Human Genetics (2024) 143:211–232
https://doi.org/10.1007/s00439-024-02644-7

ORIGINAL INVESTIGATION

Novel genotype–phenotype correlations, differential cerebellar allele-specific methylation, and a common origin of the $(ATTTC)_n$ insertion in spinocerebellar ataxia type 37

Marina Sanchez-Flores[1] · Marc Corral-Juan[1] · Esther Gasch-Navalón[1] · Davide Cirillo[2] · Ivelisse Sanchez[1] · Antoni Matilla-Dueñas[1]

Received: 26 August 2023 / Accepted: 17 January 2024 / Published online: 23 February 2024

Abstract

Spinocerebellar ataxia subtype 37 (SCA37) is a rare disease originally identified in ataxia patients from the Iberian Peninsula with a pure cerebellar syndrome. SCA37 patients carry a pathogenic intronic (ATTTC)n repeat insertion flanked by two polymorphic (ATTTT)n repeats in the Disabled-1 (DAB1) gene leading to cerebellar dysregulation. Herein, we determine the precise configuration of the pathogenic 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n SCA37 alleles by CRISPR–Cas9 and long-read nanopore sequencing, reveal their epigenomic signatures in SCA37 lymphocytes, fibroblasts, and cerebellar samples, and establish new molecular and clinical correlations. The 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n pathogenic allele configurations revealed repeat instability and differential methylation signatures. Disease age of onset negatively correlated with the (ATTTC)n, and positively correlated with the 3′(ATTTT)n. Geographic origin and gender significantly correlated with age of onset. Furthermore, significant predictive regression models were obtained by machine learning for age of onset and disease evolution by considering gender, the (ATTTC)n, the 3′(ATTTT)n, and seven CpG positions differentially methylated in SCA37 cerebellum. A common 964-kb genomic region spanning the (ATTTC)n insertion was identified in all SCA37 patients analysed from Portugal and Spain, evidencing a common origin of the SCA37 mutation in the Iberian Peninsula originating 859 years ago (95% CI 647–1378). In conclusion, we demonstrate an accurate determination of the size and configuration of the regulatory 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract, avoiding PCR bias amplification using CRISPR/Cas9-enrichment and nanopore long-read sequencing, resulting relevant for accurate genetic diagnosis of SCA37. Moreover, we determine novel significant genotype–phenotype correlations in SCA37 and identify differential cerebellar allele-specific methylation signatures that may underlie DAB1 pathogenic dysregulation.

Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
CB	Cerebellum
DAB1	Disabled-1
DMF	Differentially methylated frequency
DMR	Differentially methylated region
FC	Fibroblasts cells
GPU	Graphics processing unit
HAC	High accuracy
HMW	High molecular weight
IGV	Integrative genomics viewer
IQR	Interquartile range
LLR	Log-likelihood ratio
NGS	Next-generation sequencing
OLS	Ordinary least squares regression
PBL	Peripheral blood leukocytes
SCA	Spinocerebellar ataxia
SCA37	Spinocerebellar ataxia type 37
SD	Standard deviation
SHAP	SHapley additive exPlanations
SNP	Single-nucleotide polymorphism
SUP	Super accuracy
TRE	Tandem repeat expansion

Marina Sanchez-Flores and Marc Corral-Juan have contributed equally to this work.

✉ Antoni Matilla-Dueñas
amatilla@igtp.cat

[1] Neurogenetics Unit, Department of Neuroscience, Germans Trias i Pujol Research Institute (IGTP), Universitat Autònoma de Barcelona-Can Ruti Campus, Carretera de Can Ruti, Camí de les Escoles s/n, 08916 Badalona, Spain

[2] Barcelona Supercomputing Center (BSC), Barcelona, Spain

Background

Autosomal-dominant spinocerebellar ataxias (SCAs) are rare inherited movement neurodegenerative disorders mainly characterized by progressive cerebellar ataxia variably associated with ophthalmoplegia, pyramidal and extrapyramidal signs, dementia, pigmentary retinopathy, seizures, lower motor neuron signs, and peripheral neuropathy (Dueñas et al. 2006; Durr 2010; Klockgether et al. 2019). Disease onset is typically in adulthood, albeit some clinical signs can appear earlier. Currently, 48 well-defined dominant ataxia subtypes have been described evidencing the high clinical, genetic, and neuropathological heterogeneity. Genetic defects have been associated in 42 subtypes, playing a prominent role underlying SCAs physiopathology, with tandem repeat expansions (TREs) or conventional mutations triggering toxic gain- or loss-of-function events underlying neurodegeneration (Durr 2010; Jayadev and Bird 2013; Matilla-Dueñas et al. 2014; Tan et al. 2023). TREs, in both coding and non-coding genomic regions, represent one third of all the genetic defects associated with the SCAs and up to 60 monogenic disorders (Wen et al. 2023). They are highly polymorphic and reveal high instability in repeat length. Assessing their exact number is difficult by PCR amplification because of reduced amplification efficiency, allele dropout, and replication slippage defects limiting clinical-genetic correlations and genetic diagnosis (Polak et al. 2013; Hommelsheim et al. 2014; Potapov and Ong 2017; Kacher et al. 2021). Long pathogenic TREs are difficult to characterize by next-generation sequencing (NGS) based on short reads limited to fragments of 150 bp for their difficulty to be aligned to the reference genome leading to inaccurate sizing and misinterpretation.

We previously described the spinocerebellar ataxia subtype 37 (SCA37) in four unrelated Spanish kindreds caused by an unstable intronic (ATTTC)n pentanucleotide repeat insertion, located within an (ATTTT)n repeat tract in the 5'non-coding regulatory region of the *DAB1* gene encoding the Reelin adaptor protein, which is implicated in neuronal migration (Serrano-Munuera et al. 2013; Corral-Juan et al. 2018). The heterozygous (ATTTC)n repetitive insertion was identified in affected patients of all SCA37 families described to date from the Southeast of the Iberian Peninsula, four Spanish and six Portuguese, and, more recently, in a German kindred (Serrano-Munuera et al. 2013; Seixas et al. 2017; Corral-Juan et al. 2018; Rosenbohm et al. 2022). Normal non-pathogenic alleles show a repetitive ATTTT stretch of 7–400 units, which is not interrupted by ATTTC units, though 3% may be interrupted by AT-rich motifs; most (94%) bear ≤ 30 ATTTT repeats (Loureiro et al. 2019; Rosenbohm et al. 2022). Age-dependent penetrant pathogenic alleles include an insertion of 31–102 (ATTTC) n repeats flanked by adjacent ATTTT repeats larger than 58 units, 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n, being the 3'(ATTTT)n the largest described containing up to 420 repeat units (Rosenbohm et al. 2022), and without interruptions neither in the (ATTTT)n or the (ATTTC)n tracts (Loureiro et al. 2019). The ATTTC repeat inserted mutation dysregulates *DAB1* expression up-regulating Reelin-DAB1 and PI3K/AKT signalling in the SCA37 cerebellum (Corral-Juan et al. 2018).

Herein, we implement unbiased PCR-free amplification of targeted CRISPR/Cas9-mediated enrichment of the SCA37 altered region in combination with nanopore long-read sequencing to accurately determine the exact size and configuration of the 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n tract, obtain the methylation signatures of the SCA37 alleles, and generate predictive models for the age of onset and disease evolution using machine learning approaches. Moreover, we demonstrate a common origin of SCA37 chromosomes in the south of the Iberian Peninsula originating 859 years ago.

Methods

Patients and samples

DNA, genetic and clinical data of 24 individuals from four previously reported Spanish SCA37 unrelated kindreds (Serrano-Munuera et al. 2013; Corral-Juan et al. 2018), and 5 additional SCA37 individuals from four unrelated Spanish kindreds were included in this study. Informed consents were obtained for all individuals, and the study was approved by the clinical ethical board of the University Hospital Germans Trias i Pujol (HUGTP) in Badalona. Genetic and clinical data of 34 individuals from six previously reported Portuguese SCA37 kindreds (Seixas et al. 2017) were included to generate genotype–phenotype correlations and establish the origin and date of the SCA37 mutation in the Iberian Peninsula.

Genetic and genomic studies

Genomic DNAs (gDNA) were isolated from peripheral blood leukocytes (PBL) from 29 SCA37 Spanish patients using Chemagen Magnetic Separation Module I automated system (Perkin Elmer). Additionally, high molecular weight (HMW) DNAs from one PBL, fibroblast cells (FC) from two patients and two cerebellar (CB) samples were extracted using the HMW DNA Extraction Kit, either for cells and blood (New England Biolabs (NEB), Cat. no. T3050S) or for tissue (NEB, Cat. No. T3060S). All 29 affected SCA37 members from eight Spanish SCA37 kindreds were genotyped for the non-pathogenic (ATTTT)n and the (ATTTC)

n expanded insertion with a modified SCA37 long-PCR amplification protocol using LA Taq DNA polymerase (TaKaRa, Cat. no. RR002AG) and Sanger sequencing. Primers sequences (Suppl. Table 1) and PCR conditions are included in Supplementary Information.

Unbiased long-read nanopore sequencing of the SCA37 region containing the $(ATTTC)_n$ insertion

For nanopore sequencing of 11 SCA37 Spanish patients, the wild-type (ATTTT)n and SCA37 5'(ATTTT)n–(ATTTC)n–3′(ATTTT)n genomic regions were targeted enriched by CRISPR–Cas9 to obtain high coverage using six crRNAs complementary to strands flanking the SCA37 5′(ATTTT)n–(ATTTC)n–3'(ATTTT)n tract spanning a total of 22.29 kb within intron 11 of the *DAB1* gene (chr1: 57353671–57375963; hg38; Suppl. Figure 1a). Guides sequences and the detailed procedure of CRISPR–Cas9 assembled RNP complexes are described in Supplementary Table 2. DNA libraries were prepared according to previously published nCATS protocol (Gilpatrick et al. 2020) following the manufacturer's protocol for SQK-LSK109 kit (Oxford Nanopore Technologies, ONT) with the following modifications. Briefly, up to 7 µg of gDNA and 120 units of thermolabile proteinase K (NEB, Cat. no. P8111S) were used to remove the Cas9 protein bound to DNA, which could interfere with adapters ligation, as recently suggested by Keraite and collaborators (Keraite et al. 2022). The library was eluted in 30 µL of elution buffer. The detailed protocol is described in Supplementary Information. DNA libraries were sequenced during 72 h in a PromethION sequencer (ONT), using a FLO-PRO002 flow cell (R9.4.1) per sample. Fast5 (electronic raw signal) and FASTQ (base called data) files were generated in real time with MinKNOWN software v22.08.6 and used for downstream bioinformatics analysis. Base calling modes Fast, Hac (high accuracy) and Sup (super high accuracy) from Guppy v 6.2.11 software were used to analyse the effect of these three different data conversions (fast5 to FASTQ) modes on sequence accuracy of the CRISPR–Cas9 targeted region. Sup mode requires a higher computing infrastructure for optimal base calling performance using Graphics processing unit (GPU). Raw signal was base called using the Sup mode in Guppy software v 6.2.11. All sequenced reads were aligned to the hg38 human reference genome with Minimap2 v2.17-r941 (Li 2018). SAMtools v1.10 (Danecek et al. 2021) was used for BAM files generation and Alfred v0.2.6 (Rausch et al. 2019) in combination with NanoStat v1.6.0 (De Coster et al. 2018) software was used for nanopore sequencing quality control and assessment of multiple-guide CRISPR–Cas9 enrichment. We calculated the genome-wide off-target depth of coverage per every 5000 bp using Mosdepth v0.3.4 in combination with the Cas-OFFinder tool and the RepeatMasker track (Tarailo-Graovac and Chen 2009; Bae et al. 2014; Pedersen and Quinlan 2018). On-target sequencing reads were visualized with Integrative Genomics Viewer (IGV) v2.12.0 (Robinson et al. 2011). Since STRique, a Python package to analyse repeat tracts, was not designed to count complex repeats such as those in SCA37 (Erdmann et al. 2023), MarginPhase (Ebler et al. 2019) and WhatsHap (Martin et al. 2016) software for phasing genomic variants were used in combination with Python3-based scripts Repeat Analysis Tools (https://github.com/PacificBiosciences/apps-scripts/tree/master/RepeatAnalysisTools) to classify the reads aligned to the repeat tract region and determine their repeat tract configuration. The mean of the standard deviation from interquartile ranges previously reported in Rosenbohm and collaborators (Rosenbohm et al. 2022) was calculated with interquartile range (IQR) formula (SD = IQR/1.35) (Wan et al. 2014a). Absolute repeat instability index was calculated as previously described (Lee et al. 2010) considering a threshold of more than one read contiguously present in reads distribution. Modifications outlined by Nakamori (Nakamori et al. 2020) were applied to calculate relative instability index in the cerebellum or skin fibroblasts compared to peripheral blood lymphocytes from the same patient.

