

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/86871/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Keum, Jae Whan, Shin, Aram, Gillis, Tammy, Mysore, Jayalakshmi Srinidhi, Elneel, Kawther Abu, Lucente, Diane, Hadzi, Tiffany, Holmans, Peter, Jones, Lesley, Orth, Michael, Kwak, Seung, MacDonald, Marcy, Gusella, James F. and Lee, Jong-Min 2016. The HTT CAG expansion mutation determines age at death but not disease duration in Huntington's Disease. *American Journal of Human Genetics* 98 (2), pp. 287-298. 10.1016/j.ajhg.2015.12.018

Publishers page: <https://doi.org/10.1016/j.ajhg.2015.12.018>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



## **The *HTT* CAG Expansion Mutation Determines Age at Death but not Disease Duration in Huntington's Disease**

Jae Whan Keum,<sup>1</sup> Aram Shin,<sup>1</sup> Tammy Gillis,<sup>1</sup> Jayalakshmi Srinidhi Mysore,<sup>1</sup> Kawther Abu Elneel,<sup>1</sup> Diane Lucente,<sup>1</sup> Tiffany Hadzi,<sup>2</sup> Peter Holmans,<sup>3,9</sup> Lesley Jones,<sup>3,9</sup> Michael Orth,<sup>4,9</sup> Seung Kwak,<sup>5,9</sup> Marcy E. MacDonald,<sup>1,6,7,9</sup> James F. Gusella,<sup>1,7,8,9</sup> and Jong-Min Lee<sup>1,6,7,9,\*</sup>

<sup>1</sup> Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA

<sup>2</sup> GNS Healthcare, Inc., One Charles Park, Cambridge, MA 02142, USA

<sup>3</sup> Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff, United Kingdom

<sup>4</sup> Department of Neurology, University of Ulm, Germany

<sup>5</sup> CHDI Foundation, Princeton, NJ 08540, USA

<sup>6</sup> Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

<sup>7</sup> Medical and Population Genetics Program, the Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142, USA

<sup>8</sup> Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

<sup>9</sup> Founding GeM-HD Consortium investigators

\* Corresponding author: [jlee51@mgh.harvard.edu](mailto:jlee51@mgh.harvard.edu)

**Abstract**

Huntington's disease (HD) is caused by an expanded *HTT* CAG repeat that leads in a length-dependent, completely dominant manner to onset of a characteristic movement disorder. HD also displays early mortality, so we tested whether the expanded CAG repeat exerts a dominant influence on age at death and on the duration of clinical disease. We found that, as with clinical onset, HD age at death is determined by the expanded CAG repeat length with no contribution from the normal CAG allele. Surprisingly, disease duration is independent of the mutation's length. It is also unaffected by a strong genetic modifier of HD motor onset. These findings suggest two parsimonious alternatives: 1) HD pathogenesis is driven by mutant huntingtin but, before or near motor onset, sufficient CAG-driven damage has occurred to permit CAG-independent processes to then lead to eventual death. In this scenario, some pathological changes and their clinical correlates may still worsen in a CAG-driven manner after disease onset but these CAG-related progressive changes do not themselves determine duration. Alternatively, 2) HD pathogenesis is driven by mutant huntingtin acting in a CAG-dependent manner with different time courses in multiple cell-types, and the cellular targets that lead to motor onset and to death are different and independent. In this scenario, *HTT* CAG length-driven processes lead directly to death but not via the striatal pathology associated with motor manifestations. Each scenario has important ramifications for the design and testing of potential therapeutics, especially those aimed at preventing or delaying characteristic motor manifestations.

## Introduction

Huntington's disease (HD) is a dominantly inherited disorder (OMIM #143100)<sup>1; 2</sup> whose characteristic neurological symptoms result from an expanded CAG repeat of greater than 35 units in the huntingtin-encoding sequence of *HTT* (OMIM # 613004).<sup>2-5</sup> It has long been known that the age of onset of HD motor symptoms is negatively correlated with expanded *HTT* CAG repeat length,<sup>2; 6-9</sup> but the question of whether the length of the normal CAG allele also plays a role in determining onset remained uncertain until recently. Two publications reported that a complex interaction between the CAG repeat lengths of the normal and expanded alleles played a significant role in determining an individual's age at motor onset.<sup>10; 11</sup> However, a more comprehensive study recently uncovered issues with the statistical analysis of age of onset vs. CAG correlation, developed a route to robust statistical analysis of such data, and concluded that there is no significant impact of the normal allele CAG length, either alone or in interaction with the expanded CAG repeat, on age at motor onset.<sup>12</sup> Consistent with this conclusion, age at motor onset of HD subjects with two expanded *HTT* alleles is determined by the longer of their two CAG repeats and is similar to subjects with a single mutant allele of comparable size.<sup>12</sup> Thus, the expanded *HTT* CAG repeat triggers HD pathogenesis (defined here as the underlying abnormal process that leads to pathology, defined in turn as disease phenotype) and the length of the mutation largely determines the rate of the pathogenic process that leads to clinical motor signs in a fully dominant fashion. The motor disturbances are thought to result from dysfunction and degeneration of neurons in the striatum which, based upon pathological examination of brains post-mortem, is also correlated with CAG repeat length.<sup>13-16</sup> The relationship between CAG repeat length, age at motor onset, and inferred rate of striatal pathology has informed the potential for treatment by *HTT* gene-silencing, since the lack of an effect of either the normal allele or of an interaction between normal and expanded alleles implies that one can target the single mutant copy of *HTT* in the striatum to delay onset and worsening of motor symptoms.

Another important feature of HD that has been reported to show a strong correlation with the expanded CAG repeat length is age at death.<sup>6; 9; 17</sup> HD shows early mortality compared to the general population,<sup>18</sup> but it is not clear whether this reflects a direct or an indirect effect of mutant huntingtin on vital processes. In

principle, early mortality in HD could be due to neurodegenerative changes in the regions most associated with HD clinical symptoms or, alternatively, to some other independent effect of mutant huntingtin.<sup>16; 19-25</sup> In order to inform potential treatments for preventing or delaying disease onset and early death of HD subjects, we have performed a comprehensive statistical assessment of the impact of the expanded and normal CAG repeats on determining the timing of death and the relationship of this timing with the prior onset of diagnostic motor signs, an interval that we define as 'disease duration'. Our findings support a dominant effect for the expanded CAG repeat on age at death, like that on age at motor onset, with no impact of the normal CAG allele. However, they also indicate a surprising lack of contribution of either the mutant or normal *HTT* allele to determining the duration of clinical disease from onset to death. We propose two alternative explanations for the relationship between motor onset and death that have quite different implications for designing therapies in HD.

## Material and Methods

### Study subjects

We analyzed DNA samples from 4,448 HD heterozygous subjects with either known age of onset of motor signs and/or age at death: 4,161 subjects had age at onset data, 1,165 subjects had age at death data, and 878 subjects had both. Subjects with recorded age at death were primarily identified from HD brain banks: the Harvard Brain Tissue Resource Center (McLean Hospital, Belmont MA), the New York Brain Bank (Columbia University, New York City, NY), the National Neurological Research Specimen Bank (Department of Veterans Affairs Medical Center, Los Angeles CA), and the Harvard NeuroDiscovery Center Advanced Tissue Resources Core (Massachusetts General Hospital, Charlestown MA). HD subjects with recorded age at onset were described previously.<sup>12</sup> The *HTT* CAG repeat lengths for each DNA sample were determined by a polymerase chain reaction assay, against sequenced allele size standards as described previously.<sup>26</sup> Primary analysis was based on HD subjects who carried one expanded *HTT* CAG repeat (CAG > 35). The means of expanded and normal CAG repeat lengths of our study subjects were 45.1 (range, 36-120; median, 44) and 18.45 (range, 9-35; median, 18), respectively. Age at death, age at onset of motor signs, and disease duration (the difference between age at onset and age at death) were the primary dependent variables in the statistical models. Motor onset ages were from a clinician rater estimate, family member report, and/or self-report. When multiple onset ages were available, priority was given first to the expert rater estimate and then to the family member report. In addition to ascertained subjects for onset and/or death, we also independently analyzed European Huntington's Disease Network (EHDN) Registry observational cohort data to construct survival models. CAG repeats of EHDN samples were determined by the same method as above. The familial relationships of our samples were not known due to the lack of pedigree information in our de-identified data. Considering the tendency of related individuals to cause statistical inflation, the lack of significant influence in our results would not change dramatically if familial relationship could be included in the model. The study was approved by the Partners HealthCare Institutional Review Board.

### Statistical analysis of age at death

We previously described a robust statistical analysis method for age at onset of motor signs,<sup>12</sup> and have used it in this study for the analysis of age at death. Natural log transformed age at death data from 1,165 subjects were modeled as a function of 1) expanded CAG length, 2) normal CAG repeat length, and 3) gender (Table 1, Model 3). In addition, quality control analyses were performed to identify a subset of data that met the requirements of linear regression modeling, including 1) normality, 2) equal variance, and 3) absence of disproportionately influential observations. First, we evaluated data normality for each CAG repeat length and found that the distribution of age at death approximates a normal distribution for subjects with 40-52 CAG repeats, representing 90.3% of the samples (Figure S1). Variance in age at death was not constant but rather was larger for subjects with smaller expanded CAG repeats potentially implicating a greater role for modifiers in this range (Figure S2A). We addressed the non-constant variance by log transformation of the data as shown in Figure S2B, which also rendered the relationship between expanded CAG repeat length and age at death close to linear (Figure S2C). Lastly, we identified outliers to exclude using a standard inter-quartile outlier detection method (Figure S2C). Briefly, for each CAG repeat size, age at death data were sorted to obtain an inter-quartile value (i.e., the difference between 75 and 25 percentile data points). Then, 1.5 times of the inter-quartile was added to the 75 percentile data point and subtracted from the 25 percentile data points. Any age at death values outside of this range were identified as potential outliers. These procedures resulted in a subset of 1,019 subjects (87.5% of all those with age at death data) for a robust analysis of models that included gender along with both expanded and normal CAG length (Model 1) and only expanded CAG length (Model 2) as summarized in Table 1. In addition, Model 3 was fitted to all data without any exclusion of samples to evaluate the effect of outlier removal. The model behavior was evaluated by checking 1) residuals vs. fitted values, 2) quantile-quantile plots, and 3) residuals vs. leverage (Figure S3).

We also performed three additional tests to establish the robustness of the conclusion from Table 1. 1) Using the minimal adequate model based on only QC-passed data (Table 1, Model 2), we calculated the residual of age at death for all subjects with 40-52 CAGs including outlier subjects (total 1,052 subjects) and subsequently used it as a dependent variable to test the effect of outliers previously excluded via our quality control (QC) pipeline (Figure S4). 2) We compared the normal CAG repeat length distributions of HD

subjects who represent the top or bottom 10% of residual age at death values (105 individuals each) (Figure S5). 3) We determined the effects of samples excluded from the model due to repeats longer than 52 CAGs by constructing an independent model using only the 97 HD subjects excluded from the analysis (Figure S6). None of these additional tests supported any effect of the normal CAG repeat length on age at death.

### **Statistical analysis of disease duration**

Eight hundred and seventy eight HD subjects were fully ascertained for both age at motor onset and age at death. For each subject, disease duration was calculated by subtracting age at onset from their age at death. For these fully ascertained subjects, the range, mean, standard deviation, and median of duration values are 0-46, 15.4, 6.8, and 15 years, respectively. Approximately 72.9% of the HD subjects have duration values within one standard deviation around the mean (8.6-22.2 years), forming a non-normal distribution with a sharp peak and long-tails indicating positive excess kurtosis and prompting our non-parametric statistical analysis. Since duration values of 10-20 years were highly enriched in the data regardless of the length of expanded CAG repeats, quality control analyses that were applied to age at death data and age at onset data could not identify a subset of data points that were normally distributed (Figure S7-S8). Therefore, instead of using parametric regression models, we performed non-parametric modeling (i.e., generalized additive model and Spearman's rank correlation analysis) to test the significance of CAG repeat lengths and gender on duration. In addition, we compared disease duration of HD subjects with expanded CAG < 43 to that of HD subjects with expanded CAG > 45 to test whether HD subjects with longer expanded CAG repeats display a shorter duration compared to HD subjects with shorter expanded CAG repeats. In a conceptually similar manner, expanded CAG repeat lengths of HD subjects with the top 10% duration values were compared to those of HD subjects with the bottom 10% of duration values to test whether extreme duration subjects vary by expanded CAG repeat length. Finally, disease duration of adult onset HD subjects (age at onset > 20) was compared to juvenile onset HD subjects (age at onset < 21) (Mann-Whitney U test).

### **Survival analysis of duration for an observational study cohort and for fully ascertained samples**



In addition to non-parametric analyses of duration using HD subjects fully ascertained for both onset and death, survival analysis was performed using data from the European Huntington's Disease Network (EHDN) Registry observational cohort. In this data set, 1,314 HD subjects had only recorded onset age, and 115 HD subjects had both onset age and age at death data. A non-parametric survival analysis using a Cox Proportional Hazards model was performed by setting onset age as time 0 and duration as time to event. For subjects without recorded age at death, age at onset subtracted from age at last study visit was used as time to event with a censoring indicator. Duration values were modeled as a function of CAG (continuous variable) and gender using the "coxph" function in the "survival" R package. A similar survival analysis approach (but without censoring) using CAG repeat size as a continuous predictor variable and gender was applied to the data for 855 fully ascertained HD subjects that were used for our primary non-parametric analysis. In addition, based on sorted CAG repeat sizes of fully ascertained samples, two groups of HD subjects with the top or bottom 10% extreme CAG repeats (85 subjects in each group, 50-63 CAGs and 38-41 CAGs, respectively) were identified to be compared for duration through survival analysis.

