Long somatic DNA repeat expansions drive neuropathology in Huntington’s disease

Seva Kashin [1,2,,§], Robert E. Handsaker [1,2,§],
Nora Reed [1,2], Steven Tan [1,2], Won-Seok Lee [1,2], Tara McDonald [1,2],
Kiely French [3], Nolan Kamitaki [1,2], Christopher Mullally [1,2], Neda Morakabati [3],
Melissa Goldman [1,2], Elisabeth Lawton [3], Marina Hogan [1,2], Kiku Ichihara [1,2],
Sabina Berretta [1,3-5,*], Steven A. McCarroll [1,2,5,*]

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

'§ equal contributions; * jointly supervised this work

draft 2023.03.18
please email suggestions to (mccarroll at hms.harvard.edu)

Huntington's Disease (HD) is a fatal genetic brain disorder in which most of a person's striatal projection neurons (SPNs) degenerate and die. Science has long sought to understand why SPNs are so vulnerable in HD, why this pathology follows decades of apparent health, and how the disease-causing inherited DNA repeat (CAG_n, $n > 36$) in the *huntingtin* (*HTT*) gene leads to this neurodegeneration. This DNA repeat exhibits somatic mosaicism (variable length); we developed a way to measure its length together with genome-wide RNA expression in the same individual cells. We found that, in persons with typical inherited HD-causing alleles (of <50 CAG repeats), the CAG-repeat tract routinely expanded to 100-500+ CAG repeats in SPNs but rarely if ever did so in striatal interneurons or glia. Surprisingly, gene expression in these persons' individual SPNs exhibited no apparent relationship to those SPNs' CAG-repeat lengths across a wide range (36-150 repeats). In contrast, sparse SPNs with longer (150-500+) CAG repeats had profound gene-expression distortions which affected hundreds of genes, escalated alongside further repeat expansion, and culminated in widespread gene de-repression and expression of senescence/apoptosis genes. Across stages of HD, these "SLEAT" SPNs (with somatic long expansions and asynchronous toxicity) appeared in proportion to rates of SPN loss. Our experiments, analyses, and simulations suggest that individual SPNs undergo decades of biologically quiet DNA repeat expansion, then asynchronously enter a brief toxicity phase before dying. We conclude that, at any moment in time, most SPNs in persons with HD actually have a benign (but somewhat unstable) *huntingtin* gene; and that HD is a DNA process for almost all of a neuron's life.

Introduction

Huntington's Disease (HD) is a fatal genetic brain disorder. Most persons with HD are healthy for decades, then develop uncontrolled movements (chorea) and cognitive and psychiatric symptoms; the motor symptoms worsen over 10-20 years to severe impairment, rigidity, and lethality. Persons with HD have atrophy of the striatum and lose its principal neurons, striatal projection neurons (SPNs, also called medium spiny neurons or MSNs). No treatments are known to prevent or slow the progression of HD.

HD segregates in families, exhibiting autosomal dominant inheritance; the typically mid-life (post-reproductive) onset of HD has resulted in large families in which many individuals are affected and many more are at risk. The genetic cause of HD was discovered in 1993 to involve inherited alleles in which a DNA repeat within the *Huntingtin* (*HTT*) gene (CAG_n, encoding polyglutamine) had expanded beyond 36 CAGs (MacDonald *et al.*, 1993). All persons with HD have inherited an allele with 36 or more CAG repeats (36-55 in 98% of cases), versus the 15-30 repeats present on common *HTT* alleles (Gusella, Lee and MacDonald, 2021). Longer repeats associate with earlier HD onset: inherited *HTT* alleles with 40 CAG repeats associate with motor symptom onset at an average age of 60 years, while inherited alleles with 46 repeats associate with motor onset at 40 years.

It is unknown why stretches of 36 or more CAGs lead to HD, though a natural hypothesis has long held that polyglutamine stretches >36 residues in length cause continuous, cumulative damage to which SPNs are, for unknown reasons, particularly sensitive, and that longer polyglutamine stretches are more damaging than shorter ones.

Three profound scientific mysteries about HD involve its cell-type-specific pathology, its typical onset in mid-to-late life, and the unknown events by which inherited alleles lead to neurodegeneration.

HD pathology is cell-type-specific: Most SPNs are lost, while nearby striatal interneurons and glia survive without apparent degeneration. Since all these cells express HTT, the cell-type-specific pathology in HD has long presented a central mystery.

HD onset is typically preceded by decades over which persons who have inherited HD-causing alleles are healthy (Tabrizi *et al.*, 2022). Persons who inherit the more common HD-causing alleles (38-48 repeats) generally reach adulthood with normal striatal volumes by neuroimaging and normal scores on cognitive and motor tests; sensitive neuroimaging and fluid biomarkers begin to detect subtle changes only about a decade before the onset of motor symptoms (Tabrizi *et al.*, 2022).

It has been challenging to discover how inherited repeat-expansion alleles lead to pathology in HD. The *HTT* gene is expressed in almost all cell types and tissues, and the encoded protein (HTT) has many biological functions. Loss, over-expression, and genetic manipulation of *HTT* produce diverse phenotypes in many species and cell types; this has made it hard to identify which biological process drives disease. A diverse range of biological hypotheses are still considered plausible for HD, with recent studies focusing on embryonic development (Braz *et al.*, 2022), mitochondria (Lee *et al.*, 2020), vascular cells (Garcia *et al.*, 2022), microglia (Wilton *et al.*, 2023), and long-range circuitry effects (Pressl *et al.*, 2024). However, the specific pathophysiological process that brings about HD in humans has remained mysterious.

An important clue may reside in the length of the disease-causing CAG_n repeat in *HTT*. Longer inherited DNA-repeat alleles lead to earlier HD onset. The length of this repeat also exhibits mosaicism within and between the tissues of persons with HD; this mosaicism is pronounced in the brain (Telenius *et al.*, 1994; Kennedy *et al.*, 2003), is greater in neurons than in glia (Shelbourne *et al.*, 2007), and is greater in persons with earlier-than-expected HD onset (Swami *et al.*, 2009). Since these observations 17-30 years ago, the biological significance of mosaicism in *HTT* has been debated, with a common view holding that somatic expansion is an "epiphenomenon" or simply increases the inherent toxicity of inherited mutant HTT (mHTT) alleles. However, the recent human-genetic discovery of common modifier alleles, which influence the age at which HD symptoms commence (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium., 2019), supports the idea that somatic expansion may accelerate HD onset (Hong *et al.*, 2021). First, HD motor onset is delayed by a common, protective, synonymous CAG->CAA variant within the disease-causing CAG repeat; this variant reduces the repeat's instability without changing the encoded polyglutamine (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium., 2019). Second, age of HD onset is shaped by common genetic variation at many genes with roles in DNA maintenance, including *MSH3*, *FAN1*, *MLH1*, *LIG1*, *PMS1,* and *PMS2* (Genetic Modifiers of Huntington's Disease (GeM-HD)

Consortium, 2015; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium., 2019). Proteins encoded by these genes, and other proteins that form complexes with these proteins, affect DNA-repeat instability in cultured cells and/or mice (Wheeler *et al.*, 2003; Dragileva *et al.*, 2009; Kovalenko *et al.*, 2012; Pinto *et al.*, 2013; Tomé *et al.*, 2013; Kim *et al.*, 2020; Loupe *et al.*, 2020; Goold *et al.*, 2021). This has led to a hypothesis that somatic expansion accelerates HD onset in many if not all patients (Hong *et al.*, 2021).

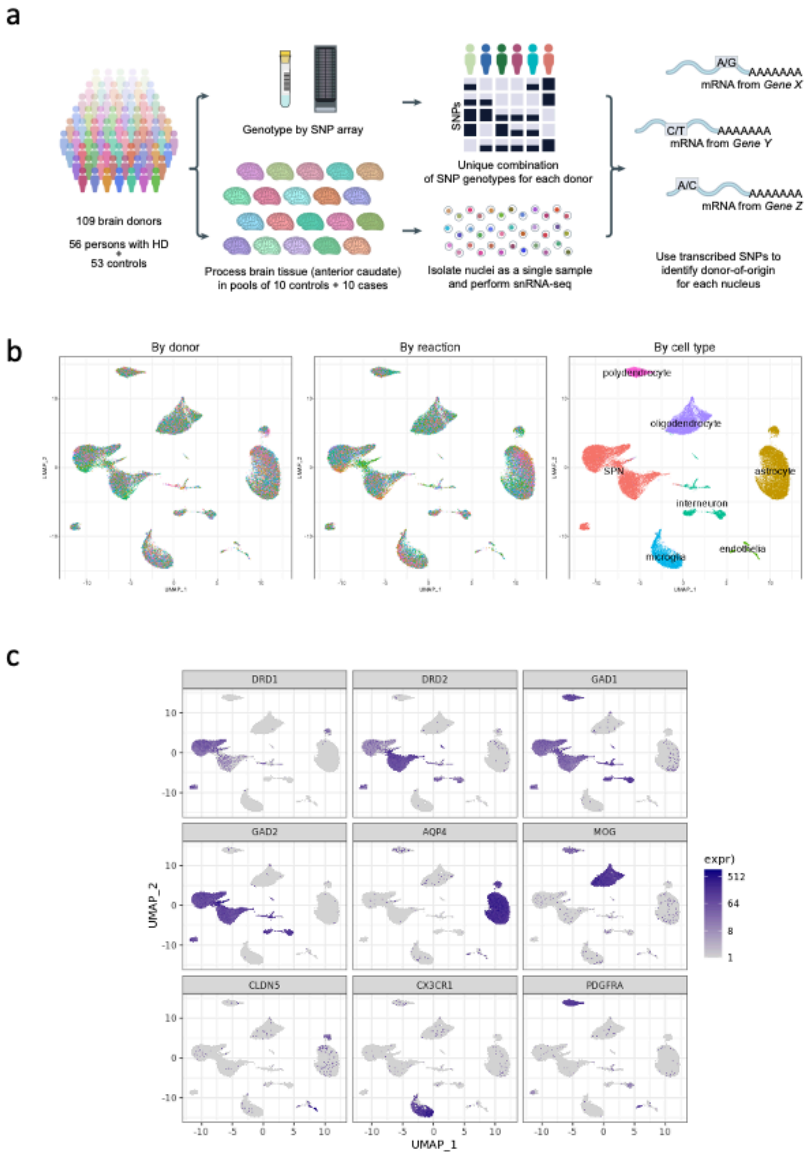
To better understand the pathophysiological process in HD, we used droplet-based single-nucleus RNA-seq (snRNA-seq) (Macosko *et al.*, 2015) to measure RNA expression in more than 600 thousand individual nuclei sampled from the caudate nucleus, the largest component of the striatum, of 56 persons with HD and 53 unaffected individuals. To relate length of the *HTT*-CAG repeat to cell types and their biological states, we then developed a laboratory method to measure its length at single-cell resolution, concurrent with the same cells' genome-wide RNA expression.

Results

Cell-type-specific vulnerability in HD

We first used conventional snRNA-seq to analyze RNA expression in 613 thousand individual nuclei sampled from the caudate from 56 persons with HD and 53 controls (mean 5,630 per donor). Each nucleus was assigned to one of seven major cell classes, based on its genome-wide pattern of RNA expression (**Supplementary Fig. 1**).

Sampling cells and RNA expression from so many persons with HD made it possible to quantify the vulnerability of SPNs in relation to each donor's age and the length of their inherited *HTT* allele. In anterior caudate from control donors, 46% (+/- 6%) of the nuclei sampled were derived from SPNs. This fraction had greatly declined in persons with HD (**Fig. 1a**).



Supplementary Figure 1. Single-nucleus RNA-seq analysis of brain tissue from 56 persons with HD and 53 controls. (**a**) "Cell village" workflow by which we perform snRNA-seq on sets of 20 donors at once. Image is only lightly modified from (Ling, Nemesh, Goldman, Kamitaki, Reed, Handsaker, Genovese, Vogelgsang, Gerges, Kashin, Ghosh, Esposito, Morris, *et al.*, 2024), where we describe this approach. (**b**) Expression profiles for the 613 thousand striatal nuclei were projected into a two-dimensional space based on similarities in their patterns of RNA expression (using UMAP), and then colored based on their donor-of-origin (left), reaction-of-origin (center), or assigned cell type (right) based on their genome-wide RNA-expression patterns. (**c**) The accuracy of these assignments can be visualized via the expression patterns of known cell-type-specific marker genes.

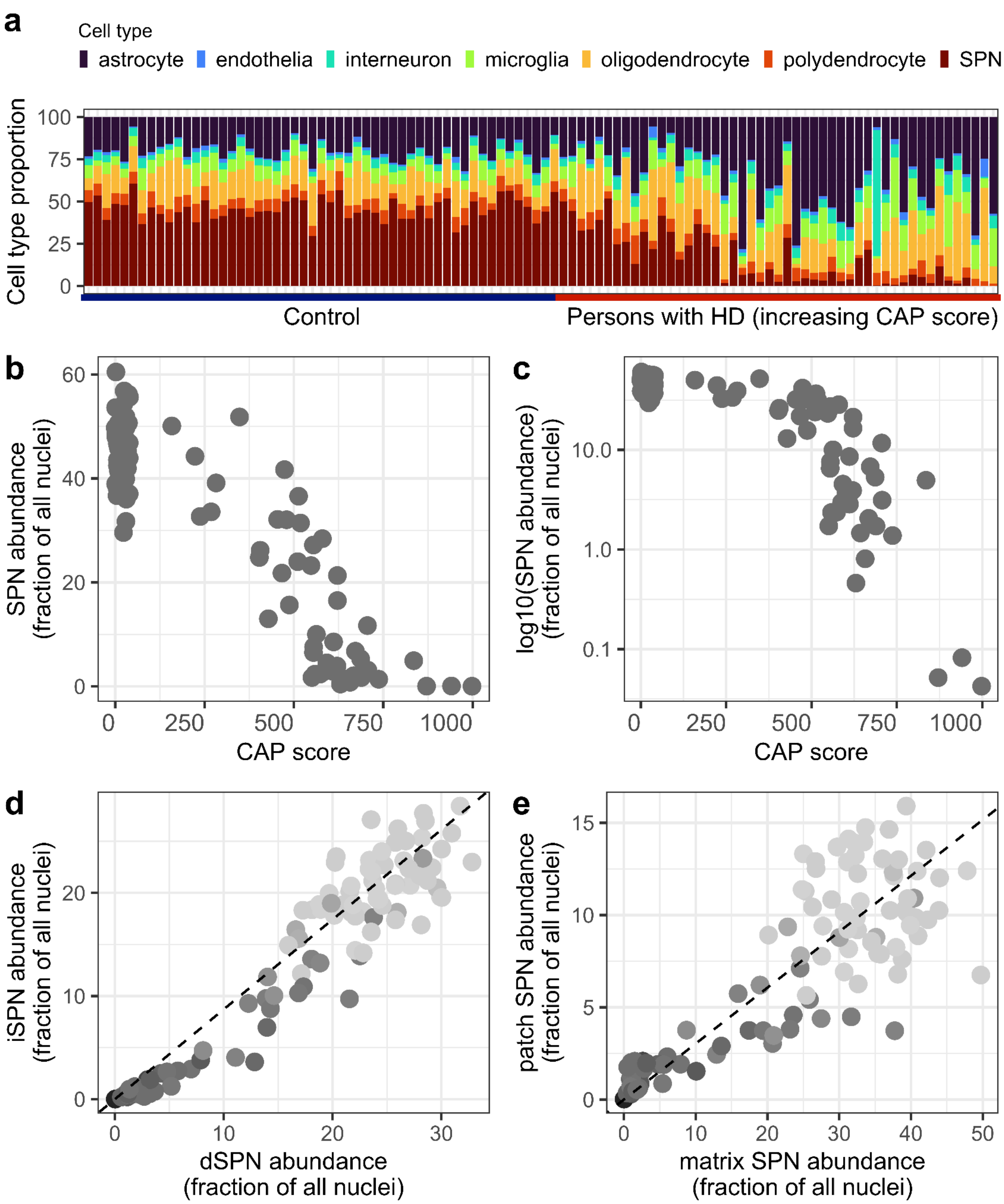


Figure 1. Trajectories of SPN loss in persons with HD. (**a**) Caudate cell-type proportions in each donor, show the loss of SPNs in persons with HD. In the figure, control donors are in the top half, and persons with HD are ordered from top to bottom by increasing CAP score. (**b,c**) Relationship of SPN loss to increasing CAP score, with SPN abundance shown on both a linear scale (b) and a log scale (c). (**d**) iSPNs (D2 SPNs) are lost earlier (on average) than dSPNs (D1 SPNs), potentially contributing to the prominence of chorea as an early HD symptom. (**e**) Patch/striosome SPNs are lost earlier (on average) than matrix SPNs.

Onset and progression in HD are estimated using CAP score, a function of age and inherited CAG-repeat length (calculated as *age* * (*inheritedCAGlength* - 33.66)). (The centrality of inherited CAG length in the formula reflects that longer inherited CAG repeats lead to earlier onset.) The use of CAP score allows persons with many different ages and inherited CAG-repeat lengths to be combined into one analysis.

As a fraction of all nuclei in the anterior caudate, SPNs declined with HD onset and progression as indexed by increasing CAP score (**Fig. 1b**). Persons with CAP scores up to 300 (corresponding to 37 years of age in a donor with an inherited 42-CAG allele) tended to have SPN proportions within the range sampled in controls (**Fig. 1b**). Loss of SPNs appeared to then greatly accelerate in donors with CAP score greater than 300, as evidenced by an increasingly steep downward slope in the relationship of SPN abundance to CAP score (**Fig. 1b).** Almost all persons with HD with CAP score greater than 600 appeared to have lost >80% of their SPNs. These results are consistent with neuroimaging findings that caudate atrophy commences about 10-15 years before the onset of motor symptoms, then greatly accelerates (Tabrizi *et al.*, 2022).

To estimate how the cell-intrinsic vulnerability of SPNs (their *rate* or probability of loss) changes over time, the abundance of SPNs can be considered on a log-scale against disease progression (**Fig. 1c**). The slope of the resulting relationship was modest before HD onset (CAP scores of 0-300), then became increasingly negative (**Fig. 1c**). This downward slope steepened with disease progression (increasing CAP score), suggesting that SPN-intrinsic vulnerability only increases.

Two canonical types of SPNs are defined by their connectivity and gene expression – direct-pathway SPNs (dSPNs) and indirect-pathway SPNs (iSPNs). iSPNs comprised 47% (+/- 6%) of the SPN population in controls, but a smaller fraction in persons with HD ($p = 8 \times 10^{-6}$, Wilcox test, **Fig. 1d**), indicating that iSPNs had tended to become vulnerable earlier (on average) than dSPNs, though in overlapping windows of time. Since iSPNs inhibit motor programs while dSPNs initiate them, the earlier average loss of iSPNs (which is consistent with stereological measurements (Albin *et al.*, 1990)) may underlie the prominence of chorea (involuntary movements) as an early motor symptom in HD (Albin *et al.*, 1990).

SPNs are also categorized based on their spatial locations to patches (striosomes) or the extra-striosomal matrix within the caudate (Graybiel and Ragsdale, 1978). Striosomal (patch)

SPNs were generally a reduced fraction of all SPNs in persons with HD (**Fig. 1e**), suggesting they were on average vulnerable earlier than extrastriosomal (matrix) SPNs, and also consistent with neuroanatomical measurements (Hedreen and Folstein, 1995). Since striosomal (patch) SPNs receive inputs from cognitive and limbic structures (such as amygdala, the anterior cingulate gyrus and orbitofrontal cortex), whereas extrastriosomal (matrix) SPNs receive more sensory and motor information (Graybiel and Matsushima, 2023), earlier vulnerability of striosomal (patch) SPNs could in principle help explain HD's early cognitive and psychiatric symptoms, which often precede motor symptoms but are less definitive diagnostically (Tabrizi *et al.*, 2013).

HTT expression

Since a longstanding hypothesis for HD pathology invokes continuous damage from lifelong exposure of cells to a toxic mutant HTT (mHTT) protein, we sought to better understand whether *HTT* expression levels could help explain the profound vulnerability of SPNs or the more-modest relative vulnerabilities of iSPNs (relative to dSPNs) or patch/striosome SPNs (relative to matrix SPNs). Expression levels 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 glia (**Fig. 3a**). *HTT* expression levels in dSPNs and iSPNs were indistinguishable (p = 0.56, paired t-test) Patch (striosome) SPNs (which appear more vulnerable than matrix SPNs, **Fig. 1e**) in fact exhibited a nominally lower *HTT* expression level than matrix SPNs did (p = 0.01, paired t-test) (**Fig. 2b**).

