frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 14 November 2023
DOI 10.3389/fgene.2023.1296614

OPEN ACCESS

EDITED BY
Gyaneshwer Chaubey,
Banaras Hindu University, India

REVIEWED BY
Jorge Diogo Da Silva,
University of Minho, Portugal
Sandra Martins,
Universidade do Porto, Portugal
Douglas Langbehn,
The University of Iowa, United States

*CORRESPONDENCE
Laura Bannach Jardim,
ljardim@hcpa.edu.br

RECEIVED 18 September 2023
ACCEPTED 27 October 2023
PUBLISHED 14 November 2023

CITATION
Sena LS, Lemes RB, Furtado GV, Saraiva-Pereira ML and Jardim LB (2023), A model for the dynamics of expanded CAG repeat alleles: *ATXN2* and *ATXN3* as prototypes.
Front. Genet. 14:1296614.
doi: 10.3389/fgene.2023.1296614

A model for the dynamics of expanded CAG repeat alleles: *ATXN2* and *ATXN3* as prototypes

Lucas Schenatto Sena[1,2], Renan Barbosa Lemes[3], Gabriel Vasata Furtado[2], Maria Luiza Saraiva-Pereira [1,2,4,5] and Laura Bannach Jardim [1,2,4,6]*

[1]Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, [2]Centros de Pesquisa Clínica e Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil, [3]Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil, [4]Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil, [5]Departamento de Bioquímica, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, [6]Departamento de Medicina Interna, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Background: Spinocerebellar ataxia types 2 (SCA2) and 3 (SCA3/MJD) are diseases due to dominant unstable expansions of CAG repeats (CAGexp). Age of onset of symptoms (AO) correlates with the CAGexp length. Repeat instability leads to increases in the expanded repeats, to important AO anticipations and to the eventual extinction of lineages. Because of that, compensatory forces are expected to act on the maintenance of expanded alleles, but they are poorly understood.

Objectives: we described the CAGexp dynamics, adapting a classical equation and aiming to estimate for how many generations will the descendants of a *de novo* expansion last.

Methods: A mathematical model was adapted to encompass anticipation, fitness, and allelic segregation; and empirical data fed the model. The arbitrated ancestral mutations included in the model had the lowest CAGexp and the highest AO described in the literature. One thousand generations were simulated until the alleles were eliminated, fixed, or 650 generations had passed.

Results: All SCA2 lineages were eliminated in a median of 10 generations. In SCA3/MJD lineages, 593 were eliminated in a median of 29 generations. The other ones were eliminated due to anticipation after the 650th generation or remained indefinitely with CAG repeats transitioning between expanded and unexpanded ranges.

Discussion: the model predicted outcomes compatible with empirical data - the very old ancestral SCA3/MJD haplotype, and the *de novo* SCA2 expansions -, which previously seemed to be contradictory. This model accommodates these data into understandable dynamics and might be useful for other CAGexp disorders.

KEYWORDS

allele dynamics, Machado-Joseph disease, mathematical model, polyglutamine diseases, spinocerebellar ataxia type 2, spinocerebellar ataxia type 3, selective forces

1 Introduction

CAG repeat expansions (CAGexp) are a major genetic cause of neurological diseases. When they occur within a codon region, the corresponding expansion of the polyglutamine tract (polyQ) in the expressed protein is thought to be neurotoxic. Mechanisms include post-transcriptional and -translational modifications and autophagic disturbances (Adegbuyiroa et al., 2017; Bunting et al., 2021). Each CAGexp or mutant polyQ targets different populations of neurons, causing distinct diseases that are commonly referred to as polyQ diseases. They include Huntington disease (HD [MIM: 143100]), the spinocerebellar ataxia type 1 (SCA1, [MIM: 164400]), type 2 (SCA2, [MIM: 183090]), type 3 (also known as Machado-Joseph disease, SCA3/MJD, [MIM: 109150]), type 6 (SCA6, [MIM: 183086]), type 7 (SCA7, [MIM: 164500]), type 17 (SCA17, [MIM: 607136]), Dentatorubropallidoluysian atrophy (DRPLA, [MIM: 125370]), and spinobulbar muscular atrophy (SBMA, also known as Kennedy's disease, [MIM: 313200]) (Figure 1) (Orr HT et al., 1993; La Spada, 1999; the Huntington's Disease Collaborative Research Group, 1993; Kawaguchi et al., 1994; Komure et al.,

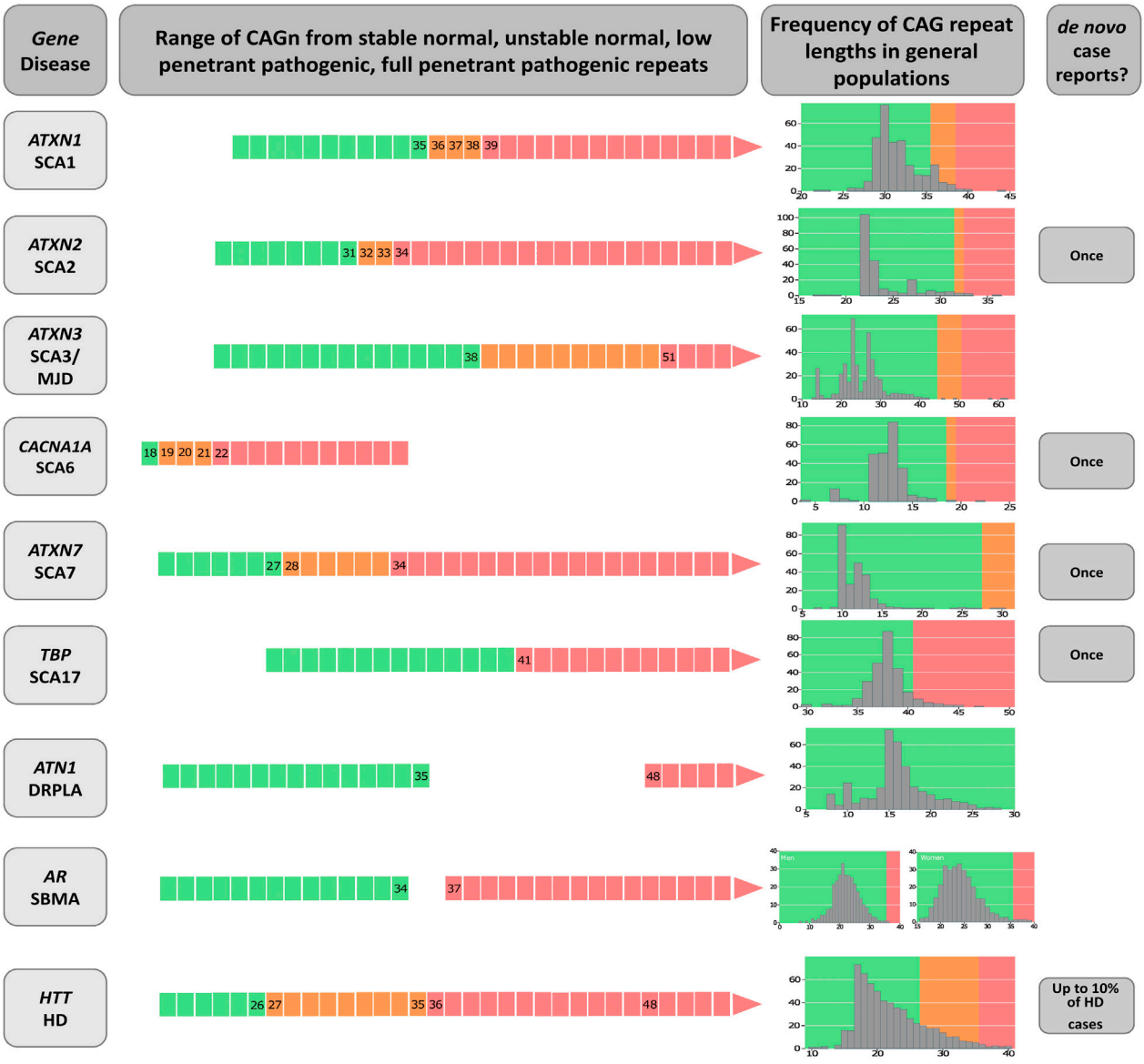

FIGURE 1
Diagram on the genetic characteristics of diseases related to polyglutamine expansions. The gene/disease column summarizes the disease-causing genes and the abbreviations of disease names of spinocerebellar ataxia types 1, 2, 3, 6, 7 and 17 (SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17, dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA), and Huntington's disease (HD). Data on CAGn presented in the second column was retrieved from Opal and Ashizawa, 1998, de Castilhos et al., 2014, Gu et al., 2004, Casey and Gomez, 1998, Shizuka et al., 1998, La Spada et al., 1991, Mittal et al., 2005, Toyoshima et al., 2005, Bech et al., 2010, Carroll et al., 2018, La Spada, 1998; Caron et al., 1998. *De novo* case reports were Futamura et al., 1998; Shizuka et al., 1998; Stevanin et al., 1998; Bech et al., 2010; Kay et al., 2018. The column called "Frequency of CAG repeat lengths in populations" shows the histograms of the alleles found in a normal population (data adapted from Gardiner et al., 2019). Green means the range of normal alleles; orange, the range of intermediate alleles; and red, the range of pathological alleles.