### **Simulation analysis**

The capacity for CAG length to explain duration was not significant in the observed data set (Figure 3). To estimate whether different levels of explanatory power of the CAG repeat for duration could have been visualized in such data, we first randomly permuted the duration values of the 855 HD subjects. Then, without data replacement, we sampled duration values randomly until the model's R square value (model: duration ~ CAG) reached either 20, 10, 5, 2, or 1%. Once the model based on permuted data generated the pre-specified R square value, the mean of duration for each CAG length based on that simulation set was recorded for plotting against CAG.

### **Software package for statistical analysis**

All statistical analyses were performed using R (version, 3.0.2).

## Results

### **Age at death in HD is strongly correlated with expanded *HTT* CAG repeat length**

We initially compared the relationships between 1) expanded CAG repeat length and age at onset of motor signs (age at onset, hereafter), 2) expanded CAG repeat length and age at death, and 3) age at onset and age at death. Consistent with previous findings,<sup>6; 9; 12</sup> analysis using all data points showed that age at motor onset (Figure 1A) and age at death (Figure 1B) are both inversely correlated with the expanded CAG repeat length (adjusted R-squared, 65.4% and 74.5%, respectively). However, there is substantial variance in age at onset and age at death that is not explained by the expanded CAG repeat, suggesting that these outcomes are modified by other factors. In addition, age at death is strongly correlated with age at onset (Figure 1C), indicating that the former can be predicted relatively accurately from the latter (adjusted R-squared, 77.5%). Interestingly, the correlation of expanded CAG length with age at death appears stronger, as age at onset is more variable. This may be because age at onset involves a subjective assessment on the part of an expert clinician whereas age at death is objectively recorded. It may also be that age at onset is more easily modified by other genetic or environmental factors than is age at death. In any event, the strong correlation of age at death with the expanded CAG length indicates that, averaged across the HD population, the length of the mutation is the primary factor determining at what age an individual with HD dies.

### **Age at death is not influenced by the length of the normal *HTT* CAG repeat**

Next, we tested whether variance in age at death of HD is influenced by the normal allele CAG repeat or gender. The expanded CAG repeats in 1,165 HD subjects ranged from 36 to 120 CAGs; the median of the expanded and normal CAG repeat lengths was 44 and 18, respectively. We have reported previously on the importance of conforming to data quality assumptions in parametric statistical analysis of CAG repeat relationships to avoid spurious results when testing for potentially subtle effects of modifiers and have proposed a QC pipeline for robust statistical analysis.<sup>12</sup> Among other factors, important requirements to

confirm when performing parametric regression analysis include 1) data normality, 2) equal variance, and 3) absence of disproportionately influential data points. We performed QC analyses, as described in Lee et al.<sup>12</sup> and in the Material and Methods (Figures S1-S2), to generate a data set of 1,019 subjects with 40-52 CAG repeats that conforms to the assumptions of regression analysis (Figure S3). This data set was used to generate reliable statistical models, Model 1 and Model 2, which include or exclude the normal CAG repeat, respectively (Table 1). Only the expanded CAG repeat length and gender explained a significant portion of the variance in age at death (Model 1; adjusted R-squared, 66.9%) since omitting the normal CAG repeat length from the model (Model 2; R-squared, 66.9%) made no significant difference (ANOVA model comparison, p-value, 0.1348). Exclusion of outliers did not change the conclusion since Model 3, fitted to all data without any sample exclusion, also revealed significance only for expanded CAG and gender (Table 1; Figure S3). Similarly, tests involving inclusion of the outliers excluded by the QC pipeline or specific examination of individuals with > 52 CAGs failed to reveal any significant impact of the normal CAG repeat length (Figures S4-S6). Thus, the size of the expanded CAG repeat and gender have predictive power for age at death, but the normal CAG length has no discernible impact.

### **Relationship between age at onset and age at death: disease duration**

The construction of statistical models relating the expanded CAG repeat length to both age at onset<sup>12</sup> and age at death (Table 1, Model 2) allowed us to estimate indirectly the average time between the initial presentation of diagnostic motor signs and death in male and female HD subjects. Age at onset was not different between the two genders (Figure 2A, dashed lines), but there was a significant ~1 year difference between males and females in age at death (Figure 2A, red and blue solid lines). Interestingly, the age at death model (solid) and the age at onset model (dashed) were largely parallel for CAG ranges associated with onset in adulthood, indicating that the disease duration from diagnosis to death is similar regardless of expanded CAG repeat size. This implies that gender-associated disease duration is independent of the length of the expanded CAG repeat, which we have suggested previously from a limited post-mortem brain study.<sup>5;9</sup>

We next used several approaches to test this proposition in the 878 HD subjects where disease duration could be measured directly because both age at motor onset and age at death were ascertained, a much larger dataset than analyzed previously. The median of the expanded and normal alleles was 44 and 18 CAGs, respectively. As suggested from the statistical models, although the duration values were highly variable for any given CAG repeat length, the median of measured disease duration for these HD subjects appeared similar regardless of expanded CAG repeat length (Figure 2B). It was not possible to construct a reliable parametric linear regression model relating disease duration and expanded CAG length since the duration values are not normally distributed but instead are highly enriched for values between 10 and 20 years, quite different from the distributions of age at onset<sup>12</sup> and age at death (Figures S7 and S8). Therefore, as a first analysis we instead performed non-parametric statistical analysis, using a generalized additive model and Spearman's rank correlation, which does not require normal distribution of the data, to test whether disease duration is associated with CAG repeat length. This generalized additive model based on 855 subjects with those CAG sizes represented by at least three HD subjects indicated that gender was a significant predictor of disease duration with male HD subjects having a slightly shorter duration compared to female HD subjects (p-value, 0.00015). However, there was no significant association or correlation between disease duration and either the expanded or normal CAG repeat lengths (Figure 3).

We also specifically examined duration values for HD subjects with shorter versus longer expanded CAG repeats. The median CAG repeat in our duration data was 44, so we excluded 3 consecutive CAGs bins around the median CAG in order to generate two arbitrary groups of HD subjects with 1) similar sample sizes and 2) distinct CAG repeat lengths. Disease duration for HD subjects with expanded CAG < 43 (Figure S9A, blue; 247 subjects) and > 45 (Figure S9A, red; 305 subjects) was not significantly different (Figure S9B; Mann-Whitney U test, p-value, 0.484). Furthermore, mutant alleles in HD subjects representing the top 10% (Figure S9C, red; median duration, 27; 87 subjects) and bottom 10% (Figure S9C, blue; median duration, 5; 87 subjects) of duration values show similar expanded CAG repeat sizes (Figure S9D; Mann-Whitney U test, p-value, 0.8975). Finally, we determined whether juvenile onset HD subjects (age at onset < 21 years old, n=55) had different duration values compared to adult onset HD subjects (age at onset > 20 years old, n=823). This analysis revealed the expected significant differences in expanded

CAG repeat lengths, age at onset, and age at death, but no difference in median disease duration between adult onset and juvenile onset HD subjects (Figure 4D; Mann-Whitney U test p-value, 0.75). In our data, the mean and median duration of those 11 juvenile onset HD subjects with CAG repeat size greater than 69 were 9.1 and 9, respectively. Therefore, although disease duration overall in juvenile onset HD individuals (age at onset < 21 years) is similar to that of adult onset HD subjects, those with extreme juvenile onset HD (e.g., age at onset < 10) may have a shorter disease duration.

Consistent with the lack of an effect of the expanded CAG repeat on disease duration in HD heterozygotes, 5 additional HD individuals with 2 expanded CAG repeat alleles had disease duration within one standard deviation (8.6 - 22.2 years) of the mean. Although based due to their rarity on a very small cohort of 'HD homozygote' individuals, this suggests further the lack of a dosage effect of mutant huntingtin on disease duration.

All of the above comparisons were consistent in supporting the conclusion that in HD, disease duration is independent of the length of the expanded CAG repeat. Since a lack of effect of the disease mutation on the duration of clinical disease is counterintuitive, we were concerned that an actual correlation with CAG length might be obscured by some ascertainment bias. There were two sources of ascertainment for our subjects: 1) cohort studies in which HD patients were ascertained for age at onset and subsequently, a subset of individuals died, and 2) brain banks where ascertainment was based upon death. Although our ultimate conclusion was derived from individuals from both groups where duration could be measured directly because both parameters were known, it was conceivable that individuals in observational cohort studies with shorter CAG repeat lengths and potentially longer duration were missed because they have not yet died. Consequently, we performed a survival analysis for our major source of cohort samples with motor onset, 1,426 heterozygous HD individuals from the European Huntington's Disease Network (EHDN) Registry study. Survival analysis based on this observational cohort with censoring indications for surviving HD subjects revealed a significant impact on age at death for expanded CAG (p-value < 2E-16), but not normal CAG (p-value, 0.0926). We then tested whether CAG repeat length and gender influence survival after onset using non-parametric Cox's proportional hazards analysis. We found no effect on disease

duration for either expanded CAG repeat length (exponentiated coefficient per one unit of expanded CAG, 1.007; p-value, 0.754) or of gender (exponentiated coefficient for male, 1.178; p-value, 0.385). Tests for proportional hazards assumption using the "cox.zph" function confirmed that neither CAG repeat size nor gender violated the assumption; p-values for expanded CAG and gender were 0.504 and 0.680, respectively.

A similar analysis of the 10% extremes of CAG repeat length from 855 HD subjects of the MA HD Center Without Walls collection where both age at onset and age at death were known showed an influence only of gender (exponentiated coefficient for low CAG group, 0.8216; p-value, 0.22; exponentiated coefficient for male, 1.2582; p-value, 0.146). Survival analysis using all 855 ascertained subjects and CAG as a continuous variable also revealed no significant contribution of CAG to duration (p-value, 0.665).

Next, we restricted analysis of duration only to those cases ascertained through death, i.e., obtained from HD brain banks. Like the full set of 878 individuals with a direct measurement of duration, this set of 524 brain-bank derived cases showed no significant correlation between CAG repeat length and the length of time between onset and death (Spearman's rank correlation p-value, 0.2738; Pearson's correlation p-value, 0.4529).

Finally, in order to judge the degree to which we had power to visualize an effect of the expanded CAG repeat length on duration, we performed simulation studies based upon the characteristics of our actual data set of 855 cases of known duration (described in the legend for Figure 3). We simulated distributions of the duration data across the CAG repeat range, based upon CAG repeat length explaining 1, 2, 5, 10 or 20% of the variance in duration (Figure S10). Briefly, a data matrix of true CAG repeat size and duration was used to calculate R square value to evaluate the variance in duration that was explained by CAG repeat in observed data (Figure S10A). Duration values were then randomly shuffled to obtain the explanatory power of permuted CAG on duration. Although CAG repeat length accounts for more than 65% of the variance in age at onset and age at death, it has no detectable influence on duration (Figure S10A) while these simulations (Figure S10B-S10F) suggest that even a 2% contribution of CAG repeat length to determining duration could have been detected (Figure S10E).

### **Impact of *HTT* haplotypes and of a genetic modifier of age at onset on duration**

In the absence of an effect of expanded CAG length on HD duration, we next assessed two additional candidate modifiers. We first tested whether common *HTT* haplotypes, which we have shown do not modify age at onset,<sup>27</sup> might instead have an effect on disease duration. *HTT* haplotypes were defined based on 21 common single nucleotide polymorphisms (SNPs) that show significant differences in allele frequencies between HD subjects and normal controls.<sup>27</sup> ANOVA models in which duration was modeled as a function of *HTT* haplotype, either on the disease chromosome (p-value, 0.363) or on the normal chromosome (p-value, 0.091), provided no evidence for a significant association between duration and *HTT* haplotype.

Recently, through a genome-wide association strategy we discovered genetic loci significantly associated with the difference between observed age at onset of motor signs and that expected based upon the CAG repeat length of the individual subjects. The genome-wide significant locus with the largest effect size, detected by SNP rs146353869 on chromosome 15, was estimated to accelerate motor onset by ~6 years suggesting hastening of HD pathogenesis, at least prior to diagnosis. The same SNP was significantly associated with age at death corrected for CAG repeat length (p-value, 9.3E-5). We therefore evaluated the potential influence of this strong age at onset modifier locus on disease duration.<sup>28</sup> As expected and shown in Figure 5, HD subjects with a minor allele for rs146353869 had onset significantly earlier than those without such an allele (Mann-Whitney U test p-value, 0.010). However, disease duration was similar between subjects with or without a rs146353869 minor allele (Mann-Whitney U test p-value, 0.370). Thus, the functional variant tagged by rs146353869 modifies CAG-driven HD pathogenesis but does not influence the length of the CAG-independent period between onset and death.

## Discussion

Our robust statistical analysis based on a large data set firmly establish that, like age at diagnostic onset, the age at death of HD individuals is determined primarily by the size of their *HTT* CAG expansion mutation, with no significant impact from the normal CAG allele. Indeed, the correlation of CAG repeat length is slightly better with age at death than with age at motor onset, due to greater variance in the latter which reflects either the subjectivity of determining onset or its greater susceptibility to modification. It has been suggested from the effects of precise genetic replicas of the CAG mutation in the endogenous mouse orthologue (formerly *Hdh*, now *Htt*) and the fact that these "knock-in" alleles rescue the embryonic lethality of a null *Htt* allele that mutant huntingtin may precipitate pathogenesis by enhancing or dysregulating some normal activity of huntingtin<sup>29</sup>. At least some normal activities of human huntingtin vary with changes in CAG length even in the normal size range, revealing the CAG repeat to be a polymorphism with functional consequences.<sup>30</sup> Our data with respect to age at motor onset and now age at death indicate that any variations in the activity of normal huntingtin occasioned by changes in the normal allele repeat size do not have an effect on the rate of pathogenesis leading to motor onset or to that leading to death.

