frontiers in Genetics

METHODS
published: 05 February 2019
doi: 10.3389/fgene.2019.00038

A Pipeline to Assess Disease-Associated Haplotypes in Repeat Expansion Disorders: The Example of MJD/SCA3 *Locus*

Inês P. D. Costa[1,2], Beatriz C. Almeida[1,2,3], Jorge Sequeiros[1,4,5], António Amorim[1,2,3] and Sandra Martins[1,2]*

[1] i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal, [2] IPATIMUP – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal, [3] Faculdade de Ciências da Universidade do Porto, Porto, Portugal, [4] IBMC – Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal, [5] ICBAS – Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal

OPEN ACCESS

Edited by:
Emmanuelle Genin,
INSERM UMR1078 Génétique,
Génomique Fonctionnelle et
Biotechnologies, France

Reviewed by:
Dusanka Savic Pavicevic,
University of Belgrade, Serbia
Hong Jiang,
Central South University, China

***Correspondence:**
Sandra Martins
smartins@ipatimup.pt

Specialty section:
This article was submitted to
Genetic Disorders,
a section of the journal
Frontiers in Genetics

Received: *30 October 2018*
Accepted: *18 January 2019*
Published: *05 February 2019*

Citation:
Costa IPD, Almeida BC, Sequeiros J, Amorim A and Martins S (2019) A Pipeline to Assess Disease-Associated Haplotypes in Repeat Expansion Disorders: The Example of MJD/SCA3 Locus. Front. Genet. 10:38. doi: 10.3389/fgene.2019.00038

At least 40 human diseases are associated with repeat expansions; yet, the mutational origin and instability mechanisms remain unknown for most of them. Previously, genetic epidemiology and predisposing backgrounds for the instability of some expanding *loci* have been studied in different populations through the analysis of diversity flanking the respective pathogenic repeats. Here, we aimed at developing a pipeline to assess disease-associated haplotypes at oligonucleotide repeat *loci*, combining analysis of single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs). Machado-Joseph disease (MJD/SCA3), the most frequent dominant ataxia worldwide, was used as an example of a detailed procedure. Thus, to identify genetic backgrounds that segregate with expanded/mutated alleles in MJD, we selected a set of 26 SNPs and 7 STRs flanking the causative CAG repeat. Key criteria and steps for this selection are described, and included (1) haplotype blocks minimizing the occurrence of recombination (for SNPs); and (2) match scores to increase potential for polymorphic information content of repetitive sequences found in Tandem Repeats Finder (for STRs). To directly assess SNP haplotypes in phase with MJD expansions, we optimized a strategy with preferential amplification of normal over expanded alleles, in addition to SNP allele-specific amplifications; this allowed the identification of disease-associated SNP haplotypes, even when only the proband is available in a given family. To infer STR haplotypes, we optimized a multiplex PCR, including 7 STRs plus the MJD_CAG repeat, followed by analysis of segregation or the use of the PHASE software. This protocol is a ready-to-use tool to assess MJD haplotypes in different populations. The pipeline designed can be used to assess disease-associated haplotypes in other repeat-expansion diseases. This should be of great utility to study (1) genetic epidemiology (population-of-origin, age and spreading routes of mutations) and (2) mechanisms responsible for *de novo* expansions, in these neurological diseases; (3) to detect predisposing haplotypes and (4) phenotype modifiers; (5) to help solving cases of apparent homoallelism (two same-size normal alleles) in diagnosis; and (6) to identify the best targets for the development of allele-specific therapies in ethnically diverse patient populations.

Keywords: haplotype, repeat instability, CAG expansion, mutation origin, Machado-Joseph disease, SCA3, SNP, STR

INTRODUCTION

Repetitive DNA sequences with a capacity to expand (sometimes up to hundreds or thousands of repeats) are found in the human genome in non-coding and exonic regions. They are currently known to be associated with approximately 40 human diseases (reviewed in Paulson, 2018). Trinucleotide repeats were the first to be identified, but an intensive search over the last decades has shown tetra-, penta- and hexanucleotide repeats also expanding above a normal polymorphic range in some human subjects (Paulson, 2018). Most of the repeat-associated disorders manifest with neurological, neuropsychiatric or neuromuscular symptoms. Among them are the spinocerebellar ataxias (SCA1, DRPLA, SCA2, MJD/SCA3, SCA6, SCA7, SCA8, SCA10, SCA12, SCA17, SCA31, and SCA36), Huntington's disease, myotonic dystrophy, spinal-bulbar muscular atrophy and fragile X syndrome. While each of these diseases is rare worldwide, together they are one of the commonest causes of hereditary neurological pathology (Sequeiros et al., 2011; Paulson, 2018).

Non-human primates have short repeat alleles at all these *loci* (Andrés et al., 2004), which implies that expansion into pathogenic ranges occurred after Homo-Pan split. If we are able to identify the genetic backgrounds where *de novo* expansions occurred (place-of-birth for these mutations), many new possibilities will open for the study of mutational origins and spread (Sequeiros et al., 2011; Obayashi et al., 2015; Lee et al., 2016; Bampi et al., 2017; Kay et al., 2017), as well as mechanisms of repeat instability (Martins et al., 2006; Falush, 2009; Warby et al., 2009; Ramos et al., 2010) and genetic modifiers for these diseases (Filippova et al., 2001; Libby et al., 2008; Becanovic et al., 2015).

To characterize the genetic background of expanded alleles, polymorphic markers are crucial given their capacity to distinguish alleles identical-by-state. Single nucleotide polymorphisms (SNPs) are the most common source of genetic variation in the human genome, but their biallelic nature reduces their informativeness. Short tandem repeats (STRs) have a very high polymorphism information content (PIC), but their mutation rate ranging 10^{-6}–10^{-2} *per* locus *per* generation (Fan and Chu, 2007) causes a high rate of recurrence. A combined stepwise analysis with both SNPs and STRs may then be the key to overcome both problems.