1995; David et al., 1997; Zhuchenko et al., 1997; Koide et al., 1999; Pulst et al., 2005; Margolis et al., 2005).

PolyQ diseases are rare, progressive, and fatal, and share several clinical and genetic characteristics (Lieberman et al., 2019). Most are autosomal dominant diseases; the exception is SBMA, an X-linked disorder restricted to males due to the limited expression of the androgen receptor in females. Specific critical thresholds separate normal repeats from pathogenic ones (Figure 1). The age at onset (AO) or age at which the first neurologic manifestation was noted by the subject or their relatives, is usually in adulthood. Larger CAGexp lengths determine earlier AO and faster rates of disease progression: in most polyQ disorders, the length of the expanded repeat explains around 50% of the AO variability (de Mattos et al., 2018; Gusela et al., 2021). Given that instabilities of the CAGexp tracts tend to increase them after meiosis, subsequent anticipation (earlier disease onset in affected offspring than in their parents) is quite frequent and may end up causing onset of symptoms in childhood. The general tendency for symptoms to appear earlier and earlier may result in progressive reductions of the reproductive periods of each subsequent generation, until reproduction is completely prevented. Thus, successive anticipations would end up eliminating a lineage of expanded alleles from the population pool. Clinical presentations of homozygous patients are not very different from those of heterozygotes (Cubo et al., 2019). Most of these characteristics support the hypothesis that a toxic and fully dominant gain of function (Lee et al., 2012) underlies the pathogenesis - although loss of function at some steps of the pathogenetic pathway has been reported in some polyQs diseases (Adegbuyiro et al., 2017; Bunting et al., 2022). The aspect of polyQs diseases that interests us in the present study is how they are maintained in the population, since, as we said before, repeated reductions of the reproductive period per generation, due to anticipation, might drive many polyQ lineages to extinction. One way to address this issue is to model for how many generations each lineage that appears after a *de novo* CAGexp expansion would last.

ATXN2 and *ATXN3* are the genes related to SCA2 and SCA3/MJD, respectively. Both disorders are characterized by gait ataxia, pyramidal signs, a dystonic and/or rigid extrapyramidal syndrome, sensory losses, amyotrophy, and progressive external ophthalmoplegia (Paulson and Shakkottai, 1998; Pulst, 1998). AO, anticipation, neurologic manifestations, and survival after onset are similar across both diseases, so that they can be only distinguished with confidence by molecular testing (Bird, 1998; Diallo et al., 2018).

There are some notable differences between both diseases. SCA3/MJD shows a relevant gap between the length of normal and expanded alleles (Figure 1), lacks *de novo* expansions, and preferentially segregates the expanded allele on meiosis. Evidence in favor of a few and old ancestral haplotypes have been obtained in several SCA3/MJD populations by robust studies (for instance, see Li et al., 2019). *ATXN2* alleles show no gap between normal and expanded, *de novo* expansions have been described, and preferentially segregates the normal allele on meiosis (Sena et al., 2021a; 2021b). Reconstruction of ancestral haplotypes using single nucleotide polymorphisms (SNP) was restricted until recently to two markers rs695871 and rs695872, and to the finding of a unique C-C haplotype worldwide (Choudhry et al., 2001; Ramos et al., 2010; Sonakar et al., 2021). We have now used five SNPs rs9300319, rs3809274, rs695871, rs12369009, and rs593226 and found at least eleven ancestral SCA2 haplotypes, just in South American families (Sena et al., 2023). These differences indicate that the biological contexts of *ATXN2* and *ATXN3* are quite diverse.

The similarities between SCA2 and SCA3/MJD clinical characteristics suggests that social and psychological impacts over their carriers should also be similar. This allows the presumption that the differences between SCA2 and SCA3/MJD transmissions to the offspring should not be attributed to distinct psychological or social pictures, but to the biological context of the expansions in *ATXN2* and *ATXN3*. Due to that, SCA2 and SCA3/MJD are probably good prototypes to test the dynamics of the CAGexp in general and to answer the question: "for how many generations will the descendants of a *de novo* expansion last?" This was the aim of the present study. The specific aims were to adapt a classical equation on allele dynamics in population genetics to be used in the case of dominant alleles related to late onset neurodegenerative disorders; and then, to test if the results of the model match with the existing epidemiological evidence on ancestral lineages and anticipation, in SCA2 and SCA3/MJD.

2 Subjects and methods

2.1 Subjects

The following human populations were used in this work: EUROSTAT (European Statistical Office) 2019 data were used to establish fertility rates of normal women stratified by life year. Measures of fitness, segregation distortion, CAGexp instability and anticipation related to SCA2 or SCA3/MJD subjects were obtained from two meta-analyses published elsewhere (Sena et al., 2021a; 2021b). To clarify, genetic fitness (or reproductive success) is a concept related to the reproductive success of a given allele or phenotype and is usually measured by the ratio between the median number of children of affected subjects over the median number of children of the unaffected subjects; segregation distortion is the phenomenon in which genotypes deviate from expected Mendelian ratios. Individual participant data (IPD) were obtained from other two original publications to estimate the reduction in AO attributable to each additional CAG repeat in CAGexp in SCA3/MJD (1,112 individuals) (de Mattos et al., 2018), and in SCA2 (93 individuals) (Pereira et al., 2015). Data on fitness of SCA2 and SCA3 carriers was obtained from two other IPD sources (Souza et al., 2016; Sena et al., 2019). Finally, we used the stability/instability of the unexpanded allele with intermediate *ATXN2* length, observed in the general population and described in 57 individuals (Almaguer-Mederos et al., 2018).

2.2 Methods

To estimate the fate of CAGexp transmissions after a *de novo* expansion, one hypothetical expanded allele was assigned to a hypothetical founder of a lineage, and Monte Carlo methods were used to assign a random genotype to each descendant in several generations, based on segregation of the parental alleles, in a way similar to the gene dropping methodology (MacCluer et al.,

1986) [30]. The frequencies of the expanded alleles across generations, and the probabilities of extinction of these expanded alleles, were then assessed in a hypothetical population. The model assumed the absence of *de novo* mutations, genetic drift, and gene flow. In contrast, the effect of three mechanisms that could change the frequency of the expanded alleles were included: differential fitness of carriers of the expanded allele compared to non-carriers; transmission probabilities according to distortion in the segregation of the expanded allele; and anticipation.

2.2.1 Adaptation of classical method on natural selection effect

The classical equation from population genetics theory encompasses fitness and distortion in allelic segregation as two selective forces of interest:

$$p' = \frac{p^2 w_{11} + 2kpqw_{12}}{\overline{W}}$$

Where:

p' = Frequency of p allele in the subsequent generation.

p^2 = Frequency of individuals with the p allele in homozygotes.

w_{11} = Fitness of the p allele in homozygotes.

k = segregation coefficient, where 0.5 represents Mendelian segregation.

p = frequency of the p allele.

q = frequency of the q allele.

w_{12} = fitness of heterozygotes.

$\overline{W}$ = average fitness.

Adaptations were made to match the original equation to the specific characteristics of polyQs diseases, as follows.

First, an anticipation coefficient (*antcoeff*) was included to account for the influence of anticipation on the allele frequency at each generation of a lineage. The *antcoeff* ranged from zero to one, where zero corresponds to symptoms starting before the beginning of the fertile life (proposed as being 12 years of age) and one is related to symptoms that begin after the end of the fertile life (proposed as being 50 years of age). The extreme values represent the worst and the best scenarios for reproductive life, respectively, and the anticipation coefficient values were the mathematical expression of the relation between AOfs and the fertility rate of the ages' interval arising from the new cutoff in the reproductive period imposed by the AOfs, in a given generation, in the studied lineage.

The next adjustments aimed to simplify the model. As the expanded alleles have a very low frequency, homozygosity is so rare that is practically non-existent: as $p^2 \approx 0$, then it was removed from the equation. The expanded alleles have complete or near complete dominance, without any clear dosage effect when present in double dose (Sanpei et al., 1996; Saute and Jardim, 2015). Since penetrance is close to 100%, the expanded allele cannot "protect" itself from the action of natural selection when it is in heterozygosity. This implies that the frequency of the expanded allele in the subsequent generation (p') is essentially modulated by the intrinsic selective forces associated with p only, and not with q (the frequency of non-expanded allele). Due to that and to the fact that q frequency is close to 1, q was also removed from the equation.

The last adjustment was to directly use the relative w fitness into the equation instead of using the carriers' W fitness divided by the general W fitness of the population, or in other words, replacing the expression $\frac{w_{12}}{\overline{W}}$ with its results w

With this, we arrived at:

$$p' = p.w.(antcoeff).2k$$

Where:

p = frequency of the expanded allele.

p' = frequency of the expanded allele in the subsequent generation.

w = relative fitness.

antcoeff = anticipation coefficient.

k = segregation coefficient.

