Unveiling the role of untranslated and methylation prone regions in the clinical variability of Machado-Joseph disease/Spinocerebellar ataxia type 3

Tese de Doutoramento

Ana Rosa Vieira Melo

Doutoramento em

Biologia

Ponta Delgada
2022

Unveiling the role of untranslated and methylation prone regions in the clinical variability of Machado-Joseph disease/Spinocerebellar ataxia type 3

Tese de Doutoramento

Ana Rosa Vieira Melo

Supervisores

Professora Doutora Maria Manuela de Medeiros Lima
Doutora Maria da Conceição Félix Bettencourt

Thesis specially elaborated to obtain the degree of Doctor in Biology
Tese especialmente elaborada para obtenção do grau de Doutor em Biologia

INSTITUTIONS AND FINANCIAL SUPPORT

Funded by:

- Doctoral Grant awarded by Fundação para a Ciência e a Tecnologia (FCT) (SFRH/BD/129547/2017).

Developed at:

- Faculdade de Ciências e Tecnologia, Universidade dos Açores, Portugal.
- Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Portugal.

With the support of:

- FEDER- Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operacional Programme for Competitiveness and Internationalization (POCI) (POCI-01-0145-FEDER-016592), Portugal 2020.
- "EXOS3 - Whole-exome sequencing of discordant and concordant affected sib pairs in spinocerebellar ataxia type 3 (SCA3): a tool to identify novel modifier genes and highlight disrupted molecular pathways (PTDC/DTP-PIC/2638/2017), granted by Fundação para a Ciência e Tecnologia (FCT)

DECLARATION OF AUTHORSHIP

I, Ana Rosa Vieira Melo, hereby declare that this thesis and the work presented in it was developed by me as a student of the Doctoral Program in Biology (3CBIO) at the University of the Azores.

I further declare that data and results of Chapter 2 were partially published over the course of this thesis project. The published paper has the following reference:

Melo ARV, Raposo M, Ventura M, Martins S, Pavão S, Alonso I, *et al.* Genetic Variation in *ATXN3* (Ataxin-3) 3'UTR: Insights into the Downstream Regulatory Elements of the Causative Gene of Machado-Joseph Disease/Spinocerebellar Ataxia Type 3. Cerebellum. 2022. doi: 10.1007/s12311-021-01358-0.

Assinado por: **Ana Rosa Vieira Melo**
Num. de Identificação: 13214349
Data: 2022.07.31 15:50:46+00'00'

CHAVE MÓVEL

Ana Rosa Vieira Melo

Para os meus pais

e para a minha tia Fátima

"What transforms this world is- knowledge. Do you see what I mean? Nothing else can change anything in this world." – Yukio Mishima in "The Temple of the Golden Pavilion"

"For true success ask yourself these four questions: Why? Why not? Why not me? Why not now?" - James Allen

ACKNOWLEDGMENTS

A presente tese permitiu-me crescer imenso enquanto ser humano e como investigadora, tornando-me numa melhor versão do que era quando iniciei este caminho. Este crescimento apenas foi possível graças ao apoio de muitas pessoas a quem gostaria de agradecer.

Em primeiro lugar às minhas orientadoras por serem um exemplo de como fazer ciência, e ao mesmo tempo conjugar a parte pessoal. O meu enorme obrigada pelas imensas discussões e reuniões científicas que me fizeram aprender, e acima de tudo, me fizeram acreditar que era possível continuar com o trabalho, mesmo quando deixei de acreditar que tal era possível. Quero também agradecer pelo vosso apoio incondicional, disponibilidade e imensa paciência quando eu teimava em não vos pedir ajuda, ou fazia perguntas tolas. Obrigada por não terem desistido de mim.

À minha orientadora, Professora Doutora Manuela Lima, quero agradecer, do fundo do coração, por ter acreditado em mim desde o início, reconhecendo em mim capacidades que eu duvidava que tinha e por me ter motivado quando "esmoreci". O meu enorme obrigada por todos os seus conselhos, profissionais e pessoais :) Obrigada por ter aceitado percorrer este longo caminho comigo, ter estado sempre (mesmo sempre) "lá" e por me permitir fazer parte da sua equipa. Este facto permitiu-me aprender mais, e aprofundar os meus conhecimentos e competências nesta área. A minha integração na sua equipa permitiu que eu conhecesse de mais perto a realidade desta doença, o que me enriqueceu enquanto ser humano e me motivou a lutar por "eles".

À minha co-orientadora, Doutora Conceição Bettencourt, o meu enorme obrigada por ter aceitado ser minha orientadora mesmo conhecendo-me mal. Obrigada pela disponibilidade, discussões científicas, paciência nas minhas "lutas" com o R e pelos seus conselhos mais pessoais, naqueles momentos em que me fui abaixo. Obrigada pela sua preocupação, e por ter a palavra certa no momento certo.

À Doutora Patrícia Maciel, o meu enorme obrigada por me ter recebido de braços abertos no ICVS durante o mês em que lá estive, fazendo-me sentir em casa. O meu obrigada também a toda a sua equipa, principalmente à Daniela Campos (de quem fui "sombra" durante uns dias), à Marta, à Bruna, à Andreia Castro, à Sara (prometo da próxima trabalhar com os ratinhos), ao Jorge, à Joana e à Andreia Carvalho.

Às minhas colegas de grupo, Mafalda e Ana. Obrigada por todo o apoio, disponibilidade, conversas científicas, conselhos e pelos bons e intensos momentos que

passamos juntas. Foram momentos únicos e que guardo com muito carinho. Ao Luís Teves, que chegou mais tarde ao grupo, mas que sempre me motivou quando me via mais cansada.

À Sara e à Filipa, que durante algum tempo também foram colegas de grupo, o meu enorme obrigada pelas conversas científicas, desabafos e brincadeiras. Ao Pedro Mântua, pelas longas conversas, conselhos e pela sua amizade, nos corredores e nos laboratórios.

A todos os membros do GAIN (Professora Manuela, Doutora. Teresa Kay, Dr. Carlos González e Dr. João Vasconcelos), à Dra. Paula Pires e ao Dr. João Lemos (Hospital de Santo Espírito da Ilha Terceira) e ao nosso grupo de investigação. As nossas visitas aos participantes do ESMI, nas Flores, Graciosa e Terceira, foram das situações que mais me marcaram. Apesar do desgaste emocional, todos nos unimos com um único objetivo de ajudar os portadores da SCA3. Essas visitas mostraram-me que a união faz mesmo a força. Nada é impossível quando a equipa se une, e essa união permitiu também que, em termos emocionais conseguisse ficar bem.

Em S. Miguel, as visitas continuam a ser marcantes, e por isso quero agradecer ao Dr. João Vasconcelos não só por me permitir partilhar com ele a avaliação clínica dos portadores da SCA3, mas também por me ensinar imensa coisa. Quero também agradecer à Dra. Ana Pedro, ao Dr. Carlos, à Rita e à Márcia, pela simpatia, espírito de entreajuda e sobretudo pela boa disposição com que sempre me brindaram.

A todos os portadores da mutação causal da SCA3, e às suas famílias, pela sua disponibilidade, confiança e por não desistirem, mesmo quando estavam em dias menos bons. Obrigada pelo privilégio que me deram de poder conviver convosco. Guardo todos esses momentos, que fizeram de mim um ser humano melhor e me motivaram a dar o meu melhor para conseguir aumentar o conhecimento sobre esta doença. Aprendi que uma pequena gota no oceano, pode fazer a diferença na vida de alguém.

Às minhas amigas Carla, Carolina, Rita, Andreia, Helga, Ritinha, Mafalda Maria e Isabel por toda a força que me deram, pela paciência, pelo carinho, pelas aventuras vividas e pelas palavras sábias no momento certo. Conhecemo-nos na Universidade já há uns bons anos, mas parece que foi ontem :). Gosto mesmo muito de vocês, miúdas!

À minha amiga Renata, por me aturar à mais de metade da nossa existência. Estamos longe, mas ao mesmo tempo perto. Mesmo longe continuas a perceber quando estou menos bem.

Aos meus amigos Manuel, Pedro, Rui e João, que apesar de longe, estão sempre perto. Obrigada por me "empurrarem" para a frente, mesmo nos momentos mais complicados. Obrigada por aturarem o meu mau humor e "dramas", e por me enviarem vídeos e memes para me fazer rir.

Se houve algo com que a vida e este doutoramento me agraciou foi com a chegada das pessoas certas, no momento certo. Serei sempre grata por ter sido abençoada com amigas extraordinárias. À minha amiga Ana, com a qual partilho o nome, o signo, o mês de nascimento, e também o quarto nas nossas visitas aos doentes. O meu obrigada pela sua amizade, força, conselhos, carinho, piadas e cantorias (acho que ainda vamos gravar um disco juntas). És muito especial, e sei que estás sempre lá para o que for preciso. Obrigada por me puxares as orelhas, e me lembrares de que eu precisava de ser gentil comigo. À minha amiga Elisabete pelo seu carinho, amizade, pelo ombro sempre disponível, pela sua paciência nas minhas crises existenciais e por estar sempre lá nos bons e nos momentos menos bons. À minha amiga Zita pela sua constante preocupação, força, amizade, conselhos e abastecimento de gulodices :). À minha amiga Anita, uma caixinha de surpresas sempre pronta a ajudar, e a dar força e conselhos. Obrigada às quatro pela vossa sincera e pura amizade, e por todos os momentos que passamos juntas.

À minha amiga Mafalda, o meu obrigada pela amizade, apoio, conselhos e estar "sempre lá". Obrigada por toda a força que me deste.

Às minhas amigas Cláudia e Marta, o meu obrigada pela vossa amizade, pelos conselhos e pelos momentos divertidos que passamos juntas.

À minha família, por acreditar em mim e me apoiar desde pequena.

Aos meus pais, um agradecimento muito especial, por me terem apoiado incondicionalmente. Obrigada por me terem ensinado tanto, por sempre acreditarem em mim e por me fazerem ter orgulho em ser vossa filha. Amo-vos!

Ao meu namorado Luís, meu amor, meu amigo e companheiro, o meu enorme obrigada por tudo. Obrigada por me teres ensinado tanto, e apoiares mas, acima de tudo, por teres compreendido todas as minhas ausências. Obrigada pela força, conselhos e momentos passados juntos. Amo-te!

Por fim, o meu enorme obrigada ao Universo por me ter dado uma segunda oportunidade e ter permitido que conseguisse dirigir-me para o caminho certo.

Index

Abstract

Spinocerebellar ataxia type 3 (SCA3), also known as Machado-Joseph disease (MJD) is a rare late-onset neurodegenerative disorder, which belongs to the polyglutamine (polyQ) group of diseases. The causal gene of this disease, *ATXN3*, presents CAG repeat expansion as the causative mutation. This gene encodes the protein ATXN3 an ubiquitous protein that can be involved in transcription mechanisms.

Untranslated regions (UTRs), 5' and 3' UTRs, along with *cis*-acting elements (such as gene promoters), are essential for the regulation of gene expression. Also, DNA methylation has an important role in such regulation.

The main goal of this thesis was to analyse the impact of untranslated and methylation prone regions in the clinical variability of SCA3. For this purpose, we studied the variation found in 3'UTR of *ATXN3* and analysed the impact of these variants in the variance of the onset of SCA3 patients.

We reported nine variants present in the 3'UTR of *ATXN3* that can have a role in its own expression regulation. In the Azorean SCA3 cohort studied, none of the variants that were identified presented a significant modulatory effect on the disease onset and we were not able to replicate results for rs709930 (and rs910369) obtained by other group of researchers. In this study, we were able to identify disease-associated haplotypes that can discriminate expanded allele; this fact can be relevant for the design of new allele-specific silencing strategies.

We also studied the variation found in the UTRs of different genes that was obtained through an extended exome sequencing and performed genotype-phenotype correlations to evaluate the impact of such variants/genes in the onset of SCA3 patients. In this study we identified four candidate modifier genes of the disease onset: *GLI4*, *TGFA*, *ERBB4* and *MPHOSPH9*. The alternative alleles of the variants in *GLI4*, presented a predicted modification in transcription factors binding sites and in transcription factors, that bind to those sites; this gene showed a delay of three years in SCA3 onset. The remaining modifier genes (*TGFA*, *ERBB4* and *MPHOSPH9*) identified in this study, are known *ATXN3* interactors, and significantly influence the onset of the disease. These genes presented variants in the 3'UTRs. We were also able to obtain eight statistically enriched pathways, two of them containing genes with a modifier effect on the onset (*TGFA* and *ERBB4*): HIF1α Signaling

and Neuregulin Signaling. These pathways can be important for further therapeutic applications in this disease.

Finally, we performed an epigenome-wide pilot study. As far as we are aware, is the first study investigating genome-wide DNA methylation changes in SCA3 patients compared to matched controls. As the sample size was small, we were not expecting genome-wide significant differentially methylated CpG sites, between patients and controls. Enrichment analysis of the obtained list of 1279 genes showed an enrichment for pathways involved in neurodegenerative diseases such as Calcium signaling pathway. We also performed an analysis of epigenetic clocks (Hannum and Horvath) for SCA3 patients and matched healthy controls. We were not able to find differences between patients and controls for DNA methylation age for both epigenetic clocks. For DNA methylation age acceleration difference and DNA methylation age acceleration difference residuals we were not able to find differences between the two groups for both clocks. These results can be explained by sample size and the limited power of this study.

This PhD thesis tried to contribute to the unveiling the role of untranslated and methylation prone regions in the clinical variability of SCA3. For this purpose, we obtained: a) variants with predicted impact in 3'UTR of *ATXN3* gene; b) candidate modifier genes, with variants in UTRs; c) no significant results for the first genome-wide study on DNA methylation comparing patients and controls; further studies in a higher number of patients and controls are required to obtain a methylation profile for SCA3.

Resumo

A Ataxia espinocerebelosa do tipo 3 (SCA3), também conhecida como doença de Machado-Joseph (MJD), é uma doença neurodegenerativa rara de manifestação tardia e pertencente ao grupo das doenças de poliglutamina (poliQ). O gene causal desta doença, *ATXN3*, apresenta como mutação causal a expansão do motivo repetitivo CAG. Este gene codifica a proteína ATXN3, uma proteína ubíqua que pode estar envolvida nos mecanismos de transcrição.

As regiões não traduzidas (UTRs), 5' e 3' UTRs, juntamente com elementos *cis-acting* (tais como os promotores dos genes), são essenciais para a regulação da expressão génica. Além disso, também a metilação do ADN tem um papel importante na referida regulação.

O principal objetivo desta tese de doutoramento foi analisar o impacto das regiões não traduzidas e das regiões propensas à metilação, na variabilidade clínica da SCA3. Para este efeito, estudámos a variação encontrada em 3'UTR do gene *ATXN3* e analisámos o impacto destas variantes na variação da idade de início da doença nos doentes com SCA3.

Reportámos nove variantes presentes em 3'UTR do gene *ATXN3* que podem ter um papel na sua própria regulação génica. Na coorte Açoriana, composta por doentes com SCA3, nenhuma das variantes que foram identificadas apresentou um efeito modulador significativo na variação da idade de início da doença, e não fomos capazes de replicar os resultados para a variante rs709930 (e rs910369) obtidos por outro grupo de investigadores. Neste estudo, conseguimos identificar haplótipos associados à doença e que podem fazer a discriminação dos alelos expandidos; este facto pode ser relevante para o desenvolvimento de novas estratégias de *allele-specific silencing*.

Estudámos também a variação encontrada nas UTRs de diferentes genes que foi obtida através da sequenciação alargada do exoma e efetuámos correlações genótipo-fenótipo para avaliar o impacto de tais variantes/genes na variação da idade de início dos doentes com SCA3. Neste estudo identificámos quatro genes candidatos a modificadores da variação da idade de início da doença: *GLI4*, *TGFA*, *ERBB4* e *MPHOSPH9*. Os alelos alternativos das variantes no gene *GLI4*, apresentaram uma modificação prevista nos locais de ligação dos factores de transcrição e nos factores de transcrição, que se ligam a esses locais; este gene apresentou ainda uma variação na idade de início da SCA3 de três anos (a idade de início é mais tardia). Os restantes genes modificadores (*TGFA*, *ERBBB4* e *MPHOSPH9*) identificados neste estudo, apresentam variantes em 3'UTRs, são conhecidos como interatores do gene *ATXN3*, e influenciam significativamente a variação na idade de

início da doença. Também conseguimos obter oito *pathways* estatisticamente enriquecidas, duas delas contendo genes com efeito modificador (*TGFA* e *ERBBB4*): *HIF1α Signaling* e *Neuregulin Signaling*. Estas *pathways* podem ser importantes para outras aplicações terapêuticas nesta doença.

Por fim, efetuámos um estudo piloto de metilação *genome-wide*. Tanto quanto nos é possível saber, este é o primeiro estudo que investiga as alterações de metilação do ADN em doentes com SCA3 e em controlos emparelhados. Como o tamanho da amostra foi pequeno, não esperávamos a ocorrência de CpGs diferencialmente metiladas, entre doentes e controlos. A análise de enriquecimento da lista de 1279 genes que obtivemos, mostrou um enriquecimento para *pathways* envolvidas em doenças neurodegenerativas, tais como a *Calcium signaling pathway*. Também realizámos uma análise com *epigenetic clocks* (Hannum e Horvath) para doentes com SCA3 e controlos emparelhados. Não conseguimos encontrar diferenças entre doentes e controlos para a idade de metilação do ADN, para ambos *epigenetic clocks* . No caso da diferença da idade de metilação do ADN e a diferença da idade de metilação residual do ADN, também não fomos capazes de encontrar diferenças entre os dois grupos para ambos os *epigenetic clocks*. Estes resultados podem ser explicados pelo tamanho da amostra e pelo poder limitado deste estudo.

Esta tese de doutoramento tentou contribuir para a descoberta do papel das regiões não traduzidas e às regiões propensas à metilação na variabilidade clínica da SCA3. Para este efeito, obtivemos: a) variantes com impacto previsto em 3'UTR do gene *ATXN3*; b) genes candidatos a modificadores, com variantes em UTRs; c) nenhum resultado significativo para o primeiro estudo a nível do genoma sobre a metilação do ADN, comparando doentes e controlos; serão necessários mais estudos, num número maior de doentes e controlos para obter um perfil de metilação para a SCA3.

Chapter 1: General introduction

1.1. Polyglutamine diseases: General aspects

Polyglutamine diseases (PolyQ) are dominantly inherited neurodegenerative diseases characterized by the expansion of a cytosine-adenine-guanine (CAG) triplet repeat in the coding region of the respective causative genes, which translates into an abnormally long polyglutamine tract in the encoded protein (revised in [1]). PolyQ diseases include Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA) and six spinocerebellar ataxias (SCA 1, 2, 3, 6, 7 and 17), all inherited as autosomal dominant, except for SBMA, which presents an X-linked recessive pattern of inheritance. In these diseases, an expansion of a CAG repeat appears in the genes *HTT*, *ATN1*, *AR*, *ATXN1, ATXN2*, *ATXN3*, *CACNA1A, ATXN7* and *TBP,* respectively (revised in [2]); noteworthy, the critical threshold for the CAG expansion (i.e. the number of repeats needed for the manifestation of the pathological phenotype) is disease-specific (Table 1.1).

PolyQ diseases show variation on age at onset (onset) – the age of the appearance of the first symptoms – which is dependent on the size of the $(CAG)_n$ tract (revised in [1,3]); the number of $(CAG)_n$ repeats explaining from 26to 88% of the variation in the onset (Table 1.1) [4]. However, the size of the $(CAG)_n$ tract only partially correlates with the onset, supporting an influence of other genetic, epigenetic and/or environmental factors on the variable expressivity of polyQ diseases (revised in [4,5]). For many of these diseases, intergenerational instability of the $(CAG)_n$ tract size [6–13] as well as somatic mosaicism have been reported [14–16]. Almost all of the polyQ diseases are characterized by anticipation- the presence of an earlier onset and more severe symptoms from one generation to the next [17,18], which occurs due to intergenerational instability of the $(CAG)_n$ tract size of the expanded allele; this means that in successive generations, individuals tend to carry a $(CAG)_n$ tract size in the expanded allele which is higher than the one harbored by their affect parent. Somatic mosaicism is a phenomenon in which $(CAG)_n$ tract size may vary according to the tissue and even within the same tissue from which the DNA is extracted; the molecular basis of somatic mosaicism is the mitotic instability of the expanded CAG tract [19].

A myriad of pathogenic mechanisms are involved in polyQ diseases, including enhanced DNA damage, protein misfolding and aggregation, formation of intracellular aggresomes, as well as transcriptional and proteasomal dysregulation (revised in [1]). PolyQ diseases are typically late-onset, frequently appearing in midlife and presenting progressive neuronal dysfunction and neuronal loss (revised in [3]). It is interesting to note that a

particular subset of neurons is targeted for each of these diseases, regardless the expression of the expanded/mutant proteins throughout the brain and in non-neural tissues (revised in [3,20]).

Spinal and bulbar muscular atrophy (SBMA) was the first repeat expansion disease to be described (revised in [3]) and the only polyQ disease whose causative mutation is located in a sex chromosome (Table 1.1.), as previously referred. SBMA presents a prevalence of 2.5 cases for 100,000 individuals [21] and is caused by a polymorphic CAG repeat expansion in the N-terminal domain of the androgen receptor (AR) gene, in exon 1 (revised in [3]). The AR is a nuclear receptor and its protein product is a member of the steroid receptor-thyroid of receptor superfamily (revised in [3]); the N-terminal domain is involved in the regulation of transcription of target genes (revised in [22]). Expanded triplet repeats in SBMA are mostly unstable but, contrarily to other polyQ diseases, they present a tendency for the contraction of a few CAGs from one generation to the next (revised in [22]); also, as referred above, a correlation of the $(CAG)_n$ tract size of the expanded allele with onset and the disease severity is observed in patients (revised in [22]). SBMA is characterized by a late onset and progressive degeneration of lower motors neurons in spinal cord and bulbar region of brainstem (revised in [3]). Despite of the low explanation of onset provided by the size of the $(CAG)_n$ tract (~30%, Table 1.1), only one study has been performed regarding genetic modifiers and their impact in SBMA onset [23]. Although different studies in animal models [24,25] as well as clinical trials have been conducted for SBMA [26–28], it remains without treatment (revised in [3]).

Huntington's disease (HD) is the most frequent polyQ disease, presenting a prevalence of 4-12.3 cases for 100,000 individuals (revised in [29]). This autosomal-dominant disease is caused by an abnormal $(CAG)_n$ tract size in exon 1 at the coding region of the huntingtin (*HTT*) gene (revised in [30]), encoding for the huntingtin protein ; huntingtin is often a cytoplasmic protein that can also present some degree of nuclear localization (revised in [31]). This protein is involved in embryonic development [32–35] and can act as a protein scaffold (revised in [31]); moreover, it plays a role in transcriptional regulation [36–38]. Noteworthy, it is now known that HTT is needed for the formation of cortical and striatal excitatory synapses [39]. In HD intermediate alleles are common and the individuals who present CAGs in the expanded allele ranging from 27 to 35 repeats, will not develop the disease; however, an offspring of individuals with 34 or 35 repeats, can develop the disease due to intergenerational instability which would generate, through expansion of the CAG

tract, alleles with 36 repeats.

Table 1.1. Polyglutamine diseases: main genetic and clinical features.

PolyQ disease	**Causative gene [3]**	**Chromosomal location**	**Normal $(CAG)_n$ range***	**Expanded $(CAG)_n$ range***	**Main clinical features [1,18]**	**Variance in onset explained by the $(CAG)_n$ tract size [40]**
Spinal and bulbar muscular atrophy (SBMA)	*AR*	Xq12	9-39	40-68	Fasciculations, neurogenic muscle atrophy and weakness	29%
Huntington's disease (HD)	*HTT*	4p16.3	6-35	36-180	Chorea, other progressive motor disability symptoms	50-60%
Dentatorubral pallidoluysian atrophy (DRPLA)	*ATN1*	12p13.31	<26	49-88	Ataxia, chorea, dystonia	50-68%
Spinocerebellar ataxia type 1 (SCA1)	*ATXN1*	6p22.3	6-34	41-83	Ataxia, ophthalmoplegia, spasticity, sensory symptoms	64-76%
Spinocerebellar ataxia type 2 (SCA2)	*ATXN2*	12q24.12	14-30	33-77	Ataxia, peripheral neuropathy, sensory symptoms	50-80%
Spinocerebellar ataxia type 3 (SCA3)	*ATXN3*	14q32.12	12-40	51-87	Ataxia, spasticity, ophthalmoplegia, pyramidal signs, sensory symptoms	46-76%
Spinocerebellar ataxia type 6 (SCA6)	*CACNA1A*	19p13.13	4-18	20-33	Ataxia, nystagmus	26-52%
Spinocerebellar ataxia type 7 (SCA7)	*ATXN7*	3p14.1	4-27	34-200	Ataxia, ophthalmoplegia, spasticity, pyramidal signs	71-88%
Spinocerebellar ataxia type 17 (SCA17)	*TBP*	6q27	25-42	41-66	Ataxia, dystonia, chorea, spasticity, pyramidal signs	unknown

*Based on http://neuromuscular.wustl.edu/ataxia/domatax.html

HD can appear in adult-onset as well as in juvenile forms; in the adult form, which is more frequent, individuals present 40 to 50 CAG repeats in the expanded allele, while in the juvenile form heterozygous patients often present more than 60 CAG repeats (revised in [31]) in the expanded chromosomes. HD is characterized by a triade constituted by cognitive, motor and psychiatric features (revised in [29]). Anticipation and intergenerational instability are also a hallmark of HD (revised in [31]). A strong yet incomplete inverse correlation between $(CAG)_n$ tract size and onset is observed in HD (Table 1.1) (revised in [31]), with the $(CAG)_n$ tract size explaining 50%-60% of the variance in onset [40]; as previously referred for polyQ diseases in general, the remaining variance should be under the influence of genetic and epigenetic factors, as well as environmental factors. Different studies have shown the impact of different genetic modifiers in this disease, such as the DNA repair genes *FAN1* and *MLH1* [41–46], highlighting the importance of studying those factors. Emerging therapies for HD have as focus the mutant *HTT* pre-mRNA and the mature mRNA which can be targeted by antisense oligonucleotides (ASOs) and interfering RNA (RNAi), respectively; these strategies have the aim to enhance early degradation and lowering of the mutant *HTT* levels [47]. Also, the mutated gene can be targeted by zinc finger proteins which repress transcription [47].

Dentatorubral-pallidoluysian atrophy (DRPLA) is a very rare disease presenting a global prevalence strongly related to the ethnic background; in Japan, the disease reaches its highest worldwide prevalence with 0.48 cases for 100,000 individuals whereas in the Caucasian population, only 25 families have been reported [3,48–56]. DRPLA is caused by an unstable CAG repeat in the atrophin-1 gene (*ATN1*) (Table 1.1) that encodes for the atrophin 1 protein, a transcription regulator that was identified in nuclear inclusions found in neurons and glia of the affected brain regions (revised in [3]). DRPLA is characterized by ataxia (revised in [3,57]), and its onset varies from infancy to adulthood [48]; this disease also presents genetic anticipation (revised in [58]). DRPLA mutated alleles are fully penetrant when the $(CAG)_n$ tract size is in the pathological range characteristic of disease (Table 1.1), explaining 50-68% of onset variance (Table 1.1.); noteworthy, incomplete penetrance has been reported for intermediate alleles (revised in [58]). To the best of our knowledge, no studies regarding genetic modifiers of DRPLA have been conducted.

Spinocerebellar ataxias belonging to the polyQ diseases group (SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17) are dominantly inherited and present as main features cerebellar ataxia, manifested by impairment of motor coordination, namely of gait [3].

Spinocerebellar ataxia type 1 (SCA1) represents approximately 6% of all autosomal dominant cerebellar ataxias (ADCAs) worldwide (revised in [3]). The CAG repeat expansion occurs in the coding region of *ATXN1* gene (Table 1.1) that encodes the protein ataxin 1, a protein involved in the transcription mechanism [59,60].

Spinocerebellar ataxia type 2 (SCA2) accounts for nearly 13% of the cases of ADCAs (revised in [3]); its causative mutation occurs in the coding region of *ATXN2* gene (Table 1.1), which encodes the protein ataxin 2, involved in RNA metabolism [61,62].

Spinocerebellar ataxia type 3 (SCA3), which is the focus of the present thesis, represents 20-50% of families with SCA [63] and is the most common form of SCA worldwide [64,65]. The *ATXN3* gene (Table 1.1) harbours the causative mutation of the disease (revised in [66,67]), encoding the ataxin 3 protein which, in its native form, presents a de-ubiquitylating activity (revised in [64]).

Spinocerebellar ataxia type 6 (SCA6) represents 13-15% of the SCAs [63]; its causative gene, *CACNA1A* (Table 1.1), encodes the protein voltage-dependent P/Q-type calcium channel subunit alpha-1A which is involved in a variety of calcium-dependent processes such as, muscle contraction, neurotransmitter release and gene [68].

Spinocerebellar ataxia type 7 (SCA7) and Spinocerebellar ataxia type 17 (SCA17) represent only 5% and 0.3% of families with SCAs (Table 1.1), respectively [69]. The causative gene of SCA7, *ATXN7*, encodes the ataxin-7 protein that is involved in the transcription mechanism [70] while the gene that causes SCA17, *TBP*, encodes the protein TATA-box-binding protein with transcription factor activity [71].

1.2. A paradigm of polyQ diseases: Spinocerebellar ataxia type 3/Machado-Joseph Disease (SCA3/MJD)

General characterization

Spinocerebellar ataxia type 3 (SCA3), also known as Machado-Joseph disease (MJD) was originally described in families of Azorean ancestry living in North-America [72–74]. As previously referred, SCA3 is the most common SCA worldwide [18], with a prevalence of 3.1/100,000 [75], and the second most common polyQ disease [18,63]. SCA3 (OMIM #109150; ORPHA # 98757) represents more than 50% of the cases of SCAs in Portugal (revised in [76]) and 78% of the cases of SCAs in Brazil [77]. In the Azorean archipelago (Portugal), the prevalence of this disease was 39/100,000 in 2016 [75]. SCA3 reaches its highest prevalence in the Azorean Island of Flores (Portugal) with a value of 633/100,000 (1/158) [75].

SCA3, a late onset disease, is characterized by an asymptomatic phase (a preclinical phase where the subjects do not present symptoms) and a symptomatic phase (a clinical, overt disease phase), The clinical phase of SCA3 exhibits a high degree of clinical variability, which represent the variable involvement of the cerebellar, motor neuron and oculomotor, peripheral, pyramidal and extrapyramidal systems [78–81]. While gait ataxia, dysarthria and ophthalmoparesis are symptoms that are frequently present in SCA3 patients, bulging eyes, blepharospasm and dystonia are less commonly observed [82]. Also, non-motor clinical features are reported for SCA3, including cramps [83,84], cognitive disturbances [85–89], depressive symptoms [90–93], olfactory impairment [86], sleep disturbances [94–98], and urinary dysfunction [99,100]. The first symptom of SCA3 is usually gait ataxia; however, although less frequent, diplopia has also been described as a first symptom [81]. The mean onset of SCA3 is around 40 years [50,101], with extremes of 4 [102] and 78 years [50]. The earliest onset that was reported by Carvalho and collaborators (Table 1.2) referred to an homozygous girl; the homozygosity present in this patient led to a more severe phenotype [102], as it has also been reported for other homozygous cases described [103], revealing that the *ATXN3* displays a dosage effect [19].

A few studies reporting homozygous SCA3 patients, revealed that the most common features include early onset, reduced survival time and overall, more severe phenotype (Table 1.2). Molecular studies using homozygous mice confirm particular molecular modifications such as increased aggregation [104]. Not all early onset/juvenile

cases of SCA3 are explained by homozygosity of the expanded allele; indeed, Paula Coutinho reported a case from the Azorean Island of Flores, with onset in infancy, that was explained by one very large expanded allele [81,105].

Table 1.2. Studies reporting homozygous cases in SCA3.

Reference	*ATXN3* genotype	Onset (in years)	Clinical features
Sobue *et al*, 1996 [106]	67/67	28	Ataxia, bulging eyes, dystonia, ophthalmoplegia, pyramidal findings
Lerer *et al*., 1996 [107]	64/70 65/66 67/68 65/69 67/68 65/69	29 37 25 17 30 36	Dysarthria, gait ataxia, lower motor neuron disease Gait ataxia, lower motor neuron disease Dysarthria, lower motor neuron disease, ophthalmoplegia
Fukutake *et al*, 2002 [108]	60/60 63/69	43 ?	Psychiatric and sleep problems at age 43; gait ataxia at age 50 Abnormal behaviour, sleep disturbances, hallucinations
Gordon *et al*., 2003 [109]	65/66	36	Ataxia, dysarthria, and oculomotor dysfunction
Carvalho *et al*, 2008 [102]	66/72	4	Ataxia, bulging eyes, dystonia, pyramidal findings
Lysenko *et al*., 2010 [103]	60/63	29	Mild ataxia, pyramidal findings
Zeng *et al*,, 2015 [110]	71/71	18	Abnormal eye movement, pyramidal signs, severe ataxia
Shang *et al*., 2018 [111]	64/67 57/65 57/65	33 38 33	Ataxia, bulging eyes, oculomotor disturbance, pyramidal signs Ataxia, oculomotor disturbances Ataxia, dystonia oculomotor disturbances, pyramidal signs,
Li *et al*., 2020 [112]	63/67 64/65 62/64 60/63 62/63 55/56 56/58	25 35 35 49 45 41 59	Cognitive impairment, dysarthria, gait ataxia, nystagmus Dysphagia, gait ataxia, nystagmus, peripheral neuropathy Bulging eyes, dysphagia, nystagmus, peripheral neuropathy Dysarthria, gait ataxia, nystagmus, peripheral neuropathy Dysarthria, gait ataxia, nystagmus Gait ataxia Cramps, gait ataxia, nystagmus
Chen *et al*., 2021 [113]	62/62	37	Nystagmus, severe dysarthria, wheelchair bound

In SCA3, imaging and neuropathology-based studies have reported gray matter atrophy in cerebellar, neostriatum, frontal lobe and parietal lobe [114–119]. The severity of the brain atrophy (loss of volume) was reported to be correlated with clinical heterogeneity;

this correlation showed that more atrophy of the brain leads to a wider variability of symptoms [120–122].

Genetics and genotype-phenotype correlations in SCA3

The causative gene of SCA3, *ATXN3*, is located at the long arm of chromosome 14 (14q32.12) [66,123] and contains, in exon 10, an unstable CAG repeat motive (Figure 1.1); *ATXN3* is considered to exhibit almost full penetrance (~98%) [19].

Normal *ATXN3* alleles present 12 to 44 CAG repeats; consensually, alleles that display above 52 repeats [76] cause SCA3.

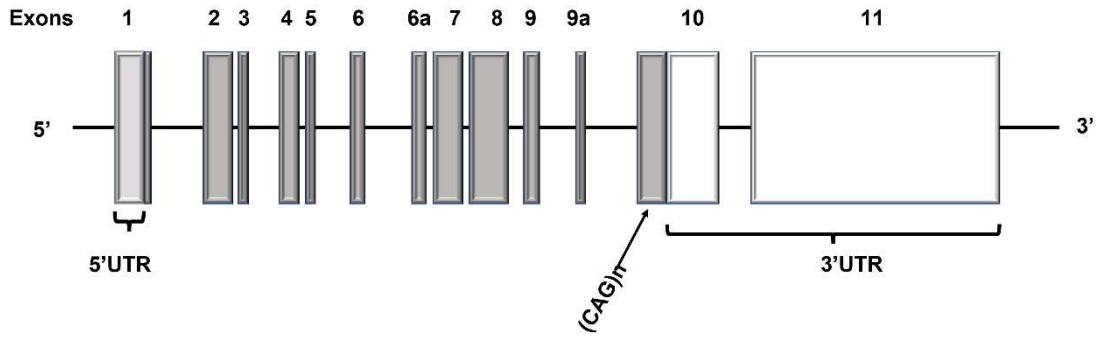


Figure 1.1. Representation of the causative gene of SCA3 - *ATXN3*. The figure presents exons (boxes in dark grey), numbered from 1 to 11, and UTRs (5'UTR is represented in a light grey box and 3'UTR is represented in white boxes). The $(CAG)_n$ is located in the exon 10 (adapted from [124]).

Less frequently found are intermediated alleles that vary between 45 and 54 CAG repeats [125–130] and can be related, or not, with the SCA3 phenotype (Table 1.3); in fact, the behaviour of these alleles is not well understood.

Table 1.3. Studies presenting SCA3 subjects with intermediate alleles and corresponding phenotype.

Reference	$(CAG)_n$	SCA3 phenotype
Takiyama *et al.* ,1997 [125]	56	Present
van Schaik *et al.*,1997 [126]	54	Present
Maciel *et al.*, 2001 [127]	51	Absent
van Alfen *et al.*, 2001[128]	53 and 54	Present
Gu *et al.*, 2004 [129]	51	Present
Padiath *et al.*, 2005[130]	45	Present

In SCA3, an inverse correlation between the $(CAG)_n$ tract size and onset is observed [115,131–134]. Also, similarly to the remaining polyQ diseases, anticipation has been reported in SCA3 (revised in [76]). Anticipation, now considered the hallmark of PolyQ diseases, is linked to intergenerational instability, as previously referred, specifically due to the occurrence of further expansions, which in the case of SCA3 are more frequent in male meiosis (revised in [76]). Noteworthy, instability is an attribute of mutated *ATXN3* alleles, with normal alleles consistently being stable upon transmission [135].

Besides the occurrence of intergenerational instability, *ATXN3* $(CAG)_n$ tract also exhibits somatic mosaicism, a reflex of the mitotic instability of the $(CAG)_n$ [19]; this instability leads to modifications in CAGs repeat length and consequently creates contractions or expansions [19]. Somatic mosaicism, which results in population of cells presenting different repeat sizes has been described in different studies (see amongst others [15,16,136–139]). Noteworthy, the pattern of somatic mosaicism for mutated SCA3 alleles is identical in the various areas of central nervous system, which present high tracts; the exception is the cerebellar cortex which presents reduced tracts [136,137]. Notwithstanding, studies performed failed to demonstrate a correlation between the degree of mosaicism and selective neuronal vulnerability; this finding has been interpreted as putatively indicating that neurons with larger CAGs can have already died [19]. Somatic mosaicism also contributes to limiting the precision of the sizing of the $(CAG)_n$ tract, as it leads to differences in the size of CAG triplets among lymphocytes and central nervous cells [19]. Technical limitations have prevented the accurate study of somatic mosaicism in SCA3 and in other polyQ diseases. Novel next-generation sequencing (NGS)-based pipelines have recently emerged and will likely lead to important advances in the understanding of the impact of somatic mosaicism in polyQ diseases.

The identification and direct detection of *ATXN3* mutation allowed the development of a molecular test to confirm the status of carrier/non-carrier. The gold-standard for the molecular test of SCA3 is fragment analysis, whose laboratorial workflow includes DNA extraction, PCR amplification and capillary electrophoresis. According to the guidelines for SCAs [140], a difference of ±1 CAG repeat for the normal allele is considered acceptable, when the same sample is genotyped in different laboratories; for the expanded allele, a difference of ±3 CAG repeats is considered acceptable. The molecular test is offered in several contexts, to individuals either with a clinical diagnosis of cerebellar ataxia (molecular diagnosis), or to healthy at-risk subjects (pre-symptomatic test). Subjects genotyped as

homozygous for the normal allele need to be further studied using a complementary technique, the triplet repeat primed PCR (TP-PCR), which is able to confirm the absence of an expanded allele (expanded alleles are harder to amplify, and there is a probability that allelic-drop-out occurs). TP-PCR for SCA3 was developed by Melo and collaborators [141], leading to a higher accuracy of the molecular test. Prenatal as well as Preimplantation Genetic Diagnosis are also available for SCA3 (revised in [76]).

