Mutational Mechanism for *DAB1* $(ATTTC)_n$ insertion in SCA37: ATTTT repeat lengthening and nucleotide substitution

Joana R Loureiro[1, 2, 3], Cláudia Oliveira[1, 2], Carolina Mota[1, 2, 4], Ana Filipa Castro[1, 2, 5], Cristina Costa[6], José Leal Loureiro[2,7,8], Paula Coutinho[2,7], Sandra Martins[9, 10], Jorge Sequeiros[2, 3, 7], Isabel Silveira[1, 2]

[1] Genetics of Cognitive Dysfunction Laboratory, i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4200-135 Porto, Portugal; [2] IBMC- Institute for Molecular and Cell Biology, Universidade do Porto, 4200-135 Porto, Portugal; [3] ICBAS, Universidade do Porto, 4050-313 Porto, Portugal; [4] Universidade de Aveiro, 3810-193 Aveiro, Portugal; [5] Faculdade de Ciências da Universidade do Porto, 4169-007 Porto, Portugal; [6] Department of Neurology, Hospital Pr. Dr. Fernando Fonseca, E.P.E, 2720-276 Amadora, Portugal; [7] UnIGENe, i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4200-135 Porto, Portugal; [8] Department of Neurology, Hospital São Sebastião, 4520-211 Feira, Portugal; [9] Population Genetics & Evolution, i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4200-135 Porto, Portugal; [10] IPATIMUP - Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal.

Correspondence: Isabel Silveira, PhD
Genetics of Cognitive Dysfunction Laboratory, i3S-Instituto de Investigação e Inovação em Saúde and IBMC – Institute for Molecular and Cell Biology, Universidade do Porto, Portugal
Rua Alfredo Allen, 208
4200-135 Porto, Portugal
E-mail: isilveir@ibmc.up.pt
Phone: +351-22-6074928

ABSTRACT

Dynamic mutations by microsatellite instability are the molecular basis of a growing number of neuromuscular and neurodegenerative diseases. Repetitive stretches in the human genome may drive pathogenicity, either by expansion above a given threshold, or by insertion of abnormal tracts in nonpathogenic polymorphic repetitive regions, as is the case in spinocerebellar ataxia type 37 (SCA37). We have recently established that this neurodegenerative disease is caused by an $(ATTTC)_n$ insertion within an $(ATTTT)_n$ in a noncoding region of *DAB1*. We now investigated the mutational mechanism that originated the $(ATTTC)_n$ insertion within an ancestral $(ATTTT)_n$. Approximately 3% of nonpathogenic $(ATTTT)_n$ alleles are interspersed by AT-rich motifs, contrarily to mutant alleles that are composed of pure $(ATTTT)_n$ and $(ATTTC)_n$ stretches. Haplotype studies in unaffected chromosomes suggested that the primary mutational mechanism, leading to the $(ATTTC)_n$ insertion, was likely one or more T>C substitutions in an $(ATTTT)_n$ pure allele of approximately 200 repeats. Then, the $(ATTTC)_n$ expanded in size, originating a deleterious allele in *DAB1* that leads to SCA37. This is likely the mutational mechanism in three similar $(TTTCA)_n$ insertions responsible for familial myoclonic epilepsy. Because $(ATTTT)_n$ tracts are frequent in the human genome, many loci could be at risk for this mutational process.

INTRODUCTION

Spinocerebellar ataxia type 37 (SCA37; MIM# 615945) is an autosomal dominant neurodegenerative disease, characterized by onset of dysarthria, followed by progressive gait and limb ataxia (Seixas et al., 2017; Serrano-Munuera et al., 2013). We have recently reported that SCA37 is caused by an $(ATTTC)_n$ insertion in the middle of a nonpathological ATTTT repeat, located in a 5′UTR intron of *DAB1*.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Unaffected subjects carry $(ATTTT)_n$ alleles ranging from seven to 400 units, while affected individuals showed an allele with the configuration $(ATTTT)_{60-79}$ $(ATTTC)_{31-75}$ $(ATTTT)_{58-90}$ (Seixas et al., 2017). We have previously shown that pathogenic SCA37 alleles present intergenerational instability toward expansion; the $(ATTTC)_n$ increases in size in 81% of parent–offspring transmissions, whereas repeat insertion contractions have never been detected in SCA37 pedigrees (Seixas et al., 2017). In pathogenic alleles, the flanking 5'-ATTTT repeat is also unstable during transmission; the 3'-ATTTT sequencing is incomplete, but for two branches with full sequencing the ATTTT length shows instability upon transmission. In SCA37 pedigrees, the large non- pathogenic $(ATTTT)_n$ alleles have also shown instability upon transmission; in one family branch, two siblings inherited from their unaffected parent nonpathogenic alleles with different sizes (Seixas et al., 2017).

Over the last three decades, expansions of microsatellites with tri-, tetra-, penta-, and hexanucleotide repeats have been identified as the cause of more than 25 neurological or neuromuscular diseases (Loureiro, Oliveira, & Silveira, 2016). More recently, however, knowledge has been gathered regarding a cryptic subclass of dynamic mutations consisting of an insertion of a repetitive motif adjacent or within a nonpathological polymorphic repeat in noncoding regions. In this type of dynamic mutation, the complex repeat without the insertion is not pathological (Seixas et al., 2017). In addition to SCA37, this type of mutation underlies SCA31 (MIM# 117210) and benign adult familial myoclonic epilepsy (BAFME1; MIM# 601068, 6 and 7) (Cen et al., 2018; Ishiura et al., 2018; Sato et al., 2009; Seixas et al., 2017).

Microsatellite repetitive tracts are often interrupted by other nucleotide motifs that result from nucleotide substitutions (Ananda et al., 2014). In some diseases, as SCA1 (MIM# 164400), SCA2 (MIM# 183090), fragile-X syndrome (MIM# 300624), myotonic dystrophy type 2 (DM2; MIM# 602668), and Friedrich ataxia (FRDA; MIM# 229300), nonpathogenic alleles are more likely to be interrupted than expanded alleles, whereas in SCA8 (MIM# 608768), SCA10 (MIM# 603516), and DM1 (MIM# 160900) interruptions are more expected in expanded alleles (Braida et al., 2010; Chung et al., 1993; Hu et al., 2017; Imbert et al., 1996; Landrian et al., 2017; Liquori et al., 2003; Maia et al., 2017; Martins et al., 2005; Menon et al., 2013; Montermini et al., 1997; Moseley et al., 2000; Musova et al., 2009; Ramos et al., 2010; Yrigollen et al., 2014; Zuhlke et al., 2002). Depending on the gene, interruptions in the repetitive

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

tract may be randomly positioned, as in *ATXN2*, *ATXN8OS*, and *ATXN10*, or follow a pattern (Chung et al., 1993; Landrian et al., 2017; Yrigollen et al., 2014). In *FMR1*, AGG interruptions are usually located every nine or 10 CGGs, whereas in *ATXN1* interrupted alleles have one to three CAT interruptions in the middle of the allele intercalated with one CAG, following a configuration from $(CAG)_n$ CAT $(CAG)_n$ to $(CAG)_n$ CAT CAG CAT CAG CAT $(CAG)_n$ (Chung et al., 1993; Yrigollen et al., 2014). Repeat interruptions in non-pathogenic alleles stabilize the repeat tract (Choudhry, Mukerji, Srivastava, Jain, & Brahmachari, 2001; Chung et al., 1993; Yrigollen et al., 2012); in expanded alleles at specific loci interruptions may decrease penetrance or delay age-of-onset, as in SCA1 and DM1 (Botta et al., 2017; Matsuyama, Izumi, Kameyama, Kawakami, & Nakamura, 1999; Menon et al., 2013; Zuhlke et al., 2002). In SCA10, however, interruptions in expanded alleles may turn them more deleterious, lowering age-of-onset, and associating with seizures (Matsuura et al., 2006).