2.2.2 The anticipation coefficient *antcoeff*

To infer the impact of anticipation on carrier's fitness, we developed the anticipation coefficient or *antcoeff* based on the premise that only a proportion of children is born after onset of symptoms, and on the premise that the reproductive period starts in adolescence. Therefore, it was important to impute what this proportion of births would be after the onset of symptoms, and what the fertility rates would be in the age groups still included before the AO of each new generation - data that will be described in the next paragraph. The *antcoeff* itself was the result of a five-step operation: first, the average CAGexp size of a given generation was estimated from the available data about CAGexp instability in the disease of interest; then, this new CAGexp length was related to its average AO; then, the new length of the reproductive period due to this new anticipation of the AO was imputed; measures of SCA2 and SCA3/MJD fitness were obtained from two meta-analyses published elsewhere (Sena et al., 2021a; 2021b); as total fitness must have an age-dependent distribution, we also weight this factor, according to what is observed in the EUROSTAT (European Statistical Office) 2019 data on fertility rates of normal women stratified by life year. And so on. The mathematical expression of the *antcoeff is* at Supplementary Material S1.

We arbitrated that the *antcoeff* would be equal to the average reproduction rate of a given generation, in the lineage of a common ancestor. Each generation of carriers would be prone to show modifications in their reproduction rate, due to the change of their average reproductive period, provoked, in turn, by the average anticipation of that generation.

We assume that each new anticipation will be associated with an additional reduction in the fertile period. Although symptomatic people might continue to have children during the early years of their illness, at some point, their clinical state will interfere with the reproductive capacity - either because of the children's threats related to motor incapacity of a parent, or because of the reduced opportunities of sexual relationships required for reproduction. The conceptual relationship between anticipation and reproduction reduction has been studied and discussed elsewhere (Prestes et al., 2008; Sena et al., 2019). We came back to those datasets and were able to estimate that 92% and 91.7% of children were born before the onset of symptoms of their of SCA2 and SCA3/MJD parents, at a mean (SD) of 13.69 (12.04) and 12.72 (12.84) years before the AO of their parents (data not shown). Then an addition of 8% and 8.3% to the new estimated birth rate per generation was done to bring the *antcoeff* closer to reality of the reproduction rates.

TABLE 1 Variables related to CAG repeats at *ATXN2* and *ATXN3*, and used in the present adapted model. Data related to expanded repeats was obtained from heterozygous carriers; data related to normal repeats was obtained from non-carriers. Data is presented as means (standard deviation).

Allele	Fitness, w	Segregation distortion, k	Instability during meiosis	AO reduction due to each CAG repeat added in the expanded repeat
ATXN2, expanded allele	1.50 (0.25)[a] Sena et al., 2019	0.404 (0.085)[a] Sena et al., 2019	2.42 (5.655) Sena et al., 2021a	1.877 (1.86)
ATXN3, expanded allele	1.45 (0.25)[a] Prestes et al., 2008	0.640 (0.085) Sena et al., 2021b	1.23 (5.126) Sena et al., 2021b	1.652 (1.729)
ATXN2, normal allele	1.00 (0.25)[a]	0.596 (0.085)[a] Sena et al., 2019	0.23 (0.468) Almaguer-Mederos et al., 2018	
ATXN3, normal allele	1.00 (0.25)[a]	0.360 (0.085) Sena et al., 2021b	0.00[a] (0.468)[a]	

[a]Imputed values. Reasons for each imputation were described in the text.

To define how reductions in the reproductive period modify the *antcoeff* in each generation, fertility rates in different age groups of a general population were established, using EUROSTAT data on the fertility of European women in 2019. In order to know how much each age group contributed to the overall fertility rate of that population in a given year, the total area below fertility function was calculated (Supplementary Material S3, Supplementary Figure S1A) and normalized to the value of 1, equal to the best anticipation coefficient, that was also equal to 1 when the disease onset is after the end of the reproductive period and 0 when symptoms begin before the start of reproductive period (Supplementary Material S3, Supplementary Figure S1B). Each reduction in the reproductive period due to anticipation then reduced the area of the plot and reduced the value of the anticipation coefficient.

The next estimation to be made was that of the AOfs variation with each new generation, to extrapolate the corresponding reduction in the reproductive period.

All cohorts published to date were biased in favor of high anticipations (Sena et al., 2021a; 2021b). Therefore, the anticipation data retrieved from the literature was not directly used. Instead, we used a combination of two other pieces of information: the average instability of CAGexp transmission in each generation, and how much an increase of one CAG repeat reduces the AOfs. Measures of CAGexp instability were obtained from two meta-analyses (Sena et al., 2021a; 2021b), while the reduction in AO attributable to each additional CAG repeat in CAGexp were estimated from IPD data retrieved from two publications (Pereira et al., 2015; de Mattos et al., 2018). Studies on CAGexp transmissions were more representative than those that report only AOfs, since they commonly include asymptomatic as well as symptomatic individuals (Souza et al., 2016; Sena et al., 2019). The product of [instability of transmission x AOfs-per-additional-CAGexp] estimated the anticipation in the subsequent generation. This multiplication produced smaller anticipations than the direct measurements; therefore, this procedure reduced distortions due to the literature bias in favor of excessive anticipations.

A linear regression performed for SCA3/MJD carriers from the IPD mentioned before (de Mattos et al., 2018) showed that each additional CAG repeat at expanded *ATXN3* was associated to a 1.652-year reduction in AOfs. To calculate the same for SCA2, we used the IPD from a publication on Brazilian, Peruvian, and Uruguayan individuals (Pereira et al., 2015). Each additional CAGexp at *ATXN2* was related to a reduction of 1.877 years in AOfs (Table 1 and Supplementary Material S2).

Contractions in the CAGexp repeat length can also occur, reducing the size of CAGexp, although very rarely documented (Cruz-Mariño et al., 2014). After a contraction, the originally expanded allele of *ATXN2* and *ATXN3* might be transmitted as an unexpanded allele, causing the selective forces associated with SCA2 and SCA3/MJD diseases to no longer influence the dynamics of these descendant alleles. Although direct data were not available, contractions were taken into consideration in the resulting simulation model.

Ultimately, we could express these relationships above in a way that the anticipation coefficient varies from 0 to 1, where 0 means that the carriers will not be able to reproduce, and 1 means their reproductive period will not be affected by the onset of symptoms. A simple causal chain can be defined as:

$$\uparrow \text{CAGexp} \rightarrow \downarrow \text{AO} \rightarrow \downarrow \text{reproductive period} \rightarrow \downarrow antcoeff$$

2.2.3 Other variables to be included in the model

In addition to anticipation, the model needed to include fitness and segregation distortion, as mechanisms potentially associated with the long-term maintenance of polyQ diseases. Data on these forces were collected from previous systematic reviews (Sena et al., 2021a; 2021b). They are summarized in Table 1. Standard deviation (SD) values were needed to run the simulations. As there was no information about the SD of the segregation distortion in SCA2, the same SD found in SCA3/MJD segregation was assigned to the SCA2 model. Likewise, the SD of the fitness of both SCA2 and SCA3/MJD were lacking. In this case, the arbitrary value of 0.25 was imputed to them, as this value, although parsimonious, would allow for some overlap of individual fitness values between carriers and non-carriers.

Fitness, segregation rates and unstable transmissions associated with non-pathogenic alleles - those originated from contractions as well as the wildtype alleles - also need to be considered in the model. The fitness of the normal alleles is an *a priori* concept and is equal to one; the same SD of 0.25 were arbitrarily imputed to them. Transmission of intermediate-sized unexpanded *ATXN2* alleles was studied in the Cuban population (Almaguer-Mederos et al., 2018). Studies on the transmission of the normal *ATXN3* alleles were not found; a neutral value of zero was imputed to the mean, as there

is no evidence that this allele expands or contracts significantly in meiosis of the general population; its standard deviation was arbitrated as of 0.468, like that found in the *ATXN2*. The data obtained from observational studies and used in the models were also systematized in Table 1.

Although the assumed variances seem reasonable, several of them were not obtained from observational data. The arbitrated SD of fitness, in particular, could add uncertainty to the inputs. Different SD values of fitness were then imputed in a second round to check if outputs would be distorted, in a sensitivity analysis of the model.

2.3 Computer simulations on the dynamics of the expanded alleles

Means and SD of the variables considered so far, were used in the simulations on what occurs in the successive generations of a lineage. The emphasis on the use of SD was decided, in order that results could capture any potential scenario of real life. Thus, at each generational step, the simulated values brought the component of randomness to our results.

Measures of central tendency of descriptive variables were eventually presented as means (range) or as medians (range), according to the pattern of their distributions.

The simulations were performed in the R Statistical Package. The hypothetical initial frequency of the expanded allele was proposed to be 0.000001 for both *ATXN2* and *ATXN3* - sufficiently low to be considered plausible. The original ancestral expanded allele was proposed to correspond to the smallest length of the symptom-associated CAG repetitive sequences found in the datasets described in section 2.1 Materials - 34 and 54 CAGexp for SCA2 and SCA3/MJD, respectively (de Mattos et al., 2018; Sena et al., 2019). Despite that, the model attributed pathogenicity to tracts with 34 CAG repeats in ATXN2 and with 51 CAG repeats in ATXN3, following the information described in Figure 1. The AOfs attributed for the first ancestor carrying this allele was the average AOfs found for the length of this expansion, in the same IPDs obtained from other studies - 55 and 65 years of age for SCA2 and SCA3/MJD, respectively (de Mattos et al., 2018; Sena et al., 2019).

