nature medicine

Article

https://doi.org/10.1038/s41591-024-03424-6

Somatic CAG repeat expansion in blood associates with biomarkers of neurodegeneration in Huntington's disease decades before clinical motor diagnosis

Received: 2 August 2024

Accepted: 15 November 2024

Published online: 17 January 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Huntington's disease (HD) is an autosomal dominant neurodegenerative disease with the age at which characteristic symptoms manifest strongly influenced by inherited *HTT* CAG length. Somatic CAG expansion occurs throughout life and understanding the impact of somatic expansion on neurodegeneration is key to developing therapeutic targets. In 57 HD gene expanded (HDGE) individuals, ~23 years before their predicted clinical motor diagnosis, no significant decline in clinical, cognitive or neuropsychiatric function was observed over 4.5 years compared with 46 controls (false discovery rate (FDR) > 0.3). However, cerebrospinal fluid (CSF) markers showed very early signs of neurodegeneration in HDGE with elevated neurofilament light (NfL) protein, an indicator of neuroaxonal damage (FDR = 3.2×10^{-12}), and reduced proenkephalin (PENK), a surrogate marker for the state of striatal medium spiny neurons (FDR = 2.6×10^{-3}), accompanied by brain atrophy, predominantly in the caudate (FDR = 5.5×10^{-10}) and putamen (FDR = 1.2×10^{-9}). Longitudinal increase in somatic CAG repeat expansion ratio (SER) in blood was a significant predictor of subsequent caudate (FDR = 0.072) and putamen (FDR = 0.148) atrophy. Atypical loss of interruption *HTT* repeat structures, known to predict earlier age at clinical motor diagnosis, was associated with substantially faster caudate and putamen atrophy. We provide evidence in living humans that the influence of CAG length on HD neuropathology is mediated by somatic CAG repeat expansion. These critical mechanistic insights into the earliest neurodegenerative changes will inform the design of preventative clinical trials aimed at modulating somatic expansion. ClinicalTrials.gov registration: NCT06391619.

Huntington's disease (HD) is a devastating condition characterized by loss of striatal medium spiny neurons (MSNs) and striatal neurodegeneration[1] leading to impaired motor, cognitive and neuropsychiatric function which typically manifests in middle age, with clinical diagnosis defined by the appearance of unequivocal HD-related motor signs. There are currently no disease-modifying treatments[2].

HD is an autosomal dominant disorder and is caused by an expanded CAG repeat ≥40 in the huntingtin gene (*HTT*) coding for polyglutamine in the mutant huntingtin protein (mHTT), which is the presumed toxic entity leading to neuronal dysfunction and death. It is well established that inherited CAG repeat length has a strong influence on age at clinical motor diagnosis[3]. Notably, the *HTT* repeat is somatically unstable[4] and expansion of tens or even hundreds of repeats are observed in the most vulnerable striatal neurons[5–8]; greater somatic expansion occurs with longer initial CAG length. Evidence indicating that faster individual-specific rates of somatic expansion in brain are

e-mail: s.tabrizi@ucl.ac.uk

associated with earlier clinical motor diagnosis and faster disease progression[9] strongly suggests that somatic expansion is a key mechanism explaining the CAG effect on disease progression. Indeed, it has been suggested that somatic expansion is required to generate pathology, and that HD involves two thresholds as follows: first, the inherited CAG length that leads to further somatic expansion, and second, the intracellular pathogenic threshold above which neuronal dysfunction and death occur[10–13]. Consistent with this, a recent postmortem study suggests that neurons may experience decades of 'biologically quiet' somatic CAG repeat expansion with neuronal damage triggered by a cascade of repeat-length dependent transcriptional dysregulation events only when the CAG reaches a threshold of ~150 repeats[8]. Further understanding the dynamics of somatic expansion directly in the brain is hampered by the nonavailability of brain biopsy material from young living participants. Although somatic CAG expansion is clearly cell-type dependent[6–8], faster individual-specific rates of somatic expansion in blood DNA are also associated with earlier clinical motor diagnosis[14], suggesting that individual-specific somatic expansion rates in blood DNA are at least partially predictive of individual-specific somatic expansion rates in the brain. This hypothesis is supported by genetic modifier studies that reveal a panoply of DNA repair gene variants as modifiers of both *HTT* somatic expansion and HD clinical phenotypes[13–17].

The polyglutamine-encoding CAG repeat tract in *HTT* is followed just downstream with a polymorphic polyproline-encoding CCG repeat. Typically, the intervening sequence between the CAG and CCG repeat tracts is comprised of a glutamine-encoding CAACAG cassette and a proline-encoding CCGCCA cassette. However, a number of atypical *HTT* repeat structures have been identified with loss of either or both of the intervening CAACAG or CCGCCA cassettes associated with an earlier age at clinical motor diagnosis; conversely, duplication of the CAACAG cassette delays this milestone[13,14,17–19]. These data reveal that both HD age at clinical motor diagnosis and the somatic expansion potential of the repeat are best predicted by pure CAG repeat length, rather than encoded polyglutamine length, providing additional support for a key role for somatic expansion in driving disease onset[13,14,18].

The monogenic nature of HD and the existence of diagnostic and predictive testing for at-risk family members makes it a tractable disease and much progress has been made towards developing disease modification treatments[2]. The first phase 1/2 trial of an antisense oligonucleotide (ASO), tominersen, showed dose-dependent lowering of mutant huntingtin levels[20]. Although the subsequent phase 3 trial was halted early due to adverse safety concerns[21], a phase 2 study to better establish safety and tolerability earlier in disease progression is ongoing (ClinicalTrials.gov registration: NCT05686551). Alternative approaches such as allele-specific huntingtin-lowering, protein splicing modulation, and gene therapy are also currently being trialed (reviewed in ref. 22). Additionally, somatic expansion and proteins, such as MSH3 and FAN1, are now being actively pursued as therapeutic targets in HD. A key question in using such therapies will be determining the optimal timing for treatment. The appearance of HD motor signs is already accompanied by substantial striatal neurodegeneration, and earlier treatment seems likely to produce greater clinical benefit. However, all the studies to date have relied on postmortem brain analyses to model the link between CAG repeat expansion to the earliest pathological progression of the disease. Understanding the triggers of the neurodegenerative process is vital in the search for future therapies and identifying the best time to treat to provide therapeutic intervention.

The greatest opportunity to influence disease progression lies in early treatment, with the goal of delaying or preventing clinical motor diagnosis. Numerous large observational studies show that brain changes occur decades from predicted clinical motor diagnosis[23–25] and that subtle cognitive and motor signs emerge as HD gene expanded (HDGE) individuals approach clinical motor diagnosis. The recent introduction of the HD Integrated Staging System (HD-ISS) provides a new empirical framework for classifying people with HD throughout life[26], with stage 0 being the HDGE group with striatal volumes within the general population range, stage 1 being the presence of a biomarker of pathogenesis (caudate and/or putamen volume change), stage 2 being the presence of motor and/or cognitive signs and stage 3 being marked by the onset of functional impairment[26]. Cohorts in the earliest stages will likely gain the most benefit from preventative therapies.

A key challenge in delivering preventative treatments is to identify and validate robust measures in HD-ISS stages 0 and 1, where the absence of outward signs of impairment renders established motor and cognitive testing batteries insensitive. HD Young Adult Study (HD-YAS) is a unique cohort, ~23 years from predicted clinical motor diagnosis at baseline with deep phenotyping including biofluid, imaging, clinical, cognitive and motor assessments. Our cross-sectional baseline data demonstrated subtle elevations in biofluid biomarkers, such as cerebrospinal fluid (CSF) neurofilament light (NfL), accompanied by slightly smaller putamen volumes in the HDGE group compared to unaffected controls[25]. Despite this, there was no difference in functional performance between the groups. This cohort, therefore, spans an optimum window for investigating the potential of interventions to delay or prevent symptoms.

Here we present 4.5-year follow-up data from HD-YAS, a deep-phenotyped longitudinal study of young stages 0 and 1 HDGE adults, ~19 years before clinical motor diagnosis. We hypothesized that the effects of somatic expansion in the brain might be detected long before clinical motor onset and tested this hypothesis through detailed longitudinal analysis of preclinical HD phenotypes, biomarkers of neurodegeneration and somatic expansion in blood DNA. We examined change over time in a range of assessments with the aim of identifying ongoing neuropathology and associations with somatic CAG expansion in blood DNA and *HTT* repeat structures, decades before predicted clinical motor diagnosis, and biomarkers of disease progression, which may have utility in future prevention trials.

Results

Participant characteristics

A total of 131 (64 HDGE and 67 controls) participants attended at baseline and 103 (57 HDGE and 46 controls) returned for follow-up ~4.5 years later (see Extended Data Fig. 1 for reasons for dropout). To account for those not returning, we recruited 23 new participants (9 HDGE and 14 controls) giving a total of 154 participants (73 HDGE and 81 controls). At baseline, 44 (81%) participants of the cohort were in HD-ISS stage 0, 9 (17%) in stage 1 and 1 (2%) in stage 2 (Fig. 1a). Over 4.5 years, 10 (~23%) participants moved from stage 0 to stage 1; there was no progression to stage 2. The transition in staging within the HD-YAS cohort is depicted by overlaying the probability matrix for each HD-ISS stage across different ages for individuals with a mean CAG repeat length of 42, comparable to the mean CAG repeat length of our cohort (Fig. 1b). Here we describe further longitudinal results from the participants; cross-sectional results, updated from the original baseline study, are provided in Supplementary Results and Discussion.

There were no significant differences (false discovery rate (FDR) < 0.15) between the HDGE and control groups in age, sex, interval between visits, education score or National Adult Reading Test (a measure of premorbid intelligence; Extended Data Table 1).

Cognitive and neuropsychiatric assessments

There was no significant longitudinal disease-related decline in any of the comprehensive cognitive (FDR > 0.8; Fig. 1c) or neuropsychiatric (FDR > 0.3; Fig. 1d) assessments, demonstrating that change in the HDGE group was no different from matched controls. Cross-sectional results are shown in Supplementary Fig. 1 and summary statistics are provided in Supplementary Tables 1 and 2 for longitudinal and Supplementary Tables 3 and 4 for cross-sectional results.

https://doi.org/10.1038/s41591-024-03424-6

Neuroimaging

After quality control, longitudinal data were available for 88 (54 HDGE and 34 controls) participants for volumetric imaging, 83 (50 HDGE and 33 controls) for diffusion-weighted imaging (DWI) and 75 (43 HDGE and 32 controls) for multiparametric mapping (MPM). As left-handed participants were excluded, 70 (43 HDGE and 27 controls) participants were available for the structural connectivity analysis. See Supplementary Table 5 and Supplementary Methods for further details.

The HDGE group showed significantly greater rates of atrophy in putamen ($P = 4.0 \times 10^{-10}$, FDR = 1.2×10^{-9}) and caudate ($P = 1.1 \times 10^{-10}$, FDR = 5.5×10^{-10}). There were also significant group differences for gray matter ($P = 7.5 \times 10^{-3}$, FDR = 9.4×10^{-3}), white matter ($P = 1.4 \times 10^{-2}$, FDR = 1.4×10^{-2}) and whole brain ($P = 7.1 \times 10^{-4}$, FDR = 1.2×10^{-3}) with associated ventricular expansion ($P = 3.9 \times 10^{-5}$, FDR = 9.8×10^{-5}; Fig. 2). Caudate, putamen and white matter loss were significantly predicted by age and CAG ($P = 2.1 \times 10^{-7}$, FDR = 1.0×10^{-6}; $P = 1.5 \times 10^{-8}$, FDR = 8.9×10^{-8}; $P = 0.01$, FDR = 0.012, respectively).

DWI demonstrated elevated rates of longitudinal change in all diffusion and neurite orientation and dispersion density imaging metrics across multiple regions of interest in the HDGE group compared to controls (FDR < 0.15). The splenium of the corpus callosum, the anterior capsule and the external capsule showed associations with age and CAG (FDR < 0.15). There were no significant between-group differences in the rate of change for any of the structural connectivity (all FDR > 0.4) or MPM measures (all FDR > 0.3), nor any evidence of an influence of age and CAG (all FDR > 0.15).

Neuroimaging results suggest that across HD-ISS stages 0 and 1, there are already elevated rates of brain atrophy accompanied by subtle microstructural white matter changes. See Extended Data Tables 2 and 3 for summary statistics for longitudinal volumetric and diffusion results, respectively. Summary statistics for remaining longitudinal metrics are provided in Supplementary Tables 6 and 7 and cross-sectional data in Supplementary Tables 8–11.

Biofluids

A total of 216 biofluid samples were collected across baseline and follow-up visits over the 4.5-year interval. Paired fasting CSF and plasma samples were acquired in 86 (53 HDGE and 33 controls) of the 103 (83.5%) longitudinal participants.

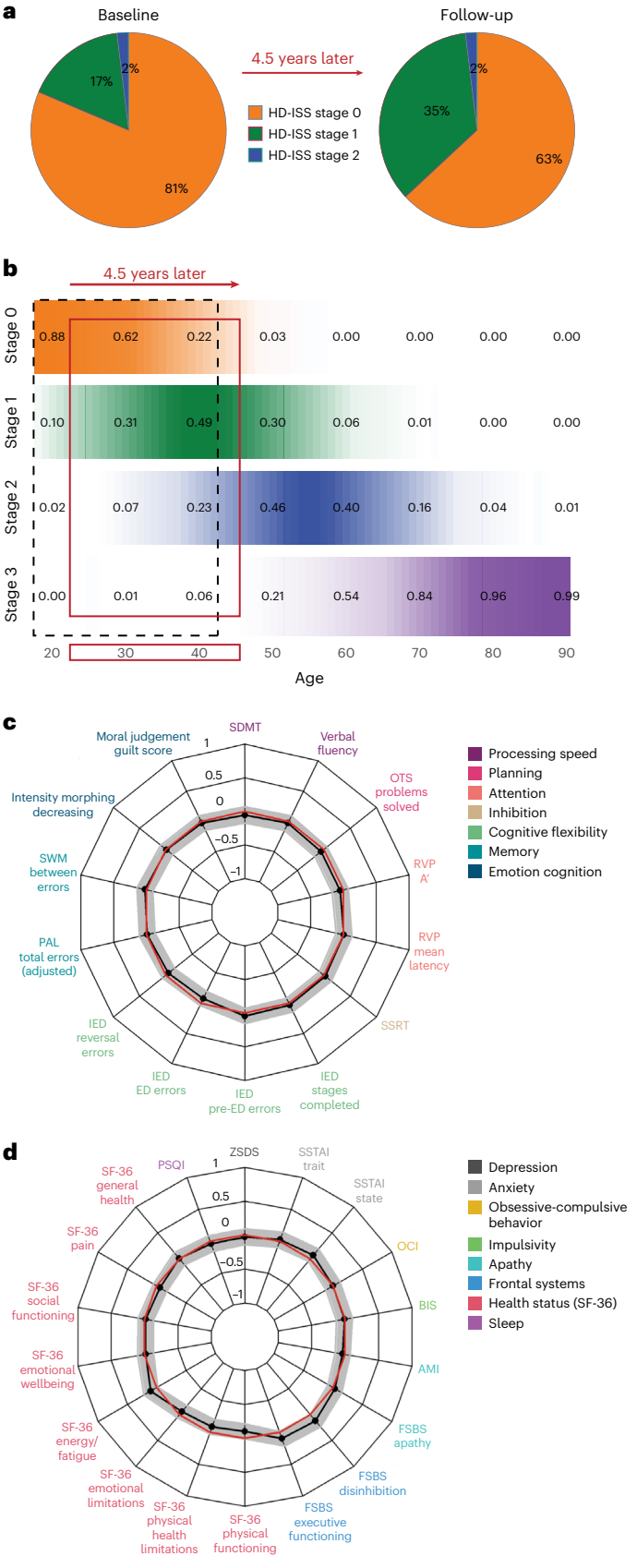


Fig. 1 | Longitudinal change in clinical, cognitive and neuropsychiatric measures. a, The distribution of HD-ISS stages at baseline and at follow-up 4.5 years later. **b**, A probability matrix for being in each HD-ISS stage across different ages for individuals with a mean CAG repeat length of 42, which is comparable to the HD-YAS cohort. These probabilities are derived from data in the Enroll-HD, PREDICT-HD and TRACK-HD studies, which were used to develop the HD-ISS[26]. The black dashed box highlights the HD-YAS cohort at baseline, while the red box indicates their position at follow-up after 4.5 years. **c**, A radar plot showing group differences in longitudinal changes in cognitive measures. **d**, A radar plot showing group differences in longitudinal changes for neuropsychiatric and functional measures. The black line represents the standardized mean difference between the HDGE and control groups, with conventional frequentist 95% CI shaded in gray. The red circle denotes no difference between means; values within this circle indicate greater change over time in the HDGE group. After FDR correction for multiple comparisons, there were no significant longitudinal group differences in any cognitive or neuropsychiatric measures. Further details on longitudinal changes in cognitive measures can be found in Supplementary Table 1 and neuropsychiatric measures in Supplementary Table 2. Cross-sectional changes in cognitive measures are presented in Supplementary Table 3 and neuropsychiatric changes in Supplementary Table 4. Cross-sectional findings are visualized in Supplementary Fig. 1. AMI, Apathy Motivation Index; BIS, Barratt Impulsivity Scale; CI, confidence interval; ED, extra dimensional; FSBS, Frontal Systems Behavioral Scale; IED, intra–extra-dimensional set shifting; OCI, Obsessive-Compulsive Inventory; OTS, One Touch Stockings; PAL, paired associates learning; PSQI, Pittsburgh Sleep Quality Index; RVP, rapid visual processing; RVP A', a signal detection theory measure of target sensitivity and mean response latency; SDMT, Symbol Digit Modalities Test; SF-36, 36-item self-report survey; SSRT, stop-signal reaction time; SSTAI, Speilberger State-Trait Anxiety Inventory; SWM, spatial working memory; ZSDS, Zung Self-rating Depression Score.

