The American Journal of Human Genetics

Supplemental Data

The *HTT* CAG-Expansion Mutation Determines Age at Death but Not Disease Duration in Huntington Disease

Jae Whan Keum, Aram Shin, Tammy Gillis, Jayalakshmi Srinidhi Mysore, Kawther Abu Elneel, Diane Lucente, Tiffany Hadzi, Peter Holmans, Lesley Jones, Michael Orth, Seung Kwak, Marcy E. MacDonald, James F. Gusella, and Jong-Min Lee

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

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

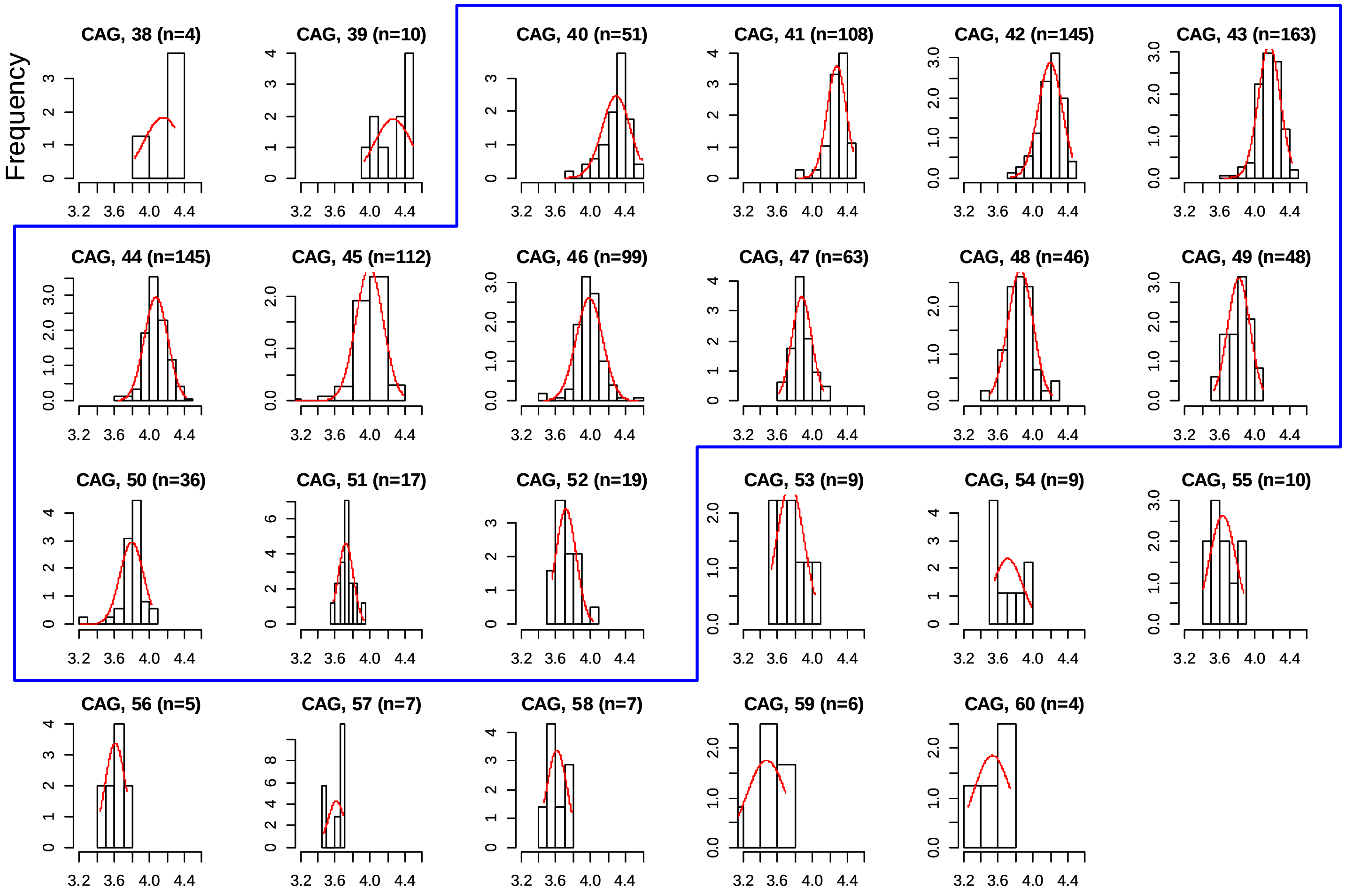


Figure S2. Variance and outliers in age at death data.

