Selective constraints and expression pattern of the *ataxin-3 like 1* retrogene

Andreia Patrícia de Sousa Pinto
Mestrado em Genética Forense
Departamento de Biologia
2013

Orientador
Doutora Sandra Martins, Investigadora Post-doc no Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP)

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

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

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

Quero começar por prestar os meus agradecimentos à primeira pessoa responsável pela minha jornada no IPATIMUP, o Professor Amorim, por quem eu nutro uma grande admiração, não só pelo seu incontestável mérito profissional como pela sua personalidade única e de moderada irreverência.

O maior agradecimento deve ser entregue à minha orientadora, Sandra Martins, entusiasta, dedicada, persistente, apaixonada, prestável, paciente… Por todas estas características fantásticas e pelo elevado peso que tem tido na minha educação, ensinando-me mais sobre Ciência e sobre o que é ser cientista.

A todos os elementos do grupo de Genética Populacional do IPATIMUP que atenderam a diversas dúvidas e necessidades de ajuda por que passei e também pelas chamadas de atenção nos momentos em que errava, sempre de forma informal e a transparecer companheirismo. Um especial obrigado à Alexandra Lopes pela sua colaboração neste trabalho.

À Dra. Susana Seixas pela sua colaboração e ajuda dispensada em benefício deste trabalho.

Ao IPATIMUP pelo financiamento e apoio científico sem o qual o sucesso do trabalho não teria sido possível.

Aos colegas de Genética Forense 2011/2012 pelo companheirismo e aos alunos de mestrado desse mesmo ano pela receção acolhedora e por todo o acompanhamento para mim indispensável e irretribuível, ficando inevitavelmente em dívida para com vocês. Um especial obrigado à Inês Martins por me ter orientado nos primeiros passos do meu trabalho.

Às minhas grandes amigas da *Generation Y Dance Crew* que fazem questão de me acompanhar em todos os meus passos que não sejam só os de dança, mostrando-se sempre orgulhosas de mim em todas as minhas conquistas e obrigando-me a ser sempre melhor em tudo para terem um bom exemplo a seguir.

Aos meus pais que me permitem estudar, um “luxo” que lhes foi retirado aos 10 anos de idade, e engrandecer na vida sempre com humildade.

E por fim, mas não menos importante, à minha avó que no passado ano, já no seu leito de morte, me confessou que levaria consigo um desgosto: ser analfabeta. Talvez por isso, sabia melhor que ninguém a importância da educação e foi, outrora, um dos grandes incentivos a que eu, ainda na escola primária, definisse o meu grande objetivo de vida: obter formação superior na área das Ciências Biológicas, que em tão inocente idade eu entendia como - “quero ir para a universidade para cuidar dos animais”.

Abstract

The *ataxin-3* gene (*ATXN3*; 14q32.1) encodes a ubiquitously expressed deubiquitinating enzyme, conserved throughout eukaryotes, with homologues among metazoans and found also in plants and protozoans. In humans, *ATXN3* may expand and encode an abnormally long polyglutamine stretch, polyQ, responsible for the most common dominant ataxia worldwide: Machado-Joseph disease (MJD/SCA3). The expanded protein gains toxic properties but the normal ataxin-3 seems also to influence the disease pathogenesis. *ATXN3* has two poorly studied paralogues, originated by transposition events: *ataxin-3 like* (*ATXN3L*; Xp22.2) and *LOC100132280* (8q23.2), here named *ATXN3L1* and *ATXN3L2*, respectively. Interestingly, a recent *in vitro* study has shown the ability of ATXN3L1 to cleave ubiquitin substrates, with its protease domain significantly more efficient than the ATXN3 domain. Regarding *ATXN3L2*, *in silico* analyses previously performed in our lab predicted only short reading frames for its translation. Still, if transcribed, these short mRNA molecules may have functional relevance by regulating the expression of the parental gene, as demonstrated for other transcripts originated from retrocopies.

Our aims were (1) to study the mechanisms that led to the human-specific $(CAG)_n$ expansion in the parental *ATXN3*, but not in other lineages or paralogues; (2) to estimate the onset of the retrotransposition events; (3) to compare the rates of evolution and selective constraints among all three genes in the primate lineage; and (4) to assess the mRNA expression pattern of the retrocopy that conserved an intact ORF (*ATXN3L1*) in human tissues.

We have found a *ATXN3L1* CAG repeat poorly polymorphic and highly interrupted across the primate lineage, whereas a pure $(CAG)_n$ was observed in *ATXN3L2* of most species, with humans and chimpanzees showing a polymorphic $(CGGCAG)_n$. Instability encompassing this hexanucleotide may have resulted from a CAG to CGG mutation occurred after the Gorilla-Pan split, followed by unequal crossover. We have also found evidence of unequal crossover events in *ATXN3* of orangutan. Phylogenetic results suggested that *ATXN3L1* and *ATXN3L2* arose by two independent retrotransposition events, in Haplorrhini (~63 MYA), and before the Platyrrhini-Catarrhini split (~43 MYA), respectively. Branch models performed indicated that *ATXN3L1* has been under selective constraints throughout primate evolution, reinforcing its functional relevance, whereas *ATXN3L2* gained premature stop codons that likely turned it into a processed pseudogene. In fact, we confirmed by Reverse Transcriptase PCR, that *ATXN3L1* is transcriptionally-active in humans, with mRNA expression in testis, placenta, brain, spleen, and

thymus. Further studies will be done to confirm the presence of endogenous protein *in vivo* and to better understand the functional diversification of *ATXN3L1*.

Keywords: *Ataxin-3*; Retrotransposition; Paralogue genes; Machado-Joseph disease; Evolution

Resumo

O gene a*taxin-3* (*ATXN3*; 14q32.1) codifica uma enzima de expressão ubíqua pertencente ao grupo das desubiquitinases. A ataxina-3 é conservada nos eucariotas, com ortólogos descritos nos metazoários, em plantas e, também, em protozoários. Em humanos, o gene *ATXN3* pode apresentar uma região repetitiva expandida que codifica uma proteína poliglutamínica mais longa do que o normal. A doença de Machado-Joseph (DMJ/SCA3), considerada a ataxia dominante mais comum em todo o mundo, é causada por ganho de função da proteína expandida, mas a ataxina-3 normal parece ser um modificador da apresentação clínica da doença. O gene *ATXN3* tem dois genes parálogos ainda pouco estudados, originados a partir de eventos de retrotransposição: *ataxin-3 like* (*ATXN3L*; Xp22.2) e *LOC100132280* (8q23.2), denominados por nós como *ATXN3L1* e *ATXN3L2*, respetivamente. Relativamente a *ATXN3L1*, é interessante referenciar um estudo *in vitro* realizado recentemente que demonstra a capacidade da ATXN3L, através do seu domínio proteolítico, quebrar ligações entre ubiquitinas mais eficazmente do que a ataxina-3 parental. Quanto a *ATXN3L2*, análises *in silico* já realizadas por nós, demonstraram a aquisição de vários codões stop prematuros, o que se reflete na existência de *open reading frames* (ORFs) muito interrompidas. Ainda assim, não se pode negligenciar a importância deste parálogo pois, tal como demonstrado para outros transcritos orginados a partir de retrocópias, as pequenas moléculas de mRNA podem desempenhar funções ao nível da regulação da expressão do gene parental.

Os nossos objetivos foram (1) o estudo dos mecanismos que levaram à expansão do $(CAG)_n$ no gene *ATXN3* humano; (2) a determinação da origem dos eventos de retrotransposição que levaram ao surgimento dos dois parálogos em estudo; (3) a comparação das taxas de evolução e das pressões seletivas dos 3 genes em estudo ao longo das diversas linhagens de primatas; e (4) a determinação do padrão de expressão do gene *ATXN3L1* em tecidos humanos, uma vez que se apresenta como o gene mais conservado.

De acordo com os nossos resultados, o motivo repetitivo CAG do gene *ATXN3L1* apresenta-se pouco polimórfico e altamente interrompido ao longo das linhagens de primatas, enquanto que o gene *ATXN3L2* possui um motivo repetitivo puro em quase todas as espécies estudadas, excetuando, por exemplo, os humanos e os chimpanzés que apresentam o motivo $(CGGCAG)_n$. A instabilidade inerente a este hexanucleotídio poderá ter resultado a partir de uma mutação de CAG para CGG logo após a divergência gorila-chimpanzé, seguida de recombinação desigual. Análises filogenéticas sugerem dois eventos de retrotransposição

distintos na origem de *ATXN3L1* e *ATXN3L2*, ocorridos, respectivamente, há cerca de 63 milhões de anos nos haplorrínios e 43 milhões de anos antes da divergência platirrínios-catarrínios. Os valores de omega (ω) calculados para *ATXN3L1* indicam que se trata de um gene sob pressões seletivas purificadoras, sugerindo, assim, tratar-se de um retrogene com relevância funcional. O RT-PCR (reverse transcriptase) realizado por nós confirma isso mesmo, mostrando que o gene *ATXN3L1* é transcrito pelo menos em testículo, placenta, cérebro, baço e timo. Subsequentemente, terão de ser elaborados mais estudos para confirmar a presença da proteína *in vivo* e perceber melhor a função desta proteína.

Palavras-chave: Ataxina-3; Retrotransposição; Genes parálogos; Doença de Machado-Joseph; Evolução

Table of Contents

Agradecimentos 1

Abstract 3

Resumo 5

Table of Contents 7

Tables 9

Figures 11

Abbreviations 13

Introduction 15

1. Origin of new genes 17

1.1 Retrotransposition events 17

1.1.1 Retrotrasposons as genetic markers 19

2. *ATXN3* paralogues 20

3. J D proteins 21

3.1 Ataxin-3 21

3.1.1 Function 22

3.2 ATXN3L 24

3.3 Josephins 24

4. Polyglutamine diseases 25

5. Machado-Joseph disease 26

Objectives 29

Materials and Methods 33

1. DNA samples 35

2. Whole genome DNA amplification 36

3. Amplification of *ATXN3L1* and *ATXN3L2* 36

4. Amplification of parental *ATXN3* 38

5. Detection of amplified products 39
6. Sequencing 39
7. Sex determination 40
8. Multiple sequence alignment and neighbor-joining phylogenetic trees....41
9. Phylogenetic analysis of selection (ω ratio calculations)....42
10. Transcriptional pattern of *ATXN3L1* in human tissues....43
10.1 RT-PCR....43
10.2 Amplification....43
Results....45
1. Evolution of the $(CAG)_n$ tract in *ATXN3*, *ATXN3L1* and *ATXN3L2* 47
2. Sex determination 49
3. Intraspecific *ATXN3L1* duplication events....50
4. *ATXN3L2* in *Aotus trivirgatus*....53
5. Origin of the retrotransposition events 54
6. Phylogenetic analysis of selection 56
6.1 Branch models 57
6.2 Site models 59
6.3 Branch-site model 60
7. Nucleotide diversity of *ATXN3L1* and *ATXN3L2* 61
8. Transcriptional pattern of *ATXN3L1*....62
Discussion 63
Future perspectives 73
References 77
Appendix....85

Tables

Table 1: Non-human primate species sequenced in our lab. 35

Table 2: Sequences retrieved from public databases. 35

Table 3: Primers designed to specifically amplify *ATXN3L1* and *ATXN3L2* genes. 36

Table 4: Summary of the PCR conditions and set of primers used to amplify our samples. 38

Table 5: Summary of the PCR conditions and set of primers used to amplify our samples. 38

Table 6: Primers designed to specifically amplify the *ATXN3* gene. 38

Table 7: Annealing temperatures used to amplify polymorphic markers located in the X chromosome for gender determination of some analyzed samples. 41

Table 8: Primers designed to specifically amplify *ATXN3L1* in human cDNA. 43

Table 9: *ATXN3*, *ATXN3L1* and *ATXN3L2* $(CAG)_n$ tracts obtained both from public databases and our sequencing of several primates. 48

Table 10: Capillary electrophoresis results for four polymorphic markers within or flanking the *androgen receptor* gene. 49

Table 11: Heterozygozity pattern and polymorphic positions of *M. fuscata ATXN3L1*. 50

Table 12: *ATXN3L1* polymorphic positions found in *Papio anubis*. 52

Table 13: Parameter estimates and likelihood scores under different branch models. Model comparisons of variable ω ratios among branches. 57

Table 14: Parameter estimates and likelihood scores under different site models. Model comparisons of variable ω ratios among sites. 59

Table 15: Parameter estimates and likelihood scores under different branch-site models. Model comparisons of variable ω ratios among sites and branches. 60

Table 16: Inferred ancestral alleles for *ATXN3L1* in *P. anubis* based on orthologous sequences observed in other primates. 70

Figures

Figure 1: Classification of the different types of retrotransposons. 17

Figure 2: Structure of mammalian transposable elements. 18

Figure 3: Comparison of the ATXN3 and ATXN3L1 structures. 20

Figure 4: Domain architecture, structure and post-translation modifications of ATXN3. 22

Figure 5: ATXN3 activity, interactions and proposed biological roles. 23

Figure 6: Schematic representation of A. *ATXN3L1*; and B. *ATXN3L2* genes, with distances between selected primers. 37

Figure 7: General PCR protocol with length of time and temperatures for the amplification of *ATXN3L1* and *ATXN3L2* loci. 37

Figure 8: Amplification conditions of the $(CAG)_n$ region in the *androgen receptor* gene with A. My Taq™ HS MIX; and B. QIAGEN Multiplex PCR Kit. 41

Figure 9: *ATXN3* gene structure and primers insertion location. 44

Figure 10: Alignment of *AMELX* and *AMELY* sequences with A. a male chimpanzee control sample; and B. a female owl monkey control sample. 49

Figure 11: Example of two *ATXN3L1* variant positions in *M. fuscata*. 51

Figure 12: The different patterns of heterozygozity found in *M. fuscata* samples while sequencing *ATXN3L1* 51

Figure 13: Neighbor-Joining phylogenetic tree for *ATXN3* and *ATXN3L1* showing the two sequences retrieved from UCSC to *Papio anubis*. 53

Figure 14: Neighbor-Joining phylogenetic tree showing *ATXN3* and *ATXN3L2* in several primates, and evidencing the origin of *Aotus trivirgatus ATXN3L2* and *ATXN3L?* sequences. 54

Figure 15: Neighbor-Joining phylogenetic tree for *ATXN3*, *ATXN3L1*, and *ATXN3L2*. 55

Figure 16: Phylogenetic tree of primates. 56

Figure 17: Phylogeny of the 8 primate species analyzed. 58

Figure 18: *ATXN3L1* transcriptional status of 20 human tissues and chimpanzee brain using E7-8 primers 62

Figure 19: Schematic representation of the unequal crossing over events that may have led to the present $(CAG)_n$ tract in A. *ATXN3L2* of *Pan troglodytes*; and B. *ATXN3* of *Pongo abelii*. 66

Figure 20: Phylogenetic tree showing the species where duplications of *ATXN3L1* seem to have occurred. 68

Figure 21: *ATXN3L1* (L1) and *ATXN3L1 duplicate* (dL1) haplotypes for A. positions c.561T>G, c.743A>G, c.923A>C and c.995C>T in *M. fuscata* individuals; and B. c.801A>C and c.995C>T in *M. fascicularis* males. 69

Abbreviations

A – Adenine

APS - ammonium persulfate

ATXN1 – Ataxin-1

ATXN1L – Ataxin-1 like

ATXN3 – Ataxin-3

ATXN3L1 – Ataxin-3 like 1

ATXN3L2 – Ataxin-3 like 2

bp – base pairs

C – Citosine

CAMP - Cyclic adenosine monophosphate

CBP - CREB-binding protein

cDNA – complementary DeoxyriboNucleic Acid

CREB - cAMP response element-binding protein

dL1 - *Ataxin-3 like 1* duplicate

DNA – DeoxyriboNucleic Acid

dNTP – deoxyNucleotide TriPhosphate

DRPLA – Dentatorubral-pallidoluysian atrophy

DUB – DeUbiquitinating enzyme

ExoFastAP – Exonuclease I and Thermosensitive Alkaline Phosphatase

G – Guanine

gDNA – genomic DeoxyriboNucleic Acid

HAT – Histone Acetiltransferase

HD – Huntington's Disease

HDAC - Histone deacetilases

Herv – Human Endogenous Retroviruses

JD – Josephin Domain

JOSD1 – Josephin domain containing protein 1

JOSD2 – Josephin domain containing protein 2

L1 – Long interspersed nuclear element 1

LINE – Long Interspersed Nucleotide Element

LTR – Long Terminal Repeat

min - minutes

MJD – Machado-Joseph Disease

MMP – Matrix MetalloProteinase

mRNA – messenger Ribonucleic Acid

MYA – Million Years Ago

NCBI – National Center for Biotechnology Information

NcoR - Nuclear receptor co-repressor 1

NII – Nuclear inclusions

ORF – Open Reading Frame

PCAF - P300/CBP-associated factor

PCR – Polymerase Chain Reaction

PolyQ – poly glutamine

RNA – Ribonucleic Acid

rpm – rotations per minute

RSP - restriction site polymorphisms

RT – Reverse Transcriptase

SBMA - Spinal and Bulbar Muscular Atrophy

SCA – SpinoCerebellar Ataxia

sec - seconds

SINE – Short Interspersed Nucleotide Element

SNP – Single Nucleotide Polymorphism

STR – Short Tandem Repeat

T – Timine

TEMED – Tetramethylethylenediamine

TSD – Target Site Duplication

Ub – Ubiquitin

UCSC –University of California, Santa Cruz

UIM – Ubiquitin Interaction Motif

UPP – Ubiquitin-proteasome pathway

UTR - Untranslated Region

Xp – short arm chromosome X

Yq – long arm chromosome Y

Introduction

1. Origin of new genes

The emergence of new genes is fundamental to the evolution of species-specific traits [1]. The major mechanism providing raw material for the origin of new genes is gene duplication. There are two types of gene duplication: direct duplication of genomic DNA and retropositional events [2]. In the first case, gene copies present often the same exon-intron organization and similar expression patterns when compared to parental genes: duplication of chromosomal segments [1]. Gene duplicates may, however, lose their core promoters and their protein-coding potential if tandem duplication or uneven crossing-over is in their origin [3]. In the case of retrotransposition, re-integration of reverse transcribed mRNA molecules occurs in the genome [4]. This is the mechanism by which our genes of interest have arisen, and for this reason, in next topics, we will explain it in more detail.

1.1 Retrotransposition events

Transposable elements have contributed greatly to what we now realize to be a highly dynamic genome. An example are the retrotransposons which encode a reverse transcriptase (RT) activity and move by a “copy and paste” process involving an RNA intermediate [5].

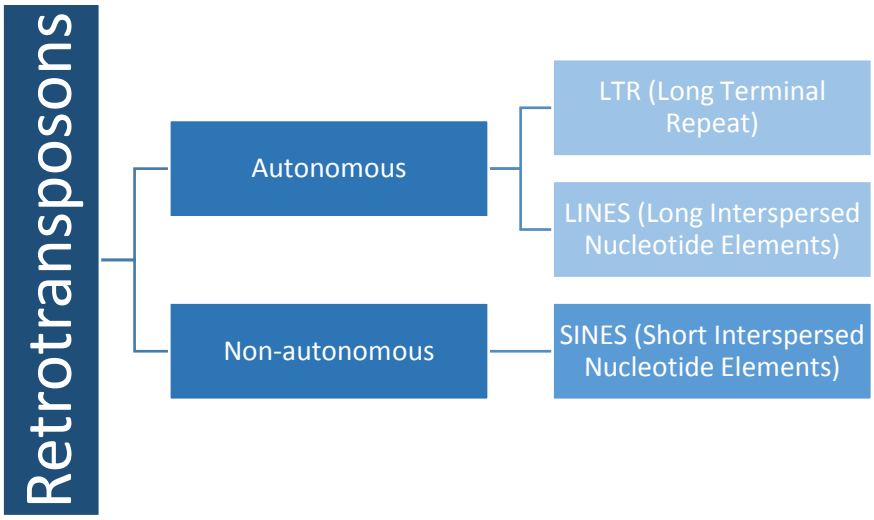
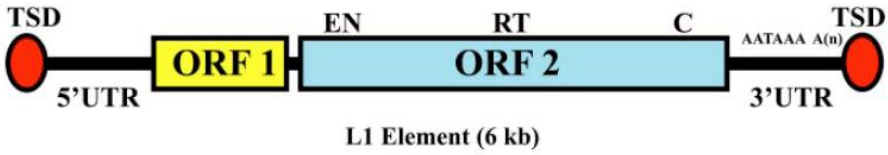


Figure 1: Classification of the different types of retrotransposons.

L1 elements are autonomous non-LTR (Long Terminal Repeat) retrotransposons classified as LINEs (Long Interspersed Nucleotide Elements) and are widely present in mammalian genomes (about 25% of their genome), being responsible for a burst on the number of retrocopies in

mammalian lineages [4]. Besides duplicating themselves, L1 elements likely have been responsible for the genomic expansion of the non-autonomous retrotransposons SINES (Short Interspersed Nucleotide Elements) since they do not encode any proteins and nor possess an independent mobility. Among SINEs, we can find Alu elements, processed pseudogenes, and other elements in the human genome [5].

A.



B.

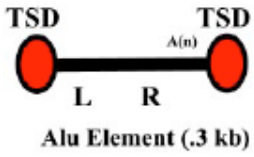


Figure 2: Structure of mammalian transposable elements. A. L1 consists of a 5' untranslated region (5'UTR), two ORFs (Open Reading Frames) separated by a short intergenic region, a 3'UTR, a polyA signal (AATAAA), and a polyA tail ($A_{(n)}$). L1 elements are often flanked by 7–20 bp target site duplications (TSD). In ORF2, both RT (reverse transcriptase) and EN (endonuclease) domains, as well as a conserved cysteine-rich motif (C) are annotated; B. Alu elements contain two similar sequences, the left (L) and right monomers (R) and end in a polyA tail. (Adapted from Ostertag, 2001) [5].