From a significantly increased baseline, CSF NfL (Fig. 3a) and CSF YKL-40 (also known as chitinase-3 like-protein-1 (CHI3L1)) (Fig. 3c) rose more rapidly in HDGE compared to controls ($P = 3.2 \times 10^{-13}$, FDR = 3.2×10^{-12} and $P = 0.01$, FDR = 0.056, respectively). New to this timepoint, proenkephalin (PENK), a surrogate marker for striatal MSN state, measured in CSF, showed a significant longitudinal reduction in HDGE individuals compared to controls ($P = 4.4 \times 10^{-4}$, FDR = 2.6×10^{-3}; Fig. 3b). An increase in plasma NfL was nonsignificant ($P = 0.336$, FDR = 0.669; Extended Data Fig. 2).

Cross-sectionally, log concentrations of both CSF NfL ($P = 5.4 \times 10^{-30}$, FDR = 6.5×10^{-29}) and PENK ($P = 1.7 \times 10^{-7}$, FDR = 1.0×10^{-6}) were highly associated with age, CAG length and their interaction. There was also evidence for an influence on longitudinal change in CSF NfL ($P = 0.027$, FDR = 0.322) and PENK ($P = 0.0547$, FDR = 0.328). Plasma NfL had a similar cross-sectional association ($P = 7.2 \times 10^{-7}$, FDR = 2.9×10^{-6}) but no significant longitudinal association with age and CAG. Regression coefficients are reported in Supplementary Tables 12–14.

Slightly higher annualized rates of change in NfL in CSF and plasma were observed in the HDGE group at stage 0 compared to stage 1 on follow-up but did not reach the threshold of significance (FDR > 0.15). Mean CSF NfL levels (across both visits) were higher in HD-ISS progressor (stages 0 to 1—mean = 6.89 pg ml^{-1}, log scale) compared to nonprogressors (stage 0 to 0—mean = 6.11 pg ml^{-1}, log scale; stage 1 to 1—mean = 6.37 pg ml^{-1}, log scale; Supplementary Table 15). After adjusting for age, sex and their interaction, the difference between stage 0 to 1 progressors and stage 0 nonprogressors was statistically significant ($P = 0.0004$). Similarly, the difference between stage 0 to 1 progressors and stage 1 nonprogressors was significant ($P = 0.045$) when controlling for age and sex, but nonsignificant without these adjustments. No significant differences were observed for plasma NfL levels (Supplementary Table 16).

CSF mHTT levels were notably very low than later disease stages[27], with only 38.3% ($n = 41/107$) of samples exceeding the lower limit of quantification and demonstrating an acceptable coefficient of variation below 30% (Supplementary Fig. 2).

The rate of change in other biofluid markers, including plasma NfL, CSF and plasma tau, CSF and plasma glial fibrillary acidic protein (GFAP), CSF and plasma ubiquitin carboxyl-terminal hydrolase L1 (UCH-L1), and CSF interleukin-6 (IL-6) and IL-8, showed no significant differences between groups (Extended Data Fig. 2). Additionally, none of the fluid biomarkers, including NfL, had an association with age, CAG or age-by-CAG interaction (FDR > 0.15). See Supplementary Table 17 for longitudinal and Supplementary Table 18 for cross-sectional summary statistics.

Somatic expansion ratios in blood

Significant longitudinal increases in the somatic expansion ratio (SER) were detected in blood DNA in the HDGE group over 4.5 years ($P = 2.0 \times 10^{-8}$), with SER clearly increasing as early as HD-ISS stage 0 (Fig. 4a). SER rates of change were strongly influenced by an accelerating effect of CAG repeat length ($P = 3.0 \times 10^{-5}$).

HTT allele structures

The majority of the HDGE group exhibited the typical *HTT* repeat structure on their expanded allele ($n = 66$, 91.6%), while a small subset ($n = 6$) showed atypical allelic variations (Fig. 5a). Specifically, the CAACAG duplication was observed in 1 (1.4%) participant, the CAACAGCCGCCA double loss was found in 4 (5.6%) and 1 (1.4%) had the CCGCCA loss.

Predictors of progression

Baseline NfL, both plasma and CSF, and CSF PENK were predictors of atrophy over time in all brain regions (all FDR < 0.04), even after controlling for the effect of age and CAG (all FDR < 0.12; Extended Data Table 4). Rate of change in caudate and putamen was most strongly associated with change in CSF NfL ($P = 3.0 \times 10^{-4}$, FDR = 0.003 and $P = 2.2 \times 10^{-4}$, FDR = 0.003, respectively) and plasma NfL ($P = 0.002$, FDR = 0.01 and $P = 0.03$, FDR = 0.06, respectively) and the association remained after controlling for age and CAG effects (all FDR < 0.09). Rates of change in caudate and putamen were also associated with longitudinal change in CSF PENK before ($P = 2.0 \times 10^{-4}$, FDR = 0.003 and $P = 1.0 \times 10^{-4}$, FDR = 0.001, respectively) and after ($P = 0.002$, FDR = 0.021 and $P = 9.0 \times 10^{-4}$, FDR = 0.011, respectively) age-by-CAG correction.

Longitudinal increase in SER was a significant predictor of the rate of subsequent caudate volume change before ($P = 0.01$, FDR = 0.04) and after age-by-CAG correction ($P = 0.03$, FDR = 0.07; Fig. 4b). Longitudinal increase in SER was also a significant predictor of the rate of subsequent putamen volume change before ($P = 0.02$, FDR = 0.07) and after ($P = 0.049$, FDR = 0.148) age-by-CAG correction (Fig. 4c). Baseline SER was strongly associated with cross-sectional levels of CSF NfL ($P = 2.5 \times 10^{-12}$, FDR = 2.9×10^{-11}; Fig. 4d) and CSF PENK ($P = 8.4 \times 10^{-5}$, FDR = 3.4×10^{-4}; Fig. 4e) before age-by-CAG correction. However, these associations did not remain significant after the correction (CSF NfL—$P = 0.827$, FDR = 0.956; CSF PENK—$P = 0.908$, FDR = 0.956).

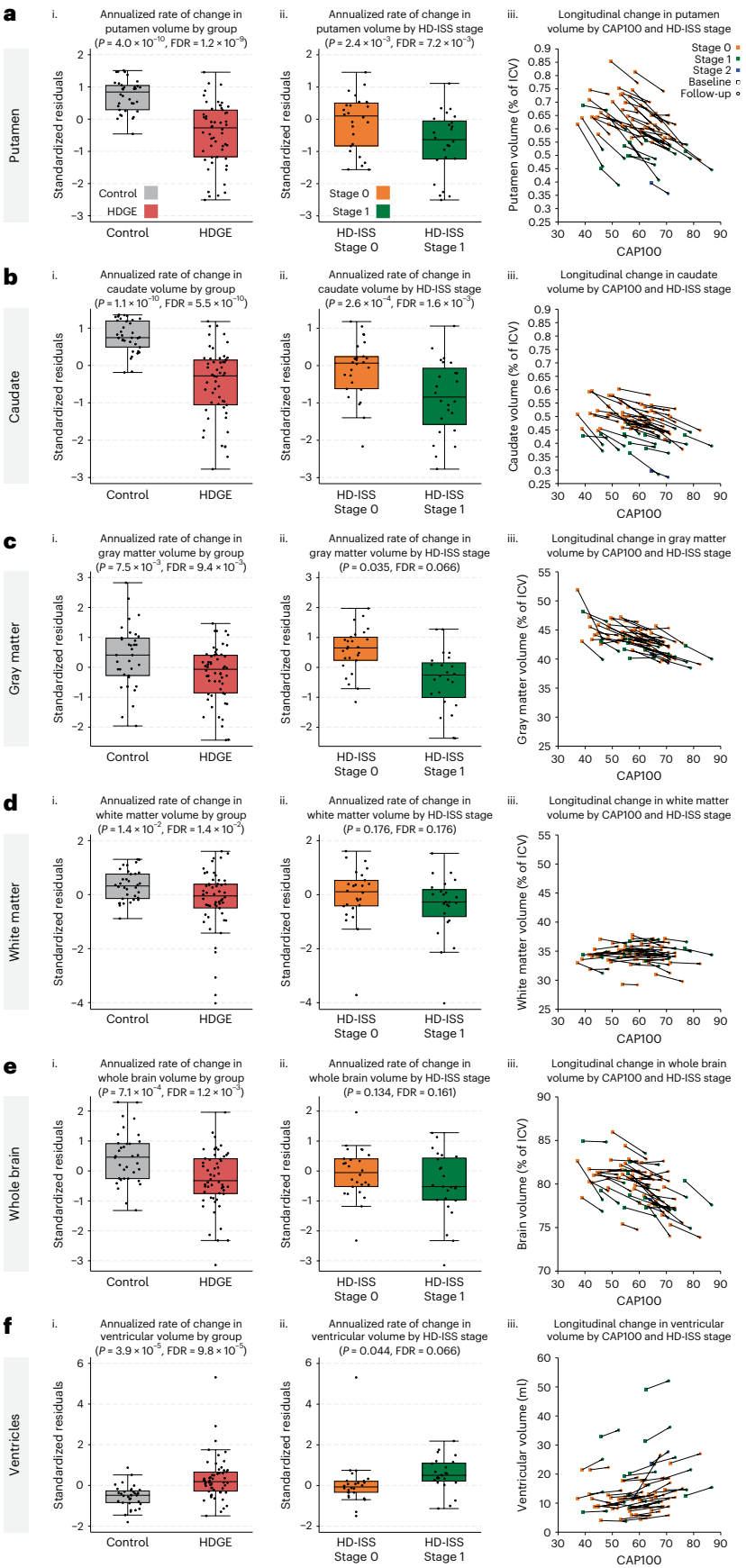
After controlling for CAG, age, age-by-CAG, sex and SER effects, compared to typical allele structure, the loss of CAACAG CCGCCA atypical allele had significant effects on rates of caudate ($P = 1.90 \times 10^{-5}$; Fig. 5b.i) and putamen ($P = 0.007$; Fig. 5b.ii) atrophy as well as cross-sectional CSF NfL ($P = 0.002$; Fig. 5b.iii) and CSF PENK ($P = 0.001$; Fig. 5b.iv) levels, with the loss of the intervening CAACAG CCGCCA associated with an accelerated neurodegenerative course (Extended Data Fig. 3). Notably, after correction for pure CAG length, there was no detectable association between atypical allele structure and SER (FDR > 0.15).

Sample size calculations

Extended Data Table 5 shows hypothetical sample size calculations for those variables with significant longitudinal effects in the HDGE group. For a 50% treatment effect over 2 years in stages 0 and 1, total sample sizes would be 232, 282 and 326 for rates of change in CSF NfL levels, caudate and putamen volume, respectively. For a 3-year trial, these numbers would be reduced to 104, 126 and 146, respectively.

Fig. 2 | Annualized changes in volumetric measures longitudinally. **a**–**f**, Putamen (**a**), caudate (**b**), gray matter (**c**), white matter (**d**), whole brain (**e**) and ventricles (**f**) are shown. For each structure, we present (i) comparison of standardized residuals (age- and sex-adjusted) for the annualized rate of change in HDGE ($n = 54$; red) and control ($n = 34$; gray) groups, (ii) comparison of standardized residuals for annualized rate of change within HDGE by follow-up HD-ISS stage 0 (orange) and stage 1 (green) and (iii) scatterplots of volume by CAP100 score, colored by HD-ISS stage within HDGE. Repeated visits per participant are connected by black lines, with baseline shown as squares and follow-up as circles. HD-ISS stages are represented as follows: stage 0 (orange), stage 1 (green) and stage 2 (blue). Negative standardized residuals denote a rate of change below the adjusted mean across groups. Each box plot displays the median (horizontal line), interquartile range (box) and whiskers extending to 1.5× IQR. Sample sizes (n) reflect biological replicates per group, with $n = 54$ for HDGE and $n = 34$ for controls; data represent longitudinal measures per participant, with no technical replicates. Volumetric change analyses for brain structures, excluding the putamen, used a single boundary-shift integral measure or voxel-based morphometry measure of scan pairs per participant (baseline to follow-up) converted to annual rates and modeled by ordinary least squares regression. Putamen changes were calculated by subtracting baseline MALP-EM segmentations from follow-up segmentations and dividing the result by the follow-up duration. Analysis results and residual adjustments reflect control for baseline age, sex and their interaction. Statistical two-sided group comparisons were adjusted for multiple comparisons using the FDR, with P values, degrees of freedom and confidence limits provided in Extended Data Table 2. CAP, CAG-Age Product; ICV, intracranial volume; IQR, interquartile range.

https://doi.org/10.1038/s41591-024-03424-6



https://doi.org/10.1038/s41591-024-03424-6

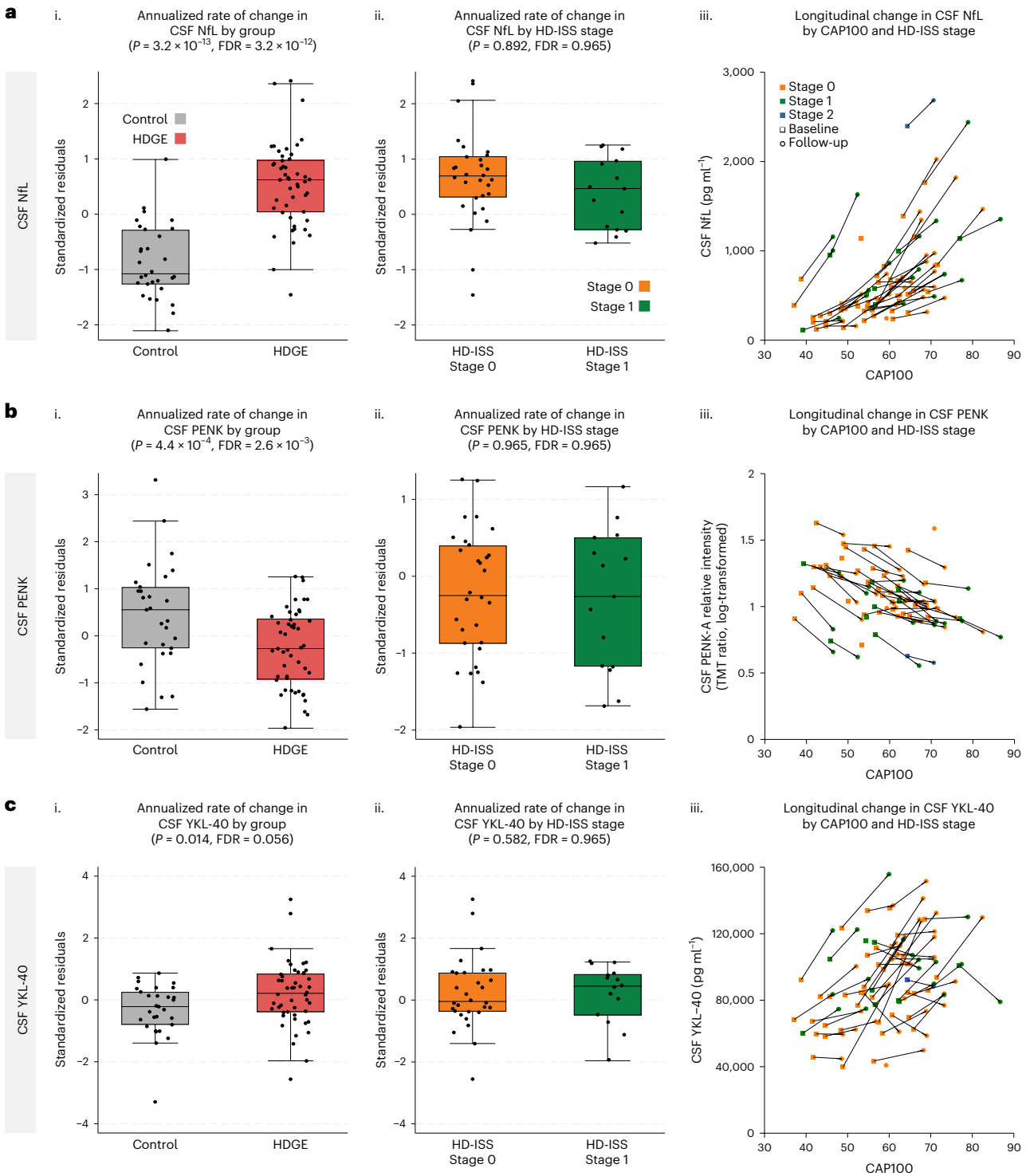


Fig. 3 | Annualized changes in biofluid markers longitudinally. a–**c**, CSF NfL (**a**), CSF PENK (**b**) and CSF YKL-40 (**c**) are shown. For each biofluid biomarker, we present (i) comparison of standardized residuals (age- and sex-adjusted) for the annualized rate of change in HDGE ($n = 48$; red) and control ($n = 30$; gray) groups, (ii) comparison of standardized residuals for annualized rate of change within HDGE by HD-ISS stage 0 (orange) and stage 1 (green) and (iii) scatterplots of biofluid marker levels by CAP100 score, colored by HD-ISS stage within HDGE. Repeated visits per participant are connected by black lines, with baseline shown as squares and follow-up as circles. HD-ISS stages are represented as follows: stage 0 (orange), stage 1 (green) and stage 2 (blue). Negative standardized residuals denote a rate of change below the adjusted mean across groups. Each box plot displays the median (horizontal line), interquartile range (box) and whiskers extending to 1.5× IQR. Sample sizes (n) reflect biological replicates per group, with $n = 48$ for HDGE and $n = 30$ for controls; data represent longitudinal measures per participant. All statistical analyses were conducted using mixed-effect linear models with a participant-specific random effect, controlling for age, sex and their interaction. Natural log-transformed concentrations served as the outcomes in these models. Statistical two-sided group comparisons were adjusted for multiple comparisons using the FDR, with P values, degrees of freedom and confidence limits provided in Supplementary Table 17. Please note one prominent outlier in the control group with marked NfL elevation, as previously reported at baseline[25]. This outlier showed no additional cause on further investigation, with normal T1 MRI brain scan and normal CSF white and red cell counts. Additionally, this control participant did not deviate from other biofluid or cognitive parameters and was, therefore, not excluded from the analysis.

