Portuguese and Taiwanese Machado-Joseph disease haplotype backgrounds

Beatriz Lopes Columbano Marques Almeida

Genética Forense

Faculdade de Ciências

2017

Orientador

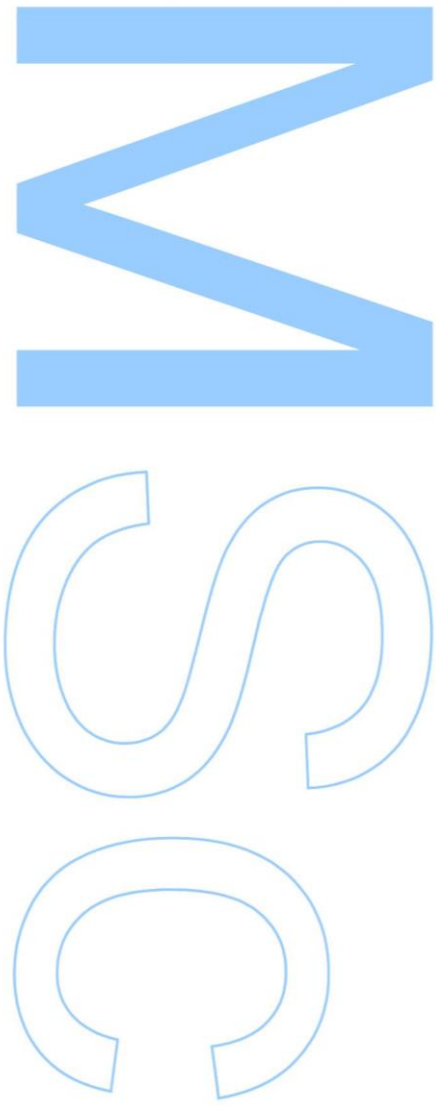
Sandra Martins, PhD, i3S/ IPATIMUP

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



"Toda a conquista começa com a decisão de tentar"

Gail Devers

Agradecimentos

Foram várias as pessoas que, direta ou indiretamente, ajudaram na elaboração deste trabalho, às quais gostaria de agradecer.

Ao Professor Amorim, pela disponibilidade e facilidades concedidas durante a realização deste trabalho.

À minha orientadora, Sandra Martins, pela sua dedicação, paciência, sentido, crítico e rigor científico depositados no decorrer deste trabalho.

Ao Centro de Genética Preditiva e Preventiva (CGPP) – i3S/IBMC, pela disponibilização das amostras.

A todos os doentes e seus familiares que colaboraram neste estudo, dirijo um agradecimento especial.

À Inês Costa por toda ajuda e amizade durante este último ano.

Aos meus companheiros de casa e amigos, Margarida e António, por todo o apoio.

Às minhas amigas, Bruna e Lara, por terem estado sempre disponíveis em qualquer momento e por sempre tornarem tudo melhor.

Ao Ricardo, pelas palavras de carinho e positividade.

Aos meus pais, por todo o apoio que sempre me deram, por todas as palavras de força e por tudo o que fazem por mim para que tudo seja possível.

A todos, muito obrigada!

Abstract

Machado-Joseph disease (MJD), also called spinocerebellar ataxia type 3 (SCA3), is a late-onset neurodegenerative disorder with a very pleomorphic clinical presentation. Among all autosomal dominant SCAs, MJD is the most frequent worldwide, and it is associated with a $(CAG)_n$ expansion in the *ATXN3* gene (14q24.3-q32.1). Patients usually carry one allele harbouring between 61 and 87 CAGs while the range for normal tract is 12-44. Contrarily to other repeat-associated disorders, there is a marked gap between normal and expanded ranges in MJD, from 45 to 60 CAGs, with the role of intermediate alleles on disease pathogenesis still unclear. To date, the molecular mechanism often proposed as causative of repeat instability of both intermediate and expanded alleles is the aberrant replication of DNA due to the formation of unusual conformations.

The analysis of MJD haplotypes defined by $(CAG)_n$-flanking SNPs and STRs allowed the identification of two independent mutational events: the worldwide spread Joseph lineage (TTA-$(CAG)_{exp}$-CAC) of Asian origin; and a more recent Machado lineage on the GTG-$(CAG)_{exp}$-GCA background. The authors analysed stable SNP haplotype backgrounds, which allowed them to distinguish alleles identical-by-descent from alleles identical-by-state flanking the CAG repeat, essential if the phylogenetic relationships among $(CAG)_n$ alleles are to be determined.

We now aimed at refining MJD haplotype studies to discern whether other independent-origin mutations may be underlying the previously identified MJD lineages. We extended the analysis to genotype a battery of 30 polymorphic markers (23 SNPs and 7 STRs) flanking the CAG repeat in 43 MJD families of Portuguese (n=21) and Taiwanese (n=22) origins.

Nineteen of the 23 analysed SNPs distinguished the two main Machado and Joseph lineages. In eight families from Taiwan, the SNP rs56268847 further discerned a more recent common ancestor for a Joseph-derived haplotype, previously described by us, in Australian aborigines and other Asian families. Based on flanking markers, genetic diversity is higher among Asian Joseph-derived families than in the most common Joseph lineage, which led us to raise alternative hypotheses to the origins of MJD. Next, our analysed fast-evolving markers will allow us to estimate more accurately the age of mutations and time for their introduction in different populations.

In addition to the study of worldwide spread MJD mutations, our analyses are important to identify the most informative SNPs for allele-specific down-regulation of expanded alleles during siRNA therapies.

Keywords: Neurodegenerative disorder, Machado-Joseph disease, *ATXN3* gene, PCR optimization, SNP, STR, haplotype, origins.

Resumo

A doença Machado-Joseph (DMJ), também designada por ataxia espinocerebelosa do tipo 3, é uma doença neurodegenerativa com início tardio e caracterizada por sintomas clínicos muito pleomórficos. Entre todas as ataxias espinocerebelosas autossómicas dominantes, a DMJ é a mais frequente em todo o mundo, e está associada a uma expansão de repetições do trinucleótido $(CAG)_n$ no gene *ATXN3* (14q24.3-q32.1). Os doentes com DMJ são, normalmente, portadores de um alelo que contém entre 61 a 87 repetições, enquanto que os alelos normais possuem repetições entre 12 a 44 CAGs. Contrariamente a outras doenças associadas a repetições, existe um notório intervalo entre o número de repetições dos alelos normais e expandidos, entre os 45 e os 60 CAGs. O papel destes alelos intermédios na patogénese da DMJ continua ainda por determinar. Atualmente, o mecanismo molecular comummente proposto como causador da instabilidade nas repetições trinucleotídicas (em alelos intermédios e expandidos) é a replicação anormal do ADN, levando à formação de conformações invulgares.

As análises dos haplótipos associados à DMJ definidos por SNPs e STRs que flanqueiam o $(CAG)_n$ permitiram identificar dois eventos mutacionais independentes: a linhagem Joseph, a mais comum no mundo (TTA-$(CAG)_{exp}$-CAC) de origem asiática; e uma linhagem mais recente – Machado – com o haplótipo GTG-$(CAG)_{exp}$-GCA. Assim, o uso de haplótipos estáveis de SNPs permitiram a distinção entre alelos idênticos-por-descendência e alelos idênticos-por-estado, essencial para determinar as relações filogenéticas entre os alelos $(CAG)_n$.

Atualmente o nosso objetivo é melhorar os estudos haplotípicos associados à DMJ de forma a descobrir se outras mutações com origem independente poderão estar subjacentes às duas linhagens previamente identificadas. Assim, aumentamos a análise para genotipar 30 marcadores polimórficos (23 SNPs e 7 STRs) que falqueiam a região trinucleotídica $(CAG)_n$ em 43 famílias DMJ com origens Portuguesa (n=21) e Taiwanesa (n=22).

Dezanove dos 23 SNPs analisados neste estudo distinguem as duas linhagens Machado e Joseph. Em 8 famílias de Taiwan, o SNP rs56268847 discerniu um ancestral comum mais recente para um haplótipo derivado da linhagem Joseph, anteriormente descrito pelo nosso grupo de investigação, em aborígenes australianos em outras famílias asiáticas. Através da análise de marcadores flanqueantes, a diversidade genética é maior nas famílias Asiáticas pertencentes à linhagem “Joseph-derived” do que na linhagem Joseph mais comum, o que nos levou a formular novas

hipóteses para as origens mutacionais da DMJ. Posteriormente, esses marcadores de evolução rápida irão permitir que consigamos estimar com mais precisão a idade das mutações e o tempo decorrido até à introdução das mesmas em diferentes populações.

Além do estudo das mutações associadas à DMJ em todo o mundo, as nossas análises serão importantes para identificar os SNPs mais informativos para o silenciamento específico de alelos expandidos durante a terapia génica usando a técnica associada ao siRNA.

Palavras-chave: Doenças neurodegenerativas, doença Machado-Joseph, gene *ATXN3*, optimização de PCR, SNP, STR, haplótipos, origens.

Contents

List of figures

List of tables

Supplementary material

List of abbreviation

The following list contemplates all the abbreviations used throughout this dissertation intending to facilitate its reading. Units and symbols from the metric system and following the International System of Units will not be included in the list, since publications from recognized authorities can be consulted regarding these topics.

AO	Age-of-onset
bp	Base pair
CpG	CG sites
dbSNP	Data base SNP
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
ID	Identification
Indel	Insertion or deletion
Kb	Kilobase
MJD	Machado-Joseph disease
PCR	Polymerase chain reaction
polyQ	Polyglutamine
RNA	Ribonucleic acid
RT	Room temperature
SCA	Spinocerebellar ataxia
siRNA	Small interference RNA
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
TEMED	Tetramethylethylenediamine

Introduction

1. Human Genomic Variation

Over human evolution, the DNA has changed as a consequence of different types of mutations, which occur in the germ line and are, therefore, heritable. Due to the high fidelity of DNA polymerases and DNA repair mechanisms, these mutations occur at low rates, still inevitable in every generation, which justify the current genetic diversity observed among individuals. On the other hand, meiotic recombination generates new combinations of allelic states from different polymorphic sites, resulting in new haplotypes, which consequently increase the genetic diversity among populations (Jobling *et al.* 2014). This way, genetic diversity represents the total genomic variation among individuals within a population (Barrandeguy & García 2014).

Human genetic diversity has been studied since blood groups were discovered (Landsteiner 1900). Early methods for studying this diversity were indirect, based on immunological reactions and electrophoretic analysis of gene products. Later, approaches evolved toward the direct detection of DNA variation. Some technological advances have further increased the pace of diversity studies, such as: i) the development of capillary DNA sequencing; ii) the subsequent determination of the human genome reference sequence; iii) precise methods for high-throughput genome-wide single nucleotide polymorphism (SNP) typing; iv) consistent methods for the measurement of structural and copy-number variation; and, v) the development of next-generation and third-generation sequencing (Jobling *et al.* 2014).

Several measures of genetic diversity have been developed over the years (Barrandeguy & García 2014). Perhaps the simplest way to describe the amount of diversity is to count the number of haplotypes present. This can be done either within many populations for a single locus, allowing a comparison of diversity between populations, or at many loci within a single population, allowing a comparison among loci (Jobling *et al.* 2014).

The Nei's gene diversity statistic is normally used to measure genetic diversity observed among members of a population (Nei 1987; Jobling *et al.* 2014). If we consider that a certain genetic locus has K distinct alleles, with positive frequencies P_1, P_2, …P_K, then $\sum_{i=1}^{k} Pi=1$, and the gene diversity at this locus is defined by

$$H=1-\sum_{i=1}^{K} P^2 i$$

Thus, *H* is the probability that two members, randomly chosen from a population, will show different genetic alleles at a specific site in the genome (Kang 2015). Consequently, for diploid loci this statistic is referred to as a measure of heterozygosity (Jobling *et al.* 2014).

The genetic diversity, generated as a consequence of mutations, is the ultimate source of variation for the long-term evolution of organisms, an important parameter in evolutionary biology, which provides perception into the demographic history of human populations (Osada 2015). Additionally, the study of human genetic diversity is essential to understand the mutational mechanisms and disease susceptibility at both population and individual levels (Lu *et al.* 2014; Osada 2015).

2. Molecular Markers

A molecular marker is a nucleotide sequence corresponding to a particular known or unknown physical location in the genome, which allows to obtain information of genetic diversity and also differentiation of individuals at the DNA level (Barrandeguy & García 2014). In less than half a century, molecular markers have totally changed our view of biology because of their ability to differentiate two or more genotypes (Schlotterer 2004; Grover & Sharma 2014). Also, in the last decades, progress in molecular biology has increased the capacity to detect molecular genetic markers. The ease of handling DNA markers improved substantially with the development of techniques based on nucleic acid hybridization, polymerase chain reaction (PCR) and DNA sequencing (Grover & Sharma 2014). These molecular markers have been analysed to answer many biological questions, ranging from gene mapping to population genetics, phylogenetic reconstruction, paternity testing and forensic applications (Schlotterer 2004).

The first true molecular markers to be established were allozymes since previously, various factors restricted the use of secondary metabolites as a marker due to their instability (Grover & Sharma 2014). The principle behind allozymes is based on protein variants distinguished by native gel electrophoresis. However, allozymes were an indirect and insensitive method to detect variation in DNA and, for this reason, the arrival of DNA manipulation techniques promoted a shift from enzyme-based to DNA-based markers (Schlotterer 2004). The discovery of restriction fragment length polymorphism (RFLPs) improved the analysis of non-coding sequences, and the survey of changes in coding sequence, contributing also to the first DNA-based genetic maps (Vignal *et al.* 2002; Schlotterer 2004). Later, the extremely high polymorphic level

of minisatellites revolutionized the genetic identification of individuals with the technique of DNA fingerprinting. In 1988, the invention of PCR improved the PCR-based markers, such as random amplified polymorphic DNA (RAPDs), inter-simple sequence repeats (ISSRs), inter-retrotransposon amplified polymorphisms (IRAPs), amplified fragment length polymorphisms (AFLPs) and microsatellites. These molecular markers allowed the amplification and analysis of virtually all genomic regions (Schlotterer 2004). Therefore, PCR, as well as fluorescent sequencing and fragment analysis technologies, have catalysed a revolution in the development of genetic markers for the analysis of populations (Brumfield *et al.* 2003). The latest major innovation in detecting genetic variation in populations was the analysis and identification of SNPs (Goodwin *et al.* 2011). More recently, advances in next generation sequencing (NGS) technologies have contributed towards high throughput discovery of even larger numbers of SNPs, revolutionizing again genetic diversity assessment projects and genome-wide association studies (GWAS) (Grover & Sharma 2014).

2.1 Microsatellites

Short Tandem Repeats (STRs, also called microsatellites) are a class of simple repetitive DNA sequences composed of small motifs between 1 and 7 nucleotides repeated in tandem, typically from 10 – 30 times, forming series with lengths of up to 100 nucleotides (Jobling *et al.* 2014; Kouniaki *et al.* 2015).

Microsatellites are more mutable than other parts of the human genome- 10^{-3} to 10^{-4} *versus* 10^{-9} mutations per locus per generation (Jobling *et al.* 2014; Kouniaki *et al.* 2015). Differences in mutation rates may however, differ according to several factors like: repeat number, repeat sequence, repeat configuration, and flanking sequence, in addition to the gender of the transmitting progenitor (Bhargava & Fuentes 2010). Several mechanisms have been suggested to explain this high mutation rate (Fan & Chu 2007; Bhargava & Fuentes 2010); however, it is commonly accepted that microsatellite mutations occur as a result of DNA replication slippage (Figure 1), even though direct evidence of the mechanism remains elusive (Jobling *et al.* 2014).

STRs appear distributed throughout the human genome, accounting for about 3% of the entire genome, with one STR per 2,000 bp (Fan & Chu 2007; Eckert & Hile 2009). Their distribution within chromosomes is not uniform and recent evidence pointed to their non-random genomic distribution (Eckert & Hile 2009), with most STRs found in noncoding regions (compared to only 8% located within exons) (Kouniaki *et al.*

2015). The length of common – mono-, di- and tetranucleotide – microsatellite alleles is dependent upon genome location. Generally, alleles within noncoding regions are longer than those within exons (Eckert & Hile 2009). In coding regions, STRs tend to occur in genes with roles in transcriptional regulation, DNA binding, protein-protein binding, and developmental processes. In these regions, STR mutations are generally in-frame additions or subtractions of repeat units, probably due to selective pressures. At some loci, STRs can expand above a given threshold, therefore leading to several neurological diseases as a consequence of dramatically expanded/deleterious alleles (Press *et al.* 2014). Many of these disease-associated STR expansions behave as dominant gain-of function mutations (Fan & Chu 2007; Press *et al.* 2014).

