Figure e-1

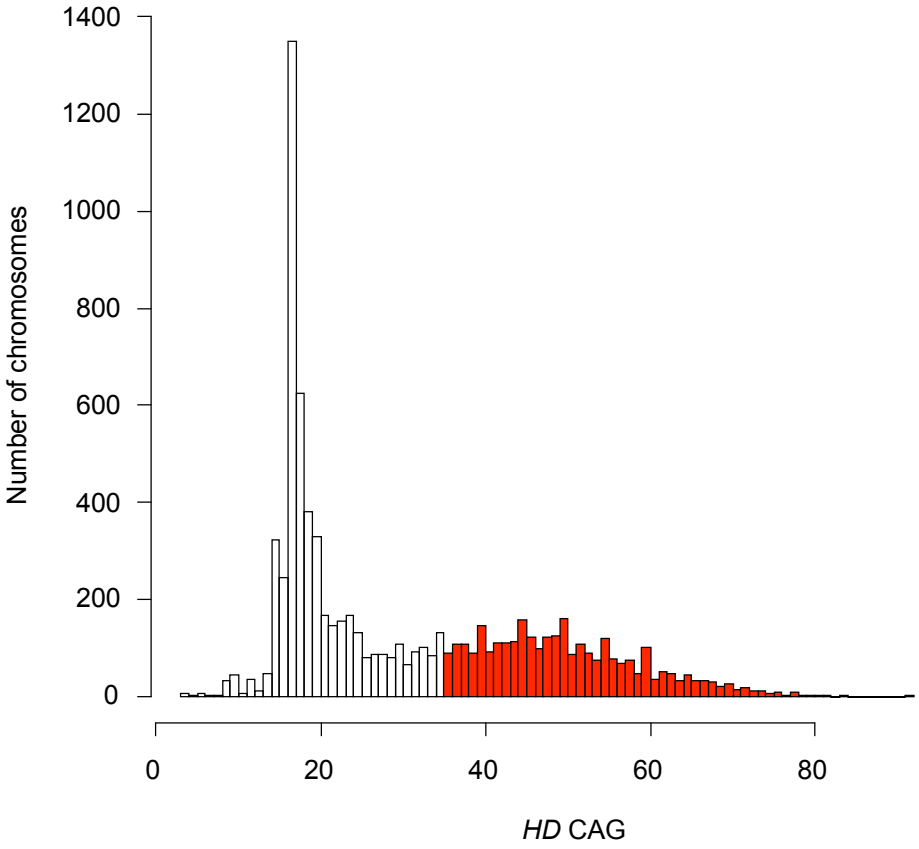

Figure e-1. Distributions of expanded (mutant) and normal *HD* CAG repeat lengths in subjects used in this study.

Frequencies of chromosomes for each CAG repeat length of normal (open) and mutant allele (expanded; red) were plotted.