https://doi.org/10.1038/s41591-024-03424-6

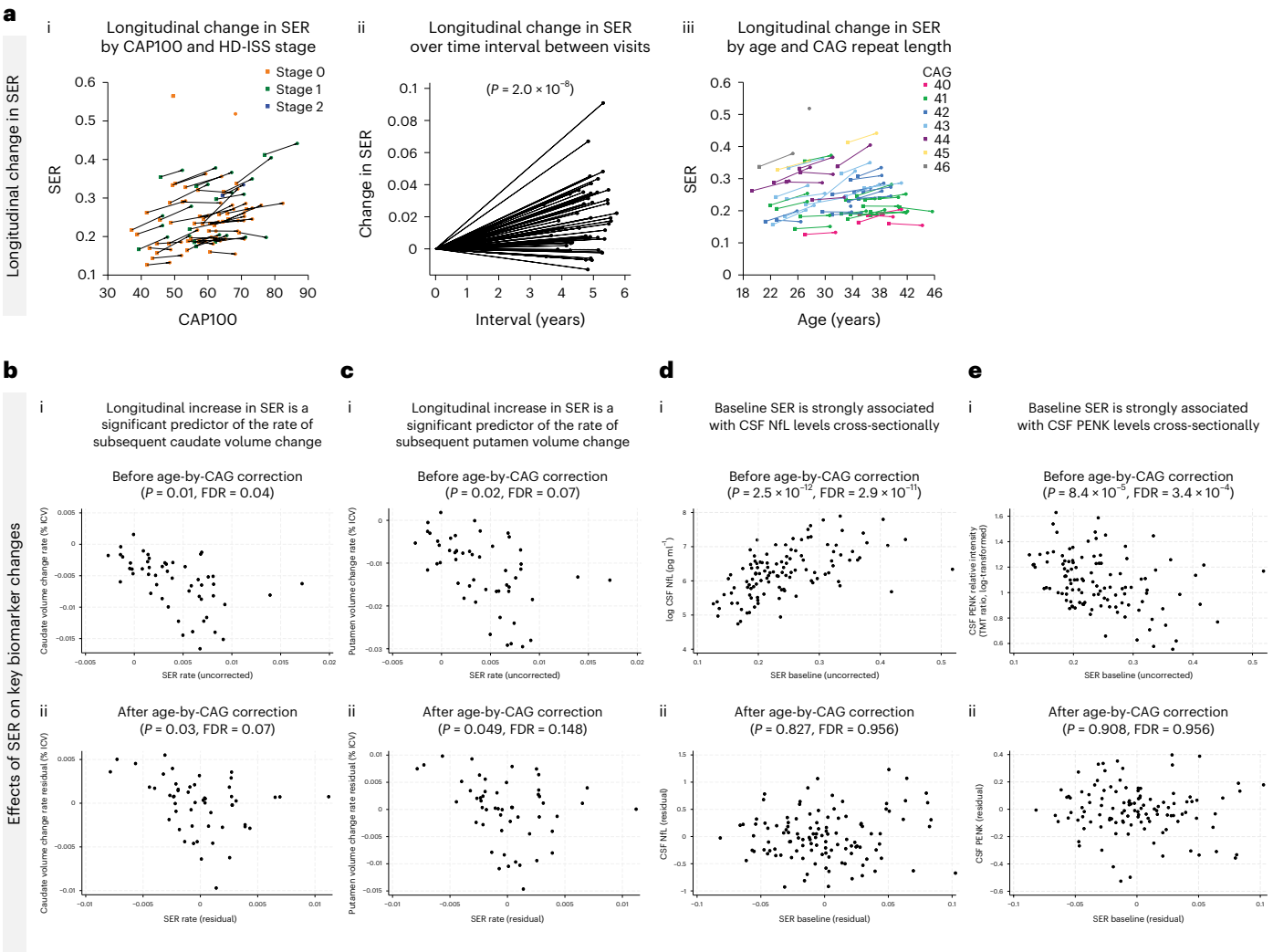


Fig. 4 | Effects of somatic expansion. a, Longitudinal changes in SER—(i) SER trajectories by CAP100 and HD-ISS stage with baseline visits represented by squares and follow-up visits by circles, and lines connecting data from the same individual, where stage 0 is shown in orange, stage 1 in green, and stage 2 in blue; (ii) changes in SER between visits and (iii) changes in SER by age and CAG repeat length. **b**,**c**, Associations between longitudinal SER increase and caudate (**b**) and putamen (**c**) volume change, (i) before and (ii) after age-by-CAG correction. **d**,**e**, Associations of baseline SER with cross-sectional CSF NfL (**d**) and CSF PENK (**e**) levels, with (i) before and (ii) after age-by-CAG correction. Associations were modeled via mixed effects regression using the measure on the vertical axis as the outcome and controlled for age, sex and age-by-sex interaction. Longitudinal caudate change based on a single boundary-shift integral measure per participant was an exception where an analogous ordinary least-squares model was employed.

Discussion

We have used state-of-the-art multimodal measures of cognition, neuroimaging, genetics and biofluid markers in a new assessment battery to study a unique cohort of young adult HDGE who were at baseline, on average, approximately 23 years before predicted clinical motor diagnosis, comparing them to matched controls in an unprecedented level of detail. Our baseline cross-sectional data identified early signs of neurodegeneration despite the maintenance of intact brain function[25] and here we present 4.5-year follow-up data with important new mechanistic insights into what drives neurodegeneration in humans carrying the HD mutation (Fig. 6).

Our data highlight the role of inherited CAG repeat length and somatic expansion on neurodegeneration, decades before clinical motor diagnosis. We identify brain atrophy, elevated levels of CSF NfL, a marker of neuronal damage, and reduced levels of CSF PENK, a marker of striatal MSN state, in the earliest adult HD cohort studied to date. Despite evidence for the start of the neurodegenerative process, there is an absence of any decline in cognitive, motor or neuropsychiatric function at HD-ISS stages 0 and 1. Notably, we show that somatic CAG repeat expansion measured longitudinally in blood, a validated measure of somatic expansion in living patients[14,17], is a predictor of the effect of CAG repeat length on striatal markers of very early neurodegeneration.

Consistent with the elevated levels of CSF NfL we reported at baseline[25], we now show substantially greater rates of increase in CSF NfL in HDGE compared to controls, indicating accelerating neuroaxonal injury from the earliest stages. Most notably, the rate of change in CSF NfL in HD-ISS stage 0 was at least as fast as in stage 1, suggesting rapid neuroaxonal injury increases even before reaching the threshold of caudate or putamen volumetric loss cutoff for stage 1. Interestingly, mean CSF NfL levels were higher in HD-ISS stage 0 to 1 progressors compared to nonprogressors in both stage 0 and stage 1. The annualized rates of increase in CSF NfL across the whole HDGE group (mean = 63.38 pg ml^{-1} yr^{-1}) are slightly lower than those reported in the previous HD-CSF cohort (mean = 79.16 pg ml^{-1} yr^{-1})[27], which is consistent with the HD-YAS cohort being towards the beginning of the neurodegenerative process.

Axonal damage and injury lead to leakage of NfL into the CSF[28–30] and are elevated in active inflammation[31]. NfL is a nonspecific marker of neuronal injury, and elevated levels have been reported in other

https://doi.org/10.1038/s41591-024-03424-6

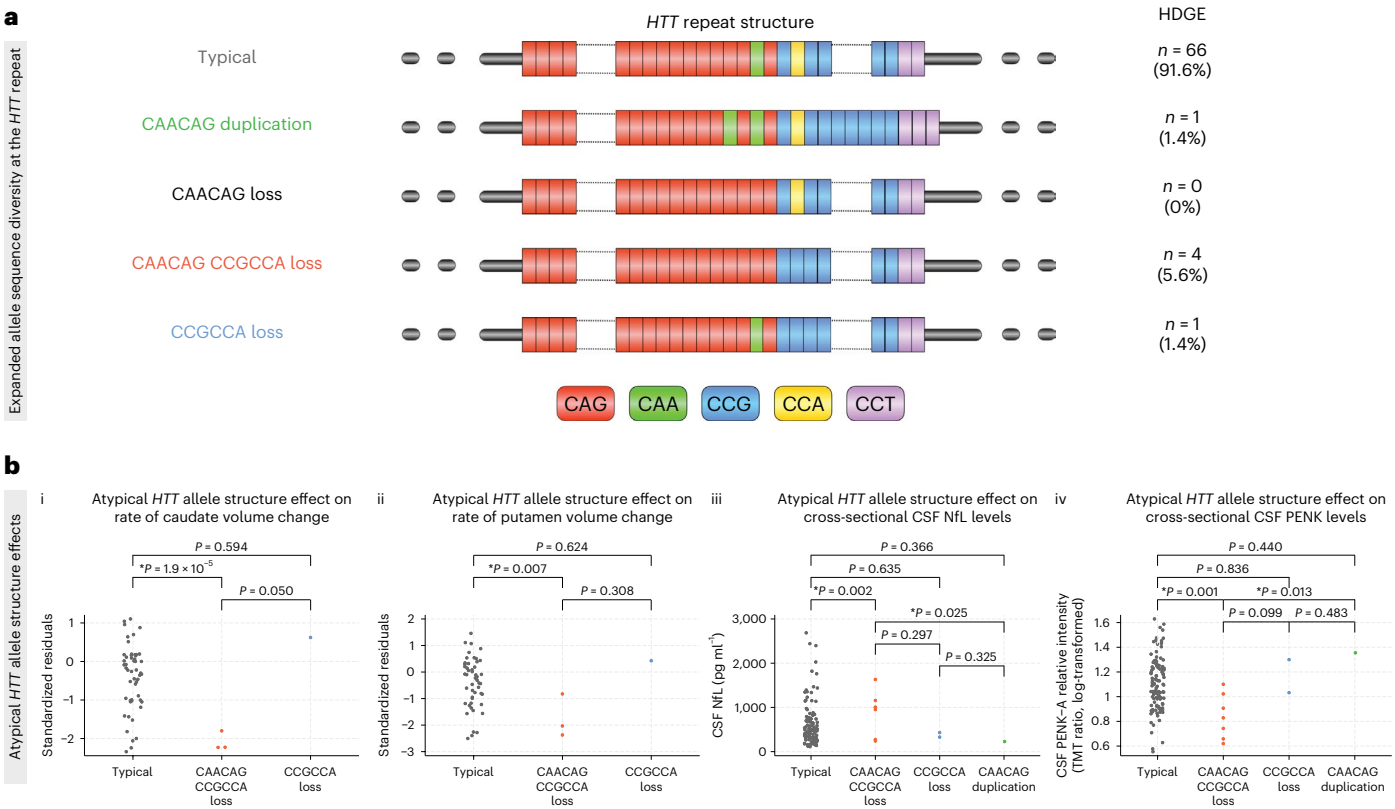


Fig. 5 | Effects of CAG architecture and allelic variants. a, Illustration of the *HTT* repeat structure and allelic variations in the HDGE cohort ($n = 72$), including the typical structure and four atypical variants—CAACAG duplication (green), CAACAG loss (black, not observed in cohort), CAACAG CCGCCA loss (red) and CCGCCA loss (blue). **b**, Illustration of atypical allele differences for key biomarker measures. The ANOCOVA models controlled for CAG length, sex and age, including age interactions with CAG and sex. Participant-specific random effects were included except for caudate change based on one boundary-shift integral per participant. (i) The effect on caudate volume change, with significant differences between typical alleles and CAACAG CCGCCA loss ($P < 0.0001$) and a trend with CCGCCA loss ($P = 0.050$). (ii) The effect on putamen volume change, with significant differences between typical alleles and CAACAG CCGCCA loss ($P = 0.007$). (iii and iv) Cross-sectional effects on CSF NfL (iii) and CSF PENK (iv) levels, respectively, with significant differences noted for CAACAG CCGCCA loss. Statistically significant comparisons ($P < 0.05$) are indicated by asterisks. Please note no longitudinal imaging for CAACAG duplication, hence no plot in (i) or (ii).

neurodegenerative conditions[32–39]. Increases in CSF NfL are not necessarily attributable to neuronal death and could result from other degenerative processes such as leaky axons. Nevertheless, it is a clear marker of neuroaxonal pathology and therefore understanding CSF NfL temporal dynamics and kinetics can provide valuable insights into mechanisms in neurodegenerative diseases[29].

Previously, cross-sectional studies have revealed lower levels of CSF PENK in manifest HD compared to other neurodegenerative conditions[40], as well as compared to HDGE before clinical motor diagnosis and controls[41,42]. Our longitudinal findings in a larger cohort, and our demonstration of a significant association between PENK levels and striatal imaging measures, serve to substantially strengthen the rationale for using PENK as a surrogate marker for striatal MSN state.

Astrocytes are implicated in disease processes through both cell-autonomous and non-cell-autonomous mechanisms[43,44], with one key study identifying a core signature of astrocyte genes with expression altered by mHTT in both humans and mouse models[44]. A recent study provided the first evidence of mHTT-induced alterations in basal pro-inflammatory cytokine production in microglia without immune stimulation, along with a reduction in endocytic and phagocytic activity in mHTT-bearing microglia under basal conditions, suggesting a possible role for microglial cell-autonomous inflammation and activity in the early stages of HD[45]. Consistent with our previous findings of elevated microglial marker CSF YKL-40 levels at baseline[25], we now show greater rates of increase in CSF YKL-40 longitudinally in the HDGE group compared to controls. However, we do not observe significant longitudinal changes in pro-inflammatory cytokine markers IL-6 and IL-8, which are components of the innate immune system, nor in GFAP, an intermediate filament protein of astrocytes associated with astroglial activation[46]. It is known that mHTT is expressed in microglia[47] and that microglial activation correlates with severity later in the disease[48], where mHTT-induced dysfunction of central nervous system (CNS) immune cells is closely linked to pathogenesis[49]. We postulate that the isolated elevation of YKL-40 may be due to both cell-autonomous and non-cell-autonomous mechanisms at play with activation driven by mHTT dysregulation of astrocytes, rather than general gliosis, which would be additionally indicated by a concomitant rise in GFAP. Our findings suggest that astrocytic dysfunction is more prominent than any abnormal innate immune response at this stage of the disease, as IL-6 and IL-8 levels, which are upregulated in HD and correlate with disease progression[49,50], remained unchanged longitudinally, reinforcing the importance of treating early at this stage, before widespread neuroinflammation occurs.