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

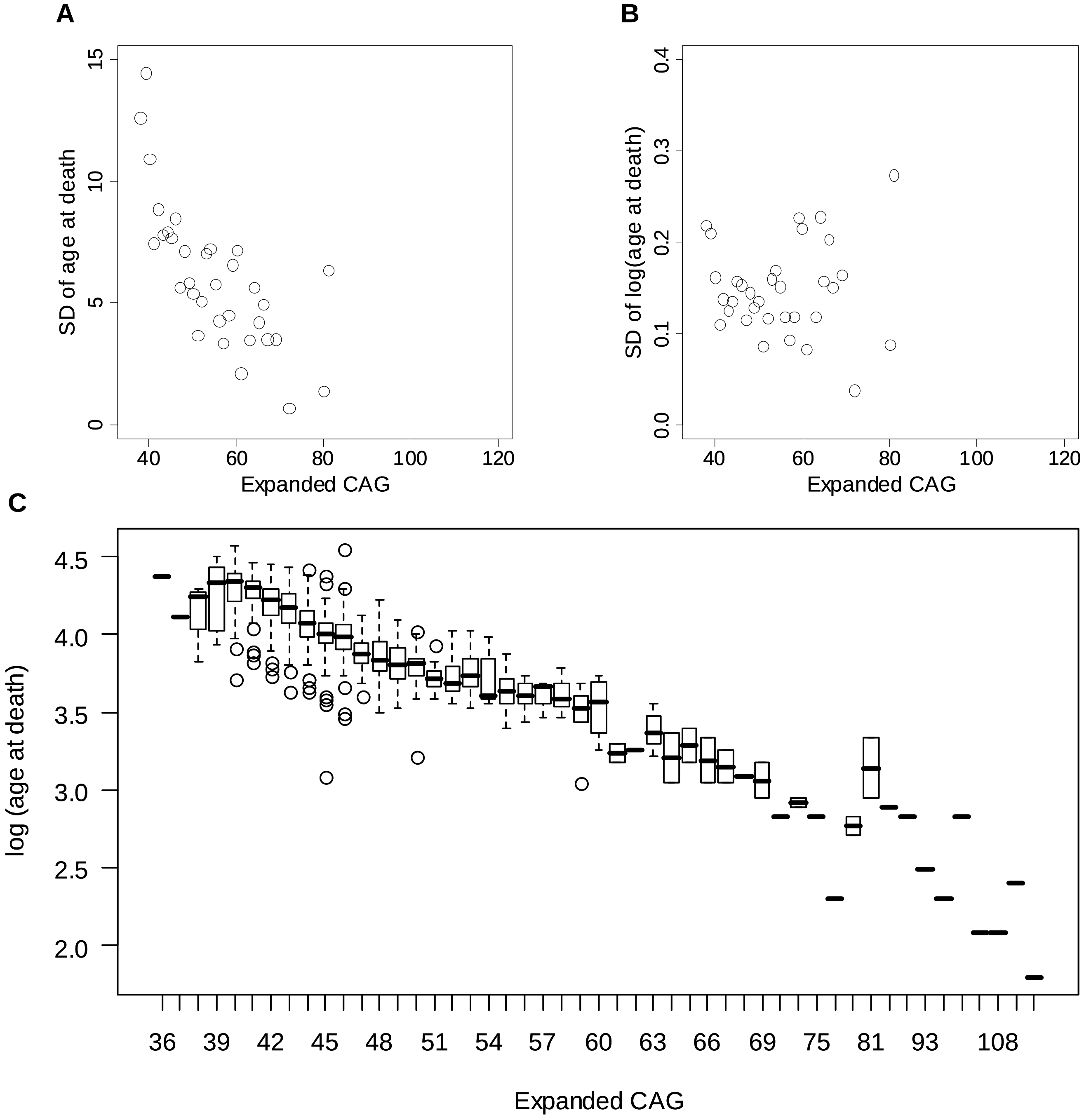


Figure S3. Model diagnostic plots.

