Identification of *ATXN3* polymorphisms with na influence on the age-at-onset of Machado-Joseph disease patients

Nicole Mineiro
Mestrado em Genética Forense
Departamento de Biologia
2018
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, _____/_____/_____

Acknowledgments

To Professor Amorim for allowing me to execute this project in his research group.

To my orientator, Dr. Sandra Martins, for her tireless support and patience, contagious enthusiasm, admirable work ethic and guidance throughout this project

To Inês Costa who worked on this project with me and who was always available to answer my endless quentions.

To Luísa, Ana Luísa, Sofia, and all the girls from Cientuna for helping me become a better version of myself.

To my childhood friends, Jessica, Márcia, Daniel, Rui and Carlos, who have been there through it all.

To Fred, for being my rock.

And last but never least, to my parents, for always believing in me, being there no matter what and always making a plan.

To all on this list a big thank you.

Abstract

Machado-Joseph disease (MJD), also known as spinocerebellar ataxia type 3 (SCA3), is a dominantly inherited, late-onset neurodegenerative disorder (mean onset at 40.2 years). MJD is currently known to be the most common of SCAs worldwide. The cause of this disease is the pathological expansion of a CAG trinucleotide repeat on exon 10 of the *ATXN3* gene (14q32.1). Numerous studies in populations of diverse ethnic origins have shown that healthy individuals have usually 12-44 triplets whereas in MJD patients, the range is between 61 and 87 CAGs. A gap is, therefore, displayed in the repeat length of *ATXN3*-$(CAG)_n$, with the role of intermediate alleles on disease pathogenesis still unclear.

Even though the disease is caused by the expanded CAG repeat, the size of the tract only explains about 48-60% of MJD age-at-onset (AO) variability. Previous studies revealed that the other half of the variability could be explained by unknown environmental causes and familial AO modifiers such as normal allele size (previously shown to have a small but significant effect on MJD phenotype) or other genetic modifiers in *cis* or *trans* with the deleterious expansion.

The aim of this study was to identify SNPs within the causative gene *ATXN3* that, in addition to the CAG repeat, have an influence on MJD AO. We analyzed a total of 120 MJD patients from different Portuguese families and searched for polymorphisms in regulatory and exonic regions of *ATXN3* that could potentially modulate disease phenotype. To achieve this, we started by optimizing multiplex reactions to amplify our regions of interest, followed by sequencing and alignment of the fragments obtained. Next, we advanced to the detection of significantly relevant polymorphisms by analysis of covariance to remove the known influence of the repeat size in the expanded allele. Lastly, we attempted to quantify the effect of those variants on disease onset.

In our heterogeneous cohort of MJD patients, expanded CAG repeat size accounted for 53% of AO variation (model A). The analysis of variations in *ATXN3* resulted in the identification of 18 SNPs- our initial set of candidate genetic modifiers of MJD clinical presentation. Analysis of covariance then showed that 3 of these SNPs (rs3814834, rs4904834 and rs10146519) contributed to the explanation of an additional 7.9% of AO variability in our cohort, besides the already identified effect of the expansion length. The cumulative influence of this set of SNPs on MJD AO led us to develop a more complex new model for AO/$(CAG)_n$ linear regression (model B). This model allowed for a more

accurate AO prediction for most patients when compared to the predictions obtained based solely on the CAG tract length of expanded alleles.

Regulatory variant rs3814834, linked to an earlier onset, is situated in the 5'UTR region of *ATXN3* and may, therefore, affect the binding of transcription factors and modify *ATXN3* transcription. To confirm this polymorphism's effect on the ATXN3 protein we could, in the future, conduct a gene expression assay to verify if there is a change on *ATXN3* transcriptional activity in the presence of different rs3814834 alleles. The functional relevance of intronic variants, rs4904834 and rs10146519 is more difficult to assess, nevertheless, one may hypothesize their influence on splicing efficiency of the *ATXN3* gene as the majority of known functional intronic polymorphisms activate a cryptic splice site or are located within splice junctions, or even within an intron splice enhancer or branchpoint site.

To further validate our claims, this study would benefit if we increased sample size to analyze a more heterogeneous cohort of patients, either by including patients from different populations or carrying MJD alleles in the extremes of the expanded CAG repeat distribution.

Relevant SNPs and respective allele frequencies can then, in addition to improving disease prognosis, be useful to develop MJD therapies using RNAi by inhibiting mutant *ATXN3* expression which would correct MJD by targeting the initial cause of the disorder.

Keywords: Machado-Joseph disease, Age-at-onset, CAG repeat, SNP, *ATXN3* gene.

Resumo

A doença de Machado-Joseph (DMJ), também conhecida como ataxia espinocerebelosa do tipo 3 (SCA3), é uma doença neurodegenerativa com transmissão dominante, de início tardio (média, 42.2 anos), conhecida atualmente como a SCA mais frequente a nível mundial. A causa desta doença é a expansão de uma repetição do trinucleotídeo CAG no exão 10 do gene *ATXN3* (14q32.1). Vários estudos de populações com diversas origens étnicas mostraram que indivíduos saudáveis têm, geralmente, 12-44 tripletos enquanto que doentes com DMJ possuem entre 61 e 87 CAGs. Há, portanto, um intervalo entre os comprimentos de alelos *ATXN3*-$(CAG)_n$ normais e expandidos, mas o papel dos alelos intermédios na patogenicidade da doença mantém-se incerto.

Apesar da doença ser causada pela expansão do motivo repetitivo CAG, este apenas explica cerca de 48-60% da variabilidade da idade de início da DMJ. Estudos prévios revelaram que a restante variabilidade poderá ser explicada por fatores ambientais ainda desconhecidos e por modificadores genéticos, tais como o tamanho do alelo normal (cuja contribuição, apesar de pequena, é significativa) ou outros fatores em *cis* ou *trans* com a expansão deletéria.

O objetivo deste estudo foi identificar SNPs localizados no gene responsável pela DMJ, *ATXN3*, que em conjunto com o motivo repetitivo CAG, influenciem a idade de início da DMJ. Analisamos um total de 120 doentes de famílias portuguesas diferentes. Procuramos polimorfismos nas regiões reguladoras, exónicas e respetivas sequências flanqueantes do gene *ATXN3* com potencial para modular a progressão da doença. Para tal, começamos por otimizar reações de multiplex para a amplificação das regiões de interesse, seguindo-se a sequenciação e alinhamento dos fragmentos obtidos. Em seguida, avançamos para a deteção de polimorfismos significativamente relevantes, tendo em consideração a influência conhecida do tamanho da repetição no alelo expandido. Por fim, tentamos quantificar o efeito das variantes encontradas no fenótipo da doença, nomeadamente na idade de início.

No nosso grupo heterogéneo de doentes com DMJ, o tamanho do motivo repetitivo CAG expandido explicou cerca de 53% da variação da idade de início (modelo A). A análise das variantes encontrados no gene *ATXN3* resultou na identificação de 18 SNPs- o nosso conjunto inicial de candidatos a estudar como modificadores genéticos da apresentação clínica da DMJ. Uma análise de covariância mostrou, então, que três dos

nossos SNPs (rs3814834, rs4904834 e rs10146519) contribuem para a explicação adicional de 7.9% da variabilidade da idade de início no nosso grupo de doentes, além do já identificado efeito do tamanho da expansão. O efeito cumulativo deste conjunto de SNPs levou-nos a desenvolver um novo modelo mais complexo para a regressão linear AO/$(CAG)_n$ (modelo B). Este modelo permitiu uma previsão mais precisa da idade de inico para a maior parte dos doentes quando comparado com as previsões obtidas com base apenas no tamanho do motivo repetitivo.

A variante rs3814834, associada a uma idade de início da DMJ mais precoce, está localizada na região reguladora do gene, num local previamente identificado como sendo de ligação a seis fatores de transcição. Deste modo. Para confirmar o efeito deste polimorfismo na possível alteração da transcrição da proteína ATXN3 poderíamos, no futuro, efetuar um ensaio de expressão génica para verificar se de facto isto acontece. Por outro lado, a relevância funcional das variantes intrónicas, rs4904834 e rs10146519, é mais difícil de avaliar. Podemos, no entanto, hipotetizar sobre a sua influência na eficácia do *splicing* do gene *ATXN3,* já que a grande maioria de polimorfismos intrónicos funcionais conhecidos ativa um local de *splicing* críptico ou encontra-se em locais de *splicing*, ou até em *splice enhancers* ou *branchpoints.*

Para validar as nossas hipóteses, seria útil aumentar o tamanho da amostra com o objetivo de estudar um conjunto mais heterogéneo de doentes com DMJ. Tal poderia ser realizado através da inclusão de amostras de doentes de origens diferentes e/ou de doentes portadores de alelos $(CAG)_n$_*ATXN3* nos extremos da distribuição do tamanho da repetição.

Posteriormente, os SNPs relevantes e as respetivas frequências alélicas podem, para além de melhorar o prognóstico da doença, ser úteis no desenvolvimento de terapias da doença que usam o RNAi para inibir a expressão do *ATXN3* mutado, o que poderia diminuir a toxicidade na DMJ.

Palavras-Chave: Doença de Machado-Joseph, Idade de início, Repetição CAG, SNP, Gene *ATXN3*

Contents

Agradecimentos……...v

Abstract……vi

Resumo…… viii

List of figures……xiii

List of tables……xiii

List of abbreviatios……xiv

Introduction……1

1. Genetic diversity……1
2. Molecular markers……2
   a. SNPs……2
   b. Microsatellites……3
3. Repeat expansion disorders……4
   a. Trinucleotide repeat expansions……5
   b. Spinocerebellar ataxias……5
4. Machado-Joseph disease……6
   a. Epidemiology……6
   b. Clinical presentation……7
   c. *ATXN3* gene……7
5. Haplotype analyses in MJD……9
   a. Genotype/phenotype correlations (intergenerational instability and age-at-onset)……9
   b. Ancestral mutational origins……10
6. Targets for siRNA therapies……11

Goals……12

Subjects and Methods……13

1. Subjects……13
2. Methods……13
   a. Primer design……13

b. PCR optimization……15
c. Detection of amplified products……16
d. DNA sequencing……..……..……...18
e. Statistical and *in silico* analyses…….……..…..20

Results……...…….21

1. Clinical and genotyping data…….……...21
2. Correlation between the expanded CAG repeat length and MJD AO…….……...22
3. The influence of *ATXN3* SNPs on MJD AO……...24
4. Addition of polymorphisms effect to the AO/CAG linear regression model……..25
5. Influence of SNP minor alleles on mean AO……...26
6. Phenotypic influence of *ATXN3* polymorphisms……28

Discussion…….……30

Concluding Remarks and Future Perspectives…….……..……34

References…….……35

List of figures

Figure 1: Schematic representation of the *ATXN3* gene structure. Exons are numbered from 1 to 11 and are presented as boxes. Filled blue boxes indicate the coding regions. The location of the polymorphic $(CAG)_n$ tract is indicated (Adapted from Bettencourt & Lima 2011) …..……...…….8

Figure 2 – Clinical and genotyping data for our cohort of 120 Portuguese MJD patients. Distribution of patients according to (A) gender (N=99), (B) diagnosed AO (N=120), (C) CAG repeat size in the normal (N=120) and (D) expanded (N=120) *ATXN3* alleles. Variables AO and CAG repeat size in expanded alleles follow a normal distribution……..21

Figure 3 - Linear regression between CAG repeat length in the MJD chromosome of 120 patients and disease AO. A linear coefficient of determination of r^2=0.5297 was obtained…….……22

Figure 4 – Scatter plot graph displaying the ratio between observed and predicted MJD AO by models A and B…….29

Figure 5 - Location of variants rs3814834, rs4904834 and rs10146519 in the *ATXN3* gene (Adapted from Bettencourt *et al.* 2012) ……29

List of Tables

Table 1 - Clinical subtypes in MJD (Adapted from Bettencourt and Lima, 2011) …….6

Table 2 - Clinical information and genotyping data of the $(CAG)_n$ and SNPs found in analyzed regions of *ATXN3*. E – amplified exon……14

Table 3 - Primers for amplification and sequencing of exonic and regulatory regions of *ATXN3*……15

Table 4 - PCR reagents for the optimized pentaplex reaction……15

Table 5 - PCR protocol for the optimized pentaplex reaction……15

Table 6 - PCR reagents for the optimized triplex reaction……16

Table 7 - PCR protocol for the optimized triplex reaction……16

Table 8 - Reagents used in the polyacrylamide gel……16

Table 9 - Steps of the polyacrylamide gel coloration with silver staining……17

Table 10 - Reagents and quantities for PCR pre-sequencing purification ……...18

Table 11 - Protocol for PCR pre-sequencing purification ……18

Table 12 - Reagents and volumes for sequencing reactions……18

Table 13 - Sequencing reaction conditions……19

Table 14 - Analyses of conformity with Hardy-Weinberg (HW) equilibrium expectations……...22

Table 15 - SNPs found in the analyzed regions of the *ATXN3* gene and their influence on the AO of MJD patients after accounting for the known effect of the CAG expansion size. Sequence alterations refer to the reverse strand (transcript *ATXN3*-247)…….23

Table 16 - Mean, standard deviation (SD) and t-test results of MJD AO when the presence or absence of SNP minor alleles is observed ……25

Table 17 - Mean, standard deviation (SD) and t-test and/or ANOVA results of MJD AO for each SNP genotype found and corresponding mean expansion size.....……...25

List of abbreviations

ANCOVA	Analysis of covariance
AO	Age-at-onset
ATXN3	Ataxin-3
bp	Base pairs
DNA	Deoxyribonucleic acid
dNTP	Deoxybonucleotide triphosphate
MJD	Machado-Joseph disease
PCR	Polymerase chain reaction
polyQ	Polyglutamine
RNA	Ribonucleic acid
RNAi	RNA interference
siRNA	Small interfering RNA
RT	Room temperature
SCA	Spinocerebellar ataxia
SNP	Single nucleotide polymorphism
UTR	Untranslated region

Introduction

1. Genetic diversity

The survival and evolution of all organisms mostly depends on their adaptability to environmental changes and is essentially obtained through genetic diversity.