Ataxin-3 protein and disease pathogenesis

ATXN3 gene encodes the protein ataxin-3, a deubiquitinase (DUB) belonging to the cysteine-protease family, with a molecular weight of 42kDa and total of 361 amino acid residues in its native form [142]. Ataxin-3 is composed by a N-terminal domain – Josephin domain – and a carboxyl (C)-terminal tail [143]; this tail contains several ubiquitin interacting motifs that bind to polyubiquitin chains as well as the polyglutamine tract [144]. The Josephin domain presents a glutamine and a cysteine in the N-terminal, and an histidine and an asparagine in the C-terminal tail; these three last amino acids constitute a catalytic triad, characteristic of cysteine proteases (revised in [65]). In the mutated protein, the polyglutamine tract is present in its elongated form (61 to 87 glutamines) (revised in [65]).

Although there is selective neuronal degeneration in SCA3 patients, the normal and mutated ataxin-3 are expressed in a ubiquitous form, being located in the cytoplasm and nucleus of the cells [142,145]. Different ataxin-3 isoforms have been described in blood, as a consequence of alternative splicing [146]. Ataxin-3 is reported to play a role in transcription, being able to interact with diverse transcriptional regulators (revised in [147]). Ataxin-3 can also target chromatin through direct binding to DNA and histones [148]. Conformational changes caused by mutated CAG tract size lead to the misfolding of the protein and consequently affect its stability, degradation, subcellular localization, and molecular interactions with other proteins [148].

The gain of function or the partial loss of the normal ataxin-3 function can be a consequence of the abnormal interactions of mutated ataxin-3 with its native partners (revised in [149]). The mutated protein can be cleaved in a proteolytical manner leading to the formation of smaller aggregation prone fragments; these aggregates can form soluble monomers, oligomers and/or insoluble aggregates located in nucleus and in cytoplasm

(revised in [149]). Misfolding of mutated ataxin-3 may disturb various processes, such as: autophagy, transcriptional regulation, DNA repair, and mitochondrial function (revised in [150]). Indeed, different molecular chaperones presented dysregulated expression levels in SCA3 models and showed an associated pattern related to disease progression (revised in [151]). In the fibroblasts of SCA3 patients [152] and in some mouse models [153–155] the impairment of autophagy have been reported; during disease progression an accumulation of mutant ataxin-3 occurs that correlates with a failure of protein clearance pathways (revised in [156]). Furthermore, the disease pathogenesis can be influenced by the dysregulation of transcription (revised in [147]), which should occur due to the abnormal interactions with coactivators and transcription factors (revised in [147]). The knowledge concerning molecular pathways involved in the pathogenesis of SCA3 provides clues for the development of therapeutic approaches. Hence, the chaperon system, the ubiquitin-proteosome system and autophagy have been explored as therapeutic targets (revised in [156]).

Genetic modifiers of SCA3

In SCA3, the frequency of the different *ATXN3* CAG alleles can exhibit inter-populations differences. In the Azorean cohort, the largest expanded allele reported contained 79 CAG triplets [101], whereas in a Japanese cohort an allele harbouring 86 CAG repeats was identified [50]; the Brazilian cohort has the largest expanded alleles reported so far, the largest with 91 CAGs [157]. As previously referred, for SCA3, the $(CAG)_n$ tract size in the expanded allele only explains 46% to 76% of the variation found in the onset [4], which implies that non-CAG factors such as genetic modifiers, may contribute to the phenotype of the disease (revised in [4]).

Genetic modifiers are defined as genes that can modify the phenotype of a disease and/or molecular expression of other genes; also, can affect penetrance, dominance, expressivity and pleiotropy [158,159]. In SCA3, two major strategies have been undertaken to attempt to identify genetic modifiers: i) candidate gene approaches, and ii) genome-wide strategies (revised in [4]). A candidate gene approach is used when a study on a specific gene or a small number of genes is performed to identify genetic modifiers. A candidate gene can be the disease causative gene itself (if variation other than the causative mutation itself is being investigated), a gene that is known to play a role in the pathogenesis of the disease (interacting directly or indirectly with the causative gene) or a gene selected for

being previously reported as a disease-modifier for other similar neurodegenerative diseases, such as HD, for example. On the other hand, a genome-wide strategy does not depend on "a priori" hypothesis and implies, for example, an unbiased search by whole-exome sequencing or using arrays of markers (such as arrays of single nucleotide variants-SNVs). Whole-genome methylation arrays, also provide an unbiased strategy to find genetic modifiers.

Regardless of the strategy, effects of genetic modifiers in SCA3 have been difficult to replicate in the different cohorts, a fact that is usually attributed to the study design, sample size or frequencies of the candidate loci in a specific population [4,160]. Although genome-wide association studies (GWAS) hold the promise of preventing some of these problems by identifying several modifier variants and their interaction, as well as pinpointing modifying pathways related to those modifiers [160] difficulties of this strategies have also been acknowledged.

A summary of the main studies of genetic modifiers for SCA3 is displayed in Table 1.4 (only studies published based on cohorts including 100 or more patients are shown).

In general terms, candidate gene approaches performed so far for SCA3 can be grouped by the underlying rational of the study: i) studies concerning variation at the causative *ATXN3* locus (other than the expanded CAG repeats); ii) studies that targeted other expansion loci as modifiers of SCA3; iii) studies with analysed other genes/variants based on the fact that such genes belong to pathways which are altered in SCA3.

Different studies analysed the variation at the *ATXN3* locus (other than the expanded CAG repeats). França and colleagues [161] evaluated if the normal *ATXN3* allele can be a genetic modifier of SCA3. In fact, they reported that the size of the normal *ATXN3* allele presents a small yet significant influence on the onset [161]. A few studies searching for modifying factors of SCA3 targeted the non-coding regions, of *ATXN3*, such as its untranslated regions (UTRs). A study conducted by Bettencourt and colleagues analysed a variant found in the 5'UTR of *ATXN3*, rs3814834, which showed no improvements in the explanation of the variance in onset of SCA3 patients [162]. Long and colleagues studying the 3'UTR of *ATXN3* reported that the presence of the alternative allele for two variants in *linkage disequilibrium* (LD), rs709930 and rs910369, was associated with an earlier onset for the SCA3 patients [163]. Later, Melo and colleagues [124] (for more details see Chapter 2), in a study describing the genetic variation found in the 3'UTR of *ATXN3*, reported that the presence of alternative allele for rs10151135 correlates with a higher SARA (Scale for the

Assessment and Rating of Ataxia) axial sub score. However, this study was not able to replicate the results from Long and colleagues [124,163].

Studies that targeted other expansion loci as modifiers have had as hypothesis that native proteins encoded by different polyQ loci interacts with mutant ataxin-3. Jardim and colleagues [164] reported that the $(CAG)_n$ tract size of the SCA2 gene presented a significant association with the presence of fasciculations in SCA3 patients. Tezenas Du Montecel and colleagues [50] studied the seven expansion polyQ loci and reported that *HTT*, *ATN1 ATXN2* genes were modifiers of onset in SCA3 patients. Patients with shorter *HTT* alleles presented a delay in onset, while SCA3 patients with larger *ATN1* alleles presented an earlier onset; patients with an intermediate allele presented an earlier onset than the ones with a shorter ATXN2 alleles [50]. Raposo *et al*. [101] were not able to replicate the findings from Tezenas Du Montecel and colleagues [50]; notwithstanding, a negative correlation of the longer allele at *ATXN1* with onset was observed in the SCA3 Azorean cohort, increasing the explanation of onset by 1.5% [101]. In response to the research published by Tezenas Du Montecel and colleagues [50], Chen and collaborators [165], studying a large cohort of Chinese Han patients confirmed that the size of the normal CAG tract size in other polyQ loci (*ATXN1* and *ATXN2*) influences SCA3 onset.

A study conducted by Bettencourt and colleagues evaluated the potential of variants in DNA repair genes as modulators of the onset in different polyQ diseases [166], including SCA3. This study revealed that rs3512, a variant that is localized at the 3'UTR of the *FAN1* (FANCD2 And FANCI Associated Nuclease 1) gene can modulate the onset of different SCAs and HD. This same variant sparked the interest of Mergener and colleagues [167]; these researchers showed homozygosity for the alternative allele of rs3512 led to a decrease of onset of 2.44 years. In the same study, the combination of being homozygous for rs3512 alternative allele and carrying a short *ATXN2* allele led to delay in onset of 4.25 years [167].

Besides variation in the nuclear genome, the mitochondrial genome has also been investigated with the purpose of identifying modulators of SCA3 phenotype. mtDNA haplogroups have been associated with manifestation of symptoms in different polyQ diseases [168], being potential contributors of the variance in the onset. A study of mtDNA haplogroups conducted by Ramos and colleagues [168] showed that haplogroups W and X, seem to have a protective effect in SCA3, causing a delay in onset in the Portuguese population.

Several genetic modifiers studies targeted genes that had already implicated in other neurodegenerative diseases. *APOE* ε2 allele, for example, has been associated with an earlier onset in HD [169]. Similarly, the study performed by Bettencourt and colleagues reported that the presence of the *APOE* ε2 allele is associated with a decrease of 5 years in onset [170]. Later, two different studies, both analysing Chinese cohorts, attempted to replicate the findings of Bettencourt and colleagues. Whereas the study conducted by Peng and colleagues [171], which studied the Chinese Han Population, also reported a decrease (of 4 years) in the onset of the carriers of the *APOE* ε2 allele, in accordance with results from Bettencourt and colleagues [170], the study with 403 SCA3 patients from eastern and southeaster China by Zhou and collaborators failed to identify a modulator effect of *APOE* on SCA3 onset [172]. The glucocerebrosidade gene, *GBA*, has been associated with Parkinson disease. A study by Siebert and colleagues [173], using samples from the Brazilian and Azorean cohort reported that variation found at this gene can play a role as a SCA3 modifier.

The growing knowledge concerning pathways altered in SCA3 has allowed to investigate the modifier potential of genes belonging to some of these pathways. Inflammation, for example, is considered an important mechanism in several neurodegenerative diseases, including SCA3 [174]. Raposo and colleagues studied the variation in the promoter of several genes which encode pro-inflammatory molecules, namely Interleukin 6 (*IL6*); they reported that for the *IL6* rs1800795 the presence of C allele had a significant impact in decreasing the onset (an average of 4 years earlier onset) when comparing with carriers of the G allele [174]. Also, in the same study it was shown that the presence of the *APOE* ε2 allele seems to anticipate onset in an average of 10 years for patients which also carry the C allele in *IL6* [174].

Autophagy is one of the pathways impaired in SCA3 [145]. Several genes are known to play a relevant role in this pathway, namely *BECN1* [145]. Decreased levels of beclin-1 have been identified in affected brain regions of patients with Alzheimer disease (AD), early in the disease process [175]. Kazachkova and colleagues [176] investigated the presence of rs60221525 alternative allele (C) localized in the promoter of *BECN1*. A tendency for a decrease in *BECN1* expression levels was observed for the patients presenting the alternative allele of rs60221525 [176]; moreover, a tendency for an earlier onset and a more severe clinical presentation of the disease in the presence of this allele, after taking into account the size of the $(CAG)_n$ tract size, was observed.

As unbiased approaches querying the entire exome/genome became more accessible, the first studies attempting to identify modifiers using such approaches emerged. A SNV array study, conducted by Akçimen and colleagues [177], using a large multicentric SCA3 cohort (N=786) revealed the presence of nine variants, *THSD7B* (rs62171220), rs2067390 (*INPP1*), *LOC105379062* (rs144891322), *LOC107984326* (rs11529293), *LOC105378956* (rs585809), *ATP11A* (rs72660056), *LRRC28* (rs11857349) and *NFAM1* (rs8141510) that were involved in the modulation of SCA3 onset by increasing or decreasing the onset (Table 1.4). Profiting from the extensively studied Azorean SCA3 genealogies, Raposo and collaborators [160] used a concordant and discordant sib-pair design to filter variants identified by WES and select potential modifiers of SCA3. In this study a total of nine genes (*PARD3, NFKB1, CHD5, ACTG1, CFAP57, DLGAP2, ITGB1, DIDO1 and CERS4*) were reported as modulators of onset in SCA3 patients, in addition to the $(CAG)_n$ tract size.

Noteworthy, the majority of studies that found genetic modifiers in SCA3 were focused on coding regions of the respective genes. However, variation found in the UTRs of genes, other than *ATXN3,* has also the potential to modulate their expression and thus, disease onset. Indeed, the 5'UTRs of different genes present a sequence with a high diversity of transcription factors binding sites, that are crucial for gene expression by allowing or not, the binding of transcription factors, which are responsible for the regulation of transcription initiation and stability [178,179]. On the other hand, the 3'UTRs of genes are also important components of the gene expression process, due to the presence of i) microRNAs binding sites, that allow or not the binding of a specific microRNA and consequently have an impact in transcription; ii) recognition motifs for RNA-binding proteins, which determined the binding of proteins to those binding sites; iii) alternative polyadenylation signals, that are present in specific sites, and lead to different sizes of the 3'UTR in a gene [180].

Epigenetic mechanisms, such as DNA methylation, histone modifications and non-coding RNA (ncRNA) – associated gene silencing play an important part in gene activity and transcription regulation [181] and are potentially involved in the phenotypic modulation of disease. Indeed, DNA methylation has been reported to play a role in some neurodegenerative diseases, such as Alzheimer's disease [182], Parkinson's disease [183] and Huntington disease [184].

In SCA3, two studies conducted in the promoter region of *ATXN3* revealed that DNA methylation in this region may contribute to the onset [185,186] as well as to the instability of the $(CAG)_n$ tract [186]. Also, a study concerning different SNVs in DNA methylation-related genes (i.e. genes encoding methyl-CpG binding domain protein 1 - MBD1) has revealed significant differences in the distribution of rs12957023 between patients and controls for allelic and genotypic frequencies [187]. A study conducted by Zhao and colleagues [188] was the first to report a pair of monozygotic twins with SCA3. This study revealed that the twins presented identical $(CAG)_n$ tract size and genotypes for a set of reported genetic modifiers, whereas the onset and the severity of the symptoms were different. The methylation study of blood DNA samples revealed differences in DNA methylation levels of several CpG sites between the twins [188], raising the hypothesis that differences in clinical presentation would be linked to methylation levels [188].

More recently, a DNA methylation study performed by Li and colleagues [189] analysed its impact in onset. The researchers performed an epigenome-wide study in SCA3 patients and then applied Horvath Pan Tissue epigenetic clock [190] to obtain DNA methylation age for each SCA3 patient. DNA methylation age acceleration, obtained by the difference between DNA methylation age and chronological age was reported and was then correlated with the onset of the disease. Although several limitations were acknowledged in the study, the results seem to indicate that an increase in DNA methylation age acceleration is associated to an earlier onset [189].

Table 1.4. Main published studies searching for genetic modifiers of SCA3, using onset as phenotypic variable. Only studies including over 100 participants and producing statistically significant results are shown.

Reference	Year	Gene (s)	Variant (s) with significant results	Type of variant (s)	Sample size	Impact in the onset
Ding *et al.*, 2022 [191]	2022	*NPY*	rs16139, rs3037354, rs2234759, rs11100494	Missense, upstream variant, 5'UTR, intronic	527	NA
Raposo *et al.*, 2021 [160]	2022	*PARD3B, NFKB1, CHD5, ACTG1, CFAP57, DLGAP2, ITGB1, DIDO1* and *CERS4*	rs11009651 (*PARD3),* rs4648050 (*NFKB1),* rs2273034 (*CHD5),* rs1139405 (*ACTG1*), rs2483688 (*CFAP57*), rs2293909 (*DLGAP2*), rs2298141 (*ITGB1*), rs1883848 (*DIDO1*), rs1883847 (*DIDO1),* rs28334 *(CERS4),* rs36247 (*CERS4)*	Synonym, intronic	282	Earlier onset: *PARD3B* Later onset: *NFKB1, CHD5, ACTG1, CFAP57, DLGAP2, ITGB1, DIDO1* and *CERS4*
Martins *et al*,, 2021 [192]	2021	*CAST*	rs27852, rs1559085	Intronic, upstream variant	287	Later onset
Akçimen *et al.*, 2020 [177]	2020	*THSD7B, INPP1 LOC105379062, LOC107984326, LOC105378956, ATP11A, LRRC28, NFAM1*	rs62171220 *(THSD7B),* rs2067390 (*INPP1),* rs144891322 (*LOC105379062*), rs11529293 (*LOC107984326*), rs585809 (*LOC105378956*), rs72660056, *(ATP11A*), rs11857349 *(LRRC28),* rs8141510 (*NFAM1*)	Intronic	786	Earlier onset: *LOC107984326, LOC105378956, ATP11A, LRRC28* Later onset: *THSD7B, INPP1, LOC105379062, NFAM1*
Mergener *et al.*, 2020 [193]	2019	*FAN1*	rs3512	3'UTR	144	Earlier onset
Bettencourt *et al.*, 2016 [166]	2016	*FAN1*	rs3512	3'UTR	397	Later onset
Long *et al.*, 2015 [163]	2015	*ATXN3*	rs709930, rs910369	3'UTR	170	Earlier onset
Tezenas du Montcel *et al.*, 2014 [50]	2014	*ATXN2, ATN1, HTT*	*	Indel	403	Earlier onset: *ATXN2* Later onset: *ATN1, HTT*
Peng *et al.*, 2014 [194]	2014	*APOE*	rs429358, rs7412	Missense	155	Earlier onset
Bettencourt *et al.*, 2011 [170]	2011	*APOE*	rs429358, rs7412	Missense	192	Earlier onset

NA- information not available

*several dbSNP IDs can identify the mutation

1.3. Regulation of gene expression – the importance of untranslated regions and dna methylation

Gene expression is a coordinated and highly regulated process and it is of crucial importance for a better understanding of the molecular mechanisms on the basis of different diseases [195,196]. Despite all human cells presenting essentially the same genome (except when somatic mosaicism is present), the proportion of information that is expressed varies in different cell types and at different life stages [195]. Modifications of gene expression patterns play fundamental roles in cell differentiation, and may also contribute to pathological responses and/or disease states and processes [195]. Gene expression regulation occurs in the context of gene interactions that can be shaped by allelic variability, and by environmental factors which may lead to epigenetic modifications of the DNA [195,197].

Transcription is the process that allows the formation of a RNA copy of a DNA sequence (Figure 1.2) and can be affected by chromatin structure, RNA polymerase activity, transcription factors availability and accessibility, as well as DNA methylation. In transcription, the formation of a preinitiation complex on the core promoter is necessary [195]. This complex is formed by RNA polymerase II and general transcription factors (GTFs) [195]. Transcription is initiated with the recognition of a promoter by RNA polymerase II. Promoter sequences are localized near the transcription starting site (TSS) and present a region rich in adenine and thymine whose function is to determine the exact transcription starting site and which strand will be the template for this process [198,199]. Promoters are divided in two main types: focused and dispersed [195]. Dispersed promoters are formed by multiple and weak transcription starting sites, lack canonical core promoter sequence elements, and frequently are fixed in CpG islands; focused promoters, on the other hand, contain a single predominant transcription starting site and, containing combinations of certain canonical nucleotides sequence motifs [195,196]. The pre-initiation complex can be characterized by a multistep recruitment process of the general transcription factors (TFs) and the RNA polymerase II, which can be facilitated or inhibited by regulatory proteins [195,196]. The general transcription factors of polymerase II, namely TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH act in core promoter recognition; noteworthy these factors are involved in the polymerase II recruitment, the interaction with regulatory factors, transcription starting site recognition as well as promoter clearance [195,196]. After the pre-initiation complex is

reorganized, TFIIH will catalyze ATP-dependent uncoil of around ten base pairs of promoter DNA; polymerase II localizes the transcription starting site on the DNA template strand and then RNA synthesis (Figure 1.2) is initiated [195,196]. After this step, the clearance of the promoter occurs and then polymerase II moves downstream the DNA template strand and the elongation of the RNA transcript continues [195].

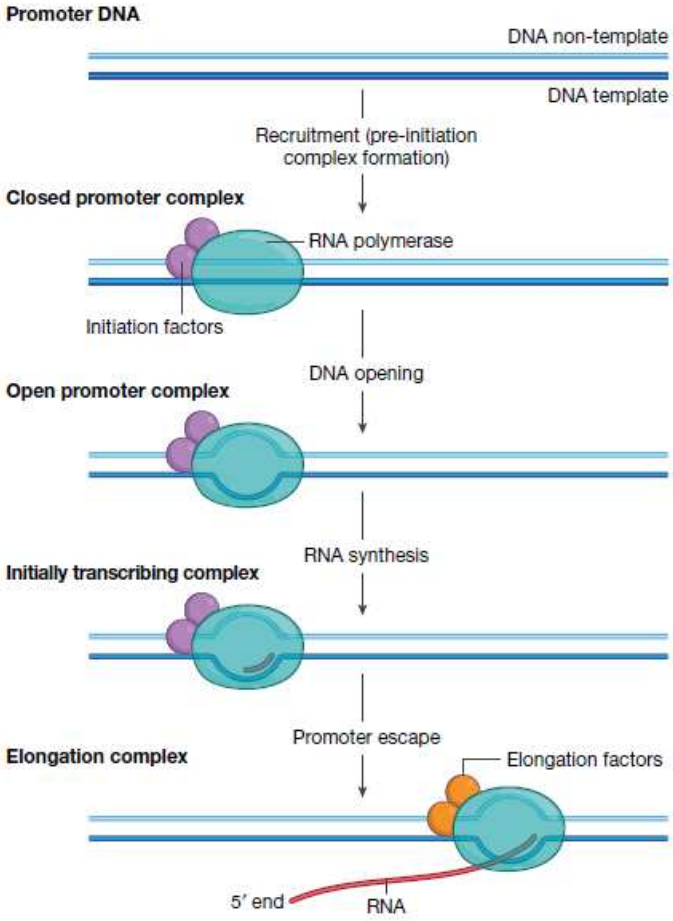


Figure 1.2. Fundamental steps that occur during gene transcription. RNA polymerase associates with initiation factors to perform the recognition of DNA promoter and form a pre-initiation complex. The opening of DNA converts the closed promoter complex to the open promoter complex that contains the template DNA in the polymerase active site [198]. An elongation complex, which can bind to elongation factors, is formed due the escaping of polymerase, after the higher grow of RNA. The last step of this process, (not shown on the figure), is the dissociation of polymerase from DNA and RNA that leads to the termination of the transcription cycle. Figure adapted from [198].

After the synthesis of the nascent RNA, polymerase II frequently pauses (promoter proximal pausing) without releasing the nascent transcript; this only occurs when additional signals lead to the transition to productive elongation [195,196]. The regulation of the escape from this pause leads to a fast response and synchronous gene activation in the neighboring cells of a tissue; the promoter proximal pausing is also linked to the 5'capping of the nascent pre-mRNA and is developed by protein factors, namely, negative elongation factor and sensitivity-inducing factor The transcript suffers cleavage at the polyadenylation site (polyA site) located in 3'UTR. This cleavage leads to the release of the mRNA and the addition of polyA tail [195,196]

Transcription is regulated through the binding of proteins to DNA specific motifs; transcription factors and RNA polymerase constitute the basal transcription complex which is necessary for high levels of transcription [197–199]. Enhancers are located far away from the promoter and generally contain binding sites for diverse transcription factors [198].

The binding of RNA polymerase occurs by the recognition of the motifs in the binding sites of transcription factors, that lead to the unfolding of DNA template strand; after this process, transcription occurs through the gene in the direction 3' to 5' , until RNA polymerase finds the terminal sequence and releases the DNA strand as well as the RNA strand that is synthetized [197–199]; other molecules – silencers- can further influence transcription leading to its repression [197–199].

The regulation of transcription is accomplished by modifying the abundance and/or activity of transcription factors, chromatin remodelers, coregulators, and other proteins [195,196]. The signals that can affect gene expression include amongst others, hormones, metabolic signals, thermic stress, oxidative stress, DNA lesions, hypoxia and organelle stress [195]. Gene stimulatory signals can change due to diverse factors, such as: i) transcription precursor proteolytic cleavage that can affect the active transcription factor; ii) disruption of protein-protein interactions that transform the inactivity of transcription factor in an active state; iii) increased transcription factor stability; and iv) transcription factor dimerization [195,196]. Nuclear receptors, which are ligand modulated transcriptions factors, can bind to *cis*-regulatory DNA sequence elements affecting the transcription of a target gene [195,196].

Besides transcription factors, coregulators are also important to the control of expression levels and its regulation [195].

It is known that the dysregulation of the transcription process has a role in the SCA3 pathogenesis and that this dysregulation occurs by two different processes : i) sequestration of transcription factors to polyQ rich aggregates; and ii) irregular interactions with transcription factors [148]. As ataxin-3 is able to interact with diverse transcriptional regulators (revised in [147]) and to target chromatin structure by binding histones or DNA directly [148], the study of the regulation of gene expression by UTRs as well as the analysis of global DNA methylation in the presence of mutated ataxin-3 is pertinent.

Five prime untranslated regions (5'UTRs) are DNA regulatory regions localized at the upstream end of all protein-coding genes; this region contains diverse transcription factors binding sites which together with transcription factors have a regulatory role [178]. Transcription factors can be ubiquitously expressed or be restricted to a specific tissue, and bind to specific motifs in 5'UTRs – transcription factor binding sites [200]. The transcription of a gene can be increased due to site-specific transcription factors, which bind directly to DNA in a proximal or distal manner with respect to a specific transcription starting site [200]. The mechanism of transcription can also be down-regulated by transcription factors that present a repressive function, which can interfere with the activation of transcription factor binding, preventing the recruitment of the transcription machinery [200]. The most common transcription factors are formed by multiple domains with distinct functions, such as dimerization with other transcription factors, contact with DNA, and contact through co-factors the basal transcription machinery [200]. Transcription factors are commonly classified based on their DNA-binding domain, being its major types: i) zinc finger; ii) helix-turn helix; and iii) homeodomain [200].

The availability of bioinformatic tools that can readily identify modifications in transcription factors binding sites caused by genetic variation has promoted an increased knowledge in this field. Predictive internet-based tools such PROMO [201], allow the screening of DNA sequences with databases of sequence motifs; however, in silico screening tends to largely overpredict binding sites [200], a fact that needs to be taken into consideration when analyzing and interpreting data derived from these tools.

Three primed untranslated regions (3'UTRs) are more conserved than other noncoding regions such as promoters (revised in [180]). 3'UTR are involved in diverse regulatory processes, such as transcript cleavage, stability and polyadenylation; these regions are important to determine the fate of mRNA molecules [178], presenting different

binding sites that may play a role in the transcription process. 3'UTRs contains microRNAs binding sites, binding sites for RNA binding proteins and polyA sites with specific signals.

Cis-acting regulatory DNA elements, which are present in 5' and 3' UTRs, are elements that are on the same chromosome as the gene whose expression they regulate [196]; these elements are extremely linked to the identification of RNA binding proteins which bind and mediate some functions of 3'UTR (revised in [202]). MicroRNA binding sites and AU-rich elements are the best studied *cis-elements*, that control the stability of mRNA [203]. RNA binding proteins when bind to the respective recognizing motifs in 3'UTR, may recruit decapping or deadenylation factors that will destabilize the mRNA [203]. If a gene originates alternative 3'UTRs, the alternative 3'UTR isoforms may control the levels of protein abundance; this is possible also by the polyA signals and by the action of microRNAs [203].

In addition to the gene expression regulation dependent on the DNA sequence such as the sequence of untranslated regions, epigenetic mechanisms are known to be major determinants of gene expression. Epigenetic mechanisms and elements, which include DNA methylation, histone modifications and non-coding RNAs, refer to a variety of processes that have heritable effects on gene expression without changing the DNA sequence [204].

Histone modifications include acetylation, methylation and others [205]. These modifications can contribute to the gene expression control by influencing the compaction of chromatin or by giving signals to other protein complexes [206]. Acetylated histones are frequently associated with active promoters; acetyltransferases modifications lead to the weaking of the interaction histone-DNA [205]. Histone acetyltransferases, that are associated with diverse protein complexes, transfer an acetyl group from acetyl-CoA to the ε-amino group of lysine side chains [205]. Histone methylation, that originally was thought to be irreversible, can occur on residues of arginine, lysine and histidine; the most extensively studied histone methylation sites include histone H3 lysine 4 (H3K4), H3K9, H3K27, H3K36, H3K79 and H4K20 [206]. In this mechanism, methyltransferases and demethylases which have a role in the addition and removal of methyl groups from different lysine residues on histones have been identified [206].

Non-coding RNAs such as long non-coding RNAs and microRNAs (miRNAs) are also involved in the epigenetic regulation of gene expression. These RNAs can be classified in two major groups: housekeeping non-coding RNAs and regulatory non-coding RNAs, in

which are included microRNAs and long non-coding RNAs (lncRNAs) [207]. MicroRNAs are abundant and originated from transcribed hairpin loop structures; these non-coding RNAs regulate gene expression in nucleus and cytoplasm by different mechanisms and mediate gene silencing [207]. LncRNAs are transcripts which, based on their location with respect to protein-coding genes can be divided into five subcategories: i) long intergenic ncRNAs, that are transcribed from both strands in intergenic regions; ii) long intronic ncRNAs, which are transcribed from introns of protein-coding genes; iii) sense lncRNAs that are transcribed from the sense strand and contain exons of protein-coding genes; iv) antisense lncRNAs which are transcribed from the antisense strand, and overlap with exonic or intronic regions and/or cover the protein-coding sequence; and v) bidirectional lncRNAs that are localized in the proximity of a coding transcript on the opposite strand [207]. It is of note that due to their regulatory effects on DNA sequences, lncRNA can be classified as *cis*-acting lncRNA and *trans*-acting lncRNAs; *cis*-lncRNAs regulate the expression of the genes that are close to these elements, while trans-lncRNAs regulates genes that are distant [207].

DNA methylation consists in the addition of a methyl group to the DNA molecule, commonly to cytosines in CpG dinucleotides [205]. This DNA modification is important for the regulation of tissue-specific expression, genomic imprinting and X chromosome inactivation [208]. Several factors can impact DNA methylation, leading to a different state (methylated/unmethylated), such as lifestyle and environmental factors. A modification in DNA methylation can contribute to a diversity of human diseases from single-gene diseases to cancer or sporadic metabolic diseases; this mechanism also showed to affect different neurodegenerative diseases such as HD [184]. Methylation varies with the aging mechanism [209]. increasing with aging. The methylation state of selected CpG sites emerged as a promising tool to estimate biological age. For that purpose, different "epigenetic clocks" were developed considering a specific number of CpG sites for each clock. Some clocks are more specific for blood samples, such as Hannum blood based epigenetic clock [210] while others can be used for different tissues such as Horvath Pan tissue epigenetic clock [190]. Globally these clocks allow the comparison between methylation age and chronological age, providing information of the presence of age acceleration or deceleration, of particular relevance in the context of disease.

1.4. Aims and pertinence of the thesis

In SCA3, as in the remaining polyQ diseases, genetic modifiers will help to explain the percentage of onset variance that is not accounted by the size of the CAG tract. The identification of SCA3 genetic modifiers will thus allow a better prediction of disease onset and contribute to the identification of altered molecular pathways, which might be relevant for the discovery of novel therapeutic targets. Also, as clinical trials are emergent in SCA3, modifier identification can improve patients' stratification and increase trial robustness; moreover, as soon as a specific intervention shows a disease-modifying effect, the administration of a compound with a therapeutic potential will present a higher efficiency if administrated prior but yet close to the disease onset, and therefore the ability to predict this point in time is of extreme relevance.

Although some progresses have been made in the quest for SCA3 modifiers, most studies have focused on coding variants, as the impact caused by the variation in these regions is more straightforward to interpret. However, candidate modifiers genes can also harbour relevant variation (i.e. variants that aid in the explanation of onset variance) in their untranslated regions, as described previously in this chapter.

Few studies concerning the variation in 5'UTR and 3'UTR of different genes and their influence in the disease onset have been performed for SCA3 patients. As these regions are important for the regulation of gene expression, their study is important to obtain more information in the context of SCA3. As the role of DNA methylation in neurodegenerative diseases is unveiled, the pertinence of its study in SCA3 is unquestionable; however, few studies have been performed so far and the methylation profile of SCA3 remains unexplored.

The main goal of this thesis is to analyse the impact of untranslated and methylation prone regions in the clinical variability of SCA3. For the concretization of this purpose, we studied the variation found in 3'UTR of *ATXN3* and analysed the impact of the variants found in the variance of the onset of SCA3 patients. We further studied the variation found in the UTRs of different genes that were identified through an extended exome sequencing and performed genotype-phenotype correlations to evaluate the impact of such variants/genes in the onset of SCA3 patients. In these two approaches, we aimed to identify genetic modifiers that can be of use for further studies and therapeutic approaches.

Finally, we performed an epigenome-wide pilot study using DNA samples from SCA3 patients and matched healthy controls to obtain the methylation profile of SCA3 and

to perform estimations of methylation age using the Horvath Pan Tissue Epigenetic Clock [190] and the Hannum Blood Based Epigenetic Clock [210]

This thesis is composed of five different chapters: an introductory chapter (Chapter 1) in which an overview of the genetics of polyglutamine diseases and more specifically of SCA3, is provided; also, general aspects of transcription regulation are described. In Chapter 2, the study of the variation in 3'UTR of *ATXN3* gene and of its impact on SCA3 onset is presented. Chapter 3 is dedicated to analyse the variants in the UTRs of different genes, obtained from the extended whole-exome sequencing data from a previous study, from Raposo and collaborators [160] and to evaluate their role as potential modifiers of SCA3. In Chapter 4, a pilot study to identify the blood-based methylation signature of SCA3 is presented. Finally, Chapter 5 summarizes the major conclusions of this thesis.

This study was approved by the Ethics Committee of the University of the Azores (Parecer 5/2018). All participants in this study provided written informed consent.

Chapter 2: Genetic variation found in *ATXN3* (ataxin-3) 3'UTR: understanding the downstream regulatory elements of the Spinocerebellar ataxia type 3/Machado-Joseph disease causative gene

2.1. Introduction

Five prime (5') and three prime (3') untranslated regions (UTRs) of genes, along with *cis*-acting elements (such as gene promoters), are essential for the regulation of gene expression [178,211]. Transcription factors (TFs) bind to the numerous binding sites (TFBSs) located in 5'UTRs and are responsible for regulating transcription initiation and stability [178,179]. 3'UTRs are also important regulatory regions which contain: a) targets for microRNAs (miRNAs) that can act by physically blocking the translation mechanism and by inducing the degradation of transcripts; b) motifs that are recognized by RNA binding proteins (RBPs) which are involved in the regulation of protein expression levels; and c) alternative polyadenylation (APA) signals, which can alter the length of 3'UTRs and consequently affect the binding of regulatory elements to those regions [180]. Disease associated phenotypes can be potentially modulated by the genetic variation that occurs in the UTRs of disease-causing genes, which can present an important impact on the regulation of their expression in an allele-specific manner [212]. Additionally, as the therapeutic modulation of mRNA expression has been emerging as a promising strategy for the treatment of many hereditary diseases [212], the understanding of genetic variation in these regions is crucial.

As previously referred (Chapter 1) SCA3 causative gene, *ATXN3*, is located at 14q32.12, and encodes for ataxin-3 [213]. *ATXN3* contains 13 exons [213,214] covering a genomic region of approximately 48 kilobases (kb). Its 3'UTR may have variable length, reflecting the alternative usage of exons 10 and 11 [65]. The $(CAG)_n$ repeat, encoding the polyQ tract, which is abnormally expanded in SCA3 mutation carriers/patients, is localized in exon 10. As mentioned in Chapter 1, normal *ATXN3* alleles contain between 12 and 44 CAG repeats, whereas expanded alleles display consensually from 52 to 86 CAG repeats [65]. The CAG tract size in the expanded allele incompletely explains variance on onset (45-70%) [215], the remaining explanation being provided by other factors, such as genetic modifiers [101,160,166,167,170,177,216] (Chapter 1). Genetic variation at 3'UTR of the causative gene of SCA3, and its potential impact on onset have only been investigated in a few studies [163]; however, none of such studies has completely covered the whole extension of this region. Previous studies on the interactions of endogenous miRNAs (particularly miR-25) with *ATXN3* 3'UTR, using animal and cell models, suggested that when specific miRNAs are overexpressed, this contributes to the reduction of mutant ataxin-3 protein levels, and to the decrease in protein aggregates [217,218]. To investigate this evidence, Long and

colleagues [163] studied two substitution variants (rs910369 and rs709930) located upstream and downstream a miR-25 binding site at the 3'UTR of *ATXN3* in a Chinese cohort of SCA3 patients. This study showed an association between rs709930 and the disease onset. Such findings, however, have not been replicated in other studies with independent cohorts [124]

The major aims of the present chapter were to describe genetic variation found at the *ATXN3* 3'UTR, to predict the functional effects of those variants, to explore the impact of 3'UTR variants on the SCA3 onset, and to build $(CAG)_n$ allele-specific haplotypes with the variants found [124].

2.2. Subjects and methods

Subjects

One hundred DNA samples (total cohort: 50 males and 50 females), from peripheral blood of Azorean SCA3 patients, molecularly confirmed as described by Bettencourt and colleagues [219] were used in this study. Data from an extended whole-exome sequencing was available for a sub-group of nineteen molecularly confirmed Azorean SCA3 patients (discovery sample) [160]. Onset was considered the age of the first manifestation of the disease (gait ataxia or diplopia), reported by each patient and/or caregiver. In the total cohort, the average onset was 37.37 ± 11.24 (mean ± standard deviation) and the average size of the CAG tract was 21.42 ± 4.59 for the normal allele, and 70.56 ± 3.63 for the expanded allele [124]. The scores of the Scale for the Assessment and Rating of Ataxia (SARA), and for the Neurological Examination Score for Spinocerebellar Ataxia (NESSCA) were available for a subset of patients (N=34 and N=65, respectively). For these patients the average scores were: 14.56 ± 8.00 and 12.25 ± 5.11, for SARA and NESSCA, respectively. For total cohort (N=100) the mean disease duration was 21.99 ± 9.76 years [124].

Genetic variation at 3'UTR of the ATXN3 gene

For the discovery sample (N=19), an extended whole-exome sequencing (WES) covering the UTRs was performed [160]. For this purpose, *Agilent SureSelect XT V5 + UTR* 75Mb capture kit was used and the sequencing was run in the Illumina *HiSeq4000* platform with a 100 bp paired end reads. Sequence alignment and variant calling were performed against the Reference Consortium Human Build 37 (GRCh37) using the Burrows-Wheeler Aligner (bwa-mem 0.7.17) [220], and the Genome Analysis Toolkit (GATK 3.7) [221]. A mean target coverage of 54.40X was obtained for 3'UTR and all substitution variants (from now on designated as variants) were covered at least 10X [124].

Variants found in *ATXN3* 3'UTR on the extended WES were subsequently validated by Sanger sequencing: variants were genotyped, after the alignment of the sequences with the sequence NM 004993.4. Genotypes for 81 additional samples (total of 100 samples – total cohort) were also obtained by Sanger sequencing. Primers sequences are available in Supplementary Table 2.1.

Hardy-Weinberg *equilibrium* tests with Bonferroni correction, *linkage disequilibrium* test between the identified variants, and allelic and genotypic frequencies were determined

using Arlequin software v.3.5 [222]. The polymorphism information content (PIC), which is a quantitative measure of the degree of polymorphism [223], was calculated for all the observed variants, using a codominant model in Gene-Calc tool (https://gene-calc.pl/pic) [223]. PIC values for co-dominant systems range between 0 (monomorphic, consequently, not informative) and 1 (very highly informative, presenting several alleles of equal frequency) [223].

Genotypic frequencies were used to performed population differentiation tests that compared the total cohort of Azorean patients with healthy individuals from European populations ancestry (HapMap-CEU and AFD_EUR_PANEL; frequencies assessed on November 17th, 2020 in dbSNP - https://www.ncbi.nlm.nih.gov/snp/) – European controls [124]. These comparisons were only performed for populations presenting an N>20. For the rs910369 and rs709930 variants, the genotypic frequencies calculated for the total cohort of Azorean patients were also compared with the ones reported by Long and collaborators for Chinese SCA3 patients and controls [163].

The allelic phase between each of the nine analysed variants and the normal/expanded *ATXN3* allele was inferred, whenever possible, by segregation analysis in the SCA3 families [124]. Haplotypes that were in *cis* with the expanded CAG repeat were directly assessed in the patients that were homozygous for all variants.

