HGGA, Volume *3*

Supplemental information

A probable *cis*-acting genetic modifier of Huntington disease frequent in individuals with African ancestry

Jessica Dawson, Fiona K. Baine-Savanhu, Marc Ciosi, Alastair Maxwell, Darren G. Monckton, and Amanda Krause

Supplemental Data:

Supplemental figures:

Supplemental tables:

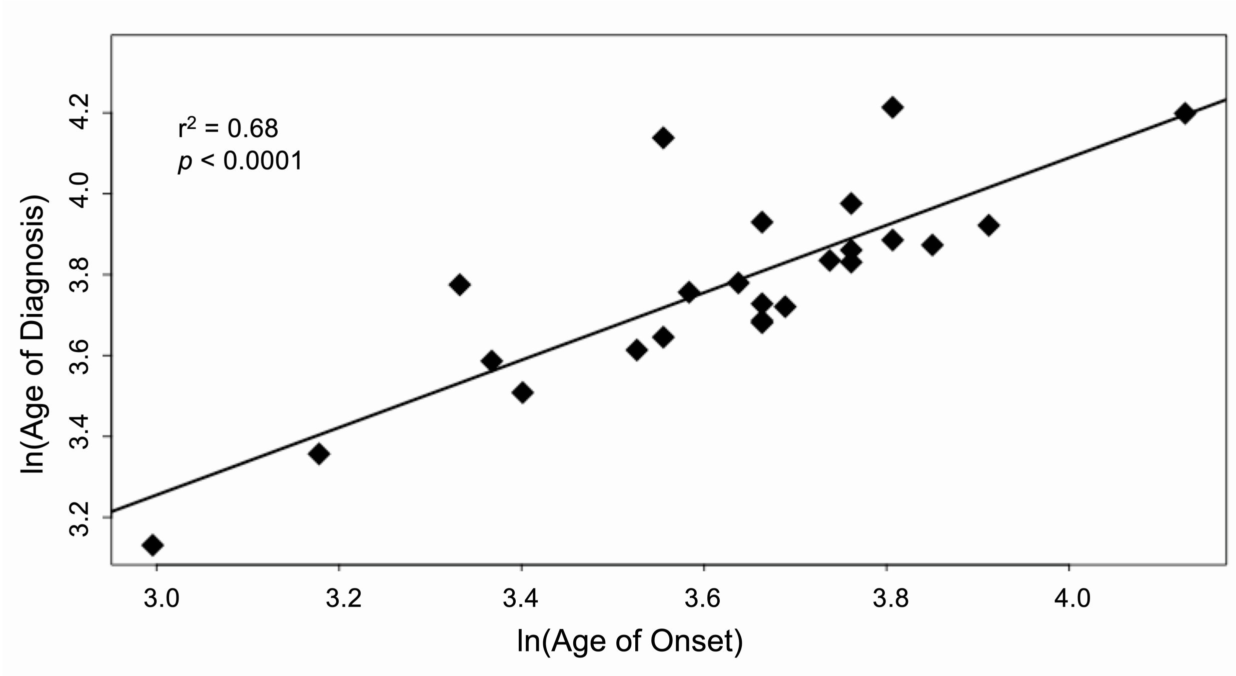

Figure S1. Linear regression analysis testing the association between the log transformed AoD and AoO.

When studying the HD phenotype, the AoO of motor symptoms is often used as it is the most well characterised measure of disease severity. However, in our individuals affected with HD, less than 50% had AoO information. The relationship between the natural log transformed AoD and AoO, for 24 of the 68 individuals affected with HD for whom AoO data was available was assessed. The R-square and *p*-values show a highly significant association ($r^2 = 0.68$, $p = 8 \times 10^{-7}$), indicating AoD can be used as an acceptable proxy for the AoO.