Retrocopies lack many of the genetic features of their parental genes, such as introns and regulatory elements [6]; for a long time, this has contributed to the idea that only non-functional duplicate gene copies could result from gene retrotransposition. Subsequent analyses suggested, however, that retroposition had efficiently sown the seeds of evolution in genomes [7]; and more recent genome-wide searches of retroposed genes have confirmed the importance of gene retrotransposition in the origin of new functional genes in fly and primates [8-14]. In humans, although selection generally acted against the insertion and high transcription of retrocopies located inside other genes, it favored the emergence of a substantial number of new genes with diverse gene structures and functions [1].

Retrogenes have for long time been described as pseudogenes, however, since raising evidence of retrocopies functionality, the term retrogene has now been widely used to describe

functional retrocopied genes. On the other hand, pseudogenes consist in genomic DNA sequences lacking the protein-coding capability of their paralogous counterpart because of frameshifts or premature stop codons [15]. Thus, the term “pseudogene” implies non-functionality, frequently viewed as a molecular fossil and used to measure background genomic substitution rates [16, 17].

Retrogenes may directly inherit promoters from their parental transcripts as has been observed for individual cases [18], but they often use the regulatory elements of host genes that surround the insertion site (suggesting that retrocopy transcription is frequently driven by the machinery of nearby genes). In fact, regions surrounding retrogenes are transcriptionally more active than regions flanking silent retrocopies. New retrogenes may also emerge from truncated regions coding small RNAs that may regulate other genes expression [1].

The transcription of a retrogene is not, however, sufficient evidence of biological function. Additional confirmation of the functionality of these retrocopies is their ancient origin and strong conservation throughout evolution [15]. The analysis of conserved retrogenes has shown that 50% are indeed conserved across millions of years of primate evolution (far fewer are though to be conserved between species more distant to human, such as rodents) [3].

1.1.1 Retrotransposons as genetic markers

As revealed by pioneering studies on humans, primates and other non-primate groups, retrotransposons afford several advantages that make them very powerful tools as genetic markers for studying human and non-human primate evolutionary histories [19-22].

When compared to retrotransposons, the most often used genetic markers (such as single nucleotide polymorphisms (SNPs), restriction site polymorphisms (RSPs) and short tandem repeats (STRs)) suffer from higher probabilities of homoplasy and greater difficulty of confidently establishing the ancestral marker state. On the other hand, LINE and SINE insertions have two uniquely valuable properties as markers for phylogenetic and population genetic analyses: first, they are virtually free of homoplasy since every observed insertion of the same element sequence is identical by descent; second, the ancestral state of the locus is known to be the absence of the insertion [23].

Retrotransposon insertion polymorphisms have also being used as forensic tools, for example, in species specific DNA detection and quantitation, analysis of complex biomaterials, human gender determination and inference of geographic origin of human samples [24].

2. *ATXN3* paralogues

Ataxin-3 (*ATXN3*; 14q32.2) was described as containing 11 exons, resulting in at least four different mRNAs due to alternative splicing. More recently, two additional exons (6a and 9a) and 56 alternative transcripts have been described [25, 26]. The $(CAG)_n$ tract, located in exon 10, displays usually the $(CAG)_2$ CAA AAG CAG CAA $(CAG)_n$ configuration [27]. In most transcripts, including the widely frequent *ATXN3-001*, this repeat is in the 5' part of the exon, with a single adenine nucleotide before the reiteration site, at the beginning of the exonic region [28].

Orthologues and paralogues are types of homologous genes that are related by speciation or duplication, respectively [29, 30]. The *ATXN3* gene has two paralogues in the genome of several primates. Both *ataxin-3 like* (Xp22.2) and *LOC100132280* (8q23.2) (here named *ataxin-3 like 1*, *ATXN3L1*; and *ataxin-3 like 2*, *ATXN3L2*, respectively) are intronless copies of *ATXN3*, which hint retrotransposition for their origin. In humans, *ATXN3* and *ATXN3L1* share 85% sequence identity in their catalytic Josephin domains and high similar crystal structures [31].

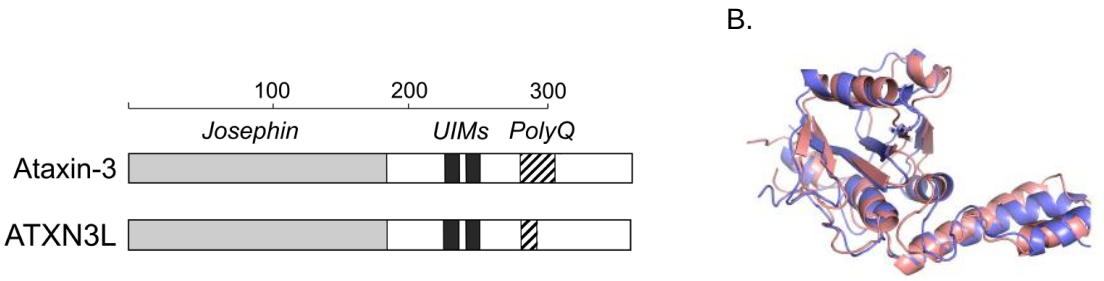


Figure 3: Comparison of the ATXN3 and ATXN3L1 structures. A. Schematic representation of ATXN3 and ATXN3L1 proteins. Josephin domains are shown in light gray, the UIMs are shown in dark gray, and the polyglutamine (polyQ) repeat regions are shown as cross-hatched. The sizes of the UIMs and polyQ regions are not shown to scale. The scale at top shows length in amino acids; B. Superposition of the ATXN3L1 crystal structure (blue) and free ATXN3 solution structure. (Adapted from Weeks et al., 2011) [31].

The other *ATXN3* paralogue, *ATXN3L2* shares about 70% homology with the parental gene, in humans. This homologous region presents an interrupted ORF, suggesting that this is a processed pseudogene of *ATXN3* [32].

Regarding that the two human copies of *ATXN3* described above remain poorly studied, it is of great importance to better characterize these retrocopies and to further explore, in the case of *ATXN3L1*, its potential role in the pathogenesis of Machado-Joseph Disease.

3. J D proteins

Josephin Domain (JD) containing proteins are the smallest family of deubiquitinating enzymes (DUBs), all sharing a highly conserved JD catalytic domain. They are found in a wide range of phylogenetic groups, including protists, plants, and metazoans. The term MJD class has also been used after the identification of the neurodegenerative Machado-Joseph disease (MJD), caused by the best-studied protein of this class, ATXN3. This protein family comprises two defined subgroups named ataxins and Josephins, based on the sequence homology of their JD. In almost all species, Josephins (JOSD1 and JOSD2) consist of JD alone, while ataxins (ATXN3 and ATXN3L1) contain an additional C-terminus with two or three ubiquitin interaction motifs (UIMs) and a polyQ tract [33].

Little is known about ATXN3L, JOSD1 and JOSD2, and only very recent studies have provided some insights into their characterization [31, 34].

3.1 Ataxin-3

Ataxin-3 is the best-characterized JD protein [26]. It presents an approximate molecular weight of 40-43 kDa in normal individuals and structurally is composed by a globular N-terminal Josephin domain with a papain-like fold, combined with an unstructured C-terminus that contains two or three UIMs and the polymorphic polyQ tract [33]. The catalytic pocket consists of a glutamine (Q9) and a cysteine (C14) residue located in the N-terminal part of the JD, and a histidine (H119) and an asparagine (N134) in the JD C-terminal part [26]. When ATXN3 is not bound to other cellular components, its structure is much less compact than when interacting with other cellular partners, which stabilizes the fold of the C-terminus [35]. The N-terminal domain and the catalytic cysteine residue consist mainly of alpha helices, while the catalytic histidine and the orienting glutamine residue are found in a beta-strand context [36].

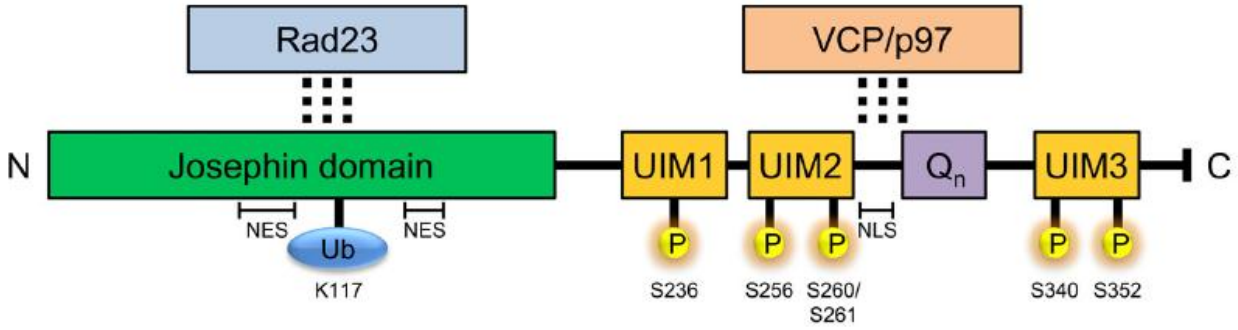


Figure 4: Domain architecture, structure and post-translation modifications of ATXN3. ATXN3 is mainly composed of a globular N-terminal catalytic domain, the JD, with DUB activity, followed by a flexible C-terminal tail containing 2 or 3 UIMS (depending on the alternative transcripts) and a polyQ sequence of variable length. Five serine residues (S236, S256, S260/S261, S340, S352) have been identified in the UIMs as phosphorylation sites, and a ubiquitinatable lysine residue was mapped to aminoacid 117, in the JD. One functional nuclear localization signal (NLS) was described in the region linking the second UIM to the polyQ region and two nuclear export signals (NES) were reported in the JD. Rad23 and VCP/p97, the two most frequently described interacting partners of ATXN3, bind to the JD and the C-terminal region of the protein, respectively [37].

3.1.1 Function

It has been suggested that the wild-type ATXN3 protein could act mainly as a deubiquitinating enzyme in the ubiquitin-proteasome pathway (UPP) [38]. DUBs, as well as other regulators of the UPP, have been linked to neurodegenerative diseases that are characterized by protein misfolding and aggregation [31]. Ubiquitin (Ub) is an 8-kDa protein covalently attached to lysine residues in specific substrates. It is implicated in diverse cellular processes and pathways including transcription, protein degradation, cell cycle progression, viral infection, and immune response [39, 40]. Once attached, ubiquitin can be removed by DUB enzymes, proteases that specifically remove adducts from the C-termini of ubiquitin molecules [31].

The catalytic activity of ATXN3 affects besides its own protein levels and ubiquitination pattern, its subcellular localization [41]. ATXN3 is ubiquitously expressed, existing in the nucleus and predominantly in cytoplasm of various cell types [42]. Nuclear export of ataxin-3 is dependent on a complex ATXN3 motif, located in the N-terminal portion of the protein, which requires the Josephin domain plus the ubiquitin-interacting motifs [41].

Many cell models have suggested a role for ATXN3 in transcriptional repression via inhibition of histone acetyltransferase (HAT) activity of the major transcriptional coactivators, cAMP response, element-binding protein (CREB), CREB binding protein (CBP), p300, and p300/CBP-associated factor (PCAF) [43, 44]. In addition, Evert et al. (2006) have shown that ATXN3 binds to specific sites in chromatin regions of the matrix metalloproteinase-2 (MMP-2) gene promoter and interacts with the transcriptional corepressors HDAC3 and NCoR, resulting in deacetylation

of histones and reduced binding of the transcription factor GATA-2 to target regions of the MMP-2 promoter [44].

Some other functions have been suggested for ataxin-3 as its importance for a correct cytoskeletal organization and muscle differentiation through the regulation of the integrin signaling transduction pathway [26]. More recently, Zhou et al. (2013) showed ataxin-3 to act as a novel anti-oxidative regulator in cooperation with mitochondrial Bcj-2 family proteins under oxidative stress conditions [45].

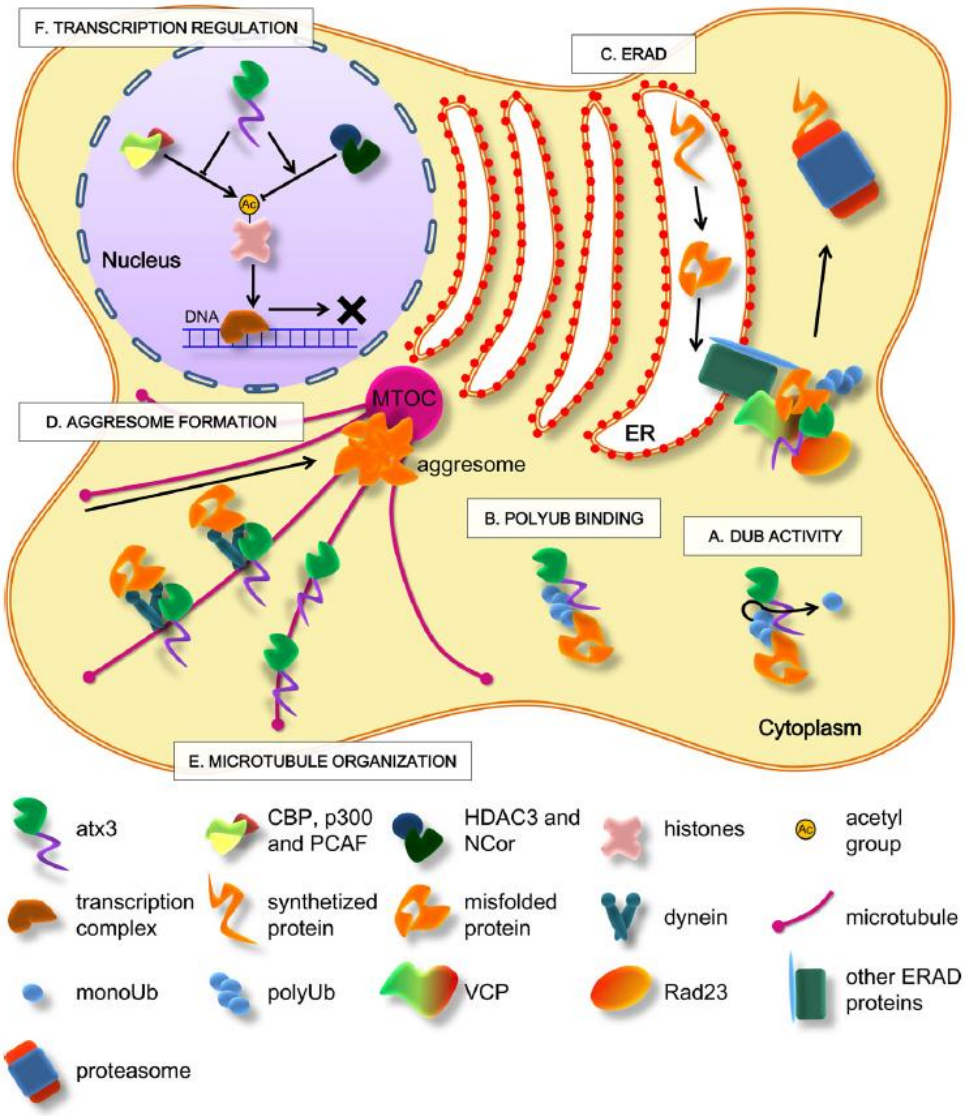


Figure 5: ATXN3 activity, interactions and proposed biological roles. ATXN3 displays DUB activity (A) and has been shown to interact with polyUb chains (B), two features that suggest its involvement in protein homeostasis systems using Ub signals, such as the UPP. ATXN3 was also shown to interact with Rad23 and VCP/p27 and to participate in a protein homeostasis mechanism in which these proteins are involved – the ERAD (C). ATXN3 has been further linked to aggresome formation (D), through interaction with cytoskeleton proteins (tubulin and dynein), and to the organization of the cytoskeleton itself (E). Interactions with regulators of histone acetylation support many reports describing ATXN3 as a participant in transcription regulation (F) [37].

3.2 ATXN3L

Studies focusing on the modulation of parental genes by the respective expressed paralogues may be important to gain insight into the mechanisms by which *ATXN3* retrocopies (especially *ATXN3L1*) may be functional relevant.

The gene responsible for spinocerebellar ataxia type 1 (SCA1), *ATXN1*, has an evolutionary conserved paralogue called *ATXN1-like*. Recent studies in fly and mouse models have shown that the overexpression of ATXN1L partially suppresses the neuropathology caused by the polyglutamine-expanded ATXN1, by inducing sequestration of polyglutamine-expanded ATXN1 into nuclear inclusions [46, 47]. They also propose that a combination of toxic gain-of-function and mild loss-of-function mechanisms contribute to SCA1 pathogenesis, with the partial loss-of-function of ATXN1 being sufficient to cause some transcriptional changes that are pathogenic in the cerebellum. Also, it has been previously reported that reduction of normal functions of genes involved in other polyglutamine diseases results in enhanced pathology, providing evidence for concomitant gain- and loss-of function mechanisms in polyglutamine disorders [48-52]. This raises the hypothesis that some conserved paralogue genes may compensate a partial loss of function of their parental genes.

Regarding our gene of interest *ATXN3L1*, it has been shown in an *in vitro* study that its JD presents higher DUB activity than the parental ATXN3 JD [31]. This led the authors to suggest that evolution has selectively maintained the *ataxin-3* sequence in a less active state (and possibly maintained *ATXN3L1* in a more active state). Although the crystal structure of the ATXN3L1 JD is highly similar to solution structures of the ATXN3 JD, the ubiquitin substrate appears to bind at different positions in the two proteins, suggesting a possible explanation for the difference between their efficiency. Nevertheless, three single-site mutations into the ATXN3 JD are sufficient to increase its DUB activity near ATXN3L1 levels [31].

3.3 Josephins

Josephin domain containing proteins 1 and 2 (JOSD1 and JOSD2) are thus named once they contain only the catalytic josephin domain. Despite their structural similarity, JOSD1 and JOSD2 differ with respect to base-line catalytic activity, modification by ubiquitin, and subcellular localization. JOSD1 is regulated by ubiquitination, suggesting that ubiquitination-dependent regulation may be a more common regulatory mechanism for DUBs than previously anticipated. At the subcellular level, JOSD1 appears most concentrated at the plasma membrane, especially

when ubiquitinated, where it influences membrane dynamics, cell motility and endocytosis. In alternative, JOSD2 is diffusely localized in the cytosol [34].

4. Polyglutamine diseases

The discovery of human genetic disorders caused by repeat expansions has brought a new interest regarding the evolutionary processes that underlie these repeat tracts [53]. Polyglutamine (polyQ) diseases constitute a group of hereditary neurodegenerative disorders caused by the expansion of a $(CAG)_n$ above a certain threshold encoding a protein with several consecutive glutamines [37]. There are nine dominantly inherited neurodegenerative disorders caused by expanded polyQ tracts: Huntington's disease (HD), spinal and bulbar muscular atrophy (SBMA), dentatorubropallidoluysian atrophy (DRPLA), and six spinocerebellar ataxias (SCA1–3, 6, 7 and 17) [54-60]. To date, around 30 types of SCAs are known and 16 genes have been identified. Ataxia, borrowed from a Greek word meaning "loss of order," is used clinically to describe aberrant regulation of limb movements with poor coordination between limbs. Spinocerebellar ataxia is caused by anomalous function of the spinocerebellum, the part of the cerebellar cortex that receives somatosensory input from the spinal cord [61].

All these diseases are characterized by a selective neuronal loss [37]. Pathogenic polyQ proteins can poison the axonal transport system and induce nuclear accumulation of inclusions leading to apoptosis [62]. Chaperones may act in order to suppress neurodegeneration by channeling toxic protein into nontoxic aggregates [63] but not always in the proper time that allows the normal development to proceed [62]. Several genetic studies have revealed that loss of the involved proteins in humans and mice does not cause neurodegeneration, leading to the conclusion that the polyglutamine expanded protein causes disease by a dominant gain-of-function mechanism whereby it confers toxic properties to the host proteins [64-70]. However, the severe gain-of-function effects might mask any subtle loss-of-function component, thus confounding the interpretation of the results. In fact, loss of normal endogenous function of mutant proteins may also play a role in dominant neurodegenerative diseases caused by gain-of-function mutations [71].

Several polyglutamine disease proteins repress transcription; therefore, characterization of the mechanisms of transcriptional repression is critical for understanding cellular pathology as well as for identifying potential targets for therapeutic intervention [72].

5. Machado-Joseph disease

Machado-Joseph disease (MJD), also known as spinocerebellar ataxia type 3 (SCA3), is one of the late-onset autosomal dominant spinocerebellar ataxias [73]. The gene associated with this disease is *ATXN3*, which was mapped to the long arm of chromosome 14 (14q24.3–q32) [74]. The causative mutation is an expansion of a CAG repeat located in the coding region of *ATXN3*. Healthy individuals have triplets ranging from 12 to 44 units whereas patients present usually one allele with 61 to 87 repeats [25]. Intermediate size alleles are rare and their pathogenicity still unclear [26].

MJD was first described in Azorean-Portuguese families living in the United States but, meanwhile, many patients from different ethnic backgrounds have been reported and MJD appears currently to be the most common SCA worldwide. The highest frequencies have been reported in Brazil (69-92%) [75], Portugal (58-74%) [76], Singapore (53%) [77], China (48-49%) [78], the Netherlands (44%) [79], Germany (42%) [80], and Japan (28-63%) [81], followed by Canada (24%) [82], United States (21%) [83], Mexico (12%) [84], Australia (12%) [85], and India (5-14%) [86], whereas it is considered as relatively rare in South Africa (4%) [87] and Italy (1%) [88]. MJD is highly pleomorphic in its clinical presentation, which led to the definition of the following 3 subphenotypes: type 1, the most severe form, is characterized by an earlier onset (mean age: 24.3 years), marked spasticity, dystonic features, cerebellar ataxia and ophthalmoplegia; type 2, the most common, is characterized by onset in midadulthood (mean age: 40.5 years), cerebellar ataxia, ophthalmoplegia, and pyramidal signs; type 3, corresponding to the mildest form, is characterized by a later onset (mean age: 46.8 years) and marked peripheral amyotrophies in addition to the main signs [89].