Figure e-2

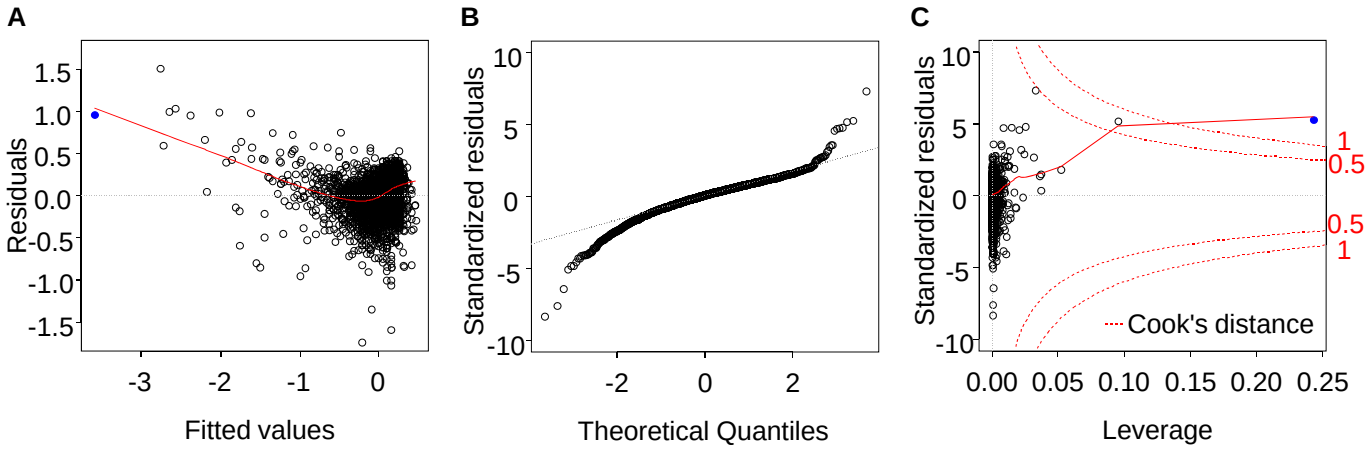

Figure e-2. The initial model violates statistical assumptions of linear regression

The traditional linear regression analysis used often in the field yielded a model that violated statistical assumptions of constant variance and normally distributed error, as demonstrated in panels A-C.

(A) To test the assumption of equal variance, we compared residuals calculated from the model to fitted values (predicted centered log onset age).

(B) To assess the normality of the initial model, we compared actual residuals to theoretical residuals from a normal distribution in a quantile-quantile plot.

(C) To identify influential data points in the initial model, leverage and the Cook's distance were plotted against the residuals. Leverage is commonly used to identify observations that have a large 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 model fit by measuring the effect of deleting a given observation.

Blue circle in Panels A and C represents the HD subject with an expanded *HD* repeat length of 120 CAG trinucleotides.