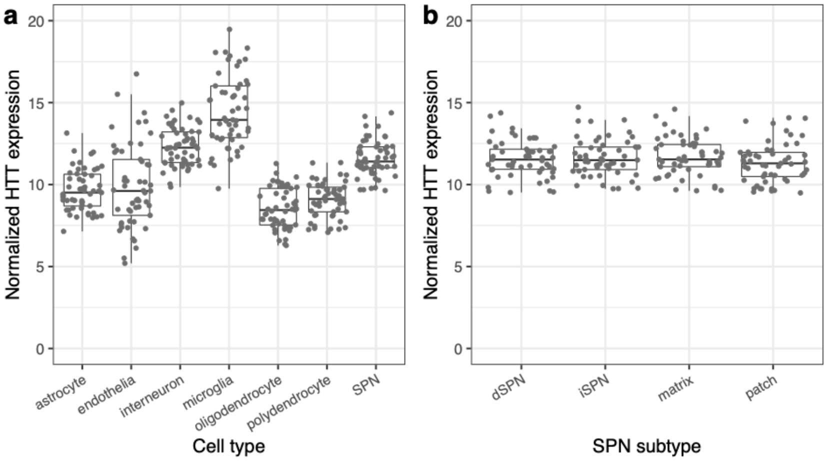


Figure 2. Expression of *HTT* transcripts (units: UMIs per 100k) in the nuclei of (a) striatal cell types and (b) SPN subtypes, among 53 control (unaffected) donors. Re.

HTT expression levels also exhibited inter-individual variation, but persons with accelerated SPN loss (relative to CAP score) were in general not those with higher expression levels of *HTT*.

Case-control differences

A conventional approach to descriptive functional genomics involves comparing gene-expression data between cases and controls to arrive at a list of "differentially expressed genes" (DEGs). We found that, even when we applied a conservative statistical approach (a non-parametric Wilcox test comparing the 53 cases to the 57 controls) to identifying differentially expressed genes, every caudate cell type – including all types of neurons, glia, and vascular cells – exhibited thousands of DEGs whose expression levels differed (on average) between cases and controls (**Supplementary Note 1**). This broadly altered gene expression in every cell type potentially reflected the profound consequences of HD, which causes atrophy of the entire caudate, death of its principal neuronal population, and greatly changed life circumstances. As described in **Supplementary Note 1**, analyses led us to conclude that the vast majority of the gene-expression changes identified from case-control comparisons are potentially secondary to neuronal loss, caudate atrophy, and reactive cellular responses by neurons and glia.

Measuring somatic CAG-repeat expansion alongside RNA expression

Since HD is caused by the DNA repeat in *HTT*, since longer repeats result in earlier HD onset, and since this DNA repeat exhibits mosaicism that could in principle be made biologically informative, we turned to investigating whether there were CAG-length-dependent cell-autonomous (inherent) gene expression changes that associated with a cell's own somatic expansion rather than with overall caudate atrophy. If CAG-repeat expansion indeed affects the toxicity of mutant HTT, then the direct effects of mutant HTT could in principle be manifest in gene expression changes in individual cells with longer repeats, relative to nearby cells with shorter repeats.

To explore this possibility, we developed a laboratory approach for sequencing the CAG-repeat of *HTT* transcripts alongside genome-wide RNA expression in the same cell nuclei. Our sn(RNA+repeat)-seq technology creates two molecular libraries from each set of nuclei: one library samples genome-wide RNA expression ("transcriptome library"), and another library specifically captures the 5' region of *HTT* transcripts ("*HTT*-CAG library") (**Fig. 3a**). 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 (**Fig. 3a**). Key aspects of creating these *HTT*-CAG libraries included the use of *HTT*-targeting primers at multiple steps of snRNA-seq; *HTT*-targeted amplification and purification steps; preservation of 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.

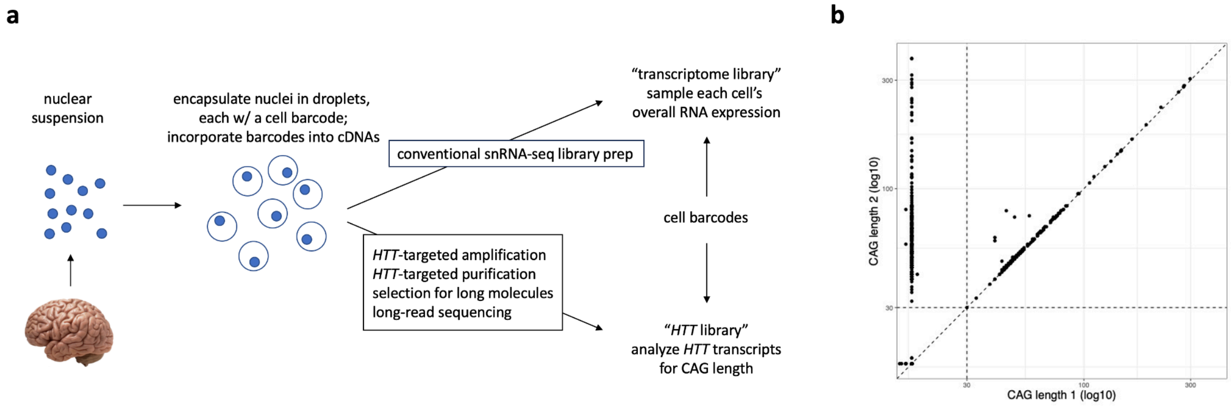
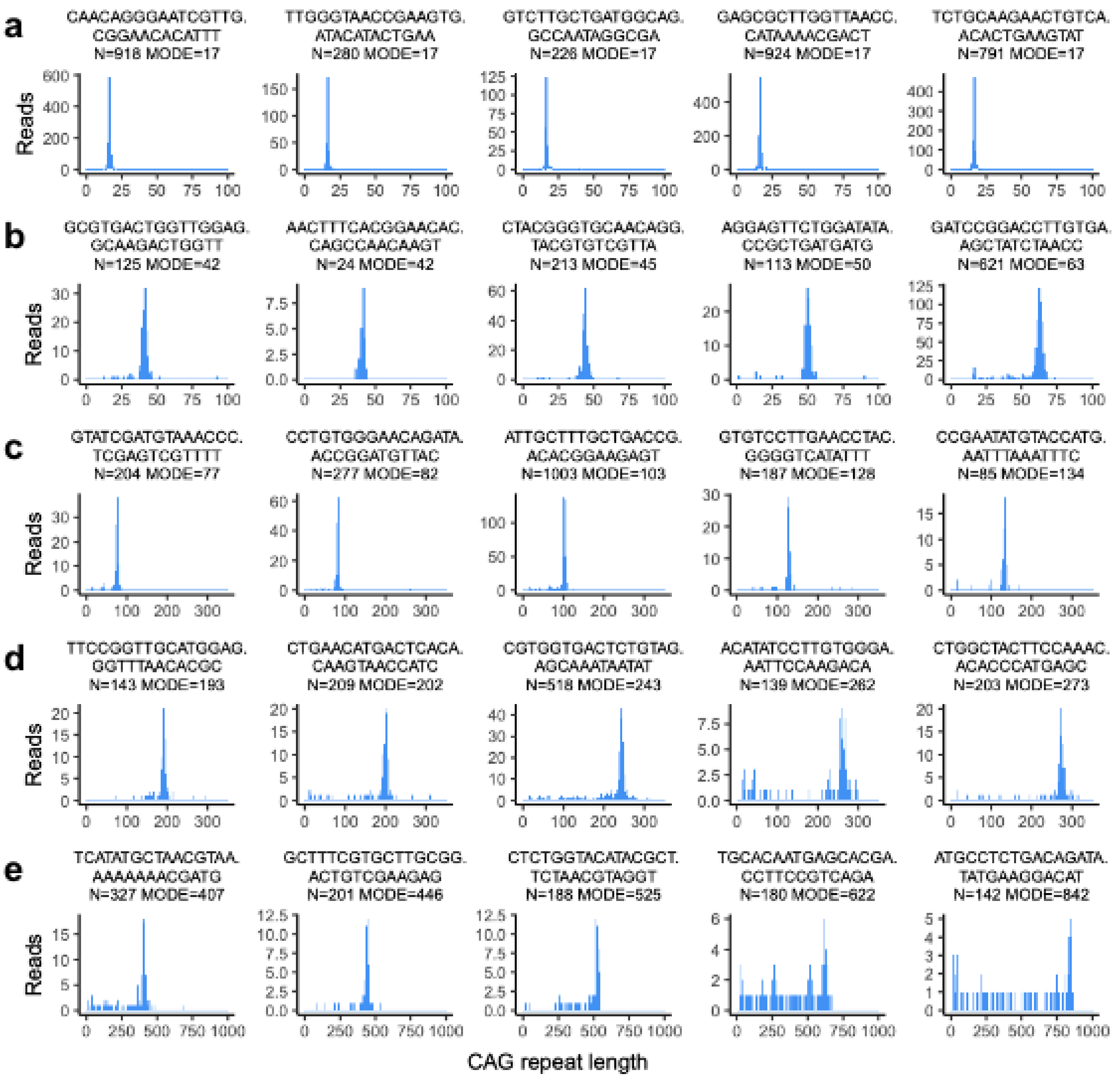

a
nuclear suspension
encapsulate nuclei in droplets, each w/ a cell barcode; incorporate barcodes into cDNAs
conventional snRNA-seq library prep
"transcriptome library" sample each cell's overall RNA expression
cell barcodes
HTT-targeted amplification
HTT-targeted purification
selection for long molecules
long-read sequencing
"HTT library" analyze HTT transcripts for CAG length
b
CAG length 2 (log10)
300
100
30
CAG length 1 (log10)
30
100
300

Figure 3. A new experimental method, sn(RNA+repeat)-seq, makes it possible to measure *HTT* CAG-repeat length concurrently with genome-wide RNA expression in the same nuclei. (**a**) Laboratory workflow. Two sequencing libraries are prepared from the same set of barcoded nuclear cDNAs. The first is a conventional snRNA-seq library. The second samples the CAG repeat in the first exon of the *HTT* gene. The presence of a shared set of 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 when sampled from different *HTT* RNA transcripts in the same nucleus. For each such measurement-pair, the longer of the two CAG-repeat measurements is shown on the y-axis. Dashed lines demarcate three apparent cases: cases in both transcripts are from the HD-causing allele (upper right); cases in which both transcripts are from the normal allele (lower left); and cases in which the two transcripts are from distinct alleles (upper left). The cases in which both measurements are from the HD-causing allele (upper right) make it possible to measure the precision and error rate of our approach.

We also developed computational approaches to analyze the data that this molecular approach produced. We found that, despite the well-known distorting effects of PCR upon DNA repeats and molecular-size distributions, sequence reads with the same cell barcode and molecular barcode (i.e. that were putatively derived from the same *HTT* transcript in the same cell) exhibited informative consensus on the CAG length of the *HTT* transcript (**Supplementary Fig. #**). When CAG-length could be measured on distinct *HTT* transcripts from the same cell, these measurements generally agreed across a wide range of CAG-repeat lengths (**Fig. 5b**).



Supplementary Fig. #. CAG-repeat length measurement from set of sequence reads derived from individual *HTT* RNA transcripts. Long reads 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 (a) for short ("wild-type") HD alleles (b) for HD-causing alleles with modest somatic expansion in the range of 40-60 CAGs (c) for longer somatic expansions in the range of 80-150 CAGs (d) for UMIs with somatic expansions in the range of 150-300 CAGs and (e) for UMIs showing very long somatic expansions beyond 300 CAGs. Note that each row uses a different x-axis scale. For transcripts with long somatically acquired CAG-repeat expansions, the PCR amplification performed in the course of library preparation creates a left-tailed distribution, favoring 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 **Fig. 3b**. 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.

Long somatic expansions are specific to SPNs

The *HTT* CAG repeat exhibited profoundly different length distributions in different cell types. Astrocytes, oligodendrocytes, microglia, endothelial cells, and interneurons exhibited modest CAG-repeat instability, in which the great majority of cells had CAG-repeat lengths within a few units of the inherited length (**Fig. 4a**). However, almost all SPNs exhibited extensive somatic expansion of the HD-causing allele, having acquired a median of ##-## additional CAG repeats in six persons with HD (**Fig. 4a**).

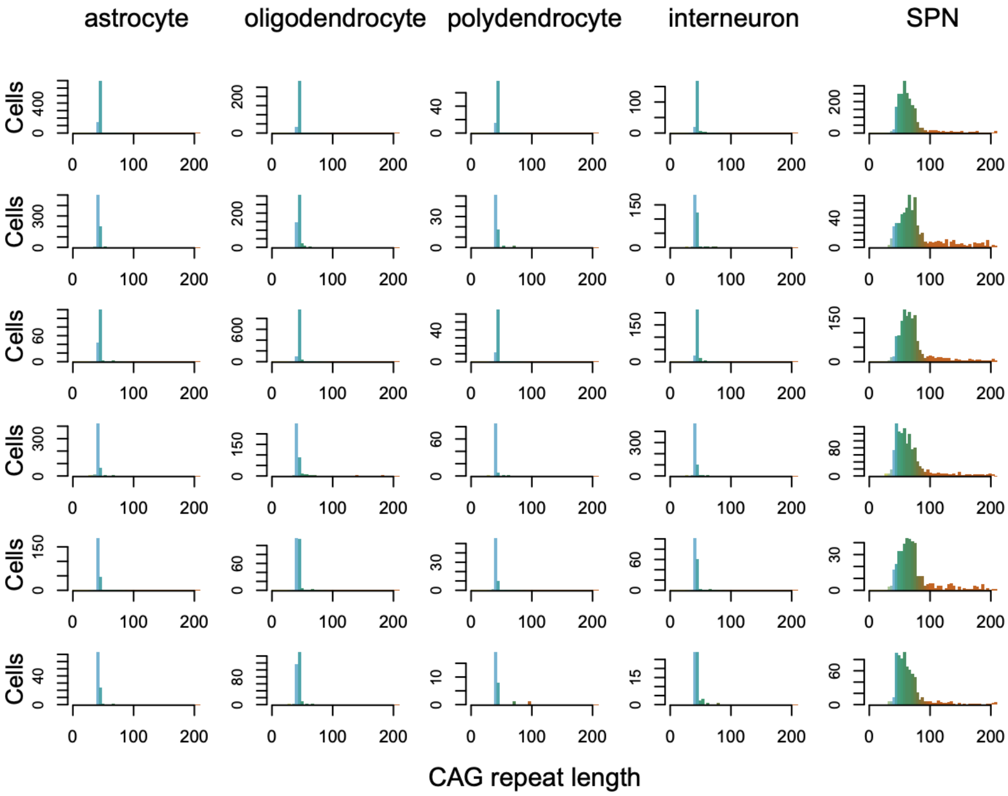
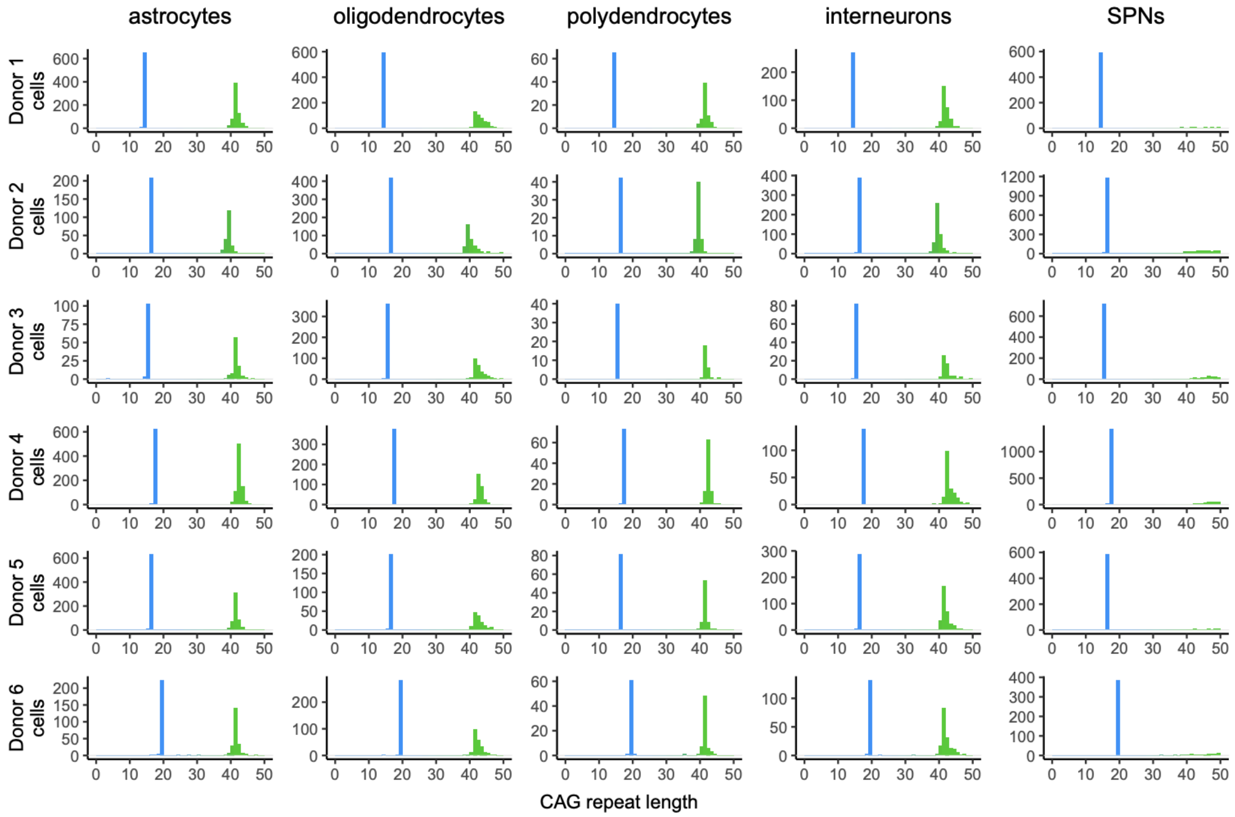


Figure 4. Distributions of CAG-repeat lengths for the HD-causing *HTT* allele in cell types of the striatum (anterior caudate). Data shown are from six persons with HD, with each person corresponding to one row of the figure.

This pattern – of modest instability in most cell types, and pervasive expansion among SPNs – was present in all the persons with HD we deeply analyzed by sn(RNA+repeat)-seq (**Fig. 4**). The distinction between SPNs and striatal interneurons was particularly notable, since all are inhibitory (GABAergic) neurons that arise from a shared developmental lineage. (Among interneurons, cholinergic interneurons exhibited more expansion than other interneurons, though we found that cholinergic neurons exhibited far less expansion than SPNs did (**Supplementary Note 2**).)

Relative to the CAG-length distributions ascertained in most sequencing studies, we detected far more molecules with long DNA-repeat expansions (**Fig. 4,5**), including many nuclei with 200-1000 CAG repeats. Notably, a 2003 study, which had utilized small-pool PCR to address the distorting effects of PCR – the tendency of PCR to under-amplify or fail to amplify longer molecules (and GC-rich sequences such as long CAG repeats) (Hommelsheim *et al.*, 2014) in competition with shorter molecules – had also identified molecules with long repeat expansions in the brain tissue of persons with HD (Kennedy *et al.*, 2003). However, much subsequent work, which used bulk PCR approaches that are vulnerable to amplification bias, had not observed these long repeats, and many studies of human HD brain have focused entirely on repeats in the 36-100 range. Our data strongly confirm the observations of Kennedy et al. (Kennedy *et al.*, 2003).

Somatic expansion appeared to be allele-specific: it strongly affected the HD-causing allele but not the other inherited allele (**Supplementary Fig. #**), suggesting that the susceptibility of an *HTT* allele to somatic expansion is regulated by its own CAG length, and consistent with the hypothesis (Hong *et al.*, 2021) that 36 CAG repeats (the threshold for an allele to be HD-causing) might be a potential threshold for appreciable somatic expansion in SPNs.