The presence of neuronal damage within HD-ISS stage 0 is further supported by the evidence of substantially elevated rates of brain atrophy and a corresponding reduction in CSF PENK levels. Stages 0 to 1 progressors also had substantially higher elevations in CSF NfL than stage 0 nonprogressors. The substantially higher rates of caudate and putamen atrophy and global brain measures and their association with disease burden suggest that neurodegenerative processes are already occurring across our cohort and at the earliest ages observed in this study. This atrophy was measurable in those with basal ganglia

https://doi.org/10.1038/s41591-024-03424-6

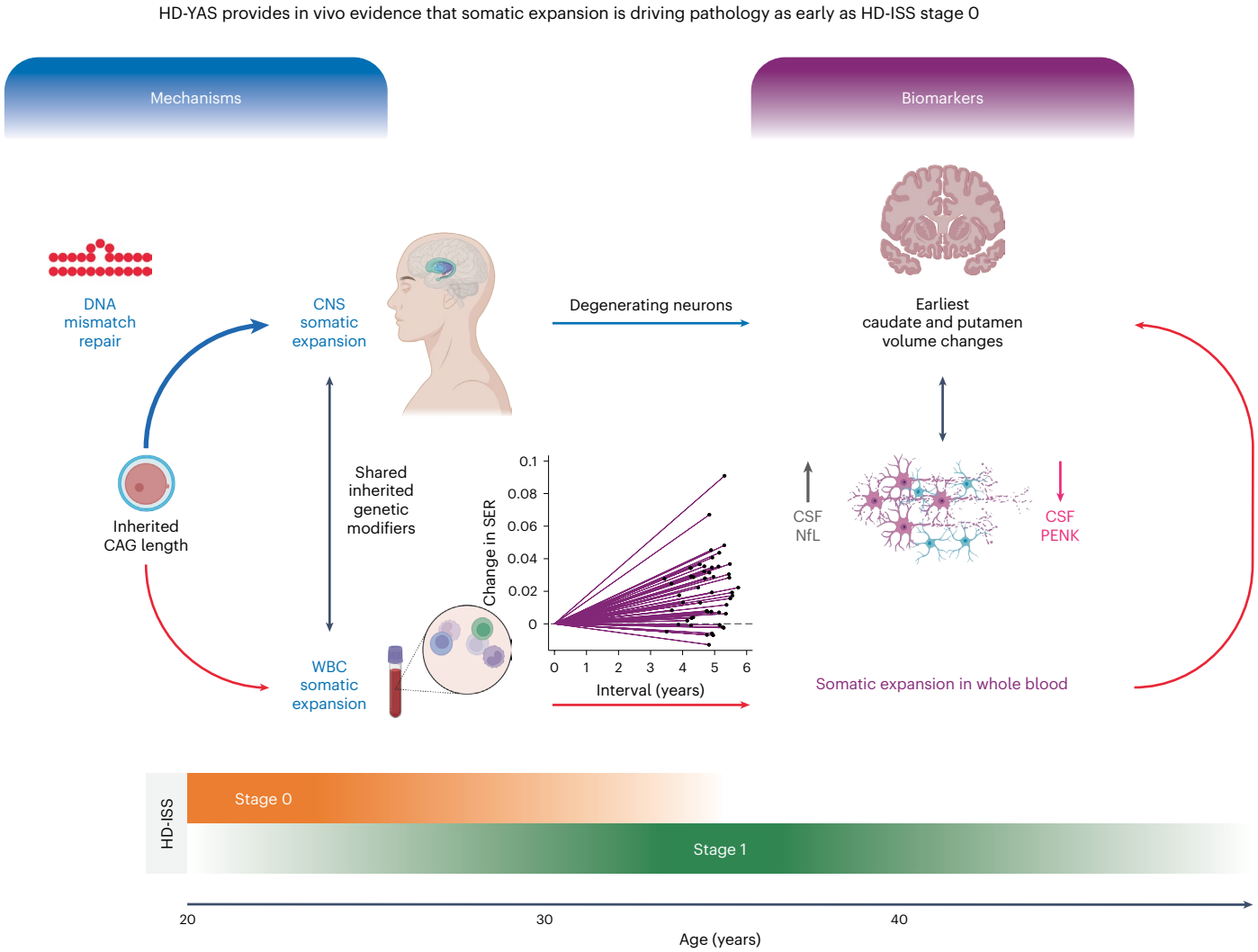


Fig. 6 | Graphical abstract. This graphical abstract illustrates the proposed pathways linking somatic expansion and its effects on biomarkers in HD-YAS. Inherited CAG repeat length is identified as the primary driver of disease progression in HD[57]. Red arrows represent observed data associations in HD-YAS and blue arrows reflect assumed causal relationships. The black bidirectional arrow under mechanisms indicates somatic expansion in WBCs as a proxy for CNS expansion, based on shared inherited genetic modifiers[14,17]. The black bidirectional arrow under biomarkers shows associations between elevated CSF NfL (marker of neuroaxonal damage) and reduced CSF PENK (surrogate of striatal MSN state) with the earliest caudate and putamen volume changes. Somatic expansion is influenced by inherited CAG repeat length, age and DNA mismatch repair gene variants[14,17]. DNA mismatch repair, highlighted in the schematic and shown as a repeat loop-out mismatch icon, is a key mechanism linking inherited CAG repeat length to somatic expansion[16] with repair activity increasing with longer CAG repeat lengths[58]. Within the CNS, somatic expansions substantially contribute to disease progression, as supported by recent postmortem research[8]. While brain biopsy would be the gold standard for direct assessment of CNS somatic expansion in vivo, WBC-derived somatic expansion is detectable peripherally, showing early and longitudinal changes by HD-ISS stages 0 and 1. The biomarkers section shows associations of somatic expansion with the earliest caudate and putamen volume changes, and CSF NfL and PENK levels. The age continuum from HD-ISS stage 0 to stage 1 illustrates early detection of somatic expansion and biomarker changes, influencing pathology from stage 0 onward. HD-YAS provides in vivo evidence that somatic expansion drives early pathology during these early stages, highlighting its potential as a promising therapeutic target in proof-of-concept clinical trials at HD-ISS stages 0 and 1 to slow or prevent further neurodegeneration before clinical motor diagnosis. WBC, white blood cell. The figure is created with BioRender.com.

volumes distributed throughout the volume range observed in unaffected controls, implying the beginning of detectable neurodegeneration. In addition to these changes seen at the macrostructural level, diffusion imaging provides evidence that there is ongoing very early microstructural white matter damage. The strong predictive power of baseline NfL (in both plasma and CSF) for subsequent atrophy in all brain regions further supports the suggestion that there is early neuroaxonal damage which leads to macroscopic effects such as brain atrophy.

Despite the evidence of ongoing pathological changes in our stages 0 and 1 cohort, neurodegeneration is not yet impacting measurable function as we saw no significant disease-related decline in any of the cognitive, neuropsychiatric or functional measures. Previous work has shown that such changes only become evident from HD-ISS stage 2 (ref. 51).

We demonstrate the accumulation of somatic expansion of the *HTT* CAG repeat in blood DNA over time in HD-ISS stages 0 and 1 and, critically, show that it is associated with both brain atrophy and CSF NfL, a marker of neuronal–axonal injury, and CSF PENK, a surrogate marker of striatal MSN state. A higher inherited CAG length was associated with a faster increase in SER over time. SER was associated with caudate and putamen atrophy, both cross-sectionally and longitudinally, even after controlling for age-by-CAG interactions. Baseline SER was strongly associated with cross-sectional levels of CSF NfL and CSF PENK before

age-by-CAG correction; however, these associations did not remain significant after the correction. We postulate that bioassay measurements demonstrate higher variability and noise compared to striatal volume measurements. Therefore, the lack of significance in associations with CSF NfL and PENK does not undermine the significant association between the longitudinal increase in SER and volumetric changes in the caudate and putamen. Additionally, the statistical strength of the influence of CAG length on atrophy was weakened in models also controlling for blood SER. Assuming that, via the common baseline CAG length effects and shared genetic modifiers, SER measured in blood is an indirect quantifiable indicator of the greater somatic expansion occurring in neurons, these results may be seen as providing in vivo evidence for the key role of somatic CAG repeat expansion in very early HD pathology in humans (Extended Data Fig. 4), reinforcing the putative pathological role of somatic expansion as a critical factor in disease progression[5–9].

If the recent suggestion from HD postmortem brains that asynchronous somatic expansion leads to asynchronous stochastic crossing of the transcriptional dysregulation threshold and asynchronous neuronal death[8] is correct, then our data would support the hypothesis that somatic expansion is already an active process in the brain and that some neurons have already crossed a critical repeat length threshold ~20 years before clinical motor diagnosis. Indeed, this phenomenon is both predicted by the stochastic models and consistent with autopsy observations of early neuronal loss[8]. This would suggest that suppressing CAG repeat somatic expansion from this point in the disease process could prevent additional neurons from passing the neuronal toxicity threshold and reduce neurodegeneration before functional deficits are manifest. Therapeutic agents targeting DNA repair proteins that modify somatic expansion show great potential, with MSH3 as a particularly attractive target for HD and other repeat expansion disorders[52], and various MSH3-targeting therapeutics are currently under development[53,54]. To this end, somatic expansion of CAG repeats in blood DNA could be a useful biomarker to demonstrate target engagement of somatic expansion-suppressing therapies with peripheral exposure.

Within our cohort, a small number of individuals carried atypical CCGCCA or CAACAG CCGCCA loss of intervening sequence *HTT* alleles. These atypical structures have a high potential to cause mid-estimation of the CAG repeat length[13,14,17–19], and using the MiSeq-derived CAG lengths changed the mean baseline years to predicted clinical motor diagnosis in HD-YAS from 24 to 23 years. After correcting for pure CAG length, these structures have previously been associated with earlier clinical motor diagnosis[13,14,17–19]. Consistent with this, we find those participants with the loss of intervening sequence structures exhibit higher rates of caudate and putamen atrophy, and have some of the greatest elevations in CSF NfL and reductions in CSF PENK, which together suggest an acceleration of the degenerative process (Fig. 5b). Detecting these effects in such small numbers so early in the course of disease suggests these synonymous DNA structural differences are exerting a substantial influence on the rate of neuropathological change. Interestingly, after correcting for pure inherited CAG there was no residual association between these allele structure variants and SER. This is consistent with previous work in other cohorts in blood, postmortem brains and cell lines[14,17,19] showing that the loss of the intervening CAACAG CCGCCA does not increase the rate of CAG expansion over and above the effects of pure CAG length. Relevant available brain data is limited so it is still possible that the CAACAG CCGCCA loss increases CAG expansions in brain but not blood. An alternative hypothesis is that, after correcting for pure CAG, the residual disease-modifying mechanism of the CAACAG CCGCCA loss is independent of somatic expansion of the *HTT* repeat via effects on RNA transcription, RNA stability, or canonical or repeat-associated non-ATG translation (Extended Data Fig. 3). Regardless, these variants clearly have a profound impact on the disease course.

This work not only provides evidence to support the potential of therapies targeting somatic expansion but also identifies robust markers of disease progression, which may have utility as likely surrogates for future preventative clinical trials. CSF NfL, PENK and brain atrophy measures have the potential to monitor disease progression in HD-ISS stages 0 and 1, where clinical endpoints are not applicable. Change in CSF NfL level has previously been used as an outcome measure for a trial of the ASO nusinersen[55] in children with spinal muscular atrophy. Earlier treatment initiation was also associated with a larger decrease in CSF NfL levels, underscoring the importance of early intervention to preserve neuronal health.

At this stage of the disease, CSF mHTT levels are very low, with only 38.3% of samples in the HDGE group exceeding the detection level. These findings underscore the limitations of available CSF mHTT assays and confirm there is an urgent need for a reliable assay capable of detecting very low concentrations of mHTT in HDGE, ideally at attomolar levels, if HTT-lowering therapies are to be pursued in stage 0 and 1 HDGE cohorts.

Our extensive phenotypic characterization of HD-ISS stages 0 and 1 may allow us to enrich recruitment for future preventative trials. For example, we demonstrate that baseline NfL and PENK levels predict subsequent brain atrophy, and the potential to establish cutoffs for enriching HD-ISS stage 0 based on these biofluids holds significant promise. Harmonization of HD-YAS with existing cohorts across the disease spectrum such as HD-CSF and HDClarity (ClinicalTrials.gov: NCT02855476) will help to establish reliable cutoffs for inclusion. Another important consideration in clinical trial design is that atypical repeat structures, although infrequent, substantially affect disease progression and may additionally impact therapeutic efficacy. Identification of these rare cases through MiSeq will be important to control for these effects and more accurately assess treatment efficacy.

If these biomarkers can serve as likely surrogate outcomes, sample size calculations suggest feasible numbers for clinical trials in an HD-ISS stage 0/1 cohort given sufficiently large treatment effects. For example, in a clinical trial over 3 years with a 50% treatment effect, 104 participants would be required with CSF NfL as an outcome measure, with 126 for caudate and 146 for putamen atrophy. Notably, the caudate boundary-shift integral measure of change we use here is already well-validated and has previously been used in the laquinimod trial in HDGE with a clinical motor diagnosis[56].

In summary, the results presented strongly support the hypotheses that individual-specific somatic expansion in blood DNA predicts individual-specific somatic expansion in the brain. We show in living participants, decades before clinical motor diagnosis, that somatic expansion of the CAG repeat appears to be an important driver of the earliest pathological disease processes, as evidenced by its association with striatal atrophy rates and CSF NfL and PENK levels. Somatic expansion of repeats underlying disease pathogenesis is likely relevant to many repeat expansion diseases, where similar DNA repair mechanisms may play a role. With new therapies in development to target the DNA repair proteins that are known to influence somatic expansion, our results are timely in demonstrating its association with measurable disease markers. By intervening with therapies targeting somatic CAG repeat expansion at the start of the neurodegenerative process, that is, HD-ISS stages 0 and 1 decades before clinical motor diagnosis, while function remains intact, there is the very real possibility that treatments can delay or even prevent the appearance of clinical signs. To this end, we have identified robust measures of early pathology with potential to act as possible biomarker surrogates of disease progression, and identified the ideal cohort for intervention to delay or prevent clinical motor diagnosis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03424-6.

References

1. Vonsattel, J. P. et al. Neuropathological classification of Huntington's disease. *J. Neuropathol. Exp. Neurol.* **44**, 559–577 (1985).
2. Tabrizi, S. J. et al. Potential disease-modifying therapies for Huntington's disease: lessons learned and future opportunities. *Lancet Neurol.* **21**, 645–658 (2022).
3. Lee, J.-M. et al. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* **78**, 690–695 (2012).
4. Monckton, D. G. The contribution of somatic expansion of the CAG repeat to symptomatic development in Huntington's disease: a historical perspective. *J. Huntingtons Dis.* **10**, 7–33 (2021).
5. Kennedy, L. et al. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum. Mol. Genet.* **12**, 3359–3367 (2003).
6. Shelbourne, P. F. et al. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. *Hum. Mol. Genet.* **16**, 1133–1142 (2007).
7. Mätlik, K. et al. Cell-type-specific CAG repeat expansions and toxicity of mutant Huntingtin in human striatum and cerebellum. *Nat. Genet.* **56**, 383–394 (2024).
8. Handsaker, R. E. et al. Long somatic DNA-repeat expansion drives neurodegeneration in Huntington disease. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.17.592722 (2024).
9. Swami, M. et al. Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.* **18**, 3039–3047 (2009).
10. Kaplan, S., Itzkovitz, S. & Shapiro, E. A universal mechanism ties genotype to phenotype in trinucleotide diseases. *PLoS Comput. Biol.* **3**, e235 (2007).
11. Donaldson, J. et al. What is the pathogenic CAG expansion length in Huntington's disease? *J. Huntingtons Dis.* **10**, 175–202 (2021).
12. Hong, E. P. et al. Huntington's disease pathogenesis: two sequential components. *J. Huntingtons Dis.* **10**, 35–51 (2021).
13. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. CAG repeat not polyglutamine length determines timing of Huntington's disease onset. *Cell* **178**, 887–900.e14 (2019).
14. Ciosi, M. et al. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine* **48**, 568–580 (2019).
15. Moss, D. J. H. et al. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol.* **16**, 701–711 (2017).
16. Rajagopal, S. et al. Genetic modifiers of repeat expansion disorders. *Emerg. Top. Life Sci.* **7**, 325–337 (2023).
17. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium & Lee, J.-M. et al. Genetic modifiers of somatic expansion and clinical phenotypes in Huntington's disease reveal shared and tissue-specific effects. Preprint at *bioRxiv* https://doi.org/10.1101/2024.06.10.597797 (2024).
18. Wright, G. E. B. et al. Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. *Am. J. Hum. Genet.* **104**, 1116–1126 (2019).
19. Dawson, J. et al. A probable *cis*-acting genetic modifier of Huntington disease frequent in individuals with African ancestry. *HGG Adv.* **3**, 100130 (2022).
20. Tabrizi, S. J. et al. Targeting huntingtin expression in patients with Huntington's disease. *N. Engl. J. Med.* **380**, 2307–2316 (2019).
21. McColgan, P. et al. Tominersen in adults with manifest Huntington's disease. *N. Engl. J. Med.* **389**, 2203–2205 (2023).
22. Estevez-Fraga, C., Tabrizi, S. J. & Wild, E. J. Huntington's disease clinical trials corner: March 2024. *J. Huntingtons Dis.* **13**, 1–14 (2024).
23. Paulsen, J. S. et al. Detection of Huntington's disease decades before diagnosis: the predict-HD study. *J. Neurol. Neurosurg. Psychiatry* **79**, 874–880 (2008).
24. Tabrizi, S. J. et al. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol.* **11**, 42–53 (2012).
25. Scahill, R. I. et al. Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington's disease Young Adult Study (HD-YAS): a cross-sectional analysis. *Lancet Neurol.* **19**, 502–512 (2020).
26. Tabrizi, S. J. et al. A biological classification of Huntington's disease: the Integrated Staging System. *Lancet Neurol.* **21**, 632–644 (2022).
27. Rodrigues, F. B. Mutant huntingtin and neurofilament light have distinct longitudinal dynamics in Huntington's disease. *Sci. Transl. Med.* **12**, eabc2888 (2020).
28. Shaw, G. et al. Hyperphosphorylated neurofilament NF-H is a serum biomarker of axonal injury. *Biochem. Biophys. Res. Commun.* **336**, 1268–1277 (2005).
29. Paterson, R. W. et al. SILK studies—capturing the turnover of proteins linked to neurodegenerative diseases. *Nat. Rev. Neurol.* **15**, 419–427 (2019).
30. Khalil, M. et al. Neurofilaments as biomarkers in neurological disorders. *Nat. Rev. Neurol.* **14**, 577–589 (2018).
31. Zetterberg, H. et al. Neurochemical aftermath of amateur boxing. *Arch. Neurol.* **63**, 1277–1280 (2006).
32. Sjögren, M. et al. Cytoskeleton proteins in CSF distinguish frontotemporal dementia from AD. *Neurology* **54**, 1960–1964 (2000).
33. Hall, S. et al. Accuracy of a panel of 5 cerebrospinal fluid biomarkers in the differential diagnosis of patients with dementia and/or parkinsonian disorders. *Arch. Neurol.* **69**, 1445–1452 (2012).
34. Delaby, C. et al. Differential levels of neurofilament light protein in cerebrospinal fluid in patients with a wide range of neurodegenerative disorders. *Sci. Rep.* **10**, 9161 (2020).
35. Bech, S. et al. Amyloid-related biomarkers and axonal damage proteins in parkinsonian syndromes. *Parkinsonism Relat. Disord.* **18**, 69–72 (2012).
36. Feneberg, E. et al. Multicenter evaluation of neurofilaments in early symptom onset amyotrophic lateral sclerosis. *Neurology* **90**, e22–e30 (2018).
37. Abdo, W. F. et al. CSF neurofilament light chain and tau differentiate multiple system atrophy from Parkinson's disease. *Neurobiol. Aging* **28**, 742–747 (2007).
38. Dhiman, K. et al. Cerebrospinal fluid neurofilament light concentration predicts brain atrophy and cognition in Alzheimer's disease. *Alzheimers Dement.* **12**, e12005 (2020).
39. Vrillon, A. et al. Comparison of CSF and plasma NfL and pNfH for Alzheimer's disease diagnosis: a memory clinic study. *J. Neurol.* **271**, 1297–1310 (2024).
40. Barschke, P. et al. Cerebrospinal fluid levels of proenkephalin and prodynorphin are differentially altered in Huntington's and Parkinson's disease. *J. Neurol.* **269**, 5136–5143 (2022).
41. Niemela, V. et al. Proenkephalin decreases in cerebrospinal fluid with symptom progression of Huntington's disease. *Mov. Disord.* **36**, 481–491 (2021).

