Universidade do Porto

Faculdade de Engenharia da Universidade do Porto

Instituto de Ciências Biomédicas Abel Salazar

Instituto de Investigação e Inovação em Saúde

Instituto de Patologia e Imunologia Molecular da Universidade do Porto

Influence of variants in *JOSD1*, *JOSD2* and *ATXN3L* on Machado-Joseph disease age-at-onset

Filipe Manuel Amorim Costa

Masters in Bioengineering – Molecular Biotechnology

Porto, September 2016

Influence of variants in *JOSD1*, *JOSD2* and *ATXN3L* on Machado-Joseph disease age-at-onset

Filipe Manuel Amorim Costa

Thesis for Masters in Bioengineering

FEUP – Faculdade de Engenharia da Universidade do Porto

ICBAS-UP – Instituto de Ciências Biomédicas de Abel Salazar

Supervisor:

Sandra Martins, Ph.D. **i3S/IPATIMUP**

i3S – Instituto de Investigação e Inovação em Saúde

IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto

ABSTRACT

Machado-Joseph disease (MJD) proteases are a subfamily of deubiquitinating enzymes (DUBs) that comprise 4 proteins: ataxin-3 (ATXN3), ataxin-3 like (ATXN3L), Josephin domain-containing 1 (JOSD1) and Josephin domain-containing 2 (JOSD2). ATXN3 is the most studied protein of this subfamily; even though its precise biologic function still remains unclear, ATXN3 seems to be important in the proteasomal protein degradation pathway. When the polyglutamine region of ATXN3 is expanded beyond 61 glutamines, individuals present the inherited neurodegenerative disease MJD. The pathogenicity associated to this disorder seems to be linked to different toxic protein entities, particularly the soluble amyloid-like oligomers formed from the proteolytic cleavage of the aggregation-prone mutated protein.

The size of the expanded *ATXN3* allele explains only 45-76% of the variability found in MJD age-at-onset (AO). The evidenced existence of other familial AO modifiers, together with the observed protective role of wild-type ATXN3 in MJD neurotoxicity, suggested that normal alleles may play a modulatory role in disease pathogenesis. Also, the proved DUB activity of the catalytic triad shared by all MJD proteases led us to raise the hypothesis that ATXN3L, JOSD1 and JOSD2 may exert a similar function to ATXN3 and therefore compensate for the decrease levels of its wild-type form in MJD patients.

The main aim of this work was to search for variants in regulatory and exonic regions of our candidate genes (*ATXN3L*, *JOSD1* and *JOSD2*) to further test and discuss their contribution as genetic AO modifiers of MJD. In order to achieve this purpose, we started by optimizing multiplex reactions to amplify the regions of interest of all three genes, to then sequence and align fragments obtained from 100 MJD Portuguese families. Subsequently, we proceeded to the detection of significantly relevant polymorphisms by analysis of covariance, after removing the known influence of the repeat size in the expanded allele. Finally, we hypothesized on the effect of those variants in MJD phenotype.

Our cohort was shown to be representative of a heterogeneous MJD population, with expanded CAG repeat size accounting for 54.7% of the AO variation (model A). The analysis of our candidate genes as possible genetic modifiers of MJD phenotype resulted in the identification of 16 single nucleotide polymorphisms (SNPs): 3 in *JOSD1*, 8 in *JOSD2* and 5 in *ATXN3L*. Analyses of covariance have shown that 5 of these SNPs (COSM3713736, rs550589305, rs141848929, rs796424878 and

rs16999010) contributed to elucidate 8.4% of the AO variability in our cohort, in addition to the already identified effect caused by the expansion size. The cumulative influence of this set of SNPs on MJD AO led us to design a new and more complex template for the linear regression AO/$(CAG)_{exp}$ (model B). This model allowed a better estimation of the AO for the 6 patients who carried the rare allele of analysed SNPs when compared to the prediction based on the single influence of the CAG tract length in expanded alleles. Variant rs16999010, associated to an earlier AO, is involved in an amino acid change in ATXN3L that aligns closely to a key residue for the binding of polyubiquitinated substrates in ATXN3. Regulatory variants rs550589305 and rs141848929, linked to a later AO, may affect the binding of transcription factors and modify *JOSD2* transcription. This way, JOSD2 dosage may be higher and compensate the decrease of normal ATXN3 in MJD patients.

To further confirm the relevance of the described results, we will increase our sample size (particularly in the extremes of the expanded CAG repeat distribution), extend the study to other populations, and undergo functional assays for the statistically relevant SNPs.

RESUMO

As proteases da doença de Machado-Joseph (DMJ) são uma subfamília de enzimas deubiquitinases (DUBs) que abrange 4 proteínas: ataxina-3 (ATXN3), ataxina-3 like (ATXN3L), Josephin-domain containing 1 (JOSD1) e Josephin-domain containing 2 (JOSD2). ATXN3 é a proteína mais estudada desta subfamília; apesar da sua exata função biológica ainda continuar por esclarecer, ATXN3 parece ser importante para a via de degradação proteassomal. Quando a poliglutamina da ATXN3 está expandida para além de 61 unidades, pacientes apresentam DMJ, uma doença neurodegenerativa hereditária. A patogenicidade desta desordem parece estar ligada a diferentes entidades proteicas tóxicas, particularmente a proteína mutada que tem propensão para se agregar.

O tamanho do alelo expandido da ATXN3 explica 45-76% da variabilidade encontrada na idade de início (ii) da DMJ. A prova da existência de outros modificadores familiares da ii, em conjunto com o papel protectivo observado pela forma normal da ATXN3 na neurotoxicidade da DMJ, sugere que os alelos normais poderem ter um papel modulador na patogenicidade da doença. Além disso, a comprovada actividade DUB da tríade catalítica partilhada pelas protéases DMJ levaram-nos a levantar a hipótese da ATXN3L, JOSD1 e JOSD2 poderem ter uma função similar à ATXN3 e, assim, compensar os níveis diminuídos da forma normal nos pacientes com DMJ.

O principal objectivo deste trabalho foi a procura de variantes nas regiões regulatórias e exónicas dos genes candidatos para depois testar e discutir a sua contribuição como modificadores genéticos da ii na DMJ. Por forma a alcançar este propósito, começamos por otimizar reacções multiplex para amplificar as regiões de interesse dos genes candidatos (*JOSD1*, *JOSD2* e *ATXN3L*), para depois sequenciar e alinhar os fragmentos obtidos das 100 famílias portuguesas de DMJ. Em seguida procedemos à deteção de polimorfismos significativamente relevantes por análise de covariância após remover removendo a conhecida influência do tamanho da repetição no alelo expandido). Por último discutimos o efeito dessas variantes no fenótipo da DMJ.

O nosso grupo de indivíduos mostrou ser representativo de uma população heterogénea da DMJ, com o tamanho da expansão CAG a explicar 54.7% da variação da ii (modelo A). A análise dos nossos genes candidatos como possíveis modificadores genéticos do fenótipo da DMJ resultou na identificação de 16

polimorfismos: 3 na *JOSD1*, 8 na *JOSD2* e 5 na *ATXN3L*. Análises de covariança mostaram que 5 destes polimorfismos (COSM3713736, rs550589305, rs141848929, rs796424878 e rs16999010) contribuíram para elucidar 8.4% da variabilidade da ii no nosso grupo de doentes, além do já identificado efeito causado pelo tamanho da expansão. A influência cumulativa deste conjunto de polimorfismos na ii da DMJ levou-nos a criar um novo e mais complexo modelo para a regressão linear ii/$(CAG)_{exp}$ (modelo B). Este modelo permite uma melhor estimativa da ii dos 6 pacientes que possuem o alelo raro dos polimorfismos analisados quando comparado com a previsão baseada apenas na influência do tamanho da repetição CAG nos alelos expandidos. O variante rs16999010, associado a uma ii mais precoce, está envolvido na substituição de um aminoácido na ATXN3L que alinha perto de um resíduo importante para a ligação de substratos poliubiquitinados à ATXN3. Os variantes regulatórios rs550589305 e rs141848929, ligados a uma ii mais avançada, podem afetar a ligação de fatores de transcrição e modificar a transcrição da JOSD2. Desta forma, a dosagem de JOSD2 pode ser superior e compensar a diminuição da forma normal da ATXN3 em indivíduos com DMJ.

Para confirmar a relevância dos resultados descritos, vamos seguidamente aumentar a dimensão da amostra (particularmente nos extremos da distribuição da expansão CAG), estender este estudo a outras populações e realizar ensaios funcionais para os polimorfismos estatisticamente relevantes.

ACKNOWLEDGEMENTS

First of all, I would like to thank Professor António Amorim for giving me the opportunity to perform the following work in his research group.

My hugest gratitude goes towards Dr. Sandra Martins for her unflagging support, illustrious guidance, unconditional dedication and admirable work ethics throughout this project and for enriching my academic education with valuable knowledge from her expertise area.

To IPATIMUP and i3S, for providing the funding and scientific support required for the success of the experimental work.

To all the elements of Population Genetics & Evolution, from researchers to students, for the fellowship and assistance provided when needed.

To my friends, particularly Diana, Sarah, Catarina, Helena, Tatiana and Sílvia, for the constant kindness and for being always capable to say comfort words in the hardest periods and share joys in the happiest moments.

Finally, to my parents, my sister and Tété, for being my role models, for teaching me all the important values that make me who I am and for always being there for me no matter what.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANCOVA	Analysis of covariance
AO	Age-at-onset
ATXN3	Ataxin-3
ATXN3L	Ataxin-3 like
bp	Base pairs
dNTP	Deoxybonucleotide triphosphate
DUB	Deubiquitinating enzyme
JD	Josephin domain
JOSD1	Josephin domain-containing 1
JOSD2	Josephin domain-containing 2
MJD	Machado-Joseph disease
PCR	Polymerase chain reaction
polyQ	Polyglutamine
SCA	Spinocerebellar ataxia
SCA3	Spinocerebellar ataxia type 3
SNP	Single nucleotide polymorphism
Ub	Ubiquitin
UIM	Ubiquitin interacting motif
UTR	Untranslated region

1. Introduction

Ubiquitination and deubiquitination are two reversible regulatory mechanisms that, through regulation of proteolysis, play fundamental roles for the correct function of critical cellular processes, such as trafficking, subcellular localization, DNA repair and signal transduction [1, 2]. Ubiquitination is a cascade of processes that involves the triggering of ubiquitin (Ub) by Ub-activating enzyme E1, followed by its transport to the active site by Ub-conjugating enzyme E2 and finalized by its recognition and binding to the target substrate by Ub-ligase E3 (Figure 1) [2, 3]. This enzymatic process leads to the addition of Ubs, via isopeptide linkages, to lysines in targeted proteins, marking them for degradation by proteasomes or lysosomes [4]. Although Ub forms a covalent bond to several proteins that are rapidly degraded, ubiquitin itself is a long-lived protein *in vivo* [5] due to the action of deubiquitinating enzymes (DUBs) that efficiently remove Ub before the proteolysis of the conjugated protein. Deubiquitination is also able to negatively regulate protein degradation, regenerate monoUb from unanchored polyUb chains and process inactive Ub precursors [6]. A total of 86 DUBs have been so far identified and divided into five subfamilies based on their sequence and structural similarity [7]; one of these subfamilies comprises the 4 elements of Machado-Joseph disease (MJD) proteases.

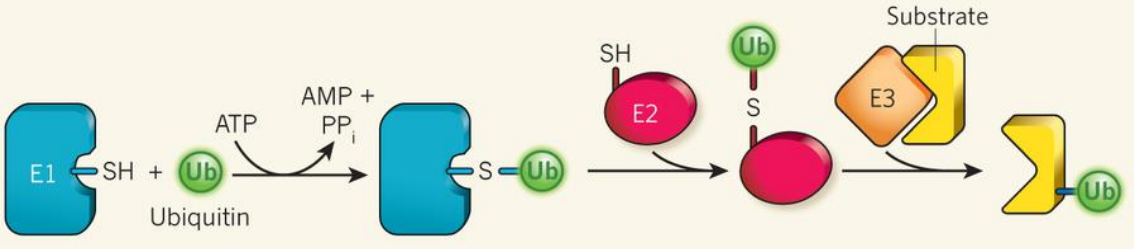
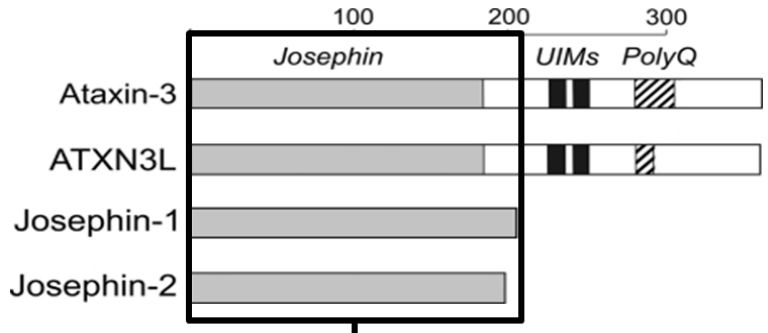


Figure 1 – Mechanism of ubiquitination involving a three-enzyme cascade. The E1 enzyme first activates the C-terminus of the Ub molecule, using energy available from the hydrolysis of an ATP molecule to AMP and pyrophosphate (PPi). After activation, ubiquitin reacts with the thiol group (SH) of the E1 active-site cysteine residue. Subsequently, ubiquitin is transported from E1 to E2, and E3 functions as an intermediary of the ubiquitin transfer to the substrate protein. Adapted from Bhogaraju and Dikic, 2016 [8].

1.1 Machado-Joseph disease proteases

The MJD proteases subfamily of DUBs includes 2 ataxins – ataxin-3 (ATXN3) and ataxin-3 like (ATXN3L) – and 2 Josephins – Josephin domain-containing 1 (JOSD1) and Josephin domain-containing 2 (JOSD2). Functionally, all these proteins share Ub

protease activity, i.e. the ability to cleave isopeptide bonds between Ub monomers. This capacity was firstly predicted through an integrative bioinformatics analysis of ATXN3 amino acid sequence [9] and later confirmed in biochemical studies using model substrates and Ub protease-specific inhibitors [10]. Structurally, the unifying feature of this subfamily is the Josephin domain (JD) [11], a region with approximately 180 amino acids in which its catalytic positions are relatively conserved among all 4 proteins (Figure 2) [12].



B

		*		~	~		^	^		^		~		*	~		*		~
ATXN3	14	C	27	Y	F	77	I	Q	87	W	91	L	119	H	W	134	N	163	F
ATXN3L	14	C	27	Y	F	77	I	Q	87	W	91	I	119	H	W	134	N	163	F
JOSD1	36	C	50	A	F	103	K	R	113	T	117	G	139	H	W	154	D	185	L
JOSD2	24	C	37	L	F	89	R	R	99	P	103	G	125	H	W	140	D	171	L

Figure 2 – Structural similarities and differences in the JD of MJD proteases. (A) Schematic representation of the members of MJD class proteins. JDs are highlighted and shown in light grey, the Ub-interaction motifs (UIMs) are shown in dark grey and the polyglutamine tracts (polyQ) are shown as cross-hatched. The sizes of the UIMs and polyQ tracts are not presented to scale. Adapted from Weeks *et al.* [13]. (B) Representation of important amino acids in the JD for protease activity after protein alignment of MJD protease members. Asterisks (*) indicate the catalytic triad, circumflex accents (^) indicate key amino acids for the correct function of Ub-binding sites, and tildes (~) indicate other amino acids that should be important in the Ub-binding sites [14]. Black background indicates residue similarity.