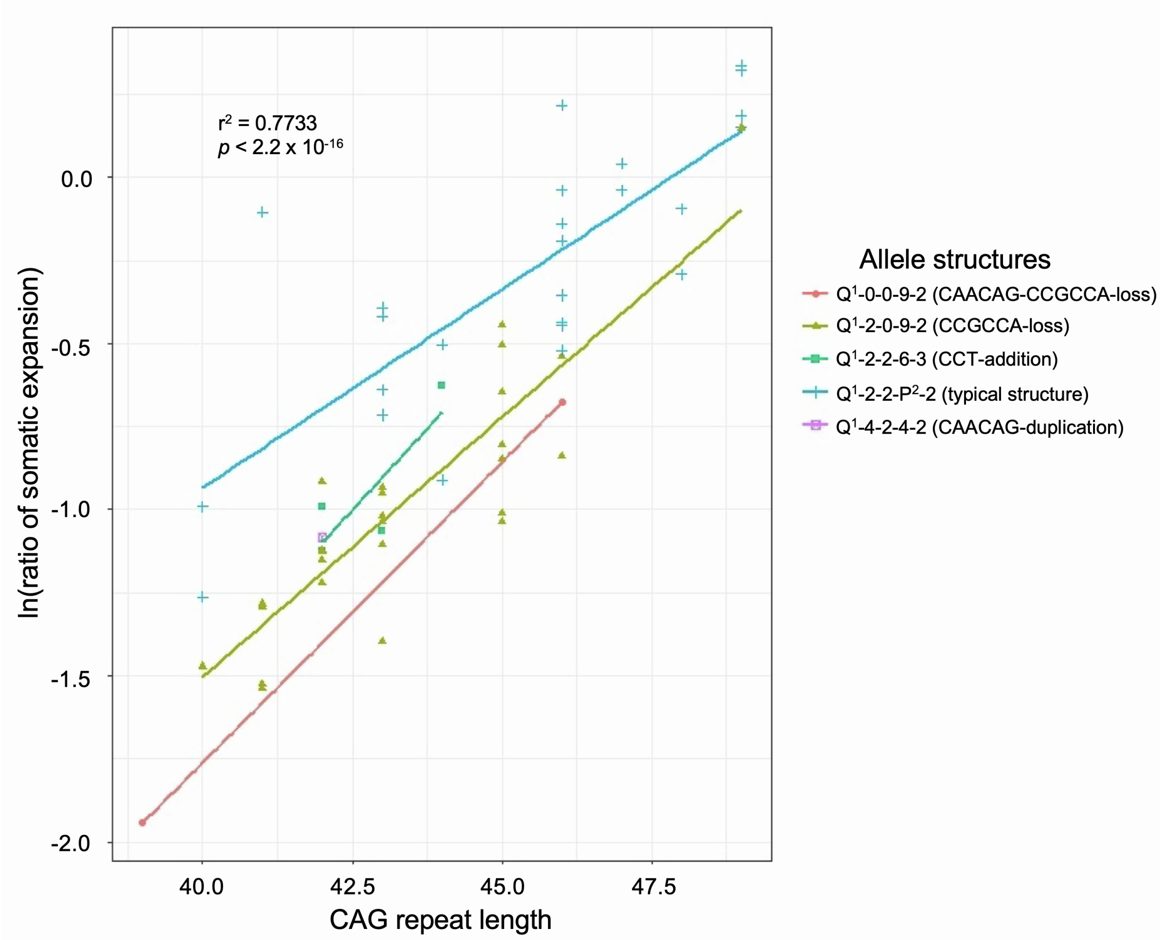

Figure S2. Linear regression analysis testing the association between the log transformed ratio of somatic expansion and inherited CAG repeat length for each disease allele structure.

The amount of somatic expansion of the CAG repeats was measured by counting the ratio of reads larger (N+1 to 10 repeats) than the progenitor CAG repeat (N). The R-square and *p*-values show a significant association ($r^2 = 0.77$, $p < 2 \times 10^{-16}$). The combined typical allele structures Q^1-2-2-P^2-2 have the highest relative ratio of somatic expansions and the lowest was present in the atypical allele structures, Q^1-2-0-9-2 and Q^1-0-0-9-2, both of which are characterised by a loss of the CCGCCA sequence (intervening proline). The Q^1-0-0-9-2 disease allele structure was excluded as it was present in two individuals.