In silico predictions for the variants found in ATXN3 3'UTR

Functional effects of variants found in the *ATXN3* 3'UTR were predicted using Functional Analysis Through Hidden Markov Models (FATHMM) (http://fathmm.biocompute.org.uk/) [224]. Genome Wide Annotation of Variants (GWAVA) (https://www.sanger.ac.uk/sanger/StatGen_Gwava) [225], and Variant Effect Predictor – VEP (http://www.ensembl.org/Tools/VEP) were also performed. For FATHMM and GWAVA, a cut-off of 0.5 was considered, with variants showing scores higher than 0.5 being predicted to have a deleterious effect, which means that variants with a score higher than 0.5 may lead to a functional impact. Predicted *ATXN3* miRNA binding sites were identified using *miRmap* [226] and *Target Scan v.7.1* tools [227]. MirSNP (http://bioinfo.bjmu.edu.cn/mirsnp/search/) [228], was used to search for variants upstream and downstream (~200 bp) of miRNA binding sites, which can have a potential impact on

miRNA binding. In this analysis, only predicted miRNAs binding sites with a score > 150 and an energy score < -7 were considered [228].

The miRNAs that were previously identified and were predicted to bind to *ATXN3*, were compared to those found in databases, using the software Ingenuity Pathways Analysis (QIAGEN, IPA). To explore RBPs motifs, the tool *RBP-Var2* (http://www.rbp-var.biols.ac.cn/) [229] was applied with default settings. PolyASite 2.0 [230] was used, with default settings, to search for alternative polyadenylation (polyA) motifs generated or abolished by the *ATXN3* 3' UTR variants.

Impact of variants found in 3'UTR in phenotype of SCA3

An analysis of covariance (ANCOVA), using the number of CAGs in the expanded *ATXN3* allele as covariate, was performed to: a) evaluate the effect of presence/absence of the identified 3'UTR variants on SCA3 onset. Patients carrying the alternative allele were compared to patients homozygous for the reference allele; b) evaluate the effect of the presence of the alternative allele in *cis* with the expanded CAG repeat, in the subset of samples with known allelic phase; c) evaluate, in a subset of samples with NESSCA and SARA scores and subscores, the effect on onset of the presence of the alternative allele. NESSCA subscores included the following items: 1) gait ataxia; 2) limb ataxia; 3) dysphagia; 4) dysarthria; 5) fasciculation and 6) distal amyotrophies. For SARA, two different subscores were considered: a) axial subscore, constituted by the items sitting, stance and gait; and b) appendicular score, constituted by the items finger chase, nose finger and fast alternative movement.

Statistical analysis was performed using *IBM SPSS Statistics v.25.0*; *p* values lower than 0.05 were considered statistically significant**.**

2.3. Results

Genetic variation at the 3'UTR of the ATXN3 and their predicted functional impact

A total of nine variants - rs1055996, rs11628764, rs55966267, rs910369, rs709930, rs10151135, rs7158238, rs3092822 and rs7158733 (Figure 2.1) were found in the 3'UTR of *ATXN3* in the discovery sample. The nine variants had been previously described in dbSNP and were subsequently investigated in the total cohort of (N=100 SCA3 patients) [124]. Allelic and genotypic frequencies for the variants found, along with the functional predictions obtained, are shown in Table 2.1. For all variants, the frequency of alternative alleles was >5%; rs55966267 showed the lowest minor allele frequency (MAF=6%). PIC values ranged between 0.13 and 0.46 (Table 2.1). Only rs11628764, rs55966267 and rs10151135 showed PIC values lower than 0.40, suggesting these three variants would be less informative when trying to discriminate normal and expanded alleles in SCA3 patients [124]. Genotypic frequencies for eight out of the nine variants (rs1055996, rs11628764, rs910369, rs709930, rs10151135, rs7158238, rs3092822 and rs7158733) were in conformity with Hardy-Weinberg equilibrium expectation.

The genotypic frequencies for rs910369 and rs10151135 differed significantly between SCA3 patients (total cohort) and Europeans controls (Supplementary table 2.1). These two variants, which are not in LD with each other (D'=0.18, r^2=0, p=0.80; Supplementary table 2.2), present higher heterozygosity in patients than in European controls and consequently higher frequencies of their corresponding minor alleles.

For the variants previously studied by Long *et al.* [163], rs910369 and rs709930, no differences in genotypic frequencies between Azorean and Chinese SCA3 patients were observed. Nonetheless, there were significant differences in the genotypic frequencies between the Chinese SCA3 cohort and both Chinese and European controls (Supplementary table 2.3).

Table 2.1. General characterization of the *ATXN3* 3'UTR variants identified in the current study [124].

Variant	Allelic Frequency	Genotypic Frequency (PIC)	Functional effect FATHMM/GWAVA (Scores)*	MirSNP - miRNA binding sites (BS)**	RBP binding - RBP-Var2 scores***
rs1055996 (N=88) (NC_000014.8: g.92527399A>G)	A: 0.71 G: 0.29	AA: 0.56 AG: 0.31 GG: 0.13 (PIC = 0.41)	Benign or neutral (0.17/0.33)	No predicted modifications	1e
rs11628764 (N=90) (NC_000014.8: g.92527977T>C)	T: 0.87 C: 0.13	TT :0.74 TC :0.26 CC: 0.00 (PIC = 0.23)	Benign or neutral (0.17/0.43)	hsa-miR-374c-5p (New BS) hsa-miR-3688-3p (- binding) hsa-miR-4694-3p (- binding)	NA
rs55966267 (N=79) (NC_000014.8: g.92529558T>C)	T: 0.94 C: 0.06	TT: 0.91 TC: 0.05 CC: 0.04 (PIC = 0.13)	Benign or neutral/Deleterious (0.18/**0.59**)	hsa-miR-335-3p (+ binding)	6
rs910369 (N=100) (NC_000014.8: g.92530282C>A)	C: 0.66 A: 0.34	CC: 0.42 CA: 0.48 AA: 0.10 (PIC = 0.45)	Benign or neutral (0.09/0.29)	NA	1e
rs709930 (N=79) (NC_000014.8: g.92530513C>T)	C: 0.65 T: 0.35	CC: 0.42 CT: 0.47 TT: 0.11 (PIC = 0.45)	Benign or neutral (0.15/0.49)	hsa-miR-586 (New BS)	1e
rs10151135 (N=78) (NC_000014.8: g.92530551G>T)	G: 0.83 T: 0.17	GG: 0.67 GT: 0.30 TT: 0.03 (PIC = 0.28)	Deleterious (**0.80/0.51**)	No predicted modifications	1e
rs7158238 (N=100) (NC_000014.8: g.92537105C>T)	C: 0.64 T: 0.36	CC: 0.43 CT: 0.41 TT: 0.16 (PIC = 0.46)	Benign or neutral 0.07/0.15	No predicted modifications	4
rs3092822 (N=100) (NC_000014.8: g.92537163T>G)	T: 0.67 G: 0.33	TT: 0.43 TG: 0.45 GG: 0.12 (PIC = 0.45)	Benign or neutral (0.15/0.16)	NA	1e
rs7158733 (N=100) (NC_000014.8: g.92537223G>T)	G: 0.66 T: 0.34	GG: 0.46 GT: 0.40 TT: 0.14 (PIC = 0.45)	Benign or neutral (0.07/0.28) ◊	NA	1e

For all substitution variants, alleles are provided on the forward (or plus) genomic strand.

*Scores in bold highlight predicted deleterious effects.

**NA – no information available; + binding: enhanced miRNA affinity for the binding site; - binding: decreased miRNA affinity for the binding site.

***Scores vary between 1 and 6: a score of 1 (a,b,c,d and e) indicates a higher impact in the RNA structure and in the modification of the binding of RBPs; score 6 is the one with a lower impact in the modification of the binding site of RBPs.

◊ According to Variant Effect Predictor (VEP) this variant was predicted to originate a stop codon.

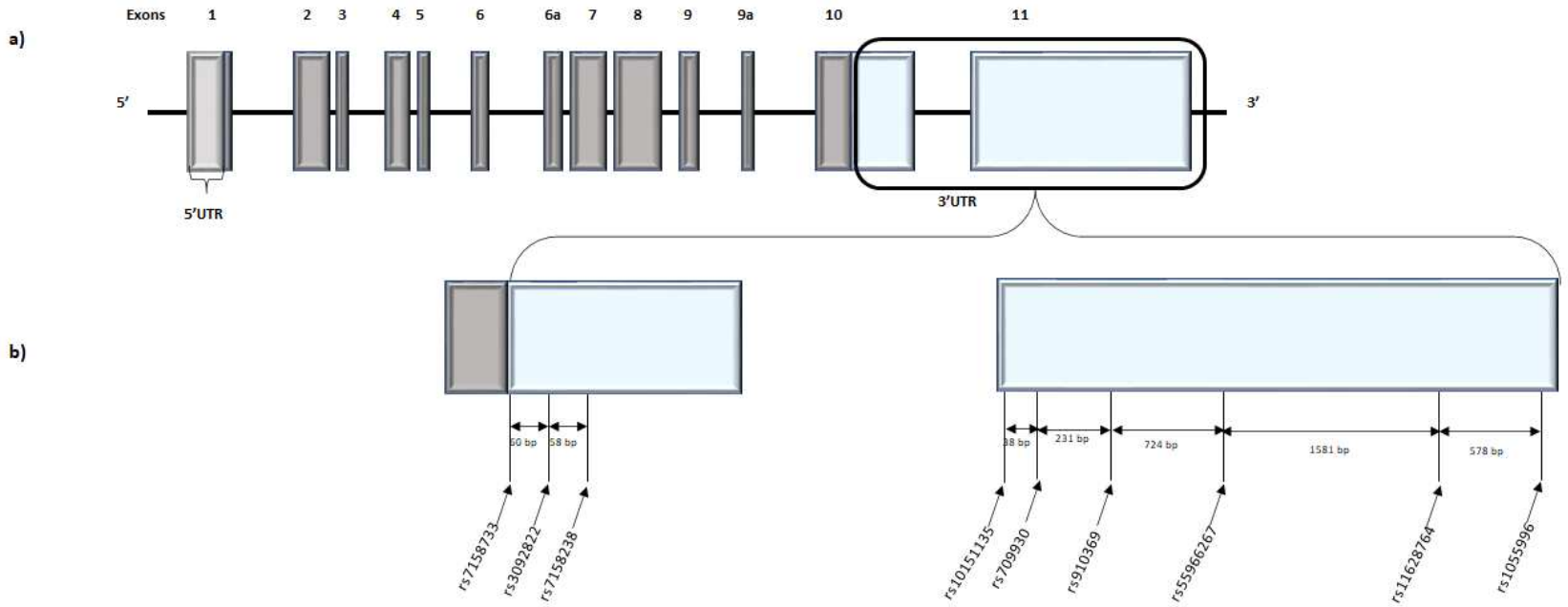


Figure 2.1. Representation of *ATXN3* gene showing the distance between the identified variants (in base pairs (bp)) in the 3'UTR of SCA3 patients. a) Exons (in medium gray) of *ATXN3* are numbered from 1 to 11. Untranslated regions (UTRs), 5'UTR and 3'UTR, are in light gray and in light blue pattern, respectively; b) Three of the nine variants (rs7158733, rs3092822 and rs7158238) are in exon 10 (initial portion of 3'UTR); the remaining six (rs10151135, rs709930, rs910369, rs55966267, rs11628764 and rs1055996) are in exon 11. Adapted from [124].

For a subset of samples, it was possible to construct haplotypes containing all nine variants analysed (Supplementary table 2.4). Ten different haplotypes were obtained and their absolute frequencies when in *cis* with the expanded or the normal allele are shown in Figure 3.2 a) and 3.2 b), respectively.

Six different haplotypes were identified as being in *cis* with the expanded allele; noteworthy, haplotype 10 (GTTATGTGT) was only observed in *cis* with the expanded allele. Furthermore, the frequency of haplotype 2 (ATTCCGCTG) in the expanded allele is 2-fold higher than in the normal allele. It is of note that nine haplotypes were associated with the normal allele; four of these were only present in the normal alleles: haplotype 4 (GCTCCTCTG), haplotype 6 (ACTCCTCGG), haplotype 8 (GTTATTTGT), and haplotype 9 (GTTCCGCTG). These results indicate that haplotype 10 might be relevant in distinguishing the expanded allele whereas haplotypes 2, 4, 6, 8 and 9 might be relevant in distinguishing the normal allele.

Concerning *in silico* predictions (Table 2.1), Variant Effect Prediction tool confirmed that variant rs7158733 leads to the formation of a premature stop codon, as previously reported [231], whereas a deleterious effect was predicted for rs10151135. None of the variants identified in the 3'UTR region of *ATXN3* is predicted to modify polyadenylation signals. A total of 1137 miRNA binding sites were identified in the *ATXN3* 3'UTR; MirSNP predicts that the presence of the variants rs11628764, rs55966267 and rs709930 can modify the affinity of miRNAs to a binding site (Table 2.1). For the variants rs11628764 and rs709930, a new miRNA binding site would be created with each variant. Except for hsa-miR-586, we were able to confirm the predictions obtained for the occurrence or not of the miRNAs binding sites using the software Ingenuity Pathways Analysis (QIAGEN, IPA).

a)

Haplotype number	rs1055996	rs11628764	rs55966267	rs910369	rs709930	rs10151135	rs7158238	rs3092822	rs7158733
1	A	T	T	C	T	G	C	T	G
2	A	T	T	C	C	G	C	T	G
3	A	T	T	C	C	T	C	T	G
4	G	C	T	C	C	T	C	T	G
5	A	T	C	C	C	G	C	T	G
6	A	C	T	C	C	T	C	G	G
7	A	T	T	A	T	G	T	G	T
8	G	T	T	A	T	T	T	G	T
9	G	T	T	C	C	G	C	T	G
10	G	T	T	A	T	G	T	G	T

b)

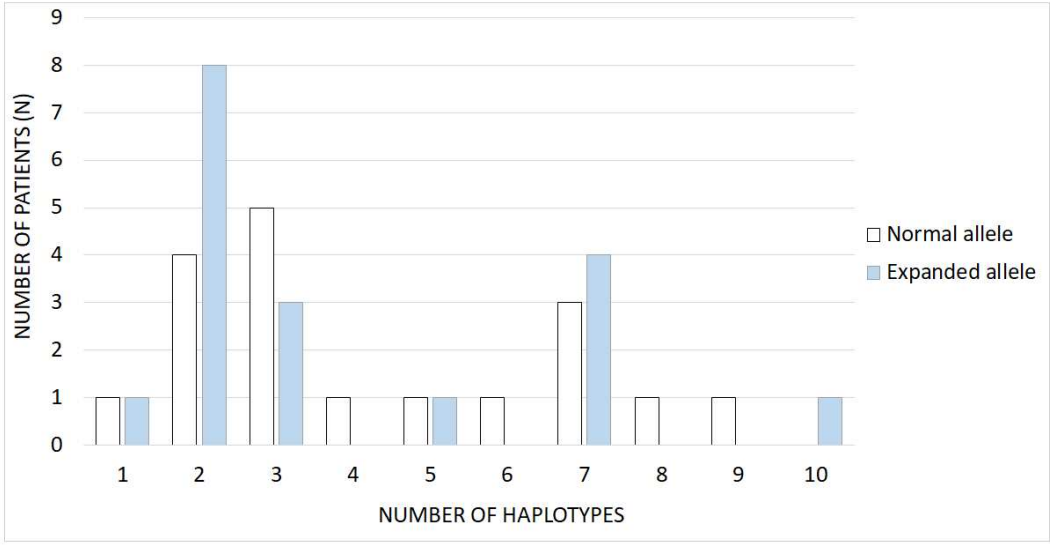


Figure 2.2. Representation of the comparison between normal and expanded alleles for the obtained haplotypes: a) haplotypes obtained for the subset of samples for which the haplotype for normal and expanded allele was complete; b) frequency (in absolute number) of haplotypes by normal and expanded allele, white and blue bars respectively (N=36 chromosomes). Adapted from [124]

.

Six out of the nine variants were predicted to affect the secondary structure of mRNA, the binding of RBPs and the expression of the RNA (score = 1e, Table 2.1). According with the same analysis, six variants map to target sites of RBPs, which are expected to alter the RNA structure (Table 2.2).

Table 2.2. RNA-binding proteins that were predicted to alter the RNA structure due to the presence of variation

RNA-binding protein	Short information	Variants
WD Repeat Domain 33 (WDR33)	Pre-mRNA 3' end processing protein that is essential for both cleavage and polyadenylation of pre-mRNA 3' ends [232]	rs910369 rs1055996
RNA-binding protein 47 (RBM47)	Presents a role in the regulation of cellular functions [233]	rs910369
RNA-binding protein FUS (FUS)	Plays a role in cellular processes such as transcription regulation, RNA splicing, RNA transport, DNA repair and damage response [234]	rs910369
Zinc Finger CCCH-Type Containing 7B (ZC3H7B),	Specific regulator of miRNA biogenesis [235]	rs910369
ELAV Like RNA Binding Protein 1(ELAVL1),	Binds to the 3'-UTR region of mRNAs and increases their stability [236],	rs910369 rs709930
Argonaute RISC Catalytic Component 2 (AGO2)	RNA-binding protein is needed for RNA-mediated gene silencing (RNAi) by the RNA-induced silencing complex [237]	rs910369 rs709930
Polypyrimidine tract binding protein 1 (PTBP1)	Has an important role in splicing [238]	rs910369 rs3092822 rs7158733
Eukaryotic Translation Initiation Factor 4A3 (eIF4AIII)	Has a role in pre-mRNA splicing as component of the spliceosome [239]	rs709930 rs10151135

Impact of the 3'UTR variants found in ATXN3 in phenotypic heterogeneity

For all variants, after adjusting for the size of expanded $(CAG)_n$ allele, no significant differences were observed on SCA3 onset between patients carrying the alternative allele (heterozygous or homozygous) and the reference allele (Table 2.3). As in heterozygous patients it is difficult to dissect allele-specific effects without knowing the allelic phase relative to the expanded $(CAG)_n$ allele, this comparison was further restricted to patients with known allelic phase which was possible to determine individually for six variants. No significant effects can be observed, in this smaller set of patients (N=47 to N=88, depending on which variant). These results can be a consequence of the limited sample size of this subset of samples.

Table 2.3. Onset (in years) adjusted for the size of expanded $(CAG)_n$ alleles observed in variants with known allelic phase [124].

Variants ID	Patients carrying the reference allele in *cis* with the expanded $(CAG)_n$ allele		Patients carrying the alternative allele in *cis* with the expanded $(CAG)_n$ allele		*p* value
	N	Adj. mean onset	N	Adj. mean onset	
rs1055996	49	36.68 ± 0.83	39	36.25 ± 1.04	*0.76*
rs910369	42	35.15 ± 1.05	18	35.31 ± 1.61	*0.94*
rs709930	33	35.38 ± 1.21	14	34.76 ± 1.86	*0.78*
rs7158238	42	37.26 ± 1.09	19	36.64 ± 1.64	*0.76*
rs3092822	45	36.71 ± 1.09	17	37.14 ± 1.79	*0.84*
rs7158733	46	35.94 ± 1.09	16	37.85 ± 1.88	*0.61*

Adj. mean onset is presented as mean ± standard error. Three variants (rs10151135, rs11628764 and rs55966267) were excluded from the analysis due to the low frequency of the minor allele.

Regarding rs709930, which was previously described by Long and colleagues [163] as being associated with SCA3 onset in Chinese patients, we were not able to replicate the observation of an earlier onset in patients that carried the alternative allele, as previously described [163].

Noteworthy, the analysis exploring the association between the variants identified and NESSCA and SARA scores and subscores revealed that for rs10151135 patients carrying the alternative allele presented a higher SARA Axial subscore than patients homozygous for the reference allele (p = 0.023).

2.4. Discussion

This study reports nine variants present in the 3' UTR of SCA3 causative gene - *ATXN3*, which can be involved in the regulation of its own expression regulation and may help to discriminate between *ATXN3* normal and expanded alleles.

Predicted functional impacts for the studied variants revealed: i) two new miRNA binding sites, which may have implications in modulation of gene expression; ii) changes in the secondary structure of mRNA, the binding of RBPs and RNA expression; and iii) the creation of a premature stop codon that leads to a short form of the ATXN3 protein, as previously described in other studies (see amongst others, [240]. These predictions confirm the complexity and potential importance of the 3'UTR regulation of *ATXN3* expression [124]. As previously mentioned, the presence of a variant can affect the binding of a miRNA leading to a change in the normal regulation of protein expression [180]. Previous studies in SCA3 revealed that the overexpression of several miRNAs (miR-9, miR-25 miR-181a and miR-494) contributed to the reduction of the levels of the ataxin-3 protein and to a decrease in aggregates of this protein [217,218]. In another study, miR-25 as well as miR-29a, miR-34b, miR-125b and miR-7014 were shown to have a potential as biomarkers of disease severity, of potential use in future studies about the pathogenesis of the disease as well as in drug development efforts [241,242]. In this study, we were able to confirm the results obtained for the miRNAs binding sites using the software Ingenuity Pathways Analysis (QIAGEN, IPA), and only one of the miRNAs was not confirmed in this analysis: hsa-miR-586. As this miRNA binding site is predicted to be created by the tool that was used, two reasons can be pointed out: i) it can be a false positive result, or ii) databases do not have information for this miRNA. None of the miRNAs predicted to target binding sites, were altered by the presence of the variants identified in the present study, have been previously reported [217,218,231,241]; although *in silico* results are consistent across prediction tools, functional studies concerning the miRNAs aspects should be performed.

A significant modulatory effect of the alternative allele at rs10151135 on SARA Axial subscore, was identified. Noteworthy, this variant is predicted to have a deleterious effect according to several prediction tools, as referred previously during this chapter, and therefore its functional effect deserves further investigation.

Despite possible functional changes/deleterious effects due to variants in 3'UTR, in the cohort studied none of the identified variants had significant modulatory effects on the

disease onset. This can be due to sample size limitations, to the low frequency of alleles with predicted deleterious effects (e.g. rs55966267 MAF = 0.06), and/or the small magnitude of effects for these variants. Also, the limited sample size of the discovery cohort can influence the results, as the remaining 81 independent samples can have additional variants that were not investigated and that can affect the onset. For rs709930, previously associated to SCA3 onset [163], we were not able to replicate the result of an earlier onset in patients whose alternative allele was in *cis* with expanded CAG. Noteworthy, in the study by Long and colleagues [163], 67.6% of the patients carried the alternative allele for rs709930 compared to 29.8% in our study. Such population-specific differences in allelic frequencies may have facilitated the detection of an effect in the Chinese SCA3 cohort as we previously reported for other modifier genes on SCA3 onset [170].

Costa and colleagues [242] highlight the importance, in the context of the development of allele-specific therapies, of describing disease-associated haplotypes that can discriminate between normal and expanded alleles. The design of therapies targeting allele-specific inhibition (affecting the expanded allele only) is based on the assumption that expanded alleles can be discriminated from their wild-type counterparts, offering advantages in relation to the non-allele-specific silencing strategy previously applied in a rat model of SCA3 [243]; an example can be the development of small interfering RNAs, which, based on the presence of a specific variant or haplotype, can discriminate between normal allele and expanded allele and silence/modulate the expression of mutant *ATXN3* [240]. In this work we were able to describe haplotypes exclusively found in *cis* with the expanded (N=1) or with the normal allele (N=4), and that should be of discriminative utility; we further identified a haplotype whose frequency is markedly higher in *cis* with the expanded allele. These findings, which are highly relevant for the design of allele-specific therapies, warrant further investigation in independent cohorts of SCA3 patients.

2.5. Conclusion

In summary, some of the variants found in this study are predicted to have functional impact on the expression of *ATXN3*. Different factors can justify our inability to find significant associations between the presence of these variants and onset for the disease, namely limitations on sample size. The access to additional cohorts of SCA3 patients should allow us to confirm this hypothesis. We were also able to identify disease-associated haplotypes with discriminative ability for the expanded allele, which can be of relevance for future design of allele-specific silencing strategies [124].

2.6. Supplementary materials

Supplementary table 2.1. Primers sequences used for the amplification of different fragments of the *ATXN3* 3'UTR.

Name	Sequence
MJD-10F	5'- AGATTTCCTAAGATCAGCACTTCC- 3'
MJD-10R	5'- CTTCATTAAAGCATACCTACAGG - 3'
MJD-11F	5'- AGGAAATGGGGATCGACCACAAAG - 3'
MJD-11R	5'- CACAAACACATTCAAACGCATCCAG - 3'
ATXN3_ 1F	5'- CAGGACAGAGTTCACATCCAT- 3'
ATXN3_1R	5'- TCTAGCAATCCTCTCCTGCC - 3'
ATXN3_2F	5'- CCCAACAGGTGATGCTATGAG - 3'
ATXN3_2R	5'- TGTGCAGCCACACACCAGGA - 3'
rs11_10_F	5'- AGCTGAGCACACACTGGATG - 3'
rs11_10_R	5' -GGTCCAATGGGCCGGTTTTA - 3'

Supplementary table 2.2. Genotypic frequencies available for five out of the nine variants found in *ATXN3* 3'UTR in Azorean SCA3 patients in comparison to those observed in European controls; for variants rs910369 and rs709930, previously analysed by Long *et al.* (2015), the reported (dbSNP Population) differences in genotypic frequencies between Chinese SCA3 patients and European controls are also shown [124].

					***p* value**			
	Azorean SCA3 patients	**European controls***	**SCA3 patients (Long *et al.*, 2015) ****	**Chinese controls (Long *et al.*,2015) ****	**Azorean SCA3 patients *vs* European controls**	**Chinese SCA3 patients *vs* Chinese controls**	**Azorean SCA3 patients *vs* Chinese MJD patients**	**Chinese SCA3 patients *vs* European controls***
rs910369	CC: 0.42 CA : 0.48 AA : 0.10 (N = 100)	CC : 0.60 CA : 0.33 AA : 0.07 (N = 111)	CC : 0.32 CT : 0.49 TT : 0.19 (N = 170)	CC : 0.45 CT : 0.48 TT : 0.07 (N = 200)	***p* = 0.038**	***p* < 0.001**	*p* = 0.079	***p* < 0.001**
rs709930	CC : 0.42 CT : 0.47 TT : 0.11 (N =79)	CC : 0.60 CT : 0.33 TT : 0.07 (N = 112)	CC : 0.32 CT : 0.49 TT : 0.19 (N = 170)	CC : 0.45 CT : 0.48 TT : 0.07 (N = 200)	*p* = 0.053	***p* < 0.001**	*p* = 0.191	***p* < 0.001**
rs10151135	GG : 0.67 GT : 0.30 TT : 0.03 (N = 78)	GG : 0.85 GT : 0.15 TT : 0.00 (N = 111)	NA	NA	***p* = 0.007**	NA	NA	NA
rs3092822	TT : 0.43 TG : 0.45 GG : 0.12 (N = 100)	TT : 0.33 TG : 0.46 GG : 0.21 (N =24)	NA	NA	*p* = 0.469	NA	NA	NA
rs7158733	GG : 0.46 GT : 0.44 TT : 0.14 (N = 100)	GG : 0.33 GT : 0.46 TT : 0.21 (N = 24)	NA	NA	*p* = 0.435	NA	NA	NA

NA – not available; *P*-values in bold indicates a statistically significant result.

*Frequencies from dbSNP (HapMap-CEU for rs910369, rs709930 and rs10151135, and AFD_EUR_PANEL for rs3092822 and rs7158733; assessed on November 17th,2020).

**Frequencies from Long *et al.* (2015); rs910369 and rs709930 found in complete LD

Supplementary table 2.3. *Linkage disequilibrium* parameters, D' and r^2, for the variants in the present study (N =100 SCA3 patients) [124].

	rs1055996	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
rs1055996		D' = 0.006 r^2 = 0.00 *p = 0.94*	D' = 1 r^2 = 0.01 *p = 0.48*	D' = 0.80 r^2 = 0.62 ***p < 0.01***	D' = 0.80 r^2 = 0.50 ***p < 0.01***	D' = 0.22 r^2 = 0.00 *p = 0.65*	D' = 0.83 r^2 = 0.52 ***p < 0.01***	D' = 0.84 r^2 = 0.60 ***p < 0.01***	D' = 0.84 r^2 = 0.64 ***p < 0.01***
rs11628764	D' = 0.006 r2 = 0.00 p = 0.94		D' = 1 r^2 = 0.00 *p = 0.78*	D' = 0.11 r^2 = 0.00 *p = 0.88*	D' = 1 r^2 = 0.02 *p = 0.24*	D' = 0.26 r^2 = 0.02 *p = 0.15*	D' = 1 r^2 = 0.02 *p = 0.13*	D' = 0.17 r^2 = 0.00 *p = 0.82*	D' = 1 r^2 = 0.02 *p = 0.18*
rs55966267	D' = 01 r^2 = 0.01 *p = 0.48*	D' = 1 r^2 = 0.00 *p = 0.78*		D' = 1 r^2 = 0.01 *p = 0.29*	D' = 1 r^2 = 0.01 *p = 0.42*	D' = 1 r^2 = 0.00 *p = 0.53*	D' = 1 r^2 = 0.01 *p = 0.36*	D' = 1 r^2 = 0.01 *p = 0.40*	D' = 1 r^2 = 0.01 *p = 0.40*
rs910369	D' = 0.80 r^2 = 0.62 ***p < 0.01***	D' = 0.11 r^2 = 0.00 *p = 0.88*	D' = 1 r^2 = 0.01 *p = 0.29*		D' = 0.88 r^2 = 0.78 ***p < 0.01***	D' = 0.18 r^2 = 0.00 *p = 0.80*	D' = 0.95 r^2 = 0.86 ***p < 0.01***	D' = 0.90 r^2 = 0.78 ***p < 0.01***	D' = 1 r^2 = 0.87 ***p < 0.01***
rs709930	D' = 0.80 r^2 = 0.50 ***p < 0.01***	D' = 1 r^2 = 0.02 *p = 0.24*	D' = 1 r^2 = 0.01 *p = 0.42*	D' = 0.88 r^2 = 0.78 ***p < 0.01***		D' = 0.25 r^2 = 0.00 *p = 0.61*	D' = 0.88 r^2 = 0.73 ***p < 0.01***	D' = 0.82 r^2 = 0.64 ***p < 0.01***	D' = 0.88 r^2 = 0.73 ***p < 0.01***
rs10151135	D' = 0.22 r^2 = 0.00 *p = 0.65*	D' = 0.26 r^2 = 0.02 *p = 0.15*	D' = 1 r^2 = 0.00 *p = 0.53*	D' = 0.18 r^2 = 0.00 *p = 0.80*	D' = 0.25 r^2 = 0.00 *p = 0.61*		D' = 0.27 r^2 = 0.00 *p = 0.57*	D' = 0.01 r^2 = 0.00 *p = 0.98*	D' = 0.27 r^2 = 0.00 *p = 0.57*
rs7158238	D' = 0.83 r^2 = 0.52 ***p < 0.01***	D' = 1 r^2 = 0.02 *p = 0.13*	D' = 1 r^2 = 0.01 *p = 0.36*	D' = 0.95 r^2 = 0.86 ***p < 0.01***	D' = 0.88 r^2 = 0.73 ***p < 0.01***	D' = 0.27 r^2 = 0.00 *p = 0.57*		D' = 0.91 r^2 = 0.70 ***p < 0.01***	D' = 1 r^2 = 0.86 ***p < 0.01***
rs3092822	D' = 0.84 r^2 = 0.60 ***p < 0.01***	D' = 0.17 r^2 = 0.00 *p = 0.82*	D' = 1 r^2 = 0.01 *p = 0.40*	D' = 0.90 r^2 = 0.78 ***p < 0.01***	D' = 0.82 r^2 = 0.64 ***p < 0.01***	D' = 0.01 r^2 = 0.00 *p = 0.98*	D' = 0.91 r^2 = 0.70 ***p < 0.01***		D' = 0.92 r^2 = 0.84 ***p < 0.01***
rs7158733	D' = 0.84 r^2 = 0.64 ***p < 0.01***	D' = 1 r^2 = 0.02 *p = 0.18*	D' = 1 r^2 = 0.01 *p = 0.40*	D' = 1 r^2 = 0.87 ***p < 0.01***	D' = 0.88 r^2 = 0.73 ***p < 0.01***	D' = 0.27 r^2 = 0.00 *p = 0.57*	D' = 1 r^2 = 0.86 ***p < 0.01***	D' = 0.92 r^2 = 0.84 ***p < 0.01***	

Supplementary table 2.4. Haplotypes obtained for the subset of samples with known allelic phase: - represents no info about the sequence; *represents estimated haplotypes and ? represents heterozygous samples which was not possible to obtain the allele in *cis* with the expanded allele [124].

Parent		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAGs (Norm/Exp)	Haplotype								
No info	No info	No info								
No info	No info	No info								
Patient 2	20	G*	T	-	A*	T*	-	T*	G*	T*
	71	A*	T	T*	C*	C*	G*	C*	T*	G*
Patient 2	20	G*	T	-	A*	T*	-	T*	G*	T*
	71	A*	T	T*	C*	C*	G*	C*	T*	G*
Patient 2	20	G*	T	-	A*	T*	-	T*	G*	T*
	71	A*	T	T*	C*	C*	G*	C*	T*	G*
Patient 2	20	G*	T	-	A*	T*	-	T*	G*	T*
	71	A*	T	T*	C*	C*	G*	C*	T*	G*
No info	No info	No info								
Patient 7	22	G*	C*	T	C	C	T	C	T	G
	71	A*	T*	T	C	C	T	C	T	G
No info	No info	No info								

Descendant		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAG (Norm/Exp)	Haplotype								
Patient 1	14	A	T	T	C	T	G	C	T	G
	68	A	T	T	C	T	G	C	T	G
Patient 2	20	G*	T	-	A*	T*	-	T*	G*	T*
	71	A*	T	T*	C*	C*	G*	C*	T*	G*
Patient 3	22	-	-	T	C	C	G	C	T	G
	71	A*	T*	T	C	C	G	C	T	G
Patient 4	14	A	T	-	C	-	G	C	T	G
	70	A	T	T*	C	C*	G	C	T	G
Patient 5	14	A	T	T	C	C	T*	C	T	G
	73	A	T	T	C	C	G*	C	T	G
Patient 6	14	A	T	T	C	C	G	C	T	G
	72	A	T	T	C	C	G	C	T	G
Patient 7	22	G*	C*	T	C	C	T	C	T	G
	71	A*	T*	T	C	C	T	C	T	G
Patient 8	25	A	T	T	C	C	T	C	T	G
	75	A	T	T	C	C	T	C	T	G
Patient 9	24	-	T	T	C	C	?	C	T	G
	68	A*	T	T	C	C	?	T*	T	G

Supplementary table 2.4. Haplotypes obtained for the subset of samples with known allelic phase: - represents no info about the sequence; *represents estimated haplotypes and ? represents heterozygous samples which was not possible to obtain the allele in *cis* with the expanded allele [124] (cont.)

Parent		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**	**Descendant**		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAGs (Norm/Exp)	Haplotype									ID	CAG (Norm/Exp)	Haplotype								
Patient 9	24	-	T	T	C	C	?	C	T	G	Patient 10	14	A	T	T	A	T	-	T	T	G
	68	A*	T	T	C	C	?	T*	T	G		69	A	T	T	C*	C*	-	T	T	G
No info	No info	No info									Patient 11	22	A	T	C	C	C	G	C	T	G
												71	A	T	C	C	C	G	C	T	G
Patient 11	22	A	T	C	C	C	G	C	T	G	Patient 12	22	A	T	-	A	-	-	C	T	G
	71	A	T	C	C	C	G	C	T	G		69	A	T	C*	C*	C*	G*	C	T	G
No info	No info	No info									Patient 13	23	A	C*	T	C	C	T*	C	G	G
												76	A	T*	T	C	C	G*	C	T*	G
Patient 13	23	A	C*	T	C	C	T*	C	G	G	Patient 14	28	A	T	T	C	C	G	C	T	G
	76	A	T*	T	C	C	G*	C	T*	G		78	A	T	T	C	C	G	C	T	G
No info	No info	No info									Patient 15	23	A	T	T	C	-	-	C	T	G
												73	A	T	T	C	C*	G*	C	T	G
Patient 15	23	A	T	T	C	-	-	C	T	G	Patient 16	22	A	C	-	A	-	-	C	T	G
	73	A	T	T	C	C*	G*	C	T	G		72	A	T*	T*	C*	C*	G*	C	T	G
Patient 15	23	A	T	T	C	-	-	C	T	G	Patient 17	23	A	T	T	C	C	G	C	T	G
	73	A	T	T	C	C*	G*	C	T	G		75	A	T	T	C	C	G	C	T	G
No info	No info	No info									Patient 18	23	A	?	T	C	C	G	C	T	G
												71	A	?	T	C	C	G	C	T	G

Supplementary table 2.4. Haplotypes obtained for the subset of samples with known allelic phase: - represents no info about the sequence; *represents estimated haplotypes and ? represents heterozygous samples which was not possible to obtain the allele in *cis* with the expanded allele [124] (cont.)

Parent		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAGs (Norm/Exp)	Haplotype								
No info	No info	No info								
No info	No info	No info								
No info	No info	No info								
Patient 21	14	A	T	-	C	-	-	C	T	G
	67	A	T	-	C	C*	G*	C	T	G
No info	No info	No info								
No info	No info	No info								
Patient 24	14	A	T	T	C	C	T*	C	T	G
	68	A	T	T	C	C	G*	C	T	G
Patient 24	14	A	T	T	C	C	T*	C	T	G
	68	A	T	T	C	C	G*	C	T	G
Patient 24	14	A	T	T	C	C	T*	C	T	G
	68	A	T	T	C	C	G*	C	T	G

Descendant		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAG (Norm/Exp)	Haplotype								
Patient 19	27	A	T	T	C	C	T	C	T	G
	69	A	T	T	C	C	T	C	T	G
Patient 20	27	A	T	-	C	-	G	C	T	G
	67	A	T	-	C	-	G	C	T	G
Patient 21	14	A	T	-	C	-	-	C	T	G
	67	A	T	-	C	C*	G*	C	T	G
Patient 22	27	A	C	-	C	C	G	C	T	G
	69	A	T*	-	C	C	G	C	T	G
Patient 23	14	A	T	T	?	?	?	?	?	?
	66	A	T	T	?	?	?	?	?	?
Patient 24	14	A	T	T	C	C	T*	C	T	G
	68	A	T	T	C	C	G*	C	T	G
Patient 25	26	A	T	T	A	T	G	T	G	T
	72	A	T	T	C*	C*	G	C*	T*	G*
Patient 26	27	-	-	T	C	C	G	C	T	G
	66	A*	T*	T	C	C	G	C	T	G
Patient 27	27	G	T	T	A	T	T	T	G	T
	71	A*	T	T	C*	C*	G*	C*	T*	G*

Supplementary table 2.4. Haplotypes obtained for the subset of samples with known allelic phase: - represents no info about the sequence; *represents estimated haplotypes and ? represents heterozygous samples which was not possible to obtain the allele in *cis* with the expanded allele [124] (cont.)

Parent		rs1055996	rs11628764	rs55966267	rs910369	rs709930	rs10151135	rs7158238	rs3092822	rs7158733	Descendant		rs1055996	rs11628764	rs55966267	rs910369	rs709930	rs10151135	rs7158238	rs3092822	rs7158733
ID	CAGs (Norm/Exp)	Haplotype									ID	CAG (Norm/Exp)	Haplotype								
No info	No info	No info									Patient 28	14	?	T	T	?	?	G	?	?	?
												63	?	T	T	?	?	G	?	?	?
No info	No info	No info									Patient 29	21	?	T	T	?	?	G	?	?	?
												68	?	T	T	?	?	G	?	?	?
No info	No info	No info									Patient 30	14	A	T	T	C*	C*	G	C*	T*	G*
												74	A	T	T	A*	T*	G	T*	G*	T*
Patient 30	14	A	T	T	C*	C*	G	C*	T*	G*	Patient 31	21	A	T	T	A	T	G	T	G	T
	74	A	T	T	A*	T*	G	T*	G*	T*		74	A	T	T	A	T	G	T	G	T
Patient 30	14	A	T	T	C*	C*	G	C*	T*	G*	Patient 32	14	G	T	T	C*	C*	G	C*	T*	G*
	74	A	T	T	A*	T*	G	T*	G*	T*		74	A*	T	T	A*	T*	G	T*	G*	T*
Patient 30	14	A	T	T	C*	C*	G	C*	T*	G*	Patient 33	21	A	T	T	A	T	G	T	G	T
	74	A	T	T	A*	T*	G	T*	G*	T*		77	A	T	T	A	T	G	T	G	T
No info	No info	No info									Patient 34	14	A*	T	T	C*	C*	T*	C*	T*	G*
												70	G*	T	T	A*	T*	G*	T*	G*	T*
Patient 34	14	A*	T	T	C*	C*	T*	C*	T*	G*	Patient 35	20	G	T	-	A	T	G	T	G	T
	70	G*	T	T	A*	T*	G*	T*	G*	T*		71	G	T	T*	A	T	G	T	G	T

Supplementary table 2.4. Haplotypes obtained for the subset of samples with known allelic phase: - represents no info about the sequence; *represents estimated haplotypes and ? represents heterozygous samples which was not possible to obtain the allele in *cis* with the expanded allele [124] (cont.).