At the clinical level, there is a strong and consistent inverse correlation between the age at onset and the length of the expansion - the bigger the repeat the earlier symptoms appear in life [90]. Expansion size, however, is not the only variant that influences age at onset. The protein context, *cis*-acting genetic factors (typically familial or ethnic) in linkage disequilibrium with the affected alleles, *trans*-acting genetic factors, the tissue context in which the cognate protein is expressed, and environmental factors seem also to influence SCAs age at onset [79]. In MJD, repeat expansions are more common than contractions [90], leading to anticipation, a clinical phenomenon characterized by an earlier onset and increased disease severity in successive generations [91]. Expanded alleles are very unstable upon transmission and increase in the size of the expanded alleles tends to be greater when the transmitting parent is the father [92-94].

Homozygosity is also a condition that clearly enhances the severity of the clinical phenotype impairment, which supports a partial loss-of-function in the pathogenesis of the disease [90].

In MJD patients, the *ATXN3* gene is widely expressed in the brain and other organs in both normal and pathogenic forms [90]. It has been suggested that the catalytically inactive form of expanded ATXN3 may accumulate more readily in neurons and therefore induce more neurotoxic [95]. The nuclear localization of ataxin-3 aggravates the formation of protein aggregates and the development of a phenotype, whereas nuclear export of the protein results in a reduced number of nuclear inclusions (NIIs) and a much milder phenotype. Therefore, keeping the expanded polyglutamine repeats in the cytoplasm might delay protein aggregation, giving easier and longer access of the toxic protein to the protein degradation machinery [96]. Live-cell imaging studies showed that expanded polyQ tracts slow the dynamics of intact ataxin-3, and that the export of expanded protein is less efficient than the export of normal ataxin-3, suggesting a defect on the nucleocytoplasmic shuttling activity of the expanded protein [41].

It is commonly believed that neuronal damage in MJD is due to a gain of function mechanism characterized by misfolding, aggregation and subsequent formation of amyloid fibers by the polyQ-containing proteins or their fragments [97]. ATXN3 UIM domain binds polyubiquitylated proteins with a strong preference for chains containing four or more ubiquitins, which is the chain length required for proteasome degradation, cleaving these polyubiquitin chains via its N-terminal Josephin domain [95, 98]. The expansion of the polyglutamine tract in ataxin-3 makes the protein highly susceptible to misfolding and subsequent aggregation, forming intraneuronal, ubiquitin-positive inclusions. In addition to the ubiquitylated mutant or misfolded protein, aggregates in disease brains recruit normal proteins, including chaperones, and proteasomes, creating larger aggregates rather than being degraded by the last ones [39, 98]. Over the course of decades this could result in inadequate degradation of many cellular regulators resulting in cellular dysfunction [98].

Expanded ATXN3 showed also altered DNA and chromatin binding, failed to accomplish functional repressor complexes on the promoter, and aberrantly activated MMP-2 expression via increased histone acetylation and GATA-2 binding. Normal and expanded ATXN3 have been observed to interact with endogenous HDAC3 and NCoR in cell and human brain extracts, but only normal ATXN3 was associated with increased deacetylase activities and deacetylation of histone H3 [44]. Expanded ATXN3 not only loses its repressing function but acquires also a new transcriptional activator function. The new transcriptional activator function is mediated via its active site, which probably disturbs the formation or maintenance of histonedeacetylating repressor complexes on target promoters. Both the loss of repressor and gain of an activator

function result in increased histone acetylation of specific target gene promoters, probably leading to altered gene expression in MJD [99, 100]. This repressor activity is dependent on its ubiquitin interaction motifs, suggesting a link between ATXN3 function in the ubiquitin/proteasome pathway and its role in transcriptional regulation [61].

Objectives

Given the ongoing evidence of retrogenes functionality, we considered of great importance to characterize *ATXN3* retrocopies, since in addition to the interest in the biological processes on their origin, they may play a role in the pathogenesis of MJD. Therefore, our main objectives were:

- to address the mechanisms that have led to the human-specific *ATXN3* $(CAG)_n$ expansion (not observed in other lineages or paralogues);
- to determine at which point in primate evolution have the retrotransposition events occurred;
- to compare the rates of evolution and selective constraints among *ATXN3* and *ATXN3L1* in the primate lineage;
- to assess the mRNA expression pattern of the retrocopy that conserves an intact ORF (*ATXN3L1*) in human tissues.

Materials and Methods

1. DNA samples

We amplified and sequenced DNA samples from 10 non-human primate species: *Pan troglodytes* (n=1), *Gorilla gorilla* (n=1), *Pygmaeus pongus* (n=1), *Hylobates* (n=1), *Macaca mulatta* (n=3), *Macaca fuscata* (n=15), *Macaca fascicularis* (n=1), *Cercopithecus aethiops* (n=1), *Aotus trivirgatus* (n=1), and *Saguinus oedipus* (n=1). Sequences obtained from these samples were analyzed together with others previously assessed by us [32] and with additional ones retrieved from public databases as summarized in tables 1 and 2.

Table 1: Non-human primate species sequenced in our lab.

Species	N
Pan troglodytes	3
Gorilla gorilla	4
Pongo abelii	2
Pygmaeus pongus	1
Hylobates	1
Papio anubis	3
Macaca mulatta	7
Macaca fuscata	15
Macaca fascicularis	4
Cercopithecus aethiops	1
Aotus trivirgatus	1
Callithrix jacchus	3
Saguinus oedipus	1

Table 2: Sequences retrieved from public databases.

Species	*ATXN3*		*ATXN3L1*		*ATXN3L2*	
	Database	Accession/Gene ID	Database	Accession	Database	Accession
H. sapiens	Ensembl	ENSG00000066427	Ensembl	ENSG00000123594	NCBI	100132280
P. troglodytes	Ensembl	ENSPTRG00000006648	Ensembl	ENSPTRG00000028316	UCSC	chr8: 109011629-109012743
P. paniscus	NCBI	100976906	NCBI	100987250	NA	NA
G. gorilla	Ensembl	ENSGGOG00000014166	NCBI	101124367	UCSC	chr8: 109925669-109926225
P. abelii	Ensembl	ENSPPYG00000006075	Ensembl	ENSPPYG00000020129	UCSC	chr8: 117579303-117579866
N. leucogenys	Ensembl	ENSNLEG00000017126	Ensembl	ENSNLEG00000018731	UCSC	GL397267: 24598565-24599123
P. anubis	UCSC	chr7: 3062	UCSC	chrX: 10617308-10618372	UCSC	chr8: 105521214-105522300
P. hamadryas	NA	NA	UCSC	scaffold3734:61188-61980	NA	NA
M. mulatta	Ensembl	ENSMMUG00000020751	Ensembl	ENSMMUG00000009370	UCSC	chr8: 112963801-112964361
S. boliviensis	NCBI	101043060	NCBI	101036136	UCSC	Genomic JH378207
C. jacchus	Ensembl	ENSCJAG00000018025	Ensembl	ENSCJAG00000023329	UCSC	chr19: 6218140-6218664
T. syrichta	Ensembl	ENSTSYG00000009902	Ensembl	ENSTSYG00000003386	NA	NA
O. garnettii	Ensembl	ENSOGAG00000008601	NA	NA	NA	NA

NA: not available

2. Whole genome DNA amplification

In order to have enough amount of DNA to complete our study, all species sequenced in this work, excluding *Macaca mulatta*, were submitted to a random amplification with GenomiPhi V2 amplification kit (GE Healthcare). Following the instructions provided by the manufacturer, 1 µL of DNA was mixed with 9 µL of sample buffer and denatured for 3 min at 95ºC; then, 1 µL of enzyme mix was added to 9 µL of the denatured product and incubated at 30ºC for 90 min; and, finally, to inactivate the enzyme, the amplified product was heated at 65ºC for 10 min.

3. Amplification of *ATXN3L1* and *ATXN3L2*

We started by amplifying the genes of interest, *ATXN3L1* and *ATXN3L2*, with primers and conditions previously optimized in our lab. Based on the alignment of gene orthologues, primers were designed to specifically amplify *ATXN3L1* and *ATXN3L2* in most primate species [32].
In cases when the available protocol did not work properly, we optimized new conditions. Both previous and newly optimized protocols were carried out in singleplex PCR reactions using 1 µL of genomic DNA, 5 µL of My Taq™ HS MIX (Bioline, London, United Kingdom), 1 µL of each primer (with a final concentration of 2.5 µM), and 2 µL of H_2O in a final volume of 10 µL. The reactions were done in 2720 Thermal Cycler (Applied Biosystems) or Thermal Cycler (BioRad). The summary of amplification parameters is described in figure 7 and table 4.

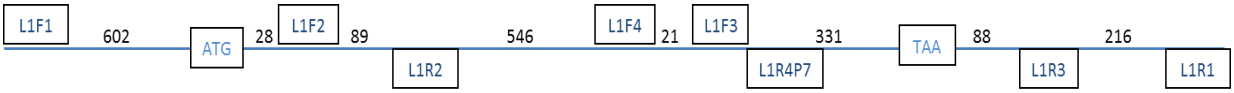
Table 3: Primers designed to specifically amplify *ATXN3L1* and *ATXN3L2* genes.

	ATXN3L1		*ATXN3L2*	
	Primer	Sequence	Primer	Sequence
Forward	L1F1	CTCTAACTAGGATACCAGCAAAG	L2F1	CATTAACCAAAGAAAGTGGGATAC
	L1F2	GTTTCCTGTGTGCTCAGCACTG	L2F1s	GTAGGAAAATAAACACAGTGAAAC
	L1F3	GAAACCAATAGAGAAGATGAA	L2F2	CCAGAGTATCAAAGGCTCAGGATC
	L1F4	AAGATGAGGAGGATTTTCAGAGG	L2F3	AGGCTCGCTTTGTGCTCAGCATTG
			L2F4P7b	GAAGAAATCTCTGGAGGGCAG
			L2P4A	GATAAGATAATTTATATGCGA
			L2P4G	GATAAGATAATTTATATGCGG
Reverse	L1R1	GGAAAAAGTTCTATGGCAAGAGC	L2R1	GGAATCCTATGCTGTAATCACAC
	L1R2	ACTGGTGACTCCTCCTTCTGCCA	L2R1s	TCACAGCTGCCTGAAAAGTGG
	L1R3	GTCTAAAAGCCTTATTTCCTCATCT	L2R2	GTGGTCAGCCTTTCACATGGATA
	L1R4P7	CCTKGCATGCTTAACTCAATAG	L2R3	TCTATTACCCAGAGTGGAGTGCAG
			L2R4	TGGTCTCGAACTCCTGGCCTCA
			L2R5	CTAAAACTCCTCCTTCTGCC

Primers L2F1s and L2R1s were specifically designed taking into account *Saguinus oedipus* sequence obtained in the lab. Similarly, L1R4P7 and L2F4P7b are specific primers to *Aotus trivirgatus* and L2P4A and L2P4G are allele-specific primers to assess the phase of *Gorilla gorilla ATXN3L2* sequences.

The following figures help to visualize the annealing site of the primers within our genes.

A.



B.

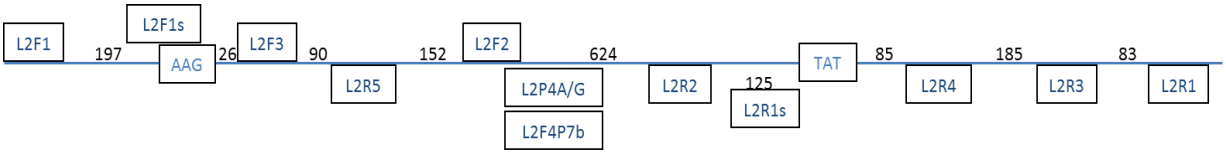


Figure 6: Schematic representation of A. *ATXN3L1*; and B. *ATXN3L2* genes, with distances (bp) between selected primers. Start and stop codons for *ATXN3L1* and the aligned homologous codons for *ATXN3L2* are also represented.

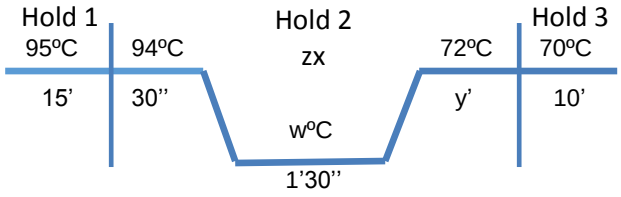


Figure 7: General PCR protocol with length of time ('- minutes; '' – seconds) and temperatures (°C) for the amplification *ATXN3L1* and *ATXN3L2* loci. x (annealing temperature) – 59 for *ATXN3L1*; and 57 for *ATXN3L2*; y (extension time) – 2 for *ATXN3L1*, and 1 for *ATXN3L2*; z (number of cycles) - 40 for *ATXN3L1*; and 35 for *ATXN3L2*.

Table 4: Summary of the PCR conditions and set of primers used to amplify our samples. w: annealing temperature; z: number of cycles.

	Species	Primers	wºC	zx
ATXN3L1	*Cercopithecus aethiops* *Macaca fascicularis* *Gorilla gorila* *Hylobates*	L1F1-L1R1	59	40
	Pan troglodytes	L1F1-L1R3	Touchdown 58-56	20-20
	Pygmaeus pongus	L1F1-L1R3	58	40
	Aotus trivirgatus	L1F1-L1R4P7 L1F2-L1R1	58 62	40
	Saguinus oedipus	L1F1-L1R1	54	40
ATXN3L2	*Cercopithecus aethiops* *Macaca fascicularis* *Gorilla gorilla* *Hylobates* *Pygmaeus pongus* *Saguinus oedipus*	L2F1-L2R1	57	35
	Pan troglodytes	L2F1-L2R1	Touchdown 58-56	10-30
	Aotus trivirgatus	L2F1s-L2R1s L2F3-L2R2 L2F4-L2R1	60 59 62	35

4. Amplification of the parental *ATXN3*

For species in which *ATXN3* sequences retrieved from databases were incomplete or poorly annotated, we used primers previously designed by MI Martins (2012) [32] to specifically amplify the parental gene.

Table 5: Summary of the PCR conditions and set of primers used to amplify our samples. w: annealing temperature (ºC); y: number of cycles; z: extension time (minutes)

Sample	w	y	z	Primers
Gorilla gorilla	60	35	1'	F2'-R3
	62	40	2'	F3-R3'
Pongo abelii	58	40	2'	F2'-R3
Macaca mulatta	61	35	2'	F1-R1
Papio anubis	61	35	1'	F18-R20

Table 6: Primers designed to specifically amplify the *ATXN3* gene.

Primer	Sequence
F1	CGTGTCCCCGGCGTTCACTC
R1	AGTCGCCCCCATGCCGATCT
F2'	GTGGTAAGCTGAGATTGCTCC
nF2*	ATTATTTAGGGGAGCCGGGCG
F3	GCAAGAAGGCTCACTTTGTGCTC
R3	CCAGCTGATGTGCAATTGAGG
R3'	TTTCCCCACCCCCACCCTTG
F18	CCACTCCTGGCCATGATAGGT
R20	GGTAAGGCCTGCTCACCATTC

*nF2 was used in sequencing

5. Detection of amplified products

To confirm specificity of DNA amplification, PCR products were submitted to electrophoresis in a polyacrylamide gel prepared with 0.375 M Tris/HCl gel buffer (pH=8.8). The gel was obtained by mixing 3 mL T9C5 (9% acrylamide, 5% N-N-metileno-bis-acrylamide), 170 µL of 2.5 % ammonium persulfate (APS) and 7 µL of TEMED as catalysis agent. Glass supports, with one side covered by a hydrophilic gel-bond film, were used to obtain a 3 mm thick gel. After loading the samples in the gel, two paper strips soaked in buffer were used at both anode and cathode to allow the horizontal run. To monitor the process of electrophoresis, we added bromophenol blue dye to the anode strip. The electrophoretic system was submitted to refrigeration at 4°C, and to a voltage between 220 and 250 V. The gel was then submitted to silver staining. The coloration method comprised (1) a fixation step of the DNA, with 10% ethanol for 10 minutes followed by 1% nitric acid for 5 minutes; (2) two washes with deionized water, of about 20 seconds each; (3) a coloration step with 0.2% silver nitrate solution, for 20 minutes; (4) two washes again with deionized water, for 20 seconds each; and finally, (5) a revelation step of the DNA fragments with a solution of 0.28 M sodium carbonate and 0.02% formaldehyde. The revelation reaction was stopped with 10% acetic acid for approximately 30 seconds. The resulting gels were washed with water and dried at room temperature.

6. Sequencing

After confirming specific amplification, an initial purification with ExoFastAP (*E. coli* exonuclease I and thermosensitive Alkaline Phosphatase, Thermo Scientific, Portugal) was performed to eliminate the remaining dNTPs and primers that have not been consumed in the amplification reaction. In this step, 1.8 µL of PCR product and 0.9 µL of ExoFastAP were used. The thermocycler program performed was 15 minutes at 37°C, followed by 15 minutes at 80°C to inactivate the enzyme. The sequencing reaction was performed in a total volume of 5 µL, including 0.5 µL of primer (2.5 µM), 0.5 µL sequencing buffer (2.5x), 1 µL of BigDye® Terminator v3.1 Cycle Sequencing Kit (normal deoxynucleotides, dye dideoxynucleotides, buffer and AmpliTaq DNA Polymerase, Applied Biosystems), 2.5 µL of purified product and 0.5 µL of water to complete the final volume. PCR sequencing process was performed for 35 cycles under the following conditions: 96ºC for 15 sec, 50ºC for 5 sec, and 60ºC for 2 min; a previous denaturation at 96ºC for 2 min and final extension at 60ºC for 10 min completed the reaction. A final purification step was performed in sequenced samples using Sephadex™ G-50 (GE Healthcare), a cross-linked dextran-gel used to separate the low (unincorporated nucleotides

and primers) from high molecular weight molecules of DNA (sequencing products). The products were purified through a Sephadex® column during 4 min at 4400 rpm, after an earlier centrifugation of 750 µL Sephadex® to form the columns. Finally, it was added 8 µL of HIDI™ formamide (Applied Biosystems) to the final product to increase the stability of single-stranded DNA for the capillary electrophoresis run in an ABI PRISM 3130x/Genetic Analyzer (Applied Biosystems). Sequences were analyzed with Sequencing Analysis v5.2 software (Applied Biosystems).

7. Sex determination

To assess the number of chromosomes analyzed of *ATXN3L1*, it was essential to know the gender of our individuals, since this gene is located in the X chromosome. Thus, we assessed other polymorphic loci located in the X chromossome, for which amplification had been previously optimized in our lab. Therefore, we genotyped seven polymorphic markers from the *androgen receptor* gene (Xq11-12): the $(CAG)_n$ region and 6 flanking STRs (short tandem repeats) [101]. Detection of PCR products has been done by capillary electrophoresis, through the use of a fluorescent forward primer for the amplification of each marker. The reaction was performed in a total volume of 10 µL containing 2x QIAGEN multiplex PCR Master Mix 1X, 0.5 µL of each primer at 2.5 µM, 0.5x Q-solution, 1 µL of DNA, and 2 µL of H_2O to complete the final volume. The $(CAG)_n$ amplification was perform using 6.25 µL of My Taq™ HS MIX, 1.25 µL of each primer, 1 µL of DNA, and 0.75 µL of H_2O to complete the final volume of 10.5 µL. Amplification conditions and primates studied are organized in table 7. To perform capillary electrophoresis, 1 µL of the amplified product was mixed with 11.5 µL of HIDI™ formamide and 0.5 µL of the internal marker GS-500 Liz (Applied Biosystems). The samples were run in an ABI-PRISM 3130 Genetic Analyzer system (Applied Biosystems) and the respective results were analyzed using GeneMapper Analysis Software v.4.0. We were not able to optimize the amplification of all samples for all markers which reduced the strength of our results. Thus, we tried to amplify the *amelogenin* gene located in the X chromosome (*AMELX* - Xp22.3-22.1) and its paralogue, the *amelogenin-like* gene (*AMELY* - Yq11), by adapting the protocol by Morrill et al. (2008) [102]; however, non-specificity did not allow us to amplify samples from our analyzed species. Then, we cut the gel fragment corresponding to our amplicon length, sequenced the re-amplified products of the *amelogenin* gene and aligned the obtained sequences with *AMELX* and *AMELY* sequences of several primates available in public databases.

A.

B.

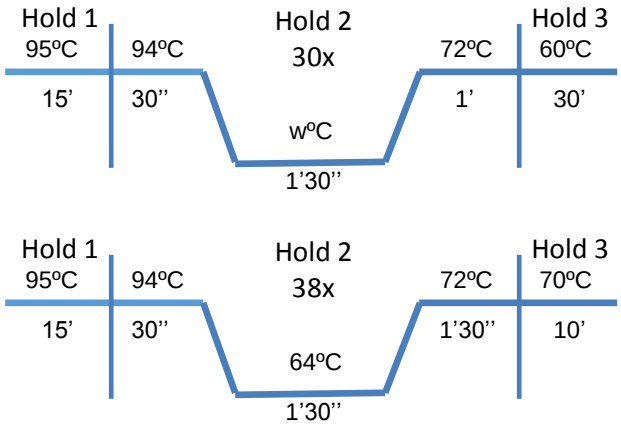


Figure 8: Amplification conditions of the $(CAG)_n$ region in the *androgen receptor* gene with A. My Taq™ HS MIX; and B. QIAGEN Multiplex PCR Kit.

Table 7: Annealing temperatures (w) used to amplify polymorphic markers located in the X chromosome for gender determination of some analyzed samples.

	STR1	STR2	STR3	STR4	DXS1111	DXS1194
Cercopithecus aetiops	62ºC	57ºC	62ºC	ND	ND	60ºC
Hylobates	59ºC	57ºC	ND	ND	60ºC	60ºC
Pygmaeus pongus	62ºC	ND	ND	ND	ND	60ºC
Saguinus oedipus	ND	ND	58ºC	ND	ND	58ºC
Macaca fuscata 2	59ºC	57ºC	62ºC	ND	ND	60ºC

(ND: Not determined)

8. Multiple sequence alignment and neighbor-joining phylogenetic trees

To study the origin of the retrotransposition events, we aligned all sequences available from the three genes under study (*ATXN3*, *ATXN3L1* and *ATXN3L2*), either obtained at the lab or retrieved from public databases; we constructed a phylogenetic tree using the neighbor-joining method and the Tamura-Nei genetic distance model. The parental gene sequence of *Canis familiaris* was used as an outgroup. The $(CAG)_n$ region was removed since it represents a source of much variation as well as some incomplete sequences. Both alignments and phylogenetic trees were performed using Geneious Pro 5.5.8 software.