Allele-specific methylation analysis and TF-binding sites enrichment

Haplotype reconstruction was used to assess allele-specific methylation with the f5c bioinformatic tool (Gamaarachchi et al. 2020). Methylation was predicted for individual CpG sites located in both strands of the target region. Outlier log-likelihood ratio (LLR) values were removed using the $1.5 \times$IQR method using R script (Yang et al. 2019). Allele-specific methylation frequencies were calculated dividing the number of methylated reads by unmethylated reads on a particular position and considering a methylated position when a LLR threshold was ≥ 2.0, as previously established (Liu et al. 2021). Differentially methylated regions (DMR) were considered when the absolute difference in the average overall methylation frequency was greater than 0.05 between WT and SCA37 alleles as previously described (Grant et al. 2022). The method "loess" in ggplot2 R-package was used for smoothed data representation of methylation fraction values with span parameter set to 0.1. Log-likelihood ratio values were also used to determine the significantly differentially methylated CpG sites in the SCA37 cerebellum with one-way ANOVA or the non-parametric Mann–Whitney *U* tests after Levene's test of homogeneity of variances using an in-house R script. The transcription factor database JASPAR 2022 CORE vertebrate collection was used to predict TF-binding sites surrounding the CpG positions

Table 1 Cas9-mediated long-read sequencing for DAB1 repeat region of Spanish wild-type and SCA37 chromosomes

Ped.	Ped. ID	Sample	Type of sample	WT-$(ATTTT)_n$			SCA37-5′$(ATTTT)$n–$(ATTTC)_n$–3′$(ATTTT)_n$							Total n of reads (WT/SCA37)
				WT-$(ATTTT)_n$ repeats by sanger	Median WT-$(ATTTT)_n$ repeats by nanopore	WT-$(ATTTT)_n$ repeats by nanopore SD	(ATTTC)n repeats by sanger	Median $(ATTTC)_n$ repeats by nanopore	$(ATTTC)_n$ repeats by nanopore SD	5′$(ATTTT)_n$ repeats by nanopore	5′$(ATTTT)_n$ repeats by nanopore SD	3′$(ATTTT)_n$ repeats by nanopore	3′$(ATTTT)_n$ repeats by nanopore SD	
AT-901	IV:9	SPA001	Blood	**9**	**9**	0.67	**48**	**48**	2.4	**66**	2.4	**83**	2.9	212 (113/99)
			Cerebellum*	**9**	**9**	0.66	**55**	**51**	3.5	**65**	2.6	**80**	5.6	482 (260/222)
	IV:10	SPA002	Blood	**9**	**9**	0.72	**46**	**48**	1.9	**66**	2.7	**83**	2.6	32 (16/16)
			Cerebellum*	**9**	**9**	0.65	**55**	**51**	3.2	**65**	2.8	**80**	4.3	441 (239/202)
	IV:4	SPA003	Fibroblasts*	**8**	**8**	0.59	**47**	**47**	2.9	**67**	2.7	**84**	3.1	644 (321/323)
AT-9012	III:3	SPB001	Blood	**8**	**8**	0.95	**51**	**50**	3.8	**67**	1.7	**85**	3.2	11 (7/4)
	IV:9	SPB002	Blood*	**11**	**11**	0.48	**58**	**59**	2.8	**69**	2.8	**84**	2.9	233 (131/102)
AT-59	IV:9	SPC001	Blood	**17**	**18**	1.18	**71**	**72**	1.7	**60**	2.2	**86**	5.4	39 (25/14)
			Fibroblasts*	**17**	**17**	1.11	**74**	**72**	4.5	**57**	2.7	**81**	3.4	907 (545/362)
AT-90	III:6	SPD001	Blood	**13**	**13**	1.21	**51**	**50**	3.3	**76**	3.6	**81**	4.9	50 (29/21)
AT-E	I:1	SPE001	Blood	**12**	**12**	0.58	**54**	**55**	2.6	**63**	1.2	**91**	2.1	28 (16/12)
AT-F	I:1	SPF001	Blood	**16**	**16**	0.62	**54**	**52**	2.5	**69**	1.9	**86**	3.4	58 (39/19)
AT-G	I:1	SPG001	Blood	**13**	**13**	0.58	**74**	**72**	3.0	**69**	3.0	**84**	2.5	17 (13/4)
AT-H	I:1	SPH001	Blood	**11**	**11**	1.05	**80**	**78**	1.5	**62**	1.4	**75**	3.2	72 (41/31)

Complete repeat lengths and conformations were accurately assessed by long-read nanopore sequencing in all DNA samples. The mean coverage depth was 185× for the total CRISPR–Cas9 targeted genomic region and 216× for the SCA37-5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract region. Wild-type alleles sequenced by long reads showed a pure WT-(ATTTT)n configuration with n ranging from 9 to 18 repeats (SD ± 0.5–1.2). Pathogenic SCA37 repeat tract consisted of 47–78 (ATTTC)n repeats (SD ± 1.5–4.5) flanked by 5′—(ATTTT)n ranging from 57 to 76 repeats (SD ± 1.2–3.6) and the 3′(ATTTT)n ranging from 75 to 91 repeats (SD ± 2.1–5.6). Repeat variability was significantly higher for the 3′(ATTTT)n (SD ± 3.9; $F_{1,15} = 24.236$, p value < 0.0001), (ATTTC)n (SD ± 3.5; $F_{1,16} = 24.478$, p value < 0.0001) and the 5′(ATTTT)n (SD ± 2.7, $F_{1,16} = 24.478$, p *value* < 0.0001) compared to the WT-(ATTTT)n repeat (SD ± 0.8). Significant correlation was observed between Sanger and Nanopore sequencing for the WT-(ATTTT)n repeat allele (Spearman's r = 0.999, p value < 0.0001 ($n = 14$) (Suppl. Figure 7a) and for the (ATTTC)n repeat in the expanded allele (Spearman's $r = 0.956$, p value < 0.0001 ($n = 14$) (Suppl. Figure 7b). Although extraction methods were not compared using the same samples, HMW DNA extraction (Sample*) performed better on average coverage than automatic Chemagen Magnetic Separation Module

Ped. Pedigree

significantly differentially methylated in SCA37 cerebellum considering f5c LLR values and identified in the top five predictive models for age of onset and disease evolution (Castro-Mondragon et al. 2022). Disease evolution was considered as the time from the onset to the time when the sample was collected.

Correlation analysis and linear regression models

All the following statistical analyses were performed using dedicated Python libraries. Two datasets were analysed. The first dataset consisted of 56 individuals from Spain and Portugal including clinical features and the repeat tract size and configuration. The second dataset consisted of nine SCA37 individuals from Spain who were sequenced by nanopore technology (eight blood samples, two skin fibroblasts samples, and two cerebellar samples). Nominal variables (gender, country/geographic area, tissue) were converted into integer labels, while numerical variables were standardized (z) considering mean (μ) and standard deviation (σ) as follows:

$$z = \frac{x - \mu}{\sigma}.$$

For regression analysis, ordinary least squares (OLS) linear regression models to examine the association between the dependent variable ("Age of onset") and all possible pairs of independent variables were used. Models with coefficient p values of less than 0.05 were selected and ranked based on the coefficient of determination (R^2) and the Bayesian Information Criterion (BIC). With n independent variables, the OLS regression model is

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \in,$$

where y is the dependent variable, β_0 is the intercept of the model, x_i corresponds to the independent variable i of the model and ε is the random error. While the intercept represents the estimated value of the dependent variable when all independent variables are equal to zero, the coefficients (slopes) represent the change in the dependent variable for each unit change in the corresponding independent variable, holding the other variables constant.

Based on the first dataset containing data of Spanish and Portuguese patients, we generated models with the dependent variables "Age of onset" and "Disease evolution", and the independent variables "Gender", "WT-(ATTTT)n", "WT total-repeat-length", "SCA37-5'(ATTTT)n", "(ATTTC)n", "SCA37-3'(ATTTT)n", "SCA37-total-repeat-length" and "Country". Based on the second dataset containing data from blood samples sequenced by long reads, we generated models with the aforementioned dependent and independent variables in addition to "Age at sample collection" and "Differences in methylation frequencies" in 86 genomic regions containing 90 CpGs identified by the methylation algorithm. Subsets of these datasets with less than three data points or with any pair of independent variables with a strong Pearson's correlation coefficient ($r > \pm 0.8$) were excluded from the model.

Appropriateness of the linear regression model was evaluated by examining the residuals versus the fitted values plot. Residuals are differences between the observed and predicted values of the dependent variable based on the estimated regression coefficients and represent the unexplained variation in the dependent variable that is not accounted for by the regression model. Regression modelling was performed using the statsmodels Python package.

It is important to note that SCA37 is a very rare disease and obtaining post-mortem samples are extremely difficult. Because of this limitation, we did not generate any predictive models from the data obtained from the cerebellar or fibroblasts samples, only from blood samples.

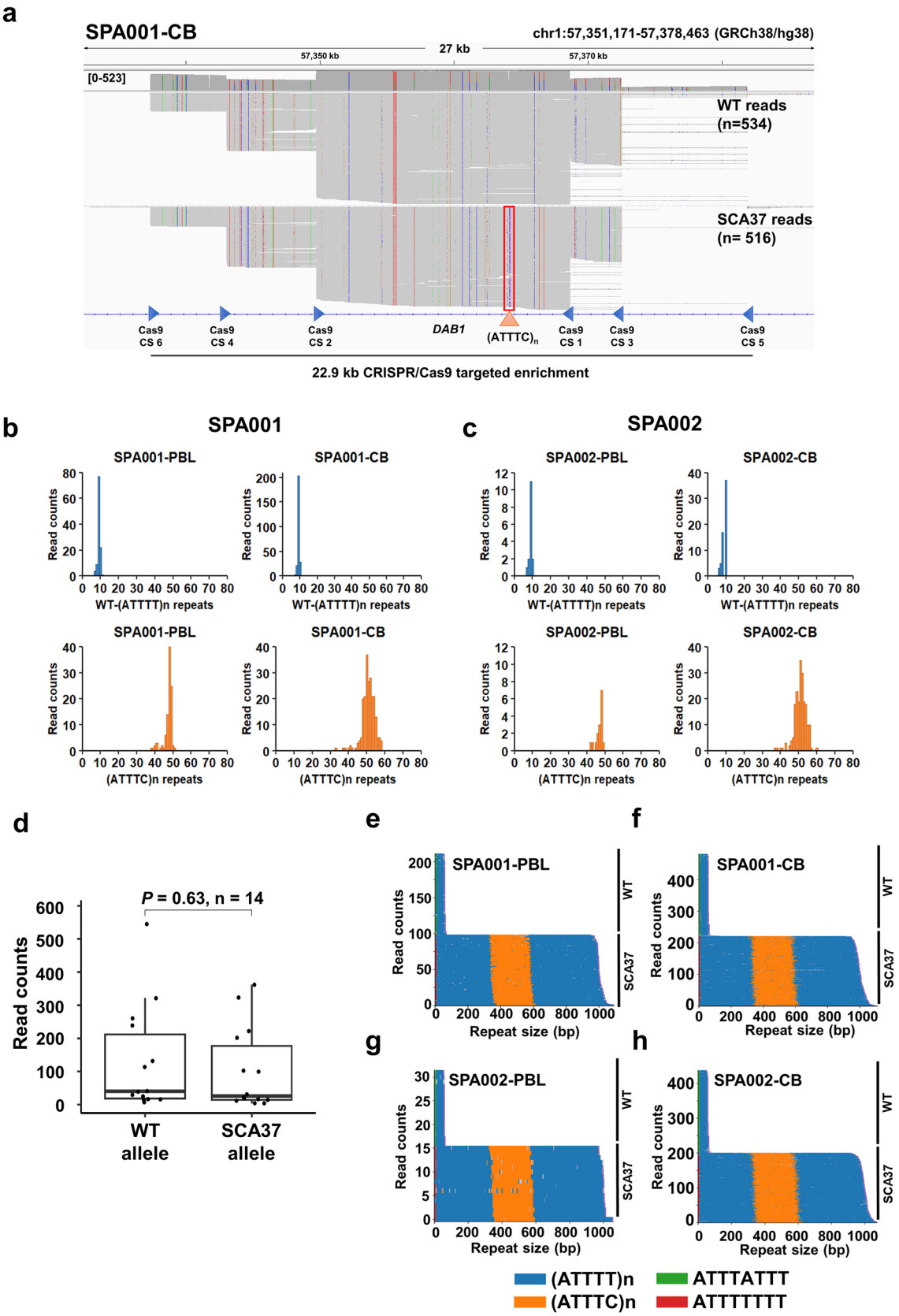
Variable importance

We determined the relative importance of the independent variables of a given regression model using the SHAP (SHapley Additive exPlanations) method (Lundberg and Lee 2017). SHAP values can be computed for individual predictions or for the overall model, allowing for global and local interpretation. The importance of a variable was determined by calculating the proportion of the model's output variability that can be attributed to each feature in the complete dataset. SHAP values were calculated using the SHAP Python library.

Estimation of the $(ATTTC)_n$ SCA37 mutation age and haplotype reconstruction

To estimate the mutations' age, 7 informative single-nucleotide polymorphisms (SNPs) located upstream and 17 located downstream of the (ATTTC)n SCA37 mutation, considering *DAB1* 5′ to 3′ direction, were used to reconstruct allele-specific haplotypes (Suppl. Figure 2). Primer sequences (Suppl. Table 1) and PCR and Sanger sequencing conditions for SNPs genotyping are included in Supplementary Information. The test for linkage disequilibrium was based on the Chi-square test.

BAM files from nanopore sequenced samples containing the on-target aligned long reads were processed with MarginPhase v1.0.0 (Ebler et al. 2019) for variant calling with a variant quality filter threshold > 30 and subsequently manually revised. WhatsHap v1.6 (Martin et al. 2016) was used for wild-type and SCA37 allele reads discrimination


a
SPA001-CB
chr1:57,351,171-57,378,463 (GRCh38/hg38)
27 kb
57,350 kb
57,370 kb
[0-523]
WT reads
(n=534)
SCA37 reads
(n= 516)
Cas9 CS 6
Cas9 CS 4
Cas9 CS 2
DAB1
(ATTTC)n
Cas9 CS 1
Cas9 CS 3
Cas9 CS 5
22.9 kb CRISPR/Cas9 targeted enrichment
b
SPA001
SPA001-PBL
SPA001-CB
Read counts
WT-(ATTTT)n repeats
(ATTTC)n repeats
c
SPA002
SPA002-PBL
SPA002-CB
d
P = 0.63, n = 14
Read counts
WT allele
SCA37 allele
e
f
g
h
SPA001-PBL
SPA001-CB
SPA002-PBL
SPA002-CB
WT
SCA37
Repeat size (bp)
(ATTTT)n
(ATTTC)n
ATTTATTT
ATTTTTTT

◂**Fig. 1** Long-read nanopore sequencing of the genomic region including the *DAB1* ATTTT/ATTTC repeat tract enriched by CRISPR/Cas9. **a** The integrative genomics viewer (IGV) showing the entire region of interest within *DAB1* intron 11 enriched by CRISPR/Cas9 successfully captured from SPA001 SCA37 patient's cerebellum. Long sequenced reads were phased using WhatsHap for haplotype reconstruction for wild-type (top) and expanded (bottom) alleles. Read counts and repeat size for WT-(ATTTT)n and SCA37-(ATTTC)n from SPA001 **b** and SPA002 **c** blood lymphocytes and cerebellar samples. **d** No allele dropout was observed in read counts of expanded alleles compared to normal alleles (two sample *t* test; *p* value = 0.63, *n* = 14). Dot point indicates outlier WT read counts for the HMW extracted SPC0001 fibroblasts. Waterfall plots generated using Guppy Sup base calling mode showed pure (ATTTT)n and (ATTTC)n repeat tracts for SPA001 and SPA002 PBLs (**e** and **g**) and cerebella **f** and **h**. No interruptions were identified in any 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeats tract (Suppl. Figure 5). Relevantly, SCA37 alleles sequenced by long reads showed an ATTTTTTT sequence preceding the 5′(ATTTT)n in SCA37 alleles in contrast to the ATTTATTT sequence preceding the WT-(ATTTT)n alleles

and haplotype reconstruction. A total of 30 SCA37 patients carrying the (ATTTC)n mutation from eight Spanish and six Portuguese unrelated kindreds were analysed. Whole-genome sequencing data previously generated from two SCA37 Spanish patients (Corral-Juan et al. 2018) were used to obtain the genotype information for 24 SNPs surrounding the (ATTTC)n mutation, spanning 6.8 Mb. Portuguese genotypes for SCA37 families (PO- M, G, R, MS, C, D) and controls were obtained from previous publication (Seixas et al. 2017). Imputation of missing genotypes and haplotype inference was performed using PHASE v2.1.1. (Stephens et al. 2001; Stephens and Donnelly 2003) with allele frequencies reported in Seixas et al. (Seixas et al. 2017) and NCBI SNP database (Sherry et al. 2001).The age of SCA37 (ATTTC)n mutation was estimated using DMLE+ v2.3 (Reeve and Rannala 2002). The relative position of the mutation was set within the haplotype considering 1 Mb = 1 cM according to physical distances given in the UCSC Genome Browser (Kent et al. 2002). Population growth rate of 0.167 was estimated based on population size, considering the oldest records from 1860 and the most recent ones for Spanish and Portuguese populations (htttp://www.ine.es/ and htttp://www.ine.pt), as reported by Russo et al. (2011). The proportion of chromosomes bearing the mutation was estimated considering a global SCA37 frequency in the Iberian Peninsula population of 106 affected chromosomes in a current population of 57959836 individuals (Slatkin and Rannala 1997). Phylogenetic networks were performed using 24 SNPs and generated with Network 10.2.0.0. (Bandelt et al. 1999) and POPTREE2 (Takezaki et al. 2010).