The pentanucleotide repeat associated with SCA37 is located in the middle poly-A of an AluJb element (Seixas et al., 2017). Several other disease-causing repeats are in middle or 3'-end poly-A regions of Alu elements, such as the repeat expansions responsible for FRDA, DM2, and SCA10, as well as the pentanucleotide repeat insertions associated with SCA31 and two types of BAFME (Cen et al., 2018; Chauhan, Dash, Grover, Rajamani, & Mukerji, 2002; Clark et al., 2004; Ishiura et al., 2018; Kurosaki et al., 2012; Kurosaki, Matsuura, Ohno, & Ueda, 2009; Montermini et al., 1997; Sato et al., 2009).

Notably, loci containing repeat insertions are similar in that (1) they have an ATTTT/TAAAA repeat in the reference genome database, (2) shorter repeats with the insertion motif have not been detected in the general population, (3) nonpathogenic ATTTT/TAAAA alleles are highly polymorphic (ranging from a few to hundreds of repeats), and (4) pathogenic alleles have a complex configuration of both nonpathogenic repeat and insertion motifs. Such a significant number of loci with repeat insertions indicates that mutational events on the origin of these deleterious alleles are recurrent in the human genome.

To gain insight into the mutational mechanisms responsible for origin, evolution, and instability of these pathogenic repeat insertions, we focused on the SCA37 locus. Thus, we (1) characterized the complexity of the pentanucleotide repeat in

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

nonpathogenic and pathogenic alleles, (2) compared orthologous sequences along the primate lineage, and (3) analyzed haplotypes flanking the repeat in humans. Our results suggest that the mutational mechanism responsible for pathogenic repeat insertions involves ATTTT repeat lengthening and nucleotide substitution, providing a simple explanation for the birth of these dynamic mutations from ancestral alleles.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

MATERIALS AND METHODS

Subjects

We used DNA samples from 44 SCA37 affected individuals, 260 control subjects from the Portuguese general population, and 394 individuals with neurodegenerative diseases without the repeat insertion, as previously published (Seixas et al., 2017). Affected individuals were referred for diagnostic purposes to the authorized Center for Predictive and Preventive Genetics (CGPP) at the Institute for Molecular and Cell Biology. This study used the de-identified, previously collected DNA samples that were stored at the CGPP biobank, as well as anonymized DNA samples from control subjects from the Portuguese general population.

Assessment of short pentanucleotide repeat alleles

Repeat size of short alleles was assessed by standard PCR with Hot-StartTaq Master Mix Kit (QIAGEN), 0.4 μM of primers 24F and 24R (Seixas et al., 2017), and 30 ng of DNA, in a final volume of 12.5 μL. PCR was performed with an initial denaturation of 15 min at 95ºC, followed by 30 cycles of amplification (95ºC for 45 s, 64ºC for 30 s, and 72ºC for 40 s), and a final extension at 72ºC for 10 min. The allele structure (AS) was determined by Sanger sequencing of PCR products, with primers 24F and 24R (Seixas et al., 2017).

Identification of pattern of interruption in large alleles

To detect large alleles or confirm homoallelism (two alleles of the same size), in samples where only one allele was detected by standard PCR, we carried out RP-PCR with a specific primer to anneal with the $(ATTTT)_n$ tract, as previously described (Loureiro, Oliveira, Sequeiros, & Silveira, 2018). Samples with a "positive" ATTTT RP-PCR were amplified with long range-PCR (Seixas et al., 2017). PCR products were

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

separated by electrophoresis, in 1% agarose gel. DNA fragments were extracted from the gel, purified with QIAquick gel extraction kit (Qiagen) and sequencing analysis was performed with the internal primers 24F or 24R4, as published (Loureiro et al., 2018; Seixas et al., 2017). For very large alleles, with over 200 repeat units or more than 1 kb, Sanger sequencing in both forward and reverse orientations did not completely overlap, but a partial overlap of the ATTTT repeat was safely achieved, and was in accordance with the estimated repeat size in agarose gel. Each allele structure found in one chromosome only was confirmed by PCR and sequencing replicates.

Haplotype analysis

Haplotypes were constructed with eight SNPs, spanning a region of 723 kb flanking the *DAB1* pentanucleotide repeat, including (1) rs1043184969 and rs954450605 upstream, and rs192485043, rs145097803, and rs929412570 downstream the repeat, as described before (Seixas et al., 2017); (2) and three additional SNPs, presenting higher minor allele frequencies and closer to the repetitive region, rs514412, rs2113453, and rs11207020; positions and physical distances are according to GRCh37/hg19. SNP regions were PCR amplified with GoTaq Green® Mastermix (Promega), 1.25 μM of primers F and R (Supporting Information Table S1) and 30 ng DNA. The initial denaturation at 95ºC for 3 min was followed by 35 cycles of amplification at 95ºC for 45 s, annealing temperature for 30 s (Supporting Information Table S1) and 72ºC for 40 s, and a final extension at 72ºC for 10 min. Allele discrimination was performed by RFLP or by Sanger sequencing (Supporting Information Table S1). Haplotype phases were assessed using PHASE v.2.1.1 (http://stephenslab.uchicago.edu/phase/download.html); only haplotypes with probability higher than 0.6 were included in further analyses. Phylogenetic relationships among unaffected short pure, large pure, and interrupted alleles were assessed by median-joining method, using the Network 5.0.0.3 software (www.fluxus-enginee ring.com).