1.1.1 Ataxin-3

The *ATXN3* gene was found to be located in the long arm of chromosome 14 (14q24.3-q32) by Takiyama's team in 1993 [15]. At the beginning of the 21st century, the genomic sequence was published and *ATXN3* was described as spanning for about 48 kb and consisting of 11 exons (Figure 3) [16].

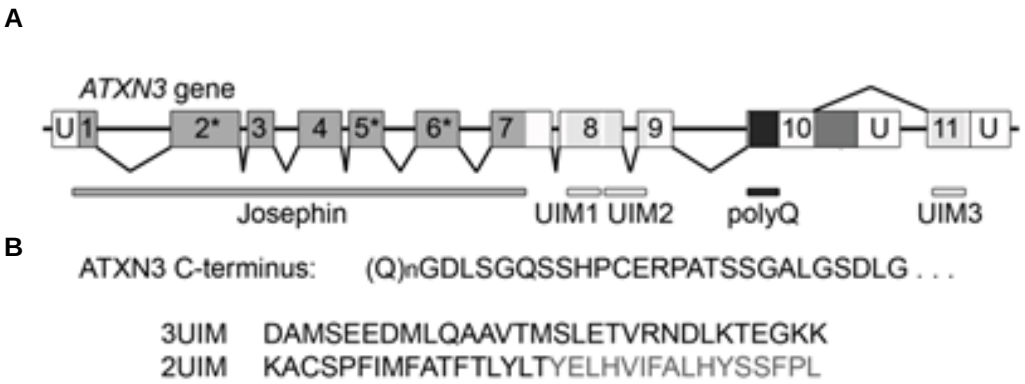


Figure 3 – Schematic representation of *ATXN3* gene structure. (A) Numbered boxes indicate exons, with asterisks designating those that include amino acids of the catalytic triad. PolyQ refers to the polyglutamine tract. Untranslated regions (U) are not to scale. (B) C-terminal protein sequence shared between 2-UIM and 3-UIM isoforms, followed by their divergent sequences: above is shown the alternative splicing that links exons 10 and 11, producing 3-UIM ATXN3; below are 2 splicing forms of the 2-UIM ATXN3 transcript (one denoted by black and the second by the entire sequence). Adapted from Harris *et al.* [17].

The *ATXN3* gene encodes ATXN3, a modular protein with an approximate molecular weight of 42 kDa that varies slightly in size depending on the length of the polymorphic glutamine repeat and on alternative splicing [17]. ATXN3 is structurally composed by a globular N-terminal JD with a papain-like fold, followed by a flexible unstructured C-terminal tail that englobes 2 or 3 Ub-interaction motifs (UIM) and a polyQ tract of variable length (Figure 4) [18]. The JD possesses 2 Ub-binding sites and a highly conserved catalytic triad, characteristic of cysteine proteases: a cysteine (C14), a histidine (H119) and an asparagine (N134) [14, 19]. In addition, a functional nuclear localization signal (NLS) has been described in the region upstream of the polyQ sequence [20].

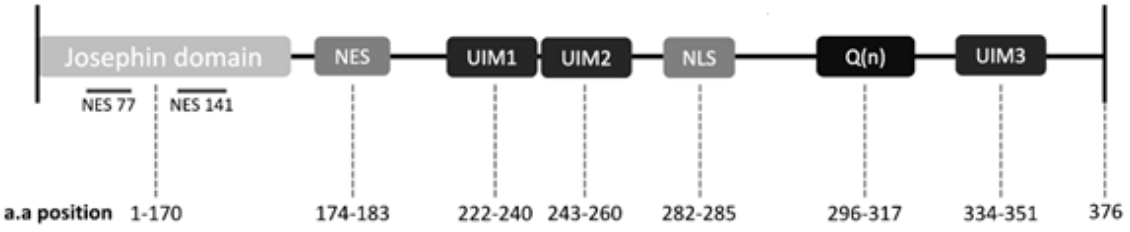


Figure 4 – Schematic structure of the ATXN3 protein, more specifically the 3-UIM isoform. ATXN3 is mainly composed of a highly conserved N-terminal region denominated Josephin domain that encodes a catalytic triad (C14, H119 and N134) and 2 identified nuclear export signals (NES), followed by a flexible C-terminal tail with a predicted NES, 2 or 3 ubiquitin-interacting motifs (UIM), a nuclear localization signal (NLS) and a polyglutamine stretch (polyQ). Adapted from Nóbrega and Pereira de Almeida *[21]*.

This protein seems to be widely, though heterogeneously, expressed within the cells, having been reported in the cytoplasm, nucleus and even mitochondria, with varying levels of predominance depending on the cell type [22, 23]. In human brain cells, ATXN3 concentrates mainly in the perikarya, but can also be found in proximal processes and axons [24]. The subcellular transport of ATXN3 from the cytoplasm to the nucleus seems to be regulated by two different and independent mechanisms: the identified NLS [22] and the casein kinase 2-dependent phosphorylation of 3 serine residues, located in the first (Ser236) and third (Ser340 and Ser352) UIMs [25].

Alternative splicing of the human *ATXN3* gene may result in the production of 56 different isoforms of ATXN3, varying at the C-terminal section of the protein [26, 27]. From those, at least 20 isoforms are likely to be translated, but only 2 have been extensively tested in detailed mechanistic and functional studies: both are full length proteins containing the polyQ stretch, only differing in the presence or absence of a third UIM in the C-terminal (Figure 2). These two splice isoforms, designated 3-UIM ATXN3 and 2-UIM ATXN3 respectively, exhibit similar enzymatic properties but differ in their stability and turnover. While 3-UIM ATXN3 is a relatively stable protein with a low turnover rate, primarily via autophagy, 2-UIM ATXN3 is highly unstable and prone to misfolding and rapid degradation, mainly via proteasome. Notably, 3-UIM ATXN3 is the predominant isoform in the central nervous system [17].

Although its precise physiologic role is still unclear, ATXN3 seems to have an effective involvement in the Ub-proteasome pathway, one of the most important mechanisms for the turnover of damaged proteins [28]. This idea was supported after observing that the inhibition of the catalytic activity of ATXN3 resulted in an increase of polyubiquitinated proteins to a similar degree found when the proteasome was inhibited [29]. The substantiation of ATXN3 role as a DUB came with the discovery of increased levels of ubiquitinated proteins observed in *ATXN3* knockout mice, when compared to wild-type animals [30]. More specifically, ATXN3 appears to function as a polyUb-editing protease that favors the shortening of polyUb chains, with preferably no less than 4 Ub monomers, in disregard of its complete disassembly. ATXN3 can bind K48-linked, K63-linked and mixed-linkage chains, but cleavage *in vitro* occurs preferentially in the two latter forms [31]. The cleavage of different types of polyUb chains by ATXN3 is achieved by a cooperative interplay between the first 2 UIMs, situated in the C-terminal region, and the 2 Ub-binding sites of the JD [10]. While the UIMs act cooperatively to recruit and bind chains, the Ub-binding sites, although presenting lower affinity to the chains, are able to adjust the position of isopeptide bonds relatively

to the catalytic site, allowing the removal of the polyUb chain from substrates prior to digestion [14].

Interestingly, ATXN3 and E3 Ub ligases, enzymes responsible for the formation of isopeptide bonds between Ub molecules and target proteins, have been speculated to transregulate each other as part of their normal function. E3 Ub ligases may be responsible for regulating ATXN3 activity, while ATXN3 seems to be able to edit polyUb chains added to other E3 ligase substrates [32]. Particularly, ATXN3 was the first DUB found to counteract, *in vitro*, the self-ubiquitination of parkin, a Parkinson disease-associated E3 Ub-ligase [33].

The enhancement of ATXN3 DUB activity in the proper cellular environment is predictably accomplished by post-translational modifications [34], therefore explaining the low activity of ATXN3 reported *in vitro* [35]. The ATXN3 monoubiquitination was found to increase its DUB activity, independently of other potential co-factors and interactors [36]. The ubiquitination of a residue alone (Lys117) of ATXN3's JD was sufficient to observe an improvement in the protein protease activity *in vitro*. This upregulation of ATXN3 activity by ubiquitination may operate as a response to higher levels of proteotoxic stress [37].

Several interacting partners of ATXN3 have been identified, enlightening for other possible roles of this protein. The proved interaction of ATXN3 with p97/valosin-containing protein (VCP) and with the 2 homologs of *Saccharomyces cerevisiae* DNA repair protein Rad23, HHR23A and HHR23B, have been linked to endoplasmic reticulum-associated degradation (ERAD). ERAD is a system that mediates the ubiquitination of misfolded proteins in the secretory pathway and their transportation to the cytosol for deterioration by the proteasome [38, 39].

ATXN3 has also been shown to cooperate with neural precursor cell expressed developmentally downregulated 8 (NEDD8), an Ub-like protein. The conjugation of NEDD8 with other proteins is called neddylation and such interaction with ATXN3 may interfere with the neddylation of other substrates, therefore influencing the processes in which NEDD8 is involved [40].

A role as a player in aggresome formation has also been proposed for ATXN3. Aggresomes are misfolded protein aggregates formed near the microtubule-organizing center (MTOC) when the proteasome is not able to deal with those proteins. Since the defective proteins are then degraded by lysosomes, aggresomes contribute for the maintenance of cellular homeostasis [41]. This theory is supported by the co-

localization of endogenous ATXN3 with aggresomes and the association of ATXN3 with dynein, a constituent of the complex charged of the transportation of misfolded proteins to MTOC [42].

The expression of ATXN3 also proved to regulate phosphatase and tensin homolog (PTEN), a tumor suppressor able to inactivate the phosphatidylinositol-3-kinase pathway (PI3K), commonly overactive in cancer. The depletion of ATXN3 leads to increased PTEN transcript, which in turn leads to higher levels of cellular PTEN protein. This induction is sufficient to downregulate PI3K signaling, leading to a decrease in cell viability. Thus, ATXN3 has been suggested as a potential therapeutic target in cancers with epigenetic downregulation of PTEN [43].

Other functions have been suggested for ATXN3, such as: a) the involvement in transcriptional regulation, since ATXN3 is able to influence the regulation of the expression of many genes [44] and repress acetylation by its interaction with histones [45]; b) the contribution to the cytoskeleton organization, as ATXN3 absence leads to disorganization of cytoskeleton constituents and a loss of cell adhesions [44, 46]; and c) the participation in myogenesis, since the *in vitro* silence of *ATXN3* in differentiating mouse myoblasts compromises their transition into muscle fibers [47].

1.1.2 Ataxin-3 like

The *ATXN3L* gene is an evolutionary conserved paralogue of *ATXN3*. This intronless gene on the X chromosome (Xp22.2) is predicted to encode the putative ATXN3L, although its presence has not yet been detected *in vivo*. ATXN3L would share 85% of sequence identity with ATXN3 and a similar crystal structure in its catalytic JD [13].

No specific function has yet been ascribed to ATXN3L; however, its JD consistently showed higher DUB activity for a range of substrates than the same domain of ATXN3, suggesting that ATXN3L might be a more active form of ATXN3. This difference may be explained by the fact that Ub appears to bind at different positions in the 2 proteins. Since only 3 mutations in ATXN3 are sufficient to nearly equalize the activity of ATXN3L, it is possible that evolution has selectively maintained the ATXN3 sequence in a less active state and a regulatory mechanism is able to keep ATXN3 activity dormant until required [13].

ATXN3L was also found to directly bind to Krüppel-like factor 5 (KLF5), a zinc finger oncogenic transcription factor that promotes cell proliferation and tumor growth. This interaction increases KLF5 protein stability since it leads to a decline in the level of ubiquitination. Functionally, knockdown of ATXN3L significantly decreased cell proliferation in 2 different breast cancer cell lines. These findings disclose a role of ATXN3L in the regulatory mechanism of KLF5 stability, in addition to the potential of the protein as a possible therapeutic target for breast cancer treatment [48]. Other studies have shed light on the relevance of ATXN3L in different cellular processes, namely: a) the potential therapeutic approach that ATXN3L may have in metastasis containment since depletion of this protein in epithelial cells leads to a decrease in cell mobility [49]; b) the capacity of ATXN3L to influence PTEN transcription, shared with ATXN3 [43]; and c) the possible role of ATXN3L in neurodegenerative disorders, suggested by the presence of a single nucleotide variant in *ATXN3L* (Gly332Asp) found in brain tissue, specific to Sporadic Alzheimer disease patients [50].

1.1.3 Josephins

JOSD1 and JOSD2 are two closely related proteins, sharing 50% of sequence identity, coded by genes located on chromosomes 22q13.1 and 19q13.33, respectively. JOSD1 was found to be widely expressed in several mouse tissues, including brain, heart, liver and skeletal muscle [11].

Josephin domain-containing proteins are structurally similar: both consist of a single JD, diverging only at the N-terminus, with JOSD1 displaying 12 additional amino acids; they differ, however, in a variety of properties, namely basal catalytic activity, modification by ubiquitination and subcellular localization. The catalytic activity of JOSD1 is, as described for ATXN3, directly regulated by ubiquitination in cells, while JOSD2 does not appear to have that same regulatory system. Also, an *in vitro* DUB activity assay showed that activity towards Ub chains is greater in JOSD2 than in JOSD1. At subcellular level, JOSD1 appears most concentrated at the plasma membrane, especially when ubiquitinated, whereas JOSD2 is diffusely present in the cytosol [11].

While the function of JOSD2 is still entirely unknown, JOSD1 was found to restrict PTEN transcription, similarly to both ATXN3 and ATXN3L [43] and have a role in the regulation of membrane dynamics, cell motility and three endocytic pathways (micropinocytosis, clathrin-mediated and caveolae-mediated endocytosis) [11].

1.2 Machado-Joseph disease (MJD)

When a mutation occurs in the *ATXN3* gene and the unstable expansion of the CAG trinucleotide repeat surpasses the pathological threshold, it produces a mutated form of ATXN3 with a prolonged polyQ tract at the C-terminus. This phenomenon is associated with MJD, also known as spinocerebellar ataxia type 3 (SCA3), a dominantly inherited neurodegenerative disorder of adult onset [51].

1.2.1 Clinical definition

MJD neurodegeneration profile involves neuronal loss in selective regions of the nervous system, including the cerebellum and substantia nigra [52]. This pathogenesis leads to a set of characteristic and pleomorphic clinical symptoms, including the hallmark progressive gait imbalance, dysarthria, dysphagia, spasticity and visual disorders (nystagmus, diplopia, progressive external ophthalmoplegia and apparent bulging eyes) [53]. Other problems have been reported in some individuals, such as dystonia, faciolingual fasciculations, neuropathy, progressive sensory loss and problems with urination [54].

The mean age of disease onset is approximately 40 years, but widely variable since first symptoms have been reported in individuals as young as 4 years and as old as 74/75 years [55, 56]. This variability discloses differences in the CAG repeat size: larger repeats cause, on average, earlier onset. For the same reason, the MJD phenotype can also be markedly distinct, which led researches to classify this variability into three clinical types: type I has the lowest mean disease AO and prominent spasticity and rigidity; type II is the most common form and begins in mid adult years; and type III is the late onset form of disease and is characterized by amyotrophy and generalized areflexia [57].

1.2.2 Epidemiology

SCAs are considered rare disorders, with an estimated prevalence of 0.3 to 2.0 patients per 100,000 individuals [58]. MJD is the most common form of dominant ataxias in most genetically characterized populations, displaying a worldwide frequency of 20-25% [59]. Even though present in many ethnic backgrounds, a strong geographic variation is associated with MJD. Among SCAs, its highest relative frequencies have

been reported in Brazil (69-92%) [60], Portugal (58-74%) [61], Japan (28-63%) [62], Singapore (53%) [63], China (48-49%) [64], the Netherlands (44%) [58] and Germany (42%) [65], with a heterogeneous distribution pattern sometimes being found even within a single country. The most representative case occurs in Portugal, where MJD is considered relatively rare in mainland (1/100,000) when compared with the high prevalence observed in the Azores (1/3,600), particularly in Flores Island (1/140) [66].