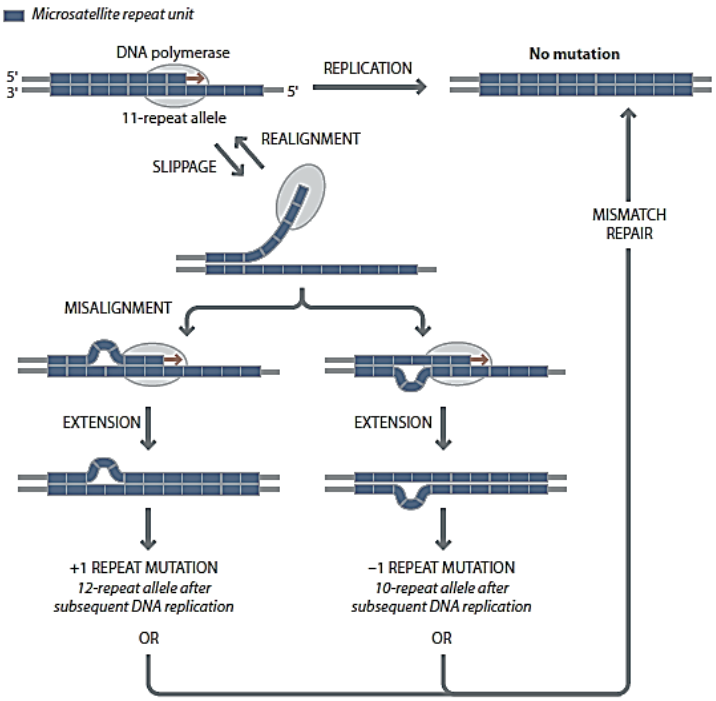


Figure 1: Model of the effects of slippage and misalignment in microsatellite mutation (adapted from Jobling *et al.* 2014).

2.2 Single Nucleotide Polymorphisms (SNPs)

SNPs are nucleotide sites in a DNA sequence where more than one allelic state exist in some populations; usually, they are binary and the minor allele frequency is 1% or greater (Brookes 1999). SNPs are found in the human genome about once in every 1,000 bp. Given that the human genome is 3,2 billion bp long, it can be estimated that there will be approximately 1 million differences between two human genomes, only due to SNPs (Goodwin *et al.* 2011). Among the principal genetic factors responsible for human phenotype variations are the SNPs, which account for approximately 90% of genetic variation observed in the human genome, with approximately 324 million SNPs

identified (dbSNP-announce at ncbi.nlm.nih.gov, Thu Apr 6 2017) (Brookes 1999; Brumfield *et al.* 2003).

Due to their low mutation rate (typically 10^{-8} mutations per nucleotide per generation), the chance of finding a SNP resulted from independent mutations in two individuals is small, reason why they tend to show identity-by-descent. Two fundamental processes lead to base substitutions: (a) misincorporation of nucleotides during replication; and (b) mutagenesis caused by the chemical modification of bases or physical damage. However, in both cases, sophisticated DNA repair processes can detect and repair a great part of base substitutions, resulting in germline mutations not transmitted to the next generation, which justifies the low SNP mutation rate (Jobling *et al.* 2014).

Mutations on the origin of SNPs are abundant and widespread in the genome (coding and non-coding regions), with an important subset occurring in genes associated with diseases or other phenotypes (Morin *et al.* 2004; Grover & Sharma 2014); nevertheless, as observed for STRs, the probability to find a SNP is higher in non-coding regions (Pinheiro 2010). In particular, in CpG dinucleotides, mutation rate is elevated because of methylation process. About 75% of CpG dinucleotides in our genome are targets of DNA methylation, a specific methyltransferase enzyme that adds a methyl group to the cytosine ring giving 5-methylcytosine (Figure 2). Spontaneous or mutagen-induced deamination (loss of an amine group) of cytosine yields uracil is not a legitimate base in DNA, being efficiently recognized and removed by uracil glycosidase. Deamination of 5-methylcytosine yields thymine, a legitimate DNA base. To fix this T-G mispair, either the T can be changed to a C, or the G to an A. Taking to account that repair machinery often makes the incorrect choice, CpG dinucleotides are hotspots for mutation (Jobling *et al.* 2014).

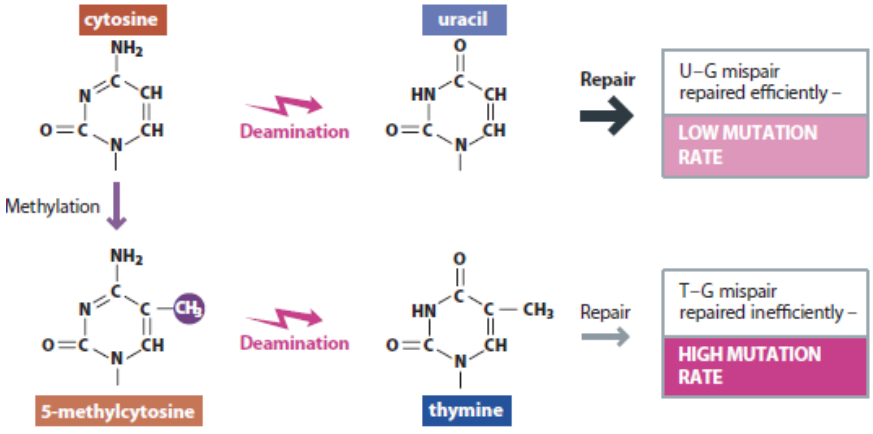


Figure 2: The CpG dinucleotide is a site for methylation (adapted from Jobling *et al*. 2014).

The biallelic state of the vast majority of these polymorphisms intrinsically limits the information that can be obtained from the analysis of any given SNP, which has been the major limiting factor for their application in forensic analysis: between 50 and 80 SNPs are required to achieve the same levels of discrimination as the current STR based methods (Goodwin *et al.* 2011); however, when DNA samples are of low quality, the analysis of SNPs may be more effective (Pinheiro 2010).

In the forensic field, SNPs have been attributed a continuously increasing importance since: a) their very low mutation rate is relevant for paternity testing; b) they are very suitable for analysis using high technologies, important for comprehensive databases; and c) they can be analysed in short amplicons, usually more desirable for the successful amplification of degraded samples (Sobrino *et al.* 2005).

3. Triplet Repeat Diseases

A new type of human genetic disease mutation was discovered more than twenty years ago – the expansion of repeated microsatellite sequences (La Spada & Taylor 2010). In humans, consequences of triplet instability on health may be severe since the expansion of trinucleotide repeats is recognized as a major cause of several neurological and neuromuscular diseases (Budworth & McMurray 2013). The understanding of the pathogenic mechanism for this diseases is unknown, but has advanced substantially in recent years (McMurray 2010; Zhao 2015). Growing evidence supports the idea that repeat expansion diseases may be the result of aberrant DNA repair (Zhao 2015).

The consequences of the expansion depend on some combination, including the location of the repeat within the affected gene, the size of the repeat unit, the number of repeats present in the deleterious allele and the sequence of the repeat (Zhao 2015). Often, repeats in intronic regions are close enough to exons to disrupt genes functions, however disease-associated trinucleotides repeats are usually inside exons (Fan & Chu 2007). Among recurrent sequence motifs are $(CAG)_n$, $(CTG)_n$, and $(GCG)_n$ motifs, coding for polyglutamine, polyleucine and polyalanine tracts, respectively (Guyenet & La Spada 2006). The largest group of triplet repeat diseases is caused by polyglutamine tracts, most of them responsible for spinocerebellar ataxias (Budworth & McMurray 2013).

3.1 Spinocerebellar Ataxias

Ataxias are a group of neurological disorders that result from variable levels of degeneration of neurons in the cerebellum, brain stem, spinocerebellar tracts, and their afferent/efferent connections. There are several different forms of inherited ataxias that strike during childhood or adulthood (Orr 2012).

Spinocerebellar ataxias (SCAs), usually denoting autosomal dominant inherited ataxias, are a heterogeneous group of neurologic disorders characterized by variable degrees of degeneration in the cerebellum, spinal tracts and brain stem (Zoghbi & Orr 2000). SCAs are considered rare disorders, with prevalence varying usually from 0.3 to 2.0 per 100,000 individuals (Bettencourt & Lima 2011). There are approximately 40 different types of SCAs described to date but not all causative mutations have been identified (Bird 1998; Orr 2012). Until now, more than twenty-seven genetic loci have been identify for SCA, namely the SCA1-8, SCA10-23, SCA25-30 and dentatorubral-pallidoluysian atrophy (DRPLA) (Gomes *et al.* 2017). According to molecular mechanisms of disease, SCAs may be organized in four groups (Shakkottai & Fogel 2013):

1. Polyglutamine (polyQ) ataxias;
2. Non coding repeats/ RNA toxicity;
3. Ataxias associated with ion-channel dysfunction; and
4. Ataxias associated with mutations in signal transduction molecules.

At the genetic level, the most common polyQ ataxias are caused by an abnormal expansion of a CAG repeat sequence that encodes an expanded tract of polyQ residues within the mutated protein (Orr 2012).

PolyQ disease proteins differ in size, cellular localization and biological function, suggesting that the toxic effect of a given polyQ expansion depends on the specific protein context, with particular details of pathogenesis probably unique to each disease (Costa & Paulson 2012). Potential mechanisms by which an elongated polyQ protein causes neuronal toxicity and further ataxia include: (a) protein misfolding resulting in altered function; (b) formation of toxic oligomeric complexes; (c) transcriptional dysregulation; (d) mitochondrial dysfunction; (e) impaired axonal transport; (f) aberrant neuronal signalling including excitotoxicity; (g) cellular protein homeostasis impairment and (h) RNA toxicity. These hypotheses are by no means mutually exclusive (Williams & Paulson 2008).

4. Machado-Joseph disease / spinocerebellar ataxia type 3

Numerous designations have been given to this disorder, namely “Machado disease”, “nigro-spinodentatal degeneration with nuclear ophthalmoplegia”, “autosomal dominant striatonigral degeneration” and “Azorean disease of the nervous system” (Bettencourt & Lima 2011; Costa & Paulson 2012). At present, the most used designation is Machado-Joseph disease (MJD) or, alternatively, spinocerebellar ataxia type 3 (SCA3).

4.1 Clinical Presentation

MJD is a multisystem neurodegenerative disorder with autosomal dominant inheritance, described by a high degree of pleomorphism in the variability of age at onset and in the neurological signs, that result in different degrees of incapacity (Bettencourt & Lima 2011).

Clinically, MJD is characterized by gait ataxia, spasticity, dystonia (repetitive muscle contractions), parkinsonism (subtype 4), myokymia (involuntary movements of facial and lingual muscles), dysarthria (motor speech disorder), dysphagia (swallowing problems), ophthamoparesis (abnormal eye movements), and diplopia (double vision) (Schöls *et al.* 2004; Costa & Paulson 2012; D'Abreu & Friedman 2016).

Many pathology reports have been published. Macroscopic brain examinations showed the pallor of the substantia nigra as well as the degeneration of the cerebellum and brainstem (Figure 3) (Park *et al.* 2015).

Figure 3: Distribution of neuronal loss. Red marks indicate regions with more prominent degeneration in MJD patient’s brains according to neuropathological studies (adapted from Saute & Jardim 2015).

The observation of three families of Azorean ancestry (Machado, Thomas and Joseph) led to the initial description, during the 1970s, of three apparently independent diseases. The first descriptions of MJD in Portugal were by Coutinho and Andrade in 1978, who studied 15 families from the Azores (Coutinho & Andrade 1978). The subsequent identification of several Portuguese families, living in the Azores islands and in mainland Portugal, led to the unification of the disease as a single genetic entity with variable phenotypic expression (Bettencourt & Lima 2011). Initially, in 1978, Coutinho and Andrade systematized the disease phenotypes into three clinical types (Coutinho & Andrade 1978). Later, two more clinical types have been proposed as described in Table 1.

Table 1: Clinical subtypes in MJD/SCA3 (Adapted from Bettencourt & Lima 2010; Saute & Jardim 2015).

MJD subtype	Clinical characteristics
Type 1 - Joseph	Early onset (progress more quickly), gait ataxia, pyramidal signs and dystonia.
Type 2 - Thomas	Most common type, intermediated onset, gait ataxia and with or without pyramidal sings.
Type 3 - Machado	Later onset, gait ataxia, associated with peripheral alterations, with or without slight pyramidal and extrapyramidal signs.
Type 4	Rare presentation with parkinsonian features, with mild cerebellar deficits and distal motor sensory neuropathy.
Type 5	Spastic paraplegia without gait ataxia (this designation has not been commonly accepted).

4.2 Genetic Epidemiology

MJD is the most common form of SCA worldwide (Teive *et al.* 2016). Among SCAs, the relative frequency of MJD is higher in countries such as Brazil- Rio Grande do Sul (92%) (Jardim *et al.* 2001), Portugal (58%) (Vale *et al.* 2010), Singapore (53%) (Zhao *et al.* 2002), China (49%) (Jiang *et al.* 2005), the Netherlands (44%) (Van de Warrenburg *et al.* 2002), Germany (42%) (Schöls *et al.* 1997), Japan (63%) (Shibata-Hamaguchi *et al.* 2009), and it is considered as relatively rare in Canada (24%) (Kraft *et al.* 2005), México and Australia (12%) (Storey *et al.* 2000; Alonso *et al.* 2007), India (14%) (Krishna *et al.* 2007), South Africa (4%) (Bryer *et al.* 2003) and Italy (1%) (Brusco *et al.* 2004). Even within each country, the geographic distribution pattern is not homogeneous. In Portugal, for example, MJD is relatively rare in mainland (1/100,000), but highly prevalent in the Azores islands (1/239) (Bettencourt & Lima 2011).

Two large studies have focused on the worldwide origin of MJD mutations. In 2001, Gaspar *et al.*, by haplotype analyses of five microsatellite markers flanking the *ATXN3* gene and three intragenic SNPs (A^{669}TG/G^{669}TG, C^{987}GG/G^{987}GG and TAA^{1118}/TAC^{1118}), found that two (ACA and GGC) MJD haplotypes, were present in 94% of all MJD families studied (Gaspar *et al.* 2001). In Azorean families, ACA haplotype was observed in Flores, while GGC haplotype was found in families from São Miguel. Their results indicated that two distinct mutational events accounted for the presence of MJD in the Azorean islands and in families of Azorean ancestry. In mainland Portugal, both haplotypes were also found. Worldwide, 72% of the families share the ACA haplotype, further supporting the idea of few mutational events responsible for MJD (Gaspar *et al.* 2001; Bettencourt & Lima 2011). Six years later, Martins *et al*. through a more extensive haplotype analyses (Martins *et al.* 2007), proposed an Asian origin for the worldwide spread MJD lineage. Their work revealed that TTACAC haplotype (or Joseph lineage) reached the highest diversity in Asia, with an estimated mutation age of ~ 7,000 years. A second *de novo* expansion, on Machado lineage (GTGGCA haplotype), is thought to be more recent (2,000 years old). The origin of Machado lineage is more controversial, but its dispersion may be mainly explained by recent Portuguese emigration (Martins *et al.* 2007; Bettencourt & Lima 2011).

4.3 *ATXN3* gene

In 1993, the disease locus for MJD was mapped to the long arm of chromosome 14, by Takiyama *et al*. (Takiyama *et al.* 1993). One year later, Kawaguchi *et al.* showed that an expansion of a CAG repeat motif at the *MJD1* gene (currently named *ATXN3*), mapped to 14q32.1, was present in all individuals with MJD disease (Kawaguchi *et al.* 1994). The genomic structure of this gene was published seven years later (Figure 4) (Ichikawa *et al.* 2001). In MJD, the CAG repeat is located in the penultimate exon of the canonical *ATXN3* transcript. The expanded repeat encodes an almost pure polyglutamine stretch interrupted by a single lysine codon due to CAG/CAA codon degeneracy $(CAG)_2$CAA AAG CAG CAA$(CAG)_n$ (Kawaguchi *et al.* 1994). Wild-type alleles range from 11 to 44 CAG repeats, whereas limits of expanded alleles usually comprise from 61 to 87 repeats units (Martins *et al.* 2012). Intermediate size alleles are rare. To date, only seven alleles have been reported, six of them associated to different clinical presentations of MJD [45 (Padiath *et al.* 2005), 51 (Gu *et al.* 2004), 53 (Van Alfen *et al.* 2001), 54 (Van Alfen *et al.* 2001), 55 (Egan *et al.* 2000), 56 (Takiyama *et al.* 1997)], and the seventh allele, with 51 repeats, has been described as being

apparently not associated with disease. This raised the possibility that low penetrance alleles, of intermediate size, may occur in MJD (Maciel *et al.* 2001; Bettencourt & Lima 2011). Recently, Takahashi *et al*. described a family with compound heterozygous individuals presenting intermediate *ATXN3* alleles (55/49) associated with progressive cerebellar ataxia and sensory axonal neuropathy. Their results suggested that two intermediate alleles carried by the same individual potentiates the phenotypic expression of the disease, suggesting multiple mechanisms for the development of MJD (Takahashi *et al.* 2017).