Genetic diversity represents the total genetic variation among individuals within a population and its ultimate source is mutation. Any alteration in the DNA sequence, from indels of a few bases and single nucleotide substitutions, to translocations of chromosomal segments, insertions, deletions, duplications and inversions of megabase segments of DNA, changes in chromosomal number, insertion of transposable elements and expansions and contractions in the number of tandemly repeated DNA motifs, is considered a mutation. To contribute to evolutionary change, mutations must be non-lethal, compatible with fertility and occur in the germline, so that they can be passed down to the next generation (Jobling *et al.* 2014).

Despite their importance in species survival, not all DNA alterations contribute in a positive way to genetic variation and in some cases subtle differences can be pathogenic and have implications on an individual's health. Thus, it is essential to study the underlying genetic diversity of disease-causing mutations so that we can have a better insight into disease susceptibility, and their implications for diagnosis, understanding and treating medical conditions.

The ideal way to analyze genetic diversity is to detect differences, such as SNPs or microsatellites, between samples of individual genomes. This comparison allows us to find causative mutations underlying human monogenic disorders, unravel the evolutionary history of individual locus, and help in the mapping of genes involved in complex disorders, through association studies (Jobling *et al.* 2014).

2. Molecular markers

Molecular markers are nucleotide sequences that correspond to a known or unknown physical location in the genome (Barrandeguy & García 2014); they allow us to detect variability among different individuals and distinguish between genotypes that are relevant to a trait of interest based on the polymorphisms found in DNA sequences (Schlotterer 2004; Grover & Sharma 2014).

The search for human genome variation and the development of methods to analyze this variation in large numbers of samples have been mainly driven by the desire to identify genes responsible for genetic disorder and to answer many biological questions, ranging from gene mapping to population genetics, phylogenetic reconstruction, paternity testing and forensic applications (Schlotterer 2004). Since the era of polymerase chain reaction (PCR), the markers most often employed have been SNPs and microsatellites (Jobling *et al.* 2014).

a. SNPs

Single nucleotide polymorphisms (SNPs) are single base pair positions in genomic DNA at which there is more than one sequence alternative in different individuals from the same species (Brooks 1999). Two fundamental processes give rise to SNPs: mis incorporation of nucleotides during replication of DNA, and mutagenesis caused by the chemical modification of bases, or physical damage due, for example, to ultraviolet or ionizing radiation. We can say that a polymorphism exists when at least two alleles are present in the population, and the minor allele is at 1% or greater frequency (Jobling *et al.* 2014).

A low mutation rate of ~10^{-8} (Nachman & Crowell 2000) is known to be on the origin of SNPs, resulting in alleles identical-by-descent, and not identical-by-state (except the rare cases when recurrent mutations occur). Thus, the presence of the same allele status in two sequences usually implies that they have inherited the base from a common ancestor. This stability also means that most SNPs belong to the class of biallelic markers (Jobling *et al.* 2014). SNPs are found in the human genome about once in every 1000 bp though some regions of the genome are richer in SNPs than others. Given that the human genome is 3.2 billion bp long, we can estimate that there will be approximately 1 million differences between two genomes that are due to SNPs. This represents approximately 85% of human genetic variation (Goodwin *et al.* 2007).

Although most of the SNPs are found within non-coding genomic regions, an important subset corresponds to mutations in genes that are associated with diseases or other phenotypes (Grover & Sharma 2014). This association is useful for linkage analysis and disease association studies

and, has been used more recently when developing siRNA-based therapies for diseases like MJD.

In the forensic field, the biallelic state of most SNPs intrinsically limits the information that can be obtained from their analysis, which has been the major limiting factor for their application. 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). A continuously increasing interest in SNP is due to a set of characteristics that do make them appropriate for forensic studies, such as their very low mutation rates (which are useful for paternity testing), their suitability for analysis using high throughput technologies (important for data basing), and the fact that they can be analyzed in short amplicons: short sizes are desirable since the size of the amplified product is critical for the successful analysis of degraded samples (Sobrino *et al.* 2005).

b. Microsatellites

Microsatellites, also known as short tandem repeats (STRs), are tandem arrays of repeat units 1-6 bp long, and are designated as mono-, di-, tri- tetra-, penta- or hexanucleotides. They are scattered evenly throughout the human genome, accounting for about 3% of the entire genome. However, their distribution within chromosomes is not quite uniform: most are found in the noncoding regions, while only about 8% locate in the codons of genes. On average, one STR occurs per each 2,000 bp in the human genome (Fan & Chu 2007).

Mutation rate for STRs is estimated to be 10^{-6} to 10^{-2} per generation (Kouniaki *et al.* 2015). The most accepted mechanism for STR mutation is replication slippage during DNA synthesis, generating new alleles with different numbers of repeating units (in more than 85% of cases occurs an increase or decrease of just a single repeat unit). The overall mutation rate increases with the increase of the array length, and below a certain number of repeats, mutation becomes undetectably slow, but mutation rate and type depend on several other aspects. For instances, dinucleotide repeat loci mutate more rapidly that tri- and tetranucleotide repeats; uninterrupted repeat arrays mutate faster than those with interrupting variant repeats; and though expansion mutations occur throughout the array size range, the rate of contraction mutation increases as alleles become larger (Jobling *et al.* 2014).

Microsatellites are abundant, ubiquitous, easy to automate, codominant, universal, robust, reliable and reproducible markers. Polymorphism patterns exhibited by microsatellites are greater than any other contemporary marker system, which make them a reliable molecular clock, useful for instances, to estimate the age of a flanking mutation. (Grover and Sharma 2014). While

most microsatellites are selectively neutral, some can clearly have phenotypic effects, such as disease-associated expanded STRs, on the origin of multiple neurological diseases.

3. Repeat expansion disorders

Alterations in simple repetitive sequences lie at the center of DNA evolution and sequence diversity that drives adaptation, however, changes in repetitive sequences can have deleterious effects on gene expression and function, causing disease (Kovtun 2008).

The disease-causing mutation, known as “repeat expansion” occurs when the number of repeats in a gene exceeds the normal threshold length defined for that specific gene. Trinucleotide repeats were the first to be identified, but over the last decades, other motifs emerged, including tetra-, penta- and hexanucleotides.

The exact mechanisms that are involved in such diseases are still unclear, but evidence suggests that it is the result of aberrant DNA repair (Zhao *et al.* 2015), DNA slippage (Sinden & Wells 1992) and/or unequal recombination (Eichler *et al.* 1995). Due to instability of repetitive DNA sequences within the human genome, the number of repeats tends to increase with transmission to offspring, resulting in the clinical phenomenon of anticipation. Anticipation is generally defined as an earlier age-of-onset and a more severe disease phenotype observed in successive generations (McMurray 2010). Several factors affect repeat instability, including the initial tract size; thus, the longer the tract the more likely it is to further expand, with anticipation being even more evident in these cases (Pearson 2005).

The effects of expanded repeats on the affected gene depend primarily on its localization within the gene. Thus, repeats in non-coding regions result in loss of protein function (as in FRDA, caused by expansion of a GAA repeat in the first intron of the gene *FRDA*) or in altered RNA function (as is the case of myotonic dystrophy, caused by the CTG or CCTG expansions in non-coding regions). On the other hand, when in a coding region, the consequence of an elongated peptide is altered protein function (as in Huntington disease and most spinocerebellar ataxias, SCAs) (Gatchel 2005).

a. Trinucleotide repeat expansions

Trinucleotide repeats (TNR) have received much attention since these have been the first disease-associated repeats to be identified. Currently, it is known that the expansion of triplet repeats is the underlying genetic defect in ~20 severe neuromuscular and neurodegenerative disorders (McMurray 2010).

The most common repeat unit is the cytosine–adenine–guanine (CAG), which encodes a polyglutamine (polyQ) stretch in most cases due to its location in exonic regions. With an average frequency of 1–10 cases per 100,000 people, this repeat tract is responsible for Huntington's disease (HD), dentatorubral pallidoluysian atrophy (DRPLA), spinal bulbar muscular atrophy X-linked type 1 (SMAX1/SBMA), and various types of spinocerebellar ataxias (Fan *et al.* 2014).

b. Spinocerebellar ataxias

Spinocerebellar ataxias (SCAs) are a group of late onset neurodegenerative disorders with symptoms caused by dysfunction of the cerebellum and brainstem, as well as their associated pathways and connections, with an autosomal dominant pattern of inheritance (Shakkottai *et al.* 2013). SCAs are a clinically heterogeneous group of disorders, but they have a few key symptoms in common. Ataxia is the predominant clinical manifestation. Typically, the patient will display gait imbalance associated with limb incoordination including problems with gross and fine motor control (Paulson 2009). Other common symptoms include oculomotor disorders, dysarthria (difficulty pronouncing words), dysmetria (inability to judge distance or scale) and kinetic tremor (Manto 2005).

To date, there are close to 30 SCAs diseases described. Globally, they are considered rare disorders, with prevalence estimates varying from 0.3 to 2.0 per 100,000 individuals (Mendonça *et al.* 2018), but locally, some types of SCAs can be as frequent as 835.2:100,000, as is the case of MJD in the Portuguese island of Flores, in the Azores (Martins & Sequeiros 2018).

Despite the high number of SCAs that are clinically well characterized, underlying mutations have only been identified for half of all known types of SCAs. So far, most of them are caused by an abnormal expansion of a repetitive sequence (CAG as the most common), but there are also SCAs caused by point mutations such as SCA5, 13, 14, 27 (Teive & Arruda 2009).

Although similar when compared in a clinical context, SCAs do not share the gene responsible for disease, nor do the genes belong to the same family. Groups of disorders are recognized due to shared molecular mechanisms of disease, such as polyglutamine ataxias, ataxias associated with ion-channel dysfunction, mutations in signal transduction molecules, and disease associated

with non-coding repeats (Shakkottai *et al.* 2013). Similar to the heterogeneous mutation spectrum underlying SCAs, genes responsible for these diseases also vary greatly regarding their function. The normal function of all SCA-causing proteins has not been fully established, but some examples of the diverse functions of normal proteins are transcription and RNA splicing (SCA1), RNA processing (SCA2) and involvement in neuron survival, neuron differentiation, and neurogenesis (SCA10) (Shakkottai *et al.* 2013).

4. Machado-Joseph disease

a. Epidemiology

MJD is considered the most common form of SCA worldwide. The relative frequency of MJD is higher in countries such as China (62.1%), Brazil (59.6%), Portugal (57.8%), Thailand (46.5%), Germany (42%), Singapore (41%), Taiwan (32%), France (32%), Japan (28%), The Netherlands (28%), Venezuela (25%), the USA (21%), and Spain (15%, the same frequency as SCA2). On the other hand, MJD seems absent in Greece, Czech Republic, Cyprus, Sri Lanka, Poland, Finland, Serbia, and rare in Turkey (0.6%) and South Africa (1.1%) (Martins & Sequeiros 2018).

The geographic distribution pattern of MJD is not homogeneous, neither between populations not even within the same country. For instances, MJD, although the most prevalent subtype of SCA in Portugal, is relatively rare in the mainland (1/100,000), but highly prevalent in the Azores, where the highest worldwide prevalence occurs in Flores (Bettencourt & Lima 2011).

Table 1 - Clinical subtypes in MJD (adapted from Bettencourt and Lima 2011).

MJD type	Age-of-onset	Symptoms
I (Joseph)	Early onset (~24.3 ears)	Gait ataxia, dystonia, pyramidal signs and a faster progression of symptoms
II (Thomas)	Intermediate onset (~40.5)	Gait ataxia and with or without pyramidal sings, most frequent
III (Machado)	Late onset (~46.8)	Gait ataxia, associated with peripheral alterations, with or without slight pyramidal and extrapyramidal signs
IV	~38-47	Parkinsonism, with mild cerebellar deficits and peripheral neuropathy
V		Spastic paraplegia without gait ataxia

b. Clinical presentation

MJD is a progressive neurodegenerative disorder transmitted in an autosomal dominant manner that affects the nervous system (Maciel *et al.* 2001). The disease is highly pleomorphic, not only in the variability in age-at-onset, but also in the neurological signs presented by different patients as well as in the resulting degree of incapacity (Bettencourt & Lima 2011).
Although clinical phenotype varies greatly from case to case, in general, MJD is characterized by cerebellar ataxia, progressive external ophthalmoplegia, dysarthria, dysphagia, pyramidal signs, dystonia, rigidity, and distal muscle atrophies (Costa & Paulson 2012). Initially, the clinical heterogeneity led three different North American families of Azorean ancestry (Machado, Thomas and Joseph) to be diagnosed with three apparently independent diseases, by three different research groups. Only upon the diagnosis of several Portuguese families from mainland and the Azores, with conditions that varied within the three forms described, was the disease considered a single genetic entity with variable phenotypic expression (Coutinho & Andrade 1978).