Based on haplotype analyses, two main ancestral mutations have been suggested to explain the present global geographic distribution of MJD. The most ancient mutational event occurred probably in Asia and has an estimated age of 5774 ± 1116 years (TTACAC or Joseph lineage), while the origin of the most recent mutation, with less than 2000 years old, is more controversial but thought to be strongly linked to Portugal and Portuguese emigration (GTGGCA or Machado lineage) [67].

1.2.3 Genetic features

While wild-type ATXN3 carries no more than 44 glutamines, MJD patients have been reported to encode mutated ATXN3 with 61-91 consecutive glutamines [68, 69]. Since MJD follows an autosomal dominant pattern of inheritance with an almost complete penetrance, all individuals who carry the expanded allele express the typical traits, and most of them are heterozygous. However, there are a few reported cases of homozygosity that show a more severe disease progression and an earlier AO, therefore suggesting a gene dosage effect on MJD clinical phenotype [55]. Furthermore, alleles with intermediate repeat length (45-56 CAGs), which are rather uncommon, present an incomplete penetrance and fall between the clearly normal and mutant ranges, thus creating discussion about their role in disease presentation [70].

In the parent-child allele transmission, $(CAG)_n$ instability has been documented with (1) repeat expansions being more common than contractions and (2) larger variations observed in paternal transmissions [71]. This intergenerational instability bias, together with the fact that longer repeats cause earlier onset disease, leads to a phenomenon denominated as anticipation, characterized by an earlier AO and more severe disease manifestations in offspring [72].

1.3 Machado-Joseph disease pathogenesis

Although the precise mechanism by which the CAG repeat expansion in MJD causes selective neuronal degeneration has not been fully clarified, growing evidence suggests that the pathogenic process occurs at the protein level. Usually, autosomal dominant genetic disorders, like MJD, can be caused by one of three mechanisms: partial deficiency of the protein, a dominant negative effect (loss-of-function) or a dominant positive effect (toxic gain-of-function). In the case of MJD, evidence is most consistent with the latter theory since the conformational modifications associated with the polyQ stretch expansion result in abnormally folded mutant ATXN3 that seems to lead to proteolytic cleavage, aggregation and altered binding properties (Figure 5) [73].

The neuropathological hallmark of MJD is the presence of intranuclear inclusions, large macromolecular aggregates that contain the mutated ATXN3 and, in MJD patients, are formed exclusively in the nucleus of neurons. Aggregation is not, however, confined to the nucleus since expanded ATXN3 also tends to form axonic and cytosolic aggregates in the brain of patients [23, 74]. Apart from polyQ-expanded ATXN3, these intracellular aggregates also contain several other constituents, including non-mutated ATXN3, transcription factors, Ub, proteasome constituents and molecular chaperones [75]. The impact of the sequestration of these molecules by aggregates in MJD pathogenesis is not consensual. While some argue that this ability leads to cytotoxicity by hindrance of transcription [76] and perturbation of Ub-dependent protein homeostasis systems [77], others suggest that this recruitment of molecules is actually a protective molecular mechanism to cope with the toxicity of the expanded proteins [78].

Furthermore, studies suggested that expanded ATXN3 tend to form aggregates as a result of polyQ expansion-induced misfolding and consequent transition to aggregation-prone conformations [78]. These ATXN3 aggregates were shown to contain β-sheet-rich fibrillary structures of amyloid nature [79]. In the conversion from ATXN3 soluble protein to amyloid aggregates, the monomeric and oligomeric intermediaries are envisioned to be the actual cytotoxic species. Research in other polyQ diseases demonstrated that β-stranded polyglutamine monomers and, particularly, soluble amyloid-like oligomers induced cytotoxicity in cultured cells [80, 81]. These structures may have a damaging effect on the cells, possibly through destabilization of membranes and/or recruitment of important proteins with consequent disruption of essential cellular systems [82]. Interestingly, the rate of aggregation was found to be

regulated by the length of the polyQ stretch of mutant ATXN3 with longer polyQ repeats being associated with increased aggregation, both *in vitro* and *in vivo* [83, 84].

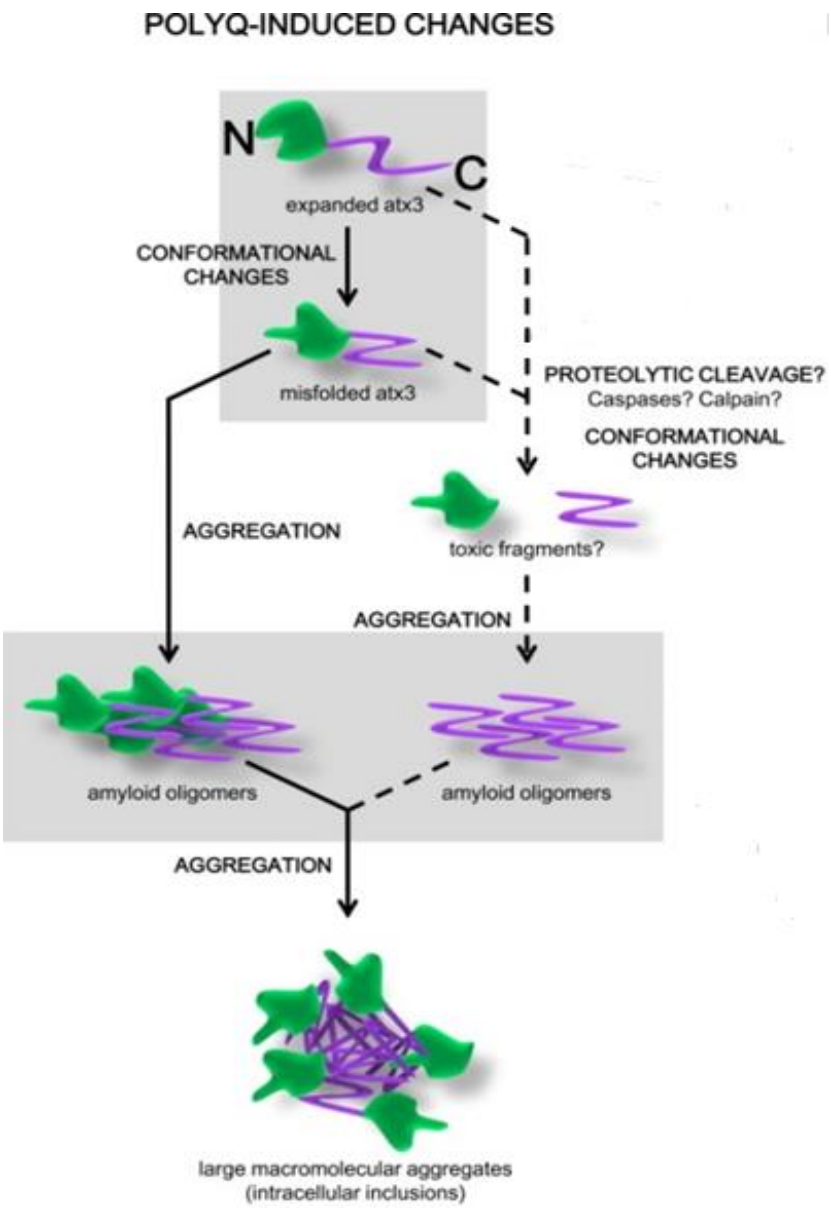


Figure 5 – Putative mechanisms mediating expanded ATXN3 toxicity. Expansion of the ATXN3 polyQ tract above 61 consecutive glutamines is known to initiate pathogenic mechanisms in neural cells leading to Machado-Joseph disease, though the actual toxic species involved are still debatable. This mutation may affect the physiological functions of ATXN3 through disturbance of the protein quality control system and transcription regulation, which may potentially lead to neuronal death and dysregulation of important cellular activities. The soluble amyloid-like oligomers that serve as intermediaries of the self-aggregation pathway may have a prejudicial effect on cells, possibly destabilizing cell membranes and/or sequestering proteins essential in various cellular systems. The proteolytic cleavage of ATXN3 and subsequent generation of toxic protein fragments may contribute for the oligomerization of the protein, eventually increasing the formation of aggregates of the expanded protein. Adapted from Matos *et al.* [85].

The possible contribution of the proteolytic cleavage of polyQ expanded ATXN3 for MJD pathogenesis has been termed as the 'toxic fragment hypothesis'. This event, reported in the brain of MJD patients [86], leads to the generation of cytotoxic and aggregation prone soluble fragments of approximately 36 kDa containing the expanded polyQ toxic entity. Members of caspase and calpain families are possibly involved in the proteolytic cleavage and consequent formation of the protein fragments [87, 88]. The expression of these fragments in transfected cell lines led to stronger aggregation and toxicity than full-length mutated ATXN3 [89]. Misfolded polyQ-expanded fragments are thought to initiate the aggregation process since the removal of the N-terminus region from mutated ATXN3 was required for *in vitro* and *in vivo* aggregation occurrence. Afterwards, these fragments may influence the JD structure or induce a misfolding event in the polyQ of full-length ATXN3, allowing their stable incorporation into the fibrillary aggregates [90].

The cellular distribution of ATXN3 is also relevant in the context of MJD pathogenesis. Several observations report the nucleus as an important player in disease development. The finding that artificially targeting mutated ATXN3 to the nucleus led to an aggravation of MJD phenotype in transgenic mice suggested that nuclear localization of the protein might be important for the manifestation of MJD symptoms *in vivo* [91]. The fact that the most conspicuous ATXN3 aggregates are found in the nucleus might be due to a less effective degradation of the protein in this compartment, when compared to the cytoplasm [89].

The polyQ expansion of ATXN3 may also produce changes in the function and interactions of the protein, possibly leading to cellular toxicity and thereby contributing for MJD development. For instance, polyQ tract expansion has been reported to increase the degree of interaction between the protein and p97/VCP, which also co-localizes with the aggregates in MJD patients [20]. This alteration likely causes a different distribution pattern for these proteins, possibly inducing endoplasmic-reticulum proteotoxic stress that may contribute to degenerative mechanisms [38].

Augmentation of the polyQ stretch also impairs the transcriptional repression that normal ATXN3 is able to perform on histone acetylase activity. This phenomenon leads to increased acetylation of total histone H3, which was indeed observed in both mutant ATXN3 overexpressing cells and human MJD brain material [92]. The prolongation of the polyQ tract in the mutated protein may also alter its binding properties, leading to augmented expression of certain genes that may be involved in MJD development.

1.4 Compensation for partial loss-of-function of normal ATXN3 in MJD

One of the most common features associated with spinocerebellar ataxias, which is also present in MJD, is the well-characterized inverse relationship between CAG repeat length in the abnormal allele and the age-at-onset (AO) of patients. Accordingly, shorter but pathological expansions are associated with elderly onset, while very large expansions can originate an onset in the early childhood years. The negative correlation between the size of the expanded allele at *ATXN3 loci* and the MJD AO is unable to fully explain all this phenotypic variability, but only 45-76% [68]. Therefore, other genetic modifiers – genes capable of affecting the phenotype of the mutant gene without having an obvious effect on the normal condition – should help to justify the remaining part of the onset diversity found in MJD.

The normal *ATXN3* repeat size has also been found to have a small but significant modulating effect of around 1% on MJD AO. Patients with identical lengths at the mutant allele but larger normal alleles will predictably have a later onset [93]. The identification of other MJD genetic modifiers resulted from the great effort of several researchers, although only three additional factors have been described: a) the presence of the ε2 allele in the alipoprotein E (APOE, a ubiquitous protein involved in lipid storage and metabolism) was found to increase the risk of earlier MJD onset in approximately 5 years, accounting for a slight but significant 1% variance in the AO [94, 95]; b) the repeat size in 3 polymorphic $(CAG)_n$-containing genes affect MJD AO, with longer alleles in *ataxin-2* (*ATXN2*) and *atrophin-1* (*ATN1*) accelerating it, while the opposite effect has been observed for *huntingtin* (*HTT*) [96]; and c) the mitochondrial NADH dehydrogenase subunit 3 (MT-ND3), with a polymorphism (rs2853826) responsible for a 3-year anticipation of the AO in male patients, corresponding to a nearly 2% AO variance (although this effect decreased to 0.7% when englobing both genders) [97].

Since these modifiers contribute to explain a small percentage of the MJD AO divergence, other genetic and/or environmental factors must also play an important role in this disease. The list of other potential MJD modifiers also include: the haplotype of 2 single nucleotide polymorphisms – rs709930 and rs910369 – in *ATXN3* 3' UTR [98], DNA methylation in the *ATXN3* promotor region [99], polymorphisms in *glutamate receptor ionotropic, kainate 2* (*GRIK2*), *interleukin-1 beta* (*IL1B*) and *neural precursor cell expressed developmentally downregulated 8* (*NEDD8*) *and 9* (*NEDD9*) genes [100], the presence of the C allele in the *interleukin-6* (IL6) [101], sequence variations

at the *glucosidase beta acid* (*GBA*) gene [102] and the haplotype background of expanded MJD alleles [103].

From all these potential modifiers, we were particularly intrigued by the described association of MJD AO and the normal *ATXN3*, which suggested a modulatory role of the wild-type allele in disease pathogenesis. Since MJD is an autosomal dominant disease, heterozygous individuals have both normal and expanded forms of ATXN3 in the same cell [104]. As suggested by the proportionality of the polyQ stretch of both normal and expanded protein and their co-aggregate formation in other polyQ disorders [105], normal ATXN3 with a longer polyQ tract might increase the efficiency of the association with mutant ATXN3. This form of ATXN3 would then be sequestrated in higher amounts to aggregates, therefore avoiding aberrant interactions with other proteins. The protective role of normal ATXN3 against polyQ neurotoxicity *in vivo* was also suggested following a *Drosophila* study that reported that flies expressing both wild-type and mutated ATXN3 did not present the phenotype found when only the expanded protein was expressed. This capacity of normal ATXN3 to mitigate MJD neuropathogenicity requires UIM functionality, Ub protease activity and normal proteasome functionality, indicating that it occurs through Ub pathways involved in quality control mechanisms [106].

Since the non-mutated ATXN3 may have a neuroprotective role in MJD development, the study and analysis of the remaining members of MJD protease subfamily in the clinical presentation of MJD becomes even more relevant. The *in vitro* corroboration that all MJD subfamily members have DUB activity allied to the residue similarity between their catalytic JDs [12] encouraged us to test the hypothesis that ATXN3L, JOSD1 and JOSD2 may exert similar functions to ATXN3 and therefore compensate for a partial loss-of-function of the normal form in MJD patients.

2. Aims

ATXN3 is the best studied protein from the MJD family of DUB, with its contribution in the form of a polyglutamine expansion for MJD manifestation being well described; nevertheless, there is still a lack of conclusive information regarding the function and regulation of the other elements of this class (JOSD1, JOSD2 and ATXN3L). Therefore, this project aimed to explore the potential relevance of the 3 less studied MJD proteases in the clinical presentation of MJD. For that purpose, DNA samples from a cohort of MJD patients from different families, with known expanded CAG repeat length and AO, were sequenced for regulatory and exonic regions of our candidate genes.

