our conclusion that HD pathogenesis is a DNA process for >95% of a neuron's life (Figure 7) suggests potentially greater focus on trying to slow somatic expansion. Experimental reduction in the function of MMR genes (including *MSH3*, *MSH2*, *MSH6*, and *PMS1*) can stabilize DNA repeats in mice and/or cultured cells[21–25,29,55] and thus might pre-empt the somatic genetic cause of HD pathology. However, much uncertainty has surrounded the therapeutic window that such an approach could have. Our results suggest that the therapeutic window is wide: if a cell spends 95% of its life in phase A, then even modestly slowing somatic expansion might substantially postpone HD symptom onset.

What about persons who already have early HD symptoms? Surprisingly, our results predict that even when a person with HD has lost 25% of their SPNs, more than 90% of still-living SPNs still have a *HTT* gene that is not yet biologically harmful (Figures 6 and 7; Methods S1 section "Repeat expansion dynamics"). Future somatic-expansion-directed therapy thus might be able to slow or stop HD progression even in persons who already have early HD symptoms. This would allow the efficacy of such therapy to be evaluated in patients with HD symptoms, a faster and more straightforward path to clinical evaluation than a long-term prevention trial.

Implications for other DNA-repeat disorders

The dynamic we have described, in which the toxic effect of a DNA repeat is acquired only after decades of somatic expansion, could in principle also apply in other DNA-repeat disorders. More than 60 human diseases are caused by inherited expansions of DNA repeats in protein-coding sequences, introns, untranslated regions, or promoters.[56–59] More than 30 of these diseases involve adult or midlife onset,[59] and several are known to involve age-associated mosaicism.[60–63] Many disorders—including myotonic dystrophy 1, X-linked dystonia Parkinsonism, Friedrich ataxia, and six forms of spino-cerebellar ataxia (SCA1, SCA2, SCA3, SCA6, SCA7, and SCA11)—are also (like HD) delayed or hastened by common genetic variation at genes that regulate somatic DNA-repeat stability.[57,64,65] If these disorders share a dynamic in which toxicity is acquired only after somatic DNA-repeat expansion, then a therapy that slows such expansion might delay or prevent many human DNA-repeat disorders.

Limitations of the study

Future work will be required to determine whether the repeat-expansion-driven dynamic that we have described here also explains HD pathology in other brain areas and in the periphery. Our analyses have also focused on the cell-autonomous effects of the CAG repeat on a cell's own biology, as measured through its gene expression, and thus do not preclude the possibility that the CAG repeat might have a non-cell-autonomous effect on other cells (e.g., via aggregation in axons[66]) without affecting a cell's own gene expression cell autonomously. Finally, we provide only indirect evidence that neuronal degeneration leads to the glial changes in HD (Methods S1 section "Case-control analyses"), although evidence from neuron-specific manipulations in mice[67] offers direct support for this idea.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Steven A. McCarroll (smccarro@broadinstitute.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data availability

Raw and processed data have been deposited at the NIH Neuroscience Multiomic Archive (NeMO, https://www.nemoarchive.org/), as accession nemo:dat-ztfn3cc. The deposition includes both open-access and controlled-access components. Data that are potentially individually identifying (such as sequencing reads that contain allelic information) are in the controlled-access components; all other data are in the open-access components.

The open-access components consist of:

- Single-cell-level count data on gene expression ("DGE" (gene-by-cell) matrices of UMI counts in h5 format).
- Metacells by cell type for snRNA-seq village experiments ("DGE" (gene-by-cell) matrices of UMI counts in h5 format).
- CAG-repeat length measurements from individual cells (in tab-delimited text format).
- Assignments of cells to individual donors and cell types (in tab-delimited text format).
- Donor meta-data (age, sex, Vonsattel grade, in tab-delimited text format).

The controlled-access components consist of:

- SNP array data (Illumina Global Screening Array) on each donor (in VCF format).
- Raw reads from snRNA-seq experiments (in Illumina FASTQ format).
- Aligned reads from snRNA-seq experiments (in bam format).
- Aligned PacBio reads from HTT-CAG experiments (in bam format).

Code availability

All original code and workflows used for processing single-cell RNA-seq data are available in GitHub at https://github.com/broadinstitute/Drop-seq.

All original code for deriving single-cell-resolution CAG-repeat length measurements from long-read (PacBio) data generated from HTT-CAG libraries as described in this manuscript is available in GitHub at https://github.com/broadinstitute/HTT-CAG-Software.

All original code for modeling of the temporal dynamics of the HTT-CAG repeat lengths is available in GitHub at https://github.com/broadinstitute/HD-CAG-Modeling.

ACKNOWLEDGMENTS

We thank the NIH NeuroBioBank for providing tissue for this project and are grateful to the patients and families whose donations of brain tissue enabled this work. This work was funded by CHDI Foundation, Inc., a nonprofit biomedical research organization exclusively dedicated to developing therapies that will substantially improve the lives of HD-affected individuals; the Ludwig Neurodegenerative Disease Seed Grants Program at Harvard Medical School; the Harvard Medical School Department of Genetics; and the National Human Genome Research Institute of the National Institutes of Health under award number R01HG006855. We are grateful to Tom Vogt, Vahri Beaumont, Jian Chen, and Cristina Sampaio for helpful advice and suggestions throughout this project's conception and execution; to Darren Monckton for helpful advice on amplifying and sequencing DNA repeats; to Sarah Tabrizi, Jim Gusella, and Marcy MacDonald for helpful advice on HD; to John Warner for helpful advice on computational modeling; to Jeff Carroll, Gillian Bates, Phil Reaper, Chris Patil, Jesse Gray, and the individuals above for comments on early manuscript drafts; and to Christina Usher for contributions to the manuscript figures.

AUTHOR CONTRIBUTIONS

Conceptualization, S.A.M., S.B., N.M.R., S.K., and R.E.H.; methodology, N.M.R., S.K., R.E.H., W.-S.L., T.M.M., N.K., S.B., and S.A.M.; software, S.K., R.E.H., and S.T.; formal analysis, S.K., R.E.H., M.G., and S.A.M.; investigation, N.M.R., W.-S.L., T.M.M., K.M., N.K., C.D.M., N.R.M., G.L., R.K., and E.L.; resources, K.M., E.L., N.R.M., and S.B.; data curation, K.M., N.R.M., M.H., and M.G.; writing – original draft, S.A.M., S.B., R.E.H., and S.K.; writing – review and editing, S.A.M., S.B., R.E.H., S.K., W.-S.L., T.M.M., and N.K.; visualization, S.K., R.E.H., and S.A.M.; supervision, S.A.M. and S.B.; project administration, M.H. and K.I.; funding acquisition, S.A.M. and S.B.

DECLARATION OF INTERESTS

Patent applications filed by the Broad Institute of MIT and Harvard related to this work include subsets of the authors as inventors. S.A.M. has received compensation for scientific advice to Roche, Pfizer, Biogen, Vertex, and LoQus23 Therapeutics.

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
  - Donor ascertainment
- METHOD DETAILS
  - Determination of inherited CAG repeat length
  - Calculation of CAG-age-product score
  - Single-nucleus RNA-seq in 20-donor "villages"
  - Isolation of nuclei from brain tissue
  - Preparation of snRNA-seq libraries
  - Single-cell measurement of CAG-repeat length
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Processing snRNA-seq data
  - Assignment of nuclei to individual donors
  - Classifying nuclei by cell type
  - Case-control differences in gene expression
  - Analysis of HTT-CAG library data
  - Somatic expansion in glia and interneurons
  - SPN gene-expression comparisons (cell groups)
  - CAG-repeat effects on gene expression
  - Identification of phase D genes
  - Modeling CAG-repeat expansion dynamics
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.11.038.

Received: June 21, 2024
Revised: September 15, 2024
Accepted: November 29, 2024
Published: January 16, 2025

REFERENCES

1. The 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.
2. Gusella, J.F., Lee, J.-M., and MacDonald, M.E. (2021). Huntington's disease: nearly four decades of human molecular genetics. Hum. Mol. Genet. *30*, R254–R263.
3. Langbehn, D.R., Brinkman, R.R., Falush, D., Paulsen, J.S., and Hayden, M.R.; International; Huntington's; Disease; Collaborative Group (2004). A

new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin. Genet. *65*, 267–277.

4. Scahill, R.I., Zeun, P., Osborne-Crowley, K., Johnson, E.B., Gregory, S., Parker, C., Lowe, J., Nair, A., O'Callaghan, C., Langley, C., et al. (2020). Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington's disease Young Adult Study (HD-YAS): a cross-sectional analysis. Lancet Neurol. *19*, 502–512.
5. Tabrizi, S.J., Schobel, S., Gantman, E.C., Mansbach, A., Borowsky, B., Konstantinova, P., Mestre, T.A., Panagoulias, J., Ross, C.A., Zauderer, M., et al. (2022). A biological classification of Huntington's disease: the Integrated Staging System. Lancet Neurol. *21*, 632–644.
6. Saudou, F., and Humbert, S. (2016). The biology of huntingtin. Neuron *89*, 910–926.
7. Braz, B.Y., Wennagel, D., Ratié, L., de Souza, D.A.R., Deloulme, J.C., Barbier, E.L., Buisson, A., Lanté, F., and Humbert, S. (2022). Treating early postnatal circuit defect delays Huntington's disease onset and pathology in mice. Science *377*, eabq5011.
8. Lee, H., Fenster, R.J., Pineda, S.S., Gibbs, W.S., Mohammadi, S., Davila-Velderrain, J., Garcia, F.J., Therrien, M., Novis, H.S., Gao, F., et al. (2020). Cell type-specific transcriptomics reveals that mutant huntingtin leads to mitochondrial RNA release and neuronal innate immune activation. Neuron *107*, 891–908.e8.
9. Garcia, F.J., Sun, N., Lee, H., Godlewski, B., Mathys, H., Galani, K., Zhou, B., Jiang, X., Ng, A.P., Mantero, J., et al. (2022). Single-cell dissection of the human brain vasculature. Nature *603*, 893–899.
10. Wilton, D.K., Mastro, K., Heller, M.D., Gergits, F.W., Willing, C.R., Fahey, J.B., Frouin, A., Daggett, A., Gu, X., Kim, Y.A., et al. (2023). Microglia and complement mediate early corticostriatal synapse loss and cognitive dysfunction in Huntington's disease. Nat. Med. *29*, 2866–2884.
11. Pressl, C., Mätlik, K., Kus, L., Darnell, P., Luo, J.-D., Paul, M.R., Weiss, A.R., Liguore, W., Carroll, T.S., Davis, D.A., et al. (2024). Selective vulnerability of layer 5a corticostriatal neurons in Huntington's disease. Neuron *112*, 924–941.e10. https://doi.org/10.1016/j.neuron.2023.12.009.
12. Telenius, H., Kremer, B., Goldberg, Y.P., Theilmann, J., Andrew, S.E., Zeisler, J., Adam, S., Greenberg, C., Ives, E.J., and Clarke, L.A. (1994). Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nat. Genet. *6*, 409–414.
13. 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. Hum. Mol. Genet. *12*, 3359–3367.
14. Shelbourne, P.F., Keller-McGandy, C., Bi, W.L., Yoon, S.-R., Dubeau, L., Veitch, N.J., Vonsattel, J.P., Wexler, N.S., Arnheim, N., et al.; US-Venezuela Collaborative Research Group (2007). Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Hum. Mol. Genet. *16*, 1133–1142.
15. 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., et al. (2024). Cell-type-specific CAG repeat expansions and toxicity of mutant huntingtin in human striatum and cerebellum. Nat. Genet. *56*, 383–394.
16. 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. Hum. Mol. Genet. *18*, 3039–3047.
17. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Electronic address: gusella@helix.mgh.harvard.edu; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium (2019). CAG repeat not polyglutamine length determines timing of Huntington's disease onset. Cell *178*, 887–900.e14.
18. 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. J. Huntingtons Dis. *10*, 35–51.
19. 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.
20. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium (2015). Identification of genetic factors that modify clinical onset of Huntington's disease. Cell *162*, 516–526.
21. Dragileva, E., Hendricks, A., Teed, A., Gillis, T., Lopez, E.T., Friedberg, E.C., Kucherlapati, R., Edelmann, W., Lunetta, K.L., MacDonald, M.E., et al. (2009). Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. Neurobiol. Dis. *33*, 37–47.
22. Wheeler, V.C., Lebel, L.-A., Vrbanac, V., Teed, A., te Riele, H., and MacDonald, M.E. (2003). Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. Hum. Mol. Genet. *12*, 273–281.
23. 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.
24. 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 Genet. *9*, e1003930.
25. Tomé, S., Manley, K., Simard, J.P., Clark, G.W., Slean, M.M., Swami, M., Shelbourne, P.F., Tillier, E.R.M., Monckton, D.G., Messer, A., et al. (2013). MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. PLoS Genet. *9*, e1003280.
26. 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. Hum. Mol. Genet. *29*, 3044–3053.
27. Kim, K.-H., Hong, E.P., Shin, J.W., Chao, M.J., Loupe, J., Gillis, T., Mysore, J.S., Holmans, P., Jones, L., Orth, M., et al. (2020). Genetic and functional analyses point to FAN1 as the source of multiple Huntington disease modifier effects. Am. J. Hum. Genet. *107*, 96–110. https://doi.org/10.1016/j.ajhg.2020.05.012.
28. Goold, R., Hamilton, J., Menneteau, T., Flower, M., Bunting, E.L., Aldous, S.G., Porro, A., Vicente, J.R., Allen, N.D., Wilkinson, H., et al. (2021). FAN1 controls mismatch repair complex assembly via MLH1 retention to stabilize CAG repeat expansion in Huntington's disease. Cell Rep. *36*, 109649.
29. 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. Nat. Genet. *23*, 471–473.
30. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.
31. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.
32. Zhang, Y., Long, J.D., Mills, J.A., Warner, J.H., Lu, W., and Paulsen, J.S.; PREDICT-HD Investigators; Coordinators of the Huntington Study Group (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. Am. J. Med. Genet. B *156b*, 751–763.
33. Albin, R.L., Reiner, A., Anderson, K.D., Penney, J.B., and Young, A.B. (1990). Striatal and nigral neuron subpopulations in rigid Huntington's disease: implications for the functional anatomy of chorea and rigidity-akinesia. Ann. Neurol. *27*, 357–365.

