The Repeat Variant in *MSH3* is not a genetic modifier for Spinocerebellar Ataxia type 3 and Friedreich's ataxia.

Short running title: *MSH3* is not a genetic modifier for SCA3 and FRDA.

Yau WY[1], Raposo M[2,3], Bettencourt C[4,5], Labrum R[6], Vasconcelos J[7], Parkinson MH[8], Giunti P[8], Wood NW[1,6], Lima M[2,3], H Houlden[1,6]

1 Department of Neuromuscular Diseases, Institute of Neurology, University College London, UK

2 Faculdade de Ciências e Tecnologia, Universidade dos Açores, Ponta Delgada, Portugal

3 Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal

4 The Queen Square Brain Bank for Neurological Disorders, UCL Queen Square Institute of Neurology, London, United Kingdom

5 Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, London, United Kingdom

6 Neurogenetics Unit, National Hospital for Neurology and Neurosurgery, London, United Kingdom

7 Serviço de Neurologia, Hospital do Divino Espírito Santo, Ponta Delgada, Portugal

8 Ataxia Centre, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, Queen Square, London, United Kingdom

Sir,

MSH3 is a DNA mismatch repair gene whose product is essential for CAG repeat expansion and whose inactivation limits the expansion (Dragileva *et al.*, 2009). The study conducted by Flower et al. demonstrates that a three-repeat allele in the mismatch repair gene *MSH3* has potential disease-modifying effect on Huntington's disease (HD) and myotonic dystrophy type 1 (DM1) (Flower *et al.*, 2019), both of which are CAG repeat expansion disorders with distinct phenotypic features. The authors postulated that these effects are mediated through the influence of the *MSH3* variant on the rate of somatic expansion. We are interested to find out whether the effect of this three-repeat allele in *MSH3* on the age at disease onset (AAO) and disease progression is applicable to other triplet repeat disorders; and we studied this in spinocerebellar ataxia type 3 (SCA3) and Friedreich's ataxia (FRDA).

SCA3 is the most common autosomal dominant ataxia with a (CAG)n expansion in *ATXN3* (Kawaguchi *et al.*, 1994; Ruano *et al.*, 2014)*,* whilst FRDA is the most common autosomal recessive ataxia with (GAA)n expansion in *FXN* (Campuzano et al., 1996). Both are progressive neurodegenerative diseases that share clinical features of ataxia with additional neurological and systemic features. The length of (CAG)n expansion above the disease threshold is inversely proportional to the AAO of SCA3, accounting for 55.2% of this variation on average (de Mattos *et al.*, 2019). The AAO of FRDA predominantly depends on the size of the smaller expanded GAA allelic length (Durr *et al.*, 1996; De Michele *et al.*, 1998).

We enrolled 132 SCA3 and 136 FRDA patients on whom we have AAO and CAG/GAA repeat size information from the Neurogenetics Unit and Ataxia Centre of the National Hospital for Neurology and Neurosurgery and the University of Azores/ Hospital of Divino Espírito Santo, Ponta Delgada (See supplementary table 1 for demographic data). Of the FRDA patients, 57 had at least two measurements of the scale for the assessment and rating of ataxia (SARA). The study was approved by the joint ethics committee of UCL institute of Neurology and the National Hospital for Neurology and Neurosurgery, UK (UCLH: 04/N034) and the ethics committee of the University of Azores (2/2016). All patients were clinically assessed by neurogenetic specialists and diagnoses confirmed through routine genetic diagnostic tests. SCA3 alleles were determined by polymerase chain reaction (PCR). PCR products were checked on a 4% agarose gel and then the fragments were resolved on an ABI 3730xl DNA analyser with a GeneScan 500 LIZ Size Standard (Thermo Fisher Scientific). We then performed fragment analysis with GeneMapper software (version 4.0, Applied Biosystems) and chose the highest peak for each allele to calculate the allele size. The allele size was subtracted from the number of extraneous bases in the PCR product outside of the repeat region (160 bp) and then divided by three to obtain repeat lengths (Supplementary table 2). We estimated the GAA repeat length of each FRDA alleles as previously described (Lamont *et al.*, 1997). Flower et al. described 16 polymorphic nine-base-pair repeats with variable sequences and lengths, ranging from three to nine repeats (Flower *et al.*, 2019). In HD and DM1, the most common alleles are 3a and 6a (three-repeat allele and six-repeat allele respectively), accounting for over 80% of all alleles in this *MSH3* region. The number of *MSH3* 3a alleles that patients carry was associated with HD and DM1 phenotypes. We identified the 3a/3b *MSH3* alleles using fragment analysis and sizing of PCR products generated using flanking PCR primers (Supplementary table 2). We distinguished