Accordingly, the main objectives of this work were:

1) to optimize multiplex PCR reactions for the amplification of exonic and regulatory regions of *JOSD1*, *JOSD2* and *ATXN3L*;
2) to search for variants in our candidate genes and calculate the percentage of MJD AO variability they may explain in addition to the known effect of the expansion size; and
3) to evaluate, *in silico*, the possible roles of *JOSD1*, *JOSD2* and *ATXN3L* variants found to affect MJD AO independently of the CAG repeat, as genetic modifiers of the MJD clinical presentation.

3. Subjects and Methods

3.1 Subjects

DNA samples from 100 MJD patients, belonging to different Portuguese families, were analysed in this study. All samples are from UnIGENe, CGPP – IBMC, where the molecular genetics diagnosis of MJD is performed; written informed consent has been previously provided by all individuals. 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 expanded allele and AO of patients are annotated in Supplementary Table A. DNA quantification was performed with Nanodrop spectrophotometer to make work aliquots with a final DNA concentration of approximately 7.5 ng/µL.

3.2 Primer design

For the amplification of our candidate genes, sequence specific primers were required. Regarding *ATXN3L*, primers previously outlined by M.I. Martins [107] were used, while for both *JOSD1* and *JOSD2* we proceeded to primer design. For that purpose, we retrieved *JOSD1* and *JOSD2* sequences from NCBI (ID: 9929 [108] and 126119 [109], respectively) and annotated regulatory and exonic regions from the canonical transcript – JOSD1-001 and JOSD2-001 (ID: ENST00000216039 [110] and ENST00000598418 [111], respectively) – on Geneious 5.5.8.

Primers were delineated in the online software Primer3web, version 4.0.0, meeting the following criteria: 19 to 22 base pairs, a percentage of GC between 45% and 65%, and a melting temperature comprised between 60ºC and 70ºC. Next, we tested for the occurrence of hairpins and primer self-dimers using OligoCalc algorithm; sequences prone to form hairpin structures with, at least, 4 bp or autodimers with 5 or more bp were discarded. The alignment tool BLAT (human genome, assembly GRCh38/hg38) was then used to guarantee the specificity of designed primers to the target sequences. Subsequently, each remaining primer was subjected to a nucleotide-nucleotide BLAST (database: nucleotide collection; organism: human; expected threshold: 20) to find somewhat similar sequences. Primers with less than 2 mismatches in the 3’ end or 3 mismatches in any position were considered inappropriate. Finally, the compatibility of multiple primer sequences was checked with

AutoDimer, a software that detects the formation of primer-dimers in short DNA oligomers, thus helping us in the development of multiplex reactions.

The list of the final selection of primers (Sigma-Aldrich) is shown in Table 1. No suitable primers for the complete amplification of JOSD1 5' UTR were found, probably due to the high GC content of this region, which complicates primer design due to high annealing temperatures and the formation of self-dimers, mismatches and secondary structures [112].

Table 1 – Primers designed to specifically amplify *JOSD1*, *JOSD2* and *ATXN3L* regulatory and exonic regions. F – primer forward; R – primer reverse; UTR – untranslated region.

JOSD1, *JOSD2* and *ATXN3L*					
Reaction	$T_{annealing}$	Gene	Region	Primer sequence (5'-3')	
Pentaplex	64ºC	*JOSD1*	Part of 5' UTR and exon 1	F	AAGATGCCAAGGAGAAGGGTC
				R	GGGGAAACGGAAACAAGGTC
			Exons 2 and 3	F	GCTCAGTGTCCCCAACCTAA
				R	CAAGTCTCAGCCCTCCTCTT
			Exon 4 and part of 3' UTR	F	TGTGACCACTGGCCTTAAGT
				R	CCGAGGGTAGAGTGAGGTTT
		JOSD2	Exon 1 (5' UTR)	F	TCCCCGGACACCAAGGAAT
				R	TATGCGAAAGAAGGGCCTCT
			Exon 2	F	CCAGAGTCCCAGCGTCTAG
				R	GTCCCAGCTCAATCAATGGC
Duplex	62ºC	*JOSD2*	Exon 3	F	GCTTCCCACATTCATGCCAA
				R	TGAGTTGTGGGGTCTGAAGC
			Exons 4 and 5 and part of 3' UTR	F	TCCATGAAGTGCTGGCCTT
				R	TTCAGCAACATTTCACCGGC
Singleplex	60ºC	*ATXN3L*	Complete CDS	F	CTCTAACTAGGATACCAGCAAAG
				R	GGAAAAAGTTCTATGGCAAGAGC

3.3 Amplification of *JOSD1*, *JOSD2* and *ATXN3L*

After optimization of polymerase chain reactions (PCRs) with the designed primers described above, one pentaplex, one duplex and one singleplex assays were performed. For the first 2 reactions, 0.5 µL of each primer (2.5 µM) and 1 µL of DNA sample (7.5 ng/µL) were mixed, together with 10 µL of Taq polymerase 2x (MyTaq H8 Mix; BIOLINE) for pentaplex PCR and 5 µL for duplex PCR. Moreover, 4 µL and 2 µL of Mol Bio grade water (5 PRIME) was added to fulfill the final volume of 20 µL for the pentaplex reaction and 10 µL for the duplex reaction, respectively. For the singleplex

assay, 0.5 µL of each primer (2.5 µM) and 2 µL of DNA sample (7.5 ng/µL) were mixed, together with 5 µL of NZYLong 2x Green Master Mix (NZYTech) and 2 µL of Mol Bio grade water to fulfill the final of 10 µL. PCR conditions are described in Figures 6 and 7.

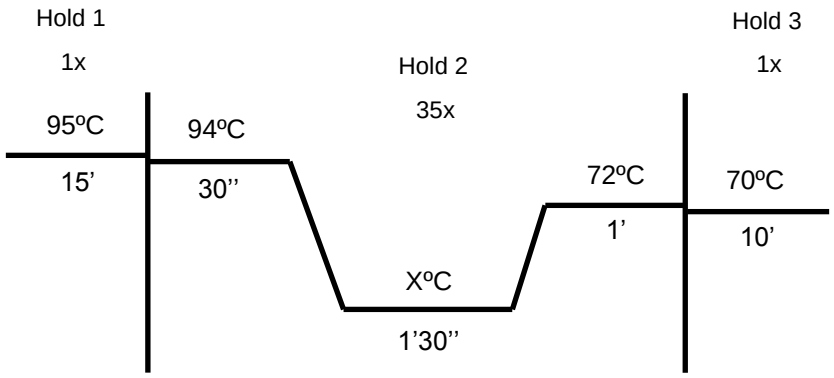


Figure 6 – PCR protocol for the amplification of *JOSD1* and *JOSD2 loci*. Length of time (' – minutes; '' – seconds) and temperature (ºC) are annotated for each step of the protocol. X (annealing temperature) varies according to the reaction: 64ºC for pentaplex and 62º for duplex.

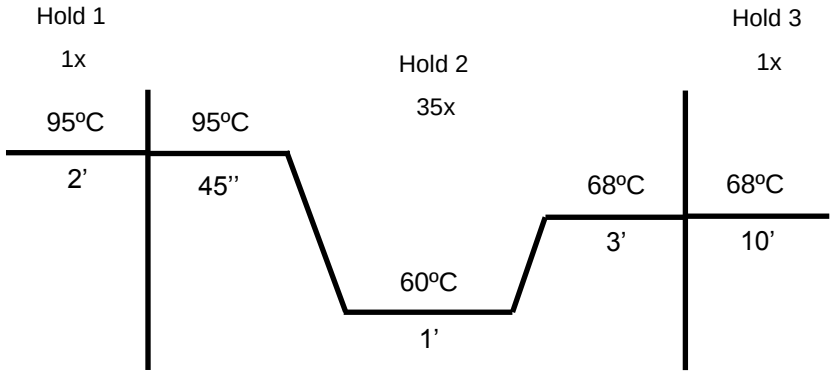


Figure 7 – PCR protocol for the amplification of *ATXN3L loci*. Length of time (' – minutes; '' – seconds) and temperature (ºC) are annotated for each step of the protocol.

3.4 Detection of amplified products

In order to confirm the presence and specificity of the amplified DNA fragments, PCR products were subjected to electrophoresis in a polyacrylamide gel. Two previously cleaned glass supports, separated by an hydrophilic gel-bond film (GEL-FIX^{TM}; SERVA), were used to produce a 3 mm thick gel composed by 3 mL of a 40% (w/v) acrylamide:bisacrylamide (19:1) solution (AccuGel 19:1; National Diagnostics, Inc.), 170 µL of the oxidizing agent 2.5% ammonium persulfate (Promega) and 7 µL of the catalysis agent TEMED (National Diagnostics, Inc.). After complete polymerization at room temperature, the gel was placed in the electrophoretic system (Multiphor II; GE

Healthcare) and 1.2 µL of each sample was loaded in the wells. To allow the horizontal run, paper strips soaked in buffer were used at both anode and cathode. The addition of bromophenol blue dye (Merck) to the anode strip allowed the monitorization of the process. With the system refrigerated by a thermostatic circulator (MultiTemp III) at 4ºC, the electrophoresis run was initially performed at 180 V, and later increased to 220 V once the samples leaved the wells and started to migrate through the gel. The electric power supply (Consort EV243) was turned off when the marked dye approached the anode strip.

Thereafter, a silver staining was performed in the gel. This coloration method involved: (1) a fixation step of the DNA, with 10% ethanol (Merck) for 10 minutes, followed by 1% nitric acid (Merck) for 5 minutes, both under agitation (GFL® Shaker 3011); (2) two washes with deionized water, for 10 seconds each; (3) a coloration step with 0.2% silver nitrate solution (Merck), for 20 minutes, under agitation and protected from the light; (4) two more washes with deionized water for 10 seconds each; and finally, (5) a revelation step of the DNA fragments with a solution of 0.28 M sodium carbonate (Applichem) and 0.02% formaldehyde (Merck) until bands started to appear in the gel. To stop the revelation reaction, 10% acetic acid (Merck) was added to the gel for approximately 10 seconds. Subsequently, the gel was left in tap water for at least 12 hours and then dried at room temperature.

3.5 DNA sequencing

After guaranteeing the specificity of the amplification, DNA sequencing was performed in the resulting fragments. Primarily, an initiatory purification was done with ExoFastAP (Thermo Scientific), a phosphatase that degrades unreacted primers and remained dNTPs. A mixture of 1.85 µL of PCR product and 0.95 µL of ExoFastAP was subjected to 37ºC for 15 minutes, followed by 15 minutes at 80ºC to activate and inactivate the enzyme, respectively. The sequencing mix was then produced by adding 2.5 µL of purified product, 0.5 µL of primer (2.5 µM), 0.6 µL of sequencing buffer (2.5x) (Applied Biosystems), 0.8 µL of BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and 0.6 µL of Mol Bio grade water, therefore completing a final volume of 5.0 µL. After an initial denaturation at 96ºC for 2 minutes, the PCR sequencing procedure consisted of 30 cycles under the following conditions: 96ºC for 15 seconds, 58ºC for 5 seconds and 60ºC for 2 minutes. The process was concluded with a final extension at 60ºC for 10 minutes. An ultimate purification was performed

using Sephadex G-50 (GE Healthcare), a cross-linked dextran matrix able to separate molecules over broad molecular weight, thus isolating the sequencing products from unincorporated nucleotides and primers. The centrifugation (Hettich Mikro 2000) of 750 µL of Sephadex during 4 minutes at 4400 rpm allowed the formation of columns. The centrifugation protocol was repeated after loading the PCR sequencing product in the center of the column, resulting in a deposit containing the final product. The last step required the addition of 12 µL of Hi-Di^{TM} formamide (Applied Biosystems) to ensure the stability of single-stranded DNA for capillary electrophoresis in an ABI PRISM 2130x/Genetic Analyzer (Applied Biosystems).

3.6 Statistical and *in silico* analyses

Sequenced fragments were compiled and examined with Sequencing Analysis v5.2 software and Geneious 5.5.8. The latter program was used to perform global alignments with free end gaps. The analysis of *JOSD1*, *JOSD2* and *ATXN3L* diversity was accomplished by annotating all variations to the reference sequence.

We calculated the proportion of gender distribution and produced histograms, with respective mean and standard deviation, for AO and CAG repeat length in normal and expanded alleles. Normality for these variables was tested using the Kolmogorov-Smirnov test. The association between the trinucleotide repeat length in the MJD allele of affected individuals and disease was assessed using a linear fitting 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 analysed *loci* and Hardy-Weinberg equilibrium was tested. Equality of variances between genotypes was verified by the Levene's test. Pearson's correlation was performed to determine variants with a linear dependence between them. An analysis of covariance (ANCOVA), using the expansion size as covariate, was conducted to compare estimated AO between variant alleles and variant genotypes. Allelic variants with a significant effect resulting from ANCOVA were used, together with the CAG repeat size in expanded alleles, to determine a new model, using multiple linear regression technique, to predict disease AO. All statistical analyses were performed in IBM SPSS Statistics 24 (© IBM Corp. Released 2016) with the exception of Hardy-Weinberg principle, executed in Arlequin 3.11 (© L. Excoffier Released 2006). A statistically significant result lower than 0.05 was considered for all tests performed.

The alterations caused by non-synonymous polymorphisms in their respective putative coding sequences were predicted and the possible impact of amino acid substitutions on the stability and function of proteins was also assessed by two different computational tools: PolyPhen-2 v2.2.2r398 software [113] and SIFT human protein algorithm [114] (© J. Craig Venter Institute).

4. Results

4.1. Clinical and genotyping data

Our cohort of 100 MJD patients included 51 females and 49 males, all from different families. The mean AO was 42.12 ± 12.59 years, with a predominance of cases (42%) between 40-50 years, as expected for this late-onset disorder (Figure 8). Regarding the *ATXN3* repetitive tract, the most common alleles in normal chromosomes displayed 14, 21, 23 and 28 CAGs (accounting 65% of cases), whereas the mean size of expanded alleles was 71.74 ± 3.78 CAGs with a major occurrence of 69 repeats (17%).

Taking into account that our cohort included more than 50 patients, Kolmogorov-Smirnov test of normality with Lilliefors correction was performed in the 3 metric independent variables. In normal alleles, the CAG repeat length was found to have a non-normal distribution (H0 cannot be rejected, $p < 0.0005$), whereas disease AO and expanded allele sizes followed a Gaussian distribution (H0 rejected, $p > 0.05$).

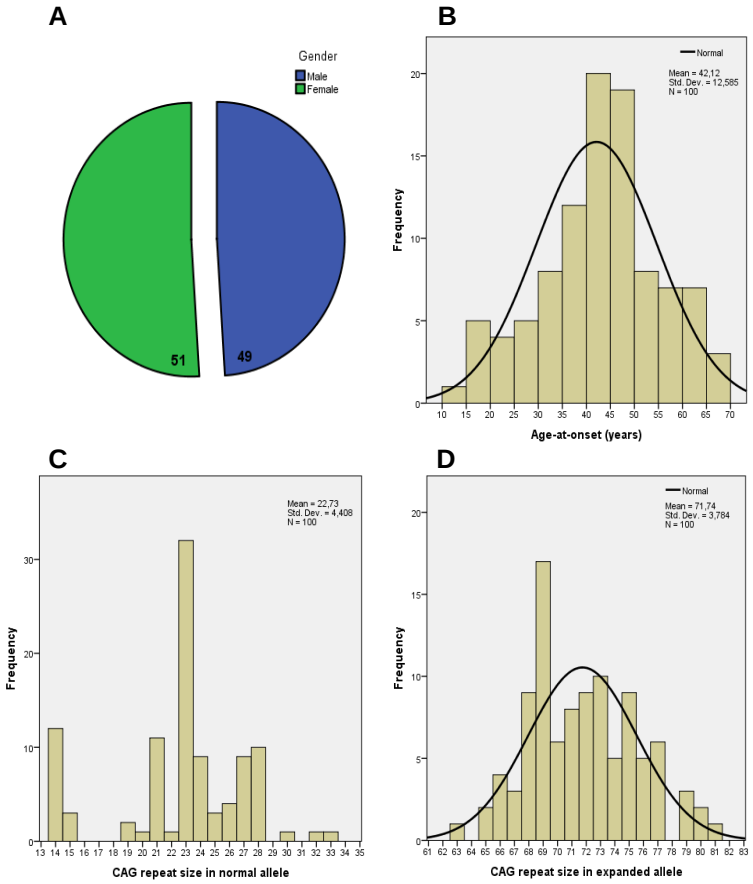


Figure 8 – Demographic, clinical and genotyping data for our MJD cohort. Distribution of the studied 100 MJD patients according to gender (A), diagnosed AO (B), CAG repeat size in the normal (C) and expanded (D) alleles. Variables AO and CAG repeat size in expanded allele follow a normal distribution.