In Machado-Joseph disease (MJD/SCA3), SNPs and STRs have been used mainly to study genetic epidemiology in this dominant ataxia, the most frequent SCA worldwide (Martins and Sequeiros, 2018); however, there is also evidence on the importance of extending haplotype analyses to study MJD instability (Martins et al., 2008). MJD is caused by an expanded $(CAG)_n$ in exon 10 of the *ATXN3* gene (14q32.12) (Kawaguchi et al., 1994). As in other SCAs, repeat instability of mutated/expanded alleles has received enormous attention, due to its importance in the clinical phenomenon of anticipation: the earlier age-at-onset (AO) and more severe symptoms in successive generations, as the repeat number tends to increase upon transmission to offspring, and repeat size correlates inversely with AO. Despite its importance, instability is not yet fully understood in MJD or other repeat-associated disorders; only a few modifiers have been identified (McGinty and Mirkin, 2018). In MJD, in addition to the length of the initial repeat tract and the gender and age of the transmitting parent (Maciel et al., 1995; Maruyama et al., 1995; Souza et al., 2016), SNP rs12895357 near the CAG repeat has been shown to affect repeat instability, the genotype $(CAG)_{exp}$-C/$(CAG)_{normal}$-G of the transmitting parent being associated with increased instability (Igarashi et al., 1996; Maciel et al., 1999; Martins et al., 2008).

Given the importance of haplotype analyses (including SNPs and STRs) to perform a comprehensive study of oligonucleotide repeat-related diseases, we designed a strategy to identify disease-associated haplotypes and show here the example of this approach to analyse the *ATXN3* locus.

MATERIALS AND METHODS

Samples

We optimized a protocol with DNA samples extracted from saliva, buccal swab and peripheral blood through different techniques, by using the QIAamp® DNA Blood Mini kit, the Citogen®Blood kit, the Chelex100 chelating resin, as well as through the standard method of salting-out. We also tested some DNA samples stored for more than two decades at 4°C. This study was carried out with anonymized DNA samples available in our laboratory, in accordance with the recommendations of international guidelines. Previous written informed consent was obtained from all subjects to use their DNA samples for research purposes, in accordance with the Declaration of Helsinki. DNA quantification was performed with Nanodrop, to prepare aliquots with a final DNA concentration of 7.5 ng/μL.

Selection of Polymorphic Markers

We selected a set of 26 SNPs within a region of 4 kb encompassing the $(CAG)_n$, based on (1) minor allele frequencies (MAF > 5%; Ensembl[1]); (2) recombination hotspots (within the same haplotype block; The International Genome Sample

[1] www.ensembl.org

TABLE 1 | STRs selected for MJD haplotype analysis, respective distances from the $(CAG)_n$ and primers used for genotyping. F - Forward primer; R - Reverse primer.

Primers	Distance from the $(CAG)_n$	Primer sequence (5′–3′)
TAT223_F^{*1}	223 kb	CCACACTTCCTTTGGACCAT
TAT223_R		GGTAGGCACCAGCTACTTGGG
GT199_F	199 kb	TTACTGGGTAGGATATACATTCC
GT199_R^{*1}		CAGCCTTCCCCCGAGTCC
ATA194_F	194 kb	CCTTATCTAACCTCCTACATCTCAGC
ATA194_R^{*2}		GCAGGGCAGGCAATGAAACACG
MJD52_F		CCAGTGACTACTTTGATTCG
CAG_R^{*3}		GTGTGAAGGTAGCGAACATGATG
AC21_F^{*1}	21 kb	CTTCAGCTCAAATGCTATCAAAC
AC21_R		CAAGGATGGCTAGTGCAGAAAT
AAAC123_F	123 kb	CAGATGGGATAGGCCACAGT
AAAC123_R^{*1}		AGTGGAGGCTTCAACCTGTT
GT190_F^{*4}	190 kb	GAGGGGACCTGGCCTACTAC
GT190_R		ACCTACAGTAACACACTTTGCAC
AC190_F^{*2}	191 kb	CTGGGAGGAGGAGGGTACAA
AC190_R		AACCCTGACTCAACTCTCGG

Fluorescent labels ^{*1}FAM (6-carboxufluorescein); ^{*2}NED; ^{*3}PET; and ^{*4}VIC

TABLE 2 | Primers designed to amplify and sequence a 4 kb flanking region of the *ATXN3*_CAG repeat. F - Forward primer; R - Reverse primer.

Name	Primer sequence (5′–3′)
MJDcloF_F	CAATTATTGGCCTTTCTGAACC
MJD52_F	CCAGTGACTACTTTGATTCG
MJD653_R	GCAAATGAGTGTTGGTTTATAGACCC
MJD716_F	ACAGAGTCTCGCTCTGTCGCCCAG
MJD1260_R	GCTGTCTGAAACATTCAAAAGTGAAG
MJD7a_R	TGCTCCTTAATCCAGGGAAATTTAG
MJD1342_F	CCACCAGTTCAGGAGCACTT
MJD1396_F	TCATGTTCGCTACCTTCACACT
MJD2109_F	GAGTTACTTTCCAGGTCTCGG
MJD2129_R	CCGAGACCTGGAAAGTAACTC
MJD2552_F	GATCCAGCAGTCCCAATCATGTA
MJD2646_R	TGCCTGGTCAGCTATAAGCA
MJD2942_F	TGGACACGGTGGCTTACGCCT
MJD3417_F	CTGGGCTGGGTGGCGGTGGCTCA
MJD3936C_R	CTAAAGGTTTTTATCTTGCTAGAC
MJDcloR_R	AGCCTTCTCTAACACCACCTTGG

Resource and The 1000 Genomes Project); and (3) sequence alignments of previously analyzed patients (to include SNPs that discriminate MJD lineages (Martins et al., 2012; Ogun et al., 2015); **Supplementary Table S1**).