34. Schilling, G., Sharp, A.H., Loev, S.J., Wagster, M.V., Li, S.H., Stine, O.C., and Ross, C.A. (1995). Expression of the Huntington's disease (IT15) protein product in HD patients. Hum. Mol. Genet. *4*, 1365–1371.

35. Landwehrmeyer, G.B., McNeil, S.M., Dure, L.S., 4th, Ge, P., Aizawa, H., Huang, Q., Ambrose, C.M., Duyao, M.P., Bird, E.D., and Bonilla, E. (1995). Huntington's disease gene: regional and cellular expression in brain of normal and affected individuals. Ann. Neurol. *37*, 218–230.

36. 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.

37. 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.

38. Ferrari Bardile, C., Garcia-Miralles, M., Caron, N.S., Rayan, N.A., Langley, S.R., Harmston, N., Rondelli, A.M., Teo, R.T.Y., Waltl, S., Anderson, L.M., et al. (2019). Intrinsic mutant HTT-mediated defects in oligodendroglia cause myelination deficits and behavioral abnormalities in Huntington disease. Proc. Natl. Acad. Sci. USA *116*, 9622–9627.

39. Hobert, O. (2021). Homeobox genes and the specification of neuronal identity. Nat. Rev. Neurosci. *22*, 627–636.

40. Gil, J., and Peters, G. (2006). Regulation of the INK4b-ARF-INK4a tumour suppressor locus: all for one or one for all. Nat. Rev. Mol. Cell Biol. *7*, 667–677.

41. Igney, F.H., and Krammer, P.H. (2002). Death and anti-death: tumour resistance to apoptosis. Nat. Rev. Cancer *2*, 277–288.

42. Herranz, N., and Gil, J. (2018). Mechanisms and functions of cellular senescence. J. Clin. Invest. *128*, 1238–1246.

43. Yuile, A., Satgunaseelan, L., Wei, J.Q., Rodriguez, M., Back, M., Pavlakis, N., Hudson, A., Kastelan, M., Wheeler, H.R., and Lee, A. (2023). CDKN2A/B homozygous deletions in astrocytomas: A literature review. Curr. Issues Mol. Biol. *45*, 5276–5292.

44. Finneran, D., Desjarlais, T., Morgan, D., and Gordon, M.N. (2023). Differential effects of neuronal Cdkn2a over-expression in mouse brain. Alzheimers Dement. *19*. https://doi.org/10.1002/alz.078485.

45. 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.

46. Iyer, R.R., and Pluciennik, A. (2021). DNA mismatch repair and its role in Huntington's disease. J. Huntingtons Dis. *10*, 75–94.

47. Phadte, A.S., Bhatia, M., Ebert, H., Abdullah, H., Elrazaq, E.A., Komolov, K.E., and Pluciennik, A. (2023). FAN1 removes triplet repeat extrusions via a PCNA- and RFC-dependent mechanism. Proc. Natl. Acad. Sci. USA *120*, e2302103120.

48. Corless, S., and Gilbert, N. (2016). Effects of DNA supercoiling on chromatin architecture. Biophys. Rev. *8*, 51–64.

49. Kaplan, S., Itzkovitz, S., and Shapiro, E. (2007). A universal mechanism ties genotype to phenotype in trinucleotide diseases. PLoS Comput. Biol. *3*, e235.

50. Tabrizi, S.J., Ghosh, R., and Leavitt, B.R. (2019). Huntingtin lowering strategies for disease modification in Huntington's disease. Neuron *101*, 801–819.

51. McColgan, P., Thobhani, A., Boak, L., Schobel, S.A., Nicotra, A., Palermo, G., Trundell, D., Zhou, J., Schlegel, V., Sanwald Ducray, P., et al. (2023). Tominersen in adults with manifest Huntington's disease. N. Engl. J. Med. *389*, 2203–2205.

52. Kingwell, K. (2021). Double setback for ASO trials in Huntington disease. Nat. Rev. Drug Discov. *20*, 412–413.

53. (2021). Promising drug for Huntington disease fails in major trial. Science. https://www.science.org/content/article/promising–drug–huntington–disease–fails–major–trial.

54. Burrus, C.J., McKinstry, S.U., Kim, N., Ozlu, M.I., Santoki, A.V., Fang, F.Y., Ma, A., Karadeniz, Y.B., Worthington, A.K., Dragatsis, I., et al. (2020). Striatal projection neurons require huntingtin for synaptic connectivity and survival. Cell Rep. *30*, 642–657.e6.

55. Ferguson, R., Goold, R., Coupland, L., Flower, M., and Tabrizi, S.J. (2023). Therapeutic validation of MMR-associated genetic modifiers in a human*ex* vivomodel of Huntington's disease. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.05.570095.

56. Paulson, H. (2018). Repeat expansion diseases. Handb. Clin. Neurol. *147*, 105–123.

57. Rajagopal, S., Donaldson, J., Flower, M., Hensman Moss, D.J., and Tabrizi, S.J. (2023). Genetic modifiers of repeat expansion disorders. Emerg. Top. Life Sci. *7*, 325–337.

58. Depienne, C., and Mandel, J.-L. (2021). 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? Am. J. Hum. Genet. *108*, 764–785.

59. Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M.A., and Friedman, J.M. (2024). Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. Nat. Rev. Genet. *25*, 476–499.

60. Monckton, D.G., Wong, L.J., Ashizawa, T., and Caskey, C.T. (1995). Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. Hum. Mol. Genet. *4*, 1–8.

61. Morales, F., Vásquez, M., Corrales, E., Vindas-Smith, R., Santamaría-Ulloa, C., Zhang, B., Sirito, M., Estecio, M.R., Krahe, R., and Monckton, D.G. (2020). Longitudinal increases in somatic mosaicism of the expanded CTG repeat in myotonic dystrophy type 1 are associated with variation in age-at-onset. Hum. Mol. Genet. *29*, 2496–2507.

62. Morales, F., Couto, J.M., Higham, C.F., Hogg, G., Cuenca, P., Braida, C., Wilson, R.H., Adam, B., del Valle, G., Brian, R., et al. (2012). Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. Hum. Mol. Genet. *21*, 3558–3567.

63. Campion, L.N., Mejia Maza, A., Yadav, R., Penney, E.B., Murcar, M.G., Correia, K., Gillis, T., Fernandez-Cerado, C., Velasco-Andrada, M.S., Legarda, G.P., et al. (2022). Tissue-specific and repeat length-dependent somatic instability of the X-linked dystonia parkinsonism-associated CCCTCT repeat. Acta Neuropathol. Commun. *10*, 49.

64. Morales, F., Vásquez, M., Santamaría, C., Cuenca, P., Corrales, E., and Monckton, D.G. (2016). A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. DNA Repair *40*, 57–66.

65. Laabs, B.-H., Klein, C., Pozojevic, J., Domingo, A., Brüggemann, N., Grütz, K., Rosales, R.L., Jamora, R.D., Saranza, G., Diesta, C.C.E., et al. (2021). Identifying genetic modifiers of age-associated penetrance in X-linked dystonia-parkinsonism. Nat. Commun. *12*, 3216.

66. DiFiglia, M., Sapp, E., Chase, K.O., Davies, S.W., Bates, G.P., Vonsattel, J.P., and Aronin, N. (1997). Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. Science *277*, 1990–1993.

67. Gangwani, M.R., Soto, J.S., Jami-Alahmadi, Y., Tiwari, S., Kawaguchi, R., Wohlschlegel, J.A., and Khakh, B.S. (2023). Neuronal and astrocytic contributions to Huntington's disease dissected with zinc finger protein transcriptional repressors. Cell Rep. *42*, 111953.

68. Vonsattel, J.P., Myers, R.H., Stevens, T.J., Ferrante, R.J., Bird, E.D., and Richardson, E.P., Jr. (1985). Neuropathological classification of Huntington's disease. J. Neuropathol. Exp. Neurol. *44*, 559–577.

69. Hedreen, J.C., Peyser, C.E., Folstein, S.E., and Ross, C.A. (1991). Neuronal loss in layers V and VI of cerebral cortex in Huntington's disease. Neurosci. Lett. *133*, 257–261.

70. Mattsson, B., Gottfries, C.G., Roos, B.E., and Winblad, B. (1974). Huntington's chorea: pathology and brain amines. Acta Psychiatr. Scand. Suppl. *255*, 269–277.

71. Greenstein, P.E., Vonsattel, J.-P.G., Margolis, R.L., and Joseph, J.T. (2007). Huntington's disease like-2 neuropathology. Mov. Disord. *22*, 1416–1423.

72. 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., et al.; TRACK-HD team (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.

73. Ciosi, M., Cumming, S.A., Chatzi, A., Larson, E., Tottey, W., Lomeikaite, V., Hamilton, G., Wheeler, V.C., Pinto, R.M., Kwak, S., et al. (2021). Approaches to sequence the HTT CAG repeat expansion and quantify repeat length variation. J. Huntingtons Dis. *10*, 53–74.

74. Warner, J.H., Long, J.D., Mills, J.A., Langbehn, D.R., Ware, J., Mohan, A., and Sampaio, C. (2022). Standardizing the CAP score in Huntington's disease by predicting age-at-onset. J. Huntingtons Dis. *11*, 153–171.

75. Ling, E., Nemesh, J., Goldman, M., Kamitaki, N., Reed, N., Handsaker, R.E., Genovese, G., Vogelgsang, J.S., Gerges, S., Kashin, S., et al. (2024). A concerted neuron–astrocyte program declines in ageing and schizophrenia. Nature *627*, 604–611.

76. Wells, M.F., Nemesh, J., Ghosh, S., Mitchell, J.M., Salick, M.R., Mello, C.J., Meyer, D., Pietilainen, O., Piccioni, F., Guss, E.J., et al. (2023). Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages. Cell Stem Cell *30*, 312–332.e13.

77. Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q., and Powell, J.E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. *20*, 264.

78. Virshup, I., Rybakov, S., Theis, F.J., Angerer, P., and Alexander Wolf, F. (2021). anndata: annotated data. Preprint at bioRxiv. 12.16.473007. https://doi.org/10.1101/2021.12.16.473007.

79. Choudhary, S., and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. *23*, 27.

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Brain tissue (anterior caudate) from persons with HD and unaffected controls	Harvard Brain Tissue Resource Center (NIH NeuroBioBank)	https://neurobiobank.nih.gov/about/
Chemicals, peptides, and recombinant proteins		
Nuclei EZ lysis buffer	MilliporeSigma	NUC101
NxGen RNase Inhibitor	Biosearch Technologies	30281
Optiprep Density Gradient Medium	MilliporeSigma	D1556
SPRIselect	Beckman Coulter	B23319
UltraRun LongRange PCR Kit	Qiagen	206442
Dynabeads MyOne Streptavidin C1	ThermoScientific	65002
iQ SYBR Green Supermix	Bio-Rad	1708880
Critical commercial assays		
Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Dual Index)	10X Genomics	PN-1000121
Pacific Biosciences SMRTbell® Express Template Prep Kit 2.0	Pacific Biosciences	PN 100-938-900
Deposited data		
Raw and processed data generated by the project	NIH Neuroscience Multi-omic Archive (NeMO), https://www.nemoarchive.org/	RRID:SCR_016152 Accession nemo:dat-ztfn3cc
Oligonucleotides		
HTT spike-in primers (used during cDNA amplification) (sequences: 5′CCCAGAGCCCCATTCATTGCC and 5′GGCGACCCTGGAAAAGCTGATG)	Integrated DNA Technologies	N/A
Primers for CAG-repeat amplification ("CAG Amp" step) (5′-/5BioagTCTCG TGGGCTCGGAGATGTGTATAAGAGACA GCCTTCGAGTCCCTCAAGTCCTTC and 5′TCGTCGGCAGCGTCAGATGTGTA TAAGAGACAGCTACACGACGCT CTTCCGATCT)	Integrated DNA Technologies	N/A
Software and algorithms		
Software code for processing single-cell/ single-nucleus RNA-seq data	GitHub	https://github.com/broadinstitute/ Drop-seq
Softwareode for deriving single-cell-resolution CAG-repeat-length measurements from HTT-CAG libraries	GitHub	https://github.com/broadinstitute/ HTT-CAG-Software
Softwareode for modeling the temporal dynamics of change in HTT-CAG repeat lengths	GitHub	https://github.com/broadinstitute/ HD-CAG-Modeling
Other		
Interactive data browser/explorer	McCarroll lab website	https://mccarrolllab.org/hd_long_ somatic_expansion

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Brain donors were recruited by the Harvard Brain Tissue Resource Center/NIH NeuroBioBank (HBTRC/NBB), in a community-based manner, across the USA. The HBTRC procedures for informed consent by the donor's legal next-of-kin and distribution

of de-identified post-mortem tissue samples and demographic and clinical data for research purposes are approved by the Mass General Brigham Institutional Review Board. Post-mortem tissue collection followed the provisions of the United States Uniform Anatomical Gift Act of 2006 described in the California Health and Safety Code section 7150 and other applicable state and federal laws and regulations. Human brain tissue was obtained (for data generation) from the HBTRC/NBB (NBB request ID# 1835). Federal regulation 45 CFR 46 and associated guidance indicates that the generation of data from de-identified post-mortem specimens does not constitute human participant research that requires institutional review board review.

The HBTRC/NBB confirmed HD diagnosis and excluded clinical comorbidity and presence of unrelated pathological findings by reviewing medical records and by formal neuropathological assessment. The 1985 Vonsattel et al. grading of neostriatal pathology[68] was used for diagnosis. Diagnosis on early cases is done using histological stainings and polyglutamine immunohistochemistry.[68–70] Positivity in pontine gray neurons rules out HD-like-2 neuropathology,[71] and cerebellar dentate neurons are mildly positive even in very early cases, while Purkinje cells are negative (unlike in cerebellar ataxia CAG expansion cases).

Donor ascertainment

Affected individuals were selected for analyses so as to represent a range of HD stages – from "at-risk" gene-expansion carriers who passed away before symptom onset, to affected persons with advanced caudate neurodegeneration. Analyses excluded donors with rare (minor allele frequency < 1%) large-effect modifier alleles or uncertain clinical history or diagnosis.

Experiments utilized fresh frozen brain tissue from each donor.

METHOD DETAILS

Determination of inherited CAG repeat length

We sequenced the CAG repeat within the *HTT* gene in each donor's genomic DNA (isolated from Brodmann Area 17) using the MiSeq assay developed by Darren Monckton's lab.[72,73]

Calculation of CAG-age-product score

We used a well-established approach to calculate CAP (CAG-age-product) score[32]

$$age * (inheritedCAGlength - 33.66)$$

We also considered a newer formula (CAP-100)[74] that weights the contribution of age and inherited CAG differently and is standardized such that CAP = 100 at the expected age of diagnosis, though we found that the Zhang formula exhibited a (modestly) stronger relationship to SPN loss across the brain donors in our analyses.

Single-nucleus RNA-seq in 20-donor "villages"

For analyses comparing *across* donors (e.g. Figure 1), to make rigorous comparisons of nuclei from many brain donors – while controlling for technical influences from extraction of nuclei, single-cell library construction, and sequencing – we processed sets of 20 brain specimens (each consisting of affected and control donors) at once as a single pooled sample, an approach we have previously described[75,76] in which we make preparations of nuclei from sets (or "villages"[76]) of 20 donors at once (Figure S1A). Specimens were allocated into batches of 20 specimens per batch. Each set of 20 tissue samples was processed as a single sample through nuclei extraction, encapsulation in droplets, library creation, and sequencing (Figure S1A). We analyzed each village using eight encapsulation reactions. We used combinations of hundreds of transcribed SNPs in each cell's sequence reads to assign each nucleus to its donor-of-origin, using the computational approach we have described previously.[75,76] This experimental approach (Figure S1A) allowed the data to be highly comparable donor-to-donor.

Note that in other experiments (Figures 2, 3, 4, 5, and 6) we deeply sampled nuclei from individual donors (generally 4-8 encapsulation reactions per donor) to increase the statistical power of within-donor comparisons of the individual nuclei (for example, to recognize relationships of CAG repeat length to gene-expression levels).

Isolation of nuclei from brain tissue

Nuclei were isolated from frozen brain tissue using approaches we have described[75,76] and deposited in protocols.io (https://www.protocols.io/view/village-nuclei-isolation-with-optiprep-36wgq3bmxlk5/v1). Briefly, in Ling et al., frozen brain tissues (20 specimens including 10 controls and 10 HD patients were pooled in a village, otherwise each specimen was processed individually for deep-dive experiment) on the glass slide was shaved off, minced, and transferred to a 6-well plate containing nuclei extraction buffer {NEB: 1% Triton X-100, 5% Kollidon VA64 in dissociation buffer (DB: 81.67 mM Na_2SO_4, 30 mM K_2SO_4, 10 mM glucose, 10 mM HEPES, 5 mM $MgCl_2$ [pH 7.4])}. Tissues were disrupted by pipetting and syringing, and filtered through a 20-micron filter and 5-micron filter serially. The filtered nuclei were resuspended in 50 mL of DB and spun down at 500 g in 4C for 10 min. After removing the supernatant, the pellet was resuspended in 1 mL of DB. Nuclei were visualized and counted by staining them with DAPI. For the density gradient-based nuclei isolation, the frozen brain tissue was transferred to dounce homogenizers filled with Nuclei EZ lysis buffer (MilliporeSigma, #NUC101) supplemented with 1 U/μL of NxGen® RNase Inhibitor (Biosearch technologies, #30281). After the tissues were homogenized by douncing, the lysates were filtered with 70 micron cell strainers and spun down at 4C with 500 g for

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

5 min. The supernatant was discarded and the pellets were resuspended in 300 μL of G30 (30% iodixanol, 3.4% sucrose, 20 mM tricine, 25mM KCl, 5 mM $MgCl_2$, [pH 7.8]). The resuspended tissue pellets were layered with 1 mL of G30 and spun down at 4C with 8000 g for 10 min. The supernatant was carefully removed and the nuclei pellet was washed twice with 1 mL of wash buffer (1% BSA in PBS supplemented with 1 U/μL NxGen® RNase Inhibitor). The nuclei were resuspended in 50 μL of the wash buffer and counted by using LUNA-FL™ Dual Fluorescence Cell Counter (Logos Biosystems).

Preparation of snRNA-seq libraries

The isolated nuclei were encapsulated into droplets and the snRNA-seq library was prepared by using M, Library & Gel Bead Kit v3.1 (10X Genomics, PN-1000121) according to the manufacturer's protocol with only minor modifications. The libraries were sequenced on Illumina NovaSeq 6000 systems platform.

Single-cell measurement of CAG-repeat length

We developed a novel approach for sequencing the CAG repeat of *HTT* transcripts in snRNA-seq experiments, and assigning these sequences to the cell from which the *HTT* transcript was derived, and thus connecting it to that cell's larger RNA-expression profile.

From each set of nuclei, our approach creates two molecular libraries: one library samples genome-wide RNA expression (''transcriptome library''), and another library specifically captures the 5′ region of *HTT* transcripts (''HTT-CAG library''). The presence of cell barcodes, shared between the two libraries, allows each CAG-length measurement to be matched to the gene-expression profile of the cell from which it is derived, and thus to the identity and biological state of that cell.

Key aspects in creating these HTT-CAG libraries include the use of *HTT*-targeting primers at multiple steps; *HTT*-targeted amplification and purification; steps to preserve long molecules throughout library preparation; careful calibration of PCR conditions to prevent the emergence of chimeric molecules during PCR; and analysis by long-read sequencing. An elaborated, expanded protocol, with illustrations and diagrams, helpful tips, potential modifications, and pausing points, is in Methods S1 section ''single-cell HTT-CAG and RNA sequencing.'' We describe here the key steps as implemented in the current work.

We begin by isolating and encapsulating nuclei in droplets as described above, using the standard 10X Genomics 3′ snRNA-seq protocol.

We modify the cDNA amplification step ("cDNA amp") of the standard 10X single-cell protocol by spiking in two primers designed to sequences 5′ of the *HTT* CAG repeat. (The sequences of these primers are in the key resources table). When using these spike-in primers, we add 2 μL of the 100 μM spike-in primers (1 μL per each primer, final concentration of each primer was 1 μM) in each reaction during the step 2.2a of the standard 3′ snRNA-seq protocol from 10X Genomics. The volume of the sample combined with the cDNA Amplification Reaction Mix is decreased accordingly to maintain 100 μL reaction volume. We then perform PCR as described in the 10X single-cell protocol.

The product of this amplification – a complex mixture of cDNAs (from all genes), with cell barcodes and UMIs incorporated into the cDNA molecules, in which each founding molecule (with a distinct cell barcode and UMI) is now represented by many copies – is then split into fractions which are used respectively to prepare the conventional ''transcriptome library'' (for single-cell analysis of genome-wide RNA expression) and the DNA-repeat (HTT-CAG) library.

The standard transcriptome sequencing library is generated from this cDNA amplification product by continuing with the standard 10X Genomics 3′ snRNA-seq protocol.

To generate the HTT-CAG library, we perform the following steps.

In the "CAG Amp" step, we seek to further amplify and isolate barcoded cDNA molecules that are informative about the *HTT* CAG repeat. In this step, we start with the purified output of the transcriptome-amplification step, which is also an intermediate created in the process of generating the 10X 3′ snRNA-seq library. Making the standard transcriptome library generally uses only some (15 μL) of this intermediate, and we use part of the rest to make the target-sequence library. (We do not use all of it, in case something goes wrong with either library and necessitates a re-do. The key thing is to use sufficient volume that almost all UMI-tagged cDNAs amplified in the previous PCR are sampled at least once. For HTT-CAG libraries, we estimate that an input of 4μL generally accomplishes this, and that use of more sample yields more-incremental increases in the number of UMIs ascertained.) We use a biotinylated gene-specific primer (designed 5′ of the target sequence, and 3′ of any spike-in primers) and another primer designed to the 10X bead sequence, to selectively amplify molecules that contain the target sequence (the CAG repeat within exon 1 of the *HTT* gene).