between 3a and 3b repeat allele using sanger sequencing to detect the polymorphism at the 12th nucleotide (cytosine/ thymine) in the three-repeat allele (Supplementary fig. 1).

Linear regression modelling of genotype-phenotype correlation was conducted in R version 3.6.1 (R_Core_Team, 2013). The number of 3a repeat alleles in *MSH3* that patients carry was categorised as 0 (reference), 1 and 2. The p-values for the effects of one or two copies of 3a repeat alleles in *MSH3* versus none and the combined effect of 3a repeat alleles in *MSH3* were calculated using analysis of variance. For age at onset analysis, the outcome variable is the age at motor onset. The predictor variables of the SCA3 model included the size of the expanded CAG repeat allele and the 3a repeat alleles. The predictor variables of the FRDA model included the sizes of both expanded GAA repeat alleles and the 3a repeat alleles (Supplementary table 3). For progression analysis, the outcome variable was annual change in SARA score and the predictor variables were the duration of the disease (categorised as 0-10 years[reference], >10-20 years and >20 years), the baseline SARA score and the 3a repeat alleles (Supplementary table 3). The parsimonious model did not contain the size of GAA repeat alleles. Power calculation was performed using R pwr package for generalised linear models.

In both of our SCA3 and FRDA cohorts, the allelic frequency of 3a repeat was 0.35. We did not identify a statistically significant association between the AAO and the number of 3a repeat allele in *MSH3* in the SCA3 cohort (p=0.45) or the FRDA cohort (p=0.18). Our results also did not confirm the inverse relationship previously seen (Fig. 1). Similarly, the annual change in SARA score did not have statistically significant association with the number of 3a repeat allele (p=0.28) (Fig. 2).

In conclusion, we were not able to replicate the effect of the 3a repeat allele in *MSH3* on the AAO of SCA3, FRDA or the progression of FRDA. The weaknesses of our study include: (1)the estimation of GAA repeat length is imprecise, especially in repeats larger than 800; (2) our sample sizes for age at onset analyses can only detect ≥10% of the proportion of variance explained in the models with 90% power and significance level at 0.05; (3) our progression analysis is hampered by the lack of complete longitudinal data on severity scores and a small sample size. However, we would still expect to see an inverse relationship between the AAO and the number of 3a repeat alleles if the 3a repeat allele is a strong disease modifier for SCA3 or FRDA. DNA repair genes have been implicated in altering the age at onset of spinocerebellar ataxias (Bettencourt *et al.*, 2016). The DNA mismatch repair protein Msh3 heterodimerises with another protein Msh2 to form MutS beta and promotes trinucleotide repeat expansion by recruiting DNA polymerase B to nascent (CAG)n hairpins for error-prone DNA synthesis (Guo *et al.*, 2016). In addition, knock-down of *MSH3* expression using short hairpin RNA slowed GAA expansion in FRDA human cellular model (Halabi *et al.*, 2012). Thus, the absence in effect of this repeat allele in *MSH3* does not preclude the mismatch repair system as a potential therapeutic target for neurodegenerative diseases driven by repeat expansions. Following recent advances in Huntington disease's research, a genome-wide association study of all the common (CAG)n repeat disorders would be an instrumental first step in understanding the genetic architecture and identifying further suitable pharmacological targets common to this group of individually rare neurogenetic disorders.