To select STRs, we used the Tandem Repeats Finder tool[2] and the following criteria for putative polymorphic repeats: consensus size of 2–5 bp; copy number above 7; matches above 80%; and proximity to the $(CAG)_n$. This search was done within the human reference *ATXN3* and 250 kb up and downstream to it (NG_008198.2). Seven STRs were selected, based on their score (>8.5): one tetra, two tri and four dinucleotide repeats, less than 223 kb from the $(CAG)_n$, thus reducing the likelihood of recombination (**Table 1**). Three of these STRs (TAT223, AC21 and GT190) have been analyzed in previous haplotype studies in several MJD populations, which will be useful for future comparisons (Martins et al., 2007).

Primer Design

We used the online software Primer3Plus to design specific primers, to amplify and sequence previous selected polymorphic markers, in addition to the MJD_CAG repetitive region (**Table 2**). Next, the alignment tool BLAST was used to guarantee the specificity of the designed primers: at maximum three mismatches (or two at 3′ end) were allowed if homologous sequences were observed. Finally, the occurrence of hairpins and primer-dimers (including self-dimers) was tested with AutoDimer (Vallone and Butler, 2004).

SNP Haplotypes

Genotyping of the SNPs was done with amplification of three fragments, not including the CAG tract, which allowed equal amplification of normal and expanded alleles. Reactions were performed using the following primers, annealing conditions, extension time, and number of cycles: MJDcloF-MJD1260R (62°C for 90 s; 90 s; 33 cycles); MJD1342F-MJD2646R (59°C for 90 s; 120 s; 35 cycles); MJD2552F-MJDcloR (61°C for 90 s; 120 s; 35 cycles). Amplification reactions were done in a total volume of 10 μL, with 0.2 μM of each primer, 1x of Taq PCR Master Mix Kit Qiagen®, 0.5x of Q-Solution for Qiagen®, and 15 ng DNA.

Phase of SNPs on expanded chromosomes was assessed through (1) amplification of two fragments encompassing the CAG repeat and either up or downstream regions (which resulted in the overrepresentation of normal alleles) and (2) allele-specific amplification of rs7142326 (amplicon length of 1384 bp). Reactions were performed using the following primers, annealing conditions, extension time and cycles: MJDcloF_F-MJD7a_R (60°C for 90 s; 120 s; 35 cycles), MJD52_F-MJD2646_R (58°C for 90 s; 120 s; 20 cycles; plus, 57°C for 90 s; 120 s; 20 cycles) and MJD1342_F-MJD3936C_R (57°C for 90 s; 120 s; 40 cycles). PCRs included 0.2 μM of each primer, 1x of Taq PCR Master Mix Kit Qiagen®, 0.5x of Q-Solution for Qiagen®and 15 ng DNA, in a final volume of 10 μL.

To sequence the amplified fragments, we started by purification with thermosensitive Alkaline phosphatase: Exonuclease I, ExoFastAP (Thermo Scientific) (1:5) at 37°C for 15 min, followed by 15 min at 80°C to inactivate the enzyme. Sequence reactions were done with 0.5 μL BigDye®Terminator Cycle kit (Applied Biosystems), according to manufacturer's instructions.

A final purification of DNA was performed using a cross-linked dextran matrix (Illustra™ Sephadex™ G-50, GE Healthcare), with centrifugation for 4 min at 4400 rpm. After loading the sequencing product, the same conditions of centrifugation resulted in a deposit containing the final product, ran in an ABI PRISM 3130x/Genetic Analyzer (Applied Biosystems) with Hi-Di™ formamide.

[2] https://tandem.bu.edu/trf/trf.html

STR Haplotypes

STRs were all amplified together with the CAG repeat, in a multiplex PCR reaction, in a final volume of 10 μL, using 0.25 μM (AAAC123, AC21, GT199, and GT190), 0.125 μM (TAT223, ATA194, and AC190), and 0.2 μM (CAG, and MJD52_F) of each primer, 1x of Taq PCR Master Mix Kit Qiagen®, 0.5x of Q-Solution for Qiagen®, and 7.5 ng DNA. The initial denaturation was performed at 95°C for 15 min, followed by 35 cycles of denaturation at 94°C for 30 s, annealing at 62°C for 90 s and extension at 72°C for 60 s; with a final extension at 70°C for 30 min. Analysis of fragment length was performed with ABI PRISM 3130x/Genetic Analyzer (Applied Biosystems). A mix of GeneScan^{TM} 500 LIZ^{TM} size standard (Thermo Scientific):Hi-Di^{TM} formamide (Applied Biosystems) (1:20) was added to 2 μL of PCR product and run in matrix G5, analyzed with GeneMapper v4.0. Allelic phases associated with the expansion were assessed by segregation and bioinformatically, using PHASE v.2.1.1 whenever DNA samples from relatives were not available.

PROTOCOL DESIGN AND APPLICATIONS

Analysis of the genetic background of expanded alleles at repetitive *loci* is important for a comprehensive study of repeat-associated neurological disorders. Most approaches, however, are rather *ad hoc* and lack a strategy to get the most from such valuable SNP and/or STR data from patients. Thus, we designed a pipeline to characterize the haplotype background of repetitive disease-associated *loci* that can be used to study any of these neurological disorders; a step-by-step protocol, optimized to assess genetic backgrounds of MJD patients, is here detailed as an example.

Process Overview

Our strategy focused first on the selection of SNPs to identify stable genetic backgrounds, defining lineages. Given the low mutation rate of SNPs (∼2.5×10^{-8}), events on the origin of these polymorphic markers are considered unique during the evolution of a species (Nachman and Crowell, 2000). It is important that SNPs are selected within a distance to the pathogenic repeat small enough to lie within a single haplotype block, thus avoiding recombination to play a relevant role on the lineages identified. Next, analysis of STRs allows differentiating a high number of haplotypes inside each lineage. Fast-evolving STRs flanking the disease *locus* should have a high potential to be pure, simple stretches, in order to be reliable as molecular clocks. By following this two-level strategy (STRs analyzed within stable SNP lineages), it is possible to achieve a high discrimination power, not compromising the discrimination of haplotypes identical-by-state (not by-descent) (**Figure 1**).