At least 1,000 different lineages with allele frequencies randomly generated in each generation, considering the values described above, were simulated per CAGexp ancestor, covering a maximum of 650 generations. One thousand runnings were done to warrant a sampling-based approach of the sensitive analysis of our outputs. The number of generations was chosen because it corresponds to circa 16,250 years, or the approximate age of the oldest SCA3/MJD SNP haplotype rs16999141, rs1048755, rs12895357, rs7158733 and rs3092822 known so far, the TTACAC or Joseph lineage (Li et al., 2019). Each of the 1,000 random lineages had their mean (SD) variables described above individually simulated in each generation. The values of the frequencies of the descendent alleles generated were re-entered into the equation to calculate the allele frequency in the subsequent generation. This process was carried out in loops until the frequency of the expanded allele reached 0 - the descendant alleles of the *de novo* expansion were eliminated -, 1 - when the descendant alleles of the *de novo* expansion reach a frequency of 100% in the population, a phenomenon called fixation -, or reached 650 generations - the proposed maximum observation time. The R source code to perform these simulations is described in the Supplementary Material S1.

3 Results

3.1 Frequencies of the expanded allele at *ATXN2* across generations

From the 1,000 lineages simulated as descendants of the ancestral expansion of 34 repeats in *ATXN2*, 933 were eliminated in a median value of 10 generations, the extinction ranging between the 2nd and 121st generations. The frequency of alleles eliminated by generation and the number of generations that the allele remained in the population are represented in Supplementary Material S3, Supplementary Figure S2A, and Figure 2A.

From these 1,000 simulated lineages, 67 were fixed after a median (range) of 60 (34–113) generations. In the generation in which the allele was fixed, the median (range) repeat length was of 32.00 (22.72–39.14) repeats - i.e., either non-pathogenic or borderline allele, in relation to SCA2 symptoms. The frequency of fixed alleles across generations are shown in Figure 3A. Figure 3B shows that all those 67 lineages in which the allele were fixed in the background population, turned out to be expanded, resulting in AOfs before the start of the reproductive period of life (Figure 3C). These lineages would then become extinct in up to the 170th generation (Table 2).

Results obtained with different imputed SDs of *ATXN2* fitness were described in Supplementary Material S4.

3.2 Frequencies of the expanded allele at *ATXN3* across generations

From the 1,000 lineages simulated as descendants of the ancestral expansion with 54 repeats in *ATXN3*, 593 were eliminated in a median of 29 generations, the extinction ranging between 3 and 649 generations. The median size of the CAG repeats when the lineages were eliminated was 84, ranging from 39.45 to 88. The frequency of alleles eliminated by generation and the histogram of the number of generations where the allele was deleted are shown in Figure 2B and in Supplementary Figure S2B.

Of the same 1,000 simulated lineages, 50 were fixed and their frequencies across the generation are shown in Figure 4A. Fixation occurred at a median (range) of 19.50 (14–63) generations. Figure 4B shows the repeat lengths until the allele was fixed. In all the fixed lineages, the allele was expanded when it became fixed: they had a mean (range) of 64.64 (53.20–75.86) CAG repeats; the mean (range) AO of their carriers was 50.34 (14.98–70.41) years. After turning fixed, further instabilities continued to occur in the descendants (Figure 4B). Of the 50 fixed *ATXN3* lineages, the 43 that continued to expand were eliminated due to severe anticipations in AO (Figure 4C); the seven lineages transmitted after the 650th generation presented a contraction, carrying a limitrophe allele between normal and pathogenic CAG repeat lengths (Figure 4B).

Finally, and more importantly, the 357 lineages where the simulated alleles were not eliminated nor fixed, had their

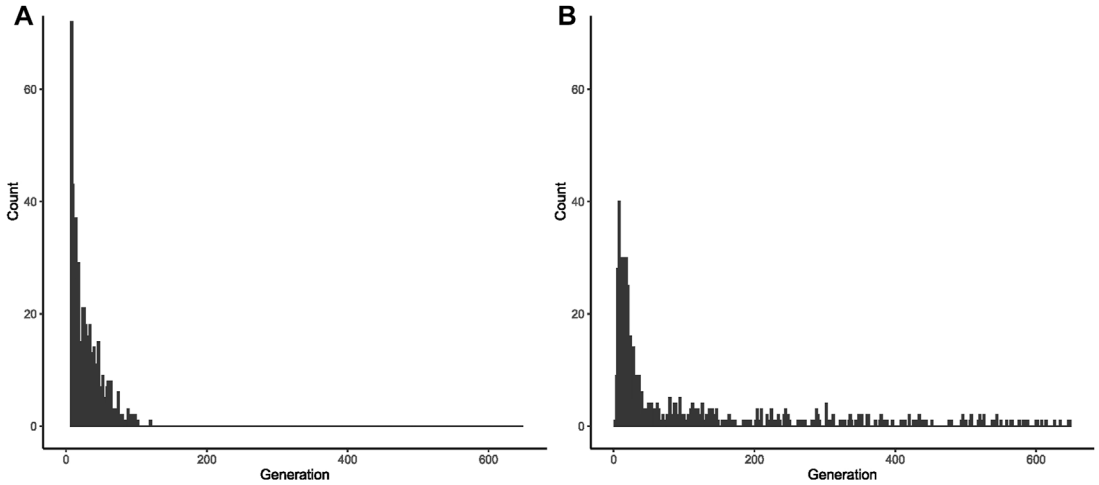

FIGURE 2
Fate of 1,000 lineages simulated as descendants of one ancestral with a CAG expansion and with an initial population frequency of 0.000001. **(A)** Proportion of descents with expanded repeats, per generation, after the first ancestor with 34 repeats in *ATXN2*. **(B)** Proportion of descents with expanded repeats, per generation, after the first ancestor with 54 repeats in *ATXN3*.

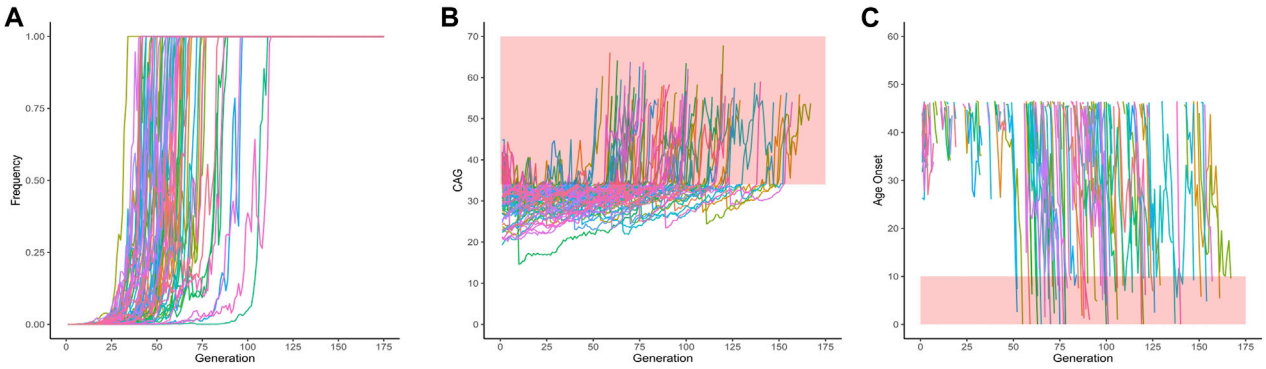

FIGURE 3
Data on the 67 lineages where fixed CAG expanded alleles at *ATXN2* were obtained after 1,000 simulations (simulations stopped when lineages were fixed). **(A)** Frequency of descendant alleles across generations until they were fixed. **(B)** The repeat lengths in the lineages where the descendant allele was fixed, per generation. The bold area in the graphic represents the pathological range of CAG repeat lengths (or CAGexp). **(C)** The predicted mean age at onset of gait ataxia per generation of each fixed lineage. The bold area in the graphic represents ages of onset younger than 10 years old. The lines describing the AO in each lineage were interrupted at 46.55 years of age, at the top of the chart, since this was the AO predicted to be related to the 34 CAG repeats, the shortest expansion in the pathological range.

TABLE 2 Comparisons between the fates of *ATXN2* and *ATXN3* lineages produced by computer simulations from their expanded ancestors, in the 650th generation.