ATXN3 gene encodes the ataxin-3 protein (~42 kDa), which belongs to the family of cysteine proteases, described as being involved in protein quality control pathways in the cell, transcriptional regulation and cytoskeleton organization (Bettencourt & Lima 2011; Matos *et al.* 2016). Different studies demonstrated that ataxin-3 is ubiquitously transcribed in neuronal and non-neuronal human tissues. Regarding its subcellular location, ataxin-3 has been observed both in the cytoplasm and the nucleus of various cell types. Thus, cellular expression of this gene is not itself sufficient to explain selective neuronal degeneration (Bettencourt & Lima 2011).

Recently, fifty-six alternative splicing variants generated by four types of splicing events (which can occur in combined way) were described for the *ATXN3* gene (Bettencourt & Lima 2011; Matos *et al.* 2011). Thus, alternative splicing may be an important mechanism regulating ataxin-3 diversity. In its mutated form, when the polyQ tract reaches the pathological threshold, the protein is thought to gain a neurotoxic function with protein aggregations formed inside the cell that, as a consequence, lead to selective neuronal cell death through a not fully understood process (Bettencourt & Lima 2011).

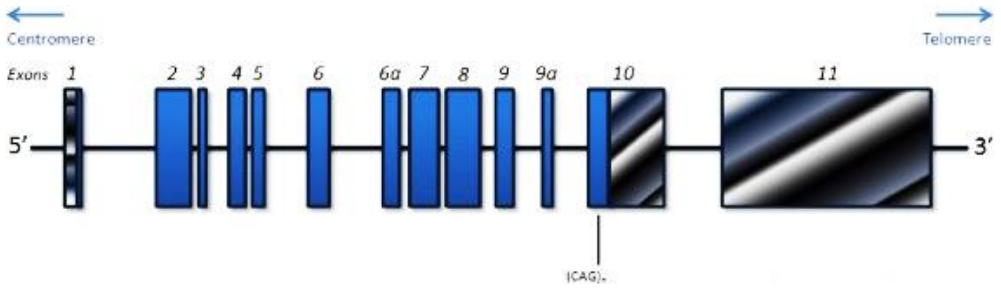


Figure 4: Schematic representation of the *ATXN3* gene structure. Exons are numbered from 1 to 11, represented as boxes. Blue boxes indicate the coding regions. The location of the polymorphic $(CAG)_n$ tract is indicated (adapted from Bettencourt & Lima 2011).

4.4 Genotype-phenotype correlation

Machado-Joseph disease shows an autosomal dominant pattern. Thus, each descendent of an affected individual has, *a priori*, a risk of 50% of being itself a carrier, with both genders having equal probabilities of receiving/transmitting the expanded allele and expressing the disease (Bettencourt & Lima 2011).

The allele size (CAG repeat length) is the major determinant of disease age-of-onset, progression and severity phenotype (Martins *et al.* 2014). An inverse correlation is found between the size of the expanded CAG repeat tract and the age-of-onset (AO) of the disease (Bettencourt & Lima 2011).

The expanded repeats are unstable, especially in male meiosis, with tendency to expand. This leads to earlier onsets and more severe phenotypes in following generations – a clinical phenomenon known as anticipation (Schöls *et al.* 2004). Such phenomenon can be explained by the dynamic process of mutation underlying triplet repeats diseases, which involves intergenerational instability (Bettencourt & Lima 2011). Most cases of childhood onset are caused by extreme expansions between generations (Schöls *et al.* 2004). CAG repeat instability can also occur in different cells from the same tissue, originating somatic mosaicism. In MJD, somatic mosaicism may occur in the brain, but larger repeats are not preferentially associated with affected brain regions (Costa & Paulson 2012).

The genetic background is another factor that seems to influence instability of repeats (Martins *et al.* 2008).The presence of several SNPs an STRs flanking the CAG tract have permitted a better understanding of the mechanisms behind repeat instability (Costa & Paulson 2012). Previously studies demonstrate that, in MJD some polymorphisms have an inter-allelic interaction on the allele from the normal chromosome that influence the intergenerational instability of expanded alleles. This was a strong evidence that elements in *trans* may affect repeat instability (Martins *et al.* 2008). Also, in 2015, two novel SNPs (rs709930 and rs910369) were analysed in Chinese patients, possibly genetic modifiers in MJD since they have been associated to a decrease of approximately 2 to 4 years in MJD AO (Long *et al.* 2015).

Cases of homozygosity are extremely rare in MJD, but the few described homozygous patients for the expansion appear to show a more severe form of disease suggesting a gene dosage effect (Carvalho *et al.* 2008; Costa & Paulson 2012); with a more severe progression and an early AO (Carvalho *et al.* 2008; Zeng *et al.* 2015). Loss of function of the normal expressed ataxin-3, or possibly aggregation of ataxin-3, may be implicated in disease mechanism (Carvalho *et al.* 2008).

Additionally, evidence of the contribution of genetic and/or environmental factors to disease presentation came in 2011, when it was proposed that epigenetic factors, such as DNA methylation in the promoter region of the *ATXN3* gene might contribute to clinical variation, with a small and positive effect on the AO (Emmel *et al.* 2011). More recently, Wang *et al.* found higher methylation levels in the first CpG island of the *ATXN3* promoter in MJD patients with earlier AO and unstable intergenerational transmissions of the CAG repeats (Wang *et al.* 2017).

4.5 Pathology

Alternative splicing may play an important role in MJD, as mentioned earlier. By analysing the domain composition of several ataxin-3 isoforms, it has been possible to predict a protective role for some isoforms, while others showed to lead to increased toxicity (Bettencourt *et al.* 2010). Independently of ataxin-3 isoform, it has been observed that ataxin-3 subcellular distribution differs in diseased brain compared to normal brain. Thus, cytoplasmic proteins are normally observed in normal brain, whereas during disease, ataxin-3 becomes concentrated in the nucleus of neurons with formations of intranuclear inclusions in many brain regions (Paulson *et al.* 1997). These neuronal inclusions, pathological hallmark of MJD, are severely ubiquitinated and contain certain heat shock molecular chaperones and proteasomal subunits, suggesting that they are repositories for aberrantly folded, aggregated proteins (Chai 2002; Schmidt *et al.* 2002). Functional interactions of wild-type ataxin-3 with other molecules have been shown to reduce its aggregation propensity and increase solubility. In the case of mutant ataxin-3, it is possible that these modifier interactions are reduced, leading to a faster rate of aggregation. Expanded ataxin-3 interacts abnormally with at least some of its native partners, impeding its own degradation and leading, in some cases, to a gain-of-function and, sometimes to partial loss of normal ataxin-3 function (Costa & Paulson 2012).

Numerous hypotheses, not mutually exclusive, have been considered while studying the potential toxic mechanism caused by misfolded mutant *ATXN3* and its altered proteins interaction: (i) formation of aggregates; (ii) failure of cellular protein homeostasis; (iii) impairment of axonal transport; (iv) transcriptional dysregulation; (v) mitochondrial dysfunction and oxidative stress; and (vi) abnormal neuronal signalling (Costa & Paulson 2012).

Objectives

Given the importance of haplotype analyses in the context of repeat-expansion disorders, our main objective for this study was to extend these analyses through the genotyping of a battery of polymorphic markers flanking our repeat of interest - *ATXN3*-$(CAG)_n$. Therefore, we aimed at designing a pipeline to perform a comprehensive haplotype analyses in MJD that will be useful to complete studies on tracing spreading routes of MJD mutations.

Our specific goals were:

1. To design a protocol to access allelic phases of expanded MJD alleles and flanking SNPs;

2. To identify the best panel of SNPs to distinguish the two main MJD lineages identified so far: Machado and Joseph;

3. To identify the most informative SNPs to distinguish normal and expanded alleles in MJD patients from different lineages, further useful for allele-specific down regulation of expanded alleles during siRNA therapies;

4. To identify *de novo* mutational origins in Machado-Joseph disease.

Subjects and Methods

1. Subjects

We studied forty-three families with MJD, twenty-one from Portugal and twenty-two from Taiwan. The Portuguese samples are from Centro de Genética Preditiva e Preventiva (CGPP) – i3S/IBMC, where the molecular diagnosis of MJD is performed; written informed consent has been previously provided by all individuals. Taiwan samples have been sent by Dr. Bing-Wen Soong from Neurological Institute, Veterans General Hospital, Taipei, Taiwan. The length of normal and expanded polyglutamine tracts was previously determined by capillary electrophoresis. DNA quantification was performed with Nanodrop spectrophotometer to make work aliquots with a final DNA concentration of approximately 7.5 ng/µL.

2. Primer design

To amplify and sequence a 4 kb region flanking exon 10 of the *ATXN3* gene, we designed sequence specific primers (Sigma) (Table 2) by using the online software Primer3Plus (Rozen & Skaletsky 1999), with the following conditions: 18 to 27 bp; a melting temperature comprised between 57°C and 63°C; a percentage of GC between 20% and 80%; in the mispriming/repeat library of the human species. Next, to test the occurrence of hairpins and primer self-dimers, we used the OligoCalc algorithm (Kibbe 2007): sequences prone to form hairpin structures were discarded. The alignment tool BLAT (Kent 2002) (human genome, assembly GRCh38/hg38) was then used to guarantee the specificity of designed primers to the target sequence, and finally, each primer was subjected to a nucleotide BLAST (Johnson *et al.* 2008) (database: nucleotide collection; organism: human; expected threshold: 10; megablast) to find similar sequences. The compatibility of multiple primer sequences was checked with AutoDimer software (Vallone & Butler 2004), which detects the formation of primer-dimers in short DNA oligomers.

Table 2: Primers designed to amplify and sequence a 4 kb region flanking the *ATXN3*-$(CAG)_n$ repeat. F - Forward primer; R – Reverse primer.

Name	Primer sequence (5'-3')	
Clo F *	F	CAATTATTGGCCTTTCTGAACC
MJDClo 653*	R	GCAAATGAGTGTTGGTTTATAGACCC
MJDClo 716*	F	ACAGAGTCTCGCTCTGTCGCCCAG
MJDClo 52 *	F	CCAGTGACTACTTTGATTCG
MJDClo 1260 *	R	GCTGTCTGAAACATTCAAAAGTGAAG
MJDClo 1342*	F	CCACCAGTTCAGGAGCACTT
MJDClo 7a *	R	TGCTCCTTAATCCAGGGAAATTTAG
MJDClo 1396	F	TCATGTTCGCTACCTTCACACT
MJDClo 2109*	F	GAGTTACTTTCCAGGTCTCGG
MJDClo 2129*	R	CCGAGACCTGGAAAGTAACTC
MJDClo 2552 *	F	GATCCAGCAGTCCCAATCATGTA
MJDClo 2646	R	TGCCTGGTCAGCTATAAGCA
MJDClo 2942*	F	TGGACACGGTGGCTTACGCCT
MJDClo 3695	R	GAGTTTTGCTCTTGTTGCCCAG
Clo R *	R	AGCCTTCTCTAACACCACCTTGG

*Primers designed in previous studies.

3. PCR Optimization

After optimization of PCRs with the primers described above, five singleplex were performed (Table 3) in order to amplify the 4 kb sequence of *ATXN3* (Figure 5). We performed two types of PCR: one including the CAG repeat, which resulted in the overrepresentation of normal over expanded alleles; and others not encompassing the repetitive CAG, which resulted in the equal amplification of both paternal and maternal inherited alleles. This way, we were able to determine the allelic phase of analysed SNPs and infer haplotypes segregating with the MJD expansion. All amplification reactions had a final volume of 10 µL, with the reagents described in Table 4, and PCR conditions in Table 3 and Table 5.

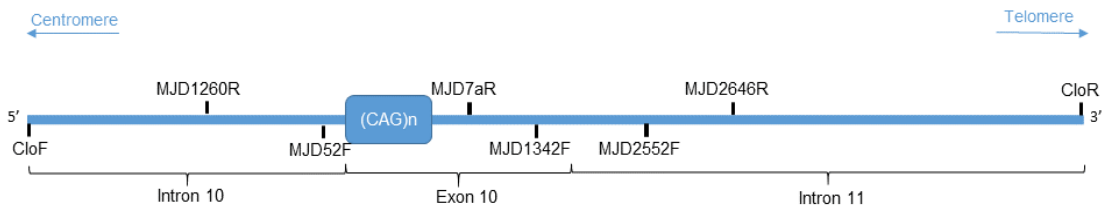


Figure 5: Schematic representation of the 4 kb *ATXN3* sequence and primers used to amplify the whole region.

Table 3: Haplotypic and genotypic PCRs, with temperatures of annealing (a) and times of extension (b).

Haplotypic PCR	**Length (bp)**	***a***	***b***	**Genotypic PCR**	**Length (bp)**	***a***	***b***
CloF-MJDClo7a	1537	60℃	2 min	CloF-MJDClo1260	1260	62°C	1 min 30 s
MJDClo52-MJDClo7a	399	62℃	1 min	MJDClo1342-MJDClo2646	1305	59℃	2 min
				MJDClo2552-CloR	1542	61℃	2 min

Table 4: PCR reagents.

Reagents	**Volume (µL)**
5x Q solution (QiaGen®)	1
Primer forward (2.5 µM)	0.5
Primer reverse (2.5 µM)	0.5
Water (ddH_2O)	1
DNA sample (7.5 ng/µL)	2
Taq polymerase (QiaGen Master Mix; QiaGen®)	5
Total	10

Table 5: PCR protocol for the amplification of different fragments within the *ATXN3* 4 kb region under analysis. a and b are the temperature of annealing and time of extension, respectively, indicated in Table 3. All PCRs had 35 cycles, except CloF-MJDClo1260 (33 cycles).

Initial denaturation	**Denaturation**	**Annealing**	**Extension**	**Final extension**
95℃	94℃	*a*	72℃	72℃
15 min	40 s	1 min 30 s	*b*	10 min
		35 cycles		

4. Detection of amplified products

In order to visualize the presence and the length of amplified DNA fragments, PCR products were subjected to electrophoresis in a polyacrylamide gel and coloration has been done with silver staining. Two glass supports previously cleaned and treated with blueslick (Serva), separated by a hydrophilic gel film (Gel-Fix™, Serva), were used to produce a thin gel composed by an acrylamide-bisacrylamide solution, ammonium persulfate and TEMED (Table 6).

Table 6: Reagents used in the polyacrylamide gel.

Reagents	**Volume**	**Function**
40% (w/v) Acrylamide:bisacrylamide (19:1) solution (AccuGel 19:1 ; National Diagnostics)	3 mL	Gel
Ammonium persulfate 2.5% (Promega)	170 µL	Oxidizing agent
TEMED (Amersco®)	7 µL	Catalysis agent

After gel polymerization was complete at room temperature, we placed the gel on an electrophoretic system (Multiphor II, GE Healthcare), previously refrigerated by a thermostatic circulator at 4ºC (MultiTemp III, Amersham Biosciences). We used 1.2 µL of each PCR product to load the wells. In order to infer the length of obtained fragments, we also loaded one well with a molecular marker (100 bp Plus; Thermo Scientific).

We used paper strips, soaked in buffer, which were placed at both anode (+) and cathode (-), to allow the horizontal run. The addition of bromophenol blue dye (Merck) to the cathode allowed the visualization and control of the electrophoresis.

The electrophoresis run initially at 180 V; after samples started to migrate, we increased the voltage to 220 V. The electrophoresis power supply (Consort EV243) was turned off when the marked dye approached the anode strip.

The coloration method involved six steps, neatly described in Table 7, all of them at room temperature and under agitation (Burgwedel).

Table 7: Steps of coloration with silver staining.

Step	**Time**	**Compounds**
(1) Fixation of the DNA	10 min	10% ethanol
	5 min	1% nitric acid (Merck)
(2) Washing	10 min, twice	Deionized water
(3) Coloration	20 min (protected from the light)	0.2% silver nitrate solution (Merck)
(4) Washing	10 s, twice	Deionized water
(5) Revelation of DNA fragments	Until bands start to appear	0.28 M sodium carbonate anhydrous (Applchem) and 0.02% formaldehyde (Merck)
(6) Stop the revelation	~10 s	10% acetic acid (Merck)

Ended the process, we left the gel in water, for at least 12 hours, and then dried at room temperature (RT).