Results

Unbiased long-read nanopore sequencing of the (ATTTT/ATTTC)n repeat genomic region within the *DAB1* gene

We sequenced the (ATTTT/ATTTC)n repeat genomic region within the *DAB1* gene in 14 DNA samples including 10 PBL, 2 cerebellar (CB) and 2 fibroblasts (FC) samples obtained from 11 SCA37 patients of eight Spanish SCA37 kindreds. The mean age of patients at the time of sample collection was 57.15 years for PBLs (range: 47–75 years; SD ± 10.50), 72.50 for cerebellum (67 and 78 years; SD ± 7.78) and 62 for skin fibroblasts (50 and 74 years; SD ± 17). The mean age at onset was 43.67 years (range: 32–64; SD ± 9.77), while the mean of disease evolution was 18.9 years (range: 4–38; SD ± 11.2, the disease evolution for patient I:1 from AT-E was not available) (Suppl. Table 3). No information was available for I:1 and IV:9 patients from family AT-H and AT-9012, respectively. The unstable $(ATTTC)_n$ pentanucleotide repeat inserted mutation located in the SCA37 genomic region within intron 11 of the *DAB1* gene was determined by long-PCR Sanger sequencing (Table 1) with (ATTTC)n ranging from 46 to 80 repeats (average = 58.42; SD ± 11.35) and a pure WT-(ATTTT)n allele ranging from 8 to 17 (average = 11.60; SD ± 3.22).

To determine the exact size and configurations of the 5'(ATTTT)n–(ATTTC)n–3'(ATTTT)n repeat tracts, we sequenced a 22.9 kb *DAB1* intronic region containing the repeat tract by enriching the targeted region with multiple-guide CRISPR–Cas9 and nanopore long-read sequencing (Fig. 1a). We evaluated enrichment efficiency by CRISPR–Cas9 using DNA samples from two healthy controls and four SCA37 patients previously genotyped for the *DAB1* repeat insertion. All guides presented a similar editing efficiency determined by qRT-PCR after 60 min of Cas9-mediated cleavage (mean cleavage efficiency = 85.8%; SD ± 4.2) (Suppl. Figure 1 and Suppl. Table 4). No significant differences were observed in Cas9-editing efficiencies between controls and SCA37 samples, suggesting that the DNA conformation arisen by the repetitive region did not interfere with Cas9-cleavage efficiency. The *DAB1* targeted loci was successfully captured with a mean coverage depth of 185 × for the CRISPR–Cas9 targeted genomic region and 216 × considering only the SCA37 repeat tract region (Suppl. Table 5). An enrichment of 521-fold above the expected background read coverage was achieved in the targeted region (2 × mean coverage depth for the whole genome; Suppl. Table 5). Genome-wide coverage analysis found the off-target reads to be distributed randomly across all chromosomes. Coverage analysis and manual inspection

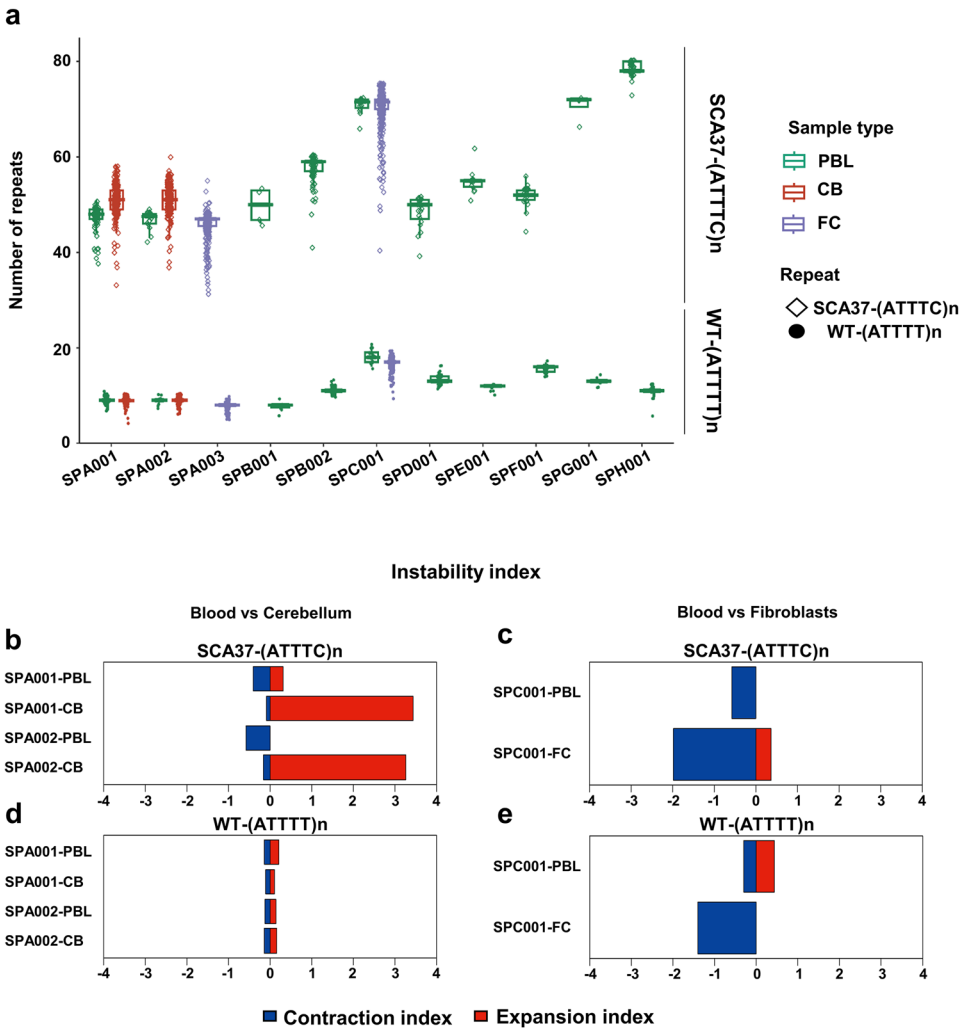


Fig. 2 Tissue-specific length, variability, and instability index for the WT-(ATTTT)n and the inserted SCA37-(ATTTC)n repeated tracts. **a** Higher repeat variability was observed for the SCA37-(ATTTC)n repeat compared to the WT-(ATTTT)n. Slightly higher instability index biased towards contraction was observed for the SCA37-(ATTTC)n repeat compared to the WT-(ATTTT)n in blood samples **b**, **d**. **b** The (ATTTC)n instability index for cerebellum (average of instability index = + 3.21) revealed an expansion-biased tissue-specific compared to blood samples (average of instability index = − 0.33). In fibroblasts, the instability index showed contraction for both the pathogenic (ATTTC)n **c** (instability index = 1.62) and the WT-(ATTTT)n (instability index = − 1.4) repeat tract **e**, compared to blood (average of instability index = − 0.33 for (ATTTC)n; average of instability index = + 0.02 for WT-(ATTT)n; Suppl. Table 8 and Suppl. Figure 10)

of off-target loci revealed that most of the regions did not encompass coding genes, but were located within centromeric, telomeric or repetitive regions, or in regions containing *in silico* predicted off-target Cas9 cuts as reported previously (Gilpatrick et al. 2020; Mizuguchi et al. 2021; Miyatake et al. 2022) (Suppl. Tables 4 and 6, Suppl. Figure 3). It is important to note that albeit the different DNA extraction methods were not compared in the same samples, high molecular weight DNAs overall performed better on sequencing coverage than samples extracted with standard automatic Chemagen Magnetic Separation (Suppl. Table 5 and Suppl. Figure 4).

DAB1 (ATTTT/ATTTC)n repeats size and configuration in SCA37

Since STRique, a Python package to analyse repeat tracts, was not designed to count complex repeats as in SCA37 (Erdmann et al. 2023), MarginPhase and WhatsHap softwares for phasing genomic variants were used in combination with RepeatAnalysisTool to determine copy numbers for the *DAB1* WT-(ATTTT)n and the pathogenic 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat alleles (Table 1; Fig. 1b, c). No allele dropout was observed in read counts of SCA37 alleles compared to normal alleles (Fig. 1d; two sample *t* test; $P = 0.63$, $n = 14$). As expected, Sup base calling mode performed with better accuracy (96.1%) compared to Hac mode (95.7%; Suppl. Table 7). The average error rates per sample were 5.8% for fast mode (SD ± 0.9), 4.3% for Hac mode (SD ± 0.5) and 3.6% for Sup mode (SD ± 0.5) (Suppl. Table 7). Waterfall plots for each sample were generated showing the repeat size, configuration and composition using Sup base calling mode in combination with RepeatAnalysisTools (Fig. 1e–h; Suppl. Figure 5). Complete repeat lengths and conformations were accurately assessed with Sup base calling mode after comparing to Hac or fast modes (Suppl. Figure 6). Wild-type alleles had a pure WT-(ATTTT)n configuration with *n* ranging from 9 to 18 repeats (SD ± 0.5–1.2). The pathogenic SCA37 repeat tract consisted of 47–78 (ATTTC)n repeats (SD ± 1.5–4.5)

flanked by 5′(ATTTT)n ranging from 57 to 76 repeats (SD ± 1.2–3.6) and the 3′(ATTTT)n ranging from 75 to 91 repeats (SD ± 2.1–5.6). Variability of repeat size number was higher for the 3′(ATTTT)n (average of the SD ± 3.9; $F_{1,15} = 24.236$, p value < 0.0001), (ATTTC)n (average of the SD ± 3.5; $F_{1,16} = 24.478$, p value < 0.0001) and 5′(ATTTT)n (average of the SD ± 2.7; $F_{1,16} = 24.478$, p value < 0.0001) all in SCA37 alleles compared to the WT-(ATTTT)n (average of the SD ± 0.8) in wild-type alleles (Table 1). By avoiding bias from PCR amplification, our strategy notably reduced repeat count variability compared with the mean standard deviations with PCR amplification and nanopore sequencing (ATTTC)n SD ± 21.86; 5′(ATTTT)n SD ± 30.61 and 3′(ATTTT)n SD ± 114.24) resulted in other studies (Rosenbohm et al. 2022).

Significant correlation was observed between Sanger and nanopore sequencing for the WT-(ATTTT)n repeat allele (Spearman's $r = 0.999$, p value < 0.0001 ($n = 14$); Suppl. Figure 7a) and for the (ATTTC)n repeat in the expanded allele (Spearman's $r = 0.956$, p value < 0.0001 ($n = 14$); Suppl. Figure 7b). No interruptions were identified in SCA37 pathogenic alleles in neither 5'(ATTTT)n, (ATTTC)n, or 3′(ATTTT)n tracts, presenting all pure repeat tracts (Fig. 1d, e; Suppl. Figure 5). Remarkably, SCA37 alleles showed an –ATTTTTTT– sequence preceding the 5′(ATTTT)n of the repeat tract. This is in contrasts with the –-ATTTATTT– sequence preceding the WT-(ATTTT)n allele (Fig. 1e–h; Suppl. Figures 5, 8). CRISPR–Cas9 targeted cleavage and long-read sequencing were able to effectively enrich the *DAB1* 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract and accurately determine the repeat size and structure in all sequenced samples.

DAB1 (ATTTT/ATTTC)n tissue-specific instability in SCA37

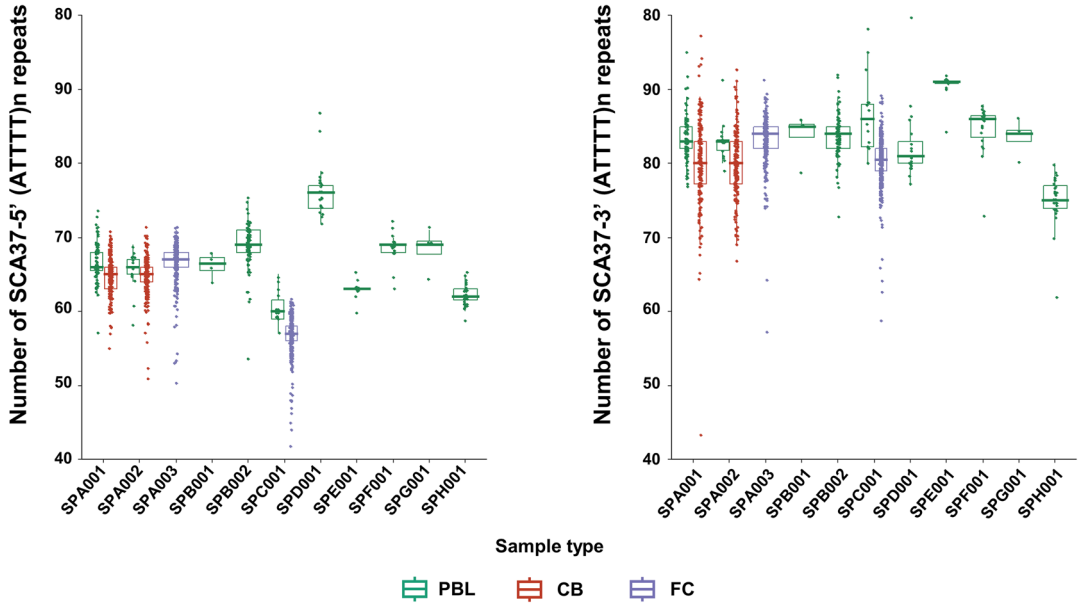
We quantified the somatic instability index for the 5'(ATTTT)n, (ATTTC)n and 3'(ATTTT)n repeat tracts in cerebellum and fibroblasts cells relative to blood cells. Higher repeat variability was observed for the (ATTTC)n repeat (average of the SD ± 3.5) compared to the WT-(ATTTT)n (average of the SD ± 0.8) (Fig. 2a; Table 1). Slightly higher instability index biased towards contraction was observed for the (ATTTC)n repeat (average of instability index =—0.33) compared with the WT-(ATTTT)n (average of instability index = + 0.02) in blood cells (Fig. 2b, d; Suppl. Figure 9 and 10,Suppl. Table 8). In addition, the (ATTTC)n instability index identified in cerebellum (average of instability index = + 3.21) revealed an expansion-biased tissue-specific compared to blood cells (average of instability index =− 0.33) (Fig. 2b). In fibroblasts the instability index showed contraction for both (ATTTC)n (instability index = 1.62) (Fig. 2c) and WT-(ATTTT)n repeats (instability index =− 1.4) (Fig. 2e) compared to blood cells (average of instability index =− 0.33 for (ATTTC)n; average of instability index = + 0.02 for WT-(ATTT)n) (Suppl. Figure 9 and 10).