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

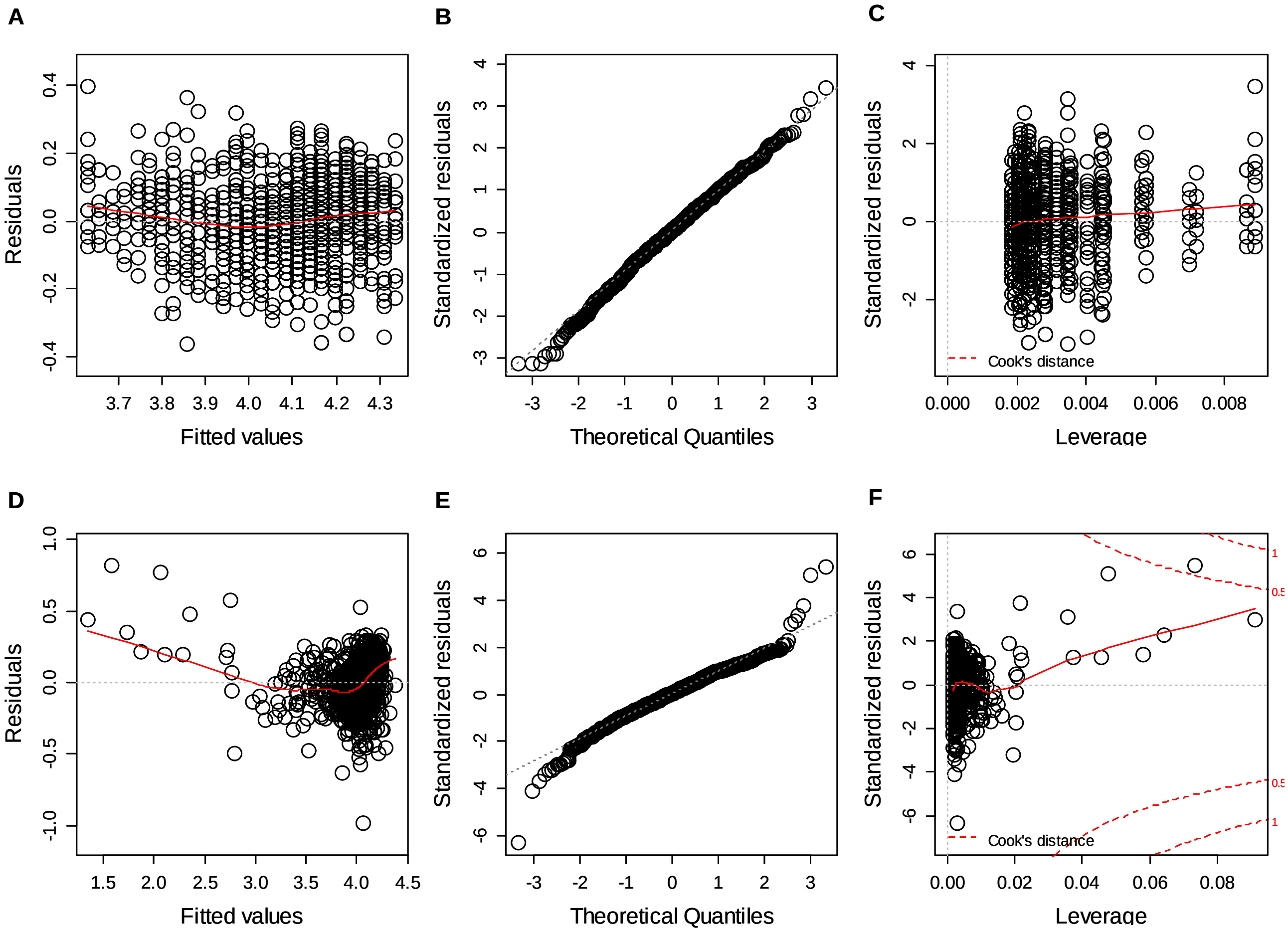


Figure S4. Normal CAG repeat does not explain age at death.