Protocol to Identify MJD Haplotype Backgrounds

We applied the strategy described to optimize a protocol to assess disease-associated haplotypes in MJD. While sequence, frequency and population data are widely available for SNPs in databases such as Ensembl and The 1000 Genomes Project, the search for STRs must rely on the potential PIC that a repetitive sequence harbors, since there are not many data available regarding allele frequencies and repeat configuration of non-deleterious STRs. For this reason, we selected potential STRs from Tandem Repeat Finder and tested their heterozygosity value by genotyping a set of random samples. After confirming their high potential to discriminate alleles identical-by-state, we sequenced at least one allele size *per* STR from two major ethnic groups: Europeans and Asians. Further analyses included exclusively pure STRs (or STRs with regular repeat configurations), without any additional source of size variation (such as indels or other tandem repeats) within the amplicon that includes the STR of interest. While optimizing the protocol, we performed three standard PCRs, to obtain SNP genotypes of MJD patients; followed by two PCRs encompassing the $(CAG)_n$ (**Figure 2A**) and an allele-specific PCR (**Figure 2B**), to assess alleles that segregated with the MJD expansion. This way, even in families with a single DNA sample available from the proband, we were able to infer directly alleles in *cis* with the expansion (i.e., lineages). For genotyping of STRs, we optimized a single multiplex reaction, to amplify all 7 STRs and the MJD_CAG repeat together (**Figure 3**), this way reducing quantity of DNA, time, reagents and sample's manipulation.

We validated our pipeline for MJD haplotyping by testing the optimized protocol in a randomly selected subset of 100 MJD families from our large cohort. Genotypes of SNPs and STRs were obtained in 92 families, a successful rate for SNP and STR amplification of 92% - taking into account the long-term storage of most DNA samples analyzed (most for over two decades), we considered 8% a low failure rate. We have also estimated success rate for haplotype inference, given the importance of assessing allelic phase of polymorphisms segregating with the pathogenic expansion: following our strategy of SNP allele-specific amplification, MJD lineages were accurately identified in all 92 families; as for flanking STRs, we were able to reconstruct haplotypes in 85% of the families genotyped (78/92); of note is the fact that 48% (44/92) of the families were composed solely by the proband and that in other 15% (14/92) only a single relative was available. Therefore, we may conclude that, following this protocol, complete extended haplotypes (SNPs-STRs) can be inferred in a larger number of MJD families, even in very small families or isolated patients.

Usefulness and Perspectives of Application

(1) **Genetic epidemiology.** Geographical differences in disease prevalence may be explained by haplotype studies, as shown for Huntington disease (HD), with the highest risk HD haplogroups being found in Europe, while absent in East Asia (Warby et al., 2011). In MJD two *de novo* expansions seem to have occurred, and, so far, its presence in remote and ethnically diverse populations has been explained

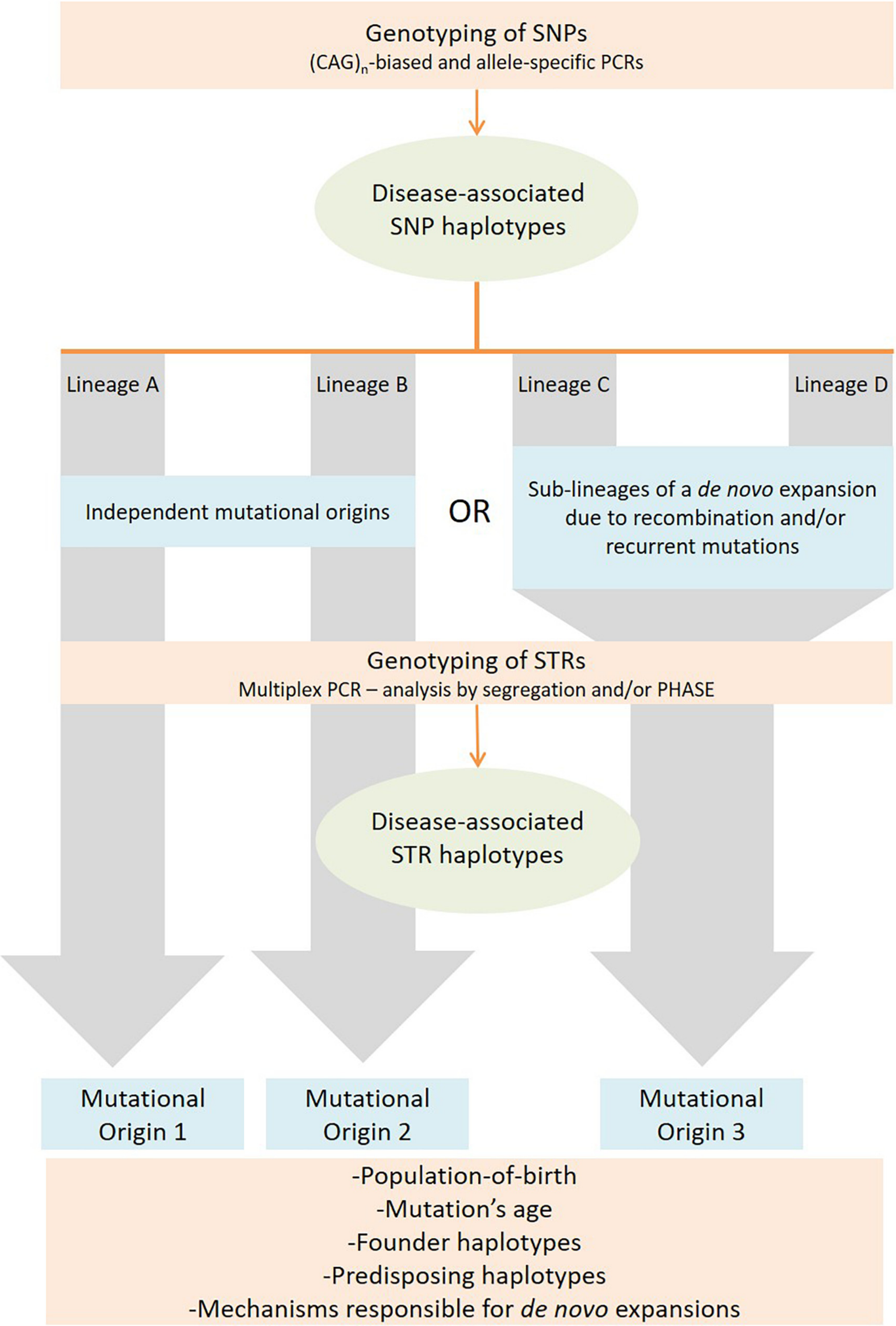
Genotyping of SNPs
(CAG)n-biased and allele-specific PCRs
Disease-associated SNP haplotypes
Lineage A
Lineage B
Lineage C
Lineage D
Independent mutational origins
OR
Sub-lineages of a *de novo* expansion due to recombination and/or recurrent mutations
Genotyping of STRs
Multiplex PCR – analysis by segregation and/or PHASE
Disease-associated STR haplotypes
Mutational Origin 1
Mutational Origin 2
Mutational Origin 3
-Population-of-birth
-Mutation's age
-Founder haplotypes
-Predisposing haplotypes
-Mechanisms responsible for *de novo* expansions