5. DNA sequencing

After analysing the obtained fragments and confirming PCR specificity, DNA sequencing was performed in each PCR product with the respective sequencing primers. This method included two purifications and a PCR sequencing reaction. An initial purification, using an exonuclease and an enzyme thermosensitive alkaline phosphatase, was done in order to degrade unreacted primers and remained dNTPs (Table 8 and 9).

Depending on the PCR performed, sequencing primers were chosen as indicated in Table 10. Together with the purified product (2.8 µL), a sequencing mix (Table 11) was subjected to a sequencing reaction with the conditions described in Table 12.

Table 8: Reagents and quantities for PCR purification.

Compounds	**Volume (µL)**
PCR product	1.85
ExoI: FastAP (1:5) (Thermo Scientific)	0.95
Total	2.80

Table 9: Protocol for PCR purification

(1) To activate the enzyme	**(2) To inactivate the enzyme**
37°C	80°C
15 min	15 min

Table 10: Primers used to perform sequencing reaction.

PCR amplified	**Sequencing primers**	
CloF-MJDClo1260	MJDClo653	R
	MJDClo716	F
CloF-MJDClo7a	MJDClo653	R
	MJDClo716	F
	MJDClo7a	R
MJD52Clo-MJDClo7a	MJDClo7a	R
MJDClo1342-MJDClo2646	MJDClo1396	F
	MJDClo2109	F
	MJDClo2129	R
MJDClo2552-CloR	MJDClo2942	F
	CloR	R

Table 11: Reagents and volumes for sequencing mix

Compounds	**Volume (µL)**
Primer (2.5 µM)	0.5
Sequencing buffer (2.5x) (Applied Biosystems)	0.75
ABI Prims® BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems)	0.5
Water (ddH_2O)	0.75
Total	2.50

Table 12: Sequencing reaction conditions.

Initial denaturation	**Denaturation**	**Annealing**	**Extension**	**Final extension**
96℃	96 ℃	58ºC	60 ℃	60℃
2 min	15 s	5 min	2 min	10 min
		30 cycles		

We did an ultimate purification to isolate the sequencing products, by using a cross-linked dextran matrix in order to separate molecules over broad molecular weight. The centrifugation (Mikro 200, Hettich Zentrifugem) of illustra™ Sephadex™ G-50 (GE Healthcare) for 4 minutes at 4400 rpm allowed the formation of columns. After loading the PCR sequencing product in the center of the column, the same conditions of centrifugation resulted in a deposit containing the final product.

To guarantee the stability of single-stranded DNA, we added 12 µL of Hi-Di™ formamide (Applied Biosystems) to the final product. This stability is necessary for capillary electrophoresis in an ABI PRISM 2130x/Genetic Analyzer (Applied Biosystems).

6. In silico analyses

All obtained sequences were aligned with Geneious 5.5.8 software (Kearse *et al.* 2012), which allowed global alignments and comparison with the reference sequence to detect sequences variants.

PHASE 2.2 software (www.stat.washington.edu/stephens/software.html) was used to reconstruct haplotypes from genotypic data when the allelic phase of SNPs and expanded MJD alleles could not be directly inferred by family segregation or by allele-specific amplification. Allele frequencies and phase-known haplotypes were taken into account, but only haplotype pairs with a probability greater than 0.6 were used for further analyses.

Phylogenetic networks were performed using the Network 5.0.0.1 software (www.fluxus-engineering.com) to gain insight into the ancestral haplotypes and

evolutionary branching trees. Since we used microsatellite data, a combined reduced median and median-joining calculation was performed with the objective to reduce reticulation and simplify the network. To draw the phylogenetic networks we used seven STRs (TAT^{223}, GT^{199}, ATA^{194}, AC^{21}, $AAAC^{123}$, GT^{190} and AC^{190}) and the SNP rs56268847; the weights for each STR was calculated from the molecular diversity data calculated by Arlequin 3.5.2.2 software (Excoffier & Lischer 2010); also we attributed the maximum weight (10) for the only analysed SNP; in networks, circle area is proportional to frequency, branch length is proportional to the number of mutations and diamonds indicate recombination.

To assess more accurately the molecular distance among the different STR haplotypes of each lineage, a pairwise analysis of the flanking haplotypes was also performed with Arlequin, applying the sum of squared size difference (R_{st}) as the distance method. This way, the estimation of evolutionary distance between each pair of haplotypes was calculated taking into account the number of presumed single-step mutation steps between the corresponding STR alleles.

Results

During the optimization process, we tried to amplify our region of interest by using several enzymes – MyTaq™ Mix (Bioline); QiaGen Multiplex PCR Kit (QiaGen®); Advantage® Genomic LA Polymerase Mix (Clontech Laboratories, Inc.); DFS–Taq DNA Polymerase (BioRon); Ranger DNA Polymerase (Bioline); LongPCR Enzyme Mix (Thermo Scientific); and Advantage® GC Genomic LA Polymerase Mix (Clontech Laboratories, Inc.). Multiple amplification conditions have been applied, but we succeeded to amplify regularly and evenly both Portuguese and Taiwan DNA samples, when amplifying samples with the QiaGen enzyme.

Protocols optimized to cover a 4 kb region encompassing the MJD repeat allowed us to, more easily, infer haplotypes segregating with the $(CAG)_n$ expansion of all families studied, in the both Portuguese and Taiwanese populations. However, some PCRs, performed in some samples did not amplify the target sequence; in these cases, we adopted three strategies, following the order described below:

1. To decrease 2°C in the annealing temperature (Tm);

2. To increase DNA concentration (15 ng/µL), and;

3. To perform both procedures above, (1) and (2), at the same time.

Depending on the PCR type (either encompassing or not the CAG expansion) and PCR specificity (preferential amplification of the normal allele over the expanded one), we obtained a different length of amplified DNA fragments in a polyacrylamide gel (Figure 6).

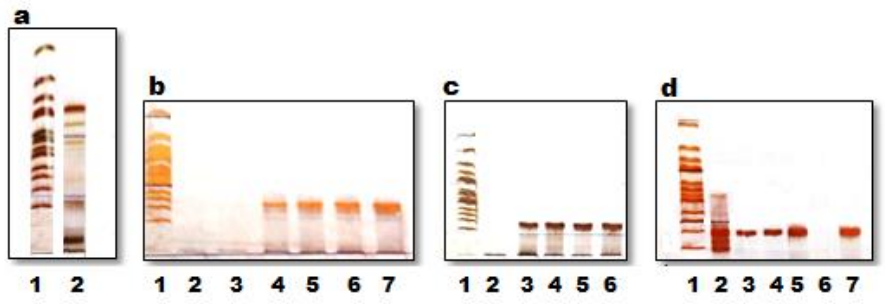


Figure 6: Polyacrylamide gel image showing the amplified products resulted from different PCRs. (a) Unspecific PCR, by using MJD52 and MJD7a primers; (b) Specific PCR obtain with CloF and MJD7a primers, in which it is possible to distinguish both normal and expanded alleles; (c) Specific PCR obtained with MJD1342 and MJD 2646 primers; (d) PCR specific for the amplification with primers MJD2552F and CloR (non-specific PCR for sample N12-6384 (well 2) probably due to excess of DNA).

Before using the Geneious software to align our sequences with the reference sequence, we interpreted results taking into account the PCR from which sequences had been originated: either $(CAG)_n$-biased PCR (with the normal allele more frequently represented over the expanded one) or regular PCR (with sequences from both homologous chromosomes equally amplified). When the PCR included the CAG repeat (CloF-MJD7a and MJD52-MJD7a), we first verified if both alleles (normal and expanded) were amplified (Figure 7, Figure 8 and Figure 9). If this was the case, we next identified the existence of slippage, as demonstrated in Figure 9, so that there was no misinterpretation of the downstream polymorphic positions.

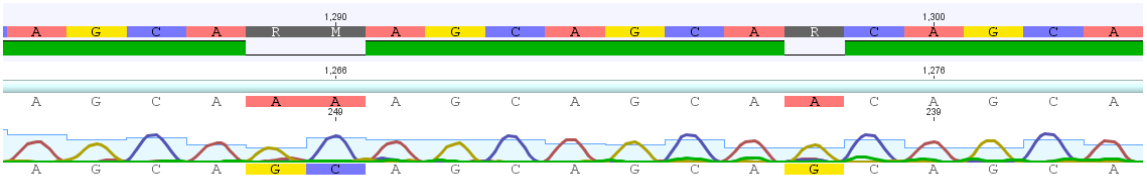


Figure 7: Electropherogram of a sequencing reaction encompassing the CAG repeat region in which only the expanded allele was amplified.

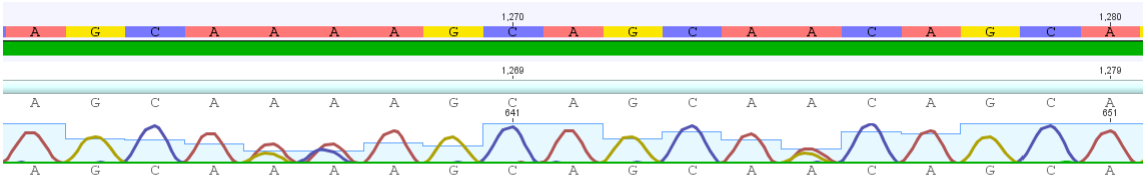


Figure 8: Electropherogram of a sequencing reaction encompassing the CAG repeat region, where both normal and expanded MJD alleles have been amplified.

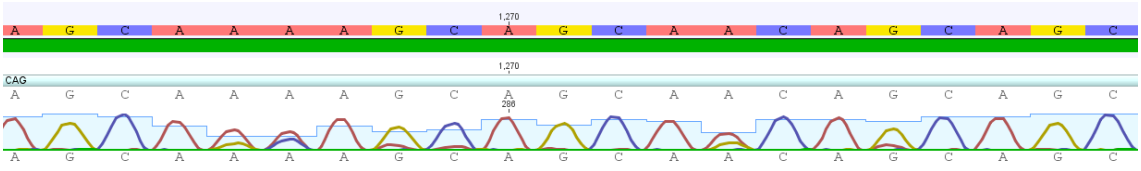


Figure 9: Electropherogram of a sequencing reaction encompassing the CAG repeat region with both normal and expanded MJD alleles amplified and in which slippage of the normal allele is also detected.

Therefore, we proceeded to the genotypic analysis of the polymorphic positions. When both $(CAG)_n$ alleles were amplified, we could face three situations for the SNP under study: i) the detection of a single base (Figure 10), with the allele segregating with the expansion directly inferred; ii) two equally represented peaks (Figure 11), which did not allow us to infer the allele in phase with MJD expansion; and, iii) two unbalanced peaks (Figure 12), with the lowest peak representing the SNP allele segregating with the expansion.

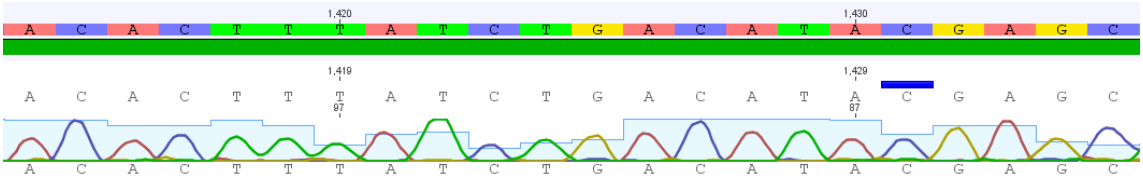


Figure 10: Sequence of a patient homozygous for the SNP rs7158733; amplification done with MJD-CloF and MJD7aR primers, and sequencing with MJD 7a primer.

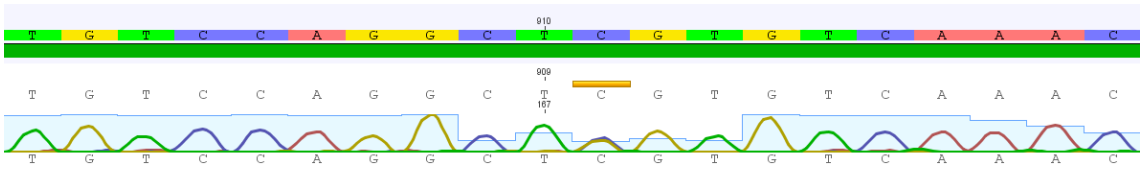


Figure 11: Sequence of a patient heterozygous for the SNP rs10467857; amplification done with MJD-CloF and MJD7aR primers, and sequencing with MJD 716 primer.

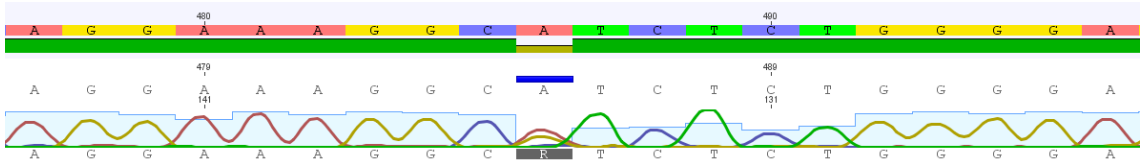


Figure 12: Sequence of a patient heterozygous for the SNP rs56268847; amplification with MJD-CloF and MJD7aR primers, and sequenced with MJD 653 primer.

In cases where the PCR did not encompass the repetitive MJD tract (CloF-MJD1260, MJD1342-MJD2646 and MJD2552-CloR), interpretation of results was straightforward, with clear homozygous and heterozygous individuals for the analysed SNPs, with single peaks or equally amplified and detected peaks, respectively.

Whenever allelic phases could not be inferred based on both $(CAG)_n$-biased or regular PCRs, we genotyped other family members to obtain the complete MJD SNP-based haplotype by family segregation. Our cohort of 43 families includes 65 MJD patients, of which 36 of Taiwanese and 29 of Portuguese origin. To complete haplotype inference by segregation, we also genotyped some non-affected relatives of different Taiwanese families. Genotypes for all studied individual are detailed in Supplementary material 1 and 2. For isolated cases or non-informative families, we resorted to the PHASE software to reconstruct haplotypes (Supplementary material 3 and 4).

We analysed all variable sites comprised within our region of interest, however, we focus on 23 SNPs that showed to be polymorphic in, at least, one of our analysed populations. With the exception of rs56268847 (A/G.485) and rs111735934 (G/A.3770), all analysed SNPs were polymorphic in both populations. As for

rs56268847 (A/G.485), all Portuguese individuals genotyped by us were homozygous for the A allele, in agreement with allele frequencies for the European control population (A/G: 1/0; Table 13). On the contrary, our results for SNP rs111735934 (G/A.3770) have shown a fairly high heterozygosity in our Portuguese cohort (53%), whereas reported value for the control European population was 0.002 (Table 13). Among studied polymorphic positions, we highlighted 19 SNPs that distinguish the Machado and Joseph lineages (underlined in Table 14), and three SNPs, rs10467857 (C/G.910), rs2006047 (A/G.3645) and rs7142326 (A/G.3912), that have shown a high polymorphic information content values (PICs of 67%, 68% and 89%, respectively; Table 13 and Table 14).

Our cohort of Portuguese samples (21 MJD families) was chosen based on their STR diversity (8 different STR haplotypes, previously assessed by Martins *et al.*, 2007 and Inês Costa, personal communication) since our aim was to study families more likely to have different ancestors. As for the Taiwanese families, taking into account their high STR diversity (13 STR-haplotypes found in 22 families), we included all 22 families in our SNP analysis.

Table 13: Heterozygosity of Portuguese and Taiwanese MJD patients for SNPs flanking the $(CAG)_n$ repeat at *ATXN3*. H|T: number of heterozygous patients by total number of analysed patients, with percentages for total (%T), Portuguese (%PT) and Taiwanese (%TW) MJD populations. European (EUR) and East Asian (EAS) control populations.