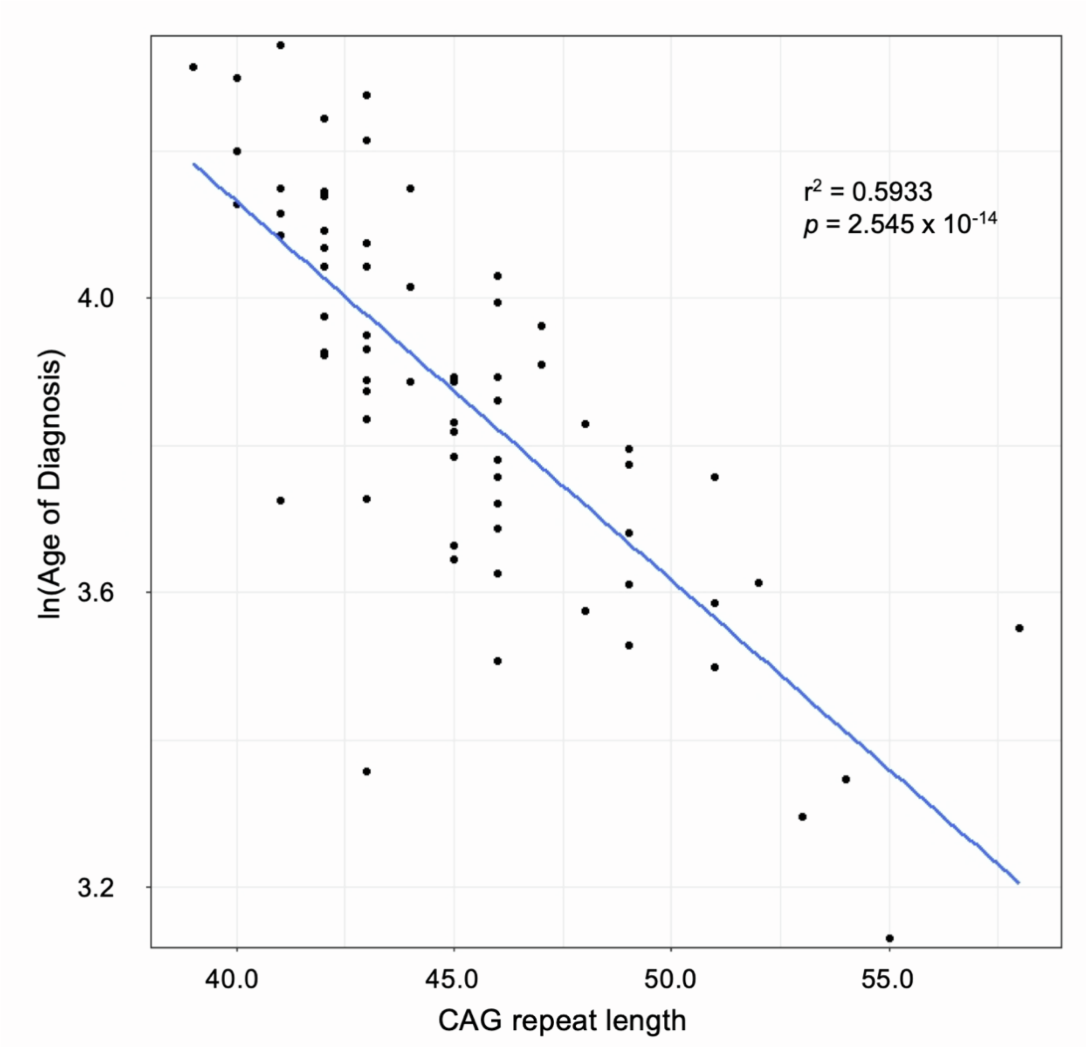

Figure S3. Linear regression analysis testing the association between the log transformed AoD and the disease associated inherited CAG repeat length.

The R-square and p-values show a significant association ($r^2 = 0.59$, $p = 2 \times 10^{-14}$), indicating that the CAG repeat length accounts for most of the variation in the HD phenotype. The CAG repeat length was the contiguous number of CAG repeats.