FIGURE 1 | Strategy followed to identify mutational origins in repeat-associated diseases by assessing SNP and STR haplotype backgrounds associated to expanded alleles.

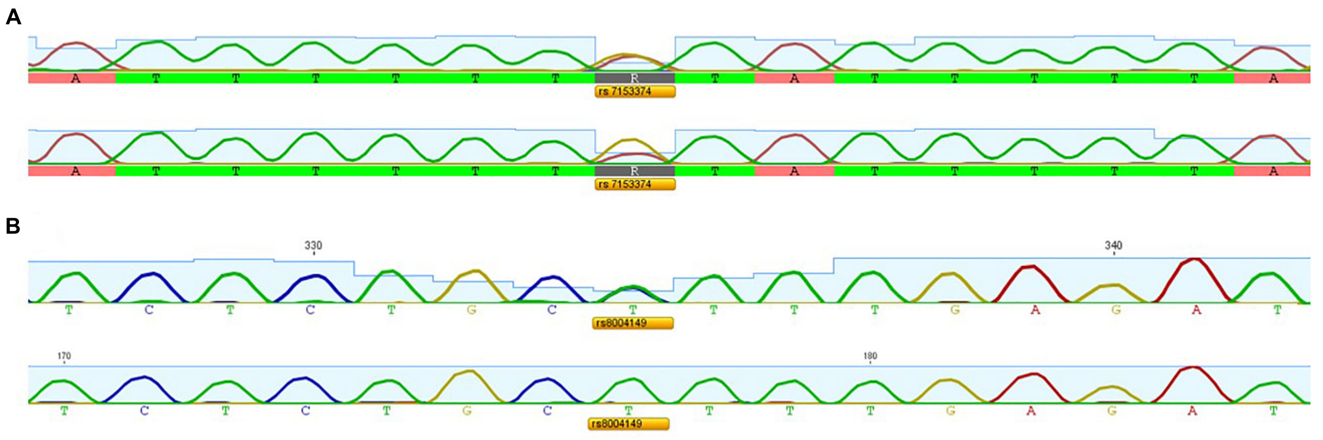

FIGURE 2 | Electropherograms showing **(A)** SNP rs7153374 genotyped by a standard PCR *versus* $(CAG)_n$-biased PCR and **(B)** SNP rs8004149 genotyped by a standard PCR *versus* allele-specific PCR.