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Pentanucleotide repeat evolution

Time of divergence between human and primate species was calculated using the node time searching data in the public knowledge-based TimeTree (Hedges, Dudley, & Kumar, 2006; Kumar, Stecher, Suleski, & Hedges, 2017). Data from nucleotide to pentanucleotide repeat evolution in *DAB1* was obtained from Vertebrate Multiz Alignment & Conservation, UCSC genome browser (Blanchette et al., 2004). Data included whole genome assemblies from shotgun sequencing on 454 GS FLX and GS FLX large read platforms from *Pan troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), *Pongo abelii* (orangutan), *Nomascus leucogenys* (gibbon), *Macaca fascicularis* (crab eating macaque), *Macaca mulatta* (rhesus), *Papio hamadryas* (baboon), *Chlorocebus sabaeus* (green monkey), *Callithrix jacchus* (marmoset), *Saimiri boliviensis* (squirrel monkey), and *Otolemur garnettii* (bushbaby); the reference genome for each primate specie includes a single allele.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

RESULTS

Unaffected alleles may be interrupted

To characterize the pentanucleotide repeat sequence in *DAB1*, we performed Sanger sequencing. We found that both unaffected $(ATTTT)_n$ and pathogenic SCA37 $(ATTTT)_n$ $(ATTTC)_n$ $(ATTTT)_n$ alleles were always preceded, at the 5'-end, by two ATTTs that were not polymorphic. From the 1,308 unaffected chromosomes studied, we assessed the purity or interrupted nature in 1,293 alleles; 1,249 (96.6%) were composed by pure repetitive stretches of $(ATTTT)_n$, one of which had only one ACTTT at the 3'-end of the repeat, while 44 alleles (3.4%) that ranged from 16 to about 400 repeat units had AT-rich interruptions (Table 1). The complete allele structure of interrupted alleles was only assessed for 30 alleles (Figure 1). SCA37 alleles had no interruptions in neither $(ATTTT)_n$, nor $(ATTTC)_n$ tracts (Figure 1).

In non-disease associated $(ATTTT)_n$ chromosomes, we detected seven types of interruption motifs in the 30 fully sequenced alleles. These interruptions included a single A or $A(T)_n$, varying from di- to tetra- or hexa- to octanucleotide (Figure 1); the most common was ATTT (46.6%), followed by AT (29.3%) (Figure 2); an ATTTTTTT was found only in one allele. The ATTTC pentanucleotide was not found interrupting nonpathogenic alleles. Regarding the position of the interruptions, we found nine different allele configurations; 17 alleles (56.7%) had an allele structure (AS) F; the G and J structures were detected in three alleles each (10%), and two alleles (6.7%) had the AS C (Figure 1). The remaining interrupted alleles did not follow any common interruption pattern.

Interruptions are associated with ATTTT repeat size

To investigate whether repeat interruptions in unaffected chromosomes were associated with specific allele sizes, we divided them in three subgroups based on repeat length: alleles with 30 ATTTTs or less, alleles ranging from 31 to 79 repeats and alleles with 80 or more repeats (Table 1). In 1,226 alleles with 30 repeats or less, only two, with 16 and 17 ATTTTs, had one interruption (0.2%); whereas in alleles of 31 to 79 ATTTTs, 27 had interruptions (69.2%); and in alleles with 80 repeats or more, 15 were interrupted (53.6%) (Table 1). The frequency of interrupted alleles

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

was higher in alleles ranging from 31 to 79 repeats than in alleles with 30 repeats or less (Fisher's exact test, $P < 0.0001$) and was also higher in alleles with 80 or more repeats than in alleles with 30 repeats or less (Fisher's exact test, $P < 0.0001$). The difference in frequency between interrupted alleles ranging from 31 to 79 and with 80 repeats or more was not statistically significant.

Regarding the number of allele interruptions (one to eight) in the fully sequenced alleles, they increased with repeat size: interrupted alleles with less than 30 repeats had only one, while larger alleles had from two to four or seven to eight interruptions for repeat sizes below or 80 units and above, respectively (Figures 1 and 3).

Two types of AT-rich interruptions were exclusively detected in alleles of 31–79 repeats (A and ATTTTT); and another three types of interruptions were only detected in alleles over 80 repeats (ATT, ATTTTTT, and ATTTTTTT; Figure 1).

Evolution of pentanucleotide alleles in primate lineage

In *DAB1*, the ATTTT repeat is in the middle poly-A region of an AluJb element encoded in the opposite strand (Seixas et al., 2017). To understand how the pentanucleotide emerged and evolved, we compared the orthologous region of primates by using reference primate databases (Figure 4). We found that the AluJb element was inserted in the *DAB1*-opposite strand before New World monkeys divergence at approximately 43.2 million years ago (mya). The Alu middle poly-A, in the *DAB1*-opposite strand, was composed of a pure stretch of adenines, as observed in New World monkeys. Before the Old World monkeys divergence, approximately 29.44 mya, the Alu middle poly-A increased its size and started mutating. In Old World monkeys, the $(ATTTT)_n$ in the *DAB1*-oriented strand is preceded by stretches of thymines and adenines and by an ATTT. The pure stretches of Ts and As mutated before the Hominoids divergence, approximately 20.19 mya. In Hominoids, the repeat is composed by one or two ATTTs, followed by a pure or interrupted $(ATTTT)_n$. We did not find ATTTC repeat motifs in any primate genome data available.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

The SCA37 haplotype is very rare in Portuguese unaffected chromosomes

SCA37 is associated with a rare haplotype shared by all affected individuals studied so far (Seixas et al., 2017). To investigate mutational mechanisms leading to the origin of the $(ATTTC)_n$ insertion, we constructed SNP haplotypes of nonpathogenic pure and interrupted $(ATTTT)_n$ alleles and compared them to the single SCA37 haplotype.

We found eight haplotypes to be associated with nonpathogenic alleles, identified from I to VIII (Table 2 and Supporting Information Table S2). Six (haplotypes I, II, III, V, VI, and VII) were associated with short pure alleles (<100 ATTTTs). Interrupted alleles shared three haplotypes (I, IV, and V; Figure 3), whereas large pure alleles (>100 ATTTTs) segregated with only two haplotypes (I and VIII). Haplotype I was shared by short pure, large pure, and interrupted alleles, while haplotype V was common to short pure and interrupted alleles. Haplotype VIII, the one associated with the $(ATTTC)_n$ insertion in SCA37, was seen only in two unaffected chromosomes, both carrying a pure allele with approximately 200 ATTTT repeats.

To explore the origin of the SCA37 mutation, we designed a haplotype network to visualize the genetic distance among short pure, large pure, and interrupted alleles (Figure 5). Short pure alleles were phylogenetically closer to each other, as interrupted alleles were; large pure alleles with haplotype VIII were genetically closer to interrupted alleles than to short pure repeats. As the SCA37 haplotype is very rare in the Portuguese population and genetically distant from the haplotypes found in nonpathogenic alleles, we hypothesized it could have been introduced from another population where it may have a higher frequency. Therefore, we analysed data from phase 3 database of The 1000 Genomes Project (The 1000 Genomes Project Consortium 2015); although several SNP alleles from the SCA37 haplotype have been observed in some persons of African, European, and Asian ancestry, the full SCA37-associated haplotype was not detected in those populations (Supporting Information Table S3).