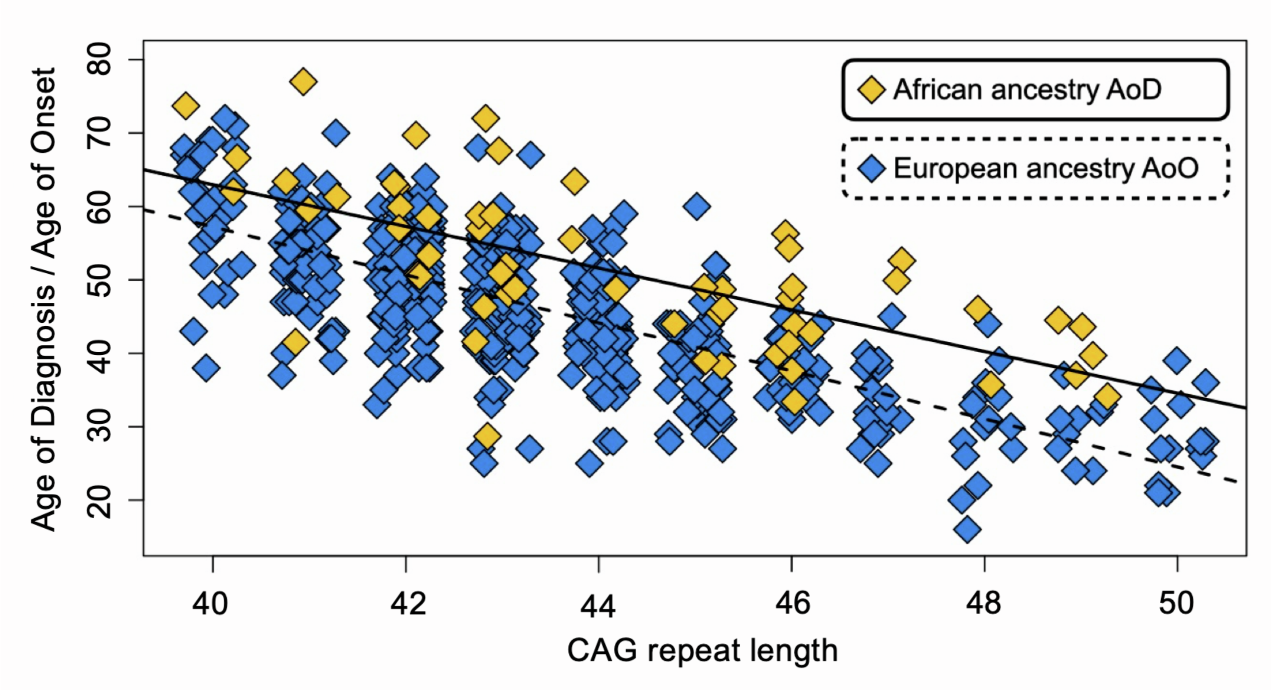

Figure S4. Comparison of the association between CAG repeat length and AoD in the African ancestry HD population, and AoO in a previously reported European ancestry HD population.

The African ancestry HD population are shown as mustard diamonds and the previously reported European ancestry HD population are shown as blue diamonds. Note that the lines of best fit for the two datasets run broadly parallel to each other with age at diagnosis in the African ancestry population (continuous line) shifted ~ 7 years later than age at onset in the European ancestry population (dashed line).

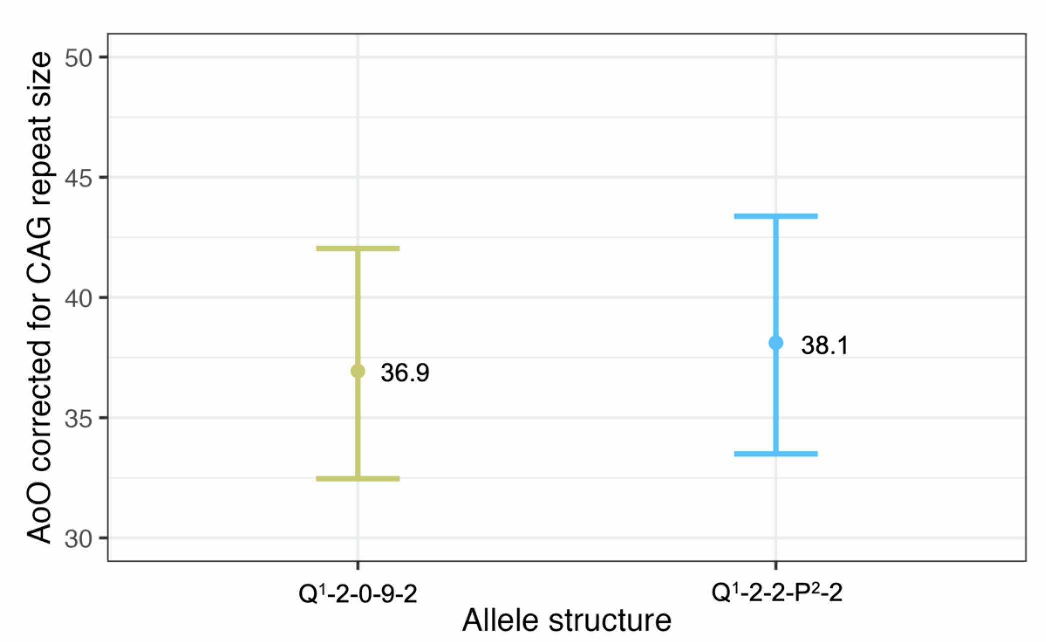

Figure S5. Estimated marginal mean of the AoO (in years) for the disease allele structures, corrected for CAG repeat size.

The earliest mean AoO was identified for the Q^1-2-0-9-2 allele structure. The estimated marginal mean AoO for the allele structures were as follows; Q^1-2-0-9-2: 36.9years (N = 12, 95% CI = 32.5 to 42.0) and Q^1-2-2-P^2-2: 38.1 years (N = 12, 95% CI = 33.5 to 43.4).