Parent		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAGs (Norm/Exp)	Haplotype								
No info	No info	No info								
No info	No info	No info								
No info	No info	No info								
No info	No info	No info								
Patient 39	27	-	-	-	?	-	-	T	?	T
	65	A*	-	-	?	-	-	T	?	T

Descendant		**rs1055996**	**rs11628764**	**rs55966267**	**rs910369**	**rs709930**	**rs10151135**	**rs7158238**	**rs3092822**	**rs7158733**
ID	CAG (Norm/Exp)	Haplotype								
Patient 36	23	A	?	-	?	-	-	?	?	?
	70	A	?	-	?	-	-	?	?	?
Patient 37	14	?	T	T	C	C	G	?	?	?
	73	?	T	T	C	C	G	?	?	?
Patient 38	23	A	?	T	?	?	G	?	?	?
	71	A	?	T	?	?	G	?	?	?
Patient 39	27	-	-	-	?	-	-	T	?	T
	65	A*	-	-	?	-	-	T	?	T
Patient 40	23	A	?	-	?	-	-	C	?	G*
	72	A	?	-	?	-	-	T*	?	T*

Supplementary table 2.5. Adjusted onset (in years) in SCA3 patients carrying the alternative allele of analysed variants (homozygous for the alternative allele and heterozygous) *versus* patients homozygous for the reference allele [124].

Variant ID	Carriers of the reference allele		Carriers of the alternative allele		*p* value
	N	Adj. mean onset	N	Adj. mean onset	
rs1055996	49	37.54 ± 1.02	39	37.02 ± 1.13	*0.94*
rs11628764	67	37.52 ± 0.86	23	36.08 ± 1.46	*0.59*
rs55966267	72	37.83 ± 0.82	7	37.71 ± 2.64	*0.35*
rs910369	42	36.53 ± 1.09	58	37.84 ± 0.92	*0.36*
rs709930	33	37.38 ± 1.22	46	38.21 ± 1.02	*0.24*
rs10151135	53	37.15 ± 0.95	25	39.09 ± 1.39	*0.23*
rs7158238	43	36.30 ± 1.06	57	38.03 ± 0.92	*0.22*
rs3092822	45	36.38 ± 1.04	55	38.04 ± 0.94	*0.24*
rs7158733	46	36.24 ± 1.02	54	38.18 ± 0.95	*0.17*

Onset is adjusted to the number of CAG repeats; Adj. mean onset is presented as mean ± standard error.

Chapter 3: Extended whole-exome sequencing reveals variants in untranslated regions of modifiers of Spinocerebellar ataxia type3 (SCA3) age at onset

3.1. Introduction

Variable expressivity, which is considered a hallmark of polyQ diseases [244] can be manifested in the different degrees of severity displayed by individual patients sharing a causative polyQ mutation, by distinct progression rates or by different ages at onset (onset). An important part of this heterogeneity is explained by the size of the $(CAG)_n$ tract size; this explanation varies from a minimum of 26% to a maximum of 88%, depending on the polyQ disease (for more details see Chapter 1 and [4]). The number of CAG repeats, thus, only partially correlates with onset, indicating that other genetic factors besides the CAG tract, as well as yet largely unstudied environmental factors likely have an important influence in the expressivity of polyQ diseases [4,245].

Identifying genetic modifiers of polyQ diseases, namely of SCA3, remains an important goal; in fact, as SCA3 remains without treatment, the identification of disease-modifying genes and their corresponding pathways may provide insights into novel molecular targets for drug development [160]. Moreover, as already stated in Chapter 1, the identification of such modifiers should further allow a better prediction of onset, of utility not only in the context of genetic counseling but also when attempting to stratify patients for clinical trials.

Genetic modifiers, defined as genes that contribute to the modulation of a phenotype whose presence is primarily determined by a major causative gene (Chapter 1, [158], have been extensively searched in SCA3 [50,101,124,160,161,163,165,166,170,171,173,174,176,177,185,191–193,246–251] as detailed in Chapter 1. As shown by different studies [245,252,253], individual genetic modifiers are part of complex systems of genes that can cause small to moderate phenotypic alterations. So far, the search for genetic modifiers of SCA3, performed using candidate approaches as well as, most recently, unbiased genome-wide strategies (for a revision see Chapter 1), has been querying coding variants in specific genes whose presence/absence can, in addition to the number of CAG repeats, increase the explanation of onset variance (see amongst others [101,160,170,192,193]). Little is known, however, about the potential that non-coding variants in regulatory regions may have as modifiers [124,163,250]. For the purpose of this study, we will only consider untranslated regions (UTRs).

It is acknowledged that the variation found in UTRs of disease-causing genes can significantly impact the regulation of their own expression and have the potential to modulate disease phenotype. Likewise, variation in the UTRs of other genes can impact the

phenotype under the dependence of the primary gene and thus such variation needs to be considered when trying to identify genetic modifiers.

UTRs can also have a similar impact as coding sequences by modifying the original function of a gene (revised in [254]). However, there still a lack of information about UTRs sequences and its variation, despite a rising number of studies concerning these regions (revised in [254]).

Studies targeting UTRs in an attempt to identify modifiers of SCA3 are limited and focused on the variation found in *ATXN3*, the disease causative gene (see [124,163,250]; Chapter 1). Associations between variants in 5´and 3´UTR of candidate genes and onset in SCA3 have, so far, been scarce. Bettencourt and colleagues [166] as well as Mergener and collaborators [193], reported that a variant (rs3512) found in UTRs of *FAN1* (FANCD2 And FANCI Associated Nuclease 1), a gene implicated in the DNA repair process, modulated onset in SCA3 and other SCAs [166,193]. The role of inflammation in SCA3 has been investigated either in brain as in blood [255–257]. With this hypothesis in mind, a study conducted by Raposo and colleagues [174] reported that variants found in the promoter region of *IL6* (Interleukin 6) support modifications in cytokines expression and modulate onset [174]. In diseases where accumulation of the mutant protein is a hallmark, as SCA3, autophagy is a key mechanism. Following this idea, Kazachkova and colleagues [176] have investigated the variation at the promoter of *BECN1* (Beclin 1). *BECN1* is a gene related with autophagy that encodes for beclin-1; this protein presents a decrease in its levels in affected brain areas of animals models from different neurological diseases, such as SCA3 [176]. This study confirmed that the promotor of *BECN1* is not conserved. Also, *BECN1* expression levels presented a tendency for a decrease in samples that presented rs60221525 C allele , and for an increase in samples with the rs116943570 T allele ; SCA3 patients with rs60221525 C allele presented a tendency for an earlier estimated onset [176]. The finding that a variant in the promoter of the *BECN1* gene influences onset needs to be confirmed in a larger cohort [176].

Recently, as referred in Chapter 2 a whole-exome sequencing (WES) based search of genetic modifiers of SCA3 was performed in a discovery sample of pairs of first-degree relatives SCA3 patients, which were onset concordant and onset discordant [160]. Several coding variants whose genotypes were shared by both members of each concordant pair but differed between members of each discordant pair were identified, and were tested for associations with onset, leading to the proposal of nine candidate modifier genes [160]. In

addition to the variants described by Raposo and collaborators [160], variants in 5' and 3'UTRs of several genes were also retrieved; the impact of those variants remained to be studied.

In this chapter, we analysed the variants in the UTRs of several genes, obtained from filtering the extended WES data from the study of Raposo and colleagues [160], to evaluate their putative role as modifiers of SCA3. In order to identify potential onset modifiers genes, we expanded the genotyping of those UTRs variants in a large multicentre cohort of SCA3 patients. Additionally, we performed in silico predictions to evaluate the functional impact of those variants and, finally we performed genotype-phenotype correlations to infer their effect on SCA3 onset.

3.2. Patients and methods

Patients and samples

Genomic DNA from peripheral blood samples of a total of 303 SCA3 patients corresponding to four different cohorts was used (Table 3.1). Samples were provided by M. Lima (Azores), J. Sequeiros (Mainland Portugal), L. Jardim (Brazil) and C. Bettencourt (UK).

Table 3.1. Characterization of the four different cohorts (Azores, Mainland Portugal, Brazil and UK) used in this study considering number of samples, gender, onset (in years) and the number of CAG repetitions in the expanded allele (CAG_{exp}).

	Azores	Mainland Portugal	Brazil	UK	Total
N	101	66	96	40	303
Gender (F/M)	49.5% / 50.5%	40% /60%	53.1% / 46.9%	40% / 60%	48.2% / 51.8%
Onset	37.65 ± 12.00	35.20 ± 10.81	35.82 ± 10.17	35.20 ± 10.81	37.66 ± 11.78
CAG_{exp}	71.40 ± 3.72	70.83 ± 4.55	74.11 ± 3.68	70.83 ± 4.55	72.03 ± 4.09

All patients included in the study had a clinical diagnosis of SCA3 and were molecularly confirmed as carrying an expanded *ATXN3* CAG repeat; the sizing of the CAG tract in *ATXN3* was performed at a single laboratory (SEEBMO-Serviço Especializado de Epidemiologia e Biologia Molecular). Onset was considered as the age of appearance of the first manifestation of the disease (gait ataxia or diplopia) reported by the patient or relative/caregiver.

Extended whole-exome sequencing (WES) *and variants selection*

In the present chapter a total of sixty-one variants in UTRs mapping to forty genes, were analysed. The list of variants was obtained from the analysis of the extended WES data from Raposo and colleagues [160]. Briefly, extended WES was performed in genomic DNA samples from four onset discordant and four onset concordant affected first-degree pairs of Azorean SCA3 patients; the members of a discordant pair were defined as first degree relatives presented different onset (range: 7–11 years) and the same number of CAG repeats in expanded allele (±1 CAG), whereas members of a concordant pair had similar onset (range: 1–4 years) and the same number of CAG repeats in the expanded allele (±1 CAG) [160].

Extended WES analysis of these samples were performed at Macrogen, Inc (for more details see Raposo and colleagues [160]). Variants in UTRs, whose genotype differed between members of discordant pairs, but was the same for members of concordant pairs were retrieved (n=94-list 1, Supplementary Table 3.1), using a script from R language. Only the variants in genes expressed in the brain, including cerebellum (data accessed at GTEx project [258]) were considered.

Variant's genotyping

From the ninety-four variants identified in UTRs, sixty-four were selected to be genotyped. From these variants, a total of sixty-one variants were successfully genotyped (n=61 – list 2, Supplementary Table 3.1).

From the sixty-one variants identified, thirty-nine variants were genotyped successfully in the 303 patients from the four different cohorts (Table 3.1) by multiplexed sequencing at Eurofins Genomics (https://eurofinsgenomics.eu/), as described by Raposo and collaborations [160]. Twenty-four variants were successfully genotyped in 177 SCA3 samples from Azorean, Mainland Portugal, and Brazilian cohorts (samples that remain from the 303 samples of the initial group), using *MassArray Agena Bioscience* (*Sequenom*) at CeGen (http://www.usc.es/cegen/); variant specific primers were designed using Agena Assay Design Suite software. After primer design, DNA was amplified using iPLEX® Gold technology [259]. One variant was genotyped by Sanger sequencing in 92 samples from the Azorean cohort samples remaining from the previously genotyped group (conditions in Supplementary Table 3.2).

In silico predictions for the functional impact of variants found in different UTRs

Functional effects of variants found in 5'UTRs and 3'UTRs were predicted using Genome Wide Annotation of Variants (GWAVA) (https://www.sanger.ac.uk/sanger/StatGen_Gwava) [225]; a cut-off of 0.5 was considered, with variants showing scores higher than 0.5 being predicted to have a deleterious effect (an effect that implies the modification in the sequence and can have a consequence).

For variants in 5'UTR, to detect differences in transcription factors binding sites (TFBSs) and alterations in transcription factors (TFs) affinity due to the presence of variation found in 5'UTRs, an *in silico* analysis using PROMO tool (http://alggen.lsi.upc.es/cgi-

bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3) was performed [201]. The similarity index (similarity of the sequence of TFBSs to the matrix and variant sequence) was measured using the method of Quandt and collaborators [260] as implemented in this tool. A dissimilarity threshold (the parameter that controls the similarity of a sequence to the matrix to be reported as a hit) of 5% was used, which indicates 95% of similarity.

For 3'UTR variants, predicted miRNA binding sites were identified using *MirSNP* (http://bioinfo.bjmu.edu.cn/mirsnp/search/) [228]; only predicted miRNAs binding sites with a stringent score > 150 and an energy score < -7, as recommended, were considered valid. For the exploration of RBPs motifs, an analysis was performed using *RBP-Var2* tool (http://www.rbp-var.biols.ac.cn/) [229]; the default settings were considered in this analysis. To search for the abolishment or generation of alternative polyadenylation (polyA) motifs, analysis were performed with PolyASite 2.0 [230] using the default settings.

Analysis of the genotyped variants

Determination of allelic and genotypic frequencies, conformity with Hardy-Weinberg equilibrium, and linkage disequilibrium were obtained in Genepop [261] and Arlequin software v.3.5 [222]. Exact tests of population differentiation using genotypic frequencies of the several variants were performed to compare the different cohorts.

An ANCOVA, using the number of CAGs in the expanded allele as covariate, was performed to calculate and compare adjusted onset (adjonset) between genotypes. Three different groups of genotypic classes were established for those analysis: REF|REF (homozygosity for reference allele), REF|VAR (heterozygosity: reference allele and alternative allele) and VAR|VAR (homozygosity for the alternative allele). We pooled together two genotypic classes and compared REF|REF *versus* REF|VAR+VAR|VAR or REF|REF+REF|VAR *versus* VAR|VAR, according to the number of samples presented in each genotypic class and to guarantee a more equilibrated analysis, following the strategy used by Raposo and colleagues [160]. Statistical analyses were performed in *IBM SPSS Statistics v.25.0* for all the tests, *p* values lower than 0.05 were considered statistically significant.

Pathway and interaction-network analyses were generated with the software Ingenuity Pathways Analysis (IPA; https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/) [262]. Canonical pathways were obtained for the forty genes that were studied. Also, interaction networks between ATXN3 and candidate modifiers genes that presented a genotype-

phenotype correlation with a *p-value* <0.05 were investigated using the "Path Explorer" tool, which calculated the "Shortest Path" between two molecules. All the analysis were performed under default settings, except for the confidence level which was set to only include experimentally observed interactions to be more stringent.

3.3. Results

Sixty-four variants located in UTRs of genes known to be expressed in the human brain, were identified from the WES data generated by Raposo and colleagues [160]. From the sixty-four variants, three failed to be genotyped. All variants identified are present in dbSNP (https://www.ncbi.nlm.nih.gov/snp/). The general characterization of the sixty-one variants successfully genotyped is described in Supplementary Table 3.3a and Supplementary Table 3.3b.

Nine substitution variants were found in the 5'UTRs of eight genes; forty-six substitution variants were found in 3'UTR, of thirty-one genes. Also, in 3'UTR a total of eight indels were found in seven genes. Thirteen genes presented more than one variant. *GLI4*, presented two variants in 5'UTR, while *LEMD2* presented two variants in 3'UTR and one variant in 5'UTR. The other genes presented variants in theirs 3'UTRs. Eight genes presented two variants in 3'UTR, while the remaining presented more than two variants (Figure 3.1).

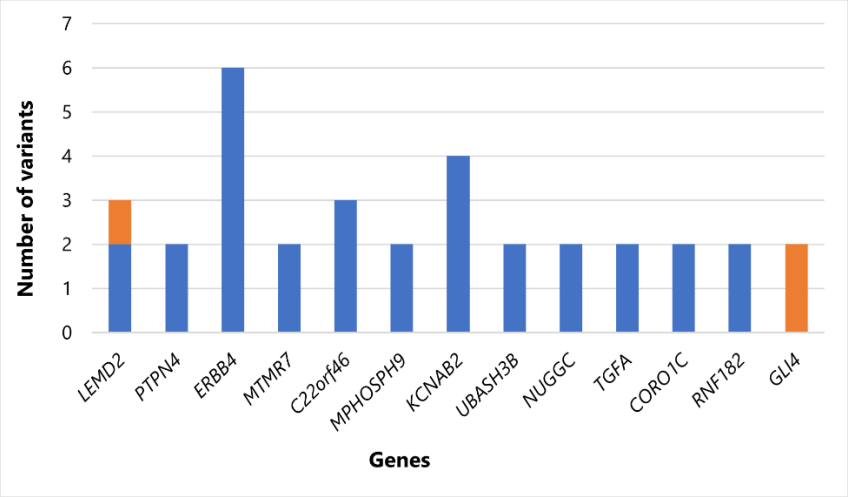


Figure 3.1. Genes with more than one variant. The columns in blue represent the variants in 3'UTR, while the columns in orange represent the variants in 5'UTR.

The main results obtained for the *in-silico* analysis of the variants found in 5'UTR of candidate genes, are shown in Table 3.1. One of the nine variants, rs7029012, did not present modifications in transcription factors binding sites when the alternative allele is present, according to PROMO predictions. In the other variants, the presence of the alternative allele was predicted to modify the binding sites for transcription factors; for three variants (rs17553593, rs7387207 and rs28375915) the loss of transcription factor

binding sites occurs and for the remaining variants the binding of new transcription factors is predicted. Concerning the scores obtaining for GWAVA, only two variants presented a predicted deleterious effect which can mean that the original effect of those region suffers a modification.

Table 3.1. Main results of *in silico* analysis obtained for the variants found in 5'UTRs of different genes

Variant	Gene	GWAVA score	PROMO	New TFs that binds
rs4711351 (NC_000006.11: g.33754592G>A)	*LEMD2* (LEM Domain Nuclear Envelope Protein 2)	0.44	Maintain number TFs and number TFBS	C/EBPalpha TFII1
rs2272624 (NC_000008.10: g.12809667A>T)	*TRMT9B* (*KIAA1456*) (TRNA Methyltransferase 9B (Putative)	0.15	(+) one TFBS (+) one TFs	C/EBPbeta
rs17553593 (NC_000008.10: g.12990734C>A)	*DLC1* (DLC1 Rho GTPase Activating Protein)	0.32	(-) all TFBS (-) all TFs	NA
rs3812433 (NC_000008.10: g.144357479C>G)	*GLI4* (GLI Family Zinc Finger 4)	0.32	(+) one TFBS (+) one TF	T3R-beta1
rs7387207 (NC_000008.10: g.144357772T>C)	*GLI4* (GLI Family Zinc Finger 4)	0.32	(-) one TFBS (-) one TFs	NA
rs7029012 (NC_000009.11: g.2717698C>G)	*KCNV2 (Potassium Voltage-Gated Channel Modifier Subfamily V Member 2)*	**0.52**	No TFBS No TFs	NA
rs1549168 (NC_000019.9: g.11529958C>T)	*RGL3* (Ral Guanine Nucleotide Dissociation Stimulator Like 3)	**0.79**	(+) one TFBS (+) one TF	AP-2alphaA
rs132770 (NC_000022.10: g.42017264A>G)	*XRCC6* (X-Ray Repair Cross Complementing 6)	0.47	(-) one TFBS (-) one TF	GR-alpha
rs28375915 (NC_000022.10: g.50354255G>A)	*PIM3* (Pim-3 Proto-Oncogene, Serine/ Threonine Kinase)	0.50	(-) three TFBSs (-) three TFs	NA

(+) – gain; (-) loss; NA – not applicable. Scores in bold indicate a deleterious effect.

Regarding the variation found in 3'UTR, when the alternative allele of three variants, rs3134430 (*UBASH3B*), rs8139993 (*DESI1*) and rs1051434 (*MPHOSPH9*). is present a predicted impact was simultaneously found for GWAVA score, miRNAs binding sites and RBP binding sites (Table 3.2). Concerning the effect that the variation in study can have in polyA binding sites, none of the variants presented an impact in polyA signals. Also, a total of eleven variants presented a GWAVA score superior to 0.5, meaning that were predicted to have a deleterious effect. Relatively to miRNA binding sites, a total of forty variants can break or create a binding site for miRNAs and/or increase/decrease the binding of a miRNA to that specific site (for more details see Table 3.2). Finally, for RNA binding protein binding sites a total of seventeen variants presented a score of 1e which means that they were predicted to affect the secondary structure of RNA.

Table 3.2. Main results of *in silico* analysis obtained for the variants found in 3'UTRs of different genes

Variant	Gene	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs1051322 (NC_000017.10: g.1946503C>G)	*DPH1* (Diphthamide Biosynthesis 1)	0.47	hsa-miR-4316 (break) hsa-miR-4649-3p (create) hsa-miR-4710 (break) hsa-miR-582-3p (+) hsa-miR-93-3p (create)	1e
rs11274444 (NC_000002.11: g.212245092_212245093insAAAATAGGATTG)	*ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4)*	No info	hsa-miR-587 (enhance)	No info
rs1595065 (NC_000002.11: g.212242641G>A)	*ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4)*	0.46	hsa-miR-1244 (create) hsa-miR-199a-3p (create) hsa-miR-199b-3p (create) hsa-miR-3129-5p (create)	No info
rs34217661 (NC_000002.11: g.212244761_212244762insG)	*ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4)*	No info	hsa-miR-302a-5p (break)	No info
rs1972820 (NC_000002.11: g.212243422G>A)	*ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4)*	0.41	hsa-miR-4633-5p (break) hsa-miR-3144-3p (create)	No info
rs1836724 (NC_000002.11: g.212244952G>A)	*ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4)*	0.43	hsa-miR-28-5p (break) hsa-miR-665 (break) hsa-miR-708-5p (break)	No info
rs12475523 (NC_000002.11: g.212245284A>G)	*ERBB4 (Erb-B2 Receptor Tyrosine Kinase 4)*	0.46	hsa-miR-378i (break)	No info
rs11542388 (NC_000001.11: g.6100638A>C)	*KCNAB2* (Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2)	0.31	hsa-miR-182-5p (-) hsa-miR-3614-3p (break)	No info
rs1056860 (NC_000001.10: g.6160876 A>G)	*KCNAB2* (Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2)	0.28	hsa-miR-1193 (-) hsa-miR-323b-3p (+) hsa-miR-5193 (break) hsa-miR-875-5p (break)	2b
rs1051434 (NC_000012.11: g.123641200C>T)	*MPHOSPH9* (M-Phase Phosphoprotein 9)	**0.54**	hsa-miR-4474-3p (create)	1e
rs3800373 (NC_000006.11: g.35542476C>A)	*FKBP5* (FKBP Prolyl Isomerase 5)	0.40	hsa-miR-505-3p (+) hsa-miR-652-5p (break)	1e
rs2147201 (NC_000006.11: g.13979526A>G)	*RNF182* (Ring Finger Protein 182)	0.4	has-miR-4307 (decrease)	2d
rs11794 (NC_000006.11: g.25701718G>A)	*SCGN* (Secretagogin, EF-Hand Calcium Binding Protein)	**0.53**	hsa-miR-138-5p (break) hsa-miR-3186-3p (break) hsa-miR-363-5p (break) hsa-miR-3656 (create/break) hsa-miR-3665 (create) hsa-miR-4456 (break) hsa-miR-4524b-3p (-)	No info
rs3732253 (NC_000002.11: g.70676098G>A)	*TGFA* (Transforming Growth Factor Alpha)	**0.51**	hsa-miR-199a-5p (+) hsa-miR-199b-5p (+) hsa-miR-4669 (break) hsa-miR-548ac (+)	0
rs71710818 (NC_000002.11: g.70676770delA)	*TGFA* (Transforming Growth Factor Alpha)	No info	hsa-miR-3915 (+)	4
rs3793428 (NC_000008.10: g.17144020C>A)	*VPS37A* (VPS37A Subunit Of ESCRT-I)	0.35	No info	1e
rs2111212 (NC_000012.12: g.108646018A>T)	*CORO1C* (Coronin 1C)	**0.53**	No info	2d
rs2111211 (NC_000012.11: g.109039869T>C)	*CORO1C* (Coronin 1C)	0.46	hsa-miR-191-3p (create) hsa-miR-5008-3p (break) hsa-miR-548v (+)	2d
rs35117663 (NC_000022.10: g.42091027_42091028insAG)	*C22orf46* (Chromosome 22 Putative Open Reading Frame 46)	No info	hsa-miR-4768-5p (create)	6

(-) binding decrease, (+) binding increase.

Table 3.2. Main results of *in silico* analysis obtained for the variants found in 3'UTRs of different genes (cont.)

Variant	Gene	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs3793428 (NC_000008.10: g.17144020C>A)	*VPS37A* (VPS37A Subunit Of ESCRT-I)	0.35	No info	1e
rs2187795 (NC_000022.10: g.42092225C>A)	*C22orf46* (Chromosome 22 Putative Open Reading Frame 46)	No info	hsa-miR-619 (decrease)	1e
rs739135 (NC_000022.11: g.41694065A>C)	*C22orf46* (Chromosome 22 Putative Open Reading Frame 46)	0.30	No info	1e
rs6981281 (NC_000008.10: g.17156904T>C)	*MTMR7* (Myotubularin Related Protein 7)	0.43	hsa-miR-1229 (decrease) hsa-miR-1285-5p (decrease)	1e
rs2246121 (NC_000008.10: g.17156293G>C)	*MTMR7* (Myotubularin Related Protein 7)	0.35	hsa-miR-4763-3p (break)	3a
rs9654238 (NC_000004.11: g.156264942G>A)	*MAP9* (Microtubule Associated Protein 9)	0.42	hsa-miR-4424 (create) hsa-miR-4774-3p (-)	No info
rs10046724 (NC_000008.10: g.27880155G>A)	*NUGGC* (Nuclear GTPase, Germinal Center Associated)	0.48	hsa-miR-186-5p (create)	No info
rs921597 (NC_000008.10: g.27880826C>T)	*NUGGC* (Nuclear GTPase, Germinal Center Associated)	0.48	has-miR-3913-5p (-)	No info
rs4348512 (NC_000008.10: g.144333434G>A)	*ZFP41*(ZFP41 Zinc Finger Protein)	0.47	hsa-miR-1269a (break) hsa-miR-1269b (break) hsa-miR-127-5p (break) hsa-miR-378c (break) hsa-miR-378f (break) hsa-miR-422a (break) hsa-miR-423-5p (-) hsa-miR-4430 (-) hsa-miR-4740-5p (break)	No info
rs34338282 (NC_000015.9:g.59389096C>T)	*RNF111* (Ring Finger Protein 111)	0.33	hsa-miR-218-5p (+) hsa-miR-643 (break)	5
rs11539756 (NC_000015.9:g.59429111G>A)	*MYO1E* (Myosin IE)	0.44	hsa-miR-2467-3p (+) hsa-miR-3141 (break) hsa-miR-3158-5p (+) hsa-miR-3616-3p (create) hsa-miR-4736 (create) hsa-miR-622 (-) hsa-miR-4436b-3p (create)	3a
rs1427793 (NC_000012.11: g.106458238T>C)	*NUAK1* (NUAK Family Kinase 1)	0.39	hsa-miR-455-3p (break) hsa-miR-4690-3p (create) hsa-miR-4727-5p (-)	1e
rs139561 (NC_000022.10: g.42195410T>C)	*MEI1* (Meiotic Double-Stranded Break Formation Protein 1)	**0.54**	hsa-miR-4474-3p (break)	No info
rs3134430 (NC_000011.9: g.122683382A>T)	*UBASH3B* (Ubiquitin Associated And SH3 Domain Containing B)	**0.54**	hsa-miR-5093 (-)	1e
rs3134428 (NC_000011.9: g.122682156A>G)	*UBASH3B* (Ubiquitin Associated And SH3 Domain Containing B)	0.4	has-miR-338-3p (-) hsa-miR-501-5p (breahashsa-miR-5196-3p (create)	1e
rs895412 (NC_000002.11: g.113973964T>C)	*PAX8* (Paired Box 8)	0.33	hsa-miR-4687-5p (+)	No info
rs1056667 (NC_000006.11: g.26510564T>C)	*BTN1A1* (Butyrophilin Subfamily 1 Member A1)	0.42	hsa-miR-4694-3p (+) hsa-miR-5093 (-)	No info
rs1061242 (NC_000001.10: g.220142070C>T)	*EPRS11* (Glutamyl-Prolyl-TRNA Synthetase 1)	0.48	hsa-miR-3163 (break)	1e
rs4673336 (NC_000002.11: g.206481573A>G)	*PARD3B* (Par-3 Family Cell Polarity Regulator Beta)	**0.56**	No info	No info
rs8139993 (NC_000022.10: g.41995335T>C)	*DESI1* (Desumoylating Isopeptidase 1)	**0.51**	hsa-miR-3607-5p (create) hsa-miR-3611 (create) hsa-miR-3611 (create)	1e

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Table 3.2. Main results of *in silico* analysis obtained for the variants found in 3'UTRs of different genes (cont.)

Variant	Gene	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs1048169 (NC_000009.11: g.19055965T>C)	*HAUS6* (HAUS Augmin Like Complex Subunit 6)	**0.52**	No modifications	1e
rs16920581 (NC_000008.10: g.57353530T>C)	*PENK* (Proenkephalin)	**0.55**	No modifications	No info
rs143599604 (NC_000003.11: g.5222491_5222492insGGGATAGTAAG)	*ARL8B* (ADP Ribosylation Factor Like GTPase 8B)	No info	hsa-miR-3127-3p (create)	1e
rs2296744 (NC_000006.11: g.33740382G>A)	*LEMD2* (LEM Domain Nuclear Envelope Protein 2)	0.47	hsa-miR-3929 (+) hsa-miR-4298 (-) hsa-miR-4419b (-) hsa-miR-4438 (-) hsa-miR-4478 (+) hsa-miR-873-5p (break)	1e
rs2296743 (NC_000006.11: g.33740025C>T)	*LEMD2* (LEM Domain Nuclear Envelope Protein 2)	0.42	hsa-miR-1913 (break) hsa-miR-1915-5p (create) hsa -miR-324-3p (break) hsa -miR-502-5p (create) hsa-miR-362-5p (create) hsa-miR-500b (create)	1e
rs11259151 (NC_000010.10: g.14562793C>T)	*FAM107B* (Family With Sequence Similarity 107 Member B)	0.50	hsa-miR-3916 (-) hsa-miR-583 (+)	5
rs35508493 (NC_000022.10: g.41574973_41574975delGTA)	*EP300* (E1A Binding Protein P300)	No info	hsa-miR-300 (break)	5
rs6633 (NC_000012.11: g.123745809G>T)	*CDK2AP1* (Cyclin Dependent Kinase 2 Associated Protein 1)	**0.64**	hsa-miR-3616-3p (enhance) hsa-miR-4503 (decrease) hsa-miR-4670-3p (create) hsa-miR-548p (decrease)	No info

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Almost all of the variants were in Hardy-Weinberg *Equilibrium*; the exceptions were rs921597 (*NUGGC*) and rs1051434 (*MPHOSPH9*). All of the variants localized in genes presenting more than one variant (Figure 3.1, Supplementary table 3.4) were in Linkage *disequilibrium* with the variants studied in the same gene.

Gene enrichment analysis, using the forty genes from List 2 (Supplementary table 3.1), resulted in eight overrepresented pathways (Table 3.3), including HIF1α signaling (1% overlapping genes), neuroregulin signaling (2%), Nucleotide Excision Repair, Enhanced Pathway (NER) (2%), the Granzyme A signaling (5%) and DNA Double-Strand Break Repair by Non-Homologous End Joining (7%).

Table 3.3. Eight statistically significant canonical pathways were obtained using all the candidate genes in study containing variants further analysed in the cohorts of study. The -log (Benjamini-Hochberg *p*-value) greater than 1.3 was considered as statistically significant.

Ingenuity Canonical Pathways	**Overlaps with dataset***	**Molecules**
RHOGDI Signaling	2/215 (1%)	DLC1, EP300
HIF1α Signaling	2/208 (1%)	EP300, TGFA
Granzyme A Signaling	1/19 (5%)	EP300
Telomere Extension by Telomerase	1/15 (7%)	XRCC6
DNA Double-Strand Break Repair by Non-Homologous End Joining	1/14 (7%)	XRCC6
Neuregulin Signaling	2/117 (2%)	ERBB4, TGFA
NER (Nucleotide Excision Repair, Enhanced Pathway)	2/103 (2%)	EP300, RNF111
TGF-β Signaling	2/96 (2%)	EP300, RNF111

*Number and percentage of molecules that overlap with the total number of genes that belong to that pathway.

Candidate modifier variants of onset in SCA3 were identified

We tested the main effects of each of the sixty-one variants on onset in the four cohorts. However, despite the population differentiation test revealing differences between the populations, the ANCOVA performed was adjusted to CAGs and all the cohorts were pooled together and treated like large cohort.

Twelve variants in four genes- *GLI4* (GLI Family Zinc Finger 4), *TGFA* (Transforming Growth Factor Alpha), *MPHOSPH9* (M-Phase Phosphoprotein 9) and *ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4) showed a significant effect in the SCA3 onset (Figure 3.2). Noteworthy, multiple variants in each of those genes were in LD (Supplementary table 3.4) with a D' ~1; consequently, for each gene, only one variant was selected to represent the effect of each gene (Figure 3.2.). The variants rs7387207 (GLI Family Zinc Finger 4 gene), rs3732253 (Transforming Growth Factor Alpha gene), rs7127313 (M-Phase Phosphoprotein 9 gene) and rs1595065 (Erb-B2 Receptor Tyrosine Kinase 4) were selected as their *p-value* is the lower when comparing with the other variants in the same gene.

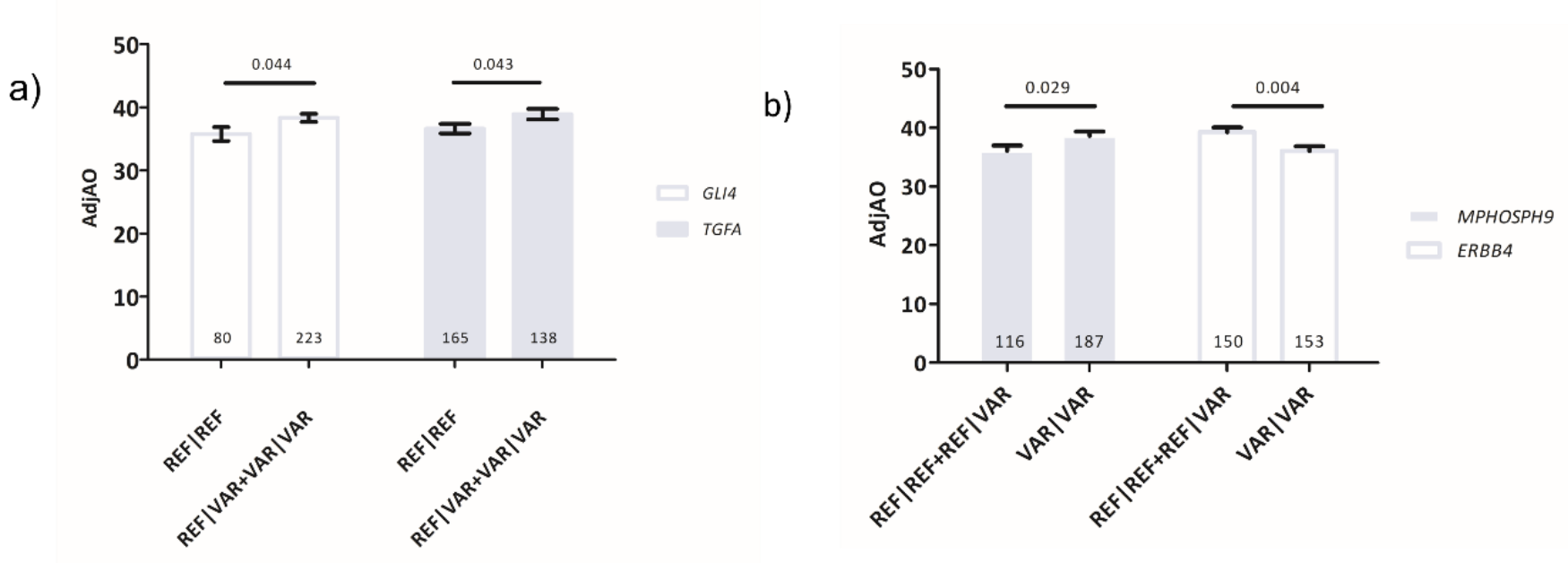


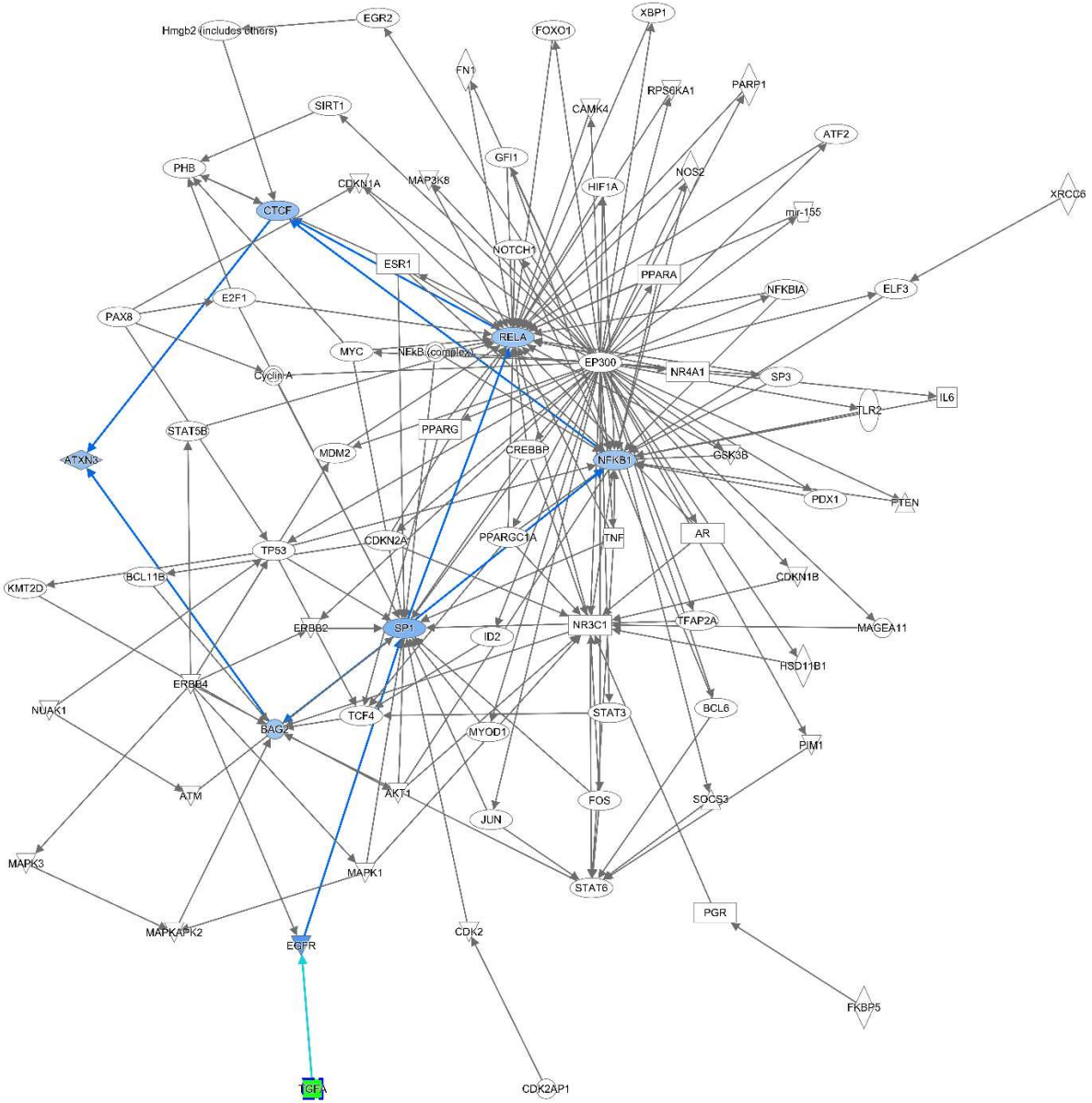
Figure 3.2. Effect of the candidate variants, in the candidate genes, on adjusted (Adj) onset (mean±standard error), after adjusting for CAGs in expanded allele: a) variants in *GLI4* (rs7387207) and *TGFA* (rs3732253), revealed to be statistically significant ($p<0.05$) when comparing the groups REF|REF *versus* REF|VAR+VAR|VAR, and b) variants in *MPHOSPH9* (rs7127313) and *ERBB4*, (rs1595065) revealed to be statistically significant ($p<0.05$) when comparing the groups REF|REF+ REF|VAR *versus* VAR|VAR. The number of SCA3 patients by allelic variant and gene are represented inside the bars

The variants in *TGFA*, *MPHOSPH9* and *GLI4* were associated with a later onset; for the two first genes, this variation was ~2 years, and for *GLI4* is ~3 years. All the six variants in *ERBB4* were associated with an earlier onset (~3 years).