Figure e-3

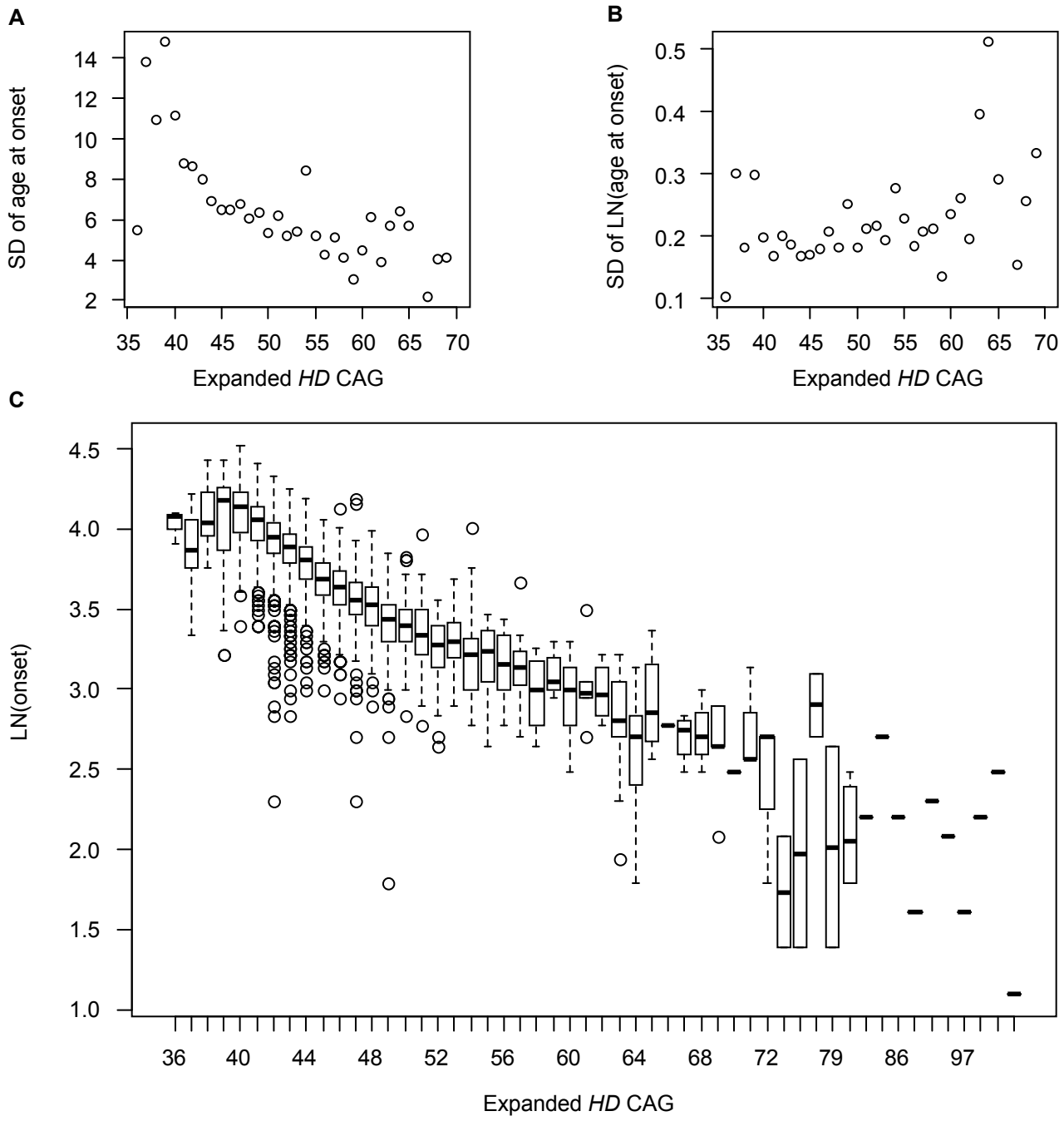

Figure e-3. Variance and normality of age at motor onset distribution for HD subjects.
The problems of non-constant variance and non-normal error were evident in the observations of age at onset at individual expanded allele CAG repeat lengths.
(A) Variance of age at onset, represented by standard deviation (SD) was not constant.
(B). Natural log transformation of the age at onset (LN(onset)) provided a partial remedy for the range of adult-onset associated repeat lengths that form the bulk of the sample.
(C) A Box plot of natural log-transformed age at onset against expanded allele showed the existence of outliers (open circles) defined by a standard quartile method (outside of 1.5 times interquartile range from first or third quartile).