4.2 Model A: Correlation of expanded CAG repeat length and AO

After confirming several key assumptions associated with linear regression (linear relationship, multivariate normality, no multicollinearity, no autocorrelation and homoscedasticity), we observed a negative correlation between the CAG repeat size in the expanded allele of MJD patients and AO (N = 100, r = - 0.74, $p < 0.0005$) (Figure 9). In this model (model 1), the explanation of the AO variance provided by the expanded allele size was 54.7% ($F(1,98) = 118.535$, $p < 0.0005$). The regression line associated to this analysis is represented by the equation (AO) = - 2.461 x (CAG_n in expanded allele) + 218.675. No relationship was found between AO and gender ($p = 0.519$) or CAG repeat size in normal alleles ($p = 0.325$).

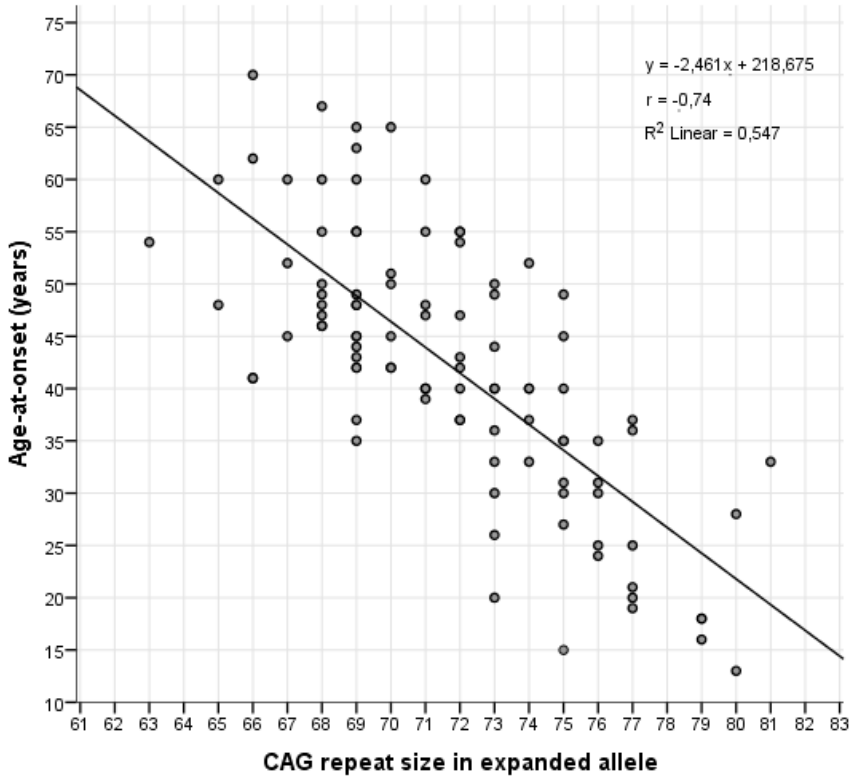


Figure 9 – Linear regression between CAG repeat length in the MJD chromosome of 100 patients and disease AO. A linear coefficient of determination of $r^2 = 0.547$ was obtained.

4.3 Nucleotide variation

We next analysed the genetic variation of our candidate genes - *JOSD1*, *JOSD2* and *ATXN3L* - by sequencing regulatory and exonic regions, as well as the areas surrounding exons that may affect splicing. Table 2 summarizes data found for the 16 genotyped single nucleotide polymorphisms (SNP), distributed by the 3 genes;

genotypes for all 100 MJD patients is detailed in Supplementary Table A. In *JOSD1*, all 3 polymorphisms detected in our cohort are in coding regions: two synonymous (rs2072800 and COSM3713736) and one non-synonymous (rs6001200), resulting in the change of a serine to an arginine in position 48 of the JOSD1 protein. In the case of *JOSD2*, half of the 8 identified SNPs were regulatory variants (rs550589305 in 5'UTR; rs376569327, rs117711212 and rs4830842 in 3'UTR). The 4 remaining variants include two synonymous alterations (rs376400138 and rs11548260), one intronic deletion (rs796424878) and a splice region variant (6 bp upstream of exon 2 and 23 bp upstream of *JOSD2* start codon; rs141848929). Regarding *ATXN3L*, 5 SNPs were found in our analysis: two 3'UTR variants (rs58878737 and rs41297291), one synonymous change (rs17322143) and two non-synonymous substitutions. Of the latter group, presence of rs16999010*A allele induces an amino acid alteration from leucine to phenylalanine at residue 266, while rs4830842*T leads to the conversion of glycine to aspartic acid in position 322 of ATXN3L.

4.4 Relevance of analysed SNPs on AO modulation

All examined *loci* were found to be in conformity with Hardy-Weinberg equilibrium expectation, with the exception of rs17322143 and rs4830842 (data not shown). This disequilibrium may be due to a reduced number of samples included in the test of these SNPs since both are in the *ATXN3L* gene, located on the X chromosome, with only 51 females analysed. We discarded the hypothesis of systematic genotyping errors since all 3 genotypes (homozygous for the most common allele, heterozygous and homozygous for the rare allele) were found in the analysis of both *loci*.

We next tested the equality of variances between different genotypes of each polymorphism to further perform ANCOVA, an analysis that assume this homogeneity: for all 16 SNPs, the null hypothesis (H0) was rejected by Levene's test ($p > 0.05$).

The effect of the minor allele of each polymorphism on the disease AO, while having in consideration the already known effect of the expanded $(CAG)_n$, was tested by ANCOVA (Table 2). The pair of alleles rs550589305*C / rs141848929*A ($F(1,98) = 6.009$, $p < 0.05$) in *JOSD2* and rs16999010*A allele ($F(1,98) = 5.271$, $p < 0.05$) in *ATXN3L* proved to have a significant impact on AO modification.

We next wanted to test the association between SNPs to evaluate the relevance of correlated SNPs in the modulation of the AO of patients. Pearson product-moment

correlation coefficient revealed a complete positive association between rs550589305 and rs141848929 (r = 1.00, $p < 0.0005$), both in *JOSD2*. Relevantly, rs2072800 (r = 0.241, $p < 0.05$), rs11548260 (r = 0.812, $p < 0.01$) and rs41297291 (r = 0.704, $p < 0.01$) also showed a significant relationship with the aforementioned pair. In addition, variants COSM3713736 and rs796424878 in *JOSD1* and *JOSD2*, respectively, were also found to be significantly correlated (r = 0.571, $p < 0.01$). Thus, when we tested the statistically correlated polymorphisms together, COSM3713736*A allele ($F(1,98) = 4.447$, $p < 0.05$) and rs796424878 deletion ($F(1,98) = 7.665$, $p < 0.01$) also proved to have influence in AO variance. No significant results were achieved when the other associated SNPs were tested in simultaneous (data not shown).

The effect of the genotype (hemizygous genotype for *ATXN3L* polymorphisms in male individuals) was also assessed by ANCOVA when all 3 genotypes were observed in our cohort (rs17322143 and rs4830842) but no different outcome was achieved (data not shown).

4.5 AO/$(CAG)_{exp}$ linear regression model B: Addition of polymorphisms effect

To test the cumulative effect of the 5 SNPs shown to influence the AO in previous tests (rs16999010, rs550589305 and rs141848929 when SNPs were analysed individually, and the correlated COSM3713736 and rs796424878 SNPs), we applied a new model (model B), which takes into consideration this set of SNPs in addition to the CAG repeat size in the expanded allele. All variables together were shown to predict 63.2% of AO variation in our cohort ($F(5,94) = 32.267$, $p < 0.0005$). Taking into account that 54.7% is explained by the expansion size alone (model A), an improvement on the model was observed when the presence/absence of these SNPs was added, with their contribution elucidating additional 8.4% to the variance of AO (N = 100, F change = 5.390, $p = 0.001$).

Table 2 – Genetic variations found in the analysed regions of *JOSD1, JOSD2* and *ATXN3L* and their influence on the AO of MJD patients after controlling for the known effect of the expansion size . Base positions of all 3 genes start in the first base of the canonical transcript (JOSD1-001, JOSD2-001 and ATXN3L-001, respectively). The number of chromosomes analysed was 200 for both *JOSD1* and *JOSD2* fragments and 151 for *ATXN3L*. Sequence alterations refer to the (+) strand. Types of polymorphisms are identified by color: green for synonymous SNPs, yellow for nonsynonymous SNPs, brown for untranslated SNPs and red for intronic SNPs. ANCOVA test results (F ratios and respective *p*-values) assuming full dominance for the minor allele. F ratios assuming 98 degrees of freedom. NA – not applicable.

Gene	Region	SNP ID	Variation name	Base position in transcript	Sequence alteration	Absolute allele frequency	Relative allele frequency	ANCOVA test results	
								F ratio	*p*-value
JOSD1	Exon 1	rs2072800	JOSD1-001:c.6A>G	686	A/G	193/7	0.965/0.035	3.268	0.074
		rs6001200	JOSD1-001:c.144G>C, p.Ser48Arg	824	G/C	199/1	0.995/0.005	1.696	0.196
	Exon 2	COSM3713736	JOSD1-001:c.246G>A	926	G/A	199/1	0.995/0.005	0.397	0.530
JOSD2	Exon 1	rs550589305	JOSD2-001:c.-244G>C	11	G/C	198/2	0.990/0.010	6.009	0.016
	Intron 1	rs141848929	JOSD2-001:c.-17-6G>A	NA	T/A	198/2	0.990/0.010	6.009	0.016
	Exon 3	rs376400138	JOSD2-001:c.243G>A	497	G/A	199/1	0.995/0.005	0.563	0.455
	Intron 3	rs796424878	JOSD2-001:c.272+40_272+41delTG	NA	TG/-	197/3	0.985/0.015	3.517	0.064
	Exon 4	rs11548260	JOSD2-001:c.360C>T	614	G/A	197/3	0.985/0.015	1.699	0.195
	3'UTR	rs376569327	JOSD2-001:c.*4A>G	825	A/G	199/1	0.995/0.005	0.658	0.419
		rs117711212	JOSD2-001:c.*9C>T	830	C/T	197/3	0.985/0.015	0.010	0.922
		rs3087582	JOSD2-001:c.*39C>T	860	C/T	184/16	0.920/0.080	0.213	0.646
ATXN3L	CDS	rs16999010	ATXN3L-001:c.796G>A,p.Leu266Phe	1261	G/A	150/1	0.993/0.007	5.271	0.024
		rs17322143	ATXN3L-001:c.939C>T	1404	C/T	134/17	0.887/0.113	2.089	0.152
		rs4830842	ATXN3L-001:c.995C>T,p.Gly322Asp	1460	C/T	97/54	0.642/0.358	0.010	0.919
	3'UTR	rs58878737	ATXN3L-001:c.*147T>G	1680	T/G	150/1	0.993/0.007	2.743	0.101
		rs41297291	ATXN3L-001:c.*246T>C	1779	T/C	150/1	0.993/0.007	2.591	0.111

To estimate the AO taking into account the above mentioned group of SNPs alongside with the CAG expansion size, we applied the following formula: (AO) = 219.424 – 2.477 x ($(CAG)_n$ in expanded allele) – 21.431 x (presence/absence of COSM3713736*A allele) + 14.624 x (presence/absence of rs550589305*C / rs141848929*A alleles) + 16.432 x (presence/absence of rs796424878 deletion) – 18.614 x (presence/absence of rs16999010*A allele). For this model B, the influence of $(CAG)_n$ expansion in AO reduces 0.1% (54.6% in the new model), while the presence of COSM3713736*A allele (*JOSD1*), rs550589305*C/rs141848929*A alleles (*JOSD2*), rs796424878 deletion (*JOSD2*) and rs16999010*A allele (*ATXN3L*) improves the model by 1.93%, 2.66%, 3.39% and 2.19%, respectively.

4.6 Impact of model B in the MJD AO prediction

In the study of our cohort of 100 MJD patients, 6 individuals presented at least one variant, shown to modulate AO in model B, in one of our candidate genes *ATXN3L*, *JOSD1*, and *JOSD2*. The interception of expansion size and disease AO of these subjects is highlighted by color in Figure 10. The observed AO can be compared with the one predicted by the linear regression of model A (when only one variable, the CAG repeat length of expanded alleles, was taken into consideration); the two dashed lines represent the upper and lower values of a 30% variation of this predicted AO given any $(CAG)_n$. When we analyse the 6 MJD patients from our cohort who presented SNP alleles significantly relevant to the AO according to model B, 5 of them showed a large deviation from the tendency line (even falling beyond the dashed lines).

We then looked for the effect of the presence or absence of the SNP minor alleles on AO means and standard deviations (Table 4). An older onset was found to be associated to the COSM3713736*A allele (4 years), pair rs550589305*C / rs141848929*A (10 years) and rs796424878 deletion (13 years). An opposite influence was found by rs169999010*A allele since the patient who carried this allele presented an AO 22 years earlier than the mean onset of the most common genotype. The low occurrence of the rare allele in these polymorphisms leads to a discrepancy in the distribution of both analysed groups, not allowing the use of a statistical test to infer the difference between means.

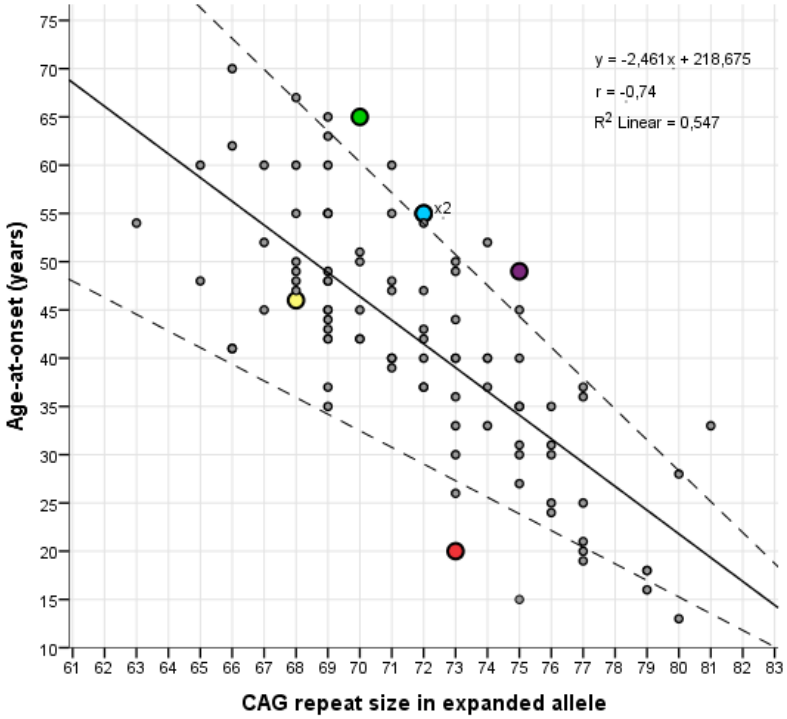


Figure 10 – Assessment of the AO/$(CAG)_{exp}$ interception in patients that present SNPs that significantly affect AO in our cohort. Full line represents the linear regression between CAG repeat length in the MJD chromosome of 100 patients and disease AO, while dashed lines represent 30% variation, both upwards and downwards, on disease AO. Colored dots denote the 6 subjects who carry alleles in analysed SNPs in *JOSD1*, *JOSD2* and *ATXN3L*, shown to influence AO variance in this cohort: patient #31 (blue) - presence of rs550589305*C and rs141848929*A alleles; patient #39 (yellow) – presence of COSM3713736*A allele and rs796424878 deletion; patient #45 (blue) - presence of rs796424878 deletion; patient #57 (green) - presence of rs796424878 deletion; patient #62 (violet) - presence of rs550589305*C and rs141848929*A alleles; and patient #89 (red) - presence of rs16999010*A allele.

