

# **Clinical and genetic association studies reveal within-individual variability as an integral part of Huntington disease**

N. Ahmad Aziz<sup>1,2</sup>

<sup>1</sup> *Population & Clinical Neuroepidemiology, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*

<sup>2</sup> *Department of Neurology, Faculty of Medicine, University of Bonn, Germany*

## **Corresponding author:**

Dr. N. Ahmad Aziz, MD PhD

Population & Clinical Neuroepidemiology, German Center for Neurodegenerative Diseases (DZNE) & Department of Neurology, Faculty of Medicine, University of Bonn

Venusberg-Campus 1/99, 53127 Bonn, Germany

Phone: +49-228-43302-954

Email: [ahmad.aziz@dzne.de](mailto:ahmad.aziz@dzne.de)

Web: [www.dzne.de/aziz](http://www.dzne.de/aziz)

## **Running Title:**

The vacillating nature of Huntington disease

## **Competing Interests:**

The author reports no competing interests relevant to the manuscript.

## **Key Words:**

Huntington disease, within-individual variability, age-at-onset, disease progression, genetic modifiers, repeat expansion disorders, free energy principle, active inference, Bayesian brain theory

## Abstract

Huntington disease (HD) is one of the most common repeat expansion disorders. Clinically, HD patients exhibit considerable visit-to-visit variability in symptoms and signs independent of measuring method or rater, yet neither the precise nature nor the determinants of this clinical variability are known. Leveraging detailed genetic and longitudinal clinical data from large cohorts of HD patients, this work 1) establishes within-individual variability in disease expression as an integral part of HD semiology, demonstrating that it increases with disease duration, mutation size, younger age-at-onset, and lower body weight, 2) provides a mathematical framework linking higher phenotypic variability to increased predictive entropy, revealing a fundamental relation between within-individual variability and energy expenditure, and 3) identifies novel genetic modifiers of both within-individual variability and age-at-onset in HD. Thus, accounting for within-individual variability in HD could facilitate the discovery of pathogenic mechanisms and outcomes directly relevant to the development of disease modifying therapies.

## Introduction

Huntington disease (HD) is an autosomal dominantly inherited neurodegenerative disorder caused by a cytosine-adenine-guanine (CAG) repeat expansion in the first exon of the huntingtin (*HTT*) gene. Involuntary movements, cognitive decline, psychiatric and behavioural disturbances as well as progressive weight loss characterise the disease.<sup>1,2</sup> We and others have previously demonstrated that the CAG repeat size of the mutant allele is the strongest determinant of both age-at-onset and rate of disease progression in HD.<sup>3</sup> Apart from indicating progressive global multi-domain impairment, clinical rating scores in HD also exhibit substantial visit-to-visit within-individual variability, which is generally assumed to reflect either limitations of the applied rating scales or intra- and interrater variability.<sup>4</sup> However, seeing and treating HD patients over many years, I have the impression that disease expression itself can exhibit large visit-to-visit variability on both subjective and (semi-)quantitative metrics *independent* of the measuring methods employed or the rater, yet neither the precise nature nor the determinants of this clinical variability have been investigated so far in HD.

Emerging evidence indicates that not only the mean trajectories of (clinical) biomarkers, but also their intraindividual variability in time constitute risk factors for a range of different diseases.<sup>5</sup> For example, visit-to-visit variability in blood pressure has been associated with an increased risk of mortality, coronary heart disease, stroke, cognitive decline and dementia,<sup>6-9</sup> while glycemic variability in diabetic patients has been related to an increased risk of cardiovascular complications.<sup>10</sup> Similarly, increased body weight variability has been linked to an increased risk of cardiovascular diseases and cognitive decline.<sup>11,12</sup> In patients with common age-associated neurodegenerative disorders like Alzheimer disease and Parkinson disease, increased visit-to-visit variability in body weight and blood pressure have been associated with a faster rate of disease progression.<sup>13,14</sup> However, whether intraindividual variability in body weight and other clinical metrics is a characteristic of HD and related to mutation size or disease progression is unknown.

Visit-to-visit within-individual variability of disease expression could reflect the breakdown of central homeostatic mechanisms in HD.<sup>3</sup> In this scenario, larger mutation size could not only be expected to result in a younger age-at-onset and a faster rate of disease progression, but also in a

larger degree of visit-to-visit variability in measures of disease expression, reflecting more extensive underlying homeostatic dysregulation. Therefore, leveraging extensive longitudinal data from the largest cohort study of HD patients, as well as recent advances in the statistical modeling of within-individual variability trajectories across time, this work aims to: 1) quantify both the extent and characteristics of within-individual variability in clinical measures of disease expression in HD, 2) assess whether mutation size is associated with longitudinal within-individual clinical variability, 3) provide a Bayesian brain theory-based mathematical framework in which within-individual variability could be understood as a direct consequence of increased predictive entropy due to progressive neuropathology, 4) elucidate the genetic determinants of within-individual variability in HD, and 5) illustrate how treating variability as a core feature of HD could provide insights into the pathogenesis of weight loss, as well as facilitate the discovery of novel genetic modifiers of age-at-onset. The findings indicate that within-individual variability in disease expression is indeed an integral part of HD semiology, with the extent of this variability increasing with both disease duration and mutation size. Moreover, genome-wide association studies (GWAS) identified several genetic modifiers of within-individual variability, indicating that investigating this thus far hidden but integral feature of HD could facilitate the identification of new pathogenic mechanisms with direct relevance for the development of disease modifying therapies for this devastating disorder.

## **Methods**

### *Study Cohort*

Monitored data from the Enroll-HD study were used, which also included longitudinal data on a subset of individuals who had previously participated in the Registry study.<sup>15</sup> Enroll-HD is a global clinical research platform designed to facilitate clinical research in HD. All sites are required to obtain and maintain local ethics committee approvals. Further details are available on the study's website: [enroll-hd.org](http://enroll-hd.org). All data from the Enroll-HD website were retrieved on 24 December 2020, and included all participants with a mutant *HTT* CAG repeat size between 36 and 65 (to exclude allele counts < 10), body mass index (BMI) < 50 kg/m<sup>2</sup>, and at least three study visits, resulting in an analytical sample size of 7583 individuals (54% female).

### *Clinical Measurements*

For assessing disease expression, apart from BMI, the Unified Huntington's Diseases Rating Scale (UHDRS) total motor score, total functional capacity (TFC), scores on the Symbol Digit Modality Test (SDMT) and Stroop Word Reading Test (SWRT), as well as a composite clinical outcome measure derived from these individual UHDRS domain scores as described previously,<sup>4</sup> were used. Except for the UHDRS total motor score, higher scores indicate better clinical performance.

### *Genotype data*

Individual-level imputed genotype data from the largest GWAS of residual age-at-onset in HD were obtained from the Genetic Modifiers of Huntington Disease (GeM-HD) Consortium.<sup>16</sup> This GWAS included data on a total of 9064 HD patients. Residual age-at-onset was defined as the difference in years between observed and expected age of motor onset based on mutant CAG repeat size.<sup>3, 16</sup> Individual-level imputed genotype data were available for a subset of 2230 HD patients from the Enroll-HD study, 1312 of whom were also included in the Gem-HD GWAS. Details of the genotyping and imputation procedures have been reported previously.<sup>16</sup>

### *Genome-wide association study of mean and within-individual variability of clinical measures*

In the subset of individuals from Enroll-HD in whom genotype data were available, two separate GWAS studies were conducted to identify the genetic modifiers of both mean and within-individual variability in the UHDRS total motor score, SDMT score, and BMI. Only participants with at least three study visits and a mutant *HTT* CAG repeat size between 38 and 52 (in order to exclude allele counts < 10) were included. Additionally, extensive per individual and per variant quality control (QC) were performed using plinkQC,<sup>17</sup> applying the following (default) filters for inclusion: genotyping rate > 0.97, heterozygosity rate < 3 standard deviations from the sample mean, proportion of identity-by-descent of < 0.1875 for pairs of individuals, European descent (defined as < 3 times the maximum Euclidean distance from the center of the HapMap European reference samples), variant missing rate < 0.01, Hardy-Weinberg equilibrium ( $p \geq 1E-5$ ), minor allele frequency  $\geq 0.05$  and imputation score  $\geq 0.5$ . After filtering and QC, 1292 individuals and 6,625,212 variants remained for analysis.

For the within-individual variability GWAS analyses two different methods were employed. Initially, TrajGWAS, a modified linear mixed effects model-based approach for assessing genetic effects on both the mean and within-individual variability of continuous outcomes<sup>5</sup> (as implemented in the TrajGWAS.jl Julia package, version 0.4.3, Julia version 1.8.5), was used. A per single-nucleotide polymorphism (SNP) analysis was conducted, using the default parameters and 10 runs with an additive model and a score test with a saddle-point approximation for estimating effect sizes and p-values for each genetic variant. The models included age, mutant CAG repeat size, an interaction term between age and mutant CAG repeat size (to adjust for the effect of CAG repeat size on the rate of disease progression), sex, and the first 10 genetic principle components (to account for population stratification) as fixed effects, a random intercept and slope for age (to account for repeated intraindividual measurements), and an intercept only for estimating the remaining within-subject variance. However, given that the TrajGWAS-based estimates of within-individual variability effects were unstable and showed a high degree of genomic inflation ( $\lambda \geq 1.3$ ), these results are not reported here. Instead, a more conservative method was employed for estimating the genetic determinants of within-individual variability in HD, as follows: 1) using the entire analytical Enroll-HD dataset (N=7583), a random intercept and slope linear mixed effects model was fitted with sex, age, age squared, mutant CAG repeat size, and an interaction term between age and mutant CAG repeat size as fixed effects, and age as a random effect, and the clinical measures as the dependent variables, 2) the remaining within individual variability was extracted as the mean of the sum of the squared errors around the fitted random regression lines for each outcome measure, 3) a logarithmic transformation (given the right-skewed distribution) and scaling to a standard normal distribution was performed, and 4) these estimates were used as the dependent variable in a mixed-effects model GWAS employing Genome-wide Efficient Mixed Model Association (GEMMA, version 0.98.5).<sup>18</sup> Briefly, in GEMMA, within-individual variability estimates were modeled as a function of minor allele dosage, sex, mutant CAG repeat size, and the first four genetic principle components, while also adjusting for remaining population and sample structure in a univariate linear mixed effects model as previously applied by the Gem-HD consortium.<sup>16</sup> A similar approach was used to perform a GWAS on the means of the clinical measures, except now employing a random intercept only linear mixed effects model to estimate the mean levels of

each clinical outcome measure (adjusted for sex, age, age squared, CAG repeat size, and its interaction with age) for each individual during the follow-up period.

### *Genome-wide association study of age of onset*

To account for mutant CAG repeat size-dependent increases in phenotypic variability, a GWAS of *relative* residual age-at-onset (defined as residual age-at-onset divided by the expected age of onset) was performed using an otherwise identical statistical model and the same dataset (N = 9064) as previously reported by the Gem-HD consortium.<sup>16</sup>

### *Genomic risk loci and gene-based analysis*

Functional Mapping and Analysis of GWAS (FUMA) was used to identify genomic risk loci.<sup>19</sup> Genome-wide significant SNPs in relatively high linkage disequilibrium (LD) (i.e.,  $r^2 > 0.6$ ) with nearby SNPs were used to define genomic risk loci. Variant annotation for each locus was based on lead SNPs and candidate SNPs, defined as those SNPs in LD with the lead SNP within a window of 250 kb and nominally significant at  $p < 0.05$ . Functionally annotated SNPs were subsequently mapped to genes based on physical position (positional mapping) and expression quantitative trait loci (eQTL mapping). MAGMA was used to conduct gene and gene-set analysis,<sup>20</sup> summarizing SNP-trait associations at the level of genes and gene sets and mapping these to biological pathways, using a Bonferroni-correction for multiple testing. This was followed by tissue enrichment analysis to investigate tissue specificity using the GTEx database (version 8). For pathway enrichment analysis, mapped genes were further investigated using the GENE2FUNC tool in FUMA,<sup>21</sup> applying the Benjamini-Hochberg false discovery rate (FDR) method for multiple comparisons correction.

### *Statistical Analysis*

To assess the association of disease duration and *HTT* CAG repeat size with visit-to-visit clinical variability a recently developed regression method was utilized, i.e., within-subject variance estimator by robust regression (WiSER) (as implemented in the WiSER.jl Julia package, version 0.2.3), which is specifically designed to assess the effects of both time-dependent and time-independent predictors on longitudinal within-individual variability.<sup>22</sup> WiSER is robust against misspecifications of the distributions of the conditional outcomes and random effects.<sup>22</sup> As

outcomes we used the composite score, as well as its constituents (including the UHDRS total motor score, TFC, and SDMT and SWRT score tests), and BMI. As predictors, age at the time of each visit, a squared term for age to account for non-linear progression, mutant *HTT* CAG repeat size, an interaction term between age and mutant *HTT* CAG repeat size, and sex were included, applying a random intercept and slope model to account for repeated measurements. Age was used as the time indicator to allow for simultaneous adjustment for possible age effects on disease progression, which would not have been possible with inclusion of disease duration due to a high degree of collinearity between age and disease duration.<sup>3</sup> To assess the association between within-individual variability and age-at-onset, residual age-at-onset was added as an independent variable to the previous model. The robustness of the findings was assessed in two additional sensitivity analyses. In the first sensitivity analysis, the same approach as detailed in the previous section (*'Genome-wide association study of mean and within-individual variability of clinical measures'*) was used for calculating the individual level variability as the mean of the sum of the squared errors around the fitted random intercept and slope regression lines for each outcome measure, comparing the values across different mutant CAG repeat size categories using the non-parametric Wilcoxon's rank sum test. In a second sensitivity analysis, the variation independent of mean (VIM) metric was calculated as described previously.<sup>9</sup> For calculating the VIM of composite score, first a constant of 10 was added to offset the composite scores as VIM is undefined over negative values.<sup>9</sup> After regressing out the effects of sex, age at baseline, and domain-specific disease severity at baseline (e.g., for motor score, the motor score at baseline was used as a measure of disease severity) from the VIM scores, the residuals were compared across different categories of mutant CAG repeat size using the non-parametric Wilcoxon's rank sum test. A similar approach was employed to visualize the association between residual age-at-onset and VIM of the different outcomes, where an additional adjustment for mutant CAG repeat size and its interaction with age at baseline were included. Programming was performed in R (base version 4.2.1) and Julia (version 1.8.5). Two-tailed p-values < 0.05 were considered statistically significant.