Figure e-4

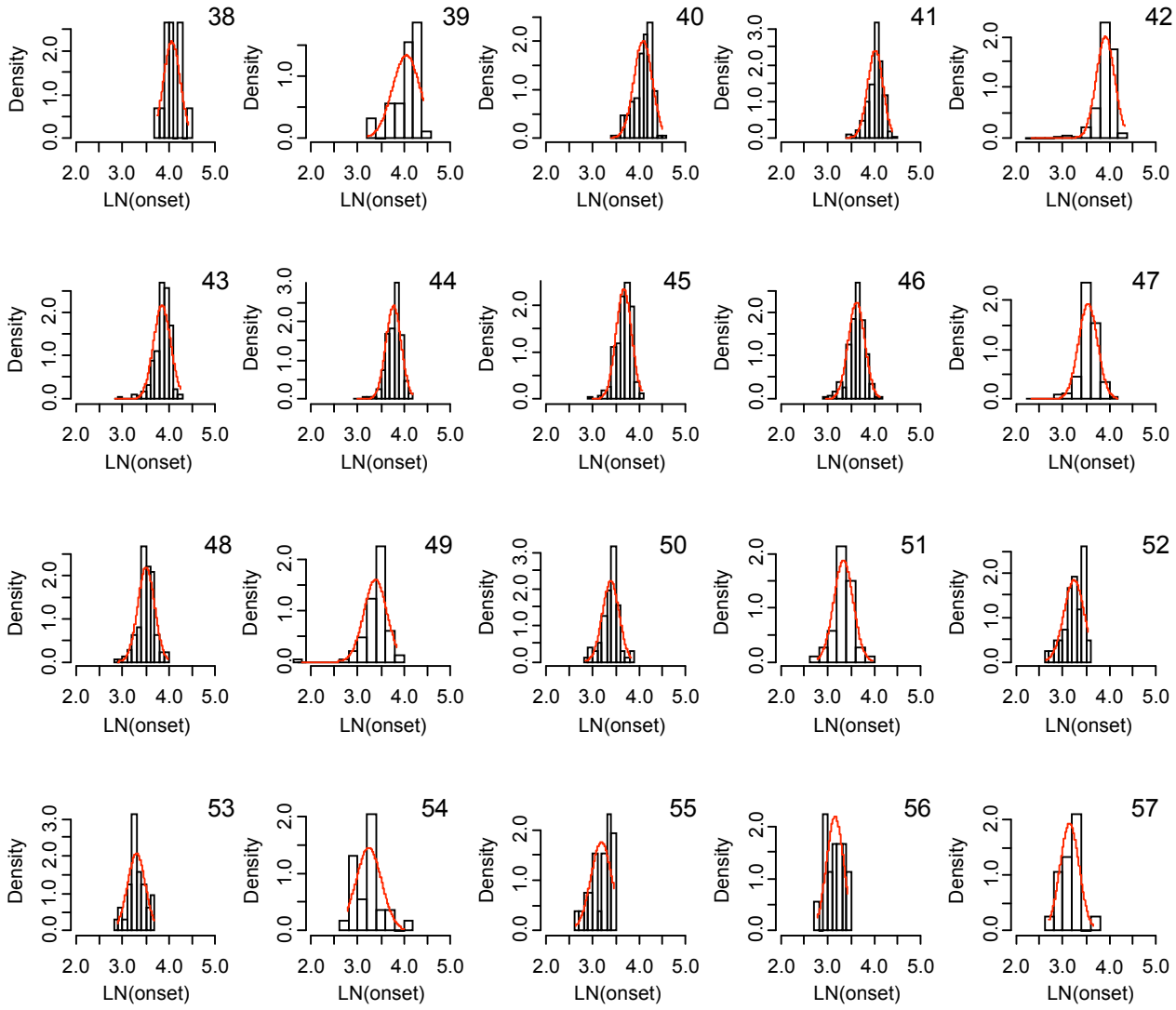

Figure e-4. Distributions of age at onset by individual expanded CAG repeat allele.
Exclusion of the outliers identified statistically in Figure e-3 resulted in distributions of log-transformed age at onset that resembled the normal distribution for adult onset-associated CAG repeat lengths: histograms of log-transformed age at onset for each expanded allele CAG repeat length were overlaid by a theoretical normal distribution (red line). Numbers represent expanded allele CAG repeat length. Number at top right represents expanded allele CAG repeat length. LN(onset), natural log transformed age at onset.