After the unification of MJD, attempts at systematization of symptoms to facilitate clinical classification and diagnosis were made by Coutinho & Andrade. At first, three phenotypic subtypes were proposed in Portuguese patients, based on age of onset, presence/absence of important extra pyramidal signs and the presence/absence of peripheral signs (Coutinho 1992). Subsequently, two other rare types of MJD were proposed as shown on Table 1.

c. *ATXN3* gene

The cause of MJD is the pathological expansion of a CAG trinucleotide repeat on exon 10 of the *ATXN3* gene (14q32.1) (Kawaguchi *et al.* 1994). The 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). The genomic structure of *ATXN3* spans ~48kb, with the most frequent transcripts containing 11 exons. Two additional exons have been recently described: 6a and 9a (Figure 1) (Bettencourt & Lima 2011).

Numerous studies in populations of diverse ethnic origins have shown that healthy individuals have 12-44 triplets whereas in MJD patients, the range is between 61 and 87 CAGs (Maciel *et al.* 2001). A gap is, therefore, displayed in the repeat length of *ATXN3*-$(CAG)_n$ alleles (van Alfen *et al.* 2001). Extremely rare intermediate alleles range between 45 and 60 repeats; only 7 cases have been described to date in the literature, six of them presenting clinical MJD phenotype (Takiyama *et al.* 1997; van Schaik *et al.* 1997; Egan *et al.* 2000; van Alfen *et al.* 2001;

Gu *et al.* 2004; Padiath *et al.* 2005) and one where the presence of the intermediate allele is not associated to the disease (Maciel *et al.* 2001).

The *ATXN3* gene encodes for the protein named ataxin-3, which belongs to the family of cysteine proteases and participates in protein quality control pathways (Todi *et al.* 2009), transcriptional regulation (Costa *et al.* 2010) and cytoskeleton organization (Rodrigues *et al.*, 2010). The gene is ubiquitously transcribed in neuronal and non-neuronal human tissues existing both in the cytoplasm and the nucleus of various cell types (Bettencourt & Lima 2011). When in its mutated form, with polyQ tract above the pathological threshold, the protein is thought to gain a neurotoxic function that, consequently, leads to selective neuronal cell death through a not fully understood process (Bettencourt & Lima 2011; Matos *et al.* 2016).

There are 56 alternative splicing variants described for the *ATXN3* gene. The variants are caused by four types of splicing events (exon skipping, new exons, usage of alternative 5' or 3' splice sites), occurring in a simple or combined way (Bettencout *et al.* 2010), meaning that beyond the genomic DNA sequence, alternative splicing may be an important mechanism in the regulation/cause of ataxin-3 diversity (Bettencourt & Lima 2011).

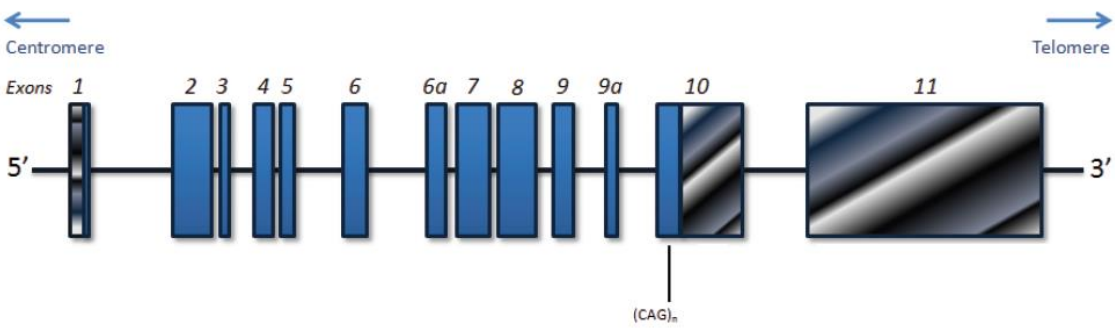


Figure 1 - Schematic representation of the *ATXN3* gene structure. Exons are numbered from 1 to 11 and are presented as boxes. Filled blue boxes indicate the coding regions. The location of the polymorphic $(CAG)_n$ tract is indicated (adapted from Bettencourt & Lima 2011).

5. Haplotype analyses in MJD

Interestingly, two mutational events may explain the fact that MJD is spread worldwide: the oldest mutation probably originated in Asia, later spreading throughout Europe, with a founder effect explaining the high prevalence in Portugal; while the second mutational event is perhaps explained by Portuguese emigration (Gaspar *et al.* 2001; Martins *et al.* 2007; Martins & Sequeiros 2018).

a. Genotype/phenotype correlations (intergenerational instability and age-at-onset)

MJD has autosomal dominant pattern. Hence, siblings of an affected individual, or an asymptomatic carrier, have an a priori risk of 50% of being carriers, with equal probabilities for both genders of receiving and transmitting the mutated allele and expressing the disease (Bettencourt & Lima 2011).

Disease age-of-onset, progression and severity in phenotype are mostly determined by the CAG repeat size since there is a strong inverse correlation between the expanded repeat size and AO as well as the clinical presentation (Maciel *et al.* 1995).

The repeat may be of different lengths in progenitors and offspring in MJD families because of the intergenerational instability of the CAG repeat. Paternal mutant alleles are slightly more unstable than maternal ones and are more prone to expand or contract when transmitted to the next generation (Costa & Paulson 2012). This explains the phenomenon of anticipation observed MJD families, in which affected offspring tend to manifest disease earlier and more severely than an affected parent, a hallmark of repeat expansion diseases (Paulson 2009). In MJD, instability of the CAG repeat may also occur in different cells from the same tissue, this is known as somatic mosaicism. This phenomenon may occur in the brain, but so far, larger repeats have not been preferentially associated with affected brain regions (Maciel *et al.* 1997).

Cases of homozygosity are extremely rare in MJD. The special characteristics of homozygote patients suggest that gene dosage is an important determinant of age of onset and clinical phenotype in MJD, as they tend to show signs of the disease much earlier and with greater severity (Lang *et al.* 1994; Zeng *et al.* 2015).

Genetic background is another factor that seems to influence instability of repeats. Specific SNPs neighboring the repeat are shown to influence the intergenerational instability of expanded alleles. It has been suggested that an inter-allelic interaction between normal and expanded alleles in MJD drive the instability at the *ATXN3* locus (Martins *et al.* 2008).

Moreover, DNA methylation in the promoter region of the *ATXN3* gene was recently proposed to have a small, positive effect on the age at onset of MJD patients, indicating that epigenetic factors might contribute to clinical variability in MJD. This could mean that genetic and/or environmental factors also contribute to disease presentation (Emmel *et al.* 2011).

b. Ancestral mutational origins

As mentioned before, MJD shows noticeable geographic differences in prevalence worldwide. The study of ancestral origins and spreading routes of MJD mutational events has contributed to explain such differences. At least two independent *de novo* MJD expansions occurred in distinct haplotype backgrounds: TTACAC and GTGGCA (Joseph and Machado lineages, respectively). The most ancient Joseph lineage, probably of Asian origin, has been introduced recently in Europe, where founder effects are responsible for the high MJD prevalence, as occurs in the Portuguese/Azorean island of Flores and northeastern mainland. Machado lineage is geographically more restricted, with most known families in Portugal (island of São Miguel and along the Tagus valley). The hypothesis of other mutational origins has been raised, namely to explain the disease among Australian aborigines; however, a comprehensive haplotype study suggested the introduction of the Joseph lineage in that community via Asia (Martins & Sequeiros 2018).

6. Targets for siRNA therapies

In MJD, the CAG expansion confers a toxic gain-of-function on the mutant protein, leading to the formation of neuronal intranuclear inclusions. There is currently no available treatment for the disease, but recent studies have demonstrated that RNAi is a promising approach for treating autosomal dominant disorders (Matos *et al.* 2018).

Therapeutics targeting the aberrant protein products or RNA using siRNA are the most rapidly evolving technologies in the triplet field as siRNA is a powerful tool for specific silencing of disease-causing genes. When introduced into cells, the interfering RNAs cause the sequence-specific cleavage of transcripts containing the complementary sequences (Fiszer *et al.* 2012).
To date, *in vivo* studies performed with siRNAs do not discriminate between the wild-type and mutant alleles. However, the loss of function of wild-type ataxin-3, which has been shown to play a role in ubiquitin-mediated proteolysis, might also affect disease pathogenicity (Alves *et al.* 2008).

Human disease alleles often differ from their wild-type counterparts by SNPs that do not themselves cause the disease phenotype. By targeting the SNP isoform present in the disease allele, single nucleotide–specific siRNA might be used to reduce selectively expression of the disease-causing allele without altering expression of the wild-type allele (Shwartz *et al.* 2006). Allele-specific inhibition of the mutant gene expression is considered a safer and more promising strategy for RNAi therapies not only because the wild-type normal protein expression is required for cellular function, but also the effect of the efficient silencing of both alleles of a specific gene in human cells is difficult to predict (Fiszer *et al.* 2012).
With improvement of mutant allele selectivity (preserving the expression of the wild-type allele), target specificity and safety, as well as optimization of delivery methods, the efficacy of siRNA-based therapeutic applications is likely to improve (Budworth & McMurray 2013).

Goals

The impact of CAG repeat length on age-at-onset (AO) of expansion-association diseases, their phenotype and progression is well known and well-studied. Nevertheless, even though there is an inverse correlation between the size of the expanded CAG repeat tract and disease AO, this does not explain the heterogeneous nature of MJD to full extent. As such, it is important to consider other genetic variations as a possible source of variation of the clinical features of MJD. This project aimed to explore SNPs found in the *ATXN3* gene as a source of MJD clinical variability.

Our specific goals were:

- To optimize multiplex PCRs for the amplification of exonic and regulatory regions of *ATXN3*;
- To search for variants in the *ATXN3* gene and calculate the percentage of their influence on MJD age-of-onset variability, after removing the known effect of the expansion size;
- To identify SNPs that may be useful in distinguishing between the wild type allele from the expanded disease-causing allele, for future applications, such as siRNA therapies.

Subjects and Methods

1. Subjects

We studied 120 MJD patients, belonging to different Portuguese families. The 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. Age-at-onset (AO), defined as the age of appearance of gait disturbance and/or diplopia, reported by the patient and/or a close relative, was recorded during clinical assessments. The length of normal and expanded polyglutamine tracts was previously determined by capillary electrophoresis. Sample ID, family, gender, length of CAG repeats in the normal and expanded alleles and AO of patients are annotated in Table 2. DNA quantification was performed with Nanodrop spectrophotometer to make work aliquots with a final DNA concentration of approximately 7.5 ng/µL.

2. Methods

a. Primer design

To amplify and sequence exons 1, 4, 5, 6 ,7, 8, 9, 10 (containing the CAG repeat) and 11 of the *ATXN3* gene, we used primers previously outlined in our lab (Martins *et al.* 2012) (Table 3). Additional primers where designed to amplify exons 2 and 3 using the online software Primer3 (Rozen & Skaletsky 1999), with the following conditions: 18 to 27 bp; a melting temperature comprised between 57°C and 65°C; and a percentage of GC between 20% and 80%. 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 tools BLAT (Kent 2002) (human genome, assembly GRCh38/hg38) and BLAST (Johnson *et al.* 2008) were then used to guarantee the specificity of designed primers to the target sequence (database: nucleotide collection; organism: human; expected threshold: 10; megablast). The compatibility of multiple primer sequences was checked with AutoDimer software (Vallone & Butler 2004).

Table 2 - Clinical information and genotyping data of the $(CAG)_n$ and SNPs found in analyzed regions of *ATXN3*. E – amplified exon.