Primer sequences are in the key resources table. We utilize the UltraRun LongRange PCR Kit for amplification; the number of PCR cycles should be carefully calibrated to the sample, with the PCR ended while in exponential phase (sometimes also called ''log phase''), to prevent late cycles in which incompletely replicated molecules then act as primers in subsequent PCR cycles; this priming by incompletely replicated molecules causes cell barcodes and UMIs to appear in association with the wrong cDNAs. In Methods S1 section ''single-cell HTT-CAG and RNA sequencing'' we provide diagrams illustrating a way to calibrate the number of PCR cycles using real-time quantitative PCR.

We then purify the resulting PCR product on streptavidin beads (Dynabeads™ MyOne™ Streptavidin C1 (ThermoScientific, #65002)) to enrich the library for *HTT* CAG sequences generated by the biotinylated primer. Since the gene-targeting (5′ HTT) primer we used in the CAG amp step is biotinylated, the streptavidin beads will bind the target molecules. The bead-associated molecules are the molecules we elute and carry forward into downstream steps.

We then separate the resulting product into long ("L") and short ("S") molecular libraries from the same PCR reactions by using SPRIselect beads (Beckman Coulter, #B23319).

The next step (Indexing and further amplification) further amplifies the target molecules (the cDNAs that contain *HTT* CAG-repeat sequences) and adds molecular indexes so that libraries from multiple samples can be pooled for sequencing. We use standard Nextera indexing primers and the UltraRun LongRange PCR Kit. As with the target-sequence enrichment step, the number of PCR cycles is ideally calibrated to the sample, with the PCR ended while in exponential ("log") phase, as described in Methods S1, section "single-cell HTT-CAG and RNA sequencing." We purify this product using SPRIselect beads as described above. This product can optionally be sequenced using Illumina short reads.

The final step involves further preparing libraries for long-read sequencing. The PacBio libraries were generated by using SMRTbell® express template prep kit 2.0 (Pacific Biosciences, #100-938-900). The "L" and "S" libraries were sequenced on different flow cells on the SEQUEL IIe platform (Pacific Biosciences). To improve sequencing yield, we used the adaptive loading feature of the sequencing platform to target a concentration of 100 pM.

An extended protocol with diagrams and further discussions of each step's optimization (to other biological samples), stopping points, and troubleshooting tips is in Methods S1 section "single-cell HTT-CAG and RNA sequencing." We also suggest checking the McCarroll lab website (additional resources) for future protocol updates and improvements.

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing snRNA-seq data

Raw sequencing reads were aligned to the hg38 reference genome with the standard Drop-seq (v2.4.1) workflow. Reads were assigned to annotated genes if they mapped to exons or introns of those genes. Ambient / background RNA were removed from digital gene expression (DGE) matrices with CellBender (v0.1.0) remove-background.

All classification models for cell assignments were trained using scPred[77] (v1.9.2). DGE matrices were processed using the following R and python packages: Seurat (v3.2.2), SeuratDisk (v0.0.0.9010), anndata (v0.8.0),[78] numpy (v1.17.5), pandas (v1.0.5), and Scanpy (v1.9.1).

Assignment of nuclei to individual donors

Individual nuclei were assigned to their donor-of-origin using combinations of hundreds of transcribed SNPs in each single-nucleus RNA-expression profile; we did this using the Dropulation software, which we have described and used previously[75,76]). Single-droplet expression profiles (corresponding to individual cell barcodes) were excluded from downstream analyses if they were determined by Dropulation to be likely (based on combinations of transcribed SNPs) to be doublets (a combination of nuclei from two donors), or if they could not be confidently assigned to one of the donors identified by our ascertainment strategy (above). Nuclei from three donors were excluded based on low RNA ascertainment (UMIs per nucleus). Data from 103 donors (53 unaffected controls and 50 persons with HD, Table S1) were carried forward into analyses.

Classifying nuclei by cell type

Inference of the cell-type-of-origin of each nucleus was made from its RNA-expression profile using scPred. Data used in analysis were filtered to remove any doublets as detected by DoubletFinder (which is also incorporated into the DropSeq analysis workflow) and any cells for which the likelihood of a correct cell type assignment (max.prob) was less than 0.8.

Case-control differences in gene expression

A common approach to functional genomics in human disease involves comparing gene-expression data between disease-affected (case) and unaffected (control) individuals to arrive at a list of "differentially expressed genes" (DEGs). To do this, we applied a conservative statistical approach – a non-parametric Wilcoxon test comparing the 50 persons with HD to the 53 controls – to identify differentially expressed genes in each of principal caudate cell types. We found that, even with this conservative inferential approach, every caudate cell type – including all types of neurons and glia – exhibited thousands of DEGs whose expression levels differed (on average) between cases and controls, as described in more detail in Methods S1 section "case-control analyses." This broadly altered gene expression in every cell type potentially reflects the profound consequences of HD, which causes atrophy of the entire caudate, neuronal death, and devascularization – changes that are likely to affect the biology of every cell type. Consistent with the idea that DEGs are largely responses to caudate atrophy, an HD-cases-only analysis found that the magnitude of these gene-expression changes was strongly correlated with the extent of a donor's earlier SPN loss (Methods S1 section "case-control analyses"). For these reasons, and given our interest in finding the first-order effects of the *HTT* CAG repeat, we focused the rest of our analyses on trying to identify gene-expression changes that associated with a cell's own CAG-repeat length. Methods S1 section "case-control analyses" includes an extended discussion comparing the results of case-control and CAG-length analyses, and may be of interest to readers planning deeper work in this area or wondering how the results of the CAG-length analysis (which we emphasize here) intersect with the results of conventional case-control analysis.

Please cite this article in press as: Handsaker et al., Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease, Cell (2025), https://doi.org/10.1016/j.cell.2024.11.038

Analysis of HTT-CAG library data

After sequencing, we processed the reads from each PacBio flowcell using a standard workflow for circular-consensus-sequencing (CCS) base calling and alignment using the following programs: (1) "ccs" (Pacific Biosciences) either version 6.0.0 or version 6.3.0 with standard options (2) "extracthifi" (Pacific Biosciences) to extract only reads with QV >= 20 and (3) "pbmm2" either version 1.4.0 or version 1.10.0 with —preset ISOSEQ and using the GRCh38 "no alt" reference genome (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz). Some of the data generated early in this study used one version of the above referenced software while data generated later in the study used updated versions of the same software, but we did not observe any functional differences between the different software versions.

After base calling and alignment, we further processed the data from each flowcell using a custom analysis pipeline we developed (https://github.com/broadinstitute/HTT-CAG-Software). This pipeline consisted of the following steps.

Each read was analyzed ("decoded") based on the expected layout based on the library construction protocol (a schematic depiction of the layout is in Methods S1 section "single-cell HTT-CAG and RNA sequencing"). The decoding algorithm searched each read for a particular set of landmarks, identifying the landmark using a sensitive Smith-Waterman alignment algorithm designed to accommodate base-level errors and base insertions or deletions in the input reads. Based on the recognition of these landmarks, the read was divided into segments capturing features of the read used in downstream analysis, including index sequences used for multiplexing multiple experiments on the same flowcell, cellular barcodes and unique molecular identifiers, and the sequence of the CAG repeat itself (Methods S1 section "single-cell HTT-CAG and RNA sequencing").

We decoded both reads that were aligned to exon1 of the *HTT* gene and unaligned reads. We found it was important to decode both aligned and unaligned reads as the aligner (pbmm2) exhibited bias in the success rate for correctly aligning reads to *HTT* exon 1 based on the CAG repeat length in the decoded read. When decoding aligned reads, we used the strand indicated by the alignment. For unaligned reads, we attempted to decode the read both as recorded in the input file and reverse-complemented and used the most confident decoding in the analysis.

Each read contained two index sequences (i5 and i7) which were used to identify the input snRNA-seq library ("reaction") when multiple reactions are multiplexed together on one PacBio flowcell. Given a list of input reactions and the pair of indexes used for each, we employed fuzzy matching to find the best match for each index sequence in the decoded read and assigned the matching input reaction if one or both indexes were able to unambiguously identify the input reaction. After determining the input reaction for each read, the cell barcode (CBC) and unique molecular identifier (UMI) were extracted from the decoded reads and reads from the same reaction containing the same CBC and UMI were aggregated. The CAG repeat length was estimated as the consensus length across all reads sharing the same CBC and UMI using the half-sample mode (hsm function in R library modeest). This yielded one repeat length measurement for each CBC+UMI combination.

Several quality control steps were then applied. To remove low quality reads, reads were dropped if the decoded UMI was longer than 28 base pairs or when the CAG repeat "purity" (fraction of bases not matching a pure CAG-triplet motif) was less than 90%. A CBC+UMI combination was retained only if it was supported by 10 or more PacBio reads. Within each reaction, we computed the Levenshtein edit distance (allowing indels) between all pairs of CBC+UMIs (R function "stringdist", method="lv"). To avoid double counting, if any CBC+UMI had another CBC+UMI from the same reaction with an edit distance of less than 4 changes, only one measurement was retained, preferring the one supported by the most reads. In principle, we could have performed error correction on the CBC+UMIs, but in practice we found that among "nearby" CBC+UMI values (likely generated by sequencing error) usually only one CBC+UMI was supported by a preponderance of the reads. The reaction from which the HTT-CAG library originated also underwent transcriptome analysis and QC. We retained only measurements where the CBC was found to (exactly) match a CBC from the same reaction that had passed all transcriptome QC.

After QC, in the small fraction of cases where we made multiple repeat-length measurements for the same cell (CBC) but with distinct UMIs, we used these for quality assessment (Figure 2B), but then retained only one measurement, keeping the UMI with the largest repeat measurement. This ensured that if we measured both the short and long alleles from a cell, we used the long allele in downstream analyses.

Somatic expansion in glia and interneurons

Though we observed the greatest somatic expansion in SPNs, we also sought to compare our data with the findings of a recent study that reported high rates of somatic expansion in cholinergic interneurons in caudate[15] To do this, we classified the caudate interneurons in our data set as either cholinergic or non-cholinergic, based on expression of the cholinergic marker genes *SLC5A7*, *SLC18A3*, *CHAT* and *LHX8*. We then quantified the degree of somatic expansion among different cell types based both on (a) the overall distributions of repeat length and (b) using a somatic instability index (Methods S1 section "somatic expansion in caudate cell types"). We found that the cholinergic interneurons accounted for the majority of the observed somatic expansion among all interneurons, but that cholinergic interneurons exhibited considerably less somatic expansion than SPNs did. An extended discussion of these approaches and analyses is in Methods S1 section "somatic expansion in caudate cell types."

SPN gene-expression comparisons (cell groups)

Analyses of the effect of CAG-repeat length on genome-wide gene expression in which we compared subsets of SPNs defined by CAG-repeat length ranges (Figures 3B, 3C, and S4B–S4E) utilized a Wilcoxon rank-sum test (wilcox.test() function in R), in which the individual SPNs (in two groups of SPNs defined by CAG repeat-length ranges) were ranked by their expression level of a gene (as a fraction of all UMIs detected in a cell), and then the distribution of ranks were compared between the two groups. We utilized the Wilcoxon test for these specific analyses because (i) this test makes no assumptions about parametric properties of the expression data, and (ii) the *p* value distributions it generates were appropriately null in control permutation analyses in which CAG-repeat lengths were permuted randomly across the individual cells. We note that such tests are still limited by the need to compare discrete groups of cells; to quantitatively estimate the functional effect of CAG-repeat length across its entire range, we developed regression-based approaches (next section).