Supplementary Fig. #. Short CAG-repeat alleles (blue) that do not cause HD are somatically stable, even in persons with HD. Distributions of CAG-repeat length showing both the short (wild-type) allele and the long HD-causing allele, 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.

The distributions of SPN CAG-repeat lengths in persons with clinically apparent HD visually resembled armadillos, with a large body and a long, slowly tapering tail (**Fig. 5**). The bulk of the distribution (the armadillo's "body") indicated significant somatic expansion in almost all SPNs: the *HTT*-CAG length exhibited modes of 47-68 CAGs (6-26 repeats longer than the same donors' inherited HTT alleles), with a distribution consistent with having histories of many incremental stochastic mutations. Thus, almost all SPNs appeared to have experienced substantial CAG-repeat expansion across each donor's lifespan.

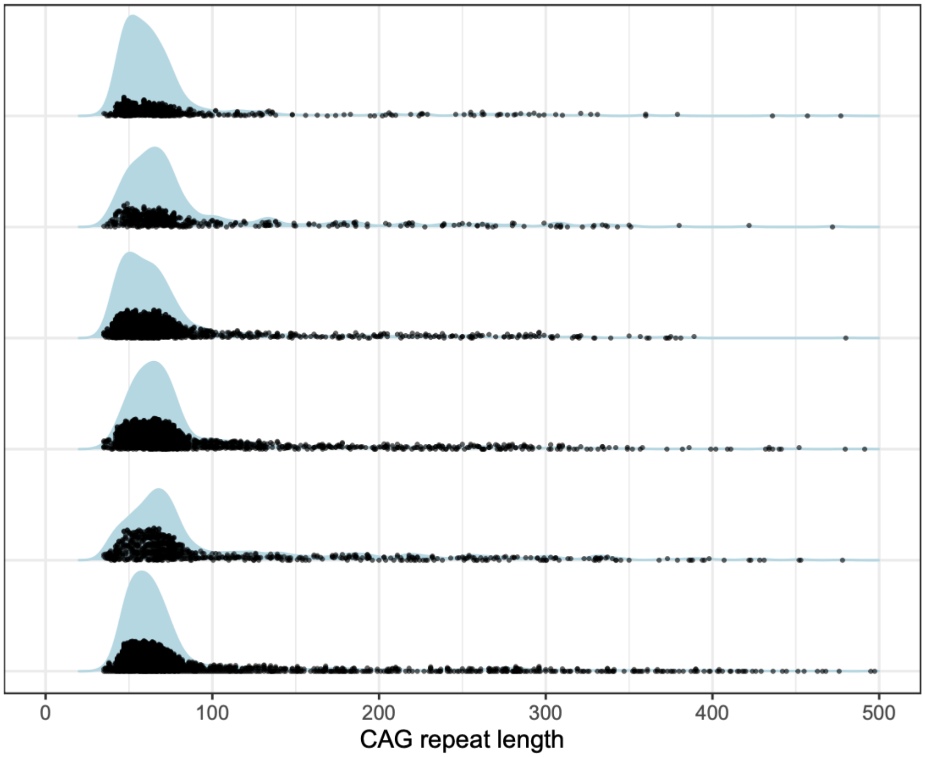


Figure 5. Length distributions for the HD-causing CAG repeat in SPNs, in six persons with HD.

The second feature (the armadillo's tail) involved a visible minority of SPNs that, at the time of analysis, had far-longer expansions (100-500+ CAGs) (**Fig. 5**). In all the donors we analyzed, this long right tail commenced at about 90 repeat units and tapered only slowly across a wide range (100 to 500+ repeat units) (**Fig. 5**). We evaluate in a later section the possibility that these two parts of the distribution – the "body" (36 to 80 repeats) and the "tail" (100 to 500+) repeats – reflect two distinct phases of somatic expansion (phase A and phase B), with the rate of expansion greatly increasing as the repeat expands beyond about 80-100 repeat units.

Biologically silent CAG-repeat expansion to 150 repeats

To recognize whether and how *HTT* CAG-repeat length affects gene expression in SPNs, we identified "allelic series" of individual SPNs within each of six donors. Each allelic series consisted of (467 to 2,337) SPNs from within the caudate of an individual person with HD, spanning a wide range of CAG-repeat lengths (35 to 842). By performing each allelic-series analyses within-person rather than across people, we controlled for the profound effects of each donor's brain atrophy (**Supplementary Note 1**), age, genetic background, and inherited *HTT* allele, enabling SPNs from the same donor's tissue sample to serve as controls for one another.

We first compared gene expression in SPNs with 35-65 CAGs to SPNs with 66-150 CAGs from the same person. Surprisingly, these SPN populations exhibited no apparent differences in gene expression (**Fig. 6a**). In contrast, SPNs with long expansions (150+ CAG repeats) differed profoundly in gene expression from SPNs with more-modest CAG-repeat expansions (40-150 CAG repeats) (**Fig. 6b**). All persons with HD exhibited a similar pattern, in which we detected no significant repeat-length-associated changes in gene expression among groups of SPNs with 36-150 CAGs, but detected profound differences in any comparisons that involved SPNs with repeat expansions longer than 150 CAG repeats (**Supplementary Fig. 2-4**).

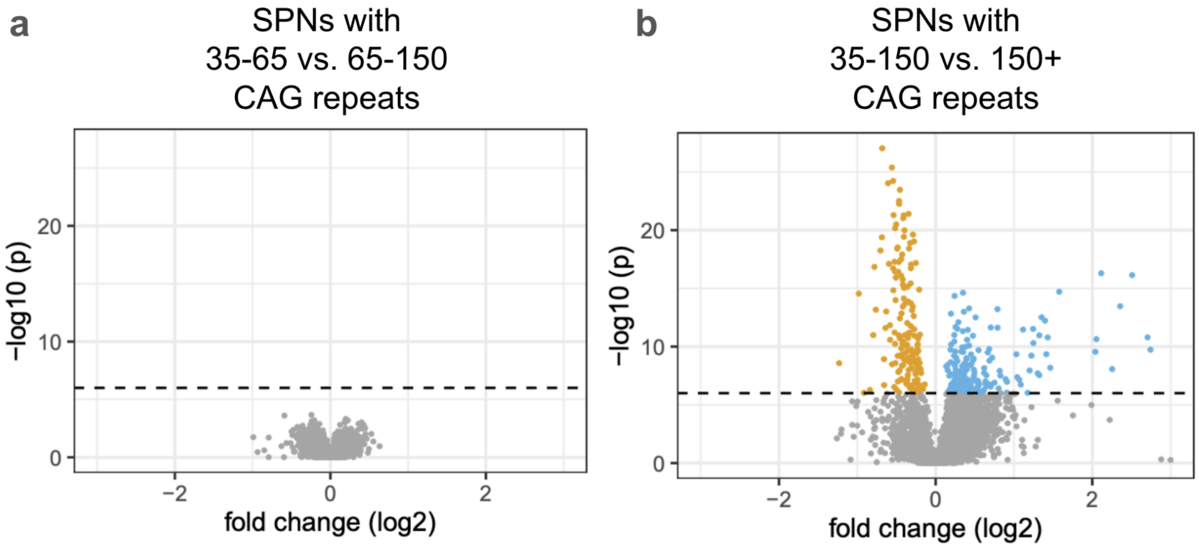
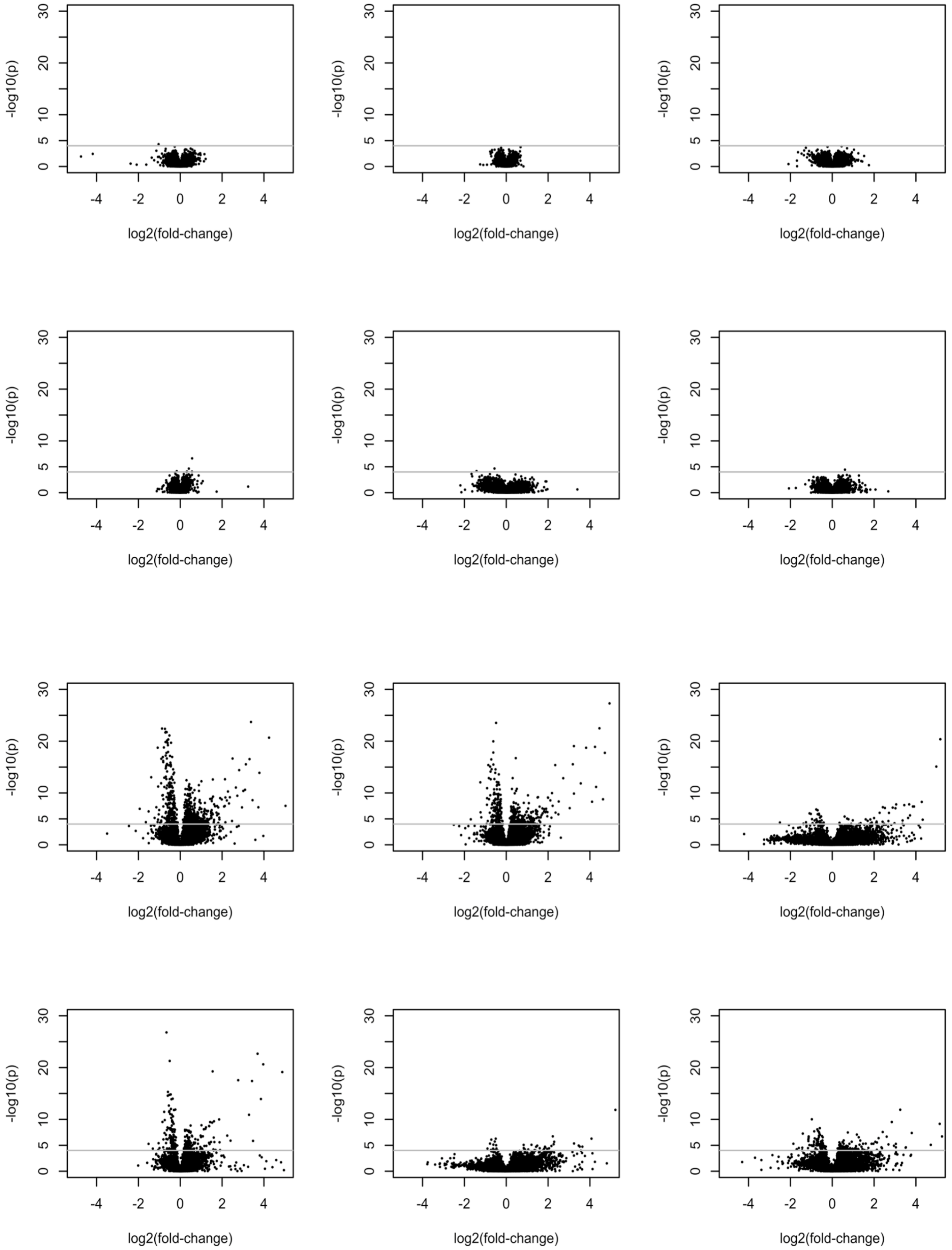
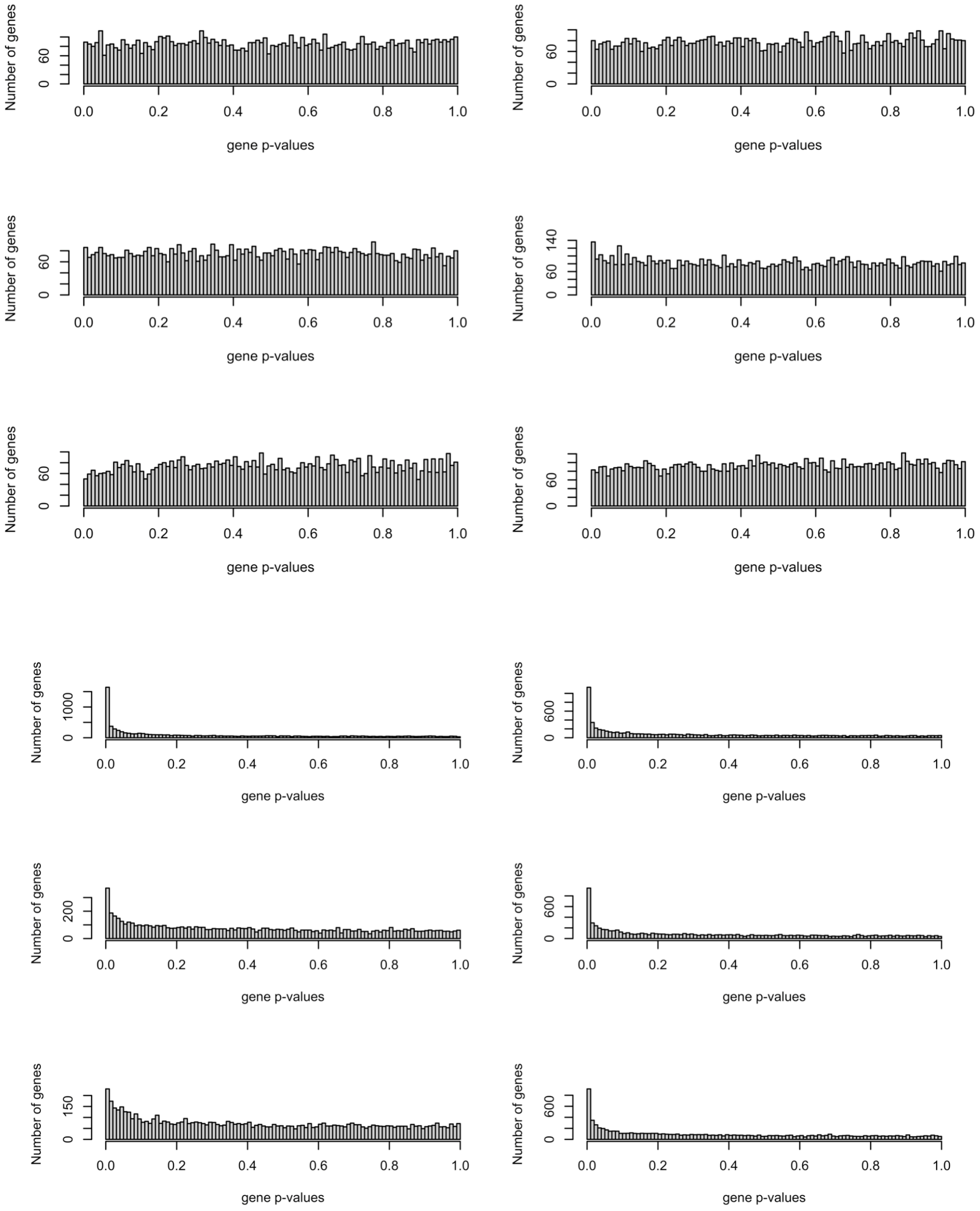


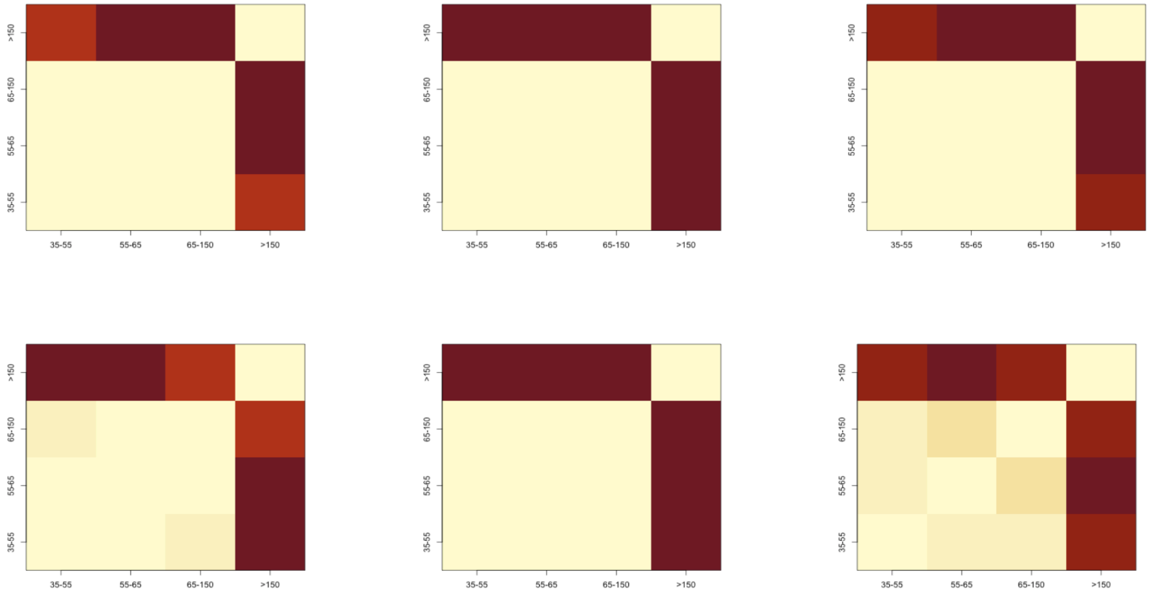
Figure 6. Comparisons (volcano plots) of gene expression between distinct sets of SPNs, sampled from the same brain area (anterior caudate) of the same person with HD but with different CAG-repeat lengths. p-values (y-axis) are derived from a Wilcoxon text across the individual SPNs in each group. Fold-changes (x-axis) are the ratio of the group medians. These and other analyses of gene expression in SPNs sampled from six different persons with HD are in **Supplementary Fig 2-4**.



Supplementary Fig. 2. (**a**) Comparisons of gene expression (volcano plots) of SPNs with 35-65 to SPNs with 66-150 CAGs, in six persons with HD. Dashed lines show the thresholds for genome-wide significance. (**b**) Comparisons of gene expression (volcano plots) of SPNs with 35-150 to SPNs with>150 CAGs, in six persons with HD.



Supplementary Figure 3. (**a**) Comparisons of gene expression (p-value distributions) of SPNs with 35-65 to SPNs with 66-150 CAGs, in six persons with HD. (**b**) Comparisons of gene expression p-value distributions) of SPNs with 35-150 to SPNs with>150 CAGs, in six persons with HD. Same donors as in Supplementary Fig. 2.



Supplementary Figure 4. Heatmaps showing differences in gene expression (one minus the Pearson correlation) between sets of SPNs defined by their CAG-repeat length, for six persons with HD (same donors as in Supplementary Fig. 2 and 3). Darker shades of red indicate larger differences (lower correlation).

Gene-expression changes with CAG-repeat expansion beyond 150 CAGs

The specific gene-expression changes that appeared in SPNs with long somatic CAG-repeat expansions were almost identical from person to person (**Fig. 7**). This indicates that *HTT*-CAG-repeat expansion beyond 150 CAGs changes SPN gene expression in a way that is both consistent across individual SPNs in the same person, and consistent across different persons with HD. (This high level of concordance from person to person, which is unusual for human post mortem studies, may be because these specific changes are the direct, cell-autonomous effects of repeat expansion – controlling for downstream sequelae of atrophy and other disease conditions that may be more specific to a donor's disease stage, age and health history.)

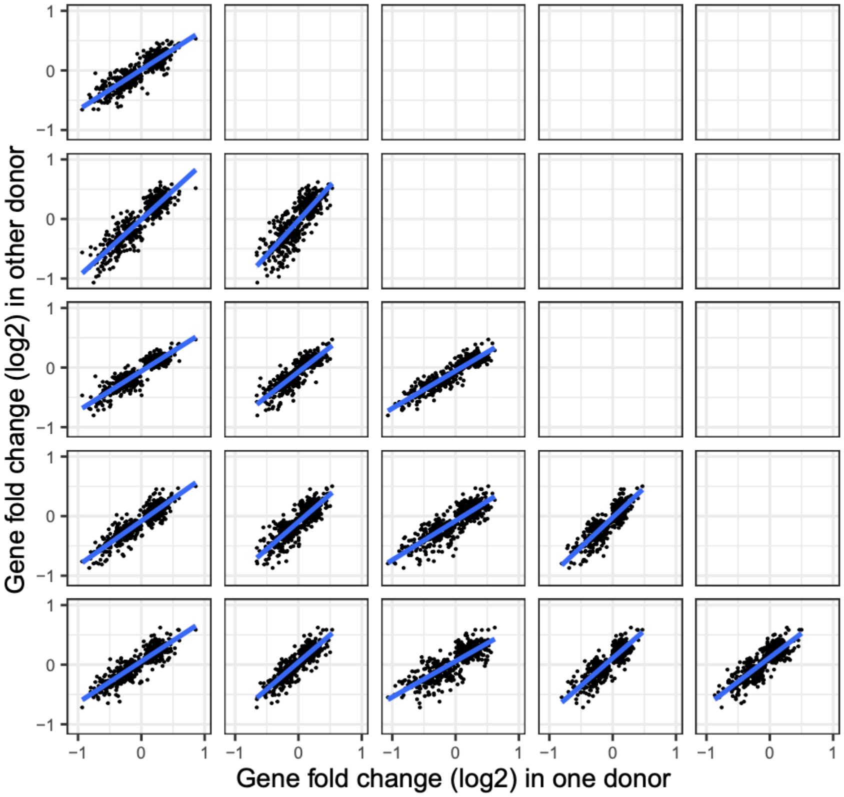


Figure 7. Consistency of long(150)-repeat-expansion associated gene-expression changes across individual persons with HD. Each panel is a pairwise comparison of SPN gene-expression data involving two persons with HD (x- and y-axes), in which the values on the two axes are plotted are the log2-fold-changes in gene expression when comparing (within-tissue) SPNs with >150 CAGs to SPNs with <150 CAGs.