Sample ID	Gender	$(CAG)_n$		Age-at-onset (years)	Polymorphisms																		
					ATXN3																		
					E1	E3		E4/5				E6	E7			E9			E10				E11
		Expanded Allele	Normal Allele		rs3814834	rs7493524	rs12147767	rs1997920	rs1997919	rs1997917	rs4904834	rs8003041	rs16999141	rs10146519	rs1048755	rs761553	rs761552	rs761551	rs12896583	rs12895357	rs7158733	rs3092822	rs150082297
1		70	23	45		C/T	G/A					G/A	C/T	G/T	G/A								
2		77	25	37	C/C			C/T	G/A	G/A	G/C	G/A	C/T		G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
3		77	21	21								G/A	T/T		G/A								
4	M	71	15	48	C/C			C/T	G/A	G/A	G/C	G/A	C/T		G/A	G/C	C/T	C/A	A/A	G/G	A/A	A/A	T/T
5	F	66	28	41	C/C	T/T	G/G	T/T	A/A	G/G	G/G	G/G	T/T		A/A	C/C	T/T	A/A	A/A	C/C	C/C	A/A	T/T
6	F	67	25	60	C/C	C/T	G/A	C/T	G/A	G/A	G/C	G/A	C/T		G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
7	F	73	23	33		C/T	A/A					A/A	C/T		G/G								
8	F	79	23	18	C/C			C/C	G/G	A/A	G/C	A/A	T/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
9	M	75	21	35	C/C	C/T	A/A					A/A	C/T		G/G	G/C	C/T	C/A	A/A	G/C	A/A	C/C	T/T
10	M	71	15	39		C/T	A/A					A/A	C/T		G/G								
11	M	74	24	40	C/C			C/C	G/G	A/A	G/C		C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
12	F	68	23	55	C/C			C/C	G/G	A/A	G/C		T/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
13	M	74	28	40	C/T			C/C	G/G	A/A	G/C		T/T		G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
14	M	71	23	60	C/C			C/C	G/G	A/A	G/C		C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
15	F	69	27	43	C/C	C/T	G/A	C/T	G/A	G/A	G/G		T/T		G/A	C/C	T/T	A/A					T/T
16	F	75	23	40									T/T		G/A								
17	F	77	24	25	C/C			C/T	G/A	G/A	G/C		C/T		G/G	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
18	M	68	14	49	C/C	C/T	G/A	C/T	G/A	G/A	G/C		C/T		G/A	G/C	C/T	C/A	A/A	G/G		A/A	T/C
19	M	65	22	48	C/C	C/T	G/A	C/C	G/G	A/A	G/G	G/A	T/T		G/G	G/G	C/C	C/C	G/G	G/G	C/C	A/A	T/T
20	M	71	24	40	C/C	C/T	A/A						C/T		G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
21	M	73	24	26	C/C			C/T	G/A	G/A	G/C	G/A	C/T	G/T	G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
22	M	70	24	51								A/A	C/T	T/T	G/G								
23	M	76	15	31				C/C	G/G	A/A	G/G		T/T		G/A	G/C	C/T	C/A	A/A	C/C	A/A	C/C	T/T
24	F	70	23	50	C/C			C/T	G/A	G/A	G/G		T/T		G/G	G/C	C/T	C/A	A/A	C/C	A/A	C/C	T/T
25	F	69	23	48	C/C	C/T	G/G	C/C	G/G	A/A	G/C		T/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
26	F	73	28	50	C/C	C/C	G/A					G/A	C/T		G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/A	T/T
27	M	73	26	40	C/C	T/T	G/G	C/T	G/A	G/A	G/C		T/T		A/A	G/C	C/T	C/A	A/A	G/C	A/C	A/A	T/T
28	F	76	14	24	C/C	C/T	G/A	T/T	A/A	G/G	G/G	G/A	C/T		G/A	C/C	T/T	A/A	A/A	C/C	A/A	C/C	T/T
29	M	73	28	30	C/C	C/T	A/A	T/T	A/A	G/G	G/G	A/A	T/T		A/A	C/C	T/T	A/A	A/A	C/C	A/A	C/C	T/T
30	M	77	19	20	C/T	C/T	G/A	C/T	G/A	G/A	G/C		T/T	G/T	G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
31	M	72	14	55	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/C	A/C		T/T
32	M	79	28	16	C/C			C/T	G/A	G/A	G/G	A/A	T/T		G/A	G/C	C/T	C/A	A/A	G/G	C/C	A/A	T/T
33	F	69	23	35	C/C	C/T	A/A	C/C	G/G	A/A	G/C	G/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
34	M	69	24	44	C/C	T/T	G/A	C/C	G/G	A/A		G/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
35	F	72	28	37	C/C	C/T	A/A	C/T	G/A	G/A	G/G	A/A	T/T		G/G	G/C	C/T	C/A	A/A	G/G	A/A	C/C	T/T
36	M	80	19	13	C/C	C/T	A/A	C/T	G/A	G/A	G/G	A/A	T/T	G/T	G/A	G/C	C/T	C/A	A/A				T/T
37	M	73	14	36								A/A	C/T	T/T	G/G								T/T
38	M	69	23	42		C/T	A/A					A/A	C/T		G/G								T/T
39	F	68	23	46	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
40	F	69	14	45	C/C	T/T	G/A	C/C	G/G	A/A	G/C	G/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
41	F	80	23	28	C/C			C/C	G/G	A/A	G/C	A/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
42	F	77	27	36	C/C	C/T	G/A	C/T	G/A	G/A	G/G	G/A	T/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
43	M	69	24	48	C/C			C/C	G/G	A/A	G/C	A/A	C/T	G/T	G/A	G/G	C/C	C/C		G/G	C/C	A/A	T/T
44	M	73	30	44	C/T	C/T	A/A	C/T	G/A	G/A	G/G	A/A	T/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
45	F	72	23	55		C/T	G/A					G/A	C/T	T/T	G/G	G/G	C/C	C/C					T/T
46	F	68	23	48	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
47	M	72	23	43	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
48	F	67	23	52	C/C							A/A	C/T	T/T	G/G	G/G	C/C	C/C					T/T
49	M	71	23	40	C/C			C/C	G/G	A/A	G/C	G/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
50	F	71	23	47	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	T/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
51	F	69	21	45	C/C			C/T	G/A	G/A	G/G	G/A	T/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
52	F	72	14	42	C/C	C/T	G/A					G/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
53	M	69	23	42	C/C			C/T	G/A	G/A	G/C	G/A	C/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
54	M	76	27	25	C/C		A/A	C/C	G/G	A/A	G/C	A/A	T/T	G/T	G/A	G/G	C/C	C/C		G/C	A/C	A/C	T/T
55	F	65	14	60	C/C			C/C	G/G	A/A	G/C	G/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
56	F	70	23	65	C/C	T/T	G/G	C/C	G/G	A/A	G/C	G/G	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
57	F	68	27	60	C/C	T/T	G/A	C/C	G/G	A/A	G/C	G/A	T/T	G/T	G/A	G/G	C/C	C/C		G/G	C/C	A/A	T/T
58	M	73	28	49	C/C	C/T	A/A	C/T	G/A	G/A	G/G	A/A	C/T	T/T	G/G	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
59	M	72	21	40	C/C			C/T	G/A	G/A	G/G	G/A	T/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
60	F	69	28	55		C/T	A/A					G/A	T/T		G/G								T/T
61	F	75	32	49	C/C			C/T	G/A	G/A	G/G	G/G	T/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
62	F	76	21	35	C/C			T/T	A/A	G/G	G/G	A/A	T/T	G/G	A/A	C/C	T/T	A/A		C/C	A/A	C/C	T/T
63	F	69	33	49	C/C			C/T	G/A	G/A	G/C	G/A	T/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
64	M	75	23	45	C/C			C/C	G/G	A/A	G/C	G/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
65	M	68	23	47	C/C	T/T	G/A	C/T	G/A	G/A	G/C	G/A	C/T	G/T	G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
66	M	70	26	42	C/C	T/T	G/A	C/C	G/G	A/A	G/C	G/A	T/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
67	M	69	14	55	C/C	C/T	G/A	C/C	G/G	A/A	G/C	G/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
68		82	23	15	C/C		G/A	C/T	G/A	G/A	G/C	G/A	T/T	G/T	G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
69	F	74	14	37	C/C			C/C	G/G	A/A	G/C	A/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
70	F	73	24	40	C/C	T/T	G/A	C/C	G/G	A/A	G/C	G/A	T/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
71	F	76	21	31	C/C	C/C	A/A	C/T	G/A	G/A	G/G	A/A	T/T	G/T	G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
72	F	63	21	54	C/C			C/T	G/A	G/A	G/G	A/A	T/T	G/T	G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
73	F	69	14	37	C/C		G/A	C/C	G/G	A/A	G/C	G/A	C/T	T/T	G/A	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
74	F	70	27	42	C/C	T/T	G/G	C/T	G/A	G/A	G/G	G/G	T/T	G/T	G/G	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
75	F	81	23	33	C/C			C/T	G/A	G/A	G/C	G/A	C/T		G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
76	M	66	23	62	C/C	C/T	G/A	C/C	G/G	A/A	G/C	G/A	T/T	T/T	G/A	G/G	C/C	C/C		G/G	C/C	A/A	T/T
77	M	75	24	27	C/C			C/C	G/G	A/A	G/C	A/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
78	M	67	21	45	C/C	C/T	G/G	C/T	G/A	G/A	G/G	G/A	T/T		G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
79		72	28	44	C/T	T/T	G/G	C/T	G/A	G/A	G/G	G/G	T/T	G/T	G/G	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
80	M	65	26	41	C/T	C/T	A/A	C/C	G/G	A/A	G/G	A/A	T/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
81	F	68	14	50	C/C	C/T	A/A	C/C	G/G	A/A	G/C		C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
82	F	75	14	35	C/C			C/T	G/A	G/A	G/C	G/A	C/T		G/A	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
83	M	79	21	18	C/C	C/T	A/A	C/T	G/A	G/A	G/G	A/A	T/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
84	F	72	27	54	C/C			T/T	A/A	G/G	G/G	A/A	T/T		A/A	C/C	T/T	A/A		C/C	A/A	C/C	T/T
85	F	72	23	37	C/C			C/T	G/A	G/A	G/C	A/A	C/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
86	F	71	23	55	C/C			C/C	G/G	A/A	G/C	G/A	C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
87	F	68	27	67	C/C			T/T	A/A	G/G	G/G	A/A	T/T		A/A	C/C	T/T	A/A	A/A	C/C	A/A	C/C	T/T
88	F	73	21	20	C/C			T/T	A/A	G/G	G/G		T/T	G/G	A/A	C/C	T/T	A/A	A/A	C/C	A/A	C/C	T/T
89	M	68	27	46	C/T			C/C	G/G	A/A	G/G	A/A	T/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
90	M	69	23	60	C/C			C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
91	M	69	23	55	C/C			C/C	G/G	A/A	G/C		C/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
92	M	75	27	30	C/T			C/T	G/A	G/A	G/G	G/A	C/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
93		76	22	34	C/C	C/T	G/A	C/C	G/G	A/A	G/C	G/A	T/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
94		71	23	65	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	T/T	T/T	G/G	G/G	C/C	C/C					T/T
95		77	23	38	C/C	C/T	A/A					A/A	C/C	T/T	G/G	G/G			A/A	G/G	C/C	A/A	T/T
96		67	23	60	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C					T/T
97		73	23	47	C/C	C/T	A/A					A/A	C/T			G/G	C/C	C/C		G/G	C/C	A/A	T/T
98	F	76	25	30	C/T	C/T	G/A	C/C	G/G	A/A	G/C	G/A	C/T		G/A	G/C	C/T	C/A		G/G	C/C	A/A	T/T
99	M	69	28	65								A/A	T/T		G/G								T/T
100		74	20	44	C/C			T/T	A/A	G/G	G/G	G/G	T/T			C/C	T/T	A/A		C/C	A/A	C/C	T/T
101		74	23	43	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	T/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
102		73	28	46	C/C			C/T	G/A	G/A	G/C	G/A	T/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
103		66	25	66	C/C	C/T	G/A	C/T	G/A	G/A	G/C	G/A	T/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
104		70	14	48	C/C			C/T	G/A	G/A	G/C	G/A	C/T	G/T	G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
105		73	30	42	C/C		A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
106		72	14	58	C/C			C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
107		77	21	40	C/C			C/T	G/A	G/A	G/G	G/A	T/T		G/G	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
108		77	28	29	C/C			C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
109		74	23	46	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
110		68	27	48	C/C	T/T		T/T	A/A	G/G	G/G	G/G	T/T	T/T	A/A	C/C	T/T	A/A		C/C	A/A	C/C	T/T
111	M	66	28	70	C/C			C/C	G/G	A/A	G/G	A/A	T/T		G/A	G/C	C/T	C/A		G/C	A/C	A/C	T/T
112	M	75	26	15	C/T	C/T	A/A	C/C	G/G	A/A	G/G	A/A	T/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
113	F	72	23	47	C/C	T/T	G/A	C/T	G/A	G/A	G/C	G/A	C/T		G/G	G/C	C/T	C/A	A/A	G/C	A/C	A/C	T/T
114	F	69	23	63	C/C			C/C	G/G	A/A	G/C	G/A	T/T		G/A	G/G	C/C	C/C		G/G	C/C	A/A	T/T
115	M	74	20	33	C/C			C/C	G/G	A/A	G/C	A/A	T/T		G/G	G/G	C/C	C/C	A/A	G/G	C/C	A/A	T/T
116		75	30	38	C/C	C/C		C/C	G/G	A/A		A/A	C/C	T/T	G/G	G/G	C/C	C/C		G/G	C/C	A/A	T/T
117		70	14	30	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/C		G/G	G/G	C/C	C/C		G/G	C/C	A/A	T/T
118		72	14	56	C/C			C/T	G/A	G/A	G/C	A/A	C/T		G/G	G/C	C/T	C/A		G/C	A/C	A/C	T/T
119		69	14	50	C/C	C/T	A/A	C/C	G/G	A/A	G/C	A/A	C/T	T/T	G/G	G/G	C/C	C/C		G/G	C/C	A/A	T/T
120		72	23	45	C/C			C/C	G/G	A/A	G/C	A/A	T/T		G/G	G/G	C/C	C/C		G/G	C/C	A/A	T/T

b. PCR optimization

In this work, we optimized two multiplex PCRs with the primers described above to cover coding and regulatory sequences of *ATXN3*: one pentaplex (to analyze exons 1, 4, 5, 9, 10 and 11), and one triplex (to analyze exons 2, 3, 6, 7 and 8). For both, a touchdown strategy was performed to increase specificity and avoid the amplification of nonspecific sequences. The pentaplex amplification reaction had a final volume of 13 µl, with the reagents described in Table 4, and PCR conditions in Table 5. The triplex amplification reaction had a final volume of 12 µl, with the reagents described in Table 6, and PCR conditions in Table 7.