SNP	**refSNP ID**	**H|T**	**%T**	**%PT**	**%TW**	**Heterozygous genotype frequency(Yates *et al.* 2016)**		**Allele frequency(Yates *et al.* 2016)**			
						EUR	**EAS**	**EUR**		**EAS**	
C/T.124	rs12586535	29|55	53	42	61	0.316	0.466	C: 78%	T: 22%	C: 56%	T: 44%
C/T.248	rs12586471	24|50	50	33	66	0.316	0.468	C: 22%	T: 78%	C: 44%	T: 56%
A/G.485	rs56268847	11|58	19	0	32	0	0.103	A: 100%	G: 0%	A: 95%	G: 5%
G/A.868	rs10467858	28|54	52	42	60	0.316	0.468	G: 78%	A: 22%	G: 56%	A: 44%
C/G.910	rs10467857	37|55	67	76	60	0.398	0.476	G: 34%	C: 66%	G: 53%	C: 47%
T/C.921	rs10467856	27|54	50	38	60	0.316	0.468	T: 78%	C: 22%	T: 56%	C: 44%
C^{987}GG/G^{987}GG	rs12895357	30|55	55	46	56	0.414	0.470	G: 73%	C: 27%	G: 58%	C: 42%
TAA^{1118}/TAC^{1118}	rs7158733	30|56	54	44	61	0.316	0.468	C: 78%	A: 22%	C: 56%	A: 44%
C^{1178}/A^{1178}	rs3092822	28|57	49	42	55	0.316	0.468	A: 78%	C: 22%	A: 56%	C: 44%
G/A.1548	rs7158238	20|45	44	39	50	na*	na*	G: na*	A: na*	G: 47%	A: 53%
A/G.1694	rs12588287	26|52	50	42	58	0.316	0.466	A: 78%	G: 22%	A: 58%	G: 42%
G/A.1714	rs7153615	26|52	50	42	58	0.314	0.468	G: 78%	A: 22%	G: 56%	A: 44%
G/A.1737	rs7153603	26|52	50	42	58	0.314	0.468	G: 78%	A: 22%	G: 56%	A: 44%
C/T.1857	rs7153696	26|52	50	42	58	0.316	0.468	C: 78%	T: 22%	C: 56%	T: 44%
G/A.1886	rs7153374	26|52	50	42	58	0.316	0.468	G: 78%	A: 22%	G: 56%	A: 44%
G/T.1959	rs7153193	26|52	50	42	58	0.316	0.470	G: 78%	T: 22%	G: 56%	T: 44%
T/C.2992	rs4904833	13|25	52	43	56	0.316	0.468	T: 78%	C: 22%	T: 56%	C: 44%
C/T.3137	rs7146985	14|27	52	43	55	0.316	0.468	C: 78%	T: 22%	C: 56%	T: 44%
T/A.3209	rs113572439	14|31	45	27	55	0.316	0.468	T: 78%	A: 22%	T: 56%	A: 44%
A/G.3645	rs2006047	30|44	68	79	60	0.398	0.476	A: 66%	G: 34%	A: 47%	G: 53%
G/A.3738	rs8004149	21|43	49	39	56	0.316	0.468	G: 78%	A: 22%	G: 56%	A: 44%
G/A.3770	rs111735934	10|44	23	53	0	0.002	0	G: 99.9%	A: 0.1%	G: 100%	A: 0%
A/G.3912	rs7142326	25|38	89	19	58	0.398	0.476	A: 66%	G: 34%	A: 47%	G: 53%

*****na - not available

X – ancestral alleles are underlined

Table 14: SNP-based haplotypes of Portuguese and Taiwanese MJD families with the most frequently observed Joseph and Machado lineages.

SNP	refSNP ID	Distance from the $(CAG)_n$ (bp)	Location in *ATXN3*	Machado lineage	Joseph lineage		Joseph-derived lineage:
				PT	PT	TW	A/G.485
C/T.124	rs12586535	1249	Intron 10	C	T	T	T
C/T.248	rs12586471	1125	Intron 10	T	C	C	C
A/G.485	rs56268847	888	Intron 10	A	A	A	G
G/A.868	rs10467858	505	Intron 10	G	A	A	A
C/G.910	rs10467857	463	Intron 10	G	G	G	G
T/C.921	rs10467856	452	Intron 10	T	C	C	C
$(CAG)_{exp}$							
C^{987}GG/G^{987}GG	rs12895357	1	Exon 10	G	C	C	C
TAA^{1118}/TAC^{1118}	rs7158733	132	Exon 10	C	A	A	A
C^{1178}/A^{1178}	rs3092822	192	Exon 10	A	C	C	C
A/G.1548	rs7158238	250	Exon 10	G	A	A	A
A/G.1694	rs12588287	396	Intron 11	A	G	G	G
A/G.1714	rs7153615	416	Intron 11	G	A	A	A
A/G.1737	rs7153603	439	Intron 11	G	A	A	A
A/G.1857	rs7153696	559	Intron 11	C	T	T	T
G/A.1886	rs7153374	588	Intron 11	G	A	A	A
T/G.1959	rs7153193	661	Intron 11	G	T	T	T
T/C.2992	rs4904833	1694	Intron 11	T	C	C	C
C/T.3137	rs7146985	1839	Intron 11	C	T	T	T
T/A.3209	rs113572439	1911	Intron 11	T	A	A	A
A/G.3645	rs2006047	2347	Intron 11	G	G	G	G
G/A.3738	rs8004149	2440	Intron 11	G	A	A	A
G/A.3770	rs111735934	2472	Intron 11	A*	G	G	G
A/G.3912	rs7142326	2614	Intron 11	G	G	G	G

***(Martins *et al.* 2012)**

Of the 21 Portuguese MJD families analysed, 62% belong to Machado lineage, and the remaining 38% are from the Joseph background. All Taiwanese families (n=22) analysed by us shared most SNP alleles with the commonly observed Joseph lineage. In this population, we observed allelic states of SNP rs56268847 (A/G.485) in phase with the expansion to differ among families. Eight MJD Taiwanese families shared the Joseph-derived haplotype, i.e., Joseph background with a G variant in the SNP A/G.485 (Table 15).

Table 15: SNP-based lineages for the analysed MJD families.

Machado families		**PT**	P52; P57; P63; P67; P72; P83; P97; P110; P111; P114; N13; N14; N15
Joseph families		**PT**	P10; P91; P96; P99; P112; P117; N12; N19
		TW	S2; S3; S7; S8; S9; S10; S13; S15; S19; S20; S21; S22; S23; S25
Joseph-derived	**A/G.485**	**TW**	S1; S4; S5; S6; S12; S17; S18; S24

A phylogenetic network was performed with all studied Taiwanese families based on their previously assessed STR haplotype in order to find phylogenetic distances between Joseph and Joseph-derived families. We can notice four evolutionary branches evolving from a possibly common ancestor. A clear separation is noticed for Joseph and Joseph-derived lineages, with at least 6 mutation steps apart between them, at the following polymorphic markers: GT^{199}, TAT^{223}, AC^{190} and AC^{21} (H10-H12); Unexpectedly, the H7 haplotype seems to be closer to Joseph-derive families, even if this family shares all SNPs analysed by us with the most common Joseph lineage (patient 3.S9 is homozygous A/A for the SNP A/G.485; Figure 13; Supplementary material 5). However, if we compared the haplotypes H6 and H7 we can count 4 mutations steps between them, while the H23 and H7 have 5 mutations steps. The reason why the software draw a network including the haplotypes H7 and H23 in the same branch, may be justified because of the maximum parsimonious principle on which it is based this software and because the weight that we gave at each marker analyse. In pairwise H6-H7 the mutation occur at GT^{190} marker which have higher weight (5) than the GT^{199} marker (4), where occur 4 of 5 step mutations that distinguish the haplotypes in pairwise H7-H23.

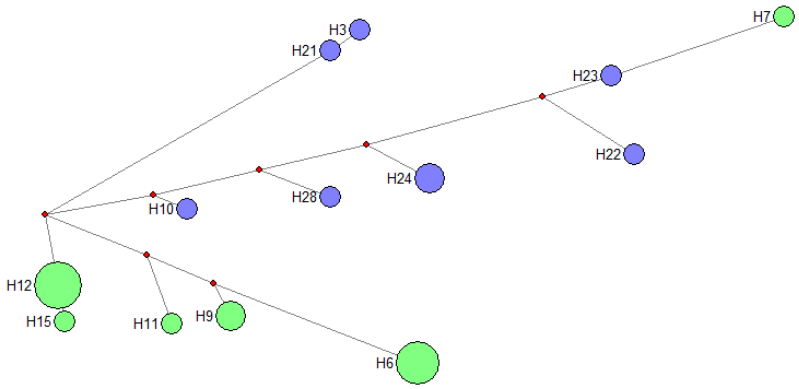


Figure 13: Phylogenetic network showing the most parsimonious relationships among STR-based haplotype in Taiwanese families belonging to Joseph (green; n=15) and Joseph-derived SNP backgrounds (blue; n=8).

In order to increase the specificity of relationships among haplotypes we included the A/G.485 SNP marker in the analysis to draw a new phylogenetic network (Figure 14). Thus, we obtained a new network with more step mutations (7) separating both Joseph and Joseph-derived haplotypes and with an older common ancestral haplotype; the H7 haplotype remained, however, closer to Joseph-derived than to Joseph backgrounds.

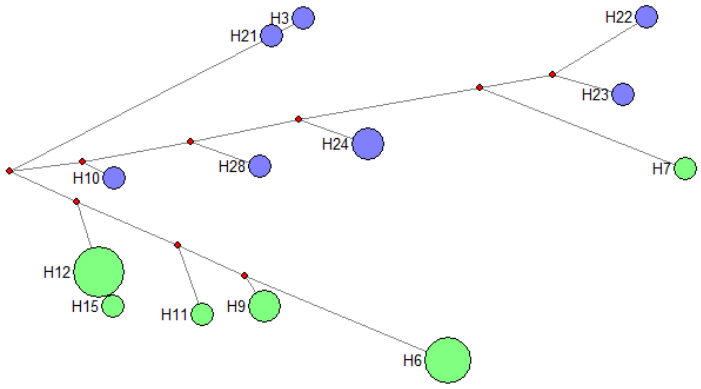


Figure 14: Phylogenetic network showing the most parsimonious relationships among STR haplotypes and including also the SNP A/G.485. Different colours denote Taiwanese Joseph (green) and Joseph-derived (blue) families.

Next, in order to picture a broader scenario, we draw a phylogenetic networks with STR haplotypes from all Asian families available in the lab, namely from Taiwanese (n=26), Japanese (n=17), Indian (n=3), Australians (aborigines; n=2), Chinese (n=1) and Cambodia (n=1). We obtained an unrooted phylogenetic network which makes difficult to infer a common ancestor. By analysing this network, we observer, however, the H7 haplotype from the Joseph lineage, sharing a common ancestor with haplotype H1.

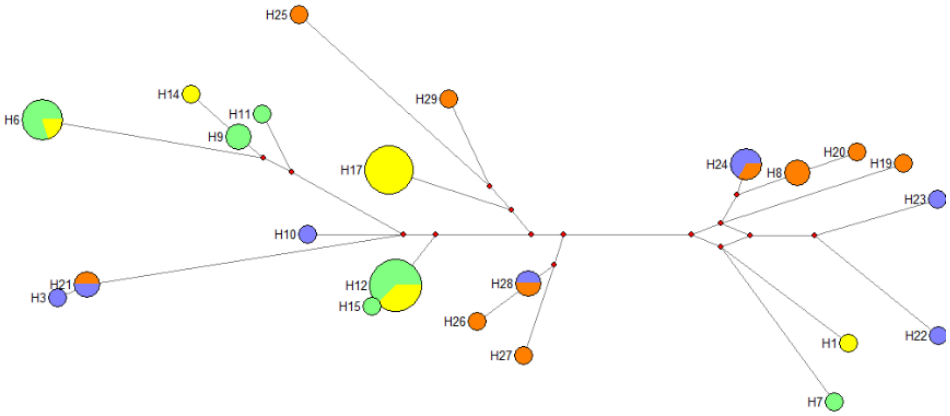


Figure 15: Phylogenetic network showing the most parsimonious relationships among STR-based haplotypes of Joseph Asian families. Joseph families from Taiwan and other Asian populations are coloured in green and yellow, respectively. Joseph-derived families from Taiwan and other Asian populations are coloured in blue and orange, respectively.

After analysing carefully the previous network, we noticed that the source of genetic distance came mostly from the STR TAT^{223}. Taking into account this is our most distant marker from the *ATXN3*-(CAG)$_n$, we raised the hypothesis that recombination could be increasing unaccounted diversity in the obtained network, thus hampering the inference of the haplotype evolutionary direction. For this reason, we chose to remove the first marker (TAT^{223}) from the STR-based haplotype to next draw a new phylogenetic network, reducing the effect of recombination (Figure 16).

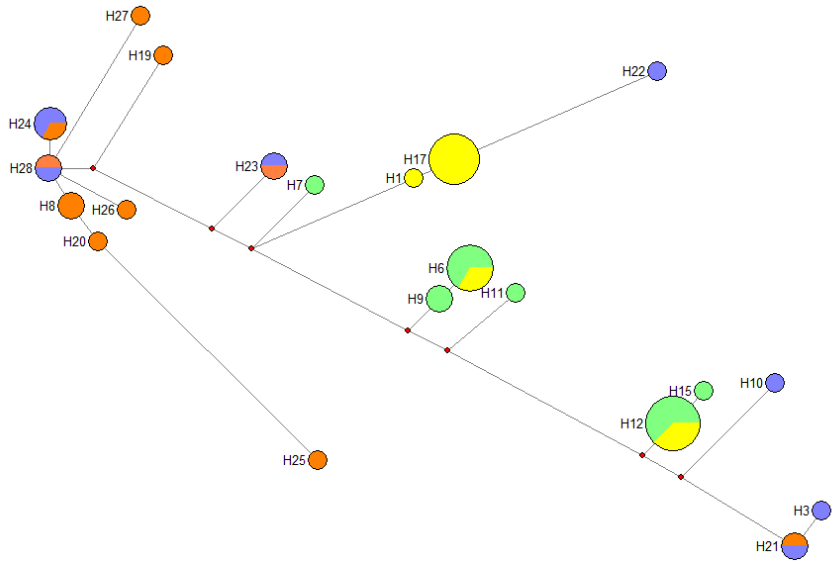


Figure 16: Phylogenetic network showing the most parsimonious relationships among STR-based haplotype without TAT^{223} marker. Joseph families from Taiwan and other Asian populations are coloured in green and yellow, respectively. Joseph-derived families from Taiwan and other Asian populations are coloured in blue and orange, respectively. The H29 and H14 haplotypes was included in same circle of H23 and H6 haplotype, respectively.

In fact, after removing the most distant marker from our STR-haplotype analysis, the interpretation of the network became clearer. The ancestral haplotype in Asian MJD populations seems to be H28, from Joseph-derived lineage (Figure 16; Supplementary material 5). Consequently, this lineage also displays more haplotype diversity than the Joseph lineage. Again, to obtain more accurate genetic distances among MJD backgrounds, we followed the same strategy as before, by including the SNP A/G.485 in network analysis. Results shown in Figure 17 suggested that the origin of Joseph lineage is posterior to the origin of the Joseph-derived lineage in Asian populations. This led us to question the ancestral background in which a *de novo* expansion occurred in ATXN3 and the type of recurrent mutation at A/G.485: A>G or G>A.

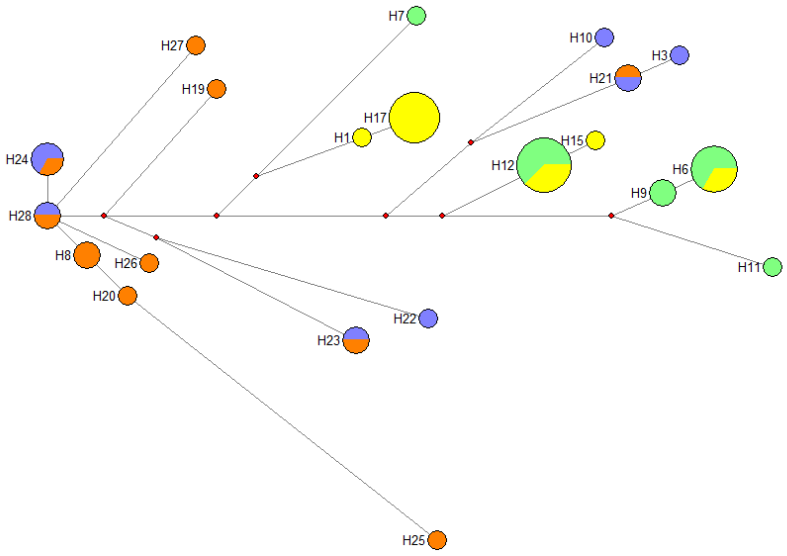


Figure 17: Phylogenetic network showing the most parsimonious relationships among STR haplotypes and including also the SNP A/G.485, without TAT^{223} marker. Joseph families from Taiwan and other Asian populations are coloured in green and yellow, respectively. Joseph-derived families from Taiwan and other Asian populations are coloured in blue and orange, respectively. The H29 and H14 haplotypes was included in same circle of H23 and H6 haplotype, respectively.

Taking account the obtained results, we aimed to know how Asian families are related between them, separately, in both Joseph and Joseph derived lineages. Thus, we draw two phylogenetic networks with STR haplotypes, without the TAT^{223} marker and without the SNP A/G.485 (Figure 18 and Figure 19).