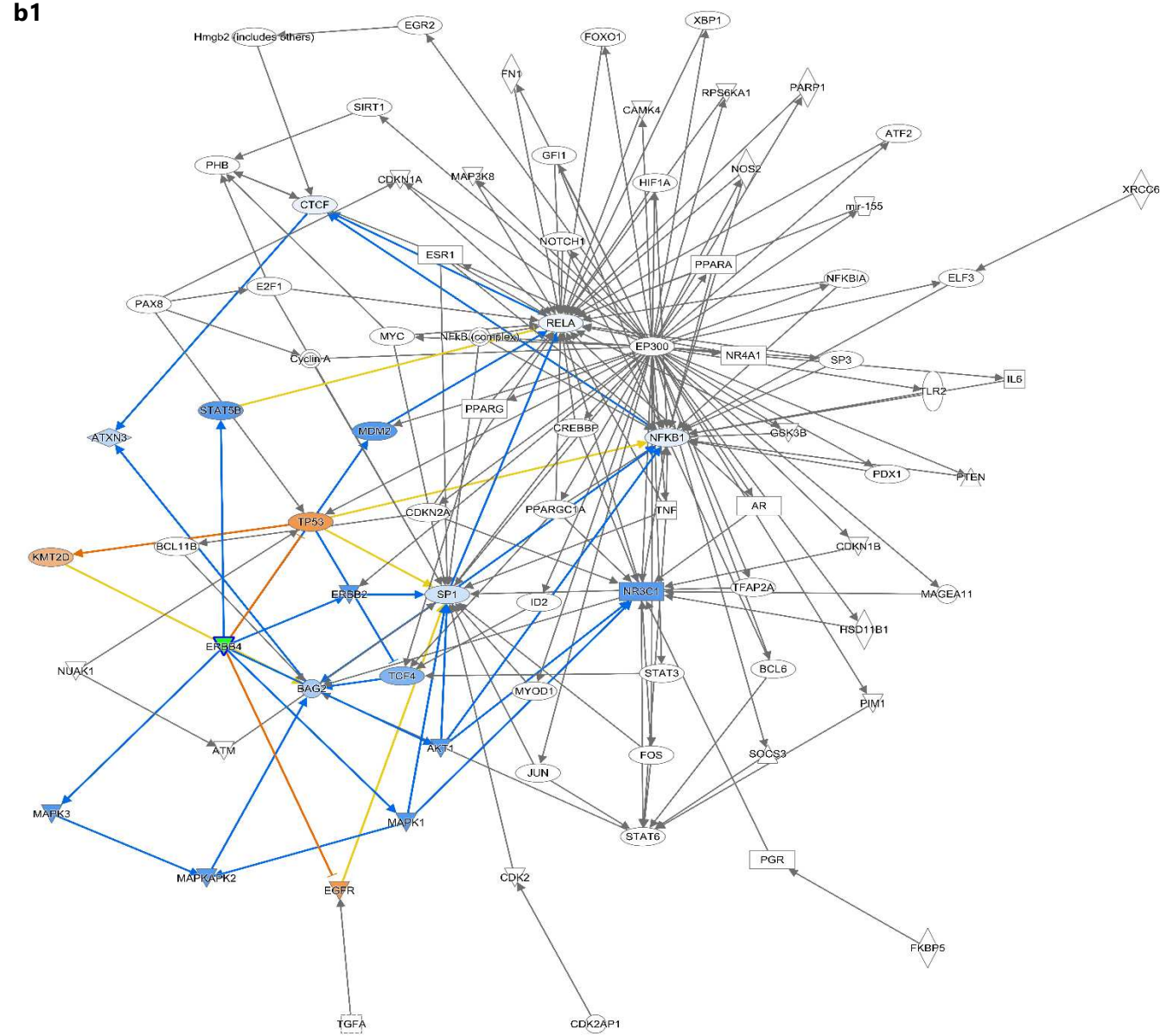
Network analysis revealed interactions between ATXN3 and TGFA and ERBB4

A network analysis was performed with *ATXN3* and the interacting candidate modifier genes. *TGFA* and *ERBB4* interact with *NFKB1*, which in turn, directly interacts with *ATXN3*. *TGFA* and *ERBB4* presented an association with the onset and interacts with *ATXN3*, their study was more valorised. To evaluate the impact of these molecules in ATXN3 an analysis of inhibition/activation were performed (see Figure 3.3) When *TGFA* and *ERBB4* were activated, ATXN3 suffers activation, while their inhibition causes the inhibition of *ATXN3*.

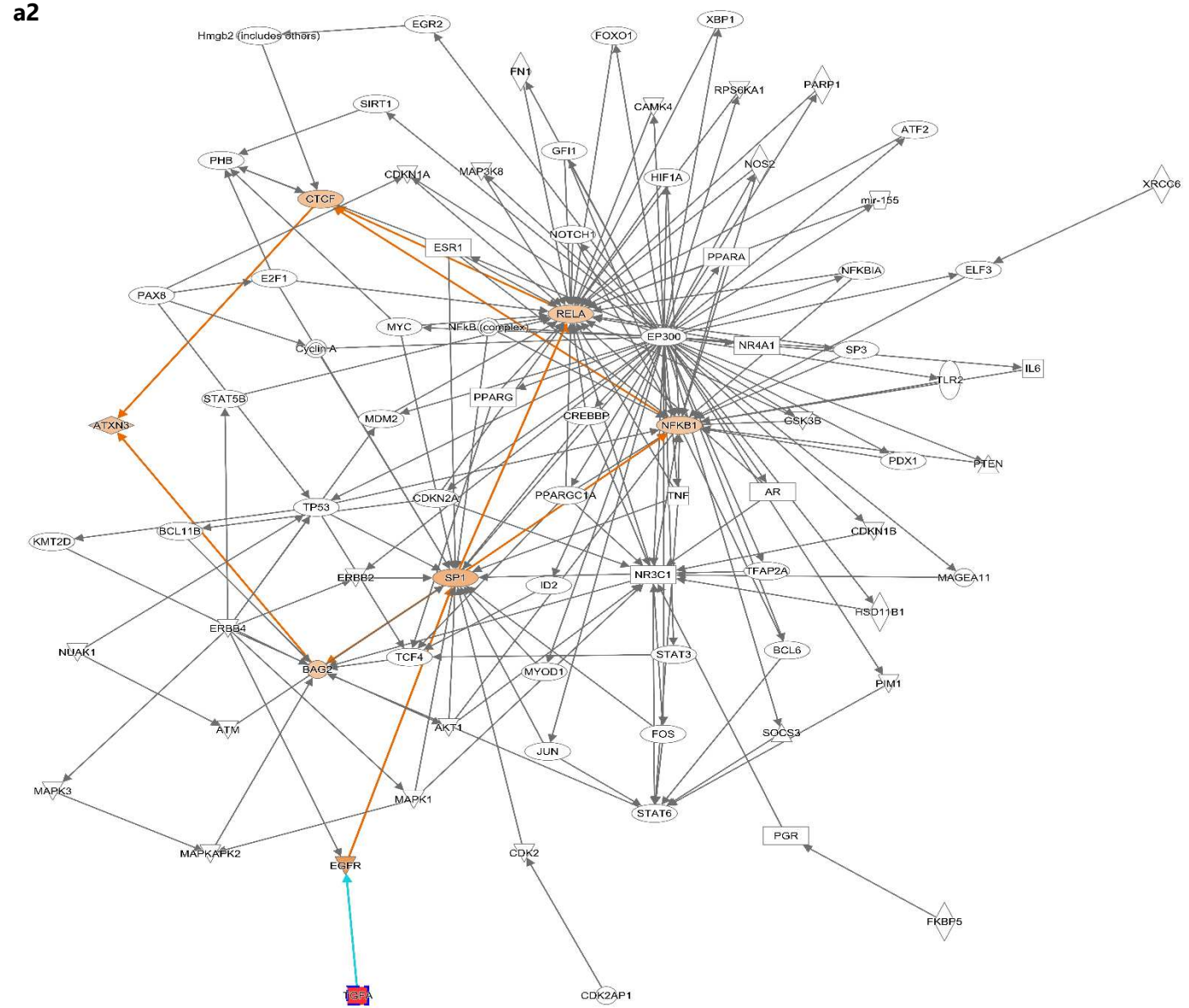
a1



b1



a2



b2

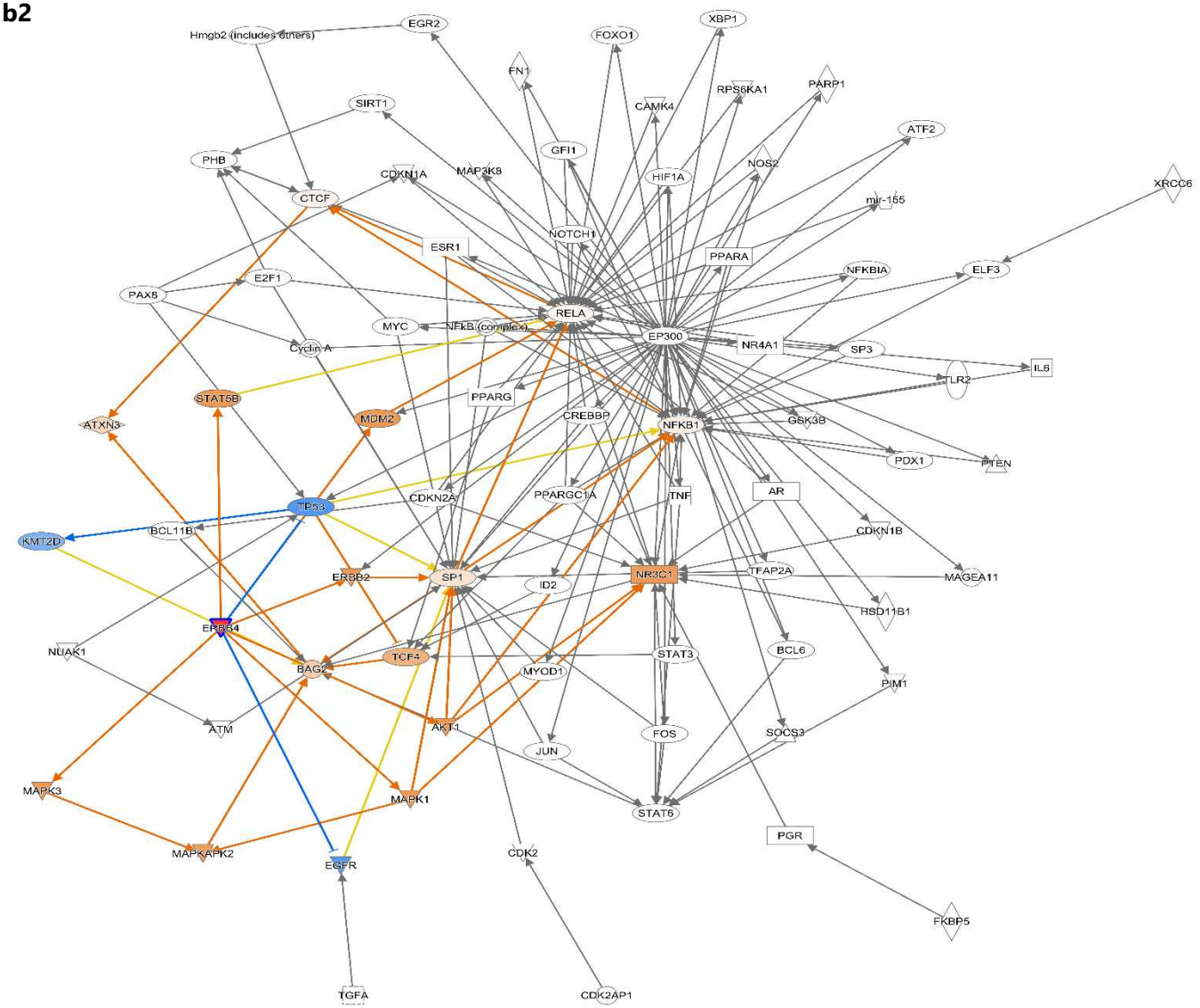


Figure 3.3. Interaction networks showing the relationship between *TGFA* and *ATXN3* (a1 and a2) and between *ERBB4* and *ATXN3* (b1 and b2). Network analysis was performed in IPA software with the tool Path Explorer. The inhibition of the two candidate modifier genes with a genotype-phenotype correlation (*TGFA* and *ERBB4*) was represented in green (a1 and b1); the activation of *TGFA* and *ERBB4* was represented in red (a2 and b2). Arrows in blue indicates the inhibition of the molecules, while arrows in orange indicates the activation of the molecules. The two candidate modifiers also have an impact in *NFKB1*, which interact directly with *ATXN3* [160].

3.4. Discussion

Genetic modifiers are known to be difficult to identify, especially in diseases that are rare such as SCA3 and other polyglutamine diseases. We accessed WES data to identify modifiers of SCA3 and focused our analysis on variants in UTRs.

In this study we showed that six variants in 5'UTR present the gain (rs2272624, rs3812433 and rs1549168) or loss (rs17553593, rs7387207, rs132770 and rs28375915) of transcription factor binding sites and transcription factors. The variant rs7029012 (*KCNV2)*, according to GWAVA score, was considered more likely to have a functional consequence, as its score was 0.52 [225]; however in PROMO tool, this variant did not present modifications in transcription factors binding sites, nor in transcription factors. Although we have restricted the dissimilarity threshold, to guarantee that the transcription factors and transcription factors binding sites that were found are likely to be real, a study have proved that PROMO can present false results [263]. To evaluate our result for this variant, further studies with the *KCNV2* and functional studies with interference RNA in *C.elegans* should be performed, to observe the effect of the silencing of this gene in the phenotype of the mutant model.

For the variation found in 3'UTRs, a total of forty variants presented modifications in the binding sites for miRNAs (Table 3.2) and seventeen variants presented a score of 1e for modifications of the binding sites for RNA-binding proteins.

Using sixty-one variants we have obtained a total of eight statistically enriched pathways. From these pathways, neuroregulin signaling, HIF1α signaling and DNA Double-Strand Break Repair by Non-Homologous End Joining were particularly interesting due to their roles in the nervous system and in the repair of DNA damage.

The neuroregulin signaling pathway presents an important function in the nervous system and was previously reported in the study of Raposo and colleagues [160] as containing modifier genes of onset in SCA3 patients. In our study, *ERBB4* and *TGFA* were the genes involved in this pathway. This pathway presents different roles in central and peripheral nervous systems, being involved in the GABAergic neuron migration, synapse formation, neuromuscular junction and development, maturation and myelination of Schwann cells [264]. Also, the neuroregulin signaling pathway has been reported as playing a role in important processes for neurodegeneration, such as neuroinflammation and oxidative stress [265].

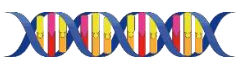
The HIF1α signaling pathway presents an important role in neurodegenerative diseases as the hypoxia-ischemia that occurs in brain can lead to depolarization, excitotoxicity, oxidative stress, inflammation, and apoptosis [266]. Also, hypoxia can lead to neurodegeneration and neuroprotection depending on the conditions: studies have revealed that due to brain plasticity, hypoxia can ameliorate some functions of an injured brain [267].

The DNA Double-Strand Break Repair by Non-Homologous End Joining pathway, is the main pathway for the repair of the DNA double-strand breaks, which is an extremely cytotoxic form of DNA damage [268]. This pathway involves synapsis and ligation of the broken strands. Consequently, we can hypothesis that a dysregulation in this pathway may led to extremely consequences in the damaged DNA.

In the group of the candidate genes with variants found in 5'UTR, only *GLI4* variant presented simultaneously a predicted impact in *in silico* analysis and an effect on onset in genotype-phenotype analysis (delaying onset by 3 years). *GLI4* belongs to the group of zinc finger proteins, a group of proteins that were predicted to enable DNA binding transcription factor activity, and can be involved in the regulation of transcription [269,270]. This gene did not interact with ATXN3 and was not present in the enrichment pathways that we have obtained. Nevertheless, in a study conducted in *C.elegans* in the Lab of Patrícia Maciel (personal communication) it was observed an amelioration of the transgenic model phenotype when *GLI4* is inactivated. Consequently, and for a better understanding of the mechanism involved with this gene and SCA3 onset variance further studies need to be performed.

Three genes with variants in 3'UTRs that presented a predicted impact in the *in silico* analysis and, simultaneously, have an effect on onset: *TGFA*, *ERBB4* and *MPHOSPH9*. *TGFA* are associated with a signaling pathway for cell proliferation, development and differentiation; this cytokine is involved in the regulation of reactions of immune system [271]. Inflammation processes have been described in SCA3 related to cytokines expression levels, as reported by different studies [174,255,256,272]. *ERBB4* is a transmembrane tyrosine kinase that binds and is activated by neuregulins and other different factors [273,274]. This gene is also responsible for the induction of diverse cellular responses, such as mitogenesis and differentiation [273,274] and is associated to amyotrophic lateral sclerosis [273,274]. Neuroregulins are important in neural development but also in adult brain [273,274]. *MPHOSPH9* is a phosphoprotein related to brain volume, being associated

to splicing abnormalities [275]. As these genes presented an effect in variance of SCA3 onset, a further study considering them as disease-modifying targets can be of value to this disease.



3.5. Conclusions

In this study we proposed four different SCA3 modifying genes with variants in UTRs. We have also identified pathways that can be relevant for further studies concerning disease-modifying targets; one of these pathways, neuroregulin signaling pathway, have been recently reported by Raposo and colleagues in the context of a genome-wide study considering variation in coding regions of different genes and their impact in SCA3 onset [160].

3.6. Supplementary material

Supplementary Table 3.1. List of the variants/genes found which were selected for the study.

List 1 (n=94)		List 2 (n=61)	
rs4148552	*ABCC4*		
rs143599604	*ARL8B*	rs143599604	*ARL8B*
rs3215660	*BTN1A1*		
rs1056667	*BTN1A1*	rs1056667	*BTN1A1*
rs1056668	*BTN1A1*		
rs12217414	*C10ORF68*		
rs739135	*C22orf46*	rs739135	*C22orf46*
rs35117663	*C22orf46*	rs35117663	*C22orf46*
rs2187795	*C22orf46*	rs2187795	*C22orf46*
rs6633	*CDK2AP1*	rs6633	*CDK2AP1*
rs2111212	*CORO1C*	rs2111212	*CORO1C*
rs2111211	*CORO1C*	rs2111211	*CORO1C*
rs12463238	*CRX*		
rs10667709	*DENND4C*		
rs3830756	*DESI1*		
rs8139993	*DESI1*	rs8139993	*DESI1*
rs17553593	*DLC1*	rs17553593	*DLC1*
rs1051322	*DPH1*	rs1051322	*DPH1*
rs1272944	*DPP6*		
rs35508493	*EP300*	rs35508493	*EP300*
rs1061248	*EPRS1*		
rs1061242	*EPRS1*	rs1061242	*EPRS1*
rs1595065	*ERBB4*	rs1595065	*ERBB4*
rs1972820	*ERBB4*	rs1972820	*ERBB4*
rs34217661	*ERBB4*	rs34217661	*ERBB4*
rs1836724	*ERBB4*	rs1836724	*ERBB4*
rs11274444	*ERBB4*	rs11274444	*ERBB4*
rs12475523	*ERBB4*	rs12475523	*ERBB4*
rs11259151	*FAM107B*	rs11259151	*FAM107B*
rs3800373	*FKBP5*	rs3800373	*FKBP5*
rs1126827	*GDF10*		
rs2853839	*GDF10*		
rs1540253	*GDF10*		
rs3812433	*GLI4*	rs3812433	*GLI4*
rs7387207	*GLI4*	rs7387207	*GLI4*
rs1048169	*HAUS6*	rs1048169	*HAUS6*
rs112943240	*HIST1H2AA*		
rs4711095	*HIST1H2AA*		
rs2072242	*HS3ST3B1*		
rs9434641	*KCNAB2*		
rs60141764	*KCNAB2*	rs60141764	*KCNAB2*
rs11542388	*KCNAB2*	rs11542388	*KCNAB2*
rs1056860	*KCNAB2*	rs1056860	*KCNAB2*
rs538	*KCNAB2*	rs538	*KCNAB2*
rs7029012	*KCNV2*	rs7029012	*KCNV2*
rs2272624	*KIAA1456*	rs2272624	*KIAA1456*
rs2272623	*KIAA1456*		
rs149099560	*LEMD2*		
rs2296743	*LEMD2*	rs2296743	*LEMD2*
rs2296744	*LEMD2*	rs2296744	*LEMD2*
rs4711351	*LEMD2*	rs4711351	*LEMD2*

Supplementary Table 3.1. Identification of the variants/genes found which were selected for the study (cont.).

List 1 (n=94)		List 2 (n=61)	
rs9654238	*MAP9*	rs9654238	*MAP9*
rs139561	*MEI1*	rs139561	*MEI1*
rs1727313	*MPHOSPH9*	rs1727313	*MPHOSPH9*
rs1051434	*MPHOSPH9*	rs1051434	*MPHOSPH9*
rs2246121	*MTMR7*	rs2246121	*MTMR7*
rs6981281	*MTMR7*	rs6981281	*MTMR7*
rs11853828	*MYO1E*		
rs11539756	*MYO1E*	rs11539756	*MYO1E*
rs8779	*NHP2L1*		
rs5787350	*NKX2-3*		
rs888208	*NKX2-3*		
rs1064959	*NUAK1*		
rs1427793	*NUAK1*	rs1427793	*NUAK1*
rs10046724	*NUGGC*	rs10046724	*NUGGC*
rs921597	*NUGGC*	rs921597	*NUGGC*
rs4673336	*PARD3B*	rs4673336	*PARD3B*
rs2019137	*PAX8*		
rs895412	*PAX8*	rs895412	*PAX8*
rs67776659	*PAX8*		
rs16920581	*PENK*	rs16920581	*PENK*
rs4738502	*PENK*		
rs28375915	*PIM3*	rs28375915	*PIM3*
rs1265099	*PSORS1C2*		
rs554190572	*PTPN4*	rs554190572	*PTPN4*
rs11676058	*PTPN4*	rs11676058	*PTPN4*
rs11633629	*PYGO1*		
rs9607434	*RAC2*		
rs1549168	*RGL3*	rs1549168	*RGL3*
rs34338282	*RNF111*	rs34338282	*RNF111*
rs2147200	*RNF182*	rs2147200	*RNF182*
rs2147201	*RNF182*	rs2147201	*RNF182*
rs11794	*SCGN*	rs11794	*SCGN*
rs6576	*SETD8*		
rs3732253	*TGFA*	rs3732253	*TGFA*
rs71710818	*TGFA*	rs71710818	*TGFA*
rs2022065	*TRIM10*		
rs3134428	*UBASH3B*	rs3134428	*UBASH3B*
rs3134430	*UBASH3B*	rs3134430	*UBASH3B*
rs3793428	*VPS37A*	rs3793428	*VPS37A*
rs132770	*XRCC6*	rs132770	*XRCC6*
rs4348512	*ZFP41*	rs4348512	*ZFP41*
rs55636053	*ZNF286B*		
rs2494703	*ZNF322*		

Supplementary Table 3.2. PCR conditions for the amplification of the variant rs28375915 (*PIM3*). The sanger sequencing for the samples was performed as a service at Eurofins Genomics (https://eurofinsgenomics.eu/).

Primers
Primer forward - PIM3_F: 5'-CATTGGCTCGTGCAGTCCC-3'
Primer reverse - PIM3_R: 5'GAACTTGGAGAGCAGCATCGC-3'
PCR conditions
Reaction mixture
0.3 µM of each primer
0.3 µM of dNTPs
1.5 mM of $MgCl_2$
10x reaction buffer
10% of dimethyl sulfoxide (DMSO)
1 U of BIOTAQ™ DNA polymerase (Bioline)
200 ng of genomic DNA
Thermocycle conditions
Initial denaturation: 95ºC, 5 min
Denaturation: 94 ºC, 30 sec Annealing: 61ºC, 30 sec Extension: 72ºC, 30 sec — 24 cycles
Final extension: 72 ºC, 7 min

Supplementary Table 3.3a. General characterization of the genotyped candidate variants in 5'UTRs (N=9) provided by extended WES analysis (List 2).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	PROMO	New TFs that binds
rs4711351 (NC_000006.11: g.33754592G>A)	*LEMD2* (LEM Domain Nuclear Envelope Protein 2)	substitution	G: 0.71 A: 0.29 (N=177)	GG: 0.474 GA: 0.463 AA: 0.063	0.44	Maintain number TFs and number TFBS	C/EBPalpha TFII1
rs2272624 (NC_000008.10: g.12809667A>T)	*TRMT9B* (*KIAA1456*) (TRNA Methyltransferase 9B (Putative)	substitution	A: 0.55 T: 0.45 (N=177)	AA: 0.280 AT: 0.531 TT: 0.189	0.15	(+) one TFBS (+) one TFs	C/EBPbeta
rs17553593 (NC_000008.10: g.12990734C>A)	*DLC1* (DLC1 Rho GTPase Activating Protein)	substitution	C: 0.79 A: 0.21 (N=177)	CC: 0.623 CA: 0.331 AA: 0.046	0.32	(-) all TFBS (-) all TFs	NA
rs3812433 (NC_000008.10: g.144357479C>G)	*GLI4* (GLI Family Zinc Finger 4)	substitution	C: 0.50 G: 0.50 (N=303)	CC: 0.264 CG: 0.472 GG: 0.264	0.32	(+) one TFBS (+) one TF	T3R-beta1
rs7387207 (NC_000008.10: g.144357772T>C)	*GLI4* (GLI Family Zinc Finger 4)	substitution	T: 0.49 C: 0.51 (N=303)	TT: 0.254 TC: 0.465 CC: 0.281	0.32	(-) one TFBS (-) one TFs	NA
rs7029012 (NC_000009.11: g.2717698C>G)	*KCNV2* (Potassium Voltage-Gated Channel Modifier Subfamily V Member 2)	substitution	C: 0.47 G: 0.53 (N=177)	CC: 0.251 CG: 0.429 GG: 0.320	**0.52**	No TFBS No TFs	NA
rs1549168 (NC_000019.9: g.11529958C>T)	*RGL3* (Ral Guanine Nucleotide Dissociation Stimulator Like 3)	substitution	C: 0.61 T: 0.39 (N=177)	CC: 0.343 CT: 0.543 TT: 0.114	**0.79**	(+) one TFBS (+) one TF	AP-2alphaA
rs132770 (NC_000022.10: g.42017264A>G)	*XRCC6* (X-Ray Repair Cross Complementing 6)	substitution	A: 0.20 G: 0.80 (N=303)	AA: 0.026 AG: 0.337 GG: 0.637	0.47	(-) one TFBS (-) one TF	GR-alpha
rs28375915 (NC_000022.10: g.50354255G>A)	*PIM3* (Pim-3 Proto-Oncogene, Serine/ Threonine Kinase)	substitution	G: 0.51 A: 0.49 (N = 92)	GG: 0.293 GA: 0.435 AA: 0.272	0.50	(-) three TFBSs (-) three TFs	NA

(+) – gain; (-) loss; NA – not applicable. Scores in bold indicate a deleterious effect.

Supplementary table 3.3b. General characterization of the genotyped candidate variants in 3'UTRs (N=52) provided by extended WES analysis (List 2).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs1051322 (NC_000017.10: g.1946503C>G)	*DPH1* (Diphthamide Biosynthesis 1)	substitution	C: 0.81 G: 0.19 (N=303)	CC: 0.647 CG: 0.32 GG: 0.03	0.47	hsa-miR-4316 (break) hsa-miR-4649-3p (create) hsa-miR-4710 (break) hsa-miR-582-3p (+) hsa-miR-93-3p (create)	1e
rs11274444 (NC_000002.11: g.212245092_212245093insAAAATAGGATTG)	*ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4)	indel	-: 0.35 insAAAATAG GATTG:0.65 (N=303)	- -: 0.139 - ins: 0.479 ins ins: 0.383	No info	hsa-miR-587 (enhance)	No info
rs1595065 (NC_000002.11: g.212242641G>A)	*ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4)	substitution	G:0.30 A: 0.70 (N=303)	GG: 0.096 GA: 0.399 AA: 0.505	0.46	hsa-miR-1244 (create) hsa-miR-199a-3p (create) hsa-miR-199b-3p (create) hsa-miR-3129-5p (create)	No info
rs34217661 (NC_000002.11: g.212244761_212244762insG)	*ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4)	indel	-: 0.35 insG: 0.65 (N=303)	--: 0.135 - ins: 0.455 ins ins: 0.409	No info	hsa-miR-302a-5p (break)	No info
rs1972820 (NC_000002.11: g.212243422G>A)	*ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4)	substitution	G: 0.35 A: 0.65 (N=303)	GG:0.129 GA: 0.446 AA:0.426	0.41	hsa-miR-4633-5p (break) hsa-miR-3144-3p (create)	No info
rs1836724 (NC_000002.11: g.212244952G>A)	*ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4)	substitution	G: 0.38 A: 0.62 (N=303)	GG: 0.139 GA: 0.472 AA: 0.389	0.43	hsa-miR-28-5p (break) hsa-miR-665 (break) hsa-miR-708-5p (break)	No info
rs12475523 (NC_000002.11: g.212245284A>G)	*ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4)	substitution	A: 0.37 G: 0.63 (N=303)	AA: 0.132 AG: 0.479 GG: 0.389	0.46	hsa-miR-378i (break)	No info
rs60141764 (NC_000001.10: g.6160692_6160693insC)	*KCNAB2* (Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2)	indel	-: 0.25 insC: 0.75 (N=303)	--: 0.063 - ins: 0.383 ins ins: 0.554	No info	No info	6
rs11542388 (NC_000001.11: g.6100638A>C)	*KCNAB2* (Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2)	substitution	A: 0.75 G: 0.25 (N=303)	AA: 0.063 AG 0.383 GG: 0.554	0.31	hsa-miR-182-5p (-) hsa-miR-3614-3p (break)	No info
rs538 (NC_000001.10: g.6160958 A>G)	*KCNAB2* (Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2)	substitution	A: 0.26 G: 0.74 (N=303)	AA: 0.063 AG: 0.389 GG: 0.548	0.22	No info	No info

(-) binding decrease, (+) binding increase.

Supplementary table 3.3b. General characterization of the genotyped candidate variants in 3'UTRs (N=52) provided by extended WES analysis (List 2) (cont.).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs1056860 (NC_000001.10: g.6160876 A>G)	*KCNAB2* (Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2)	substitution	A: 0.27 G: 0.73 (N=303)	AA: 0.069 AG:0.399 GG: 0.531	0.28	hsa-miR-1193 (-) hsa-miR-323b-3p (+) hsa-miR-5193 (break) hsa-miR-875-5p (break)	2b
rs1727313 (NC_000012.11:g.123640853 C>G)	*MPHOSPH9* (M-Phase Phosphoprotein 9)	substitution	C: 0.20 G: 0.80 (N=303)	CC: 0.020 CG: 0.366 GG: 0.614	0.47	No info	2d
rs1051434 (NC_000012.11: g.123641200C>T)	*MPHOSPH9* (M-Phase Phosphoprotein 9)	substitution	C: 0.38 T: 0.62 (N=303)	CC: 0.125 CT:0.515 TT: 0.360	**0.54**	hsa-miR-4474-3p (create)	1e
rs3800373 (NC_000006.11: g.35542476C>A)	*FKBP5* (FKBP Prolyl Isomerase 5)	substitution	C: 0.33 A: 0.67 (N=177)	CC: 0.103 CA: 0.451 AA: 0.446	0.4	hsa-miR-505-3p (+) hsa-miR-652-5p (break)	1e
rs2147200 (NC_000006.11: g.13978854G>T)	*RNF182* (Ring Finger Protein 182)	substitution	G: 0.61 T: 0.39 (N=303)	GG: 0.363 GT: 0.492 TT: 0.145	0.39	No info	2d
rs2147201 (NC_000006.11: g.13979526A>G)	*RNF182* (Ring Finger Protein 182)	substitution	A: 0.61 G: 0.39 (N=303)	AA: 0.366 AG: 0.488 GG: 0.145	0.4	has-miR-4307 (decrease)	2d
rs11794 (NC_000006.11: g.25701718G>A)	*SCGN* (Secretagogin, EF-Hand Calcium Binding Protein)	substitution	G: 0.77 A: 0.23 (N=303)	GG: 0.587 GA: 0.373 AA: 0.040	**0.53**	hsa-miR-138-5p (break) hsa-miR-3186-3p (break) hsa-miR-363-5p (break) hsa-miR-3656 (create/break) hsa-miR-3665 (create) hsa-miR-4456 (break) hsa-miR-4524b-3p (-)	No info
rs3732253 (NC_000002.11: g.70676098G>A)	*TGFA* (Transforming Growth Factor Alpha)	substitution	G: 0.74 A: 0.26 (N=303)	GG: 0.545 GA: 0.386 AA: 0.069	**0.51**	hsa-miR-199a-5p (+) hsa-miR-199b-5p (+) hsa-miR-4669 (break) hsa-miR-548ac (+)	0
rs71710818 (NC_000002.11: g.70676770delA)	*TGFA* (Transforming Growth Factor Alpha)	indel	A: 0.74 delA: 0.26 (N=303)	AA: 0.548 A delA: 0.380 delA delA: 0.073	No info	hsa-miR-3915 (+)	4

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Supplementary table 3.3b. General characterization of the genotyped candidate variants in 3'UTRs (N=52) provided by extended WES analysis (List 2) (cont.).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs3793428 (NC_000008.10: g.17144020C>A)	*VPS37A* (VPS37A Subunit Of ESCRT-I)	substitution	C: 0.76 A: 0.24 (N=303)	CC: 0.574 CA: 0.373 AA: 0.053	0.35	No info	1e
rs2111212 (NC_000012.12: g.108646018A>T)	*CORO1C* (Coronin 1C)	substitution	A: 0.46 T: 0.54 (N=303)	AA: 0.201 AT: 0.505 TT: 0.294	**0.53**	No info	2d
rs2111211 (NC_000012.11: g.109039869T>C)	*CORO1C* (Coronin 1C)	substitution	T: 0.47 C: 0.53 (N=303)	TT: 0.211 TC: 0.508 CC: 0.281	0.46	hsa-miR-191-3p (create) hsa-miR-5008-3p (break) hsa-miR-548v (+)	2d
rs35117663 (NC_000022.10: g.42091027_42091028insAG)	*C22orf46* (Chromosome 22 Putative Open Reading Frame 46)	indel	-: 0.28 insAG: 0.72 (N=303)	--: 0.086 -insAG: 0.380 insAG insAG: 0.535	No info	hsa-miR-4768-5p (create)	6
rs2187795 (NC_000022.10: g.42092225C>A)	*C22orf46* (Chromosome 22 Putative Open Reading Frame 46)	substitution	C: 0.18 A: 0.82 (N=303)	CC: 0.026 CA: 0.307 AA: 0.667	No info	hsa-miR-619 (decrease)	1e
rs739135 (NC_000022.11: g.41694065A>C)	*C22orf46* (Chromosome 22 Putative Open Reading Frame 46)	substitution	A: 0.190 C: 0.810 (N = 303)	AA: 0.026 AC: 0.310 CC: 0.663	0.30	No info	1e
rs6981281 (NC_000008.10: g.17156904T>C)	*MTMR7* (Myotubularin Related Protein 7)	substitution	T: 0.76 C: 0.24 (N= 303)	TT: 0.554 TC: 0.409 CC: 0.036	0.43	hsa-miR-1229 (decrease) hsa-miR-1285-5p (decrease)	1e
rs2246121 (NC_000008.10: g.17156293G>C)	*MTMR7* (Myotubularin Related Protein 7)	substitution	G: 0.74 C: 0.26 (N=303)	GG: 0.548 GC: 0.386 CC: 0.066	0.35	hsa-miR-4763-3p (break)	3a
rs9654238 (NC_000004.11: g.156264942G>A)	*MAP9* (Microtubule Associated Protein 9)	substitution	G: 0.47 A: 0.53 (N=177)	GG: 0.229 GA: 0.491 AA: 0.280	0.42	hsa-miR-4424 (create) hsa-miR-4774-3p (-)	No info
rs10046724 (NC_000008.10: g.27880155G>A)	*NUGGC* (Nuclear GTPase, Germinal Center Associated)	substitution	G: 0.42 A: 0.58 (N=177)	GG: 0.160 GA: 0.514 AA: 0.326	0.48	hsa-miR-186-5p (create)	No info
rs921597 (NC_000008.10: g.27880826C>T)	*NUGGC* (Nuclear GTPase, Germinal Center Associated)	substitution	C: 0.74 T: 0.26 (N=177)	CC: 0.160 CT: 0.514 TT: 0.326	0.48	has-miR-3913-5p (-)	No info

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Supplementary table 3.3b. General characterization of the genotyped candidate variants in 3'UTRs (N=52) provided by extended WES analysis (List 2) (cont.).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs4348512 (NC_000008.10: g.144333434G>A)	*ZFP41*(ZFP41 Zinc Finger Protein)	substitution	G: 0.51 A: 0.49 (N=303)	GG: 0.274 GA: 0.482 AA: 0.244	0.47	hsa-miR-1269a (break) hsa-miR-1269b (break) hsa-miR-127-5p (break) hsa-miR-378c (break) hsa-miR-378f (break) hsa-miR-422a (break) hsa-miR-423-5p (-) hsa-miR-4430 (-) hsa-miR-4740-5p (break)	No info
rs34338282 (NC_000015.9:g.59389096C>T)	*RNF111* (Ring Finger Protein 111)	substitution	C: 0.73 T: 0.27 (N=177)	CC: 0.543 CT: 0.371 TT: 0.086	0.33	hsa-miR-218-5p (+) hsa-miR-643 (break)	5
rs11539756 (NC_000015.9:g.59429111G>A)	*MYO1E* (Myosin IE)	substitution	G: 0.76 A: 0.24 (N=177)	GG: 0.571 GA: 0.377 AA: 0.051	0.44	hsa-miR-2467-3p (+) hsa-miR-3141 (break) hsa-miR-3158-5p (+) hsa-miR-3616-3p (create) hsa-miR-4736 (create) hsa-miR-622 (-) hsa-miR-4436b-3p (create)	3a
rs1427793 (NC_000012.11: g.106458238T>C)	*NUAK1* (NUAK Family Kinase 1)	substitution	T: 0.77 C: 0.23 (N=177)	TT: 0.617 TC: 0.314 CC: 0.069	0.39	hsa-miR-455-3p (break) hsa-miR-4690-3p (create) hsa-miR-4727-5p (-)	1e
rs139561 (NC_000022.10: g.42195410T>C)	*MEI1* (Meiotic Double-Stranded Break Formation Protein 1)	substitution	T: 0.82 C: 0.18 (N=177)	TT: 0.669 TC: 0.303 CC: 0.029	**0.54**	hsa-miR-4474-3p (break)	No info
rs3134430 (NC_000011.9: g.122683382A>T)	*UBASH3B* (Ubiquitin Associated And SH3 Domain Containing B)	substitution	A: 0.38 T: 0.62 (N=177)	AA: 0.149 AT: 0.457 TT: 0.394	**0.54**	hsa-miR-5093 (-)	1e
rs3134428 (NC_000011.9: g.122682156A>G)	*UBASH3B* (Ubiquitin Associated And SH3 Domain Containing B)	substitution	A: 0.37 G:0.63 (N=177)	AA: 0.149 AG: 0.451 GG: 0.400	0.4	has-miR-338-3p (-) hsa-miR-501-5p (breahashsa-miR-5196-3p (create)	1e
rs895412 (NC_000002.11: g.113973964T>C)	*PAX8* (Paired Box 8)	substitution	T: 0.51 C: 0.49 (N=177)	TT: 0.274 TC: 0.474 CC: 0.251	0.33	hsa-miR-4687-5p (+)	No info

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Supplementary table 3.3b. General characterization of the genotyped candidate variants in 3'UTRs (N=52) provided by extended WES analysis (List 2) (cont.).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs1056667 (NC_000006.11: g.26510564T>C)	*BTN1A1* (Butyrophilin Subfamily 1 Member A1)	substitution	T: 0.48 C: 0.52 (N=177)	TT: 0.223 TC: 0.514 CC: 0.263	0.42	hsa-miR-4694-3p (+) hsa-miR-5093 (-)	No info
rs1061242 (NC_000001.10: g.220142070C>T)	*EPRS11* (Glutamyl-Prolyl-TRNA Synthetase 1)	substitution	C: 0.58 T: 0.42 (N=177)	CC: 0.360 CT: 0.440 TT: 0.200	0.48	hsa-miR-3163 (break)	1e
rs4673336 (NC_000002.11: g.206481573A>G)	*PARD3B* (Par-3 Family Cell Polarity Regulator Beta)	substitution	A: 0.35 G: 0.65 (N=303)	AA: 0.119 AG: 0.472 GG: 0.409	**0.56**	No info	No info
rs8139993 (NC_000022.10: g.41995335T>C)	*DESI1* (Desumoylating Isopeptidase 1)	substitution	T: 0.79 C: 0.21 (N=177)	TT: 0.623 TC: 0.343 CC: 0.034	**0.51**	hsa-miR-3607-5p (create) hsa-miR-3611 (create) hsa-miR-3611 (create)	1e
rs1048169 (NC_000009.11: g.19055965T>C)	*HAUS6* (HAUS Augmin Like Complex Subunit 6)	substitution	T: 0.57 C: 0.43 (N=177)	TT: 0.331 TC: 0.474 CC: 0.194	**0.52**	No modifications	1e
rs16920581 (NC_000008.10: g.57353530T>C)	*PENK* (Proenkephalin)	substitution	T: 0.69 C: 0.31 (N=303)	TT: 0.455 TC: 0.469 CC: 0.076	**0.55**	No modifications	No info
rs143599604 (NC_000003.11: g.5222491_5222492insGGGATAGTAAG)	*ARL8B* (ADP Ribosylation Factor Like GTPase 8B)	indel	-: 0.63 InsGGGATAGT AAG: 0.40 (N=303)	--: 0.409 - ins: 0.442 ins ins: 0.149	No info	hsa-miR-3127-3p (create)	1e
rs2296744 (NC_000006.11: g.33740382G>A)	*LEMD2* (LEM Domain Nuclear Envelope Protein 2)	substitution	G: 0.54 A: 0.46 (N=177)	GG: 0.274 GA: 0.526 AA: 0.200	0.47	hsa-miR-3929 (+) hsa-miR-4298 (-) hsa-miR-4419b (-) hsa-miR-4438 (-) hsa-miR-4478 (+) hsa-miR-873-5p (break)	1e
rs2296743 (NC_000006.11: g.33740025C>T)	*LEMD2* (LEM Domain Nuclear Envelope Protein 2)	substitution	C: 0.71 T: 0.29 (N=177)	CC: 0.474 CT: 0.463 TT: 0.063	0.42	hsa-miR-1913 (break) hsa-miR-1915-5p (create) hsa -miR-324-3p (break) hsa -miR-502-5p (create) hsa-miR-362-5p (create) hsa-miR-500b (create)	1e

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Supplementary table 3.3b. General characterization of the genotyped candidate variants in 3'UTRs (N=52) provided by extended WES analysis (List 2) (cont.).