9. Phylogenetic analysis of selection (ω ratio calculations)

Three different phylogenies were built using MEGA v5.1 software – one with *ATXN3L1* sequences alone, the second with *ATXN3* sequences and a third with all sequences used in the previous phylogenies. *ATXN3L2* displays a disrupted ORF, with maximum length fragments of 200 bp predicted as alternative ORFs; for this reason this gene was not considered for the selection analysis. We discarded incomplete sequences and highly related species, leaving only those with *ATXN3* and *ATXN3L1* complete sequences (*H. sapiens*, *P. troglodytes*, *G. gorilla*, *P. anubis*, *P. abelii*, *M*. *mulatta*, *N. leucogenys*, and *C. jacchus*). To carry out a comprehensive analysis of our genes, all the sequences were verified for frameshift mutations and positions that did not align properly with any triplet were removed.

Maximum likelihood estimates of d_N/d_S (ω) were carried out using the codeml program from the package Phylogenetic Analysis by Maximum Likelihood - PAML version 4.2. To test for variable selective pressures among branches, we performed the branch model using (1) the null or one-ratio model, which assumes the same ω parameter for all branches; (2) the two-ratio model, which assumes one ω for the branch that predates the duplication event (i. e. *ATXN3*) and a second ω for the branch that follows the duplication event (i. e. *ATXN3L1*); and (3) the free-ratio model, which assumes an independent ω for all branches in the phylogeny, involving as many ω parameters as the number of branches [103, 104]. In addition, we performed site models, which allow the ω value to vary among sites of the protein and compared models of neutrality with positive selection (M1 *vs* M2 and M7 *vs* M8) [105, 106]. Finally, to evaluate lineage-specific changes at amino acid sites, we performed the branch-site model. This model assumes that the branches on the phylogeny are divided *a priori* into foreground (*ATXN3L1*) and background (*ATXN3)* and allows ω to vary both among sites in the protein and across branches [107]. The significance of each nested model was obtained from twice the variation of likelihoods (2Δl) using a χ^2 statistic.

10. Transcriptional pattern of *ATXN3L1* in human tissues

10.1 RT-PCR

To evaluate if *ATXN3L1* was transcribed, we first performed a transcriptase reverse PCR using QuantiTect® Reverse Transcription kit (QIAGEN) and mRNA samples of 20 different human tissues namely: ovary, bladder, trachea, esophagus, thymus, thyroid, colon, kidney, skeletal muscle, testis, small intestine, heart, spleen, prostate, liver, brain, placenta, cervix, lung, and adipose tissue (Ambion® FirstChoice® Human Total RNA Survey Panel).

We started with the genomic DNA elimination by mixing 1 µg of our RNA with 2 µL of gDNA wipeout buffer 7x and 11 µL of RNase-free water, obtaining a total volume of 14 µL that was incubated for 2 min at 42ºC. To this template RNA we added a reverse-transcription master mix (1 µL of Quantiscript Reverse Transcriptase, which contained RNase inhibitor also; 4 µL of Quantiscript RT Buffer 5x, containing Mg^{2+} and dNTPs, 1 µL of RT Primer Mix) and incubated at 42ºC for 30 min. Finally, to inactivate the Quantiscript Reverse Transcriptase, samples were submitted to 95ºC for 3 min. All the reactions were performed on ice.

10.2 Amplification

Primers were previously design in regions flanking the exon junctions of the corresponding *ATXN3* exons determined by aligning the *ATXN3L1* sequence with the concatenated exons of parental *ATXN3*; this way, the exons' limits of the parental gene were annotated in the retrocopy [32].

Table 8: Primers designed to specifically amplify *ATXN3L1* in human cDNA.

Primer	Sequence	Amplicon size (bp)
E7-8F	CAGTGTCGAAGAGATGGATAC	140
E7-8R	CCTCATCTTGGTCTGATGTTCCAGACT	

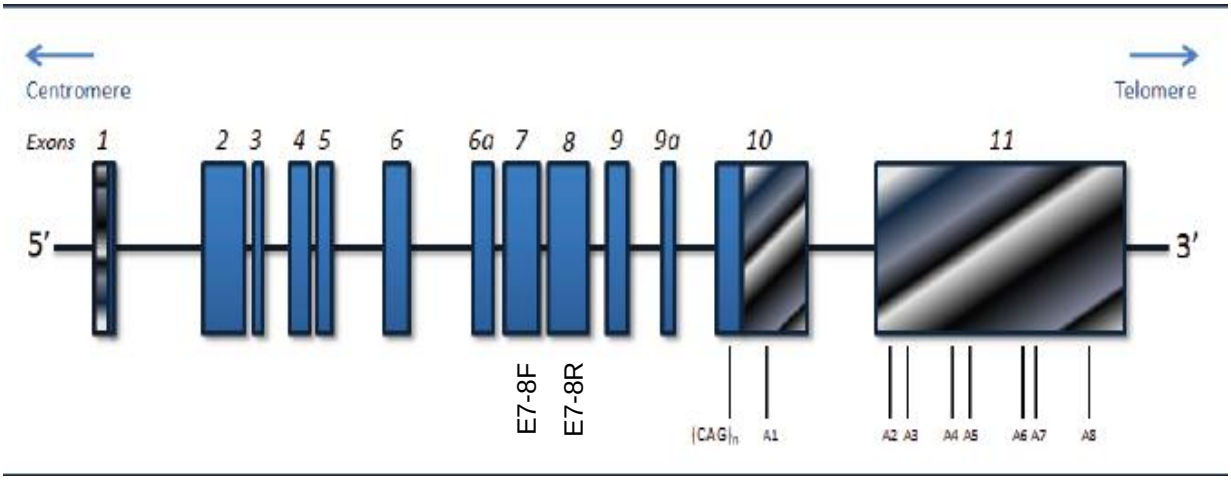


Figure 9: *ATXN3* gene structure and primers insertion location. Transcribed regions are represented in blue, and not-transcribed and UTR regions in grey. (Adapted from Bettencourt et al., 2011) [26].

We performed a duplex PCR amplification by using 5 µL of My Taq™ HS MIX, 0.5 µL of each primer (2.5 µM), 1 µL of our cDNA and 2 µL of H_2O to complete the final volume of 10 µL. The PCR conditions included an initial denatured process at 95ºC for 15 min, a final extension at 60ºC for 30 min and 35 intermediate cycles at 94ºC for 30 sec, 59ºC for 90 sec and 72ºC for 30 sec. The *GAPDH* (*glyceraldehyde-3-phosphate dehydrogenase*) gene was used as positive control since it is ubiquitously expressed and one of the most commonly used housekeeping genes in comparisons of gene expression data [108]. Finally, we confirmed specific amplifications by polyacrylamide gel electrophoresis followed by silver staining.

Results

1. Evolution of the $(CAG)_n$ tract in *ATXN3*, *ATXN3L1* and *ATXN3L2*

To gain insight into the processes by which the $(CAG)_n$ tracts have been accumulating variation throughout the evolution of each paralogue, we analyzed sequences from all primates obtained from databases and in our lab (Table 9).

While analyzing the compiled sequences, we observed the $(CAG)_n$ tract conserved in the parental gene along the primate lineage, except in the two orangutans (*Pongo abelii* and *Pygmaeus pongus*), which presented several interruptions of triplets coding lysine or phenylalanine. In humans (*Homo sapiens*), chimpanzees (*Pan troglodytes* and *Pan paniscus*), gorilla (*Gorilla gorilla*), baboon (*Papio anubis*), rhesus monkey (*Macaca mulatta*), and African green monkey (*Cercopithecus aetiops*), a $(CAG)_2$ CAA AAG CAG CAA tract is followed by AAG and/or $(CAG)_n$; in cynomolgous monkey (*Macaca fascicularis*) some point mutations originated the tract $(CAG)_2$ CAA $(CAG)_2$ AAG $(CAG)_7$. Both configurations give rise to an almost pure polyQ tract, with a single lysine (K) interruption (two in gorilla, baboon and African green monkey). The $(CAG)_n$ tract in bushbaby has a $(CAG)_n$ CAA $(CAG)_n$, followed by a AAG. The protein tract is almost pure in this species presenting only a terminal lysine. Gibbon (*Nomascus leucogenys*) and marmoset (*Callithrix jacchus*) present an almost pure polyQ stretch with one and two glutamic acids (E) interruptions, respectively. Squirrel monkey (*Saimiri boliviensis*) and tarsier (*Tarsius syrichta*) are the only species analyzed with a pure polyQ stretch.

For *ATXN3L1*, the repetitive region is also conserved among primates, with a $(CAG)_2$ GAA CAG AAG $(CAG)_2$ $(CAA)_2$ CAG tract observed in several species. The resulting putative protein sequence is interrupted, generally with the Q_2 E Q K Q_5 configuration. Gorilla and orangutan have a slightly different configuration which gives rise to an additional leucine (L) before the last group of four glutamines. The owl monkey (*Aotus trivirgatus*), squirrel monkey, marmoset, and cotton-top tamarin (*Saguinus oedipus*) present a highly interrupted tract: GAG CAA GAA CAA $(CAG)_2$ AGG AAA CAA AAG, giving rise to the interrupted polyQ stretch E Q E Q_3 R K Q K. Tarsier is the only species presenting a small pure polyQ tract for *ATXN3L1* of just four glutamines. The chimpanzee sequence obtained from Ensembl present a different $(CAG)_n$ tract from the other 3 individuals sequenced by us, with a CAC instead of a CAA in the 3'-end of the repeat.

As for *ATXN3L2*, translated products have not been considered since it presents a disrupted ORF. Thus, we observed a pure $(CAG)_n$ tract in gorilla, orangutan and gibbon, whereas in humans and chimps it is highly interrupted by CGG triplets. For the first time, a hexanucleotide repeat is observed in *ataxin-3* paralogues, instead of a trinucleotide pattern. By comparing both

retrocopies, *ATXN3L1* acquired more interruptions in its CAG repeat region than *ATXN3L2*, which turned it more stable. Accordingly, the almost pure *ATXN3L2* tract in species which diverged earlier in primate evolution, and the polymorphic $(CGG\ CAG)_n$ in humans and chimps show the higher instability associated to this *locus* in comparison to both parental *ATXN3* and *ATXN3L1* retrogene.

Table 9: *ATXN3*, *ATXN3L1* and *ATXN3L2* $(CAG)_n$ tracts obtained both from public databases and our sequencing of several primates. N represents the number of chromosomes analyzed.

	ATXN3		*ATXN3L1*		*ATXN3L2**	
	N	Sequence	N	Sequence	N	Sequence
Homo sapiens	98	(CAG)2 CAA AAG CAG CAA (CAG)n Q3 K Qn	36	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	58	(CAG)3 (CGG CAG)4-11
Pan troglodytes	1	(CAG)2 CAA AAG CAG CAA (CAG)12 Q3 K Q14	4	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	6	CAG (CGGCAG)3-9 GGG CAG
			1	(CAG)2 GAA CAG AAG (CAG)2 CAA CAC CAG Q2 E Q K Q3 H Q	1	CAG (CGGCAG)5 GGG CAG (CGGCAG)3 GGG CAG
Pan paniscus	1	(CAG)2 CAA AAG CAG CAA (CAG)11 Q3 K Q13	1	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5		
Gorilla gorilla	2	(CAG)2 CAA AAG CAA AAG (CAG)4 Q3 K Q K Q4	3	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	9	(CAG)6-17
			2	(CAG)2 GAA CAG AAG CTG CAG (CAA)2 CAG Q_2 E Q K L Q4		
Pongo abelii	1	(CAG)2 CAA AAG CAG CAA (CAG)2 CCG CAA AAG CAG CAA (CAG)2 CCG CAA (CAG)9 Q3 K Q4 P Q K Q4 P Q10	1	(CAG)2 GAA CAG AAG CTG CAG (CAA)2 CAG Q2 E Q K L Q4	4 1	(CAG)6 (CAG)7
Pygmaeus pongus	2	(CAG)2 CAA AAG CAG CAA (CAG)2 CCG CAA (CAG)2 CAA (CAG)2 CCG CAA (CAG)7 Q3 K Q4 P Q6 P Q8	1	(CAG)2 GAA CAG AAG CTG CAG (CAA)2 CAG Q2 E Q K L Q4	2	(CAG)6
Nomascus leucogenys	1	(CAG)3 CAA (CAG)5 GAG (CAG)3 Q9 E Q3	1	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	1	(CAG)8
Hylobates			1	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	2	(CAG)3 CAA (CAG)2
Papio anubis	2	(CAG)2 CAA AAG CAG CAA (CAG)2 AAG (CAG)7 Q3 K Q4 K Q7	4	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	3 3	(CAG)7 CAC CAG (CAG)8 CAC CAG
Papio hamadryas			1	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5		
Macaca mulatta	1	(CAG)2 CAA (CAG)2 AAG (CAG)7 Q5 K Q7	11	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	13 2	(CAG)6 CAC CAG (CAG)7 CAC CAG
Macaca fuscata			30	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	30	(CAG)6 CAC CAG
Macaca fascicularis	2	(CAG)2 CAA (CAG)2 AAG (CAG)7 Q5 K Q7	6	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	5 3	(CAG)6 CAC CAG (CAG)8 CAC CAG
Cercopithecus aethiops	2	(CAG)2 CAA AAG CAG CAA (CAG)13 AAG (CAG)6 Q3 K Q15 K Q6	1	(CAG)2 GAA CAG AAG (CAG)2 (CAA)2 CAG Q2 E Q K Q5	2	(CAG)6 CAC CAG
Aotus trivirgatus			1	GAG CAA GAA CAA (CAG)2 AGG AAA CAA AAG E Q E Q3 R K Q K	2	GAG (CAG)3 CAA
Saimiri boliviensis	1	CAG CAA (CAG)5 (CAA)5 CAG CAA (CAG)12 Q26	1	GAG CAA GAA CAA (CAG)2 AGG AAA CAA AAG E Q E Q3 R K Q K	1	GAG (CAG)4 CAA
Callithrix jacchus	1	CAG CAA (CAG)5 CAA (CAG)3 GAG (CAG)3 GAG (CAG)3 Q11 E Q3 E Q3	6	GAG CAA GAA CAA (CAG)2 AGG AAA CAA AAG E Q E Q3 R K Q K		
Saguinus oedipus			1	GAG CAA GAA CAA (CAG)2 AGG AAA CAA AAG E Q E Q3 R K Q K	2	GAG (CAG)3 CAA
Tarsius syrichta	1	(CAA)2 (CAG)2 Q4	1	CAG CAA (CAG)2 Q4		
Otolemur garnettii	1	(CAG)5 CAA (CAG)5 AAG Q11 K				

**ATXN3L2* presents a disrupted ORF

2. Sex determination

We determined the gender of primates for which this information was lacking. First, we genotyped the *androgen receptor* gene based on the theoretical basis of observing homozygozity in males and heterozygozity in females, since *AR* is located in the X chromosome. The successful genotyping of four *AR* polymorphic markers by capillary electrophoresis was not enough to determine the sex of all the primates (Table 10).

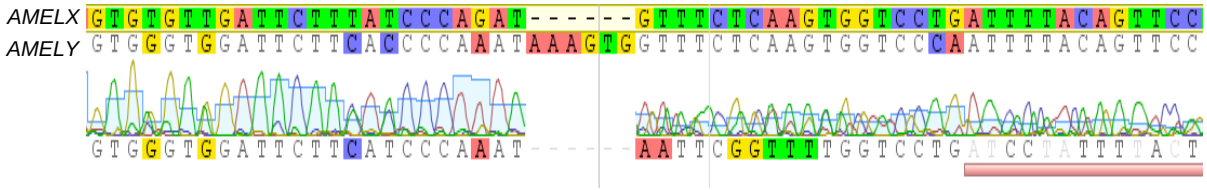
Then, we tried to differentiate the allelic states of the X/Y polymorphism in the *amelogenin* gene by electrophoresis; the presence of one or two fragments 6 bp distance of each other would identify a female or a male, respectively. All 5 samples analyzed by us appeared to present a single length fragment which led us to suspect that our electrophoresis conditions did not allow the correct visualization of the two allelic states of this Indel. Thus, we proceeded with the sequencing of the *AMELX*/*Y* amplified products and aligned with reference sequences (Figure 10). The identification of a 6 bp insertion in *AMELY*, together with additional *AMELX/Y* specific point variants across the sequence, allowed us to determine the gender of our primate samples as shown in the appendix section (Table A1).

Table 10: Capillary electrophoresis results for four polymorphic markers within or flanking the *androgen receptor* gene.

	CAG	STR1	STR3	DXS1194
Cercopithecus aetiops	h	H	h	H
Hylobates	h	ND	ND	H
Pygmaeus pongus	h	h	ND	h
Saguinus oedipus	h	ND	ND	ND
Macaca fuscata 2	h	H	H	H

H – heterozygous sequence; h – homozygous sequence; ND – not determined.

A.



B.

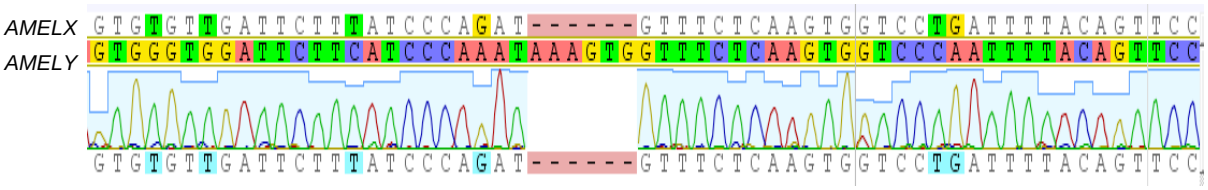


Figure 10: Alignment of *AMELX* and *AMELY* sequences with A. a male chimpanzee control sample; and B. a female owl monkey control sample.

3. Intraspecific *ATXN3L1* duplication events

After confirming the gender of the 15 *Macaca fuscata* samples, we noticed that all individuals showed to be heterozygous for at least one *ATXN3L1* position, including all 5 males. In total, we identified 7 different patterns with 4 SNPs as represented in table 11 and figure 11. Results have been confirmed in all specimens by sequencing in both forward and reverse directions. We therefore, hypothesized that a duplication of the *ATXN3L1* locus occurred in this species or in a common recent ancestor of Old World monkeys. Interestingly, we noticed that in some sequences there was a different proportion of each nucleotide (Figure 11- B). This could be due to preferential annealing of specific primers designed for *ATXN3L1* in the amplification or in the sequencing reaction. Yet, the patterns remained constant through the several PCR and sequencing replications.

Table 11: Heterozygozity pattern and polymorphic positions of *M. fuscata ATXN3L1*.

		Variant				
Specimen	Sex	c.561T>G	c.743A>G	c.923A>C	c.995C>T	Pattern
M. fuscata 1	Female	T	G	**A**/C	C/T	1
M. fuscata 2	Female	T	G	**A**/C	C/T	1
M. fuscata 3	Male	T	G	A	C/T	2
M. fuscata 4	Male	T	A/G	A/C	C/T	3
M. fuscata 5	Female	T	G	A	C/T	2
M. fuscata 6	Male	T	A/G	A/C	C/T	3
M. fuscata 7	Female	T	G	**A**/C	C/T	1
M. fuscata 8	Female	T	G	A	C/T	2
M. fuscata 9	Female	T/G	G	A	C/**T**	7
M. fuscata 10	Female	T	G	**A**/C	C/T	1
M. fuscata 11	Female	T	G	A	**C**/T	4
M. fuscata 12	Male	T	G	A	C/T	2
M. fuscata 13	Female	T	A/**G**	**A**/C	C/T	6
M. fuscata 14	Female	T	G	A	C/**T**	5
M. fuscata 15	Male	T	A/G	A/C	C/T	3

*in bolt are represented alleles with the highest intensity in sequencing.

A.

B.

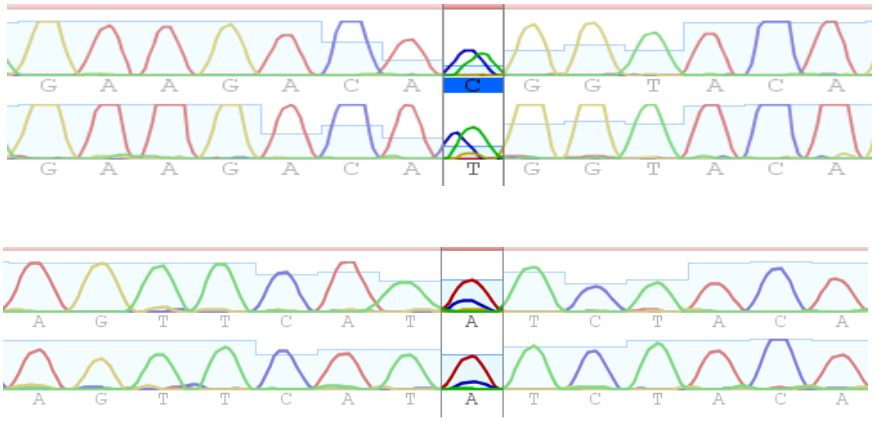


Figure 11: Example of two *ATXN3L1* variant positions in *M. fuscata*. The two sequences represented in each image (A and B) were sequenced at different time courses and by using alternative primers (forward and reverse) to assure the reproducibility of results; A. Heterozygous position c.995C>T in *M. fuscata* 8; B. Heterozygous position c.923A>C in *M. fuscata* 10 showing the two nucleotides unequally represented.