## Results

### **Part 1: Within individual variability is an integral feature of HD**



The baseline characteristics of the study sample are included in **Supplementary Table 1**. As expected, age and CAG repeat size, as well as their interaction, were highly significantly associated with disease progression across all six outcome measures. On average, the TFC score and BMI were higher, and the SDMT score was lower, in men compared to women (**Table 1**).

*Within-individual variability increases with disease duration and CAG repeat size*

Visit-to-visit within-individual variability in the composite, motor and functional scores and BMI markedly increased with both age and mutant CAG repeat size, whereas cognitive test score variability decreased for SDMT with higher age and CAG repeat size (**Table 1**). Importantly, the effects of age and mutant CAG repeat size on visit-to-visit variability were synergistic across all outcome measures as indicated by highly significant interaction effects with similar directions to the main effects (**Table 1**). Men exhibited lower BMI variability, but otherwise no sex differences in variability of disease expression were observed. Except for cognitive scores, both sensitivity analyses confirmed the findings based on the WiSER method, showing a progressively higher within-individual variability with increasing mutant CAG repeat size for the composite, motor and functional scores, as well as BMI (**Figure 1** and **Supplementary Figure 1**). However, the sensitivity analyses, which employed groupings based on mutant CAG repeat size, indicated that within-individual variability in cognitive scores (SDMT and SWRT scores) also increased with larger mutant CAG repeat size. The difference between the WiSER-based estimates and the mutant CAG repeat size-stratified analyses for cognitive scores could be due to the levelling off of the variability for larger mutation sizes (**Figure 1** and **Supplementary Figure 1**), which may reflect floor/ceiling effects of the cognitive testing batteries in HD.<sup>23</sup>

*Within-individual variability increases with younger age-at-onset*

A lower residual age-at-onset was robustly associated with higher visit-to-visit within-individual variability in all clinical outcome measures, including the UHDRS composite, motor, functional and cognitive scores, as well as BMI, independent of sex, age or mutation size (**Table 2** and **Figure 2**).

**Part 2: Within-individual variability within the Bayesian brain theory framework**

Increased variability in phenotype across time in a given individual may reflect the progressive breakdown of homeostatic mechanisms, and thus increased entropy, due to accruing pathology at molecular, cellular, tissue, organ or system level. To assess the validity of this hypothesis, the ‘free energy principle’-based Bayesian brain model, which is thought to provide a unified theory of brain function,<sup>24</sup> was leveraged. Within this framework, phenotype was defined here as the “action” of the agent, because this is the only perceptible part of the variational free energy minimization process that could be measured by an external observer (e.g., a clinical rater).

#### *Analytical derivation of the relation between entropy and phenotypic variability*

Specifically, it was assessed how action is related to changes in the entropy of the underlying distribution, which for Gaussian distributions is directly related to the variance  $\sigma$  (Note: for consistency, the same terminology and symbols as used by Buckley et al.<sup>25</sup> are employed, and for clarity of presentation, the derivation is restricted to the univariate case). Let  $x$  be a random variable with a Gaussian distribution:

$$x \sim \mathcal{N}(\mu, \sigma) \quad \text{Eq. 1}$$

Then the differential entropy ( $H$ ) of  $x$  is given by<sup>26</sup>:

$$H = \frac{1}{2} \ln 2\pi\sigma + \frac{1}{2} \quad \text{Eq. 2}$$

For a simple agent-based model possessing only a single brain state variable ( $\mu$ ) and a single sensory channel ( $\varphi$ ), the organism believes its sensory input is generated by some generative function  $g$  parametrized by  $\theta$ :

$$\varphi = g(\mu; \theta) + z \quad \text{Eq. 3}$$

With  $z$  denoting random noise with zero mean and variance  $\sigma_z$ , and  $\mu$  denoting the organism’s beliefs about how the environmental states are generated, as follows:

$$\mu = \bar{\mu} + w \quad \text{Eq. 4}$$

Here  $\bar{\mu}$  represents some fixed parameter, while  $w$  denotes random noise with zero mean and variance  $\sigma_w$ . Assuming that with accruing neuropathology the estimates of the organism’s beliefs regarding its environment become increasingly unstable and unpredictable, which can be taken to imply that the variance of the distribution from which  $w$  arises increases, in the following I will thus specifically seek to derive the explicit function describing the relation between action ( $a$ ) and  $\sigma_w$ . For a simple agent-based model,  $a$  is given by<sup>25</sup>:

$$a = -K_a \frac{d\varphi}{da} \frac{dE(\mu, \varphi)}{d\varphi} \quad \text{Eq. 5}$$

From which it follows that:

$$-K_a E(\mu, \varphi) = \int a da = \frac{1}{2} a^2 + c_1 \quad \text{Eq. 6}$$

Here  $K_a$  denotes the learning rate associated with the action (with the negative sign indicating that the direction of action is such to oppose departures from steady-state), and  $E(\mu, \varphi)$  represents the Laplace-encoded energy given by<sup>25</sup>:

$$E(\mu, \varphi) = -\ln p(\varphi|\mu) - \ln p(\mu) = \frac{1}{2\sigma_z} \varepsilon_z^2 + \frac{1}{2\sigma_w} \varepsilon_w^2 + \frac{1}{2} \ln(\sigma_z \sigma_w) \quad \text{Eq. 7}$$

Where:

$$\varepsilon_z \stackrel{\text{def}}{=} \varphi - g(\mu; \theta) = z \quad \text{and} \quad \varepsilon_w \stackrel{\text{def}}{=} \mu - \bar{\mu} = w$$

Therefore, to assess how action changes as a function of entropy, one needs to take the partial derivative of Eq. 7 with respect to  $\sigma_w$ :

$$\frac{dE(\mu, \varphi)}{d\sigma_w} = \frac{d}{2d\sigma_w} \left( \frac{\varepsilon_w^2}{\sigma_w} + \ln \sigma_w \right) \quad \text{Eq. 8}$$

Now, let:

$$Z = \frac{\varepsilon_w}{\sqrt{\sigma_w}} \quad \text{Eq. 9}$$

Substituting Eq. 9 in Eq. 8 yields:

$$\frac{dE(\mu, \varphi)}{d\sigma_w} = \frac{d}{2d\sigma_w} (Z^2 + \ln \sigma_w) \quad \text{Eq. 10}$$

Given that  $Z$ , by definition, is scaled to the standard normal distribution, i.e.,  $Z \sim \mathcal{N}(0, 1)$ , and thus independent of the choice of  $\sigma_w$ , Eq. 10 simplifies to:

$$\frac{dE(\mu, \varphi)}{d\sigma_w} = \frac{1}{2\sigma_w} \quad \text{Eq. 11}$$

From which it follows that:

$$E(\mu, \varphi) = \int \frac{1}{2\sigma_w} d\sigma_w = \frac{1}{2} \ln \sigma_w + c_2 \quad \text{Eq. 12}$$

Combining Eq. 6 & Eq. 12 (and setting the integration constants  $c_1 = c_2 = 0$ ) obtains:

$$a^2 = K_a \ln \sigma_w \Rightarrow a = \pm \sqrt{K_a \ln \sigma_w} \quad \text{Eq. 13}$$

Eq. 13 thus indicates that the *magnitude* of the action towards the steady-state is a monotonously increasing function of  $\sigma_w$  (with its direction opposite to that of the deviation from the steady-state), and by implication (given Eq. 2) also a monotonously increasing function of the entropy of the underlying distribution from which  $w$  (or  $\varepsilon_w$ ) arises. As  $\mathbb{E}(a) = 0$  during steady-state (because, by definition (Eq. 4), during a steady-state any deviation from the set-point ( $\bar{\mu}$ ) will be due to random noise  $w$ , which the action  $a$  tends to oppose), the variance of the action  $a$  can simply be derived as:

$$\text{Var}(a) = \mathbb{E}(a - \mathbb{E}(a))^2 = \mathbb{E}(a)^2 = a^2 \quad \text{Eq. 14}$$

From Eq. 13 it then follows that:

$$\text{Var}(a) = K_a \ln \sigma_w \quad \text{Eq. 15}$$

Therefore, Eq. 15 demonstrates that variability of action, a measure of within-individual variability, is a monotonously increasing function of  $\sigma_w$ , and thus of the entropy of the encoded brain states. This implies that within-individual variability in phenotype could be regarded as an increase in the entropy of the output of the underlying neuronal substrates encoding the organism's beliefs about its environment.

### *Simulations using a Bayesian agent support link between entropy and phenotypic variability*

To illustrate the analytically derived relations above, the previously published code for simulating a simple but complete agent-based model that utilizes the free-energy principle to perform perceptual inference<sup>25</sup> was leveraged (the R code of the implementation is provided in the **Supplementary Materials**). Briefly, this agent's environment consists of a unidimensional line and a single temperature source. The agent's position on this line is represented by the environmental variable  $\vartheta$  and its temperature  $T$  depends on its position as follows:

$$T(\vartheta) = \frac{T_0}{(\vartheta^2 + 1)^2} \quad \text{Eq. 16}$$

This agent is presumed to sit on a frictionless line and be stationary in the absence of action. The agent is allowed to set its own velocity, which is defined as its action  $a$ :

$$a = \frac{d\vartheta}{dt} \quad \text{Eq. 17}$$

Further details are provided in the original publication.<sup>25</sup> Here the behavior of this agent was simulated to assess whether an increase of the entropy of the distribution that underlies the

agent's beliefs about its environment (parametrized through  $\sigma_w$ ), would lead to an increase in the variability of the agent's movements (represented by  $Var(a)$ ), as predicted by *Eq. 15*. To this end, the agent's desired and initial temperature were both set to 20 °C, and its action for a time period of 1000 arbitrary units was simulated for different values of  $\sigma_w$ , ranging from 1 to 3 (the exact initialization parameters/values and further details to reproduce all simulation results can be found in the R script provided in the **Supplementary Materials**). The magnitude of action was scaled to the variance of the sensory input  $\sigma_z$ . **Figure 3** illustrates the agent's action as a function of time for three different values of  $\sigma_w$ , indicating increased variability of action for higher values of  $\sigma_w$ . This latter association was closely in line with the analytical predictions and could be directly confirmed by calculating the non-parametric Spearman's correlation coefficients between  $\sigma_w$  and estimates of the variability of the simulated actions (Spearman's  $\rho = 0.95$ ,  $p < 2.2e-16$ ; **Supplementary Figure 3** and **Figure 4** (top panel), all correlations were calculated for the steady-state defined as time  $\geq 750$ ). Given that, in this particular case, the agent's actions wholly consisted of 'measurable' movements, it was also possible to readily calculate the work performed by the agent, defined as the cumulative sum of the product of acceleration and distance traveled over time. Importantly, this analysis demonstrated that the amount of work required to maintain steady-state is strongly associated with  $\sigma_w$  (Spearman's  $\rho = 0.97$ ,  $p < 2.2e-16$ ; **Figure 4**, bottom panel). Thus, these simulations confirmed the analytically derived relations presented in the previous section, illustrating that increased entropy of brain states not only will lead to higher phenotypic variability, but will also increase the energy costs required to maintain steady-states.

### **Part 3: Leveraging variability to gain insights into Huntington disease pathogenesis**

This section seeks to demonstrate the utility of accounting for phenotypic variability as an integral part of HD. Specifically, it is illustrated how considering variability as a core feature of HD could provide new insights into the pathogenesis of weight loss, as well as facilitate the uncovering of novel genetic modifiers of the disease.

#### *Weight loss*

Weight loss is a hallmark of HD, is seen in both patients and most genetic (rodent) models of the disease, can precede the onset of motor signs and symptoms by many years, and is a robust

predictor of the rate of disease progression.<sup>27, 28</sup> However, its precise underlying causes are poorly understood.<sup>29</sup> As detailed in the previous section, increased within-individual variability in phenotype is inevitably related to higher energy expenditure, because more work will be required to maintain steady-states in the face of higher entropy of the brain's predictive machinery as a result of pathology. Applied to HD, this theoretical framework predicts that increased phenotypic variability would result in increased energy expenditure, and thus a lower body weight and a higher rate of weight loss. To directly probe this predicted association, the within-individual variability in the UHDRS total motor score was assessed in relation to both body weight and the rate of weight loss in the Enroll-HD cohort. Average body weight (estimated through a random intercept linear mixed-effects model with adjustment for age, age<sup>2</sup>, sex, mutant CAG repeat size, as well as its interaction with age, *and total motor score*) decreased with higher within-individual variability of total motor score (**Figure 5**, left panel). Moreover, the rate of body weight loss (represented by the random slope coefficients obtained from a random intercept and slope linear mixed-effects model with age, age<sup>2</sup>, sex, mutant CAG repeat size, as well as its interaction with age, *and total motor score*, as fixed effects, and an intercept and age as random effects) also increased with higher variability of total motor score (**Figure 5**, right panel). These findings thus indicate that higher energy expenditure in HD may not be necessarily due to an inherent defect in systemic energy regulation, but could be an inevitable consequence of increased energy requirements to maintain homeostasis in the face of a progressively inefficient capacity for predictive inference.

#### *Genetic modifiers of within-individual variability*

The GWAS of within-individual variability of BMI identified a genome-wide significant intergenic locus on chromosome 22, tagged by the lead SNP rs62231080 (standardized  $\beta = -0.260 \pm 4.57E-2$ ,  $p = 1.48E-8$ ) (**Table 3**, **Figure 6**, and **Supplementary Figure 3**). Positional and eQTL analyses mapped this locus to the protein coding genes *NUP50*, *FAM18A*, and *RIBC2*, and the lincRNA genes *CTA-217C2.1* and *CTA-268H5.14* (**Figure 6**). The GWAS of mean BMI levels revealed two additionally novel loci on chromosomes 16 and 18, tagged by the lead SNPs rs200717776 and rs11874212, respectively (**Table 3**, **Figure 7**, and **Supplementary Figure 4**). Positional and eQTL analyses mapped the chromosome 16 locus to the protein coding *HEATR3* gene, and the antisense lincRNA *RP11-429P3.5*, and the chromosome 18 locus to the protein

coding *TSHZ1* and *RP11-17M16.1* genes (**Figure 7**). Notably, two independently significant SNPs (rs11150911 and rs12964145) at the chromosome 18 locus have previously also been associated with BMI and body weight in population-based cohorts.<sup>30, 31</sup> Moreover, MAGMA gene-based analysis identified *EIF3F* as a gene associated with mean BMI levels in HD patients (gene-based p-value = 2.03E-6 (< Bonferroni-corrected threshold of 0.05/19208 = 2.60E-6), while another gene on chromosome 14 (*KLHL33*) showed only suggestive statistical significance (p = 3.48E-6) (**Supplementary Figure 5**). Tissue and pathway enrichment analysis did not yield any additional significant associations (data not shown).