Table 3 – Mean and standard deviation (SD) of MJD AO when the presence or absence of SNP minor alleles is observed. Types of polymorphisms are identified by color: green for synonymous SNPs, yellow for nonsynonymous SNPs, brown for untranslated SNPs and red for intronic SNPs.

Gene	SNP ID	Minor allele	Age-at-onset (years)		N
			Mean	SD	
JOSD1	COSM3713736	Presence	46	-	1
		Absence	42.08	12.643	99
JOSD2	rs550589305 / rs141848929	Presence	52	4.243	2
		Absence	41.92	12.626	98
	rs796424878	Presence	55.33	9.504	3
		Absence	41.71	12.483	97
ATXN3L	rs16999010	Presence	20	-	1
		Absence	42.34	12.448	99

Next, we evaluated models A and B by comparing the AO predicted by each one with the observed values (Table 5). When only the expansion size is considered to affect MJD AO (model A), the predicted AO is 19 years later for patient #89, 5 years later for patient #39, 19 years earlier for patient #57 and 14 years earlier for patients #31, #45 and #62. On the other hand, all AO projected by model B varied between 0.96 and 1.05 in the predicted/observed AO ratio.

Table 4 – Comparison between AO predicted by the 2 models and observed AO, and respective ratios. The color associated to patient ID relates to Figure 9.

Patient ID	$(CAG)_n$	Age-at-onset (years)			Ratio between predicted and observed AO
		Predicted		Observed	
#31	72	Model A	41.483	55	0.754236
		Model B	55.704		1.0128
#39	68	Model A	51.327	46	1.115804
		Model B	46.465		1.010109
#45	72	Model A	41.483	55	0.754236
		Model B	57.512		1.045673
#57	70	Model A	46.405	65	0.713923
		Model B	62.466		0.961015
#62	75	Model A	34.100	49	0.695918
		Model 2	48.273		0.985163
#89	73	Model A	39.022	20	1.9511
		Model B	19.989		0.99945

4.7 Phenotypic influence of polymorphisms

After showing the effect of a set of polymorphisms in our candidate genes in the improvement of a model to predict MJD AO, we next attempted to justify their role on disease phenotype.

The variant rs16999010 (ATXN3L-001:c.796G>A,p.Leu266Phe) is a non-synonymous alteration in ATXN3L, resulting in a switch from a leucine (present in most patients) to a phenylalanine (coded by the allele found to influence MJD AO). Both amino acids are neutral, non-polar and hydrophobic but while leucine is non-aromatic, phenylalanine has an aromatic conformation. Moreover, both are thought to be rarely involved in protein functions, but frequently are important in substrate recognition. Since this position is closely aligned (5 bp upfront) with the second UIM of ATXN3, a relevant impact on ubiquitin-binding can be hypothesized. Two prediction tools of the

non-synonymous variations effect on protein function have shown slightly different conclusions: while PolyPhen-2 predicts the variation to be benign (i.e., most likely lacks any phenotypic effect) to the protein sequence (with a score of 0.359), SIFT classifies the polymorphism as tolerated (with a score of 0.07), but close to the deleterious limit (inferior to 0.05).

Variants rs550589305 (JOSD2-001:c.-244G>C) and rs141848929 (JOSD2-001:c.-17-6G>A) are situated in the regulatory region of *JOSD2*, which may result in alterations on the expression of this gene. Therefore, we analysed the large collection of ChIP-seq experiments performed by the ENCODE project to gain insight into the transcription factors that could bind to the regions encompassing our candidate SNPs: more than 60 transcription factors were found to bind the region where rs550589305 variant is present; rs141848929 is located in the binding site of POLR2A, REST and NR3C1 transcription factors.

The small deletion polymorphism rs796424878 (JOSD2-001:c.272+40_272+41delTG) found in 3 MJD patients is located in an intronic region of *JOSD2*, which may be functionally relevant if affecting e.g. RNA splicing, but no evidence allow us to speculate on this.

Lastly, the variant COSM3713736 (JOSD1-001:c.246G>A) leads to a base alteration but with no amino acid modification on position 82 of JOSD1 protein.

5. Discussion

This study was performed in 100 Portuguese MJD patients with no suspected recent genetic relatedness. Taking into account that only 2 mutational origins have been described in the Portuguese families (TTACAC and GTGGCA haplotypes, named Joseph and Machado lineages, respectively), virtually all MJD patients from each lineage should share a common ancestor. We have selected patients from different families, bearing in mind that a hard work has previously been done by our team in order to link patients and relatives and reconstruct genealogies as complete as possible based on patients' reports. This way, we attempted to study patients who do not share a recent common ancestor in order to focus on the impact of AO genetic modifiers and remove the bias effect that could have been introduced by other familial factors segregated intergenerationally.

In our cohort, the *ATXN3* repeat was highly polymorphic, both in normal (14-33 repeats) and expanded (63-81 repeats) MJD size ranges; repeat length in the expanded allele was normally distributed, whereas normal alleles seems to present a bimodal distribution, with 14 and 23 CAGs as the most frequent as described in the literature, particularly for the European population [115]. Unlike other repeat expansion disorders, no overlap was found between the ranges of normal and MJD repeat sizes, which facilitated the MJD molecular diagnosis in these patients. With a mean average of approximately 42 years, AO was also found to be vastly variable in our cohort, ranging from 13 to 70 years, which is in accordance with the values referred in the literature [68, 69]. By observing all these parameters, we may state that our cohort seems to be representative of a heterogeneous population of MJD patients.

The instability of the expanded CAG repeat provides a molecular basis to explain AO great variability in MJD. The significant negative correlation found in our cohort between the expanded repeat size in the MJD gene and AO (model A) supports the influence of the CAG tract length as the main contributor in the determination of disease presentation. In our study, the repeat size accounts for about 55% of the variation in disease onset (r^2 = 0.547); in other studies, this percentage ranged from 45% to 76% [116], always suggesting that other intervening factors, either genetic or environmental, should have a major influence in the time of appearance of first symptoms in MJD. Some of the genetic factors that justify the phenotypic variance have already been described; still, the relevance of searching for additional genetic influences is supported by previous findings revealing that 45.5% of AO variance in MJD should be explained by familial factors [116, 117].

The exact contribution of the expanded repeat size to the AO is, however, imprecise since several parameters that affect this correlation are difficult to control accurately, such as (1) MJD AO since information given by the patient or their relatives may be inaccurate (for instance, increased awareness of MJD in the family alert for the first symptoms, which usually results in earlier onsets); (2) the exact CAG repeat length since molecular diagnosis relies often on PCR-based techniques, which introduce stutter bands due to slippage of the Taq polymerase; and (3) somatic mosaicism, typical of expanded alleles since the studied cells and tissues do not necessarily carry the same expansion sizes as the ones involved in disease presentation and progression.

The influence of *ATXN3* normal alleles on MJD phenotype has been a controversial topic in the literature. Igarashi *et al.* were the first to report an inter-allelic interaction between normal and expanded *ATXN3* alleles (involving the $C^{987}GG/G^{987}GG$ polymorphism at the 3' end of the CAG repeat) that affected the intergenerational stability in MJD patients [118]. The direct effect of this SNP on disease phenotype after controlling for the expansion size has never been studied even if this could explain the reason why the effect of normal *ATXN3* alleles is observed in some populations but not replicated by other authors while studying MJD families from different ethnic backgrounds. Interestingly, if we consider that $C^{987}GG/G^{987}GG$ alleles (or others in linkage disequilibrium with them) in normal chromosomes may affect *ATXN3* expression, it would be easily to understand an effect of the $C^{987}GG/G^{987}GG$ genotype on disease presentation based on the importance of the normal ATXN3 loss-of-function to cell toxicity in MJD. Therefore, this finding raised awareness of the possible relationship between onset age of MJD and the number of CAG repeats in the normal chromosome. Several studies were not able to find any association [71, 119, 120], but others – namely Durr *et al.* and França *et al.* in French and Brazilian patients, respectively – reported a small but significant modulating influence of the normal CAG repeat on AO [52, 93]. The authors explained these contradictory results by the differences in the cohort size since a relatively small sample may produce underpowered outcomes. On the other hand, taking into account that $C^{987}GG/G^{987}GG$ allele frequencies vary between populations, if this SNP modulates MJD AO, such reduced effects may be lost when a low frequency of modifier alleles is observed. In our cohort, the analysis of 100 MJD patients from the Portuguese population failed to show a modulating effect of the CAG repeat length in normal *ATXN3* alleles on MJD AO.

From the 16 variants found in our genes of interest, we identified a significant association between AO of MJD patients and the genotype of 5 different SNPs; for 3 of them (rs16999010, rs550589305 and rs141848929), the effect has been noted for each SNP individually (though the last 2 are 100% correlated in our cohort), while COSM3713736 and rs796424878 were only found to significantly affect AO when analysed together. The addition of these 5 polymorphisms to the previous association - model B - proved to have a statistically additive effect of about 8% on the AO variation of our cohort (part $r^2 = 0.084$), with different contributions from each variant. Relevantly, rs796424878 deletion was found to have the most influence in the model improvement (3.39%).

Model B has, however, a series of limitations regarding its correct employment and posterior extrapolation into MJD populations. (1) The observed minor allele frequencies of these 5 polymorphisms were considerably low (MAFs < 0.02), which decreases the statistical power of the discovery (other studies on genetic modifiers of AO in MJD have shown SNPs with higher MAFs [98, 101]). (2) As a consequence of a limited number of patients carrying the rare SNP allele, genotype distributions cannot be assorted and no similar standard deviations can be found when comparing different homozygous; this way, homoscedasticity assumption is violated and we may obtain false positive results. (3) Since the amelioration from model A to model B is based on the influence of SNPs found in only 6 individuals from our cohort of 100 MJD patients, a larger set of subjects still need to be tested to assure the fidelity of the model given any MJD Portuguese patient.

The impact of model B was evaluated by comparing the differences of means associated to the presence or absence of the rare SNP allele and by comparing the AO observed to the ones predicted by both models. Patient #89, who carries the C allele for rs16999010, presented an onset 22 years earlier than the average found for patients homozygous for the G allele; model A predicted a similar onset to the most common genotype, while model B was able to correct the AO of this individual. The pair of alleles rs550589305*C and rs141848929*A, genotyped in 2 subjects (#31 and #62), was found to retard AO by 10 years; model A predicted 14/15 years of an earlier onset, while model B successfully improved the forecast. A mean postponement of 13 years was observed for the patients who carried the rs796424878 deletion; these 3 individuals had a different value for the interception $AO/(CAG)_{exp}$ when compared to model A tendency line since 2 of them were well above the line and the other was barely below. Patients #45 and #57 are predicted by model A to have an earlier AO by 13 and 18 years, respectively, while model B prediction approximates to the observed

value with an error of 2 years. Patient #39, who has an AO inferior to model A regression by 5 years, also carries the COSM3713736*A allele, which presence was associated to a 4 years delay on AO. The balance of the opposite impact of these SNPs in model B (the positive effect of the rs796424878 deletion and the negative influence of COSM3713736) allowed us to revise the projected AO for patient #39 to a more accurate value, closer to the observed AO. All these observations confirm the better capacity of model B to accurately forecast the AO in patients carrying the rare allele in these SNPs.

In our study, the most promising polymorphisms, more likely to explain part of the AO variation, are SNPs that may alter the protein sequence or expression. Variant *ATXN3L*-001:c.796G>A,p.Leu266Phe affects an amino acid closely aligned to the second UIM in ATXN3; this sequence motif contains the residue Ser256 that predominantly mediates the binding of polyubiquitinated substrates [25]. Interestingly, both original and modified residues (coded by the ancestral and derived alleles of this SNP, respectively) are usually involved in substrates recognition [121]. Also, in our model B of linear regression, this polymorphism is associated to an aggravated effect on AO, which increases the importance of this position while considering the hypothesis of ATXN3L compensation for the partial loss-of-function of the normal ATXN3 in MJD patients. Computational methods anticipate slightly different outcomes: PolyPhen-2 predicts the effect of this amino acid substitution to be benign, whereas SIFT score is close to the deleterious border; this difference may be due to the fact that SIFT uses an algorithm mainly based on the evolutionary conservation of amino acids within protein families whereas PolyPhen-2 takes into account information provided by the protein structure [122]. This evidence justifies the relevance of performing functional studies for this polymorphism.

The presence of the 2 variants in regulatory regions (JOSD2-001:c.-244G>C and JOSD2-001:c.-17-6G>A) may modify the binding capacity of several transcription factors to JOSD2, thus affecting JOSD2 levels of expression. Considering the association of both SNPs to a later disease onset, they may eventually improve the binding of activators or bloke the binding of repressors in order to increase JOSD2 transcription. Under this scenario, a higher expression of JOSD2 protein (if able to exert a similar function to normal ATXN3) would compensate the decrease of normal ATXN3 protein levels in MJD patients and, therefore, improve their phenotype.

6. Concluding Remarks and Future Perspectives

Proposed aims for this work have been mostly achieved, yet a few improvements can be performed to reinforce the obtained results and validate our findings. The main conclusion taken from this study is that genetic variants from *JOSD1*, *JOSD2* and *ATXN3L* may have significant influence in the clinical presentation of MJD. In order to fully validate this hypothesis, some adjustments and additional assessments should be performed.

The broadest approach would be the replication of this work in a larger sample of MJD patients, with the inclusion of individuals from different populations, which would allow the study of additional haplotype background patterns. This would likely increase the number and frequency of SNPs found in our candidate genes and creates a fully representative cohort of worldwide MJD patients. The strategy to perform this study could also be optimized by the design of a SNaPShot multiplex that would allow the genotyping of all 5 relevant SNPs in a single reaction, therefore reducing time, effort, sample requirements and well-to-well variability.

Recently, Tezenas du Montcel *et al.* identified a quadratic effect of the expanded CAG repeat size on the onset age of MJD patients, reflecting the possible existence of a nonlinear relationship of the expanded CAG length in both extremes of AO [96]. Therefore, an interesting project would involve a large number of MJD patients from those extremes (for example, onsets earlier than 20 years and later than 60 years) to evaluate if the presence of minor alleles of candidate would help to justify a very diverse MJD clinical presentations in patients with extreme CAG repeat sizes.