Since diagnosis of HD is primarily based upon characteristic motor signs, the length of time that an individual actually displays HD, i.e., the period between motor onset and death, is defined here as disease duration. In the literature, HD disease duration varies greatly with reported ranges of 1-45 years and medians of 16.2-21.4 years, although in some cases analyses have included individuals ascertained for either onset or death but not both.<sup>16; 31-33</sup> In our analyses, we have been able to focus on individuals where both age at motor onset and age at death had been ascertained, permitting a direct assignment of disease duration values. Across this data set, disease duration: 1) is not correlated with the expanded CAG repeat length, 2) is similar between HD subjects with low CAG or high CAG expansions and 3) is comparable between adult onset and juvenile onset HD. In addition, expanded CAG lengths are similar between HD subjects with shorter or longer duration. As our studies were based exclusively on individuals who have already displayed disease onset, it remains formally possible that individuals with expanded CAG repeats who are pre-manifest will, in the future, display a disease duration that is correlated with CAG repeat length.



However, we think that this theoretical possibility is unlikely. The potential bias of studies excluding pre-manifest individuals was first considered in examining the CAG-onset relationship based on a survival model by Langbehn et al.<sup>34</sup> The "Langbehn et al. survival model" for CAG-onset relationship was similar to 1) our regression model based on a large cohort of manifest HD subjects (> 4,000) (data not shown), and 2) predictions from previously published formulae for the adult onset CAG repeat range.<sup>35</sup> The similarity in the two models when the regression model data set is large indicates that the exclusion of pre-manifest individuals does not substantially bias the estimate of the CAG-onset relationship. We would expect the same to be true of the CAG-death relationship and that our estimation of duration based on a large sample of subjects fully ascertained for both onset and death is not likely to be substantially biased by the absence of pre-manifest subjects.

Although sometimes assumed, it has not been clearly demonstrated whether juvenile onset HD subjects (< 21 years of age) generally have a different duration from adult onset individuals (> 20 years of age). In one study, disease duration was similar across HD subjects for onset in the juvenile and adult age ranges, with a somewhat shorter duration for those with onset age over 50.<sup>32</sup> A second report based on age at onset of first sign and age at death, or on age alone if HD subjects were still living, noted that extremes of onset, either juvenile or elderly, were associated with significantly shorter durations than typical adult onset.<sup>33</sup> Juvenile onset subjects had a 1-2 year shorter duration (median, 20.0) than those with typical adult onset between 20 and 49 years (median of 21.3 for age at onset between 20 and 34; 22.1 for age at onset between 35 and 49).<sup>33</sup> The lack of consistency in disease duration may be due to different definitions and types of onset.. However, breaking down the juvenile onset category to only those most extreme cases with the longest CAG repeats who developed clinical symptoms before 10 years of age has been reported to reveal a significantly shorter disease duration (range, 2-15 years; mean, 6.6 years) than other HD subjects.<sup>36; 37</sup>

Nevertheless, our data indicate strongly that the duration of the period in which clinical manifestations of HD are expressed, i.e., the period between diagnosis and death, is not altered in the vast majority of HD individuals by the expanded CAG repeat, the normal CAG repeat, *HTT* haplotype, or a strong age at onset

modifier. Our statistical confirmation of this seemingly counterintuitive finding argues against the simple scenario in which a toxic CAG length-determined property of mutant huntingtin drives a single pathogenic process that leads in a sequential manner first to onset and then directly from onset to death. Rather it suggests two alternative scenarios for the HD disease process that should be considered in designing, testing and interpreting the results of therapeutic interventions, particularly those aimed at gene-silencing of the mutant CAG expanded *HTT* allele.

It is well established that the rate of the pathogenic process that leads to clinical diagnosis is determined primarily by the length of the expanded CAG tract. The onset of diagnostic motor abnormalities is thought to ensue from the dysfunction and eventual loss of striatal neurons and it has been estimated that at the time of clinical onset, 30% or more of medium spiny neurons in the striatum have been lost, with the remainder already compromised<sup>38</sup> by a neurodegenerative process that continues as clinical manifestations progressively worsen (commonly referred to as “progression”). Indeed, in studies of post-mortem HD brains, Hadzi et al. have shown that the extent of striatal pathology (pathological grade) is correlated both with CAG repeat length and duration.<sup>16</sup> The former correlation is consistent with a CAG-driven pathogenic process causing striatal pathology. The latter correlation makes sense because the longer the duration of disease, the longer that CAG-driven process has to act, and consequently the greater the extent of striatal pathology. However, we show the lack of correlation of CAG repeat length with duration, suggesting that the extent of striatal pathology *per se* is not a determinant of death and that the reasons for correlation of each of these two parameters with striatal pathology are distinct. In analyses of the 310 brains from our study for which pathological grade is known, we find, like Hadzi et al.,<sup>16</sup> that higher pathological grade associates with longer CAG repeat, earlier onset, and earlier death, all of which are consistent with a more rapid rate of HD pathogenesis with increasing mutation size. The higher pathological grade also associates with longer duration in these individuals, reinforcing the conclusion that disease duration is not determined by CAG repeat length or by extent of striatal pathology.

Like motor onset, age at onset of cognitive and to a lesser extent of psychiatric signs are correlated with the expanded CAG length and may reflect contributions of additional neuropathological changes in cortical

regions. A simple view of pathogenesis, in which the same CAG-driven process in the striatum or in these cortical regions that leads to diagnostic onset simply continues with worsening manifestations and ultimate death, is not a viable hypothesis because of the CAG-independence of disease duration. Instead, we propose two alternative parsimonious explanations for these data.

In the first scenario, the pathogenic process is driven by mutant huntingtin, with motor onset resulting when the coping capacity of the most vulnerable structures, particularly the striatum, has reached its limit. Around or before motor onset, the damage has so weakened the homeostatic mechanisms of the individual that catastrophic failure of the regions associated with onset enables susceptibility to one or more CAG-independent processes that contribute to causing early death, on average 15 years later. In this “two-stage” scenario, CAG-driven pathological changes, including those affecting other cells and tissues, and consequent worsening of symptoms continue to occur after onset, but these CAG-driven changes are not critical to causing death. Instead, death results from one or more separate CAG-independent processes acting on the background of the initial CAG-dependent pathogenesis having reached a critical juncture. The potential CAG-independent processes that may determine duration are not necessarily limited to intrinsic biological risk factors, as many external factors (nutrition, medical care, nursing home care, infectious exposure, etc.) could play a contributory role.

In the second scenario, a pathogenic process is driven by mutant huntingtin independently in multiple cell types in the brain, or in other organs, each of which has a distinct coping potential. The striatum is most vulnerable and so succumbs first and produces clinical manifestations that progress over time. However, neither the occurrence of diagnostic symptoms nor their progression due the underlying CAG length-dependent pathology determines viability of the individual. Rather, some other cells are critical for viability and when these essential cells independently reach their coping limit, averaging 15 years after motor onset, the subject dies. In this second scenario, unlike the first, an *HTT* CAG length-driven pathogenic process leads directly to death in a manner that is not a downstream consequence of the pathology associated with characteristic neurological manifestations.

The leading causes of death of HD subjects as reported in the literature are pneumonia and heart disease.<sup>18; 39</sup> While not directly connected to the brain, each of these causes could occur on the background of physiological changes occasioned by the neuronal dysfunction that becomes evident at the time of HD onset (scenario 1) or be precipitated by direct effects of mutant huntingtin in peripheral cells, such as immune system components or cardiac muscle (scenario 2). Indeed, HD subjects display a number of peripheral abnormalities that could represent either direct or indirect effects of mutant huntingtin, including abnormal energy metabolism, extreme weight loss, diabetes, and reduced pulmonary function.<sup>21; 23; 24; 40-42</sup> These all suggest that although HD is classified as a neurodegenerative disorder, the overall impact of the *HTT* CAG expansion is ultimately rather widespread. Given that HD CAG-driven neuropathological changes in symptom-associated brain regions continue after onset, the early mortality in HD may result from effects of the expanded CAG repeat in vital brain regions not associated with early clinical symptoms (scenario 2), may be due to CAG repeat effects outside the brain (scenario 2), or may be due to intrinsic or extrinsic factors unrelated to the CAG repeat (scenario 1).

Our findings and the two distinct explanations proposed for them have implications for developing and testing disease-modifying therapeutic modalities in HD. Our results clearly indicate that only the expanded *HTT* CAG repeat length has significant power for predicting both age at onset and age at death but that there is remaining variance in these measures unexplained by the CAG mutation that must be due to other causes. One approach that we and others are taking to reveal validated therapeutic targets for HD is to identify genetic modifiers of motor onset in human patients, i.e., genetic factors that alter the rate of CAG-driven HD pathogenesis and therefore can identify proteins and processes to target for development of traditional small molecule therapeutics. An example of such a factor is the as yet undefined functional variant tagged by SNP rs146353869 that dramatically hastens motor onset. In both scenario 1 and 2, drugs based upon modification of CAG-driven processes would be expected to be effective in delaying both onset and death if delivered sufficiently prior to onset. In scenario 2, such interventions might also be expected to delay death even if delivered after motor onset, whereas in scenario 1 they may alleviate the progression of clinical symptoms without actually delaying death. The CAG-independence of disease duration suggests that if the first scenario is correct, a different array of modifiers and potential therapeutic targets may be

effective in delaying death compared to those involved in CAG-dependent processes. Indeed, both alternative scenarios also allow for modifiers that act in a cell or tissue-specific fashion, with the potential of modifying predominantly age at onset or disease duration, but most likely not both.

A fundamentally different route to therapeutic intervention currently being explored is suppression of mutant huntingtin expression through nucleic acid-based gene silencing strategies. Planning for initial clinical trials has typically considered delivery of the therapeutic agent to the brains of HD individuals early after clinical diagnosis, with the goal of delaying the progression of clinical symptoms and associated neuropathology. In our first scenario, such treatment can be very effective in limiting the progression of clinical symptoms and improving quality of life without necessarily altering disease duration since the latter is independent of CAG repeat length. Consequently, although it presents greater regulatory hurdles and difficulties in clinical trial design, the ideal time to treat using a *HTT* gene-suppression strategy is significantly prior to diagnostic onset since an effective treatment would then be expected to delay both onset and death. In the second scenario, treatment of the brain may have no impact on preventing death if the gene silencing strategy is not being delivered to the cells or organ responsible for maintaining viability. Thus, in either scenario, our findings suggest that a *HTT* gene-silencing strategy in the brain that is successful in reducing the progression of symptoms may improve the quality of life of HD individuals without necessarily preventing their early death.

Although it is not possible currently to distinguish between our two scenarios for HD pathogenesis leading to early death, the CAG-independence of disease duration suggests that the results of treatment trials measuring some aspects of clinical progression after onset may not be predictive for the efficacy of the same treatments in preventing motor onset or early death. Fortunately, ongoing large natural history studies of HD offer a route to more fully explore the biological basis for early death in HD in order to distinguish between these explanations and to guide therapeutic development. Most importantly, the application of the continuous CAG analysis strategy, applied here to onset and death, to defining and distinguishing other phenotypic measures (molecular, imaging, neurological, etc.) tied to the pathogenic process offers the hope

of enabling clinical trials before motor onset, when disease-modifying treatments based upon CAG-length dependent effects are expected to be most broadly effective.

## **Supplemental Data**

Supplemental Data include ten figures.

## **Acknowledgments**

The authors acknowledge and thank the Harvard Brain Tissue Resource Center, McLean Hospital (Belmont MA, USA), Dr. Francine Benes, Director, the New York Brain Bank, Columbia University (New York City, NY), Dr. Jean-Paul Vonsattel, Director, the National Neurological Research Specimen Bank (Department of Veterans Affairs Medical Center, Los Angeles CA), Dr. Wallace Tourtellotte Director, and the Neuropathology Core of the MA Alzheimer Disease Research Center (Matt Frosch, Director) and the Harvard NeuroDiscovery Center Advanced Tissue Resources Core (Massachusetts General Hospital, Charlestown MA), Dr. Charles Vanderburg, Director, for providing HD postmortem brain tissues. We also thank Dr. Richard H. Myers (Boston University School of Medicine) for helpful comments on this manuscript. This work was supported by NIH grants NINDS NS16367, U01NS082079, R01NS091161 and the CHDI Foundation, Inc. The authors declare that no competing interests exist.