The GWAS of within-individual variability of UHDRS total motor score revealed a locus on chromosome 16 (rs772972) at suggestive significance (p = 4.10E-7), which was located in an intronic region of the *GSG1L* gene (**Supplementary Figure 6**). Interestingly, this locus has previously been associated with residual age-at-onset in HD.<sup>16</sup> The GWAS of mean UHDRS total motor score did not identify genome-wide significant loci (**Table 3**). The GWAS of within-individual variability of UHDRS SDMT revealed a locus on chromosome 1 (rs4101233) at suggestive significance (p = 1.26E-7), which was located in an intergenic region (**Supplementary Figure 7**). Notably, this locus could be mapped to the protein coding genes *ZNF644* and *CDC7*, the latter of which encodes a DNA replication enzyme that may be related to *FANI* activity.<sup>32</sup> Only the well-characterized *MSH3* locus on chromosome 5, which has previously been associated with residual age-at-onset in HD,<sup>16</sup> reached genome-wide significance in the GWAS of mean SDMT levels (**Table 3**).

#### *Genetic modifiers of age of onset*

Given that age of onset in HD is usually recorded as a single measurement, it is not possible to directly assess within-individual variability of age of onset. However, in Part I of the results section, it was shown that within-individual variability is strongly associated with mutant CAG repeat size. Given that within-individual variability is also likely to account for at least part of the between-individual variability, here the association between CAG repeat size and between-individual variability in age of onset is examined in closer detail, using the same dataset consisting of 9064 individuals compiled by the GemHD consortium.<sup>16</sup> At first glance, the between-individual variability of residual age-at-onset appears to *decrease* for larger CAG repeat

sizes (**Figure 8**, left panels; Spearman's  $\rho = -0.91$ ,  $p < 0.001$ ). However, given the strong inverse association between mutant CAG repeat size and age of onset, this association is likely to be confounded by the substantially younger age of onset of the participants with larger CAG repeat sizes. To account for this confounding by age, residual age-at-onset was divided by the predicted age of onset (based on mutant CAG repeat size). Indeed, after accounting for predicted age of onset, the variability of the *relative* residual age-at-onset (i.e., expressed as a fraction of the predicted age of onset) strongly *increased* with larger mutant CAG repeat size (**Figure 8**, right panels; Spearman's  $\rho = 0.81$ ,  $p < 0.001$ ).

Due to the strong association between CAG repeat size and the rate of disease progression (or age of onset as a proxy for the rate of disease progression),<sup>3</sup> it is obvious that delaying or hastening of age of onset by a certain number of years cannot be assumed to indicate the same degree of disease modification for individuals with different mutant CAG repeat sizes (see **Appendix** for an explicit analytical explanation). For example, assuming that pathological burden increases linearly from birth onwards, to delay the expected age of onset by 10 years in two premanifest mutation carriers with expected ages of onset of 30 and 40 years, the rate of disease progression would need to be lowered by 25% and 20%, respectively (**Appendix**). Therefore, using the relative difference between the actual and expected age of onset (as opposed to the absolute difference between actual and expected age of onset as utilized previously<sup>16</sup>), is likely to result in more accurate estimates of potential genetic modifiers of age of onset in HD.

To directly assess this postulate, a GWAS of the *relative* residual age-at-onset in HD was conducted, using the same dataset and otherwise exact same covariates and approach as previously published by the GemHD consortium.<sup>16</sup> Indeed, this approach resulted in the identification of a substantially larger number of genome-wide significant SNPs than reported previously (i.e., 488 vs. 328 (excluding those on chromosome 4 as these may reflect artefacts due to rare CAG repeat sequence haplotypes<sup>16</sup>)) (**Figure 9**). Interestingly, almost all (~98%) genome-wide significant SNPs identified previously were also detected using relative residual age-at-onset as the outcome (**Supplementary Figure 8**), suggesting a sizable increase in statistical power.



To allow direct comparisons, for all potential genetic modifiers reported previously,<sup>16</sup> the effect estimates and associated p-values were compared between the two methods (**Table 4**). Although all effect estimates were directionally consistent and mostly similar between the two methods, using relative residual age-at-onset resulted in the identification of an additional novel genome-wide significant locus tagged by a rare variant (rs138433183) on chromosome 7, located in an intronic region of the *ESYT2* gene (**Figure 10**). Moreover, three other loci that failed to reach genome-wide significance before using residual age-at-onset as the outcome (including the *PMS1* and *PMS2* loci, as well as another locus on chromosome 19 (19AM3)), reached genome-wide significance using relative residual age-at-onset as the outcome (**Table 4**). Conversely, the *SYT9* locus on chromosome 11 did not reach genome-wide significance using the current method, suggesting a potential false positive hit in the previous GWAS. More in-depth characterization of these loci is beyond the scope of this paper and can be accomplished by interested researchers using the provided FUMA output files and summary statistics (**Supplementary Materials**).

## Discussion

This is the first study to assess visit-to-visit within-individual variability and its clinical and genetic correlates in HD. Leveraging detailed genetic and longitudinal clinical data from large cohorts of HD patients, this work 1) establishes within-individual variability in disease expression as an integral part of HD semiology, demonstrating that it increases across all clinical domains with disease duration, mutation size, younger age-at-onset, and lower body weight, 2) provides a mathematical framework linking higher phenotypic variability to increased predictive entropy, revealing a fundamental relation between within-individual variability and energy expenditure, and 3) identifies novel genetic modifiers of both within-individual variability and age-at-onset in HD, detailing new concepts and approaches that could also increase the statistical power of future genetic association studies in HD.

Despite the clinical impression that HD patients exhibit considerable variability in symptoms and signs independent of the measuring method or rater, no previous studies have targeted this particular aspect of HD. Here it was found that all clinical domains of HD, including motor, functional, cognitive and metabolic features of the disease, exhibit pronounced visit-to-visit

variability. Moreover, within-individual variability robustly increased with larger mutant CAG repeat size for the composite, motor and functional scores, as well as BMI. For cognitive scores, within-individual variability tended to increase for CAG repeat sizes in the lower pathogenic range, followed by a leveling off of the effect for larger CAG repeat sizes. Given the strong association between mutation size and cognitive deterioration,<sup>3</sup> this latter finding may thus be due to floor/ceiling effects of the cognitive testing batteries in HD.<sup>23</sup> In addition, an age of onset lower than expected based on mutation size was also robustly associated with a higher degree of within-individual variability across all clinical domains of HD. Collectively, these findings thus establish within-individual variability in disease expression as an integral part of HD semiology.

Interestingly, motor impersistence, referring to the inability to sustain voluntary muscle contractions, has long been recognized as a cardinal feature of HD, and is thought to underlie other well-known clinical signs of the disease, including the ‘milkmaid’s sign’ (rhythmic squeezing of the examiner’s fingers), and the ‘fly-catcher’s tongue’ (inability to maintain tongue protrusion beyond lips).<sup>33</sup> Albeit the precise cause of motor impersistence in HD is not fully elucidated, it could also be regarded as a manifestation of increased within-individual variability, though on a much smaller time-scale. Indeed, higher quantitative motor measures of variability – including grip and tongue force variability, as well as variability in inter-onset intervals during speeded tapping – have been reported in HD mutation carriers.<sup>34</sup> In line with the findings of the current paper, increased variability on these quantitative motor measures was shown to be predictive of a faster rate of progression in early HD, while higher variability in inter-onset intervals during speeded tapping was also associated with a higher risk of disease onset in premanifest HD mutation carriers.<sup>34</sup>

Leveraging the free energy principle, which conceptualizes action as a means through which the brain attempts to minimize sensory prediction errors (i.e., ‘active inference’),<sup>24</sup> the current work also provides a useful theoretical framework for understanding the potential causes of phenotypic variability in HD. Specifically, both analytically and through simulation experiments, it demonstrates that an increase in predictive entropy of the brain’s machinery (e.g., due to gradual breakdown of homeostatic mechanisms as a consequence of progressive neuronal dysfunction and loss) will inevitably lead to increased variability in phenotype. Although an exhaustive

discussion of the free energy formulation is beyond the scope of this paper, here it is important to stress that within this framework ‘action’ is not necessarily limited to macroscopic, observable movements, but could encompass any active motion, including, e.g., secretory activity of internal (hormonal) glands, gastrointestinal motility, as well as modulation of the oculomotor and cardiorespiratory system. Interestingly, previously we found a higher ‘approximate entropy’ of prolactin secretion in early-stage HD patients compared to age, sex and BMI matched healthy control subjects,<sup>35</sup> further supporting the notion of within-individual variability as a general feature of the disease. It should be noted though that the proposed Bayesian framework for linking entropy of brain states to (endo)phenotypic variability will only hold under the assumption that the agent will have retained the capacity to engage in active inference. For example, once patients have become bedridden in advanced stages of the disease, motor variability will be low, not because of reduced predictive entropy, but simply due to a dramatic loss of motor control.

One common but still little understood feature of HD is progressive weight loss, which is strongly associated with mutation size, already starts in the premanifest stages of the disease despite adequate dietary intake, and has been associated with higher resting energy expenditure.<sup>36-38</sup> Although a higher BMI is robustly associated with a slower rate of disease progression in HD,<sup>27</sup> in a previous Mendelian Randomization study we demonstrated that this association is not causal, but likely a *consequence* of a slower rate of disease progression.<sup>29</sup> Consistent with the proposed framework linking a higher predictive entropy to increased phenotypic variability, and consequently to increased energy expenditure, in the present study a higher variability of motor features of HD was indeed associated with both lower body weight and a faster rate of weight loss, *independent* of the severity of motor impairment. These findings thus indicate that higher energy expenditure in HD may not necessarily be due to inherent defects in systemic energy regulation, but rather could be an inevitable consequence of increased energy requirements to maintain homeostasis in the face of a progressively inefficient capacity of a diseased brain for predictive inference. In fact, this could also account for the ubiquity of weight loss in other neurodegenerative diseases, in which emaciation generally accompanies a faster rate of disease progression.<sup>39</sup> Therefore, in the context of HD and other brain disorders, BMI and weight loss could be regarded as proxies for the underlying rate of neuropathology.

This work also demonstrates that accounting for phenotypic variability can accelerate the discovery of novel genetic modifiers of HD. First, a GWAS of within-individual BMI variability identified a novel genome-wide significant locus (rs62231080) near the *NUP50* gene, which was also associated with the expression levels of this gene. Interestingly, variants in *NUP50*, encoding the nucleopore basket protein NUP50, were recently also associated with amyotrophic lateral sclerosis, with loss of NUP50 affecting splicing regulation, neuronal survival and motor function in multiple animal models.<sup>40</sup> Moreover, disruption of nucleocytoplasmic transport is a consistent feature of HD models and was recently also linked to neurodevelopmental defects in cellular and *Drosophila* models of HD.<sup>41-43</sup> Notably, the rs62231080 locus was distinct from three other novel loci associated with mean BMI levels. Interestingly, *EIF3F*, which was identified through gene-based analysis and encodes a subunit of the eukaryotic translation initiation factor eIF3, has also been linked to neurodevelopmental defects.<sup>44</sup> Importantly, eIF3F favors repeat-associated non-ATG (RAN) translation, a key mechanism through which repeat expansions could induce neurodegeneration, with its knockdown reducing the levels of RAN proteins.<sup>45</sup> Given that both BMI and its variability could reflect the rate of neuropathology as argued above, these loci thus do not necessarily need to be involved in systemic energy regulation per se, but may also exert their influence by affecting neurodegeneration in HD. Similarly, GWASs of within-individual variability in the UHDRS TMS and SDMT scores, identified other novel loci, although these latter loci did not reach genome-wide significance likely due to the relatively small sample size available for the longitudinal genetic association studies. Second, by accounting for the association between mutation size and age of onset variability through the derivation of a new metric – i.e., relative residual age-at-onset – it was demonstrated that also the statistical power of GWASs of age of onset in HD can be increased considerably. Indeed, utilizing relative instead of absolute residual age-at-onset resulted in the identification of 49% more genome-wide significant SNPs, as well as the discovery of an additional genome-wide significant locus in the *ESYT2* gene, encoding the extended synaptotagmin 2 protein that is important for calcium-dependent lipid binding and transport. Interestingly, gene fusion between *HTT* and *ESYT2* has previously been linked to breast cancer.<sup>46</sup> Additionally, given that extended synaptotagmin 2 mediates endocytosis of activated fibroblast growth factor receptors,<sup>47</sup> and fibroblast growth factor treatment has been shown to be

neuroprotective in HD animal models,<sup>48</sup> alterations in the growth factor signaling pathway might partly underlie these observed effects.

Although due to the strong association between mutant *HTT* CAG repeat size and the rate of disease progression, as well as availability of large and comprehensive datasets, HD provided an excellent model for deriving and testing the theoretical framework presented in this paper, there is no pertinent reason to assume that the association between increased predictive entropy and phenotypic variability would be *specific* to HD. Indeed, the proposed theory provides an intuitive framework for explaining a range of other little understood phenomena across different neurological diseases, including the association of increased visit-to-visit variability in body weight and blood pressure and a faster rate of disease progression in patients with Alzheimer and Parkinson disease.<sup>13, 14</sup> weight loss preceding neurodegeneration by many years in other neurodegenerative diseases,<sup>39</sup> the association between body weight variability and a higher risk of cognitive impairment and dementia,<sup>11</sup> as well as the robust associations of many neuron-specific genes and BMI.<sup>49</sup> Similarly, the concept of relative age-at-onset is likely to be useful in the study of other disorders in which there is a strong association between genotype and age of onset, especially other repeat expansion disorders exhibiting an inverse association between mutation size and age of onset. Finally, it should be noted that the concept of a ‘Bayesian agent’ is not necessarily limited to multicellular organisms and could be extended to other complex model systems capable of active inference, including cells and organoids. Therefore, the proposed link between predictive entropy and phenotypic variability could also be probed in model systems of HD and other (neurodegenerative) diseases where any dynamically regulated ‘phenotype’ could serve as a potential read-out. An exhaustive discussion of these other topics is beyond the remit of the current paper, but will be elaborated on in future work.