	Lineages eliminated from the population	Fixed alleles		Lineages that remained in the population
		Lineages extinct after fixation	Lineages held as fixed	
ATXN2	933	67	0	0
ATXN3	593	43	7	357
p	<0.001	ns	ns	<0.001

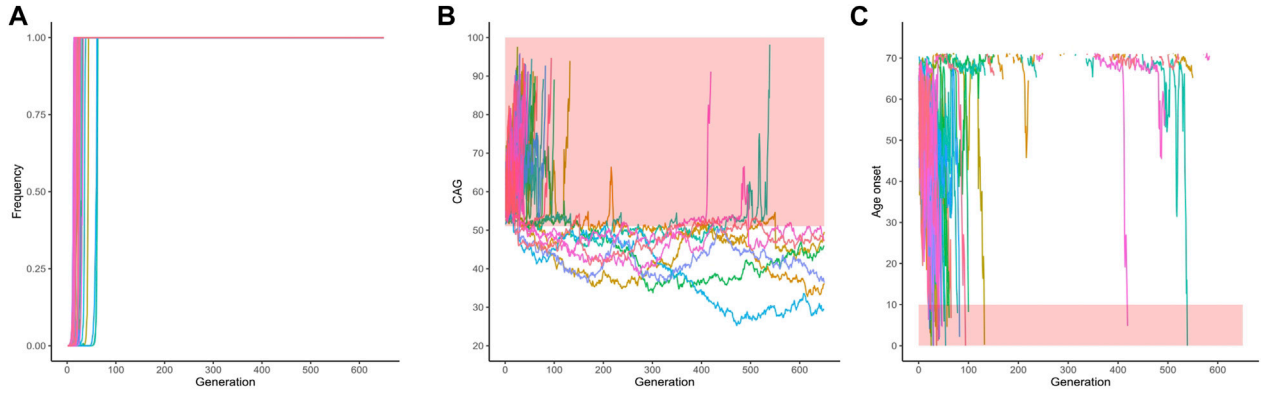

FIGURE 4
Data on the fifty lineages where fixed CAG expanded alleles at *ATXN3* were obtained after 1,000 simulations (simulations stopped when lineages were fixed). **(A)** Frequency of descendant alleles across generations until they were fixed. **(B)** The repeat lengths in the lineages where the descendant allele was fixed, per generation. The bold area in the graphic represents the pathological range of CAG repeat lengths (CAGexp). **(C)** The predicted mean age at onset of gait ataxia per generation of each fixed lineage. The bold area in the graphic represents ages of onset younger than 10 years old. The lines describing the AO in each lineage were interrupted at 71.14 years of age, at the top of the chart. They were related to CAG repeat sizes that were outside the pathological range, according to the regression.

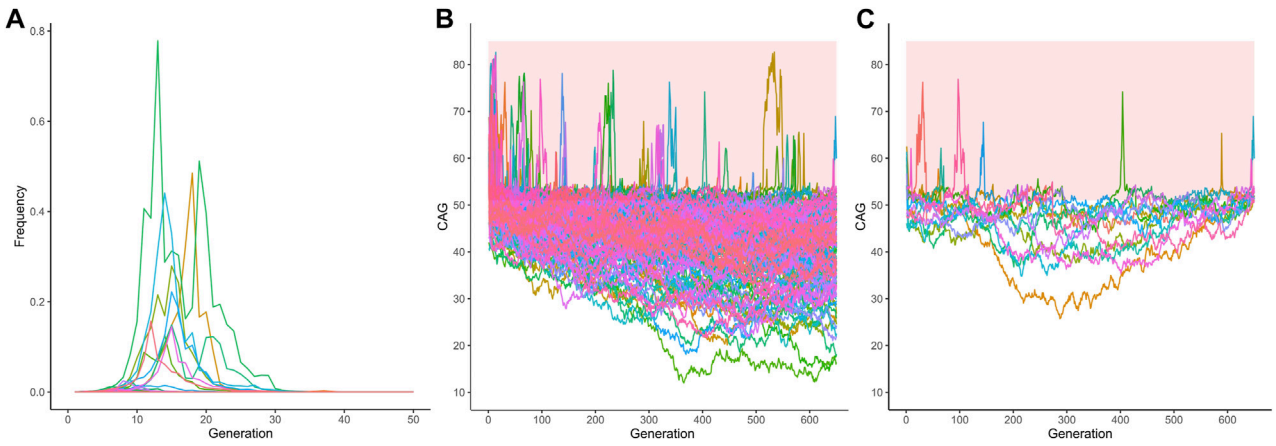

FIGURE 5
The 351 lineages of the CAG expanded alleles at *ATXN3* that remained non-fixed in the 1,000 simulations. **(A)** Frequency of descendant alleles until the 650th generation. Most have residual frequencies, very close to zero, and are not discernible on the graph. **(B)** The repeat lengths in these lineages, per generation. Note the transitions between the pathogenic and non-pathogenic CAG repeat lengths. The bold area in the graphic represents the pathological range of CAG repeat lengths (CAGexp). **(C)** The repeat lengths in the 13 lineages that presented expanded alleles in generation 650. The bold area in the graphic represents the pathological range of CAG repeat lengths (CAGexp). The repeat lengths alternate between the expanded and unexpanded ranges in generations that antecede the 650th one.

frequencies across generations shown in Figure 5A and Figure 5B. These 357 lineages that remained in the population without fixed alleles showed CAG repeats transitioning between the expanded (equal or larger than 51 repeats) and non-expanded ranges (Figure 5C), suggesting that this phenomenon might make a lineage to survive for many centuries. Of note, the allele frequencies of the expanded repeats (51 repeats or more) remained very low and reached a median (IQR) 2.7e-107 (1.960994e-108) of in the 650th generation. When the 13 lineages that presented expanded alleles in generation 650 are presented separately, this alternation between the expanded and non-expanded allele can be more clearly observed (Figure 5D). Table 2 summarizes these different fates of SCA3/MJD lineages and compares them to those of SCA2 lineages.

Results obtained with different imputed SDs of *ATXN3* fitness were described in Supplementary Material S4.

4 Discussion

Data on prevalence and anticipation have been difficult to put together in a unified biological explanation for polyQ diseases, as they are in contradiction with each other. Phenomena such as increased fitness and preferential segregation of the mutant alleles

were then proposed to balance anticipation. But the empirical results, either because they were sparse or heterogeneous, kept the explanatory hypotheses in abeyance. The present model on the dynamics of expanded CAG alleles obtained compatible scenarios with current epidemiology, precisely using the empirical data on anticipation available now. Our model, like other population genetics models, was intended to capture some approximations of reality. Thus, our results, as they are closer to empirical evidence, can help to develop an acceptable and understandable explanation about why polyQ diseases can be present in populations for long periods of time.

The term "dynamics of the CAGexp" means the intrinsic, mutation-driven pattern of change in time of CAGexp which should be, by its nature, a multifactorial event. These dynamics might depend on the repeat motif itself (length, interruptions, etc.), on the surrounding sequence, and on other factors that interplay with this surrounding context (sex and parental age, for instance) (Andrés et al., 2003). To model the dynamics of the CAGexp, we have used computer simulation, adding to so many other applications related to population structure and evolutionary genetics (Hoban et al., 2012).

Available empirical data on the evolutionary mechanisms at work on polyQs are certainly incomplete. Despite this, with data already available, the model predicted intergenerational dynamics that ended up being compatible with both apparently incoherent empirical data from SCA3/MJD - the few ancestral lineages with a long survival - and also the relatively more coherent facts associated with SCA2 - the serious anticipations due to dramatic expansions described in the literature and the multiple ancestral lineages (Sena et al., 2023), both compatible with short survivals of its lineages.

In common, the CAGexp dynamics first went through an increase followed by a decrease in the frequencies of expanded alleles in successive generations. But the CAGexp alleles at *ATXN2* and *ATXN3* followed quite different trajectories (Table 2). The *ATXN3* allele might remain longer in the population, a fate due to the bias in favor of the expanded allele in the segregation of gametes, to favorable fitness, and to the less intense instability and anticipation than in relation to the expanded allele in *ATXN2* (Supplementary Figure S3).

In contrast, the rise and fall of frequencies were quite sharp in SCA2. The model predicted that any real expansion in *ATXN2* would have a chance close to 98.6% of becoming extinct approximately 10 generations after its appearance. In this scenario, SCA2 recurrence in human populations distant as those of India, Cuba, and others, would depend upon *de novo* expansions.

The normal $(CAG)_{22}$ allele in *ATXN2* is the most prevalent in the population (Laffita-Mesa et al., 2012; Gardiner et al., 2019). CAG repeats at *ATXN2* are those with the lowest variance and allelic heterozygosity, between the *loci* related to polyQ diseases (Andrés et al., 2002). These characteristics largely stem from the $(CAG)_{22}$ allele being favored in meiotic segregation (Yu et al., 2005; Chen et al., 2013). The internal sequence of this allele $(CAG)_{22}$ contains two CAA interruptions, an important factor of stability in CAG repeats (Choudhry et al., 2001).

Given the positive selection of the $(CAG)_{22}$ allele and the tendency of the expanded alleles to be rapidly withdrawn after they appear, SCA2 lineages should quickly disappear, and *de novo* expansions should be the most likely reason for the maintenance of SCA2 in populations. Our finding of at least eleven different ancestral SCA2 haplotypes among South American families are in line with this interpretation (Sena et al., 2023). This might be a case of mutation–selection balance. Intermediate or pre-expanded alleles associated with an unstable haplotype are the main sources of *de novo* expansions in HD (Warby et al., 2009). The same might happen in SCA2. Indeed, an intermediate 32-repeat allele was detected in the asymptomatic father of a sporadic ataxic subject carrying 35 CAGexp in *ATXN2* (Futamura et al., 1998). Although this was a unique confirmed report in the literature, it is worth reminding that *de novo* expansions most probably give rise to mild expansions and therefore to clinical manifestations very late in life, when parents are more frequently deceased and there is no way to document the phenomenon. Although the event might be very rare, it is necessary to clarify whether *de novo* expansions in *ATXN2* would come from predisposing haplotypes, with normal CAG repeats prone to instabilities and expansions when crossing meiosis. Comparisons among species suggest that $C^{rs695871}$-$C^{rs695872}$ (CC) is the oldest ancestral haplotype (Choudhry et al., 2001). We can speculate that GT haplotypes - and mostly with $(CAG)_{22}$ - might be associated with more stable repeats than the CC haplotypes but analyzes with more markers are necessary to answer this question.