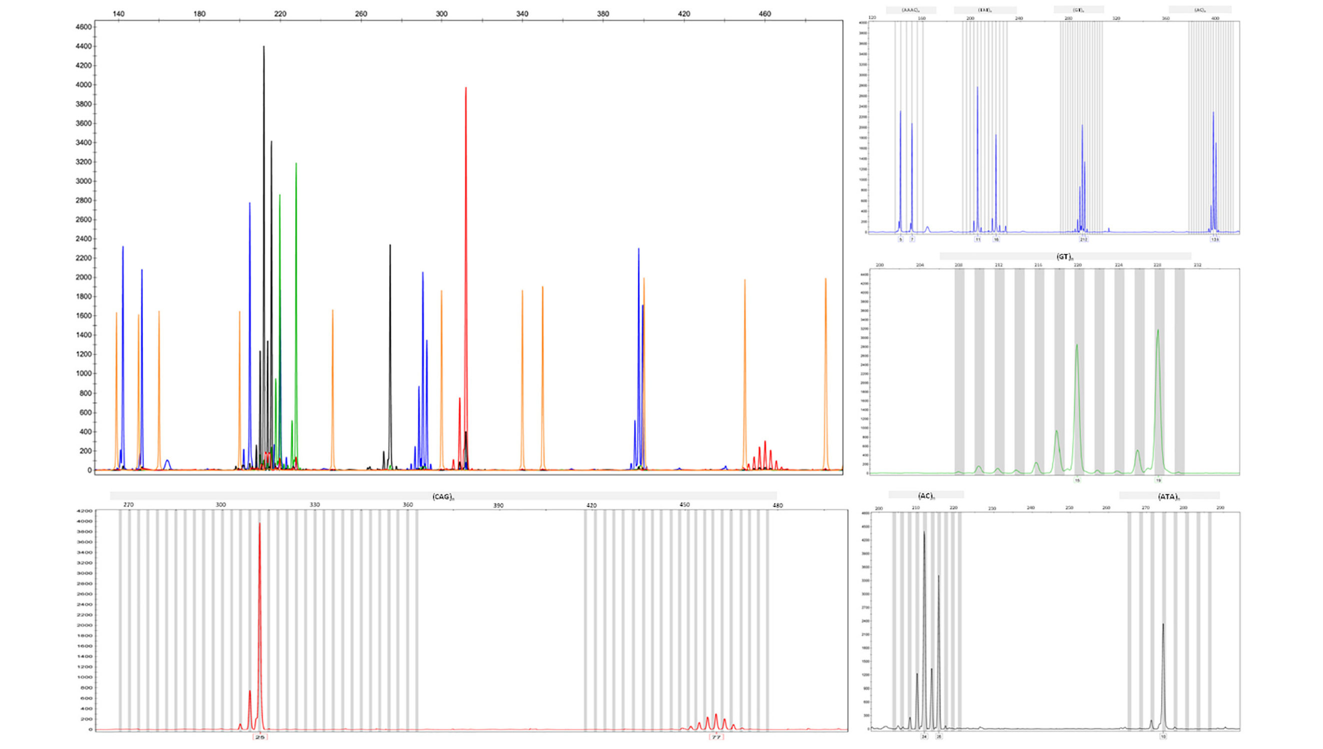

FIGURE 3 | Electropherograms showing amplified products from the optimized multiplex PCR for the analysis of the CAG repeat of *ATXN3* and seven flanking STRs.

by genetic drift and founder effects (Gaspar et al., 2001; Martins et al., 2007, 2012; Ogun et al., 2015); This optimized protocol could clarify remaining questions about the origins and history of MJD mutations. By following the same pipeline to study other expanding *loci*, one could understand better disease prevalence and provide a clearer scenario about the occurrence of *de novo* expansions (Venkatesh et al., 2018).

(2) **Mechanisms of *de novo* expansion.** Analysis of SNPs and STRs flanking *ATXN3* suggested a multistep mutation mechanism for the evolution of the CAG repeat responsible for MJD (Martins et al., 2006). Also based on the analysis of flanking haplotypes, other authors explained the origin of a rare intermediate MJD allele (45 CAGs) after a gene conversion event (Mittal et al., 2005). In other expansion diseases, risk SNP haplotypes seem to predispose to large jumps,

namely large expansions into the pathogenic range in *HTT* (responsible for HD) (Warby et al., 2009) and in *C9orf72* (the most common known genetic cause of amyotrophic lateral sclerosis and frontotemporal lobar degeneration) (Xi et al., 2015) or large contractions into the normal range in fragile X (Maia et al., 2017).

(3) **Instability of expanded alleles.** Once expanded, the haplotype background seems to affect $(CAG)_n$ intergenerational instability, since *cis*-elements in different haplotypes may regulate instability at that *locus*. In paternal transmission of MJD, opposite biases towards further expansion or contraction have been observed on specific SNP backgrounds: TTACAC *versus* GTGGCA haplotypes (Martins et al., 2008); this shows the relevance of further exploring the effect of flanking sequences on repeat instability. Haplotypes can serve as tags of *cis* or *trans* elements influencing *de novo* expansion or biasing towards further expansion or contraction, but can also have a direct effect. In the SCA7 gene, CTCF binding-sites *cis* to the expansion were shown to regulate *ATXN7*_CAG hyperinstability (Libby et al., 2003, 2008). Also, a haplotype study in fragile X has shown the direct influence of a flanking SNP on the replication of CGG repeats, due to its location within a medium reiterative element 1B sequence, proved to affect chromatin structure (Ennis et al., 2007). Through the identification of haplotypes, either with direct or indirect effect on repeat instability, we can improve genetic counseling, by predicting further expansion or contraction in offspring, but also understand better the mechanisms behind such events.

(4) **Phenotype modifiers.** Expanded repeat disorders frequently show an inverse correlation between expansion size and AO; however, this correlation explains only part of AO variation (Du Montcel et al., 2014). Haplotype background may play a direct role in disease expressivity; even after correcting for the known effect of expansion size; this is the case when a given SNP lies in the promoter or regulatory sequences of repeat-associated *loci*, which can affect binding of transcription factors, thus altering gene expression and disease presentation. In HD, SNP rs13102260 was identified as a modifier of AO. This SNP is located in a NF-κB binding site that regulates *HTT* promoter transcriptional activity. *In vitro* data showed a direct effect of rs13102260 on NF-κB binding and huntingtin expression, this way affecting disease presentation (Becanovic et al., 2015).

(5) **Clinical laboratory diagnosis.** The analysis of polymorphisms flanking the disease-causing repeat may also be helpful in diagnosis. A new genetic tool using 13 STRs flanking the expansion repeat has been proposed to help fragile X detection, avoiding ambiguity due to allele dropout (Rajan-Babu et al., 2017). Analysis of genetic markers is also important for predictive testing, namely to solve cases of apparent homoallelism, i.e., when a single peak may represent two normal alleles of the same size or a second allele expanded beyond the range of detection of the assay (Maciel et al., 2001; Smith et al., 2013).

(6) **Allele-specific therapies.** The urgency to develop direct therapies to prevent or slow progression of neurodegeneration led some authors to propose the sub-expression of the causative gene as a potential approach (Evers, 2015). In the case of *ataxin-3*, however, depletion of the gene results in cell death, by accumulation of ubiquitinated material, cytoskeletal disorganization and loss of cell adhesion (Alves et al., 2008). Therefore, the development of small interfering RNAs, based on the presence of a SNP that discriminates between wild-type and mutant transcripts, could be an efficient strategy for treatment of MJD. There are several SNPs described flanking the *ATXN3*_CAG repeat; however, for allele-specific therapy, patients must be heterozygous for the target SNP. The protocol optimized here for MJD may be used to identify the most informative SNPs for each population; the same pipeline can be followed to perform similar analysis in other repeat-associated diseases.