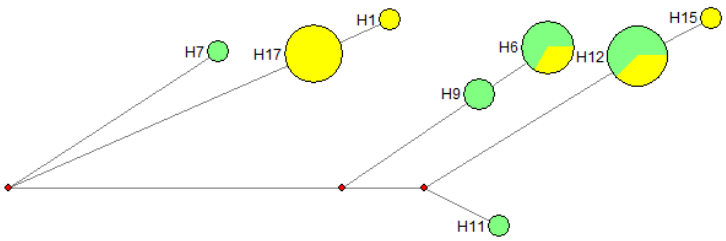


Figure 18: Phylogenetic network showing the most parsimonious relationships among STR haplotypes. Joseph families from Taiwan and other Asian populations are coloured in green and yellow, respectively. The haplotype H14 was included in same circle of haplotype H6.

All phylogenetic networks were draw, as already referred above, based in the maximus parsimonious principle. Thus, the reticulation obtained in the new network with STR haplotypes from all Joseph-derived Asian families (Figure 19) may be justified due the few haplotypes analysed when compared the high number of haplotypes existent, once in previous network (Figure 16 and Figure 17) do not exist reticulations between Joseph-derived haplotypes. We obtained an unrooted phylogenetic network, however, the haplotype H28 seems to be the older common ancestral with the high number of haplotypes directly related.

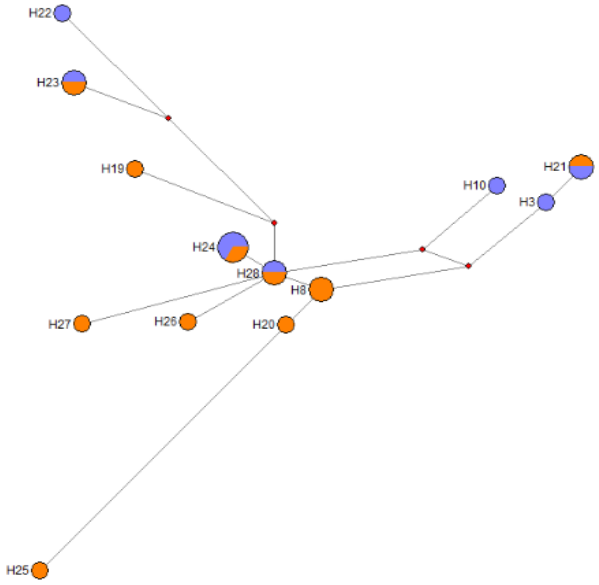


Figure 19: Phylogenetic network showing the most parsimonious relationships among STR haplotypes. Joseph-derived families from Taiwan and other Asian populations are coloured in blue and orange, respectively. The haplotype H29 was included in same circle of haplotype H23.

Finally, we assessed the genetic diversity among these MJD families by using the Arlequin software: Asian Joseph families are less diverse (0.83±0.04) while those carrying the G variant in the A/G.485 (0.96±0.03). Also, we used the Arlequin software to measure the genetic distance among all Joseph Portuguese and Asian families, and Joseph-derived families (Table 16). A higher genetic difference has been found between Portuguese and Joseph-derived families (0.67) than any other pairwise analysis, whereas differences were observed between Asian and Joseph-derived families (0.24).

Table 16: Pairwise genetic distances between Asian Joseph-derived families (n=19), Asian Joseph families (n=27) and Portuguese Joseph families (n=40).

Families	Asian Joseph-derived	Portuguese Joseph	Asian Joseph
Asian Joseph-derived	0.00000		
Portuguese Joseph	0.66983	0.00000	
Asian Joseph	0.23657	0.26580	0.00000

Discussion

Genetic variation and molecular evolution can often be explained by molecular population genetics, which is based on population genetics principles (Casillas & Barbadilla 2017).

The mutational origin of Machado-Joseph disease was initially suggested to be Portuguese mainly due to first descriptions among Azorean/Portuguese emigrants living in the United States and also to the current high number of MJD families diagnosed in our population. Following this scenario, it has been proposed that the original event occurred in mainland Portugal, spread to the Azores during its colonization, and from there to North and South of America, while the Portuguese seaman could explain its presence in Asia (Sequeiros & Coutinho 1993). Later, after the identification of the causative expansion in *ataxin-3*, molecular studies and population genetics principles have not supported this first idea. More accurately is a strategy based on the assumption that populations, places-of-birth for a given mutation, would have accumulated more molecular diversity on flanking regions of the respective disease-causing locus of interest than those where the mutation was later introduced. Therefore, based on this principle, a high disease frequency in a given population does not give us the place-of-birth for a mutation. Following this principle, an Asian mutational origin has been proposed for the most frequent and worldwide spread Joseph MJD lineage. A further analysis performed with STRs-haplotypes under the light of this assumption gave insight into the main spreading routes of Joseph lineage (Martins *et al.* 2007).

As concerns Machado MJD lineage, the place-of-birth for this mutation is more controversial and still unclear; however, recently, our research group suggested a possible Portuguese origin for this lineage. Among a total of 69 Machado families studied (Supplementary material 6 and 7), 59 of Portuguese ancestry shared very similar STR-haplotypes. The remaining 10 non-Portuguese families (5 Ibero-American, 4 North-American and one of unknown origin) showed a greater STR-haplotype diversity[*]. Following the same strategy based used for Joseph lineage, we suggested here an Iberian origin for the Machado MJD lineage. Moreover, based on the geographical distribution of families carrying this SNP background, we suggested that the principal route of dispersal for this mutation has been from the Iberian Peninsula to North and South America.

[*] Inês Costa, personal communication

When studying repeat-expansion disorders, the high intergenerational instability of their mutated alleles adds an extra difficulty to distinguish alleles identical-by-descent (derived from a common ancestor) from alleles identical-by-state (resulted from independent origin events). For this reason, to trace the history of a given disease-associated mutation, the analysis of haplotype backgrounds associated to pathogenic alleles is not only useful, but an essential tool (Martins 2007). By assessing allelic phases of analysed flanking polymorphic markers and the expansion, one can avoid the overlapping of normal alleles' evolutionary history and our expanded locus of interest. The inference of haplotypes segregating with the expanded allele are also of utmost importance to study the molecular pathway underlying repeat instability.

Igarashi *et al*., in 1996, studied the intergenerational instability of the $(CAG)_n$-repeat in MJD. They suggested, based on study of SNP $C^{987}GG/G^{987}GG$, that an inter-allelic interaction between normal and expanded alleles should affecting the instability at the *ataxin-3* locus, involving probably the occurrence of gene conversion (Igarashi *et al.* 1996). Almost ten years later, gene conversion was proposed by Mittal *et al.* as the mechanism involved in the origin of a rare intermediate allele of MJD. The authors observed a $(CAG)_{45}$ allele within the Joseph haplotype, carrying nonetheless variant at the sixth repeat (CAG instead of CAA), only observed in smaller alleles that were significantly associated with a different SNP background: the Machado MJD lineage (Mittal *et al.* 2005). Gene conversion has also been proposed to explain *de novo* generation of expanded alleles in other triplet-associated disorders, as in SCA17 (Tomiuk *et al.* 2007). Also, in 2008, Martins *et al.* studied the influence of the MJD lineage on integrational instability and they raised the hypothesis of an effect of the haplotype background on *ATXN3*-$(CAG)_n$ instability in the paternal lineage. They observed that TTACAC (Joseph) lineage usually had a tendency for CAG expansion, whereas the GTGGCA (Machado) lineage was more frequently associated with contractions (Martins *et al.* 2008). In summary, this highlights the importance of studying the haplotype backgrounds associated to expanded alleles to search for new answers on different areas, such as: molecular mechanism of expansion, mutational origin, and disease pleomorphism.

In this study we analysed a 4 kb region encompassing the CAG repeat of *ataxin-3,* and genotyped 19 polymorphic positions which allowed to enlarge the haplotypes backgrounds analysed in our MJD patients:

Machado lineage: CTAGGT $(CAG)_{exp}$ GCAGAGGCGGTCTGGAG;

Joseph lineage: TCAAGC $(CAG)_{exp}$ CACAGAATATCTAGAGG;

Joseph derived lineage: TCGAGC $(CAG)_{exp}$ CACAGAATATCTAGAGG.

Based on the high haplotype diversity observed in Asian populations, we formulated three hypotheses:

(1) The Joseph-derived lineage (Joseph lineage with a G variant at SNP A/G.485) was originated by a recurrent mutation (485. A>G) on the assumed common Joseph haplotype background shortly after the mutational origin. This could justify the high haplotype diversity of this Joseph-derived lineage (0.9649±0.0276). This is in accordance with the reported low frequency of G allele (5%) on East Asian control populations;

(2) The haplotype background with a G variant at SNP A/G.485 is more ancient that the Joseph lineage, thus being the most ancestral origin for MJD. In this case, the common Joseph lineage would have been originated by a recurrent mutation in SNP A/G.485 SNP (G>A), also shortly after the mutational origin. This would also explain (i) the high diversity of (what we named) Joseph-derived lineage, (ii) STR haplotypes from these lineages spread throughout the different branches of the networks, and (iii) the lower genetic diversity of Joseph lineage in comparison to Joseph-derived;

(3) Both lineages underlie two different mutational origins for MJD in Asian populations. This hypothesis justifies high genetic diversity observed in both lineages, with Joseph worldwide spread and common due to genetic drift and founder effects.

SNP genotyping data we obtained from the study of MJD origins can, in addition, be useful to find the best strategy to distinguish normal and expanded alleles. This strategy may be useful in the context of MJD therapeutics. To date, no effective treatment exists for ataxia. Usually, the bases of treatment of degenerative cerebellar ataxia are physiotherapy, occupational therapy, and speech therapy, but the level of success is variable for each intervention (D'Abreu & Friedman 2016). Taking into account the pathogenesis of MJD, considered to be caused by “gain of toxic function” of the expanded protein, a most effective and simple gene therapeutic approach for MJD requires the reduction of the mutant protein. Furthermore, given the important role of wild-type ATXN3 towards cell survival, the reduction of mutant ataxin-3 must be allele-specific, leaving wild-type protein intact. Thus, the first gene therapy suggested for MJD by Yi Li *et al*. in 2004, was the use of small interference RNA (siRNA). The advantage of the presence of a C variant in $\underline{C}^{987}GG/\underline{G}^{987}GG$ SNP in linkage disequilibrium with the disease-causing CAG expansion (Miller *et al.* 2003; Li *et al.* 2004) would, however, benefit MJD patients belonging to Joseph lineage. Therefore, in order to find the best SNPs to distinguish normal and expanded in the highest possible

number of MJD patients for down-regulation of expanded alleles during siRNA therapies, we started by searching for a polymorphic positions in which all MJD patients shared the same allele status in cis with the expansion. Taking this into account, we highlighted the SNPs rs10467857 (C/G.910). For this polymorphic position fifty-five MJD patients were genotyped, and in all heterozygous patients (thirty-seven), the G variant was always associated with the expanded allele. Furthermore, we did not find any patient that be homozygous for allele C (i.e., none MJD patients from our cohorts had the genotype C/C at C/G.910). We performed this same analysis in two other polymorphic positions flanking the MJD repeat: rs2006047 (A/G.3645) and rs7142326 (A/G.3912). In any of our Taiwanese or Portuguese patients, we found the A/A for A/G.3645 or A/G.3912 (Supplementary material 1 and 2); simultaneously, by analysing our PHASE results (Supplementary material 3 and 4), the G allele of both SNPs was always associated with expanded allele. Also, in a previous study, it is described that none of these SNPs distinguish Machado and Joseph lineages, with the G nucleotide always observed to segregate with the expanded alleles (Martins *et al.* 2012)[†]. By analysing all 19 described alternative *ATXN3* transcripts, however, none of these 3 SNPs (rs10467857, rs2006047 and rs71423226) is located within transcribed regions. This way, our strategy should be further applied to search for other $(CAG)_n$ flanking variants known to be transcribed in humans.

[†] Sandra Martins, personal communication

Conclusions and Future perspectives

Under a haplotype-based approach, this study focused on four specific goals that were mostly achieved:

(1) we optimized 5 PCRs that allowed the inference of allelic phases of expanded MJD alleles;

(2) we identified 19 SNPs flanking the $(CAG)_n$-*ATXN3* that distinguish the two main MJD lineages identified so far: Machado and Joseph;

(3) we suggested 3 SNPs to distinguish normal and expanded alleles in MJD patients from different lineages; and

(4) we raised the hypothesis of another *de novo* mutational origin for Machado-Joseph disease to explain the high genetic diversity in a third observed MJD SNP-background.

Next, to emphasise and clarify the obtained results and to validate our findings, future work may include:

(i) the increasement of the number of Asian samples as well the number of populations to analyse in order to find new MJD lineages;

(ii) the optimization of a single 4 kb PCR with allele-specific on different SNPs to achieve more accurately the allelic phases of expanded MJD alleles, and;

(iii) the increasement of the number of haplotypes obtained by familiar segregation, diminishing, consequently, the need of inferring allelic phases through informatics tools. Thus, it would be necessary to assess a higher number of relatives in cases of isolated MJD patients, to verify the existence of possibly informative non-affected relatives.

References

Alonso E., Martínez-Ruano L., De Biase I., Mader C., Ochoa A., Yescas P., Gutiérrez R., White M., Ruano L., Fragoso-Benítez M., Ashizawa T., Bidichandani S.I. & Rasmussen A. (2007) Distinct distribution of autosomal dominant spinocerebellar ataxia in the Mexican population. *Movement Disorders* **22**, 1050-3.

Barrandeguy M.E. & García M.V. (2014) Quantifying genetic diversity: the starting point for population genetic studies using molecular markers. *Journal of genetics* **93**, 587-9.

Bettencourt C. & Lima M. (2011) Machado-Joseph Disease: from first descriptions to new perspectives. *Orphanet J Rare Dis* **6**, 35.

Bettencourt C., Santos C., Montiel R., Costa M.d.C., Cruz-Morales P., Santos L.R., Simões N., Kay T., Vasconcelos J., Maciel P. & Lima M. (2010) Increased transcript diversity: novel splicing variants of Machado–Joseph Disease gene (ATXN3). *neurogenetics* **11**, 193-202.

Bhargava A. & Fuentes F.F. (2010) Mutational dynamics of microsatellites. *Mol Biotechnol* **44**, 250-66.

Bird (1998) Hereditary Ataxia Overview URL https://www.ncbi.nlm.nih.gov/books/NBK1138/.

Brookes A.J. (1999) The essence of SNPs. *Gene* **234**, 177-86.

Brumfield R.T., Beerli P., Nickerson D.A. & Edwards S.V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18**, 249-56.

Brusco A., Gellera C., Cagnoli C. & et al. (2004) Molecular genetics of hereditary spinocerebellar ataxia: Mutation analysis of spinocerebellar ataxia genes and cag/ctg repeat expansion detection in 225 italian families. *Arch Neurol* **61**, 727-33.

Bryer A., Krause A., Bill P., Davids V., Bryant D., Butler J., Heckmann J., Ramesar R. & Greenberg J. (2003) The hereditary adult-onset ataxias in South Africa. *Journal of the Neurological Sciences* **216**, 47-54.

Budworth H. & McMurray C.T. (2013) A brief history of triplet repeat diseases. *Methods Mol Biol* **1010**, 3-17.

Carvalho D.R., La Rocque-Ferreira A., Rizzo I.M., Imamura E.U. & Speck-Martins C.E. (2008) Homozygosity Enhances Severity in Spinocerebellar Ataxia Type 3. *Pediatric Neurology* **38**, 296-9.

Casillas S. & Barbadilla A. (2017) Molecular population genetics. *Genetics* **205**, 1003-35.

Chai Y. (2002) Live-cell imaging reveals divergent intracellular dynamics of. **99**, 9310-5.

Costa M.d.C. & Paulson H.L. (2012) Toward understanding Machado–Joseph disease. *Progress in Neurobiology* **97**, 239-57.

Coutinho P. & Andrade C. (1978) Autosomal dominant system degeneration in Portuguese families of the Azores Islands A new genetic disorder involving cerebellar, pyramidal, extrapyramidal and spinal cord motor functions. *Neurology* **28**, 703-.

D'Abreu A. & Friedman J.H. (2016) Spinocerebellar ataxia type 3. *Neurology MedLink*.

Eckert K.A. & Hile S.E. (2009) Every Microsatellite is Different: Intrinsic DNA Features Dictate Mutagenesis of Common Microsatellites Present in the Human Genome. *Molecular carcinogenesis* **48**, 379-88.

Egan R.A., Camicioli R. & Popovich B.W. (2000) A small 55-repeat MJD1 CAG allele in a patient with Machado-Joseph disease and abnormal eye movements. *European neurology* **44**, 189-90.