Figures

Figure 1. The number of *MSH3* 3a repeat allele is not associated with SCA3 and FRDA AAO. Boxplots for AAO adjusted for the inherited CAG length in SCA3(a) and adjusted for the inherited GAA lengths in FRDA(b); the adjusted AAO is subtracted from the mean AAOs of SCA3 and FRDA cohorts respectively. The box indicates the 95% confidence intervals of the mean, the horizontal line within the box the mean and the whiskers the standard deviations.

Figure 2. The number of *MSH3* 3a repeat allele is not associated with annual change in SARA. Boxplots for annual change in SARA adjusted for the disease duration and baseline SARA. The box indicates the 95% confidence intervals of the mean, the horizontal line within the box the mean and the whiskers the standard deviations.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare no competing interests.

Supplementary materials

Supplementary figure 1. Identification of 9bp repeat allele using fragment analysis, Sanger sequencing and gel electrophoresis.

(a)&(b) demonstrate the sizing of the *MSH3* exon 1 on GeneMapper and chromatograms of the representative samples (upper: homozygous 6a alleles, mid: homozygous 3a allele, lower: heterozygous 3a/6a alleles). Solid arrow indicates the start of the repeat allele; star indicates the position of allelic nucleotide difference between 3a allele(Cytosine) and both 3b and 6a alleles(Thymine); arrowhead indicates the positions of allelic difference between 3a allele(Cytosine) and 6a allele(Guanine), followed by misaligned sequences. (c) Gel electrophoresis of PCR products of representative samples with homozygous 6a, homozygous 3a and heterozygous 3a/6a alleles. (d) The reference sequences of the 6a, 3a and 3b repeat alleles in *MSH3* reported by Flower et al.

Supplementary table 1. Basic Demographic data of Spinocerebellar Ataxia type 3 and Friedreich Ataxia cohorts

Supplementary table 2. Primers sequences and thermocycling conditions

Supplementary table 3. Regression models of the relationships between *MSH3* 3a repeat allele and the AAO of SCA3 and FRDA, and progression of FRDA. The table shows the adjusted squared coefficient of correlation (adjusted r2), and statistical significance (p) for each model, and the coefficient, standard error, t statistic and statistical significance (p), associated with each parameter in the model and the group effect of 3a repeat alleles. AAO: age at onset; CAG: inherited repeat allele length in SCA3; GAA: inherited repeat allele length in FRDA; FRDA: Friedreich's Ataxia; SARA: scale for the assessment and rating of ataxia; SCA3: spinocerebellar ataxia type 3.

References

Bettencourt C, Hensman-Moss D, Flower M, Wiethoff S, Brice A, Goizet C*, et al.* DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol 2016; 79(6): 983-90.

Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F*, et al.* Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 1996; 271(5254): 1423-7.

de Mattos EP, Kolbe Musskopf M, Bielefeldt Leotti V, Saraiva-Pereira ML, Jardim LB. Genetic risk factors for modulation of age at onset in Machado-Joseph disease/spinocerebellar ataxia type 3: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry 2019; 90(2): 203-10.

De Michele G, Filla A, Criscuolo C, Scarano V, Cavalcanti F, Pianese L*, et al.* Determinants of onset age in Friedreich's ataxia. J Neurol 1998; 245(3): 166-8.

Dragileva E, Hendricks A, Teed A, Gillis T, Lopez ET, Friedberg EC*, et al.* Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. Neurobiol Dis 2009; 33(1): 37-47.

Durr A, Cossee M, Agid Y, Campuzano V, Mignard C, Penet C*, et al.* Clinical and genetic abnormalities in patients with Friedreich's ataxia. N Engl J Med 1996; 335(16): 1169-75.

Flower M, Lomeikaite V, Ciosi M, Cumming S, Morales F, Lo K*, et al.* MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. Brain 2019.

Guo J, Gu L, Leffak M, Li GM. MutSbeta promotes trinucleotide repeat expansion by recruiting DNA polymerase beta to nascent (CAG)n or (CTG)n hairpins for error-prone DNA synthesis. Cell Res 2016; 26(7): 775-86.