CONCLUSION

Identification of mutational haplotype backgrounds is key to unravel the mechanisms behind repeat-associated diseases. The strategy proposed was based on the analysis of fast evolving STR markers, placed within stable SNP-based lineages. By using the example of MJD, we have shown the procedure to assess allelic phases of both SNPs and STRs segregating with expanded alleles. Haplotype definition is crucial in dominant diseases, since normal alleles are not on the mutational background where *de novo* expansion(s) took place but are inherited from the non-affected parent. This is also relevant for recessive diseases, since two mutations (expanded alleles) in homozygosity may not share a common ancestral origin. When the gene of interest is on the X chromosome, that allows to infer haplotypes directly in males (e.g., when analyzing the CAG expansion in the *androgen receptor*, responsible for SBMA) (Santos et al., 2014).

AUTHOR CONTRIBUTIONS

SM conceived and designed the study. IC, BA, and SM developed the methodology. IC and BA analyzed the genotype. IC and SM drafted the manuscript. JS, AA, and SM critically revised the manuscript. AA and SM obtained funding and supervised the study.

FUNDING

This work was financed by FEDER-Fundo Europeu de Desenvolvimento Regional funds through the COMPETE

2020-Operacional Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the project "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274). SM is funded by FCT (IF/00930/2013). This article is a result of the project NORTE-01-0145-FEDER-000008, supported by Norte Portugal Regional Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00038/full#supplementary-material

REFERENCES

Alves, S., Nascimento-Ferreira, I., Auregan, G., Hassig, R., Dufour, N., Brouillet, E., et al. (2008). Allele-specific RNA silencing of mutant ataxin-3 mediates neuroprotection in a rat model of machado-joseph disease. *PLoS One* 3:e3341. doi: 10.1371/journal.pone.0003341

Andrés, A. M., Soldevila, M., Lao, O., Volpini, V., Saitou, N., Jacobs, H. T., et al. (2004). Comparative genetics of functional trinucleotide tandem repeats in humans and apes. *J. Mol. Evol.* 59, 329–339. doi: 10.1007/s00239-004-2628-5

Bampi, G. B., Bisso-Machado, R., Hünemeier, T., Gheno, T. C., Furtado, G. V., Veliz-Otani, D., et al. (2017). Haplotype study in SCA10 families provides further evidence for a common ancestral origin of the mutation. *NeuroMol. Med.* 19, 501–509. doi: 10.1007/s12017-017-8464-8

Becanovic, K., Nørremølle, A., Neal, S. J., Kay, C., Collins, J. A., Arenillas, D., et al. (2015). A SNP in the HTT promoter alters NF-κB binding and is a bidirectional genetic modifier of huntington disease. *Nat. Neurosci.* 18, 807–816. doi: 10.1038/nn.4014

Du Montcel, S. T., Durr, A., Bauer, P., Figueroa, K. P., Ichikawa, Y., Brussino, A., et al. (2014). Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain* 137, 2444–2455. doi: 10.1093/brain/awu174

Ennis, S., Murray, A., Brightwell, G., Morton, N. E., and Jacobs, P. A. (2007). Closely linked cis-acting modifier of expansion of the CGG repeat in high risk FMR1 haplotypes. *Hum. Mutat.* 28, 1216–1224. doi: 10.1002/humu.20600

Evers, M. M. (2015). *Developing Genetic Therapies for Polyglutamine Disorders*. 's-Hertogenbosch: Uitgeverij BOXPress.

Falush, D. (2009). Haplotype background, repeat length evolution, and huntington's disease. *Am. J. Hum. Genet.* 85, 939–942. doi: 10.1016/j.ajhg.2009.11.002

Fan, H., and Chu, J. Y. (2007). A brief review of short tandem repeat mutation. *Genomics Proteom. Bioinforma.* 5, 7–14. doi: 10.1016/S1672-0229(07)60009-6

Filippova, G. N., Thienes, C. P., Penn, B. H., Cho, D. H., Hu, Y. J., Moore, J. M., et al. (2001). CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat. Genet.* 28, 335–343. doi: 10.1038/ng570

Gaspar, C., Lopes-Cendes, I., Hayes, S., Goto, J., Arvidsson, K., Dias, A., et al. (2001). Ancestral origins of the machado-joseph disease mutation: a worldwide haplotype study. *Am. J. Hum. Genet.* 68, 523–528. doi: 10.1086/318184

Igarashi, S., Takiyama, Y., Cancel, G., Rogaeva, E. A., Sasaki, H., Wakisaka, A., et al. (1996). Intergenerational instability of the CAG repeat of the gene for machado-joseph disease (MJD1) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat. *Hum. Mol. Genet.* 5, 923–932. doi: 10.1093/hmg/5.7.923

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., Hayden, M. R., and Leavitt, B. R. (2017). *Epidemiology of Huntington Disease*, 1st Edn, Vol. 144. Amsterdam: Elsevier B.V, 31–46.