There is a continuum in the distribution of CAG alleles in *ATXN2* found in controls and in SCA2 carriers. The lack of a gap between normal and SCA2-associated alleles and the occurrence of *de novo* cases are peculiar characteristics that occur not only in SCA2 but also in other polyQ diseases, such as SCA6, SCA7, and HD (Figure 1). It is possible that the dynamics of the expanded allele of these other polyQs might be like that described here for SCA2.

In contrast, and as said before, SCA3/MJD is a polyQ disorder somehow different from most others, combining few ancestral haplotypes with a long-term permanence across generations (Martins et al., 2007). These facts were in apparent contradiction with the serious anticipations registered in several SCA3/MJD cohorts, in such a way that one data or the other could be viewed with some doubt (Maciel et al., 1995; Souza et al., 2016). Given its segregation distortion, fitness and *antcoeff* data, our mathematical model predicted that any expansion in *ATXN3* would have at least a 36.4% chance of lasting up to 650 generations. There was a substantial variability in the length of the transmitted repeats modeled here, and 13 of the 1,000 simulations reached the 650th generation as expanded. In fact, the mathematical model generated *ATXN3* lineages that appear to be able to stay indefinitely. We interpret that the segregation that favors the expanded *ATXN3* allele and the high fitness of SCA3/MJD are sufficient factors to explain the preservation of ancient ancestral haplotypes, such as the (at least) 16,000 years old TTACAC or Joseph lineage (Martins and Sequeiros, 2018; Li et al., 2019), until today. It is necessary to reflect that, although the actual ancient TTACAC alleles have not yet disappeared, it might be a matter of time for that to happen. In any case, our mathematical model not only supports the verisimilitude of the empirical data collected so far on SCA3/MJD, but also unifies them into a coherent set.

Our simulations have also revived an old hypothesis to explain the antiquity of the ancestral generations of SCA3/MJD: the existence of a haplotype that predisposes to expansions (Maciel et al., 1999). Although the TTACAC alleles carried by patients nowadays

appear to have been passed on from a common ancestor for 650 generations - 16,000 years or more - this is not to say that ancestors with ataxic symptoms were frequent. Figure 4C, 5C and 5D denote that most of the expanded *ATXN3* lineages in the 650th generation previously oscillated between 37 and 50 CAG repeats, a range in which repeats do not produce symptoms. Although our model well supports the hypothesis that a haplotype might predispose for expansions, clinical and laboratory data published so far are less favorable, since SCA3/MJD carriers without a biological parent carrying one expansion have never been described in the literature. As far as we are aware, alleles in the range between normal and pathogenic have been described twice, but the authors did not clarify whether these alleles were non-penetrant (intermediate) or penetrant (pathogenic). Two subjects were detected in a collateral branch of a SCA3/MJD family, asymptomatic at 35 and 67 years, and carrying one *ATXN3* allele of 51 repeats (Maciel et al., 2001). Another study measured the *ATXN3* CAG repeats in 16,547 subjects from five European population-based cohorts and detected alleles with 46 and 49 CAG repeats (Gardiner et al., 2019); the authors mentioned that they "had no long-term follow-up data on the participants", being "unable to confirm if carriers would have developed disease symptoms".

The lineages modeled here refer to the descendants of a first expansion carrier. It is interesting to consider the population environment as well, to understand the uncommon occurrence or even the lack of intermediate alleles in *ATXN3*. We have seen that short *ATXN3* alleles are transmitted preferentially in meiosis in the general population (Sena et al., 2021a). This is the opposite of what happens in affected individuals, where the expanded allele is preferentially transmitted. Therefore, a disruptive selection takes place in *ATXN3*, that is, a selection in favor of extreme characteristics - favoring short alleles and expanded alleles and creating a gap between them. This gap would also explain the absence or ultra-rarity of intermediate alleles. In addition, it would also partially explain why *de novo* mutations have not been described in SCA3/MJD to date.

In any case, the fact that CAG tracts in *ATXN2*, *ATXN7* and *HTT*, among others, are prone to *de novo* expansion, while CAG tracts in *ATXN3* do not appear to be, needs to be further elucidated by observational and/or experimental studies. Discovering the reason for this discrepancy can have an impact even for future therapeutic or preventive management. One might suspect, for instance, that the CAG repeat of *ATXN3* has structural features in wild type alleles that confers a strong protection against instabilities. Or that the preferential segregation of the shortest allele in the presence of two normal alleles is a force to prevent a novel expansion of a normal allele originated from a contraction of an originally expanded allele.

To date, there is no indication whether the dynamics of the expanded allele in *ATXN3* finds similes among other polyQs. The best candidates to share these dynamics would be the polyQs for which no intermediate alleles were found, such as in DRPLA and in SBMA (Figure 1). In the case of SBMA, X-linked inheritance adds complexity to the description of intermediate alleles. DRPLA has additional similarities with SCA3/MJD: the gap between the normal range and the pathogenic range of CAGn is large (Figure 1); CAGexp on *ATN1* is favorably segregated (Ikeuchi et al., 1996); there appear to be very few ancestral haplotypes (Martins et al., 2003); and the existence of neurological subtypes, where manifestations are qualitative different, depending on AO.

Fixation of some descendant CAGexp alleles was a counterintuitive result of our model, but occurred in a minority of simulations, i.e., in 6.7% and 5% of *ATXN2* and *ATXN3*, respectively. The average CAG lengths of fixed lineages were 32 and 64 CAGexp for *ATXN2* and *ATXN3*, which are related to the late onset (or non-penetrant range, in *ATXN2* case) of the disease; part of them did not expand during fixation, thus not reducing the reproductive period of their carriers. When the simulations were continued, all fixed *ATXN2* lineages were eliminated due to the severe anticipation (Figure 3C). The 14 fixed *ATXN3* lineages at generation 650 had an unexpanded CAG repeat size; in some situations, their descents transited to the expanded range later (Figure 4B). Indeed, these scenarios seem highly hypothetical.

It is important to point out some weaknesses due to the lack of concrete data to include in the model. By running the model a large number of times, we tried to reduce the uncertainty of the outputs. But some inputs can still raise concerns. First, although we set a cutoff of 650 generations for the simulations, our fertility parameters were based on available data on the contemporary European populations. It is almost certain that in the past people had their children at an earlier age than today. The precocity of the general reproductive period would have meant a relaxation of the selective forces that acted to eliminate CAGexp from the population. As a result, perhaps the lineages of any CAGexp lasted longer in the deep past than we calculate here. This bias, however, would have occurred equivalently in the ATXN2 and ATXN3 lineages.

Second, some arbitrary values were included in the model (Table 1). Of those, the SDs of the fitness for the four categories of alleles seemed to be potentially problematic. However, after running the model with other imputed variances of the fitness, the results presented very similar fates to those obtained with the initial SD values, resulting in the extinction of all SCA2 lineages and in the survival of several SCA3/MJD lineages until the 650th generation (Supplementary Material S4).

One can also speculate whether, in the past, other CAG repeats *loci* would have undergone pathogenic expansions causing maladaptive phenotypes. And that these alleles could have become extinct, so that the phenomenon would go unrecorded and be lost in human history.

Finally, as a theoretical work, the present study raised not direct evidence, but probabilities from existing empirical data on selective forces that converged with current epidemiology. It is worth emphasizing that we were interested in clarifying the effects of past history on present and not on future prevalences. Even so, further studies on prevalence and on ancestral haplotypes are still required to amplify these generalizations. Prevalence of SCA3/MJD in the Azores archipelago increased between 1981 and 2015 (de Araújo et al., 2016). Similarly, prevalence of SCA1 in the Sakha (Yakut) people of Eastern Siberia increased between 1994 and 2013 (Platonov et al., 2016). Monitoring the frequencies of polyQ diseases is relevant to keep track of eventual changes and to clarify the effectiveness of our model, in perspective.