## **Web Resources**

Online Mendelian Inheritance in Man (OMIM), <http://www.omim.org/>

## References

1. Huntington, G. (1872). On chorea. *Med Surg Rep* 26, 320-321.
2. HD-CRG. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* 72, 971-983.
3. Cattaneo, E., Zuccato, C., and Tartari, M. (2005). Normal huntingtin function: an alternative approach to Huntington's disease. *Nature reviews Neuroscience* 6, 919-930.
4. Bates, G.P. (2005). History of genetic disease: the molecular genetics of Huntington disease - a history. *Nature Reviews Genetics* 6, 766-773.
5. Gusella, J., and MacDonald, M. (2002). No post-genetics era in human disease research. *Nature reviews Genetics* 3, 72-79.
6. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., et al. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature genetics* 4, 398-403.
7. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature genetics* 4, 387-392.
8. Snell, R.G., MacMillan, J.C., Cheadle, J.P., Fenton, I., Lazarou, L.P., Davies, P., MacDonald, M.E., Gusella, J.F., Harper, P.S., and Shaw, D.J. (1993). Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nature genetics* 4, 393-397.
9. Persichetti, F., Srinidhi, J., Kanaley, L., Ge, P., Myers, R.H., D'Arrigo, K., Barnes, G.T., MacDonald, M.E., Vonsattel, J.P., Gusella, J.F., et al. (1994). Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. *Neurobiology of disease* 1, 159-166.
10. Aziz, N.A., Jurgens, C.K., Landwehrmeyer, G.B., van Roon-Mom, W.M., van Ommen, G.J., Stijnen, T., and Roos, R.A. (2009). Normal and mutant HTT interact to affect clinical severity and progression in Huntington disease. *Neurology* 73, 1280-1285.
11. Djousse, L., Knowlton, B., Hayden, M., Almqvist, E.W., Brinkman, R., Ross, C., Margolis, R., Rosenblatt, A., Durr, A., Dode, C., et al. (2003). Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *American journal of medical genetics* 119A, 279-282.
12. Lee, J.M., Ramos, E.M., Lee, J.H., Gillis, T., Mysore, J.S., Hayden, M.R., Warby, S.C., Morrison, P., Nance, M., Ross, C.A., et al. (2012). CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* 78, 690-695.
13. Guo, Z., Rudow, G., Pletnikova, O., Codispoti, K.E., Orr, B.A., Crain, B.J., Duan, W., Margolis, R.L., Rosenblatt, A., Ross, C.A., et al. (2012). Striatal neuronal loss correlates with clinical motor impairment in Huntington's disease. *Movement disorders : official journal of the Movement Disorder Society* 27, 1379-1386.
14. Paulsen, J.S., Nopoulos, P.C., Aylward, E., Ross, C.A., Johnson, H., Magnotta, V.A., Juhl, A., Pierson, R.K., Mills, J., Langbehn, D., et al. (2010). Striatal and white matter predictors of estimated diagnosis for Huntington disease. *Brain Res Bull* 82, 201-207.
15. Rosas, H.D., Goodman, J., Chen, Y.I., Jenkins, B.G., Kennedy, D.N., Makris, N., Patti, M., Seidman, L.J., Beal, M.F., and Koroshetz, W.J. (2001). Striatal volume loss in HD as measured by MRI and the influence of CAG repeat. *Neurology* 57, 1025-1028.
16. Hadzi, T.C., Hendricks, A.E., Latourelle, J.C., Lunetta, K.L., Cupples, L.A., Gillis, T., Mysore, J.S., Gusella, J.F., MacDonald, M.E., Myers, R.H., et al. (2012). Assessment of cortical and striatal involvement in 523 Huntington disease brains. *Neurology* 79, 1708-1715.
17. Pekmezovic, T., Svetel, M., Maric, J., Dujmovic-Basuroski, I., Dragasevic, N., Keckarevic, M., Romac, S., and Kostic, V.S. (2007). Survival of Huntington's disease patients in Serbia: longer survival in female patients. *European journal of epidemiology* 22, 523-526.
18. Lanska, D.J., Lavine, L., Lanska, M.J., and Schoenberg, B.S. (1988). Huntington's disease mortality in the United States. *Neurology* 38, 769-772.
19. Morton, A.J. (2013). Circadian and sleep disorder in Huntington's disease. *Experimental neurology* 243, 34-44.
20. Petersen, A., and Gabery, S. (2012). Hypothalamic and Limbic System Changes in Huntington's Disease. *Journal of Huntington's disease* 1, 5-16.
21. Reyes, A., Cruickshank, T., Ziman, M., and Nosaka, K. (2014). Pulmonary function in patients with Huntington's disease. *BMC pulmonary medicine* 14, 89.
22. Andreassen, O.A., Dedeoglu, A., Stanojevic, V., Hughes, D.B., Browne, S.E., Leech, C.A., Ferrante, R.J., Habener, J.F., Beal, M.F., and Thomas, M.K. (2002). Huntington's disease of the endocrine pancreas: insulin deficiency and diabetes mellitus due to impaired insulin gene expression. *Neurobiology of disease* 11, 410-424.



23. Hu, Y., Liang, J., and Yu, S. (2014). High prevalence of diabetes mellitus in a five-generation Chinese family with Huntington's disease. *Journal of Alzheimer's disease : JAD* 40, 863-868.
24. Djousse, L., Knowlton, B., Cupples, L.A., Marder, K., Shoulson, I., and Myers, R.H. (2002). Weight loss in early stage of Huntington's disease. *Neurology* 59, 1325-1330.
25. Kremer, H.P., and Roos, R.A. (1992). Weight loss in Huntington's disease. *Archives of neurology* 49, 349.
26. Perlis, R.H., Smoller, J.W., Mysore, J., Sun, M., Gillis, T., Purcell, S., Rietschel, M., Nothen, M.M., Witt, S., Maier, W., et al. (2010). Prevalence of incompletely penetrant Huntington's disease alleles among individuals with major depressive disorder. *Am J Psychiatry* 167, 574-579.
27. Lee, J.M., Gillis, T., Mysore, J.S., Ramos, E.M., Myers, R.H., Hayden, M.R., Morrison, P.J., Nance, M., Ross, C.A., Margolis, R.L., et al. (2012). Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene region. *Am J Hum Genet* 90, 434-444.
28. Genetic Modifiers of Huntington's Disease Consortium. Electronic address, g.h.m.h.e., and Genetic Modifiers of Huntington's Disease Ge, M.H.D.C. (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* 162, 516-526.
29. White, J.K., Auerbach, W., Duyao, M.P., Vonsattel, J.P., Gusella, J.F., Joyner, A.L., and MacDonald, M.E. (1997). Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nature genetics* 17, 404-410.
30. Seong, I.S., Ivanova, E., Lee, J.M., Choo, Y.S., Fossale, E., Anderson, M., Gusella, J.F., Laramie, J.M., Myers, R.H., Lesort, M., et al. (2005). HD CAG repeat implicates a dominant property of huntingtin in mitochondrial energy metabolism. *Hum Mol Genet* 14, 2871-2880.
31. Pridmore, S.A. (1990). Age of death and duration in Huntington's disease in Tasmania. *The Medical journal of Australia* 153, 137-139.
32. Roos, R.A., Hermans, J., Vegter-van der Vlis, M., van Ommen, G.J., and Bruyn, G.W. (1993). Duration of illness in Huntington's disease is not related to age at onset. *Journal of neurology, neurosurgery, and psychiatry* 56, 98-100.
33. Foroud, T., Gray, J., Ivashina, J., and Conneally, P.M. (1999). Differences in duration of Huntington's disease based on age at onset. *Journal of neurology, neurosurgery, and psychiatry* 66, 52-56.
34. Langbehn, D.R., Brinkman, R.R., Falush, D., Paulsen, J.S., Hayden, M.R., and International Huntington's Disease Collaborative, G. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical genetics* 65, 267-277.
35. Langbehn, D.R., Hayden, M.R., Paulsen, J.S., and Group, P.-H.I.o.t.H.S. (2010). CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 153B, 397-408.
36. Byers, R.K., and Dodge, J.A. (1967). Huntington's chorea in children. Report of four cases. *Neurology* 17, 587-596.
37. Nahhas, F.A., Garbern, J., Krajewski, K.M., Roa, B.B., and Feldman, G.L. (2005). Juvenile onset Huntington disease resulting from a very large maternal expansion. *American journal of medical genetics* 137A, 328-331.
38. Aylward, E.H., Sparks, B.F., Field, K.M., Yallapragada, V., Shpritz, B.D., Rosenblatt, A., Brandt, J., Gourley, L.M., Liang, K., Zhou, H., et al. (2004). Onset and rate of striatal atrophy in preclinical Huntington disease. *Neurology* 63, 66-72.
39. Lanska, D.J., Lanska, M.J., Lavine, L., and Schoenberg, B.S. (1988). Conditions associated with Huntington's disease at death. A case-control study. *Archives of neurology* 45, 878-880.
40. Berent, S., Giordani, B., Lehtinen, S., Markel, D., Penney, J.B., Buchtel, H.A., Starosta-Rubinstein, S., Hichwa, R., and Young, A.B. (1988). Positron emission tomographic scan investigations of Huntington's disease: cerebral metabolic correlates of cognitive function. *Annals of neurology* 23, 541-546.
41. Young, A.B., Penney, J.B., Starosta-Rubinstein, S., Markel, D., Berent, S., Rothley, J., Betley, A., and Hichwa, R. (1987). Normal caudate glucose metabolism in persons at risk for Huntington's disease. *Archives of neurology* 44, 254-257.
42. Young, A.B., Penney, J.B., Starosta-Rubinstein, S., Markel, D.S., Berent, S., Giordani, B., Ehrenkauf, R., Jewett, D., and Hichwa, R. (1986). PET scan investigations of Huntington's disease: cerebral metabolic correlates of neurological features and functional decline. *Annals of neurology* 20, 296-303.

## Figure Legends

### Figure 1. Correlation between expanded *HTT* CAG repeat length and age at death.

A) Age at onset of motor signs plotted against the expanded CAG repeat length.

B) Age at death plotted against the expanded CAG repeat length.

C) Age at death plotted against age at onset for individuals where both are known.

In each panel, each circle represents a unique HD subject. The red trend line represents a statistical model based on all data points prior to quality control analysis, describing the relationship between natural log transformed age at onset and expanded CAG repeat length (A), between natural log transformed age at death and expanded CAG repeat length (B) and between age at death and age at onset (C). A summary of a model, including model formula, sample number (N), and model's adjusted R-squared value (Adj.R<sup>2</sup>) is provided inside of each plot. The larger variance in age at onset (A) compared to age at death (B) is not due to statistical artifacts related to sample size as the R-squared values for age at onset models based on randomly picked 1,165 subject sample sets (mean, 0.6537) were similar to that of the original model using all data points.

### Figure 2. Distribution of disease duration by *HTT* CAG repeat length.

A) Models for age at death (Table 1, Model 2) and age at onset<sup>12</sup> were constructed using normally distributed samples with gender covariate. For the age at onset model, gender was included in the model described previously.<sup>12</sup> Blue and red lines represent statistical models for males and females, respectively. Solid and dotted lines means the CAG - age at death model and the CAG - age at onset model, respectively.

B) Distribution of disease duration for each expanded CAG repeat was plotted using boxplot format. Open circles are outliers defined by a standard interquartile outlier identification method.

**Figure 3. Non-parametric analyses to test effects of CAG repeat size on duration.**

A) Sample sizes were plotted against expanded CAG repeats for statistical modeling of duration.

Duration values were plotted against either expanded CAG (B), or normal CAG (C).

D) A generalized linear model (GAM) was constructed to test effects of expanded CAG repeat, normal CAG and gender on duration in the non-normally distributed data. Expanded CAGs with sample sizes greater than 2 were used (CAG 38-63). P-values of independent variables are provided. E) In addition, Spearman's rank correlation analysis was performed to determine whether duration values correlated with the sizes of either expanded or normal CAG repeats.

**Figure 4. HD subjects with adult onset and juvenile onset have similar duration.**

HD subjects with adult age at onset (>20 years; N=823) or juvenile age at onset (< 21 years; N=55) were compared for their expanded CAG repeats (A), age at onset (B), age at death (C) and duration (D) using the Mann-Whitney U test. In each panel, a box plot summarizes the distribution of the test object (left) with a summary statistics table provided (right).

**Figure 5. Duration is not associated with rs146353869.**

Based on recent genome-wide association (GWA) analysis results, we tested whether duration is altered by a strong genetic modifier tagged by SNP rs146353869 on chromosome 15. HD subjects with the minor allele of this SNP developed clinical symptoms significantly (~6 years) earlier than subjects with comparable expanded CAG length but without the minor allele of rs146353869.

A) Among samples used in our GWA analysis aimed at identifying age at onset modifiers, 654 individuals have both age at onset and age at death data. Disease duration (i.e., age at death minus age at onset) was

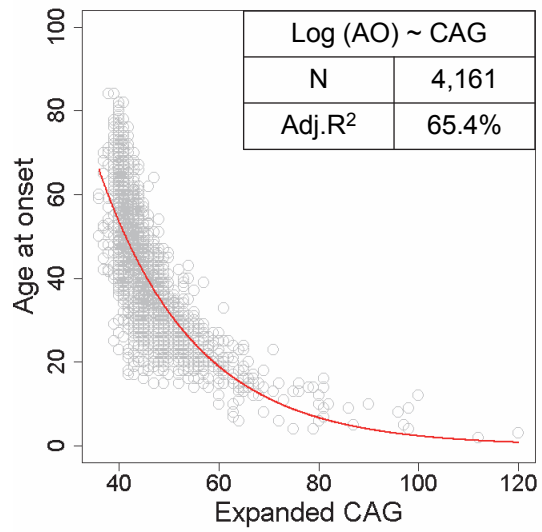
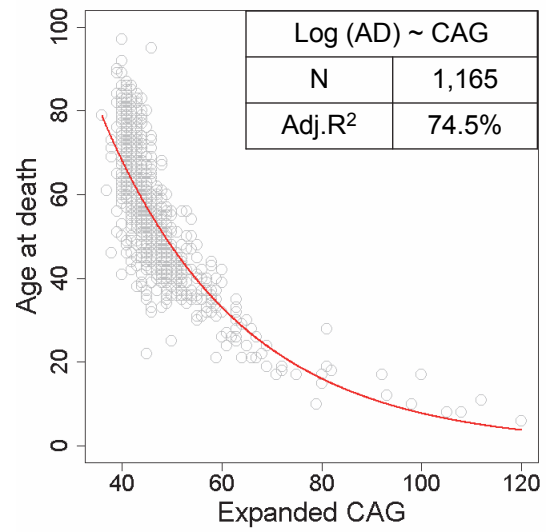
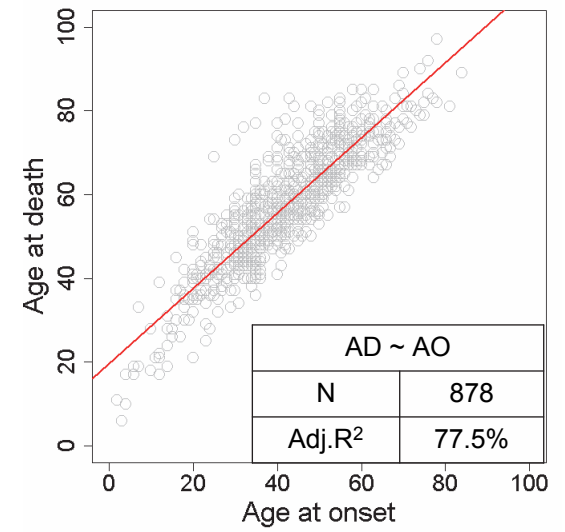
plotted against residual age at onset. Grey (636 subjects) and red circles (18 subjects) represent HD subjects without and with a minor allele for rs146353869.