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

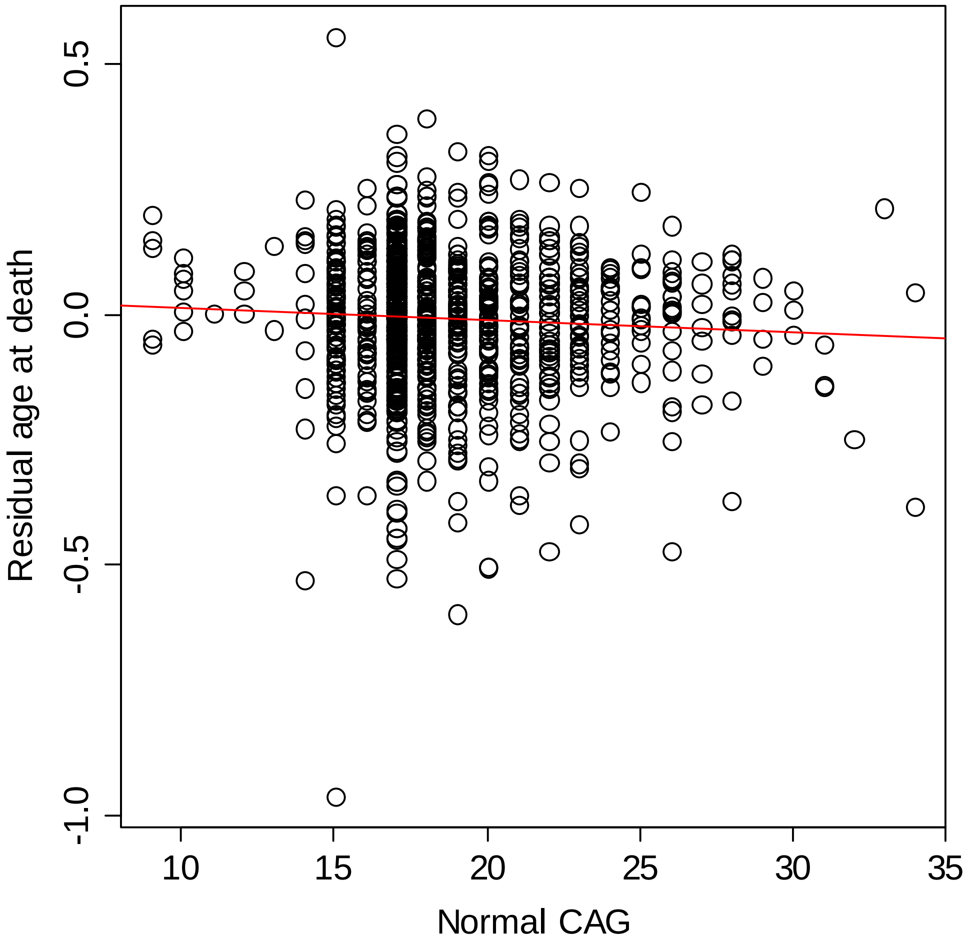


Figure S5. Extreme age at death samples do not differ in normal CAG repeat lengths.