The current work also has several potential limitations. First, the *intraindividual* rate of disease progression was modeled linearly in time by using mixed effects models with random intercepts and slopes for age, because inclusion of higher order random terms for age led to model non-convergence, likely due to the relative sparsity of within-individual longitudinal measurements to enable stable estimates of higher order random effects. However, all models included a quadratic term for age as a fixed effect, which accounts for non-linear changes of the mean of the outcome

measures over time. Moreover, an alternative method based on VIM, which uses a non-linear exponential model to estimate variability independent of the mean,<sup>9</sup> yielded similar results. Therefore, it is highly unlikely that increases of within-individual variability with age or mutant CAG repeat size could be accounted for by exponential increases in the rate of disease progression. Second, the visit-to-visit within-individual variability estimates were solely based on clinical outcome measures, including motor, functional and cognitive measures and BMI, obtained at about yearly or longer intervals. More objective, quantitative measures, such as the Q-Motor testing battery and other wearable motor and cognitive testing paradigms,<sup>34, 50, 51</sup> applied at shorter intervals and more frequently, are thus likely to be more sensitive and yield more accurate estimates of within-individual variability profiles over time. Nevertheless, given that even the relatively crude clinical outcome measures utilized here could robustly capture at least part of the within-individual variability in phenotype, the current results are likely to represent underestimates of the true magnitude of within-individual variability in HD. Third, the available sample size for the GWAS of within-individual variability in HD was relatively small. Despite this, a genome-wide significant locus associated with within-individual variability of BMI was identified, while several other loci exhibited suggestive significance for within-individual variability of BMI, as well as motor and cognitive scores. Genotyping of a larger portion of the Enroll-HD cohort will be highly instrumental in identifying more genetic modifiers of within-individual variability in HD.

In conclusion, within-individual variability in disease expression is an integral part of HD semiology, accounting for which could facilitate the discovery of pathogenic mechanisms and outcomes directly relevant to the development of disease modifying therapies. Future studies should assess the utility of within-individual variability as an additional outcome measure: 1) For the (re-)interpretation of the findings of past and current clinical trials, as well as for the design of future (clinical) studies, 2) As a translatable outcome measure in cell and animal studies, as by definition it provides a scale invariant read-out sensitive to the earliest stages of pathology, and 3) In the context of other diseases, especially other repeat expansion and neurodegenerative diseases.

## **Acknowledgements**

This paper is dedicated to all study participants, care givers, medical professionals, researchers and other support staff who contributed to the unique datasets which enabled this work. Data used in this work was generously provided by the participants in the Enroll-HD study and made available by CHDI Foundation, Inc. Enroll-HD is a global clinical research platform intended to accelerate progress towards therapeutics for Huntington's disease; core datasets are collected annually on all research participants as part of this multi-center longitudinal observational study. Enroll-HD is sponsored by CHDI Foundation, Inc. Enroll-HD would not be possible without the vital contribution of the research participants and their families. Data used in this work was generously provided by the combined efforts of the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium (dbGaP accession number phs000371). NAA is partly supported by a European Research Council Starting Grant (Number: 101041677).

## References

1. Roos RA. Huntington's disease: a clinical review. *Orphanet JRareDis*. 2010 2010;5(1):40. doi:10.1186/1750-1172-5-40
2. Walker FO. Huntington's disease. *Lancet*. 1/20/2007 2007;369(9557):218-228. doi:10.1016/S0140-6736(07)60111-1
3. Aziz NA, van der Burg JM, Tabrizi SJ, Landwehrmeyer GB. Overlap between age-at-onset and disease progression determinants in Huntington disease. *Neurology*. 2018;90(24):e2099-e2106. doi:10.1212/01.wnl.0000334276.09729.0e
4. Schobel SA, Palermo G, Auinger P, et al. Motor, cognitive, and functional declines contribute to a single progressive factor in early HD. *Neurology*. Dec 12 2017;89(24):2495-2502. doi:10.1212/wnl.0000000000004743
5. Ko S, German CA, Jensen A, et al. GWAS of longitudinal trajectories at biobank scale. *American journal of human genetics*. Mar 3 2022;109(3):433-445. doi:10.1016/j.ajhg.2022.01.018
6. Mahinrad S, Bennett DA, Sorond FA, Gorelick PB. Blood pressure variability, dementia, and role of antihypertensive medications in older adults. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. Jul 2023;19(7):2966-2974. doi:10.1002/alz.12935
7. Sible IJ, Nation DA. Blood Pressure Variability and Cognitive Decline: A Post Hoc Analysis of the SPRINT MIND Trial. *Am J Hypertens*. Feb 24 2023;36(3):168-175. doi:10.1093/ajh/hpac128
8. Goldbourt U, Grossman E. Blood pressure variability at midlife is associated with all-cause, coronary heart disease and stroke long term mortality. *J Hypertens*. Sep 2020;38(9):1722-1728. doi:10.1097/hjh.0000000000002447
9. Sible IJ, Nation DA. Visit-to-Visit Blood Pressure Variability and CSF Alzheimer Disease Biomarkers in Cognitively Unimpaired and Mildly Impaired Older Adults. *Neurology*. Jun 14 2022;98(24):e2446-e2453. doi:10.1212/wnl.00000000000200302
10. Zhou JJ, Koska J, Bahn G, Reaven P. Fasting Glucose Variation Predicts Microvascular Risk in ACCORD and VADT. *J Clin Endocrinol Metab*. Mar 25 2021;106(4):1150-1162. doi:10.1210/clinem/dgaa941
11. Zonneveld MH, Noordam R, Sabayan B, et al. Weight loss, visit-to-visit body weight variability and cognitive function in older individuals. *Age Ageing*. Jan 8 2023;52(1)doi:10.1093/ageing/afac312
12. Massey RJ, Siddiqui MK, Pearson ER, Dawed AY. Weight variability and cardiovascular outcomes: a systematic review and meta-analysis. *Cardiovasc Diabetol*. Jan 9 2023;22(1):5. doi:10.1186/s12933-022-01735-x
13. Urso D, van Wamelen DJ, Batzu L, et al. Clinical trajectories and biomarkers for weight variability in early Parkinson's disease. *NPJ Parkinsons Dis*. Aug 2 2022;8(1):95. doi:10.1038/s41531-022-00362-3
14. de Heus RAA, Olde Rikkert MGM, Tully PJ, Lawlor BA, Claassen J. Blood Pressure Variability and Progression of Clinical Alzheimer Disease. *Hypertension*. Nov 2019;74(5):1172-1180. doi:10.1161/hypertensionaha.119.13664
15. Landwehrmeyer GB, Fitzer-Attas CJ, Giuliano JD, et al. Data Analytics from Enroll-HD, a Global Clinical Research Platform for Huntington's Disease. *Movement Disorders Clinical Practice*. 2017;4(2):212-224. doi:10.1002/mdc3.12388



16. GeM\_HD\_consortium. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*. Aug 8 2019;178(4):887-900.e14. doi:10.1016/j.cell.2019.06.036
17. meyer-lab-cshl/plinkQC: plinkQC 0.3.2 (v0.3.2). doi: 10.5281/zenodo.3934294
18. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*. Jun 17 2012;44(7):821-4. doi:10.1038/ng.2310
19. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. Nov 28 2017;8(1):1826. doi:10.1038/s41467-017-01261-5
20. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol*. Apr 2015;11(4):e1004219. doi:10.1371/journal.pcbi.1004219
21. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. Dec 23 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004
22. German CA, Sinsheimer JS, Zhou J, Zhou H. WiSER: Robust and scalable estimation and inference of within-subject variances from intensive longitudinal data. *Biometrics*. Dec 2022;78(4):1313-1327. doi:10.1111/biom.13506
23. Mestre TA, Bachoud-Lévi A-C, Marinus J, et al. Rating scales for cognition in Huntington's disease: Critique and recommendations. *Movement Disorders*. 2018;33(2):187-195. doi:10.1002/mds.27227
24. Friston K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*. 2010/02/01 2010;11(2):127-138. doi:10.1038/nrn2787
25. Buckley CL, Kim CS, McGregor S, Seth AK. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*. 2017;81:55-79. doi:10.1016/j.jmp.2017.09.004
26. Misra N, Singh H, Demchuk E. Estimation of the entropy of a multivariate normal distribution. *Journal of Multivariate Analysis*. 2005/02/01/ 2005;92(2):324-342. doi:10.1016/j.jmva.2003.10.003
27. van der Burg JMM, Gardiner SL, Ludolph AC, Landwehrmeyer GB, Roos RAC, Aziz NA. Body weight is a robust predictor of clinical progression in Huntington disease. *Ann Neurol*. Aug 05 2017;82(3):479-483. doi:10.1002/ana.25007
28. van der Burg JM, Bacos K, Wood NI, et al. Increased metabolism in the R6/2 mouse model of Huntington's disease. *Neurobiol Dis*. 1/2008 2008;29(1):41-51. Not in File.
29. van der Burg JMM, Weydt P, Landwehrmeyer GB, Aziz NA. Effect of Body Weight on Age at Onset in Huntington Disease: A Mendelian Randomization Study. *Neurology Genetics*. Aug 2021;7(4):e603. doi:10.1212/nxg.0000000000000603
30. Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nature genetics*. Oct 2021;53(10):1415-1424. doi:10.1038/s41588-021-00931-x
31. Kichaev G, Bhatia G, Loh PR, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *American journal of human genetics*. Jan 3 2019;104(1):65-75. doi:10.1016/j.ajhg.2018.11.008
32. Deshmukh AL, Porro A, Mohiuddin M, et al. FAN1, a DNA Repair Nuclease, as a Modifier of Repeat Expansion Disorders. *Journal of Huntington's disease*. 2021;10(1):95-122. doi:10.3233/jhd-200448

33. Shin C, Kim R, Yoo D, et al. A Practical Guide for Clinical Approach to Patients With Huntington's Disease in Korea. *J Mov Disord*. Apr 2024;17(2):138-149. doi:10.14802/jmd.24040
34. Tabrizi SJ, Scahill RI, Owen G, et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol*. 7/2013 2013;12(7):637-649. Not in File.
35. Aziz NA, Pijl H, Frolich M, Roelfsema F, Roos RA. Altered thyrotropic and lactotropic axes regulation in Huntington's disease. *ClinEndocrinol(Oxf)*. 10/2010 2010;73(4):540-545. Not in File.
36. Aziz NA, Pijl H, Frolich M, et al. Systemic energy homeostasis in Huntington's disease patients. *JNeurolNeurosurgPsychiatry*. 8/14/2010 2010;81(11):1233-1237. Not in File.
37. Mochel F, Haller RG. Energy deficit in Huntington disease: why it matters. *JClinInvest*. 2/2011 2011;121(2):493-499. Not in File.
38. Aziz NA, van der Burg JM, Landwehrmeyer GB, et al. Weight loss in Huntington disease increases with higher CAG repeat number. *Neurology*. 11/4/2008 2008;71(19):1506-1513. Not in File.
39. Aziz NA, van der Marck MA, Pijl H, Olde Rikkert MG, Bloem BR, Roos RA. Weight loss in neurodegenerative disorders. *JNeurol*. 12/2008 2008;255(12):1872-1880. Not in File.
40. Megat S, Mora N, Sanogo J, et al. Integrative genetic analysis illuminates ALS heritability and identifies risk genes. *Nature communications*. Jan 20 2023;14(1):342. doi:10.1038/s41467-022-35724-1
41. Das MR, Chang Y, Anderson R, et al. Repeat-associated non-AUG translation induces cytoplasmic aggregation of CAG repeat-containing RNAs. *Proceedings of the National Academy of Sciences of the United States of America*. Jan 17 2023;120(3):e2215071120. doi:10.1073/pnas.2215071120
42. Dubey SK, Lloyd TE, Tapadia MG. Disrupted nuclear import of cell cycle proteins in Huntington's/PolyQ disease causes neurodevelopment defects in cellular and Drosophila model. *Heliyon*. Feb 29 2024;10(4):e26393. doi:10.1016/j.heliyon.2024.e26393
43. Lange J, Wood-Kaczmar A, Ali A, et al. Mislocalization of Nucleocytoplasmic Transport Proteins in Human Huntington's Disease PSC-Derived Striatal Neurons. *Frontiers in cellular neuroscience*. 2021;15:742763. doi:10.3389/fncel.2021.742763
44. Hüffmeier U, Kraus C, Reuter MS, et al. EIF3F-related neurodevelopmental disorder: refining the phenotypic and expanding the molecular spectrum. *Orphanet J Rare Dis*. Mar 18 2021;16(1):136. doi:10.1186/s13023-021-01744-1
45. Ayhan F, Perez BA, Shorrock HK, et al. SCA8 RAN polySer protein preferentially accumulates in white matter regions and is regulated by eIF3F. *The EMBO journal*. Oct 1 2018;37(19)doi:10.15252/embj.201899023
46. Yoshihara K, Wang Q, Torres-Garcia W, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*. Sep 10 2015;34(37):4845-54. doi:10.1038/onc.2014.406
47. Jean S, Mikryukov A, Tremblay MG, et al. Extended-synaptotagmin-2 mediates FGF receptor endocytosis and ERK activation in vivo. *Dev Cell*. Sep 14 2010;19(3):426-39. doi:10.1016/j.devcel.2010.08.007
48. Jin K, LaFevre-Bernt M, Sun Y, et al. FGF-2 promotes neurogenesis and neuroprotection and prolongs survival in a transgenic mouse model of Huntington's disease. *Proceedings of the National Academy of Sciences of the United States of America*. Dec 13 2005;102(50):18189-94. doi:10.1073/pnas.0506375102