Further evidence that cell-autonomous CAG-length-driven gene-expression changes arise at long CAG-repeat lengths was offered by regression analyses (using negative binomial regression), in which the expression level of each gene was fit to a combination of donor effects, SPN-subtype effects, and CAG-repeat-length effects. Hundreds of genes had expression levels whose expression levels associated with CAG-repeat length. These genes systematically showed stronger relationships to a "hinge function" – in which CAG-repeat length had no effect until reaching 150 units – than to a simple linear function in which CAG-repeat length affected gene expression across its entire range of expansion (**Supplementary Fig. #**). Notably, this analysis identified no substantial set of "dissenting" genes that associated more strongly with the linear model than with the hinge model (**Supplementary Fig. #**). The model with a hinge at 150 also out-performed models with hinges at 120 or 180 repeats (**Supplementary Fig. #**).

The nearly identical nature of these repeat-expansion-associated gene-expression changes from person to person with HD allowed us to use all of the donors' data together to identify genes whose expression levels were affected by CAG-repeat length. Even by stringent criteria, we identified hundreds of genes whose expression levels changed in a repeat-length-dependent manner (**Fig. 8a**). Notably, 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. A second set of genes exhibited discrete and dramatic changes in a specific subset of these SPNs. We further describe these two sets of genes and expression changes below.

Continuous changes with expansion beyond 150 units

More than 200 genes exhibited incipient and escalating gene-expression distortion to the extent the repeat had beyond 150 units (**Fig. 8, Supplementary Fig. 7**). We refer to this as Phase C (continuous change), and to the affected genes as C- (down-regulated) and C+ (up-regulated) genes.

Repeat-length-associated expression changes were extremely modest at 150-180 repeats, but analyses that drew upon all of the genes together indicated that these changes had commenced by about 150 repeats (**Fig. 8b**). This pattern – of the absence of any clear cell-autonomous biological change before 150 units, then progressively escalating change with expansion beyond 150 repeats – was shared across each of the individual persons with HD we analyzed (**Fig. 8b**). This pattern was also shared by direct and indirect SPNs, and by patch and matrix SPNs. Though individual persons with HD varied in the fraction of their SPNs that had attained long repeats at the end of life, all appeared to share the high threshold (of about 150 repeats) at which these same gene-expression changes had commenced in individual SPNs (**Fig. 8**).

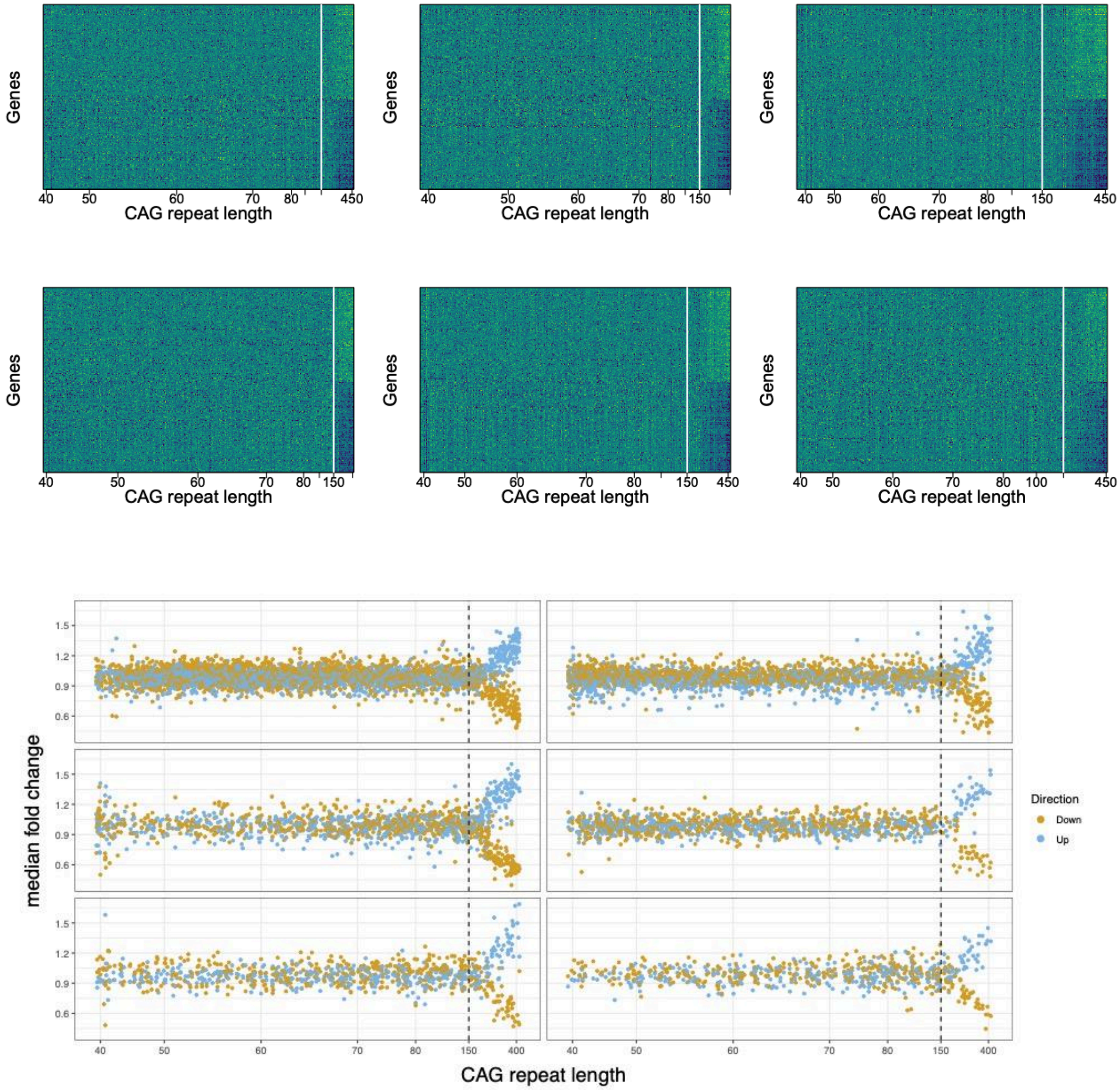
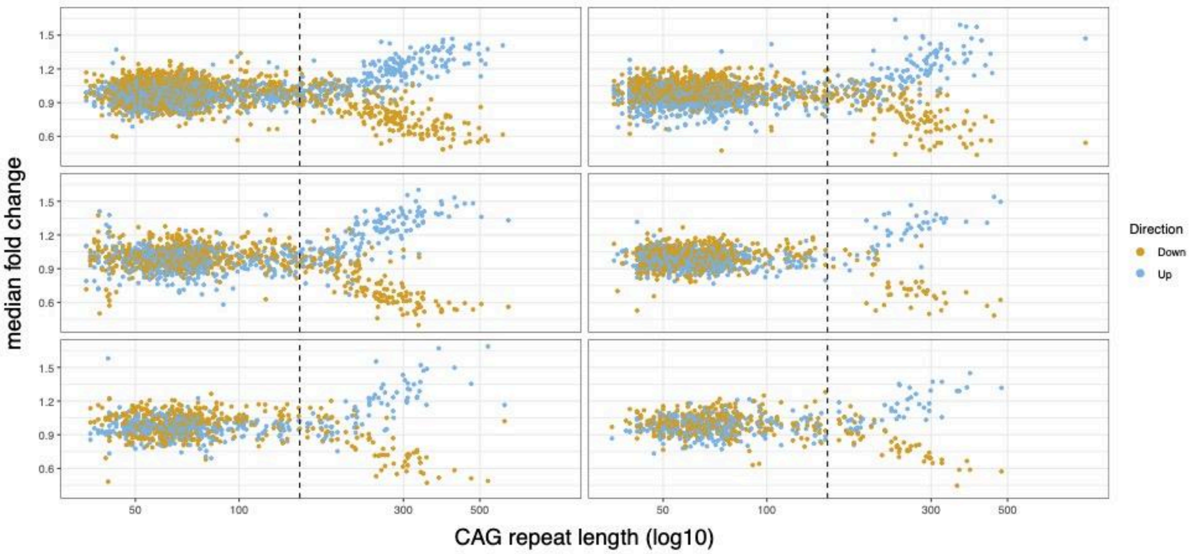
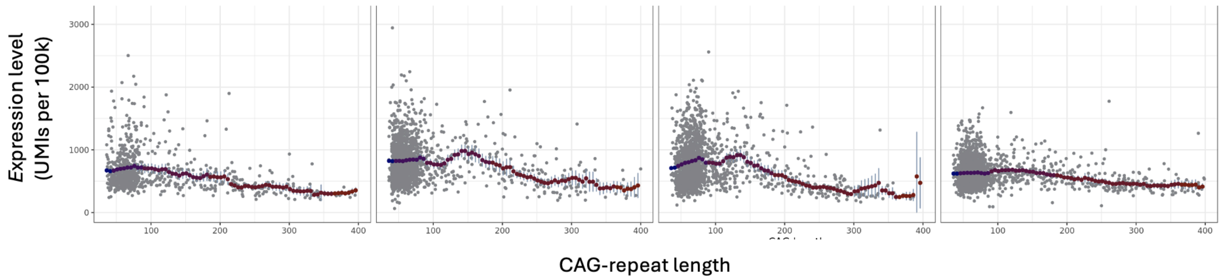
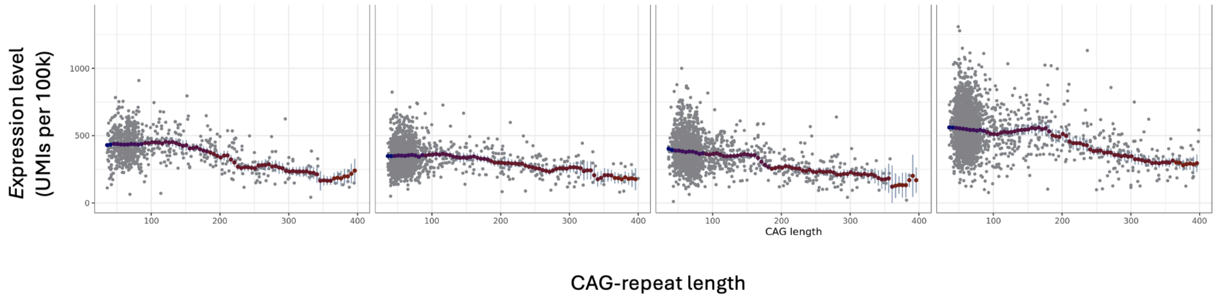


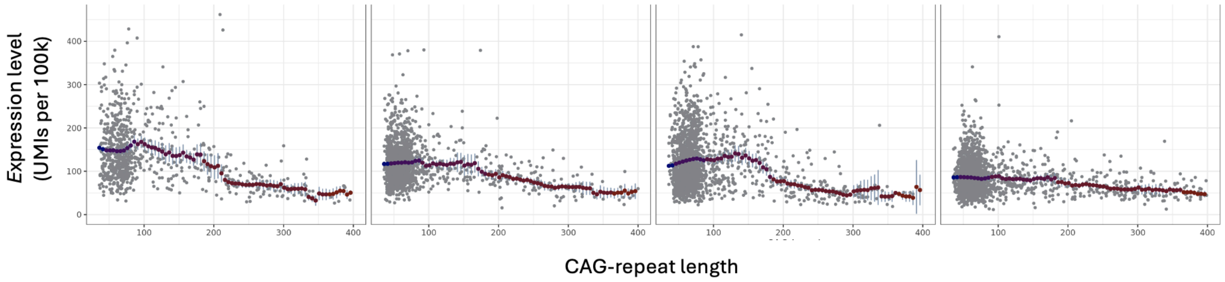
Figure 8. Gene-expression change in SPNs with somatic expansion beyond 150 CAG repeats. Each set of six plots represents gene expression data from SPNs sampled from six persons with HD. (**a**) On each plot, a specific donor's individual SPNs are ordered from left to right by their CAG-repeat length (thus corresponding to the columns of the heatmap). Each row shows single-cell RNA expression data for a specific gene in each of these SPNs. (The genes shown are genes found to change in expression concurrently with further repeat expansion beyond 150 units.) Shades of each facet show the level of expression of that gene in that SPN, relative to the average SPN in that donor (yellow: elevated expression; blue: reduced expression). **(b)** As in **a**, on each plot, a specific person's individual SPNs are ordered from left to right by their CAG-repeat length. Orange points show the average fold-change of a set of ### genes that decrease in expression with repeat expansion (C- genes). Blue points show the average fold-change of a set of ### genes that increase in expression with repeat expansion (C+ genes).

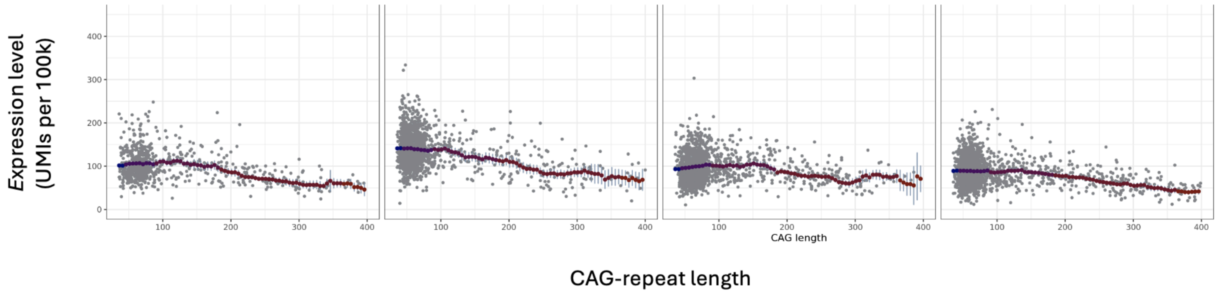


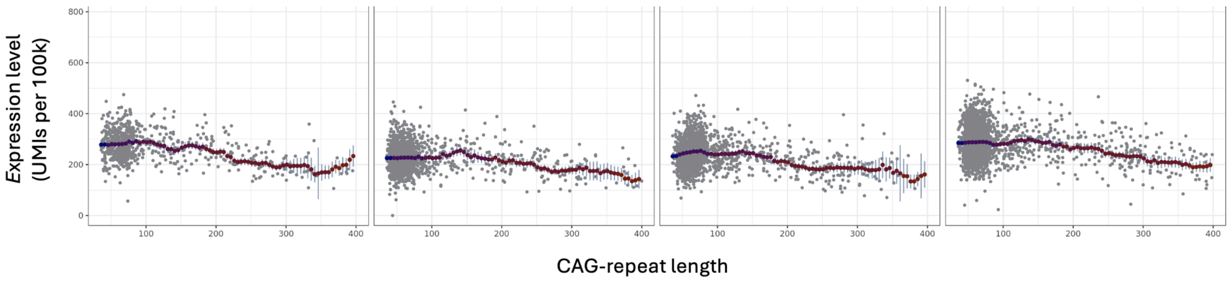
Supplementary Figure 6. As in **Fig. 8,** but here with CAG-repeat length on a log scale. Orange points show the median fold-change of a set of genes that decrease in expression with repeat expansion (C- genes). Blue points show the average fold-change of a set of genes that increase in expression with repeat expansion (C+ genes).


Expression level
(UMIs per 100k)
3000
2000
1000
0
100
200
300
400
CAG-repeat length


Expression level
(UMIs per 100k)
1000
500
0
100
200
300
400
CAG length
CAG-repeat length

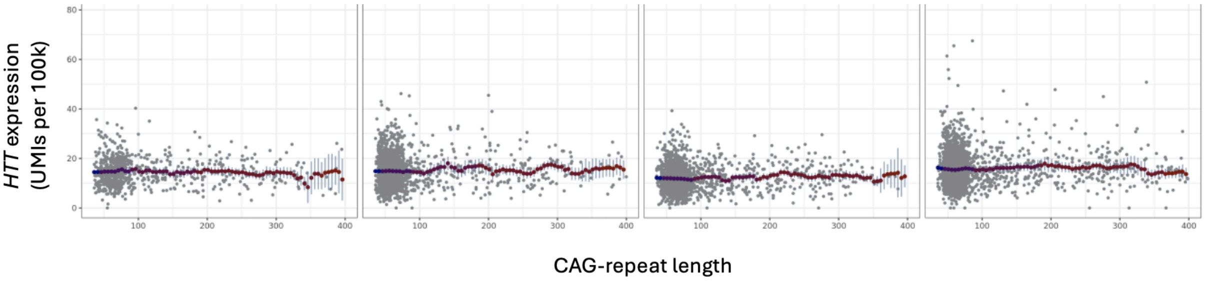

Expression level
(UMIs per 100k)
400
300
200
100
0
100
200
300
400
CAG-repeat length


Expression level
(UMIs per 100k)
400
300
200
100
0
100
200
300
400
CAG length
CAG-repeat length


Expression level
(UMIs per 100k)
800
600
400
200
0
100
200
300
400
CAG-repeat length

Supplementary Fig. 7. Expression levels of six example phase C genes (corresponding to specific rows of **Fig. 8a**), in the individual SPNs of four persons with HD whose caudate was deeply analyzed by sn(RNA+repeat)-seq. Gray points represent individual SPNs. Red bars and confidence intervals (blue) average the data in bins of ##% of repeat-length to reduce the measurement noise inherent in single-cell measurements. The genes shown are *PRKCB, PHACTR1, SLC35F3, ATP2B1*, and *CELF2*. All are genes that normal SPNs 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 are pronounced in a still-smaller fraction (those in which the CAG repeat has expanded even further) – so they appear as tiny, insignificant changes in bulk and sorted-cell-type analyses. Second, most of these genes, like most human genes, exhibit normal inter-individual variation in expression levels (evan at baseline), further obscuring (in case-control comparisons) effects that are clear in within-donor, within-tissue comparisons of individual cells.



Supplementary Fig. 8. Expression levels of *HTT*, as measured by snRNA-seq, in the individual SPNs of four persons with HD whose caudate was deeply analyzed by sn(RNA+repeat)-seq. Gray points represent individual SPNs. Red bars and confidence intervals (blue) average the data in bins of ##% of repeat-length to reduce the measurement noise inherent in single-cell measurements.

We sought to find biological patterns shared by the genes whose expression levels were affected by CAG-repeat expansion in SPNs. A profound pattern related to these genes' expression levels in SPNs relative to other types of neurons: the declining (C-) genes were among the most strongly expressed genes in SPNs, and also tended systematically to be genes that were more strongly expressed in normal SPNs than in other types of inhibitory neurons (**Fig. 9**). These included *PDE10A, PPP2R2B, PPP3CA, PHACTR1, RYR3*, and more than 100 other genes that normal SPNs express more strongly than striatal interneurons do (**Fig. 9**). This suggests that a core biological change in Phase C involves the steady, quantitative erosion of gene-expression features that distinguish normal SPNs from other kinds of inhibitory neurons – presumably, gene expression features that had been acquired late in the development and maturation of SPNs.