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

DISCUSSION

This study provides a plausible mutational mechanism for the origin of the ATTTC repeat insertion within a nonpathological $(ATTTT)_n$, in SCA37. Haplotype analysis allowed us to identify a class of large nonpathogenic alleles at *DAB1* characterized by (1) the acquisition of A or $A(T)_n$ repeat interruptions and (2) repeat size instability. This analysis showed a haplotype shared among nonpathogenic short pure, interrupted and large pure alleles, indicating that both nucleotide changes and repeat size instability are likely in the origin of the polymorphic pentanucleotide repeat at *DAB1*. The haplotype network (Figure 5) shows a clear distance among short pure stable, large, and interrupted unstable alleles, suggesting that genetic variants flanking the mutant allele act as cis elements influencing repeat instability. The evolution of the repetitive region in primates (Figure 4) further supports this observation. Haplotype studies in nonpathogenic and pathogenic chromosomes suggest that the mutational mechanism towards the acquisition of the deleterious ATTTC repeat insertion in SCA37 was one or more T>C substitutions in a large pure $(ATTTT)_n$ allele.

Haplotype VIII, associated with pathogenic SCA37 and nonpathogenic large pure alleles, is genetically very distant from other haplotypes found in the Portuguese population, indicating that the disease-associated haplotype could have been introduced from other populations. Since we did not find this haplotype in The 1000 Genomes Project database, it may be very rare worldwide or present in populations with no genetic data available. Indeed, the Iberian Peninsula, where the SCA37 families have been identified so far (Corral-Juan et al., 2018; Seixas et al., 2017), has been inhabited by people originated from populations in Asia and North Africa not completely assessed yet by The 1000 Genomes Project. Haplotype VIII could either be predisposed for both nucleotide and repeat size instability, thus a risk haplotype for the generation of de novo $(ATTTC)_n$ insertions or, alternatively, the common ancestral haplotype of SCA37 alleles.

The identification of haplotype VIII in large pure alleles of Portuguese ancestry, strongly suggests that the $(ATTTC)_n$ insertion originated in a large pure $(ATTTT)_n$ allele. It is unlikely that these large pure $(ATTTT)_n$ alleles with haplotype VIII have been originated by an $(ATTTC)_n$ deletion, because (1) ATTTC repeat contractions

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

have never been detected in 22 parent-to-offspring transmissions (Corral-Juan et al., 2018; Seixas et al., 2017) and (2) a complete $(ATTTC)_n$ deletion would result in an allele with approximately 120 repeats, instead of the 200 repeats found in the two unrelated unaffected individuals with haplotype VIII. Thus, our hypothesis for the pathogenic allele evolution includes, first, an $(ATTTT)_n$ increase and, then, the birth of the ATTTC, likely by one or more T>C substitutions in the middle of the ancestral repeat, similar to the T>A changes leading to birth of the *DAB1* ATTTT repeat during primate evolution; then the ATTTC repeat became unstable.

Interestingly, this mutational mechanism has previously been suggested in the origin of pure repetitive stretches of (CCG)n in the middle of the expanded (CTG)n, in DM1. The CCG interruptions may appear individually, in runs of CCGCTG or in a continuously block (Braida et al., 2010; Cumming et al., 2018; Pešović et al., 2017). The CCG blocks may be as large as 40 repeat units (Pešović et al., 2017). It was suggested that CCG blocks originate, first, by base substitution (T>C) followed by small changes in length, which later results in longer runs of CCGs (Braida et al., 2010; Cumming et al., 2018).

It is puzzling why small ATTTC repeats were not detected in the unaffected Portuguese population; however, this may be explained by the fact that haplotype VIII is very rare; small ATTTC repeats may come to be identified in a larger sample or in other populations with a higher frequency of this haplotype. Another hypothesis to explain why alleles containing a small number of ATTTC repeats are missing would be the insertion of the ATTTC repeat stretch in *DAB1* from another genomic region, as happens, for example, in the generation of pseudogenes (Ewing et al., 2013). In this scenario, an RNA containing the repeat would be reverse transcribed and then inserted in the middle of the *DAB1* ATTTT repeat. However, when we observed the SCA37 pathogenic alleles, we found that the repeat insertion region only contains the ATTTC repeat track, and it is not flanked by DNA sequences suggestive of a retrotransposition mechanism. This argument, in the addition to lack of literature regarding retrotransposition of microsatellites alone, gives strength to the hypothesis that the ATTTC repeat insertion was originated by T>C nucleotide substitution followed by size increase.

The pure (ATTTT)n in *DAB1*-oriented strand (at the SCA37 locus) emerged after several A>T substitutions in the Alu poly-A (Alu orientation), which originated the

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

AT-rich stretches identified in various species from the primate lineage. Then, the pure tracts of adenines and thymines seem to have disappeared by the occurrence of more A>T or T>A substitutions or indels, likely arisen by DNA polymerase errors or slippage, followed by an increase in the number of pentanucleotides, originating the pure pentanucleotide alleles found in chimpanzee and humans. Thus, the primate genomes show that the evolution of the pure $(ATTTT)_n$ resulted from numerous mutational events in pure poly-A, to finally originate the pure highly unstable pentanucleotide repeat found nowadays that later suffered T>C transition(s), which increased in size and became the SCA37 causing variant.

Different cellular mechanisms have been associated with gain and loss of genetic material in repeat loci (Gomes-Pereira et al., 2014; Martins et al., 2014; Santos, Pimenta, Wong, Amorim, & Martins, 2014; Slean et al., 2016). The repetitive nature of these genomic regions burden normal cell mechanisms such as replication and transcription, leading to an increase in DNA breaks in repeats (Krasil- nikova & Mirkin, 2004; Zhang et al., 2012). The mechanisms responsible for DNA break-repair have been largely associated with repeat size instability (Axford et al., 2013; Gomes-Pereira et al., 2014; Lee et al., 2010; Polleys, House, & Freudenreich, 2017) and may have had a role in the highly polymorphic nature of this pentanucleotide repeat.

The ultimate mutational mechanism leading to an $(ATTTC)_n$ insertion seems to be recurrent in the human genome. Similar $(TTTCA)_n$ mutations have been identified in three types of BAFME (Cen et al., 2018; Ishiura et al., 2018). The primary mutational feature, common to these diseases, is likely the T>C substitution(s) in the last thymine of the ATTTT or TTTTA repeats, followed by size increase of the $(ATTTC)_n$ or $(TTTCA)_n$, as a result of its unstable nature. This new type of mutational mechanism should be taken into consideration when searching new disease-causing variants.

In conclusion, the repeat region associated with SCA37 is highly polymorphic, mutable, and unstable, and the AluJb seems to have had a role in these features, as the $(ATTTC)_n$ insertion was probably originated by one or more T>C substitutions in the Alu element. Further- more, this type of events seems to be recurrent in Alu ATTTT motifs in the human genome, suggesting that other AT-rich repeats can be implicated in other brain diseases.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Acknowledgements

We are grateful to the families and individuals who participated in this work. We thank Patricia Ribeiro for technical assistance. This study was financed by Fundo Europeu de Desenvolvimento Regional (FEDER), through the COMPETE 2020 Operational Pro- gram for Competitiveness and Internationalization (POCI) of Portugal 2020, and by the Fundacão para a Ciência e a Tecnologia (FCT) and Ministério da Ciência, Tecnologia e Ensino Superior (Portugal), in the framework of the project POCI-01-0145-FEDER-029255; (PTDC/MED-GEN/29255/2017) to I.S. J.R.L. and C.L.O. were sup- ported by scholarships from PEst-C/SAU/LA0002/2013. S.M. is funded by the project IF/00930/2013/ CP1184/CT0002 from FCT. This work was also funded by the Porto Neurosciences and Neurologic Disease Research Initiative at the Instituto de Investigação e Inovação em Saúde (Norte-01-0145-FEDER-000008), supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTU- GAL 2020 Partnership Agreement with FEDER.