Figure e-5

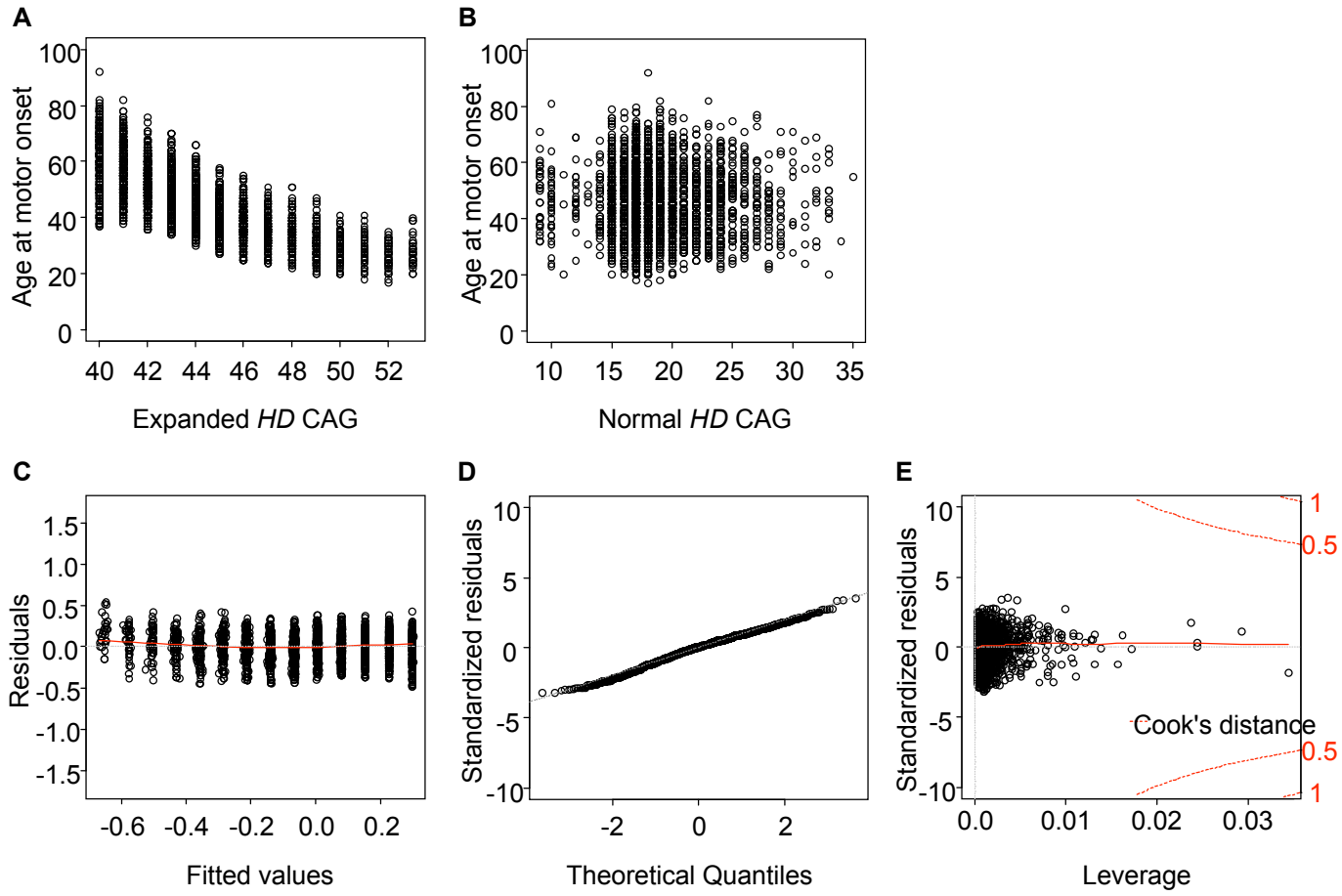

Figure e-5. An updated model with a well-behaved subsample of the dataset

To update the model to conform with the assumptions for linear regression, we excluded data points contributing to non-normal error or non-constant variance. We used 3,674 subjects with expanded allele CAG repeat lengths between 40 and 53, after excluding 198 and 70 subjects with higher or lower repeats, respectively and 126 outliers. The final well-behaved subset comprised 90.3% of the original sample.

(A) The relationship between expanded allele CAG repeat length and age at onset of motor signs.

(B) The relationship between normal allele CAG repeat length and age at onset of motor signs.

(C) Residuals calculated from the updated model compared to fitted values.

(D) Comparison of actual residuals to theoretical residuals from a normal distribution in a quantile-quantile plot.

(E) Leverage and the Cook's distance plotted against the residuals.