49. Hemerich D, Svenstrup V, Obrero VD, et al. An integrative framework to prioritize genes in more than 500 loci associated with body mass index. *American journal of human genetics*. May 9 2024;doi:10.1016/j.ajhg.2024.04.016
50. Gordon MF, Grachev ID, Mazeh I, et al. Quantification of Motor Function in Huntington Disease Patients Using Wearable Sensor Devices. *Digit Biomark*. Sep-Dec 2019;3(3):103-115. doi:10.1159/000502136
51. Reilmann R, Schubert R. Motor outcome measures in Huntington disease clinical trials. *Handbook of clinical neurology*. 2017;144:209-225. doi:10.1016/b978-0-12-801893-4.00018-3

**Table 1. Summary of model estimates**

	Composite score <sup>a</sup>		Motor score <sup>a</sup>		TFC score <sup>a</sup>		SDMT score <sup>a,c</sup>		SWRT score <sup>a,c</sup>		BMI <sup>a</sup>	
	Effect estimate <sup>b</sup>	p-value	Effect estimate <sup>b</sup>	p-value	Effect estimate <sup>b</sup>	p-value	Effect estimate <sup>b</sup>	p-value	Effect estimate <sup>b</sup>	p-value	Effect estimate <sup>b</sup>	p-value
<i>Fixed-effect estimates</i>												
Intercept	<b>-0.136</b> (0.040)	7e-4	<b>0.092</b> (0.022)	<1e-4	<b>-0.090</b> (0.025)	3e-4	<b>-0.023</b> (0.024)	0.350	<b>-0.042</b> (0.027)	0.119	-0.102 (0.039)	0.009
Sex (m)	0.038 (0.026)	0.149	-0.026 (0.014)	0.062	<b>0.053</b> (0.015)	6e-4	<b>-0.036</b> (0.015)	0.017	-0.009 (0.017)	0.603	<b>0.099</b> (0.023)	<1e-04
Age	<b>-1.012</b> (0.016)	<1e-99	<b>1.073</b> (0.014)	<1e-99	<b>-1.002</b> (0.014)	<1e-99	<b>-0.957</b> (0.010)	<1e-99	<b>-0.995</b> (0.011)	<1e-99	<b>-0.065</b> (0.012)	<1e-6
Age <sup>2</sup>	<b>-0.237</b> (0.011)	<1e-99	<b>0.258</b> (0.006)	<1e-99	<b>-0.275</b> (0.007)	<1e-99	<b>-0.109</b> (0.007)	<1e-56	<b>-0.207</b> (0.007)	<1e-99	<b>-0.134</b> (0.009)	<1e-46
CAG size	<b>-1.034</b> (0.023)	<1e-99	<b>1.207</b> (0.017)	<1e-99	<b>-1.047</b> (0.018)	<1e-99	<b>-1.021</b> (0.015)	<1e-99	<b>-1.110</b> (0.017)	<1e-99	<b>-0.291</b> (0.018)	<1e-58
Age × CAG	<b>-0.448</b> (0.018)	<1e-99	<b>0.577</b> (0.011)	<1e-99	<b>-0.500</b> (0.013)	<1e-99	<b>-0.386</b> (0.012)	<1e-99	<b>-0.496</b> (0.013)	<1e-99	<b>-0.187</b> (0.014)	<1e-40
<i>WS variability estimates</i>												
Intercept	<b>-1.683</b> (0.086)	<1e-84	<b>-2.377</b> (0.069)	<1e-99	<b>-1.918</b> (0.069)	<1e-99	<b>-2.607</b> (0.048)	<1e-99	<b>-2.134</b> (0.051)	<1e-99	<b>-1.486</b> (0.075)	<1e-85
Sex (m)	-0.037 (0.054)	0.492	-0.031 (0.046)	0.533	-0.010 (0.048)	0.826	-0.024 (0.033)	0.458	0.033 (0.033)	0.313	<b>-0.523</b> (0.052)	<1e-23
Age	<b>0.173</b> (0.031)	<1e-07	<b>0.946</b> (0.055)	<1e-64	<b>0.866</b> (0.046)	<1e-77	<b>-0.397</b> (0.020)	<1e-91	<b>-0.095</b> (0.021)	<1e-05	<b>0.120</b> (0.041)	0.003
CAG size	<b>0.499</b> (0.065)	<1e-13	<b>1.068</b> (0.059)	<1e-72	<b>0.941</b> (0.050)	<1e-78	<b>-0.280</b> (0.025)	<1e-32	0.033 (0.026)	0.208	<b>0.296</b> (0.044)	<1e-10
Age × CAG	<b>0.462</b> (0.030)	<1e-52	<b>0.096</b> (0.020)	<1e-05	<b>0.139</b> (0.020)	<1e-09	<b>-0.102</b> (0.017)	<1e-09	0.001 (0.017)	0.942	<b>0.101</b> (0.028)	3e-4

<sup>a)</sup> To enable comparison of the effect estimates across the different outcome measures, as well as to reduce collinearity among the predictor variables, all outcome measures and continuous predictor variables were standardized to a mean of zero and a standard deviation of one before inclusion in the models.

<sup>b)</sup> All effect estimates were obtained using the within-subject variance estimator by robust regression (WiSER) Julia package with default settings and 100 iterations.<sup>22</sup> For each of the six clinical outcome measures a separate model was run including sex, age, age<sup>2</sup>, mutant CAG repeat size and its interaction with age as fixed-effects, an intercept and slope for age as random effects, and sex, age, mutant CAG repeat size and its interaction with age as within-subject variability predictors. Effect estimates represent standardized regression coefficients (standard error) with all statistically significant effect estimates highlighted in bold.

<sup>c)</sup> Both Unified Huntington Disease Rating Scale cognitive scores indicate the total number of correct items (in 1 minute).

**Abbreviations:** BMI = body mass index, CAG = cytosine-adenine-guanine, SD = standard deviation, SDMT = symbol digit modalities test, SWRT = Stroop word reading test, WS = within-subject

**Table 2. The association between residual age-at-onset and within-subject variability**

	Residual age-at-onset			
	<i>Fixed-effect estimates</i>		<i>WS variability estimates</i>	
	<i>Effect estimate<sup>b</sup></i>	<i>p-value</i>	<i>Effect estimate<sup>b</sup></i>	<i>p-value</i>
<b>Composite score<sup>a</sup></b>	<b>0.595 (0.016)</b>	<1e-99	<b>-0.057 (0.029)</b>	0.046
<b>Motor score<sup>a</sup></b>	<b>-0.483 (0.026)</b>	<1e-78	<b>-0.707 (0.065)</b>	<1e-27
<b>TFC score<sup>a</sup></b>	<b>0.745 (0.019)</b>	<1e-99	<b>-0.084 (0.027)</b>	0.002
<b>SDMT score<sup>a,c</sup></b>	<b>0.217 (0.026)</b>	<1e-15	<b>-0.488 (0.133)</b>	2e-4
<b>SWRT score<sup>a,c</sup></b>	<b>0.390 (0.023)</b>	<1e-63	<b>-0.368 (0.098)</b>	2e-4
<b>BMI<sup>a</sup></b>	-0.025 (0.028)	0.369	<b>-0.183 (0.072)</b>	0.011

<sup>a)</sup> All outcome measures and continuous predictor variables were standardized to a mean of zero and a standard deviation of one before inclusion in the models.

<sup>b)</sup> All effect estimates were obtained using the within-subject variance estimator by robust regression (WiSER) Julia package with default settings and 100 iterations (except for BMI, where 4 iterations were used, because more iterations resulted in non-convergence of the model).<sup>22</sup> For each of the six clinical outcome measures a separate model was run including sex, age, age<sup>2</sup>, mutant CAG repeat size and its interaction with age, and residual age at onset as fixed-effects, an intercept and slope for age as random effects, and sex, age, mutant CAG repeat size and its interaction with age, and residual age-at-onset as within-subject variability predictors. Effect estimates represent standardized regression coefficients (standard error) with all statistically significant effect estimates highlighted in bold.

<sup>c)</sup> Both Unified Huntington Disease Rating Scale cognitive scores indicate the total number of correct items (in 1 minute).

**Abbreviations:** BMI = body mass index, CAG = cytosine-adenine-guanine, SD = standard deviation, SDMT = symbol digit modalities test, SWRT = Stroop word reading test, WS = within-subject

**Table 3. Genetic loci associated with within-individual variability or mean levels of body mass index, total motor score or symbol digit modalities test scores in Huntington disease.**

Trait	Outcome	Chr	Lead SNP	BP (hg19)	Effect allele	MAF (%)	Effect size <sup>a</sup>	SE	P-value	Candidate modifier genes
<b>BMI</b>	<i>variability</i>	<b>22</b>	<b>rs62231080</b>	<b>45484152</b>	A	<b>21.8</b>	<b>-0.260</b>	<b>4.57E-02</b>	<b>1.48E-08</b>	<i>NUP50</i>
	<i>mean</i>	<b>16</b>	<b>rs200717776</b>	<b>50145334</b>	CT	<b>10.1</b>	<b>0.408</b>	<b>7.35E-02</b>	<b>3.43E-08</b>	<i>HEATR3</i>
		<b>18</b>	<b>rs11874212</b>	<b>73503347</b>	A	<b>30.1</b>	<b>0.222</b>	<b>3.90E-02</b>	<b>1.59E-08</b>	<i>TSHZ1</i>
<b>TMS</b>	<i>variability</i>	16	rs772972	27838619	T	32.4	-0.188	3.68E-02	4.10E-07	<i>GSG1L</i>
	<i>mean</i>	3	rs1984782	153289784	G	46.3	0.125	2.50E-02	6.21E-07	<i>P2RY1</i>
		6	rs36079127	99300452	G	37.9	-0.134	2.55E-02	1.93E-07	<i>FBXL4</i>
<b>SDMT</b>	<i>variability</i>	1	rs4101233	91238773	T	38.5	0.204	3.85E-02	1.26E-07	<i>CDC7, ZNF644</i>
	<i>mean</i>	<b>5</b>	<b>rs34322595</b>	<b>79998532</b>	<b>T</b>	<b>31.2</b>	<b>-0.151</b>	<b>2.64E-02</b>	<b>1.30E-08</b>	<i>MSH3</i>
		12	rs55833733	125839198	T	22.2	0.146	2.74E-02	1.16E-07	<i>TMEM132B</i>
		13	rs9559664	110471106	G	15.3	-0.175	3.41E-02	3.45E-07	<i>IRS2, RN7SKP10</i>
		15	rs55666775	100431982	A	9.5	-0.217	4.07E-02	1.09E-07	<i>CTD-2054N24.2, CTD-3076O17.1</i>

<sup>a</sup>) Effect estimates represent standardized regression coefficients. All loci that passed the genome-wide (p-value < 5E-8) or suggestive (p-value < 1E-06) statistical significance threshold are summarized. Genome-wide significant loci are marked in bold.

**Abbreviations:** Chr = chromosome, BP = base pair position in Human Genome version 19 coordinates, BMI = body mass index, MAF = minor allele frequency, SDMT = symbol digit modalities test, SE = standard error, TMS = total motor score.

**Table 4. Comparison between absolute vs. relative residual age-at-onset genetic modifiers**

Chr	Modifier	Lead SNP	BP (hg19)	Minor allele	MAF (%)	Effect size <sup>a</sup> (years)	P-value	Effect size (%)	P-value	Candidate modifier genes	
1	1AM1b	rs567500111	164283625	A	0.3	-4.4	6.9E-06	-8.2	0.000147		
2	<b>2AM1</b>	<b>rs3791767</b>	<b>190639915</b>	<b>C</b>	<b>20.7</b>	<b>-0.8</b>	<b>6.3E-08</b>	<b>-1.8</b>	<b>3.76E-09</b>	<b>PMS1</b>	
3	3AM1	rs1799977	37053568	G	31.0	0.8	5.1E-10	1.6	8.92E-10	<b>MLH1</b>	
5	5AM1	rs701383	79913275	A	25.7	-0.8	5.5E-10	-1.6	1.45E-08		
	5AM2	rs113361582	80086504	G	0.3	6.1	1.3E-09	1.6	5.79E-13	<b>MSH3</b>	<b>DHFR</b>
	5AM3	rs1650742	79990883	G	33.1	0.6	1.6E-06	1.2	2.84E-06		
	5BM1	rs79727797	145886836	A	2.4	2.3	3.8E-10	5.5	8.9E-12	<b>TCERG1</b>	
7	<b>7AM1</b>	<b>rs74302792</b>	<b>6079993</b>	<b>A</b>	<b>15.9</b>	<b>0.8</b>	<b>7.4E-08</b>	<b>2.0</b>	<b>6.21E-09</b>	<b>PMS2</b>	
	<b>7AM2</b>	<b>rs138433183</b>	<b>158535496</b>	<b>G</b>	<b>0.2</b>	<b>-7.5</b>	<b>6.6E-08</b>	<b>-17.0</b>	<b>4.27E-08</b>	<b>ESYT2</b>	<b>NCAPG2</b>
8	8AM1	rs79136984	103213640	T	8.2	-1.2	3.6E-09	-2.9	8.37E-11	<b>RRM2B</b>	<b>UBR5</b>
11	11AM1	rs7936234	96106737	A	19.6	0.6	1.7E-05	1.6	4.06E-07	<b>CCDC82</b>	
	<b>11BM1b</b>	<b>rs79714630</b>	<b>7303052</b>	<b>G</b>	<b>0.1</b>	<b>-9.6</b>	<b>1.1E-08</b>	<b>-17.8</b>	<b>1.85E-06</b>	<b>SYT9</b>	
12	12AM1b	rs140253376	108992727	A	0.2	-6.1	8.3E-06	-11.4	0.000192		
15	15AM1	rs150393409	31202961	A	1.4	-5.2	1.8E-28	-11.3	2.39E-27		
	15AM2	rs35811129	31241346	A	27.5	1.3	9.4E-26	3.1	2.41E-29	<b>FAN1</b>	
	15AM3	rs151322829	31197995	T	0.7	-3.8	1.4E-08	-8.4	1.18E-08		
	15AM4	rs34017474	31230611	C	38.2	0.8	8.5E-11	1.8	2.09E-12		
16	16AM1b	rs187055476	27873637	G	0.3	-6.1	5.5E-09	-12.8	4.63E-08	<b>GSG1L</b>	
18	18AM1b	rs530017366	56126806	T	0.2	-5.1	1.2E-04	-8.5	0.003895		
19	19AM1	rs274883	48622545	G	16.7	0.9	5.3E-09	2.0	2.1E-09		
	<b>19AM2</b>	<b>rs3730945</b>	<b>48645976</b>	<b>G</b>	<b>37.1</b>	<b>-0.6</b>	<b>5.8E-07</b>	<b>-1.4</b>	<b>1.89E-08</b>	<b>LIG1</b>	
	19AM3	rs145821638	48620943	A	0.1	7.7	1.5E-06	17.4	9.48E-07		

<sup>a</sup>) These effect estimates are based on the results of the continuous analysis by the GemHD consortium (see *Table 1* in the original publication).<sup>16</sup> Loci with relevant differences between the two methods are highlighted in bold.