42. Caron, N. S. et al. Cerebrospinal fluid biomarkers for assessing Huntington disease onset and severity. *Brain Commun.* **4**, fcac309 (2022).
43. Ilieva, H., Polymenidou, M. & Cleveland, D. W. Non-cell autonomous toxicity in neurodegenerative disorders: ALS and beyond. *J. Cell Biol.* **187**, 761–772 (2009).
44. Diaz-Castro, B. et al. Astrocyte molecular signatures in Huntington's disease. *Sci. Transl. Med.* **11**, eaaw8546 (2019).
45. Stöberl, N. et al. Mutant huntingtin confers cell-autonomous phenotypes on Huntington's disease iPSC-derived microglia. *Sci. Rep.* **13**, 20477 (2023).
46. Von Bartheld, C. S., Bahney, J. & Herculano-Houzel, S. The search for true numbers of neurons and glial cells in the human brain: a review of 150 years of cell counting. *J. Comp. Neurol.* **524**, 3865–3895 (2016).
47. Shin, J.-Y. et al. Expression of mutant huntingtin in glial cells contributes to neuronal excitotoxicity. *J. Cell Biol.* **171**, 1001–1012 (2005).
48. Pavese, N. et al. Microglial activation correlates with severity in Huntington disease: a clinical and PET study. *Neurology* **66**, 1638–1643 (2006).
49. Björkqvist, M. et al. A novel pathogenic pathway of immune activation detectable before clinical onset in Huntington's disease. *J. Exp. Med.* **205**, 1869–1877 (2008).
50. Wild, E., Björkqvist, M. & Tabrizi, S. J. Immune markers for Huntington's disease? *Expert Rev. Neurother.* **8**, 1779–1781 (2008).
51. Tabrizi, S. J. et al. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol.* **10**, 31–42 (2011).
52. Flower, M. et al. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain* **142**, 1876–1886 (2019).
53. Benn, C. L., Gibson, K. R. & Reynolds, D. S. Drugging DNA damage repair pathways for trinucleotide repeat expansion diseases. *J. Huntingtons Dis.* **10**, 203–220 (2021).
54. O'Reilly, D. et al. Di-valent siRNA-mediated silencing of MSH3 blocks somatic repeat expansion in mouse models of Huntington's disease. *Mol. Ther.* **31**, 1661–1674 (2023).
55. Olsson, B. et al. NFL is a marker of treatment response in children with SMA treated with nusinersen. *J. Neurol.* **266**, 2129–2136 (2019).
56. Reilmann, R. et al. Safety and efficacy of laquinimod for Huntington's disease (LEGATO-HD): a multicentre, randomised, double-blind, placebo-controlled, phase 2 study. *Lancet Neurol.* **23**, 243–255 (2024).
57. Bates, G. P. et al. Huntington disease. *Nat. Rev. Dis. Primers* **1**, 15005 (2015).
58. Ferguson, R. et al. Therapeutic validation of MMR-associated genetic modifiers in a human ex vivo model of Huntington disease. *Am. J. Hum. Genet.* **111**, 1165–1183 (2024).

Rachael I. Scahill[1,21], Mena Farag[1,21], Michael J. Murphy[1], Nicola Z. Hobbs[1], Michela Leocadi[1], Christelle Langley[2], Harry Knights[1], Marc Ciosi[3], Kate Fayer[1], Mitsuko Nakajima[1], Olivia Thackeray[1], Johan Gobom[4,5], John Rönnholm[4], Sophia Weiner[4], Yara R. Hassan[1], Nehaa K. P. Ponraj[3], Carlos Estevez-Fraga[1], Christopher S. Parker[6], Ian B. Malone[7], Harpreet Hyare[8], Jeffrey D. Long[9], Amanda Heslegrave[10,11], Cristina Sampaio[12,13], Hui Zhang[6], Trevor W. Robbins[14], Henrik Zetterberg[4,5,11,15,16], Edward J. Wild[1], Geraint Rees[17], James B. Rowe[18,19,20], Barbara J. Sahakian[2], Darren G. Monckton[3], Douglas R. Langbehn[9] & Sarah J. Tabrizi[1]

[1]Huntington's Disease Centre, Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, UK. [2]Department of Psychiatry, University of Cambridge, Cambridge, UK. [3]School of Molecular Biosciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [4]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy at University of Gothenburg, Mölndal, Sweden. [5]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden. [6]Department of Computer Science and Centre for Medical Image Computing, University College London, London, UK. [7]Dementia Research Centre, Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, UK. [8]Department of Brain Repair and Rehabilitation, UCL Queen Square Institute of Neurology, University College London, London, UK. [9]Department of Psychiatry and Biostatistics, Carver College of Medicine and College of Public Health, University of Iowa, Iowa City, Iowa, USA. [10]Dementia Research Institute, University College London, London, UK. [11]Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, UK. [12]Faculdade Medicina da Universidade de Lisboa (FMUL), Lisbon, Portugal. [13]CHDI Management, Inc. Advisors to CHDI Foundation, Princeton, NJ, USA. [14]Department of Psychology, University of Cambridge, Cambridge, UK. [15]Hong Kong Center for Neurodegeneassociated with substantially fasterrative Diseases, Clear Water Bay, Hong Kong, China. [16]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, University of Wisconsin—Madison, Madison, WI, USA. [17]UCL Institute of Cognitive Neuroscience, University College London, London, UK. [18]Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [19]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [20]Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK. [21]These authors contributed equally: Rachael I. Scahill, Mena Farag. ✉e-mail: s.tabrizi@ucl.ac.uk

Methods

Participant characteristics

Participants were recruited across the UK and enrolled at one study site (University College London (UCL)). The inclusion criteria are detailed in the Supplementary Methods. Participants (131 in total, 64 HDGE and 67 controls) attended at baseline and 103 (57 HDGE and 46 controls) returned for follow-up approximately 4.5 years later.

The study was registered on ClinicalTrials.gov (NCT06391619) where the study protocol and the predefined statistical analysis plan are provided. All participants underwent comprehensive assessment of clinical, cognitive and neuropsychiatric function, neuroimaging, blood sampling, and optional CSF collection consistent with the baseline procedure (Supplementary Methods)[25]. See Extended Data Table 6 for a list of assessments and Supplementary Table 5 for missing data. Supplementary Methods provides further details for all assessments.

Ethics

The study received approval by the London−Bloomsbury Research Ethics Committee (22/LO/0058). All study procedures adhered to principles outlined in the Declaration of Helsinki, and before enrollment, written consent was obtained from all participants.

Clinical, cognitive and neuropsychiatric assessments

A clinical examination was performed including assessment for lumbar puncture suitability.

All the HDGE participants with longitudinal neuroimaging ($n = 54$) were staged according to the HD-ISS[26] at each visit. The longitudinal pipeline of FreeSurfer version 6 (https://surfer.nmr.mgh.harvard.edu/pub/dist/freesurfer/) was used to derive caudate and putamen segmentations to classify stages 0 and 1, and stage 2 was defined by participants reaching the age- and education-adjusted cutoffs for the Symbol Digit Modalities Test and/or Total Motor Score. Additionally, predicted years to clinical motor diagnosis were linked to the standardized CAG-Age-Product (CAP) score. A CAP score of 100 occurs at the CAG-specific expected age of motor diagnosis[59].

All cognitive and neuropsychiatric tasks from baseline were repeated at follow-up with two exceptions. Due to participant feedback, we replaced the Progressive Ratio Task from baseline with the Goals Prior Assay task at follow-up to assess the motivational domain. Additionally, Stroop interference was included within the core cognitive tasks in follow-up. See Supplementary Methods and Supplementary Figs. 3–5 for details of the cognitive and neuropsychiatric testing battery.

Neuroimaging

Scanning was performed on a 3T Siemens Prisma (Siemens Healthineers, Erlangen, Germany) and parameters were consistent between baseline and 4.5-year follow-up. Neuroimaging assessments included volumetric T1-weighted imaging (T1W), DWI and MPM. We applied the same predefined regions of interest (ROI) that were examined at baseline. Volumes were extracted from T1W images and mean values across the relevant ROI were derived for (1) standard DWI and neurite orientation and dispersion density imaging (NODDI) metrics[60], (2) structural connectivity (right-hand dominant participants only) and (3) MPMs.

Longitudinal imaging changes were derived by subtraction of values, except for direct measures of change for some volumetric ROIs. The boundary-shift integral is a direct measure of change between positionally matched (registered) serial images, which is more sensitive to longitudinal change than subtraction[61]. This technique was used for whole brain, ventricles and caudate[62]. Within-participant voxel-compression maps were derived and convolved with gray and white matter maps generated by voxel-based morphometry[24] to estimate volume change within these tissues.

Details of diffusion, structural connectivity and MPM processing pipelines are provided in Supplementary Methods and Supplementary Fig. 6.

Biofluids

Biofluids were collected at baseline and follow-up under the same standardized, well-validated conditions, methods and equipment[63]. To remove potential batch effects, biofluid samples collected at baseline were reanalyzed in parallel with the follow-up samples, employing the assays detailed in Supplementary Table 19. Quantification of analytes was performed blinded to group status.

Measurements in CSF included mHTT, NfL protein, total tau (tau), GFAP, UCH-L1, YKL-40 (also known as chitinase-3 like-protein-1 (CHI3L1)), IL-6 and IL-8. In plasma, NfL, tau, GFAP and UCH-L1 were quantified. The Neurology 4-Plex A (GFAP, NfL, Tau and UCH-L1) was measured in singlicate, yielding the following inter-plate coefficients of variation (%CV): CSF GFAP (3.7%), CSF NfL (8.2%), CSF Tau (11.1%), CSF UCH-L1 (62.8%), plasma GFAP (5.6%), plasma NfL (5.0%), plasma Tau (11.1%) and plasma UCH-L1 (63.8%). The %CVs were calculated from duplicate measurements of internal plate controls made of pooled human plasma. CSF IL-6, IL-8 and YKL-40 were measured in duplicate, with the following %CVs: CSF YKL-40 (10.4%), CSF IL-6 (29.2%) and CSF IL-8 (12.2%). CSF mHTT was measured in triplicate and the relative abundance of CSF PENK was quantified in single measurements.

Unbiased liquid chromatography-mass spectrometry (LC–MS)-based proteomics analysis was performed using the tandem mass tag (TMT) technique[64] to measure relative abundance of CSF PENK. CSF samples were prepared including an initial multi-affinity depletion step to reduce interference by high-abundant blood-derived proteins (Supplementary Methods). Following this step, samples were subjected to reduction and alkylation of cysteine residues, digestion with trypsin and endoproteinase Lys-C and isobaric labeling using TMTpro 18-plex reagents[65] (Thermo Fisher Scientific). TMT multiplex peptide samples were fractionated by high-pH reversed-phase high-performance LC (HPLC)[66,67], and analyzed by nano-HPLC (EasyLC, Thermo Fisher Scientific) coupled to a high-resolution Orbitrap hybrid mass spectrometer (Orbitrap Lumos Tribrid, Thermo Fisher Scientific). Protein identification and data processing for quantification was performed using Proteome Discoverer 2.5 (Thermo Fisher Scientific), and R Statistics. More details are provided in Supplementary Methods.

HTT CAG repeat structure and somatic expansion

DNA was extracted from whole blood using the chemagic 360-D instrument (Perkin Elmer) for automated DNA extraction. The modal length of the pure *HTT* CAG repeat, the *HTT* repeat structure and quantity of blood *HTT* somatic expansions in the HDGE group were determined by ultra-deep amplicon MiSeq sequencing[14,68]. The *HTT* MiSeq reads were processed and genotyped using ScaleHD (v1 (ref. 14)) and RGT (https://github.com/hossam26644/RGT). The SER[14] of the *HTT* CAG repeat was quantified from MiSeq data at baseline and follow-up, allowing for longitudinal assessment of somatic expansion changes during the interval between the two visits. CAG repeat length estimated using MiSeq was used for all statistical analyses of associations.

Somatic expansion mediation models

Within the HDGE group, we estimated statistical associations between white blood cell (WBC) somatic expansion ratios with NfL, PENK and with volumetric brain measures. The analyses were performed with and without correction for age, CAG and age-by-CAG interaction. The models also controlled age, sex and age-by-sex interaction as covariates. We modeled cross-sectional SER versus biomarker relationships via random effects regression as described earlier. We estimated relationships between baseline SER and longitudinal change in SER versus longitudinal change in biomarkers using ordinary least square regression as described above for other models of volumetric direct-change rates. For NfL and putamen volumes, we calculated change rates by subtraction of baseline from follow-up values.

To assess the potential causal implications of SER-biomarker associations, we compared the statistical strength of SER as a predictor of

the biomarker before and after controlling for age and CAG length (Extended Data Fig. 4 illustrates a schematic of the underlying causal reasoning). A substantially weakened SER relationship after age and CAG control would suggest that the associations may be due to mutual influence of CAG length over time without more direct causality. Conversely, we assessed the strength of the age–CAG versus biomarker relationships with and without SER control. Weakened age–CAG associations with a biomarker in the presence of both a significant predictive SER effect on the biomarker and a significant age–CAG association with SER are consistent with an intermediate causal role for CNS somatic expansion (as indirectly assessed by SER in WBC DNA). These statistical comparisons are the essence of statistical assessments of plausible causality in formal causal models. In the simple, cross-sectional case, the degree of mediation is often expressed by the relative reduction in the adjusted correlation or regression coefficient of a single distal variable. However, the analyses here involve the joint mediation of two terms—CAG length and its interaction with age. Mediation is no longer simply quantified by this approach. Hence, we instead emphasize the hypothesis testing aspect of the underlying statistical tests.

We used the same regression methods as for the somatic expansion models to assess associations between log NfL levels and volumetric outcomes. However, we attempted no causal interpretation of NfL versus volumetric relationships. Although these are both HD-related biomarkers, they are considered concurrent indicators of the same pathology and there is no well-justified conception of one of these changes 'causing' the other.

Sample size estimates

Approximate sample size estimates for clinical trials involving equal allocation to one treatment group and a placebo group were calculated based on observed HDGE group versus control longitudinal models. The group differences in longitudinal rates were converted to Cohen's *d* effect sizes using within-group longitudinal standard deviations derived from estimated random effect variances. Assumed treatment effects were then defined as a percent slowing of the group difference in longitudinal rates adjusted for the assumed trial length. These should be considered somewhat optimistic order-of-magnitude estimates. We could not estimate potential increased within-group variance over time (and resultant sample size increase) based on only two observed time points. We did not factor in trial dropout rates. On the other hand, we did not consider potential sample size reduction due to efficient (but perhaps controversial) estimates of treatment-induced slope deviation derived from repeated measures over time, but instead based calculations on net group differences at the end of a trial. See Supplementary Methods for further details.

Statistical analysis

Unless otherwise noted, all statistical analyses were performed in accordance with a prespecified statistical analysis plan. Analyses testing the relationship between caudate and putamen volumes or NfL levels to observed or predicted disease progression were defined as primary hypotheses, with a statistical significance level of $P = 0.05$. Analyses of potential disease associations for the rest of the wide-ranging test battery were considered exploratory and primarily assessed by estimated FDR calculated using the Benjamini–Hochberg method[69]. FDR was calculated separately per measurement domain (cognitive, volumetric imaging, bioassays, etc.) to examine conceptually and technically sensible sets of underlying *P* values. We suggest an FDR threshold of 15% for identifying exploratory results of interest.

Demographic group comparisons were conducted by *t* test for continuous variable, by chi-square test for sex differences and by Mann-Whitney U test for education level. Both the longitudinal and repeated-measure cross-sectional analyses were performed by maximum likelihood random effect regression modeling with (necessarily) only random intercepts per participant. The primary regression predictors of interest were group status (control versus HDGE) and in models testing age and CAG effects, terms for age, MiSeq-derived CAG length and age-by-CAG interactions nested within the HDGE. The joint significance of age and CAG effects were estimated by the maximum likelihood test for differences in model fit. The models also contained covariates for sex, a non-nested age term and sex-by-age interaction. The non-nested age term was included to make HDGE group-specific aging effects estimable. Cognitive models also included International Standard Classification of Education education level and estimated intelligence quotient via the National Adult Reading Test score. Due to the consistent skewness of measured distributions, logarithmic transformations were used on all bioassay concentration measures.

Longitudinal volumetric analyses of brain structures other than putamen relied on a single measure of volume change per participant derived from pairs of baseline and follow-up scans. These changes were converted to annual rates and modeled by ordinary least squares regression with predictor variables analogous to those listed above. All other models of longitudinal change used total values at the two visits as outcomes and longitudinal effects of primary predictors and covariates were estimated by interactions with follow-up time between the visits. Those models also retained all baseline effects for predictors and covariates. To preserve the unbiased estimation of model parameters when data is missing at random, baseline data from participants with no follow-up were included in the model.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We are committed to data sharing while maintaining confidentiality due to the sensitive and potentially identifiable nature of these data. Biofluid samples will not be shared due to the limited amount of material available. The remaining samples will be required for replication for the next HD-YAS visit. Upon reasonable request, data will be made available 24 months after the end of data collection, through application via UCL to the Principal Investigator, Professor Sarah Tabrizi. Researchers will be required to submit a proposal meeting the research criteria and must demonstrate full GDPR compliance. A data access agreement with UCL will be required.