Furthermore, the 3'(ATTTT)n located downstream of the (ATTTC)n pentanucleotide repeat insertion showed higher repeat variability (average of SD ± 3.9) compared to the 5′(ATTTT)n upstream variability (Average of SD ± 2.7) (Fig. 3a; Table 1). Remarkably, the 3'(ATTTT)n presented higher variability compared to the 5′(ATTTT)n, within and between Spanish, Portuguese, and German patients (Fig. 3b) (Seixas et al. 2017; Corral-Juan et al. 2018; Loureiro et al. 2019; Rosenbohm et al. 2022). Both the 5′(ATTTT)n and 3′(ATTTT)n instability indices in cerebellum revealed a contraction-biased tissue-specific compared to blood cells, being higher in 3′(ATTTT)n (cerebellar 5′(ATTTT)n instability index =− 1.59; blood 5'(ATTTT)n instability index = 0.11; cerebellar 3'(ATTTT)n instability index =− 2.5; blood 3'(ATTTT)n instability index = + 0.33) (Fig. 3c). Likewise, fibroblasts showed instability index towards contraction biased for both 5'(ATTTT)n and 3′(ATTTT)n repeats compared to blood cells (fibroblasts 5′(ATTTT)n instability index =− 3.07; fibroblasts 3′(ATTTT) instability index =− 1.53) (Fig. 3d; Suppl. Figure 9 and 10, Suppl. Table 8). These data evidence the sizing and configuration accuracy of the new implemented method, the absence of interruptions in SCA37 pathogenic alleles and an expansion bias of the (ATTTC)n repeat and contraction bias of the 5'(ATTTT)n and 3'(ATTTT)n in SCA37 cerebellum relative to blood.

Differential *DAB1* methylation signature in the SCA37 cerebellum

A total of 55 informative SNPs within the on-target region were identified by nanopore sequencing and used to discriminate between wild-type or SCA37 allele reads within the on-target region (Suppl. Table 9). Differential CpG methylation frequencies (DMF) between wild-type and SCA37 alleles were quantified for blood cells ($n = 10$), fibroblasts ($n = 2$) and cerebellum ($n = 2$). Methylation analysis with the f5c software identified 86 potential methylated regions including 90 CpGs within the on-target region and 30 differentially methylated CpGs between SCA37 and wild-type cerebellar alleles (p value < 0.05), with 12 of them locating upstream and 18 downstream of the 5′(ATTTT)n–(ATTTC)n–3'(ATTTT)n repeat tract (Suppl. Table 10). In cerebellum, three distinctive differentially methylated regions (DMR) adjacent to the 5'(ATTTT)n–(ATTTC)n–3'(ATTTT)n repeat tract were identified (Fig. 4a, d): one upstream hypomethylated region (DMR1 =− 8.84% ranging from chr1:57367323 to chr1:57371263) and two downstream hypermethylated regions (DMR2 = 5.51% ranging from chr1:57364049

a

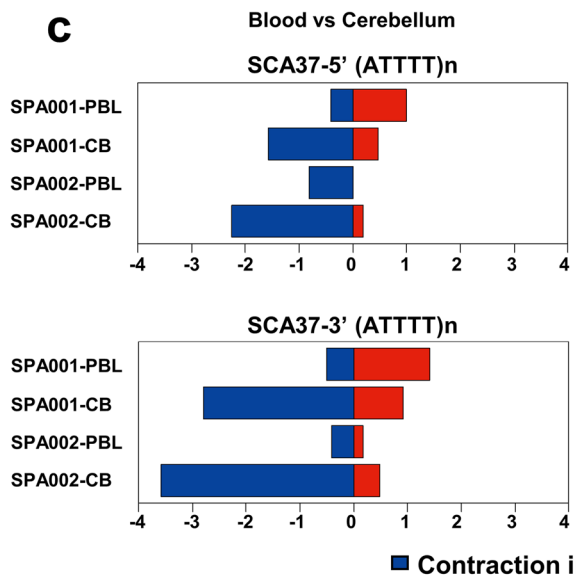


b

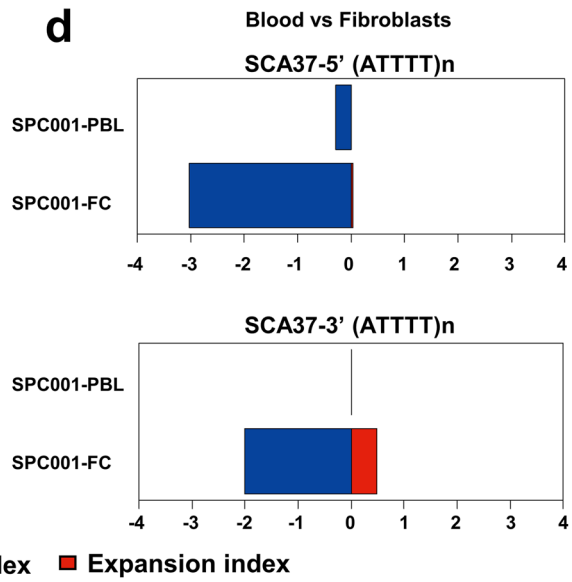
Geographic origin	WT-(ATTTT)n repeats	SCA37 5' (ATTTT)n repeats	SCA37 (ATTTC)n repeats	SCA37 3' (ATTTT)n repeats
Spanish	7-99	57-76	45-81	75-91
Portuguese [a]	7-400	60-79	31-75	58-90
German [b]	40-60	72-79	39-102	408-420

Instability Index

c



d



◂**Fig. 3** Repeat instability of the 5′(ATTTT)n upstream and 3′(ATTTT)n downstream of the inserted (ATTTC)n repeat in the mutant pathogenic allele. **a** The 3′(ATTTT)n located downstream of the (ATTTC)n pentanucleotide repeat insertion (right) showed higher repeat variability (average SD ± 3.9) compared to the upstream 5′(ATTTT)n (average SD ± 2.7) (left) in the SCA37 allele. **b** Remarkably, the 3′(ATTTT)n presented the highest repeat variability between Spanish (75–91), Portuguese (58–90), and German (408–420) cases. **c** The instability index in both the 5′(ATTTT)n and 3′(ATTTT)n flanking the (ATTTC)n in the pathogenic alleles in cerebellum revealed a contraction-biased tissue-specific compared to blood samples (cerebellar 5'(ATTTT)n instability index = − 1.59; blood 5′(ATTTT)n instability index = 0.11; cerebellar 3'(ATTTT)n instability index = − 2.5; blood 3′(ATTTT)n instability index = + 0.33). **d** In fibroblasts, the instability index also showed contraction biased for both 5′(ATTTT)n and 3′(ATTTT)n repeated tracts in pathogenic alleles compared to blood (fibroblasts 5'(ATTTT)n instability index = − 3.07; fibroblasts 3′—(ATTTT) instability index = − 1.53) (Suppl. Table 8 and Suppl. Figure 10)

to chr1:57367009; and DMR3 = 5.34% ranging from chr1:57,361,325 to chr1:57,363,731).

No relevant DMF (> 5%) between wild-type and SCA37 alleles were observed in blood in the same genomic regions (Fig. 4b, d). Fibroblasts presented a DMF of 6.44% for DMR1 (hypermethylation), a DMF of 13.43% for DMR2 (hypermethylation) and a DMF of 1.63% for DMR3, evidencing overall tissue-specific methylation (Fig. 4c, d) (Battaglia et al. 2022). Based on this evidence, we propose that the cerebellar-specific methylation signature identified in SCA37 alleles in this study may underlie DAB1 dysregulation shown in cerebellum of SCA37 patients (Corral-Juan et al. 2018). However, extending the methylation study including additional SCA37 alleles from control individuals and SCA37 patients would confirm the observed tissue-specific methylation effects identified.

Novel SCA37 genotype–phenotype correlations

We sought to assess the clinical relevance of the exact configuration of the complex repeated elements within the ATTTT/ATTTC *DAB1* genomic region. To this aim, we evaluated the presence of possible clinical relationships between "Age of onset" or "Disease evolution" with a series of genetic variables "WT-(ATTTT)n", "WT total-repeat-length", "SCA37-5'(ATTTT)n", "SCA37-(ATTTC)n", "SCA37-3'(ATTTT)n", "SCA37-total-repeat-length") and "Gender" and "Country of origin". This information was available from 56 SCA37 patients, 26 from Spain and 30 from Portugal (Suppl. Table 11). We performed this analysis for wild-type and SCA37 alleles. For the "Age of onset", negative linear correlations with the "(ATTTC)n" ($r = -0.572$; p value $= 1.452 \times 10^{-5}$; $n = 56$) and "Country of origin" (r = -0.356; p value $= 7.265 \times 10^{-3}$; $n = 56$) were found (Fig. 5a; Suppl. Table 12). Moreover, the "3′(ATTTT)n" (r = 0.458; p value $= 2.839 \times 10^{-2}$; $n = 22$) and "Gender" (r = 0.349; p value $= 9.141 \times 10^{-3}$; $n = 56$) positively correlated with "Age of onset" (Fig. 5a; Suppl. Table 12). Although no significant correlations were found for "Disease evolution" probably due to the small sample size, a moderate positive correlation with "5'(ATTTT)n" ($r = 0.371$; p value $= 6.142 \times 10^{-2}$; $n = 29$) was detected.

These results suggest the influence of the (ATTTC)n, geographic origin, gender and the 3'(ATTTT)n on the age of onset in SCA37. Remarkably, this is the first evidence of the phenotypic contribution of the 3'(ATTTT)n number size downstream the (ATTTC)n mutation in SCA37.

SCA37 regression models for predicting age of onset and disease evolution

A series of predictive linear regression models were generated using all possible combinations of up to five variables from the CpG-identified regions. These models were selected based on the significance of their coefficients (p value < 0.05) and ranked based on the coefficient of determination (R^2) and the Bayesian information criterion (BIC), which favours simpler models compared to other popular model selection approaches, such as the Akaike information criterion (AIC).

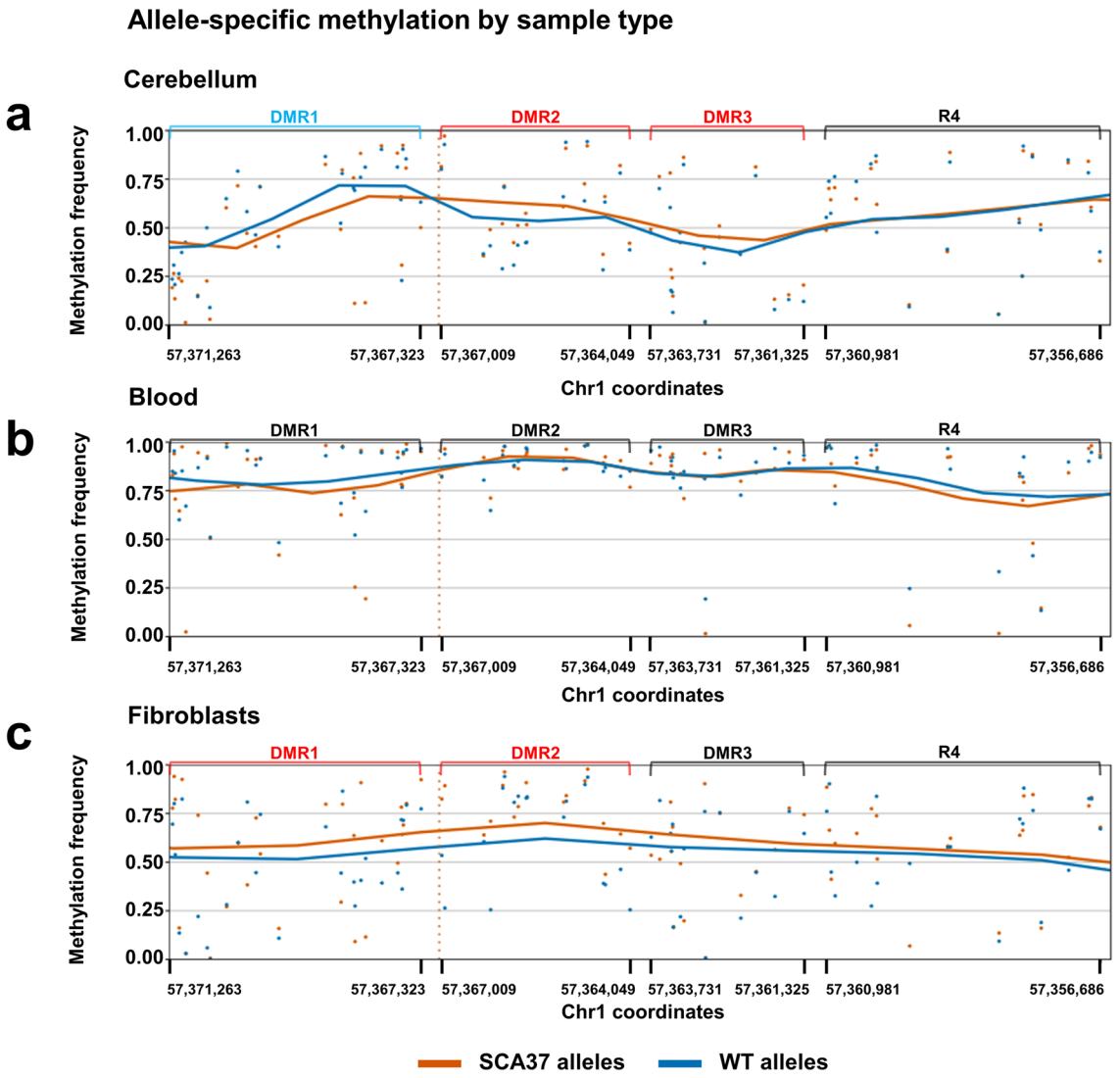
As expected, and confirming our previous observations (Corral-Juan et al. 2018), the regression analysis identified a model of association considering "Age of onset", "Gender" and the "(ATTTC)n" repeat size ($R^2 = 0.438$, $n = 56$; p value < 0.039) (Fig. 5b–d, Suppl. Table 13) with an equation model as follows:

$$Age\ of\ onset = -0.2692 + (0.7053\ x\ \ Gender) + (-0.5874\ \ x\ (ATTTC)n),$$

using a value of 0 for females and a value of 1 for males in the "Gender" variable. The analysis of the SHAP values of this model resulted in a contribution of 58.76% of the "(ATTTC)n" and 41.24% of "Gender" to the "Age of onset". Additionally, a significant association between "Age of onset" and the "Country of origin" was found with a low coefficient of determination ($R^2 = 0.127$, $n = 56$; p value < 0.046), but not in combination with other variables (Suppl. Table 13). In contrast, no significant models were found for "Disease evolution" (Suppl. Table 14).

The same analysis was applied considering the differences in methylation frequencies detected in blood samples in the specific genomic SCA37 region flanking the 5'(ATTTT)n–(ATTTC)n–3'(ATTTT)n repeat tracts as additional independent variables.