**Abbreviations:** Chr = chromosome, BP = base pair position in Human Genome version 19 coordinates, MAF = minor allele frequency.

**Figure 1. Within-individual clinical variability in Huntington disease increases with larger mutant *HTT* CAG repeat size.**

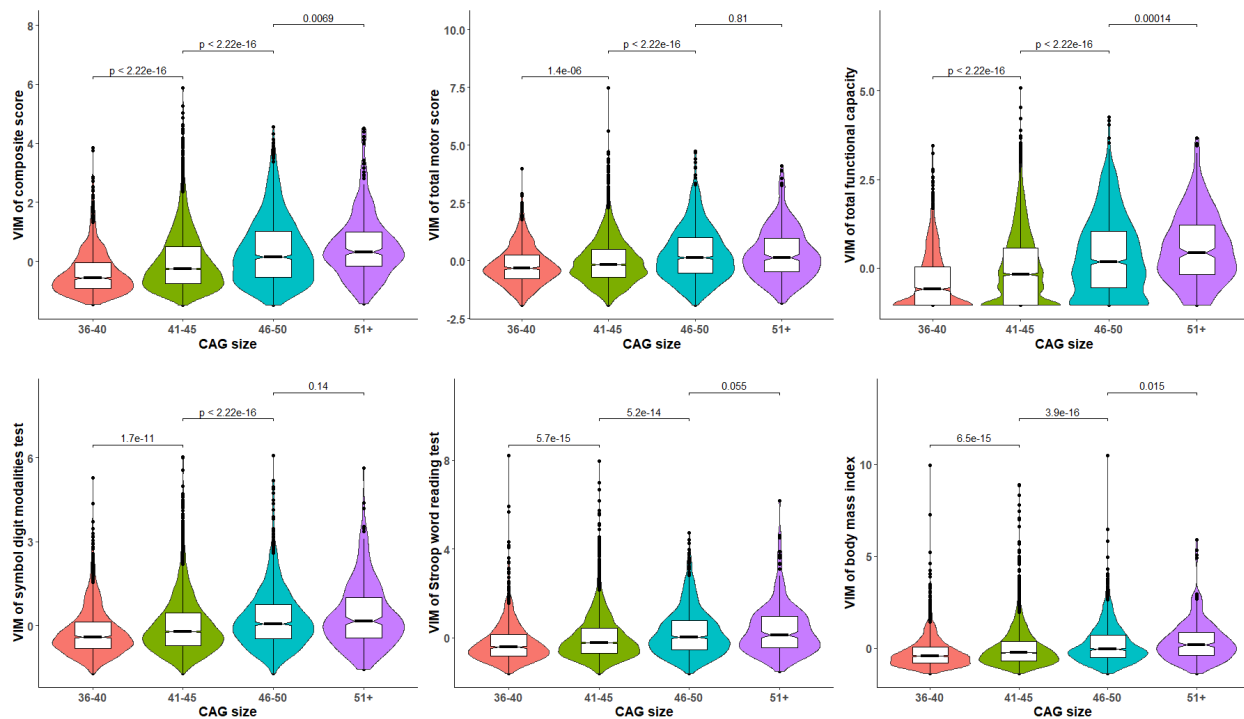
The violin plots depict the association of mutant *HTT* CAG repeat size and the standardized variance independent of the mean (VIM) metric for the Unified Huntington Disease Rating Scale composite score, total motor score, total functional capacity, and symbol digit modalities and Stroop word reading test scores, as well as body mass index.

VIM of all clinical measures significantly increased with larger mutant *HTT* CAG repeat size, although for the cognitive scores, the effect tended to level off for larger mutation sizes. The

boxes indicate the interquartile ranges around the median (thick black lines), with values deviating more than 1.5 times the interquartile range from the median represented by black dots.

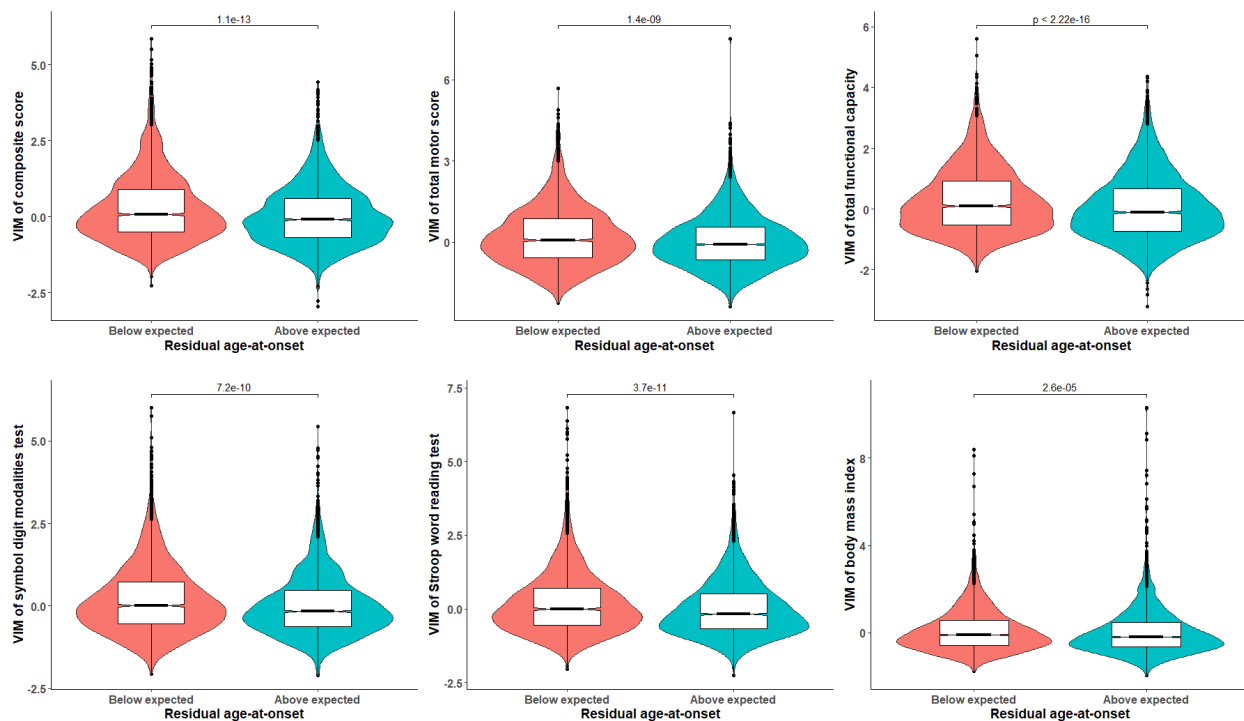
After regressing out the effects of sex, age at baseline, and domain-specific disease severity at baseline (e.g., for composite score, the composite score at baseline was used as a measure of

disease severity) from the VIM scores, the residuals were compared across different categories of mutant CAG repeat size using the non-parametric Wilcoxon's rank sum test.

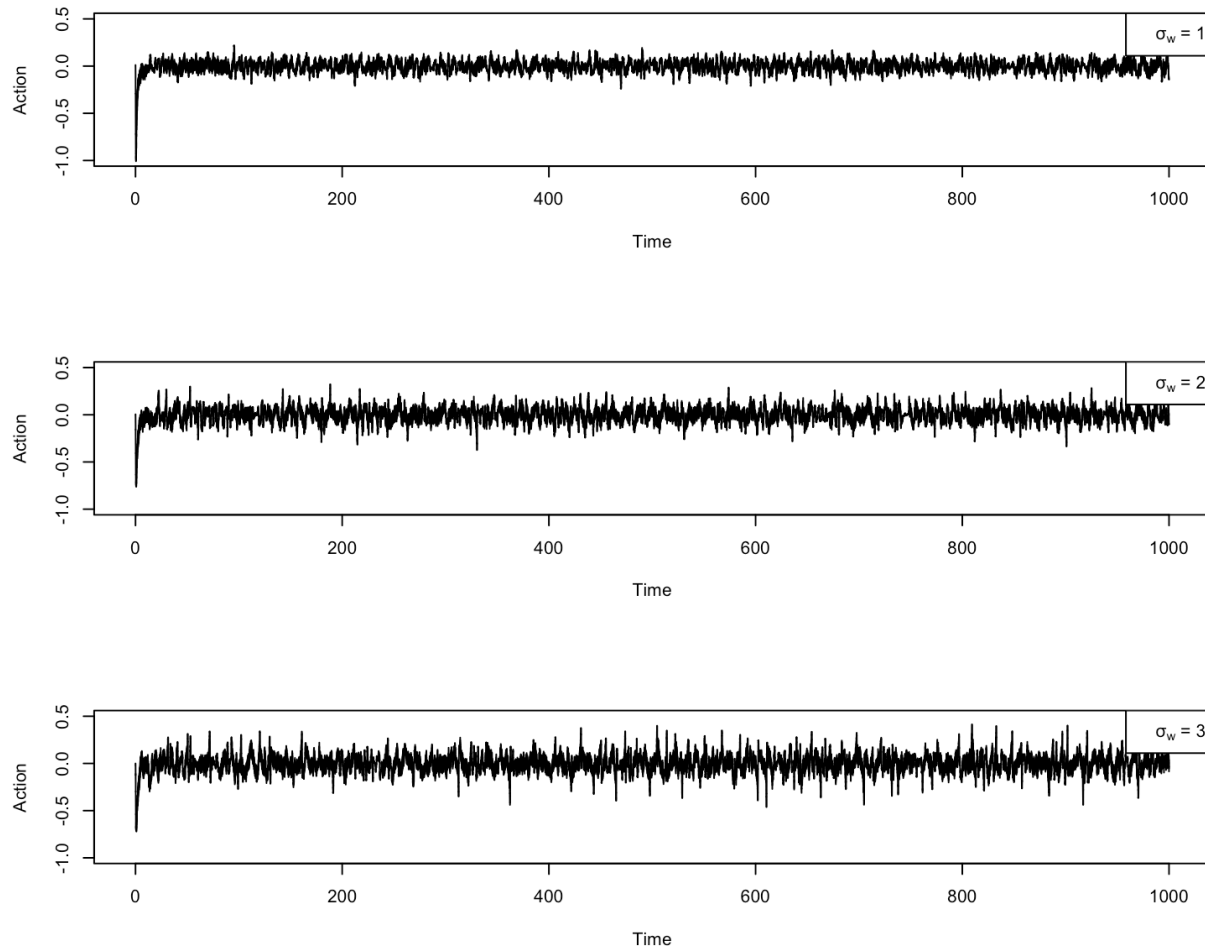




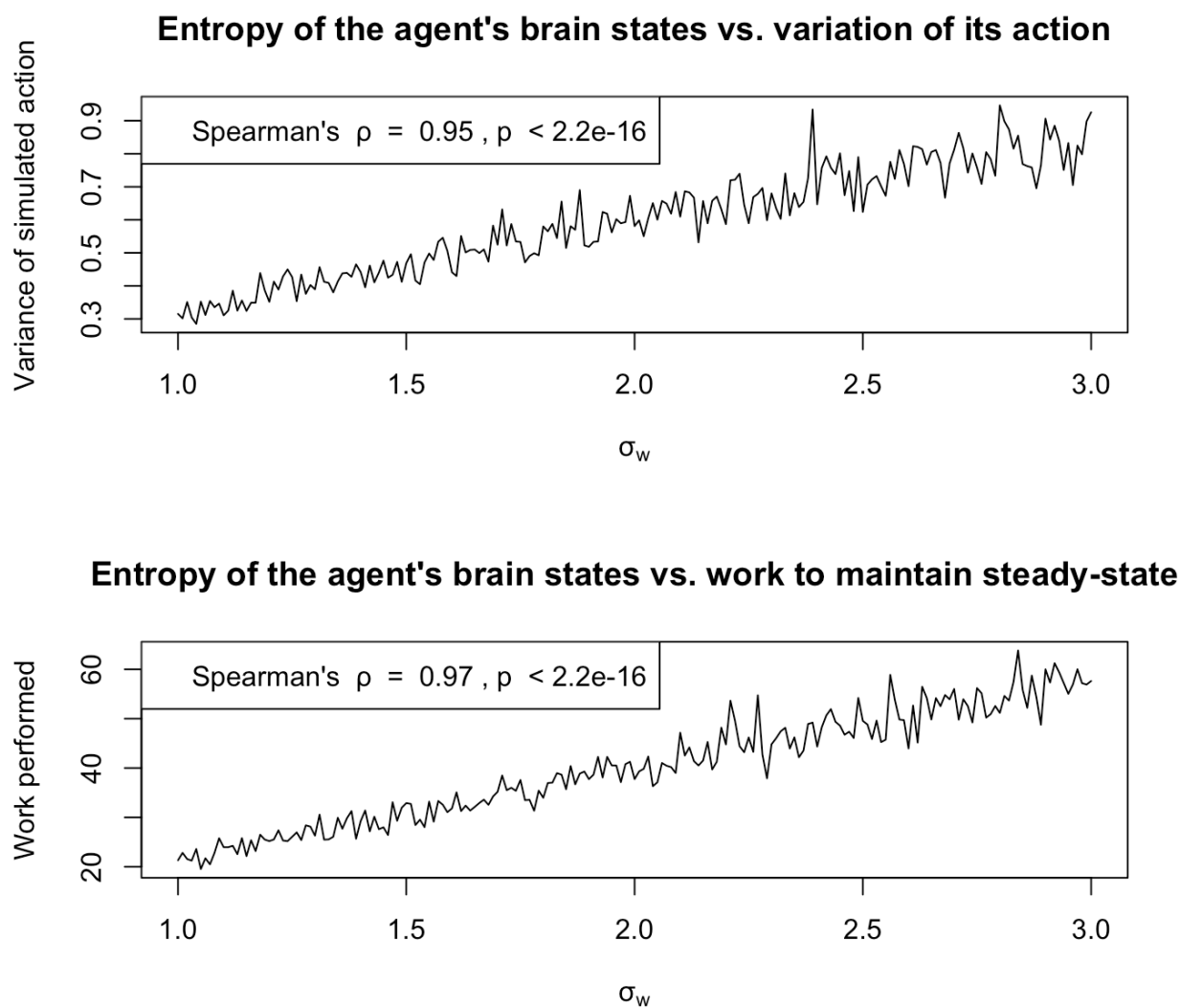
**Figure 2. Younger age-at-onset is associated with increased within-individual clinical variability in Huntington disease.** The violin plots depict the association of residual age-at-onset (i.e., whether the observed age-at-onset was below or above expected age of onset based on mutant *HTT* CAG repeat size) and the standardized variance independent of the mean (VIM) metric. The boxes indicate the interquartile ranges around the median (thick black lines), with values deviating more than 1.5 times the interquartile range from the median represented by black dots. After regressing out the effects of sex, age at baseline, mutant CAG repeat size and its interaction with age at baseline, and domain-specific disease severity at baseline from the VIM scores, the residuals were compared across the two categories of residual age-at-onset using the non-parametric Wilcoxon's rank sum test.