REFERENCES

Ananda, G., Hile, S. E., Breski, A., Wang, Y., Kelkar, Y., Makova, K. D., & Eckert, K. A. (2014). Microsatellite interruptions stabilize primate genomes and exist as population-specific single nucleotide polymor- phisms within individual human genomes. PLoS Genetics, 10, e1004498. https://doi.org/10.1371/journal.pgen.1004498

Axford, M. M., Wang, Y. H., Nakamori, M., Zannis-Hadjopoulos, M., Thorn- ton, C. A., & Pearson, C. E. (2013). Detection of slipped-DNAs at the trinucleotide repeats of the myotonic dystrophy type I disease locus in patient tissues. PLoS Genetics, 9, e1003866. https://doi.org/10.1371/journal.pgen.1003866

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., ... Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research, 14, 708–715. https:// doi.org/10.1101/gr.1933104

Botta, A., Rossi, G., Marcaurelio, M., Fontana, L., D'Apice, M. R., Brancati, F., ... Novelli, G. (2017). Identification and characterization of 5' CCG inter- ruptions in complex DMPK expanded alleles. European Journal of Human Genetics, 25, 257–261. https://doi.org/10.1038/ejhg.2016.148

Braida, C., Stefanatos, R. K. A., Adam, B., Mahajan, N., Smeets, H. J. M., Niel, F., ... Monckton, D. G. (2010). Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dys- trophy type 1 patients. Human Molecular Genetics, 19, 1399–1412. https://doi.org/10.1093/hmg/ddq015

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Cen, Z., Jiang, Z., Chen, Y., Zheng, X., Xie, F., Yang, X., ... Luo, W. (2018). Intronic pentanucleotide TTTCA repeat insertion in the SAMD12 gene causes familial cortical myoclonic tremor with epilepsy type 1. Brain, 141, 2280–2288. https://doi.org/10.1093/brain/awy160

Chauhan, C., Dash, D., Grover, D., Rajamani, J., & Mukerji, M. (2002). Origin and instability of GAA repeats: Insights from Alu ele- ments. Journal of Biomolecular Structure & Dynamics, 20, 253–263. https://doi.org/10.1080/07391102.2002.10506841

Choudhry, S., Mukerji, M., Srivastava, A. K., Jain, S., & Brahmachari, S. K. (2001). CAG repeat instability at SCA2 locus: Anchoring CAA inter- ruptions and linked single nucleotide polymorphisms. Human Molecular Genetics, 10, 2437–2446.

Chung, M.Y., Ranum, L. P., Duvick, L. A., Servadio, A., Zoghbi, H. Y., & Orr, H. T. (1993). Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nature Genetics, 5, 254–258.

Clark, R. M., Dalgliesh, G. L., Endres, D., Gomez, M., Taylor, J., & Bidichan- dani, S. I. (2004). Expansion of GAA triplet repeats in the human genome: Unique origin of the FRDA mutation at the center of an Alu. Genomics, 83, 373–383. https://doi.org/10.1016/j.ygeno.2003.09.001

Corral-Juan, M., Serrano-Munuera, C., Rábano, A., Cota-González, D., Segarra-Roca, A., Ispierto, L., ... Matilla-Dueñas, A. (2018). Clini- cal, genetic and neuropathological characterization of spinocerebel- lar ataxia type 37. Brain, 147, 1981–1997. https://doi.org/10.1093/brain/awy137

Cumming, S. A., Hamilton, M. J., Robb, Y., Gregory, H., McWilliam, C., Cooper, A., ... Monckton, D. G. (2018). De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. European Journal of Human Genetics, 26, 1635–1647. https://doi.org/10.1038/s41431-018 -0156-9

Ewing, A. D., Ballinger, T. J., Earl, D., Harris, C. C., Ding, L., & Wilson, R. K. ... Platform.(2013) Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biology, 14, R22. https://doi.org/10.1186/gb-2013-14-3-r22

Gomes-Pereira, M., Hilley, J. D., Morales, F., Adam, B., James, H. E., & Monckton, D. G. (2014). Disease-associated CAG·CTG triplet repeats expand rapidly in non-dividing mouse cells, but cell cycle arrest is insufficient to drive expansion. Nucleic Acids Research, 42, 7047–7056. https://doi.org/10.1093/nar/gku285

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: A public knowledge- base of divergence times among organisms. Bioinformatics, 22, 2971– 2972. https://doi.org/10.1093/bioinformatics/btl505

Hu, Y., Hashimoto, Y., Ishii, T., Rayle, M., Soga, K., Sato, N., ... Yokota, T. (2017). Sequence configuration of spinocerebellar ataxia type 8 repeat expansions in a Japanese cohort of 797 ataxia subjects. Journal of the Neurological Sciences, 382, 87–90. https://doi.org/10.1016/j.jns.2017.08.3256

Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J. M., ... Brice, A. (1996). Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. Nature Genetics, 14, 285–291. https://doi.org/10.1038/ng1196-285

Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M. K., Fujiyama, A., ... Tsuji, S. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nature Genetics, 50, 581–590. https://doi.org/10.1038/s41588-018-0067-2

Krasilnikova, M. M., & Mirkin, S. M. (2004). Replication stalling at Friedre- ich's ataxia (GAA)(n) repeats in vivo. Molecular and Celularl Biology, 24, 2286–2295. https://doi.org/10.1128/MCB.24.6.2286-2295.2004

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). Time- Tree: A resource for timelines, timetrees, and divergence times. Molecular Biology and Evolution, 34, 1812–1819. https://doi.org/ 10.1093/molbev/msx116

Kurosaki, T., Matsuura, T., Ohno, K., & Ueda, S. (2009). Alu-mediated acqui- sition of unstable ATTCT pentanucleotide repeats in the human ATXN10 gene. Molecular Biology and Evolution, 26, 2573–2579. https://doi.org/10.1093/molbev/msp172

Kurosaki, T., Ueda, S., Ishida, T., Abe, K., Ohno, K., & Matsuura, T. (2012). The unstable CCTG repeat responsible for myotonic dystrophy type 2 orig- inates from an AluSx element insertion into an early primate genome. PLoS One, 7, e38379. https://doi.org/10.1371/journal.pone.0038379

Landrian, I., McFarland, K. N., Liu, J., Mulligan, C. J., Rasmussen, A., & Ashizawa, T. (2017). Inheritance patterns of ATCCT repeat interrup- tions in spinocerebellar ataxia type 10 (SCA10) expansions. PLoS One, 12, e0175958. https://doi.org/10.1371/journal.pone.0175958

Lee, J. M., Zhang, J., Su, A. I., Walker, J. R., Wiltshire, T., Kang, K., ... Wheeler, V. C. (2010). A novel approach to investigate tissue- specific trinucleotide repeat instability. BMC Systems Biology, 4, 29. https://doi.org/10.1186/1752-0509-4-29

Liquori, C. L., Ikeda, Y., Weatherspoon, M., Ricker, K., Schoser, B. G. H., Dal- ton, J. C., ... Ranum, L. P. W. (2003). Myotonic dystrophy type 2: Human founder haplotype and evolutionary conservation of the repeat tract. The American Journal of Human Genetics, 73, 849–862.