CAG-repeat effects on gene expression

We sought to explore and evaluate the effect of CAG-repeat length upon gene expression in SPNs, and to systematically identify those genes whose expression levels are affected by CAG-repeat length. We found that Negative Binomial Regression (NBR) analyses made it possible to explore and critically evaluate a wide range of functional forms for the potential relationship of CAG-repeat length to gene expression in SPNs. Methods S1 section "recognizing effects of CAG-repeat length" provides a detailed description and extended discussion of the wide range of models we considered (including simple models in which the CAG-repeat had a continuous effect on gene expression throughout its length range), and the analyses leading to our final choice of model; that extended discussion and set of analyses will be of interest to readers who want to consider alternative models or to apply these kinds of approaches in other scientific contexts. We focus here on the specific model that underlies our final identification of phase C and phase D genes.

In order to identify genes whose expression changes as the HD-causing *HTT* allele's CAG-repeat expands, we considered a number of Generalized Linear Regression Models (GLMs) containing CAG-repeat length (CAG_LENGTH) as one of the predictive variables which contribute to a gene's expression level. We fitted these models using the single-cell-resolution data from the SPNs from all six (6) deeply sequenced individuals with clinically apparent HD. Some of the models were fitted on all the SPNs, while in others we considered only the SPNs with CAG-repeat length in a specific range (e.g. between 36 and 100), to better recognize any changes within this range and make sure the gene expression signal from the CAG_LENGTH term was not distorted by SPNs with far-longer CAG-repeat lengths. As described in Methods S1 section "recognizing effects of CAG-repeat length," these analyses pointed strongly toward a hinge-function relationship in which CAG repeat length had no effect on gene expression until the CAG repeat was longer than 150 CAGs.

Taking into consideration that single-cell-resolution gene expression data are consistently overdispersed,[79] we decided to utilize Negative Binomial Regression (NBR) models, which model the errors in GLMs using the Negative binomial distribution. Unlike the Poisson distribution, in which the variance υ is equal to the mean μ, in the Negative binomial distribution $\upsilon > \mu$, making it a more suitable distribution for modeling overdispersed data.

In more formal terms, for each gene *g*, we modeled the levels of its expression $E_g(c)$ (number of UMIs), in a cell c with a CAG-repeat length measurement (for the HD-causing allele) of CAG_LENGTH_c by fitting the following generic negative binomial regression (NBR) model:

$$E_g(c) \sim NB(\mu_g(c), v_g)$$

where the log mean of this negative binomial model, $\log(\mu_g)$, is a linear function L of the following cell covariates:

N_c is the total count of UMIs (RNA transcripts from any gene) in cell c

$f_{m,n}(\text{CAG\_LENGTH}_c)$ is a function family (parameterized by positive integers *m* and *n* such that $m < n$) of the CAG-repeat length of the HD allele in cell c. *f* can be a simple function of CAG_LENGTH (for example, it can be CAG_LENGTH itself), or can be defined for example as a "hinge function" (in which CAG_LENGTH has no effect until a threshold m is reached, then has a continuously escalating effect with further increase in CAG_LENGTH):

$f_{m,n}(\text{CAG\_LENGTH}_c) = 0$, if $\text{CAG\_LENGTH}_c < m$;

$f_{m,n}(\text{CAG\_LENGTH}_c) = \min(\text{CAG\_LENGTH}, n) - m$, if $\text{CAG\_LENGTH}_c \geq m$

SPN_DI_c indicates whether cell c is a direct or indirect SPN

SPN_MP_c indicates whether cell c is a matrix or patch (striosome) SPN

DONOR_c is the donor from whom cell c was sampled. Donor-level effects implicitly include the effects of age, genetic background, disease stage, and earlier caudate atrophy.

The final and most effective form of this model (m=150, n=500) that we tried involved a "hinge function" with a hinge at 150 CAGs – i.e., a model in which effects of CAG-repeat length commence at approximately 150 CAGs – and also included an additional phaseD term to capture phase D effects (Methods S1 section "recognizing effects of CAG-repeat length"). Genes for which $f_{m,n}(\text{CAG\_LENGTH}_c)$ had a highly significant regression coefficient were identified as Phase C genes; genes for which the phaseD term had a highly significant regression coefficient were identified as Phase D genes. These genes and their regression coefficients and *p* values are reported in Table S2.

We arrived at the final functional form of this model only after consideration and critical evaluation of a wide variety of functional forms, including simpler models that tried to explain gene-expression levels as simpler linear functions of CAG-repeat length. Methods S1 section "recognizing effects of CAG-repeat length" explains in detail the full set of alternative models that were evaluated, and the results of these evaluations.

Identification of phase D genes

Prominent among the genes found to change expression with CAG-repeat expansion beyond 150 CAGs were a set of genes that exhibited particularly large fold-changes; inspection of the data revealed that these large fold-changes resulted from these genes being almost completely repressed at baseline (in SPNs with <150 CAGs), and that these genes tended to have become de-repressed together in the same SPNs: generally, a specific subset of SPNs with particularly long expansions of the CAG repeat. To identify such genes systematically, we implemented a multi-stage approach, utilizing the single-SPN CAG-length and RNA-expression data from the six deeply sampled donors with manifest HD clinical motor symptoms (donors 1–6 in Table S1). In stage 1, we identified genes for which transcripts were (i) detected in fewer than 1% of those SPNs with <100 CAGs, and (ii) detected in a significantly larger fraction ($p < 10^{-5}$ by Fisher's exact test) of those SPNs with > 200 CAGs. This initial screen identified 89 genes (29 of which are in the HOX gene loci, and 60 of which are at other loci dispersed across the genome). Analysis of the expression of these genes in the single-SPN CAG-length and RNA-expression data also revealed that these genes were expressed in only a subset of those SPNs with >200 CAGs. In stage 2, we used the genes identified in stage 1 to refine our definition of the cells of interest, so that these were limited to 174 SPNs with >200 CAGs in which we also detected at least 3 UMIs from these 89 genes collectively (a "test set" of SPNs); we also defined a "control set" of 5,282 SPNs with <150 CAGs in which we detected only 0–1 UMIs from these 89 genes collectively. In stage 2, we identified genes for which transcripts were (i) detected in fewer than 1% of the "control set" SPNs, (ii) detected in at least 10 times higher fraction of the "test set" SPNs than in the "control set" SPNs and p < 10–5 by Fisher's exact test, and (iii) expressed in at least 3 "test set" SPNs. This analysis identified 107 "phase D de-repressed" genes, of which 39 were in the HOX loci. UMIs from these genes were used to recognize the phase D status of individual SPNs, a key input to the Negative Binomial Regression analyses of phase D gene expression in Methods S1 section "recognizing effects of CAG-repeat length." Note that, beyond these 107 de-repressed genes (which exhibit almost no detectable expression at baseline in SPNs), a larger set of genes was found by the NBR analysis to exhibit quantitative changes in gene expression levels in phase D SPNs. Both sets of genes are enumerated in Table S2.

Modeling CAG-repeat expansion dynamics

The observed distributions of the CAG-repeat lengths in caudate SPNs exhibited an unusual shape, with a mode on the left and a long right tail ("armadillo" shaped distributions, Figure 2C).

To better understand how these distributions might arise and evolve from the kind of simple, incremental, stepwise expansion-and-contraction process that biological studies have suggested is their primary mode of somatic mutation (reviewed in [46]), we developed stochastic models and simulations for the dynamics of the somatic expansion process (detailed in Methods S1 section "repeat expansion dynamics").

We developed stochastic models based on continuous-time Markov chains (CTMC) and evaluated several families of functions to generate the rate matrices for somatic expansion and contraction. The models are based on our growing understanding of the biological mechanisms that underlie expansion of the HTT-CAG repeat. The models take into account the loss of SPNs over time and our observation that there are few observable cell-autonomous transcriptional changes in neurons with shorter repeats (below 150 CAGs). Evaluation and simulation of these models suggested that a sufficient explanation for the armadillo-shaped distributions is a rapid increase in the rate (CAGs/year) of net somatic expansion in SPNs when the length of the repeat transitions from approximately 70 to 90 CAGs.

We evaluated and compared a variety of different models (see Methods S1 section "repeat expansion dynamics" for an extended discussion). The model selected for further analyses reported in the main text (TwoPhasePowerModel/150) models two phases of somatic expansion, in which each phase models the mutation rate as a power law function of the current repeat length, the transition point between the phases is fitted from the data, and it is assumed that cell loss occurs only at repeat lengths above 150 CAGs.

To understand the effects of inherited repeat length on the dynamics of somatic expansion (Figure 6D), we first fitted our selected somatic expansion model to the observed data for an example donor. We then ran a simulation using the fitted model parameters, changing only the length of the inherited repeat, but no other model parameters. The simulation provides a prediction of the repeat-length distribution at any time point in time, including ages prior to or later than their actual age of death.

To determine to what extent these models for the dynamic behavior of somatic repeat expansion could explain the inverse relationship between inherited repeat length and age of symptom onset (Figure 6E), we first fitted our selected somatic expansion model for each donor to their observed data. We then used as a proxy for age of symptom the age at which the fitted model would predict that 25% of that donor's SPNs would have expanded to 300 or more CAGs. These thresholds were based on (a) available medical records for age of disease onset in our donors, (b) the observation that repeat-length dependent transcriptional dysregulation begins at around 150 CAGs and (c) the small number of observed SPNs that attain repeats lengths longer than 500 CAGs. We then ran these

simulations for each donor, changing the inherited repeat length as in the analysis for Figure 6D, which yielded a curve representing the predicted relationship between inherited repeat length and our proxy for age of symptom onset.

To create animated visualizations of repeat expansion dynamics, we first fitted a specific repeat expansion model (TwoPhasePowerModel/150, Methods S1 section "repeat expansion dynamics") to model the dynamics for a specific donor and potentially a specific scenario (e.g. a hypothetical inherited repeat length) and then simulated the random walks of a large number of cells (n=3000). We initialized a vector of repeat lengths to the desired inherited allele length, then at each age from birth to the desired final length we modified the repeat length based on the transition probability matrix underlying the chosen model using a time step of 1 year. The final animations were created by plotting each individual movie frame separately (using R) as a png file and then combining the individual frames into an animated gif using the ffmpeg software (version 6.6.1).

To estimate the trajectory of the five phases of disease progression in the ELongATE model (Figure 7B) in an example donor, we fitted our selected expansion model to the observed data for that donor. We then defined approximate phase transitions for each cell in terms of the expected repeat length predicted by the model. We used a threshold of 80 CAGs for the transition from phase A to phase B, a threshold of 150 CAGs for the transition from phase B to phase C, a threshold of 250 CAGs for the transition from phase C to phase D and a threshold of 500 CAGs for the transition from phase D to phase E. Because the rate of expansion is rapid when the repeat is highly expanded (> 100 CAGs), these trajectories have limited sensitivity to the precise thresholds used for the later phases; simulations using different thresholds produced qualitatively similar results.

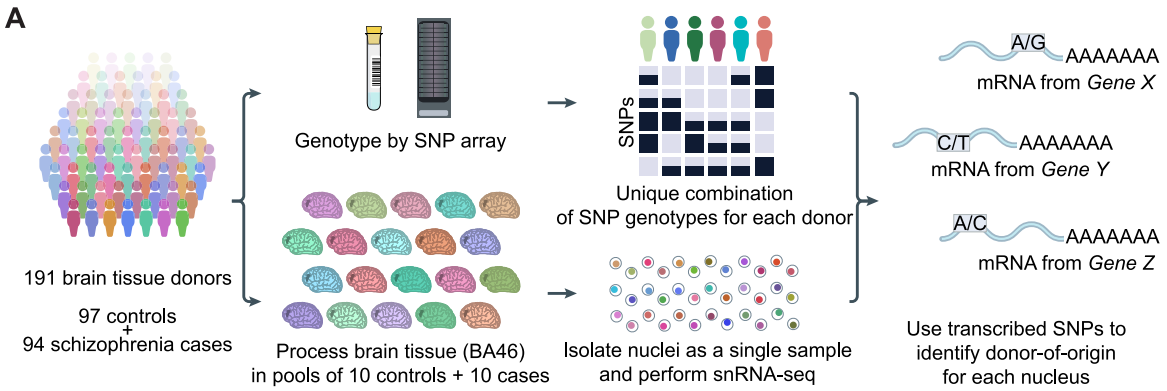
An extended discussion of our modeling choices, their rationales, and their critical evaluation is provided in Methods S1 section "repeat expansion dynamics." We have made the modeling software available in a Github repository as described in the Code Availability section.