Table 3 - Primers for amplification and sequencing of exonic and regulatory regions of ATXN3.

Name	Primer Orientation (F-forward; R-reverse)	Primer sequence (5'-3')
ATXN3 UTR5'	F	TGCTCCTTAATCCAGGGAAATTTAG
ATXN3 R1	R	AGATCGGCATGGGGGCGACT
ATXN3 F2 NM	F	ACAATTCTGTTTAACAAGTCACATTT
ATXN3 R3 NM	R	GCCTGGGATGTAAGCAACAC
ATXN3 F6	F	CACAACTAACATAGCTACACTTCC
ATXN3 R7	R	GACAGGACCTCCCTTTGTTGCC
ATXN3 F9	F	CCTGGCCAATGTGGCAAATG
ATXN3 R9	R	CACTGTCATCTAATGTGCCTG
ATXN3 F11	F	TCCAGTGTTCTGTGCTGCCTTTT
ATXN3 R12	R	GAAATCTAAAGGGAAAGCCCAC
ATXN3 F12	F	GGTTGCAGTTATTACCAGTGC
ATXN3 R11	R	AGTCGCCAACAACACAAGGACC
ATXN3 F13	F	ACGCCCAGCCAGAAGAGTAG
ATXN3 R14	R	CTCCTGACCTCAGGCAATCTG
ATXN3 F18	F	CCACTCCTGGCCATGATAGGT
ATXN3 R20'	R	TCCTCTCCTGCCTTGGTTTCCC
ATXN3 F21	F	CTGGTGGCTATCTGGGATTAGGA
ATXN3 R21	R	GGACCCTATGCTGTAATCACACAG
MJD Clo 7aR	R	TGCTCCTTAATCCAGGGAAATTTAG

Table 4 - PCR reagents for the optimized pentaplex reaction.

Reagents	Volume (µL)
2x Multiplex Master Mix (QiaGen®)	6
5x Q solution (QiaGen®)	1
Primer mix (2.5 µM)	5
DNA sample (7.5 ng/µL)	1
Total	13

Table 5 - PCR protocol for the optimized pentaplex reaction.

95ºC	**94ºC**	**63ºC**	**72ºC**	**94ºC**	**62ºC**	**72ºC**	**94ºC**	**61ºC**	**72ºC**	**70ºC**
15'	30''	1'30	1	30''	1'30	1'	30''	1'30	1'	10'
	10 cycles				10 cycles			15 cycles		

Table 6 - PCR reagents for the optimized triplex reaction.

Reagents	Volume (µL)
2x Multiplex Master Mix (QiaGen®)	5
5x Q solution (QiaGen®)	1
Each primer (x6)	0.5 (x6)
DNA sample (7.5 ng/µL)	1
H_2O	2
Total	12

Table 7 - PCR protocol for the optimized triplex reaction.

95ºC	**94ºC**	**58ºC**	**72ºC**	**94ºC**	**57ºC**	**72ºC**	**70ºC**
15’	30’’	90’’	1’30	30’’	90’’	1’30	10’
		20 cycles			20 cycles		

c. Detection of amplified products

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-FixTM, Serva), were used to produce a 3 mm thick gel composed by an acrylamide-bisacrylamide solution, ammonium persulfate and TEMED (Table 8).

Table 8 - Reagents used in the polyacrylamide gel.

Reagents	Volume	Function
40% (w/v) Acrylamide:bisacrylamide (19:1) solution (AccuGel19:1; National Diagnostics)	3 mL	Gel
Ammonium persulfate 2.5% (Promega)	170 µL	Oxidizing agent
TEMED (Amersco®)	10 µL	Catalysis agent

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

To allow the horizontal run, we used paper strips, soaked in buffer, which were placed at both anode and cathode. The addition of bromophenol blue dye (Merck) to the cathode allowed the visualization and control of the electrophoresis. The electrophoresis run at 180 V and the electrophoresis power supply (Consort EV243) was turned off when the marked dye approached the anode strip. The coloration method involved six steps, described in Table 9, at room temperature and under agitation (Burgwedel). When the process was concluded, we left the gel in water, for at least 12 hours, and then dried at room temperature (RT).

Table 9 - Steps of the polyacrylamide gel coloration with silver staining.

Step	Time	Reagents
(1) DNA fixation		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) DNA fragments revelation	Until the appearance of bands	0.28 M sodium carbonate anhydrous (Applchem) and 0.02% formaldehyde (Merck)
(6) Stop the revelation	~10 s	10% acetic acid (Merck)

d. DNA sequencing

After confirming PCR amplification and specificity, DNA sequencing was performed with the resulting fragments using the respective sequencing primers. An initial purification has done with an exonuclease and an enzyme thermosensitive alkaline phosphatase to degrade unreacted primers and remaining dNTPs (Table 8 and 9). Together with the purified product (2.8 µL), a sequencing mix (Table 10) was subjected to a sequencing reaction with the conditions described in Table 11.

Table 10 - Reagents and quantities for PCR pre-sequencing purification.

Reagents	**Volume (µL)**
ExoSap : FastAPP (1:4)	0.95
PCR product	1.85
Total	2.80

Table 11 - Sequencing reaction conditions.

96ºC	**96ºC**	**58ºC**	**60ºC**	**60ºC**
2’	15’’	5’	2’’	10’
		30 cycles		

Table 12 - Reagents and volumes for sequencing reactions.

Reagents	**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 (ddH2O)	0.75
Total	2.50

Table 13 - Protocol for PCR purification pre-sequencing purification.

Enzyme activation	**Enzyme inactivation**
37ºC	80º C
15 min	15’

An ultimate purification was performed to isolate the sequencing products, using a cross-linked dextran matrix in order to separate molecules over broad molecular weight. The centrifugation (Mikro 200, Hettich Zentrifugem) of illustraTM SephadexTM 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^{TM} formamide (Applied Biosystems) to the final product. This stability is necessary for capillary electrophoresis in an ABI PRISM 2130x/Genetic Analyzer (Applied Biosystems).

e. Statistical and *in silico* analyses

The sequences obtained were aligned with Geneious 5.5.8 software (Kearse *et al.* 2012), which allowed global alignments and comparison with the reference sequence. The analysis of *ATXN3* diversity was accomplished by annotating all alterations to the reference sequence.

We started by calculating the proportion of gender distribution of our sample, and by producing histograms for AO and CAG repeat length for both normal and expanded alleles (with respective mean and standard deviation), followed by the D'Agostino-Pearson omnibus normality test. The association between the deleterious CAG repeat length and AO was assessed by fitting with a linear regression model. The equation of the regression line, the correlation coefficient (r) and the coefficient of determination (R^2) were determined. Allele frequencies were estimated for all analyzed variants and Hardy-Weinberg equilibrium was tested, applying the Bonferroni correction. An analysis of covariance (ANCOVA) was conducted to determine the effect of the minor allele of each polymorphism on disease AO (dependent variable) using expansion size as a covariate. Allelic variants with an ANCOVA significant effect were used, together with the CAG repeat size of expanded alleles, to determine a new model to predict disease AO, by using a multiple linear regression technique. Finally, a t-test was used to see if there was a significant difference between means of AO an expansion size when the minor allele was present *versus* when the minor allele was absent.

All statistical analyses were performed in IMB SPSS Statistics 25 (IBM Corp. Released 2017) with the exception of the Hardy-Weinberg equilibrium test, executed in Arlequin 3.11 software; graphs were made in GraphPad Prism 7.05. A statistically significant result lower than 0.05 was considered for all tests performed.

Results

1. Clinical and genotyping data

Genotyping data for this work has been obtained within the scope of a Norte2020 Project (NORTE-01-0145-FEDER-000008; PI: Sandra Martins), with part of patients' genotypes obtained by Inês Costa (Research fellow).

We analyzed a total of 120 MJD patients belonging to different families of Portuguese origin. Reported AO varied between 13 and 70 years with a mean of 42,77 ± 12,37 years. Most patients were diagnosed with an AO between 40-50 years (44%) (Figure 2).

Concerning the disease-causing repeat in the *ATXN3* gene, we found that normal allele size fell between 14 and 33 repeats, with a mean size of 22,66 ± 4,61 repeats; as for the expanded alleles they fell between 63 and 82 repeats and their mean was 71,88 ± 3,80, with a major occurrence of 69 repeats, which alone accounted for 15% of patients.

We performed the D'Agostino-Pearson omnibus normality test on all 3 variables. For AO and CAG repeat size of expanded alleles, the null hypothesis could not be rejected, meaning that they displayed a normal distribution ($p > 0.05$; p-values of 0.558 and 0.352 for the AO and expansion size, respectively). On the other hand, normal sized *ATXN3* alleles did not fit into a normal distribution (p=0.0456).

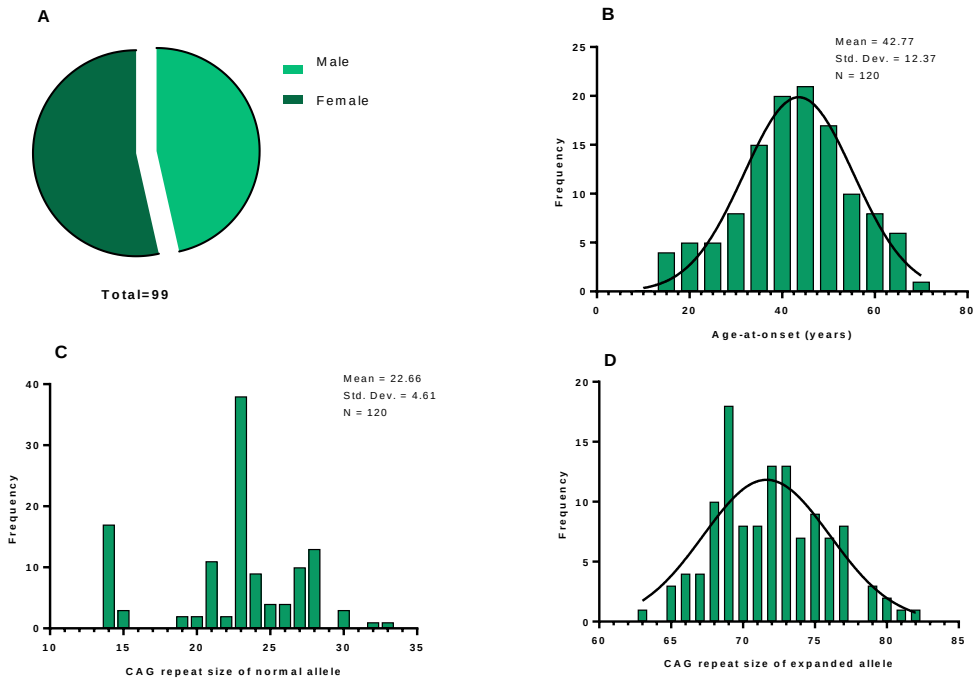


Figure 2 – Clinical and genotyping data for our cohort of 120 Portuguese MJD patients. Distribution of patients according to (A) gender (N=99), (B) diagnosed AO (N=120), (C) CAG repeat size in the normal (N=120) and (D) expanded (N=120) *ATXN3* alleles. Variables AO and CAG repeat size in expanded alleles follow a normal distribution.

2. Correlation between the expanded CAG repeat length and MJD AO

After calculating the linear regression between AO and the CAG repeat size of the expanded allele, we observed a negative correlation between the two variables (N=120; R^2=0.5297; p<0.0001), with a regression line given by: (AO) = -2.367 ($(CAG)_n$ of expanded allele) + 212.9 (Figure 3). The expanded allele size was shown to account for 52.97% of AO variance. No relationship was found between AO and gender (p=0.6339) or CAG repeat size in normal alleles (p=0.3502).