Loureiro, J. R., Oliveira, C. L., Sequeiros, J., & Silveira, I. (2018). A repeat- primed PCR assay for pentanucleotide repeat alleles in spinocerebel- lar ataxia type 37. Journal of Human Genetics, In press. https://doi.org/ 10.1038/s10038-018-0474-3

Loureiro, J. R., Oliveira, C. L., & Silveira, I. (2016). Unstable repeat expansions in neurodegenerative diseases: Nucleocytoplas- mic transport emerges on the scene. Neurobiology of Aging, 39, 174–183. https://doi.org/10.1016/j.neurobiolaging.2015.12.007

Maia, N., Loureiro, J. R., Oliveira, B., Marques, I., Santos, R., Jorge, P., & Mar- tins, S. (2017). Contraction of fully expanded FMR1 alleles to the normal range: Predisposing haplotype or rare events? Journal of Human Genet- ics, 62, 269–275. https://doi.org/10.1038/jhg.2016.122

Martins, S., Pearson, C. E., Coutinho, P., Provost, S., Amorim, A., Dube, M. P., ... Rouleau, G. A. (2014). Modifiers of (CAG)(n) instability in Machado– Joseph disease (MJD/SCA3) transmissions: An association study with DNA replication, repair and recombination genes. Human Genetics, 133, 1311–1318. https://doi.org/10.1007/s00439-014-1467-8

Martins, S., Seixas, A. I., Magalhaes, P., Coutinho, P., Sequeiros, J., & Silveira, I. (2005). Haplotype diversity and somatic instability in normal and expanded SCA8 alleles. The American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 139b, 109–114. https://doi.org/10.1002/ajmg.b.30235

Matsuura, T., Fang, P., Pearson, C. E., Jayakar, P., Ashizawa, T., Roa, B. B., & Nelson, D. L. (2006). Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: Repeat purity as a disease modifier? America Journal of Human Genetics, 78, 125–129.

Matsuyama, Z., Izumi, Y., Kameyama, M., Kawakami, H., & Nakamura, S. (1999). The effect of CAT trinucleotide interruptions on the age at onset of spinocerebellar ataxia type 1 (SCA1). Journal of Medical Genetics, 36, 546–548.

Menon, R. P., Nethisinghe, S., Faggiano, S., Vannocci, T., Rezaei, H., Pemble, S., ... Giunti, P. (2013). The role of interruptions in polyQ in the pathol- ogy of SCA1. PLoS Genetics, 9, e1003648. https://doi.org/10.1371/ journal.pgen.1003648

Montermini, L., Andermann, E., Labuda, M., Richter, A., Pandolfo, M., Caval- canti, F., ... Cocozza, S. (1997). The Friedreich ataxia GAA triplet repeat: Premutation and normal alleles. Human Molecular Genetics, 6, 1261– 1266.

Moseley, M. L., Schut, L. J., Bird, T. D., Koob, M. D., Day, J. W., & Ranum, L. P. W. (2000). SCA8 CTG repeat: En masse contractions in sperm and intergenerational sequence changes may play a role in reduced penetrance. Human Molecular Genetics, 9, 2125–2130. https://doi.org/10.1093/hmg/9.14.2125

Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., ... Sedlacek, Z. (2009). Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. The American Jour- nal of Medical Genetics, Part A, 149a, 1365–1374. https://doi.org/10.1002/ajmg.a.32987

Pešović, J., Perić, S., Brkušanin, M., Brajušković, G., Rakocˇević-Stojanović, V., & Savić-Pavićević, D. (2017). Molecular genetic and clinical char- acterization of myotonic dystrophy type 1 patients carrying vari- ant repeats within DMPK expansions. Neurogenetics, 18, 207–218. https://doi.org/10.1007/s10048-017-0523-7

Polleys, E. J., House, N. C. M., & Freudenreich, C. H. (2017). Role of recombi- nation and replication fork restart in repeat instability. DNA Repair, 56, 156–165. https://doi.org/10.1016/j.dnarep.2017.06.018

Ramos, E. M., Martins, S., Alonso, I., Emmel, V. E., Saraiva-Pereira, M. L., Jardim, L. B., ... Silveira, I. (2010). Common origin of pure and inter- rupted repeat expansions in spinocerebellar ataxia type 2 (SCA2). The American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 153, 524–531. https://doi.org/10.1002/ajmg.b.31013

Santos, D., Pimenta, J., Wong, V. C., Amorim, A., & Martins, S. (2014). Diversity in the androgen receptor CAG repeat has been shaped by a multistep mutational mechanism. The American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 165b, 581–586. https://doi.org/10.1002/ajmg.b.32261

Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T., ... Mizusawa, H. (2009). Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. The American Journal of Human Genetics, 85, 544–557. https://doi.org/10.1016/j.ajhg.2009.09.019

Seixas, A. I., Loureiro, J. R., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C. L., ... Silveira, I. (2017). A pentanucleotide ATTTC repeat insertion in the non-coding region of *DAB1*, mapping to SCA37, causes spinocerebellar ataxia. The American Journal of Human Genetics, 101, 87– 103. https://doi.org/10.1016/j.ajhg.2017.06.007

Serrano-Munuera, C., Corral-Juan, M., Stevanin, G., San Nicolas, H., Roig, C., Corral, J., ... Matilla-Duenas, A. (2013). New subtype of spinocere- bellar ataxia with altered vertical eye movements mapping to chro- mosome 1p32. JAMA Neurology, 70, 764–771. https://doi.org/10.1001/ jamaneurol.2013.2311

Slean, M. M., Panigrahi, G. B., Castel, A. L., Pearson, A. B., Tomkin- son, A. E., & Pearson, C. E. (2016). Absence of MutSbeta leads to the formation of slipped-DNA for CTG/CAG contractions at primate replication forks. DNA Repair (Amst), 42, 107–118. https://doi.org/10.1016/j.dnarep.2016.04.002

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526, 68. https://doi.org/ 10.1038/nature15393

Yrigollen, C. M., Durbin-Johnson, B., Gane, L., Nelson, D. L., Hagerman, R., Hagerman, P. J., & Tassone, F. (2012). AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with frag- ile X syndrome. Genetics in Medicine, 14, 729–736. https://doi.org/10.1038/gim.2012.34

Yrigollen, C. M., Sweha, S., Durbin-Johnson, B., Zhou, L., Berry-Kravis, E., Fernandez-Carvajal, I., ... Tassone, F. (2014). Distribution of AGG inter- ruption patterns within nine world