Halabi A, Ditch S, Wang J, Grabczyk E. DNA mismatch repair complex MutSbeta promotes GAA.TTC repeat expansion in human cells. J Biol Chem 2012; 287(35): 29958-67.

Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S*, et al.* CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet 1994; 8(3): 221-8.

Lamont PJ, Davis MB, Wood NW. Identification and sizing of the GAA trinucleotide repeat expansion of Friedreich's ataxia in 56 patients. Clinical and genetic correlates. Brain 1997; 120 (Pt 4): 673-80.

R_Core_Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing 2013.

Ruano L, Melo C, Silva MC, Coutinho P. The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. Neuroepidemiology 2014; 42(3): 174-83.
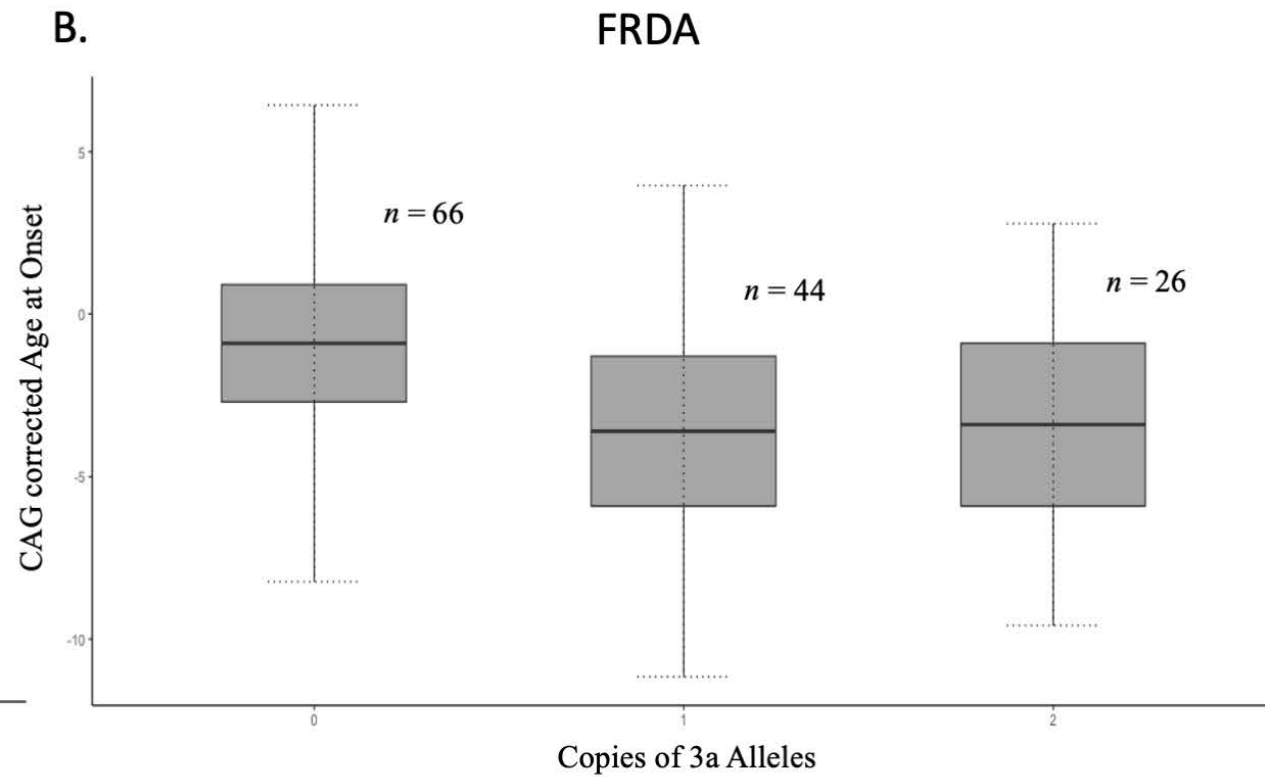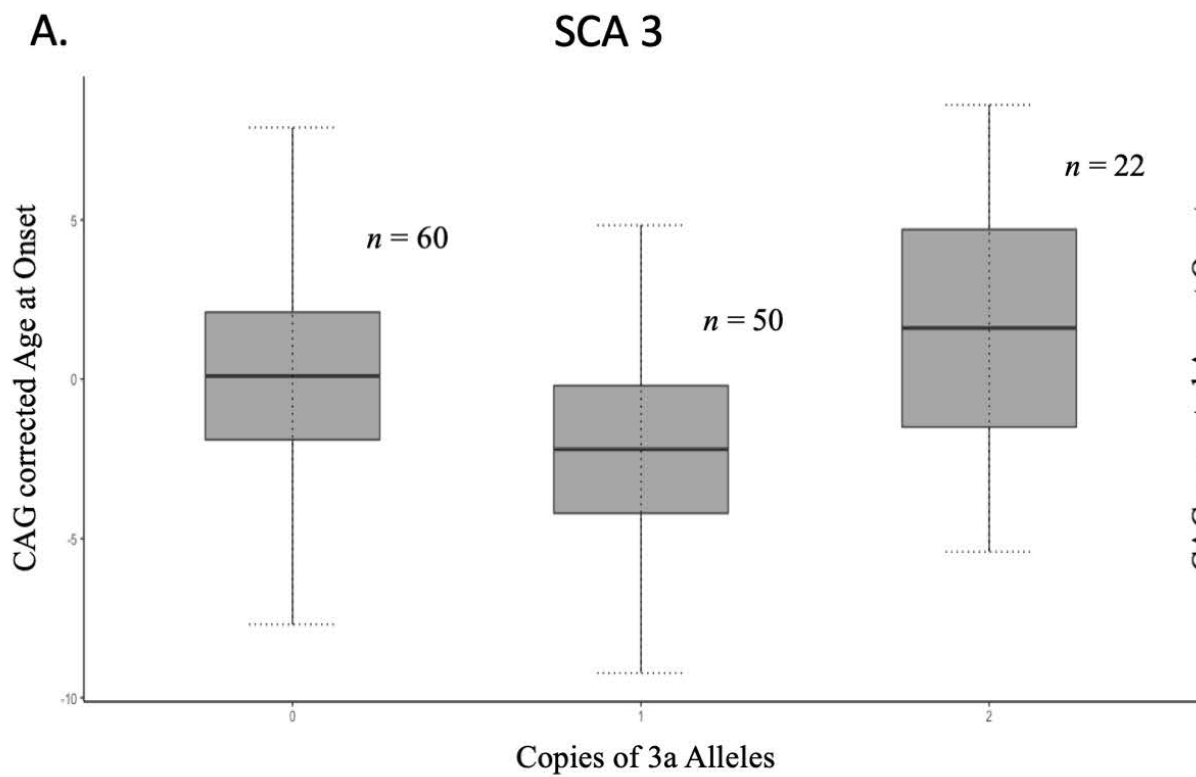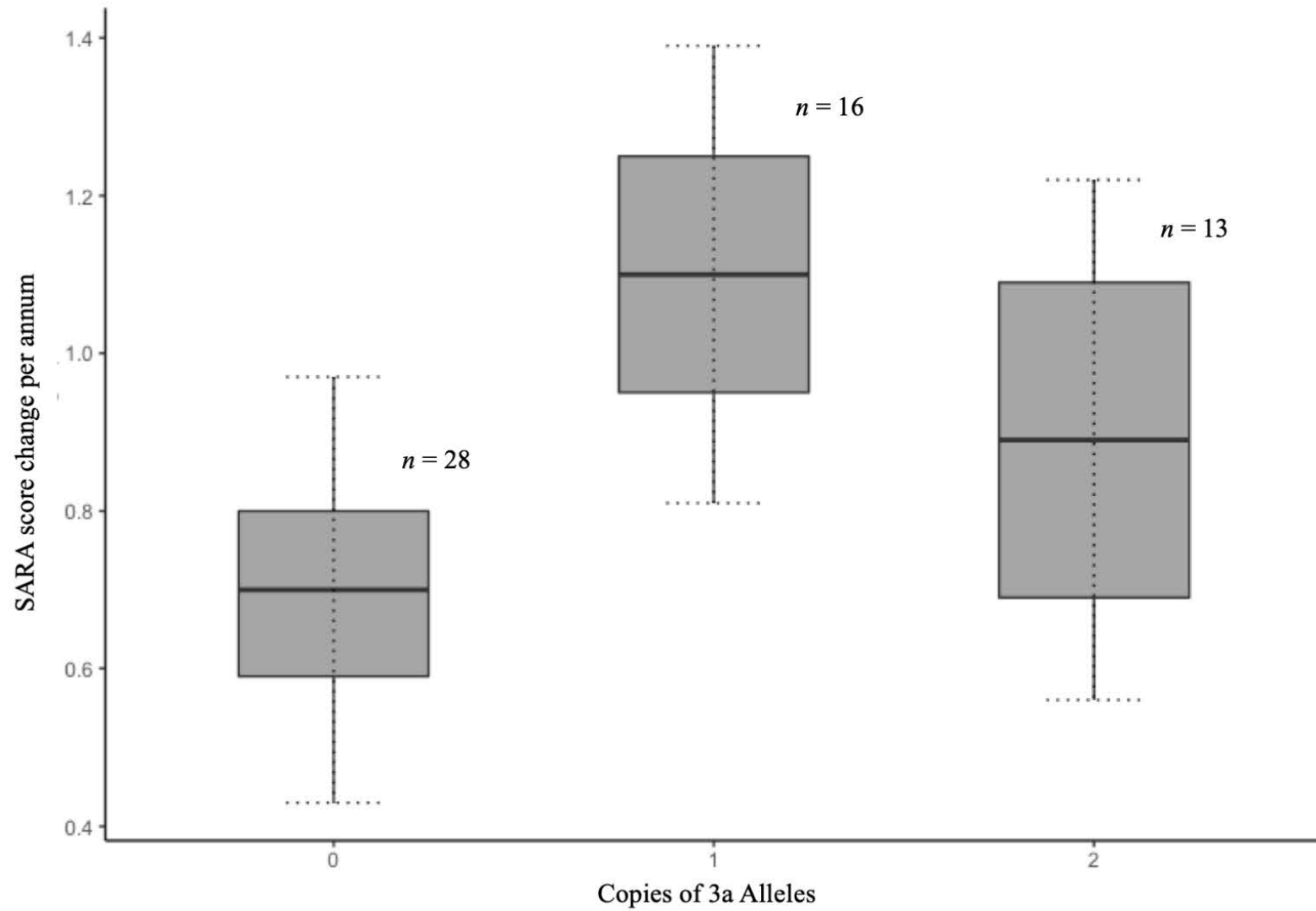
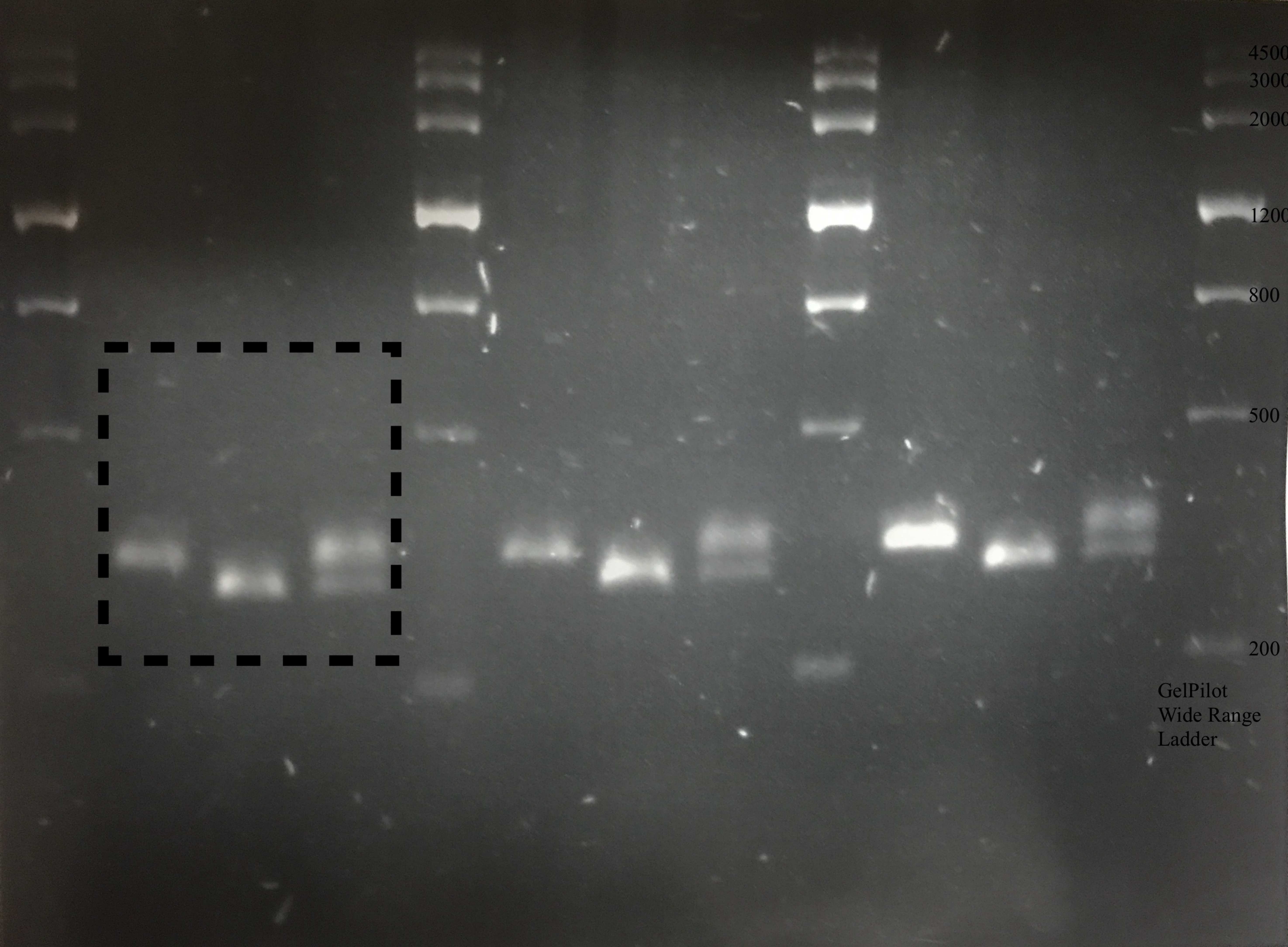|	Avg AAO in years (Range)	Gender	Ethnicity	(CAG)n expanded allele	First (GAA)n expanded allele	Second (GAA)n expanded allele	SARA score (First measurement)	SARA score (Second measurement)	Avg Duration between measurements
SCA3	36 (13-70)	M: 74 (56%), F: 58 (44%)	Portugese: 97 (71%), East Asian: 8 (6%), East African 7 (5%), West African 7 (5%), Afro-caribean 7 (5%), Brazilian 4 (3%), others 6 (6%)	70 (58-79)	-	-	-	-	-
FA	16 (3-60)	M: 764(47%), F: 72 (53%)	White european: 118 (87%), East Asian: 11 (8%), others 7 (5%)	-	810 (67-4500)	1100 (83-4500)	22 (2-39)	37 (19-70)	4