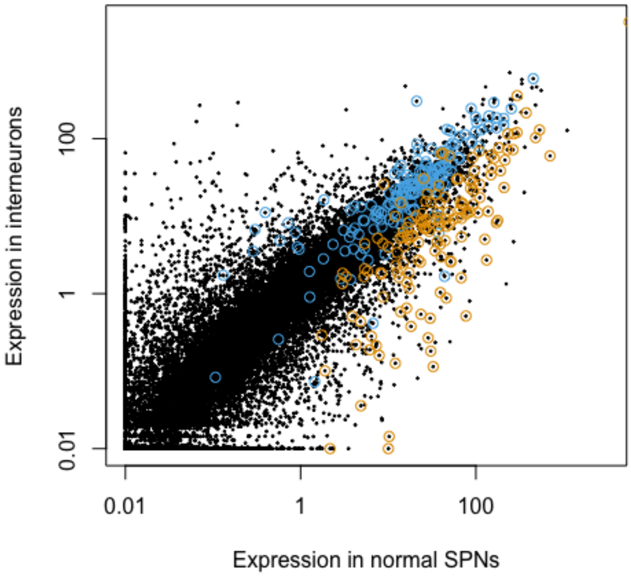


Figure 9. Gene-expression features of SPN identity and Phase-C changes. The genes whose expression levels decline in SPNs as their *HTT* CAG repeat expands further beyond 150 units (C- genes, orange circles) tend to be genes that are more strongly expressed in SPNs than in nearby striatal interneurons. Black points: all genes. Orange circles: Genes whose expression levels decline with CAG-repeat expansion beyond 150 units (C- genes), Blue circles: Genes whose expression levels increase with expansion beyond 150 units (C+ genes),

The changes that commenced at 150 repeats conceptually resembled those observed in a specific mouse HD model (Q175, with extremely long inherited CAG-repeat expansion), in which SPNs, interneurons and glia all exhibit reduced expression of genes that distinguish them from one another (Malaiya *et al.*, 2021). A notable difference is that in HD, we observed such changes only in SPNs, and only in the minority of SPNs that had acquired long expansions of the *HTT* CAG repeat. (In the Q175 mouse, all cells of all types begin life with expansions beyond 150 repeats.)

Though the primary biological property shared by the genes that declined in expression during Phase C was the way their expression distinguished normal SPNs from other cell types (**Fig. 9**), many of these genes also have important physiological functions. For example, genes encoding the potassium channel subunits KCND2, KNCQ5, KCNJ10, KCNJ16, and KCNMA1 all declined in expression during phase C, a change that might affect the physiological properties of phase-C SPNs.

HTT expression itself, as measured by snRNA-seq, did not associate with an SPN's own CAG-repeat length (**Supplementary Fig. 8**), though this does not preclude the possibility that post-transcriptional processing of *HTT* transcripts (Sathasivam *et al.*, 2013) changes in a way that snRNA-seq does not detect. *HTT* expression levels were slightly lower in the donors who had passed away with the greatest caudate atrophy (>90% SPN loss), but this decline (like thousands of other SPN gene-expression changes) appeared to be a sequela of extreme atrophy, as it did not associate with CAG-repeat length within any donor.

De-repression crisis in SPNs with even-longer CAG repeats

A distinct set of genes, that are normally repressed in SPNs, also exhibited repeat-length-dependent change, but with a different pattern. These genes had remained repressed even in most SPNs with long expansions (>150 repeats), but had become de-repressed in SPNs in which the phase-C changes had progressed the greatest degree (**Fig. 10a**). In the cells in which de-repression had occurred, this de-repression (phase D) tended to involve very many of these genes (**Fig. 10a**). The de-repression of these genes thus appeared to behave as a discrete event in which very many of the genes had become de-repressed within a short period of time.

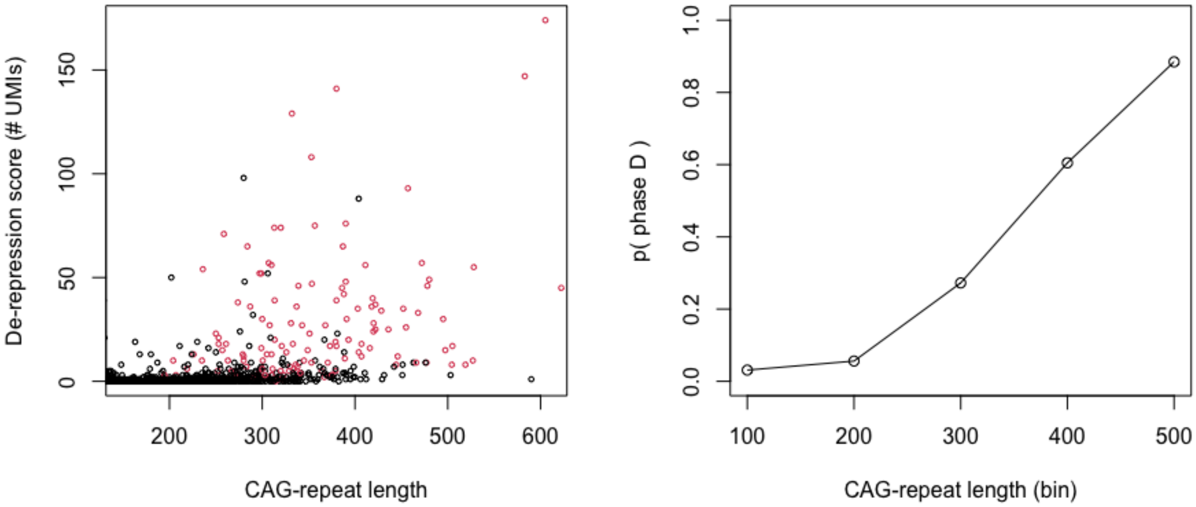


Figure 10. Widespread gene de-repression (phase D) in those SPNs with particularly long expansions and progressed phase-C gene-expression changes. (**a**) De-repression of a set of 173 genes that are normally silent in SPNs with only modest somatic expansions (40-100 repeats). Points represent individual SPNs; red points are those SPNs whose Phase-C expression changes (Fig. 9-10) have progressed beyond a threshold value. x-axis: cell's CAG-repeat length. y-axis: de-repression score, the number of transcripts (UMIs) detected from these 173 genes. (**b**) Fraction of SPNs exhibiting this de-repression phenotype, in sets of SPNs binned by CAG-repeat length (50-150, 150-250, 250-350, 350-450, 450+).

We refer to this state as "de-repression crisis" (Phase D), and to the affected genes as phase-D genes.

The likelihood that an SPN was in Phase D exhibited a strong relationship to its CAG-repeat length, and was associated with still-longer expansions than identity-softening (Phase C) was (**Fig. 10a,b**). De-repression was rare (<4%) even among SPNs with 150-250 CAG repeats, then greatly increased in frequency with further repeat expansion (**Fig. 10b**).

The 173 genes we found to be de-repressed in Phase D also had specific biological features in common. They included most genes at the HOXA, HOXB, HOXC, and HOXD loci, as well as noncoding RNAs (*HOTAIR, HOTTIP, HOTAIRM1*) at these same loci. These genes are normally expressed during early embryonic development but not in adult neurons. The de-repressed genes also included transcription-factor genes, at dozens of loci across the genome (including *FOXD1, IRX3, LHX6, LHX9, POU4F2, SHOX2, SIX1, TCF4, TBX5, TLX2, ZIC4*), that are normally expressed in other neural cell types but not in normal SPNs.

The de-repression of so many transcription-factor genes that are normally repressed in SPNs could in principle lead to expression of genes associated with other cell fates. Indeed, phase-D SPNs expressed many genes that are normally expressed in glutamatergic (excitatory) neurons (such as *SLC17A6, SLC17A7, SLC6A5*), in astrocytes (*SLC1A2*), or in oligodendrocytes (*MBP*). These changes suggested that SPNs in phase D were experiencing more-profound challenges to cell identity, and in particular were losing negative as well as positive features of cell identity.

In addition to widespread transcription associated with other neural cell types, two of the most strongly de-repressed genes were *CDKN2A* and *CDKN2B*, which encode proteins (p15(INK4b) and p16(INK4a)) that promote senescence and apoptosis in many cellular contexts (Igney and Krammer, 2002; Gil and Peters, 2006; Herranz and Gil, 2018; Yuile *et al.*, 2023). *CDKN2A* and *CDKN2B* are not normally expressed in adult neurons; we detected little if any expression of these genes (<0.1 UMIs per nucleus) in interneurons, in SPNs from control donors, or in SPNs with modest (40-150 repeats) CAG-repeat expansions from persons with HD. However, we detected substantial expression (>3 UMIs per nucleus) in phase-D SPNs. CDKN2A and CDKN2B induce senescence and/or apoptosis in many cellular contexts and play a key role in tumor suppression in the central nervous system and elsewhere (Igney and Krammer, 2002; Gil and Peters, 2006; Herranz and Gil, 2018; Yuile *et al.*, 2023); ectopic expression of Cdkn2a is toxic to adult neurons (Finneran *et al.*, 2023). The de-repression of *CDKN2A* and *CDKN2B* in phase-D SPNs could in principle be an imminent cause of their death.

A similar constellation of genes – including mouse orthologs of the same Hox genes, other transcription factors, and *Cdkn2a* and *Cdkn2b* – is de-repressed when components of the Polycomb Repressor Complex 2 (PRC2) are inactivated in adult mice (von Schimmelmann *et al.*, 2016). Inducible inactivation of PRC2 in adult mice leads within months to SPN loss, decline in motor functions and lethality (von Schimmelmann *et al.*, 2016).

SPN loss coincides with the appearance of long-CAG-repeat SPNs

The above results suggested that the transcriptional changes in long-repeat (phase C,D) SPNs might closely precede the loss of these same SPNs. Though we cannot observe the same human 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 progression (as in **Fig. 1**) in order to bring donors with a variety of ages and inherited CAG-repeat lengths into a single analysis.

Across HD onset and progression (as indexed by CAP score), the rate of SPN loss – the slope of the decline in SPN abundance – reflected closely the fraction of donors' SPNs that were in Phase C or D at these same HD stages (**Fig. 11a**), consistent with the hypothesis that phases C,D precede cell death – an interpretation that is also consistent with our observation that SPNs with long expansions (and associated transcriptional pathology) do not accumulate despite the clear population dynamic of directional repeat expansion in SPNs.

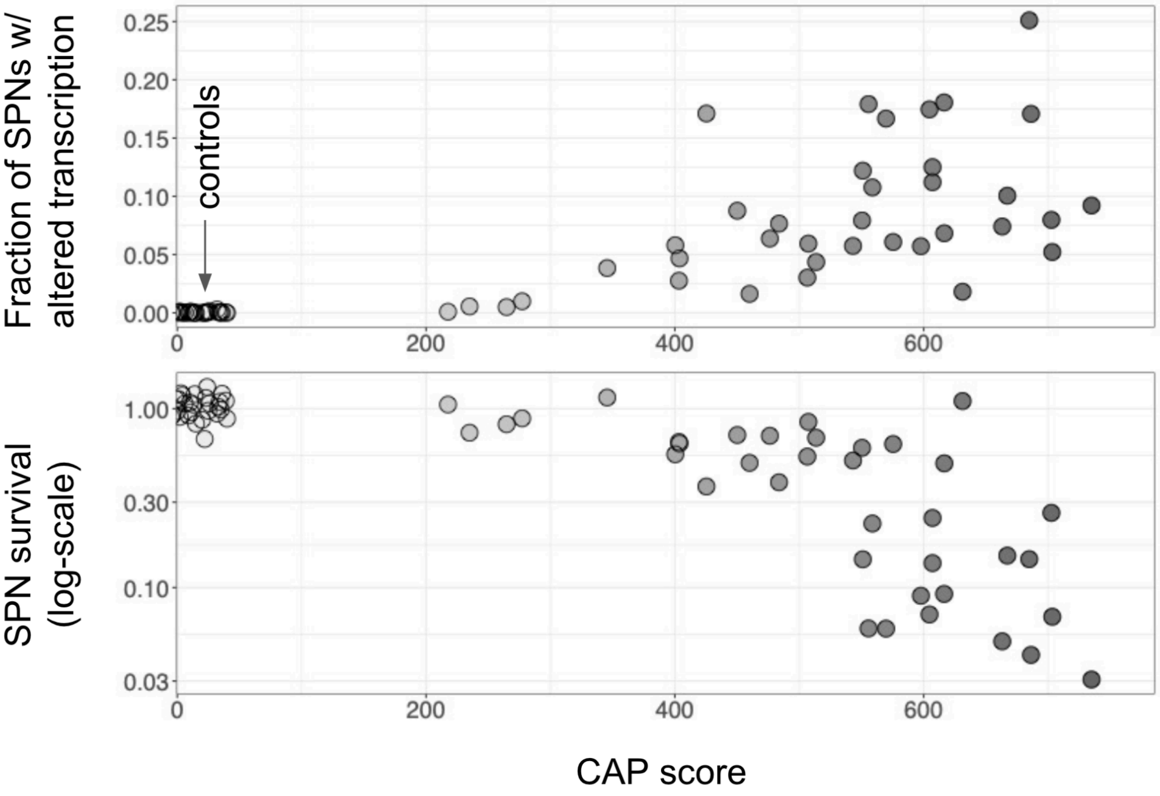


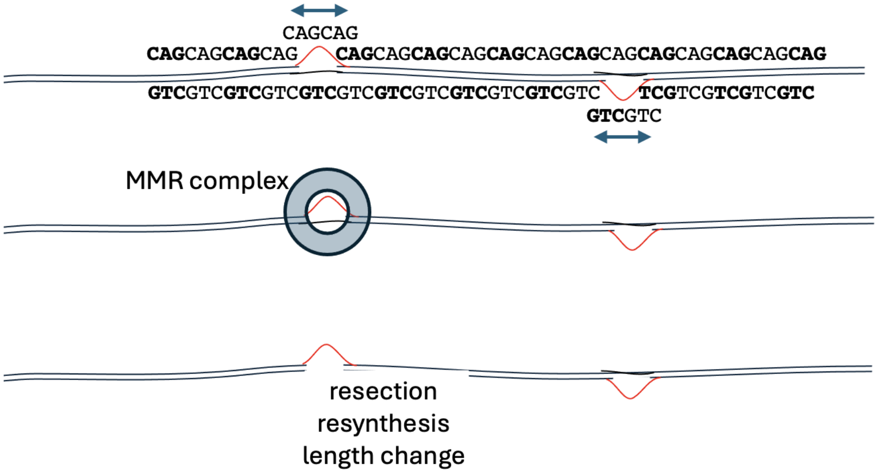
Figure 11. (**a**) Relationship to CAP score (across donors) to the fraction of each donor's SPNs that exhibited transcriptional signatures of pathological CAG-repeat expansion (phases C or D). Data from 57 control (unaffected) donors are shown on the left side of the plot with CAP scores near zero. (**b**) Relationship of SPN survival (estimated from the fraction of all nuclei sampled that were SPNs) to HD onset and progression as indexed by CAP score.

Could key relationships in HD arise from DNA repeat-expansion dynamics and a high toxicity threshold? Insights from computational modeling

Our findings were hard to reconcile with conventional biological models of HD, in which most or all SPNs endure a toxic mutant HTT *simultaneously*. Could an alternative model of *sequential* SPN toxicity – in which, at any one time, most SPNs have a benign HTT whose CAG repeat length is far below a high toxicity threshold – plausibly explain the relentless loss of SPNs in HD (**Fig. 1**)? Could the armadillo-shaped distributions of CAG-repeat lengths in SPNs arise in a simple way from known mutational mechanisms (without extravagant assumptions about single-cell heterogeneity in mutation rates)? Could the decades-long symptom-free period that typically precedes HD be reconciled with the subsequent, fast loss of SPNs (**Fig. 1**) without invoking systemic (non-cell-autonomous) disease-accelerating mechanisms? Could somatic-expansion dynamics (rather than relative toxicity) explain the well-known association of early HD onset with longer inherited DNA repeats?

To address these and other questions, and to better appreciate the dynamic processes that might give rise to clinical observations and to end-of-life single-cell DNA and RNA measurements, we turned to computational modeling of repeat-expansion dynamics over the human lifespan, seeking to understand whether simple models that adhered to an emerging understanding of DNA repeat-expansion mechanisms (Iyer and Pluciennik, 2021; Phadte *et al.*, 2023) (**Supplementary Fig. 6**) would predict repeat-length distributions and cell-loss trajectories consistent with our experimental results.

In post-mitotic cells such as neurons, length-change mutations are thought to result from occasional strand misalignment (mispairing of repeats), which can occur after transcription or after transient helix destabilization upon nucleosome disassembly (Corless and Gilbert, 2016). Mispaired repeats create extrahelical extrusions (“slip-out” structures) (**Supplementary Fig. 6**). Small extrahelical extrusions are recognized by mismatch repair (MMR) complexes, which initiate repair pathways (Iyer and Pluciennik, 2021). The repair process involves nicking, excision, and resynthesis of one of the two strands, using the other strand as a template; if the two slip-out structures are farther apart than this excision distance, then resynthesis can result in a length-change mutation – an expansion or contraction, depending on which strand has been nicked and excised. MMR complexes appear to have a bias that leads to expansions more frequently than contractions (Phadte *et al.*, 2023). Observations from mouse and cellular models indicate that expansion of the repeat tract tends to occur in small increments (Dragileva *et al.*, 2009; Goold *et al.*, 2021).



Supplementary Figure 6. Biological model of DNA-repeat expansion in non-replicating cells (from earlier work, reviewed by (Iyer and Pluciennik, 2021), which provides the framework for out simulations – in particular, that the repeat tract changes in length via small expansion and contraction events, and that the frequency of such events (initially low) increases with the length of the expanding CAG-repeat tract.

Our simulations adhered as closely as possible to this emerging understanding. We assumed that all SPNs had the inherited *HTT* allele at the time of birth; that length-change mutations were stochastic events (expansions or contractions) that changed the repeat length by one unit; that the likelihood of mutation increased as the repeat expanded; and that SPN loss occurred among SPNs with >300 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 **Supplementary Note 3**, and available in an animated format here.

The most challenging aspect of the repeat-length data for models to explain was its armadillo shape – the simultaneous presence of a large majority of SPNs with 40-100 CAG repeats, and a small minority of SPNs with far-longer repeats (100-800+). Intriguingly, all the donors we analyzed exhibited this transition across a similar CAG-repeat length range of about 70-90 repeats (**FIg. 5, Fig. 12#**). Models in which the increase in the mutation rate was a simple linear, quadratic, or higher-order polynomial 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; our models estimated this transition as occurring over a similar repeat-length interval (70-90) in each donor (**Fig. 12#**). We note that, at this length scale, otherwise-mobile slip-out structures may with increasing likelihood be separated by an intervening histone, a configuration that might greatly increase the likelihood that they are surveilled by MMR complexes before resolving on their own.

Explaining the experimental data did not require any assumptions about 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 simply from that (i) length-change mutations were initially rare events (occurring less than once per year per cell across 36-55 CAGs), and (ii) such mutations, upon occurring, increased the probability of subsequent mutations.

A fundamental relationship in HD is the long-observed association between longer inherited alleles and earlier HD onset, a relationship that is particularly steep for inherited alleles of 36-50 repeats and thus was long thought to reflect increasing mHTT toxicity in this range. We found that our simulations also produced this relationship, but for a different reason: longer inherited alleles bypass the CAG-repeat lengths (36-42) across which somatic expansion is slowest.