Further functional studies can also be made to evaluate the functional relevance of analysed SNPs on transcription regulation, splicing or translation. It would be interesting to test the effect of the residue substitution in ATXN3L caused by rs16999010 variant by measuring the enzymatic activity of ATXN3L *in vitro* in the presence of either allele rs16999010*G or rs16999010*A. On the other hand, the impact of regulatory variants rs550589305 and rs141848929 in the transcriptional activity of *JOSD2* could be studied by luciferase assays. Finally, for both rs141848929 and rs796424878, that may affect the proper RNA splicing of *JOSD2*, we could analyse alternative transcripts present in patients carrying the SNP alleles under study by reverse transcription PCRs performed after extracting RNA from blood samples of these patients.

7. References

1. Reyes-Turcu, F.E., K.H. Ventii, and K.D. Wilkinson, *Regulation and cellular roles of ubiquitin-specific deubiquitinating enzymes.* Annu Rev Biochem, 2009. **78**: p. 363-97.
2. Hershko, A. and A. Ciechanover, *The ubiquitin system.* Annu Rev Biochem, 1998. **67**: p. 425-79.
3. Callis, J., *The ubiquitination machinery of the ubiquitin system.* Arabidopsis Book, 2014. **12**: p. e0174.
4. Fang, S. and A.M. Weissman, *A field guide to ubiquitylation.* Cell Mol Life Sci, 2004. **61**(13): p. 1546-61.
5. Haas, A.L. and P.M. Bright, *The dynamics of ubiquitin pools within cultured human lung fibroblasts.* J Biol Chem, 1987. **262**(1): p. 345-51.
6. Amerik, A.Y. and M. Hochstrasser, *Mechanism and function of deubiquitinating enzymes.* Biochim Biophys Acta, 2004. **1695**(1-3): p. 189-207.
7. Eletr, Z.M. and K.D. Wilkinson, *Regulation of proteolysis by human deubiquitinating enzymes.* Biochim Biophys Acta, 2014. **1843**(1): p. 114-28.
8. Bhogaraju, S. and I. Dikic, *Cell biology: Ubiquitination without E1 and E2 enzymes.* Nature, 2016. **533**(7601): p. 43-4.
9. Scheel, H., S. Tomiuk, and K. Hofmann, *Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics.* Hum Mol Genet, 2003. **12**(21): p. 2845-52.
10. Burnett, B., F. Li, and R.N. Pittman, *The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity.* Hum Mol Genet, 2003. **12**(23): p. 3195-205.
11. Seki, T., et al., *JosD1, a membrane-targeted deubiquitinating enzyme, is activated by ubiquitination and regulates membrane dynamics, cell motility, and endocytosis.* J Biol Chem, 2013. **288**(24): p. 17145-55.
12. Tzvetkov, N. and P. Breuer, *Josephin domain-containing proteins from a variety of species are active de-ubiquitination enzymes.* Biol Chem, 2007. **388**(9): p. 973-8.
13. Weeks, S.D., et al., *Crystal structure of a Josephin-ubiquitin complex: evolutionary restraints on ataxin-3 deubiquitinating activity.* J Biol Chem, 2011. **286**(6): p. 4555-65.
14. Nicastro, G., et al., *Josephin domain of ataxin-3 contains two distinct ubiquitin-binding sites.* Biopolymers, 2009. **91**(12): p. 1203-14.
15. Takiyama, Y., et al., *The gene for Machado-Joseph disease maps to human chromosome 14q.* Nat Genet, 1993. **4**(3): p. 300-4.
16. Ichikawa, Y., et al., *The genomic structure and expression of MJD, the Machado-Joseph disease gene.* J Hum Genet, 2001. **46**(7): p. 413-22.
17. Harris, G.M., et al., *Splice isoforms of the polyglutamine disease protein ataxin-3 exhibit similar enzymatic yet different aggregation properties.* PLoS One, 2010. **5**(10): p. e13695.

18. Masino, L., et al., *Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail.* FEBS Lett, 2003. **549**(1-3): p. 21-5.
19. Nicastro, G., et al., *Understanding the role of the Josephin domain in the PolyUb binding and cleavage properties of ataxin-3.* PLoS One, 2010. **5**(8): p. e12430.
20. Boeddrich, A., et al., *An arginine/lysine-rich motif is crucial for VCP/p97-mediated modulation of ataxin-3 fibrillogenesis.* EMBO J, 2006. **25**(7): p. 1547-58.
21. Nóbrega, C. and L. Pereira de Almeida, *Machado-Joseph Disease / Spinocerebellar Ataxia Type 3*, in *Spinocerebellar Ataxia*, J. Gazulla, Editor. 2012.
22. Macedo-Ribeiro, S., et al., *Nucleocytoplasmic shuttling activity of ataxin-3.* PLoS One, 2009. **4**(6): p. e5834.
23. Pozzi, C., et al., *Study of subcellular localization and proteolysis of ataxin-3.* Neurobiol Dis, 2008. **30**(2): p. 190-200.
24. Trottier, Y., et al., *Heterogeneous intracellular localization and expression of ataxin-3.* Neurobiol Dis, 1998. **5**(5): p. 335-47.
25. Mueller, T., et al., *CK2-dependent phosphorylation determines cellular localization and stability of ataxin-3.* Hum Mol Genet, 2009. **18**(17): p. 3334-43.
26. Goto, J., et al., *Machado-Joseph disease gene products carrying different carboxyl termini.* Neurosci Res, 1997. **28**(4): p. 373-7.
27. Bettencourt, C., et al., *Increased transcript diversity: novel splicing variants of Machado-Joseph disease gene (ATXN3).* Neurogenetics, 2010. **11**(2): p. 193-202.
28. Doss-Pepe, E.W., et al., *Ataxin-3 interactions with rad23 and valosin-containing protein and its associations with ubiquitin chains and the proteasome are consistent with a role in ubiquitin-mediated proteolysis.* Mol Cell Biol, 2003. **23**(18): p. 6469-83.
29. Berke, S.J., et al., *Defining the role of ubiquitin-interacting motifs in the polyglutamine disease protein, ataxin-3.* J Biol Chem, 2005. **280**(36): p. 32026-34.
30. Schmitt, I., et al., *Inactivation of the mouse Atxn3 (ataxin-3) gene increases protein ubiquitination.* Biochem Biophys Res Commun, 2007. **362**(3): p. 734-9.
31. Winborn, B.J., et al., *The deubiquitinating enzyme ataxin-3, a polyglutamine disease protein, edits Lys63 linkages in mixed linkage ubiquitin chains.* J Biol Chem, 2008. **283**(39): p. 26436-43.
32. Wada, K. and T. Kamitani, *UnpEL/Usp4 is ubiquitinated by Ro52 and deubiquitinated by itself.* Biochem Biophys Res Commun, 2006. **342**(1): p. 253-8.
33. Durcan, T.M., et al., *The Machado-Joseph disease-associated mutant form of ataxin-3 regulates parkin ubiquitination and stability.* Hum Mol Genet, 2011. **20**(1): p. 141-54.
34. Nijman, S.M., et al., *A genomic and functional inventory of deubiquitinating enzymes.* Cell, 2005. **123**(5): p. 773-86.
35. Nicastro, G., et al., *The solution structure of the Josephin domain of ataxin-3: structural determinants for molecular recognition.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10493-8.

36. Todi, S.V., et al., *Ubiquitination directly enhances activity of the deubiquitinating enzyme ataxin-3.* EMBO J, 2009. **28**(4): p. 372-82.
37. Todi, S.V., et al., *Activity and cellular functions of the deubiquitinating enzyme and polyglutamine disease protein ataxin-3 are regulated by ubiquitination at lysine 117.* J Biol Chem, 2010. **285**(50): p. 39303-13.
38. Zhong, X. and R.N. Pittman, *Ataxin-3 binds VCP/p97 and regulates retrotranslocation of ERAD substrates.* Hum Mol Genet, 2006. **15**(16): p. 2409-20.
39. Wang, G., et al., *Ataxin-3, the MJD1 gene product, interacts with the two human homologs of yeast DNA repair protein RAD23, HHR23A and HHR23B.* Hum Mol Genet, 2000. **9**(12): p. 1795-803.
40. Ferro, A., et al., *NEDD8: a new ataxin-3 interactor.* Biochim Biophys Acta, 2007. **1773**(11): p. 1619-27.
41. Markossian, K.A. and B.I. Kurganov, *Protein folding, misfolding, and aggregation. Formation of inclusion bodies and aggresomes.* Biochemistry (Mosc), 2004. **69**(9): p. 971-84.
42. Burnett, B.G. and R.N. Pittman, *The polyglutamine neurodegenerative protein ataxin 3 regulates aggresome formation.* Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4330-5.
43. Sacco, J.J., et al., *The deubiquitylase Ataxin-3 restricts PTEN transcription in lung cancer cells.* Oncogene, 2014. **33**(33): p. 4265-72.
44. Rodrigues, A.J., et al., *Absence of ataxin-3 leads to cytoskeletal disorganization and increased cell death.* Biochim Biophys Acta, 2010. **1803**(10): p. 1154-63.
45. Li, F., et al., *Ataxin-3 is a histone-binding protein with two independent transcriptional corepressor activities.* J Biol Chem, 2002. **277**(47): p. 45004-12.
46. Neves-Carvalho, A., et al., *Dominant negative effect of polyglutamine expansion perturbs normal function of ataxin-3 in neuronal cells.* Hum Mol Genet, 2015. **24**(1): p. 100-17.
47. do Carmo Costa, M., et al., *Ataxin-3 plays a role in mouse myogenic differentiation through regulation of integrin subunit levels.* PLoS One, 2010. **5**(7): p. e11728.
48. Ge, F., et al., *Ataxin-3 like (ATXN3L), a member of the Josephin family of deubiquitinating enzymes, promotes breast cancer proliferation by deubiquitinating Kruppel-like factor 5 (KLF5).* Oncotarget, 2015. **6**(25): p. 21369-78.
49. Buus, R., et al., *Deubiquitinase activities required for hepatocyte growth factor-induced scattering of epithelial cells.* Curr Biol, 2009. **19**(17): p. 1463-6.
50. Gomez-Ramos, A., et al., *Distinct X-chromosome SNVs from some sporadic AD samples.* Sci Rep, 2015. **5**: p. 18012.
51. Paulson, H., *Machado-Joseph disease/spinocerebellar ataxia type 3.* Handb Clin Neurol, 2012. **103**: p. 437-49.
52. Durr, A., et al., *Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular, and neuropathological features.* Ann Neurol, 1996. **39**(4): p. 490-9.

53. Riess, O., et al., *SCA3: neurological features, pathogenesis and animal models.* Cerebellum, 2008. **7**(2): p. 125-37.
54. Rub, U., et al., *Spinocerebellar ataxia type 3 (Machado-Joseph disease): severe destruction of the lateral reticular nucleus.* Brain, 2002. **125**(Pt 9): p. 2115-24.
55. Carvalho, D.R., et al., *Homozygosity enhances severity in spinocerebellar ataxia type 3.* Pediatr Neurol, 2008. **38**(4): p. 296-9.
56. Zhang, B., et al., *Clinical manifestations and gene mutation in a case of Machado-Joseph disease.* Neural Regen Res, 2012. **7**(35): p. 2842-7.
57. Franca, M.C., Jr., et al., *Muscle excitability abnormalities in Machado-Joseph disease.* Arch Neurol, 2008. **65**(4): p. 525-9.
58. van de Warrenburg, B.P., et al., *Spinocerebellar ataxias in the Netherlands: prevalence and age at onset variance analysis.* Neurology, 2002. **58**(5): p. 702-8.
59. Verbeek, D.S., *Spinocerebellar ataxia type 23: a genetic update.* Cerebellum, 2009. **8**(2): p. 104-7.
60. Jardim, L.B., et al., *A survey of spinocerebellar ataxia in South Brazil - 66 new cases with Machado-Joseph disease, SCA7, SCA8, or unidentified disease-causing mutations.* J Neurol, 2001. **248**(10): p. 870-6.
61. Vale, J., et al., *Autosomal dominant cerebellar ataxia: frequency analysis and clinical characterization of 45 families from Portugal.* Eur J Neurol, 2010. **17**(1): p. 124-8.
62. Maruyama, H., et al., *Difference in disease-free survival curve and regional distribution according to subtype of spinocerebellar ataxia: a study of 1,286 Japanese patients.* Am J Med Genet, 2002. **114**(5): p. 578-83.
63. Zhao, Y., et al., *Prevalence and ethnic differences of autosomal-dominant cerebellar ataxia in Singapore.* Clin Genet, 2002. **62**(6): p. 478-81.
64. Tang, B., et al., *Frequency of SCA1, SCA2, SCA3/MJD, SCA6, SCA7, and DRPLA CAG trinucleotide repeat expansion in patients with hereditary spinocerebellar ataxia from Chinese kindreds.* Arch Neurol, 2000. **57**(4): p. 540-4.
65. Schols, L., et al., *Autosomal dominant cerebellar ataxia: phenotypic differences in genetically defined subtypes?* Ann Neurol, 1997. **42**(6): p. 924-32.
66. Sequeiros, J. and P. Coutinho, *Epidemiology and clinical aspects of Machado-Joseph disease.* Adv Neurol, 1993. **61**: p. 139-53.
67. Martins, S., et al., *Asian origin for the worldwide-spread mutational event in Machado-Joseph disease.* Arch Neurol, 2007. **64**(10): p. 1502-8.
68. Maruyama, H., et al., *Molecular features of the CAG repeats and clinical manifestation of Machado-Joseph disease.* Hum Mol Genet, 1995. **4**(5): p. 807-12.
69. Donis, K.C., et al., *Spinocerebellar ataxia type 3/Machado-Joseph disease starting before adolescence.* Neurogenetics, 2016. **17**(2): p. 107-13.
70. Paulson, H.L., *Dominantly inherited ataxias: lessons learned from Machado-Joseph disease/spinocerebellar ataxia type 3.* Semin Neurol, 2007. **27**(2): p. 133-42.

71. Maciel, P., et al., *Correlation between CAG repeat length and clinical features in Machado-Joseph disease.* Am J Hum Genet, 1995. **57**(1): p. 54-61.
72. Coutinho, P. and J. Sequeiros, *[Clinical, genetic and pathological aspects of Machado-Joseph disease].* J Genet Hum, 1981. **29**(3): p. 203-9.
73. Pulst, S.-M., *Genetics of movement disorders*. 2003, San Diego, Calif.: Academic Press. 561 p.
74. DiFiglia, M., et al., *Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain.* Science, 1997. **277**(5334): p. 1990-3.
75. Schmidt, T., et al., *Protein surveillance machinery in brains with spinocerebellar ataxia type 3: redistribution and differential recruitment of 26S proteasome subunits and chaperones to neuronal intranuclear inclusions.* Ann Neurol, 2002. **51**(3): p. 302-10.
76. Chai, Y., et al., *Live-cell imaging reveals divergent intracellular dynamics of polyglutamine disease proteins and supports a sequestration model of pathogenesis.* Proc Natl Acad Sci U S A, 2002. **99**(14): p. 9310-5.
77. Chai, Y., et al., *Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions in SCA3/MJD and suppression of polyglutamine aggregation in vitro.* Hum Mol Genet, 1999. **8**(4): p. 673-82.
78. Williams, A.J. and H.L. Paulson, *Polyglutamine neurodegeneration: protein misfolding revisited.* Trends Neurosci, 2008. **31**(10): p. 521-8.
79. Bevivino, A.E. and P.J. Loll, *An expanded glutamine repeat destabilizes native ataxin-3 structure and mediates formation of parallel beta -fibrils.* Proc Natl Acad Sci U S A, 2001. **98**(21): p. 11955-60.
80. Nagai, Y., et al., *A toxic monomeric conformer of the polyglutamine protein.* Nat Struct Mol Biol, 2007. **14**(4): p. 332-40.
81. Takahashi, J., et al., *Recruitment of nonexpanded polyglutamine proteins to intranuclear aggregates in neuronal intranuclear hyaline inclusion disease.* J Neuropathol Exp Neurol, 2001. **60**(4): p. 369-76.
82. Demuro, A., et al., *Calcium dysregulation and membrane disruption as a ubiquitous neurotoxic mechanism of soluble amyloid oligomers.* J Biol Chem, 2005. **280**(17): p. 17294-300.
83. Streets, A.M., et al., *Simultaneous measurement of amyloid fibril formation by dynamic light scattering and fluorescence reveals complex aggregation kinetics.* PLoS One, 2013. **8**(1): p. e54541.
84. Teixeira-Castro, A., et al., *Neuron-specific proteotoxicity of mutant ataxin-3 in C. elegans: rescue by the DAF-16 and HSF-1 pathways.* Hum Mol Genet, 2011. **20**(15): p. 2996-3009.
85. Matos, C.A., S. de Macedo-Ribeiro, and A.L. Carvalho, *Polyglutamine diseases: the special case of ataxin-3 and Machado-Joseph disease.* Prog Neurobiol, 2011. **95**(1): p. 26-48.