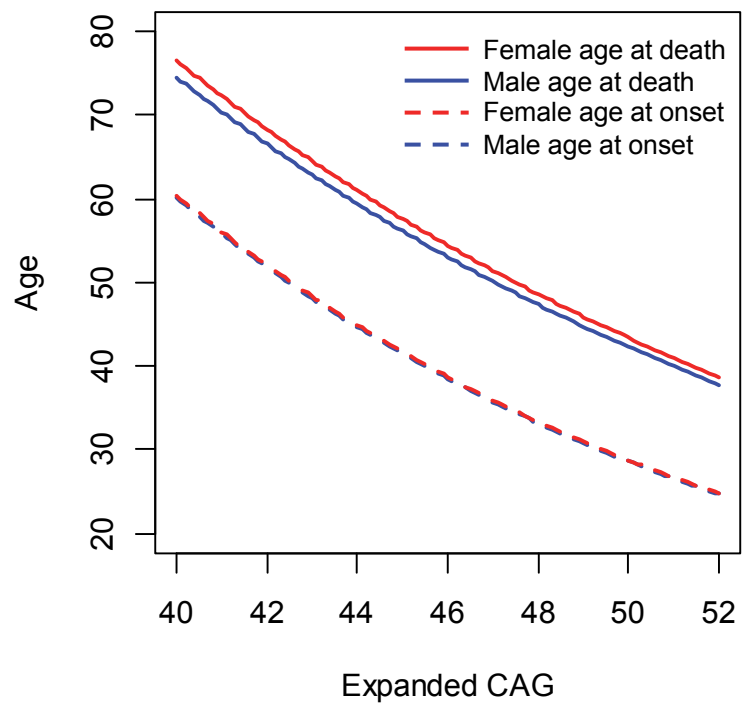
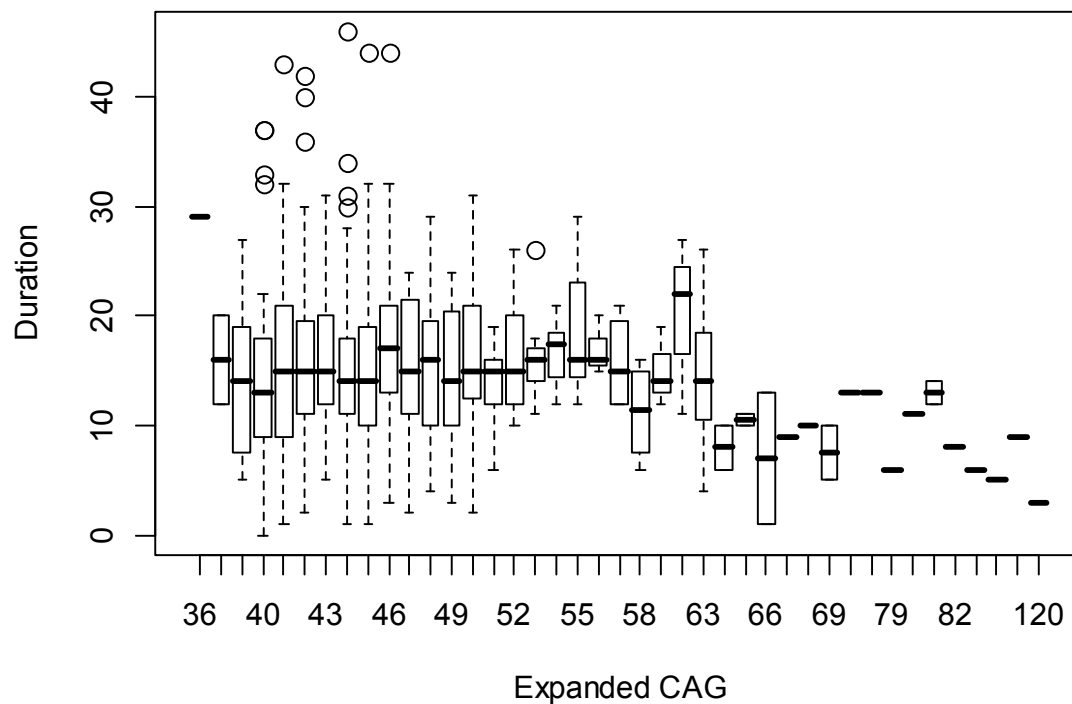
Residual age at onset (B) and duration (C) of HD subjects without (Genotype 0) or with (Genotype 1) a minor allele for rs146353869 are plotted. Black horizontal lines represent the mean. Mann-Whitney U tests were performed to compare residual age at death (p-value, 0.01) and duration (p-value, 0.37) between the two groups of HD subjects differentiated by the rs146353869 genotype.

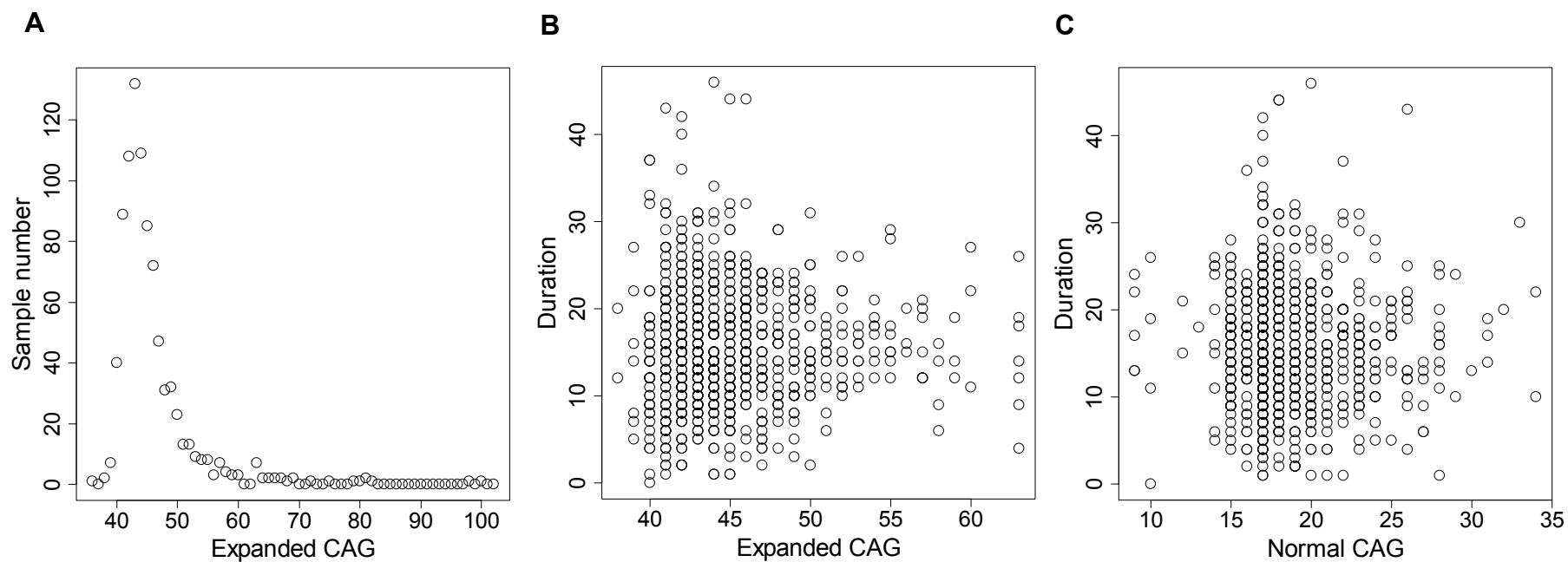
**Table 1. Summary of statistical models to test effects of CAG repeat sizes on age at death in HD.**

	Sample size	Expanded CAG Estimate (p-value)	Normal CAG Estimate (p-value)	Gender (male) Estimate (p-value)	Adjusted R <sup>2</sup> (%)
Model 1	1,019	-0.056722 ( $< 2e-16$ )	-0.001636 (0.134840)	-0.026275 (0.000288)	66.9
Model 2	1,019	-0.056723 ( $< 2e-16$ )	Not tested	-0.026240 (0.000296)	66.9
Model 3	1,165	-0.0363498 ( $< 2e-16$ )	-0.00259 (0.06007)	-0.0288 (0.00157)	74.8%

Two different statistical models were constructed to test the impact of CAG repeats on age at death of HD subjects using QC-passed data points. Natural log transformed age at death was modeled as a function of expanded CAG, normal CAG, and gender using QC passed data points (Model 1). After confirming the lack of influence of the normal CAG, Model 2 was constructed by using only the expanded CAG repeat and gender. Finally, to evaluate the impact of samples excluded in Models 1 and 2, Model 3 was fitted to data including all HD subjects with age at death data. Interaction between expanded and normal CAG repeats was not significant, and therefore was excluded from modeling.

**A****B****C**

**A****B**



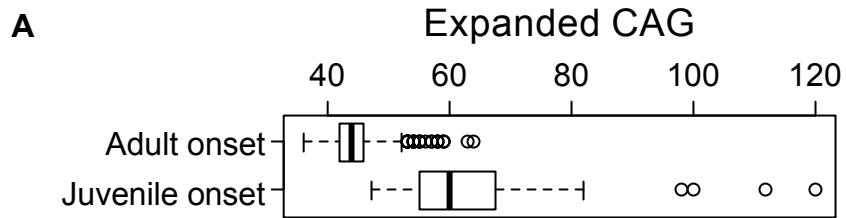
**D**

Generalized additive model	Sample size	Expanded CAG p-value	Normal CAG p-value	Gender p-value
Duration ~ s(expanded CAG) + s(normal CAG) + gender	855	0.665	0.973	0.00015

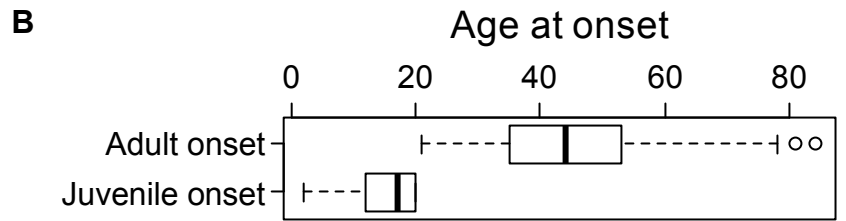
**E**

Spearman's rank correlation	Sample size	p-value
Duration and expanded CAG	855	0.1415
Duration and normal CAG	855	0.6266

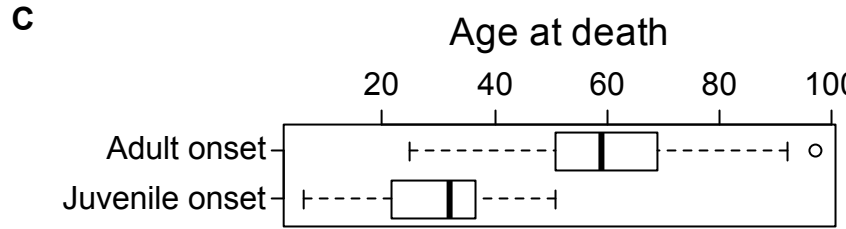




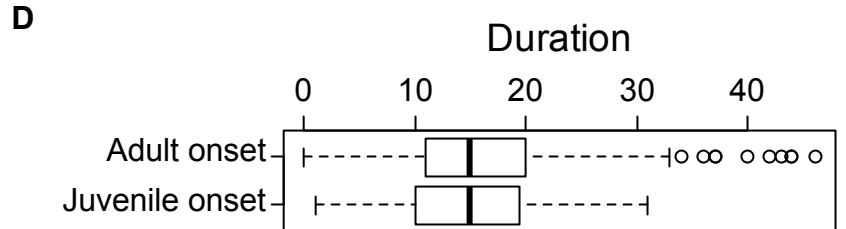
Expanded CAG	Range	Median	P-value
Adult	36-64	44	4.01E-33
Juvenile	47-120	60	