Sixty-four out of 86 potentially methylated CpG regions were identified. The variables "Disease evolution" and "Age of onset" were considered when available. A series of linear regression models were constructed with all possible



d

	DMR1 (1:57,367,323- 1:57,371,263)	DMR2 (1:57,364,049-1:57,367,009)	DMR3 (1:57,361,325-1:57,363,731)
Cerebellum	-8.84 %	5.51 %	5.34 %
Blood	-4.71 %	1.58 %	-0.91 %
Fibroblasts	6.44 %	13.43 %	1.63 %

Fig. 4 CpG methylation signatures of the SCA37 region within *DAB1* on 1p32 in cerebellum, peripheral blood cells and fibroblasts. For methylation studies, SCA37 and wild-type alleles were classified using WhatsHap software. **a** Similar allele-specific methylation signatures were present in two SCA37 cerebellar samples (SPA001-CB and SPA002-CB) compared to blood **b** and fibroblasts **c**. **d** Three global differentially methylated regions (DMR) were identified in SCA37 pathogenic alleles compared to WT-(ATTTT)n alleles. An hypomethylated region (R1) upstream of the 5'(ATTTT)n–(ATTTC)n–3′(ATTTT)n SCA37 tract showed a 8.84% mean reduction of methylation frequencies ranging from "chr1:57367323" to "chr1:57371263". In contrast, two regions, R2 and R3 downstream of the SCA37 repetitive tract, were found differentially hypermethylated compared to WT alleles, increasing 5.51% (R2; ranging from "chr1:57364049" to "chr1:57367009") and 5.34% (R3; ranging from "chr1:57361325" to "chr1:57363731") their global methylation frequencies. Blood samples did not reveal significant differences in methylation frequencies between WT and SCA37 alleles **b**. **c** Fibroblasts showed a slightly global increase of methylation frequencies in the SCA37 alleles compared to WT alleles with an increase of 6.44% in R1 (ranging from "chr1:57361325" to "chr1:57363731") and 13.44% in R2 (ranging from "chr1:57364049" to "chr1:57367009"). The y-axis represents methylation frequencies shown in percentage and the genomic positions represented on the x-axis indicates DMR coordinates. Blue and orange lines represent smoothed methylation frequencies for wild-type and SCA37 disease alleles, respectively. The position of the 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n SCA37 tract is represented with a red vertical dotted line

combinations of up to three variables to avoid combinatorial explosion and restrain model complexity. Model selection and evaluation criteria were the same as the previous analysis. Regression analysis identified several models of the dependent variable "Age of onset" (Suppl. Table 15) or "Disease evolution" (Suppl. Table 16). To navigate the sheer amount of models that were generated, we ranked those with the highest number of data points based on lowest BIC and highest R^2 (Fig. 5e, f), and selected the best among the top five models for further dominance analysis. The best model for age of onset ($R^2 = 0.998$, $n = 8$; p value < 0.0007; Fig. 5g, h) was found to include the variable 3'(ATTTT)n (contributing 31.38% to the model) and the CpG regions "chr1:57361330" in DMR3 (contributing 60.31%) and "chr1:57360976" in R4 (contributing 8.32%). The best model of disease evolution ($R^2 = 0.999$, $n = 7$; p value < 0.0008; Fig. 5 i, j) was obtained with the CpG regions "chr1:57367557" in DMR1 (contributing 25.11% to the model), "chr1:57362080" in DMR3 (contributing 28.74%), and "chr1:57360845" in R4 (contributing 46.15%).

Besides these most significant five predictive models, it is important to highlight the "Age of onset" model identified using the "(ATTTC)n" and the two CpG regions "chr1: 57,367,004" and "chr1: 57,365,681" ($R^2 = 0.932$, $n = 8$; p *value* < 0.011; Fig. 5k, l), both located in DMR2 and which were also identified contributing together in other 7 significant prediction models(Suppl. Table 15; significant prediction models 23, 31, 35, 39, 43, 45 and 64). Moreover, the analysis also revealed a model of "Disease evolution" associated with the independent variable "(ATTTC)n", but including the combination of two different CpG regions "chr1:57370049" and "chr1:57368270" both in DMR1 ($R^2 = 0.98$, $n = 7$; p value < 0.0027; Fig. 5 m and 5n). These mathematical predictive models establish the basis for a first attempt towards individualized medicine in SCA37.

Variable importance identified in linear regression models in SCA37

Considering "Age of onset" and "Disease evolution" associated with methylation signatures, we analysed each variable importance based on SHAP values after selecting the top 15 best predicting CpG regions based on F-regression, which is a rapid linear model for assessing the potential impact of individual regressors one by one, in a sequential manner, without incurring into combinatorial explosion. Prediction models of "Age of onset" and "Disease evolution" using these 15 genomic methylated signatures as independent variables were created and their relative importance based on SHAP values were calculated (Fig. 6a, b). The most important methylation CpG region in the model of "Age of Onset" identified is "chr1:57367607" in DMR1 (15.17%), followed by "chr1:57359079" in R4 (14.55%), while that in the model of "Disease evolution" is "chr1:57367636" (17.20%) and "chr1:57371220" (13.23%) both in DMR1. While the other CpG regions may not seem important in a model solely based on methylation signatures, they became highly prominent when combined with other clinical variables. In particular, CpG regions "chr1:57361330" in DMR3 and "chr1:57365870" in DMR2, significant differentially methylated in SCA37 cerebellar alleles and included in the list of important features for methylation models, also emerged in the top five models for "Age of onset" when utilizing variables of any type (Fig. 5e), such as a model of age of onset involving both CpG regions "chr1:57361330" (49.96%) and "chr1:57365870" (11.29%), and the "3'(ATTTT)n" repeat tract (38.75%) (Fig. 6c; Suppl. Table 15). The analysis of the top 15 best predicting CpG regions (Fig. 6a, b) in the JASPAR transcription factor database revealed four transcription factor binding sites at the same strand and transcription synthesis direction from the *DAB1* gene for the NOBOX, PRDM1, PAX4 and LHX3 transcription factors. Strikingly, PRDM1 transcription factor was found to interact with 82Q ATXN1, the mutant protein in spinocerebellar ataxia type 1 (SCA1), and has been found to be expressed at the granule and Purkinje cells of the cerebellum during chicken embryonic and germline development (Lim et al. 2006; Wan et al. 2014b). This supports the widely accepted hypothesis that common molecular signalling alterations underlie cerebellar neurodegeneration in spinocerebellar ataxias.

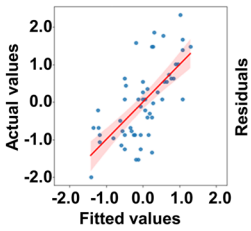
Common origin and age of the SCA37 $(ATTTC)_n$ mutation

Long-read sequencing of targeted enriched CRISPR–Cas9 SCA37 expanded alleles including the (ATTTC)n repeat mutation and the flanking regions from 11 Spanish patients from eight unrelated kindreds confirmed a common haplotype spanning a 22.29 kb region within intron 11 of the *DAB1* gene (Suppl. Table 17). A combination of SNP-genotyping and PHASE inference of 24 SNPs adjacent to the (ATTTC)n expanded insertion was used to elucidate a possible SCA37 ancestral haplotype. Additional data from four SCA37 Spanish and 16 Portuguese patients from 14 unrelated kindreds were also included for haplotype inference. Haplotype reconstruction revealed a 964-kb shared region in linkage disequilibrium flanking the (ATTTC)n mutation in SCA37 chromosomes in all SCA37 patients (Fig. 7a; linkage disequilibrium p value < 0.00001; Suppl. Table 17), evidencing a common origin of the SCA37 mutation in the Iberian Peninsula. The patient I:1 AT-F SP-F1 individual shared the complete haplotype except for SNP rs954450605, indicating that a single-nucleotide substitution occurred later in this marker. A total of five distinctive haplotypes were identified in all 30 SCA37 Iberic patients (Fig. 7a). Twenty-two patients from five Spanish and four Portuguese kindreds

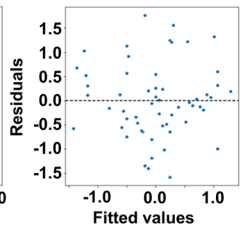
a

Dependent variable	Independent variable	Pearson's correlation coefficient (r)	*p*-value
Age of onset (years)	SCA37-(ATTTC)n	-0.572	1.45×10^{-5}
	SCA37-3'(ATTTT)n	0.458	2.84×10^{-2}
	Country [a]	-0.356	7.27×10^{-3}
	Gender [b]	0.349	9.14×10^{-3}

b



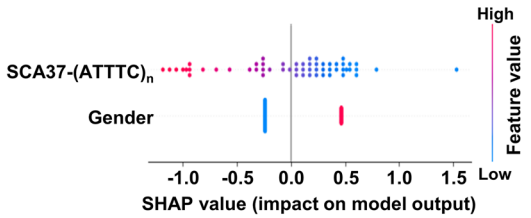
c



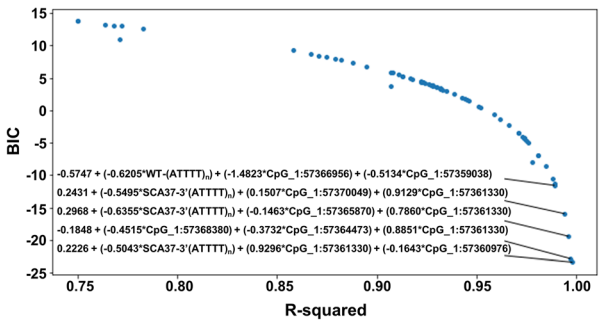
$$Age\ of\ onset = -0.27 + (0.71 \times Gender) + (-0.59 \times SCA37\text{-}(ATTTC)_n)$$

n = 56; R-squared = 0.438; BIC = 138.698

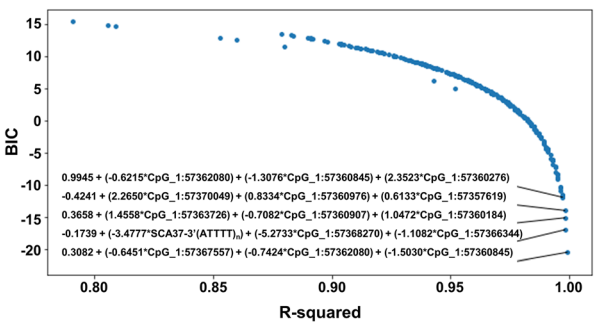
d



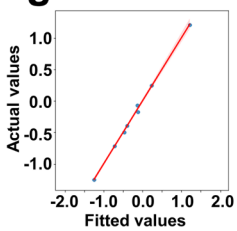
e



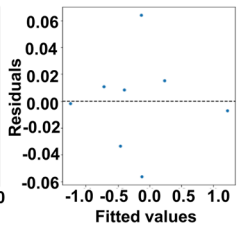
f



g



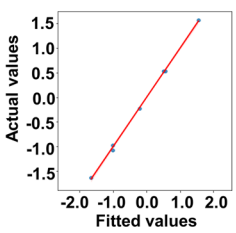
h



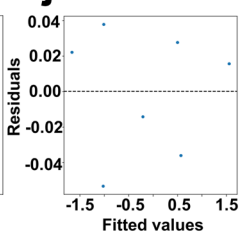
$$Age\ of\ onset = 0.22 + (-0.50 \times SCA37\text{-}3'(ATTTT)n) + (0.93 \times CpG\_1{:}57361330) + (-0.16 \times CpG\_1{:}57360976)$$

n = 8; R-squared = 0.998 ; BIC = -23.39

i



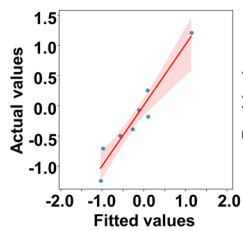
j



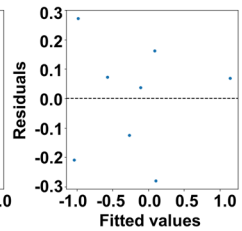
$$Disease\ evolution = 0.31 + (-0.65 \times CpG\_1{:}57367557) + (-0.74 \times CpG\_1{:}57362080) + (-1.50 \times CpG\_1{:}57360845)$$

n = 7; R-squared = 0.999; BIC = -20.45

k



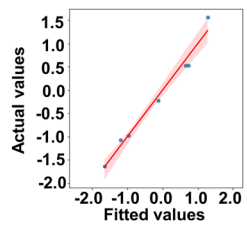
l



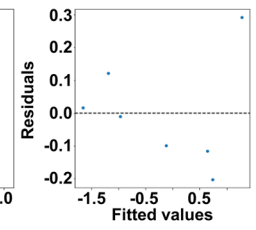
$$Age\ of\ onset = -0.52 + (-0.33 \times SCA37\text{-}(ATTTC)_n) + (0.77 \times CpG\_1{:}57367004) + (-1.01 \times CpG\_1{:}57365681)$$

n = 8; R-squared = 0.932; BIC = 3.29

m



n



$$Disease\ evolution = -1.24 + (-0.42 \times SCA37\text{-}(ATTTC)_n) + (1.22 \times CpG\_1{:}57370049) + (-3.04 \times CpG\_1{:}57368270)$$

n = 7; R-squared = 0.926; BIC = 1.39

◂**Fig. 5** Novel genotype–phenotype associations and predictive linear regression models established in SCA37. **a** Significant Pearson correlation coefficients were obtained using "Age of onset" as dependent variable and the "SCA37-(ATTTC)n" ($n=56$), "SCA37-3′(ATTTT)n" ($n=22$), "Country" (n = 56) and "Gender" ($n=56$) as independent variables. **b** and **c** Scatterplot based on the regression model of the dependent variable "Age of onset" and the independent variables "Gender" and "(ATTTC)n". **b** an additional linear regression model (red line) with confidence interval of 95% (red shadow) plotted over the observed ("actual") values of the dependent variable (z-scores) and their predicted ("fitted") values. **c** the residuals of the regression model are plotted against the predicted values of the dependent variable. **d** A bee swarm plot summarizing the distribution of SHAP values for each variable of the regression model is shown. Male gender and shorter "(ATTTC)n" have higher impact value in the age of onset prediction model (pink blots) than female gender and longer "(ATTTC)n" (blue dots). Rank of the selected models of age of onset **e** and disease evolution **f** using the dataset reporting the methylated CpG regions with the five most relevant models (lowest BIC and highest R^2) indicated. **g** and **h** The best model for age of onset ($R^2=0.998$, $n=8$; p value <0.0007) was found to include the variable 3′(ATTTT)n and the CpG regions "chr1:57361330" in DMR3 and "chr1:57360976" in R4. **i** and **j** The best model of disease evolution ($R^2=0.999$, $n=7$; p value <0.0008) was obtained with the CpG regions "chr1:57367557" in DMR1, "chr1:57362080" in DMR3, and "chr1:57360845" in R4. **k** and **l** The best prediction model for "Age of onset" considering the "(ATTTC)n", includes two CpG regions "chr1: 57367004" and "chr1: 57365681", both located in DMR2 ($R^2=0.932$, $n=8$). **m** and **n** A significant model of "disease evolution" associated with the independent variable "(ATTTC)n", and the combination of two different CpG regions "chr1:57370049" and "chr1:57368270", both located in DMR1 ($R^2=0.926$, $n=7$)

shared the most frequent haplotype in the cohort spanning 3.14 Mb in the *DAB1* genomic region (Fig. 7a). Considering a disease prevalence of $f=0.14$, an Iberic population growth rate of $r=0.17$, and considering each generation to span 25 years, the mutation was estimated to have occurred 859 years ago (95% CI 647–1,378; Fig. 7b). The estimation of the mutation age considering different demographic and disease prevalence parameters for the DMLE + 2.3 software generated comparable results (Suppl. Figure 11). Likewise, considering the southwest of the Iberian Peninsula instead of the whole Iberic region, the mutation was estimated to have occurred approximately 917 years ago (95% CI: 656–1404) (Suppl. Figure 11). Phylogenetic relationship among five different SCA37 haplotypes were generated showing the most parsimonious relationship between them considering 24 informative SNPs in 14 SCA37 families (Fig. 7c, d). A common haplotype (HAP2) was shared by five Spanish and four Portuguese families. In the Spanish AT-901 family this haplotype resulted from a recombination event from HAP1 haplotype (Supplementary Table 17; (Corral-Juan et al. 2018)) HAP3 and HAP4 haplotypes correspond to one Spanish and one Portuguese families, respectively. Relevantly, HAP5 haplotype is shared by two Spanish and one Portuguese families and appeared from a recombination event in the Spanish AT-9012 family (Supplementary Table 17; (Corral-Juan et al. 2018)).