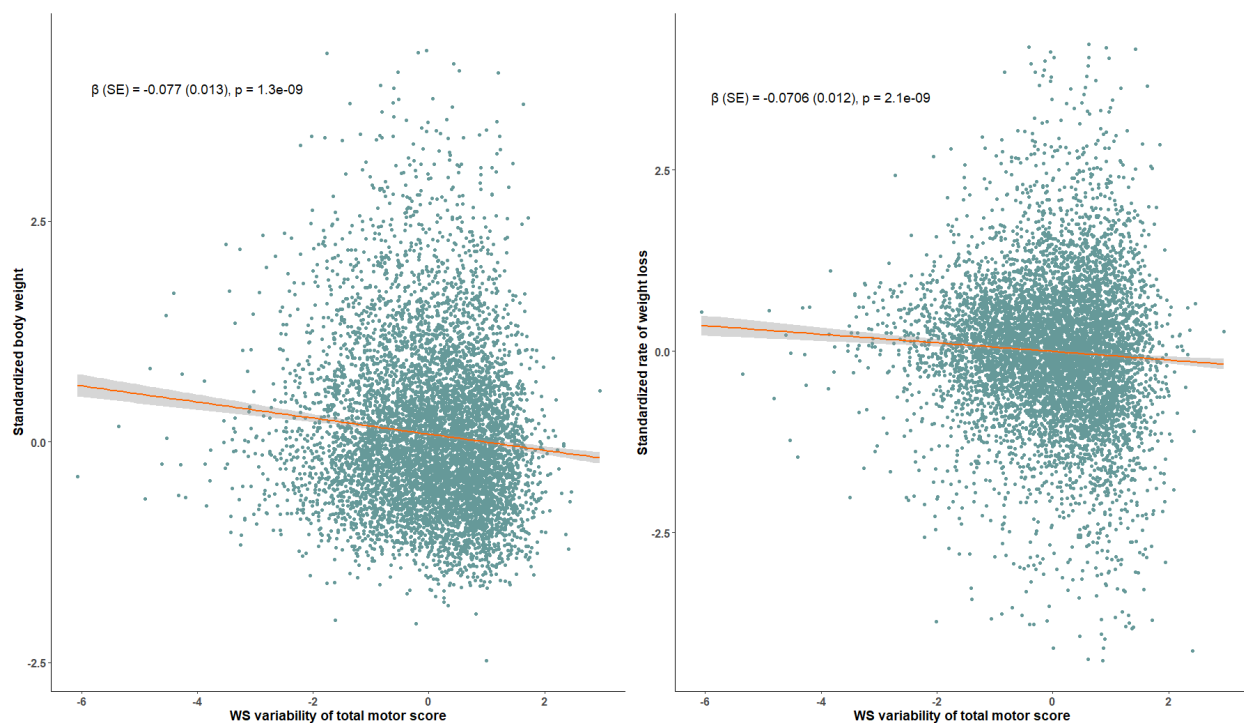
**Figure 3. Phenotypic variability increases with higher values of predictive entropy.** The figure depicts the simulated action (i.e., movements around a predefined set-point) of a simple Bayesian agent for different values of the entropy of the distribution that underlies the agent's beliefs about its environment as parametrized through  $\sigma_w$ . Variability of action is larger for higher values of  $\sigma_w$ . See main text for more details.



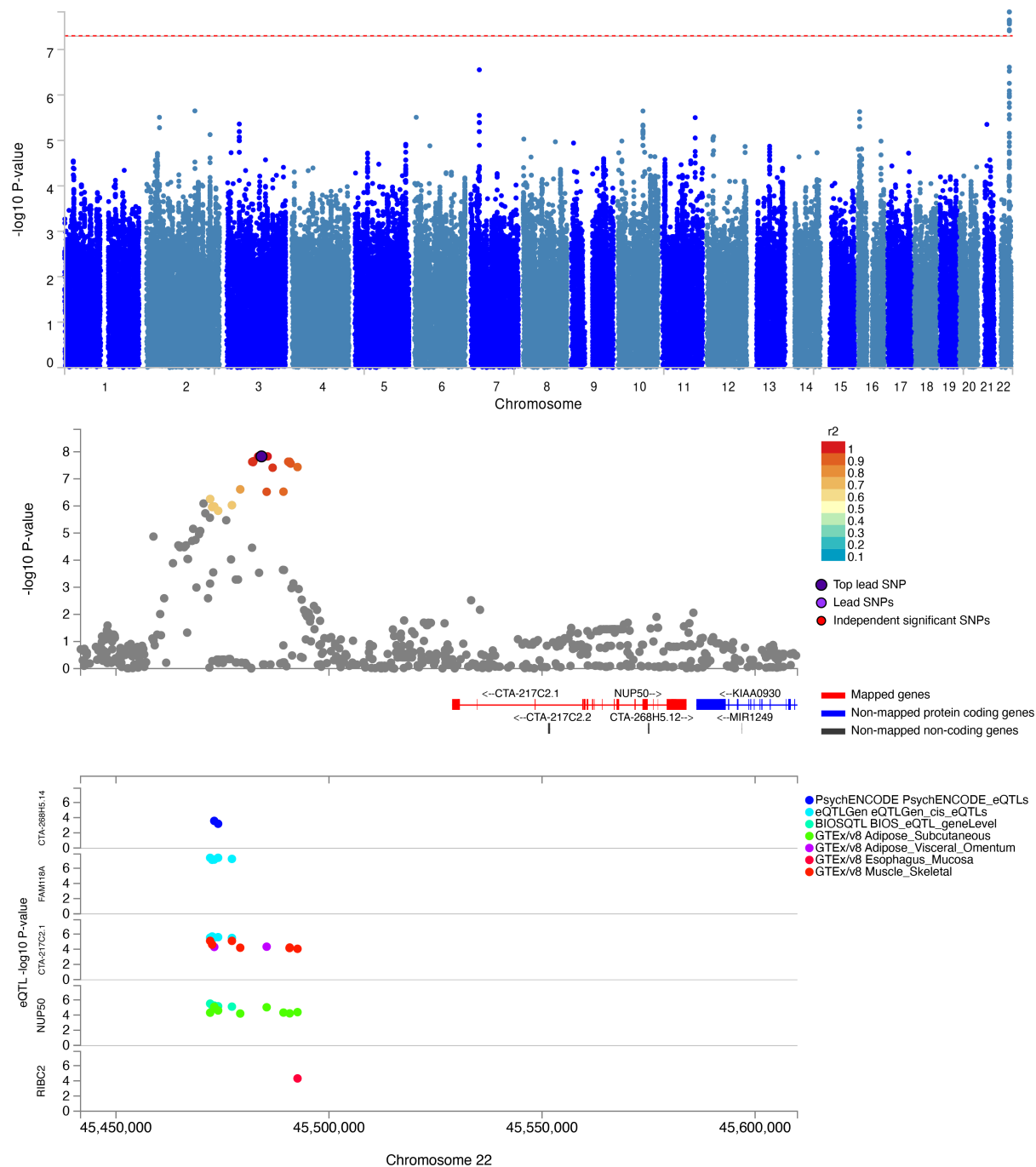
**Figure 4. Increased phenotypic variability results in higher energy expenditure.** The agent's phenotypic variability increased with higher values of the entropy of the distribution that underlies the agent's beliefs about its environment as parametrized through  $\sigma_w$  (top panel). The amount of work performed by the agent required to maintain steady-state – defined as the cumulative sum of the product of acceleration and distance traveled over time – increased with higher values of  $\sigma_w$  (bottom panel).



**Figure 5. Within-individual variability of motor impairment is associated with lower body weight and a faster rate of weight loss.** Average body weight (estimated through a random intercept linear mixed-effects model with adjustment for age, age<sup>2</sup>, sex, mutant CAG repeat size, as well as its interaction with age, and total motor score) decreased with higher within-individual variability of total motor score (left panel). Similarly, the rate of body weight loss (represented by the random slope coefficients obtained from a random intercept and slope linear mixed-effects model with age, age<sup>2</sup>, sex, mutant CAG repeat size, as well as its interaction with age, and total motor score, as fixed effects, and an intercept and age as random effects) also increased with higher variability of total motor score (right panel). The insets indicate the standardized regression estimates (standard error (SE)). The red lines represent the regression lines, with the corresponding grey areas indicating the SEs around the mean. WS = within-subject.

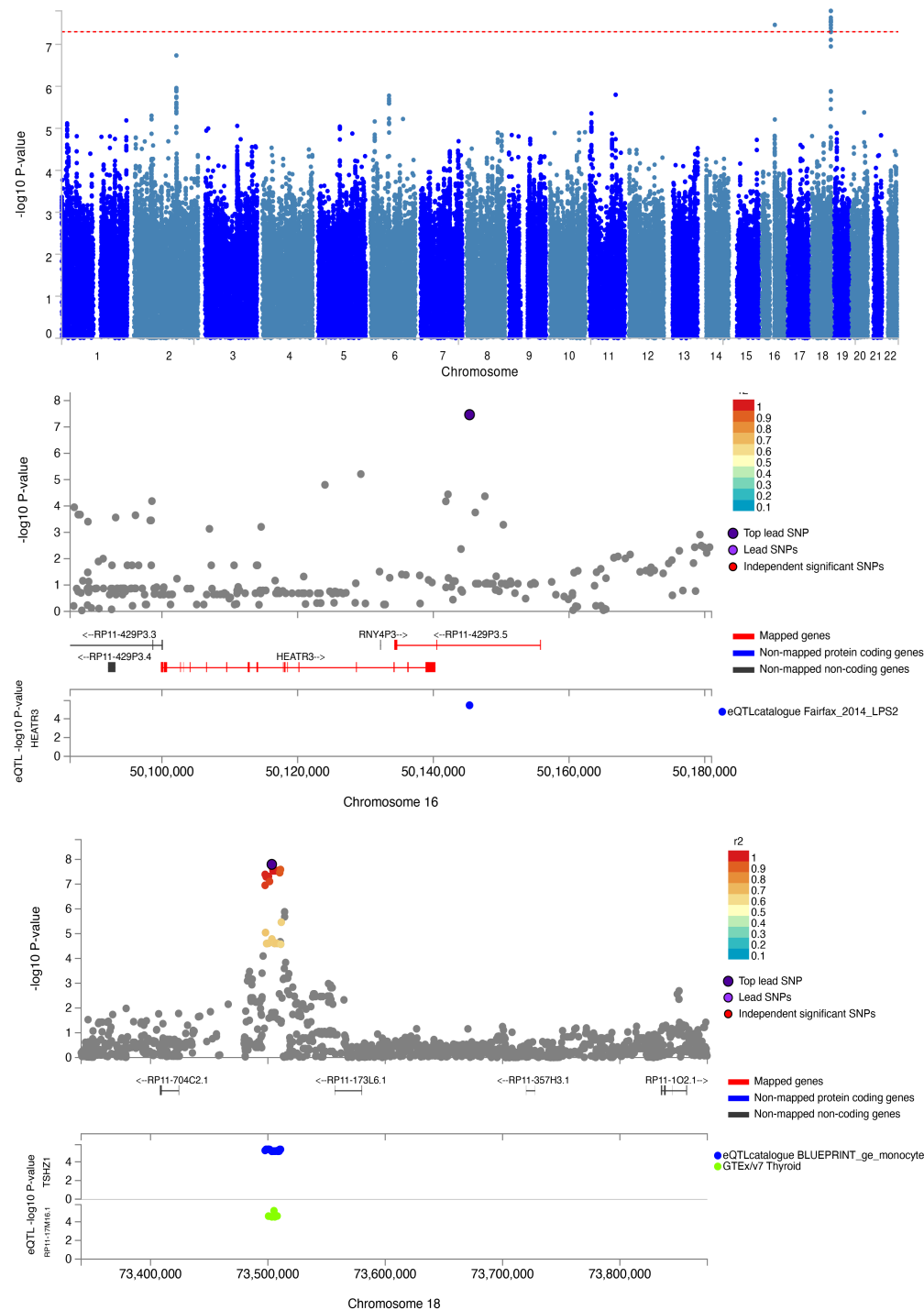


**Figure 6. Genome-wide association study of body mass index variability in Huntington disease.** The dashed red line in the Manhattan plot (top plot) indicates the threshold for genome-wide significance (i.e.,  $p = 5.0E-8$ ). An intergenic locus on chromosome 22, tagged by the lead SNP rs62231080, reached genome-wide significance (standardized  $\beta = -0.260 \pm 4.57E-2$ ,  $p = 1.48E-8$ ). Positional (middle panel) and eQTL (bottom panel) analyses mapped this locus to the genes *NUP50*, *FAM18A*, and *RIBC2*, and the lincRNA gene *CTA-217C2.1* and *CTA-268H5.14*.