Code availability

All software is freely available with the exception of the in-house MIDAS software used to generate the boundary-shift integral for caudate, whole brain and ventricles. This can be requested from Professor Nick Fox at the Dementia Research Centre, UCL, UK.

References

59. Warner, J. H. et al. Standardizing the CAP score in Huntington's disease by predicting age-at-onset. *J. Huntingtons Dis.* **11**, 153–171 (2022).
60. Zhang, H. et al. NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* **61**, 1000–1016 (2012).
61. Fox, N. C. & Freeborough, P. A. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. *J. Magn. Reson. Imaging* **7**, 1069–1075 (1997).
62. Hobbs, N. Z. et al. Automated quantification of caudate atrophy by local registration of serial MRI: evaluation and application in Huntington's disease. *Neuroimage* **47**, 1659–1665 (2009).
63. Wild, E. J. et al. Quantification of mutant huntingtin protein in cerebrospinal fluid from Huntington's disease patients. *J. Clin. Invest.* **125**, 1979–1986 (2015).

64. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).
65. Li, J. et al. Proteome-wide mapping of short-lived proteins in human cells. *Mol. Cell* **81**, 4722–4735.e5 (2021).
66. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *J. Proteome Res.* **13**, 6176–6186 (2014).
67. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
68. Ciosi, M. et al. Library preparation and MiSeq sequencing for the genotyping-by-sequencing of the Huntington disease *HTT* exon one trinucleotide repeat and the quantification of somatic mosaicism. *Protoc. Exch.* https://doi.org/10.1038/protex.2018.089 (2018).
69. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

Acknowledgements

We are grateful to all the study participants and their families who have so generously given their time to the HD Young Adult Study. We would also like to thank the staff at the Wellcome Centre for Human Neuroimaging (London, UK) and the Leonard Wolfson Experimental Neurology Centre (London, UK). LC–MS analysis was performed at the Proteomics Core Facility, Sahlgrenska Academy, Gothenburg University, with financial support from SciLifeLab and BioMS. We thank Annabelle Coleman and Alexiane Touzé for their help with blood collection and processing, Siti Binte Mohd Ikhsan for help with collecting cognitive and neuropsychiatric data, Peter McColgan for assistance with the structural connectivity pipeline and Davina Hensman-Moss and Emily Gantman for helpful discussions on the paper. Funding for this study was provided by the following: Wellcome Collaborative Award (223082/Z/21/Z to S.J.T., R.I.S., M.F., M.J.M., N.Z.H., M.L., C.L., K.F., T.W.R., J.B.R., B.J.S., D.G.M. and D.R.L.); the CHDI Foundation (a not-for-profit organization dedicated to finding treatments for Huntington's disease) funded CSF collection; the UK Dementia Research Institute (DRI; London, UK) through UK DRI, principally funded by the UK Medical Research Council; University College London Hospital/University College London (London, UK), supported by the National Institute for Health and Care Research (NIHR) University College London Hospitals Biomedical Research Centre; Wellcome Trust (220258 to J.B.R.); Medical Research Council (MC_UU_00030/14 and MR/T033371/1 to J.B.R.) and NIHR Cambridge Biomedical Research Centre (NIHR203312 to J.B.R.). H.Z. is a Wallenberg Scholar and a Distinguished Professor at the Swedish Research Council.

Author contributions

R.I.S., C.S., H.Z., E.J.W., G.R., B.J.S., D.G.M. and D.R.L. gained funding for this study. R.I.S., K.F., E.J.W. and G.R. designed the study. R.I.S., M.F., M.J.M., N.Z.H., K.F. and O.T. contributed to recruitment. R.I.S., N.Z.H., M.L., H.K., C.S.P. and I.B.M. contributed to image acquisition and analysis. M.F., M.J.M., M.N., C.E.F. and Y.H. contributed to clinical assessments. M.F. and M.J.M. contributed to biofluid collection. M.F., J.G., J.R., S.W. and A.H. contributed to biofluid processing and analysis. R.I.S., M.F., M.J.M., M.L., C.L. and K.F. contributed to cognitive and neuropsychiatric assessments. C.L., T.W.R. and B.J.S. contributed to cognitive and neuropsychiatric battery design. M.C. and N.K.P.P. contributed to somatic expansion analysis. K.F. contributed to project management. K.F. and O.T. contributed to biofluid processing. H.H. contributed to the radiological review process. J.D.L. and D.R.L. performed statistical analysis. C.S. contributed to data interpretation. H.Z. designed biofluid processing methods and analysis. J.B.R. contributed to cognitive battery design. D.G.M. contributed to the somatic expansion processing and analysis. D.R.L. wrote the Statistical Analysis Plan. R.I.S., M.F., C.L., M.C., B.J.S., D.G.M., D.R.L. and S.J.T. drafted the paper. S.J.T. was PI of the Wellcome Trust Collaborative Award and had overall responsibility for study design, project management, ethics approval, analysis and drafting the paper. All authors contributed to the interpretation of the data and reviewed the paper.

Competing interests

J.B.R. is supported by the Wellcome Trust (220258), the Medical Research Council (MC_UU_00030/14; MR/T033371/1) and the NIHR Cambridge Biomedical Research Centre (NIHR203312). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. J.B.R. has undertaken paid consultancy for Asceneuron, Astronautx, Astex, ClinicalInk, CumulusNeuro, Curasen, Invivro, Prevail and SVHealth; received research funding unrelated to the current work from AstraZeneca, Lily, GSK and Janssen; and is Chief Scientific Advisor to Alzheimer's Research UK. D.G.M. has been a scientific consultant and/or received honoraria/grants from AMO Pharma, Dyne, F. Hoffman-La Roche, Function Rx, LoQus23, MOMA Therapeutics, Novartis, Ono Pharmaceuticals, Pfizer Pharmaceuticals, Rgenta Therapeutics, Sanofi and Sarepta Therapeutics. J.G. is supported by Alzheimerfonden (AF-980746) and Stiftelsen för Gamla tjänarinnor (2022-01324). J.B.R. has appeared as an expert witness to the Medicines and Healthcare products Regulatory Agency, unrelated to the current work. D.G.M. is on the Scientific Advisory Board of the Myotonic Dystrophy Foundation and EuroDyMA (European Dystrophia Myotonica Association), is a scientific advisor to the Myotonic Dystrophy Support Group and is a vice president for research of Muscular Dystrophy UK. G.R. is a nonexecutive director of UCL Business. D.R.L. is an unpaid academic member of the Critical Path Institute HD-RSC Consortium Coordinating Committee. B.J.S. is a co-inventor of the Cambridge Neuropsychological Test Automated Battery. I.B.M. is an associate editor for Frontiers in Neurology. The other authors declare no competing interests.

Additional information

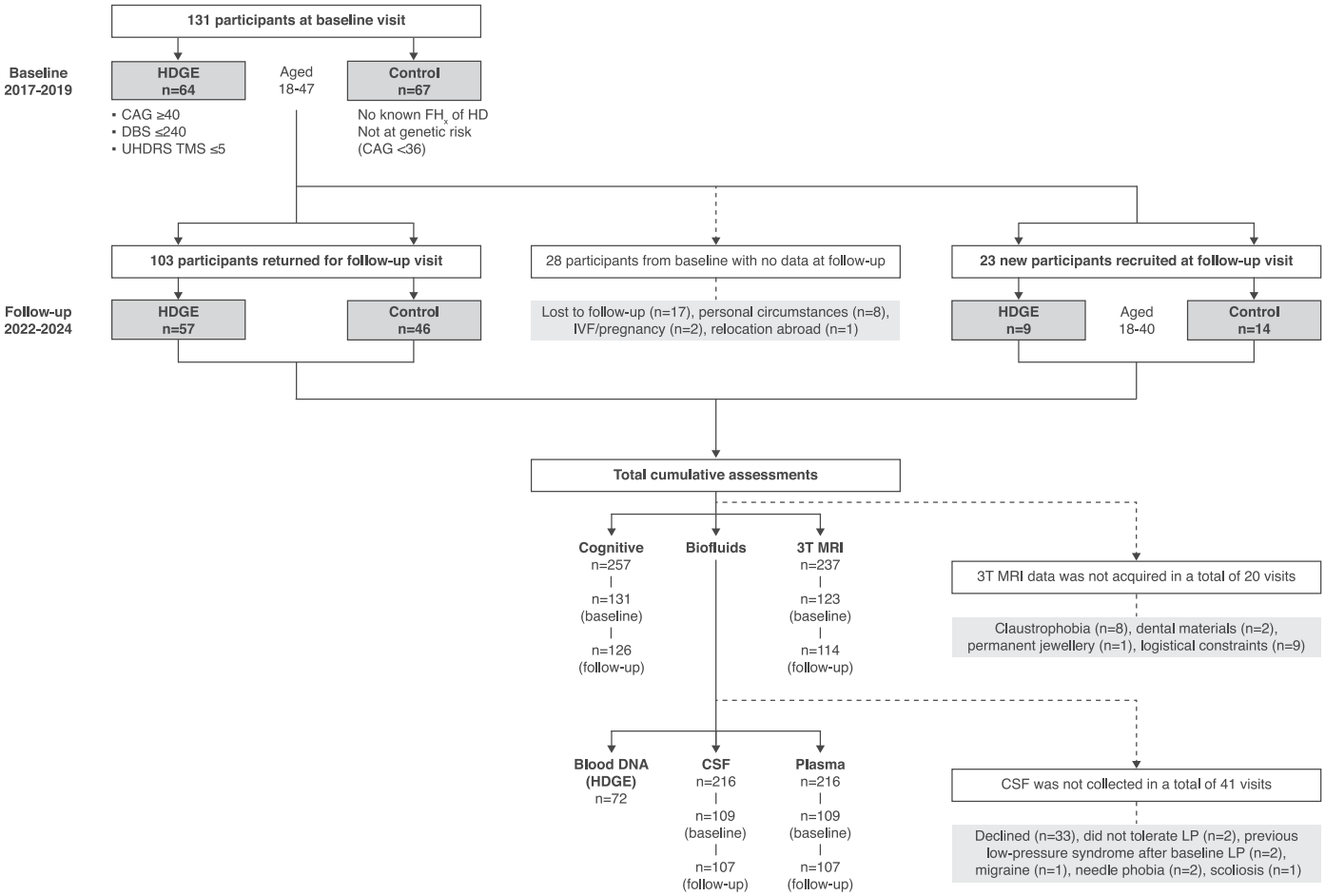
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03424-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03424-6.

Correspondence and requests for materials should be addressed to Sarah J. Tabrizi.

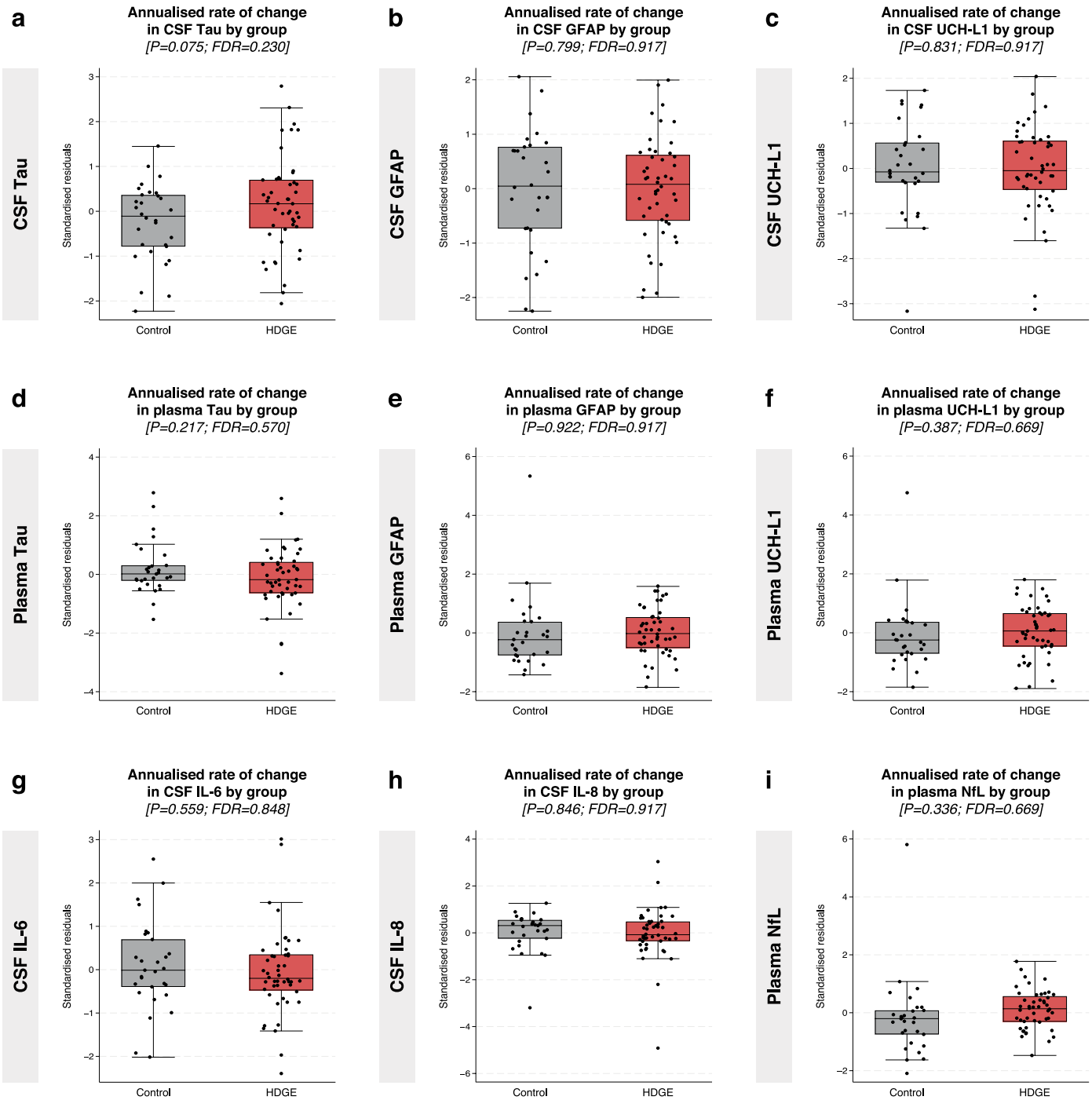
Peer review information *Nature Medicine* thanks Harry Orr and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jerome Staal, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-03424-6



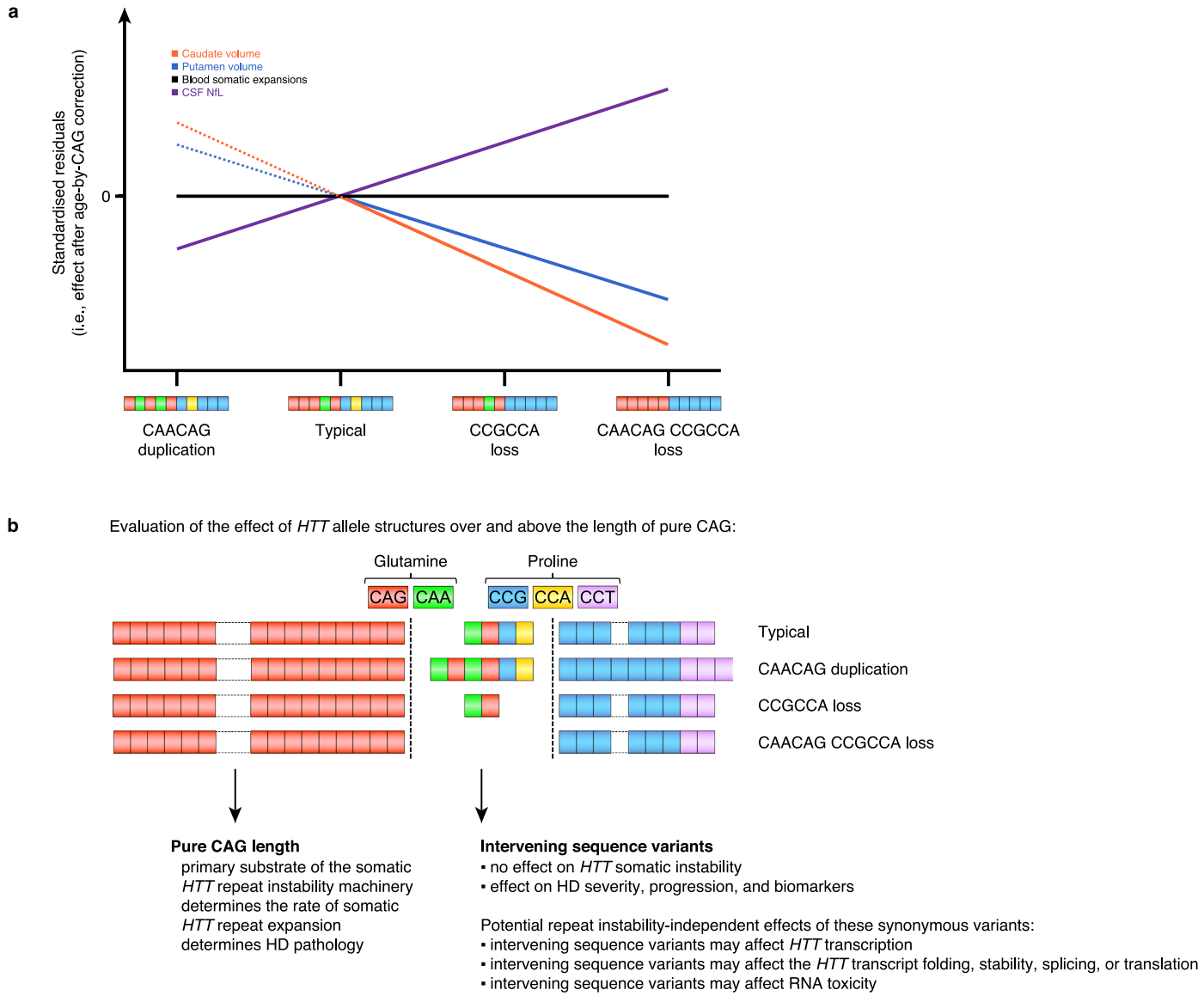
Extended Data Fig. 1 | Flowchart of recruitment, follow-up and total cumulative assessments. Flowchart detailing participant enrollment at baseline (2017–2019) and the retention and recruitment of new participants at follow-up (2022–2024) in the Huntington's Disease Young Adult Study (HD-YAS). CSF, cerebrospinal fluid; DBS, disease burden score; FH_x, family history; HDGE, HD gene expanded; IVF, in vitro fertilization; LP, lumbar puncture; TMS, Total Motor Score; UHDRS, Unified Huntington's Disease Rating Scale.