Emmel V.E., Alonso I., Jardim L.B., Saraiva-Pereira M.L. & Sequeiros J. (2011) Does DNA methylation in the promoter region of the ATXN3 gene modify age at onset in MJD (SCA3) patients? *Clinical Genetics* **79**, 100-2.

Excoffier L. & Lischer H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* **10**, 564-7.

Fan H. & Chu J.-Y. (2007) A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics & Bioinformatics* **5**, 7-14.

Gaspar C., Lopes-Cendes I., Hayes S., Goto J., Arvidsson K., Dias A., Silveira I., Maciel P., Coutinho P., Lima M., Zhou Y.X., Soong B.W., Watanabe M., Giunti P., Stevanin G., Riess O., Sasaki H., Hsieh M., Nicholson G.A., Brunt E., Higgins J.J., Lauritzen M., Tranebjaerg L., Volpini V., Wood N., Ranum L., Tsuji S., Brice A., Sequeiros J. & Rouleau G.A. (2001) Ancestral Origins of the Machado-Joseph Disease Mutation: A Worldwide Haplotype Study. *American Journal of Human Genetics* **68**, 523-8.

Gomes T., Guimaraes J. & Leão M. (2017) Investigation of Genetic Aetiology in Neurodegenerative Ataxias: Recommendations from the Group of Neurogenetics of Centro Hospitalar São João, Portugal. *Acta Médica Portuguesa* **30**, 502-12.

Goodwin W., Linacre A. & Hadi S. (2011) *An Introduction to Forensic Genetics*. Wiley.

Grover A. & Sharma P.C. (2014) Development and use of molecular markers: past and present. *Crit Rev Biotechnol* **36**, 290-302.

Gu W., Ma H., Wang K., Jin M., Zhou Y., Liu X., Wang G. & Shen Y. (2004) The shortest expanded allele of the MJD1 gene in a Chinese MJD kindred with autonomic dysfunction. *European neurology* **52**, 107-11.

Guyenet S.J. & La Spada A.R. (2006) Triplet Repeat Diseases. In: *Reviews in Cell Biology and Molecular Medicine* (Wiley-VCH Verlag GmbH & Co. KGaA.

Ichikawa Y., Goto J., Hattori M., Toyoda A., Ishii K., Jeong S.-Y., Hashida H., Masuda N., Ogata K., Kasai F., Hirai M., Maciel P., Rouleau G.A., Sakaki Y. & Kanazawa I. (2001) The genomic structure and expression of MJD, the Machado-Joseph disease gene. *J Hum Genet* **46**, 413-22.

Igarashi S., Takiyama Y., Cancel G., Rogaeva E.A., Sasaki H., Wakisaka A., Zhou Y.X., Takano H., Endo K., Sanpei K., Oyake M., Tanaka H., Stevanin G., Abbas N., Dürr A., Rogaev E.I., Sherrington R., Tsuda T., Ikeda M., Cassa E., Nishizawa M., Benomar A., Julien J., Weissenbach J., Wang G.X., Agid Y., St. George-Hyslop P.H., Brice A. & Tsuji S. (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. *Human Molecular Genetics* **5**, 923-32.

Jardim L.B., Silveira I., Pereira M.L., Ferro A., Alonso I., do Céu Moreira M., Mendonça P., Ferreirinha F., Sequeiros J. & Giugliani R. (2001) A survey of spinocerebellar ataxia in South Brazil–66 new cases with Machado-Joseph disease, SCA7, SCA8, or unidentified disease–causing mutations. *Journal of Neurology* **248**, 870-6.

Jiang H., Tang B.-s., Xu B., Zhao G.-h., Shen L., Tang J.-g., Li Q.-h. & Xia K. (2005) Frequency analysis of autosomal dominant spinocerebellar ataxias in mainland Chinese patients and clinical and molecular characterization of spinocerebellar ataxia type 6. *Chinese medical journal* **118**, 837-43.

Jobling M.A., Kivisild T. & Tyler-Smith C. (2014) *Human Evolutionary Genetics*. Garland Science.

Johnson M., Zaretskaya I., Raytselis Y., Merezhuk Y., McGinnis S. & Madden T.L. (2008) NCBI BLAST: a better web interface. *Nucleic acids research* **36**, W5-W9.

Kang N.A.R.a.J.T.L. (2015) Genetic Diversity and Societally Important Disparities.

Kawaguchi Y., Okamoto T., Taniwaki M., Aizawa M., Inoue M., Katayama S., Kawakami H., Nakamura S., Nishimura M., Akiguchi I., Kimura J., Narumiya S.

& Kakizuka A. (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet* **8**, 221-8.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P. & Drummond A. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-9.

Kent W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656-64.

Kibbe W.A. (2007) OligoCalc: an online oligonucleotide properties calculator. *Nucleic acids research* **35**, W43-W6.

Kouniaki D.I., Papasteriades C. & Tsirogianni A. (2015) Short Tandem Repeats Loci in Parentage Testing. *2015* **10**, 8.

Kraft S., Furtado S., Ranawaya R., Parboosingh J., Bleoo S., McElligott K., Bridge P., Spacey S., Das S. & Suchowersky O. (2005) Adult onset spinocerebellar ataxia in a Canadian movement disorders clinic. *Canadian journal of neurological sciences* **32**, 450.

Krishna N., Mohan S., Yashavantha B., Rammurthy A., Kumar H.K., Mittal U., Tyagi S., Mukerji M., Jain S. & Pal P.K. (2007) SCA 1, SCA 2 & SCA 3/MJD mutations in ataxia syndromes in southern India. *Indian Journal of Medical Research* **126**, 465.

La Spada A.R. & Taylor J.P. (2010) Repeat expansion disease: Progress and puzzles in disease pathogenesis. *Nat Rev Genet* **11**, 247-58.

Landsteiner K. (1900) Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutseruns und der Lymphe. *Z. Bakteriol.* **27**, 357-62.

Li Y., Yokota T., Matsumura R., Taira K. & Mizusawa H. (2004) Sequence-dependent and independent inhibition specific for mutant ataxin-3 by small interfering RNA. *Annals of Neurology* **56**, 124-9.

Long Z., Chen Z., Wang C., Huang F., Peng H., Hou X., Ding D., Ye W., Wang J. & Pan Q. (2015) Two Novel SNPs in ATXN3 3'UTR May Decrease Age at Onset of SCA3/MJD in Chinese Patients. *PloS one* **10**, e0117488.

Lu Y.-F., Goldstein D.B., Angrist M. & Cavalleri G. (2014) Personalized Medicine and Human Genetic Diversity. *Cold Spring Harb Perspect Med* **4**, a008581.

Maciel P., do Carmo Costa M., Ferro A., Rousseau M., Santos C.S., Gaspar C., Barros J., Rouleau G.A., Coutinho P. & Sequeiros J. (2001) Improvement in the molecular diagnosis of Machado-Joseph disease. *Arch Neurol* **58**, 1821-7.

Martins S. (2007) Evolutionary and Epidemiological Genetics of Machado-Joseph Disease. In: *Biology*. Porto University.

Martins S., Calafell F., Gaspar C. & et al. (2007) Asian origin for the worldwide-spread mutational event in machado-joseph disease. *Arch Neurol* **64**, 1502-8.

Martins S., Coutinho P., Silveira I., Giunti P., Jardim L.B., Calafell F., Sequeiros J. & Amorim A. (2008) Cis-acting factors promoting the CAG intergenerational instability in Machado-Joseph disease. *Am J Med Genet B Neuropsychiatr Genet* **147B**, 439-46.

Martins S., Pearson C.E., Coutinho P., Provost S., Amorim A., Dubé M.-P., Sequeiros J. & 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**, 1311-8.

Martins S., Soong B., Wong V.N. & et al. (2012) Mutational origin of machado-joseph disease in the australian aboriginal communities of groote eylandt and yirrkala. *Arch Neurol* **69**, 746-51.

Matos C.A., de Macedo-Ribeiro S. & Carvalho A.L. (2011) Polyglutamine diseases: The special case of ataxin-3 and Machado–Joseph disease. *Progress in Neurobiology* **95**, 26-48.

Matos C.A., Nóbrega C., Louros S.R., Almeida B., Ferreiro E., Valero J., Pereira de Almeida L., Macedo-Ribeiro S. & Carvalho A.L. (2016) Ataxin-3 phosphorylation decreases neuronal defects in spinocerebellar ataxia type 3 models. *The Journal of Cell Biology* **212**, 465-80.

McMurray C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet* **11**, 786-99.

Miller V.M., Xia H., Marrs G.L., Gouvion C.M., Lee G., Davidson B.L. & Paulson H.L. (2003) Allele-specific silencing of dominant disease genes. *Proceedings of the National Academy of Sciences* **100**, 7195-200.

Mittal U., Srivastava A.K., Jain S., Jain S. & Mukerji M. (2005) Founder haplotype for machado-joseph disease in the indian population: Novel insights from history and polymorphism studies. *Arch Neurol* **62**, 637-40.

Morin P.A., Luikart G., Wayne R.K. & the S.N.P.w.g. (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-16.

Nei M. (1987) *Molecular evolutionary genetics*. Columbia university press.

Orr H.T. (2012) Cell biology of spinocerebellar ataxia. *J Cell Biol* **197**, 167-77.

Osada N. (2015) Genetic diversity in humans and non-human primates and its evolutionary consequences. *Genes & Genetic Systems* **90**, 133-45.

Padiath Q.S., Srivastava A.K., Roy S., Jain S. & Brahmachari S.K. (2005) Identification of a novel 45 repeat unstable allele associated with a disease phenotype at the MJD1/SCA3 locus. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **133**, 124-6.

Park H., Kim H.-J. & Jeon B.S. (2015) Parkinsonism in Spinocerebellar ataxia. *BioMed research international* **2015**.

Paulson H.L., Perez M.K., Trottier Y., Trojanowski J.Q., Subramony S.H., Das S.S., Vig P., Mandel J.L., Fischbeck K.H. & Pittman R.N. (1997) Intranuclear Inclusions of Expanded Polyglutamine Protein in Spinocerebellar Ataxia Type 3. *Neuron* **19**, 333-44.

Pinheiro M.d.F.T. (2010) Genética Forense Perspectivas da Identificação Genética. *Porto: Edições Universidade Fernando Pessoa*.

Press M.O., Carlson K.D. & Queitsch C. (2014) The overdue promise of short tandem repeat variation for heritability. *Trends Genet* **30**, 504-12.

Rozen S. & Skaletsky H. (1999) Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics methods and protocols*, 365-86.

Schlotterer C. (2004) The evolution of molecular markers [mdash] just a matter of fashion? *Nat Rev Genet* **5**, 63-9.

Schmidt T., Lindenberg K.S., Krebs A., Schöls L., Laccone F., Herms J., Rechsteiner M., Riess O. & Landwehrmeyer G.B. (2002) Protein surveillance machinery in brains with spinocerebellar ataxia type 3: Redistribution and differential recruitment of 26S proteasome subunits and chaperones to neuronal intranuclear inclusions. *Annals of Neurology* **51**, 302-10.

Schöls L., Amoiridis G., Büttner T., Przuntek H., Epplen J.T. & Riess O. (1997) Autosomal dominant cerebellar ataxia: Phenotypic differences in genetically defined subtypes? *Annals of Neurology* **42**, 924-32.

Schöls L., Bauer P., Schmidt T., Schulte T. & Riess O. (2004) Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. *The Lancet Neurology* **3**, 291-304.

Sequeiros J. & Coutinho P. (1993) Epidemiology and clinical aspects of Machado-Joseph disease. *Advances in neurology* **61**, 139.

Shakkottai V.G. & Fogel B.L. (2013) Clinical neurogenetics: autosomal dominant spinocerebellar ataxia. *Neurol Clin* **31**, 987-1007.

Shibata-Hamaguchi A., Ishida C., Iwasa K. & Yamada M. (2009) Prevalence of Spinocerebellar Degenerations in the Hokuriku District in Japan. *Neuroepidemiology* **32**, 176-83.

Sobrino B., Brión M. & Carracedo A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International* **154**, 181-94.

Storey E., du Sart D., Shaw J.H., Lorentzos P., Kelly L., McKinley Gardner R., Forrest S.M., Biros I. & Nicholson G.A. (2000) Frequency of spinocerebellar ataxia types 1, 2, 3, 6, and 7 in Australian patients with spinocerebellar ataxia. *American Journal of Medical Genetics* **95**, 351-8.

Takahashi Y., Kanai M., Taminato T., Watanabe S., Matsumoto C., Araki T., Okamoto T., Ogawa M. & Murata M. (2017) Compound heterozygous intermediate MJD alleles cause cerebellar ataxia with sensory neuropathy. *Neurology: Genetics* **3**, e123.

Takiyama Y., Nishizawa M., Tanaka H., Kawashima S., Sakamoto H., Karube Y., Shimazaki H., Soutome M., Endo K., Ohta S., Kagawa Y., Kanazawa I., Mizuno Y., Yoshida M., Yuasa T., Horikawa Y., Oyanagi K., Nagai H., Kondo T., Inuzuka T., Onodera O. & Tsuji S. (1993) The gene for Machado-Joseph disease maps to human chromosome 14q. *Nat Genet* **4**, 300-4.

Takiyama Y., Sakoe K., Nakano I. & Nishizawa M. (1997) Machado-Joseph disease: cerebellar ataxia and autonomic dysfunction in a patient with the shortest known expanded allele (56 CAG repeat units) of the MJD1 gene. *Neurology* **49**, 604-6.

Teive H.A.G., Moro A., Arruda W.O., Raskin S., Teive G.M.G., Dallabrida N. & Munhoz R.P. (2016) Itajaí, Santa Catarina Azorean ancestry and spinocerebellar ataxia type 3. *Arquivos de Neuro-Psiquiatria* **74**, 858-60.

Tomiuk J., Bachmann L., Bauer C., Rolfs A., Schöls L., Roos C., Zischler H., Schuler M.M., Bruntner S. & Riess O. (2007) Repeat expansion in spinocerebellar ataxia type 17 alleles of the TATA-box binding protein gene: an evolutionary approach. *European journal of human genetics: EJHG* **15**, 81.

Vale J., Bugalho P., Silveira I., Sequeiros J., Guimarães J. & Coutinho P. (2010) Autosomal dominant cerebellar ataxia: frequency analysis and clinical characterization of 45 families from Portugal. *European Journal of Neurology* **17**, 124-8.

Vallone P.M. & Butler J.M. (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* **37**, 226-31.

Van Alfen N., Sinke R.J., Zwarts M.J., Gabreëls-Festen A., Praamstra P., Kremer B.P. & Horstink M.W. (2001) Intermediate CAG repeat lengths (53, 54) for MJD/SCA3 are associated with an abnormal phenotype. *Annals of Neurology* **49**, 805-8.

Van de Warrenburg B., Sinke R., Verschuuren–Bemelmans C., Scheffer H., Brunt E., Ippel P., Maat–Kievit J., Dooijes D., Notermans N. & Lindhout D. (2002) Spinocerebellar ataxias in the Netherlands Prevalence and age at onset variance analysis. *Neurology* **58**, 702-8.

Vignal A., Milan D., SanCristobal M. & Eggen A. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**, 275-306.

Wang C., Peng H., Li J., Ding D., Chen Z., Long Z., Peng Y., Zhou X., Ye W., Li K., Xu Q., Ai S., Song C., Weng L., Qiu R., Xia K., Tang B. & Jiang H. (2017) Alteration of methylation status in the ATXN3 gene promoter region is linked to the SCA3/MJD. *Neurobiology of Aging* **53**, 192.e5-.e10.

Williams A.J. & Paulson H.L. (2008) Polyglutamine Neurodegeneration: Protein Misfolding Revisited. *Trends in neurosciences* **31**, 521-8.

Yates A., Akanni W., Amode M.R., Barrell D., Billis K., Carvalho-Silva D., Cummins C., Clapham P., Fitzgerald S., Gil L., Girón C.G., Gordon L., Hourlier T., Hunt S.E., Janacek S.H., Johnson N., Juettemann T., Keenan S., Lavidas I., Martin F.J., Maurel T., McLaren W., Murphy D.N., Nag R., Nuhn M., Parker A., Patricio M., Pignatelli M., Rahtz M., Riat H.S., Sheppard D., Taylor K., Thormann A., Vullo A., Wilder S.P., Zadissa A., Birney E., Harrow J., Muffato M., Perry E., Ruffier M., Spudich G., Trevanion S.J., Cunningham F., Aken B.L., Zerbino D.R. & Flicek P. (2016) Ensembl 2016. *Nucleic acids research* **44**, D710-D6.

Zeng S., Zeng J., He M., Zeng X., Zhou Y., Liu Z., Jiang H., Tang B. & Wang J. (2015) Chinese homozygous Machado-Joseph disease (MJD)/SCA3: a case report. *J Hum Genet* **60**, 157-60.