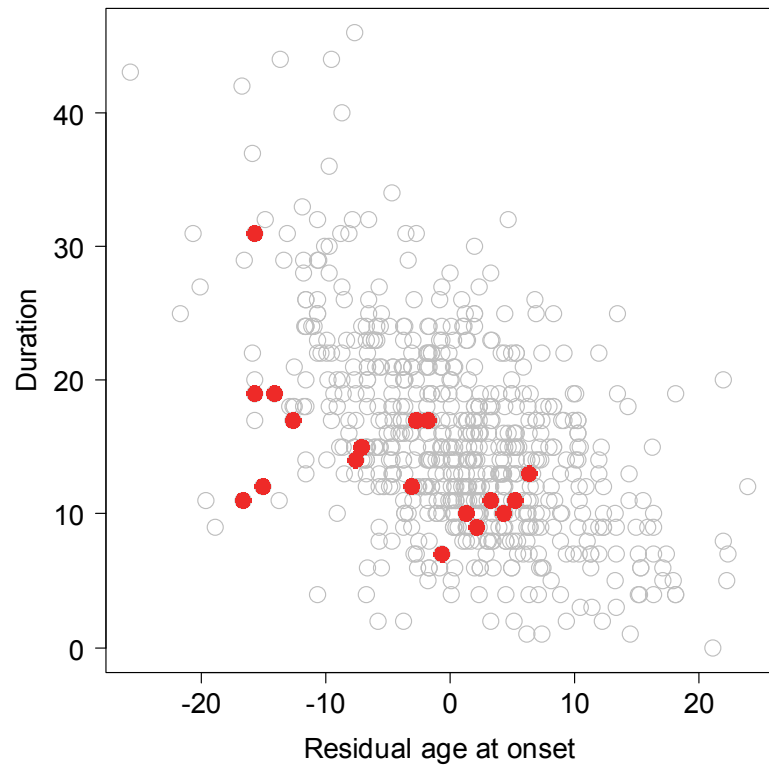
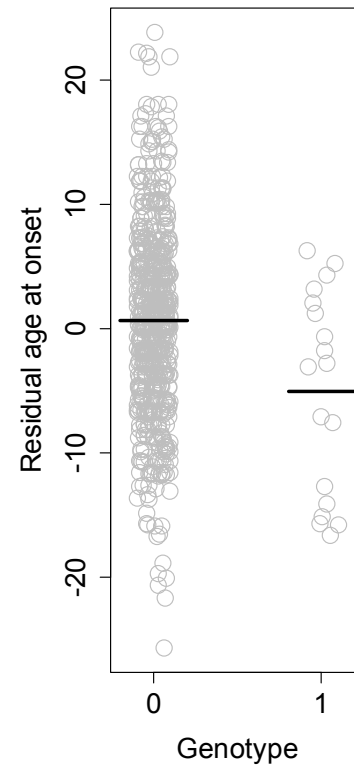
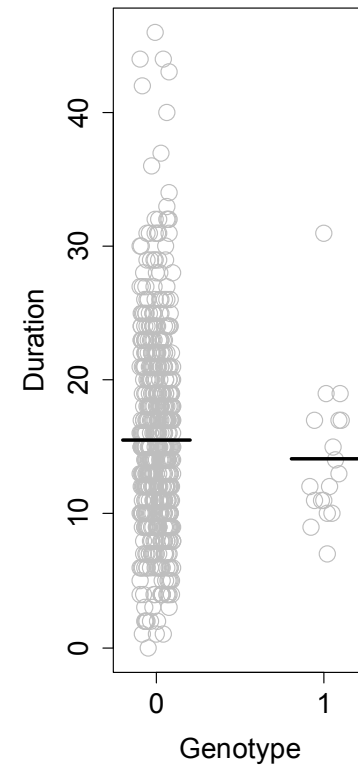
Age at onset	Range	Median	P-value
Adult	21-84	44	1.72E-35
Juvenile	2-20	17	



Age at death	Range	Median	P-value
Adult	25-97	59	6.11E-32
Juvenile	6-51	32	



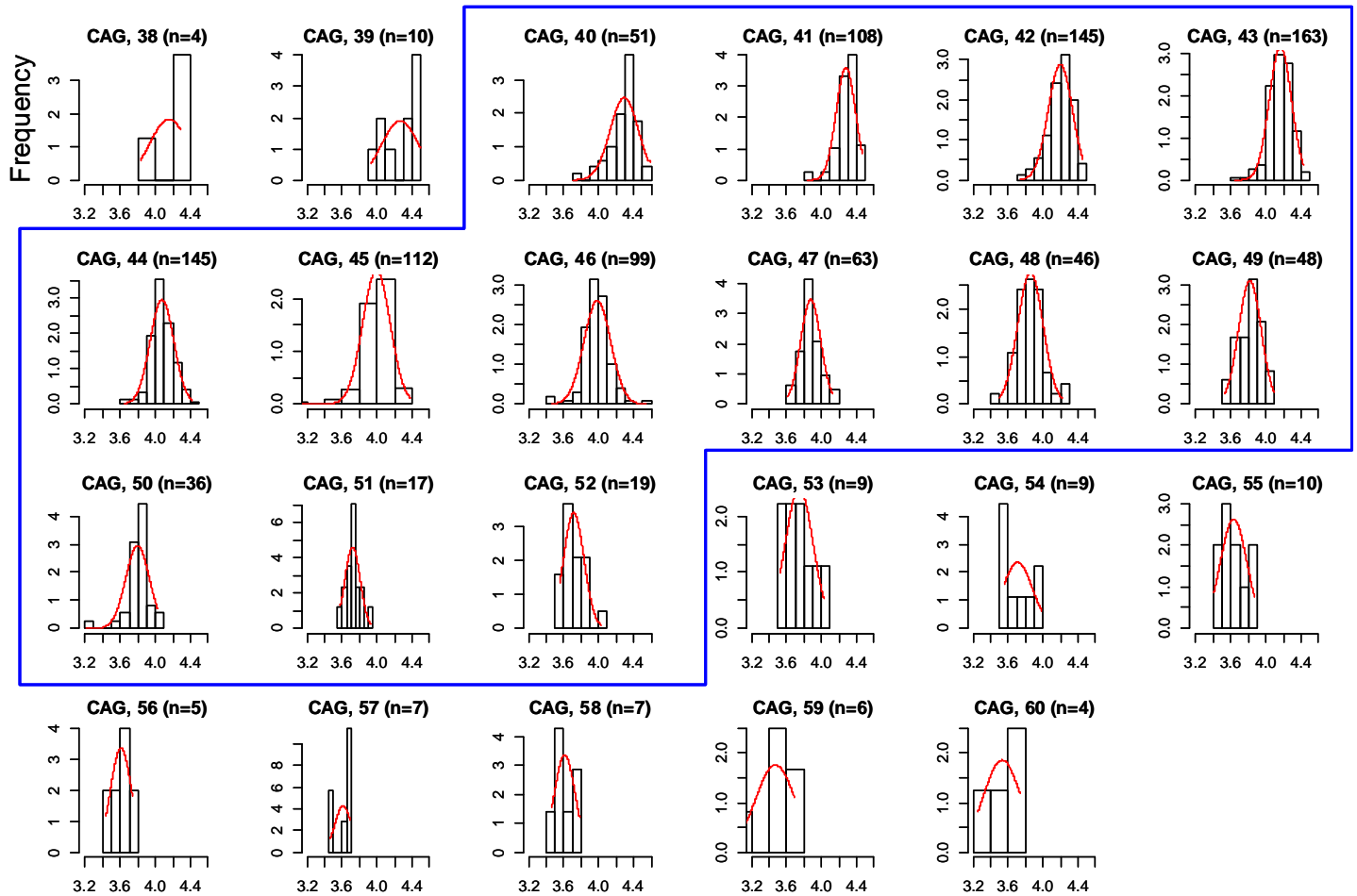
Duration	Range	Median	P-value
Adult	0-46	15	0.75
Juvenile	1-31	15	

**A****B****C**

1 **Figure S1. Evaluation of normality of age at death data.**

2 Data normality was evaluated by comparing distribution of age at death for a given expanded CAG repeat  
3 length (histogram) to a theoretical normal distribution based on the mean and standard deviation of age at  
4 death (red line). The expanded CAG repeat length and sample size are indicated at the top of each plot.  
5 Histograms inside of the boundary in blue (CAG 40-52) resembled theoretical normal distributions.

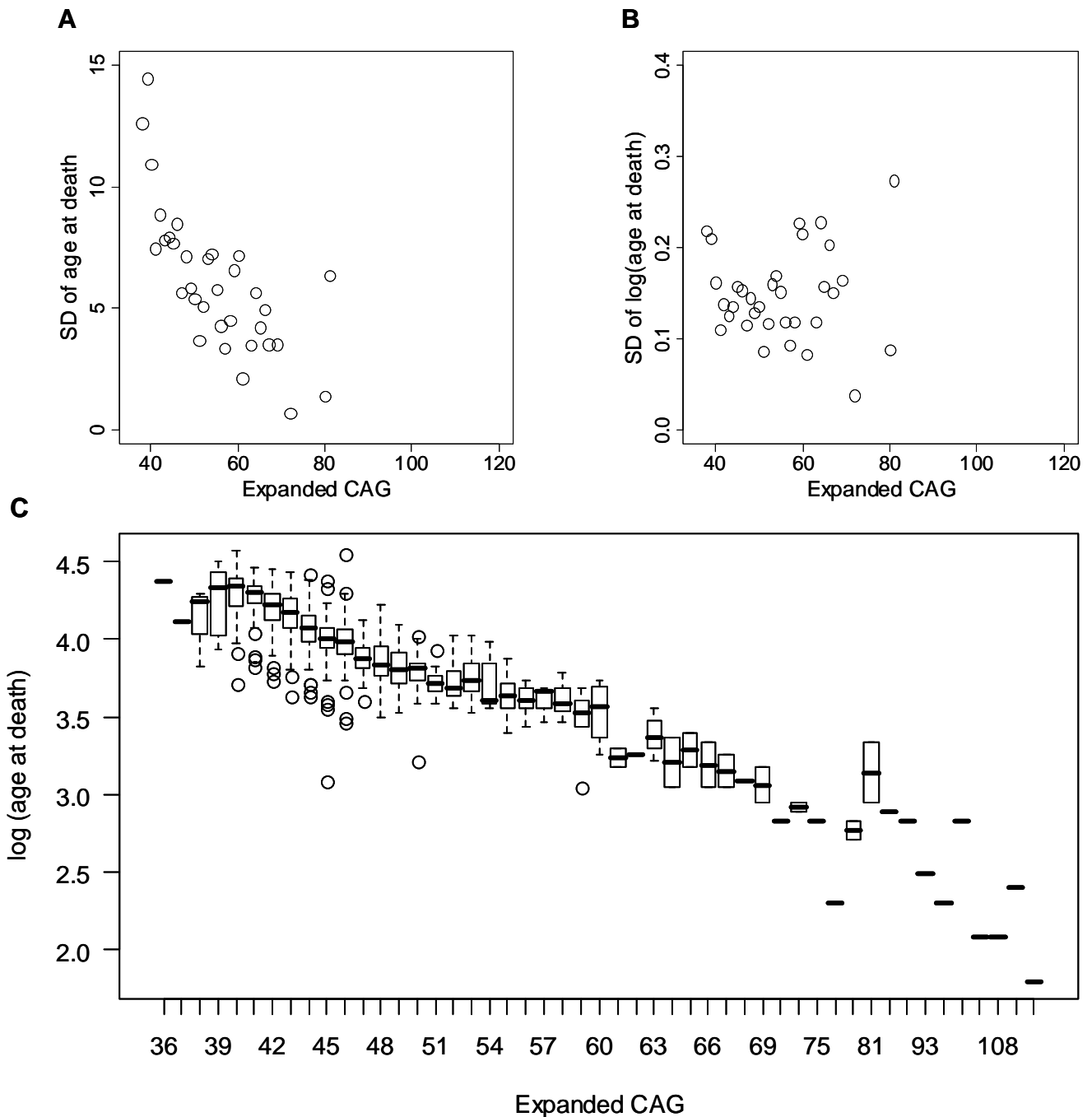
6



1 **Figure S2. Variance and outliers in age at death data.**

2 A) Variance of age at death was evaluated by plotting standard deviation of age at death against the expanded  
3 CAG repeat length. B) To resolve the non-constant variance problem in age at death data for subsequent  
4 parametric modeling, age at death was transformed into log scale (natural log), and standard deviation was re-  
5 calculated for each expanded CAG. C) Log transformed age at death was plotted against expanded CAG  
6 repeat on a box plot to identify phenotypic outliers. Outliers were identified by a standard interquartile method  
7 for each CAG repeat as described previously<sup>12</sup>, and indicated open circles.