Variant	Gene	Type	Allelic Frequencies	Genotypic Frequencies	GWAVA score	MirSNP - miRNA binding sites (BS)	RBP binding - RBP-Var2 scores
rs11259151 (NC_000010.10: g.14562793C>T)	*FAM107B* (Family With Sequence Similarity 107 Member B)	substitution	C: 0.55 T: 0.45 (N=177)	CC: 0.269 CT: 0.560 TT: 0.171	0.5	hsa-miR-3916 (-) hsa-miR-583 (+)	5
rs554190572 (NC_000002.11: g.120734737_120734738insA)	*PTPN4* (Protein Tyrosine Phosphatase Non-Receptor Type 4)	indel	A: 0.71 dupA: 0.29 (N=303)	AA: 0.102 A dupA: 0.383 dupA dupA: 0.515	No info	No info	3a
rs11676058 (NC_000002.11: g.120736882T>C)	*PTPN4* (Protein Tyrosine Phosphatase Non-Receptor Type 4)	substitution	T: 0.72 C: 0.28 (N=303)	TT: 0.545 TC: 0.356 CC: 0.099	No info	No info	6
rs35508493 (NC_000022.10: g.41574973_41574975delGTA)	*EP300* (E1A Binding Protein P300)	indel	GTA: 0.72 delGTA: 0.26 (N=303)	GTA GTA: 0.512 GTA delGTA: 0.422 delGTA delGTA:0.066	No info	hsa-miR-300 (break)	5
rs6633 (NC_000012.11: g.123745809G>T)	*CDK2AP1* (Cyclin Dependent Kinase 2 Associated Protein 1)	substitution	G:0.80 T:0.20 (N=177)	GG: 0.623 GT: 0.354 TT: 0.023	**0.64**	hsa-miR-3616-3p (enhance) hsa-miR-4503 (decrease) hsa-miR-4670-3p (create) hsa-miR-548p (decrease)	

(+) binding increase; (-) binding decrease; Scores in bold indicates a deleterious effect.

Supplementary Table 3.4. List of the variants that were in *Linkage Disequilibrium* and presented a D' > 0.98 and a *p* value < 0.05.

Linkage Disequilibrium		
Variant (Gene)	*LD* with variant(s) (Gene)	*D'*
rs3812433 (*GLI4*)	rs7387207 (*GLI4*)	0.99
rs2296744 (*LEMD2*)	rs4711351 (*LEMD2*)	1
rs1595065 (*ERBB4*)	rs34217661 (*ERBB4*)	0.99
rs1972820 (*ERBB4*)	rs34217661 (*ERBB4*)	0.99
rs1595065 (*ERBB4*)	rs1836724 (*ERBB4*)	0.99
rs35117663 (*C22orf46*)	rs2187795 (*C22orf46*)	0.99
rs739135 (*C22orf46*)	rs2187795 (*C22orf46*)	1
rs739135 (*C22orf46*)	rs35117663 (*C22orf46*)	0.99

Chapter 4: Epigenome-wide DNA methylation profiles in whole blood of Spinocerebellar ataxia type 3/ Machado-Joseph disease patients: a pilot study

4.1. Introduction

DNA methylation (DNAm) is an epigenetic mechanism that consists in the addition of a methyl group frequently to a cytosine in cytosine-phosphate-guanine sequence [276]. This mechanism plays a crucial role in different biological processes, such as tissue-specific gene expression regulation, silencing of repetitive DNA sequences, genomic imprinting and inactivation of the X chromosome [184,276]. To understand the functions of DNAm, we have to consider the distribution of methylation across the genome; the genomic location of DNAm may lead to a different impact on gene expression [277]. DNAm in genes bodies is usually associated with gene transcription, elongation and splicing whereas when methylation occurs in gene promoters it may lead to a decrease in its expression [276].

Abnormal patterns of DNAm have been associated with the pathophysiology of neurodegenerative diseases influencing their clinical and pathological expression, such as age at onset (onset) and pathological burden in brain regions affected by these diseases [276,278,279]. In the neurodegenerative diseases belonging to the polyglutamine (polyQ) group, DNAm patterns have also been associated to repeat instability [280]; it is also important to understand the impact of DNAm changes in other aspects of these diseases, such as variable disease expressivity and transcriptional changes. Throughout the years, different studies on DNAm were performed in Huntington disease (HD) and a few studies in Spinocerebellar ataxias (SCAs). In HD, studies were performed in *post-mortem* brain samples of patients, in cell and animal models and in whole blood [281–285]; overall, these studies revealed that DNAm levels presented profound changes: i) a study that analysed motor progression of HD patients and in animal model, three loci (*PEX14 GRIK4*, and *COX4I2*) revealed an association between their methylation levels and motor progression [281]; ii) a study targeting global DNAm reported elevated methylation values [282]; iii) in a study with DNA from whole-blood samples, 326 differentially methylated genes were identified between patients and controls [283]; iv) a study with brain samples, led to the establishment of a DNA methylation profile [284]; moreover, a correlation between onset and DNAm in tissues from cerebral cortex was identified. In this research, *HTT* gene locus presents tissue specific DNA methylation patterning; thirty-eight sites within the *HTT* locus were differentially methylated between matched cortex and liver samples [284]. v) in a study with culture cells, it was reported the hypermethylation of the *APEX1* promoter in HD cells when comparing to controls [285]. In SCAs, few studies in blood were performed concerning DNAm, mainly for SCA1, SCA2 and SCA3 [188,282,286–289]. Until the moment

of the writing of this thesis, no methylation studies in *post-mortem* brain samples of SCAs patients were published. Hamzeiy and colleagues [282], in a study of DNA global methylation in blood, reported that SCA1 and SCA2 patients presented a significant increase in global methylation, when compared to sex-and age matched healthy subjects, and after controlling for lifestyle factors such as smoking, alcohol consumption and physical activity. For SCA2, another study revealed that CpG methylation in *ATXN2* gene promotor was associated with disease onset in SCA2 homozygotes [287]. SCA3 methylation studies will be presented later in this section.

DNAm pattern changes in the genome, during the development due to a dynamic process involving new DNA methylation and demethylation [208]; consequently, cells that were differentiated develop an unique and stable DNAm pattern that will regulate gene transcription specific for each tissue [208]. DNAm patterns also changes with aging and with physiological and environmental *stimuli* [208]. Indeed, the lifestyle of a subject is known to have an impact on DNAm patterns; factors like smoke, alcohol, medication, and other diseases can modify these patterns [208].

Despite the increasing number of studies about DNAm and its changes in human blood and brain, reliable results require the fulfilment of very specific conditions. Storage conditions, the years of storage until the DNA extraction, age at collection (chronological age) of the biological tissue (blood, brain, etc.), the availability of human *post-mortem* brain tissue and information about the lifestyle of the participants and/or donors are crucial to avoid biased results [290–294].

Although chronological age (the age of the subject when the sample is collected – age at collection) is stable among individuals, the same is not applicable to biological age (the age that each body seems to have according with different factors such, molecular changes and lifestyle) which varies and can predict the aging acceleration or deacceleration of those individuals [295]. Different epigenetic clocks are available for different tissues and were applied in different studies and diseases [189,190,210,281,296–303]; these clocks estimate DNAm-age as a proxy of biological age, individuals can be predicted as biologically older or younger than their chronological age. Examples of epigenetic clocks applicable to blood samples include Horvath Pan-tissue epigenetic clock [190] and Hannum blood based epigenetic clock [210]. A higher biological aging might be a reflection of disease, morbidity and mortality [295].

Epigenetic studies for SCAs are scarce. For spinocerebellar ataxia type 3 (SCA3), two small methylation studies analysed the promoter of the *ATXN3* gene [186,288] revealed that methylation levels can have a modulator effect on the variance of onset of SCA3. In one of the studies, conducted by Emmel and colleagues, a methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA) was performed to assess the methylation degree at six CpG dinucleotides at the promotor of SCA3 causal gene [288], which presents two large CpG islands. In the other study, performed by Wang and colleagues [186], the levels of methylation of the two CpG islands were analyse by pyrosequencing. Ding and colleagues [187] studied polymorphisms in DNAm related genes and observed an association with $(CAG)_n$ instability also modulating the variance in onset in SCA3. More recently, the same group [286], using a mouse model (three transgenic mice and three wild-type mice), identified two DNAm genes, whose methylation status of differential methylated regions were negatively associated with their levels of transcription: *En1* (*Engrailed Homeobox 1*) and *Nkx2-1* (*NK2 Homeobox 1*); these genes may be useful for understanding the pathogenesis of this disease. Zhao and colleagues [188] performed a study with monozygotic twins that revealed differences on methylation levels at nine CpG sites between the individuals. More recently, Li and colleagues [189] suggested that DNAm age acceleration, which in this study was estimated as the difference between DNAm age and chronological age can be a modifier of SCA3 onset; an increase in DNAm age leading to an earlier onset [189].

Currently, there is no study that compares genome wide DNAm in SCA3 patients with controls. Therefore, the aims of this study were to generate genome wide DNAm data from whole-blood samples of SCA3 patients and paired-matched healthy control and to explore the presence of differentially methylated positions that can define a SCA3-specific blood-based methylation profile. Moreover, we aimed to identify the main genes/pathways involved in methylation alterations observed in SCA3. Lastly, we wanted to explore the existence of age acceleration in SCA3, as estimated by several epigenetic clocks.

4.2. Subjects and methods

Subjects

Genomic DNA from peripheral blood of sixteen samples, eight from Azorean individuals with SCA3 (patients) and eight from matched healthy Azorean (controls), was used (Table 4.1).

SCA3 patients were clinically and molecularly confirmed as carriers of the expansion at the *ATXN3* gene. The sizing of CAG tract for normal and expanded alleles was performed according to Bettencourt and colleagues [219]. Patients selected had a disease duration ranging from zero to five years; seven patients were non-smokers, and one was a smoker. Age at onset (onset) was considered the age of the first manifestation of the disease (gait ataxia or diplopia), reported by each patient and/or caregiver.

Controls were free from neurological-movement diseases and selected from non-SCA3 families and non-carriers of the *ATXN3* CAG repeat expansion. Patients and controls were matched for age, gender, smoker status and approximate DNA storage time. Age at collection (chronological age) was considered the age of the subjects (patients or controls) at the time of blood collection.

Table 4.1. General demographic characterization of the SCA3 patients and paired-matched controls used in the present study.

	N	Gender (F/M)	Onset (in years)	CAG_{exp}	Chronological age (in years)*
Patients	8	5 / 3	37.25±3.54	69.88±3.93	39.50±3.62
Controls	8	5 / 3	NA	NA	39.25±4.53
Total	16	10 / 6	NA	NA	39.38±3.96

NA- not applicable

*Chronological age - average age in years (mean±standard error)

Infinium EPIC BeadChip methylation profiling and quality control

Genomic DNA concentration was determined using a Qubit 1™ Fluorometer with dsDNA BR Assay Kit™ for Qubit. A bisulfite conversion was performed with 500ng of DNA and using the EZ DNA Methylation Kit (Zymo Research), according to the manufacturer instructions. Genome-wide DNA methylation profiling was performed at UCL Genomics, using the Infinium *HumanMethylationEPIC* BeadChip array (EPIC array; lllumina), which covers over 850,000 methylation sites throughout the genome; the covered regions can

include gene bodies, gene promoters and regulatory intergenic regions. In these regions different CpG sites can be found: i) islands, which are DNA regions rich in CpG, and are often associated with transcription starting site of a gene; can be also found in intergenic regions and gene bodies [277]; ii) open seas, that are regions located more than 4kb from CpG islands [304]; iii) shores are the regions immediately flanking CpG islands [277]; and iv) shelves are regions <2kb flanking outwards from a CpG shore these CpG sites were distributed across different genomic regions [277] Beta values obtained were used to estimate the methylation levels of each CpG site using the ratio of intensities between unmethylated and methylated alleles [278]; Beta values presented a range from 0 to 1, which represents approximately 0% to 100% methylation, respectively. For all statistical analysis M values (logarithmic transformation of the beta values) were used.

The EPIC array DNAm data was analysed using different R Bioconductor packages. The obtained raw data (idat files) were imported and pre-processed with minfi [305] and ChAMP packages [306]. Quality control was performed with minfi [305], watermelon [307] and ChAMP packages [306]. Probes with one or more of the following criteria were excluded: a) poor quality; b) cross reactive; c) included common genetic variants; and d) mapped to sex chromosomes. Blood cell type heterogeneity beta matrix and cell-type specific proportions corrections were performed using the ChAMP package [306].

Differential methylation analysis

Determination of differentially methylated positions (CpGs) between SCA3 patients and controls was performed as previously described [278]. Briefly, a multiple linear regression model was applied to identify associations between DNAm changes that occur at specific CpG sites and SCA3 status; possible confounding factors such as age, sex, and surrogate variables were taken in account in this model. Limma package [308] was used for analysis differentially methylated CpG sites; p-values <0.05 were used to select differentially methylated CpGs in the absence of genome-wide significant CpGs.

Identifying the most promising genes with differentially methylated CpGs sites

To obtain a list of genes that can present CpGs differently methylated, CpGs were ranked by an arbitrary lower nominal *p-value* ($p < 0.005$). The genes with CpGs which nominal *p-value* was lower than 0.005 were selected for functional enrichment analysis.

Enrichment analysis were performed in gProfiler [309,310]; significance Benjamini-Hochberg-FDR threshold of 0.05 were considered.

DNA methylation (DNAm) age estimates

To obtain DNAm age for each sample, epigenetic age was calculated using Horvath and colleagues tutorial and online tool [190] (https://dnamage.genetics.ucla.edu); two epigenetic clocks were used to obtain DNAm age: Horvath Pan-Tissue epigenetic clock [190], and Hannum Blood Based epigenetic clock [210]. A linear regression model was applied and DNAm age according to Horvath and Hannum epigenetic clocks was regressed on chronological age (age at collection). Age acceleration residuals was defined as the corresponding raw residuals. Age acceleration difference was calculated by subtracting chronological age to DNAm age obtained for each epigenetic clock.

A Mann-Whitney test was performed to evaluate differences between SCA3 patients and controls for: chronological age and DNAm age obtained for each of the considered epigenetic clocks. Spearman's correlations were performed between DNAm age for each epigenetic clock and chronological age (patients and controls), chronological age and disease duration (only patients), DNAm age for Horvath and Hannum epigenetic clocks and onset (patients), and age acceleration difference for Horvath and Hannum epigenetic clocks and chronological age (patients and controls).

Statistical analyses were performed in *IBM SPSS Statistics v.25.0*; *p-value*s lower than 5% were considered statistically significant.

4.3. Results

After stringent quality control, we obtained results for 724,096 CpG sites covering a total of 199,620 CpG sites mapping to genes. A nominal arbitrary *p* value of 0.005, was used as a cut-off for selection of the differentially methylated CpG sites. The resulting list contained 1398 CpG sites distributed by 1279 different genes (Figure 4.1).

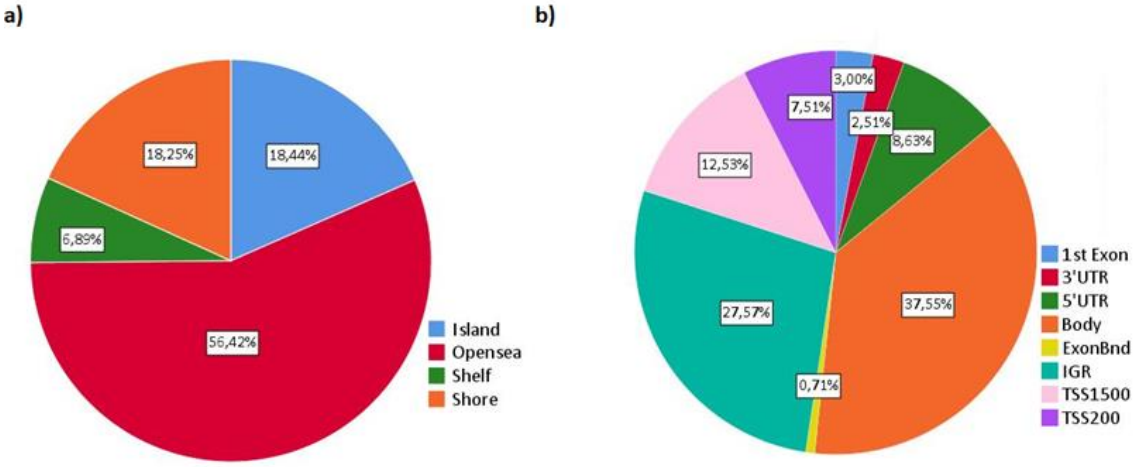


Figure 4.1. Distribution of the differentially methylated CpGs sites according to a) classification of the percentage of islands, open seas, shelves and shores was described; b) regions: 1st Exon, represents the first exon of a gene with CpG sites; 3'UTR and 5'UTR represents untranslated regions in 3' and 5', respectively; body, represents the part of genes between untranslated regions; ExonBnd represents exon boundaries; IGR, represents intergenic regions; TSS1500, represents 200-1500 bases upstream of the transcription start site; and TSS200, represents 0-200 bases upstream of the transcriptional start site.

The majority of CpGs sites are located at open seas (56.24%), with shores and islands presenting almost the same number of CpG sites (18.25% and 18.44%, respectively), whereas the lowest number of CpG sites are shelves (6.89%) (Figure 4.1 a)). Only 27.57% of the CpG sites are located in IGR, whereas the remaining (72.43%) are in gene related elements (Figure 4.1.b)).

From the top ranked 1398 CpG sites (Figure 4.1b), a total of one hundred and one (8.63%) mapped to 5'UTRs; these CpG sites were located in a total of ninety-six genes. In the gene bodies it was possible to obtain a total of 718 CpG sites (37.55%) in 662 genes.

Relatively to CpG sites located in 3'UTRs, a total of thirty-seven was mapped in thirty-seven genes (Figure 4.1b).

For *ATXN3*, the causative gene of SCA3, from a list of 24 CpG sites remaining after quality control, none presented a nominal *p* value lower than 0.005. However, we still characterised these as it may provide the ground for further studies (Figure 4.2, Table 4.1).

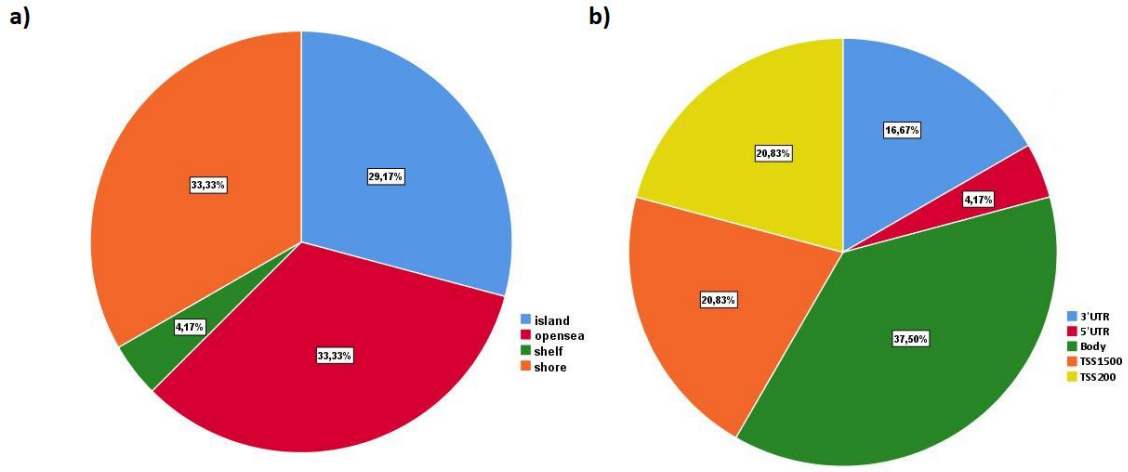


Figure 4.2. Distribution of the CpGs sites for *ATXN3* according to a) classification of the percentage of islands, open seas, shelves and shores; b) regions: 3'UTR and 5'UTR represents untranslated regions in 3' and 5', respectively; body, represents the part of genes between untranslated regions; TSS1500, represents 200-1500 bases upstream of the transcription start site; and TSS200, represents 0-200 bases upstream of the transcriptional start site.

Wang and colleagues [186] reported that the SCA3 causal gene presents two CpG islands in the promotor. From the twenty-four CpG sites found in *ATXN3*, the majority is located in gene body (37.50%); it is of note that the remaining CpG sites are localized in UTRs and TSSs (transcription starting sites). Noteworthy, only one CpG site which passed quality control is located at the 5'UTR of *ATXN3*. A detailed characterization of the 24 CpG sites for *ATXN3* is presented in table 4.1.

Table 4.1. Characterization of the 24 CpGs sites at *ATXN3* gene, after quality control.

CpGs	Map Info*	Feature**	CGI***
cg02013035	92572273	Body	shore
cg10500579	92572856	Body	island
cg26758642	92573987	TSS1500	shore
cg09996633	92566961	Body	opensea
cg02777278	92526520	3'UTR	opensea
cg09972424	92573067	TSS200	island
cg07188233	92573147	TSS200	shore
cg12939335	92525886	3'UTR	opensea
cg20089531	92573973	TSS1500	shore
cg04896976	92526934	3'UTR	opensea
cg12034871	92525005	Body	opensea
cg26081025	92570361	Body	shelf
cg06220504	92564574	Body	opensea
cg02028524	92573410	TSS1500	shore
cg00355673	92573075	TSS200	island
cg16135395	92573832	TSS1500	shore
cg09458170	92572575	Body	island
cg06216926	92572917	5'UTR	island
cg12027153	92572993	TSS200	island
cg12671658	92572097	Body	shore
cg04314597	92559815	Body	opensea
cg27159443	92526015	3'UTR	opensea
cg07468777	92573175	TSS1500	shore
cg11817993	92572978	TSS200	island

*Map info is the location of the CpG site in chromosome.

**Feature is the region where the CpG site is located in the genome.

***CGI is constituted by CpG sites grouped within clusters or islands.

The enrichment analysis, performed in g:Profiler for the genes containing the top ranked 1398 CpG sites revealed an enrichment for pathways of Morphine addiction, G protein signaling, PI3K-Akt signaling pathway, Inflammatory mediator regulation of TRP channels, amongst others (Table 4.2).

Table 4.2. Enrichment analysis performed in gProfiler. KEGG: Kyoto encyclopedia of genes and genome; WP: Wiki-Pathways. 1398 CpGs at CpG sites differentially methylated at an adjusted *p* value of 0.05.

Source	Pathway	ID	Adjusted *p* value
KEGG	Morphine addiction	KEGG:05032	0.0079
KEGG	PI3K-Akt signaling pathway	KEGG:04151	0.0100
KEGG	Inflammatory mediator regulation of TRP channels	KEGG:04750	0.0100
KEGG	Rap1 signaling pathway	KEGG:04015	0.0138
KEGG	Focal adhesion	KEGG:04510	0.0294
KEGG	ECM-receptor interaction	KEGG:04512	0.0427
KEGG	Endocrine and other factor-regulated calcium reabsorption	KEGG:04961	0.0445
KEGG	Calcium signaling pathway	KEGG:04020	0.0445
KEGG	Inositol phosphate metabolism	KEGG:00562	0.0445
KEGG	Platelet activation	KEGG:04611	0.0445
WP	G protein signaling pathways	WP:WP35	0.0004
WP	mBDNF and proBDNF regulation of GABA neurotransmission	WP:WP4829	0.0204
WP	Hippo signaling regulation pathways	WP:WP4540	0.0204

The result of our enrichment analyses also revealed that despite we were analyzing results from DNA present in blood samples, it was possible to obtain some terms related with brain (Table 4.3 and Table 4.4.).

The result of enrichment analyses provided by GO (gene ontology) which analyses aimed to perform the identification of biological processes, cellular locations as well as molecular functions that impact the condition that is studied. This analysis showed that for molecular function protein binding, cytoskeletal protein binding amongst others presented a significant result.

Table 4.3. Results of enrichment analysis performed in gProfiler. GO: gene ontology: MF: molecular function; BP: binding protein; CC: cellular component

Source	Name	ID	Adjusted *p* value
GO:MF	protein binding	GO:0005515	6.237900820797352e-28
GO:MF	cytoskeletal protein binding	GO:0008092	0.000018769236423520046
GO:MF	actin binding	GO:0003779	0.0000696710331577782
GO:MF	identical protein binding	GO:0042802	0.000269658405234293
GO:MF	ion binding	GO:0043167	0.0003402401534337794
GO:MF	transmembrane transporter binding	GO:0044325	0.0010101733304248635
GO:MF	binding	GO:0005488	0.0011778027402245788
GO:MF	protein domain specific binding	GO:0019904	0.0019269357874824383
GO:MF	signaling adaptor activity	GO:0035591	0.0023401866513908134
GO:MF	kinase activity	GO:0016301	0.0030242621476718307
GO:MF	adenyl nucleotide binding	GO:0030554	0.004492010128228928
GO:MF	adenyl ribonucleotide binding	GO:0032559	0.004759582067217344
GO:MF	phosphotransferase activity, alcohol group as acceptor	GO:0016773	0.0050643972752335755
GO:MF	ATP binding	GO:0005524	0.00669784210316821
GO:MF	actin filament binding	GO:0051015	0.00669784210316821
GO:MF	phosphatidylinositol binding	GO:0035091	0.00669784210316821
GO:MF	lipid binding	GO:0008289	0.008796261753088492
GO:MF	metal ion binding	GO:0046872	0.009332857960146585
GO:MF	transferase activity, transferring phosphorus-containing groups	GO:0016772	0.010942308133446422
GO:BP	regulation of cellular process	GO:0050794	5.4463027265219e-18
GO:BP	developmental process	GO:0032502	2.4669441125189323e-16
GO:BP	anatomical structure development	GO:0048856	1.4013683649506866e-15
GO:BP	multicellular organismal process	GO:0032501	1.973168237510035e-15
GO:BP	multicellular organism development	GO:0007275	1.99067015921706e-13
GO:BP	biological regulation	GO:0065007	3.1108483756614157e-13
GO:BP	cell communication	GO:0007154	3.1108483756614157e-13
GO:BP	system development	GO:0048731	3.733163210297515e-13
GO:BP	signaling	GO:0023052	2.0297539040311744e-12
GO:BP	regulation of biological process	GO:0050789	8.274448323734957e-12

Noteworthy for GO the four top enrichment results for binding proteins were regulation of cellular process, developmental process, anatomical process and multicellular organism development.

Table 4.4. Results of enrichment analysis performed in gProfiler. BP: binding protein; CC: celullar component.

Source	Name	ID	Adjusted *p* value
GO:BP	cellular developmental process	GO:0048869	1.6195967593631977e-11
GO:BP	localization	GO:0051179	1.7321387823714578e-11
GO:BP	cell differentiation	GO:0030154	1.7321387823714578e-11
GO:BP	cellular response to stimulus	GO:0051716	5.2451025701461626e-11
GO:BP	regulation of biological quality	GO:0065008	9.229191606828542e-11
GO:BP	regulation of localization	GO:0032879	8.514409440630333e-10
GO:BP	nervous system development	GO:0007399	1.2313933740209654e-9
GO:BP	regulation of cellular component organization	GO:0051128	1.2313933740209654e-9
GO:BP	signal transduction	GO:0007165	1.4172152459127372e-9
GO:BP	movement of cell or subcellular component	GO:0006928	2.952793457774442e-9
GO:CC	cytoplasm	GO:0005737	5.713981953894149e-16
GO:CC	cytosol	GO:0005829	2.5036764128020284e-14
GO:CC	cell junction	GO:0030054	3.735019558785243e-14
GO:CC	cell periphery	GO:0071944	3.735019558785243e-14
GO:CC	plasma membrane	GO:0005886	5.937015052028292e-13
GO:CC	cell projection	GO:0042995	1.6617898195054043e-12
GO:CC	plasma membrane region	GO:0098590	3.934160947015692e-12
GO:CC	plasma membrane bounded cell projection	GO:0120025	2.4463360961438608e-11
GO:CC	membrane	GO:0016020	1.816998196754985e-10
GO:CC	neuron projection	GO:0043005	2.067620928680382e-9
GO:CC	synapse	GO:0045202	2.2624150906605488e-9
GO:CC	anchoring junction	GO:0070161	8.049537600891896e-8
GO:CC	nucleoplasm	GO:0005654	0.00000280941825612688
GO:CC	axon	GO:0030424	0.000003289202573163949 3
GO:CC	cell projection membrane	GO:0031253	0.000004687782514317666
GO:CC	actin cytoskeleton	GO:0015629	0.000011776074200100058
GO:CC	somatodendritic compartment	GO:0036477	0.00002049389597001827
GO:CC	ruffle	GO:0001726	0.00003538239654681194
GO:CC	vesicle	GO:0031982	0.000050266626038260473
GO:CC	actin-based cell projection	GO:0098858	0.00007440261100654966

Interestingly, results by cellular components show that cytoplasm, cytosol, cell junction and cell periphery presented the lower *p*-value; also, and despite our DNA is from blood samples, we have obtained some cellular components related with nervous system, namely, neuron projection and synapse.

The epic array used in the recent study of Li *et al.* [189] include all CpGs which are used to provide information concerning methylation age (Epigenetic clocks). In the analysis of our small set of subjects, we detect no differences in chronological age between SCA3 patients and controls (p = 0.874) as expected, since our controls were age matched with SCA3 patients. When we compared DNAm age for Horvath and for Hannum epigenetic clocks between SCA3 patients and controls, no differences were observed between the two groups (Figure 4.3).

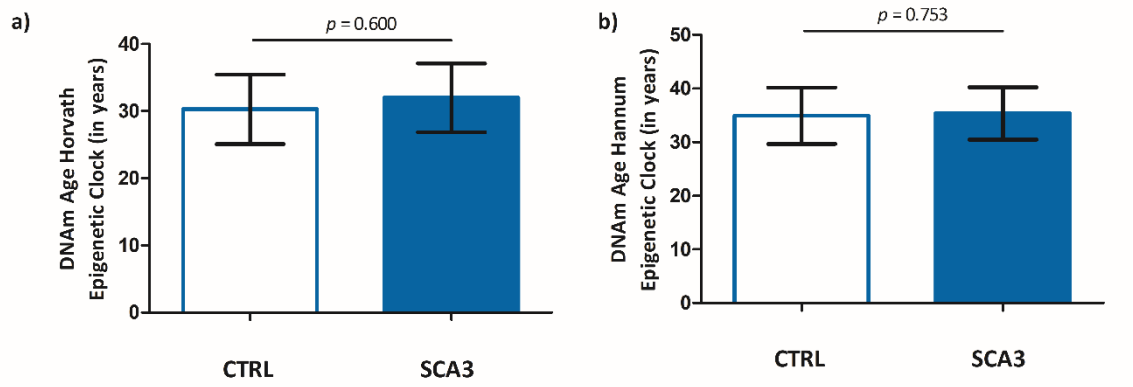


Figure 4.3. Comparisons between SCA3 patients (SCA3) and controls (CTRL) by a) DNAm age estimated from Horvath epigenetic clock; and b) DNAm age estimated obtained from Hannum epigenetic clock.

Since that were no differences for DNAm age in SCA3 patients and controls, we pooled together the two groups and correlated DNAm age obtained for each epigenetic clock with chronological age of the subjects (Figure 4.4).

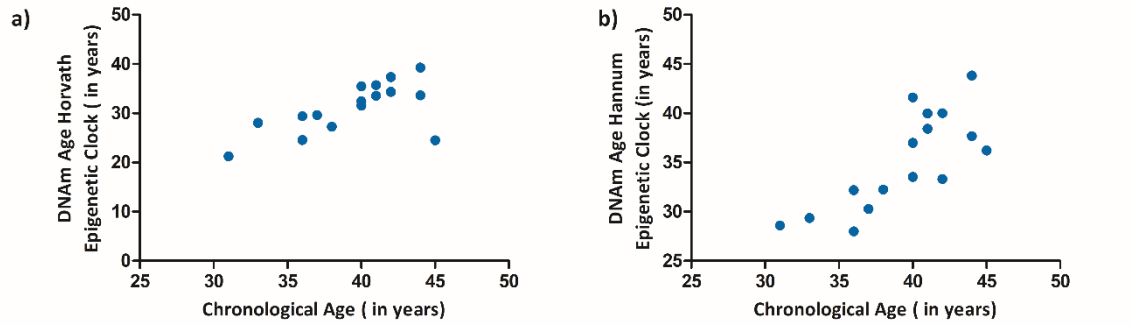


Figure 4.4. Correlation, with SCA3 patients and controls pooled together, between DNAm age for a) Horvath epigenetic clock and chronological age (rho = 0.587, p = 0.017) and b) Hannum epigenetic clock and chronological age (rho = 0.750, p= 0.001).

As expected, a positive correlation was found between DNAm age and chronological age for both clocks, the Hannum epigenetic clock being the one with a higher correlation. This result is in agreement with the fact that this epigenetic clock is specific for blood samples.

For age acceleration difference, no significant differences were obtained between SCA3 patients and controls for both clocks (Figure 4.5). Noteworthy in both groups, methylation age was lower than chronological age providing negative values of age acceleration.

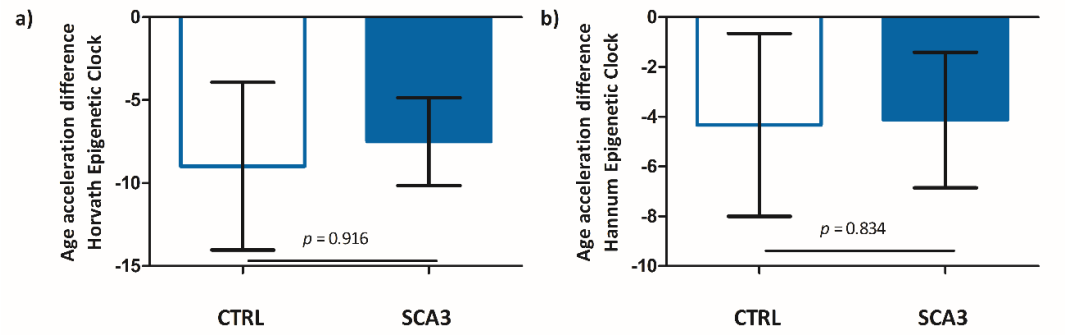


Figure 4.5. Comparisons between SCA3 patients (SCA3) and controls (CTRL) by a) Age acceleration difference from Horvath epigenetic clock; and b) Age acceleration difference from Hannum epigenetic clock.

In the case of age acceleration difference residuals, no differences were found between SCA3 patients and controls for both clocks (Figure 4.6).

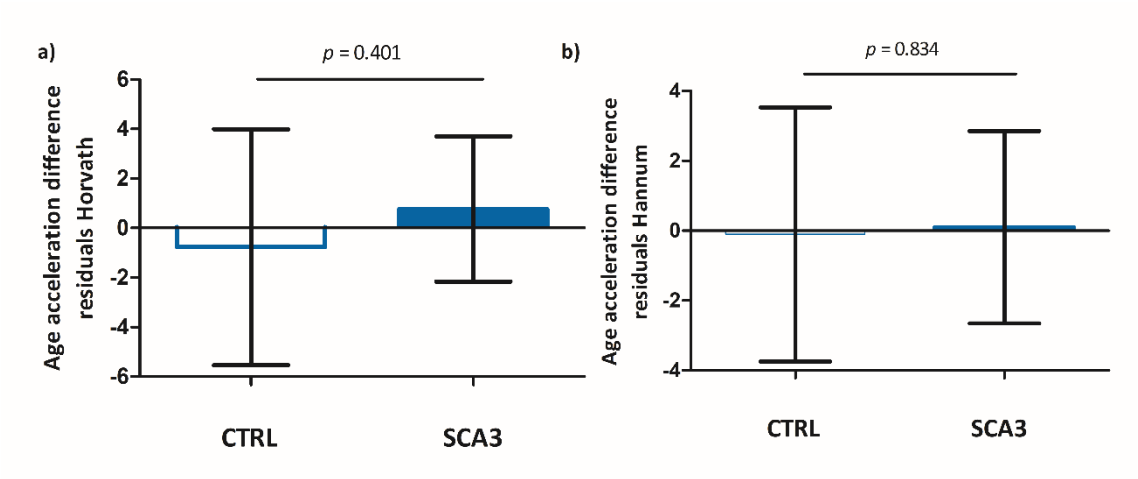


Figure 4.6. Comparisons between SCA3 patients (SCA3) and controls (CTRL) by a) Age acceleration difference residuals from Horvath epigenetic clock; and b) Age acceleration difference residuals from Hannum epigenetic clock.