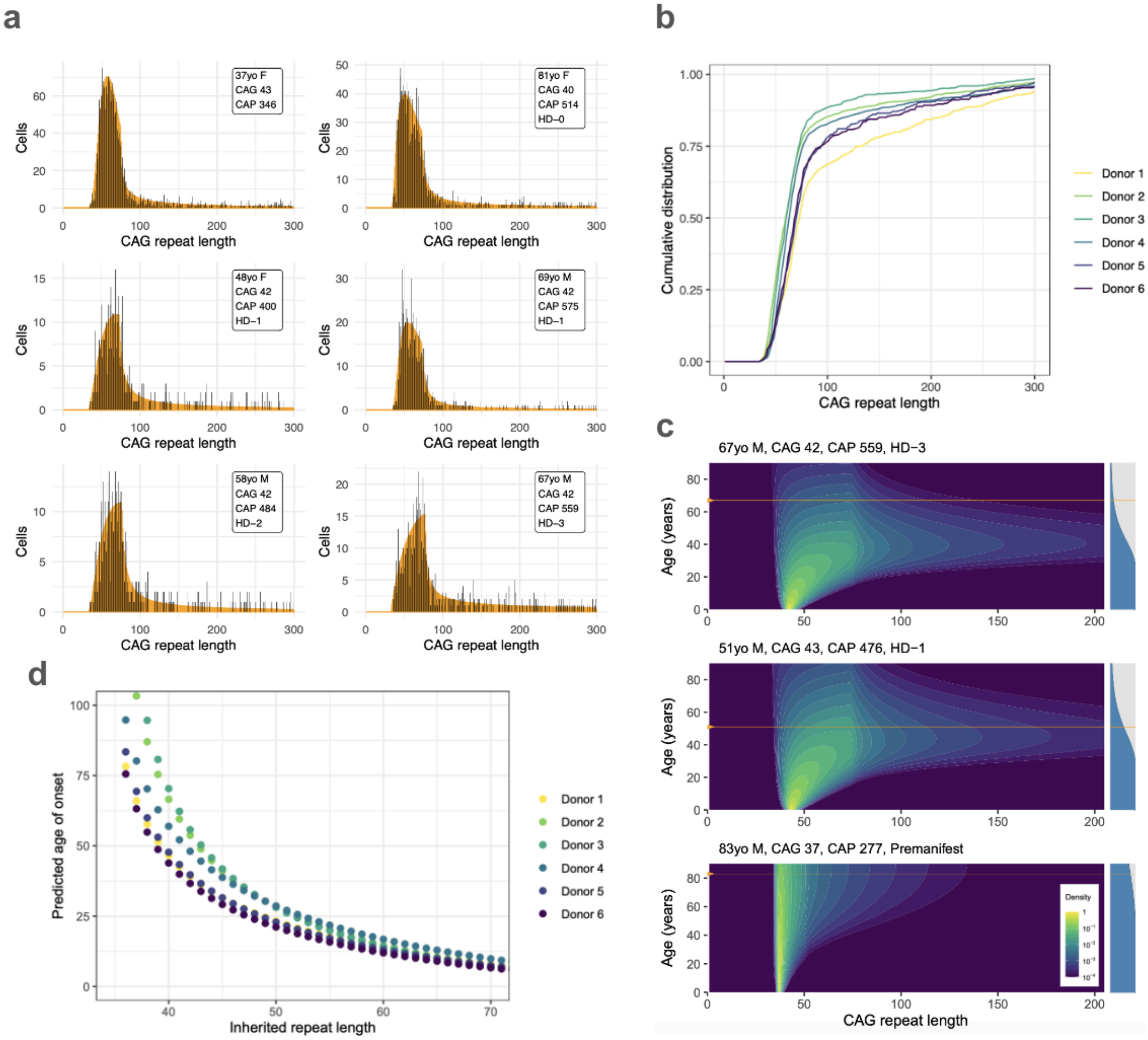

a
37yo F
CAG 43
CAP 346
81yo F
CAG 40
CAP 514
HD–0
48yo F
CAG 42
CAP 400
HD–1
69yo M
CAG 42
CAP 575
HD–1
58yo M
CAG 42
CAP 484
HD–2
67yo M
CAG 42
CAP 559
HD–3
Cells
CAG repeat length
b
Cumulative distribution
CAG repeat length
Donor 1
Donor 2
Donor 3
Donor 4
Donor 5
Donor 6
c
67yo M, CAG 42, CAP 559, HD–3
51yo M, CAG 43, CAP 476, HD–1
83yo M, CAG 37, CAP 277, Premanifest
Age (years)
CAG repeat length
Density
d
Predicted age of onset
Inherited repeat length
Donor 1
Donor 2
Donor 3
Donor 4
Donor 5
Donor 6

Figure 12 Stochastic modeling of somatic CAG repeat expansion. (a) Distributions of CAG repeat length measurements in caudate SPNs from six representative donors (black) overlaid with fitted stochastic models. (b) Cumulative distributions of the observed repeat length measurements in caudate SPNs from these same donors. (c) Representation of the predicted distribution of repeat lengths in the cells of three donors over time, based on the stochastic model fit to each donor's observed data at the time of death (orange line). Heat map on left shows the predicted density of cells with a given repeat length. Blue curve on the right of each shows the predicted fraction of SPNs remaining in that donor at each age. (d) Modeling a proxy for age of onset. As a proxy for age of onset, we used the predicted time at which 25% of a donor's SPNs have reached a repeat length of 300 or more CAGs. We estimated each donor's age of onset proxy at different hypothetical inherited repeat lengths. The shapes of the resulting curves closely approximate the known relationship between inherited repeat length and age of HD onset.

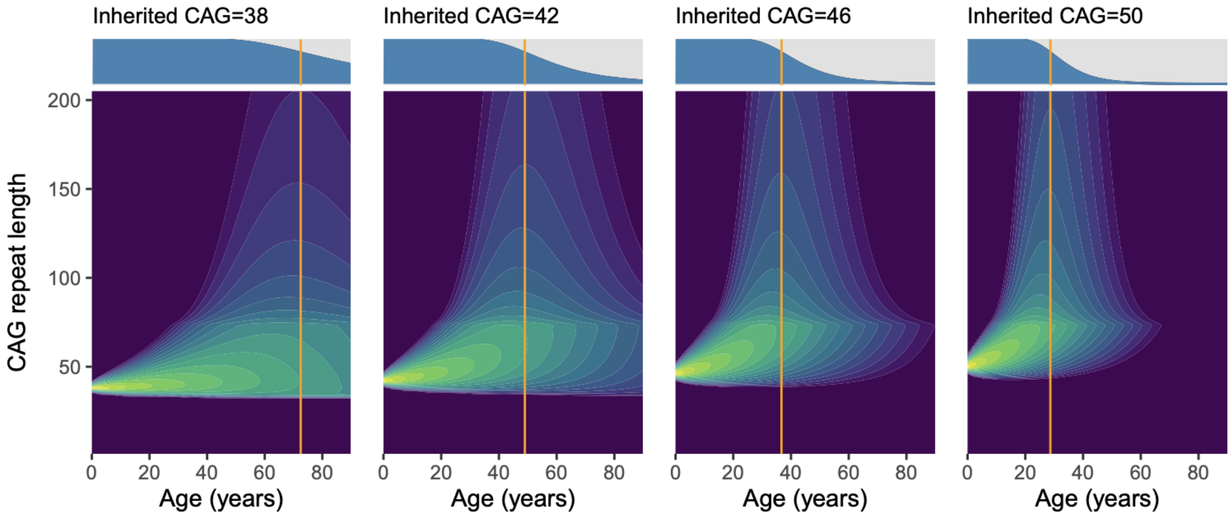
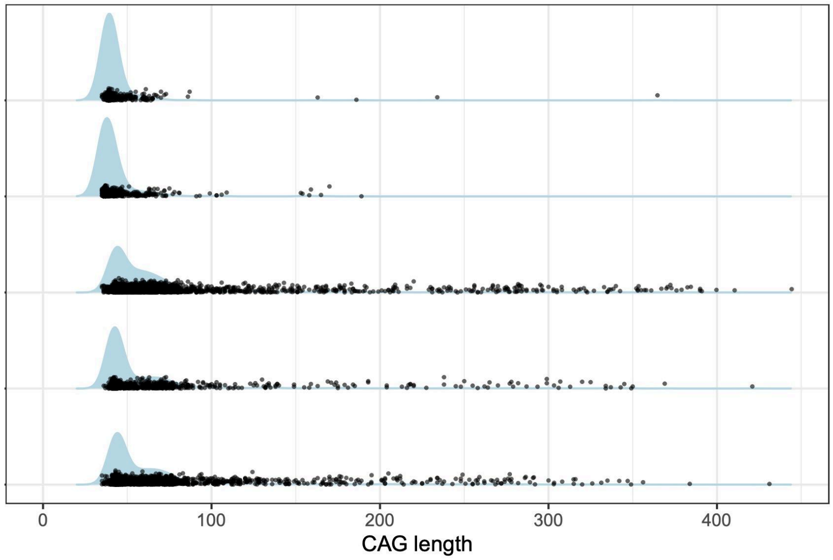


Figure 12 (new panel). The effect of adjusting the model for one typical donor (with a true inherited CAG=43) by changing the model parameter for the inherited repeat length, keeping the other fitted parameters fixed. The vertical orange line shows the age at which the model would predict 30% loss of SPNs, as a rough proxy for expected age of disease onset.

Another mystery about HD has involved the long "pre-manifest" period (generally decades) in which persons have no apparent symptoms (ISS Stage 0, see (Tabrizi *et al.*, 2022)), a period that has been attributed to slow toxicity of mHTT or to an ability of younger cells to limit this toxicity. Our simulations predicted that persons in this stage might in fact have substantial somatic expansion, but with only a small fraction of their SPNs having traversed the slow-expansion phase (phase A) and entered subsequent phases. To test this, we analyzed caudate tissue from two persons with HD who had passed away and contributed their brains for research prior to motor symptom onset. Distributions of CAG-repeat lengths in SPNs from these allele carriers appeared to confirm this prediction, as it exhibited substantial somatic expansion but very few cells with long (>100) expansions (**Supplementary Fig. 7**).



Supplementary Figure 7. SPN CAG-repeat length distributions in two HD allele carriers who passed away prior to motor symptom onset (top two rows), in comparison to three persons with manifest HD symptoms (bottom three rows).

We also found that explaining a long-puzzling feature of HD – the initially glacial, then rapidly escalating rate of caudate atrophy – did not require common assumptions of a non-cell-autonomous disease-escalating process (such as inflammation or spreading prions). Rather, the escalating period 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.

Our simulations suggest that the average SPN in a person with the most common HD-causing inherited allele (42 repeats) spends 96.4% (s.d. 2.0%) of its life below the threshold of 150 CAG repeats at which distorted gene expression appears to commence – i.e., with a biologically benign mutant *Huntingtin* gene.

A four-phase model for SPN pathology in HD

These results and analyses point to a model for HD in which the somatic expansion of a neuron's own CAG repeat is necessary and sufficient for its pathology – and is consequential only upon becoming extremely long (>150 repeats). Based on these experiments and analyses, we propose a SLEAT model (somatic long expansion, asynchronous toxicity) for the pathology of SPNs in HD, involving a series of phases, each driven cell-autonomously by each cell's own expanding *HTT* allele (**Fig. 13, Table 1**).

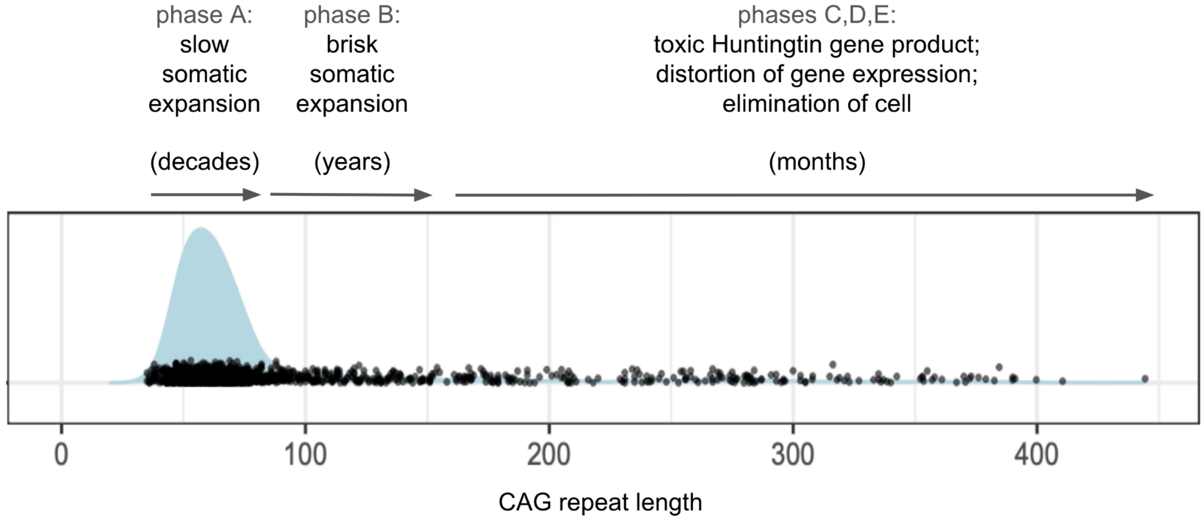


Figure 13. A SLEAT model (somatic long expansion, asynchronous toxicity) we propose for striatal neuropathology in HD. In this model, individual neurons pass asynchronously through five key pathological phases. Neurons in phases A and B have an *HTT* gene product that is benign in the sense that it has no apparent cell-autonomous effects on gene expression. Because progression through phase A takes decades and involves stochastic length-expansion mutations that then increase the probability of subsequent such mutations, individual neurons pass asynchronously from phase A to subsequent phases, in which expansion is faster, more predictable, and leads quickly to subsequent phases. Individual neurons thus experience a toxic HTT gene product at very different times. Individual neurons spend >95% of their lives in a long period of biologically silent DNA-repeat expansion (a “ticking DNA clock”, phases A and B). These phases are further described in Table 1.

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 in the same tissue)		
					Genes affected	**Dynamic**	**Functional theme**
A	slow repeat expansion	36 to 80	Slow (initially <1/yr) but increasing	Decades	–	–	–
B	brisk repeat expansion	80 to 150	Fast	Years	–	–	–
C	continuous change	> 150	Fast	Months	Hundreds	Gradual, escalating	Erosion of cell identity
D	de-repression	>> 150	Fast	Weeks	Hundreds more	Discrete, sudden	De-repression of genes normally expressed by other cell types
E	eliminated						

Table 1. Proposed phases in HD pathology at the single-neuron level. Time estimates are for persons who inherit the more common HD-causing alleles (36 to 50 CAG repeats).

In the first phase (phase A, when a neuron has 36 to about 80 repeat units), an SPN undergoes decades of slow-but-accelerating repeat expansion. We estimate that an SPN may takes ## years (on average) to expand from 40 to 60 repeats, than another ## years to expand from 60 to 80. The long time a cell spends in phase A can explain the mid-life manifestation of HD, the strong negative association between inherited CAG-repeat length and age at onset, and the strong tendency of common genetic modifiers of HD age-at-onset to involve genes with roles in DNA maintenance (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium., 2019). Phase A could be compared to a slowly and capriciously ticking DNA clock.

As a neuron enters the second phase (phase B, 80 to 150 repeat units), the rate of expansion greatly accelerates. Having taken decades to expand to 80 repeats, the CAG-repeat tract may now expand to 150 in just a few years. This acceleration accounts for the observation that a donor can simultaneously have modest expansion (36-80 repeats) in the great majority of neurons, and extremely long expansions (100-500+ repeats) in others – the long, slowly tapering tail of the armadillo-shaped distribution of SPN CAG-repeat lengths (**Fig. 5**). Still, as in phase A, the neuron's *HTT* CAG repeat does not appear to affect its own gene expression. Phase B could be compared to a more rapidly, predictably ticking DNA clock.

As a neuron enters the third phase (phase C, 150+ repeat units), hundreds of genes begin to change in expression levels. These changes escalate as the repeat further expands (**Fig. 8**), systematically eroding SPN-specific gene expression.

In its fourth phase (phase D), an SPN de-represses scores of genes that are normally (i) silenced in adult neurons or (ii) expressed in other neural cell types but not SPNs. Phase-D neurons also express *CDKN2A* and *CDKN2B*, which encode proteins that promote senescence and apoptosis.

In the final phase, an SPN is eliminated (phase E). Such cells do not appear in CAG-length and gene-expression data, though their earlier loss is apparent in the declining numbers of SPNs (**Fig. 1**)), and the effects of their cumulative loss are clear: in the atrophy and de-neuralization of the caudate, in a profoundly changed context for all remaining cells, and in gene-expression changes in remaining cells of all types, changes which systematically correlated with earlier caudate atrophy as estimated from earlier SPN attrition (**Supplementary Note 1**).

Importantly, individual SPNs appear to enter the fast phases (B,C,D,E) at different times, an asynchrony which 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 such events.

Though we have focused on human HD patients with typical midlife onset, this model might in principle also help explain the earlier, faster, more-synchronous, and less cell-type-specific pathology observed in widely used mouse models of HD (such as Q175 and R6/2, which start life with >150 repeats in all cells), as well as the fast disease progression observed in rare, childhood-onset HD patients who have inherited alleles with >70 repeats (i.e. close to or beyond the end of phase A).

Discussion

Three biological questions

Biological research on HD has long been animated by three questions. What is toxic to cells about the inherited alleles that cause HD? Why does this toxicity take so long to bring about neuronal loss? And why is this toxicity so cell-type-specific?

Our experiments, analyses and proposed model (**Fig. 13**, **Table 1**) suggest surprising answers to all three questions.

The surprising answer to the first question – the biological nature of the toxicity of inherited HD-causing alleles – is that such alleles may encode no inherent toxicity, and may remain benign even after decades of somatic expansion to 100 repeats (phase A in **Fig. 13, Table 1**). We found no effect of CAG-repeat length on gene expression across the wide range of repeat-lengths inherited by almost all patients (36-100), though we found profound, and likely quite toxic, gene-expression distortions at much-longer repeat lengths (above 150 units) that are attained only after decades of somatic expansion. A potential interpretation is that the apparent threshold for an inherited allele to be disease-causing (~36 repeats) reflects not that such alleles encode toxic RNAs or proteins, but that they are unstable (ref) – sufficiently unstable as to be likely to expand beyond 150 repeats within a human lifetime. Our model also suggests that the association of longer inherited repeats with earlier onset does not reflect their inherent toxicity; rather, such alleles require less time to expand to the toxicity threshold, by bypassing the CAG-repeat length range (36-42) in which somatic expansion is slowest.

The surprising answer to the second question – the reason for the apparently slow nature of the neuropathology in HD – is that once a neuron begins to experience toxicity from its own *Huntingtin* gene product, that neuron's decline may not be slow at all. Our analytical results, together with the paucity of such neurons at any one time, 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 *huntingtin* toxicity briefly and asynchronously.

We propose that the answer to the third question – the cell-type specificity of cell death in HD – is that the CAG repeat only reaches the high toxicity length threshold in certain cell types.

Apparent rates of expansion varied greatly across cell types (**Fig. #**), and SPNs were the only caudate cell type for which large fractions of individual cells appear to reach the high toxicity threshold in a human lifetime.

Therapeutic implications

The most significant 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, with approaches based on antisense oligonucleotides, splicing modulation, or gene editing (Tabrizi, Ghosh and Leavitt, 2019). 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 long and slow, then even a partial reduction in HTT production might greatly postpone HD onset or progression.

However, multiple HTT-lowering treatments – while substantially lowering HTT expression – have failed to show efficacy in HD clinical trials (*Nature Reviews Drug Discovery*, 2021, *Science*, 2021). Our experiments and analyses suggest a potential explanation: at any one time, very few SPNs may actually have a toxic HTT protein from whose lowering they could benefit (**Fig. 13**). Furthermore, once arriving at this phase of cell-biological toxicity (phases C,D in **Table 1**), a toxic mHTT may bring about neuronal loss quickly, in months rather than decades – a fast toxicity that may require a large, continuous reduction in mHTT expression to meaningfully slow.

The model that HD pathogenesis is a DNA process for >95% of a neuron's life (**Fig. 13**, **Table 1**) might suggest focusing on an alternative approach: slowing the ticking DNA clock. The finding that many of the common genetic variants that affect age-at-onset are in genes that encode DNA-maintenance enzymes (such as *MSH3*) had nominated this idea well before the current work (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium., 2019; Ferguson and Tabrizi, 2023), but much uncertainty has surrounded the therapeutic window that such a candidate approach would have. Our results suggest that the therapeutic window for a hypothetical therapy that slowed somatic expansion might be far wider than anticipated: if a cell spends 95% of its life in phase A, then even a modest slowing of somatic expansion might greatly postpone HD symptom onset for persons who have inherited HD-causing alleles.