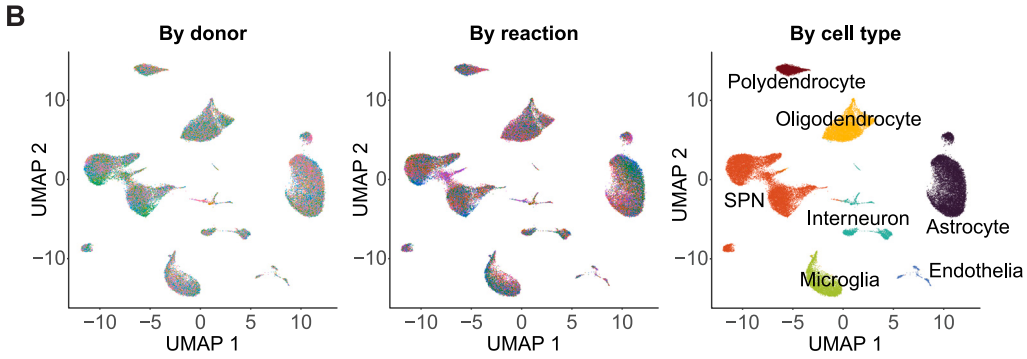
ADDITIONAL RESOURCES

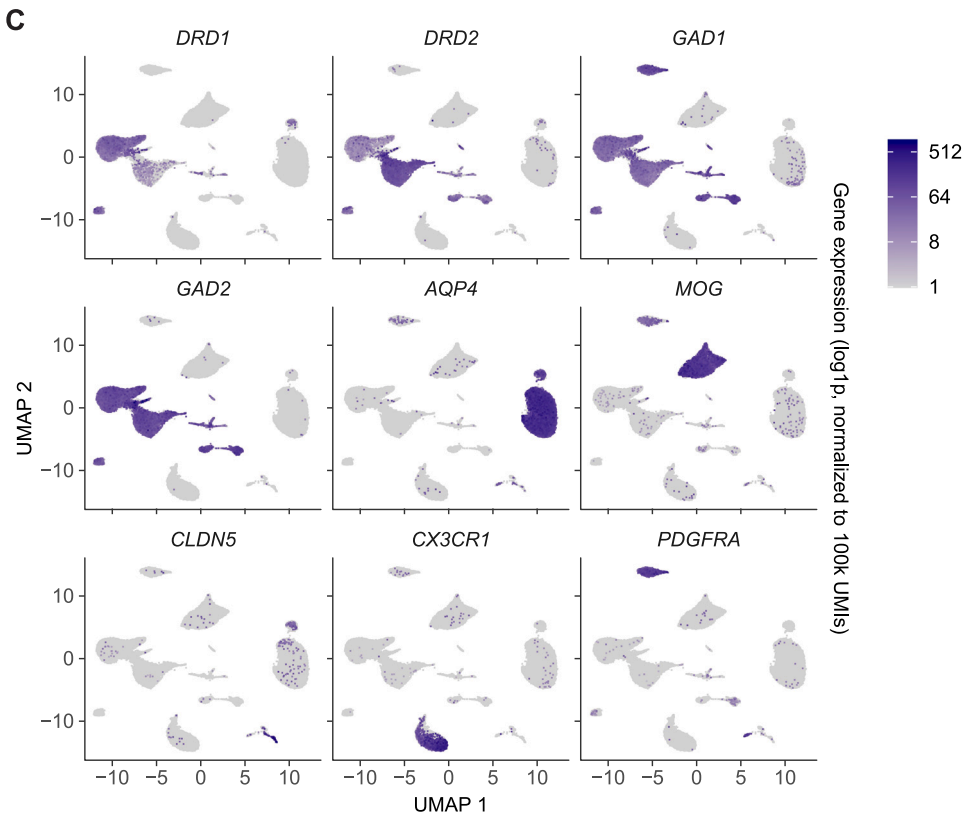
Web site with interactive data browser and other resources: https://mccarrolllab.org/hd_long_somatic_expansion

Supplemental figures

CellPress
OPEN ACCESS







(legend on next page)

Figure S1. Single-nucleus RNA-seq analysis of brain tissue from 50 persons with HD and 53 controls, related to Figure 1

(A) "Cell village" workflow by which we perform snRNA-seq on tissue from ~20 donors at once. Image is only lightly modified from Ling et al.,[75] where we describe this approach.

(B) Multi-dimensional single-cell RNA expression data for the 613 thousand striatal cell nuclei were projected into two dimensions using the uniform manifold approximation and projection (UMAP) algorithm and then colored based on their donor-of-origin (left), village-of-origin (center), or assigned cell type (right), which was based on their genome-wide RNA expression patterns.

(C) Expression patterns of known cell-type-specific marker genes, on the same UMAP as in (B).

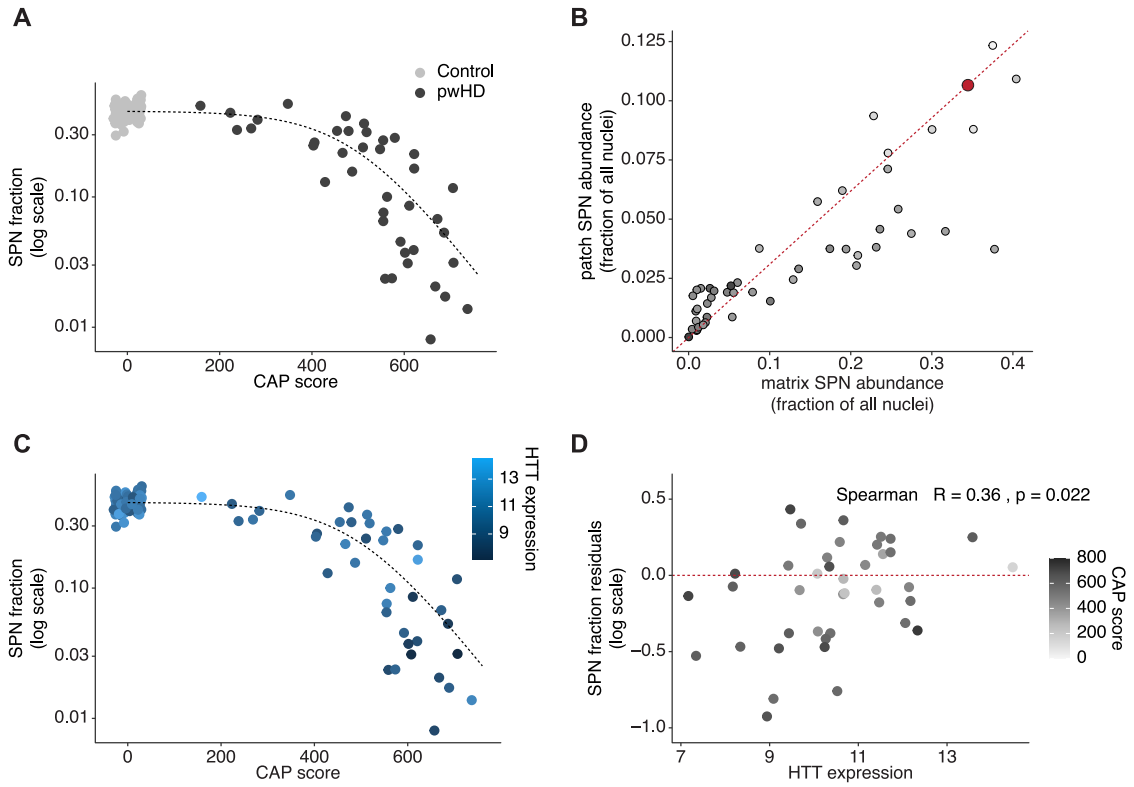


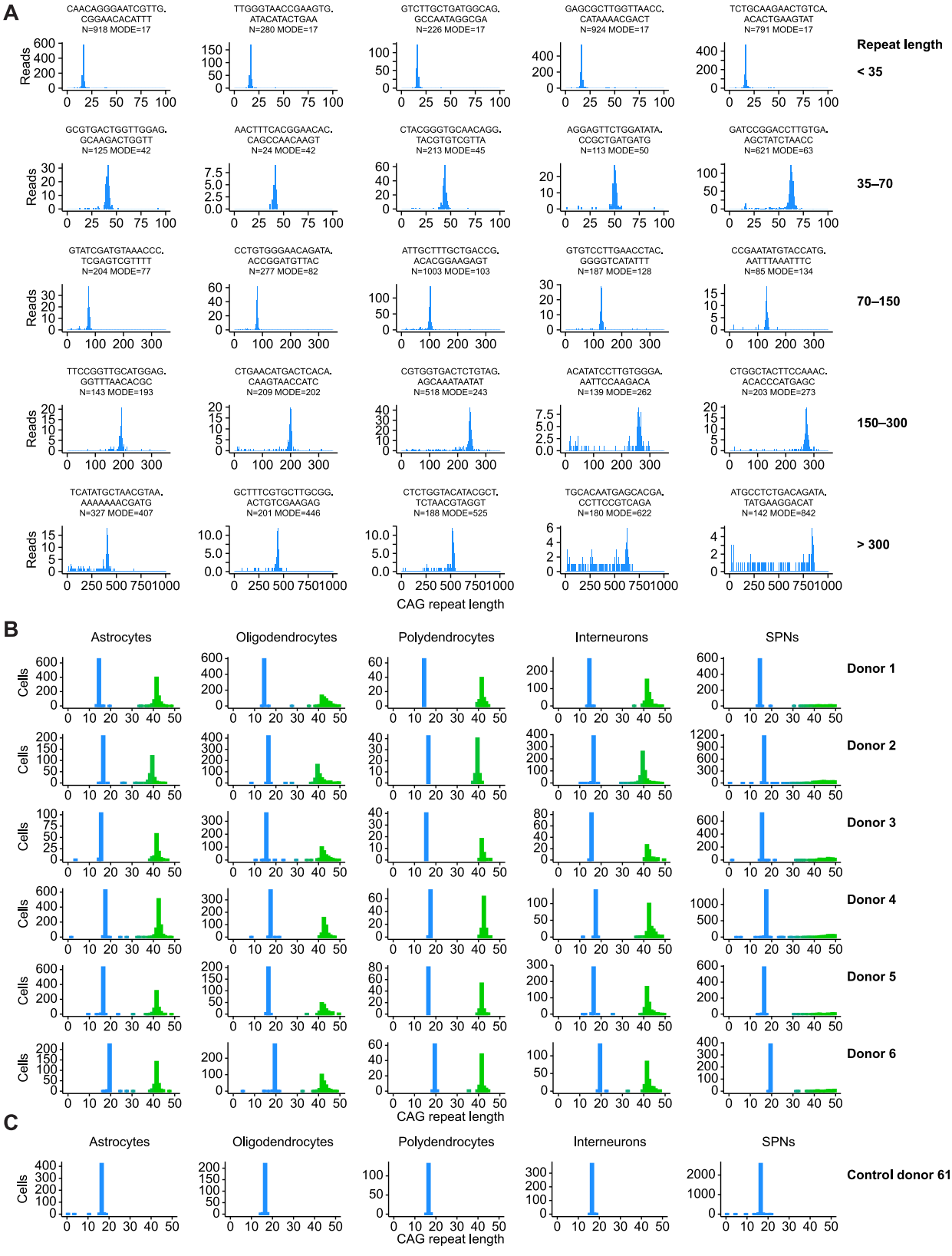
Figure S2. Relationships between SPN loss and other features of HD, related to Figure 1

(A) SPN loss with HD progression. Same data as in Figure 1B but shown here on a logarithmic scale. The slope of this relationship estimates rates of SPN loss with increasing CAP score. Unaffected control donors are shown as gray circles with CAP score of zero (jitter has been added to reduce overplotting). The dashed curve is from a fit of a logistic function to the SPN survival curve (before log transformation, i.e., as in Figure 1A).