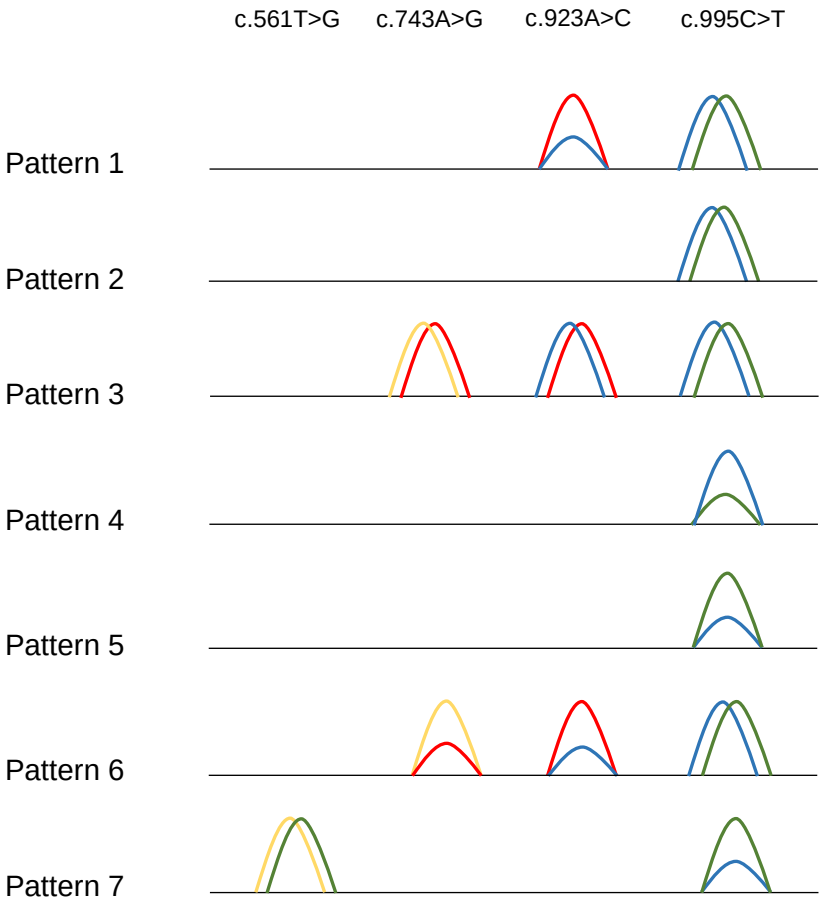


Figure 12: The different patterns of heterozygozity found in *M. fuscata* samples while sequencing *ATXN3L1*. The four nucleotides are distinguished by four different colours (red – adenine; blue – cytosine; green – timine; and yellow – guanine).

As for the other *Macaca* we found two heterozygous position in two *M. fascicularis* males sequenced by us (with c.995C>T corresponding to *M. fuscata* c.995C>T), although in this case a pattern could not be established due to the low number of analyzed chromosomes (6). On the other hand, our data did not allow us to confirm if this duplication is present or not in *M. mulatta* since no biallelic states have been found in the four males sequenced. For this species, a single heterozygous position has been observed in a female individual (c.174C>T).

We also found a duplication of the *ATXN3L1 locus* in *Papio anubis*. Using the Blat tool in UCSC Genome Browser, we found two sequences located in the X chromosome with 92% similarity to human *ATXN3L1*. By analyzing our sequencing results of three baboon samples, we observed a male with 4 positions with 2 allelic states, and noticed that these heterozygous positions corresponded to the bases varying between UCSC sequences ChrX: 10617308-10618372 and ChrX: 10636574-10637639 (Table 12). We aligned the two *P. anubis* sequences retrieved from UCSC with all other *ATXN3L1* orthologues and confirmed that both clustered with other *ATXN3L1* sequences and shared a common node, suggesting a recent duplication event of this *locus*.

Table 12: *ATXN3L1* polymorphic positions found in *Papio anubis*.

Origin	Sequences	Sex	Variant					
			c.280C>T	c.319A>G	c.987A>T	c.1006A>G	c.1007C>T	c.1011C>T
UCSC	*P. anubis* (ChrX: 10617308-10618372)	Unknown	T	G	T	A	T	C
	P. anubis (ChrX: 10636574-10637639)	Unknown	C	A	A	G	C	T
Our samples	Baboon 1	Female	C/T	A/G	A/T	A/G	C/T	C/T
	Baboon 2	Female	C/T	A/G	A/T	A/G	C/T	C/T
	Baboon 3	Male	ND	ND	A/T	A/G	C/T	C/T

ND: not determined

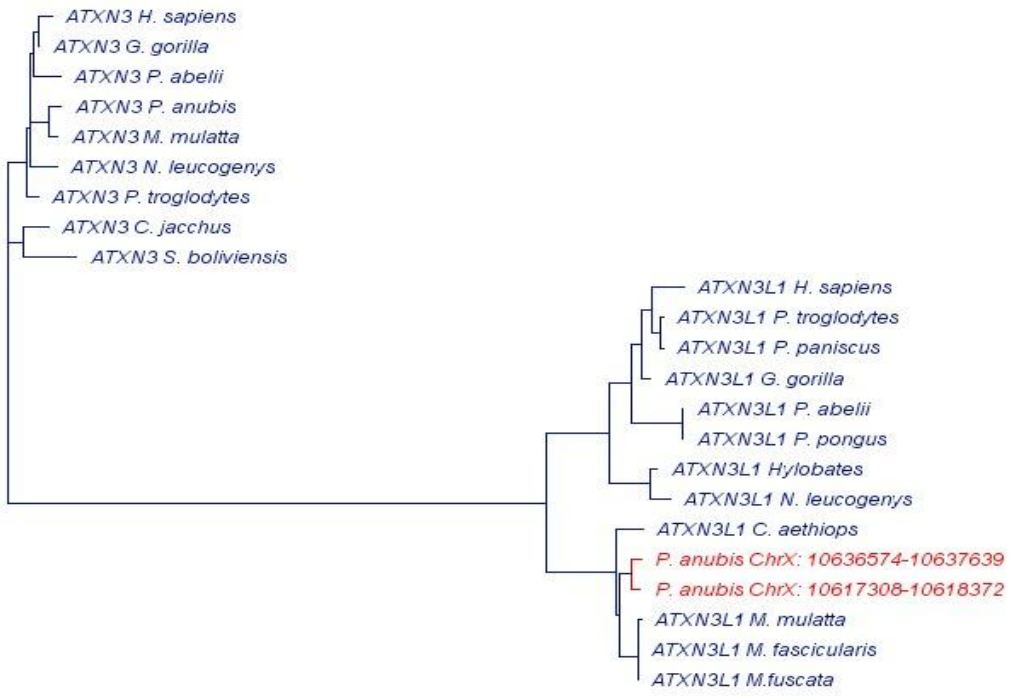


Figure 13: Neighbor-Joining phylogenetic tree for *ATXN3* and *ATXN3L1* showing the two sequences retrieved from UCSC to *Papio anubis*.

4. *ATXN3L2* in *Aotus trivirgatus*

Several sequences highly homologous to the human *ATXN3* CDS available in public databases have been identified by us as independent-origin retrocopies of *ATXN3* [32]. When using different sets of primers in the optimization of *ATXN3L2* amplification in *Aotus trivirgatus*, we found two different sequences with high similarity to human *ATXN3L2*. The multiple sequence alignment followed by neighbor-joining phylogenetic tree allowed us to identify *A. trivirgatus ATXN3L2* whereas the second sequence (here named *ATXN3L*?) showed a closer molecular distance with other recent retrocopies of *ATXN3*. Indeed, results showed *ATXN3L*? sharing a more common recent ancestor with parental genes than with *ATXN3* paralogues, as we can see in figure 14.

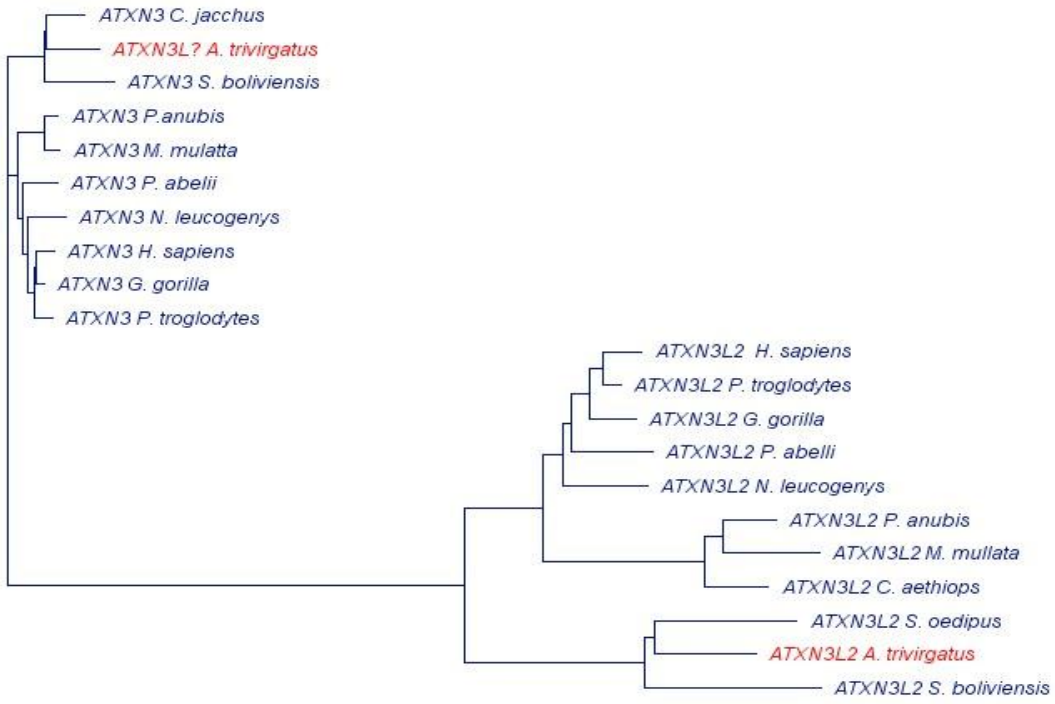


Figure 14: Neighbor-Joining phylogenetic tree showing *ATXN3* and *ATXN3L2* in several primates, and evidencing the origin of *Aotus trivirgatus ATXN3L2* and *ATXN3L?* sequences.

5. Origin of the retrotransposition events

To estimate the time for the origin of *ATXN3* paralogues, we constructed phylogenetic trees including all species in which the three genes have been identified. The alignment and respective tree has been done after removing the $(CAG)_n$ tract, since it usually represents a source of much variation generated by a still unknown process (most likely to be independent from the locus evolution). As shown in figure 15, *ATXN3L1* is present in the entire clade of Haplorrhines but absence in Strepsirrhines. The *ATXN3L2* is present only in the Anthropoid clade. Thus, *ATXN3L1* and *ATXN3L2* seem to have independent origins about at least 63 MYA (before the split between Tarsiers and Anthropoids) and 43 MYA (before Platyrrhine and Catarrhine split from their ancestor), respectively, based on the tree of species (Figure 16) and respective divergence times [109]. The most common *ATXN3* transcript, *ATXN3-001*, seems to be in the origin of *ATXN3L1* and *ATXN3L2* given the high similarity between these retrogenes and *ATXN3-001* over all the other transcripts [32].

ATXN3L2 is present in tamarin but we were not able to amplify this gene in marmosets (Figure 16). Taking into account that both are New World monkeys, problems in the annealing of our primers in the *Callithrix* species seem more likely than the absent of *ATXN3L2* in marmosets. Thus, further optimizations need to be done in order to amplify and sequence this gene in our *Callithrix jacchus* samples.

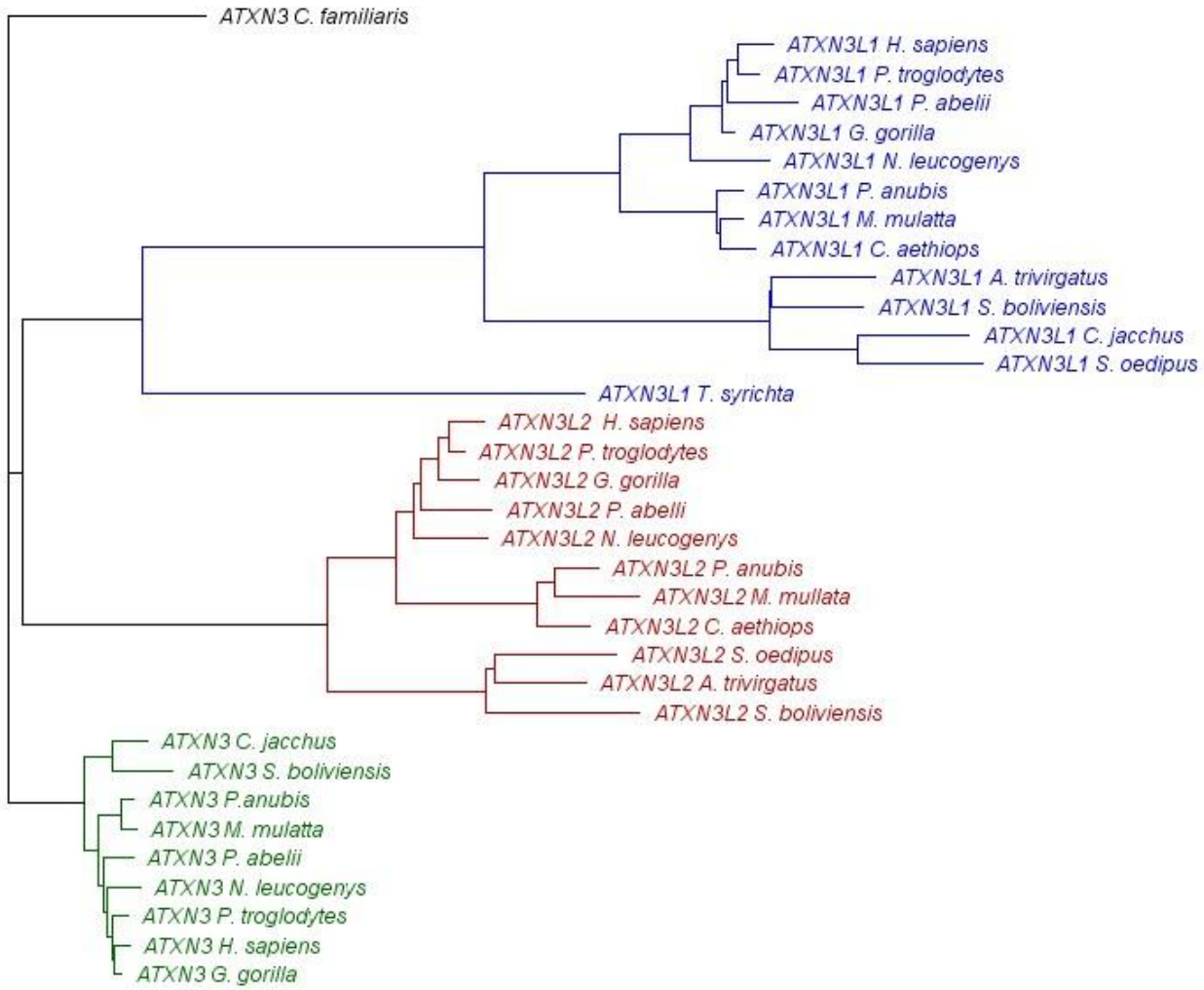


Figure 15: Neighbor-Joining phylogenetic tree for *ATXN3*, *ATXN3L1*, and *ATXN3L2*. *Canis familiaris* was used as an outgroup.

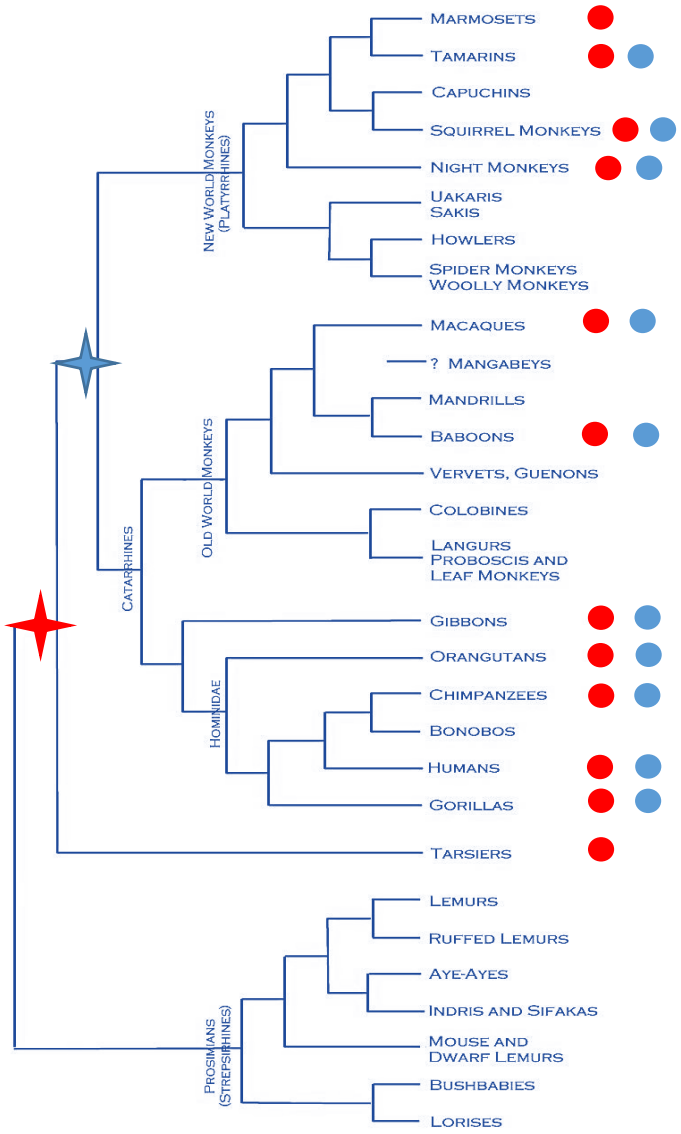


Figure 16: Phylogenetic tree of primates. The presence of *ATXN3L1* and *ATXN3L2* is marked in each species with a red and a blue dot, respectively. The blue and red stars represent the time at which *ATXN3L2* and *ATXN3L1* may have been originated. (Adapted from http://whozoo.org/mammals/Primates/primatephylogeny.htm).

6. Phylogenetic analysis of selection

Evolutionary pressures on proteins have been quantified by calculating the omega ratio (ω) for *ATXN3* and *ATXN3L1*, using the coding sequences of eight primate species: *H. sapiens*, *P. troglodytes*, *G. gorilla*, *P. abelii*, *N. leucogenys*, *P. anubis*, *M. mulatta*, and *C. jacchus*. We assured a close agreement between gene tree and species tree.

6.1 Branch models

To address the extent of the selective pressures exerted on *ATXN3* and *ATXN3L1*, we performed branch models and tested whether these genes experienced different selective pressures during primate evolution. First, we estimated a single ω for the entire phylogeny (one-ratio model), in which we assumed no differentiation in *ATXN3* and *ATXN3L1* selective constraints over evolution. The observed ω value below 1 (ω=0.30) pointed out to an overall conservation of *ATXN3* and *ATXN3L1* (Table 13). Then, to examine whether the two genes were subjected to different selective pressures, we applied a different model (two-ratio model) considering two branches within the phylogeny comprising either *ATXN3* or *ATXN3L1*. Both ω values were below 1 (ω_{ATXN3}=0.17; $\omega_{ATXN3L1}$=0.33) (Table 13) but differed significantly from the previous reported model (-2Δl=8.89; P<0.05), indicating that *ATXN3* (ω=0.17) has been under stronger constraints than *ATXN3L1* (ω=0.33).

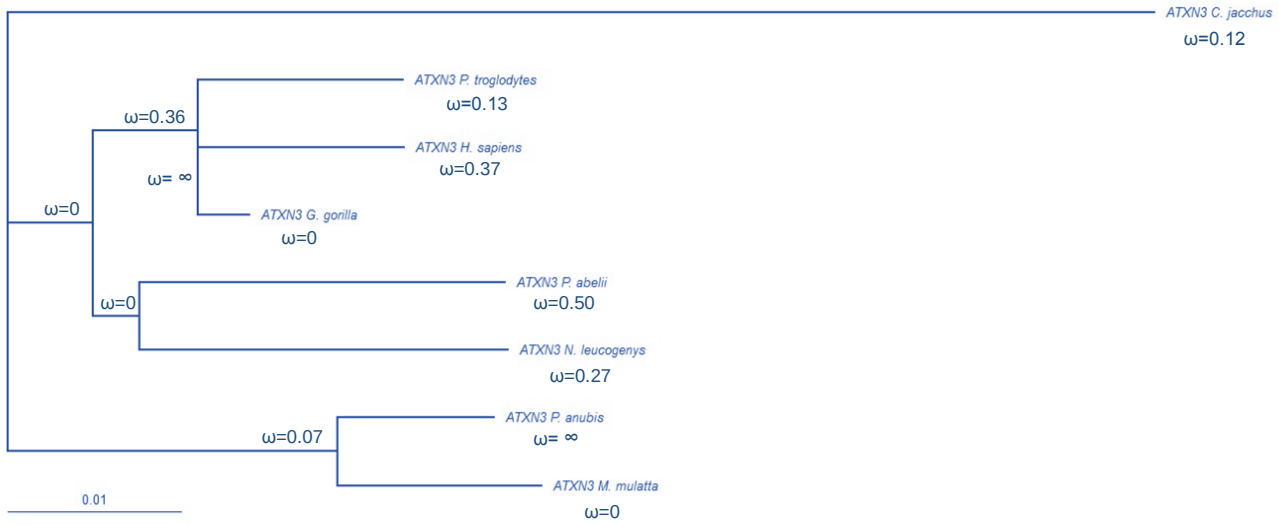
Table 13: Parameter estimates and likelihood scores under different branch models. Model comparisons of variable ω ratios among branches.