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

populations. Intractable & Rare Dis- eases Research, 3, 153–161. https://doi.org/10.5582/irdr.2014.01028

Zhang, Y., Shishkin, A. A., Nishida, Y., Marcinkowski-Desmond, D., Saini, N., Volkov, Kirill V., ... Lobachev, Kirill S. (2012). Genome-wide screen identifies pathways that govern GAA/TTC repeat fragility and expan- sions in dividing and nondividing yeast cells. Molecular Cell, 48, 254–265. https://doi.org/10.1016/j.molcel.2012.08.002

Zuhlke, C., Dalski, A., Hellenbroich, Y., Bubel, S., Schwinger, E., & Burk, K. (2002). Spinocerebellar ataxia type 1 (SCA1): Phenotype-genotype correlation studies in intermediate alleles. European Journal of Human Genetics, 10, 204–209. https://doi.org/10.1038/sj.ejhg.5200788

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Figure 1

$(ATTTT)_n$ alleles

PI	Chromosomes (n)	Repeat size	Allele structure
A	1248	7-400	7-400
B	1	13	12
C	2	16-17	15-16
D	1	41	33
E	1	42	33
F	1	43	26
G	1	44	20 16
H	17	42-52	18-23 14-22
I	3	45-48	19-21 17-21
J	1	50	41
K	1	86	15 15 11 35
L	1	90	13 16 9 42
M	1	88	14 17 8 39
N	1	88	14 15 15 37
O	3	88-95	12-13 17-18 7-8 40-46
P	1	120	13 18 78

SCA37 alleles

SCA37	44	~190-220	60-79 31-75 58-90

Legend

A AT ATT ATTT ATTTT ATTTTT ATTTTTT ATTTTTTT ACTTT

Pure ATTTC stretch Pure ATTTT stretch

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Figure 1. Schematic representation of the structure of nonpathogenic and pathogenic (SCA37) alleles with nucleotide variations. Allele size is the number of pentanucleotide repeats in each allele and is represented by white boxes. The seven types of interruption motifs, shown by colored boxes, and the $(ATTT)_2$ preceding the pentanucleotide repeat, indicated by two grey boxes, were not considered for repeat size assessment. The T>C substitution in pattern B is ss2137543903. The patterns of interruptions are shown from C to K for the 30 fully sequenced alleles at the 3' -end from the total of 44 interrupted alleles; number of $(ATTTT)_n$ and $(ATTTC)_n$ in each stretch are indicated inside the respective box. AS, allele structure

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Figure 2

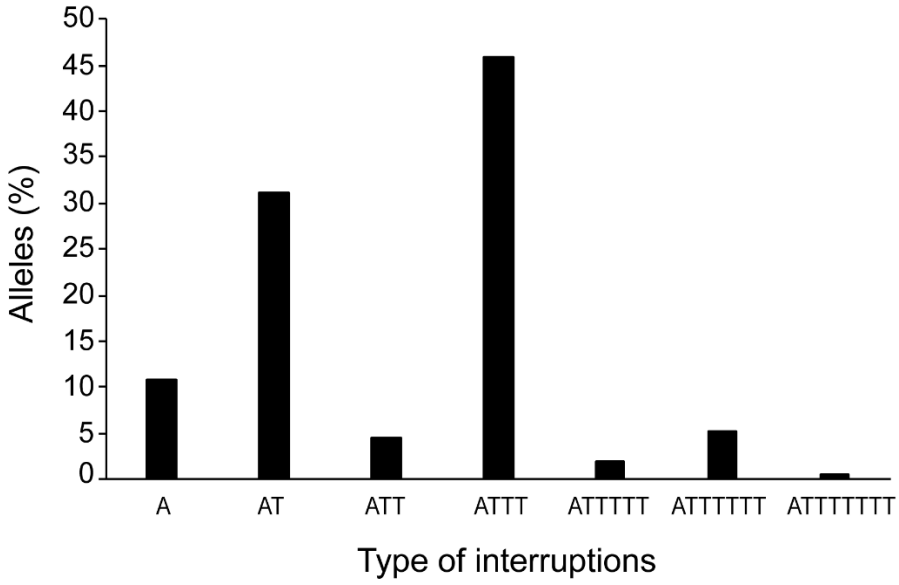

Figure 2. Frequency of each AT-rich interruption in the interrupted alleles fully sequenced (n=30).