What about persons who already have HD symptoms? Surprisingly, our results predict that, even in such patients, most future neuronal loss will occur only after future somatic expansion, since at all times, even at the end of life, most living neurons are still in Phase A (**Fig. 12, 13**). If our model is correct, a future somatic-expansion-directed therapy might be able to slow or stop HD progression even in persons who already have HD symptoms. This would allow the efficacy of such therapies 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

An exciting possibility is that the dynamic we have described here – in which an inherited allele must undergo decades of somatic expansion before acquiring toxicity – might apply in DNA-repeat disorders beyond HD. More than 40 human diseases are now known to be caused by inherited expansions of simple DNA repeats in protein-coding sequences, introns, UTRs, or promoters (Paulson, 2018; Rajagopal *et al.*, 2023). Several of these diseases involve age-associated mosaicism and mid-life onset (Monckton *et al.*, 1995; Morales *et al.*, 2020; Campion *et al.*, 2022), and many of these – including Myotonic dystrophy 1, X-linked dystonia Parkinsonism, Friedrich ataxia, and six forms of spino-cerebellar ataxia – are also (like HD) delayed or hastened by common genetic variation at DNA-repair genes including *MSH3*, *FAN1* and/or *PMS2* (Morales *et al.*, 2016; Laabs *et al.*, 2021; Rajagopal *et al.*, 2023). It will be important to recognize the biological significance of somatic expansion in these disorders. If a similar SLEAT (somatic long expansions, asynchronous toxicity) dynamic drives the onset and progression of other DNA-repeat disorders, then a therapy that slows somatic expansion could in principle prevent or slow many human diseases associated with DNA repeats.

Acknowledgments

We are grateful to the patients and families whose donations of brain tissue to the NIH NeuroBioBank enabled this work. We are grateful for support from the CHDI Foundation, the Harvard Medical School Department of Genetics, and the Harvard Ludwig Initiative in Neurodegenerative Disease. We are grateful for helpful advice and suggestions throughout this project's conception and execution from Tom Vogt, Vahri Beaumont, and Jian Chen; to Darren Monckton, for helpful advice on amplifying and sequencing DNA repeats; to Sarah Tabrizi, for helpful advice on HD; and to John Warner, for helpful advice on computational modeling.

Supplementary Note 1: Case-control RNA-expression differences

A conventional approach to descriptive functional genomics in human disease involves comparing gene-expression data between cases and controls to arrive at a list of “differentially expressed genes” (DEGs). We found that, even when we applied a conservative statistical approach (a non-parametric Wilcox test comparing the 53 cases to the 57 controls) to identifying differentially expressed genes, every caudate cell type – including all types of neurons, glia, and vascular cells – exhibited thousands of DEGs whose expression levels differed (on average) between cases and controls (**Fig. SN1.1**).

This broadly altered gene expression in every cell type potentially reflected the profound consequences of HD, which causes atrophy of the entire caudate (reduced in HD to a small fraction of its normal size), neuronal death, and greatly changed life circumstances (such as paralysis). Such changes may affect the biology of every cell type.

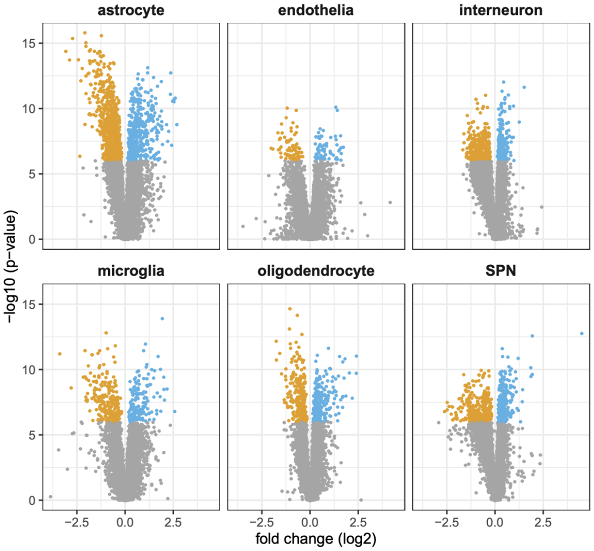


Figure SN1.1. Case-control gene-expression differences. Volcano plots representing case-control differences in gene expression in each cell type. Median fold-changes (x-axis) and p-values (y-axis) are based on comparison of 53 persons with HD to 57 controls, using a Wilcox (non-parametric) test. Blue points: genes with higher average expression levels in HD; orange points: genes with reduced average expression levels in HD.

The expression levels of these same genes (case-control DEGs) correlated strongly, in an HD-only analysis, with donors’ earlier caudate atrophy (**Fig. SN1.2**) as measured by SPN survival (estimated as in **Fig. 2**). This was true of SPN gene expression as well as gene expression in other cell types (**Fig. SN1.2**).

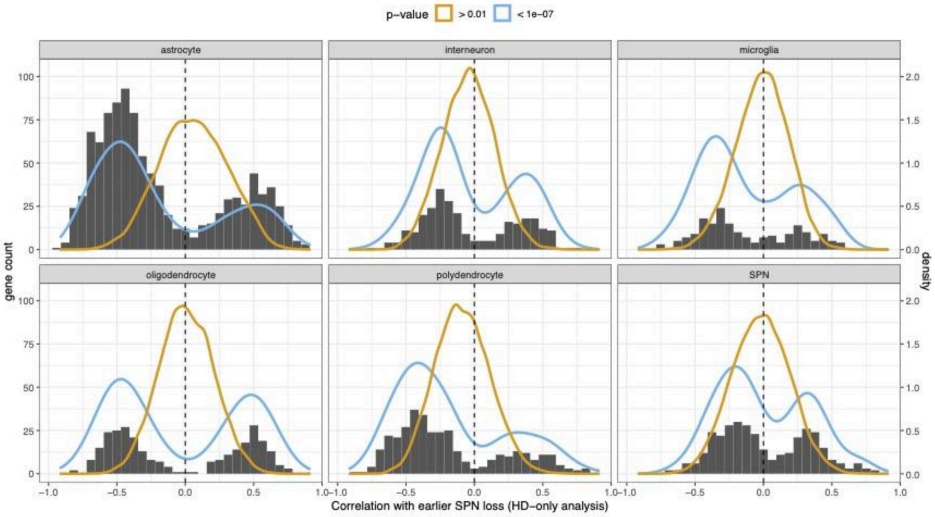


FIgure SN1.2 In each cell type, genes whose expression levels associated with case-control status (at genome-wide significance ($p < 10^{-7}$)) show a systematic pattern of strong positive or negative correlation (x-axis), in an HD-cases-only analysis, with earlier caudate atrophy as estimated from the extent to which SPN loss has already occurred (bars, summarized by blue density curve). As a negative control, genes whose expression levels did not associate with case-control status are shown as a negative control (orange density curve). These results are consistent with a model in which case-control gene-expression differences in each cell type are dominated by responses to caudate atrophy and earlier neuronal loss.

We believe that such data and analyses are useful for identifying biomarkers for HD progression and for understanding the responses of diverse cell types to neuronal death and circuitry changes.

However, inferences of cause and effect are hard to make from DEGs when so many genes change in expression levels in a highly correlated manner r and in proportion to earlier atrophy.

Supplementary Note 2: Somatic expansion in caudate cell types other than SPNs

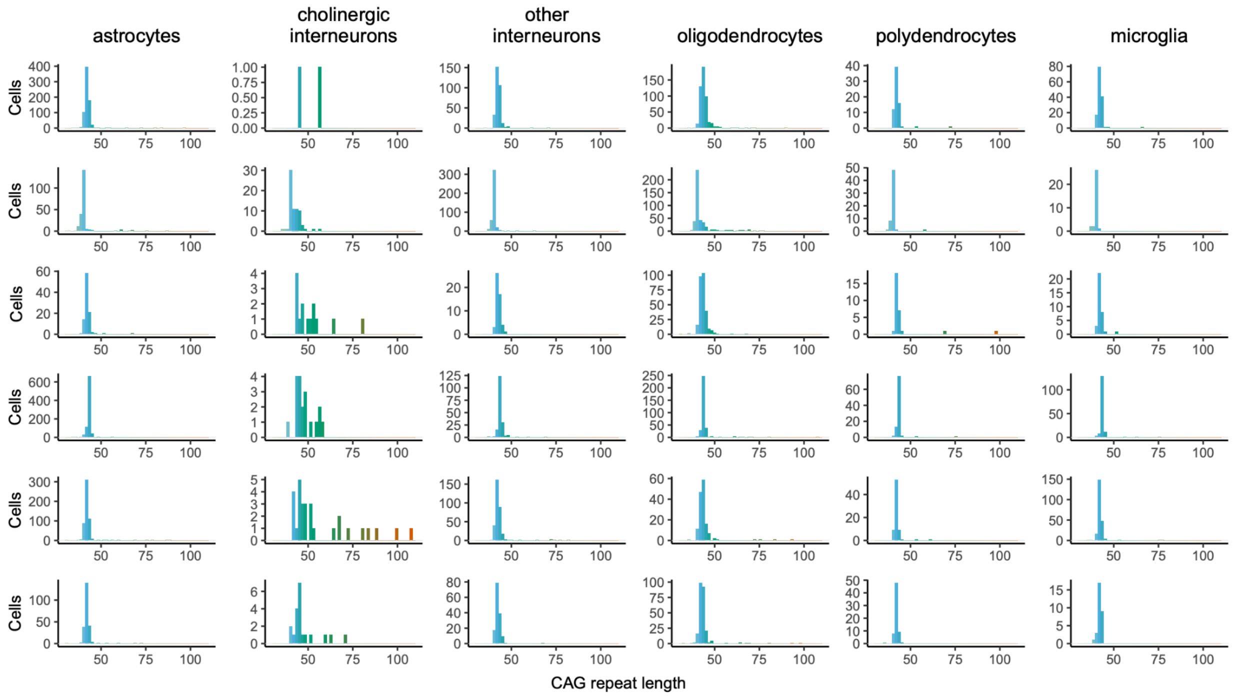


Figure SN2.1a. Somatic expansion in caudate cell types other than SPNs. In most cell types, the *HTT* CAG repeat exhibited only modest somatic instability. An exception was the cholinergic interneurons, a sparse caudate cell type (comprising between 0.5% and 3% of sampled nuclei) which exhibited more expansion than other non-SPN cell types did. Oligodendrocytes consistently exhibited more somatic expansion than other glial cell types, but less than cholinergic interneurons, which in turn exhibited far less expansion than SPNs did (see **Fig. SN2.2**). Data is shown only for long HD-risk alleles.

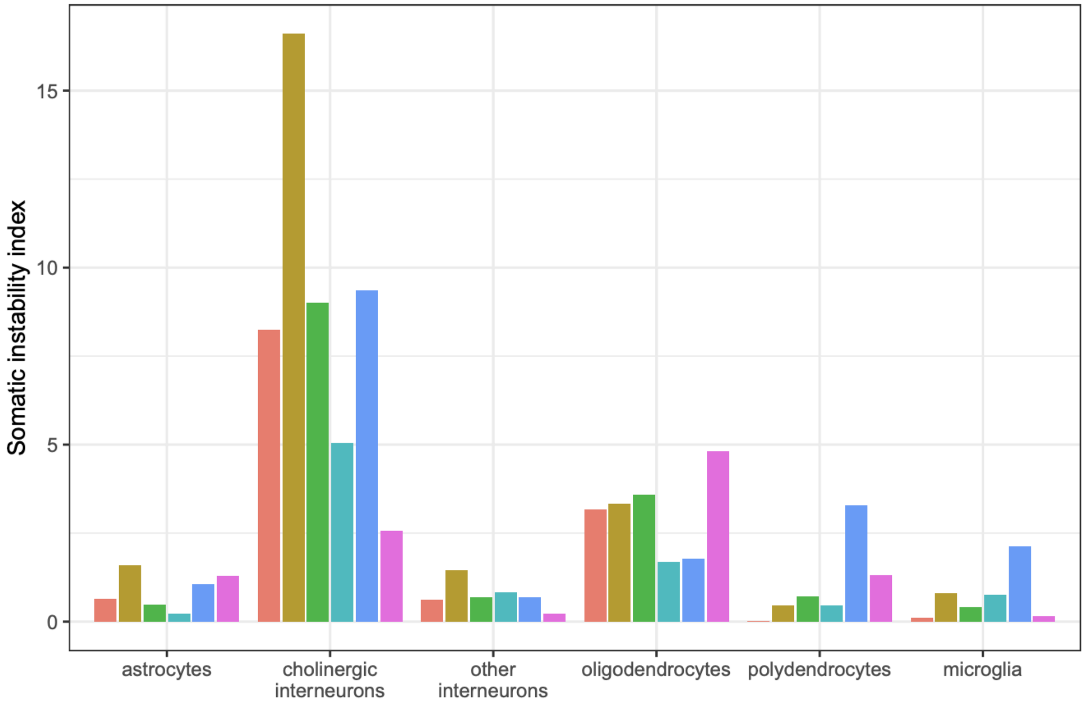


Figure SN2.1b. Somatic expansion in caudate cell types other than SPNs. Here the data for each of the six donors in SN1.3a (denoted here by the colors of the bars) are summarized with a somatic instability index which was calculated as the mean difference between the long CAG repeat allele in each cell and the inherited repeat length in each donor (uncommon outliers beyond 100 CAG repeats have been Winsorized).

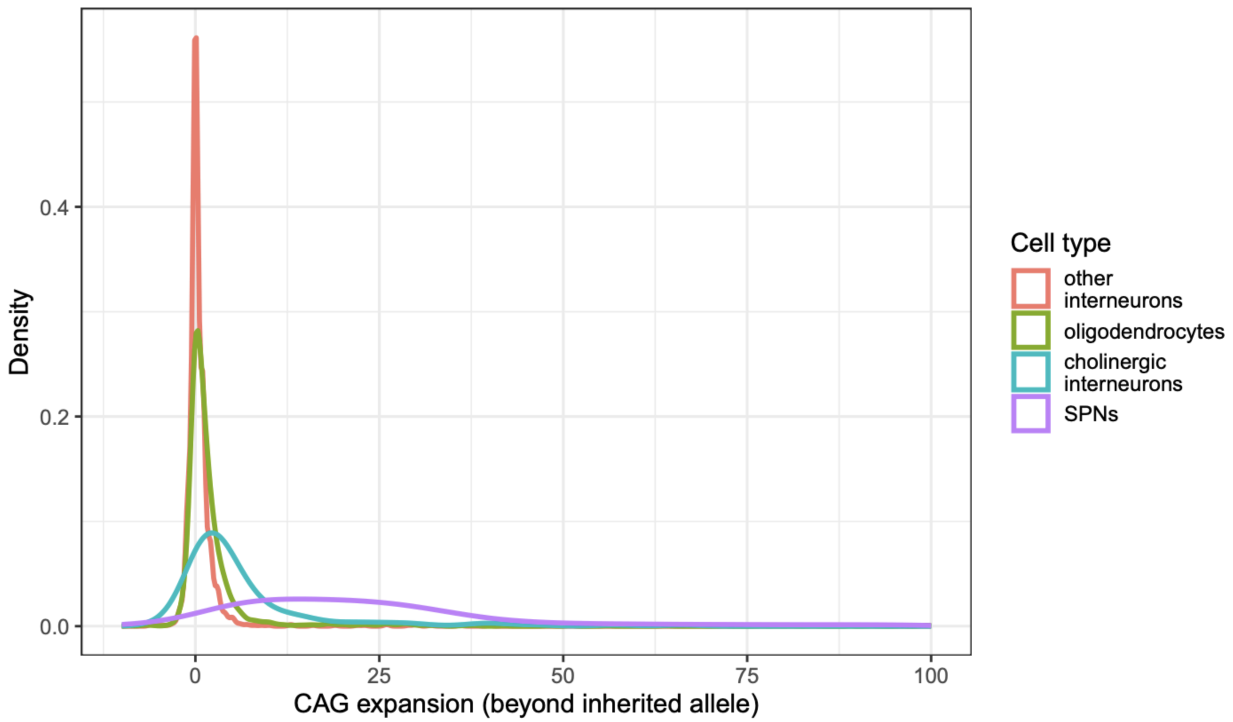


Figure SN2.2. Relative amounts of somatic expansion (density plot) in selected caudate cell types. Smoothed density plot showing the degree of somatic expansion, combined across six donors, in SPNs, oligodendrocytes, cholinergic interneurons and other caudate interneurons.

A recent study found that SPNs and cholinergic interneurons had similar amounts of somatic expansion and concluded that, therefore, somatic expansion is not sufficient to explain the vulnerability of SPNs in HD (Mätlik *et al.*, 2024). Our own data suggest that cholinergic neurons have much less somatic expansion than SPNs do (**Fig. SN2.2**), and in addition that SPNs have far more expansions in the range (150+ CAG repeats) that we consider to be biologically significant – a range that was not considered or ascertained in (Mätlik *et al.*, 2024). We also find that somatic expansion in cholinergic interneurons is variable across individual persons with HD (**Fig. SN2.1**). Finally, we consider it possible that some cholinergic interneurons are indeed lost in at least some persons with HD; in our own data, in persons with HD, slightly lower fractions of caudate nuclei were derived from cholinergic neurons, though this result was only nominally significant (p = 0.02). Our own conclusion is that SPNs have much greater somatic expansion and vulnerability than cholinergic neurons do, and that additional factors beyond somatic expansion into a biologically significant range (150+) are not needed to explain the relative vulnerability of SPNs.

Supplementary Note 3: Modeling and simulation of SPN CAG-repeat expansion dynamics

To better understand Huntington's disease progression, we investigated the dynamic behavior of somatic CAG repeat expansion over time in individual patients. While we are only able to sample cells from each patient's brain at a single time point, we sought to build simulations based on existing knowledge about the underlying mutational mechanism with minimal additional assumptions. We sought models that could explain the new observations of the long tail of cells with highly expanded repeats as well as the observed gene expression changes and the estimated degree of cell death experienced by each donor.

We created and critically evaluated a number of different models of repeat expansion. Across all of the models, we assumed an Independent mutational process in each cell, where most mutations change the repeat length by only one repeat unit. We assumed the mutation rate is a function of the current repeat length and that there is a bias for repeat expansions over contractions. The mutation rate and expansion bias may be different for each individual. For each individual, we included in their model their inherited repeat length, the age at which they were sampled, and an estimate from their data of the proportion of SPNs that they had lost. We estimated SPN loss for each donor based on the observed proportion of SPNs compared to other cell types (Supplementary Note).

An important feature of the simulations is to be able to model the high degree of SPN loss in HD. We sought to make minimal assumptions about the mechanism or timing of cell loss, beyond the observation that there were no significant gene expression changes in SPNs that have less than approximately 150 CAG repeats. We chose not to try to fit any specific functional relationship between repeat length and cell loss, assuming only that any cells that had been lost over time must have expanded beyond this minimum threshold (Supplementary Note). Figure ##-a shows examples of these models fit to the observed distributions in six representative donors.

One striking feature of the repeat length distributions is that they have a consistent shape across donors: while most of the distribution consists of cells with under 100 CAGs, there is also a long right tail containing a few percent of cells with very long repeats. This is reflected in a sharp "bend" in the cumulative distribution of the observed repeat lengths in each donor occurring around 80 CAGs (Figure ##-b). This feature of the distributions was not visible in earlier studies, but is enabled by the new ability to measure the repeat with long reads at single-cell resolution and with high throughput.

This feature of the distributions proved difficult to fit with models where the increase in the mutation rate is a simple function of repeat length. We were able to obtain better fits to the observed data utilizing models based on two phases of expansion, one slow and one fast, and we found that models lacking two phases tended to converge towards the two-phase models to the extent they were able to approximate the behavior of the two-phase models (Supplementary Note). Many two-phase models had similar predictive power and these models generated similar predictions of each donor’s expansion trajectory (Supplementary Note). For subsequent analyses, we used a two-phase model that follows a fitted power law in both phases that we found had a good tradeoff between simplicity and fit to the observed data (Figure ##-a).

Although the models are fitted to post-mortem data, they predict the dynamics of the expansion process across the patient’s entire lifetime, as well as forward in time (Figure ##-c). The models predict that for a typical adult-onset patient, repeat expansion is modest during the first decades of life, then rapidly increases as the CAG repeat gets longer. The models suggest that the average rate of expansion above 100 CAGs is XX times the rate of expansion below 70 CAGs. Figure ##-c shows three donors who died at different disease stages. Although their observed distributions are different, the models suggest a similar overall progression based on their inherited CAG allele.