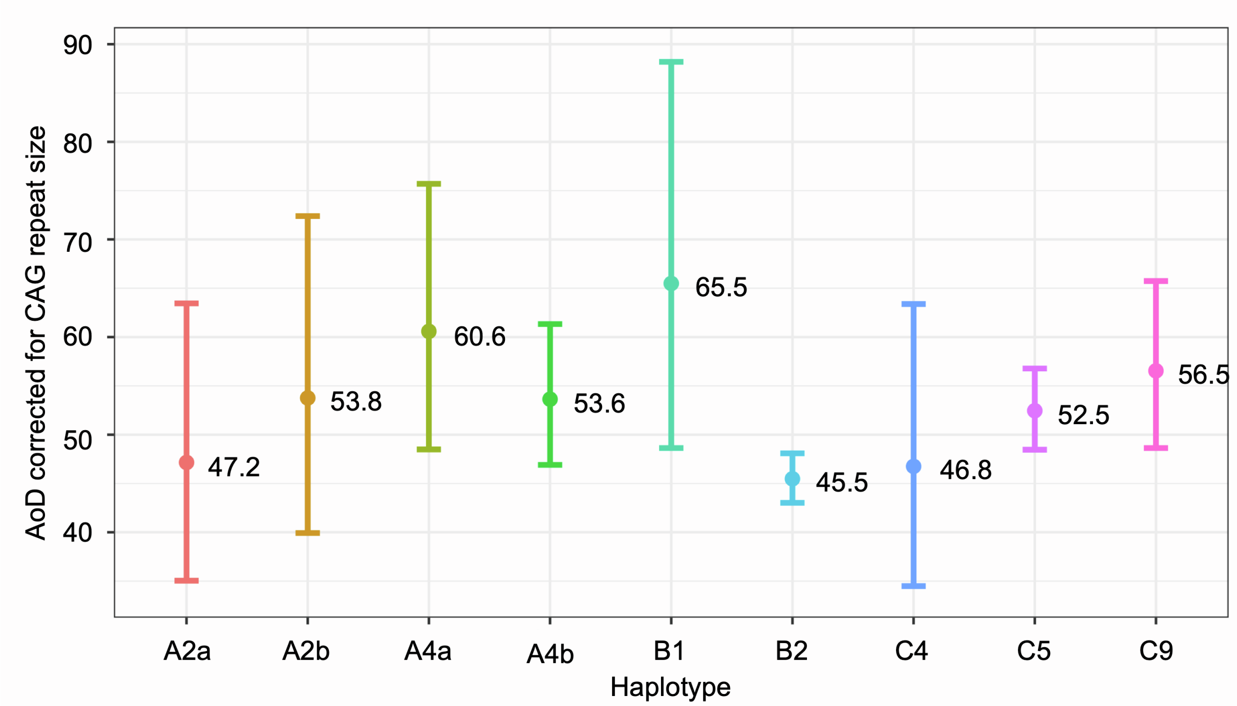

Figure S6. Estimated marginal mean of the AoD (in years) for each disease-associated haplotype.

The earliest mean AoD was identified for haplotype B2. The estimated marginal mean AoD for the haplotypes were as follows; B2: 45.5 years (N = 29, 95% CI = 43.0 to 48.1), C4: 46.8 years (N = 1, 95% CI = 34.5 to 63.4), A2a: 47.2 years (N = 1, 95% CI = 35.0 to 63.4), C5: 52.5 years (N = 19, 95% CI = 48.5-56.8), A4b: 53.6 years (N = 5, 95% CI = 46.9 to 61.3), A2b: 53.8 years (N = 1, 95% CI = 39.9 to 72.4), C9: 56.5 years (N = 4, 95% CI = 48.6 to 65.7), A4a: 60.6 years (N = 3, 95% CI = 48.5 to 75.7) and B1: 65.5 years (N = 1, 95% CI = 48.6 to 88.2).