Discussion

In the present study, we accurately determine the pathogenic *DAB1* 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract size and configuration in SCA37 using an unbiased non-PCR amplified long-read sequencing of the CRISPR–Cas9 targeted enriched locus. By sequencing long fragments of native gDNA spanning the *DAB1* repeat expansion, we were able to identify a differential SCA37 methylation signature in SCA37 alleles in cerebellum. We could also establish novel genotype–phenotype associations considering molecular and disease's variables such as "Age of onset" and "Disease evolution" by generating predictive regression models in SCA37. Long-read sequencing of the (ATTTC)n genomic region in combination with 24 SNPs genotyping proved a common haplotype spanning a 964 kb genomic region within *DAB1* intron 11 in all SCA37 expanded alleles from 30 Spanish and Portuguese SCA37 patients demonstrating a common origin of the SCA37 mutation in the Iberian Peninsula. The SCA37 mutation was estimated to have occurred 859 years ago (95% CI 647–1,378).

To date, long-PCR followed by Sanger sequencing for alleles with repeats of moderate size has been the method of choice for genetic diagnosis of SCA37 (Matilla-Dueñas and Volpini 1993). However, this is challenging when sequencing long repeat tracts. By using unbiased non-PCR amplified nanopore sequencing and CRISPR–Cas9 targeted enrichment of the SCA37 mutation region we could unequivocally size and determine the configuration of the 5′(ATTTT)n, (ATTTC)n and 3′(ATTTT)n repeat tracts, identifying a novel association between disease age of onset and the 3´(ATTTT) repeat size in SCA37 alleles. This highlights the importance of accurately determining the exact configuration of the pathogenic alleles to establish accurate genotype–phenotype correlations in SCA37. Importantly, we detected the highest sequencing coverage when fresh high molecular weight DNA was obtained during gDNA extraction. This is of relevance for obtaining consistent accurate repeat sequences and methylation signatures by nanopore sequencing.

Tandem repeat expansions (TREs) have been described to be causative of at least 60 human monogenic disorders, including psychiatric, neurodevelopmental, neuromuscular, and neurodegenerative disorders (Wen et al. 2023). Implementing an accurate method for sequencing long-tandem repeats without PCR bias removes experimental variability, overcoming the technical limitations for genetic diagnosis of those disorders and enabling genotype–phenotype correlations needed in precision medicine. A recent study has described the first SCA37 patients identified outside the Iberian Peninsula by using biased long-range PCR followed by long-read nanopore sequencing to determine the repeat size and structure of the 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n

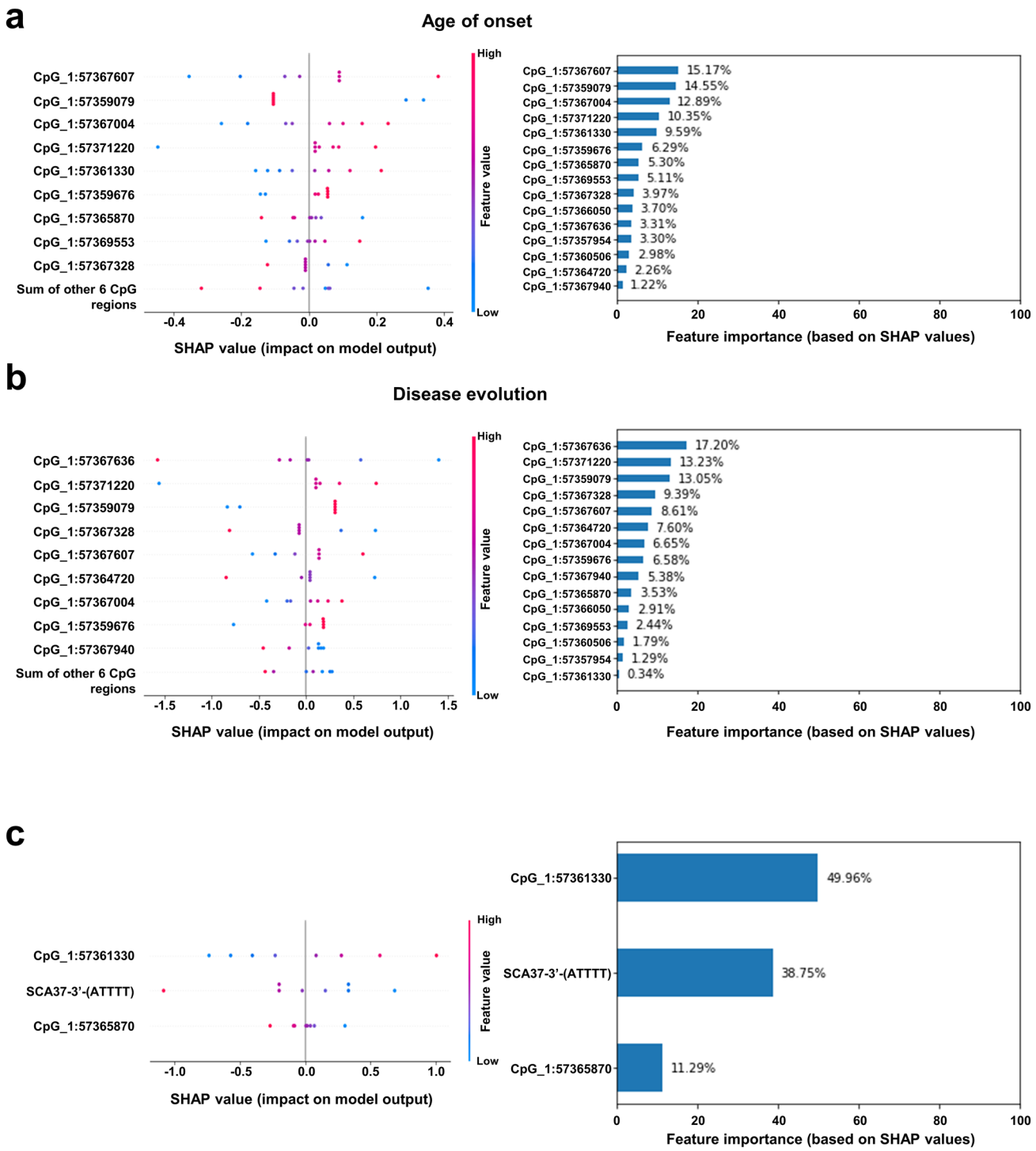


Fig. 6 Importance value and impact of methylated CpGs regions in predictive models. SHAP values and relative importance of the 15 most recurrent methylated CpG regions in regression models of "Age of onset" **a** and "Disease evolution" **b**. **c** SHAP values and relative importance of a model of age of onset using 3′(ATTTT)n and CpG regions "chr1:57365870" in DMR2 and "chr1:57361330" in DMR3 (See Fig. 5e)

repeat tract. In this case, the exact size of the repeat could not be precisely determined due to PCR amplification bias and high experimental variability (Rosenbohm et al. 2022). In contrast, in the present study targeted CRISPR–Cas9 enrichment of the native DNA including the SCA37 repeat region in combination with long-read nanopore sequencing obtained an average base accuracy rate of 96.4% with a mean sequencing error of 3.6% (range: 3.01–4.62%) for an average

521-fold enrichment. By avoiding bias from PCR amplification, our strategy notably reduced repeat count variability compared to other studies (Rosenbohm et al. 2022). Another important limitation of nanopore sequencing of PCR amplicons containing TREs is that it cannot reveal allele-specific methylation signatures such as the ones identified in our study.

Remarkably, we found that the 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tracts in all SCA37 alleles were preceded by an ATTTTTTT sequence. In contrast, the (ATTTT)n repeat tract in wild-type alleles were found to be preceded by ATTTATTT. ATTT motifs are switch regulatory elements (SREs), and were identified as tetrameric targets of POU-family transcription factors located on promoter regions acting as cis-regulatory elements repressing gene transcription (Schaffer et al. 2003; Alazard 2005; Lachman et al. 2006). Moreover, TTTATTTA sequences form highly compact native DNA structures called mini-dumbbells (MDBs) implicated in stabilizing DNA structure potentially affecting protein binding and DNA translation and replication (Guo and Lam 2016). Likewise, replacement of the TTTATTTA sequence by ATTTTTTT in SCA37 alleles would influence DNA secondary structure providing repeat instability and transcription dysregulation.

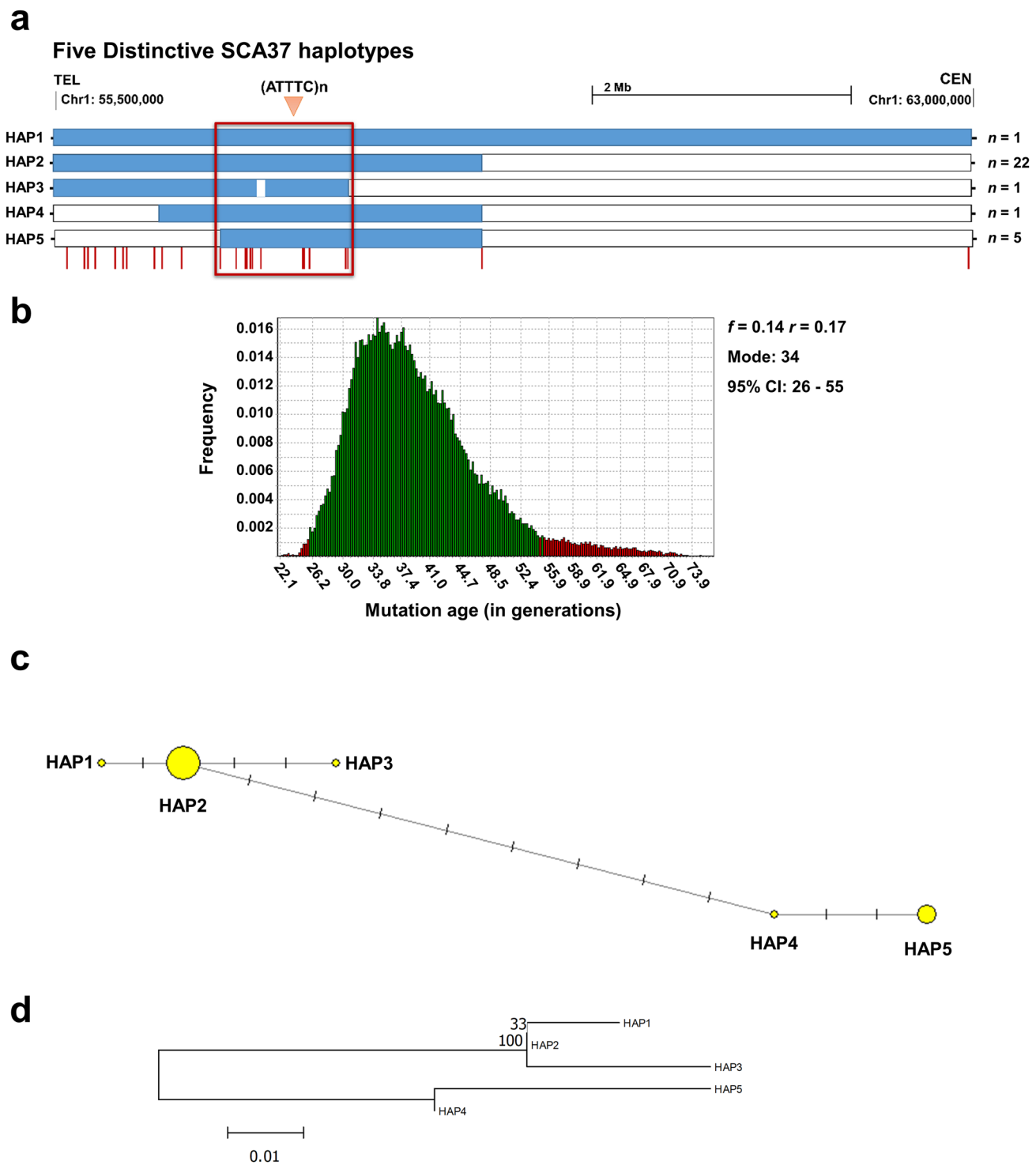
Repeat interruptions contribute to stabilizing repeat tracts of non-pathogenic alleles during somatic and germinal transmission (i.e. SCA1, SCA2, HTT, etc.), whereas in *RFC1*, *SCA10*, *SCA27B* and *SCA31* genes, interruptions are present in expanded alleles where the effects on repeat stability are unclear (Moseley 2000; Richards 2001; Sobczak and Krzyzosiak 2005; Wright et al. 2019; Wilke et al. 2023). In our study, all the (ATTTT)n and (ATTTC)n repetitive motifs in SCA37 chromosomes were uninterrupted, whereas one out of 29 of wild-type alleles presented AT-rich interruptions such as $(ATTTT)_2AT(ATTTT)_{20}A(ATTTT)_{22}(ATTT)(ATTTT)_4(ATTT)(ATTTT)_2$ similar to the one previously reported (Loureiro et al. 2019).