Model	Likelihood (ln)	Parameters
One-ratio (M0)	-3630.56	ω=0.30
	ATXN3 = -1764.34	ω=0.17
	ATXN3L1 = -2631.58	ω=0.33
Two-ratios	-3626.11	ω_{ATXN3}=0.15
		$\omega_{ATXN3L1}$=0.34
Free-ratio	*ATXN3* = -1755.55	
	ATXN3L1 = -2622.20	

Models compared	df	-2Δl	Critical χ^2 values (P=0.05)
One-ratio *vs* Two-ratios	1	**8.89***	3.84
One-ratio *vs* Free-ratio (*ATXN3*)	12	17.59	21.03
One-ratio *vs* Free-ratio(*ATXN3L1*)	12	18.75	21.03

*Extremely significant (P<0.01; χ^2 = 6.6)

A.



B.

Figure 17: Phylogeny of the 8 primate species analyzed. Branches are drawn in proportion to their lengths, defined as the expected numbers of nucleotide substitutions per codon and estimated using the one-ratio model, which assumes the same d_N/d_S ratio for all branches in the tree [110]. The ω values presented here were determined in free-ratio model for A. *ATXN3*; and B. *ATXN3L1*.

6.2 Site models

In order to test if different selective pressures had been acting along the protein sequence, we performed several site models. In these cases, variable ω ratios among sites were calculated for each gene and neutral and selection models compared (M1 *vs*. M2 and M7 *vs*. M8). We observed that in both cases, neutral models fit better the *ATXN3* and *ATXN3L1* data than positive selection models (Table 14). This shows that there may not be a large category of amino acid sites in our genes where non-synonymous substitutions occur at a higher rate than synonymous substitutions.

Table 14: Parameter estimates and likelihood scores under different site models. Model comparisons of variable ω ratios among sites.

	Model	Likelihood (ln)	Parameters
ATXN3	M1	-1764.22	p_1=0.96 ; p_2=0.04
			ω_1=0.14 ; ω_2=1.00
	M2	-1764.22	p_1=0.96 ; p_2=0.00 ; p_3=0.04
			ω_1=0.14 ; ω_2=1.00 ; ω_3=1.00
	M7	-1764.17	p_1=p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.10
			ω_1=0.00 ; ω_2=0.01 ; ω_3=0.03 ; ω_4=0.06 ; ω_5=0.10 ; ω_6=0.14 ; ω_7=0.19 ; ω_8=0.26 ; ω_9=0.37 ; ω_{10}=0.55
	M8	-1764.17	p_1=p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.10 ; p_{11}=0.00
			ω_1=0.00 ; ω_2=0.01 ; ω_3=0.03 ; ω_4=0.06 ; ω_5=0.10 ; ω_6=0.14 ; ω_7=0.19 ; ω_8=0.26 ; ω_9=0.37 ; ω_{10}=0.55 ; ω_{11}=1.00
ATXN3L1	M1	-2627.39	p_1=0.81 ; p_2=0.19
			ω=0.21 ; ω=1.00
	M2	-2627.18	p_1=0.90 ; p_2=0.00 ; p_3=0.10
			ω_1=0.25 ; ω_2=1.00 ; ω_3=1.52
	M7	-2627.89	p_1=p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.10
			ω_1=0.01 ; ω_2=0.05 ; ω_3=0.11 ; ω_4=0.18 ; ω_5=0.27 ; ω_6=0.36 ; ω_7=0.46 ; ω_8=0.58 ; ω_9=0.72 ; ω_{10}=0.88
	M8	-2627.18	p_1=p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=p_{11}=0.09
			ω_1=0.20 ; ω_2=0.22 ; ω_3=0.23 ; ω_4=0.24 ; ω_5=0.25 ; ω_6=0.26 ; ω_7=0.27 ; ω_8=0.28 ; ω_9=0.29 ; ω_{10}=0.32 ; ω_{11}=1.54

	Models compared	df	-2Δl	Critical χ^2 values (P=0.05)
ATXN3	M1 *vs* M2	2	0.00	5.99
	M7 *vs* M8		0.00	
ATXN3L1	M1 *vs* M2	2	0.41	5.99
	M7 *vs* M8		1.41	

6.3 Branch-site model

To test if some codons of *ATXN3L1* have been evolving differently in different lineages, we applied the branch-site model, in which branches on the phylogeny are divided *a priori* into foreground (*ATXN3L1* branch) and background (ancestral *ATXN3* branch) and selective pressures are allowed to vary over sites and branches. All sites seem to be constrained or neutrally evolving with ω values always below or equal 1 (Table 15). Given these results, we decided to define the duplication branch as our foreground and all the other branches as background to test whether a faster evolution has happened right after de retrotransposition event. Once more, no significant values were found.

Table 15: Parameter estimates and likelihood scores under different branch-site models. Model comparisons of variable ω ratios among sites and branches.

Model	Initial ω	Likelihood (ln)	Parameters				
H0	1 (fixed)	-3614.52	p=	0.69	0.00	0.31	0.00
			ω_{ATXN3}=	0.15	1.00	0.15	1.00
			$\omega_{ATXN3L1}$=	0.15	1.00	1.00	1.00
H1	0	-3614.52	p=	0.69	0.00	0.31	0.00
			ω_{ATXN3}=	0.15	1.00	0.15	1.00
			$\omega_{ATXN3L1}$=	0.15	1.00	1.00	1.00
	1	-3614.44	p=	0.68	0.01	0.30	0.01
			ω_{ATXN3}=	0.15	1.00	0.15	1.00
			$\omega_{ATXN3L1}$=	0.15	1.00	1.00	1.00
	2	-3614.44	p=	0.68	0.01	0.30	0.01
			ω_{ATXN3}=	0.15	1.00	0.15	1.00
			$\omega_{ATXN3L1}$=	0.15	1.00	1.00	1.00
	4	-3614.52	p=	0.69	0.00	0.31	0.00
			ω_{ATXN3}=	0.15	1.00	0.15	1.00
			$\omega_{ATXN3L1}$=	0.15	1.00	1.00	1.00
	5	-3614.44	p=	0.68	0.01	0.30	0.01
			ω_{ATXN3}=	0.15	1.00	0.15	1.00
			$\omega_{ATXN3L1}$=	0.15	1.00	1.00	1.00

Models compared	df	-2Δl	Critical χ^2 values (P=0.05)
H0 *vs* H1 (0)	1	0.00	3.84
H0 *vs* H1 (1)		0.14	
H0 *vs* H1 (2)		0.14	
H0 *vs* H1 (4)		0.00	
H0 *vs* H1 (5)		0.14	

7. Nucleotide diversity of *ATXN3L1* and *ATXN3L2*

By analyzing intraspecific nucleotide diversity, we observed that for *ATXN3L1*, gorilla (*G. gorilla*) presented the highest number of polymorphisms (26) followed by chimpanzee (*P. troglodytes*) (11), baboon (*P. anubis*) (6), japanese macaque (*M. fuscata*) (4), orangutan (*P. abelii*) (3), and cynomolgous monkey (*M. fascicularis*) (2). In gorilla, 11 non-synonymous substitutions were observed, one of them leading to a premature stop codon 21 amino acids before the end of the putative protein. From the 11 polymorphisms found in chimpanzee, 10 were registered only due to the reference sequence retrieved from Ensembl Genome Browse (with 6 of them leading to non-synonymous substitutions, including a premature stop codon). Although the high similarity between our three chimpanzee samples, they are not related to the best of our knowledge. Orangutan showed one non-synonymous substitution over the 3 polymorphisms registered. In baboon, japanese macaque and cynomolgous monkey, we were not able disentangle between the condition of *ATXN3L1* polymorphisms and sequence difference resulted from the *ATXN3L1 duplicate* since all species have a highly homologous duplicate of this retrogene. Rhesus monkey (*M. mulatta*) and marmoset (*C. jacchus*) presented only one silent and one non-synonymous polymorphism each, respectively.

Considering *ATXN3L2*, gorilla was the species showing also the highest number of polymorphisms (41 in total), 4 of them indels (3 insertions and 1 deletion). Actually, deletions have also been found in rhesus monkey (which presented 12 polymorphisms in total, the same number as cynomolgous monkeys) and in orangutans (which had 3 additional SNPs). Baboon showed the longest sequence length variant: a 52 bp insertion in one of the 7 chromosomes analyzed (in addition to 5 SNPs). Neither chimpanzee nor japanese macaque presented length polymorphisms, but only 3 SNPs in 7 and 30 chromosomes analyzed, respectively.

More details about this analysis can be retrieved from tables A3 and A4 in appendix.

8. Transcriptional pattern of *ATXN3L1*

We tested the presence of *ATXN3L1* transcripts in cDNA of different human body tissues as well as in chimpanzee brain by using primers designed to specifically amplify *ATXN3L1* cDNA. Sequencing of the RT-PCR product obtained from human brain confirmed that amplified products were specific *ATXN3L1* sequences, and not the parental *ATXN3*. Therefore, we observed that, in humans, *ATXN3L1* is transcribed in testis, placenta, brain, spleen, and thymus. Testis and placenta were the tissues for which we observed more amplified product (a similar amplification pattern for *GAPDH* in all tissues discarded the possible heterogeneity in the cDNA quantity among different tissue samples). Although in a very lower intensity, we can still notice the presence of amplified product in trachea, skeletal muscle, prostate, lung, cervix, heart, bladder, and small intestine.

Interestingly, we were also able to test transcription of *ATXN3L1* in chimpanzee brain. Regarding the sequence conservation in the annealing region of E7-8 primers between human and chimpanzee *ATXN3L1*, we tested whether a functional gene was likely to be present in this non-human primate. Indeed, we have shown that *ATXN3L1* is expressed in chimpanzee brain, although not so intensively as in humans.

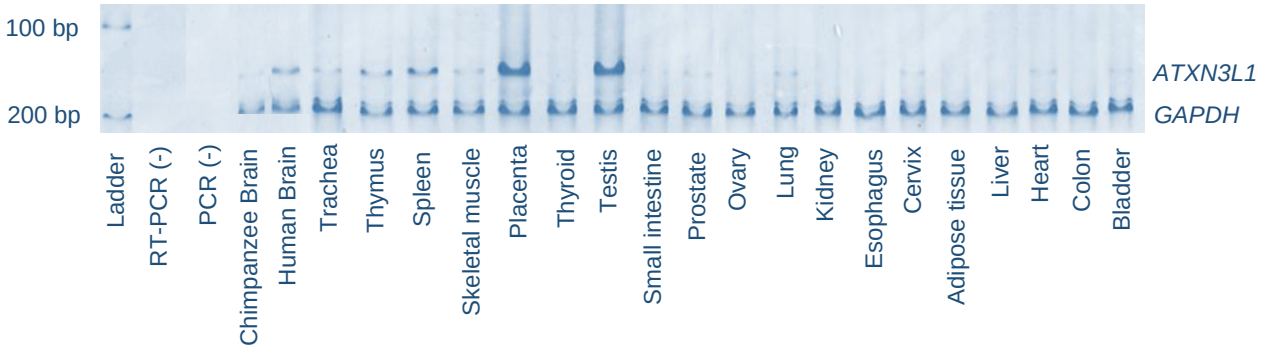


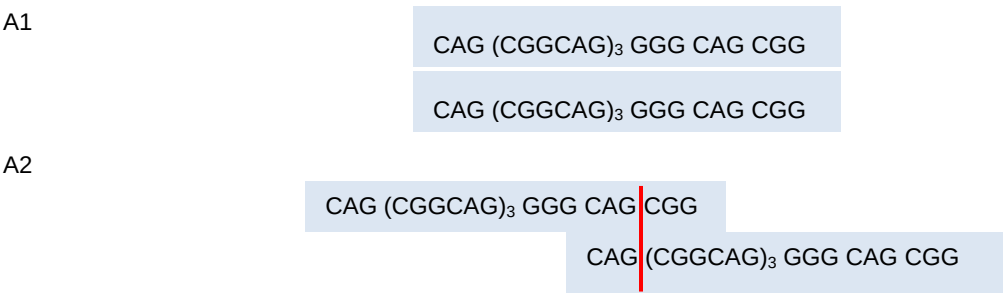
Figure 18: *ATXN3L1* transcriptional status of 20 human tissues and chimpanzee brain using E7-8 primers. *GAPDH* cDNA was amplified as a control.

Discussion

Retrotransposition and retrogenes are gaining increasing attention as recent studies have shown that they may play an important role in genome evolution and in the formation of new genes [10]. Studies focusing on the modulation of parental genes by the respective expressed paralogues may be important to gain insight into the mechanisms by which *ATXN3* retrocopies (especially *ATXN3L1*) may acquire functional relevance. In this study, we benefited from the existence of two paralogues of the *ataxin-3* gene in the primate lineage, originated through retrotransposition, to analyze the mechanisms of $(CAG)_n$ evolution. This repetitive region is shared by all three paralogues, but expanded exclusively in the parental gene (*ATXN3*) of humans. Thus, by comparing variability patterns in the normal range of three highly similar genes (regarding the CDS or homologous sequence in *ATXN3L2*), which have been under different selective pressures can allow us the testing of hypotheses on the causes of expansion and disease.

Andres et al. (2004) suggested that human capacity for expansion at repetitive loci could be explained by a higher variance and coefficient of variation observed in our species; thus, loci with increased variance may be more likely to expand and give rise to pathogenic alleles [111]. Still, one question remained: why do these disease associate repeats have higher variance than non-expanding loci? While analyzing the $(CAG)_n$ tract of our three paralogues in several primates, we noticed that a cassete-like structure is part of the *ATXN3* orangutan and *ATXN3L2* chimpanzee sequences (Figure 19). These patterns have, more likely, resulted from unequal crossover events than from successive point mutations in a pure $(CAG)_n$. Other less clear structures are found in the *ATXN3* repeat of gorilla and marmoset and additional unequal recombination may have happened in other species, but no evidence can be noticed due to a pure nature of the tract. Similarly, unequal exchange of DNA during recombination may have underlain the evolution of other disease-associated repeats and still remain unnoticed due to the lack of interruptions that allow us to detect them. Actually, in polyalanine disorders, in which repetitive tracts are highly interrupted, unequal crossing-over is often mentioned as the main mechanism for the origin of newly sized alleles [112, 113].

A.



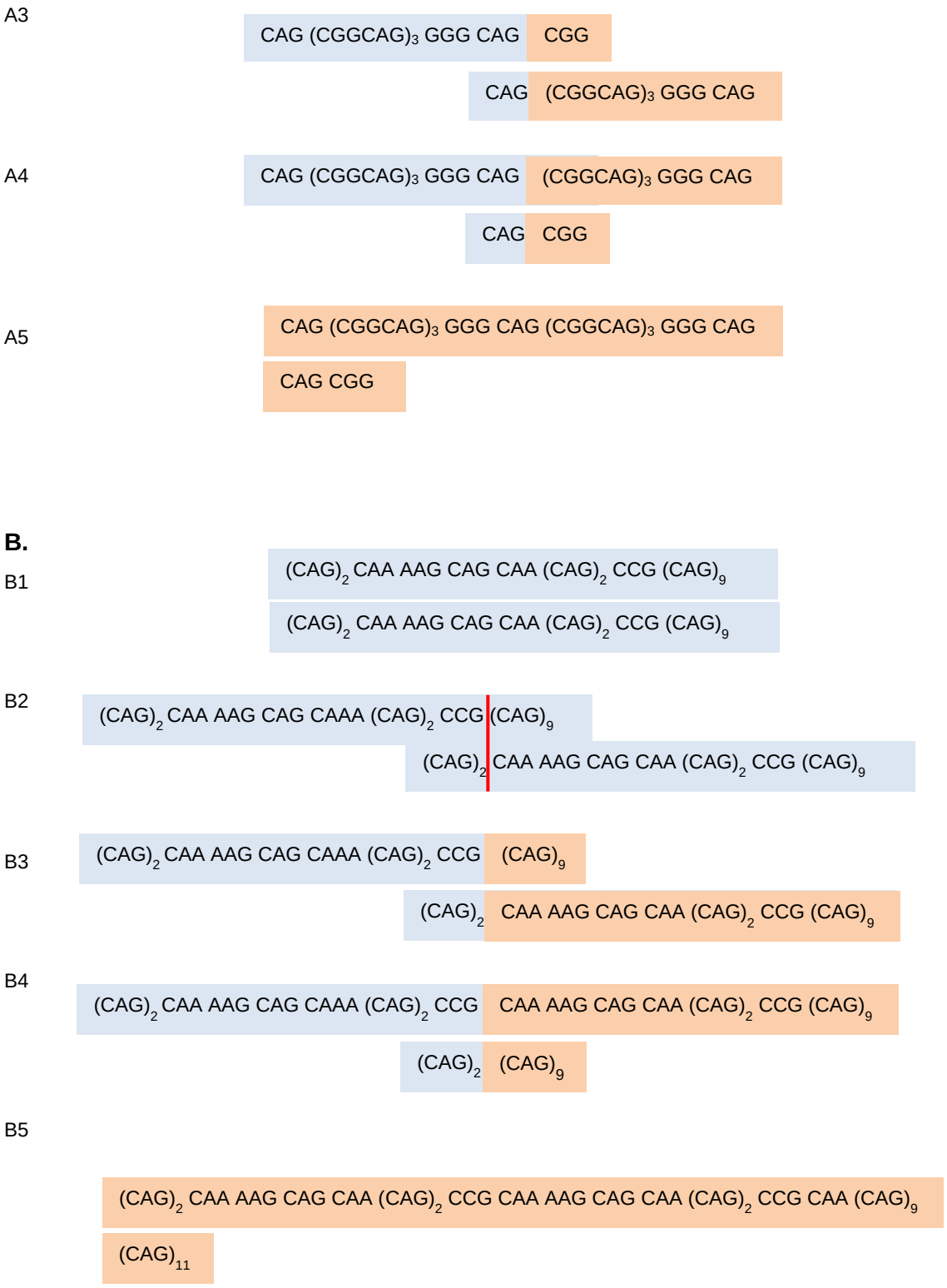


Figure 19: Schematic representation of the unequal crossing over events that may have led to the present $(CAG)_n$ tract in A. *ATXN3L2* of *Pan troglodytes*; and B. *ATXN3* of *Pongo abelii*. The red lines represent the chiasmata.

Our data suggested that the two retrocopies have independent origins: the *ATXN3L1* about 63 MYA, in the Haplorrhini clade, and *ATXN3L2* about 43 MYA, before the Platyrrhine-Catarrhine split. Currently, a short CAA_2 CAG_2 is observed in Tarsier parental gene. Similarly, the *ATXN3L1* orthologue has a CAG CAA CAG_2 motif, but the acquisition of several interruptions seems to have occurred across the primate lineage. In fact, no more than two consecutive CAGs are observed, which turns this repeat less prone to instability. Although the function of the parental ATXN3 polyQ is poorly understood, it is interesting to notice that despite the repeat interruptions, ATXN3L1 of most primates may encode 5 consecutive glutamines in the 3'-end of its polyQ stretch.

The *ATXN3L2* pseudogene presents currently an almost pure $(CAG)_n$ in several species. Uninterrupted CAGs are also common to parental rodents and porcines homologous sequences, mammalian species prior to the second retrotransposition event on the origin of *ATXN3L2* [114, 115]. This way, the repetitive tract has probably a pure nature on its origin that has been maintained throughout evolution in this non-functional gene.

Some functions have been suggested to *ATXN3* polyQ protein as transcriptional regulation, nuclear localization and protein-protein interactions [37]. A recent study has shown that species with high polyQ protein content have a higher number of proteins bearing domains with functions related to phosphatidylinositol (PI) signaling and ubiquitin-directed protein degradation [116]. As we know by then, aggregates containing proteins with an expanded polyQ stretch are ubiquitinated [95]. These facts suggest that $(CAG)_n$ tracts can actually be subjected to selective pressures, in order to maintain a possible functional role in the protein context.

Interestingly, while completing data on ataxin-3 paralogous with our own sequencing, we were able to sequence a more recent *ATXN3* retrocopy in owl monkey (*Aotus trivirgatus*). In fact, *ATXN3* gene has been on the origin of many different retrocopies along the mammalian lineage [32] maybe due to the abundance of L1 retrotransposons in mammals [4, 5].

In addition, we were able to find evidence of duplication events, occurred after the retrotransposition, in *M. fuscata*, *M. fascicularis* and *P. anubis* species. Given the high similarity between *ATXN3L1* and the respective duplicate of each species shown in the obtained phylogenetic tree, we suggest that two different duplications have occurred after speciation (Figure 20). Actually, the *ATXN3L1* genes of *Papio anubis* and *Macaca* species presented more nucleotide differences between them than between *ATXN3L1* and the respective duplicate. Thus, it is more parsimonious that two different duplications have occurred than a single event followed by subsequent selective pressures acting similarly on both copies (*ATXN3L1* and *ATXN3L1 duplicate*) of each species.

Figure 20: Phylogenetic tree showing the species where duplications of *ATXN3L1* seem to have occurred. The red stars represent the duplication events. *Hylobates* was used as an outgroup. This tree was obtained from Geneious sotware using neighbor-joining method.

In *M. fuscata*, we tried to distinguish between *ATXN3L1* and *ATXN3L1 duplicate* based on the seven different patterns found through the sequencing of 15 individuals. We started by comparing the heterozygous with homozygous non-variable positions. Then, we compared also *ATXN3L1* sequences of other primates to infer the ancestral allele, i.e. the *ATXN3L1* allele. In cases where we found different nucleotide proportions in sequencing, the ancestral allele was considered as being in *ATXN3L1* sequence and the derived allele in the duplicate sequence. This way, we were able to discern that two heterozygous positions (c.743A>G and c.923A>C) were indeed polymorphic in the duplicate, and one position (c.995C>T) was polymorphic in both *ATXN3L1* and *ATXN3L1 duplicate* (Figure 21 - A). The fluorescence imbalance in the sequencing suggests that the *ATXN3L1 duplicate* is, as well, located in the X chromosome.

We did the same exercise to *M. fascicularis* and found alleles A and T to be the ancestral alleles of c.801A>C and c.995C>T, respectively (Figure 21 - B).