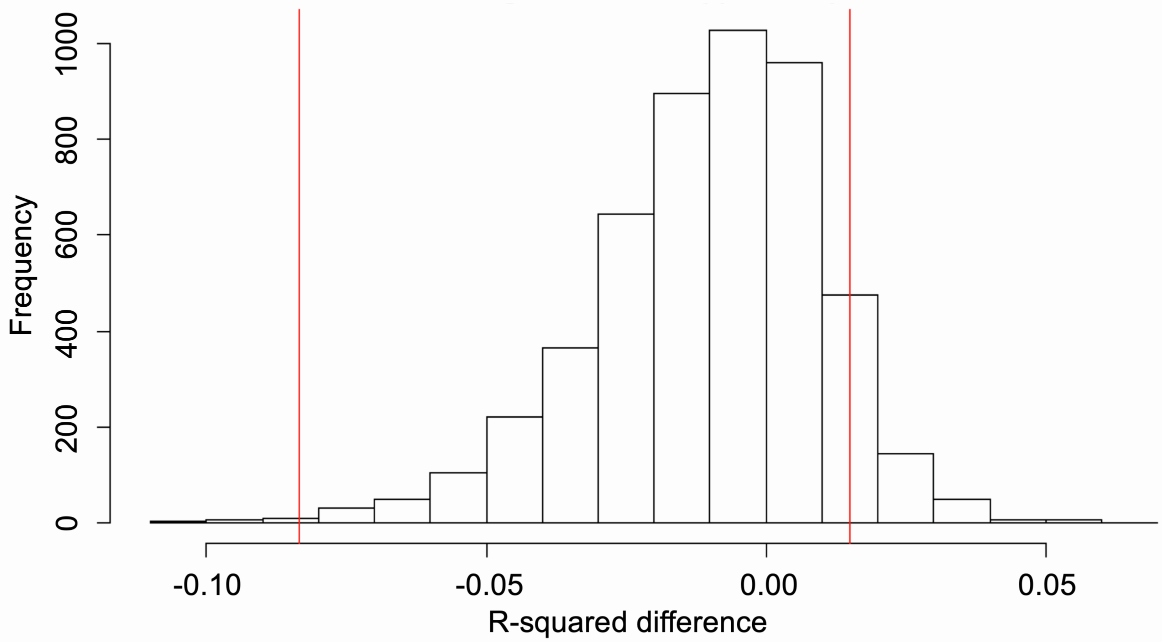

Figure S7. The frequency distribution of the R-squared difference between the Q[1]-2-0-9-2 allele structure and haplotype B2 models.

The goodness of fit test was conducted on 5,000 bootstrapped samples in R (v3.4.3). The effect on the AoD could not be separated out as the 95% confidence interval (red lines) of the R-square difference between the allele structure and haplotype models spanned zero. There is thus no statistical indication that either the local structure Q[1]-2-0-9-2, or the broader B2 haplotype, better explains the variation in AoD observed (*i.e.,* B2 is not better associated with AoD than Q[1]-2-0-9-2).

Table S1. Demographic information for individuals affected with HD.

Individual number	Disease-associated allele CAG (Q^1)	Disease associated allele structure. (Q^1-Q^2-P^1-P^2-P^3)	AoD	AoO
1	39	Q^1-0-0-9-2	74.8	NA
2	40	Q^1-2-2-10-2	73.7	NA
3	40	Q^1-2-2-9-2	62.0	NA
4	40	Q^1-2-0-9-2	66.6	62.0
5	41	Q^1-2-0-9-2	61.3	NA
6	41	Q^1-2-0-9-2	41.5	NA
7	41	Q^1-2-0-9-2	59.5	NA
8	41	Q^1-2-0-9-2	63.4	NA
9	41	Q^1-2-2-7-2	77.0	NA
10	42	Q^1-2-2-6-3	69.7	NA
11	42	Q^1-2-0-9-2	62.7	35.0
12	42	Q^1-2-2-6-3	63.1	NA
13	42	Q^1-2-0-9-2	50.7	NA
14	42	Q^1-2-0-9-2	53.3	43.0
15	42	Q^1-2-0-9-2	59.9	NA
16	42	Q^1-2-0-9-2	50.5	50.0
17	42	Q^1-4-2-4-3	58.5	NA
18	42	Q^1-2-0-9-2	57.0	NA
19	43	Q^1-2-2-6-3	57.0	NA
20	43	Q^1-2-0-9-2	58.8	NA
21	43	Q^1-2-0-9-2	41.6	39.0
22	43	Q^1-2-0-9-2	48.1	47.0
23	43	Q^1-2-0-9-2	28.7	24.0
24	43	Q^1-2-0-9-2	51.9	NA
25	43	Q^1-2-2-10-2	50.9	39.0
26	43	Q^1-2-2-10-2	58.8	NA
27	43	Q^1-2-0-9-2	46.3	42.0
28	43	Q^1-2-2-10-2	67.6	45.0
29	43	Q^1-2-0-9-2	48.8	NA
30	43	Q^1-2-2-10-2	72.0	NA
31	44	Q^1-2-2-7-2	48.7	NA
32	44	Q^1-2-2-6-3	63.4	NA
33	44	Q^1-2-2-7-2	55.5	NA
34	45	Q^1-2-0-9-2	38.3	35.0
35	45	Q^1-2-0-9-2	45.5	NA
36	45	Q^1-2-0-9-2	39.0	NA
37	45	Q^1-2-0-9-2	48.7	45.0
38	45	Q^1-2-0-9-2	49.0	NA
39	45	Q^1-2-0-9-2	44.0	NA
40	45	Q^1-2-0-9-2	46.1	43.0
41	46	Q^1-2-2-10-2	56.3	NA
42	46	Q^1-2-2-10-2	33.4	30.0
43	46	Q^1-2-2-7-2	43.8	38.0
44	46	Q^1-2-2-10-2	39.9	39.0