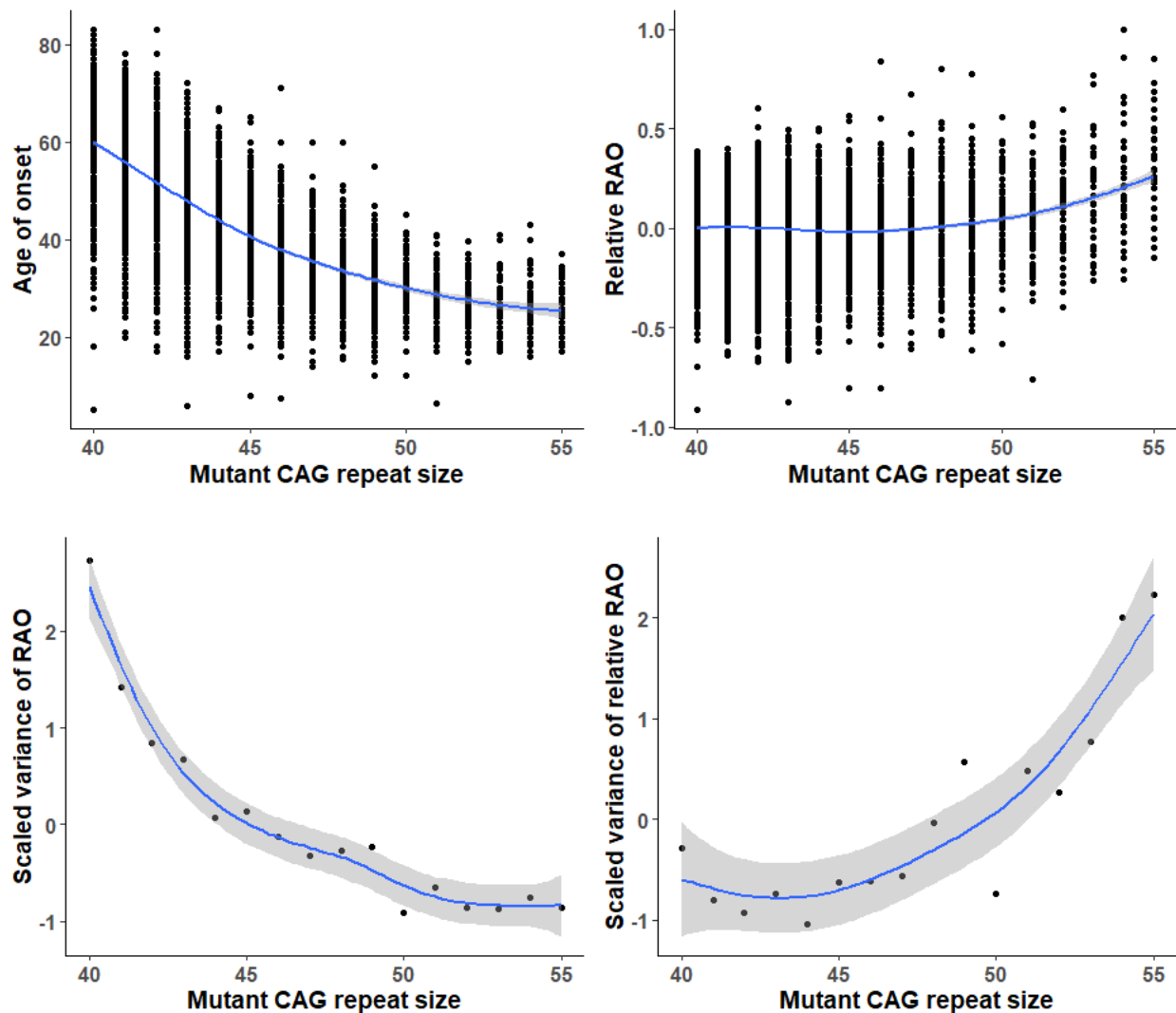


### Figure 7. Genome-wide association study of mean body mass index in Huntington disease.

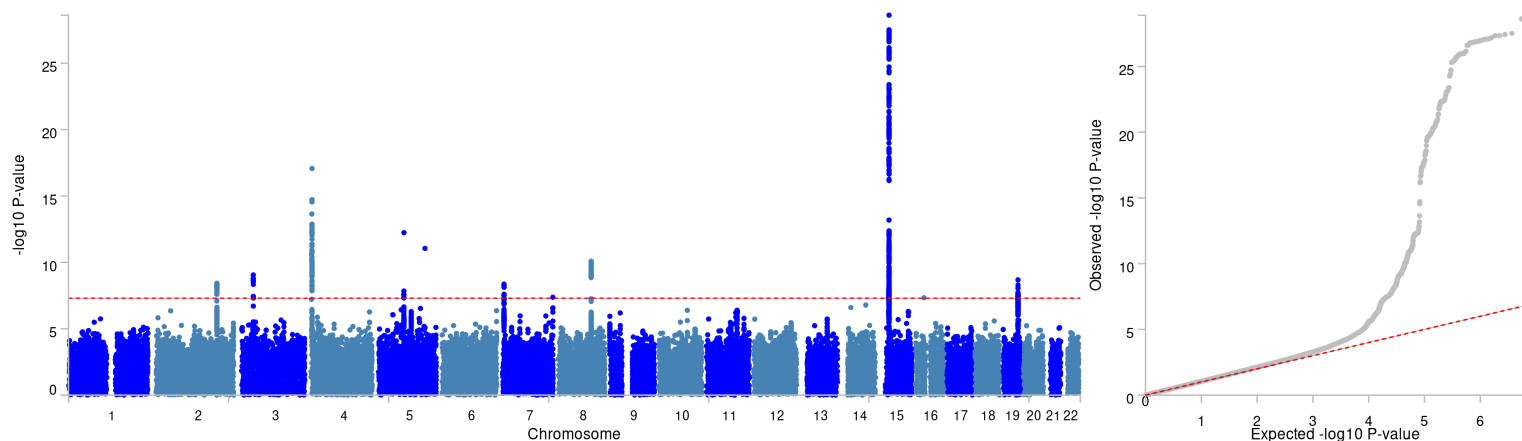
The dashed red line in the Manhattan plot (top plot) indicates the threshold for genome-wide significance (i.e.,  $p = 5.0E-8$ ). Two loci on chromosomes 16 and 18, tagged by the lead SNPs rs200717776 and rs11874212, respectively, reached genome-wide significance. Positional and eQTL analyses (lower panels) mapped the chromosome 16 locus to the protein coding *HEATR3* gene, and the antisense lncRNA *RP11-429P3.5*, and the chromosome 18 locus to the protein coding *TSHZ1* and *RP11-17M16.1* genes.



**Figure 8. Age of onset variability in Huntington disease.** Although both absolute age of onset and residual age-at-onset (RAO) variability *decrease* with higher mutant CAG repeat size (left panels), the relative RAO, as well as its variability *increase* with larger mutation sizes (right panels). The blue lines represent smoothing splines and associated standard errors around the mean indicated in grey. See main text for more details.

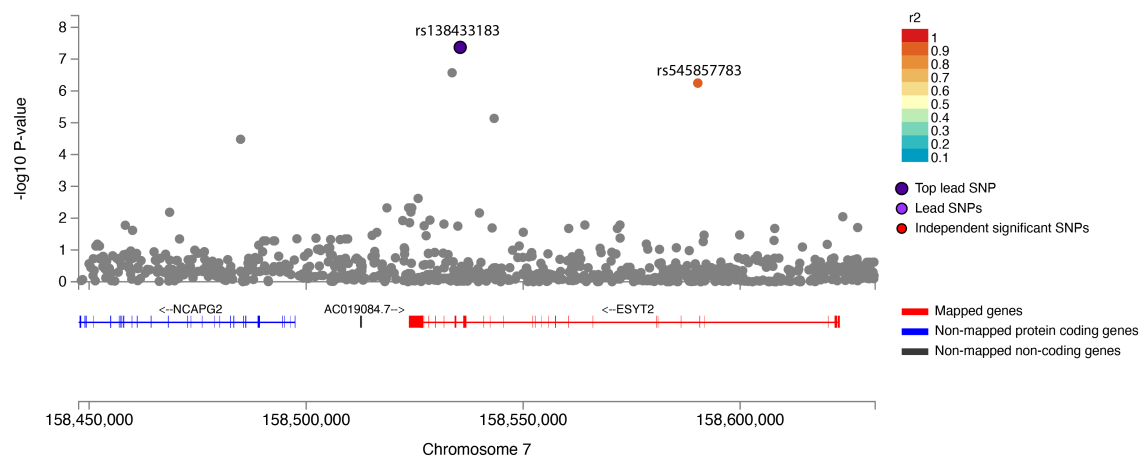


**Figure 9. Genome-wide association study of *relative* residual age-at-onset in Huntington disease.** The Manhattan (left) and corresponding quantile-quantile plot (right) are depicted. The dashed red line in the Manhattan plot indicates the threshold for genome-wide significance (i.e.,  $p = 5E-8$ ). See main text and **Table 4** for more details about the identified loci.



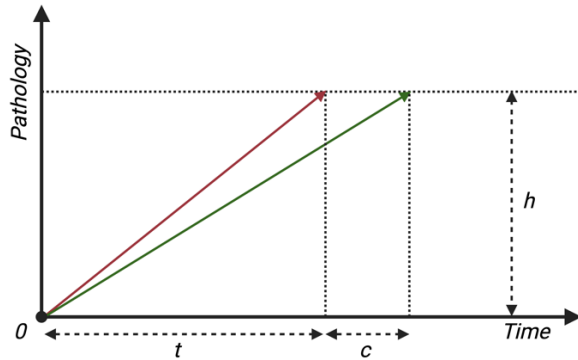


**Figure 10. Genome-wide association study of *relative* residual age-at-onset in Huntington disease identifies a novel locus.** The figure depicts the genomic location of novel rare variants on chromosome 7 within the *ESYT2* gene, tagged by the lead (rs138433183) and an independently significant (rs545857783) SNP, associated with *residual* age-at-onset in Huntington disease.



## Appendix: Comparison between absolute and relative residual age-at-onset

Let  $h$  represent the extent of pathology (in arbitrary units) required for the disease to become clinically manifest, and let  $t$  represent the expected age at clinical onset (in years), assuming a linear increase of pathological burden with time since birth (**Figure A-1**):



The average rate of *subclinical* disease progression  $r$  per year could then be defined as:

$$r(t) = \frac{h}{t} \quad \text{Eq. A-1}$$

Now suppose that a potential disease modifying genetic variant (or drug) would increase the expected age at onset by  $c$  years, i.e.:

$$r_0 = r(t) = \frac{h}{t} \quad \text{and} \quad r_c = r(t+c) = \frac{h}{t+c} \quad \text{Eq. A-2}$$

This would have to result in a corresponding decrease in  $r$  of:

$$\frac{r_c}{r_0} = \left(1 + \frac{c}{t}\right)^{-1} \quad \text{Eq. A-3}$$

Thus, *Eq. A-3* demonstrates that for smaller values of  $t$ , the *relative* decrease in the rate of disease progression needs to be greater to ensure a delay in clinical onset by a certain fixed number of years  $c$  as compared to larger values of  $t$ , and vice versa. For example, to delay the expected age of onset by 10 years in two premanifest mutation carriers with expected ages of onset of 30 and 40 years, the rate of disease progression would need to be lowered by 25% and 20%, respectively. In contrast, suppose that a potential disease modifying genetic variant (or drug) would change the expected age at onset by a factor  $f$ , then in the presence of this modifier the rate of disease progression would be given by:

$$r_f = r(ft) = \frac{h}{ft} \quad \text{Eq. A-4}$$

Which will result in a similar *relative* change in the rate of disease progression for different values of expected age at onset (and by extension, for different mutant CAG repeat sizes):

$$\frac{r_f}{r_0} = \frac{1}{f} \quad \text{Eq. A-5}$$

Please note that for the above derivation, it was assumed that pathological burden will increase linearly over time. For completeness, here follows a derivation assuming an exponential increase in pathological burden over time:

Suppose that pathological burden  $p$  will increase exponentially over time:

$$p(t) = \alpha e^{\beta t} + \gamma \quad \text{Eq. A-6}$$

For simplicity, and without loss of generalizability, let's assume  $\alpha = 1$ , and  $\gamma = 0$ . Now, suppose that a potential disease modifying genetic variant (or drug) would increase the expected age at onset by  $c$  years, resulting in a change of the growth parameter  $\beta$ , i.e.:

$$p_0(t) = e^{\beta_0 t} \text{ and } p_c(t) = e^{\beta_c t} \quad \text{Eq. A-7}$$

From:

$$h = p_0(t) = p_c(t + c) \quad \text{Eq. A-8}$$

It then follows that:

$$e^{\beta_0 t} = e^{\beta_c(t+c)} \Rightarrow \beta_c = \frac{\beta_0}{t+c} \quad \text{Eq. A-9}$$

Therefore:

$$p_c(t) = e^{\beta_0 \left(1 + \frac{c}{t}\right)^{-1}} = p_0(t)^{\frac{1}{t+c}} \Leftrightarrow \ln p_c(t) = \frac{1}{t+c} \ln p_0(t) \quad \text{Eq. A-10}$$

Given that the initial growth parameter  $\beta_0$  will decrease to a greater extent for smaller values of  $t$ , Eq. A-10 thus demonstrates that also under the assumption of an exponential increase of pathology over time, for smaller values of  $t$ , the *relative* decrease in the rate of disease progression needs to be greater to ensure a delay in clinical onset by a certain fixed number of years  $c$  as compared to larger values of  $t$ , and vice versa. On the other hand, assuming that a potential disease modifying genetic variant (or drug) would change the expected age at onset by a factor  $f$ , then in the presence of this modifier, for the rate of disease progression the following holds:

$$h = p_0(t) = p_f(ft) \quad \text{Eq. A-11}$$

From which it follows that:

$$e^{\beta_0 t} = e^{\beta_f f t} \Rightarrow \beta_f = \frac{\beta_0}{f} \quad \text{Eq. A-12}$$

Therefore:

$$p_f(t) = e^{\frac{\beta_0}{f} t} = p_0(t)^{\frac{1}{f}} \Leftrightarrow \ln p_f(t) = \frac{1}{f} \ln p_0(t) \quad \text{Eq. A-13}$$

Eq. A-13 thus shows that when assuming that a modifier will result in a *relative* change of  $t$ , the growth parameter will simply be scaled to a proportional degree, independent of  $t$ . This latter is not only biologically more plausible, but also ensures that the effect of the modifier would be easier to compare across individuals with different expected ages of onset (and by implication, also with different mutant CAG repeat sizes).