In conclusion, the general dynamics of the CAGexp alleles seems to follow an increase in frequency for a few generations, followed by a decrease in frequency. Expanded *ATXN2* alleles showed a clear and rapid tendency to be eliminated from the population. Their maintenance in human populations must be explained by *de novo* expansions. To the contrary, expanded *ATXN3* alleles

showed a tendency to remain longer in the population, a phenomenon explained at least by the favorable fitness, by the distortion in favor of the expanded allele or by a less intense instability and anticipation when compared to the expanded allele in *ATXN2*. These results contribute to the understanding of the survival of ancient origins for the *ATXN3* expansions. Finally, we think that the present mathematical model, combined with evidence of specific selective forces, can be used to simulate the dynamics of expanded alleles in other polyQ diseases.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://ec.europa.eu/eurostat/data/database doi: 10.1111/cge.13978 doi: 10.1111/cge.13888 doi: 10.1136/jnnp-2018-319200.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

LS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing–original draft. RL: Formal Analysis, Methodology, Supervision, Writing–review and editing. GF: Data curation, Investigation, Writing–review and editing. MS-P: Data curation, Supervision, Writing–review and editing. LJ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Resources, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by Financiamento e Incentivo à Pesquisa do Hospital de Clínicas de Porto Alegre (FIPE-HCPA) (grant numbers 2019-0254 and 2019-0169). LS, MS-P, and LJ were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

Acknowledgments

The present work was posted as a preprint on bioRxiv on 9 September 2023 with the doi doi.org/10.1101/2023.09.07.556735.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1296614/full#supplementary-material

References

Adegbuyiroa, A., Sedighia, F., Pilkington, A. W., Groovera, S., and Legleitera, J. (2017). Proteins containing expanded polyglutamine tracts and neurodegenerative disease. *Biochemistry* 56 (9), 1199–1217. doi:10.1021/acs.biochem.6b00936

Almaguer-Mederos, L. E., Mesa, J. M. L., González-Zaldívar, Y., Almaguer-Gotay, D., Cuello-Almarales, D., Aguilera-Rodríguez, R., et al. (2018). Factors associated with ATXN2 CAG/CAA repeat intergenerational instability in Spinocerebellar ataxia type 2. *Clin. Genet.* 94, 346–350. doi:10.1111/cge.13380

Andrés, A. M., Lao, O., Soldevila, M., Calafell, F., and Bertranpetit, J. (2003). Dynamics of CAG repeat loci revealed by the analysis of their variability. *Hum. Mutat.* 21, 61–70. doi:10.1002/humu.10151

Bech, S., Petersen, T., Nørremølle, A., Gjedde, A., Ehlers, L., Eiberg, H., et al. (2010). Huntington's disease-like and ataxia syndromes: identification of a family with a *de novo* SCA17/TBP mutation. *Park. Relat. Disord.* 16, 12–15. doi:10.1016/j.parkreldis.2009.06.006

Bird, T. D. (1998). "Hereditary ataxia overview," in GeneReviews®. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

Bunting, E. L., Hamilton, J., and Tabrizi, S. J. (2021). Polyglutamine diseases. *Curr. Opin. Neurobiol.* 72, 39–47. doi:10.1016/j.conb.2021.07.001

Caron, N. S., Wright, G. E. B., and Hayden, M. R. (1998). "Huntington disease," in GeneReviews®. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

Carroll, L. S., Massey, T. H., Wardle, M., and Peall, K. J. (2018). Dentatorubral-pallidoluysian atrophy: an update. *Tremor Other Hyperkinet Mov.* 8, 577. doi:10.7916/D81N9HST

Casey, H. L., and Gomez, C. M. (1998). "Spinocerebellar ataxia type 6," in GeneReviews. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

Chen, X. C., Sun, H., Zhang, C. J., Zhang, Y., Lin, K. Q., Yu, L., et al. (2013). Positive selection of CAG repeats of the ATXN2 gene in Chinese ethnic groups. *J. Genet. Genomics* 40, 543–548. doi:10.1016/j.jgg.2013.08.003

Choudhry, S., Mukerji, M., Srivastava, A. K., Jain, S., and Brahmachari, S. K. (2001). CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. *Hum. Mol. Genet.* 10, 2437–2446. doi:10.1093/hmg/10.21.2437

Cruz-Mariño, T., Laffita-Mesa, J. M., Gonzalez-Zaldivar, Y., Velazquez-Santos, M., Aguilera-Rodriguez, R., Estupinan-Rodriguez, A., et al. (2014). Large normal and intermediate alleles in the context of SCA2 prenatal diagnosis. *J. Genet. Couns.* 23, 89–96. doi:10.1007/s10897-013-9615-1

Cubo, E., Martinez-Horta, S. I., Santalo, F. S., Descalls, A. M., Calvo, S., Gil-Polo, C., et al. (2019). Clinical manifestations of homozygote allele carriers in Huntington disease. *Neurology* 92 (18), e2101–e2108. doi:10.1212/WNL.0000000000007147

David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., et al. (1997). Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nat. Genet.* 17, 65–70. doi:10.1038/ng0997-65

de Araújo, M. A., Raposo, M., Kazachkova, N., Vasconcelos, J., Kay, T., et al. (2016). Trends in the epidemiology of spinocerebellar ataxia type 3/machado-joseph disease in the Azores islands, Portugal. *JSM Brain Sci.* 1 (1), 1001.

de Castilhos, R. M., Furtado, G. V., Gheno, T. C., Schaeffer, P., Russo, A., Barsottini, O., et al. (2014). Spinocerebellar ataxias in Brazil-frequencies and modulating effects of related genes. *Cerebellum* 13, 17–28. doi:10.1007/s12311-013-0510-y

de Mattos, E. P., Kolbe Musskopf, M., Bielefeldt Leotti, V., Saraiva-Pereira, M. L., and Jardim, L. B. (2019). Genetic risk factors for modulation of age at onset in Machado-Joseph disease/spinocerebellar ataxia type 3: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* 90, 203–210. doi:10.1136/jnnp-2018-319200

Diallo, A., Jacobi, H., Cook, A., Labrum, R., Durr, A., Brice, A., et al. (2018). Survival in patients with spinocerebellar ataxia types 1, 2, 3, and 6 (EUROSCA): a longitudinal cohort study. *Lancet Neurol.* 17 (4), 327–334. doi:10.1016/S1474-4422(18)30042-5

Europa, (2023). Fertility indicators. https://ec.europa.eu/eurostat/databrowser/view/demo_find/default/table?lang=en.

Futamura, N., Matsumura, R., Fujimoto, Y., Horikawa, H., Suzumura, A., and Takayanagi, T. (1998). CAG repeat expansions in patients with sporadic cerebellar ataxia. *Acta Neurol. Scand.* 98, 55–59. doi:10.1111/j.1600-0404.1998.tb07378.x

Gardiner, S. L., Boogaard, M. W., Trompet, S., de Mutsert, R., Rosendaal, F. R., Gussekloo, J., et al. (2019). Prevalence of carriers of intermediate and pathological polyglutamine disease-associated alleles among large population-based cohorts. *JAMA Neurol.* 76, 650–656. doi:10.1001/jamaneurol.2019.0423

Gu, W., Ma, H., Wang, K., Jin, M., Zhou, Y., Liu, X., et al. (2004). The shortest expanded allele of the MJD1 gene in a Chinese MJD kindred with autonomic dysfunction. *Eur. Neurol.* 52, 107–111. doi:10.1159/000080221

Gusella, J. F., Lee, J. M., and MacDonald, M. E. (2021). Huntington's disease: nearly four decades of human molecular genetics. *Mol. Genet.* 30, 254–263. doi:10.1093/hmg/ddab170

Hoban, S., Bertorelle, G., and Gaggiotti, O. (2012). Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* 13, 110–122. doi:10.1038/nrg3130

Ikeuchi, T., Igarashi, S., Takiyama, Y., Onodera, O., Oyake, M., Takano, H., et al. (1996). Non-Mendelian transmission in dentatorubral-pallidoluysian atrophy and Machado-Joseph disease: the mutant allele is preferentially transmitted in male meiosis. *Am. J. Hum. Genet.* 58, 730–733.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., et al. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14Q32.1. *Nat. Genet.* 8, 221–228. doi:10.1038/ng1194-221

Kay, C., Collins, J. A., Wright, G. E. B., Baine, F., Miedzybrodzka, Z., Aminkeng, F., et al. (2018). The molecular epidemiology of Huntington disease is related to intermediate allele frequency and haplotype in the general population. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 177 (3), 346–357. doi:10.1002/ajmg.b.32618

Koide, R., Kobayashi, S., Shimohata, T., Ikeuchi, T., Maruyama, M., Saito, M., et al. (1999). A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? *Hum. Mol. Genet.* 8, 2047–2053. doi:10.1093/hmg/8.11.2047

Komure, O., Sano, A., Nishino, N., Yamauchi, N., Ueno, S., Kondoh, K., et al. (1995). DNA analysis in hereditary Dentatorubral-Pallidoluysian Atrophy – correlation between CAG repeat length and phenotypic variation and the molecular-basis of anticipation. *Neurology* 45, 143–149. doi:10.1212/wnl.45.1.143

Laffita-Mesa, J. M., Velázquez-Pérez, L. C., Santos Falcón, N., Cruz-Mariño, T., González Zaldívar, Y., Vázquez Mojena, Y., et al. (2012). Unexpanded and intermediate CAG polymorphisms at the SCA2 locus (ATXN2) in the Cuban population: evidence about the origin of expanded SCA2 alleles. *Eur. J. Hum. Genet.* 20 (1), 41–49. doi:10.1038/ejhg.2011.154

La Spada, A. (1999). "Spinal and bulbar muscular atrophy," in *GeneReviews®*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

La Spada, A. R. (1998). "Spinocerebellar ataxia type 7," in *GeneReviews®*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E., and Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 352, 77–79. doi:10.1038/352077a0

Lee, J. M., Ramos, E. M., Lee, J. H., Gillis, T., Mysore, J. S., Hayden, M. R., et al. (2012). CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* 78, 690–695. doi:10.1212/WNL.0b013e318249f683

Li, T., Martins, S., Peng, Y., Wang, P., Hou, X., Chen, Z., et al. (2019). Is the high frequency of machado-joseph disease in China due to new mutational origins? *Front. Genet.* 9, 740. doi:10.3389/fgene.2018.00740

Lieberman, A. P., Shakkottai, V. G., and Albin, R. L. (2019). Polyglutamine repeats in neurodegenerative diseases. *Annu. Rev. Pathol.* 14, 1–27. doi:10.1146/annurev-pathmechdis-012418-012857

MacCluer, J. W., VandeBerg, J. L., and Read B, R. O. A. (1986). Pedigree analysis by computer simulation. *Zoo. Biol.* 5, 147–160. doi:10.1002/zoo.1430050209

Maciel, P., Costa, M. C., Ferro, A., Rousseau, M., Santos, C. S., Gaspar, C., et al. (2001). Improvement in the molecular diagnosis of Machado-Joseph disease. *Arch. Neurol.* 58, 1821–1827. doi:10.1001/archneur.58.11.1821

Maciel, P., Gaspar, C., DeStefano, A. L., Silveira, I., Coutinho, P., Radvany, J., et al. (1995). Correlation between CAG repeat length and clinical features in Machado-Joseph disease. *Am. J. Hum. Genet.* 57, 54–61.