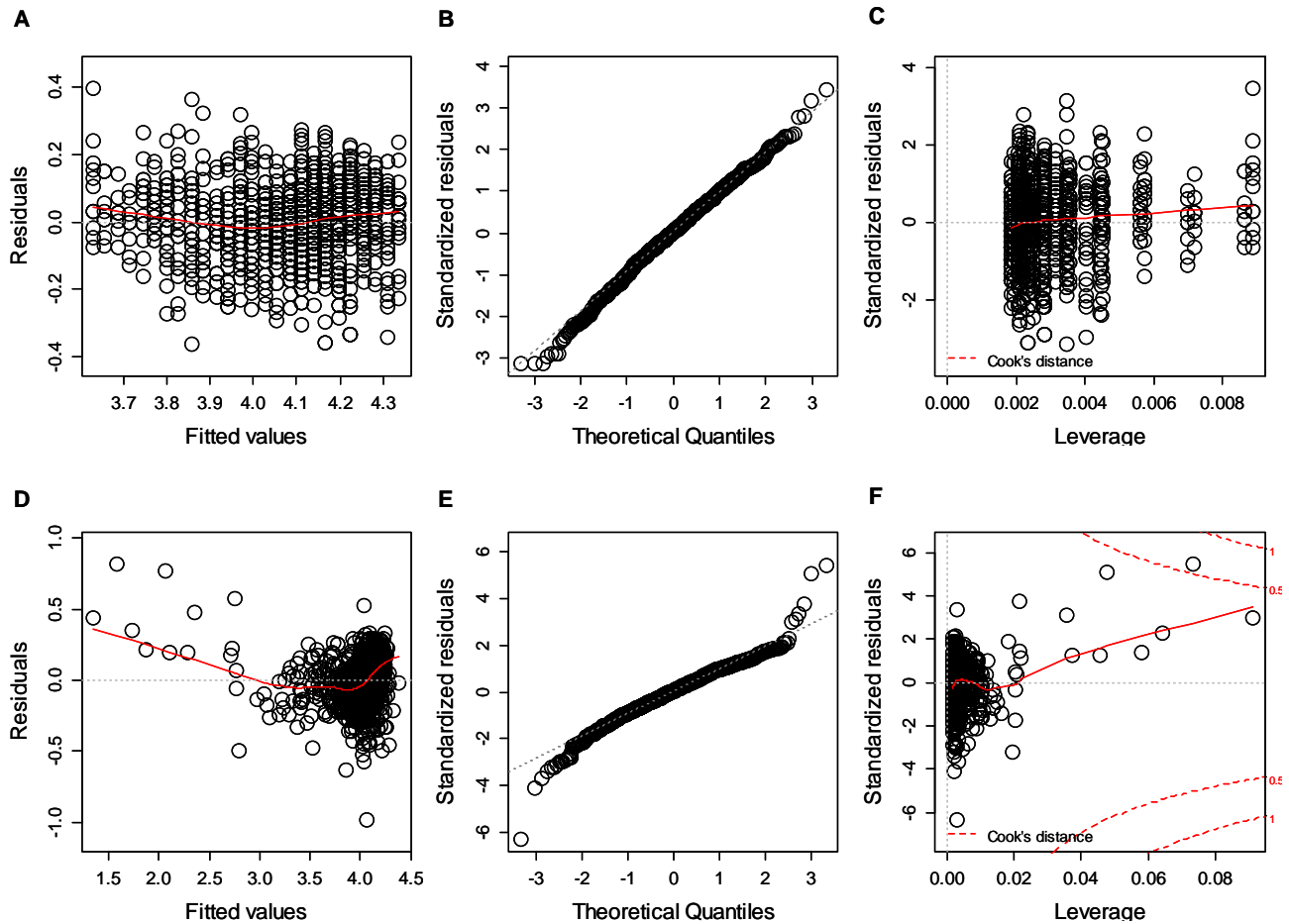
8



1 **Figure S3. Model diagnostic plots.**

2 For the QC dataset used to generate the Model 2 and a model using all samples (Model 3) in Table 1, we  
3 determined whether the requirements of linear models were met. Specifically, we checked variance (A and D),  
4 normality (B and E), and leverage (C and F). In a model using only QC-passed data, variance and normality  
5 were greatly improved compared to those of model using all data points (A vs. D; B vs. E), supporting its  
6 reliability. A and D) Residuals calculated from a model using normally distributed samples are compared to  
7 fitted values. B and E) Normality of the model using normally distributed samples was assessed by comparing  
8 actual residuals to theoretical residuals in a quantile-quantile plot. C and F) To identify influential data points in  
9 the model using normally distributed samples, standardized residuals were plotted against leverage and shown  
10 with the Cook's distance (red dotted contour lines). Leverage is commonly used to identify observations that  
11 have a disproportionate effect on the regression model, and a data point with high leverage indicates that that  
12 observation is distantly located from the center of the measurements. Cook's distance estimates the influence  
13 of data points on a model fit by measuring the effect of deleting a given observation. Red lines in plots  
14 represent LOWESS regression smoothed lines, based on locally-weighted polynomial regression models  
15 describing trends between values on the X-axis and Y-axis.

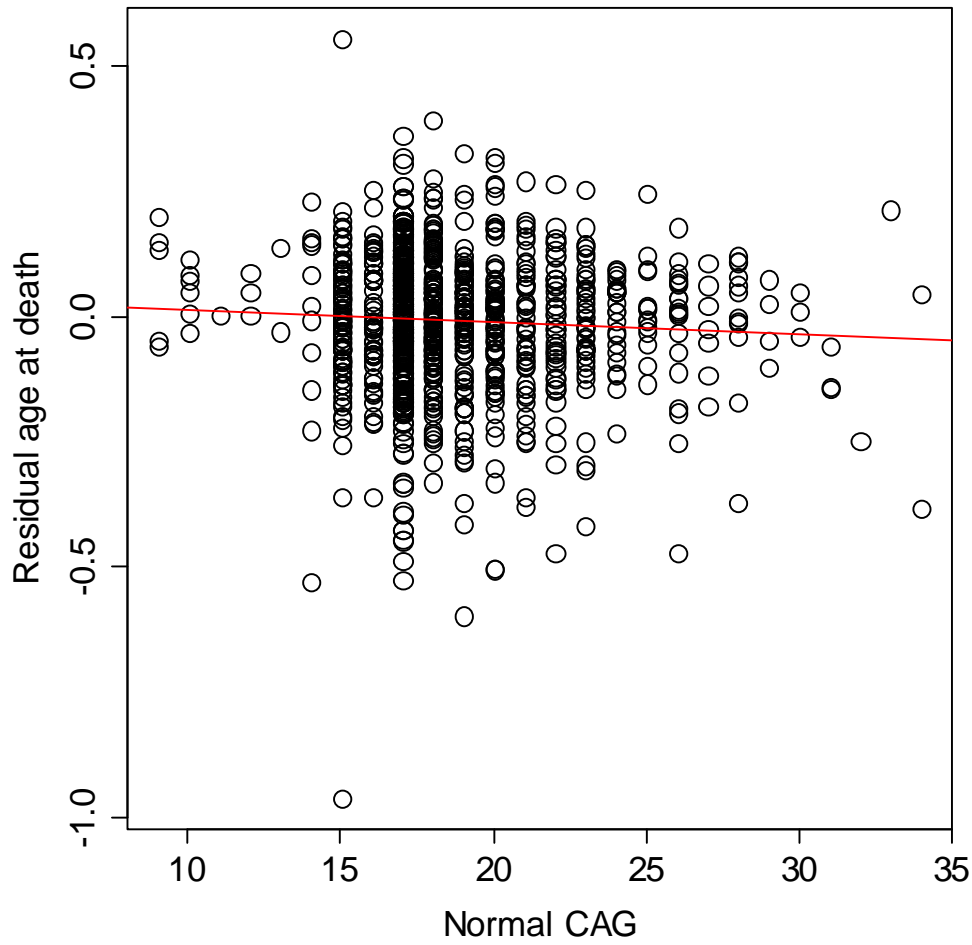
16



1 **Figure S4. Normal CAG repeat does not explain age at death.**

2 To test whether data points excluded as outliers show evidence of an influence of the normal CAG allele on  
3 age at death, residuals of all samples were calculated from the minimal adequate model described in Table 1  
4 (Model 2). Subsequently, residuals were modeled as a function of normal CAG repeat length. The red line  
5 represents the model with an adjusted R-squared value of 0.2648%, indicating that there is no significant  
6 relationship between normal CAG repeat length and age at death (p-value, 0.0521).

7

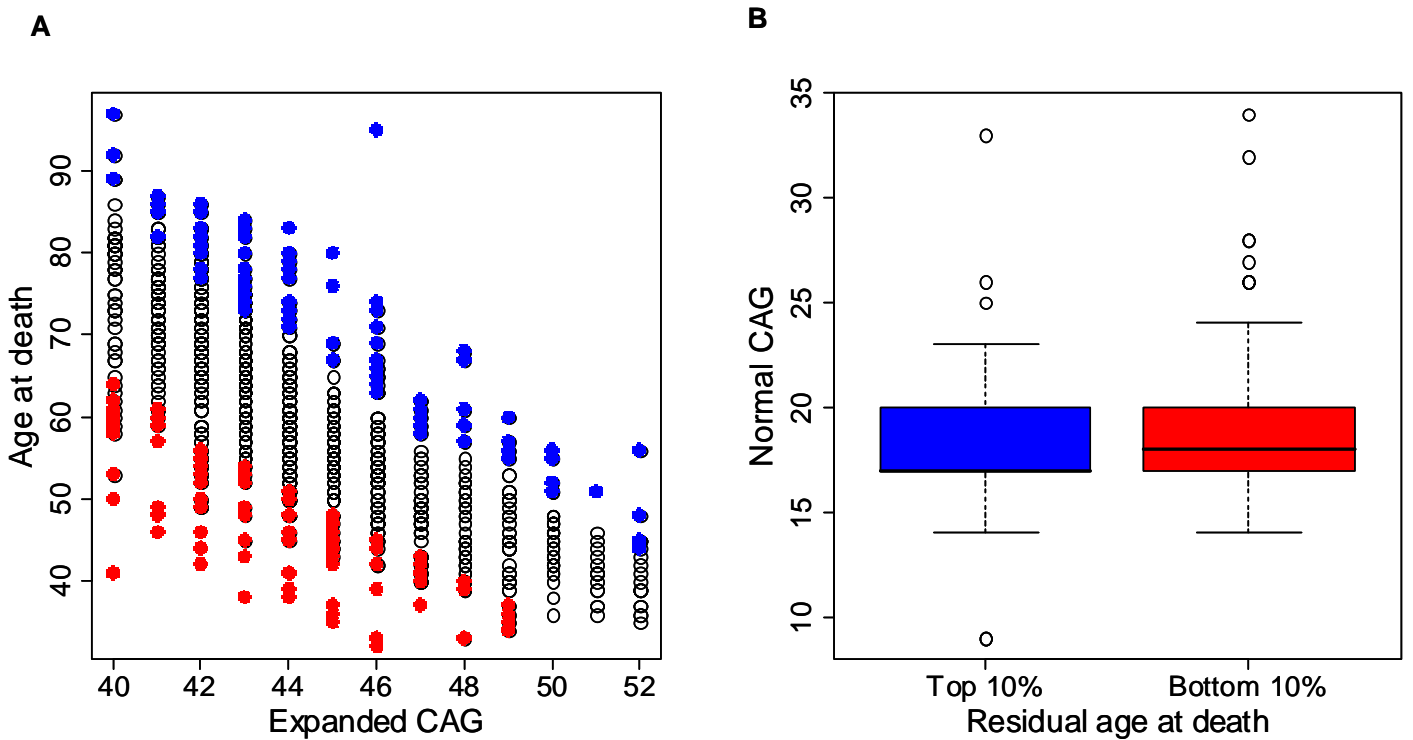


1 **Figure S5. Extreme age at death samples do not differ in normal CAG repeat lengths.**

2 To test whether the 10% extremes of residual of age at death based on the model described in Table 1 (Model  
3 2) had different normal CAG repeat lengths, 105 samples representing to top 10% and 105 samples  
4 representing the bottom 10% of residuals were identified.

5 A) Residual of age at death was plotted against expanded CAG repeat length and the 10% extremes shown as  
6 blue and red circles. B) Normal CAG repeat lengths (Y-axis) were compared between the 10% extremes from  
7 panel A, and did not differ (Mann-Whitney U test p-value, 0.06414)

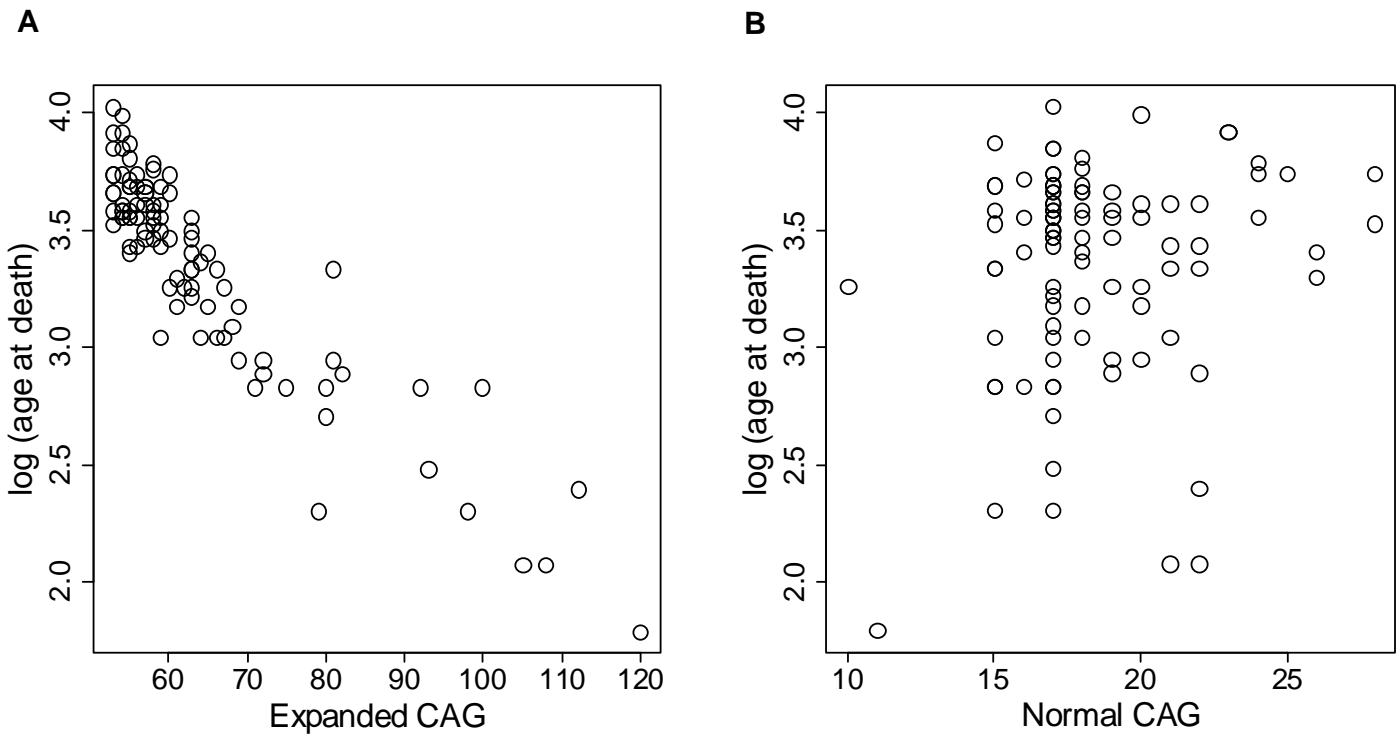
8



**Figure S6. Age at death is determined by expanded CAG repeat length in a fully dominant fashion in samples with expanded CAG > 52.**

To test whether normal CAG repeats had significant effects on age at death in HD subjects with expanded CAG repeats greater than 52 units, 97 such subjects were identified.