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

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

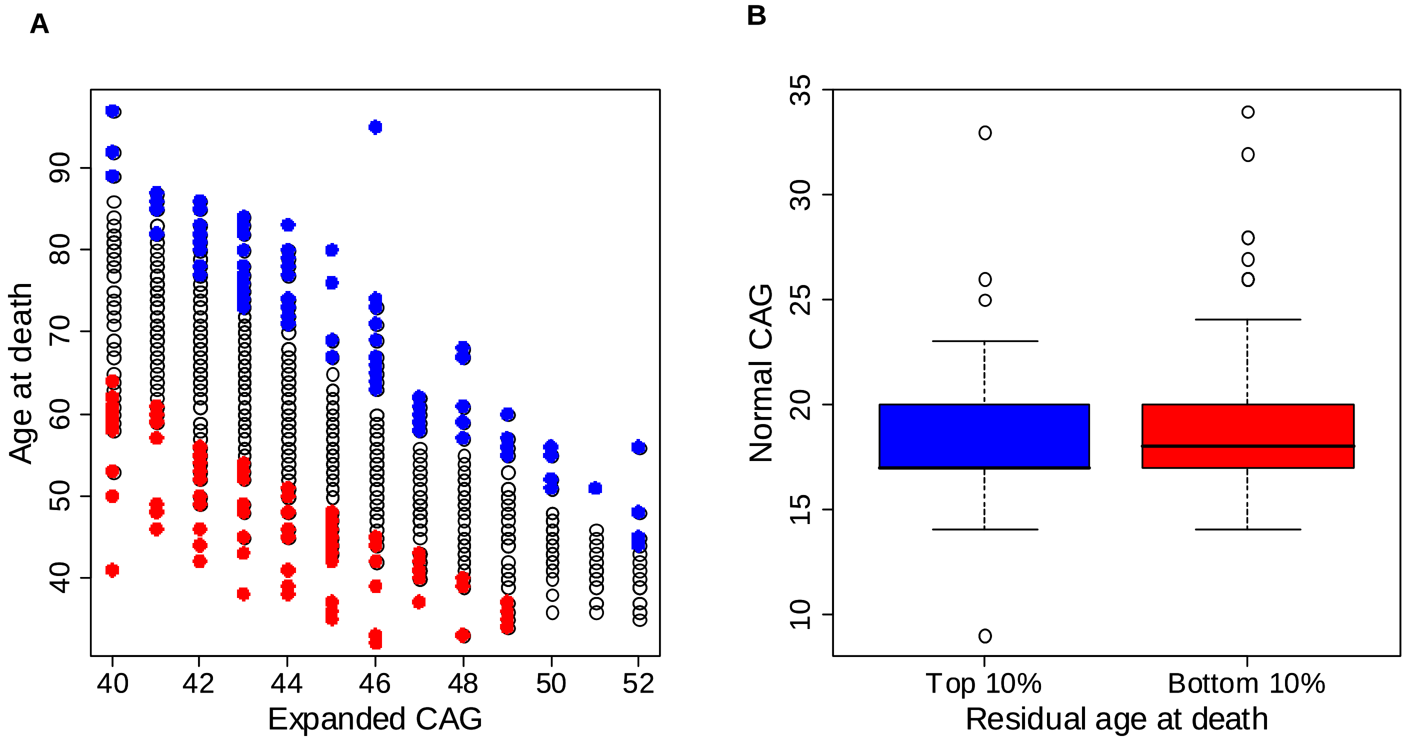
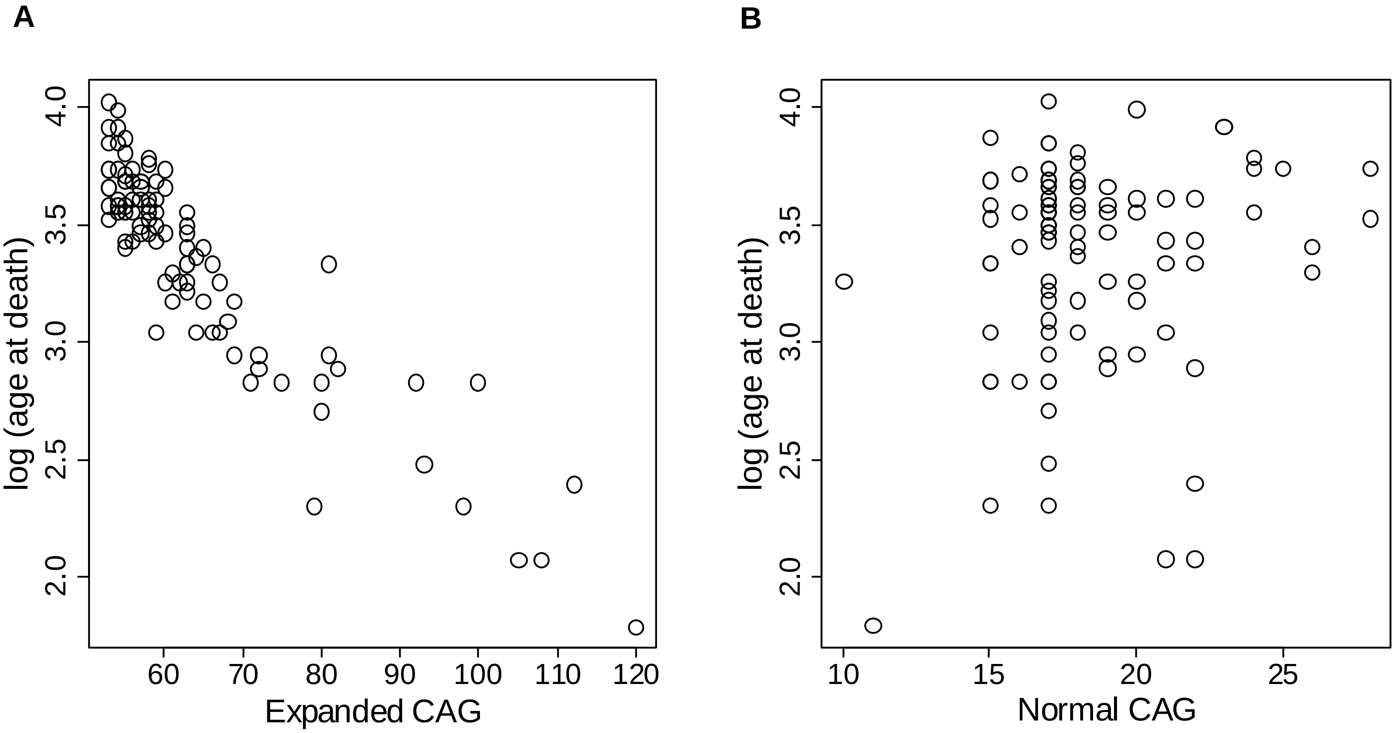