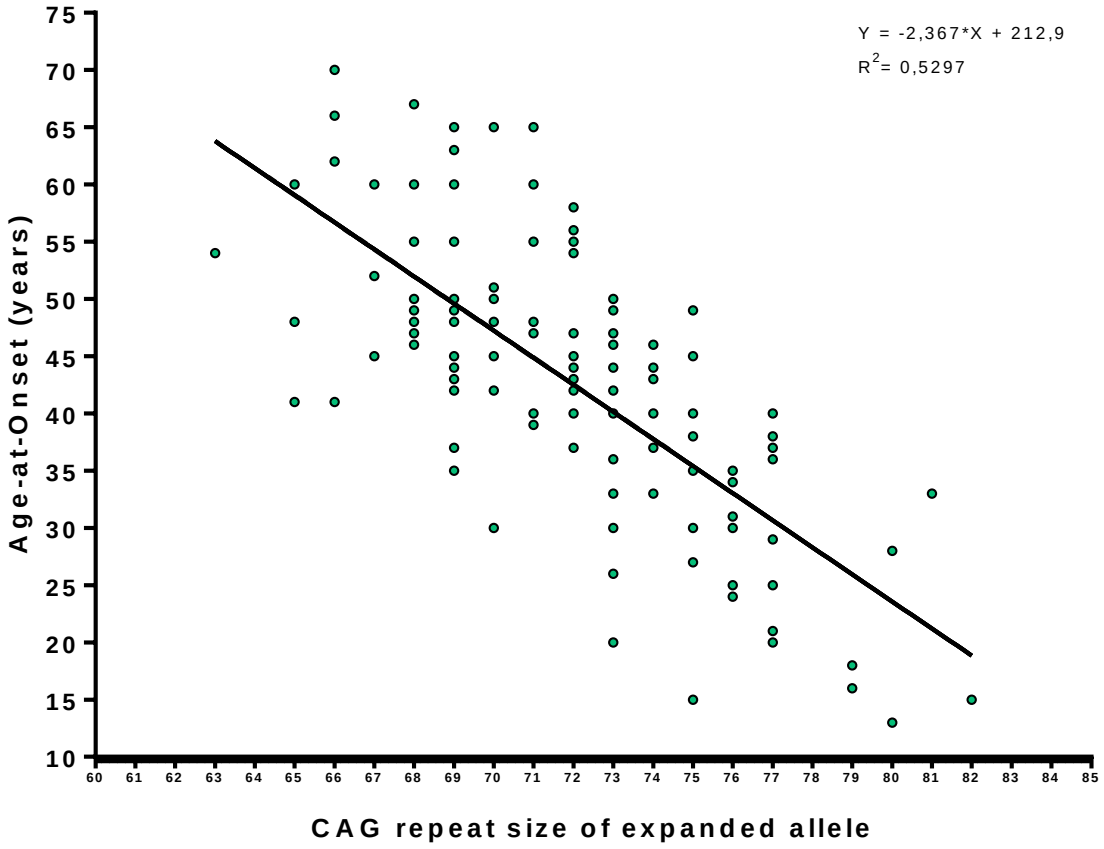


Figure 3 - Linear regression between CAG repeat length in the MJD chromosome of 120 patients and disease AO. A linear coefficient of determination of r^2=0.5297 was obtained.

3. Genetic variation of *ATXN3* in MJD patients

To analyze the genetic variation of the *ATXN3* gene, we sequenced all 11 exons, as well as the 5'UTR and intronic areas surrounding these exons that could affect splicing. In our sample of 120 MJD patients, we found 18 variants distributed throughout the analyzed regions (genotypes for the 120 patients are detailed in Table 2).

Located within exon 1 in the 5'UTR region, rs3814834 is a variant that may affect the regulation of the *ATXN3* gene. Also of interest, we found a non-coding transcript exon variant rs7493524, in the region surrounding exon 3, in addition to the intronic variant rs12147767. In the area between exons 4 and 5, we identified 4 SNPs, 3 non-coding transcript exon variants (rs1997920, rs1997919 and rs1997917) and one intronic variant (rs4904834). An intron variant was also identified in the surrounding region of exon 6 (rs8003041). Within exon 7, we identified a missense alteration which, when present, results in the substitution of a serine with a leucine (rs16999141). We also found a missense alteration that results in a change from valine to methionine within exon 8 (rs1048755) as well as an intronic polymorphism in the surrounding region (rs10146519). Three other polymorphisms were detected in the intronic region of exon 9 (rs761553 rs761552 and rs761551). As for the CAG repeat-containing exon 10, we identified 4 SNPs: one synonymous (rs12896583), one missense, resulting in a change from glycine to arginine (rs12895357), a stop codon gaining alteration (rs7158733) and a 3'UTR variant (rs3092822). Finally, we found another variant in the 3'UTR surrounding exon 11 (rs150082297) that might be also relevant for post-transcriptionally regulation of *ATXN3* expression.

4. The influence of *ATXN3* SNPs on MJD AO

Conformity with the Hardy-Weinberg (HW) equilibrium expectations was tested for each analyzed SNP (Table 14). Of the 18 SNPs included in this study, five were not in conformity with HW equilibrium, which led us to perform the Bonferroni correction (p/N=0.05/18) reducing the threshold for equilibrium to p>0.0028 (p/N=0.05/18). After the correction, rs7493524 and rs10146519 were shown to be in accordance with HW equilibrium, while three SNPs still displayed an excess (rs16999141 and rs3092822) or deficit (rs7158733) of heterozygotes. After confirming that all three genotypes (common homozygous, heterozygous and rare homozygous) have been genotyped by us for these SNPs, we excluded the possibility of allelic dropout (i.e. loss of one allele during PCR amplification of DNA). Therefore, assuming that this disequilibrium may be related to other factors rather than genotyping errors, we proceeded to next analyses with all 18 SNPs. For each SNP, we used analysis of co-variance (ANCOVA) to investigate the association of the presence of the minor allele with AO of patients, while having in consideration the known effect of the CAG expansion (Table 15). Our results have shown that rs3814834*T (F=5.404, p<0.05), rs4904834*G (F=5,615, p<0.05), and rs10146519*G (F=4,896, p<0.05), had a significant impact on MJD AO.

Table 14 – Analyses of conformity with Hardy-Weinberg (HW) equilibrium expectations.

SNP ID	Observed/expected heterozygosity	HW p-value
rs3814834	0,083/0,080	1
rs7493524	0,742/0,491	0,00005*
rs12147767	0,391/0,427	0,55777
rs1997920	0,382/0,404	0,62457
rs1997919	0,382/0,404	0,62639
rs1997917	0,382/0,404	0,62387
rs4904834	0,66/0,444	0*
rs8003041	0,427/0,407	0,8066
rs16999141	0,500/0,400	0,00575
rs10146519	0,382/0,354	0,71031
rs1048755	0,390/0,389	1
rs761553	0,418/0,422	1
rs761552	0,422/0,424	1
rs761551	0,471/0,437	0,52486
rs12896583	0/0,027	0,00678
rs12895357	0,317/0,391	0,07647
rs7158733	0,311/0,407	0,02692
rs3092822	0,281/0,384	0,00955*
rs150082297	0,009/0,009	1

*Equilibrium after Bonferroni correction

5. Addition of polymorphisms effect to the AO/CAG linear regression model

To test the cumulative effect of the 3 SNPs shown to influence the AO in previous tests (rs3814834, rs4904834 and rs10146519), we applied a different model that, in addition to the CAG repeat size in the expanded allele, also takes into account this set of SNPs. All variables analyzed together, in addition to the expansion size, were shown to predict 60.9% of AO variation in our sample (R^2=0.609). Considering that 52.97% of AO was explained by the expansion size alone, the addition of the effect of these SNPs improved the model by 7.9%.

To estimate the AO including the group of SNPs as well as the CAG expansion size, we applied the following formula: (AO) = 200.213 – 2.211 x ($(CAG)_n$ in expanded allele) – 7.822 x (presence/absence of rs3814834*T) + 2.299 x (presence/absence of rs4904834) – 1.015 x (presence/absence of rs10146519)). For this new model, the presence of rs3814834, rs4904834 and rs10146519 improve the estimation of MJD AO by 4.1%, 3.3% and 11.5% respectively.

Table 15 – SNPs found in the analyzed regions of the *ATXN3* gene and their influence on the AO of MJD patients after accounting for the known effect of the CAG expansion size. Sequence alterations refer to the reverse strand (transcript ATXN3-247).

		Sequence	ANCOVA test results	
Region	**SNP ID**	**alteration**	**F ratio**	**P-value**
Exon 1	**rs3814834**	**C/T**	**5.404**	**0.022**
	rs7493524	T/C	0.401	0.671
Exon 3	rs12147767	A/G	0.186	0.831
	rs1997920	C/T	0.691	0.504
Exon 4/5	rs1997919	G/A	0.691	0.504
	rs1997917	A/G	0.691	0.504
	rs4904834	**G/A**	**5.615**	**0.02**
Exon 6	rs8003041	A/G	0.187	0.833
Exon 7	rs16999141	T/C	0.707	0.495
	rs10146519	**T/G**	**4.896**	**0.011**
Exon 8	rs1048755	G/A	1.101	0.339
	rs761553	G/C	0.933	0.397
Exon 9	rs761552	C/T	0.846	0.432
	rs761551	C/A	0.691	0.56
	rs12896583	A/G	1.593	0.21
Exon 10	rs12895357	G/C	0.881	0.418
	rs7158733	C/A	0.719	0.49
	rs3092822	A/C	0.329	0.72
Exon 11	rs150082297	T/C	0.123	0.726

6. Influence of SNP minor alleles on mean AO

To quantify the influence of our SNPs on disease AO, we first compared the effect of the presence and absence of each minor allele on the mean AO (Table 16). An earlier onset of about 9 years was observed in the presence of rs3814834*T allele when compared to its absence, however, difference was on the edge of significance (p=0.05), which will encourage us to increase sample size for this SNP. Interestingly, rs4904834*C and rs10146519*G were found associated to a later (of about 3 years; p<0.05) and earlier (8 years; p<0.05) onset of MJD. Taking into account that this effect has been observed in the presence of allelic variants, a dominant effect of these SNPs may be underlying disease AO.

To test the hypothesis of a dominant model for the influence of our SNPs on MJD AO, we decided to analyze the effects of patients' genotypes; to control the co-variate $(CAG)_{exp}$, we included the mean expanded allele size, highlighting, this way, differences in mean AO caused by the SNP genotype, and not due to the effect of the expansion size (Table 17).

When comparing the effect of rs3814834 genotypes on disease onset, patients heterozygous and homozygous for the commonest allele have shown to have significant AO differences (9 years; p<0.305). This suggests a dominant or co-dominant model for the effect of rs3814834, though we did not find patients homozygous for the minor allele. The opposite was observed for rs4904834, since the difference between mean AO of heterozygous and major homozygous patients was not significant (5 years; p<0.05), though we also did not find any patient homozygous for the minor allele in our cohort. Interestingly, for rs10146519 (the only significant SNP for which patients homozygous for both minor and major alleles were found), when comparing mean AO for each genotype we found a gradient: heterozygous T/G patients displayed a mean AO of less 8 years than T/T patients, and homozygous a difference of additional 12 years, suggesting a dominant model with dosage effect for the influence of rs10146519 on MJD AO. Significant differences have been found when comparing AO between these three genotypes (p>0.05).

Table 16 – Mean, standard deviation (SD) and t-test results of MJD AO when the presence or absence of SNP minor alleles is observed.

Gene	SNP ID	Minor Allele	Age-at-onset			N
			Mean	SD	P-value	
ATXN3	**rs3814834**	Presence	34,44	11,27	0,05	9
		Absence	43,14	12,36		207
	rs4904834	Presence	44,32	12,32	0,198	66
		Absence	41,83	13,07		134
	rs10146519	Presence	38,08	14,03	0,003	25
		Absence	46,19	11,08		85

Next, we evaluated models A and B by comparing AOs predicted by each one with observed values (Figure 4). The most accurate model will be the one with result closest to 1 for all predictions. When only the expansion size was considered (model A), we found ratios that are further from 1 than when the effect of SNPs is also considered. Nevertheless, there does not seem to be a tenancy to over- or under-predict for either model.

Table 17 – Mean, standard deviation (SD) and t-test and/or ANOVA results of MJD AO for each SNP genotype found and corresponding mean expansion size.

Gene	SNP ID	Minor Allele	Age-at-onset			Expansion Size			N
			Mean	SD	P-value	Mean	SD	P-value	
ATXN3	**rs3814834**	C/C	43,54	12,32	0,035*	71,79	3,91	0,469*	99
		C/T	34,44	11,27		72,78	3,93		9
	rs4904834	G/G	39,41	13,52	0,071*	72,03	4,48	0,762*	34
		G/C	44,32	12,32		71,77	3,74		66
	rs10146519	T/T	48,19	9,26	0,007**	70,91	3,16	0,33**	32
		T/G	40,1	14,08		71,95	4,76		21
		G/G	27,5	10,6		74,5	2,12		2

*similar P-values have been observed with both t-test and ANOVA
**ANOVA

7. Phenotypic influence of *ATXN3* polymorphisms

After showing the effect of three *ATXN3* polymorphisms in the improvement of a model to predict MJD AO (figure 5), we next attempted to justify their role on disease phenotype.

The variant rs3814834 is situated in the regulatory region of *ATXN3*, which may result in alterations on the expression of this gene. We used UCSC Genome Browser to gain insight into transcription factors that could bind to the region that encompasses this polymorphism and found that 6 transcription factor's binding site included the rs3814834 position: SMARCA4, TFAP2C, EGR1, FOXA1, MITF and FOS. Variants rs4904834 and rs10146519 are located in intronic regions of the *ATXN3* gene and might be functionally relevant if one hypothesizes their influence on splicing efficiency of *ATXN3*, since the majority of known functional intronic polymorphisms activate a cryptic splice site or are located within splice junctions, or even within an intron splice enhancer or branchpoint site (Cooper 2010). Previously, intronic SNP rs6438552 in *GSK3B* gene has been found to affect phenotype of Parkinson's disease through the regulation of selected splice acceptor sites, causing GSK3B transcripts to lack exons 9 and 11 (Kwok *et al.* 2005).