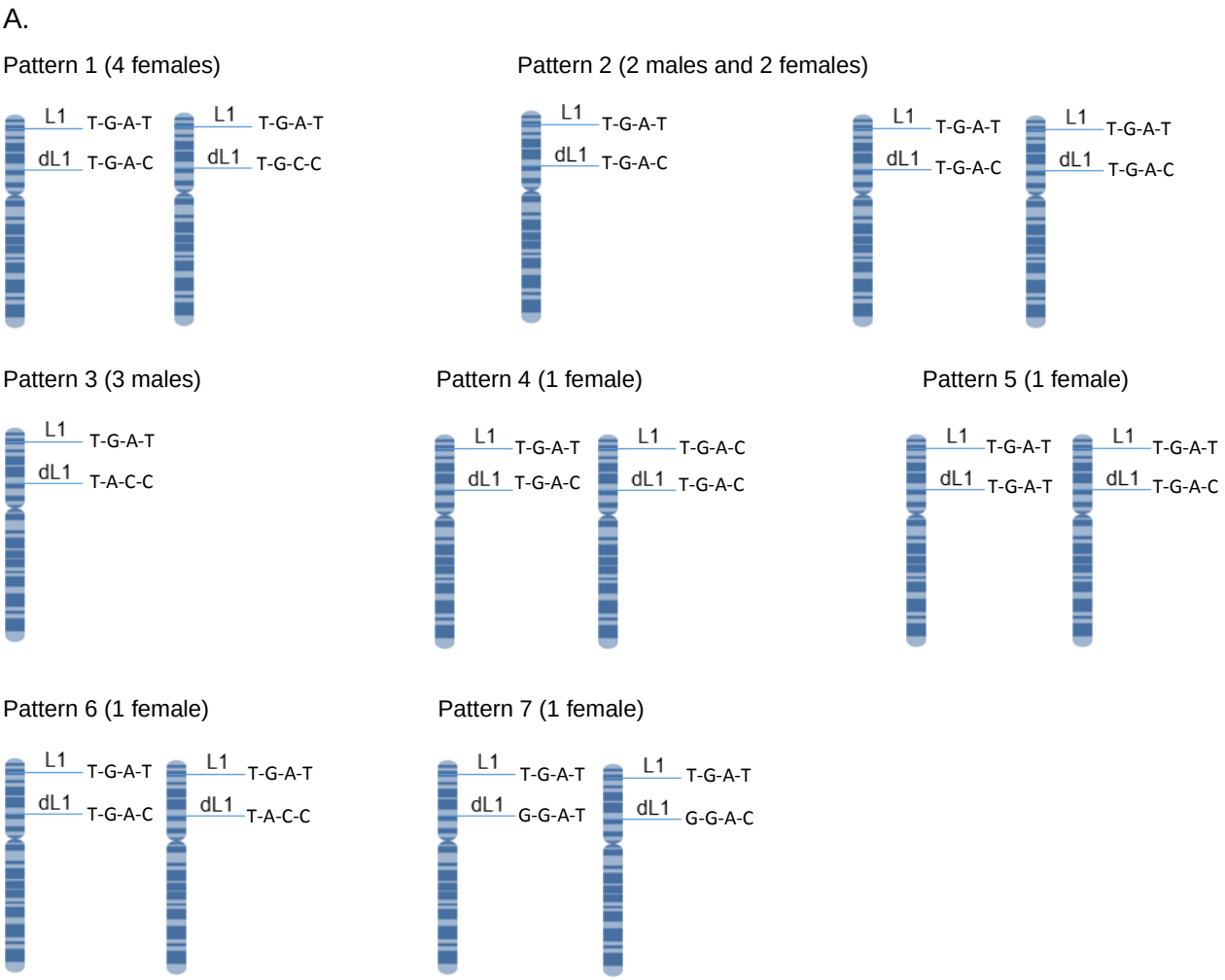


Figure 21: *ATXN3L1* (L1) and *ATXN3L1 duplicate* (dL1) haplotypes for A. positions c.561T>G, c.743A>G, c.923A>C and c.995C>T in *M. fuscata* individuals; and B. c.801A>C and c.995C>T in *M. fascicularis* males.

Regarding the divergent positions in *P. anubis*, we could not infer haplotype phases for *ATXN3L1* and *ATXN3L1 duplicate* since both sequences retrieved from the UCSC database seem to contain ancestral alleles at different positions (Table 16).

Table 16: Inferred ancestral alleles for *ATXN3L1* in *P. anubis* based on orthologous sequences observed in other primates.

Sequences	Variant					
	c.280C>T	c.319A>G	c.987A>T	c.1006A>G	c.1007C>T	c.1011C>T
P. anubis (ChrX: 10617308-10618372)	T	G	T	A	T	C
P. anubis (ChrX: 10636574-10637639)	C	A	A	G	C	T
Ancestral allele	T	A	T	G	C	C

Analyzing the nucleotide diversity of *ATXN3L1* within each of these three species (*M. fuscata*, *M. fascicularis* and *P. anubis*) no differences were found besides those related to the *ATXN3L1* duplicates. In fact, most species present low interspecific diversity, although there is the possibility of underestimating variance due to sampling from a single, localized population or inbred zoo collection, which is, for instance, the case of *M. fuscata*. According to Ensembl database, only 6 positions are indicated as polymorphic in *P. troglodytes*. We were able to find up to 10 SNPs for this species, when comparing our sequences to that from Ensembl, leading us to suppose that our samples, even unrelated, may be from a very different population. From all the species analyzed, *G. gorilla* is the only presenting a high intraspecific diversity for *ATXN3L1*. For *ATXN3L2*, a higher variance was found within species including not only SNPs but also indels for at least 4 species. This may due to the fact that *ATXN3L2* presents a highly interrupted ORF and also is located into an autosome, which is much subjected to recombination than a sexual chromosome.

The synonymous to non-synonymous substitution rate ($d_N/d_S=\omega$) is a measure of the proportion of mutations in DNA sequence that also alter amino acid sequence; it is often used to assess whether a sequence is under evolutionary constraint, being an indicator of functionality [3]. In general, a d_N/d_S ratio that is significantly lower than unity is considered to indicate functional constraint and ω values above unity indicates positive selection [72].

We performed a series of branch, site and branch-site models to test the conservative hypothesis of *ATXN3L1* and whether *ATXN3* and *ATXN3L1* experienced different selective pressures during primate evolution.

ATXN3 and *ATXN3L1* are highly conserved in the primate lineage (ω=0.17 and ω=0.33, respectively), but a scenario of functional divergence after the transposition should be considered since different selective pressures for *ATXN3* and *ATXN3L1* have been detected (two-ratio test fits our data better than one-ratio). In fact, *ATXN3* has been under stronger constraints than *ATXN3L1*. This difference does not seem to result from variation across the

sequence since we have found uniform selective pressures among sites for both genes. Actually, neutral models fit better the *ATXN3* and *ATXN3L1* data than positive selection models.

Genomic context has been found to affect the rate of duplicate gene evolution. In general, retrogenes seem to evolve faster than their parent genes, although this reflects broad trends rather than gene-by-gene studies [117]. According to our branch-site models results, all sites in both paralogues and lineages seem to be equally constrained or neutrally evolving.

We next wanted to confirm the evidence of *ATXN3L1* functional relevance by analyzing its transcription. In the NCBI – UniGene – EST Profile, *ATXN3L1* expression has been registered in human testis and brain, with higher levels of expression in testis. Our results have shown, however, that human *ATXN3L1* is also transcribed in placenta, spleen, and thymus, bringing a new insight into the functional relevance of this paralogue. It is also interesting to notice that a slight fraction of this paralogue transcript could be found in trachea, skeletal muscle, prostate, lung, cervix, heart, bladder, and small intestine. In chimpanzee, we were able to confirm *ATXN3L1* expression in brain, the most important tissue in the context of the human Machado-Joseph disease.

It has been shown that retrogenes are not randomly located on chromosomes with a significant excess of retrogenes that originate from the X chromosome and retroposed to autosomes [13, 14]. Regarding the recruitment of retrogenes, the X chromosome is the only outlier in the genome since it accepts an excess of functional copies from autosomes [14]. Functional genes disproportionately enter the X chromosome compared with nonfunctional genes most probably due to natural selection favoring the fixation and maintenance of retrogenes. Therefore, it is interesting to notice that the most conserved *ATXN3* retrocopy, *ATXN3L1*, is located on the X chromosome.

Future Perspectives

Most of the aims established for this thesis have been achieved, yet so many issues remain to reveal and so many other questions have arisen. The main conclusion taken from this study is that *ATXN3L1* is a transcriptionally-active retrogene under selective constraints supporting its functional relevance. Even so, further studies need be done to confirm the presence of endogenous protein *in vivo*, in tissues where mRNA expression has been detected by us. Thus, we will design a specific anti-ATXN3L1 antibody to detect human endogenous ATXN3L1. After having confirmed the specificity of anti-ATXN3L1, the expression levels of *ATXN3L1* will be assessed in human protein extracts, by Western blot, and in tissue sections by immunohistochemistry. Next, to better understand the functional diversification of ATXN3L1, we expect to have mammalian cell lines stably expressing ATXN3L1, which will allow us to determine ATXN3L1 subcellular location, by fluorescence, and to test its binding to ATXN3, as well as to the transcriptional coactivators and DNA repair proteins that interact with the parental ATXN3.

Further studies can be made, also, to confirm if the small *ATXN3L2* ORFs are transcribed or not, since it may have functional relevance by regulating the expression of the parental gene, as demonstrated for other transcripts originated from retrogenes. To detect mRNA expression of *ATXN3L2*, RT-PCR will be performed with allele-specific primers in cDNAs of 20 different normal human tissues (the same approach we used for ATXN3L1). To unequivocally discriminate this transcript from the homologous, primers will be designed to amplify concatenated regions corresponding to *ATXN3* exons, in gene-specific nucleotides with highly conserved sites across species within orthologues.

We can also increase the reliability of our phylogenetic analysis of selection including more species sequences. Most of the species analyzed in this work were successfully sequenced for *ATXN3L1* and *ATXN3L2*, lacking the *ATXN3* sequence. Specific primers must be design to amplify the *ATXN3* gene of species assessed for the other paralogues. Finally, a higher and widest sampling can also help us to get closer to more reliable nucleotide diversity.

References

1. Vinckenbosch, N., I. Dupanloup, and H. Kaessmann, *Evolutionary fate of retroposed gene copies in the human genome.* Proc Natl Acad Sci U S A, 2006. **103**(9): p. 3220-5.
2. Ohshima, K., *RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity.* International journal of evolutionary biology, 2013. **2013**.
3. Pink, R.C., et al., *Pseudogenes: Pseudo-functional or key regulators in health and disease?* Rna, 2011. **17**(5): p. 792-798.
4. Pan, D. and L. Zhang, *Burst of young retrogenes and independent retrogene formation in mammals.* PloS one, 2009. **4**(3): p. e5040.
5. Ostertag, E.M. and H.H. Kazazian Jr, *Biology of mammalian L1 retrotransposons.* Annual review of genetics, 2001. **35**(1): p. 501-538.
6. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nature Reviews Genetics, 2009. **10**(10): p. 691-703.
7. Brosius, J. and *Retroposons--seeds of evolution* Science, 1991. **251**(4995): p. 753.
8. Sayah, D.M., et al., *Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1.* Nature, 2004. **430**(6999): p. 569-573.
9. Babushok, D.V., et al., *A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids.* Genome research, 2007. **17**(8): p. 1129-1138.
10. Kaessmann, H., N. Vinckenbosch, and M. Long, *RNA-based gene duplication: mechanistic and evolutionary insights.* Nature Reviews Genetics, 2009. **10**(1): p. 19-31.
11. Betrán, E., et al., *Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene.* Molecular biology and evolution, 2002. **19**(5): p. 654-663.
12. Burki, F. and H. Kaessmann, *Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux.* Nature genetics, 2004. **36**(10): p. 1061-1063.
13. Betrán, E., K. Thornton, and M. Long, *Retroposed new genes out of the X in Drosophila.* Genome research, 2002. **12**(12): p. 1854-1859.
14. Emerson, J., et al., *Extensive gene traffic on the mammalian X chromosome.* Science, 2004. **303**(5657): p. 537-540.
15. Svensson, Ö., L. Arvestad, and J. Lagergren, *Genome-wide survey for biologically functional pseudogenes.* PLoS computational biology, 2006. **2**(5): p. e46.
16. Graur, D., Y. Shuali, and W.-H. Li, *Deletions in processed pseudogenes accumulate faster in rodents than in humans.* Journal of molecular evolution, 1989. **28**(4): p. 279-285.
17. Zhang, Z. and M. Gerstein, *Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes.* Nucleic acids research, 2003. **31**(18): p. 5338-5348.
18. Féral, C., G. Guellaën, and A. Pawlak, *Human testis expresses a specific poly (A)-binding protein.* Nucleic acids research, 2001. **29**(9): p. 1872-1883.
19. PERNA, N.T., et al., *Alu insertion polymorphism: a new type of marker for human population studies.* Human biology, 1992: p. 641-648.
20. Ryan, S.C. and A. Dugaiczyk, *Newly arisen DNA repeats in primate phylogeny.* Proceedings of the National Academy of Sciences, 1989. **86**(23): p. 9360-9364.
21. Shimamura, M., et al., *Molecular evidence from retroposons that whales form a clade within even-toed ungulates.* Nature, 1997. **388**(6643): p. 666-670.
22. Shedlock, A.M. and N. Okada, *SINE insertions: powerful tools for molecular systematics.* Bioessays, 2000. **22**(2): p. 148-160.
23. Witherspoon, D., et al., *Human population genetic structure and diversity inferred from polymorphic L1 (LINE-1) and Alu insertions.* Human heredity, 2006. **62**(1): p. 30-46.

24. Ray, D.A., J.A. Walker, and M.A. Batzer, *Mobile element-based forensic genomics.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2007. **616**(1): p. 24-33.
25. Bettencourt, C., et al., *Increased transcript diversity: novel splicing variants of Machado–Joseph Disease gene (ATXN3).* neurogenetics, 2010. **11**(2): p. 193-202.
26. Bettencourt, C. and M. Lima, *Machado-Joseph Disease: from first descriptions to new perspectives.* Orphanet J Rare Dis, 2011. **6**: p. 35.
27. Kawaguchi, Y., et al., *CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32. 1.* Nature genetics, 1994. **8**(3): p. 221-228.
28. Djian, P., J.M. Hancock, and H.S. Chana, *Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration.* Proceedings of the National Academy of Sciences, 1996. **93**(1): p. 417-421.
29. Fitch, W.M., *Distinguishing homologous from analogous proteins.* Systematic Biology, 1970. **19**(2): p. 99-113.
30. Fitch, W.M., *Homology: a personal view on some of the problems.* Trends in genetics, 2000. **16**(5): p. 227-231.
31. Weeks, S.D., et al., *Crystal Structure of a Josephin-Ubiquitin Complex EVOLUTIONARY RESTRAINTS ON ATAXIN-3 DEUBIQUITINATING ACTIVITY.* Journal of Biological Chemistry, 2011. **286**(6): p. 4555-4565.
32. Martins, M., *Evolution and Functional Relevance of Ataxin-3 Paralogues*, in *Secção Autónoma de Ciências da Saúde*2012, Aveiro.
33. Tzvetkov, N. and P. Breuer, *Josephin domain-containing proteins from a variety of species are active de-ubiquitination enzymes.* Biological chemistry, 2007. **388**(9): p. 973-978.
34. Seki, T., et al., *JosD1, a Membrane-targeted Deubiquitinating Enzyme, Is Activated by Ubiquitination and Regulates Membrane Dynamics, Cell Motility, and Endocytosis.* Journal of Biological Chemistry, 2013. **288**(24): p. 17145-17155.
35. Masino, L., et al., *Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail.* FEBS letters, 2003. **549**(1): p. 21-25.
36. Scheel, H., S. Tomiuk, and K. Hofmann, *Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics.* Human molecular genetics, 2003. **12**(21): p. 2845-2852.
37. Matos, C.A., S. de Macedo-Ribeiro, and A.L. Carvalho, *Polyglutamine diseases: the special case of ataxin-3 and Machado–Joseph disease.* Progress in Neurobiology, 2011. **95**(1): p. 26-48.
38. Riess, O., et al., *SCA3: neurological features, pathogenesis and animal models.* The Cerebellum, 2008. **7**(2): p. 125-137.
39. Donaldson, K.M., et al., *Ubiquitin-mediated sequestration of normal cellular proteins into polyglutamine aggregates.* Proceedings of the National Academy of Sciences, 2003. **100**(15): p. 8892-8897.
40. Nicastro, G., et al., *Understanding the role of the Josephin domain in the PolyUb binding and cleavage properties of ataxin-3.* PloS one, 2010. **5**(8): p. e12430.
41. Macedo-Ribeiro, S., et al., *Nucleocytoplasmic shuttling activity of ataxin-3.* PloS one, 2009. **4**(6): p. e5834.
42. Chai, Y., et al., *Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions in SCA3/MJD and suppression of polyglutamine aggregation in vitro.* Human molecular genetics, 1999. **8**(4): p. 673-682.
43. Li, F., et al., *Ataxin-3 is a histone-binding protein with two independent transcriptional corepressor activities.* Journal of Biological Chemistry, 2002. **277**(47): p. 45004-45012.

44. Evert, B.O., et al., *Ataxin-3 represses transcription via chromatin binding, interaction with histone deacetylase 3, and histone deacetylation.* J Neurosci, 2006. **26**(44): p. 11474-86.
45. Zhou, L., et al., *Ataxin-3 Protects Cells Against H< sub> 2</sub> o< sub> 2</sub>-Induced Oxidative Stress By Enhancing The Interaction Between Bcl-X< sub> l</sub> And Bax.* Neuroscience, 2013.
46. Crespo-Barreto, J., et al., *Partial loss of ataxin-1 function contributes to transcriptional dysregulation in spinocerebellar ataxia type 1 pathogenesis.* PLoS genetics, 2010. **6**(7): p. e1001021.
47. Bowman, A.B., et al., *Duplication of Atxn1l suppresses SCA1 neuropathology by decreasing incorporation of polyglutamine-expanded ataxin-1 into native complexes.* Nature genetics, 2007. **39**(3): p. 373-379.
48. Thomas, P.S., et al., *Loss of endogenous androgen receptor protein accelerates motor neuron degeneration and accentuates androgen insensitivity in a mouse model of X-linked spinal and bulbar muscular atrophy.* Human molecular genetics, 2006. **15**(14): p. 2225-2238.
49. Van Raamsdonk, J.M., et al., *Loss of wild-type huntingtin influences motor dysfunction and survival in the YAC128 mouse model of Huntington disease.* Human molecular genetics, 2005. **14**(10): p. 1379-1392.
50. Van Raamsdonk, J.M., et al., *Wild-type huntingtin ameliorates striatal neuronal atrophy but does not prevent other abnormalities in the YAC128 mouse model of Huntington disease.* BMC neuroscience, 2006. **7**(1): p. 80.
51. Dragatsis, I., M.S. Levine, and S. Zeitlin, *Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice.* Nature genetics, 2000. **26**(3): p. 300-306.
52. Auerbach, W., et al., *The HD mutation causes progressive lethal neurological disease in mice expressing reduced levels of huntingtin.* Human molecular genetics, 2001. **10**(22): p. 2515-2523.
53. Martins, S., et al., *A multistep mutation mechanism drives the evolution of the CAG repeat at MJD/SCA3 locus.* Eur J Hum Genet, 2006. **14**(8): p. 932-40.
54. Orr, H.T. and H.Y. Zoghbi, *Trinucleotide repeat disorders.* Annu. Rev. Neurosci., 2007. **30**: p. 575-621.
55. Gatchel, J.R. and H.Y. Zoghbi, *Diseases of unstable repeat expansion: mechanisms and common principles.* Nature Reviews Genetics, 2005. **6**(10): p. 743-755.
56. Zühlke, C., et al., *Mitotic stability and meiotic variability of the (CAG) n repeat in the Huntington disease gene.* Human molecular genetics, 1993. **2**(12): p. 2063-2067.
57. Keckarevic, D., et al., *The status of SCA1, MJD/SCA3, FRDA, DRPLA and MD triplet containing genes in patients with Huntington disease and healthy controls.* Journal of neurogenetics, 2000. **14**(4): p. 257-263.
58. Lorenzetti, D., S. Bohlega, and H.Y. Zoghbi, *The expansion of the CAG repeat in ataxin-2 is a frequent cause of autosomal dominant spinocerebellar ataxia.* Neurology, 1997. **49**(4): p. 1009-1013.
59. Sobue, G., *X-linked recessive bulbospinal neuronopathy (SBMA).* Nagoya Journal of Medical Science, 1995. **58**: p. 95-106.
60. Koide, R., et al., *Unstable expansion of CAG repeat in hereditary dentatorubral–pallidoluysian atrophy (DRPLA).* Nature genetics, 1994. **6**(1): p. 9-13.
61. Carlson, K.M., J.M. Andresen, and H.T. Orr, *Emerging pathogenic pathways in the spinocerebellar ataxias.* Current opinion in genetics & development, 2009. **19**(3): p. 247-253.
62. Gunawardena, S., et al., *Disruption of Axonal Transport by Loss of Huntingtin or Expression of Pathogenic PolyQ Proteins in< i> Drosophila</i>.* Neuron, 2003. **40**(1): p. 25-40.