Figure S6. Age at death is determined by expanded CAG repeat length in a fully dominant fashion in samples with expanded CAG > 52.

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^2
CAG > 52	97	<2e-16	0.941	81.65%

Figure S7. Non-normal distribution of duration.

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

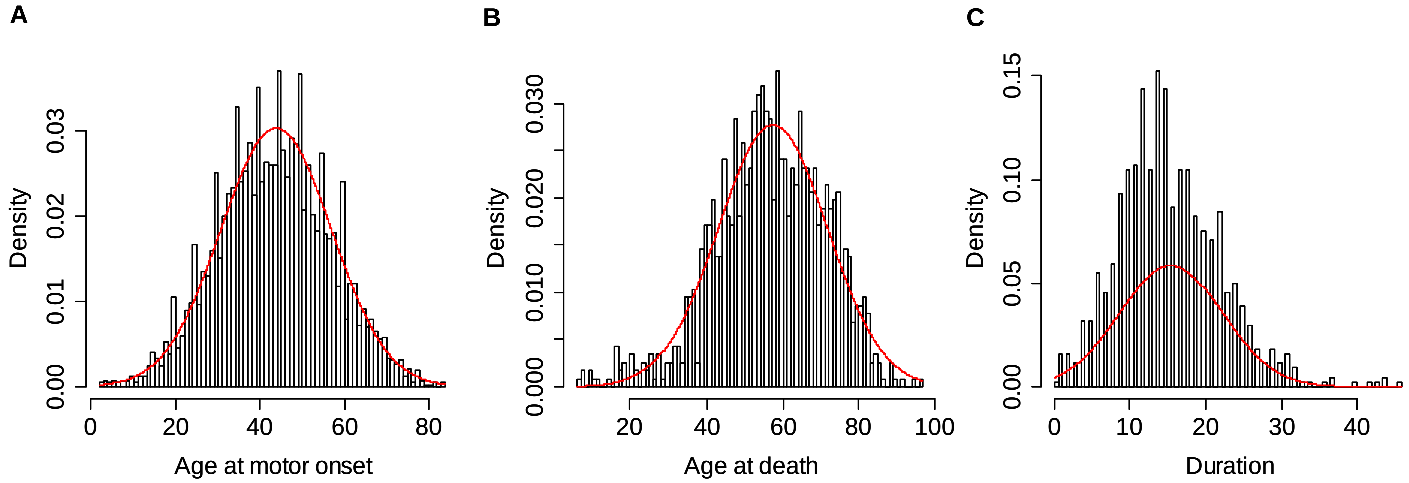


Figure S8. Evaluation of normality of duraton data.

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

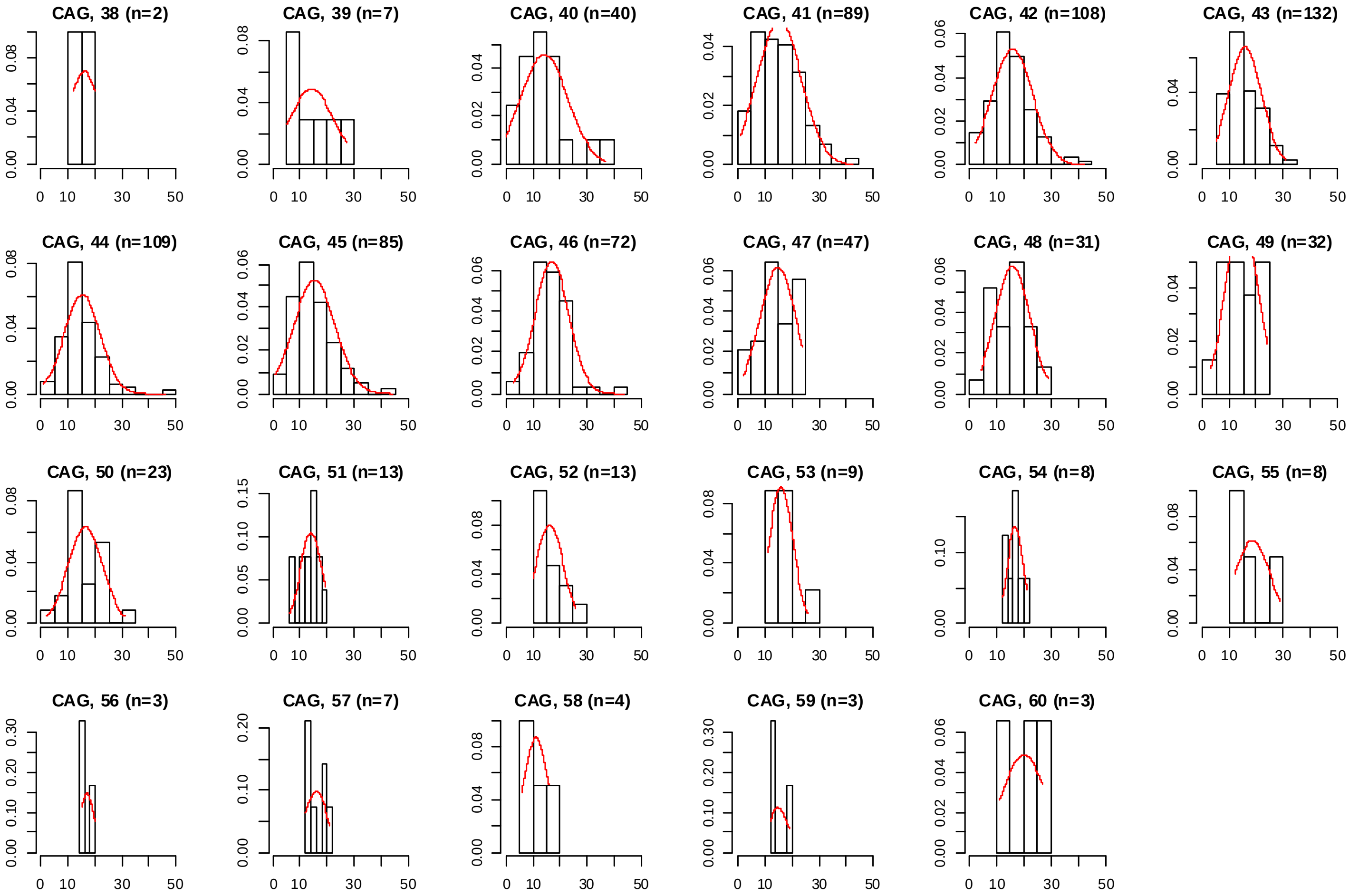


Figure S9. HD disease duration is independent of *HTT* expanded CAG repeat length.