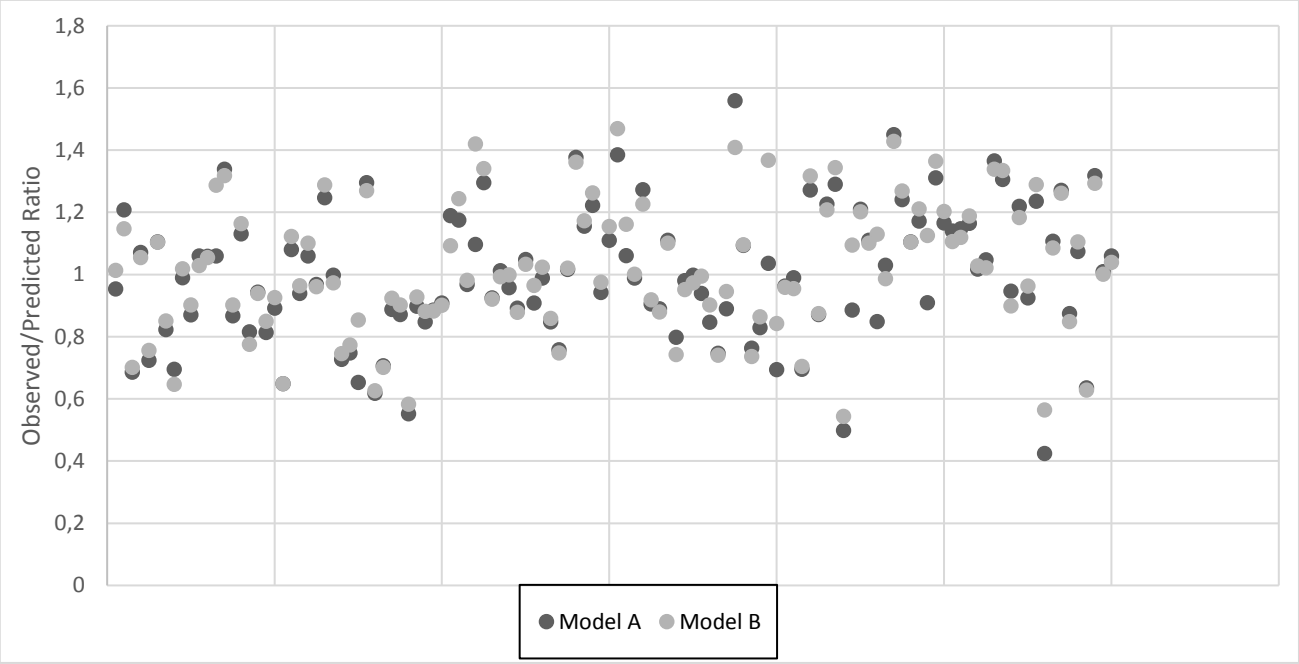


Figure 4 – Scatter plot graph displaying the ratio between observed and predicted MJD AO by models A and B.

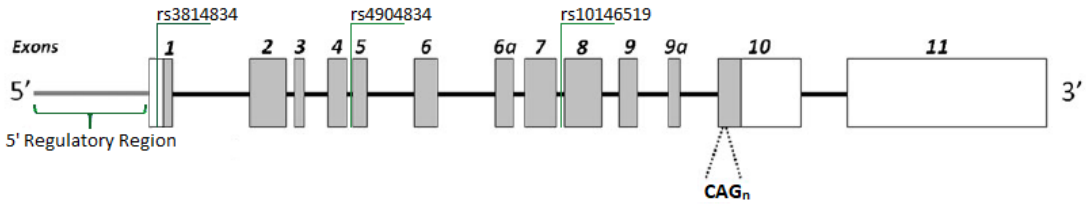


Figure 5 - Location of variants rs3814834, rs4904834 and rs10146519 in the *ATXN3* gene (Adapted from Bettencourt *et al.* 2012).

Discussion

This study was performed with DNA samples from 120 Portuguese MJD patients. To fulfill our goal, we initially collected information such as AO, gender and normal/expanded CAG repeat sizes of the causative gene *ATXN3* for all these patients. In order to focus on the impact of AO genetic modifiers and remove the bias effect that could be introduced by other familial factors segregated intergenerationally, we attempted to study patients who did not share a recent common ancestor. This task is particularly challenging in the case of MJD since, to date, only two mutational origins have been described: an ancient worldwide spread mutation of Asian origin (named Joseph lineage after the name of the third and largest identified MJD family, emigrated from the Azorean island of Flores (Gaspar *et al.* 2001), where this lineage is highly frequent); and a more recent Machado lineage, probably of Portuguese ancestry, the most common SNP haplotype in São Miguel, home of Machado family, the first to be described among Azorean emigrants living in the United States of America (Martins *et al.* 2007). With only two mutational origins identified, one can argue that all MJD patients carrying the Joseph lineage should share a common ancestor; the same for patients from the Machado MJD lineage. Thus, our attempt to avoid the bias effect of familial factors on AO was to select a cohort of patients from different families, whose genealogies have been constructed as more detailed as possible based on patients reports made to link individuals previously taken from different families.

Before analyzing the effect of *ATXN3* SNPs on MJD AO, we tested whether our cohort would be a good representation of a heterogeneous population of MJD patients. We started by analyzing parameters such as repeat size of both normal and expanded alleles in addition to the AO. In our group of patients, the length of the *ATXN3* repeat varied greatly in both normal (14-33 repeats) and expanded (63-82 repeats) alleles. No overlap between the normal and disease size ranges was found, greatly facilitating MJD molecular diagnosis, which is not always the case with other repeat expansion disorders. In MJD, intermediate alleles are extremely rare, with their role on disease pathogenicity still under debate (Takahashi *et al.* 2017). In our cohort, normal alleles displayed an apparent bimodal distribution with 14 and 23 repeats dominating the frequencies, also in conformity with previous findings (Limprasert *et al.* 1996). On the other hand, the expanded allele displayed a normal distribution and AO was found to be highly variable, ranging from 13 to 70 years with an average of 42.78 years, consistent with the literature (Long *et al.* 2015; Maruyama *et al.* 1995). Considering these parameters, we concluded

that the group of patients here analyzed seem to fall into distributions similar to others described before in different populations being, therefore, a good representation of other cohorts of MJD patients.

Like most repeat expansion disorders, MJD displays a negative correlation between AO and the number of repeats in the deleterious allele. However, despite this correlation, there is substantial variability between expected and observed AO. Among our patients, we observed the case of two patients in our cohort with identical expanded CAG lengths (77 repeats) whose AO differed more than 55 years (one with an age at onset of 15 and the other 70). In the literature, repeat length of expansions is reported to explain between 48 and 60% of this variability (Soong *et al.* 1997; Tezenas du Montcel *et al.* 2014; França *et al.* 2012; Long *et al.* 2015; van de Warrenburg *et al.* 2005); in our study, this percentage was ~52. This highlights the importance of finding other genetic and/or environmental factors that contribute to the large variability of AO. While determining these correlations, one must take into consideration, however, some underlying difficulties, such as the accurate assessment of the expansion size. The identification of the exact repeat size may be inaccurate due to PCR-based technique errors, such as slippage of Taq-polymerase. On the other hand, due to somatic mosaicism (the result of mitotic instability of large repetitive *loci*) frequently associated to expanded alleles, more than one expansion size is usually found in blood cells; in this study, we selected the peak with highest fluorescence intensity or the peak corresponding to the largest allele in cases where two same-sized peaks have been observed. Additionally, repeat sizes of expanded alleles in blood (the tissue from which DNA has been extracted to perform this and most studies on AO modifiers described in the literature) may differ from sizes in the cells of structures that are actually involved in the development of the disease (Maciel *et al.* 1995). Interestingly, a previous study tried to associate somatic mosaicism of the *ATXN3*_CAG repeat size in the central nervous system of MJD patients and neuropathological changes observed (Lopes-Cedes *et al.* 1996). The authors assessed the size of MJD expanded alleles in 20 different regions of the brain, brainstem, cerebellum, and spinal cord from 3 patients, concluding, however, that the pattern of $(CAG)_{exp}$ mosaicism reflected the developmental pathways and cell composition of different central nervous system regions, instead of being the direct cause of MJD selective cell death. Lastly, AO/$(CAG)_{exp}$ correlations may be difficult due to ambiguity while obtaining medical history of patients. An increase in MJD awareness in affected families facilitates detection of first symptoms causing a bias towards an earlier onset in younger generations and consequential imprecision in AO.

In this study, we focused the search for additional genetic (rather than environmental) influences on AO. The relevance of genetic AO MJD modifiers is supported by previous studies that reveal that 45.5% of the AO variance can be explained by interfamilial differences, supporting the importance of genetic inherited factors influencing MJD clinical presentation (De Stefano at al. 1996; Warrenburg *et al.* 1996).

We started by analyzing the influence of normal *ATXN3* alleles on disease onset, but no differences have been observed. There are studies where the normal allele was reported to have a small but significant influence on MJD AO: França *et al.* (1996) stated that patients matched for $(CAG)_{exp}$ length would have a clearly different predicted AO depending on the size of normal alleles they had inherited; interestingly, the authors noticed that this modulating effect of the normal *ATXN3* allele was more relevant for patients carrying small expansions (França *et al.* 1996).

Increasingly, disease-associated SNPs are being identified as potential modifiers of disease progression in MJD and other repeat-associated disorders (Long *et al.* 2015; Becanović *et al.* 2015). From the 18 polymorphisms found in the *ATXN3* gene, we identified a significant association between AO of MJD patients and the genotype of 3 individually analyzed SNPs: rs3814834, rs4904834 and rs10146519. The addition of the effect of these variants improved the explained AO variance by 7.9%, with rs10146519 having the most influence in model improvement (11.5%).

The impact of model B was evaluated by comparing the differences of AO means associated to the presence/absence of the minor SNP allele and, then, to each SNP genotype. When comparing only presence/absence of the minor allele, rs3814834 and rs4904834 did not show a significant difference in mean AO ($p \geq 0.05$), though rs3814834 was on the limit of significance and an increase in sample size could give a more conclusive result. As for rs10146519, we found a significant difference between means of about 8 years ($p < 0.05$), the presence of the minor allele resulted in an earlier AO when present. When comparing genotypes, the mean AO of patients homozygous for the common allele of rs3814834 was significantly different to that found in heterozygous patients ($p < 0.05$). An earlier AO was observed for heterozygous patients (9 years), however, since rare homozygous were not found in our cohort, no further conclusions could be made on whether the minor allele has a dominant or co-dominant effect on disease AO. Also for SNP rs4904834, we did not find homozygous for the rare allele, but no differences were detected between AO means ($p < 0.05$). Finally, for rs10146519, we were able to find all three genotypes and noted a gradient, starting from the latest AO for

homozygous patients carrying the major allele, with a 8 year decrease for heterozygous and additional 12 years earlier for rare homozygous. This supports a dominant effect of this SNP on disease AO with dosage effect. One should, however, take into account that only 2 rare homozygos patients were found in our cohort. In an attempt to control for the influence of CAG expansion sizes on the previous analysis, we compared mean $(CAG)_{exp}$ for each genotype but no significant differences were found.

When comparing the accuracy of AO predictions from each model, the new model that included SNP effects (model B) was in fact more accurate than model A, which takes into account only the known effect of repeat size of expanded alleles. When calculating the ratio between observed and predicted AO for each, model B had the closest predictions.

In our study, the most promising polymorphisms, more likely to explain part of the AO variation, are regulatory variants that may alter gene expression and modify disease expressivity and intronic variants that may alter RNA splicing mechanisms. Regulatory SNPs can alter the binding of transcription factors to their binding sites, changing gene transcriptional activity and consequently modifying disease phenotype and progression. Variant rs3814834 is located in the regulatory region of the *ATXN3* gene and is situated in the binding site of 6 transcription factors. Introns can also harbor functional polymorphisms that influence the expression of the genes that host them; some of these intronic variants may also confer susceptibility to disease or otherwise modulate the genotype–phenotype relationship. Variants rs4904834 and rs10146519 are located in intronic regions and could potentially affect *ATXN3* expression, but future experiments should be done to confirm these SNPs functionality or identify other variations in linkage disequilibrium with this SNP.

It is important to identify genetic modifiers in MJD not only because they may provide improved prognostic information for MJD gene carriers, but because these genes and related pathways are potential therapeutic targets for the disease. The most direct, specific and effective solution to correct MJD would be to inhibit mutant *ATXN3* expression using RNAi, thus targeting the initial cause of the disorder. Studies using RNAi in mice have been conducted but are still in very early stages (Costa *et al.* 2013). The findings in our studies are potential proof that an allele-specific RNAi therapy for MJD may have a beneficial effect on AO but larger studies including a larger number of associated SNPs is in order.

Concluding Remarks and Future Perspectives

The three proposed goals of our study were mostly achieved. We were able to optimize multiplex reactions to amplify exonic and regulatory regions of *ATXN3*; we identified 3 SNPs that, together with the expanded allele, had a significant effect on MJD AO and found one SNP that could be useful for future applications, such as RNAi therapies, due to its strategic location in the regulatory region of the gene.

We can conclude from our study that genetic variants in *ATXN3* may have a significant influence on the clinical presentation of MJD, though, a few improvements could be made to further validate our results and corroborate our findings. A larger and more heterogeneous cohort of patients would benefit this study: by including additional haplotype backgrounds and adding patients from the extremes of expanded repeat distribution, we could have a cohort that is a more realistic representation of the worldwide MJD population. This could also result in the identification of an increased number of SNPs with a significant effect on disease progression and the possibility of an AO prediction of increased accuracy.