https://doi.org/10.1038/s41591-024-03424-6



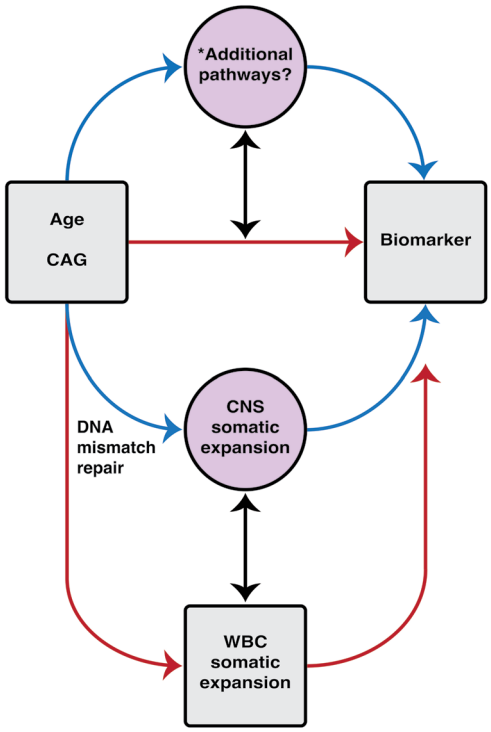
Extended Data Fig. 2 | Annualized rate of change in non-significant biofluid markers. Panel shows comparison of standardized residuals for annualized rate of change in the HDGE vs. control groups for: (**a**) CSF Tau, (**b**) CSF GFAP, (**c**) CSF UCH-L1, (**d**) plasma Tau, (**e**) plasma GFAP, (**f**) plasma UCH-L1, (**g**) CSF IL-6, (**h**) CSF IL-8 and (**i**) plasma NfL. Negative standardized residuals indicate that the rate of change was less than the adjusted mean rate of change across both groups. The horizontal lines represent medians, boxes indicate the upper and lower quartiles, and whiskers are 1.5× IQR. All statistical analyses were conducted using mixed-effect linear models with a participant-specific random effect, controlling for age, sex and their interaction. Natural log-transformed concentrations served as the outcomes in these models. Statistical two-sided group comparisons and correlations controlled for the effects of age and sex and were corrected for multiple comparisons using the FDR. CSF, cerebrospinal fluid; FDR, false discovery rate; GFAP, glial fibrillary acidic protein; HDGE, HD gene expanded; IL, interleukin; IQR, interquartile range; NfL, neurofilament light; UCH-L1, ubiquitin carboxyl-terminal hydrolase L1; YKL-40, also known as chitinase-3 like-protein-1 (CHI3L1).

https://doi.org/10.1038/s41591-024-03424-6



Extended Data Fig. 3 | The effect of atypical *HTT* allele structure on blood somatic expansions does not explain their effect on HD phenotypes and biomarkers. (**a**) Schematic representation of the absence of effect of the atypical *HTT* allele structure on blood somatic expansion in the HD-YAS cohort and of their relative effect on caudate and putamen volumes and on CSF NfL levels after age-by-CAG correction. (**b**) Graphical representation of the *HTT* allele structures observed in the HD-YAS cohort. On the left-hand side is the pure CAG tract, which is likely the primary substrate of the somatic *HTT* repeat instability machinery and is the primary determinant of the rate of somatic HTT repeat expansion and HD pathology. Center, the sequence variants intervening between the CAG and CCG tracts, which define the atypical *HTT* allele structures observed (that is, CAACAG duplication, CCGCCA loss and CAACAG CCGCCA loss).

https://doi.org/10.1038/s41591-024-03424-6



Extended Data Fig. 4 | Conceptual causal model. Conceptual model illustrating the causal relationships among age-related CAG repeat length (age, CAG), CNS somatic expansion, WBC somatic expansion and biomarkers. The squares in gray depict measured data, while the circles in purple represent unobservable quantities. The black, bidirectional arrow signifies the assumption that somatic expansion in WBCs serves as a proxy for somatic expansion in the CNS. Red arrows indicate correlations between observable data and the underlying biological processes of interest. Blue arrows are presumed to reflect true causal pathways indirectly. We note that DNA mismatch repair is a key mechanism linking inherited CAG repeat length to somatic expansion, with mismatch repair activity increasing with longer CAG repeat lengths. *Note additional pathways could be involved due to the imprecise and indirect measurement of actual CNS expansion. Abbreviations: CNS=Central Nervous System. WBC=White Blood Cell.

https://doi.org/10.1038/s41591-024-03424-6

Extended Data Table 1 | Participant demographic characteristics

Characteristic	HDGE		Controls		HDGE vs Controls Baseline, Follow-up
	Longitudinal (n=57)		Longitudinal (n=46)		
Interval (years)	4.7 (0.6)		4.7 (0.6)		0.5
	Baseline (n=64)	Follow-up (n=66)	Baseline (n=67)	Follow-up (n=60)	
Age (years)	29.0 (5.6)	33.8 (5.7)	29.1 (5.7)	33.6 (6.2)	0.9, 0.9
Male:Female (% male)	30:34 (47%)	32:34 (48%)	28:39 (42%)	28:32 (47%)	0.6, 0.8
Education (years)	16.2 (2.1)	17.0 (2.6)	16.3 (2.2)	17.2 (2.7)	0.9, 0.7
NART	102.4 (7.5)	103.7 (7.7)	103.5 (8.3)	106.5(8.7)	0.4, 0.1
CAG repeat length	42.3 (1.6)	42.2 (1.5)	N/A	N/A	N/A
CAP100 score	55.4 (8.2)	62.8 (9.0)	N/A	N/A	N/A
Estimated years to clinical motor diagnosis	23.2 (5.3)	19.2 (5.2)	N/A	N/A	N/A

Table presenting longitudinal participant demographics in HD-YAS. Values are means (SD), n (%) or median (IQR). Two-tailed group comparisons were made using *t* tests (interval, age, education, NART) and χ^2 tests (sex). CAP100, CAG–age product, whereby CAP100 is the expected age of diagnosis. HDGE, HD gene expanded; IQR, interquartile range; NA, not applicable; NART, National Adult Reading Test, an estimate for IQ; SD, standard deviation; IQ, intelligence quotient.

https://doi.org/10.1038/s41591-024-03424-6

Extended Data Table 2 | Longitudinal volumetric data

Outcomes (to ICV)	Control Mean	HDGE Mean	Estimate of Mean Difference	Df	Lower CL	Upper CL	*P* value	FDR
Caudate Change	8.00×10^{-6}	6.07×10^{-5}	5.27×10^{-5}	83	3.85×10^{-5}	6.69×10^{-5}	**1.10×10^{-10}**	**5.51×10^{-10}**
Putamen Change	-3.09×10^{-5}	-1.22×10^{-4}	-5.90×10^{-5}	98	-7.60×10^{-5}	-4.30×10^{-5}	**3.96×10^{-10}**	**1.19×10^{-9}**
Ventricular Change	3.77×10^{-5}	2.61×10^{-4}	2.24×10^{-4}	83	1.21×10^{-4}	3.26×10^{-4}	**3.91×10^{-5}**	**9.77×10^{-5}**
Whole Brain Change	3.12×10^{-4}	1.51×10^{-3}	1.19×10^{-3}	83	5.18×10^{-4}	1.87×10^{-3}	**7.13×10^{-4}**	**1.19×10^{-3}**
Grey Matter Change	5.42×10^{-4}	7.55×10^{-4}	2.13×10^{-4}	83	5.85×10^{-5}	3.68×10^{-4}	**7.53×10^{-3}**	**9.41×10^{-3}**
White Matter Change	3.21×10^{-4}	5.75×10^{-4}	2.54×10^{-4}	83	5.27×10^{-5}	4.56×10^{-4}	**1.41×10^{-2}**	**1.41×10^{-2}**

Table showing longitudinal volumetric outcomes (normalized to ICV) for HDGE participants compared to controls. Volumetric analyses for brain structure longitudinal changes, excluding the putamen, modeled a single change measure per paired participant scans (boundary-shift integral or voxel-based morphometry convolution) after conversion to annual change rates. These changes were modeled by ordinary least squares regression. Putamen changes were derived from subtraction of baseline and follow-up MALP-EM segmentations divided by follow-up length. Analysis results and residual adjustments reflect control for baseline age, sex, and their interaction. Statistical two-sided group comparisons were adjusted for multiple comparisons using the FDR, with P values, degrees of freedom, and confidence limits provided in the table. Significant values at FDR < 0.15 are highlighted in bold. CL, confidence limit; Df, degrees of freedom; FDR, false discovery rate; HDGE, HD gene expanded; ICV, intracranial volume.

https://doi.org/10.1038/s41591-024-03424-6

Extended Data Table 3 | Longitudinal diffusion data

Outcomes	Control Mean	HDGE Mean	Estimate of Mean Difference	Df	Lower CL	Upper CL	*P* value	FDR
Splenium Corpus Callosum MD	-9.87×10^{-4}	1.30×10^{-3}	2.29×10^{-3}	89	1.30×10^{-3}	3.28×10^{-3}	**1.32×10^{-5}**	**4.64×10^{-4}**
Splenium Corpus Callosum RD	-5.01×10^{-4}	1.80×10^{-3}	2.30×10^{-3}	90	1.28×10^{-3}	3.32×10^{-3}	**2.21×10^{-5}**	**4.64×10^{-4}**
Anterior Internal Capsule AD	-1.84×10^{-3}	-2.45×10^{-4}	1.60×10^{-3}	90	8.30×10^{-4}	2.37×10^{-3}	**8.21×10^{-5}**	**1.14×10^{-3}**
Splenium Corpus Callosum FA	2.77×10^{-5}	-1.30×10^{-3}	-1.33×10^{-3}	91	-1.99×10^{-3}	-6.80×10^{-4}	**1.08×10^{-4}**	**1.14×10^{-3}**
Splenium Corpus Callosum AD	-1.95×10^{-3}	3.03×10^{-4}	2.26×10^{-3}	89	1.05×10^{-3}	3.46×10^{-3}	**3.33×10^{-4}**	**2.80×10^{-3}**
Anterior Internal Capsule ODI	2.60×10^{-4}	-3.93×10^{-4}	-6.50×10^{-4}	90	-1.01×10^{-3}	-2.90×10^{-4}	**5.14×10^{-4}**	**3.60×10^{-3}**
Mid Corpus Callosum RD	-3.49×10^{-5}	2.27×10^{-3}	2.31×10^{-3}	91	9.00×10^{-4}	3.72×10^{-3}	**1.58×10^{-3}**	**8.40×10^{-3}**
Anterior Internal Capsule FA	-1.56×10^{-4}	6.71×10^{-4}	8.30×10^{-4}	88	3.20×10^{-4}	1.33×10^{-3}	**1.6×10^{-3}**	**8.40×10^{-3}**
Splenium Corpus Callosum FWF	-2.88×10^{-3}	-2.40×10^{-4}	2.64×10^{-3}	93	1.00×10^{-3}	4.29×10^{-3}	**1.97×10^{-3}**	**9.18×10^{-3}**
Mid Corpus Callosum MD	-3.93×10^{-3}	1.43×10^{-3}	1.82×10^{-3}	90	6.70×10^{-4}	2.97×10^{-3}	**2.22×10^{-3}**	**9.34×10^{-3}**
External Capsule AD	-1.85×10^{-3}	-9.37×10^{-4}	9.10×10^{-4}	90	3.20×10^{-4}	1.50×10^{-3}	**2.86×10^{-3}**	**0.011**
Splenium Corpus Callosum NDI	7.33×10^{-4}	-1.28×10^{-3}	-2.01×10^{-3}	92	-3.34×10^{-3}	-6.80×10^{-4}	**3.54×10^{-3}**	**0.012**
Mid Corpus Callosum FA	-2.45×10^{-4}	-1.91×10^{-3}	-1.67×10^{-3}	92	-2.79×10^{-3}	-5.40×10^{-4}	**4.08×10^{-3}**	**0.013**
External Capsule MD	-1.08×10^{-3}	-4.40×10^{-4}	6.40×10^{-4}	90	1.70×10^{-4}	1.11×10^{-3}	**8.48×10^{-3}**	**0.025**
Mid Corpus Callosum FWF	-1.20×10^{-3}	2.38×10^{-4}	1.44×10^{-3}	90	2.60×10^{-4}	2.61×10^{-3}	**0.017**	**0.048**
Anterior Internal Capsule MD	-9.05×10^{-4}	-3.80×10^{-4}	5.30×10^{-4}	90	7.00×10^{-5}	9.80×10^{-4}	**0.024**	**0.063**
Mid Corpus Callosum ODI	-9.74×10^{-5}	2.02×10^{-4}	3.00×10^{-4}	89	3.00×10^{-5}	5.70×10^{-4}	**0.027**	**0.067**
External Capsule RD	-6.94×10^{-4}	-1.91×10^{-4}	5.00×10^{-4}	89	3.00×10^{-5}	9.70×10^{-4}	**0.037**	**0.087**
External Capsule NDI	1.05×10^{-3}	4.33×10^{-4}	-6.20×10^{-4}	90	-1.23×10^{-3}	-1.00×10^{-5}	**0.046**	**0.103**
Genu Corpus Callosum NDI	4.99×10^{-4}	-9.20×10^{-4}	-1.42×10^{-3}	92	-2.84×10^{-3}	2.17×10^{-6}	**0.050**	**0.106**
Posterior Internal Capsule MD	-9.30×10^{-4}	-3.63×10^{-4}	5.70×10^{-4}	93	-4.00×10^{-5}	1.18×10^{-3}	**0.068**	**0.137**
Genu Corpus Callosum AD	-2.14×10^{-3}	-1.23×10^{-3}	9.10×10^{-4}	90	-9.00×10^{-5}	1.91×10^{-3}	**0.074**	**0.142**
Genu Corpus Callosum MD	-8.24×10^{-4}	-6.57×10^{-5}	7.60×10^{-4}	91	-1.00×10^{-4}	1.62×10^{-3}	**0.083**	**0.148**
Mid Corpus Callosum NDI	-6.67×10^{-4}	-2.03×10^{-3}	-1.36×10^{-3}	93	-2.91×10^{-3}	1.90×10^{-4}	**0.085**	**0.148**
Posterior Internal Capsule AD	-2.60×10^{-3}	-1.83×10^{-3}	7.70×10^{-4}	91	-1.70×10^{-4}	1.71×10^{-3}	0.105	0.171
Posterior Internal Capsule NDI	1.47×10^{-3}	6.31×10^{-4}	-8.40×10^{-4}	91	-1.86×10^{-3}	1.80×10^{-4}	0.106	0.171
Genu Corpus Callosum RD	-1.69×10^{-4}	5.30×10^{-4}	7.00×10^{-4}	91	-2.10×10^{-4}	1.61×10^{-3}	0.132	0.199
Mid Corpus Callosum AD	-1.10×10^{-3}	-2.84×10^{-4}	8.20×10^{-4}	89	-2.50×10^{-4}	1.89×10^{-3}	0.133	0.199
Posterior Internal Capsule RD	-5.58×10^{-5}	3.79×10^{-4}	4.40×10^{-4}	92	-1.90×10^{-4}	1.06×10^{-3}	0.173	0.250
Genu Corpus Callosum FA	-4.25×10^{-4}	-8.93×10^{-4}	-4.70×10^{-4}	90	-1.20×10^{-3}	2.70×10^{-4}	0.208	0.291
Genu Corpus Callosum ODI	-8.87×10^{-4}	-4.50×10^{-4}	4.40×10^{-4}	107	-3.00×10^{-4}	1.18×10^{-3}	0.245	0.332
Anterior Internal Capsule FWF	-1.08×10^{-4}	2.09×10^{-4}	3.20×10^{-4}	103	-2.90×10^{-4}	9.20×10^{-4}	0.299	0.389
Splenium Corpus Callosum ODI	-3.49×10^{-4}	-2.23×10^{-4}	1.30×10^{-4}	95	-1.20×10^{-4}	3.70×10^{-4}	0.306	0.389
Anterior Internal Capsule NDI	1.12×10^{-3}	6.82×10^{-4}	-4.30×10^{-4}	90	-1.29×10^{-3}	4.20×10^{-4}	0.317	0.391
Posterior Internal Capsule FWF	-1.15×10^{-3}	-5.36×10^{-4}	6.20×10^{-4}	118	-7.20×10^{-4}	1.96×10^{-3}	0.363	0.436
Genu Corpus Callosum FWF	2.59×10^{-5}	-3.77×10^{-4}	-4.00×10^{-4}	108	-1.59×10^{-3}	7.90×10^{-4}	0.504	0.588
External Capsule ODI	-1.50×10^{-4}	-3.39×10^{-5}	1.20×10^{-4}	95	-4.80×10^{-4}	7.10×10^{-4}	0.699	0.793
Posterior Internal Capsule ODI	-3.37×10^{-4}	-2.71×10^{-4}	7.00×10^{-5}	99	-7.20×10^{-4}	8.50×10^{-4}	0.867	0.956
Posterior Internal Capsule FA	-7.43×10^{-4}	-7.87×10^{-4}	-4.00×10^{-5}	91	-6.90×10^{-4}	6.00×10^{-4}	0.891	0.956
Anterior Internal Capsule RD	-4.15×10^{-4}	-4.41×10^{-4}	-3.00×10^{-5}	89	-4.80×10^{-4}	4.30×10^{-4}	0.911	0.956
External Capsule FA	-2.31×10^{-4}	-2.44×10^{-4}	-1.00×10^{-5}	89	-3.90×10^{-4}	3.60×10^{-4}	0.948	0.971
External Capsule FWF	2.11×10^{-4}	2.16×10^{-4}	1.00×10^{-5}	102	-3.90×10^{-4}	4.00×10^{-4}	0.979	0.979