(B) Decline in patch (striosomal) and matrix (extra-striosomal) SPNs with HD progression. Gray scale represents CAP score as in Figure 1. Red point denotes the median of 53 unaffected control donors.

(C) SPN loss and inter-individual variation in *HTT* expression levels. Same data as in (A), but points are colored to reflect individual donors' expression levels of *HTT* in SPNs.

(D) Residuals of the relationship in (C) are plotted against individual donors' *HTT* expression levels. Gray shading represents CAP score. Note that individuals' *HTT* expression levels show a weak, nominally positive relationship to SPN survival (rather than the negative relationship predicted by the "cumulative lifetime damage" model, in which individuals with higher expression levels would exhibit earlier/faster SPN loss), although this arises substantially from the donors in the lower-left part of the plot—donors with very high CAP scores and extreme caudate atrophy.



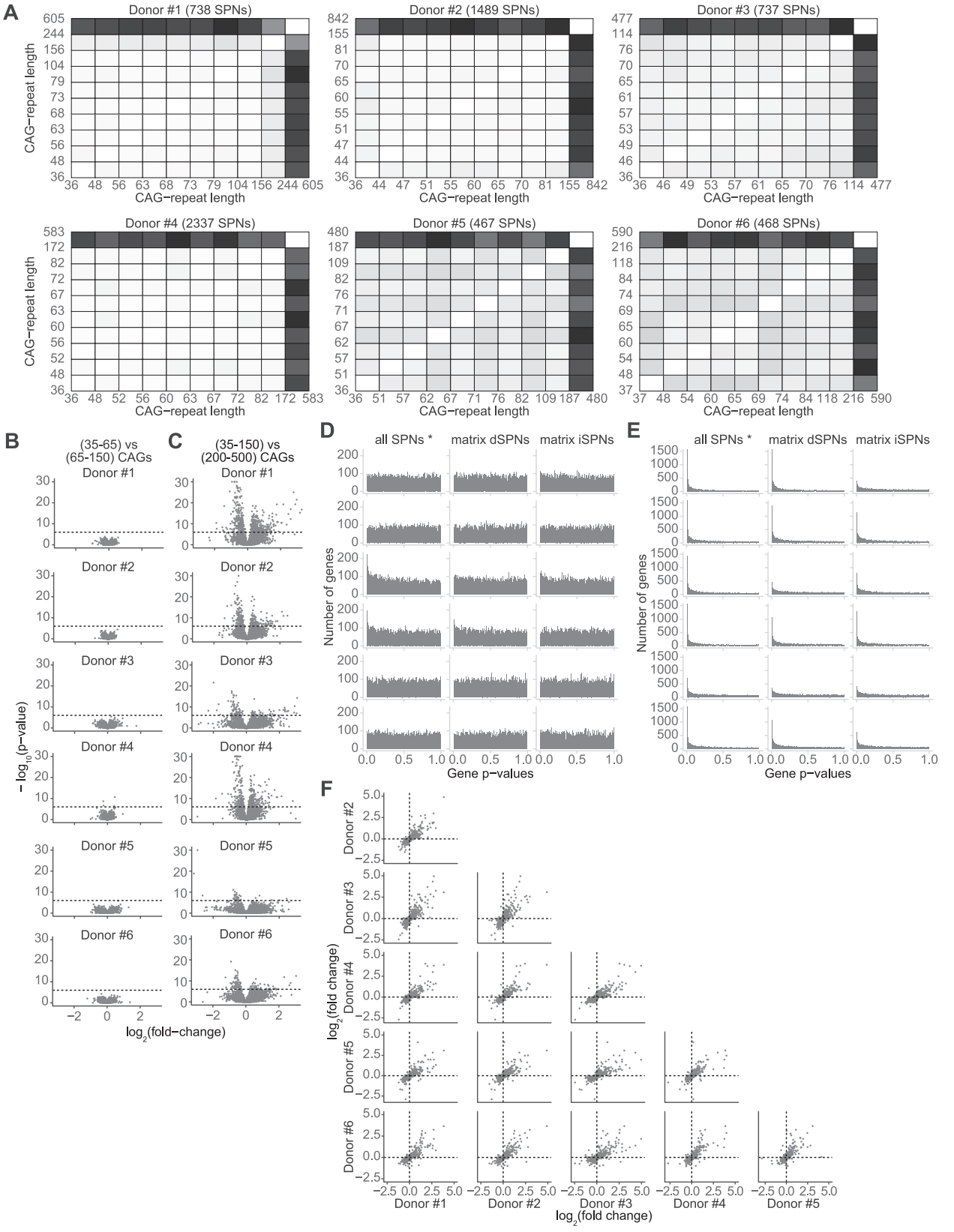
(legend on next page)

Figure S3. Determination of a consensus CAG-repeat length for individual HTT transcripts and allele specificity of somatic expansion, related to Figure 2

(A) Determination of consensus CAG-repeat length from sets of sequence reads derived from individual *HTT* RNA transcripts. Sequence reads (PCR products) originating from the same underlying RNA molecule share a common unique molecular identifier (UMI) that was applied during reverse transcription. The histograms show representative distributions of CAG-repeat lengths in the reads for individual UMIs for (top row) short, common HD alleles (<35 repeats) that do not cause HD; for HD-causing alleles with modest somatic expansion in the range of 40–60 CAGs (second row); for longer somatic expansions in the range of 80–150 CAGs (middle row); for UMIs with somatic expansions in the range of 150–300 CAGs (fourth row); and for UMIs showing very long somatic expansions beyond 300 CAGs (bottom row). Note that each row uses a different x axis scale. For transcripts with long somatically acquired CAG-repeat expansions, the PCR amplification during library preparation creates a left-tailed distribution, reflecting the way that PCR errors lead to the generation of molecules with shorter repeats and then favor these smaller molecules over longer ones. For each UMI, we use the mode (the Robertson-Cryer half-sample mode estimator, function hsm() from the R package modeest) as the consensus CAG-repeat length for that *HTT* transcript. The accuracy of these determinations is evaluated in Figure 2B. Cell barcodes on the same sequence reads make it possible to then connect each such CAG-repeat length determination to the wider RNA expression profile (cell type and cell state) of the nucleus from which it was sampled.

(B) Allele specificity of somatic expansion. Distributions of single-cell CAG-repeat length measurements for each donor/cell-type combination, showing both the short, non-HD-causing allele (<35 repeats, blue) and the longer, HD-causing allele (>35 repeats, green), here zoomed in to the 0–50 range (beyond which most SPNs have already expanded their HD-causing allele). The shorter allele (blue) appears to be somatically stable across neuronal and glial cell types in all six persons with HD.

(C) CAG-repeat length distribution from an unaffected control donor (homozygous for a repeat length of 17 CAGs). In this control donor, both alleles appear to be somatically stable.



(legend on next page)

Figure S4. Appearance of gene expression changes in SPNs with the longest CAG-repeat expansions, related to Figure 3

(A) Changes in SPN gene expression across deciles of the repeat length distribution. Correlation squares show the magnitude of gene expression differences (one minus the correlation coefficient) when comparing sets of SPNs (from the same donor) that have been grouped into deciles based on the CAG-repeat length of their HD-causing *HTT* allele. Gray scale: black indicates maximal difference observed in any comparison; white indicates no difference. Note that the CAG-repeat length thresholds for the deciles vary by donor and that data for donors whose SPNs were less deeply sampled (e.g., lower right) exhibit more statistical noise. The donor in the lower left is the same donor analyzed in Figure 3.

(B and C) Changes in SPN gene expression with somatic CAG-repeat expansion (volcano plots). (B) Comparisons of gene expression (volcano plots) of SPNs with 35–65 CAGs to SPNs with 66–150 CAGs, in six persons with HD. Dashed lines show the thresholds for genome-wide significance. (C) Comparisons of gene expression (volcano plots) of SPNs with 35–150 CAGs to SPNs with 200–500 CAGs, in six persons with HD. Note that the statistical power of the analysis varies by donor in relation to the number of SPNs sampled and with long (>150 CAGs) repeat expansion. *p* values (y axis) are derived from a Wilcoxon test across the individual SPNs in each group.

(D and E) Changes in SPN gene expression with somatic CAG-repeat expansion (*p* value distributions). *p* values are computed as in (B) and (C) (Wilcoxon test).

(D) Comparisons of gene expression (*p* value distributions) of SPNs with 35–65 to SPNs with 66–150 CAGs, in six persons with HD. Each row corresponds to a single donor with HD. The analyses in columns involve different sets of SPNs. Note that the "all SPNs" analysis (first column) involves a heterogeneous set of SPN subtypes (patch and matrix; direct and indirect); since SPN subtypes exhibit somewhat different rates of somatic expansion, this causes a small number of marker genes for these subtypes to also associate modestly with CAG-repeat length. The more rigorous comparisons in the other two columns involve specific, common SPN subtypes.

(E) Comparisons of gene expression (*p* value distributions) of SPNs with 35–150 to SPNs with 200–500 CAGs, in the same donors as in (D). Note that the statistical power of the analysis varies by donor and SPN subtype in relation to the depth of sampling of SPNs with long (>150 CAGs) repeat expansions. The y axis is truncated at 1,500 genes to aid visualization.

(F) Similarity of long-repeat-expansion-associated gene expression changes across persons with HD. Each panel is a pairwise comparison of SPN gene expression data involving two persons with HD (x and y axes), in which the values on the two axes are the $\log_2$ fold changes in gene expression when comparing (within-tissue) SPNs with >150 CAGs with SPNs with <150 CAGs. Genes whose expression levels change significantly with repeat expansion in at least one of the donors are shown.