Zhao X.N. (2015) The Repeat Expansion Diseases: the dark side of DNA repair? **32**, 96-105.

Zhao Y., Tan E.K., Law H.Y., Yoon C.S., Wong M.C. & Ng I. (2002) Prevalence and ethnic differences of autosomal-dominant cerebellar ataxia in Singapore. *Clinical Genetics* **62**, 478-81.

Zoghbi H.Y. & Orr H.T. (2000) Glutamine Repeats and Neurodegeneration. *Annual Review of Neuroscience* **23**, 217-47.

Supplementary material

Supplementary material 1: Genotyping data of SNPs found in Portuguese MJD families.

Sample ID	Family	rs12586535	rs12586471	rs56268847	rs10467858	rs10467857	rs10467856	rs12895357	rs7158733	rs3092822	rs7158238	rs12588287	rs7153675	rs7153603	rs7153696	rs7153374	rs7253193	rs4904833	rs7146985	rs113572439	rs2006047	rs8004149	rs111735934	rs7142326
U4215	P10							G/C	C/A	A/C	G/A	A/G	A/G	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	G/A	G/G	A/G
U4455	P52	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G				A/G	G/G	G/A	A/G
U4459	P57	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G				A/G	G/G	G/A	A/G
U4485	P63	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G	T/T	C/C	T/T	A/G	G/G	G/A	A/G
U4483	P63	C/T	T/C	A/A	G/A	G/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T				G/G			
U1722	P67	C/C	T/T	A/A	G/G	G/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G			T/T	A/G	G/G	G/A	A/G
U484	P72	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G	T/T	C/C	T/T	A/G	G/G	G/A	A/G
U2402	P83									A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G							
U5816	P83	C/C	T/T	A/A		C/G	T/T																	
U3196	P91	C/T		A/A	G/A	C/G	T/C	G/C	C/A	A/C		A/G	G/A	G/A	C/T	G/A	G/T							
U3301	P96																	T/C	T/C	T/A	A/G	A/G	G/G	A/G
U3302	P96	C/T	T/C	A/A	G/A	C/G	C/T	G/C	C/A	A/C	A/G	A/G	A/G	A/G	C/T	G/A	G/T						G/G	
U4121	P96																	C/C	T/T	A/A	G/G	A/A	G/G	G/G
U1419	P97	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A		A/A	G/G	G/G	C/C	G/G	G/G	T/T	C/C	T/T	A/G	G/G	G/A	A/G
U1548	P99	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	A/C		A/G	G/A	G/A	C/T	G/A	G/T							
U6178	P110	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G			T/T	A/G	G/G	G/A	
U3911	P110				G/G	C/G	T/T		C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G			T/T	A/G	G/G	G/A	
U5993	P111	C/T	T/C	A/A	G/A	G/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T				G/G	G/A	G/A	G/G
U7830	P112	C/T		A/A	G/A	G/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T				G/G	G/A	G/G	G/G
U6280	P114	C/C	T/T	A/A	G/G	G/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G							
U7901	P117	C/T		A/A	G/A	C/G		G/C	C/A	C/A	A/G	A/G	G/A	G/A	C/T	G/A	G/T							
U6262	P117	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	C/A	A/G	A/G	G/A	G/A	C/T	G/A	G/T				A/G	G/A	G/G	A/G
U7284	N12	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T				A/G	G/A	G/G	A/G
U6384	N12	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T	C/T	C/T	T/A	A/G	G/A	G/G	A/G
U7874	N13	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G							
U7628	N13	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G				A/G	G/G	G/A	
U7933	N14	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G				A/G	G/G	G/A	
U7431	N15	C/C	T/T	A/A	G/G	C/G	T/T	G/G	C/C	A/A	G/G	A/A	G/G	G/G	C/C	G/G	G/G			T/T				
U12201	N19	T/T	C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T							

X - underline nucleotides are associated to normal allele

Supplementary material 2: Genotyping data of SNPs found in Taiwanese MJD families.

Sample ID	Family	rs12586535	rs12586471	rs56268847	rs10467858	rs10467857	rs10467856	rs12895357	rs7158733	rs3092822	rs7158238	rs12588287	rs7153675	rs7153603	rs7153696	rs7153374	rs7253193	rs4904833	rs7146985	rs113572439	rs2006047	rs8004149	rs111735934	rs7142326
1	S1	C/T	T/C	A/G	G/A	C/G	T/C	G/C	C/A	C/A	A/G	A/G	G/A	G/A	C/T	G/A	G/T				A/G	G/A	G/G	A/G
3	S1	T/T	C/C	A/G				C/C	A/A	C/C									T/T	A/A				
4	S2	T/T	C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T	C/C	T/T	A/A	G/G	A/A	G/G	G/G
1	S3	T/T	C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T	C/C	T/T	A/A	G/G	A/A	G/G	G/G
3	S3	C/T	T/C	A/A	G/A	C/G	T/C	G/C		A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	A/G	G/G	A/G
6	S4	T/T	C/C	A/G	A/A	G/G	C/C	C/C	A/A	C/C								C/C	T/T	A/A	G/G	A/A	G/G	G/G
2	S5	T/T	C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T				G/G	A/A	G/G	G/G
1	S5	T/T	C/T	A/A	G/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T							
2	S6	C/T	T/C	A/G	G/A	C/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	A/G	G/G	A/G
3	S7	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	G/A	G/G	A/G
2	S8		C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T							
5	S8	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	A/C		A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	G/A	G/G	A/G
3	S9	T/T	C/T	A/A	A/A	G/G	C/C	G/C	A/C	A/C	A/G	A/G	G/A	A/G	C/T	A/G	G/T	C/C	T/T	A/A	G/G	A/A	G/G	G/G
1	S10	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	C/C		A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	G/A	G/G	A/G
3	S10	T/T	C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T							
1	S12	T/T	C/C	A/G	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T				G/G	A/A	G/G	A/G
2	S12	T/T		A/G	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T				G/G	A/A	G/G	G/G
4	S12	T/C	T/C	A/G	G/A	G/C	C/T	G/C	C/A	A/C														
1	S13	T/T	C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T	C/C	T/T	A/A		A/A	G/G	G/G
2	S15			A/A	G/A	C/G	T/C			A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A				
1	S15	C/T	T/C	A/A																				
1	S17	C/T	T/C	A/G				G/C	C/A	A/C														
2	S17	C/T		A/G	G/A	C/G	T/C	G/C	C/A	C/A	A/G							T/C	C/T	T/A	A/G	G/A	G/G	A/G
1	S18	C/T	T/C	A/G	G/A	C/G	T/C		C/A	A/C	G/A	A/G	G/A	G/A	C/T	G/A	G/T							
2	S18	C/T	T/C	A/G	G/A	C/G	T/C	G/C	C/A	A/C	G/A							T/C	C/T	T/A	A/G	G/A	G/G	A/G
1	S19										A/A	G/G	A/A	A/A	T/T	A/A	T/T	C/C	T/T	A/A	G/G	A/A	G/G	G/G
3	S19	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	C/A	A/G	A/G	G/A	G/A	C/T	G/A	G/T				A/G	G/A	G/G	
3	S20	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	C/A		A/G	G/A	G/A	C/T	G/A	G/T				A/G	G/A	G/G	A/G
1	S21	T/T	C/C	A/A				C/C	A/A	C/C														
3	S21	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A										C/T	T/A	A/G	G/A	G/G	A/G
1	S22	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	A/C	A/G	A/G	G/A	G/A	C/T	G/A	G/T				A/G	G/A	G/G	A/G
2	S22							G/C	C/A												G/G	A/A	G/G	G/G
1	S23	T/T	C/C	A/A	A/A	G/G	C/C			C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T	C/C	T/T	A/A	G/G	A/A	G/G	G/G
2	S23		C/C	A/A	A/A	G/G	C/C	C/C	A/A	C/C	A/A	G/G	A/A	A/A	T/T	A/A	T/T	C/C	T/T	A/A	G/G	A/A	G/G	G/G
3	S24	C/T	T/C	A/G	G/A	C/G	T/C	G/C	C/A	A/C		A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	G/A	G/G	A/G
1	S25	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A			A/G	G/A	G/A	C/T	G/A	G/T				A/G			
2	S25	C/T	T/C	A/A	G/A	C/G	T/C	G/C	C/A	C/A		A/G	G/A	G/A	C/T	G/A	G/T	T/C	C/T	T/A	A/G	G/A	G/G	A/G

X - underline nucleotides are associated to normal allele

Supplementary material 3: Haplotypes inferred for Portuguese MJD families based on allele-specific amplification, family segregation or PHASE software.

Machado families	rs12586535	rs12586471	rs56268847	rs10467858	rs10467857	rs10467856	rs12895357	rs7158733	rs3092822	rs7158238	rs12588287	rs7153675	rs7153603	rs7153696	rs7153374	rs7253193	rs4904833	rs7146985	rs113572439	rs2006047	rs8004149	rs111735934	rs7142326	Probabilities[a]
P52	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.930
P57	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.909
P63	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.999
P67	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.913
P72	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.996
P83/P57[#]	C	T	A			T			A	G	A	G	G	C	G	G								
P97	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.988
P110	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.912
P111	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.917
P114	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.601
N13	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.910
N14	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.919
N15	C	T	A	G	G	T	G	C	A	G	A	G	G	C	G	G	T	C	T	G	G	A	G	0.602
Joseph families																								
P10	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.862
P96	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.957
P91	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.886
P112	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.895
P117	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.915
N12	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	1.000
N19	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.917
P99	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.870

Note: The nucleotides in black represents the phase obtained from Phase software.

[a] Probabilities for the haplotypes reconstructed by PHASE 2.2.

[#] We were not able to infer the SNP haplotypes with PHASE 2.2 from P83 family; however the STR markers, previously analysed, showed that both P83 and P57 families have the same STR haplotype.

Supplementary material 4: Haplotypes inferred for Taiwanese MJD families based on allele-specific amplification, family segregation or PHASE software.

Families	rs12586535	rs12586471	rs56268847	rs10467858	rs10467857	rs10467856	rs12895357	rs7158733	rs3092822	rs7158238	rs12588287	rs7153675	rs7153603	rs7153696	rs7153374	rs7253193	rs4904833	rs7146985	rs113572439	rs2006047	rs8004149	rs111735934	rs7142326	Probabilities[a]
S1	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.956
S2	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	1.000
S3	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.995
S4	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.947
S5	T	T	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.759
S6	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	1.000
S7	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.999
S8	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.986
S9	T	T	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.835
S10	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.983
S12	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.958
S13	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.987
S15	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.892
S17	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.955
S18	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.948
S19	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.948
S20	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.962
S21	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.879
S22	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.966
S23	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.998
S24	T	C	G	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.981
S25	T	C	A	A	G	C	C	A	C	A	G	A	A	T	A	T	C	T	A	G	A	G	G	0.988

Note: The nucleotides in black represents the phase obtained from Phase software
[a] Probabilities for the haplotypes reconstructed by PHASE 2.2.

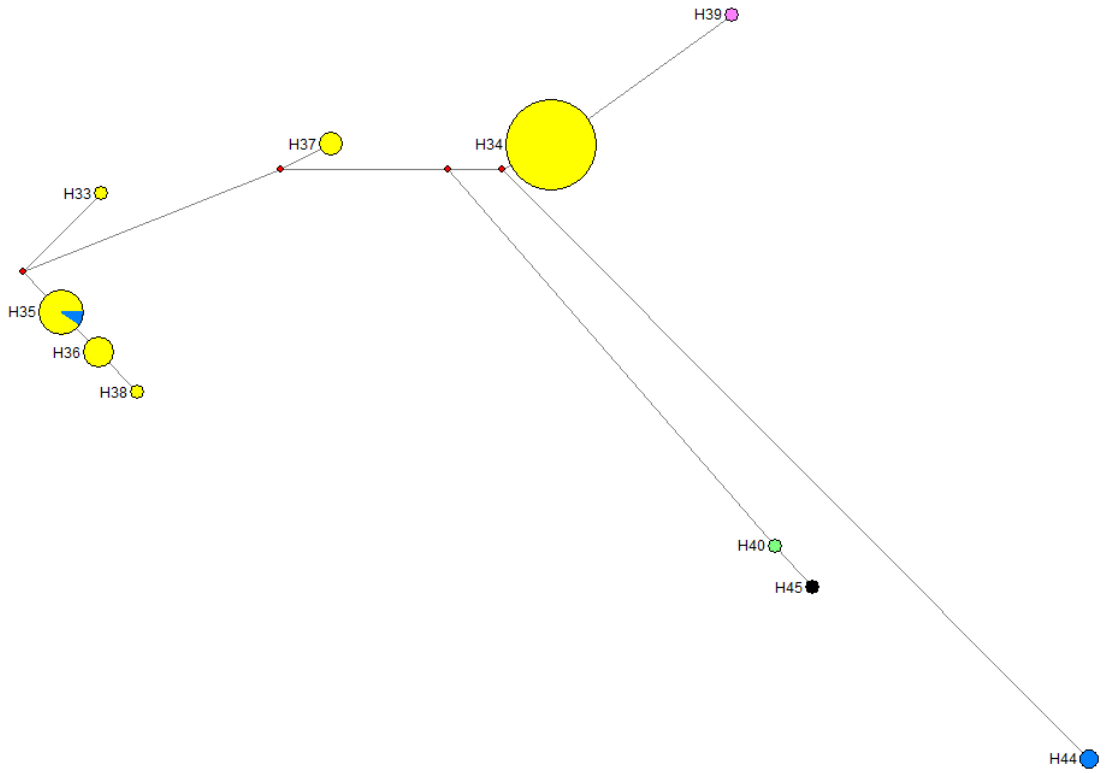
Supplementary material 5: STR-haplotype, previously obtaneid, for Asian MJD families from the Joseph lineage.

Haplotype	Number of families	TAT^{223}	GT^{199}	ATA^{194}	AC^{21}	$AAAC^{123}$	GT^{190}	AC^{190}
Asian haplotypes with A allele at A/G.485								
H1	1	17	23	10	14	8	19	23
H6	5	16	25	10	14	5	15	24
H7	1	16	25	10	14	5	19	24
H9	2	11	24	10	14	5	14	24
H11	1	11	21	9	14	5	15	24
H12	8	10	21	10	13	7	15	24
H14	1	10	25	10	14	5	15	24
H15	1	10	21	11	13	7	15	24
H17	7	8	22	10	14	8	19	23
Asian haplotypes with G allele at A/G.485								
H19	1	17	23	10	13	7	19	28
H20	1	16	23	9	12	7	19	26
H21	2	16	23	9	12	8	15	24
H22	1	16	20	11	14	8	19	25
H23	1	16	21	10	14	5	19	25
H24	3	14	23	10	12	8	19	25
H25	1	10	24	9	18	7	19	26
H26	1	10	25	10	12	7	19	25
H27	1	10	23	13	12	7	19	24
H28	2	10	23	10	12	7	19	25
H29	1	10	21	10	14	5	19	25
H8	2	15	23	9	12	7	19	25
H10	1	11	22	10	12	7	16	26
H3	1	16	23	9	12	8	15	25
Portuguese haplotypes								
H31	38	11	21	11	14	5	15	25
H32	2	17	21	10	14	5	15	24

Supplementary material 6: STR-haplotype, previously obtained, for worldwide MJD families from Machado lineage

Haplotype	Number of families	TAT^{223}	GT^{199}	ATA^{194}	AC^{21}	$AAAC^{123}$	GT^{190}	AC^{190}
Portuguese								
H33	1	16	23	9	18	7	19	24
H34	40	10	20	10	18	7	19	24
H35	9	16	25	10	18	7	19	24
H36	5	17	25	10	18	7	19	24
H37	3	11	24	9	18	7	19	24
H38	1	17	25	10	19	7	19	24
Spanish								
H39	2	10	20	10	18	7	15	24
Peruvians								
H40	2	11	21	11	14	7	19	28
Caribbean								
H41	1	16	22/26	10	14	8	19	25/27
North American								
H42	1	16	25	10	18	7	19	24
H43	1	10/11	19/21	10/13	14/18	7	19	24/28
H44	2	15	17	7	14	7	18	24
Unkonwn								
H45	1	11	21	11	14	7	18	28

Note: Underlined haplotypes were not used to reconstruct the phylogenetic network



Supplementary material 7: Phylogenetic network showing the most parsimonious relationships among STR-based haplotypes of all Machado MJD families. Families from Portugal (n=59) are coloured in yellow. Families from Spain (n=2), Peru (n=2), North-American (n=3) and from unknown origin (n=1) are coloured in pink, green, blue and black, respectively. The haplotype H42 was included in the same circle as haplotype H35.