An important feature of these simulations is that each individual neuron follows the same program of stochastic expansion asynchronously: were we able to observed the repeat length distribution at different points in time in an individual patient, we would see a distribution whose shape gradually changes over time with disease progression (Figure ##-c). At any given point in time, only a small fraction of the remaining SPNs would have very long repeat expansions above 150 CAGs, but such cells would be undergoing rapid expansion and distorted gene expression indicative of toxicity. An animated simulation illustrating this asynchronous behavior and the underlying model dynamics is available at (##URL).

An important prediction of the models is that each individual SPN spends many years in a relatively slow repeat expansion phase below 70 CAGs. We estimate that the typical SPN in a pwHD spends 96.4% (s.d. 2.0%) of its life below the critical threshold of 150 CAGs at which we first begin to observe distorted gene expression.

These models are consistent with many observed properties of HD progression. As one example, we used the models that were fit to each donor to predict how age at onset would have varied in that individual conditional on a different inherited allele. As a simple proxy for age at onset, we used the age at which 25% of a donor’s SPNs would have reached a repeat length of 300 or more CAGs. The results agree well with the known relationship between the inherited allele length at onset [PMID: 31398342] (Figure ##-d).

Methods

Brain donations and ethical compliance

Brain donors were recruited by the Harvard Brain Tissue Resource Center/NIH NeuroBioBank (HBTRC/NBB), in a community-based manner, across the USA. Human brain tissue was obtained from the HBTRC/NBB. 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. 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 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 is used for diagnosis . Diagnosis on early cases is done using histological stainings and polyglutamine immunohistochemistry (Hedreen *et al.*, 1991; Mattsson *et al.*, 1974; Vonsattel *et al.*, 1985). Positivity in pontine gray neurons rules out HD like-2 neuropathology , and cerebellar dentate neurons are mildly positive even in very early cases, while Purkinje cells are negative (unlike in cerebellar ataxia CAG expansion cases).

Affected individuals were selected for this study 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. Experiments utilized fresh frozen brain tissue from each donor.

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.

Extraction and analysis of nuclei in 20-donor “villages” for snRNA-seq

For analyses comparing *across* donors (**Fig. 1-3, 11**), 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 (Wells *et al.*, 2023; Ling, Nemesh, Goldman, Kamitaki, Reed, Handsaker, Genovese, Vogelgsang, Gerges, Kashin, Ghosh, Esposito, French, *et al.*, 2024) in which we make preparations of nuclei from sets (or “villages” (Wells *et al.*, 2023)) of 20 donors at once. ESpecimens 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 (**Supplementary Fig. 1**). 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 (Wells *et al.*, 2023; Ling, Nemesh, Goldman, Kamitaki, Reed, Handsaker, Genovese, Vogelgsang, Gerges, Kashin, Ghosh, Esposito, French, *et al.*, 2024). This experimental approach allowed the data to be highly comparable donor-to-donor (**Fig. 2**).

Nuclei were isolated from frozen brain tissue using approaches we have described (Wells *et al.*, 2023; Ling, Nemesh, Goldman, Kamitaki, Reed, Handsaker, Genovese, Vogelgsang, Gerges, Kashin, Ghosh, Esposito, French, *et al.*, 2024) and https://uploadOPTIPREPprotocol). 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 500g 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/uL 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 500g for 5 min. The supernatant was discarded and the pellets were resuspended in 300 ul of G30 (30% iodixanol, 3.4% sucrose, 20 mM tricine, 25mM KCl, 5 mM MgCl2, [pH 7.8]). The resuspended tissue pellets were layered with 1 mL of G30 and spun down at 4C with 8000g 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/uL NxGen® RNase Inhibitor). The nuclei were resuspended in 50 ul of the wash buffer and counted by using LUNA-FL™ Dual Fluorescence Cell Counter (Logos Biosystems).

Single-nucleus RNA-seq (transcriptome libraries)

The isolated nuclei were encapsulated into droplets and the snRNA-seq library was prepared by using Chromium Next GEM Single Cell 3' GEM, 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.

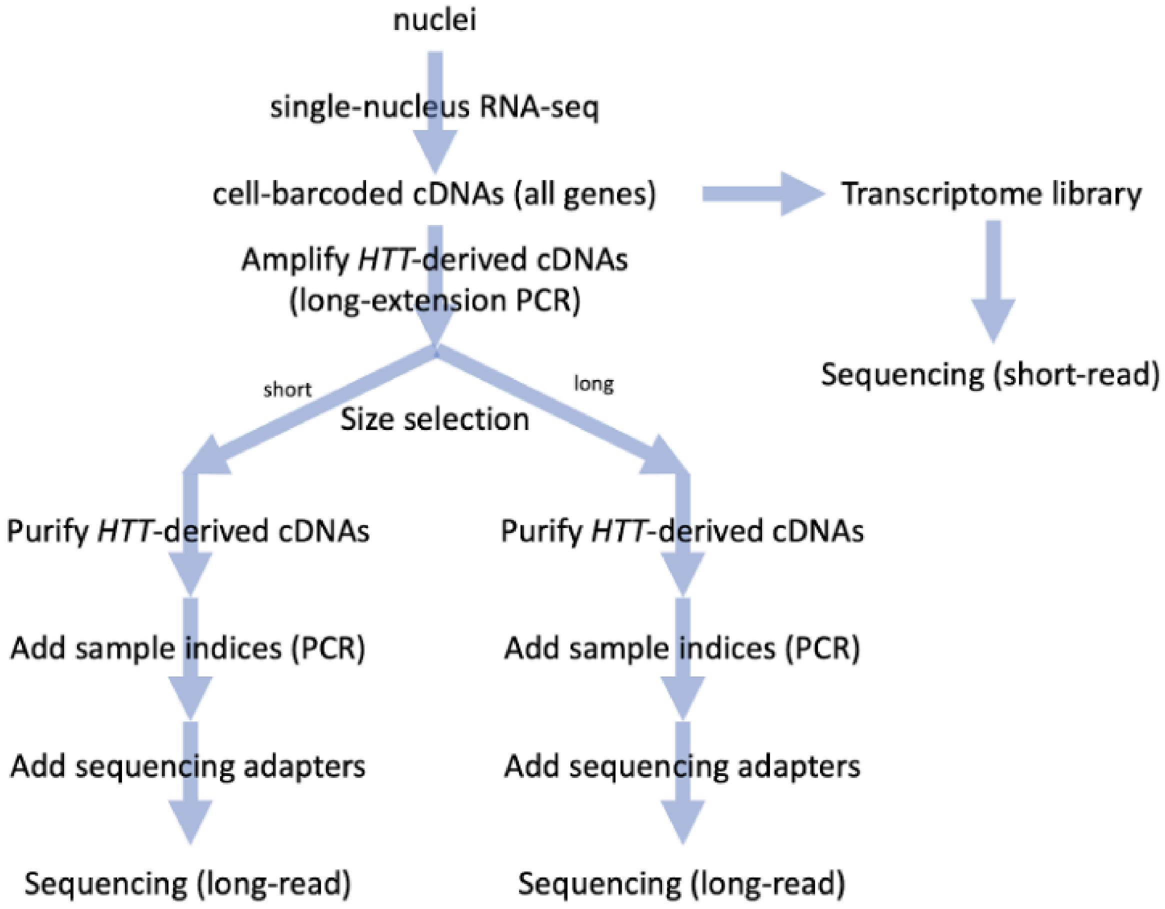
Processing snRNA-seq data

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 (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)(Virshup *et al.*, 2021), numpy (v1.17.5), pandas (v1.0.5), and Scanpy (v1.9.1).

Measurements of CAG-repeat length in individual cells

We also 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. Our approach (sn(RNA+repeat)-seq) creates two molecular libraries from each set of nuclei: one library samples genome-wide RNA expression ("transcriptome library"), and another library specifically captures the 5' region of *HTT* transcripts ("*HTT*-CAG library") (**Fig. 3**). 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 (**Fig. above**) 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, step-by-step protocol, with helpful tips, tricks, modifications and pausing points, is being prepared for release on protcols.io with the publication of this work. We include a summary here.

We begin by isolating and encapsulating nuclei in droplets as described above. To enable the capture and amplification of 5' HTT sequences, we spike in primers that target *HTT*-specific gene sequences.

We then prepare two libraries from each sample: a standard transcriptome sequencing library, and an *HTT*-CAG library. To generate the *HTT*-CAG library, 4 uL of the cDNA generated from each reaction of the encapsulated nuclei is used for PCR amplification with a biotinylated forward primer that targets *HTT* 5' of the CAG repeat and a partial read1 primer appended to Nextera adapter sequence. Purification of the resulting PCR product on streptavidin beads (Dynabeads™ MyOne™ Streptavidin C1 (ThermoScientific, #65002)) enriches the library for *HTT* CAG sequences.

We 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 involves further preparing libraries for long-read sequencing. Each purified "L" and "S" library from the same PCR reactions was indexed with the same Illumina Nextera indices (N701-712 and N501-508), and pooled separately at an equimolar ratio for generating Pacific Biosciences (PacBio) sequencing libraries. 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).

Computational analysis of HTT-CAG data

Barcode correction if any
Identification of read-families sharing a CBC and UMI
Determination of a consensus CAG-repeat length for each read family

Analyses of CAG-repeat effects on SPN gene expression – SPN groups (Wilcox text)

Analyses of CAG-repeat effects on SPN gene expression – Negative Binomial Regression

Identification of phase-D genes

Modeling and simulation of CAG-repeat expansion dynamics

References

Albin, R.L. *et al.* (1990) ‘Striatal and nigral neuron subpopulations in rigid Huntington’s disease: implications for the functional anatomy of chorea and rigidity-akinesia’, *Annals of neurology*, 27(4), pp. 357–365.

Braz, B.Y. *et al.* (2022) ‘Treating early postnatal circuit defect delays Huntington’s disease onset and pathology in mice’, *Science*, 377(6613), p. eabq5011.

Campion, L.N. *et al.* (2022) ‘Tissue-specific and repeat length-dependent somatic instability of the X-linked dystonia parkinsonism-associated CCCTCT repeat’, *Acta neuropathologica communications*, 10(1), p. 49.

Corless, S. and Gilbert, N. (2016) ‘Effects of DNA supercoiling on chromatin architecture’, *Biophysical reviews*, 8(Suppl 1), pp. 51–64.

Dragileva, E. *et al.* (2009) ‘Intergenerational and striatal CAG repeat instability in Huntington’s disease knock-in mice involve different DNA repair genes’, *Neurobiology of disease*, 33(1), pp. 37–47.

Ferguson, R. and Tabrizi, S.J. (2023) ‘Can MSH3 lowering stop HTT repeat expansion in its CAG tract?’, *Molecular therapy: the journal of the American Society of Gene Therapy*, pp. 1509–1511.

Finneran, D. *et al.* (2023) ‘Differential effects of neuronal Cdkn2a over‐expression in mouse brain’, *Alzheimer’s & dementia: the journal of the Alzheimer's Association*, 19(S13). Available at: https://doi.org/10.1002/alz.078485.

Garcia, F.J. *et al.* (2022) ‘Single-cell dissection of the human brain vasculature’, *Nature*, 603(7903), pp. 893–899.

Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium (2015) ‘Identification of Genetic Factors that Modify Clinical Onset of Huntington’s Disease’, *Cell*, 162(3), pp. 516–526.

Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium. (2019) ‘CAG Repeat Not Polyglutamine Length Determines Timing of Huntington’s Disease Onset’, *Cell*, 178(4), pp. 887–900.e14.

Gil, J. and Peters, G. (2006) ‘Regulation of the INK4b-ARF-INK4a tumour suppressor locus: all for one or one for all’, *Nature reviews. Molecular cell biology*, 7(9), pp. 667–677.

Goold, R. *et al.* (2021) ‘FAN1 controls mismatch repair complex assembly via MLH1 retention to stabilize CAG repeat expansion in Huntington’s disease’, *Cell reports*, 36(9), p. 109649.

Graybiel, A.M. and Matsushima, A. (2023) ‘Striosomes and Matrisomes: Scaffolds for Dynamic Coupling of Volition and Action’, *Annual review of neuroscience*, 46, pp. 359–380.

Graybiel, A.M. and Ragsdale, C.W., Jr (1978) ‘Histochemically distinct compartments in the striatum of human, monkeys, and cat demonstrated by acetylthiocholinesterase staining’, *Proceedings of the National Academy of Sciences of the United States of America*, 75(11), pp. 5723–5726.

Gusella, J.F., Lee, J.-M. and MacDonald, M.E. (2021) ‘Huntington’s disease: nearly four decades of human molecular genetics’, *Human molecular genetics*, 30(R2), pp. R254–R263.

Hedreen, J.C. and Folstein, S.E. (1995) ‘Early loss of neostriatal striosome neurons in Huntington’s disease’, *Journal of neuropathology and experimental neurology*, 54(1), pp. 105–120.

Herranz, N. and Gil, J. (2018) ‘Mechanisms and functions of cellular senescence’, *The Journal of clinical investigation*, 128(4), pp. 1238–1246.

Hommelsheim, C.M. *et al.* (2014) ‘PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications’, *Scientific reports*, 4, p. 5052.

Hong, E.P. *et al.* (2021) ‘Huntington’s Disease Pathogenesis: Two Sequential Components’, *Journal of Huntington’s disease*, 10(1), pp. 35–51.

Igney, F.H. and Krammer, P.H. (2002) ‘Death and anti-death: tumour resistance to apoptosis’, *Nature reviews. Cancer*, 2(4), pp. 277–288.

Iyer, R.R. and Pluciennik, A. (2021) ‘DNA Mismatch Repair and its Role in Huntington’s Disease’, *Journal of Huntington’s disease*, 10(1), pp. 75–94.

Kennedy, L. *et al.* (2003) ‘Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis’, *Human molecular genetics*, 12(24), pp. 3359–3367.

Kim, K.-H. *et al.* (2020) ‘Genetic and Functional Analyses Point to FAN1 as the Source of Multiple Huntington Disease Modifier Effects’, *The American Journal of Human Genetics*, pp. 96–110. Available at: https://doi.org/10.1016/j.ajhg.2020.05.012.

Kovalenko, M. *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(9), p. e44273.

Laabs, B.-H. *et al.* (2021) ‘Identifying genetic modifiers of age-associated penetrance in X-linked dystonia-parkinsonism’, *Nature communications*, 12(1), p. 3216.

Lee, H. *et al.* (2020) ‘Cell Type-Specific Transcriptomics Reveals that Mutant Huntingtin Leads to Mitochondrial RNA Release and Neuronal Innate Immune Activation’, *Neuron*, 107(5), pp. 891–908.e8.

Ling, E., Nemesh, J., Goldman, M., Kamitaki, N., Reed, N., Handsaker, R.E., Genovese, G., Vogelgsang, J.S., Gerges, S., Kashin, S., Ghosh, S., Esposito, J.M., Morris, K., *et al.* (2024) ‘A concerted neuron–astrocyte program declines in ageing and schizophrenia’, *Nature*, pp. 1–8.

Ling, E., Nemesh, J., Goldman, M., Kamitaki, N., Reed, N., Handsaker, R.E., Genovese, G., Vogelgsang, J.S., Gerges, S., Kashin, S., Ghosh, S., Esposito, J.M., French, K., *et al.* (2024) ‘Concerted neuron-astrocyte gene expression declines in aging and schizophrenia’, *bioRxiv : the preprint server for biology* [Preprint]. Available at: https://doi.org/10.1101/2024.01.07.574148.

Loupe, J.M. *et al.* (2020) 'Promotion of somatic CAG repeat expansion by Fan1 knock-out in Huntington's disease knock-in mice is blocked by Mlh1 knock-out', *Human molecular genetics*, 29(18), pp. 3044–3053.

MacDonald, M.E. *et al.* (1993) 'A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes', *Cell*, 72(6), pp. 971–983.

Macosko, E.Z. *et al.* (2015) 'Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets', *Cell*, 161(5), pp. 1202–1214.

Malaiya, S. *et al.* (2021) 'Single-Nucleus RNA-Seq Reveals Dysregulation of Striatal Cell Identity Due to Huntington's Disease Mutations', *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 41(25), pp. 5534–5552.

Mätlik, K. *et al.* (2024) 'Cell-type-specific CAG repeat expansions and toxicity of mutant Huntingtin in human striatum and cerebellum', *Nature genetics* [Preprint]. Available at: https://doi.org/10.1038/s41588-024-01653-6.

Monckton, D.G. *et al.* (1995) 'Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses', *Human molecular genetics*, 4(1), pp. 1–8.

Morales, F. *et al.* (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, pp. 57–66.

Morales, F. *et al.* (2020) 'Longitudinal increases in somatic mosaicism of the expanded CTG repeat in myotonic dystrophy type 1 are associated with variation in age-at-onset', *Human molecular genetics*, 29(15), pp. 2496–2507.

Nature Reviews Drug Discovery (2021) 'Double setback for ASO trials in Huntington disease', 19 May, pp. https://www.nature.com/articles/d41573–021–00088–6.

Paulson, H. (2018) 'Repeat expansion diseases', *Handbook of clinical neurology*, 147, pp. 105–123.

Phadte, A.S. *et al.* (2023) 'FAN1 removes triplet repeat extrusions via a PCNA- and RFC-dependent mechanism', *Proceedings of the National Academy of Sciences of the United States of America*, 120(33), p. e2302103120.

Pinto, R.M. *et al.* (2013) 'Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches', *PLoS genetics*, 9(10), p. e1003930.

Pressl, C. *et al.* (2024) 'Selective vulnerability of layer 5a corticostriatal neurons in Huntington's disease', *Neuron* [Preprint]. Available at: https://doi.org/10.1016/j.neuron.2023.12.009.

Rajagopal, S. *et al.* (2023) 'Genetic modifiers of repeat expansion disorders', *Emerging topics in life sciences*, 7(3), pp. 325–337.

Sathasivam, K. *et al.* (2013) 'Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease', *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), pp. 2366–2370.

von Schimmelmann, M. *et al.* (2016) ‘Polycomb repressive complex 2 (PRC2) silences genes responsible for neurodegeneration’, *Nature neuroscience*, 19(10), pp. 1321–1330.

Science (2021) ‘Promising drug for Huntington disease fails in major trial’, 23 March, p. https://www.science.org/content/article/promising–drug–huntington–disease–fails–major–trial.

Shelbourne, P.F. *et al.* (2007) ‘Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain’, *Human molecular genetics*, 16(10), pp. 1133–1142.

Swami, M. *et al.* (2009) ‘Somatic expansion of the Huntington’s disease CAG repeat in the brain is associated with an earlier age of disease onset’, *Human molecular genetics*, 18(16), pp. 3039–3047.

Tabrizi, S.J. *et al.* (2013) ‘Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington’s disease in the TRACK-HD study: analysis of 36-month observational data’, *Lancet neurology*, 12(7), pp. 637–649.

Tabrizi, S.J. *et al.* (2022) ‘A biological classification of Huntington’s disease: the Integrated Staging System’, *Lancet neurology*, 21(7), pp. 632–644.

Tabrizi, S.J., Ghosh, R. and Leavitt, B.R. (2019) ‘Huntingtin Lowering Strategies for Disease Modification in Huntington’s Disease’, *Neuron*, 101(5), pp. 801–819.

Telenius, H. *et al.* (1994) ‘Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm’, *Nature genetics*, 6(4), pp. 409–414.

Tomé, S. *et al.* (2013) ‘MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington’s disease mice’, *PLoS genetics*, 9(2), p. e1003280.

Virshup, I. *et al.* (2021) ‘anndata: Annotated data’, *bioRxiv*. Available at: https://doi.org/10.1101/2021.12.16.473007.

Wells, M.F. *et al.* (2023) ‘Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages’, *Cell stem cell*, 30(3), pp. 312–332.e13.

Wheeler, V.C. *et al.* (2003) ‘Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum’, *Human molecular genetics*, 12(3), pp. 273–281.

Wilton, D.K. *et al.* (2023) ‘Microglia and complement mediate early corticostriatal synapse loss and cognitive dysfunction in Huntington’s disease’, *Nature medicine*, 29(11), pp. 2866–2884.

Yuile, A. *et al.* (2023) ‘CDKN2A/B Homozygous Deletions in Astrocytomas: A Literature Review’, *Current issues in molecular biology*, 45(7), pp. 5276–5292.