Maciel, P., Gaspar, C., Guimarães, L., Goto, J., Lopes-Cendes, I., Hayes, S., et al. (1999). Study of three intragenic polymorphisms in the Machado-Joseph disease gene (MJD1) in relation to genetic instability of the (CAG)n tract. *Eur. J. Hum. Genet.* 7, 147–156. doi:10.1038/sj.ejhg.5200264

Margolis, R. L., Rudnicki, D. D., and Holmes, S. E. (2005). Huntington's disease like-2: review and update. *Acta Neurol. Taiwan* 14, 1–8.

Martins, S., Calafell, F., Gaspar, C., Wong, V. C., Silveira, I., Nicholson, G. A., et al. (2007). Asian origin for the worldwide-spread mutational event in Machado-Joseph disease. *Arch. Neurol.* 64, 1502–1508. doi:10.1001/archneur.64.10.1502

Martins, S., Matamá, T., Guimarães, L., Vale, J., Guimarães, J., Ramos, L., et al. (2003). Portuguese families with dentatorubropallidoluysian atrophy (DRPLA) share a common haplotype of Asian origin. *Eur. J. Hum. Genet.* 11, 808–811. doi:10.1038/sj.ejhg.5201054

Martins, S., and Sequeiros, J. (2018). Origins and spread of machado-joseph disease ancestral mutations events. *Adv. Exp. Med. Biol.* 1049, 243–254. doi:10.1007/978-3-319-71779-1_12

Mittal, U., Roy, S., Jain, S., Srivastava, A. K., and Mukerji, M. (2005). Post-zygotic *de novo* trinucleotide repeat expansion at spinocerebellar ataxia type 7 locus: evidence from an Indian family. *J. Hum. Genet.* 50, 155–157. doi:10.1007/s10038-005-0233-0

Opal, P., and Ashizawa, T. (1998). "Spinocerebellar ataxia type 1," in *GeneReviews®*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

Orr, H. T., Chung, M.-y., Banfi, S., Kwiatkowski, T. J., Servadio, A., Beaudet, A. L., et al. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.* 4, 221–226. doi:10.1038/ng0793-221

Paulson, H., and Shakkottai, V. (1998). "Spinocerebellar ataxia type 3," in GeneReviews®. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon. et al. (Seattle, Washington, D.C, USA: University of Washington).

Pereira, F. S., Monte, T. L., Locks-Coelho, L. D., Silva, A. S., Barsottini, O., Pedroso, J. L., et al. (2015). ATXN3, ATXN7, CACNA1A, and RAI1 genes and mitochondrial polymorphism A10398G did not modify age at onset in spinocerebellar ataxia type 2 patients from South America. *Cerebellum* 14, 728–730. doi:10.1007/s12311-015-0666-8

Platonov, F. A., Tyryshkin, K., Tikhonov, D. G., Neustroyeva, T. S., Sivtseva, T. M., Yakovleva, N. V., et al. (2016). Genetic fitness and selection intensity in a population affected with high-incidence spinocerebellar ataxia type 1. *Neurogenetics* 17, 179–185. doi:10.1007/s10048-016-0481-5

Prestes, P. R., Saraiva-Pereira, M. L., Silveira, I., Sequeiros, J., and Jardim, L. B. (2008). Machado-Joseph disease enhances genetic fitness: a comparison between affected and unaffected women and between MJD and the general population. *Ann. Hum. Genet.* 72, 57–64. doi:10.1111/j.1469-1809.2007.00388.x

Pulst, S. M. (1998). "Spinocerebellar ataxia type 2," in *GeneReviews®*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

Pulst, S. M., Santos, N., Wang, D., Yang, H. Y., Huynh, D., Velazquez, L., et al. (2005). Spinocerebellar ataxia type 2: polyQ repeat variation in the CACNA1A calcium channel modifies age of onset. *Brain* 128, 2297–2303. doi:10.1093/brain/awh586

Ramos, E. M., Martins, S., Alonso, I., Emmel, V. E., Saraiva-Pereira, M. L., Jardim, L. B., et al. (2010). Common origin of pure and interrupted repeat expansions in spinocerebellar ataxia type 2 (SCA2). *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 153B (2), 524–531. doi:10.1002/ajmg.b.31013

Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., et al. (1996). Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nat. Genet.* 14, 277–284. doi:10.1038/ng1196-277

Saute, J. A. M., and Jardim, L. B. (2015). Machado Joseph disease: clinical and genetic aspects, and current treatment. *Expert Opin. Orphan Drugs* 3, 517–535. doi:10.1517/21678707.2015.1025747

Sena, L. S., Castilhos, R. M., Mattos, E. P., Furtado, G. V., Pedroso, J. L., Barsottini, O., et al. (2019). Selective forces related to spinocerebellar ataxia type 2. *Cerebellum* 18, 188–194. doi:10.1007/s12311-018-0977-7

Sena, L. S., Dos Santos Pinheiro, J., Hasan, A., Saraiva-Pereira, M. L., and Jardim, L. B. (2021a). Selective forces acting on spinocerebellar ataxia type 3/Machado-Joseph disease recurrency: a systematic review and meta-analysis. *Clin. Genet.* 100, 347–358. doi:10.1111/cge.13888

Sena, L. S., Dos Santos Pinheiro, J., Saraiva-Pereira, M. L., and Jardim, L. B. (2021b). Selective forces acting on spinocerebellar ataxia type 3/Machado-Joseph disease recurrency: a systematic review and meta-analysis. *Clin. Genet.* 99, 347–358. doi:10.1111/cge.13888

Sena, L. S., Furtado, G. V., Fagundes, N. J. R., Pedroso, J. L., Barsottini, O., Ribeiro, P., et al. (2023). Spinocerebellar ataxia type 2 has multiple ancestral origins. https://www.medrxiv.org/content/10.1101/2023.09.12.23295432v1.

Shizuka, M., Watanabe, M., Ikeda, Y., Mizushima, K., Okamoto, K., and Shoji, M. (1998). Molecular analysis of a *de novo* mutation for spinocerebellar ataxia type 6 and (CAG)n repeat units in normal elder controls. *J. Neurol. Sci.* 161, 85–87. doi:10.1016/s0022-510x(98)00270-6

Sonakar, A. K., Shamim, U., Srivastava, M. P., Faruq, M., and Srivastava, A. K. (2021). SCA2 in the Indian population: unified haplotype and variable phenotypic patterns in a large case series. *Park. Relat. Disord.* 89, 139–145. doi:10.1016/j.parkreldis.2021.07.011

Souza, G. N., Kersting, N., Krum-Santos, A. C., Santos, A. S., Furtado, G. V., Pacheco, D., et al. (2016). Spinocerebellar ataxia type 3/Machado-Joseph disease: segregation patterns and factors influencing instability of expanded CAG transmissions. *Clin. Genet.* 90, 134–140. doi:10.1111/cge.12719

Stevanin, G., Giunti, P., Belal, G. D., Dürr, A., Ruberg, M., Wood, N., et al. (1998). *De novo* expansion of intermediate alleles in spinocerebellar ataxia 7. *Hum. Mol. Genet.* 7 (11), 1809–1813. doi:10.1093/hmg/7.11.1809

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell* 72, 971–983. doi:10.1016/0092-8674(93)90585-e

Toyoshima, Y., Onodera, O., Yamada, M., et al. (2005). "Spinocerebellar ataxia type 17," in *GeneReviews®*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Seattle, Washington, D.C, USA: University of Washington).

Warby, S. C., Montpetit, A., Hayden, A. R., Carroll, J. B., Butland, S. L., Visscher, H., et al. (2009). CAG expansion in the Huntington disease gene is associated with a specific and targetable predisposing haplogroup. *Am. J. Hum. Genet.* 84, 351–366. doi:10.1016/j.ajhg.2009.02.003

Yu, F., Sabeti, P. C., Hardenbol, P., Fu, Q., Fry, B., Lu, X., et al. (2005). Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLoS Genet.* 1 (3), e41. doi:10.1371/journal.pgen.0010041

Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D. W., Amos, C., et al. (1997). Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat. Genet.* 15, 62–69. doi:10.1038/ng0197-62