Diffusion (AD, FA, MD, and RD) and NODDI (FWF, ODI, and NDI) metrics were derived for the following regions of interest: corpus callosum (genu, mid and splenium), internal capsule (anterior and posterior) and external capsule. Longitudinal change was derived by subtraction of baseline from follow-up value and change was annualized. Analysis results and residual adjustments reflect control for baseline age, sex and their interaction. Statistical two-sided group comparisons were adjusted for multiple comparisons using the FDR, with P values, degrees of freedom, and confidence limits provided in the table. Significant values at FDR < 0.15 are highlighted in bold. AD, axial diffusivity; CL, confidence limit; Df, degrees of freedom; FA, fractional anisotropy; FWF, free water fraction; HDGE, HD gene expanded; MD, mean diffusivity; NDI, neurite density index; ODI, orientation dispersion index; RD, radial diffusivity.

https://doi.org/10.1038/s41591-024-03424-6

Extended Data Table 4 | Predictors of brain atrophy measures

Correction	Outcomes	Caudate (*P* value, FDR)	Putamen (*P* value, FDR)	Grey matter (*P* value, FDR)	White matter (*P* value, FDR)	Whole brain (*P* value, FDR)	Ventricles (*P* value, FDR)
No Age and CAG	Baseline CSF NfL	**2.00×10^{-6}, 1.70×10^{-5}**	**1.64×10^{-5}, 4.91×10^{-5}**	**0.020, 0.023**	**9.50×10^{-5}, 4.73×10^{-4}**	**2.54×10^{-4}, 8.48×10^{-4}**	**1.55×10^{-3}, 2.73×10^{-3}**
	Baseline plasma NfL	**2.00×10^{-3}, 2.85×10^{-3}**	**0.014, 0.015**	**0.037, 0.037**	**1.53×10^{-3}, 2.73×10^{-3}**	**5.21×10^{-3}, 6.51×10^{-3}**	**1.64×10^{-3}, 2.73×10^{-3}**
	Baseline CSF PENK	**1.26×10^{-4}, 1.26×10^{-3}**	**2.18×10^{-4}, 0.001**	**0.078, 0.128**	**0.022, 0.056**	**0.003, 0.013**	**0.019, 0.056**
With Age and CAG	Baseline CSF NfL	**8.59×10^{-4}, 4.29×10^{-3}**	**0.003, 0.007**	**0.124, 0.124**	**3.41×10^{-4}, 3.41×10^{-3}**	**0.002, 0.008**	**0.012, 0.019**
	Baseline plasma NfL	**0.025, 0.032**	0.160, 0.160	**0.100, 0.111**	**0.008, 0.018**	**0.026, 0.032**	**0.009, 0.018**
	Baseline CSF PENK	**2.03×10^{-3}, 2.03×10^{-2}**	**1.13×10^{-3}, 0.0045**	0.250, 0.348	**0.034, 0.112**	**0.010, 0.052**	**0.050, 0.125**
No Age and CAG	Change in CSF NfL	**2.97×10^{-4}, 2.97×10^{-3}**	**2.22×10^{-4}, 2.67×10^{-3}**	**0.041, 0.103**	0.493, 0.548	**0.074, 0.149**	**0.017, 0.057**
	Change in plasma NfL	**2.55×10^{-3}, 0.013**	**0.030, 0.062**	0.443, 0.548	0.626, 0.626	0.179, 0.298	0.257, 0.367
	Change in CSF PENK	**2.49×10^{-4}, 2.49×10^{-3}**	**1.08×10^{-4}, 0.001**	**0.027, 0.137**	0.665, 0.831	0.295, 0.421	0.275, 0.421
With Age and CAG	Change in CSF NfL	**0.002, 0.020**	**0.002, 0.027**	0.084, 0.169	0.210, 0.349	**0.0599, 0.1497**	**0.010, 0.040**
	Change in plasma NfL	**0.012, 0.040**	0.117, 0.205	0.741, 0.741	0.548, 0.609	0.264, 0.369	0.295, 0.369
	Change in CSF PENK	**0.002, 0.021**	**9.42×10^{-4}, 0.011**	0.075, 0.224	0.579, 0.723	0.457, 0.653	0.361, 0.602

Table summarizing the association between baseline and change in key fluid biomarkers with brain atrophy measures across different brain regions, both with and without adjustment for age and CAG repeat length. The primary regression predictors of interest were group status (control vs. HDGE) and in models testing age and CAG effects, terms for age, MiSeq-derived CAG length, and age-by-CAG interactions nested within the HDGE. The joint significance of age and CAG effects were estimated by the maximum likelihood test for differences in model fit. The models also contained covariates for sex, age and sex-by-age interaction. Longitudinal volumetric analyses of brain structures other than putamen relied on a single measure of volume change (boundary-shift integral or voxel-based morphometry convolution) per participant derived from pairs of baseline and follow-up scans. These changes were converted to annual rates and modeled by ordinary least squares regression with predictor variables analogous to those listed above. All other models of longitudinal change used the difference between the two visits as outcomes and longitudinal effects of primary predictors and covariates were estimated by interactions with follow-up time between the visits. Those models also retained all baseline effects for predictors and covariates. To preserve the unbiased estimation of model parameters when data is missing at random, baseline data from participants with no follow-up were included in the model. Significant values at FDR < 0.15 are highlighted in bold. CSF, cerebrospinal fluid; FDR, false discovery rate; NfL, neurofilament light, PENK, proenkephalin.

https://doi.org/10.1038/s41591-024-03424-6

Extended Data Table 5 | Sample size calculations

Outcome:	50% Treatment Effect		70% Treatment Effect	
Change in	2 years	3 years	2 years	3 years
CSF NfL	232	104	120	54
Caudate	282	126	144	64
Putamen	326	146	184	104
Somatic Expansion Ratio	614	274	314	140
Ventricular	776	346	396	176
Whole Brain	1208	538	452	202
Grey Matter	2164	962	1104	492
White Matter	3612	1606	1844	820

Table showing approximate sample size estimates for clinical trials involving equal allocation to one treatment group and a placebo group, which were calculated based on observed HDGE group versus control longitudinal models. The group differences in longitudinal rates were converted to Cohen's *d* effect sizes using within-group longitudinal standard deviations derived from estimated random effect variances. Assumed treatment effects were then defined as the percent slowing of the group difference in longitudinal rates multiplied by the assumed trial length. Treatment effect assumes a reduction in the difference in rate between the HDGE and control groups. Boundary-shift integral is used to measure the change in the caudate, brain and ventricles. Gray matter and white matter changes are generated by convolving baseline voxel-based morphometry-derived gray or white tissue maps with voxel-compression maps of within-participant change. Putamen change is measured by subtraction of baseline and follow-up MALP-EM segmentations. See Supplementary Methods for further details. CSF, cerebrospinal fluid; HDGE, HD gene expanded; NfL, neurofilament light.

https://doi.org/10.1038/s41591-024-03424-6

Extended Data Table 6 | Panel of assessments

Modality	Assessments
Cognition	**Core** Stroop Colour Naming Stroop Word Reading Stroop Interference* Symbol Digit Modalities Test (SDMT) Verbal Fluency (Category) **CANTAB** Intra-Extra Dimensional Shift (IED) One Touch Stockings of Cambridge (OTS) Spatial Working Memory (SWM) Paired Associate Learning (PAL) Rapid Visual Information Processing (RVP) Stop Signal (SST) **EMOTICOM** Emotion Intensity Face Morphing Moral Judgement Goals Prior Assay*
Neuropsychiatric	**Self-Report Questionnaires** Pittsburgh Sleep Quality Index (PSQI) Barratt Impulsiveness Scale (BIS-11) Toronto Empathy Questionnaire (TEQ) State-Trait Anxiety Inventory - Trait (STAI-T) Obsessive-Compulsive Inventory-Revised (OCI-R) Zung Self-Rating Depression Scale (SDS) Apathy Motivation Index (AMI) Toronto Alexithymia Scale (TAS-20) Frontal Systems Behaviour Scale (FrsBe) Short Form Health Survey (SF-36)
Neuroimaging	**Volumetric** Caudate Putamen Grey Matter White Matter Ventricles Whole Brain **DWI** Axial Diffusivity (AD) Fractional Anisotropy (FA) Mean Diffusivity (MD) Radial Diffusivity (RD) **NODDI** Free Water Fraction (FWI Neurite Density Index (NDI) Orientation Dispersion Index (ODI) **MPM** Magnetisation Transfer (MT) R1: Longitudinal Relaxation Rate R2*: Effective Transverse Relaxation Rate Proton Density (PD) **Structural Connectivity** 6 cortico-striatal connectome 14 cortico-cortical connectome
Biofluids	**CSF** Mutant huntingtin NfL Tau GFAP UCH-L1 YKL-40 IL-6 IL-8 PENK **Plasma** NfL Tau GFAP UCH-L1 **Whole Blood – DNA** Somatic Expansion Ratio (SER)

Table providing a detailed overview of assessments across the following four modalities: cognition, neuropsychiatric, neuroimaging and biofluids. *New tasks at follow-up. CANTAB, Cambridge Neuropsychological Test Automated Battery; CSF, cerebrospinal fluid; DWI, diffusion-weighted imaging; EMOTICOM, emotion, motivation, impulsivity and social cognition; GFAP, glial fibrillary acidic protein; IL, interleukin; MPM, multiparametric mapping; NfL, neurofilament light; NODDI, neurite orientation dispersion and density imaging; PENK, proenkephalin; UCH-L1, ubiquitin carboxyl-terminal hydrolase L1; YKL-40, also known as chitinase-3 like-protein-1 (CHI3L1).

nature portfolio

Corresponding author(s): Professor Sarah J Tabrizi

Last updated by author(s): Nov 12, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected between April 6, 2022, and March 21, 2024. All derived variables were generated as per descriptions in Supplementary Methods and analysed according to the pre-specified Statistical Analysis Plan (SAP) available on ClinicalTrials.gov (NCT06391619). An additional exploratory fluid biomarker, proenkephalin (PENK) was introduced due to the availability of a novel assay.
Data analysis	Statistical analysis was performed using R version 4.3.3. Imaging was analysed using freely available software as described in the Supplementary Methods. All software is freely available with the exception of the in-house MIDAS software used to generate the boundary shift integral for caudate, whole brain, and ventricles. This can be requested from Professor Nick Fox at the Dementia Research Centre, UCL.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We are committed to data sharing whilst maintaining confidentiality due to the sensitive and potentially identifiable nature of this data. Biofluid samples will not be shared due to the limited amount of material available. Remaining samples will be required for replication for the next HD-YAS visit. Upon reasonable request, data will be made available 24 months after the end of data collection, through application via UCL to the PI, Professor Sarah Tabrizi (s.tabrizi@ucl.ac.uk). Researchers will be required to submit a proposal meeting the research criteria and must demonstrate full GDPR compliance. A data access agreement with UCL will be required.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex assigned at birth reported.
Reporting on race, ethnicity, or other socially relevant groupings	Reporting on race, ethnicity, or other socially relevant groupings not relevant to study and not reported.
Population characteristics	These are provided in Online Methods, Supplementary Methods, Extended Data Table 1, and Extended Data Figure 1.
Recruitment	Participants were recruited across the UK and enrolled at one study site (UCL). The inclusion criteria have been described previously (Scahill et al., 2020; doi: 10.1016/S1474-4422(20)30143-5) and are detailed in the Supplementary Methods. Participants (131 in total, 64 HD gene expanded (HDGE); 67 controls) attended at baseline and 103 (57 HDGE; 46 controls) returned for follow-up approximately 4.5 years later.
Ethics oversight	The study received approval by the London-Bloomsbury Research Ethics Committee (22/LO/0058). All study procedures adhered to principles outlined in the Declaration of Helsinki, and prior to enrolment, written consent was obtained.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size calculations were provided by our statistician, Professor Douglas Langbehn of the University of Iowa. These were based on the assumptions that with a type 1 error rate of 5%, a sample of 60 participants/group will provide 80% power to detect a mean difference versus controls of 0.53 adjusted within group standard deviations (effect size), allowing for 5 covariates. Similarly, after allowing for 5 covariates, the sample of 60 CAG-expanded participants allows the same statistical power for detecting a partial Pearson correlation of 0.36 among outcome measures and between these measures and the CAP score or other potential predictors of Huntington's disease risk.
Data exclusions	These are provided in Supplementary Table 5 with further details of reasons for exclusions provided in Supplementary Methods.
Replication	All human plasma and cerebrospinal fluid (CSF) analytes were successfully measured. The Neurology 4-Plex A (GFAP, NfL, Tau, UCH-L1) was measured in singlicate, yielding the following inter-plate coefficients of variation (%CV): CSF GFAP (3.7%), CSF NfL (8.2%), CSF Tau (11.1%), CSF UCH-L1 (62.8%), plasma GFAP (5.6%), plasma NfL (5.0%), plasma Tau (11.1%), and plasma UCH-L1 (63.8%). The %CVs were calculated from duplicate measurements of internal plate controls made of pooled human plasma. CSF IL-6, IL-8, and YKL-40 were measured in duplicate, with the following %CVs: CSF YKL-40 (10.4%), CSF IL-6 (29.2%), and CSF IL-8 (12.2%). CSF mHTT was measured in triplicate and the relative abundance of CSF PENK was quantified in single measurements.
Randomization	This is a longitudinal observational study, not an interventional study.
Blinding	Study staff providing clinical assessments were not blinded to genetic status due to the need to obtain appropriate consent and discuss any potential clinical symptoms etc. All quality control of data, image, and biofluid analysis was performed blinded to genetic status. Data was exported and analysed by an independent statistician, Professor Douglas Langbehn of the University of Iowa.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Antibodies

Antibodies used	Biofluid assay details (for human plasma and cerebrospinal fluid analyses) available in Supplementary Table 19.
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NCT06391619
Study protocol	The full study protocol is available on ClinicalTrials.gov NCT06391619.
Data collection	Data was collected between April 6, 2022, and March 21, 2024. All derived variables were generated as per descriptions in Supplementary Methods and analysed according to the pre-specified SAP available on ClinicalTrials.gov (NCT06391619). An additional exploratory fluid biomarker, proenkephalin (PENK) was introduced due to the availability of a novel assay.
Outcomes	All outcomes were pre-specified in the SAP available on ClinicalTrials.gov NCT06391619 and are listed in Extended Data Table 6. An additional exploratory outcome was included in our biofluid analysis, CSF PENK, due to the development of a novel assay.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	Observational study of HDGE and healthy control participants.
Design specifications	Participants were recruited on the basis of being able to undergo 3-Tesla (3T) MRI. After recruitment some participants were not able to complete MRI assessments due to unexpected claustrophobia and other contraindications. >80% of the cohort underwent MRI.
Behavioral performance measures	There were no behavioural performance measures included with the MRI.

Acquisition

Imaging type(s)	Volumetric MRI, diffusion imaging, and multiparametric mapping.
Field strength	3-Tesla (3T).
Sequence & imaging parameters	These are provided in the Supplementary Methods.
Area of acquisition	Whole brain.

Diffusion MRI ☒ Used ☐ Not used

Preprocessing

Preprocessing software	FreeSurfer 7.2.0 was used for structural connectivity analyses, and FreeSurfer 6 for HD-ISS staging, as specified in Tabrizi et al., 2022 (doi: 10.1016/S1474-4422(22)00120-X). Additional software included FSL 5.0.11, DTITK 2.3.3, MIDAS 6.7, VBM (SPM12), MALP-EM 1.2, and MRtrix 3.0.4. Further details and information are available in the Supplementary Methods.
Normalization	DTITK 2.3.3, VBM (SPM12), MIDAS 6.7.
Normalization template	Study-specific template in JHU space (Mori et al., 2008; doi: 10.1016/j.neuroimage.2007.12.035).
Noise and artifact removal	Topup, eddy correct and FA ring removal using DTITK 2.3.3, bias correction using N3 (Sled et al., 1998; doi: 10.1109/42.668698) within MIDAS 6.7.
Volume censoring	Topup, eddy correct and FA ring removal using DTITK 2.3.3, bias correction using N3 (Sled et al., 1998; doi:

Volume censoring | 10.1109/42.668698) within MIDAS 6.7.

Statistical modeling & inference

Model type and settings | Linear mixed models.

Effect(s) tested | Group differences, associations with age and CAG, biofluids, and somatic expansion ratio.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Statistic type for inference (See Eklund et al. 2016) | *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction | Multiple comparisons using the False Discovery Rate (FDR).

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity | *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis | *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis | *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*