Lee, Y. C., Tsai, P. C., Guo, Y. C., Hsiao, C. T., Liu, G. T., Liao, Y. C., et al. (2016). Spinocerebellar ataxia type 36 in the han chinese. *Neurol. Genet.* 2:e68. doi: 10.1212/NXG.0000000000000068

Libby, R. T., Hagerman, K. A., Pineda, V. V., Lau, R., Cho, D. H., Baccam, S. L., et al. (2008). CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. *PLoS Genet.* 4:e1000257. doi: 10.1371/journal.pgen.1000257

Libby, R. T., Monckton, D. G., Fu, Y. H., Martinez, R. A., McAbney, J. P., Lau, R., et al. (2003). Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. *Hum. Mol. Genet.* 12, 41–50. doi: 10.1093/hmg/ddg006

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

Maia, N., Loureiro, J. R., Oliveira, B., Marques, I., Santos, R., Jorge, P., et al. (2017). Contraction of fully expanded FMR1 alleles to the normal range: predisposing haplotype or rare events? *J. Hum. Genet.* 62, 269–275. doi: 10.1038/jhg.2016.122

Martins, S., Calafell, F., Gaspar, C., Wong, V. C. N., Silveira, I., Nicholson, G. A., et al. (2007). Asian origin for the worldwide-spread mutational event in machado-joseph disease. *Arch. Neurol.* 64, 1502–1509. doi: 10.1001/archneur.64.10.1502

Martins, S., Calafell, F., Wong, V. C. N., Sequeiros, J., and Amorim, A. (2006). A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus. *Eur. J. Hum. Genet.* 14, 932–940. doi: 10.1038/sj.ejhg.5201643

Martins, S., Coutinho, P., Silveira, I., Giunti, P., Jardim, L. B., Calafell, F., et al. (2008). Cis-acting factors promoting the CAG intergenerational instability in machado-joseph disease. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 147, 439–446. doi: 10.1002/ajmg.b.30624

Martins, S., and Sequeiros, J. (2018). *Origins and Spread of Machado-Joseph Disease Ancestral Mutations Events*. Cham: Springer, 243–254. doi: 10.1007/978-3-319-71779-1_12

Martins, S., Soong, B. W., Wong, V. C. N., Giunti, P., Stevanin, G., Ranum, L. P. W., et al. (2012). Mutational origin of machado-joseph disease in the australian aboriginal communities of groote eylandt and yirrkala. *Arch. Neurol.* 69, 746–751. doi: 10.1001/archneurol.2011.2504

Maruyama, H., Nakamura, S., Matsuyama, Z., Sakai, T., Doyu, M., Sobue, G., et al. (1995). Molecular features of the CAG repeats and clinical manifestation of machado-joseph disease. *Hum. Mol. Genet.* 4, 807–812.

McGinty, R. J., and Mirkin, S. M. (2018). Cis- and trans-modifiers of repeat expansions: blending model systems with human genetics. *Trends Genet.* 34, 448–465. doi: 10.1016/j.tig.2018.02.005

Mittal, U., Srivastava, A. K., Jain, S., Jain, S., and Mukerji, M. (2005). Founder haplotype for machado-joseph disease in the indian population: novel insights from history and polymorphism studies. *Arch. Neurol.* 62, 637–640.

Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304.

Obayashi, M., Stevanin, G., Synofzik, M., Monin, M. L., Duyckaerts, C., Sato, N., et al. (2015). Spinocerebellar ataxia type 36 exists in diverse populations and can be caused by a short hexanucleotide GGCCTG repeat expansion. *J. Neurol. Neurosurg. Psychiatry* 86, 986–995. doi: 10.1136/jnnp-2014-309153

Ogun, S. A., Martins, S., Adebayo, P. B., Dawodu, C. O., Sequeiros, J., and Finkel, M. F. (2015). Machado-Joseph disease in a nigerian family: mutational origin and review of the literature. *Eur. J. Hum. Genet.* 23, 271–273. doi: 10.1038/ejhg.2014.77

Paulson, H. (2018). *Repeat Expansion Diseases*, 1st Edn, Vol. 147. Amsterdam: Elsevier B.V, 105–123.

Rajan-Babu, I. S., Lian, M., Cheah, F. S. H., Chen, M., Tan, A. S. C., Prasath, E. B., et al. (2017). FMR1 CGG repeat expansion mutation detection and linked haplotype analysis for reliable and accurate preimplantation genetic diagnosis of fragile X syndrome. *Expert Rev. Mol. Med.* 19, 1–12. doi: 10.1017/erm.2017.10

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. Part B Neuropsychiatr. Genet.* 153, 524–531. doi: 10.1002/ajmg.b.31013

Santos, D., Pimenta, J., Wong, V. C., Amorim, A., and Martins, S. (2014). Diversity in the androgen receptor CAG repeat has been shaped by a multistep mutational mechanism. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 165, 581–586. doi: 10.1002/ajmg.b.32261

Sequeiros, J., Martins, S., and Silveira, I. (2011). Epidemiology and population genetics of degenerative ataxias. *Handb. Clin. Neurol.* 103, 227–251.

Smith, D. C., Esterhuizen, A., and Greenberg, J. (2013). Caution regarding the interpretation of homoallelism in polyglutamine multiplex assays: a recommendation for confirmatory testing of homozygous alleles. *J. Mol. Diagn.* 15, 706–709. doi: 10.1016/j.jmoldx.2013.05.009

Souza, G. N., Kersting, N., Krum-Santos, A. C., Santos, A. S. P., 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

Vallone, P. M., and Butler, J. M. (2004). AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* 37, 226–231.

Venkatesh, S. D., Kandasamy, M., Moily, N. S., Vaidyanathan, R., Kota, L. N., Adhikarla, S., et al. (2018). Genetic testing for clinically suspected spinocerebellar ataxias: report from a tertiary referral centre in india. *J. Genet.* 97, 219–224.

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

Warby, S. C., Visscher, H., Collins, J. A., Doty, C. N., Carter, C., Butland, S. L., et al. (2011). HTT haplotypes contribute to differences in huntington disease prevalence between europe and east asia. *Eur. J. Hum. Genet.* 19, 561–566. doi: 10.1038/ejhg.2010.229

Xi, Z., van Blitterswijk, M., Zhang, M., McGoldrick, P., McLean, J. R., Yunusova, Y., et al. (2015). Jump from pre-mutation to pathologic expansion in C9orf72. *Am. J. Hum. Genet.* 96, 962–970. doi: 10.1016/j.ajhg.2015.04.016

Conflict of Interest Statement: 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.