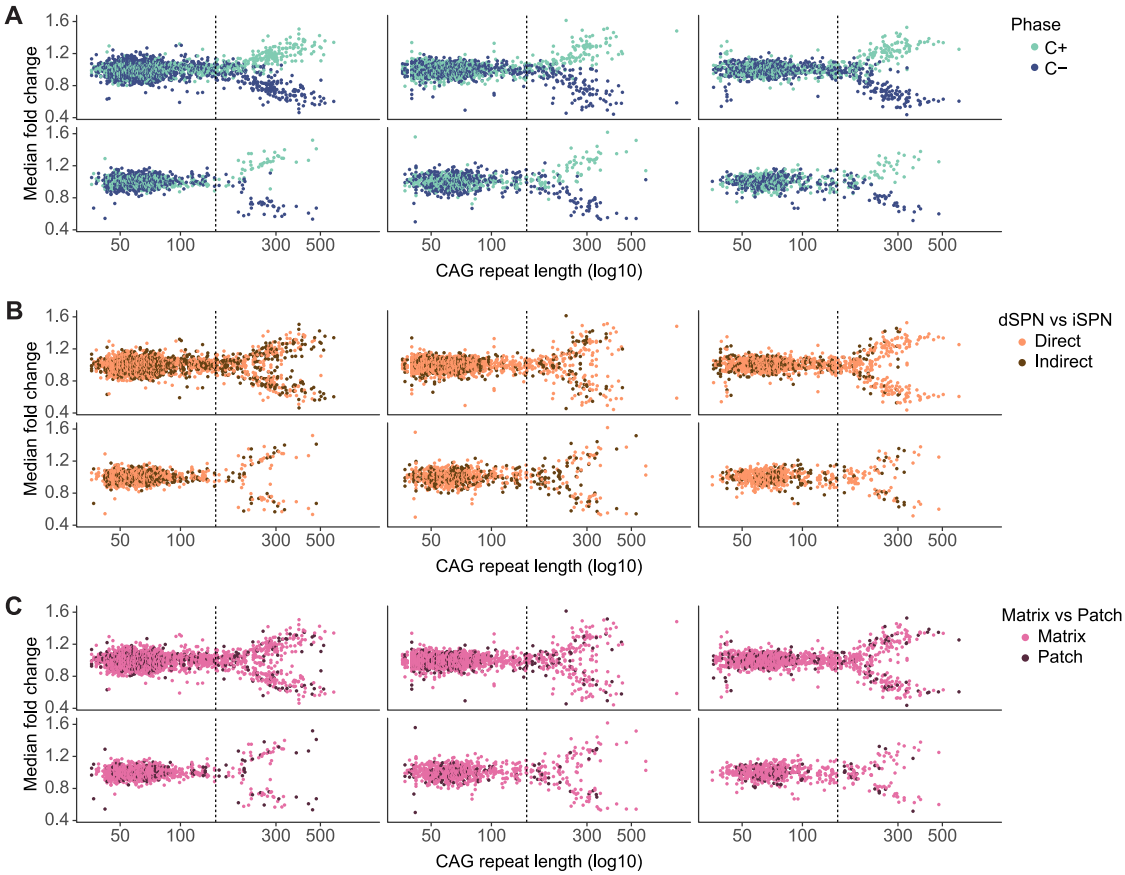
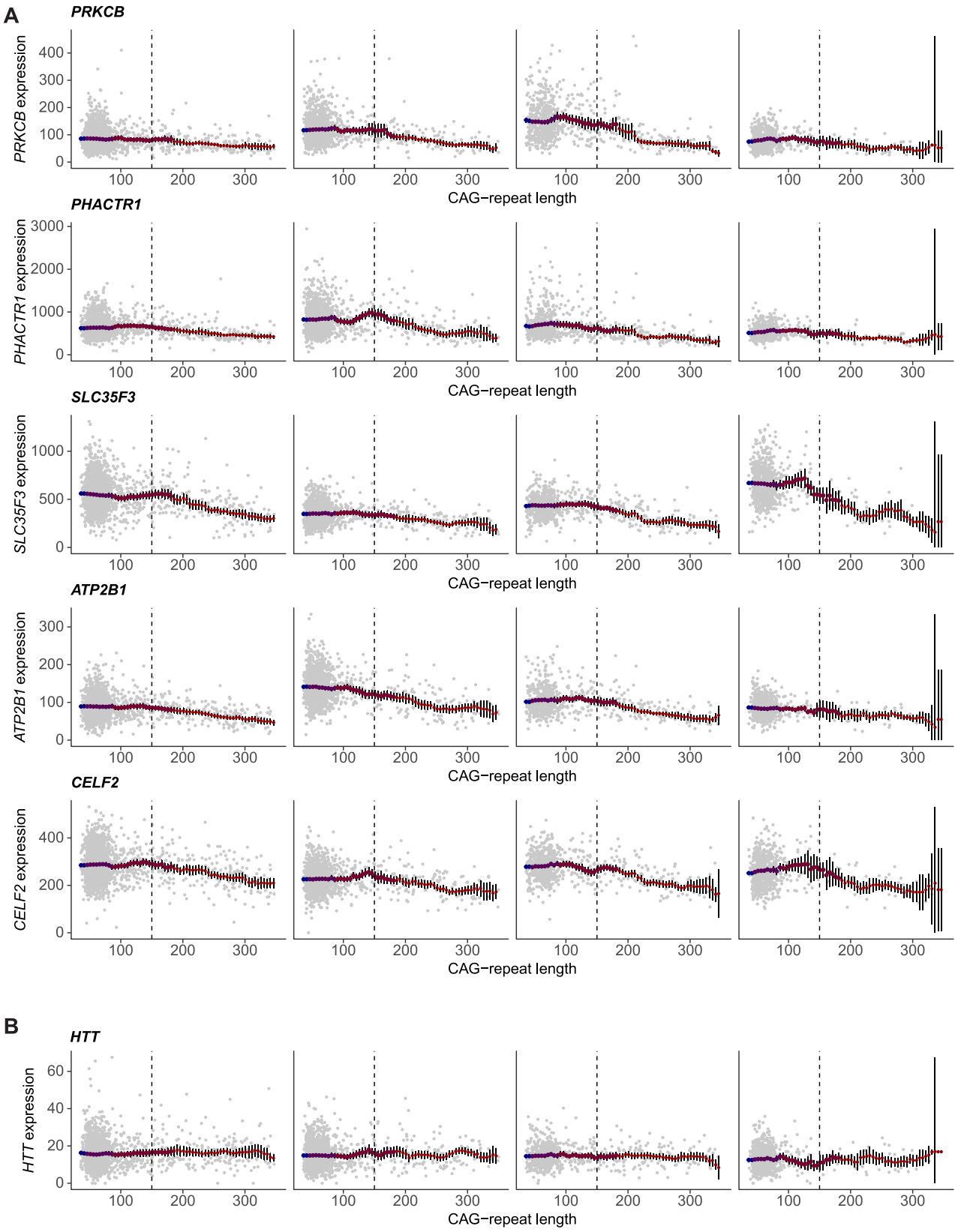


Figure S5. Shared thresholds for gene expression changes across SPN subtypes and individual persons with HD, related to Figure 4

(A) Relationship between transcriptional changes and CAG-repeat length (same data as in Figure 4B, but here with CAG-repeat length on the x axis, using a logarithmic scale). Each SPN is represented by both a blue point and a green point: blue points show the median fold change of a set of 192 genes, which decreases in expression with repeat expansion (C− genes); green points show the median fold change of a set of 274 genes, which increases in expression with repeat expansion (C+ genes).

(B and C) Same data as in (A), but here the same points are colored to show the subtype of SPN. Each SPN is represented by two points (here, in the same color) that show the expression fold change for its C+ and C− genes. In (B), "direct" and "indirect" refer to dSPNs (D1 SPNs) and iSPNs (D2 SPNs), respectively; in (C), "patch" and "matrix" refer to striosomal SPNs and extra-striosomal SPNs, respectively.



(legend on next page)

Figure S6. Expression levels of example genes as a function of CAG-repeat length, related to Figure 4

(A) Expression levels of five example phase C− genes in the individual SPNs of four persons with HD (one gene per row, one donor per column). Expression (y axis) is quantified as UMIs / 100,000. Gray points represent individual SPNs. Colored points are moving averages in windows of width $\log_2$(CAG-repeat length) to reduce the measurement noise inherent in single-cell measurement. Point color indicates window width, vertical bars denote confidence intervals. The genes shown (*PRKCB*, *PHACTR1*, *SLC35F3*, *ATP2B1*, and *CELF2*) are genes that SPNs normally express more strongly than interneurons do. This analysis also helps explain why conventional descriptive genomics analyses (to find "differentially expressed genes" in case-control comparisons) generally fail to recognize such effects in HD. First, these effects are present in just a small fraction of any donor's SPNs at any one time (those SPNs with long CAG-repeat expansions), and they are pronounced in a still smaller fraction (those in which the CAG repeat has expanded even further)—so they appear as small and often insignificant changes in bulk and sorted-cell-type analyses. Second, most of these genes, like most human genes, exhibit inter-individual variation in expression levels (at baseline), further obscuring (in case-control comparisons) effects that are clear in within-tissue comparisons of individual cells.

(B) Expression levels of *HTT* in the individual SPNs of persons with HD. Expression is quantified as in (A). Gray points represent individual SPNs. Colored points and vertical bars denote moving averages and confidence intervals, as in (A).

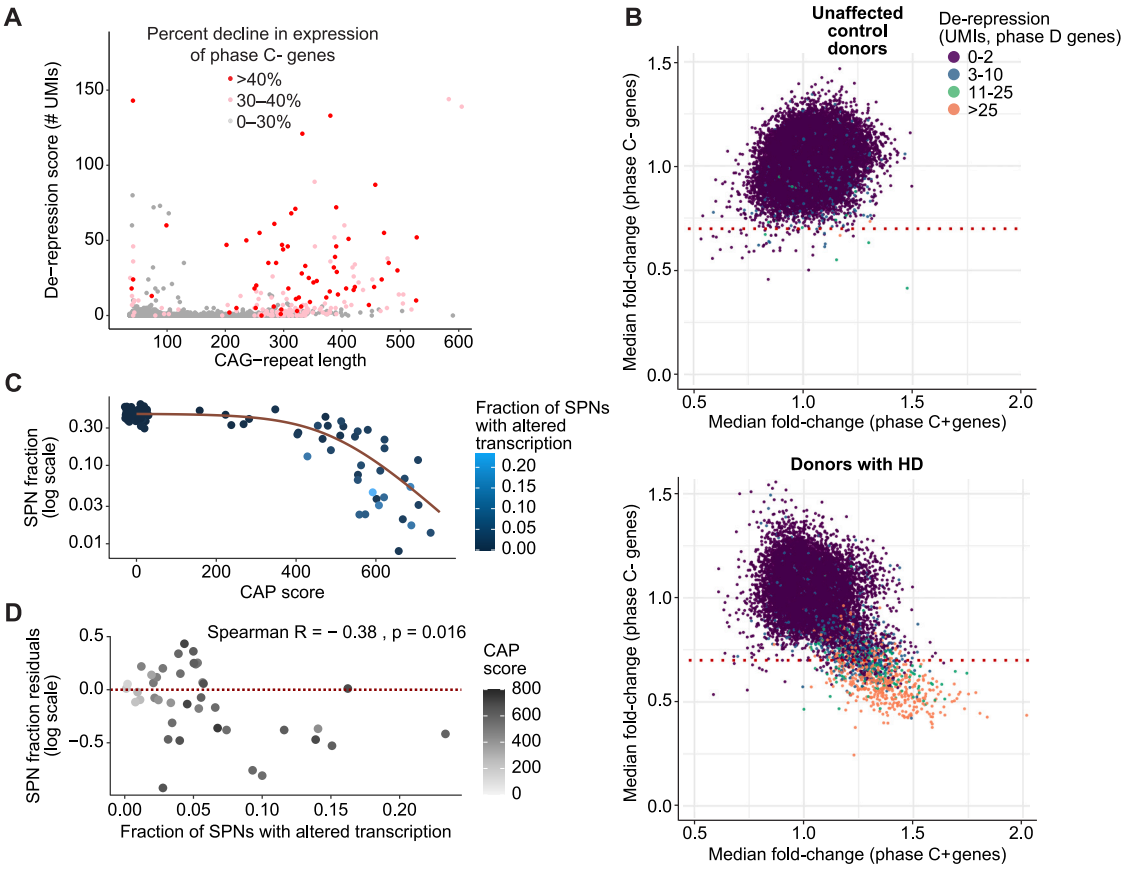


Figure S7. Gene de-repression in SPNs with long *HTT* CAG repeats and phase C gene expression changes, related to Figure 5

(A) Relationship of de-repression (y axis) to CAG-repeat length (x axis) and to progression of phase C gene expression changes (point colors). This is an alternative visualization of the relationship in Figure 5A, to make more visible the CAG-repeat lengths of most of the SPNs in which these genes have become de-repressed.

(B) Relationships (across individual SPNs, points) among phase D de-repression (indicated by point color, see figure legend) and progression of phase C gene expression changes (x axis: phase C+ [increasing-expression] genes; y axis: phase C− [decreasing-expression] genes). The upper panel shows SPNs sampled from 53 unaffected control donors; the lower panel shows SPNs sampled from 50 persons with HD. Dotted line shows the threshold used to define "altered transcription" for the analysis in Figure 5E.

(C and D) CAG-repeat-driven transcriptionopathy and rates of SPN loss. (C) SPN survival (on a logarithmic scale, y axis) is plotted against CAP score, an estimate of age-expected HD progression. The curve is from a fit of a logistic function to the SPN survival curve. Points are colored based on the fraction of each donor's SPNs that have phase C transcriptionopathy (which is plotted on the x axis in D). (D) Residuals of the relationship in (C) (negative values represent excess SPN loss relative to age-expected loss) are plotted (y axis) against the fraction of SPNs with transcriptionopathy (x axis). Gray shading represents CAP score. Note in (C) and (D) that donors for whom a larger fraction of SPNs exhibit phase C transcriptionopathy are generally donors with precocious SPN loss (relative to CAP score).