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

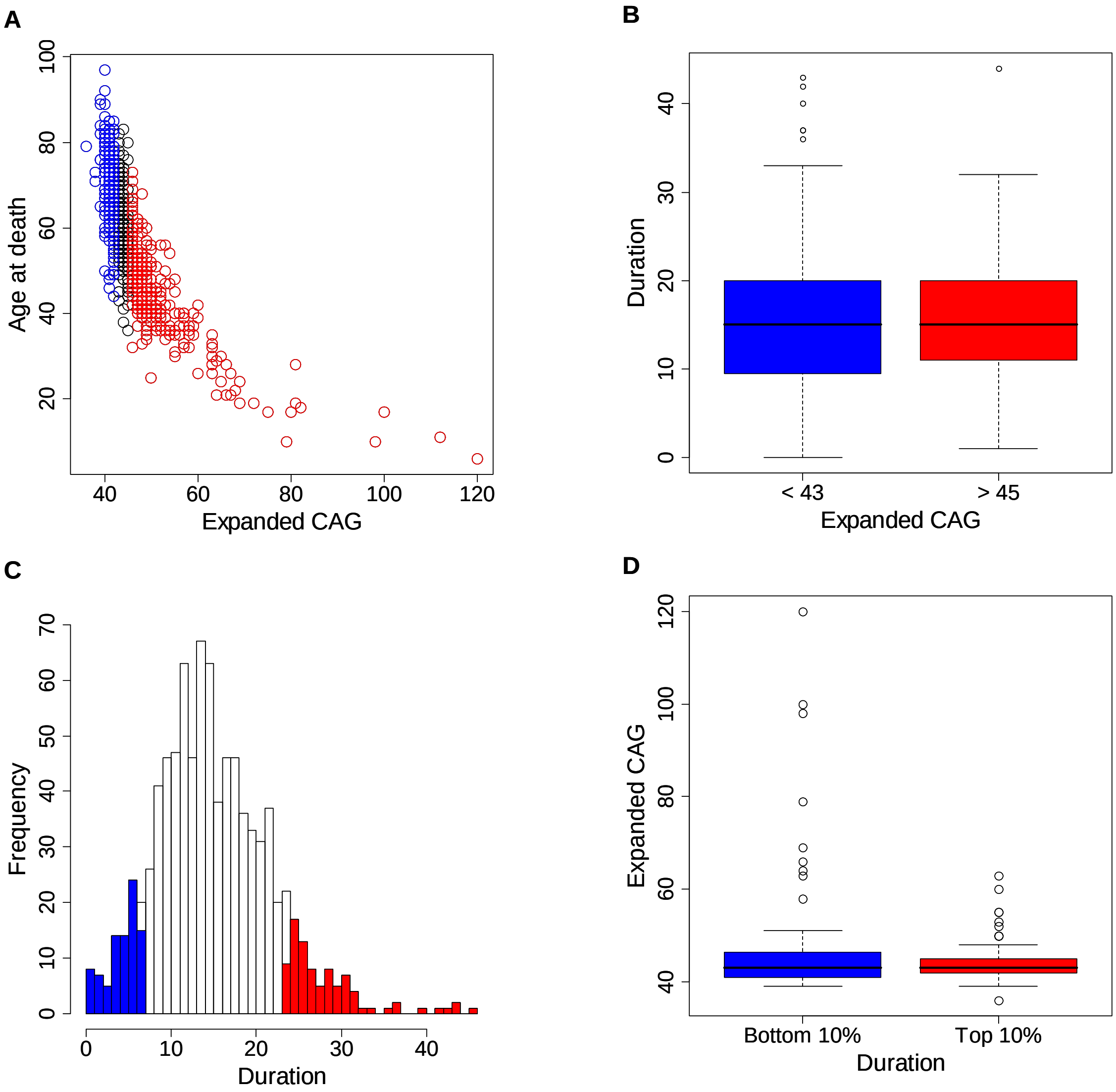
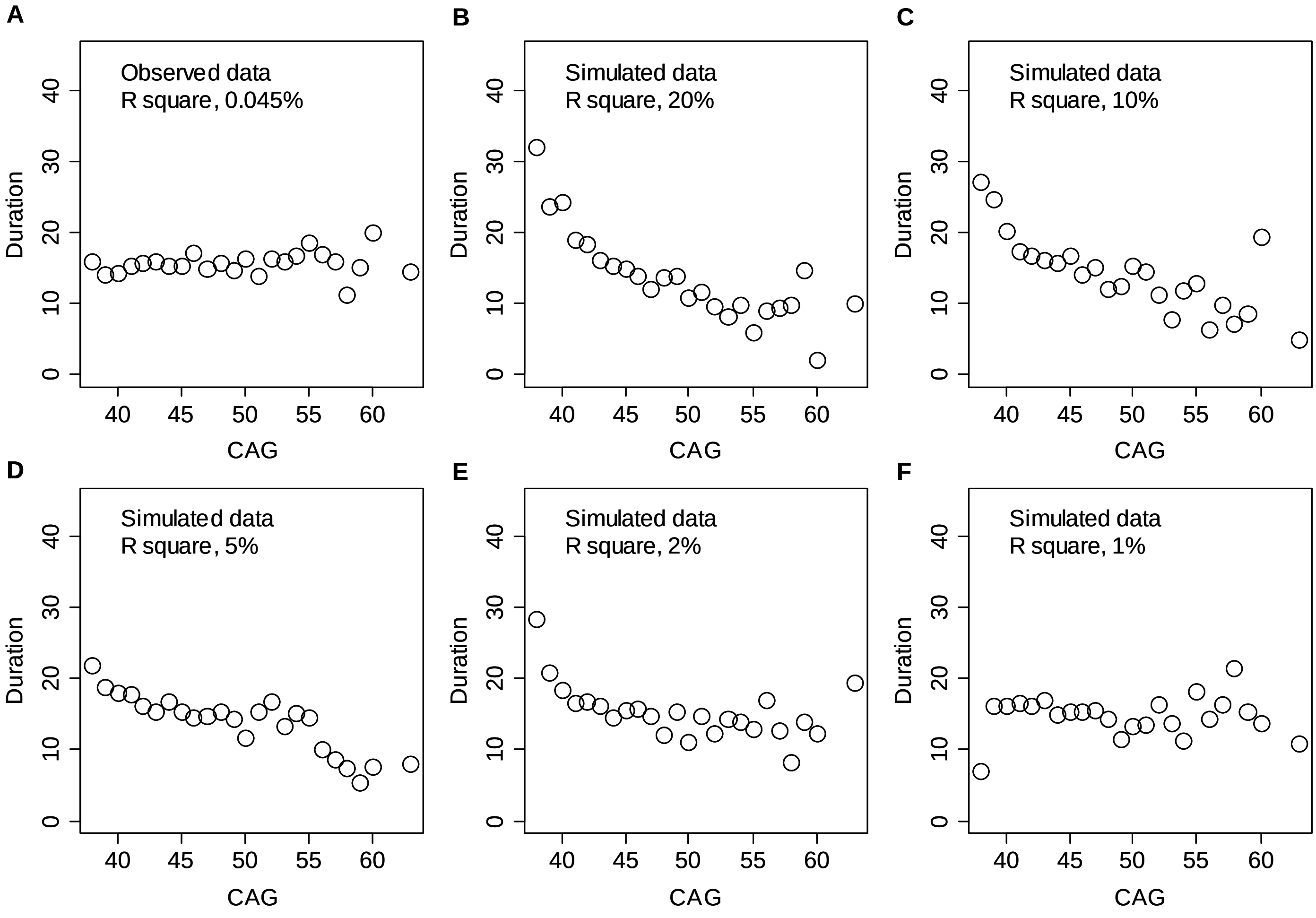


Figure S10. Simulation analysis.

Various statistical analyses consistently supported that CAG repeat length does not influence disease duration in typical adult onset HD subjects. Simulation analysis was performed in order to evaluate the pattern of relationship between CAG length and duration that would have been observed if CAG repeat length had a significant impact on duration. Duration values of 855 HD subjects (for more information refer to the legend of Figures S7) were randomly permuted to generate simulated data, in which the size of expanded CAG explains pre-specified amounts of variation in duration (B-F). A) Mean values of observed duration were plotted against CAG repeat sizes. Expanded CAG repeats explained 0.045% of variance of duration in observed data.

Data permutation was performed until pre-specified regression model's R square value was achieved (B, 20%; C, 10%; D, 5%; E, 2%; F, 1%), and then the mean of permuted duration values for a given CAG length was plotted by CAG length. Representative plots are shown. Each open circle represent mean of duration values for a given CAG length.


A
Observed data
R square, 0.045%
B
Simulated data
R square, 20%
C
Simulated data
R square, 10%
D
Simulated data
R square, 5%
E
Simulated data
R square, 2%
F
Simulated data
R square, 1%
Duration
CAG