A) Log transformed age at death of HD subjects with expanded CAG repeats greater than 52 was plotted against expanded CAG repeat length. B) Log transformed age at death of the same subjects was plotted against normal CAG repeat. C) A multiple regression model to fit the data was generated. In this model, log transformed age at death of HD subjects with expanded CAG > 52 was modeled as a function of expanded CAG repeat, and normal CAG repeat.



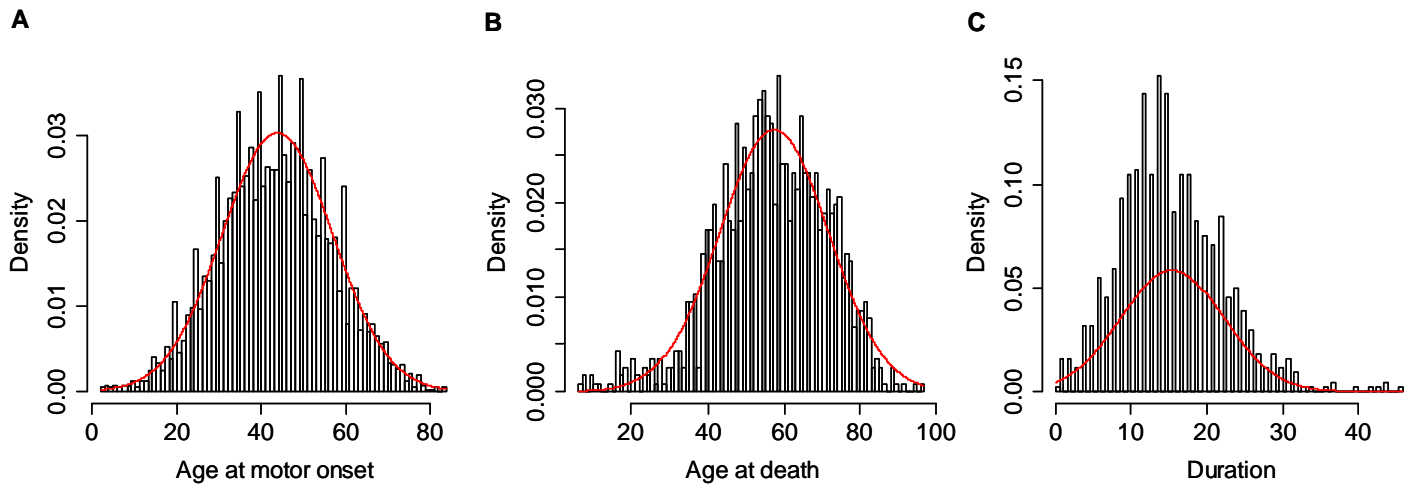
**C**

Samples	Sample size	Expanded CAG p-value	Normal CAG p-value	Adjusted R <sup>2</sup>
CAG > 52	97	<2e-16	0.941	81.65%



1 **Figure S7. Non-normal distribution of duration.**

2 Relative frequencies (density on Y axis) of age at onset of motor signs (A; 4,161 samples), age at death (B;  
3 1,165 samples), and duration (C; 878 samples) for each CAG repeat were plotted in histograms. All data  
4 without quality control analysis were plotted. Red lines represent theoretical normal distributions based on the  
5 means and standard deviations of data.

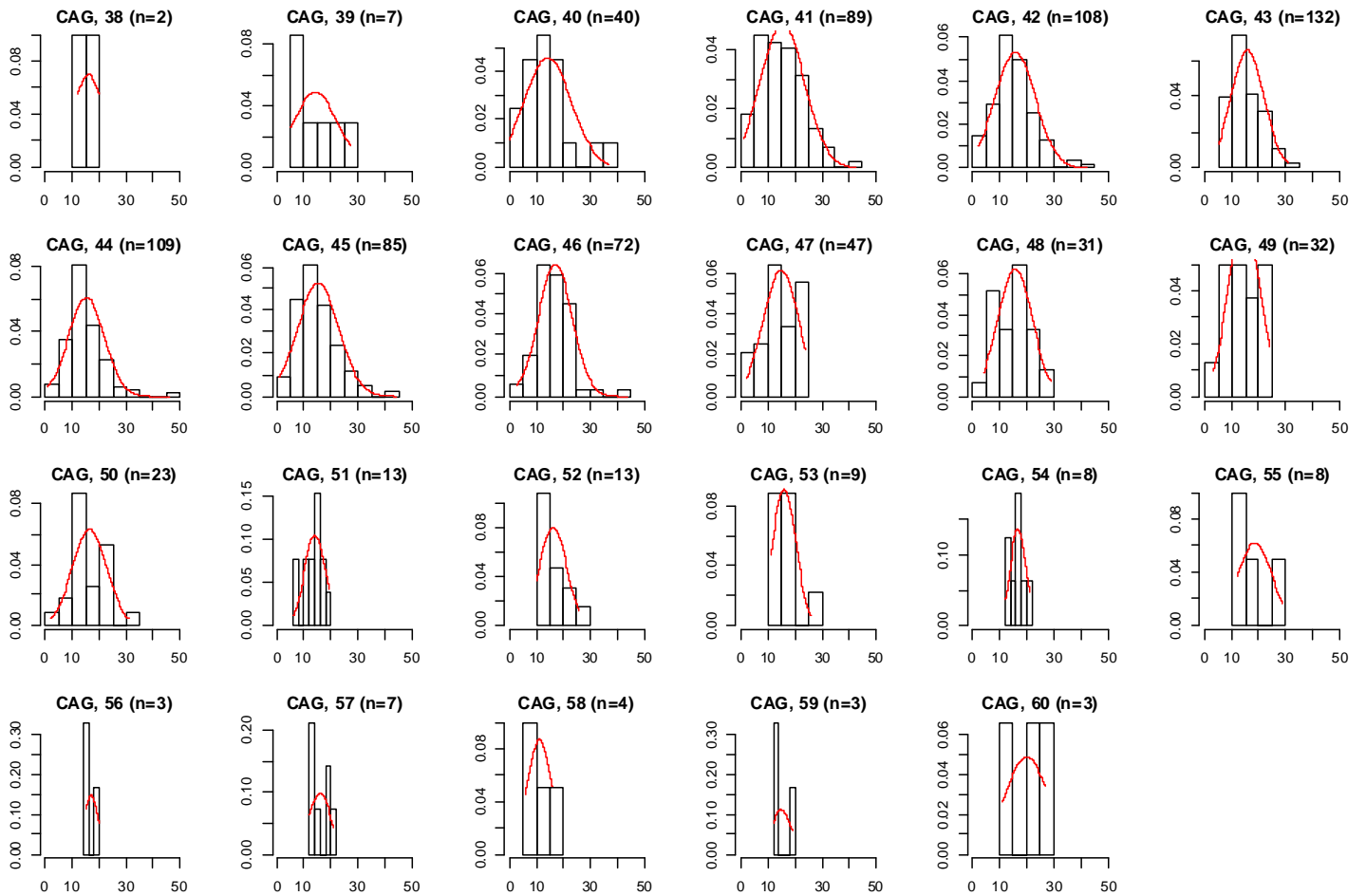


1 **Figure S8. Evaluation of normality of duraton data.**

2 Data normality for each CAG repeat was evaluated by comparing the observed distribution of duration values  
3 for a given expanded CAG repeat length to a theoretical normal distribution based on the mean and standard  
4 deviation of data (red line). The expanded CAG repeat length and sample size are indicated at the top of each  
5 plot.

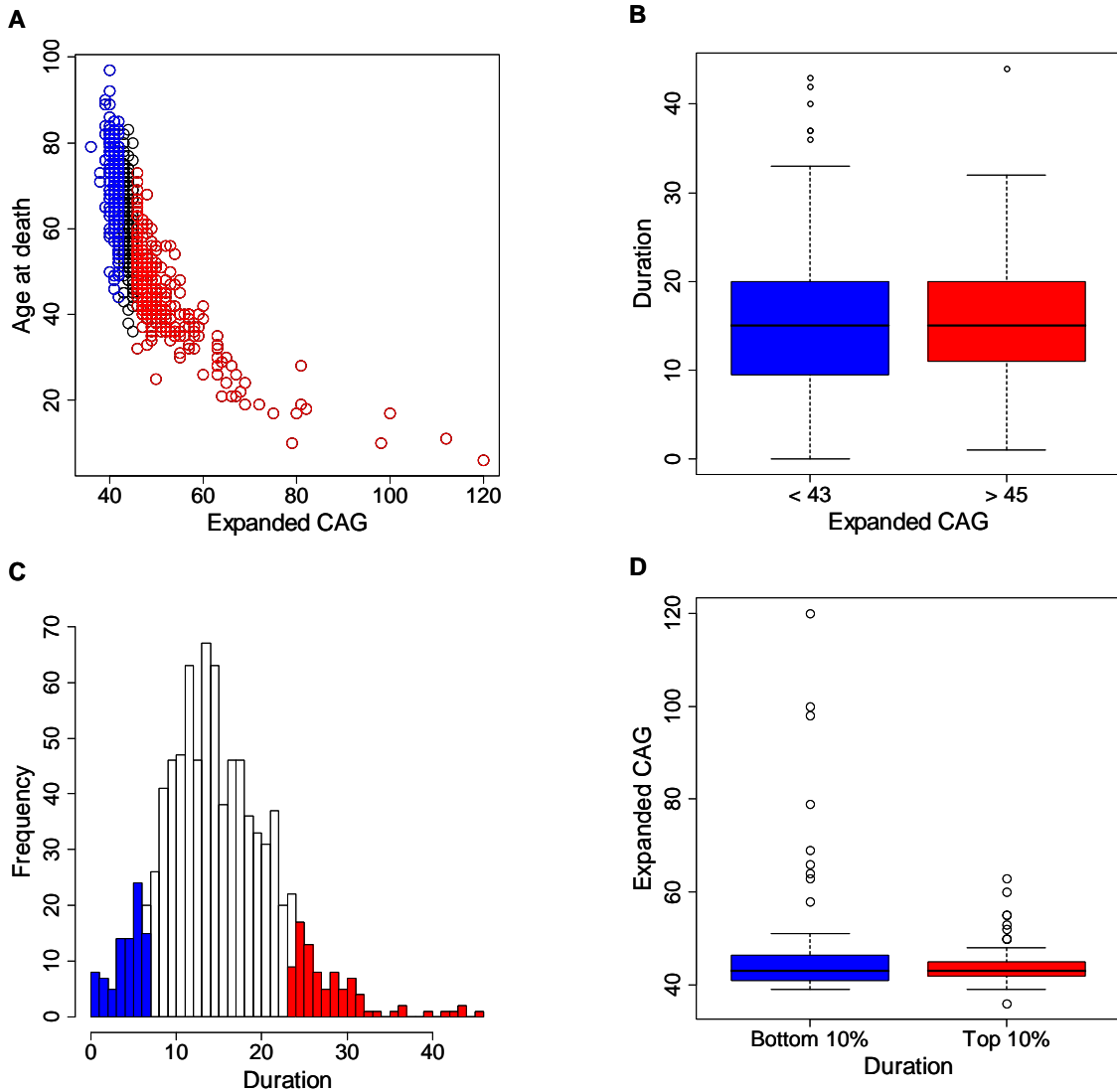
6

7



1 **Figure S9. HD disease duration is independent of *HTT* expanded CAG repeat length.**

2 A) To test whether HD subjects with smaller expanded CAG repeats had different duration values from those  
3 with larger expanded CAG repeats, disease duration was compared for 247 HD subjects with expanded CAG  
4 < 43 (blue circles) and 305 HD subjects with expanded CAG > 45 (red circles). B) Distributions of disease  
5 duration for the individuals in Panel A are summarized. A Mann-Whitney U test revealed no significant  
6 difference in disease duration between the two groups (p-value, 0.484). In addition, duration values between  
7 different CAG bins such as CAG < 44 vs. CAG > 44 or CAG < 42 vs. CAG > 46 were not significantly different  
8 (p-value, 0.96 and 0.77, respectively). C) To test whether expanded CAG repeat lengths of HD subjects in the  
9 top or bottom 10% extremes of disease duration differed, the 87 HD subjects in each group were identified.  
10 Blue and red bars represent HD subjects with the shortest and longest disease duration, respectively. D)  
11 Distributions of expanded CAG repeats in the individuals from Panel C are summarized. A Mann-Whitney U  
12 test revealed no significant difference in CAG repeat length between the two groups (p-value, 0.897).



1 **Figure S10. Simulation analysis.**

2 Various statistical analyses consistently supported that CAG repeat length does not influence disease duration  
3 in typical adult onset HD subjects. Simulation analysis was performed in order to evaluate the pattern of  
4 relationship between CAG length and duration that would have been observed if CAG repeat length had a  
5 significant impact on duration. Duration values of 855 HD subjects (for more information refer to the legend of  
6 Figures S7) were randomly permuted to generate simulated data, in which the size of expanded CAG explains  
7 pre-specified amounts of variation in duration (B-F). A) Mean values of observed duration were plotted against  
8 CAG repeat sizes. Expanded CAG repeats explained 0.045% of variance of duration in observed data.  
9 Data permutation was performed until pre-specified regression model's R square value was achieved (B, 20%;  
10 C, 10%; D, 5%; E, 2%; F, 1%), and then the mean of permuted duration values for a given CAG length was  
11 plotted by CAG length. Representative plots are shown. Each open circle represent mean of duration values  
12 for a given CAG length.

13