86. Goti, D., et al., *A mutant ataxin-3 putative-cleavage fragment in brains of Machado-Joseph disease patients and transgenic mice is cytotoxic above a critical concentration.* J Neurosci, 2004. **24**(45): p. 10266-79.
87. Haacke, A., F.U. Hartl, and P. Breuer, *Calpain inhibition is sufficient to suppress aggregation of polyglutamine-expanded ataxin-3.* J Biol Chem, 2007. **282**(26): p. 18851-6.
88. Berke, S.J., et al., *Caspase-mediated proteolysis of the polyglutamine disease protein ataxin-3.* J Neurochem, 2004. **89**(4): p. 908-18.
89. Breuer, P., et al., *Nuclear aggregation of polyglutamine-expanded ataxin-3: fragments escape the cytoplasmic quality control.* J Biol Chem, 2010. **285**(9): p. 6532-7.
90. Haacke, A., et al., *Proteolytic cleavage of polyglutamine-expanded ataxin-3 is critical for aggregation and sequestration of non-expanded ataxin-3.* Hum Mol Genet, 2006. **15**(4): p. 555-68.
91. Bichelmeier, U., et al., *Nuclear localization of ataxin-3 is required for the manifestation of symptoms in SCA3: in vivo evidence.* J Neurosci, 2007. **27**(28): p. 7418-28.
92. Evert, B.O., et al., *Ataxin-3 represses transcription via chromatin binding, interaction with histone deacetylase 3, and histone deacetylation.* J Neurosci, 2006. **26**(44): p. 11474-86.
93. Franca, M.C., Jr., et al., *Normal ATXN3 Allele but Not CHIP Polymorphisms Modulates Age at Onset in Machado-Joseph Disease.* Front Neurol, 2012. **3**: p. 164.
94. Bettencourt, C., et al., *The APOE epsilon2 allele increases the risk of earlier age at onset in Machado-Joseph disease.* Arch Neurol, 2011. **68**(12): p. 1580-3.
95. Peng, H., et al., *APOE epsilon2 allele may decrease the age at onset in patients with spinocerebellar ataxia type 3 or Machado-Joseph disease from the Chinese Han population.* Neurobiol Aging, 2014. **35**(9): p. 2179 e15-8.
96. Tezenas du Montcel, S., et al., *Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes.* Brain, 2014. **137**(Pt 9): p. 2444-55.
97. Chen, S., et al., *Mitochondrial NADH Dehydrogenase Subunit 3 Polymorphism Associated with an Earlier Age at Onset in Male Machado-Joseph disease Patients.* CNS Neurosci Ther, 2016. **22**(1): p. 38-42.
98. Long, Z., et al., *Two novel SNPs in ATXN3 3' UTR may decrease age at onset of SCA3/MJD in Chinese patients.* PLoS One, 2015. **10**(2): p. e0117488.
99. Emmel, V.E., et al., *Does DNA methylation in the promoter region of the ATXN3 gene modify age at onset in MJD (SCA3) patients?* Clin Genet, 2011. **79**(1): p. 100-2.
100. Emmel, V.E., et al., *GRIK2, IL1B, NEDD8 and NEDD9 genes may act as modifiers of the Machado Joseph Disease/SCA 3 phenotype.* The 12th International Congress of Human Genetics and the American Society of Human Genetics 61st Annual Meeting - POSTER ABSTRACTS, 2011: p. 47.

101. Raposo, M., et al., *Promoter Variation and Expression Levels of Inflammatory Genes IL1A, IL1B, IL6 and TNF in Blood of Spinocerebellar Ataxia Type 3 (SCA3) Patients.* Neuromolecular Med, 2016.
102. Siebert, M., et al., *Glucocerebrosidase gene variants in parkinsonian patients with Machado Joseph/spinocerebellar ataxia 3.* Parkinsonism Relat Disord, 2012. **18**(2): p. 185-90.
103. Martins, S., *Evolutionary and epidemiological genetics of Machado-Joseph disease*. 2007, University of Porto.
104. Paulson, H.L., et al., *Intranuclear inclusions of expanded polyglutamine protein in spinocerebellar ataxia type 3.* Neuron, 1997. **19**(2): p. 333-44.
105. Slepko, N., et al., *Normal-repeat-length polyglutamine peptides accelerate aggregation nucleation and cytotoxicity of expanded polyglutamine proteins.* Proc Natl Acad Sci U S A, 2006. **103**(39): p. 14367-72.
106. Warrick, J.M., et al., *Ataxin-3 suppresses polyglutamine neurodegeneration in Drosophila by a ubiquitin-associated mechanism.* Mol Cell, 2005. **18**(1): p. 37-48.
107. Martins, M.I., *Evolution and functional relevance of ataxin-3 paralogues*, in *Secção Autónoma de Ciências da Saúde*. 2012, University of Aveiro.
108. *JOSD1 Josephin domain containing 1 [Homo sapiens (human)] - Gene - NCBI*. [cited 2015; Available from: http://www.ncbi.nlm.nih.gov/gene/9929.
109. *JOSD2 Josephin domain containing 2 [Homo sapiens (human)] - Gene - NCBI*. [cited 2015; Available from: http://www.ncbi.nlm.nih.gov/gene/126119.
110. *Ensembl genome browser 80: Homo sapiens - Exons - Transcript: JOSD1-001 (ENST00000216039)*. [cited 2015; Available from: http://www.ensembl.org/Homo_sapiens/Transcript/Exons?db=core;g=ENSG00000161677;r=19:50505998-50511353;t=ENST00000598418.
111. *Ensembl genome browser 80: Homo sapiens - Exons - Transcript: JOSD2-001 (ENST00000598418).*
112. Kumar, A. and J. Kaur, *Primer Based Approach for PCR Amplification of High GC Content Gene: Mycobacterium Gene as a Model.* Mol Biol Int, 2014. **2014**: p. 937308.
113. *PolyPhen-2 - Prediction of functional effects of human nsSNPs*. [cited 2016; Available from: http://genetics.bwh.harvard.edu/pph2/.
114. *SIFT Human Protein*. [cited 2016; Available from: http://sift.jcvi.org/www/SIFT_enst_submit.html.
115. Limprasert, P., 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.* Hum Mol Genet, 1996. **5**(2): p. 207-13.
116. van de Warrenburg, B.P., et al., *Age at onset variance analysis in spinocerebellar ataxias: a study in a Dutch-French cohort.* Ann Neurol, 2005. **57**(4): p. 505-12.
117. DeStefano, A.L., et al., *A familial factor independent of CAG repeat length influences age at onset of Machado-Joseph disease.* Am J Hum Genet, 1996. **59**(1): p. 119-27.

118. Igarashi, S., et al., *Intergenerational instability of the CAG repeat of the gene for Machado-Joseph disease (MJD1) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat.* Hum Mol Genet, 1996. **5**(7): p. 923-32.
119. Ranum, L.P., et al., *Spinocerebellar ataxia type 1 and Machado-Joseph disease: incidence of CAG expansions among adult-onset ataxia patients from 311 families with dominant, recessive, or sporadic ataxia.* Am J Hum Genet, 1995. **57**(3): p. 603-8.
120. Zaltzman, R., et al., *Spinocerebellar ataxia type 3 in Israel: phenotype and genotype of a Jew Yemenite subpopulation.* J Neurol, 2016.
121. Betts, M.J. and R.B. Russell, *Amino Acid Properties and Consequences of Substitutions*, in *Bioinformatics for Geneticists*, M.R. Barnes and I.C. Gray, Editors. 2003, John Wiley & Sons, Ltd. p. 289-316.
122. Wei, P., X. Liu, and Y.X. Fu, *Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study.* BMC Proc, 2011. **5 Suppl 9**: p. S20.

8. Supplementary material

Supplementary Table A – Clinical information and genotyping data of *ATXN3* repeat and SNPs found in *JOSD1*, *JOSD2* and *ATXN3L* for our cohort of 100 MJD patients. Tinted background refers to genotypes with the presence of the minor allele. E – exon; I – intron.

Sample ID	Gender	$(CAG)_n$		Age-at-onset (years)	Polymorphisms																
					JOSD1			*JOSD2*									*ATXN3L*				
					E1		E2	E1	I1	E3	I3	E4	3' UTR			CDS			3' UTR	
		Expanded allele	Normal allele		rs2072800	rs6001200	COSM3713736	rs550589305	rs141848929	rs376400138	rs796424878	rs11548260	rs376569327	rs117711212	rs3087582	rs16999010	rs17322143	rs4830842	rs58878737	rs41297291
1	M	70	23	45	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
2	M	77	25	37	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T	G	C	C	T	T
3	M	77	21	21	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
4	M	71	15	48	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	T	T	T	T
5	M	66	28	41	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
6	F	67	25	60	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
7	M	73	23	33	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
8	F	79	23	18	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
9	M	75	21	35	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
10	M	71	15	39	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T	G	C	T	T	T
11	M	74	24	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
12	M	68	23	55	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	T/T	G	C	T	T	T
13	F	74	28	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
14	F	71	23	60	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
15	F	69	27	43	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/T	T/T	T/T	T/T
16	M	75	23	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
17	F	77	24	25	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/T	T/T	T/T	T/T
18	M	68	14	49	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
19	F	65	22	48	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
20	M	71	24	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
21	M	73	24	26	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
22	F	70	24	51	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
23	M	75	15	31	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
24	F	70	23	50	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/T	T/T	T/T
25	M	69	23	48	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
26	F	73	28	50	A/A	G/C	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/T	C/T	T/T	T/T
27	M	73	26	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
28	M	76	14	24	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
29	M	73	28	30	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
30	M	77	19	20	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
31	M	72	14	55	A/A	G/G	G/G	G/C	G/A	G/G	TG	G/A	A/A	C/C	C/C	G	C	T	T	C
32	M	79	28	16	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/T	C/C	G	C	T	T	T
33	F	69	23	35	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/T	T/G	T/T
34	M	69	24	44	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
35	F	72	28	37	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
36	F	80	19	13	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
37	F	73	14	36	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	T/T	T/T	T/T	T/T
38	M	69	23	42	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	T	T	T
39	F	68	23	46	A/A	G/G	G/A	G/G	G/G	G/G	-	G/G	A/A	C/C	C/C	G/G	C/C	C/T	T/T	T/T
40	M	69	14	45	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
41	F	80	23	28	A/A	G/G	G/G	G/G	G/G	G/A	TG	G/G	A/A	C/C	C/C	G/G	C/T	C/T	T/T	T/T
42	F	77	27	36	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	T/T	T/T	T/T	T/T
43	M	69	24	48	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G	C	C	T	T
44	F	73	30	44	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	T/T	T/T	T/T
45	F	72	23	55	A/A	G/G	G/G	G/G	G/G	G/G	-	G/G	A/A	C/C	C/C	G/G	C/C	C/T	T/T	T/T
46	F	68	23	48	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
47	F	79	21	18	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T	G/G	C/C	T/T	T/T	T/T
48	M	72	23	43	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T	G	C	C	T	T
49	F	67	23	52	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T
50	F	71	23	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C	G/G	C/C	C/C	T/T	T/T

Sample ID	Gender	(CAG)n		Age-at-onset (years)	Polymorphisms																
		JOSD1			*JOSD1*			*JOSD2*									*ATXN3L*				
					E1		E2	E1	I1	E3	I3	E4	3' UTR				CDS			3' UTR	
		Expanded allele	Normal allele		rs2072800	rs6001200	COSM3713736	rs550589305	rs141848929	rs376400138	rs796424878	rs11548260	rs376569327	rs117711212	rs3087582		rs16999010	rs17322143	rs4830842	rs58878737	rs41297291
51	F	71	23	47	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
52	F	69	21	45	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	T/T	T/T	T/T
53	F	72	14	42	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
54	F	69	23	42	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/G	C/C	C/T		G/G	C/C	C/C	T/T	T/T
55	M	76	27	25	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T		G	C	C	T	T
56	M	65	14	60	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
57	M	70	23	65	A/A	G/G	G/G	G/G	G/G	G/G	-	G/G	A/A	C/C	C/C		G	C	C	T	T
58	F	68	27	60	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/T	T/T	T/T
59	M	73	28	49	A/G	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
60	M	72	21	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	T	T	T
61	F	69	28	55	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/T	T/T	T/T
62	F	75	32	49	A/G	G/G	G/G	G/C	G/A	G/G	TG	G/A	A/A	C/C	C/C		G/G	C/T	C/T	T/T	T/T
63	F	76	21	35	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T		G/G	C/C	C/T	T/T	T/T
64	F	69	33	49	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	T/T	T/T	T/T	T/T
65	M	75	23	45	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	T	T	T
66	M	68	23	47	A/G	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/T	C/T		G	C	C	T	T
67	F	70	26	42	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
68	M	69	14	55	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
69	M	71	21	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
70	F	74	14	37	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
71	F	73	24	40	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
72	F	76	21	31	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
73	F	63	21	54	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/A	A/A	C/C	C/C		G/G	C/C	C/T	T/T	T/T
74	F	69	14	37	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
75	F	70	27	42	A/G	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/T	T/T	T/T	T/T
76	F	81	23	33	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/T	C/T	T/T	T/T
77	F	66	23	62	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/T	C/T	T/T	T/T
78	M	75	24	27	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
79	M	67	21	45	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	T	T	T
80	F	74	23	52	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/T	T/T	T/T
81	M	66	26	41	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
82	F	68	14	50	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/T	T/T	T/T
83	F	75	14	35	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/T	T/T	T/T
84	F	72	27	54	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
85	M	75	27	30	A/G	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
86	F	72	23	37	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T		G/G	T/T	T/T	T/T	T/T
87	M	71	23	55	A/G	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
88	F	68	27	67	A/G	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	T/T		G/G	C/C	C/C	T/T	T/T
89	M	73	21	20	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		A	C	T	T	T
90	M	68	27	46	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
91	M	69	23	60	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/T	C/T		G	C	C	T	T
92	M	69	23	55	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	T	T	T
93	M	66	28	70	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
94	M	75	26	15	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
95	F	72	23	47	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
96	F	69	23	63	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T		G/G	C/C	C/C	T/T	T/T
97	F	77	23	19	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/T		G/G	C/C	C/C	T/T	T/T
98	M	74	20	33	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T
99	F	76	25	30	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G/G	C/C	C/C	T/T	T/T
100	M	69	28	65	A/A	G/G	G/G	G/G	G/G	G/G	TG	G/G	A/A	C/C	C/C		G	C	C	T	T