Figure e-6

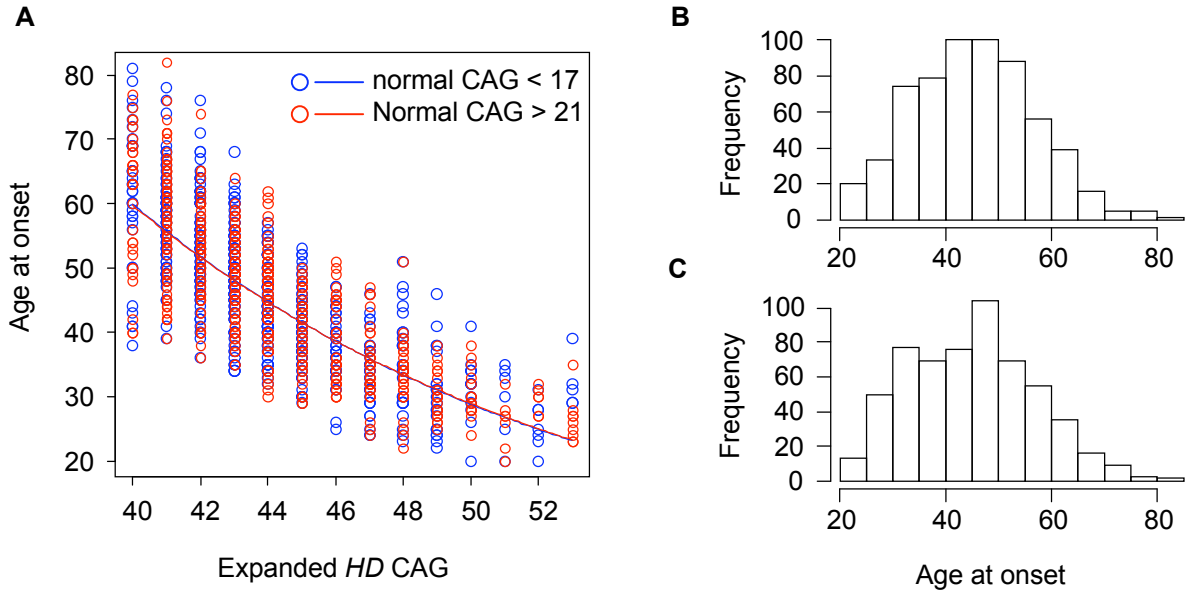

Figure e-6. Comparisons between subjects with shorter or longer normal allele CAG repeat lengths.

(A). Two linear regression models were generated using subjects with shorter normal allele CAG repeat length (normal allele CAG < 17, blue line and blue circle; 616 subjects) and subjects with longer normal allele CAG repeat length (normal allele CAG > 21, red line and red circle; 576 subjects). Log-transformed age at onset of normally distributed data points was modeled by expanded allele repeat length. Expanded allele CAG repeat lengths between the two groups were not significantly different (Mann-Whitney *U* test, p value, 0.7336).

Distributions of age at onset of subjects with shorter normal allele CAG repeat length (B) and those with longer normal allele CAG repeat length (C) were plotted. The distributions of age at onset were not significantly different between the two groups (Mann-Whitney U test, p value, 0.5488).

Table e-1. Statistical analyses of samples excluded from the updated model as phenotypic or repeat length outliers.

Model[$] (CAG range)	Number of Subjects	Intercept (p-value)	Normal CAG (p-value)	R-squared
Model using phenotypic outliers (40-53)	126	-0.48792 (0.001)	-0.00151 (0.841)	0.00032
Model using subjects with CAG > 53 (54-99)	197	0.27252 (0.136)	0.00833 (0.401)	0.00361
Model using all subjects excluded in the updated model (36-99)	393	0.16350 (0.290)	-0.00800 (0.333)	0.00240

[$] Residuals were calculated based on the minimum adequate model: the updated model (without centering) from Figure e-5 using well-behaved data points (CAG, 40-53; 3674 subjects), and then, the residual was modeled as a function of length of the normal CAG allele.