In most of the repeat expansion diseases, the expansion is unstable during somatic and germline transmissions triggered by DNA synthesis or repair errors (Matsuura et al. 2004; Kacher et al. 2021). Those errors apply for most of instable repeat tracts, depending on repeat size, configuration, genomic location or cell type (Khristich and Mirkin 2020; Mouro Pinto et al. 2020). In the present study, the analysis of the different repeat motifs conforming the 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n in SCA37 alleles revealed differential repeat instability when compared in blood, cerebellar, and fibroblasts cells. Cerebellar samples from two SCA37 patients presented higher level of somatic instability of the 5’(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract, with an increase in the number of (ATTTC)n, and a decrease of the 5′(ATTTT)n and 3′(ATTTT)n repeats, compared to PBLs. No instability was observed when comparing WT-(ATTTT)n alleles in the three different cellular types. Relevantly, fibroblasts cells presented a decreased number of all three pentanucleotide motifs in the SCA37 repeat tract in pathogenic alleles when comparing with PBLs. Tissue-specific somatic instability has been identified in a few TREs associated with other neurodegenerative diseases such as Huntington’s disease, where the CAG repeat shows a lower instability in spinal cord and cerebellum, or SCA1 and SCA3 with a lower degree of mosaicism found in the cerebellar cortex for the CAG repetitive tracts compared to other CNS regions (Cancel et al. 1998; Kraus-Perrotta and Lagalwar 2016; Mouro Pinto et al. 2020). In contrast, *FXN* GAA expanded repeats in Friedreich’s ataxia patients show an expansion bias in the cerebellum (De Biase et al. 2007), and skeletal muscle in Myotonic Dystrophy Type 1 present with much larger *DMPK* CTG expansions (Thornton et al. 1994). In the present study, somatic instability revealed tissue-specificity instability of the 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract in the affected SCA37 cerebellum. It appears that somatic instability depends on the type and length of the expanded repeat, increases with age, and is also influenced by DNA replication and the repair genes implicated in each particular cell type (Pearson et al. 2005).

A relevant and novel contribution of the present study is the identification of specific methylation signatures in SCA37 disease alleles that are supportive of dysregulated expression of the *DAB1* gene observed in SCA37 cerebellum (Corral-Juan et al. 2018). Likewise, Fragile X-associated tremor/ataxia syndrome (FXTAS) is characterized by cerebellar ataxia and presents with an unmethylated repeat expanded mutation leading to increased expression of the gene product and Purkinje cell loss in the cerebellum (Kenneson 2001).

We generate for the first time a predictive mathematical model for the age of onset and diseases evolution in SCA37 considering the importance of genetic variables such as the size and configuration of the complex 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract. To the best of our knowledge there are only a few studies using mathematical models and machine learning algorithms to predict clinical outcomes in other spinocerebellar ataxias including SCA1, SCA2, SCA3 and SCA6 (Tezenas du Montcel et al. 2014; Peng et al. 2021; Ru et al. 2022; Hatano et al. 2023). Predictive outcome models are becoming very useful tools for genetic counselling, clinical prognosis, and response follow-up of therapeutic treatments.

In this study, we have identified an identical 964 kb haplotype found in linkage disequilibrium flanking the $(ATTTC)_n$ repeat inserted mutation (chr1:56785142–57749488, hg38) shared by all SCA37 patients studied from 14 Spanish and Portuguese kindreds, indicating that an ancestral chromosome is responsible for



all SCA37 cases in the Iberian Peninsula described to date. Common ancestral origins and mutation founder effects have been described in several SCAs (Verbeek et al. 2004; Sequeiros et al. 2012), such as SCA2 in Cuba (Velázquez Pérez et al. 2009), SCA3 in Azores (Gaspar et al. 2001) or SCA36 in the Spanish Costa da Morte (García-Murias et al. 2012). The relatively high frequency of the SCA37 mutation in the south of the Iberian Peninsula, compared to other SCAs, has implications for disease prioritization during genetic diagnosis. We were able to trace back the relatively recent origin of the $(ATTTC)_n$ pathogenic insertion to 1164 (645–1376, 95% CI). Estimating the age of mutations partly depend on parameters that are difficult to exactly determine such as population growth rate or

◂**Fig. 7** Distinctive SCA37 haplotypes and common origin of the SCA37 mutation. **a** Haplotype analysis revealed the presence of a 964 kb shared region (red box) in all Iberian SCA37 patients segregating with the causative SCA37 mutation, revealing a common origin of the SCA37 mutation in the Iberian Peninsula which originated approximately 859 years ago (95% CI: 647–1378). Red bars represent informative SNPs positions. **b** DMLE + 2.3. analysis showing a posterior probability density of the mutation age for population grown rate ($r = 0.17$) and the proportion of sampled disease-bearing chromosomes ($f = 0.14$) considering an intergenerational time interval of 25 years. The estimated median age identified is 34 generations. Green bars show the 95% confidence interval between 26 and 55 generations. The frequency at which each number of generations was resulted from the iterations is shown on the y-axis. Outcome considering the total or the southwest of the Iberian Peninsula population with either 20 or 25 years/generation are included in Fig. S10 of the Additional file 2. **c** Haplotype network showing the phylogenetic relationship among five different SCA37 pathogenic alleles. Circle size is proportional to the number of chromosomes; line length is proportional to the genetic distance among haplotypes. **d** Phylogenetic reconstruction based on genetic distances (D_A) between the five haplotypes. The numbers next to nodes, represent a measure of support for the node. The line bar with 0.01 value indicates the number of genetic changes (nucleotide substitutions per site)

proportion of disease-bearing chromosomes (Rannala and Bertorelle 2001). Although these are possible limitations of this study, there is consistency in the resulted estimated age when using two different intergenerational time intervals giving reliability to our data (Fenner 2005). It would be interesting to investigate whether the SCA37 mutation found in German SCA37 patients (Rosenbohm et al. 2022) and in the Iberian Peninsula SCA37 patients share the same common haplotype. Such is the case for the SCA3/Machado-Joseph disease CAG pathogenic repeat, where a common origin chromosome was found in most families worldwide (Bettencourt et al. 2008). Interestingly, by the time the SCA37 mutation originated according to our study, the southwest region of the Iberian Peninsula comprised a part of al-Andalus occupied by a significant Arabic and North African populations. Therefore, we cannot rule out a non-European origin of the original SCA37 chromosome.

Conclusions

In this study, by accurately determining the size and configuration of the complex 5′(ATTTT)n–(ATTTC)n–3′(ATTTT)n repeat tract within the *DAB1* gene underlying SCA37 pathology by CRISPR/Cas9-mediated enrichment combined with nanopore long-read sequencing, we establish novel genotype–phenotype associations with significant implications for genetic diagnosis. Importantly, we also identify differential cerebellar hypomethylation upstream the repeat in SCA37 alleles that may account for DAB1 pathogenic cerebellar dysregulation in SCA37. Finally, this study provides evidence of a relatively recent common origin of the SCA37 mutation in the Iberian Peninsula dated to 1164 CE.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-024-02644-7.

Acknowledgements We thank all patients and family members participating in this study. We acknowledge the IGTP-HUGTP Biobank integrated in the Spanish National Biobanks Network of the Instituto de Salud Carlos III (PT13/0010/0009), and the IGTP genomics core facilities and staff for their contribution to this publication. We are indebted to CNAG Staff for assistance with Oxford Nanopore sequencing and data processing. We thank Natalia Benitez for technical assistance.

Author contributions A.M-D., M.C-J., I.S. and D.C. wrote the main manuscript. M.S-F. and E.G-N. acquired the data and M.C-J., M.S-F., D.C. and A.M-D. analysed and interpreted the data. M.C-J. and M.S-F. prepared the figures. A.M-D, I.S. and D.C. critically revised the manuscript for important intellectual content. A.M-D. obtained funding and supervised the study. All authors reviewed the manuscript.

Funding This work was funded by the National Institute of Health Carlos III (ISCIII)(PI17/00534).

Data availability The data generated and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Human rights and animal research This study was conducted according to the ethical principles for medical research involving human subjects according to the Declaration of Helsinki.

References

Alazard R (2005) Identification of the "NORE" (N-Oct-3 responsive element), a novel structural motif and composite element. Nucleic Acids Res 33:1513–1523. https://doi.org/10.1093/nar/gki284

Bae S, Park J, Kim J-S (2014) Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 30:1473–1475. https://doi.org/10.1093/bioinformatics/btu048

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Battaglia S, Dong K, Wu J, Chen Z, Najm FJ, Zhang Y, Moore MM, Hecht V, Shoresh N, Bernstein BE (2022) Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. Nat Genet 54:1504–1513. https://doi.org/10.1038/s41588-022-01188-8

Bettencourt C, Santos C, Kay T, Vasconcelos J, Lima M (2008) Analysis of segregation patterns in Machado-Joseph disease pedigrees. J Hum Genet 53:920–923. https://doi.org/10.1007/s10038-008-0330-y

Cancel G, Gourfinkel-An I, Stevanin G, Didierjean O, Abbas N, Hirsch E, Agid Y, Brice A (1998) Somatic mosaicism of the CAG repeat expansion in spinocerebellar ataxia type 3/Machado-Joseph disease. Hum Mutat 11(3):23

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Perez NM, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, Mathelier A (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res 50:D165–D173. https://doi.org/10.1093/nar/gkab1113

Corral-Juan M, Serrano-Munuera C, Rábano A, Cota-González D, Segarra-Roca A, Ispierto L, Cano-Orgaz AT, Adarmes AD, Méndez-Del-Barrio C, Jesús S, Mir P, Volpini V, Alvarez-Ramo R, Sánchez I, Matilla-Dueñas A (2018) Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. Brain 141:1981–1997. https://doi.org/10.1093/brain/awy137

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM (2021) Twelve years of SAMtools and BCFtools. Gigascience. https://doi.org/10.1093/gigascience/giab008

De Biase I, Rasmussen A, Endres D, Al-Mahdawi S, Monticelli A, Cocozza S, Pook M, Bidichandani SI (2007) Progressive GAA expansions in dorsal root ganglia of Friedreich's ataxia patients. Ann Neurol. https://doi.org/10.1002/ana.21052

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Dueñas AM, Goold R, Giunti P (2006) Molecular pathogenesis of spinocerebellar ataxias. Brain 129:1357–1370. https://doi.org/10.1093/brain/awl081

Durr A (2010) Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. Lancet Neurol 9:885–894. https://doi.org/10.1016/s1474-4422(10)70183-6

Ebler J, Haukness M, Pesout T, Marschall T, Paten B (2019) Haplotype-aware diplotyping from noisy long reads. Genome Biol 20:1–16. https://doi.org/10.1186/s13059-019-1709-0

Erdmann H, Schöberl F, Giurgiu M, Leal Silva RM, Scholz V, Scharf F, Wendlandt M, Kleinle S, Deschauer M, Nübling G, Heide W, Babacan SS, Schneider C, Neuhann T, Hahn K, Schoser B, Holinski-Feder E, Wolf DA, Abicht A (2023) Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. Brain 146:1831–1843. https://doi.org/10.1093/brain/awac377

Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415–423. https://doi.org/10.1002/ajpa.20188

Gamaarachchi H, Lam CW, Jayatilaka G, Samarakoon H, Simpson JT, Smith MA, Parameswaran S (2020) GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. BMC Bioinform 21:343. https://doi.org/10.1186/s12859-020-03697-x

García-Murias M, Quintáns B, Arias M, Seixas AI, Cacheiro P, Tarrío R, Pardo J, Millán MJ, Arias-Rivas S, Blanco-Arias P, Dapena D, Moreira R, Rodríguez-Trelles F, Sequeiros J, Carracedo A, Silveira I, Sobrido MJ (2012) "Costa da Morte" ataxia is spinocerebellar ataxia 36: clinical and genetic characterization. Brain 135:1423–1435. https://doi.org/10.1093/brain/aws069

Gaspar C, Lopes-Cendes I, Hayes S, Goto J, Arvidsson K, Dias A, Silveira I, Maciel P, Coutinho P, Lima M, Zhou YX, Soong BW, Watanabe M, Giunti P, Stevanin G, Riess O, Sasaki H, Hsieh M, Nicholson GA, Brunt E, Higgins JJ, Lauritzen M, Tranebjaerg L, Volpini V, Wood N, Ranum L, Tsuji S, Brice A, Sequeiros J, Rouleau GA (2001) Ancestral origins of the Machado-Joseph disease mutation: a worldwide haplotype study. Am J Hum Genet 68:523–528. https://doi.org/10.1086/318184

Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, Downs B, Sukumar S, Sedlazeck FJ, Timp W (2020) Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol 38:433–438. https://doi.org/10.1038/s41587-020-0407-5

Grant OA, Wang Y, Kumari M, Zabet NR, Schalkwyk L (2022) Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. Clin Epigenetics 14:62. https://doi.org/10.1186/s13148-022-01279-7

Guo P, Lam SL (2016) Minidumbbell: A new form of native dna structure. J Am Chem Soc 138:12534–12540. https://doi.org/10.1021/jacs.6b06897

Hatano Y, Ishihara T, Hirokawa S, Onodera O (2023) Machine learning approach for the prediction of age-specific probability of SCA3 and DRPLA by survival curve analysis. Neurol Genet 9:e200075. https://doi.org/10.1212/nxg.0000000000200075

Hommelsheim CM, Frantzeskakis L, Huang M, Ülker B (2014) PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci Rep 4:5052. https://doi.org/10.1038/srep05052

Jayadev S, Bird TD (2013) Hereditary ataxias: overview. Genet Med 15:673–683. https://doi.org/10.1038/gim.2013.28

Kacher R, Lejeune F-X, Noël S, Cazeneuve C, Brice A, Humbert S, Durr A (2021) Propensity for somatic expansion increases over the course of life in Huntington disease. Elife. https://doi.org/10.7554/elife.64674

Kenneson A (2001) Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. Hum Mol Genet 10:1449–1454. https://doi.org/10.1093/hmg/10.14.1449

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996–1006. https://doi.org/10.1101/gr.229102

Keraite I, Becker P, Canevazzi D, Frias-López C, Dabad M, Tonda-Hernandez R, Paramonov I, Ingham MJ, Brun-Heath I, Leno J, Abulí A, Garcia-Arumí E, Heath SC, Gut M, Gut IG (2022) A method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome. Nat Commun 13:5902. https://doi.org/10.1038/s41467-022-33530-3

Khristich AN, Mirkin SM (2020) On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J Biol Chem 295:4134–4170. https://doi.org/10.1074/jbc.rev119.007678

Klockgether T, Mariotti C, Paulson HL (2019) Spinocerebellar Ataxia Nat Rev Dis Primers 5:24. https://doi.org/10.1038/s41572-019-0074-3

Kraus-Perrotta C, Lagalwar S (2016) Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. Cerebellum Ataxias. https://doi.org/10.1186/s40673-016-0058-y

Lachman HM, Pedrosa E, Nolan KA, Glass M, Ye K, Saito T (2006) Analysis of polymorphisms in AT-rich domains of neuregulin 1 gene in schizophrenia. Am J Med Genet B Neuropsychiatr Genet 141B:102–109. https://doi.org/10.1002/ajmg.b.30242

Lee J-M, Zhang J, Su AI, Walker JR, Wiltshire T, Kang K, Dragileva E, Gillis T, Lopez ET, Boily M-J, Cyr M, Kohane I, Gusella JF, MacDonald ME, Wheeler VC (2010) A novel approach to investigate tissue-specific trinucleotide repeat instability. BMC Syst Biol 4:29. https://doi.org/10.1186/1752-0509-4-29

Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual J-F, Fisk CJ, Li N, Smolyar A, Hill DE, Barabási A-L, Vidal M, Zoghbi HY (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125:801–814. https://doi.org/10.1016/j.cell.2006.03.032

Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, Foox J, Mason C, Carroll M, Cheng A, Li S (2021) DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol 22:295. https://doi.org/10.1186/s13059-021-02510-z

Loureiro JR, Oliveira CL, Mota C, Castro AF, Costa C, Loureiro JL, Coutinho P, Martins S, Sequeiros J, Silveira I (2019) Mutational mechanism for DAB1 (ATTTC)n insertion in SCA37: ATTTT repeat lengthening and nucleotide substitution. Hum Mutat 40:404–412. https://doi.org/10.1002/humu.23704

Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schöenhuth A, Marschall T (2016) WhatsHap: fast and accurate read-based phasing. bioRxiv. https://doi.org/10.1101/085050

Matilla-Dueñas A, Ashizawa T, Brice A, Magri S, McFarland KN, Pandolfo M, Pulst SM, Riess O, Rubinsztein DC, Schmidt J, Schmidt T, Scoles DR, Stevanin G, Taroni F, Underwood BR, Sánchez I (2014) Consensus paper: pathological mechanisms underlying neurodegeneration in spinocerebellar ataxias. Cerebellum 13:269–302. https://doi.org/10.1007/s12311-013-0539-y

Matsuura T, Fang P, Lin X, Khajavi M, Tsuji K, Rasmussen A, Grewal RP, Achari M, Alonso ME, Pulst SM, Zoghbi HY, Nelson DL, Roa BB, Ashizawa T (2004) Somatic and germline instability of the ATTCT repeat in spinocerebellar Ataxia type 10. Am J Human Genet 74:1216–1224. https://doi.org/10.1086/421526

Miyatake S, Koshimizu E, Fujita A, Doi H, Okubo M, Wada T, Hamanaka K, Ueda N, Kishida H, Minase G, Matsuno A, Kodaira M, Ogata K, Kato R, Sugiyama A, Sasaki A, Miyama T, Satoh M, Uchiyama Y, Tsuchida N, Hamanoue H, Misawa K, Hayasaka K, Sekijima Y, Adachi H, Yoshida K, Tanaka F, Mizuguchi T, Matsumoto N (2022) Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. NPJ Genom Med. https://doi.org/10.1038/s41525-022-00331-y

Mizuguchi T, Toyota T, Miyatake S, Mitsuhashi S, Doi H, Kudo Y, Kishida H, Hayashi N, Tsuburaya RS, Kinoshita M, Fukuyama T, Fukuda H, Koshimizu E, Tsuchida N, Uchiyama Y, Fujita A, Takata A, Miyake N, Kato M, Tanaka F, Adachi H, Matsumoto N (2021) Complete sequencing of expanded SAMD12 repeats by long-read sequencing and Cas9-mediated enrichment. Brain 144:1103–1117. https://doi.org/10.1093/brain/awab021

Moseley ML (2000) SCA8 CTG repeat: en masse contractions in sperm and intergenerational sequence changes may play a role in reduced penetrance. Hum Mol Genet 9:2125–2130. https://doi.org/10.1093/hmg/9.14.2125

Mouro Pinto R, Arning L, Giordano JV, Razghandi P, Andrew MA, Gillis T, Correia K, Mysore JS, Grote Urtubey D-MM, Parwez CR, von Hein SM, Clark HB, Nguyen HP, Förster E, Beller A, Jayadaev S, Keene CD, Bird TD, Lucente D, Vonsattel J-PP, Orr H, Saft C, Petrasch-Parwez E, Wheeler VC, Pinto RM, Arning L, Giordano JV, Razghandi P, Andrew MA, Gillis T, Correia K, Mysore JS, Grote Urtubey D-MM, Parwez CR, von Hein SM, Brent Clark H, Nguyen HP, Förster E, Beller A, Jayadaev S, Dirk Keene C, Bird TD, Lucente D, Vonsattel J-PP, Orr H, Saft C, Petrasch-Parwez E, Wheeler VC (2020) Patterns of CAG repeat instability in the central nervous system and periphery in Huntington's disease and in spinocerebellar ataxia type 1. Hum Mol Genet 29:2551–2567. https://doi.org/10.1093/hmg/ddaa139

Nakamori M, Panigrahi GB, Lanni S, Gall-Duncan T, Hayakawa H, Tanaka H, Luo J, Otabe T, Li J, Sakata A, Caron M-C, Joshi N, Prasolava T, Chiang K, Masson J-Y, Wold MS, Wang X, Lee MYWT, Huddleston J, Munson KM, Davidson S, Layeghifard M, Edward L-M, Gallon R, Santibanez-Koref M, Murata A, Takahashi MP, Eichler EE, Shlien A, Nakatani K, Mochizuki H, Pearson CE (2020) A slipped-CAG DNA-binding small molecule induces trinucleotide-repeat contractions in vivo. Nat Genet 52:146–159. https://doi.org/10.1038/s41588-019-0575-8

Pearson CE, Edamura KN, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6:729–742. https://doi.org/10.1038/nrg1689

Pedersen BS, Quinlan AR (2018) Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34:867–868. https://doi.org/10.1093/bioinformatics/btx699

Peng L, Chen Z, Chen T, Lei L, Long Z, Liu M, Deng Q, Yuan H, Zou G, Wan L, Wang C, Peng H, Shi Y, Wang P, Peng Y, Wang S, He L, Xie Y, Tang Z, Wan N, Gong Y, Hou X, Shen L, Xia K, Li J, Chen C, Zhang Z, Qiu R, Tang B, Jiang H (2021) Prediction of the age at onset of spinocerebellar ataxia type 3 with machine learning. Mov Disord 36:216–224. https://doi.org/10.1002/mds.28311

Polak U, McIvor E, Dent SYR, Wells RD, Napierala M (2013) Expanded complexity of unstable repeat diseases. BioFactors 39:164–175. https://doi.org/10.1002/biof.1060

Potapov V, Ong JL (2017) Examining sources of error in PCR by single-molecule sequencing. PLoS ONE 12:e0169774. https://doi.org/10.1371/journal.pone.0169774

Rannala B, Bertorelle G (2001) Using linked markers to infer the age of a mutation. Hum Mutat 18:87–100. https://doi.org/10.1002/humu.1158

Rausch T, Hsi-Yang Fritz M, Korbel JO, Benes V (2019) Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. Bioinformatics 35:2489–2491. https://doi.org/10.1093/bioinformatics/bty1007

Reeve JP, Rannala B (2002) DMLE+: Bayesian linkage disequilibrium gene mapping. Bioinformatics 18:894–895. https://doi.org/10.1093/bioinformatics/18.6.894

Richards RI (2001) Dynamic mutations: a decade of unstable expanded repeats in human genetic disease. Hum Mol Genet 10:2187–2194. https://doi.org/10.1093/hmg/10.20.2187

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754

Rosenbohm A, Pott H, Thomsen M, Rafehi H, Kaya S, Szymczak S, Volk AE, Mueller K, Silveira I, Weishaupt JH, Tönnies H, Seibler P, Zschiedrich K, Schaake S, Westenberger A, Zühlke C, Depienne C, Trinh J, Ludolph AC, Klein C, Bahlo M, Lohmann K (2022) Familial cerebellar ataxia and amyotrophic lateral sclerosis/frontotemporal dementia with DAB1 and C9ORF72 repeat expansions: an 18-year study. Mov Disord 37:2427–2439. https://doi.org/10.1002/mds.29221

Ru D, Li J, Xie O, Peng L, Jiang H, Qiu R (2022) Explainable artificial intelligence based on feature optimization for age at onset prediction of spinocerebellar ataxia type 3. Front Neuroinform 16:978630. https://doi.org/10.3389/fninf.2022.978630

Russo R, Gambale A, Esposito MR, Serra ML, Troiano A, de Maggio I, Capasso M, Luzzatto L, Delaunay J, Tamary H, Iolascon A (2011) Two founder mutations in the SEC23B gene account for the relatively high frequency of CDA II in the Italian population. Am J Hematol 86:727–732. https://doi.org/10.1002/ajh.22096

Schaffer A, Kim EC, Wu X, Zan H, Testoni L, Salamon S, Cerutti A, Casali P (2003) Selective inhibition of class switching to IgG and IgE by recruitment of the HoxC4 and Oct-1 homeodomain

proteins and Ku70/Ku86 to newly identified ATTT cis-elements. J Biol Chem 278:23141–23150. https://doi.org/10.1074/jbc.m212952200

Seixas AI, Loureiro JR, Costa C, Ordóñez-Ugalde A, Marcelino H, Oliveira CL, Loureiro JL, Dhingra A, Brandão E, Cruz VT, Timóteo A, Quintáns B, Rouleau GA, Rizzu P, Carracedo Á, Bessa J, Heutink P, Sequeiros J, Sobrido MJ, Coutinho P, Silveira I (2017) A Pentanucleotide ATTTC repeat insertion in the non-coding region of DAB1, mapping to SCA37, causes spinocerebellar ataxia. Am J Hum Genet 101:87–103. https://doi.org/10.1016/j.ajhg.2017.06.007

Sequeiros J, Martins S, Silveira I (2012) Epidemiology and population genetics of degenerative ataxias. Handb Clin Neurol 103:227–251. https://doi.org/10.1016/b978-0-444-51892-7.00014-0

Serrano-Munuera C, Corral-Juan M, Stevanin G, San Nicolás H, Roig C, Corral J, Campos B, De Jorge L, Morcillo-Suárez C, Navarro A, Forlani S, Durr A, Kulisevsky J, Brice A, Sánchez I, Volpini V, Matilla-Duenãs A (2013) New subtype of spinocerebellar ataxia with altered vertical eye movements mapping to chromosome 1p32. JAMA Neurol 70:764–771. https://doi.org/10.1001/jamaneurol.2013.2311

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311. https://doi.org/10.1093/nar/29.1.308

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458

Sobczak K, Krzyzosiak WJ (2005) CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. J Biol Chem 280:3898–3910. https://doi.org/10.1074/jbc.m409984200

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169. https://doi.org/10.1086/379378

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989. https://doi.org/10.1086/319501

Takezaki N, Nei M, Tamura K (2010) POPTREE2: software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. Mol Biol Evol 27:747–752. https://doi.org/10.1093/molbev/msp312

Tan D, Wei C, Chen Z, Huang Y, Deng J, Li J, Liu Y, Bao X, Xu J, Hu Z, Wang S, Fan Y, Jiang Y, Wu Y, Wu Y, Wang S, Liu P, Zhang Y, Yang Z, Jiang Y, Zhang H, Hong D, Zhong N, Jiang H, Xiong H (2023) CAG repeat expansion in THAP11 is associated with a novel spinocerebellar ataxia. Mov Disord 38:1282–1293. https://doi.org/10.1002/mds.29412

Tarailo-Graovac M, Chen N (2009) Using repeat masker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform. https://doi.org/10.1002/0471250953.bi0410s25

Tezenas du Montcel S, Durr A, Rakowicz M, Nanetti L, Charles P, Sulek A, Mariotti C, Rola R, Schols L, Bauer P, Dufaure-Garé I, Jacobi H, Forlani S, Schmitz-Hübsch T, Filla A, Timmann D, van de Warrenburg BP, Marelli C, Kang J-S, Giunti P, Cook A, Baliko L, Melegh B, Boesch S, Szymanski S, Berciano J, Infante J, Buerk K, Masciullo M, Di Fabio R, Depondt C, Ratka S, Stevanin G, Klockgether T, Brice A, Golmard J-L (2014) Prediction of the age at onset in spinocerebellar ataxia type 1, 2, 3 and 6. J Med Genet 51:479–486. https://doi.org/10.1136/jmedgenet-2013-102200

Thornton CA, Johnson K, Moxley RT (1994) Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. Ann Neurol. https://doi.org/10.1002/ana.410350116

Velázquez Pérez L, Cruz GS, Santos Falcón N, Enrique Almaguer Mederos L, Escalona Batallan K, Rodríguez Labrada R, Paneque Herrera M, Laffita Mesa JM, Rodríguez Díaz JC, Rodríguez RA, González Zaldivar Y, Coello Almarales D, Almaguer Gotay D, Jorge Cedeño H (2009) Molecular epidemiology of spinocerebellar ataxias in Cuba: insights into SCA2 founder effect in Holguin. Neurosci Lett 454:157–160. https://doi.org/10.1016/j.neulet.2009.03.015

Verbeek DS, Piersma SJ, Hennekam EFAM, Ippel EF, Pearson PL, Sinke RJ (2004) Haplotype study in Dutch SCA3 and SCA6 families: evidence for common founder mutations. Eur J Hum Genet 12:441–446. https://doi.org/10.1038/sj.ejhg.5201167

Wan X, Wang W, Liu J, Tong T (2014a) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 14:135. https://doi.org/10.1186/1471-2288-14-135

Wan Z, Rui L, Li Z (2014b) Expression patterns of prdm1 during chicken embryonic and germline development. Cell Tissue Res 356:341–356. https://doi.org/10.1007/s00441-014-1804-1

Wen J, Trost B, Engchuan W, Halvorsen M, Pallotto LM, Mitina A, Ancalade N, Farrell M, Backstrom I, Guo K, Pellecchia G, Thiruvahindrapuram B, Giusti-Rodriguez P, Rosen JD, Li Y, Won H, Magnusson PKE, Gyllensten U, Bassett AS, Hultman CM, Sullivan PF, Yuen RKC, Szatkiewicz JP (2023) Rare tandem repeat expansions associate with genes involved in synaptic and neuronal signaling functions in schizophrenia. Mol Psychiatry 28:475–482. https://doi.org/10.1038/s41380-022-01857-4

Wilke C, Pellerin D, Mengel D, Traschütz A, Danzi MC, Dicaire M-J, Neumann M, Lerche H, Bender B, Houlden H, Faber J, Roxburgh R, Pedroso JL, Alvez PC, Barsottini O, Pane C, Saccà F, Filla A, Santorelli FM, Ricca I, Züchner S, Schöls L, Brais B, Synofzik M (2023) GAA- FGF14 ataxia (SCA27B): phenotypic profile, natural history progression and 4-aminopyridine treatment response. Brain. https://doi.org/10.1093/brain/awad157

Wright GEB, Collins JA, Kay C, McDonald C, Dolzhenko E, Xia Q, Bečanović K, Drögemöller BI, Semaka A, Nguyen CM, Trost B, Richards F, Bijlsma EK, Squitieri F, Ross CJD, Scherer SW, Eberle MA, Yuen RKC, Hayden MR (2019) Length of Uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of huntington disease. Am J Human Genet 104:1116–1126. https://doi.org/10.1016/j.ajhg.2019.04.007

Yang J, Rahardja S, Fränti P (2019) outlier detection proceedings of the international conference on artificial intelligence. Inform Process Cloud Comput. https://doi.org/10.1145/3371425.3371427

Lundberg S, Lee S-I (2017) A Unified approach to interpreting model predictions. 10.48550

Matilla-Dueñas A, Volpini V (1993) Spinocerebellar Ataxia Type 37. GeneReviews®