Additionally, functional studies can be made to evaluate the functional relevance of SNPs here found to be statistical significant on MJD AO, namely on transcription regulation, splicing or translation. For instances, it would be interesting to test the impact of regulatory variant rs3814834 on the transcriptional activity of *ATXN3* with a luciferase reporter assay. On the other hand, for rs4904834 and rs10146519, we could analyze alternative transcripts present in patients carrying the variant's alleles via reverse transcription PCRs performed using RNA from these patients.

Finally, SNP allele frequencies on the MJD population obtained in this study can be further used by others while developing gene silencing therapies such as allele-specific silencing of mutant ataxin-3: the best candidate SNPs would be those located in transcribed regions, with alleles in strong linkage disequilibrium with the expanded allele.

References

1. Alves, Sandro, *et al.* "Allele-specific RNA silencing of mutant ataxin-3 mediates neuroprotection in a rat model of Machado-Joseph disease." *PloS one* 3.10 (2008): e3341.
2. Barrandeguy, María Eugenia, & María Victoria García. "Quantifying genetic diversity: the starting point for population genetic studies using molecular markers." *Journal of genetics*93.2 (2014): 587-589.
3. Bečanović, Kristina, *et al.* "A SNP in the HTT promoter alters NF-κB binding and is a bidirectional genetic modifier of Huntington disease." *Nature neuroscience* 18.6 (2015): 807.
4. Bettencourt, Conceição, & Manuela Lima. "Machado-Joseph Disease: from first descriptions to new perspectives." *Orphanet journal of rare diseases* 6.1 (2011): 35.
5. Bettencourt, Conceição, *et al.* "Increased transcript diversity: novel splicing variants of Machado–Joseph Disease gene (*ATXN3*)." *Neurogenetics* 11.2 (2010): 193-202.
6. Brookes, Anthony J. "The essence of SNPs." *Gene* 234.2 (1999): 177-186.
7. Budworth, Helen, and Cynthia T. McMurray. "A brief history of triplet repeat diseases." *Trinucleotide Repeat Protocols*. Humana Press, Totowa, NJ, 2013. 3-17.
8. Cooper, David N. "Functional intronic polymorphisms: buried treasure awaiting discovery within our genes." (2010): 284.
9. Costa, Maria, *et al.* "Ataxin-3 plays a role in mouse myogenic differentiation through regulation of integrin subunit levels." *PLoS One* 5.7 (2010): e11728.
10. Costa, Maria, and Henry L. Paulson. "Toward understanding Machado–Joseph disease." *Progress in neurobiology* 97.2 (2012): 239-257.Coutinho, M. P. M. D. A. (1992). Doença de Machado-Joseph: tentativa de definição.
11. Coutinho, Paula, and Corino Andrade. "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* 28.7 (1978): 703-703.
12. DeStefano, Anita L., *et al.* "A familial factor independent of CAG repeat length influences age at onset of Machado-Joseph disease." *American journal of human genetics* 59.1 (1996): 119..

13. Egan, Robert A., Richard Camicioli, and Bradley W. Popovich. "A small 55-repeat MJD1 CAG allele in a patient with Machado-Joseph disease and abnormal eye movements." *European neurology* 44.3 (2000): 189-190.
14. Emmel, V. E., *et al.* "Does DNA methylation in the promoter region of the *ATXN3* gene modify age at onset in MJD (SCA3) patients?." *Clinical genetics* 79.1 (2011): 100-102.
15. Fan, Hao, and Jia-You Chu. "A brief review of short tandem repeat mutation." *Genomics, Proteomics & Bioinformatics* 5.1 (2007): 7-14.
16. Fiszer, Agnieszka, *et al.* "An evaluation of oligonucleotide-based therapeutic strategies for polyQ diseases." *BMC molecular biology* 13.1 (2012): 6.
17. França Jr, Marcondes C., *et al.* "Normal *ATXN3* allele but not CHIP polymorphisms modulates age at onset in Machado–Joseph disease." *Frontiers in neurology* 3 (2012): 164.
18. Gaspar, C., *et al.* "Ancestral origins of the Machado-Joseph disease mutation: a worldwide haplotype study." *The American Journal of Human Genetics* 68.2 (2001): 523-528.
19. Gatchel, J. R., & Zoghbi, H. Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. Nature Reviews Genetics, 6(10), 743.
20. Goodwin, William, Adrian Linacre, and Sibte Hadi. *An introduction to forensic genetics*. Vol. 2. John Wiley & Sons, 2011.
21. Grover, Atul, and P. C. Sharma. "Development and use of molecular markers: past and present." *Critical reviews in biotechnology* 36.2 (2016): 290-302.
22. Gu, Weihong, *et al.* "The shortest expanded allele of the MJD1 gene in a Chinese MJD kindred with autonomic dysfunction." *European neurology* 52.2 (2004): 107-111.
23. Jobling, Mark, Matthew Hurles, and Chris Tyler-Smith. *Human evolutionary genetics: origins, peoples & disease*. Garland Science, 2013.
24. Johnson, Mark, *et al.* "NCBI BLAST: a better web interface." *Nucleic acids research* 36.suppl_2 (2008): W5-W9.
25. Kawaguchi, Yoshiya, *et al.* "CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32. 1." *Nature genetics* 8.3 (1994): 221.
26. Kearse, Matthew, *et al.* "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data." *Bioinformatics* 28.12 (2012): 1647-1649.
27. Kent, W. James. "BLAT—the BLAST-like alignment tool." *Genome research* 12.4 (2002): 656-664.

28. Kibbe, Warren A. "OligoCalc: an online oligonucleotide properties calculator." *Nucleic acids research* 35.suppl_2 (2007): W43-W46.
29. Kovtun, Irina V., and Cynthia T. McMurray. "Features of trinucleotide repeat instability in vivo." *Cell research* 18.1 (2008): 198.
30. Kouniaki, Diamanto, Chryssa Papasteriades, and Alexandra Tsirogianni. "Short Tandem Repeats Loci in Parentage Testing." *Hospital Chronicles* 10.2 (2015).
31. Kwok, John BJ, *et al.* "GSK3B polymorphisms alter transcription and splicing in Parkinson's disease." *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 58.6 (2005): 829-839.
32. Lang, Anthony E., *et al.* "Homozygous inheritance of the Machado-Joseph disease gene." *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 36.3 (1994): 443-447.
33. Limprasert, Pornprot, *et al.* "Analysis of CAG repeat of the Machado-Joseph gene in human, chimpanzee and monkey populations: a variant nucleotide is associated with the number of CAG repeats." *Human molecular genetics* 5.2 (1996): 207-213..
34. Lopes-Cendes, Iscia, *et al.* "Somatic mosaicism in the central nervous system in spinocerebellar ataxia type 1 and Machado-Joseph disease." *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 40.2 (1996): 199-206.
35. Long, Zhe, *et al.* "Two novel SNPs in *ATXN3* 3'UTR may decrease age at onset of SCA3/MJD in Chinese patients." *PloS one* 10.2 (2015): e0117488.
36. Maciel, Patrícia, *et al.* "Correlation between CAG repeat length and clinical features in Machado-Joseph disease." *American journal of human genetics* 57.1 (1995): 54.
37. Maciel, Patricia, *et al.* "Mosaicism of the CAG repeat in CNS tissue in relation to age at death in spinocerebellar ataxia type 1 and Machado-Joseph disease patients." *American journal of human genetics* 60.4 (1997): 993.
38. Maciel, Patrícia, *et al.* "Improvement in the molecular diagnosis of Machado-Joseph disease." *Archives of neurology* 58.11 (2001): 1821-1827.
39. Manto, Mario-Ubaldo. "The wide spectrum of spinocerebellar ataxias (SCAs)." (2005): 2.
40. Martins S. (2007) Evolutionary and Epidemiological Genetics of Machado-Joseph Disease. In: Biology. Porto University.

41. Martins, Sandra, *et al.* "Asian origin for the worldwide-spread mutational event in Machado-Joseph disease." *Archives of neurology* 64.10 (2007): 1502-1508.
42. Martins, Sandra, *et al.* "Cis-acting factors promoting the CAG intergenerational instability in Machado–Joseph disease." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 147.4 (2008): 439-446.
43. Martins, Sandra, *et al.* "Modifiers of (CAG) n instability in Machado–Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes." *Human genetics* 133.10 (2014): 1311-1318.
44. Martins, Sandra, *et al.* "Mutational origin of Machado-Joseph disease in the Australian Aboriginal communities of Groote Eylandt and Yirrkala." *Archives of neurology* 69.6 (2012): 746-751..
45. Martins, Sandra, and Jorge Sequeiros. "Origins and Spread of Machado-Joseph Disease Ancestral Mutations Events." *Polyglutamine Disorders*. Springer, Cham, 2018. 243-254.
46. Matos, Carlos, Luís Pereira de Almeida, and Clévio Nóbrega. "Machado-Joseph disease/Spinocerebellar ataxia type 3: lessons from disease pathogenesis and clues into therapy." *Journal of neurochemistry* (2018).
47. Matos, Carlos A., *et al.* "Ataxin-3 phosphorylation decreases neuronal defects in spinocerebellar ataxia type 3 models." *J Cell Biol* 212.4 (2016): 465-480.
48. Maruyama, Hirofumi, *et al.* "Molecular features of the CAG repeats and clinical manifestation of Machado—Joseph disease." *Human molecular genetics* 4.5 (1995): 807-812.
49. McMurray, Cynthia T. "Mechanisms of trinucleotide repeat instability during human development." *Nature Reviews Genetics*11.11 (2010): 786.
50. Mendonça, Nuno, *et al.* "Clinical Features of Machado-Joseph Disease." *Polyglutamine Disorders*. Springer, Cham, 2018. 255-273.
51. Nachman, Michael W., and Susan L. Crowell. "Estimate of the mutation rate per nucleotide in humans." *Genetics* 156.1 (2000): 297-304.
52. Padiath, Quasar Saleem, *et al.* "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* 133.1 (2005): 124-126.
53. Paulson, Henry L. "The spinocerebellar ataxias." *Journal of neuro-ophthalmology: the official journal of the North American Neuro-Ophthalmology Society* 29.3 (2009): 227.

54. Paulson, H. L., *et al.* "Intranuclear inclusions of expanded polyglutamine protein in spinocerebellar ataxia type 3." *Neuron*19.2 (1997): 333-344.
55. Pearson, Christopher E., Kerrie Nichol Edamura, and John D. Cleary. "Repeat instability: mechanisms of dynamic mutations." Nature Reviews Genetics 6.10 (2005): 729.
56. Pinheiro M.d.F.T. Genética Forense Perspectivas da Identificação Genética. Porto: Edições Universidade Fernando Pessoa (2010).
57. Rodrigues, Ana-João, *et al.* "Absence of ataxin-3 leads to cytoskeletal disorganization and increased cell death." *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*1803.10 (2010): 1154-1163.
58. Rozen S. & Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Bioinformatics methods and protocols, 365-86 (1999).
59. Schlötterer, Christian. "The evolution of molecular markers—just a matter of fashion?." *Nature reviews genetics* 5.1 (2004): 63.
60. Shakkottai, Vikram G., and Brent L. Fogel. "Clinical neurogenetics: autosomal dominant spinocerebellar ataxia." *Neurologic clinics* 31.4 (2013): 987-1007.
61. Sobrino B., Brión M. & Carracedo A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Science International 154, 181-94.
62. Soon, Bing-wen, *et al.* "Machado-joseph disease: Clinical, molecular, and metabolic characterization in chinese kindreds." *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 41.4 (1997): 446-452.
63. Schwarz, Dianne S., *et al.* "Designing siRNA that distinguish between genes that differ by a single nucleotide." *PLoS genetics*2.9 (2006): e140.
64. 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.
65. Teive, Hélio AG, *et al.* "Itajaí, Santa Catarina–Azorean ancestry and spinocerebellar ataxia type 3." *Arquivos de neuro-psiquiatria*74.10 (2016): 858-860.
66. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes.
67. Todi, Sokol V., *et al.* "Ubiquitination directly enhances activity of the deubiquitinating enzyme ataxin-3." *The EMBO journal* 28.4 (2009): 372-382.

68. Vallone, Peter M., and John M. Butler. "AutoDimer: a screening tool for primer-dimer and hairpin structures." *Biotechniques* 37.2 (2004): 226-231.
69. Van Alfen, Nens, *et al.* "Intermediate CAG repeat lengths (53, 54) for MJD/SCA3 are associated with an abnormal phenotype." *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 49.6 (2001): 805-808.
70. Van Schaik, I. N., *et al.* "Machado-Joseph disease presenting as severe asymmetric proximal neuropathy." *Journal of Neurology, Neurosurgery & Psychiatry* 63.4 (1997): 534-536.
71. van de Warrenburg, Bart PC, *et al.* "Age at onset variance analysis in spinocerebellar ataxias: a study in a Dutch–French cohort." *Annals of neurology* 57.4 (2005): 505-512.
72. Zeng, Sheng, *et al.* "Chinese homozygous Machado–Joseph disease (MJD)/SCA3: a case report." *Journal of human genetics*60.3 (2015): 157.
73. Zhao, Xiao-Nan, and Karen Usdin. "The repeat expansion diseases: the dark side of DNA repair." *DNA repair* 32 (2015): 96-105.