In summary we were not able to detect differences between SCA3 patients and controls, for both measures of age acceleration. However, for the case of age acceleration residuals we observed a tendency for an acceleration in SCA3 patients for both clocks.

4.4. Discussion

This genome-wide DNA methylation pilot study is the first one reporting results from an EPIC array in SCA3 patients and in matched controls.

From the 1279 genes obtained for the ranking of CpG sites differentially methylated the enrichment analysis highlighted three pathways of particular interest: G protein signaling, inflammatory mediator regulation of TRP channels and Calcium signaling pathway.

G protein-coupled receptors (GPCRs), involved in the G protein signaling pathway were reported to be implicated in the pathogenesis of neurodegenerative diseases, such as Alzheimer's, Parkinson's and Huntington's diseases [311]. Several of the non-sensory GPCRs can be expressed in brain playing an important role in basic features for human survival such as appetite, pain or vision but also plays a role in the immune regulation and synaptic transmission [311,312]. As GPCR ligands include several molecules of different categories such as lipids, growth factors, hormones, peptides, ions, photons and biogenic amines, GPCRs have a high potential as targets for therapeutic applications [311].

TRP channels, which are regulated by the inflammatory mediator regulation of TRP channels, act directly on cations intracellular levels and indirectly modulate several intracellular pathways [313]. These channels were reported to be involved in cytokine production, cell differentiation and cytotoxicity [313]. As they have contributed to the transition of immune responses and inflammation, their study in SCA3 can be of interest.

Calcium signaling pathway was proposed as a common mechanism in the pathogenesis of different SCAs [314]. A study conducted by Bettencourt and colleagues [314], presented data that support this hypothesis and showed the importance of this pathway as it may explain selective neuropathology and are relevant for Purkinje cells. These cells are important targets in the pathogenesis of SCAs, being the signaling of calcium important to their normal function [314]. Also, abnormal levels of calcium in Purkinje cells are thought to activate toxic cascades that can lead to cellular death.

Interestingly, the Morphine addiction pathway presents a low *p-value*. Opioid receptors constitute a group of G protein-coupled receptors and its activation by ligands induces significant molecular modifications in the cell, such as the activation of potassium channels and the reduction of calcium conductance, amongst others [315]. To the best of our knowledge, the role of opioids in SCA3 is not yet explored.

DNAm has already been the target of a few studies in SCA3, as already mentioned. In particular, a recent study from Li and colleagues [189], for genome-wide DNAm approach to samples from SCA3 patients using the same array we have used in this work. However, there are some constrains in their study: a) they failed to include controls subjects and therefore were not able to provide data on the blood based methylome profile of SCA3; and b) they have not taken in consideration any kind of lifestyles factors such as smoking, and therefore their results might be biased by such factors. Using a set of 40 patients, Li and colleagues [189] have reported the existence of age acceleration in SCA3.

In our study, we were not able to identify differences for DNAm age estimated between patients and controls for both epigenetic clocks that were applied. Although this can be due to our sample size, it is unclear whether the lack of consideration for the smoking habits in the study of Li and colleagues [189], can have biased their results. This can only be confirmed in further studies with a higher number of DNA samples from individuals with detailed lifestyle factors. Such studies are currently being planned in our group. Also, the study of DNAm patterns in *post-mortem* brain samples of SCA3 patients is mandatory since these patterns should be tissue specific. Such studies are currently being developed in our group.

4.5. Conclusions

In summary**,** this study provided an overview of DNA methylation variation in SCA3 patients and matched controls. As no significant results were found for the first genome-wide study on DNA methylation comparing patients and controls, further studies with a higher number of patients and controls should be conducted to obtain a methylation profile for SCA3.

4.6. Supplementary material

Supplementary Table 4.1. DNA methylation (DNAm) age according to Horvath and Hannum and colleagues epigenetics clocks for blood DNA samples from SCA3 patients and controls. NA: not applicable.

ID	Status	Onset	Chronological Age	DNAm Age Horvath	Age Acceleration Difference Horvath	DNAm Age Hannum	Age Acceleration Difference Hannum
C18	Control	NA	31	21,265	-9,735	28,614	-2,386
DMJ118	Patient	38	40	31,573	-8,427	33,529	-6,471
DMJ092	Patient	39	41	33,552	-7,448	38,425	-2,575
FL20	Control	NA	36	29,395	-6,605	28,007	-7,993
DMJ086	Patient	37	42	34,364	-7,636	33,328	-8,672
C93	Control	NA	37	29,628	-7,372	30,284	-6,716
C91	Control	NA	41	35,739	-5,261	39,992	-1,008
DMJ127	Patient	42	42	37,371	-4,629	40,018	-1,982
DMJ140	Patient	32	33	28,073	-4,927	29,369	-3,631
C37	Control	NA	40	35,491	-4,509	36,981	-3,019
C16	Control	NA	40	32,441	-7,559	41,612	1,612
DMJ089	Patient	33	36	24,573	-11,427	32,187	-3,813
FL7	Control	NA	44	33,662	-10,338	37,671	-6,329
DMJ093	Patient	36	38	27,279	-10,721	32,248	-5,752
C27	Control	NA	45	24,518	-20,482	36,222	-8,778
DMJ098	Patient	41	44	39,282	-4,718	43,821	-0,179

Chapter 5: General conclusions

5.1. General conclusions

The search for the genetic basis of disease, namely in what concerns genetic modifiers can be facilitated with the use of a more homogeneous cohort [101]. In the present thesis, the Azorean SCA3 cohort, due to its features, such as a regular follow-up of patients, which result in a rigorous date of onset, together with the registration of the lifestyle components such as smoking habits, was studied to identify genetic and epigenetic modifiers in SCA3.

To perform these studies, together with the samples from the Azorean SCA3 cohort, samples from other cohorts (Mainland Portugal, United Kingdom, and Brazil) were added.

The analysis of variation in *ATXN3* besides the CAG repeats, is an approach to find genetic modifiers. In Chapter 2, we reported nine variants presented in the 3'UTR of *ATXN3* gene, which can have a role in its own expression regulation. The predicted functional impacts for these variants support the complexity and importance of the 3'UTR regulation on *ATXN3* expression. However, none of the variants that were identified presented a significant modulatory effect on the disease onset. Moreover, when we compare our results for rs709930 (and rs910369) in *ATXN3* gene, with the ones reported by Long and colleagues [163], we were not able to replicate the result of an earlier onset in our patients whose alternative allele was in *cis* with expanded CAG tract. In the study of Long and colleagues [163], the percentage of patients carrying the alternative allele for rs709930 was more than two-fold than in our cohort of patients. This population-specific differences in allelic frequencies can have increased the power to detect an effect in the Chinese cohort of patients, a fact that we also reported for other modifier genes of the SCA3 onset. Noteworthy, we were able to identify disease-associated haplotypes that can discriminate the expanded allele; this fact can be relevant for the design of new allele-specific silencing strategies. Various factors may limit the statistical power to detect significant associations between the presence of the nine variants and SCA3 onset for the disease, such as small sample effect size. Further studies with additional cohorts of SCA3 patients would be needed to confirm these results.

In Chapter 3, we obtained associations between the presence of some variants in the UTRs of different genes and SCA3 disease onset, when using a multicenter cohort of SCA3 patients. The four different cohorts from the Azores, Mainland Portugal, United Kingdom, and Brazil, were pooled together for genotype-phenotype correlations, adjusting for the CAGs in the expanded allele. This analysis allowed us to identify four candidate

modifier genes of the disease onset: *GLI4*, *TGFA*, *ERBB4* and *MPHOSPH9*. *GLI4*, presented two variants in 5'UTR with a predicted modification in transcription factors binding sites and in transcription factors that bind to those sites, when the alternative allele is present. This gene also showed a delay of three years in SCA3 onset, in the genotype-phenotype analysis. *GLI4* has not been reported as a direct or indirect interactor of *ATXN3,* but this gene can be involved in reported transcription regulation [269,270]. Noteworthy, a study conducted by P. Maciel Lab (personal communication) has revealed that the silencing of this gene leads to the amelioration of a *C. elegans* transgenic model phenotype, which could be an indicator that indeed this gene could be related with SCA3 pathogenesis. More studies will be needed to understand the impact of this gene. The remaining candidate modifier genes (*TGFA*, *ERBB4* and *MPHOSPH9*) identified in this study, which presented variants in 3'UTRs are known *ATXN3* interactors, and significantly influence the onset of the disease. Variants in *TGFA*, a cytokine involved in the regulation of reactions of immune system [271] and have a role in the inflammation process as described for other cytokines expression levels, in different SCA3 studies [174,255,256,272], are associated with a later onset in SCA3 patients. Variants in *ERBB4,* a transmembrane tyrosine kinase that is responsible for the induction of diverse cellular responses, such as mitogenesis and differentiation [273,274], are associated with an earlier onset of the disease. Variants in *MPHOSPH9,* a phosphoprotein related to brain volume and associated to splicing abnormalities in schizophrenia [275] are associated with a delayed onset. We were also able to obtain eight statistically enriched pathways, two of them containing genes with a modifier effect on the onset (*TGFA* and *ERBB4*): HIF1α Signaling and Neuregulin Signaling. These pathways can be important for further therapeutic applications.

In Chapter 4, we performed a pilot study that, as far as we are aware, is the first study investigating genome-wide DNA methylation changes in SCA3 patients compared to matched controls, not only by age and sex, but also by smoking habits. As the sample size was small, we were not expecting genome-wide significant differentially methylated CpG sites, between patients and controls. After considering an arbitrary cut-off for nominal *p* value for differentially methylated CpG sites, a list of genes was obtained. Enrichment analysis of this list of 1279 genes showed an enrichment for pathways involved in neurodegenerative diseases such as Calcium signaling pathway. Alterations in DNA methylation age have been reported for several neurodegenerative diseases. Whether these alterations occur in SCA3 was not stablished. We performed an analysis of epigenetic clocks

(Hannum and Horvath) for SCA3 patients and matched healthy controls. We were not able to find differences between patients and controls for DNA methylation age for both epigenetic clocks. Also, for DNA methylation age acceleration difference and DNA methylation age acceleration difference residuals we were not able to find differences between the two groups for both clocks. The sample size and the limited power of this study can affect these results.

This thesis tried to contribute to unveiling the role of untranslated and methylation prone regions in the clinical variability of SCA3. We found variants with predicted impact in 3'UTR of *ATXN3* gene. We were able to obtain candidate modifier genes, with variants in UTRs and performed the first genome-wide study on DNA methylation comparing patients and controls, providing an approach to DNA methylation landscapes in SCA3. However, further studies in a higher number of patients and controls are required to obtain a methylation profile for SCA3.

REFERENCES

1. Bunting EL, Hamilton J, Tabrizi SJ. Polyglutamine diseases. Curr Opin Neurobiol [Internet]. Elsevier Ltd; 2022;72:39–47. Available from: https://doi.org/10.1016/j.conb.2021.07.001

2. Jones L, Houlden H, Tabrizi SJ. DNA repair in the triplet repeat diseases. 2017; Available from: http://discovery.ucl.ac.uk/1534545/1/Jones_DNA repair in the trinucleotide repeat disorders.pdf

3. Stoyas CA, La Spada AR. The CAG–polyglutamine repeat diseases: a clinical, molecular, genetic, and pathophysiologic nosology [Internet]. 1st ed. Handb. Clin. Neurol. Elsevier B.V.; 2018. Available from: http://dx.doi.org/10.1016/B978-0-444-63233-3.00011-7

4. Chen Z, Sequeiros J, Tang B, Jiang H. Genetic modifiers of age-at-onset in polyglutamine diseases. Ageing Res Rev. 2018;48:99–108.

5. Liu H, Tang T-S, Guo C. Epigenetic profiles in polyglutamine disorders. Epigenomics. England; 2018;10:9–25.

6. Chung MY, Ranum LP, Duvick LA, Servadio A, Zoghbi HY, Orr HT. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nat Genet. United States; 1993;5:254–8.

7. Nagai Y, Azuma T, Funauchi M, Fujita M, Umi M, Hirano M, et al. Clinical and molecular genetic study in seven Japanese families with spinocerebellar ataxia type 6. J Neurol Sci [Internet]. Elsevier; 1998;157:52–9. Available from: https://doi.org/10.1016/S0022-510X(98)00044-6

8. Silveira I, Alonso I, Guimarães L, Mendonça P, Santos C, Maciel P, et al. High germinal instability of the (CTG)n at the SCA8 locus of both expanded and normal alleles. Am J Hum Genet. 2000;66:830–40.

9. Shimazaki H, Takiyama Y, Sakoe K, Amaike M, Nagaki H, Namekawa M, et al. Meiotic instability of the CAG repeats in the SCA6/CACNA1A gene in two Japanese SCA6 families. J

Neurol Sci. Netherlands; 2001;185:101–7.

10. Libby RT, Monckton DG, Fu Y-H, Martinez RA, McAbney JP, Lau R, et al. Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. Hum Mol Genet. England; 2003;12:41–50.

11. Martins S, Pearson CE, Coutinho P, Provost S, Amorim A, Dubé MP, et al. Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. Hum Genet. 2014;133:1311–8.

12. Almaguer-Mederos LE, Mesa JML, González-Zaldívar Y, Almaguer-Gotay D, Cuello-Almarales D, Aguilera-Rodríguez R, et al. Factors associated with ATXN2 CAG/CAA repeat intergenerational instability in Spinocerebellar ataxia type 2. Clin Genet. Denmark; 2018;94:346–50.

13. Du Y-C, Ma Y, Shao Y-R, Gan S-R, Dong Y, Wu Z-Y. Factors Associated with Intergenerational Instability of ATXN3 CAG Repeat and Genetic Anticipation in Chinese Patients with Spinocerebellar Ataxia Type 3. Cerebellum. United States; 2020;19:902–6.

14. Chong SS, McCall AE, Cota J, Subramony SH, Orr HT, Hughes MR, et al. Gametic and somatic tissue-specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. Nat Genet. United States; 1995;10:344–50.

15. Lopes-Cendes I, Maciel P, Kish S, Gaspar C, Robitaille Y, Clark HB, et al. Somatic mosaicism in the central nervous system in spinocerebellar ataxia type 1 and Machado-Joseph disease. Ann Neurol. United States; 1996;40:199–206.

16. Tanaka F, Sobue G, Doyu M, Ito Y, Yamamoto M, Shimada N, et al. Differential pattern in tissue-specific somatic mosaicism of expanded CAG trinucleotide repeats in dentatorubral-pallidoluysian atrophy, Machado-Joseph disease, and X-linked recessive spinal and bulbar muscular atrophy. J Neurol Sci. Netherlands; 1996;135:43–50.

17. Gouw LG, Castañeda MA, McKenna CK, Digre KB, Pulst SM, Perlman S, et al. Analysis of the dynamic mutation in the SCA7 gene shows marked parental effects on CAG repeat transmission. Hum Mol Genet. England; 1998;7:525–32.

18. Klockgether T, Mariotti C, Paulson HL. Spinocerebellar ataxia. Nat Rev Dis Prim [Internet]. Springer US; 2019;5:1–21. Available from: http://dx.doi.org/10.1038/s41572-019-0074-3

19. Lima M, Armas J, Bettencourt C. Non-Mendelian Genetic Aspects in Spinocerebellar Ataxias (SCAS): The Case of Machado-Joseph Disease (MJD). In: Gazulla J, editor. Spinocerebellar Ataxia. InTech; 2012. p. pp 27-40.

20. Zoghbi HY, Orr HT. Glutamine Repeats and Neurodegeneration. Annu Rev Neurosci [Internet]. Annual Reviews; 2000;23:217–47. Available from: https://doi.org/10.1146/annurev.neuro.23.1.217

21. Francini-Pesenti F, Vitturi N, Tresso S, Sorarù G. Metabolic alterations in spinal and bulbar muscular atrophy. Rev Neurol (Paris). 2020;176:780–7.

22. Hashizume A, Fischbeck KH, Pennuto M, Fratta P, Katsuno M. Disease mechanism, biomarker and therapeutics for spinal and bulbar muscular atrophy (SBMA). J Neurol Neurosurg Psychiatry. England; 2020;91:1085–91.

23. Rocchi A, Milioto C, Parodi S, Armirotti A, Borgia D, Pellegrini M, et al. Glycolytic-to-oxidative fiber-type switch and mTOR signaling activation are early-onset features of SBMA muscle modified by high-fat diet. Acta Neuropathol. 2016;132:127–44.

24. Katsuno M, Adachi H, Doyu M, Minamiyama M, Sang C, Kobayashi Y, et al. Leuprorelin rescues polyglutamine-dependent phenotypes in a transgenic mouse model of spinal and bulbar muscular atrophy. Nat Med. 2003;9:768–73.

25. Lieberman AP, Yu Z, Murray S, Peralta R, Low A, Guo S, et al. Peripheral Androgen Receptor Gene Suppression Rescues Disease in Mouse Models of Spinal and Bulbar

Muscular Atrophy. Cell Rep [Internet]. The Authors; 2014;7:774–84. Available from: http://dx.doi.org/10.1016/j.celrep.2014.02.008

26. Banno H, Katsuno M, Suzuki K, Takeuchi Y, Kawashima M, Suga N, et al. Phase 2 trial of leuprorelin in patients with spinal and bulbar muscular atrophy. Ann Neurol. 2009;65:140–50.

27. Katsuno M, Banno H, Suzuki K, Takeuchi Y, Kawashima M, Yabe I, et al. Efficacy and safety of leuprorelin in patients with spinal and bulbar muscular atrophy (JASMITT study): A multicentre, randomised, double-blind, placebo-controlled trial. Lancet Neurol [Internet]. Elsevier Ltd; 2010;9:875–84. Available from: http://dx.doi.org/10.1016/S1474-4422(10)70182-4

28. Fernández-Rhodes LE, Kokkinis AD, White MJ, Watts CA, Auh S, Jeffries NO, et al. Efficacy and safety of dutasteride in patients with spinal and bulbar muscular atrophy: A randomised placebo-controlled trial. Lancet Neurol. 2011;10:140–7.

29. Ghosh R, Tabrizi SJ. Huntington disease [Internet]. 1st ed. Handb. Clin. Neurol. Elsevier B.V.; 2018. Available from: http://dx.doi.org/10.1016/B978-0-444-63233-3.00017-8

30. Nance MA. Genetics of Huntington disease. Handb Clin Neurol. Netherlands; 2017;144:3–14.

31. Jimenez-Sanchez M, Licitra F, Underwood BR, Rubinsztein DC. Huntington's Disease: Mechanisms of Pathogenesis and Therapeutic Strategies. Cold Spring Harb Perspect Med. 2017;7.

32. Nguyen GD, Gokhan S, Molero AE, Mehler MF. Selective roles of normal and mutant huntingtin in neural induction and early neurogenesis. PLoS One. 2013;8:e64368.

33. Lo Sardo V, Zuccato C, Gaudenzi G, Vitali B, Ramos C, Tartari M, et al. An evolutionary recent neuroepithelial cell adhesion function of huntingtin implicates ADAM10-Ncadherin. Nat Neurosci. United States; 2012;15:713–21.

34. Nasir J, Floresco SB, O'Kusky JR, Diewert VM, Richman JM, Zeisler J, et al. Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. Cell. United States; 1995;81:811–23.

35. Zeitlin S, Liu JP, Chapman DL, Papaioannou VE, Efstratiadis A. Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. Nat Genet. United States; 1995;11:155–63.

36. McFarland KN, Huizenga MN, Darnell SB, Sangrey GR, Berezovska O, Cha J-HJ, et al. MeCP2: a novel Huntingtin interactor. Hum Mol Genet. 2014;23:1036–44.

37. Zuccato C, Tartari M, Crotti A, Goffredo D, Valenza M, Conti L, et al. Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. Nat Genet. United States; 2003;35:76–83.

38. Kegel KB, Meloni AR, Yi Y, Kim YJ, Doyle E, Cuiffo BG, et al. Huntingtin is present in the nucleus, interacts with the transcriptional corepressor C-terminal binding protein, and represses transcription. J Biol Chem. United States; 2002;277:7466–76.

39. McKinstry SU, Karadeniz YB, Worthington AK, Hayrapetyan VY, Ozlu MI, Serafin-Molina K, et al. Huntingtin is required for normal excitatory synapse development in cortical and striatal circuits. J Neurosci Off J Soc Neurosci. 2014;34:9455–72.

40. Bettencourt C, Hensman-Moss D, Flower M, Wiethoff S, Brice A, Goizet C, Stevanin G, Koutsis G, Karadima G, Panas M, Yescas-Gómez P, García-Velázquez LE, Alonso-Vilatela ME, Lima M, Raposo M, Traynor B, Sweeney M, Wood N, Giunti P; SPATAX Network, Durr A, JL. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol. 2016;79:983–90.

41. Ciosi M, Maxwell A, Cumming SA, Hensman Moss DJ, Alshammari AM, Flower MD, et al. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes.

EBioMedicine [Internet]. Elsevier B.V.; 2019;48:568–80. Available from: https://doi.org/10.1016/j.ebiom.2019.09.020

42. Wright GEB, Collins JA, Kay C, McDonald C, Dolzhenko E, Xia Q, et al. Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. Am J Hum Genet. 2019;104:1116–26.

43. Lee JM, Correia K, Loupe J, Kim KH, Barker D, Hong EP, et al. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell. 2019;178:887-900.e14.

44. Moss DJH, Tabrizi SJ, Mead S, Lo K, Pardiñas AF, Holmans P, et al. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. Lancet Neurol. 2017;16:701–11.

45. Flower M, Lomeikaite V, Ciosi M, Cumming S, Morales F, Lo K, et al. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. Brain. 2019;142:1876–86.

46. Kim KH, Hong EP, Shin JW, Chao MJ, Loupe J, Gillis T, et al. Genetic and Functional Analyses Point to FAN1 as the Source of Multiple Huntington Disease Modifier Effects. Am J Hum Genet [Internet]. ElsevierCompany.; 2020;107:96–110. Available from: https://doi.org/10.1016/j.ajhg.2020.05.012

47. Bashir H. Emerging therapies in Huntington's disease. Expert Rev Neurother. England; 2019;19:983–95.

48. Grimaldi S, Cupidi C, Smirne N, Bernardi L, Giacalone F, Piccione G, et al. The largest caucasian kindred with dentatorubral-pallidoluysian atrophy: A founder mutation in italy. Mov Disord. 2019;34:1919–24.

49. Morgado J, Reimão S, Coelho M, Rosa MM, Ferreira JJ, Correia Guedes L. Eye of the Tiger Sign and Very Late Onset in Dentatorubral-Pallidoluysian Atrophy. Mov Disord Clin

Pract [Internet]. John Wiley and Sons Inc.; 2015;2:313–5. Available from: https://pubmed.ncbi.nlm.nih.gov/30838235

50. Tezenas du Montcel S, Durr A, Bauer P, Figueroa KP, Ichikawa Y, Brussino A, et al. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. Brain. 2014;137:2444–55.

51. Tsuji S. Dentatorubral – pallidoluysian atrophy [Internet]. 1st ed. Ataxic Disord. Vol. 103. Elsevier B.V.; 2012. Available from: http://dx.doi.org/10.1016/B978-0-444-51892-7.00041-3

52. Wardle M, Morris HR, Robertson NP. Clinical and Genetic Characteristics of Non-Asian Dentatorubral-Pallidoluysian Atrophy : A Systematic Review. 2009;24:1636–40.

53. Vinton A, Fahey MC, Brien TJO, Shaw J, Storey E, Sart D, et al. Clinical Report Dentatorubral-Pallidoluysian Atrophy in Three Generations , With Clinical Courses From Nearly Asymptomatic Elderly to Severe Juvenile , in an Australian Family of Macedonian Descent. 2005;204:201–4.

54. Silveira I, Coutinho P, Maciel P, Gaspar C, Hayes S, Dias A, et al. CAG Repeats in 48 Portuguese Ataxia Families. 1998;138:134–8.

55. Nielsen JE, Sørensen SA, Hasholt L, Nørremølle A. Dentatorubral-pallidoluysian atrophy. Clinical features of a five-generation Danish family. Mov Disord. United States; 1996;11:533–41.

56. Warner TT, Williams LD, Walker RW, Flinter F, Robb SA, Bundey SE, et al. A clinical and molecular genetic study of dentatorubropallidoluysian atrophy in four European families. Ann Neurol. United States; 1995;37:452–9.

57. Tsuji S. Dentatorubral-pallidoluysian atrophy. Handb Clin Neurol. Netherlands; 2012;103:587–94.

58. Carroll LS, Massey TH, Wardle M, Peall KJ. Dentatorubral-pallidoluysian Atrophy: An Update. Tremor Other Hyperkinet Mov (N Y). 2018;8:577.

59. Finsterer J. Ataxias with Autosomal , X-Chromosomal. Can J Neurol Sci. 2009;36:191–6.

60. Martins Junior CR, de Borba FC, Martinez ARM, de Rezende TJR, Cendes IL, Pedroso JL, et al. Twenty-five years since the identification the first SCA gene: History, clinical featureand perspectives for SCA1. Arq Neuropsiquiatr. 2018;76:555–62.

61. Gispert S, Twells R, Orozco G, Brice A, Weber J, Heredero L, et al. Chromosomal assignment of the second locus for autosomal dominant cerebellar ataxia (SCA2) to chromosome 12q23-24.1. Nat Genet. United States; 1993;4:295–9.

62. Lopes-Cendes I, Andermann E, Attig E, Cendes F, Bosch S, Wagner M, et al. Confirmation of the SCA-2 locus as an alternative locus for dominantly inherited spinocerebellar ataxias and refinement of the candidate region. Am J Hum Genet. 1994;54:774–81.

63. Hersheson J, Haworth A, Houlden H. The inherited ataxias: Genetic heterogeneity, mutation databases, and future directions in research and clinical diagnostics. Hum Mutat. 2012;33:1324–32.

64. Saute JAM, Jardim LB. Planning future clinical trials for Machado-Joseph disease. Adv Exp Med Biol. 2018;1049:321–48.

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

66. Takiyama Y, Nishizawa M, Tanaka H, Kawashima S, Sakamoto H, Karube Y, et al. The gene for Machado-Joseph disease maps to human chromosome 14q. Nat Genet. United States; 1993;4:300–4.

67. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. United States; 1994;8:221–8.

68. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, et al. Autosomal

dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. Nat Genet. United States; 1997;15:62–9.

69. Kazachkova N, Lima M. Towards A Therapeutic Intervention in Polyglutamine Ataxias: From Models to Clinical Trials. Atta-ur-Rahman, editor. Front. Clin. Drug Res. - CNS Neurol. Disord. Bentham Science Publishers; 2015.

70. Benomar A, Krols L, Stevanin G, Cancel G, LeGuern E, David G, et al. The gene for autosomal dominant cerebellar ataxia with pigmentary macular dystrophy maps to chromosome 3p12-p21.1. Nat Genet. United States; 1995;10:84–8.

71. Nakamura K, Jeong S-Y, Uchihara T, Anno M, Nagashima K, Nagashima T, et al. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum Mol Genet [Internet]. 2001;10:1441–8. Available from: https://doi.org/10.1093/hmg/10.14.1441

72. Nakano KK, Dawson DM, Spence A. Machado disease. A hereditary ataxia in Portuguese emigrants to Massachusetts. Neurology. United States; 1972;22:49–55.

73. Woods BT, Schaumburg HH. Nigro-spino-dentatal degeneration with nuclear ophthalmoplegia. A unique and partially treatable clinico-pathological entity. J Neurol Sci. Netherlands; 1972;17:149–66.

74. Rosenberg RN, Nyhan WL, Bay C, Shore P. Autosomal dominant striatonigral degeneration. A clinical, pathologic, and biochemical study of a new genetic disorder. Neurology. United States; 1976;26:703–14.

75. Araújo MA De, Raposo M. Trends in the Epidemiology of Spinocerebellar Ataxia Type 3 / Machado-Joseph Disease in the Azores Islands , Portugal. JSM Brain Sci. 2016;1:1001.

76. Bettencourt C, Lima M. Genetics at different levels in machado-joseph disease (MJD/SCA3): Cause, modifiers and therapy. Ataxia Causes, Symptoms Treat. 2012. p. 123–38.

77. de Castilhos RM, Furtado GV, Gheno TC, Schaeffer P, Russo A, Barsottini O, et al.

Spinocerebellar ataxias in Brazil--frequencies and modulating effects of related genes. Cerebellum. United States; 2014;13:17–28.

78. Coutinho P, Andrade C. Autosomal dominant system degeneration in Portuguese families of the Azores Islands. Neurology [Internet]. 1978;28:703 LP – 703. Available from: http://n.neurology.org/content/28/7/703.abstract

79. Coutinho P, Sequeiros J. [Clinical, genetic and pathological aspects of Machado-Joseph disease]. J Genet Hum. Switzerland; 1981;29:203–9.

80. Coutinho P, Guimarães A, Scaravilli F. The pathology of Machado-Joseph disease. Report of a possible homozygous case. Acta Neuropathol. Germany; 1982;58:48–54.

81. Coutinho P. Doença de Machado-Joseph [Internet]. Tese. University of Porto; 1992. Available from: http://hdl.handle.net/10216/10229

82. Lima L, Coutinho P. Clinical criteria for diagnosis of Machado-Joseph disease: report of a non-Azorena Portuguese family. Neurology. United States; 1980;30:319–22.

83. Pedroso JL, Bezerra MLE, Braga-Neto P, Pinheiro DS, Minett T, do Prado GF, et al. Is neuropathy involved with restless legs syndrome in Machado-Joseph disease? Eur Neurol. Switzerland; 2011;66:200–3.

84. França MCJ, D'Abreu A, Nucci A, Lopes-Cendes I. Muscle excitability abnormalities in Machado-Joseph disease. Arch Neurol. United States; 2008;65:525–9.

85. Lopes TM, D'Abreu A, França MCJ, Yasuda CL, Betting LE, Samara AB, et al. Widespread neuronal damage and cognitive dysfunction in spinocerebellar ataxia type 3. J Neurol. Germany; 2013;260:2370–9.

86. Braga-Neto P, Felicio AC, Hoexter MQ, Pedroso JL, Dutra LA, Alessi H, et al. Cognitive and olfactory deficits in Machado–Joseph disease: A dopamine transporter study. Parkinsonism Relat Disord [Internet]. 2012;18:854–8. Available from: https://www.sciencedirect.com/science/article/pii/S1353802012001472

87. Braga-Neto P, Pedroso JL, Alessi H, Dutra LA, Felício AC, Minett T, et al. Cerebellar cognitive affective syndrome in Machado Joseph disease: core clinical features. Cerebellum. United States; 2012;11:549–56.

88. Kawai Y, Takeda A, Abe Y, Washimi Y, Tanaka F, Sobue G. Cognitive impairments in Machado-Joseph disease. Arch Neurol. United States; 2004;61:1757–60.

89. Maruff P, Tyler P, Burt T, Currie B, Burns C, Currie J. Cognitive deficits in Machado-Joseph disease. Ann Neurol. United States; 1996;40:421–7.

90. Martinez ARM, Nunes MB, Faber I, D'Abreu A, Lopes-Cendes Í, França MC. Fatigue and Its Associated Factors in Spinocerebellar Ataxia Type 3/Machado-Joseph Disease. Cerebellum. 2017;16:118–21.

91. Cecchin CR, Pires AP, Rieder CR, Monte TL, Silveira I, Carvalho T, et al. Depressive symptoms in Machado-Joseph disease (SCA3) patients and their relatives. Community Genet. Switzerland; 2007;10:19–26.

92. Zawacki TM, Grace J, Friedman JH, Sudarsky L. Executive and emotional dysfunction in Machado-Joseph disease. Mov Disord. United States; 2002;17:1004–10.

93. Sakai T. Effects of tetrahydrobiopterin on ataxia in Machado-Joseph disease may be based upon the theory of "cerebellar long-term depression". Med Hypotheses. United States; 2001;57:180–2.

94. LaGrappe D, Massey L, Kruavit A, Howarth T, Lalara G, Daniels B, et al. Sleep disorders among Aboriginal Australians with Machado-Joseph Disease: Quantitative results from a multiple methods study to assess the experience of people living with the disease and their caregivers. Neurobiol sleep circadian Rhythm. 2022;12:100075.

95. Pedroso JL, Braga-Neto P, Martinez ARM, Martins CRJ, Rezende Filho FM, Sobreira-Neto MA, et al. Sleep disorders in Machado-Joseph disease. Curr Opin Psychiatry. United States; 2016;29:402–8.

96. Pedroso JL, Braga-Neto P, Felício AC, Minett T, Yamaguchi E, Prado LBF do, et al. Sleep disorders in Machado-Joseph disease: a dopamine transporter imaging study. J Neurol Sci. Netherlands; 2013;324:90–3.

97. Pedroso JL, Braga-neto P. Sleep Disorders in Machado – Joseph Disease : Frequency , Discriminative Thresholds , Predictive Values , and Correlation with Ataxia-Related Motor and Non-Motor Features. 2011;291–5.

98. D'Abreu A, França MJ, Conz L, Friedman JH, Nucci AM, Cendes F, et al. Sleep symptoms and their clinical correlates in Machado-Joseph disease. Acta Neurol Scand. Denmark; 2009;119:277–80.

99. Musegante AFA, Almeida PNS, Barboza AL, Barroso UJ. Urinary symptoms and urodynamic findings in patients with Machado-Joseph disease. J Neurol. Germany; 2011;258:623–6.

100. Sakakibara R, Uchiyama T, Arai K, Yamanishi T, Hattori T. Lower urinary tract dysfunction in Machado-Joseph disease: a study of 11 clinical-urodynamic observations. J Neurol Sci. Netherlands; 2004;218:67–72.

101. Raposo M, Ramos A, Bettencourt C, Lima M. Replicating studies of genetic modifiers in spinocerebellar ataxia type 3: can homogeneous cohorts aid? Brain. England; 2015. p. e398.

102. Carvalho DR, La Rocque-Ferreira A, Rizzo IM, Imamura EU, Speck-Martins CE. Homozygosity Enhances Severity in Spinocerebellar Ataxia Type 3. Pediatr Neurol. 2008;38:296–9.

103. Lysenko L, Grewal RP, Ma W, Peddareddygari LR. Homozygous Machado Joseph Disease: a case report and review of literature. Can J Neurol Sci Le J Can des Sci Neurol. England; 2010;37:521–3.

104. do Carmo Costa M, Luna-Cancalon K, Fischer S, Ashraf NS, Ouyang M, Dharia RM, et al. Toward RNAi Therapy for the Polyglutamine Disease Machado–Joseph Disease. Mol Ther

[Internet]. 2013;21:1898–908. Available from: https://www.sciencedirect.com/science/article/pii/S1525001616322560

105. Maciel P, Gaspar C, DeStefano AL, Silveira I, Coutinho P, Radvany J, et al. Correlation between CAG repeat length and clinical features in Machado-Joseph disease. Am J Hum Genet. 1995;57:54–61.

106. Sobue G, Doyu M, Nakao N, Shimada N, Mitsuma T, Maruyama H, et al. Homozygosity for Machado-Joseph disease gene enhances phenotypic severity. J. Neurol. Neurosurg. Psychiatry. 1996. p. 354–6.

107. Lerer I, Merims D, Abeliovich D, Zlotogora J, Gadoth N. Machado-Joseph disease: correlation between the clinical features, the CAG repeat length and homozygosity for the mutation. Eur J Hum Genet. England; 1996;4:3–7.

108. Fukutake T, Shinotoh H, Nishino H, Ichikawa Y, Goto J, Kanazawa I, et al. Homozygous Machado-Joseph disease presenting as REM sleep behaviour disorder and prominent psychiatric symptoms. Eur J Neurol. England; 2002;9:97–100.

109. Gordon CR, Joffe V, Vainstein G, Gadoth N. Vestibulo-ocular arreflexia in families with spinocerebellar ataxia type 3 (Machado-Joseph disease). J Neurol Neurosurg Psychiatry. 2003;74:1403–6.

110. Zeng S, Zeng J, He M, Zeng X, Zhou Y, Liu Z, et al. Chinese homozygous Machado-Joseph disease (MJD)/SCA3: a case report. J Hum Genet. England; 2015;60:157–60.

111. Shang X-J, Xu H-L, Yang J-S, Chen P-P, Lin M-T, Qian M-Z, et al. Homozygote of spinocerebellar Ataxia type 3 correlating with severe phenotype based on analyses of clinical features. J Neurol Sci. Netherlands; 2018;390:111–4.

112. Li Q, Zhao J. Clinical features and genetic characteristics of homozygous spinocerebellar ataxia type 3. 2020;1–9.

113. Chen Y, Li D, Wei M, Zhou M, Zhang L, Yu J, et al. Homozygous spinocerebellar ataxia

type 3 in China: a case report. J Int Med Res. 2021;49:3000605211021370.

114. Pedroso JL, França MCJ, Braga-Neto P, D'Abreu A, Saraiva-Pereira ML, Saute JA, et al. Nonmotor and extracerebellar features in Machado-Joseph disease: a review. Mov Disord. United States; 2013;28:1200–8.

115. D'Abreu A, França MCJ, Yasuda CL, Campos BAG, Lopes-Cendes I, Cendes F. Neocortical atrophy in Machado-Joseph disease: a longitudinal neuroimaging study. J neuroimaging Off J Am Soc Neuroimaging. United States; 2012;22:285–91.

116. D'Abreu A, França MCJ, Yasuda CL, Souza MSA, Lopes-Cendes I, Cendes F. Thalamic volume and dystonia in Machado-Joseph disease. J neuroimaging Off J Am Soc Neuroimaging. United States; 2011;21:e91-3.

117. Riess O, Rüb U, Pastore A, Bauer P, Schöls L. SCA3: neurological features, pathogenesis and animal models. Cerebellum. United States; 2008;7:125–37.

118. Rüb U, Brunt ER, Deller T. New insights into the pathoanatomy of spinocerebellar ataxia type 3 (Machado-Joseph disease). Curr Opin Neurol. England; 2008;21:111–6.

119. Guo J, Chen H, Biswal BB, Guo X, Zhang H, Dai L, et al. Gray matter atrophy patterns within the cerebellum-neostriatum-cortical network in SCA3. Neurology [Internet]. 2020;95:e3036 LP-e3044. Available from: http://n.neurology.org/content/95/22/e3036.abstract

120. Eichler L, Bellenberg B, Hahn HK, Köster O, Schöls L, Lukas C. Quantitative assessment of brain stem and cerebellar atrophy in spinocerebellar ataxia types 3 and 6: impact on clinical status. AJNR Am J Neuroradiol. 2011;32:890–7.

121. Abe Y, Tanaka F, Matsumoto M, Doyu M, Hirayama M, Kachi T, et al. CAG repeat number correlates with the rate of brainstem and cerebellar atrophy in Machado-Joseph disease. Neurology [Internet]. 1998;51:882 LP – 884. Available from: http://n.neurology.org/content/51/3/882.abstract

122. Onodera O, Idezuka J, Igarashi S, Takiyama Y, Endo K, Takano H, et al. Progressive atrophy of cerebellum and brainstem as a function of age and the size of the expanded CAG repeats in the MJD1 gene in Machado-Joseph disease. Ann Neurol. United States; 1998;43:288–96.

123. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994;8:221–8.

124. Melo ARV, Raposo M, Ventura M, Martins S, Pavão S, Alonso I, et al. Genetic Variation in ATXN3 (Ataxin-3) 3′UTR: Insights into the Downstream Regulatory Elements of the Causative Gene of Machado-Joseph Disease/Spinocerebellar Ataxia Type 3. The Cerebellum [Internet]. 2022; Available from: https://doi.org/10.1007/s12311-021-01358-0

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

126. van Schaik IN, Jöbsis GJ, Vermeulen M, Keizers H, Bolhuis PA, de Visser M. Machado-Joseph disease presenting as severe asymmetric proximal neuropathy. J Neurol Neurosurg Psychiatry. 1997;63:534–6.