Supplementary table 2. Primers sequences and thermocycling conditions

	primers	PCR	Cycling
ATXN3 Flanking PCR	Fw: 6-FAM-5'-CCAGTGACTACTTTGATT-3' Rv: 5'-TGGCCTTTCACATGGATGTGAA-3'	Promega 5X buffer 10µl dNTPs (2mM) 5µl DMSO 5µl Primers 10pmol/µl each gDNA 50ng GoTaq G2 Taq polymerase 1.25 units H_2O to 50µl	94°C 5mins [94°C 30s, 50°C 30s, 72°C 30s] X 30 72°C 7mins 4°C
FRDA Long PCR	Fw: 6-FAM-5'-GGGATTGGTTGCCAGTGCTTAAAAGTTAG-3' Rv: 5'-GATCTAAGGACCATCATGGCCACACTTGCC-3'	Phusion Hi Fidelity 2X Mastermix buffer 12.5µl Primers 2.5pmol/µl each gDNA 50ng H_2O to 25µl	98°C 30s [98°C 5s, 72°C 3mins] X10 [98°C 5s, 72°C 3mins+20s/cycle) X20 72°C 10mins 4°C
MSH3Ex1 Flanking PCR	Fw: 6-FAM-5'-CGAGGCAAGCGGTTTTGAG-3' Rv: 5'-TTCCCACCTTCCCCTTCTTC-3'	Faststart mastermix 2X 7.5 µl Primers 10pmol/µl each gDNA 50ng H_2O to 15µl	95°C 4mins [95°C 30s, 56°C30s, 72°C 1min] X35 72°C 1min 4°C

SCA3 spinocerebellar ataxia, FRDA Friedreich's ataxia, Fw forward, Rv reverse, dNTP deoxyribose nucleoside triphosphate, DMSO dimethyl sulfoxide, gDNA genomic DNA

	Model	Adjusted R-squared	Model p-value	Parameter		Coefficient	Standard error	t-statistic	Parameter p-value	p-values for the group effect of 3a repeat alleles
1a	*AAO = α0 + α1X(CAG.CTG) + α2 (Variant Allele one copy) + α3(Variant Allele two copies)*	0.44	<2.2e-16	*Intercept*	α0	168.3	12.8	12.8	<2e-16	
	SCA3, n=132			*CAG*	α1	-1.87	0.18	0.18	<2e-16	
				Variant Allele one copy	α2	-1.47	1.66	1.66	0.38	0.45
				Variant Allele two copies	α3	1.15	2.17	2.17	0.6	
1b	*AAO = α0 + α1Xlog(GAA.TCC)1 + α2Xlog(GAA.TCC)2 + α3(Variant Allele one copy) + α4(Variant Allele two copies)*	0.55	<2.2e-16	*Intercept*	α0	84.1	7.42	11.3	<2e-16	
	FRDA, n=136			*GAA First allele*	α1	-12.7	1.23	-10.3	<2e-16	
				GAA Second allele	α2	2.23	1.36	1.65	0.1	
				Variant Allele one copy	α3	-2.64	1.43	-1.84	0.07	0.18
				Variant Allele two copies	α4	-1.73	1.78	-0.98	0.33	
2a	*Change in SARA per annum = α0 + α1(Variant Allele one copy) + α2(Variant Allele two copies) + α3(Disease Duration)1 + α4(Disease Duration)2 + α5(SARA)*	0.08	0.1	*Intercept*	α0	0.81	0.32	2.5	0.02	
	FRDA SARA, n=57			*Variant Allele one copy*	α1	0.44	0.26	1.7	0.1	0.28
				Variant Allele two copies	α2	0.13	0.27	0.47	0.63	
				Disease Duration1 (>10-20 years)	α3	0.65	0.31	2.1	0.04	
				Disease Duration2 (>20 years)	α4	0.35	0.35	1	0.32	
				SARA score (first)	α5	-0.02	0.01	-1.9	0.06	

Supplementary table 3. Regression models of the relationships between *MSH3* 3a repeat allele and the AAO of SCA3 and FRDA, and progression of FRDA. The table shows the adjusted squared coefficient of correlation (adjusted r2), and statistical significance (p) for each model, and the coefficient, standard error, t statistic and statistical significance (p), associated with each parameter in the model and the group effect of 3a repeat alleles. AAO: age at onset; CAG: inherited repeat allele length in SCA3; GAA: inherited repeat allele length in FRDA; FRDA: Friedreich's Ataxia; SARA: scale for the assessment and rating of ataxia; SCA3: spinocerebellar ataxia type 3.