45	46	Q^1-2-2-10-2	NA	NA
46	46	Q^1-2-0-9-2	41.3	40.0
47	46	Q^1-2-0-9-2	54.3	NA
48	46	Q^1-0-0-9-2	37.6	NA
49	46	Q^1-2-2-10-2	47.5	43.0
50	46	Q^1-2-2-7-2	49.0	NA
51	46	Q^1-2-2-7-2	42.8	36.0
52	47	Q^1-2-2-10-2	52.6	NA
53	47	Q^1-2-2-10-2	49.9	NA
54	48	Q^1-2-2-7-2	46.0	NA
55	48	Q^1-2-2-10-2	35.7	NA
56	49	Q^1-2-2-10-2	39.7	39.0
57	49	Q^1-2-2-10-2	43.6	28.0
58	49	Q^1-2-2-10-2	44.5	NA
59	49	Q^1-2-2-10-2	37.0	NA
60	49	Q^1-2-0-9-2	34.1	NA
61	51	Q^1-2-0-9-2	33.1	NA
62	51	Q^1-2-2-7-2	42.8	NA
63	51	Q^1-2-2-10-2	36.1	29.0
64	52	Q^1-2-2-7-2	37.1	34.0
65	53	Q^1-2-2-10-2	27.0	NA
66	54	Q^1-2-2-10-2	28.4	NA
67	55	Q^1-2-2-7-2	22.9	20.0
68	58	Q^1-2-0-9-2	34.9	NA

Demographic information of individuals affected with HD showing the disease associated CAG repeat length (Q^1), allele structures (Q^1-Q^2-P^1-P^2-P^3), age of diagnosis (AoD) and age of onset (AoO). The age of onset information was only available for 24 individuals, whereas the age of diagnosis was available in all except for one individual.

Table S2. The tag-SNPs used to construct the *HTT* haplotypes.

Tag-SNP number	rs number	Location on chromosome 4
1	rs2857936	3060583
2	rs762855	3073068
3	rs3856973	3078446
4	rs10015979	3107715
5	rs363075	3135947
6	rs363064	3139683
7	rs363102	3147289
8	rs4690073	3158423
9	rs363099	3160329
10	rs363096	3178294
11	rs2276881	3229934
12	rs362307	3240118
13	rs1006798	3256646

Location on chromosome 4: *Homo sapiens* (human) genome assembly GRCh38.p12 from Genome Reference Consortium.

Table S3. Multiple linear models testing the association between the ratio of somatic expansion and various explanatory variables.

	Model	r^2	*p*-value for model	Parameter values			
				Sample size	Explanatory variable	Effect to ratio of SE	*p*-value for explanatory variable
1	Ln (RSE) ~ CAG + Age at sampling + CAG*Age at sampling	0.758	$< 2 \times 10^{-16}$	60	*CAG*	*0.131*	*8×10^{-16}*
					Age at sampling	*0.008*	*1.8×10^{-3}*
					CAG*Age at sampling	0.000	0.764
2	Ln (RSE) ~ CAG + Age at sampling + CAG*Age at sampling + Allele structures	0.851	$< 2 \times 10^{-16}$	60	*CAG*	*0.114*	*2.6×10^{-3}*
					Age at sampling	0.018	0.521
				2	*Q^1-0-0-9-2*	*-0.257*	*7.7×10^{-4}*
				30	*Q^1-2-0-9-2*	*-0.168*	*1×10^{-5}*
				4	*Q^1-2-2-6-3*	*-0.151*	*0.014*
				1	Q^1-4-2-4-3	0.196	0.180
					CAG*Age at sampling	0.000	0.613

The statistically significant explanatory variables are indicated in *italics*. Ratio of somatic expansion (RSE)

Model 1. Linear model testing the association of the CAG repeat length and age at sampling on the RSE. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.79$, $p < 2 \times 10^{-16}$). The CAG repeat length and age at sampling also had a significant association. Model 2. Linear model testing the association of the CAG repeat length, age at sampling and the allele structures on the RSE, relative to a reference (the grouped typical allele structure, Q^1-2-2-P^2-2). The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.85$, $p < 2 \times 10^{-16}$). The CAG repeat length and the allele structures Q^1-0-0-9-2, Q^1-2-0-9-2 and Q^1-2-2-6-3 also had a significant association.