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Figure 3

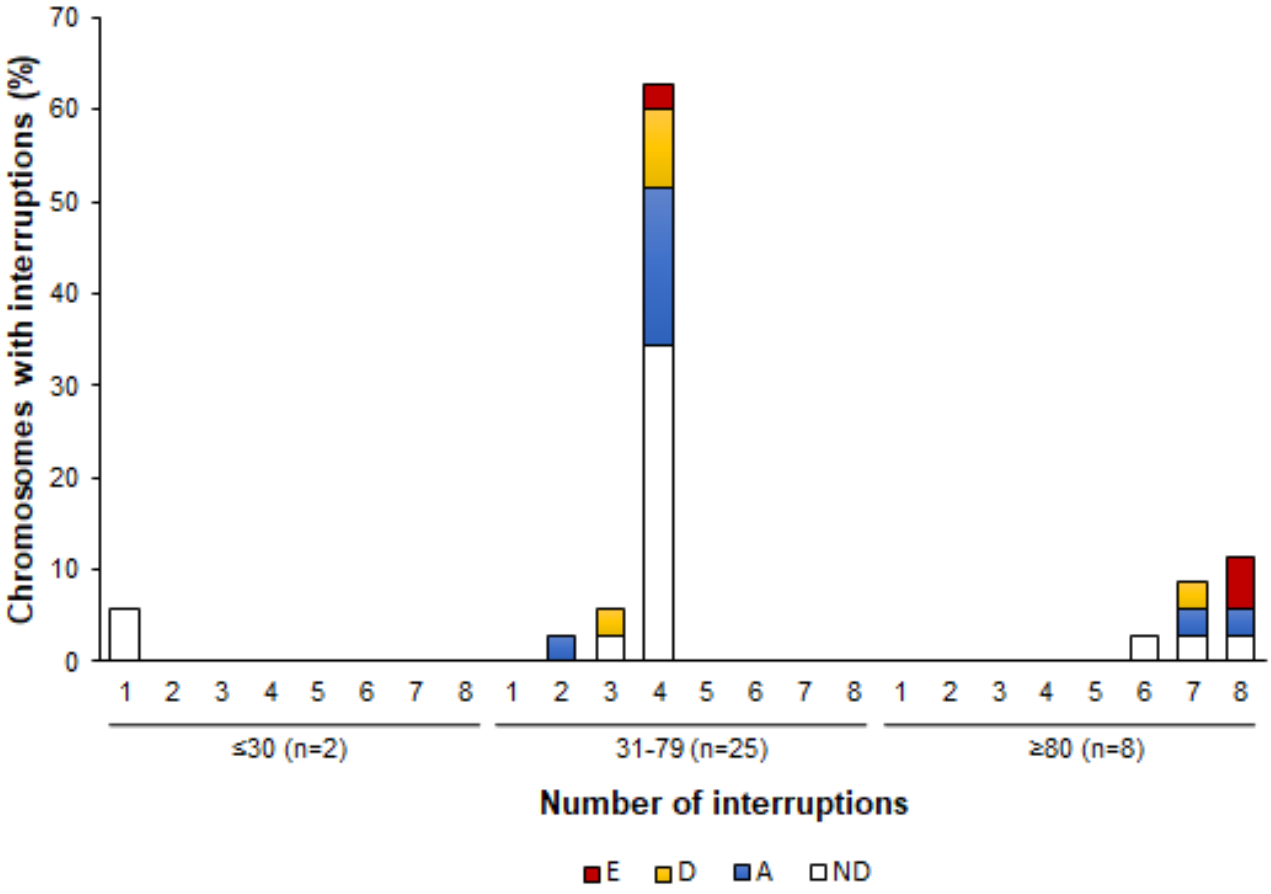

Figure 3. Number of interruptions associated with repeat size in the interrupted alleles fully sequenced (n = 30). Haplotypes associated with each type of repeat alleles are represented with different colors; n.d., not determined

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Figure 4

Figure 4. Evolution of the *DAB1* ATTTT repeat in primate lineages. AluJb retrotransposition event in *DAB1*-opposite strand is represented by a black box. For simplicity, nucleotide composition of Alu poly-A in primate species is represented in *DAB1*-oriented strand. The length of the phylogenetic tree branches represents the evolutionary distance among species; M, million years ago

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Figure 5

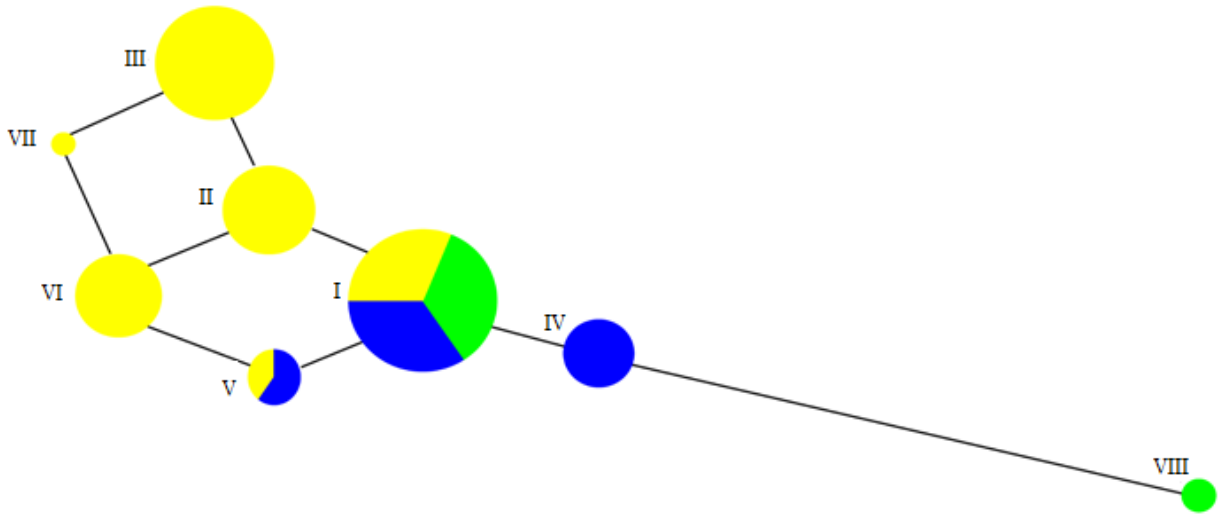

Figure 5. Close phylogenetic relationship among nonpathogenic *DAB1* alleles. Haplotype network showing the phylogenetic relationship among nonpathogenic (short pure, large pure and interrupted) alleles. Circle size is proportional to the number of chromosomes; line length is proportional to the genetic distance among haplotypes. Short pure alleles (<100 ATTTTs) are represented in yellow; large pure alleles in green (>100 ATTTTs) and interrupted alleles in blue.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Table 1. Frequency of pure and interrupted *DAB1* alleles in 1,308[a] unaffected chromosomes

Allele size	Allele type	Normal chromosomes	
		n	%
≤30	Pure	1,224	99.8%
	Interrupted	2	0.2%
	Total	1,226	
31-79	Pure	12	30.8%
	Interrupted	27	69.2%
	Total	39	
≥80	Pure	13	46.4%
	Interrupted	15	53.6%
	Total	28	
Total	Pure	1,249	96.6%
	Interrupted[b]	44	3.4%
	Total*	1,293	

[a]15 additional ATTTT alleles were studied, but could not be fully sequenced at the 3'-end for purity assessement.

[b]The allele structure was completely characterized at 3'-end in only 30 alleles.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt

Table 2. Haplotypes in pure and interrupted *DAB1* nonpathogenic alleles.

							Haplotypes											
SNP	Position chr1 (hg19)	Distance from $(ATTTT)_n$ (bp)	Ancestral allele	MAF All[a]	MAF EUR[a]	MAF PT[b]	$(ATTTC)_n$ insertion	Pure $(ATTTT)_{>100}$		Interrupted $(ATTTT)_n$			Pure $(ATTTT)_{<100}$					
rs1043184969	57,491,966	-340,750	T	*	*	G 0.0007	G	G	T	T	T	T	T	T	T	T	T	T
rs514412	57,551,569	-281,147	G	A 0.1949	A 0.1998	A 0.3530	G	G	G	G	G	A	G	G	G	A	A	A
rs954450605	57,551,605	-281,111	G	*	*	A 0.0013	A	A	G	G	G	G	G	G	G	G	G	G
$(ATTTT)_n$	57,832,716	0																
rs2113453	57,833,689	+973	C	C 0.4157	T 0.4750	C 0.4157	T	T	T	T	T	T	C	T	T	T	T	C
rs11207020	57,834,035	+1,319	C	T 0.2171	T 0.2843	T 0.2170	T	T	T	T	T	T	C	C	T	C	T	C
rs192485043	57,926,567	+93,851	T	A 0.0002	A 0.0010	A 0.0015	A	A	T	T	T	T	T	T	T	T	T	T
rs145097803	58,201,704	+368,988	T	A 0.0032	*	A 0.0046	A	A	T	T	T	T	T	T	T	T	T	T
rs929412570	58,215,160	+382,444	C	*	*	G 0.0007	G	G	C	C	G	C	C	C	C	C	C	C
Abs frequency							44	2	11	10	8	3	25	17	11	10	2	1
Haplotype ID							VIII	VIII	I	I	IV	V	III	II	I	VI	V	VII

[a] Minor allele frequency from 1000g phase 3; [b] Minor allele frequency in Portugal: Seixas et al., 2017 and present study; EUR, Europe; PT, Portugal.
*Not available or not found.

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal
+351 220 408 800
info@i3s.up.pt
www.i3s.up.pt