63. Muchowski, P.J., et al., *Hsp70 and hsp40 chaperones can inhibit self-assembly of polyglutamine proteins into amyloid-like fibrils.* Proceedings of the National Academy of Sciences, 2000. **97**(14): p. 7841-7846.
64. Yoo, S.-Y., et al., *SCA7 knockin mice model human SCA7 and reveal gradual accumulation of mutant ataxin-7 in neurons and abnormalities in short-term plasticity.* Neuron, 2003. **37**(3): p. 383-401.
65. Watase, K., et al., *A Long CAG Repeat in the Mouse< i> Sca1</i> Locus Replicates SCA1 Features and Reveals the Impact of Protein Solubility on Selective Neurodegeneration.* Neuron, 2002. **34**(6): p. 905-919.
66. Watase, K., et al., *Spinocerebellar ataxia type 6 knockin mice develop a progressive neuronal dysfunction with age-dependent accumulation of mutant CaV2. 1 channels.* Proceedings of the National Academy of Sciences, 2008. **105**(33): p. 11987-11992.
67. Yeh, S., et al., *Generation and characterization of androgen receptor knockout (ARKO) mice: an in vivo model for the study of androgen functions in selective tissues.* Proceedings of the National Academy of Sciences, 2002. **99**(21): p. 13498-13503.
68. Zeitlin, S., et al., *Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue.* Nature genetics, 1995. **11**(2): p. 155-163.
69. Matilla, A., et al., *Mice lacking ataxin-1 display learning deficits and decreased hippocampal paired-pulse facilitation.* The Journal of neuroscience, 1998. **18**(14): p. 5508-5516.
70. Kiehl, T.-R., et al., *Generation and characterization of Sca2 (ataxin-2) knockout mice.* Biochemical and biophysical research communications, 2006. **339**(1): p. 17-24.
71. Crespo-Barreto, J., et al., *Partial loss of ataxin-1 function contributes to transcriptional dysregulation in spinocerebellar ataxia type 1 pathogenesis.* PLoS Genet, 2010. **6**(7): p. e1001021.
72. Li, F., et al., *Ataxin-3 is a histone-binding protein with two independent transcriptional corepressor activities.* J Biol Chem, 2002. **277**(47): p. 45004-12.
73. Limprasert, P., et al., *Analysis of CAG repeat of the Machado-Joseph gene in human, chimpanzee and monkey populations: a variant nucleotide is associated with the number of CAG repeats.* Hum Mol Genet, 1996. **5**(2): p. 207-13.
74. Takiyama, Y., et al., *The gene for Machado–Joseph disease maps to human chromosome 14q.* Nature genetics, 1993. **4**(3): p. 300-304.
75. Jardim, L.B., et al., *A survey of spinocerebellar ataxia in South Brazil–66 new cases with Machado-Joseph disease, SCA7, SCA8, or unidentified disease–causing mutations.* Journal of neurology, 2001. **248**(10): p. 870-876.
76. Vale, J., et al., *Autosomal dominant cerebellar ataxia: frequency analysis and clinical characterization of 45 families from Portugal.* European Journal of Neurology, 2010. **17**(1): p. 124-128.
77. Zhao, Y., et al., *Prevalence and ethnic differences of autosomal-dominant cerebellar ataxia in Singapore.* Clinical genetics, 2002. **62**(6): p. 478-481.
78. Tang, B., et al., *Frequency of SCA1, SCA2, SCA3/MJD, SCA6, SCA7, and DRPLA CAG trinucleotide repeat expansion in patients with hereditary spinocerebellar ataxia from Chinese kindreds.* Archives of neurology, 2000. **57**(4): p. 540.
79. Van de Warrenburg, B., et al., *Spinocerebellar ataxias in the Netherlands Prevalence and age at onset variance analysis.* Neurology, 2002. **58**(5): p. 702-708.
80. Schöls, L., et al., *Autosomal dominant cerebellar ataxia: phenotypic differences in genetically defined subtypes?* Annals of neurology, 1997. **42**(6): p. 924-932.

81. Maruyama, H., et al., *Difference in disease-free survival curve and regional distribution according to subtype of spinocerebellar ataxia: A study of 1,286 Japanese patients.* American journal of medical genetics, 2002. **114**(5): p. 578-583.
82. Kraft, S., et al., *Adult onset spinocerebellar ataxia in a Canadian movement disorders clinic.* The Canadian Journal of Neurological Sciences, 2005. **32**(4): p. 450-458.
83. Moseley, M.L., et al., *Incidence of dominant spinocerebellar and Friedreich triplet repeats among 361 ataxia families.* Neurology, 1998. **51**(6): p. 1666-1671.
84. Alonso, E., et al., *Distinct distribution of autosomal dominant spinocerebellar ataxia in the Mexican population.* Movement disorders, 2007. **22**(7): p. 1050-1053.
85. Storey, E., et al., *Frequency of spinocerebellar ataxia types 1, 2, 3, 6, and 7 in Australian patients with spinocerebellar ataxia.* American journal of medical genetics, 2000. **95**(4): p. 351-358.
86. Saleem, Q., et al., *Molecular analysis of autosomal dominant hereditary ataxias in the Indian population: high frequency of SCA2 and evidence for a common founder mutation.* Human genetics, 2000. **106**(2): p. 179-187.
87. Bryer, A., et al., *The hereditary adult-onset ataxias in South Africa.* Journal of the neurological sciences, 2003. **216**(1): p. 47-54.
88. Brusco, A., et al., *Molecular genetics of hereditary spinocerebellar ataxia: mutation analysis of spinocerebellar ataxia genes and CAG/CTG repeat expansion detection in 225 Italian families.* Archives of neurology, 2004. **61**(5): p. 727.
89. Coutinho, P. and C. Andrade, *Autosomal dominant system degeneration in Portuguese families of the Azores Islands A new genetic disorder involving cerebellar, pyramidal, extrapyramidal and spinal cord motor functions.* Neurology, 1978. **28**(7): p. 703-703.
90. Carvalho, D.R., et al., *Homozygosity enhances severity in spinocerebellar ataxia type 3.* Pediatric neurology, 2008. **38**(4): p. 296-299.
91. Sobczak, K. and W.J. Krzyzosiak, *Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability.* Hum Mutat, 2004. **24**(3): p. 236-47.
92. Maciel, P., et al., *Correlation between CAG repeat length and clinical features in Machado-Joseph disease.* American journal of human genetics, 1995. **57**(1): p. 54.
93. Maruyama, H., et al., *Molecular features of the CAG repeats and clinical manifestation of Machado-Joseph disease.* Human molecular genetics, 1995. **4**(5): p. 807-812.
94. Takiyama, Y., et al., *Evidence for inter-generational instability in the CAG repeat in the MJD1 gene and for conserved haplotypes at flanking markers amongst Japanese and Caucasian subjects with Machado-Joseph disease.* Human molecular genetics, 1995. **4**(7): p. 1137-1146.
95. Todi, S.V., et al., *Cellular turnover of the polyglutamine disease protein ataxin-3 is regulated by its catalytic activity.* Journal of Biological Chemistry, 2007. **282**(40): p. 29348-29358.
96. Mueller, T., et al., *CK2-dependent phosphorylation determines cellular localization and stability of ataxin-3.* Human molecular genetics, 2009. **18**(17): p. 3334-3343.
97. Tzvetkov, N. and P. Breuer, *Josephin domain-containing proteins from a variety of species are active de-ubiquitination enzymes.* Biol Chem, 2007. **388**(9): p. 973-8.
98. Burnett, B., F. Li, and R.N. Pittman, *The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity.* Hum Mol Genet, 2003. **12**(23): p. 3195-205.
99. Evert, B.O., et al., *Inflammatory genes are upregulated in expanded ataxin-3-expressing cell lines and spinocerebellar ataxia type 3 brains.* The Journal of Neuroscience, 2001. **21**(15): p. 5389-5396.

100. Evert, B.O., et al., *Gene expression profiling in ataxin-3 expressing cell lines reveals distinct effects of normal and mutant ataxin-3.* Journal of Neuropathology & Experimental Neurology, 2003. **62**(10): p. 1006-1018.
101. Santos, D., *Mecanismos mutacionais associados aos alelos (CAG)n normais do gene receptor de androgénio*, in *Biology Department*2010, Porto.
102. Morrill, B.H., L.F. Rickords, and H.J. Schafstall, *Sequence length polymorphisms within primate amelogenin and amelogenin-like genes: usefulness in sex determination.* American journal of primatology, 2008. **70**(10): p. 976-985.
103. Yang, Z., *Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.* Molecular biology and evolution, 1998. **15**(5): p. 568-573.
104. Bielawski, J.P. and Z. Yang, *Maximum likelihood methods for detecting adaptive evolution after gene duplication*, in *Genome Evolution*. 2003, Springer. p. 201-212.
105. Nielsen, R. and Z. Yang, *Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.* Genetics, 1998. **148**(3): p. 929-936.
106. Yang, Z., et al., *Codon-substitution models for heterogeneous selection pressure at amino acid sites.* Genetics, 2000. **155**(1): p. 431-449.
107. Yang, Z. and R. Nielsen, *Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.* Molecular biology and evolution, 2002. **19**(6): p. 908-917.
108. Barber, R.D., et al., *GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues.* Physiological genomics, 2005. **21**(3): p. 389-395.
109. Jobling, M.A., M. Hurles, and C. Tyler-Smith, *Human evolutionary genetics: origins, peoples & disease*. 2004. p. 204.
110. Goldman, N. and Z. Yang, *A codon-based model of nucleotide substitution for protein-coding DNA sequences.* Molecular biology and evolution, 1994. **11**(5): p. 725-736.
111. Andres, A.M., et al., *Comparative genetics of functional trinucleotide tandem repeats in humans and apes.* J Mol Evol, 2004. **59**(3): p. 329-39.
112. Arai, H., et al., *De novo polyalanine expansion of PHOX2B in congenital central hypoventilation syndrome: unequal sister chromatid exchange during paternal gametogenesis.* Journal of human genetics, 2007. **52**(11): p. 921-925.
113. Brown, L.Y. and S.A. Brown, *Alanine tracts: the expanding story of human illness and trinucleotide repeats.* TRENDS in Genetics, 2004. **20**(1): p. 51-58.
114. Madsen, L.B., et al., *Identification of the porcine homologous of human disease causing trinucleotide repeat sequences.* Neurogenetics, 2007. **8**(3): p. 207-218.
115. Bakalian, A., N. Delhaye-Bouchaud, and J. Mariani, *Quantitative analysis of the Purkinje cell and the granule cell populations in the cerebellum of nude mice.* Journal of neurogenetics, 1995. **9**(4): p. 207-218.
116. Schaefer, M.H., E.E. Wanker, and M.A. Andrade-Navarro, *Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks.* Nucleic acids research, 2012. **40**(10): p. 4273-4287.
117. McCole, R.B., et al., *A CASE-BY-CASE EVOLUTIONARY ANALYSIS OF FOUR IMPRINTED RETROGENES.* Evolution, 2011. **65**(5): p. 1413-1427.

Appendix

Table A1: Gender of all samples analyzed in this work. (The gender was determined by us (see Results section, "Sex determination"))

Samples	Sex	Samples	Sex
Chimpanzee 1	male	Japanese macaque 3	male
Chimpanzee 2	female	Japanese macaque 4	male
Chimpanzee 3	male	Japanese macaque 5	female
Gorilla 1	female	Japanese macaque 6	male
Gorilla 2	male	Japanese macaque 7	female
Gorilla 3	male	Japanese macaque 8	female
Gorilla 4	female	Japanese macaque 9	female
Orangutan 1	male	Japanese macaque 10	female
Orangutan 2	male	Japanese macaque 11	female
Orangutan 3	male	Japanese macaque 12	male
Gibbon	female	Japanese macaque 13	female
Baboon 1	female	Japanese macaque 14	female
Baboon 2	female	Japanese macaque 15	male
Baboon 3	male	Cynomolgous monkey 1	male
Rhesus monkey 1	male	Cynomolgous monkey 2	female
Rhesus monkey 2	male	Cynomolgous monkey 3	female
Rhesus monkey 3	male	Cynomolgous monkey 4	male
Rhesus monkey 4	male	African green monkey	female
Rhesus monkey 5	female	Owl monkey	female
Rhesus monkey 6	female	Marmoset 1	female
Rhesus monkey 7	female	Marmoset 2	male
Japanese macaque 1	female	Marmoset 3	female
Japanese macaque 2	female	Cotton-top tamarin	female

Table A2: Common name of all the species analyzed in this work.

Species	Common name
Homo sapiens	Human
Pan troglodytes	Chimpanzee
Pan paniscus	Pigmy chimpanzee
Gorilla gorilla	Gorilla
Pongo abelii	Sumatran orangutan
Pygmaeus pongus	Orangutan
Nomascus leucogenys	Northern white-cheeked gibbon
Hylobates	Gibbon
Papio anubis	Olive Baboon
Papio hamadryas	Baboon
Macaca mulatta	Rhesus monkey
Macaca fuscata	Japanese macaque
Macaca fascicularis	Cynomolgous monkey
Cercopithecus aethiops	African green monkey
Aotus trivirgatus	Owl monkey
Saimiri boliviensis	Squirrel monkey
Callithrix jacchus	Marmoset
Saguinus oedipus	Cotton-top tamarin
Tarsius syrichta	Tarsier
Otolemur garnettii	Bushbaby

Table A3: Parameters used in codeml for phylogenetic analysis of selection.

Models	Parameters					
	run mode	model	Nsites	ncatG	ω	$\omega_{initial}$
Branch models						
one-ratio	0	0	0	2	0	1.6
two-ratio	0	2	0	2	0	1.6
free-ratio	0	1	0	2	0	1.6
Site models						
M1	0	0	1	2	0	1.6
M2	0	0	2	2	0	1.6
M7	0	0	7	2	0	1.6
M8	0	0	8	2	0	1.6
Branch-site models						
H0	0	2	2	4	1	1
H1.0	0	2	2	4	0	0
H1.1	0	2	2	4	0	1
H1.2	0	2	2	4	0	2
H1.4	0	2	2	4	0	4
H1.5	0	2	2	4	0	5

Table A4: Nucleotide diversity found in *ATXN3L1*. Absolute frequencies, MAF (Minor Allele Frequency) and amino acid change were not determined for *P. anubis*, *M. fuscata* and *M. fascicularis* due to heterozygous positions found in males.

Species	Variation ID	Base position	Number of chromosomes analyzed	Base change	Absolute frequencies	MAF	Amino acid change
P. troglodytes	c.295C>T	295	5	C/T	1/4	0.8	Y/P
	c.296C>A	296		C/A	1/4	0.8	
	c.356A>G	356		A/G	1/4	0.8	H/R
	c.363C>T	363		C/T	1/4	0.8	F/F
	c.368C>T	368		C/T	1/4	0.8	I/T
	c.397C>T	397		C/T	1/4	0.8	L/L
	c.409C>T	409		C/T	1/4	0.8	L/L
	c.420A>G	420		A/G	1/4	0.8	P/P
	c.443C>T	443		C/T	1/4	0.8	L/P
	c.460C>T	460		C/T	1/4	0.8	R/stop
	c.664A>G	664	4	A/G	3/1	0.75	G/G
G. gorilla	c.37C>T	37	7	C/T	5/2	0.71	L/L
	c.307A>G	307		A/G	5/2	0.71	S/G
	c.378C>T	378		C/T	6/1	0.86	F/F
	c.414A>G	414		A/G	3/4	0.57	A/A
	c.417T>G	417		T/G	3/4	0.57	G/G
	c.427A>G	427		A/G	3/4	0.57	I/V
	c.435C>T	435		C/T	2/5	0.71	D/D
	c.472C>G	472		C/G	2/5	0.71	E/Q
	c.587C>T	587		C/T	2/5	0.71	L/S
	c.606C>T	606		C/T	2/5	0.71	H/H
	c.636A>G	636		A/G	5/2	0.71	V/V
	c.678C>G	678		C/G	2/5	0.71	E/D
	c.694A>G	694		A/G	2/5	0.71	A/T
	c.737A>C	737		A/C	2/5	0.71	A/D
	c.738C>T	738		C/T	2/5	0.71	
	c.777C>T	777		C/T	5/2	0.71	S/S
	c.852A>G	852		A/G	5/2	0.71	E/E
	c.865A>C	865		A/C	5/2	0.71	K/Q
	c.887A>T	887		A/T	5/2	0.71	Q/L
	C.915C>T	915		C/T	5/2	0.71	G/G
	c.916C>T	916		C/T	5/2	0.71	H/Y
	c.928A>C	928		A/C	2/5	0.71	L/I
	c.934A>G	934		A/G	1/6	0.86	E/K
	c.962C>T	962		C/T	1/6	0.86	T/I
	c.972C>T	972		C/T	2/5	0.71	D/D
	c.1003C>T	1003		C/T	2/5	0.71	Stop/Q
P. abelii	c.435C>T	435	3	C/T	1/2	0.67	D/D
	c.989A>G	989		A/G	1/2	0.67	S/N
	c.1045C>T	1045		C/T	1/2	0.67	L/L
P. anubis	c.280C>T	280	5	C/T			
	c.319A>G	319		A/G			
	c.987A>T	987	6	A/T			
	c.1006A>G	1006		A/G			
	c.1007C>T	1007		C/T			
	c.1011C>T	1011		C/T			
M. mulatta	c.174 C>T	174	11	C/T	7/4	0.64	Y/Y
M. fuscata	c.561T>G	561	25	T/G			
	c.743A>G	743		A/G			
	c.923A>C	923		A/C			
	c.995C>T	995		C/T			
M. fascicularis	c.801A/C	801	6	A/C			
	c.975C/T	975		C/T			
C. jacchus	c.930A>T	930	6	A/T	3/3	0,5	D/E

A – adenine; C – cytosine; G – guanine; T – timine; MAF - maximum allele frequency; V - Valine; T – Threonine; E - Glutamic acid; R – Arginine; K – Lysine; N – Asparagine; M- Methionine; Y – Tyrosine; D - Aspartic acid; G – Glycine; L - Leucine

Table A5: Nucleotide diversity found in *ATXN3L2*.

Species	Variation ID	Base position	Number of chromosomes analyzed	Base change	Absolute frequencies	MAF
P. troglodytes	c.434G>T	434	7	G/T	5/2	0.71
	c.577C>T	577		C/T	5/2	0.71
	c.1021C>G	1021		C/G	5/2	0.71
G. gorilla	c.56A>G	56	9	A/G	7/2	0.78
	c.65A>T	65		A/T	1/8	0.89
	c.122A>T	122		A/T	2/7	0.78
	c.185A>T	185		A/T	2/7	0.78
	c.190A>C	190		A/C	7/2	0.78
	c.195C>G	195		C/G	2/7	0.78
	c.198C>T	198		C/T	2/7	0.78
	c.218T>G	218		T/G	6/3	0.67
	c.223.1T	223		insT	4/5	0.56
	c.224C>G	224		C/G	8/1	0.89
	c.225C>G	225		C/G	8/1	0.89
	c.225.4T	225		ins4T	1/8	0.89
	c.225.4T.1G.1C.4T	225		ins4TGC4T	1/8	0.89
	c.302A>G	302		A/G	2/7	0.78
	c.310A>G	310		A/G	7/2	0.78
	c.331A>G	331		A/G	7/2	0.78
	c.344C>T	344		C/T	2/7	0.78
	c.362A>G	362		A/G	6/3	0.67
	c.366C>T	366		C/T	4/5	0.56
	c.371A>T	371		A/T	7/2	0.78
	c.434T-del	434		delT	2/7	0.78
	c.530C>T	530		C/T	7/2	0.78
	c.574C>T	574		C/T	7/2	0.78
	c.575C>T	575		C/T	7/2	0.78
	c.581A>G	581		A/G	7/2	0.78
	c.592C>T	592		C/T	7/2	0.78
	c.647A>G	647		A/G	7/2	0.78
	c.651C>T	651		C/T	8/1	0.89
	c.654A>G	654		A/G	7/2	0.78
	c.668A>G	668		A/G	2/7	0.78
	c.669A>G	669		A/G	2/7	0.78
	c.679A>T	679		A/T	7/2	0.78
	c.684C>T	684		C/T	1/8	0.89
	c.694C>G	694		C/G	7/2	0.78
	c.722A>G	722		A/G	2/7	0.78
	c.727A>G	727		A/G	1/8	0.89
	c.779A>G	779		A/G	2/7	0.78
	c.862A>G	862		A/G	2/7	0.78
	c.875A>G	875		A/G	7/2	0.78
	c.983A>G	983		A/G	2/7	0.78
	c.1056A>T	1056		A/T	7/2	0.78
P. abelii	c.231_236GCTTTT-del	231_236	5	delGCTTTT	2/3	0.6
	c.658C>T	658		C/T	4/1	0.8
	c.691C>T	691		C/T	4/1	0.8
	c.734A>G	734		A/G	1/4	0.8

Table A5 (continuation): Nucleotide diversity found in *ATXN3L2*.

P. anubis	c.49_101indel	49_101	7	indel	1/6	0.86
	c.102T>G	102		T/G	4/3	0.57
	c.373T>G	373		T/G	6/1	0.86
	c.556T>G	556		T/G	6/1	0.86
	c.987A>G	987		A/G	1/6	0.86
	c.1049T>G	1049		T/G	2/5	0.71
M. mulatta	c.29_31AAG-del	28_31	15	delAAG	1/14	0.93
	c.36 A>G	36		A/G	1/14	0.93
	c.73A>G	73		A/G	7/8	0.53
	c.261C>T	261		C/T	1/14	0.93
	c.267A>G	267		A/G	14/1	0.93
	c.341A>C	341		A/C	10/5	0.67
	c.420A>G	420		A/G	1/14	0.93
	c.680A>G	680		A/G	1/14	0.93
	c.740C>G	740		C/G	14/1	0.93
	c.795A>G	795		A/G	14/1	0.93
	c.864A>G	864		A/G	10/5	0.67
	c.1015C>T	1015		C/T	6/9	0.6
M. fuscata	c.575A>G	575	30	A/G	25/5	0.83
	c.579A>G	579		A/G	4/26	0.87
	c.581C>T	581		C/T	29/1	0.97
M. fascicularis	c.73A>G	73	8	A/G	4/4	0.63
	c.134A>G	134		A/G	3/5	0.63
	c.180A>C	180		A/C	1/7	0.88
	c.341A>C	341		A/C	4/4	0.63
	c.471C>T	471		C/T	3/5	0.63
	c.495A>G	495		A/G	5/3	0.63
	c.507A>G	507		A/G	4/4	0.63
	c.549A>G	549		A/G	5/3	0.63
	c.627T>G	627		T/G	3/5	0.63
	c.821C>T	821		C/T	4/4	0.63
	c.864A>G	864		A/G	4/4	0.63
	c.1048A>G	1048		A/G	1/7	0.88

A – adenine; C – cytosine; G – guanine; T – timine; MAF - maximum allele frequency.