Table S4. Multiple linear models testing the association between the HD phenotype and various explanatory variables.

Model		r^2	*p*-value for model	Parameter values			
				Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
1	Ln (AoD) ~ CAG + CAACAG + CCGCCA + CCG + CCT	0.609	1.44×10^{-10}	64	*CAG*	*-2.902*	*6.11×10^{-12}*
				0=2, 2=61, 4=1	CAACAG	-1.554	0.498
				0=31, 2=33	*CCGCCA*	*4.029*	*7.34×10^{-4}*
				7/9/10=59, 4/6=5	CCG	-0.258	0.799
				2=59, 3=5	CCT	2.175	0.680
2	Ln (AoO) ~ CAG + CAACAG + CCGCCA + CCG + CCT	0.458	5.80×10^{-3}	24	*CAG*	*-1.836*	*5.23×10^{-3}*
				2=24	CAACAG	NA	NA
				0=12, 2=12	CCGCCA	0.634	0.757
				7=4, 9=12, 10=8	CCG	-0.076	0.963
				2=24	CCT	NA	NA
3	Ln (AoO) ~ CAG + Allele structures	0.458	1.624×10^{-3}	24	*CAG*	*-1.825*	*2.36×10^{-3}*
				12	Q^1-2-0-9-2	-1.191	0.752
4	Ln (AoD) ~ CAG + Allele structures	0.610	1.36×10^{-10}	64	*CAG*	*-2.910*	*5.71×10^{-12}*
				2	Q^1-0-0-9-2	-5.681	0.288
				29	*Q^1-2-0-9-2*	*-7.133*	*8.15×10^{-4}*
				4	Q^1-2-2-6-3	3.691	0.396
				1	Q^1-4-2-4-3	-2.489	0.743
5	Ln (AoD) ~ CAG + Haplogroups	0.587	2.054×10^{-10}	64	*CAG*	*-2.903*	*2.09×10^{-11}*
				9	*A*	*8.551*	*0.014*
				18	*C*	*6.202*	*0.022*
				4	*C-SA*	*11.752*	*0.012*
6	Ln (AoD) ~ CAG + Haplotypes	0.643	5.412×10^{-9}	64	*CAG*	*-3.069*	*3.19×10^{-11}*
				1	A2a	1.864	0.811
				1	A2b	9.225	0.275
				2	*A4a*	*16.826*	*0.018*
				5	*A4b*	*9.086*	*0.029*
				1	*B1*	*22.293*	*0.019*
				1	C4	1.419	0.857
				17	*C5*	*7.771*	*6.76×10^{-3}*
				4	*C9*	*12.323*	*7.85×10^{-3}*

The statistically significant explanatory variables are indicated in *italics*.

Model 1. Linear model testing the association of the individual components of the *HTT* repeat tract on the AoD. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.61$, $p = 1 \times 10^{-10}$), the CAG repeat length and CCGCCA sequence were also individually significant. Model 2. Linear model testing the association of the individual components of the *HTT* repeat tract on the AoO.The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.46$, $p = 5.8 \times 10^{-3}$), the CAG repeat length was also individually significant. The CAACAG sequence and the CCT repeat had no variation in the 24 individuals for which AoO information was available as indicated by NA. Model 3. Linear model testing the association of the allele structures on the AoO, relative to the grouped typical allele Q^1-2-2-P^2-2. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.46$, $p = 1.6 \times 10^{-3}$), and the CAG repeat length also had a significant association. Model 4. Linear model testing the association of the allele structures on the AoD, relative to the grouped typical allele Q^1-2-2-P^2-2. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.61$, $p = 1 \times 10^{-10}$), the CAG repeat length and Q^1-2-0-9-2 disease allele structure also had a significant association Model 5. Linear model testing the association of the background haplogroup on the AoD, relative to the most common haplogroup B. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.587$, $p = 2 \times 10^{-10}$), the CAG repeat length; haplogroup A, C and the haplogroup variant C-SA also had a significant association. Model 6. Linear model testing the association of the background haplotype on the AoD, relative to the most common haplotype B2. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.643$, $p = 5 \times 10^{-9}$), the CAG repeat length, haplotype A4a, A4b, B1, C5 and C9 had a significant association.