127. Maciel P, Costa MC, Ferro A, Rousseau M, Santos CS, Gaspar C, et al. Improvement in the molecular diagnosis of Machado-Joseph disease. Arch Neurol. United States; 2001;58:1821–7.

128. van Alfen N, Sinke RJ, Zwarts MJ, Gabreëls-Festen A, Praamstra P, Kremer BP, et al. Intermediate CAG repeat lengths (53,54) for MJD/SCA3 are associated with an abnormal phenotype. Ann Neurol. United States; 2001;49:805–7.

129. Gu W, Ma H, Wang K, Jin M, Zhou Y, Liu X, et al. The shortest expanded allele of the MJD1 gene in a Chinese MJD kindred with autonomic dysfunction. Eur Neurol. Switzerland;

2004;52:107–11.

130. Padiath QS, Srivastava AK, Roy S, Jain S, Brahmachari SK. Identification of a novel 45 repeat unstable allele associated with a disease phenotype at the MJD1/SCA3 locus. Am J Med Genet Part B, Neuropsychiatr Genet Off Publ Int Soc Psychiatr Genet. United States; 2005;133B:124–6.

131. Maciel P, Gaspar C, Destefano AL, Silveira I, Coutinho P, Radvany J, et al. Correlation between CAG Repeat Length and Clinical Features in Machado-Joseph Disease. 1995;54–61.

132. Dürr A, Stevanin G, Cancel G, Duyckaerts C, Abbas N, Didierjean O, et al. Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular, and neuropathological features. Ann Neurol. United States; 1996;39:490–9.

133. Jardim LB, Pereira ML, Silveira I, Ferro A, Sequeiros J, Giugliani R. Neurologic findings in Machado-Joseph disease: relation with disease duration, subtypes, and (CAG)n. Arch Neurol. United States; 2001;58:899–904.

134. Costa C. Recent therapeutic prospects for Machado – Joseph disease. 2020;519–26.

135. Bettencourt C, Santos C, Kay T, Vasconcelos J, Lima M. Analysis of segregation patterns in Machado-Joseph disease pedigrees. J Hum Genet. England; 2008;53:920–3.

136. Cancel G, Gourfinkel-An I, Stevanin G, Didierjean O, Abbas N, Hirsch E, et al. Somatic mosaicism of the CAG repeat expansion in spinocerebellar ataxia type 3/Machado-Joseph disease. Hum Mutat. United States; 1998;11:23–7.

137. Hashida H, Goto J, Kurisaki H, Mizusawa H, Kanazawa I. Brain regional differences in the expansion of a CAG repeat in the spinocerebellar ataxias: dentatorubral-pallidoluysian atrophy, Machado-Joseph disease, and spinocerebellar ataxia type 1. Ann Neurol. United States; 1997;41:505–11.

138. Maciel P, Lopes-Cendes I, Kish S, Sequeiros J, Rouleau GA. Mosaicism of the CAG repeat in CNS tissue in relation to age at death in spinocerebellar ataxia type 1 and

Machado-Joseph disease patients. Am. J. Hum. Genet. 1997. p. 993–6.

139. Cancel G, Abbas N, Stevanin G, Dürr A, Chneiweiss H, Néri C, et al. Marked phenotypic heterogeneity associated with expansion of a CAG repeat sequence at the spinocerebellar ataxia 3/Machado-Joseph disease locus. Am J Hum Genet. 1995;57:809–16.

140. Sequeiros J, Martindale J, Seneca S, Giunti P, Kämäräinen O, Volpini V, et al. EMQN Best Practice Guidelines for molecular genetic testing of SCAs. Eur J Hum Genet. 2010;18:1173–6.

141. Melo ARV, Ramos A, Kazachkova N, Raposo M, Bettencourt BF, Rendeiro AR, et al. Triplet Repeat Primed PCR (TP-PCR) in Molecular Diagnostic Testing for Spinocerebellar Ataxia Type 3 (SCA3). Mol Diagn Ther. New Zealand; 2016;20:617–22.

142. Schmidt T, Landwehrmeyer GB, Schmitt I, Trottier Y, Auburger G, Laccone F, et al. An isoform of ataxin-3 accumulates in the nucleus of neuronal cells in affected brain regions of SCA3 patients. Brain Pathol. 1998;8:669–79.

143. Masino L, Musi V, Menon RP, Fusi P, Kelly G, Frenkiel TA, et al. Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail. FEBS Lett. England; 2003;549:21–5.

144. Tzvetkov N, Breuer P. Josephin domain-containing proteins from a variety of species are active de-ubiquitination enzymes. Biol Chem. Germany; 2007;388:973–8.

145. McLoughlin HS, Moore LR, Paulson HL. Pathogenesis of SCA3 and implications for other polyglutamine diseases. Neurobiol Dis. 2020;134:104635.

146. Bettencourt C, Raposo M, Ros R, Montiel R, Bruges-Armas J, Lima M. Transcript diversity of Machado-Joseph disease gene (ATXN3) is not directly determined by SNPs in exonic or flanking intronic regions. J Mol Neurosci. United States; 2013;49:539–43.

147. Evers MM, Toonen LJA, van Roon-Mom WMC. Ataxin-3 protein and RNA toxicity in spinocerebellar ataxia type 3: current insights and emerging therapeutic strategies. Mol

Neurobiol. 2014;49:1513–31.

148. Evert BO, Araujo J, Vieira-Saecker AM, de Vos RAI, Harendza S, Klockgether T, et al. Ataxin-3 represses transcription via chromatin binding, interaction with histone deacetylase 3, and histone deacetylation. J Neurosci Off J Soc Neurosci. 2006;26:11474–86.

149. Saute JAM, Jardim LB. Machado Joseph disease: Clinical and genetic aspects, and current treatment. Expert Opin Orphan Drugs. 2015;3:517–35.

150. Matos CA, de Almeida LP, Nóbrega C. Machado-Joseph disease/spinocerebellar ataxia type 3: lessons from disease pathogenesis and clues into therapy. J Neurochem. England; 2019;148:8–28.

151. Costa M do C, Paulson HL. Toward understanding Machado-Joseph disease. Prog Neurobiol. 2012;97:239–57.

152. Onofre I, Mendonça N, Lopes S, Nobre R, De JB, Almeida LP De. Fibroblasts of Machado Joseph Disease patients reveal autophagy impairment. Nat Publ Gr [Internet]. Nature Publishing Group; 2016;1–10. Available from: http://dx.doi.org/10.1038/srep28220

153. Cunha-Santos J, Duarte-Neves J, Carmona V, Guarente L, Pereira de Almeida L, Cavadas C. Caloric restriction blocks neuropathology and motor deficits in Machado-Joseph disease mouse models through SIRT1 pathway. Nat Commun. 2016;7:11445.

154. Nascimento-Ferreira I, Nóbrega C, Vasconcelos-Ferreira A, Onofre I, Albuquerque D, Aveleira C, et al. Beclin 1 mitigates motor and neuropathological deficits in genetic mouse models of Machado-Joseph disease. Brain. England; 2013;136:2173–88.

155. Nascimento-Ferreira I, Santos-Ferreira T, Sousa-Ferreira L, Auregan G, Onofre I, Alves S, et al. Overexpression of the autophagic beclin-1 protein clears mutant ataxin-3 and alleviates Machado-Joseph disease. Brain. England; 2011;134:1400–15.

156. Li X, Liu H, Fischhaber PL, Tang T-S. Toward therapeutic targets for SCA3: Insight into the role of Machado-Joseph disease protein ataxin-3 in misfolded proteins clearance. Prog

Neurobiol. England; 2015;132:34–58.

157. Souza GN, Kersting N, Santos ASP, Furtado G V. Spinocerebellar ataxia type 3 / Machado – Joseph disease : segregation patterns and factors influencing instability of expanded CAG transmissions. 2016;134–40.

158. Rahit KMTH, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. Genes (Basel). 2020;11:239.

159. Nadeau JH. Modifier genes in mice and humans. Nat Rev Genet. England; 2001;2:165–74.

160. Raposo M, Bettencourt C, Melo ARV, Ferreira AF, Alonso I, Silva P, et al. Novel Machado-Joseph disease-modifying genes and pathways identified by whole-exome sequencing. Neurobiol Dis. United States; 2021;162:105578.

161. França MCJ, Emmel VE, D'Abreu A, Maurer-Morelli C V, Secolin R, Bonadia LC, et al. Normal ATXN3 Allele but Not CHIP Polymorphisms Modulates Age at Onset in Machado-Joseph Disease. Front Neurol. 2012;3:164.

162. Bettencourt C, Raposo M, Kazachkova N, Santos C, Kay T, Vasconcelos J, et al. Sequence analysis of 5′ regulatory regions of the Machado-Joseph disease gene (ATXN3). Cerebellum. 2012;11:1045–50.

163. Long Z, Chen Z, Wang C, Huang F, Peng H, Hou X, et al. Two novel SNPs in ATXN3 3′ UTR may decrease age at onset of SCA3/MJD in Chinese patients. PLoS One. 2015;10:e0117488.

164. Jardim L, Silveira I, Pereira ML, Do Céu Moreira M, Mendonça P, Sequeiros J, et al. Searching for modulating effects of SCA2, SCA6 and DRPLA CAG tracts on the Machado-Joseph disease (SCA3) phenotype. Acta Neurol Scand. 2003;107:211–4.

165. Chen Z, Zheng C, Long Z, Cao L, Li X, Shang H, et al. (CAG)n loci as genetic modifiers of age-at-onset in patients with Machado-Joseph disease from mainland China. Brain.

England; 2016. p. e41.

166. Bettencourt C, Hensman-Moss D, Flower M, Wiethoff S, Brice A, Goizet C, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Ann Neurol. 2016;79:983–90.

167. Mergener R, Furtado GV, de Mattos EP, Leotti VB, Jardim LB, Saraiva-Pereira ML. Variation in DNA Repair System Gene as an Additional Modifier of Age at Onset in Spinocerebellar Ataxia Type 3/Machado–Joseph Disease. NeuroMolecular Med [Internet]. Springer US; 2020;22:133–8. Available from: https://doi.org/10.1007/s12017-019-08572-4

168. Ramos A, Planchat M, Vieira Melo AR, Raposo M, Shamim U, Suroliya V, et al. Mitochondrial DNA haplogroups and age at onset of Machado-Joseph disease/spinocerebellar ataxia type 3: a study in patients from multiple populations. Eur J Neurol. England; 2019;26:506–12.

169. Li X-Y, Zhang Y-B, Xu M, Cheng H-R, Dong Y, Ni W, et al. Effect of Apolipoprotein E Genotypes on Huntington's Disease Phenotypes in a Han Chinese Population. Neurosci Bull. 2019;35:756–62.

170. Bettencourt C, Raposo M, Kazachkova N, Cymbron T, Santos C, Kay T, et al. The APOE ε2 allele increases the risk of earlier age at onset in Machado-Joseph disease. Arch Neurol. United States; 2011;68:1580–3.

171. Peng H, Wang C, Chen Z, Sun Z, Jiao B, Li K, et al. The APOE ε2 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 [Internet]. Elsevier Ltd; 2014;35:2179.e15-2179.e18. Available from: http://dx.doi.org/10.1016/j.neurobiolaging.2014.03.020

172. Zhou Q, Ni W, Dong Y, Wang N, Gan SR, Wu ZY. The role of apolipoprotein E as a risk factor for an earlier age at onset for machado-joseph disease is doubtful. PLoS One. 2014;9:9–12.

173. Siebert M, Donis KC, Socal M, Rieder CRM, Emmel VE, Vairo F, et al. Glucocerebrosidase gene variants in parkinsonian patients with Machado Joseph/spinocerebellar ataxia 3. Parkinsonism Relat Disord. England; 2012;18:185–90.

174. Raposo M, Bettencourt C, Ramos A, Kazachkova N, Vasconcelos J, Kay T, 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. United States; 2017;19:41–5.

175. Pickford F, Masliah E, Britschgi M, Lucin K, Narasimhan R, Jaeger PA, et al. The autophagy-related protein beclin 1 shows reduced expression in early Alzheimer disease and regulates amyloid beta accumulation in mice. J Clin Invest. 2008;118:2190–9.

176. Kazachkova N, Raposo M, Ramos A, Montiel R, Lima M. Promoter Variant Alters Expression of the Autophagic BECN1 Gene: Implications for Clinical Manifestations of Machado-Joseph Disease. Cerebellum. The Cerebellum; 2017;16:957–63.

177. Akçimen F, Martins S, Liao C, Bourassa C V, Catoire H, Nicholson GA, et al. Genome-wide association study identifies genetic factors that modify age at onset in Machado-Joseph disease. Aging (Albany NY). 2020;12:4742–56.

178. Barrett LW, Fletcher S, Wilton SD. Untranslated Gene Regions and Other Non-coding Elements. 2013.

179. Liu H, Yin J, Xiao M, Gao C, Mason AS, Zhao Z, et al. Characterization and evolution of 5′ and 3′ untranslated regions in eukaryotes. Gene [Internet]. Elsevier B.V.; 2012;507:106–11. Available from: http://dx.doi.org/10.1016/j.gene.2012.07.034

180. Wanke KA, Devanna P, Vernes SC. Understanding Neurodevelopmental Disorders: The Promise of Regulatory Variation in the 3′UTRome. Biol Psychiatry. United States; 2018;83:548–57.

181. Murthy M, Cheng YY, Holton JL, Bettencourt C. Neurodegenerative movement

disorders: An epigenetics perspective and promise for the future. Neuropathol Appl Neurobiol. England; 2021;47:897–909.

182. Chouliaras L, Kenis G, Visser PJ, Scheltens P, Tsolaki M, Jones RW, et al. DNMT3A moderates cognitive decline in subjects with mild cognitive impairment: replicated evidence from two mild cognitive impairment cohorts. Epigenomics. England; 2015;7:533–7.

183. Liao X, Zhan Z, Luo Y, Li K, Wang J, Guo J, et al. Association study between SNP rs150689919 in the DNA demethylation gene, TET1, and Parkinson's disease in Chinese Han population. BMC Neurol [Internet]. 2013;13:196. Available from: https://doi.org/10.1186/1471-2377-13-196

184. Zsindely N, Siági F, Bodai L. DNA Methylation in Huntington's Disease. Int J Mol Sci. 2021;22.

185. Emmel VE, Alonso I, Jardim LB, Saraiva-Pereira ML, Sequeiros J. Does DNA methylation in the promoter region of the ATXN3 gene modify age at onset in MJD (SCA3) patients? Clin. Genet. Denmark; 2011. p. 100–2.

186. Wang C, Peng H, Li J, Ding D, Chen Z, Long Z, et al. Alteration of methylation status in the ATXN3 gene promoter region is linked to the SCA3/MJD. Neurobiol Aging. United States; 2017;53:192.e5-192.e10.

187. Ding D, Wang C, Chen Z, Peng H, Li K, Zhou X, et al. Polymorphisms in DNA methylation-related genes are linked to the phenotype of Machado-Joseph disease. Neurobiol Aging. United States; 2019;75:225.e1-225.e8.

188. Zhao H, Yang L, Dong Y, Wu Z-Y. Phenotypic variance in monozygotic twins with SCA3. Mol Genet genomic Med. 2020;8:e1438.

189. Li J, Shu A, Sun Y, Yang W, Tang X, Pu H, et al. DNA methylation age acceleration is associated with age of onset in Chinese spinocerebellar ataxia type 3 patients. Neurobiol Aging. United States; 2022;113:1–6.

190. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol [Internet]. 2013;14:3156. Available from: https://doi.org/10.1186/gb-2013-14-10-r115

191. Ding D, Chen Z, Wang C, Tang X, Zhang L, Fang Q, et al. A Variant in Genes of the NPY System as Modifier Factor of Machado-Joseph Disease in the Chinese Population. Front Aging Neurosci. 2022;14:822657.

192. Martins AC, Rieck M, Leotti VB, Saraiva-Pereira ML, Jardim LB. Variants in Genes of Calpain System as Modifiers of Spinocerebellar Ataxia Type 3 Phenotype. J Mol Neurosci. United States; 2021;71:1906–13.

193. Mergener R, Furtado GV, de Mattos EP, Leotti VB, Jardim LB, Saraiva-Pereira ML. Variation in DNA Repair System Gene as an Additional Modifier of Age at Onset in Spinocerebellar Ataxia Type 3/Machado-Joseph Disease. Neuromolecular Med. United States; 2020;22:133–8.

194. Peng H, Wang C, Chen Z, Sun Z, Jiao B, Li K, et al. APOE ε2 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. United States; 2014;35:2179.e15-8.

195. Bonet ML, Palou A. Chapter 3 - Regulation of Gene Expression. In: Caterina RDE, Martinez JA, Kohlmeier MBT-P of N and N, editors. Academic Press; 2020. p. 17–25. Available from: https://www.sciencedirect.com/science/article/pii/B9780128045725000033

196. Brent L. Fogel DHG. Neurology in Clinical Practice. In: Robert B. Darof, Joseph Jankovic, John C. Mazziotta SLP, editor. Bradley's Neurol Clin Pract. Seventh. London, New York, Oxford, Philadelphia, St Louis, Sydney, Toronto: Elsevier Inc; 2016. p. 648–73.

197. Regateiro FJ. Manual de genética médica. 1ª. Coimbra: Imprensa da Universidade de Coimbra; 2007.

198. Cramer P. Organization and regulation of gene transcription. Nature. England; 2019;573:45–54.

199. Strachan, T., Goodship, J., & Chinnery P. Genetics and Genomics in Medicine. 1st ed. Owen E, editor. New York and London: Garland Science; 2014.

200. Carlberg C, Molnár F. Transcription Factors BT - Mechanisms of Gene Regulation. In: Carlberg C, Molnár F, editors. Dordrecht: Springer Netherlands; 2016. p. 57–73. Available from: https://doi.org/10.1007/978-94-017-7741-4_4

201. Messeguer X, Escudero R, Farré D, Núñez O, Martínez J, Albà MM. PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics. England; 2002;18:333–4.

202. Mayr C. What Are 3′ UTRs Doing? Cold Spring Harb Perspect Biol. 2019;11.

203. Mayr C. Evolution and Biological Roles of Alternative 3′UTRs. Trends Cell Biol. 2016;26:227–37.

204. Sepehri Z, Beacon TH, Osman FDS, Jahan S, Davie JR. Chapter 2 - DNA methylation and chromatin modifications. In: Ferguson BSBT-NE, editor. Transl Epigenetics [Internet]. Academic Press; 2019. p. 13–36. Available from: https://www.sciencedirect.com/science/article/pii/B9780128168431000023

205. Minarovits J, Banati F, Szenthe K, Niller HH. Epigenetic Regulation. Adv Exp Med Biol. United States; 2016;879:1–25.

206. Greer EL, Shi Y. Histone methylation: a dynamic mark in health, disease and inheritance. Nat Rev Genet [Internet]. 2012;13:343–57. Available from: https://doi.org/10.1038/nrg3173

207. Zhang P, Wu W, Chen Q, Chen M. Non-Coding RNAs and their Integrated Networks. J Integr Bioinform. 2019;16.

208. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology [Internet]. 2013;38:23–38. Available from: https://doi.org/10.1038/npp.2012.112

209. Saul D, Kosinsky RL. Epigenetics of Aging and Aging-Associated Diseases. Int J Mol Sci. 2021;22.

210. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67.

211. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol. 2002;3:REVIEWS0004.

212. Asrani KH, Farelli JD, Stahley MR, Miller RL, Cheng CJ, Subramanian RR, et al. Optimization of mRNA untranslated regions for improved expression of therapeutic mRNA. RNA Biol. 2018;15:756–62.

213. Ichikawa Y, Goto J, Hattori M, Toyoda A, Ishii K, Jeong S-Y, et al. The genomic structure and expression of MJD, the Machado-Joseph disease gene. J Hum Genet [Internet]. 2001;46:413–22. Available from: https://doi.org/10.1007/s100380170060

214. Bettencourt C, Santos C, Montiel R, Costa C, Cruz-morales P, Santos LR, et al. Increased transcript diversity : novel splicing variants of Machado – Joseph Disease gene (ATXN3). 2010;193–202.

215. de Mattos EP, Leotti VB, Soong B-W, Raposo M, Lima M, Vasconcelos J, et al. Age at onset prediction in spinocerebellar ataxia type 3 changes according to population of origin. Eur J Neurol. 2019;26.

216. De Mattos EP, Kolbe Musskopf M, Bielefeldt Leotti V, Saraiva-Pereira ML, Jardim LB. Genetic risk factors for modulation of age at onset in Machado-Joseph disease/spinocerebellar ataxia type 3: A systematic review and meta-analysis. J Neurol Neurosurg Psychiatry. 2019;90:203–10.

217. Huang F, Zhang L, Long Z, Chen Z, Hou X, Wang C, et al. miR-25 alleviates polyQ-mediated cytotoxicity by silencing ATXN3. FEBS Lett. 2014;588:4791–8.

218. Carmona V, Cunha-Santos J, Onofre I, Simões AT, Vijayakumar U, Davidson BL, et al. Unravelling Endogenous MicroRNA System Dysfunction as a New Pathophysiological Mechanism in Machado-Joseph Disease. Mol Ther [Internet]. Elsevier Ltd.; 2017;25:1038–55. Available from: http://dx.doi.org/10.1016/j.ymthe.2017.01.021

219. Bettencourt C, Fialho RN, Santos C, Montiel R, Bruges-Armas J, Maciel P, et al. Segregation distortion of wild-type alleles at the Machado-Joseph disease locus: a study in normal families from the Azores islands (Portugal). J Hum Genet. England; 2008;53:333–9.

220. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

221. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

222. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. England; 2010;10:564–7.

223. Serrote CML, Reiniger LRS, Silva KB, Rabaiolli SM dos S, Stefanel CM. Determining the Polymorphism Information Content of a molecular marker. Gene [Internet]. 2020;726:144175. Available from: https://www.sciencedirect.com/science/article/pii/S0378111919308340

224. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536–43.

225. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11:294–6.

226. Vejnar CE, Zdobnov EM. MiRmap: comprehensive prediction of microRNA target

repression strength. Nucleic Acids Res. 2012;40:11673–83.

227. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;4.

228. Liu C, Zhang F, Li T, Lu M, Wang L, Yue W, et al. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics [Internet]. 2012;13:661. Available from: https://doi.org/10.1186/1471-2164-13-661

229. Mao F, Xiao L, Li X, Liang J, Teng H, Cai W, et al. RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins. Nucleic Acids Res [Internet]. 2015/12/03. Oxford University Press; 2016;44:D154–63. Available from: https://pubmed.ncbi.nlm.nih.gov/26635394

230. Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. Nucleic Acids Res. 2020;48:D174–9.

231. Shi Y, Huang F, Tang B, Li J, Wang J, Shen L, et al. MicroRNA profiling in the serums of SCA3/MJD patients. Int J Neurosci. England; 2014;124:97–101.

232. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR 3rd, et al. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol Cell. 2009;33:365–76.

233. Radine C, Peters D, Reese A, Neuwahl J, Budach W, Jänicke RU, et al. The RNA-binding protein RBM47 is a novel regulator of cell fate decisions by transcriptionally controlling the p53-p21-axis. Cell Death Differ [Internet]. 2020;27:1274–85. Available from: https://doi.org/10.1038/s41418-019-0414-6

234. Deleon J, Miller BL. Chapter 27 - Frontotemporal dementia. In: Geschwind DH, Paulson HL, Klein CBT-H of CN, editors. Neurogenetics, Part II [Internet]. Elsevier; 2018. p. 409–30. Available from: https://www.sciencedirect.com/science/article/pii/B9780444640765000272

235. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. United States; 2004;36:40–5.

236. Rothamel K, Arcos S, Kim B, Reasoner C, Lisy S, Mukherjee N, et al. ELAVL1 primarily couples mRNA stability with the 3′ UTRs of interferon-stimulated genes. Cell Rep [Internet]. 2021;35:109178. Available from: https://www.sciencedirect.com/science/article/pii/S2211124721005234

237. Li X, Wang X, Cheng Z, Zhu Q. AGO2 and its partners: a silencing complex, a chromatin modulator, and new features. Crit Rev Biochem Mol Biol [Internet]. Taylor & Francis; 2020;55:33–53. Available from: https://doi.org/10.1080/10409238.2020.1738331

238. Zhu W, Zhou B-L, Rong L-J, Ye L, Xu H-J, Zhou Y, et al. Roles of PTBP1 in alternative splicing, glycolysis, and oncogensis. J Zhejiang Univ Sci B. 2020;21:122–36.

239. Choe J, Ryu I, Park OH, Park J, Cho H, Yoo JS, et al. eIF4AIII enhances translation of nuclear cap-binding complex–bound mRNAs by promoting disruption of secondary structures in 5′UTR. Proc Natl Acad Sci [Internet]. Proceedings of the National Academy of Sciences; 2014;111:E4577–86. Available from: https://doi.org/10.1073/pnas.1409695111

240. Prudencio M, Garcia-moreno H, Jansen-west KR, Al-shaikh RH, Gendron TF, Heckman MG, et al. Toward allele-specific targeting therapy and pharmacodynamic marker for spinocerebellar ataxia type 3. :1–12.

241. Hou X, Gong X, Zhang L, Li T, Yuan H, Xie Y, et al. Identification of a potential exosomal biomarker in spinocerebellar ataxia Type 3/Machado-Joseph disease. Epigenomics. England; 2019;11:1037–56.

242. Costa IPD, Almeida BC, Sequeiros J, Amorim A, Martins S. A pipeline to assess disease-associated haplotypes in repeat expansion disorders: The example of MJD/SCA3 locus. Front Genet. 2019;10:1–9.

243. Alves S, Nascimento-Ferreira I, Auregan G, Hassig R, Dufour N, Brouillet E, et al. Allele-specific RNA silencing of mutant ataxin-3 mediates neuroprotection in a rat model of Machado-Joseph disease. PLoS One. 2008;3:e3341.

244. Cohen-Carmon D, Meshorer E. Polyglutamine (polyQ) disorders: the chromatin connection. Nucleus. 2012;3:433–41.

245. Gusella JF, MacDonald ME, Lee J-M. Genetic modifiers of Huntington's disease. Mov Disord. 2014;29:1359–65.

246. Yau WY, Raposo M, Bettencourt C, Labrum R, Vasconcelos J, Parkinson MH, et al. The repeat variant in MSH3 is not a genetic modifier for spinocerebellar ataxia type 3 and Friedreich's ataxia. Brain. 2020;143:e25.

247. Flower M, Lomeikaite V, Holmans P, Jones L, Tabrizi SJ, Monckton DG. Reply: The repeat variant in MSH3 is not a genetic modifier for spinocerebellar ataxia type 3 and Friedreich's ataxia. Brain. England; 2020. p. e26.

248. Laffita-Mesa JM, Nennesmo I, Paucar M, Svenningsson P. A Novel Duplication in ATXN2 as Modifier for Spinocerebellar Ataxia 3 (SCA3) and C9ORF72-ALS. Mov Disord. 2021;36:508–14.

249. Ramos A, Planchat M, Vieira Melo AR, Raposo M, Shamim U, Suroliya V, et al. Mitochondrial DNA haplogroups and age at onset of Machado-Joseph disease/spinocerebellar ataxia type 3: a study in patients from multiple populations. Eur J Neurol. England; 2019;26:506–12.

250. Bettencourt C, Raposo M, Kazachkova N, Santos C, Kay T, Vasconcelos J, et al. Sequence analysis of 5′ regulatory regions of the Machado-Joseph disease gene (ATXN3). Cerebellum. United States; 2012;11:1045–50.

251. Jardim L, Silveira I, Pereira ML, do Céu Moreira M, Mendonça P, Sequeiros J, et al. Searching for modulating effects of SCA2, SCA6 and DRPLA CAG tracts on the Machado-

Joseph disease (SCA3) phenotype. Acta Neurol Scand. Denmark; 2003;107:211–4.

252. Chen Z, Sequeiros J, Tang B, Jiang H. PT SC. Ageing Res Rev [Internet]. Elsevier B.V.; 2018; Available from: https://doi.org/10.1016/j.arr.2018.10.004

253. Génin E, Feingold J, Clerget-Darpoux F. Identifying modifier genes of monogenic disease: Strategies and difficulties. Hum Genet. 2008;124:357–68.

254. Li X, Montgomery SB. Detection and impact of rare regulatory variants in human disease. Front Genet. 2013;4:67.

255. Evert BO, Vogt IR, Kindermann C, Ozimek L, de Vos RA, Brunt ER, et al. Inflammatory genes are upregulated in expanded ataxin-3-expressing cell lines and spinocerebellar ataxia type 3 brains. J Neurosci Off J Soc Neurosci. 2001;21:5389–96.

256. Evert BO, Vogt IR, Vieira-Saecker AM, Ozimek L, de Vos RAI, Brunt ERP, et al. Gene expression profiling in ataxin-3 expressing cell lines reveals distinct effects of normal and mutant ataxin-3. J Neuropathol Exp Neurol. England; 2003;62:1006–18.

257. da Silva Carvalho G, Saute JAM, Haas CB, Torrez VR, Brochier AW, Souza GN, et al. Cytokines in Machado Joseph Disease/Spinocerebellar Ataxia 3. Cerebellum. United States; 2016;15:518–25.

258. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.

259. Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet. United States; 2009;Chapter 2:Unit 2.12.

260. Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 1995;23:4878–84.

261. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour. England; 2008;8:103–6.

262. Krämer A, Green J, Pollard JJ, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30:523–30.

263. Zogopoulos L. V, Spaho K, Ntouka C, Lappas A. G, Kyranis I, Bagos G. P, et al. TFBSPred: A functional transcription factor binding site prediction webtool for humans and mice. Int J Epigen [Internet]. Centre of Systems Biology, Biomedical Research Foundation, Academy of Athens, 11527 Athens, Greece Department of Computer Science, Hellenic Open University, 26335 Patras, Greece Department of Computer Science and Biomedical Informatics, University of Thes; 2021;1:9. Available from: https://doi.org/10.3892/ije.2021.9

264. Kataria H, Alizadeh A, Karimi-Abdolrezaee S. Neuregulin-1/ErbB network: An emerging modulator of nervous system injury and repair. Prog Neurobiol. England; 2019;180:101643.

265. Xu J, Hu C, Chen S, Shen H, Jiang Q, Huang P, et al. Neuregulin-1 protects mouse cerebellum against oxidative stress and neuroinflammation. Brain Res. Netherlands; 2017;1670:32–43.

266. Merelli A, Repetto M, Lazarowski A, Auzmendi J. Hypoxia, Oxidative Stress, and Inflammation: Three Faces of Neurodegenerative Diseases. J Alzheimers Dis. Netherlands; 2021;82:S109–26.

267. Burtscher J, Mallet RT, Burtscher M, Millet GP. Hypoxia and brain aging: Neurodegeneration or neuroprotection? Ageing Res Rev. England; 2021;68:101343.

268. Reid DA, Keegan S, Leo-Macias A, Watanabe G, Strande NT, Chang HH, et al. Organization and dynamics of the nonhomologous end-joining machinery during DNA double-strand break repair. Proc Natl Acad Sci U S A. 2015;112:E2575-84.

269. Bu S, Lv Y, Liu Y, Qiao S, Wang H. Zinc Finger Proteins in Neuro-Related Diseases Progression. Front Neurosci [Internet]. 2021;15. Available from: https://www.frontiersin.org/article/10.3389/fnins.2021.760567

270. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of

functional annotations within the Gene Ontology consortium. Brief Bioinform. 2011;12:449–62.

271. Bialek K, Czarny P, Watala C, Wigner P, Talarowska M, Galecki P, et al. Novel association between TGFA, TGFB1, IRF1, PTGS2 and IKBKB single-nucleotide polymorphisms and occurrence, severity and treatment response of major depressive disorder. PeerJ [Internet]. PeerJ Inc.; 2020;8:e8676–e8676. Available from: https://pubmed.ncbi.nlm.nih.gov/32140313

272. Evert BO, Schelhaas J, Fleischer H, de Vos RAI, Brunt ER, Stenzel W, et al. Neuronal intranuclear inclusions, dysregulation of cytokine expression and cell death in spinocerebellar ataxia type 3. Clin Neuropathol. Germany; 2006;25:272–81.

273. Ou G, Lin W, Zhao W. Neuregulins in Neurodegenerative Diseases. Front Aging Neurosci [Internet]. 2021;13. Available from: https://www.frontiersin.org/article/10.3389/fnagi.2021.662474

274. Woo R-S, Li X-M, Tao Y, Carpenter-Hyland E, Huang YZ, Weber J, et al. Neuregulin-1 Enhances Depolarization-Induced GABA Release. Neuron [Internet]. 2007;54:599–610. Available from: https://www.sciencedirect.com/science/article/pii/S0896627307002590

275. Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell. 2019;179:750-771.e22.

276. Wang M, Ngo V, Wang W. Deciphering the genetic code of DNA methylation. Brief Bioinform. 2021;22.

277. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. England; 2012;13:484–92.

278. Bettencourt C, Foti SC, Miki Y, Botia J, Chatterjee A, Warner TT, et al. White matter DNA methylation profiling reveals deregulation of HIP1, LMAN2, MOBP, and other loci in

multiple system atrophy. Acta Neuropathol [Internet]. Springer Berlin Heidelberg; 2020;139:135–56. Available from: https://doi.org/10.1007/s00401-019-02074-0

279. Sanchez-Mut J V, Heyn H, Vidal E, Moran S, Sayols S, Delgado-Morales R, et al. Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. Transl Psychiatry. 2016;6:e718.

280. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J Biol Chem. 2020;295:4134–70.

281. Lu AT, Narayan P, Grant MJ, Langfelder P, Wang N, Kwak S, et al. DNA methylation study of Huntington's disease and motor progression in patients and in animal models. Nat Commun. 2020;11:4529.

282. Hamzeiy H, Savaş D, Tunca C, Şen NE, Gündoğdu Eken A, Şahbaz I, et al. Elevated Global DNA Methylation Is Not Exclusive to Amyotrophic Lateral Sclerosis and Is Also Observed in Spinocerebellar Ataxia Types 1 and 2. Neurodegener Dis. Switzerland; 2018;18:38–48.

283. Zadel M, Maver A, Kovanda A, Peterlin B. DNA Methylation Profiles in Whole Blood of Huntington's Disease Patients. Front Neurol. 2018;9:655.

284. De Souza RAG, Islam SA, McEwen LM, Mathelier A, Hill A, Mah SM, et al. DNA methylation profiling in human Huntington's disease brain. Hum Mol Genet. England; 2016;25:2013–30.

285. Mollica PA, Reid JA, Ogle RC, Sachs PC, Bruno RD. DNA Methylation Leads to DNA Repair Gene Down-Regulation and Trinucleotide Repeat Expansion in Patient-Derived Huntington Disease Cells. Am J Pathol. United States; 2016;186:1967–76.

286. Ding D, Wang C, Chen Z, Xia K, Tang B, Qiu R, et al. Polyglutamine-expanded ataxin3 alter specific gene expressions through changing DNA methylation status in SCA3/MJD. Aging (Albany NY). 2020;13:3680–98.

287. Laffita-Mesa JM, Bauer PO, Kourí V, Peña Serrano L, Roskams J, Almaguer Gotay D, et al. Epigenetics DNA methylation in the core ataxin-2 gene promoter: novel physiological and pathological implications. Hum Genet. Germany; 2012;131:625–38.

288. Emmel VE, Alonso I, Jardim LB, Saraiva-Pereira ML, Sequeiros J. Does DNA methylation in the promoter region of the ATXN3 gene modify age at onset in MJD (SCA3) patients? Clin Genet. 2011;79:100–2.

289. Bauer PO, Zumrova A, Matoska V, Mitsui K, Goetz P. Can ataxin-2 be down-regulated by allele-specific de novo DNA methylation in SCA2 patients? Med Hypotheses. United States; 2004;63:1018–23.

290. Fitzgerald KN, Hodges R, Hanes D, Stack E, Cheishvili D, Szyf M, et al. Potential reversal of epigenetic age using a diet and lifestyle intervention: a pilot randomized clinical trial. Aging (Albany NY). 2021;13:9419–32.

291. Dugué P-A, Jung C-H, Joo JE, Wang X, Wong EM, Makalic E, et al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. Epigenetics. 2020;15:358–68.

292. McGinnis KA, Justice AC, Tate JP, Kranzler HR, Tindle HA, Becker WC, et al. Using DNA methylation to validate an electronic medical record phenotype for smoking. Addict Biol. 2019;24:1056–65.

293. Huang L-H, Lin P-H, Tsai K-W, Wang L-J, Huang Y-H, Kuo H-C, et al. The effects of storage temperature and duration of blood samples on DNA and RNA qualities. PLoS One. 2017;12:e0184692.

294. Bulla A, De Witt B, Ammerlaan W, Betsou F, Lescuyer P. Blood DNA Yield but Not Integrity or Methylation Is Impacted After Long-Term Storage. Biopreserv Biobank. United States; 2016;14:29–38.

295. Bergsma T, Rogaeva E. DNA Methylation Clocks and Their Predictive Capacity for

Aging Phenotypes and Healthspan. Neurosci insights. 2020;15:2633105520942221.

296. Horvath S, Langfelder P, Kwak S, Aaronson J, Rosinski J, Vogt TF, et al. Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. Aging (Albany NY). 2016;8:1485–512.

297. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging (Albany NY). 2015;7:1130–42.

298. Oblak L, van der Zaag J, Higgins-Chen AT, Levine ME, Boks MP. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. Ageing Res Rev. England; 2021;69:101348.

299. Xu M, Zhu J, Liu X-D, Luo M-Y, Xu N-J. Roles of physical exercise in neurodegeneration: reversal of epigenetic clock. Transl Neurodegener. 2021;10:30.

300. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11:303–27.

301. Zhao W, Ammous F, Ratliff S, Liu J, Yu M, Mosley TH, et al. Education and Lifestyle Factors Are Associated with DNA Methylation Clocks in Older African Americans. Int J Environ Res Public Health. 2019;16.

302. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10:573–91.

303. Quach A, Levine ME, Tanaka T, Lu AT, Chen BH, Ferrucci L, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. Aging (Albany NY). 2017;9:419–46.

304. Qu Y, Lennartsson A, Gaidzik VI, Deneberg S, Bengtzén S, Arzenani MK, et al. Genome-Wide DNA Methylation Analysis Shows Enrichment of Differential Methylation in "Open Seas" and Enhancers and Reveals Hypomethylation in DNMT3A Mutated Cytogenetically Normal AML (CN-AML). Blood [Internet]. 2012;120:653. Available from: https://www.sciencedirect.com/science/article/pii/S0006497118533449

305. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics [Internet]. 2014;30:1363–9. Available from: https://doi.org/10.1093/bioinformatics/btu049

306. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;33:3982–4.

307. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics [Internet]. 2013;14:293. Available from: https://doi.org/10.1186/1471-2164-14-293

308. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

309. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res [Internet]. 2016;44:W83–9. Available from: https://doi.org/10.1093/nar/gkw199

310. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res [Internet]. 2019;47:W191–8. Available from: https://doi.org/10.1093/nar/gkz369

311. Huang Y, Todd N, Thathiah A. The role of GPCRs in neurodegenerative diseases: avenues for therapeutic intervention. Curr Opin Pharmacol. England; 2017;32:96–110.

312. Vassilatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, Mortrud MT, et al. The G protein-coupled receptor repertoires of human and mouse. Proc Natl Acad Sci U S A. 2003;100:4903–8.

313. Parenti A, De Logu F, Geppetti P, Benemei S. What is the evidence for the role of TRP

channels in inflammatory and immune cells? Br J Pharmacol. 2016;173:953–69.

314. Bettencourt C, Ryten M, Forabosco P, Schorge S, Hersheson J, Hardy J, et al. Insights From Cerebellar Transcriptomic Analysis Into the Pathogenesis of Ataxia. JAMA Neurol [Internet]. 2014;71:831–9. Available from: https://doi.org/10.1001/jamaneurol.2014.756

315. Listos J, Łupina M, Talarek S, Mazur A, Orzelska-Górka J, Kotlińska J. The Mechanisms Involved in Morphine Addiction: An Overview. Int J Mol Sci. 2019;20.

UNIVERSIDADE DOS AÇORES
Faculdade de Ciências e Tecnologia

Rua da Mãe de Deus
9500-321 Ponta Delgada
Açores, Portugal