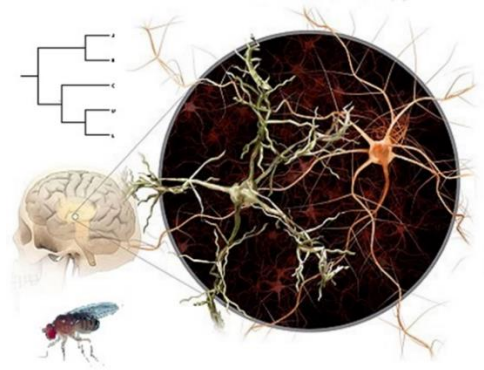


An evolutionary approach to understand how interactomes change when expansions of the polyglutamine (polyQ) tracts occur: implications for the human polyQ diseases



André Manuel Dias Sousa

Molecular and Cell Biology

Department of Biology – Faculty of Science of the University of Porto (FCUP)
2019

Supervisor

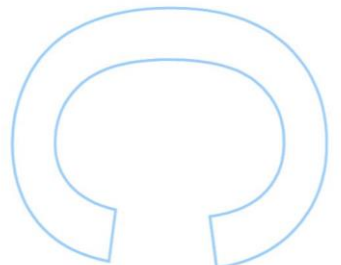
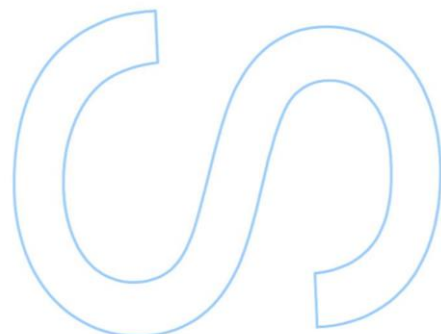
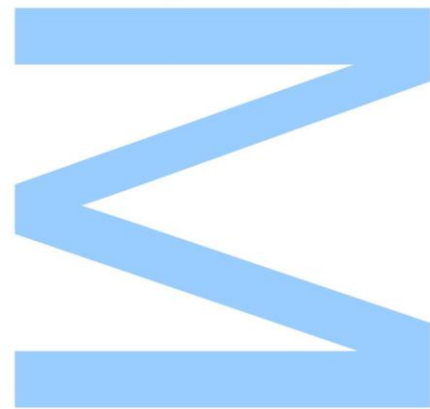
Cristina Vieira, Assistant Researcher, Instituto de Investigação e Inovação em Saúde (I3S); Instituto de Biologia Molecular e Celular (IBMC)

Co-Supervisor

Jorge Vieira, Principal Researcher, I3S/IBMC

Co-Supervisor

Sara Rocha, Post-Doc Researcher, I3S/IBMC

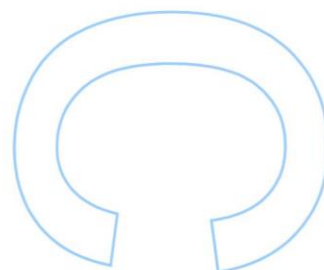
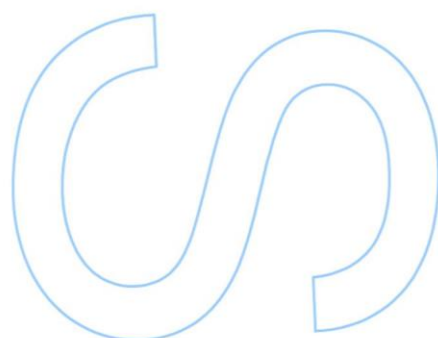
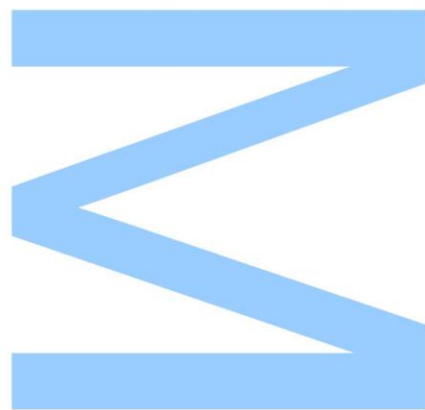




Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____ / ____ / ____



Author: André Manuel Dias Sousa

Institution Address: Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

E-mail: andre.sousa@ibmc.up.pt

Telephone: 933275016

Supervisor: Cristina Alexandra Gonçalves Paula Vieira; Assistant Researcher

Co-Supervisor: Jorge Manuel de Sousa Basto Vieira; Principal Researcher

Co- Supervisor: Sara Maria de Jesus Freitas Rocha; Post-Doc Researcher

Institution Address: Instituto de Investigação e Inovação em Saúde (I3S), Rua Alfredo Allen, 208 4200-135 Porto, Portugal

E-mail: cgvieira@ibmc.up.pt

Telephone: +351 220 408 800

Plagiarism declaration:

I, André Manuel Dias Sousa, student number 201208669, currently undertaking a Cell and Molecular Master's Degree, edition 2018/2019, hereby declare that this thesis was written by myself and using my own words, not contemplating quotations from published and unpublished sources indicated and acknowledged as such. I am aware of any consequence as a result of plagiarism.

Porto, 27 September 2019

Agradecimentos

A realização da presente dissertação seria-me impossível sem a contribuição de diversas pessoas durante este percurso, demonstrando, desde já, a minha gratidão. Desde o início desta dissertação, contei com a confiança, suporte e o apoio de inúmeras pessoas.

À Professora Doutora Cristina Vieira, orientadora desta dissertação, agradeço toda a formação, disponibilidade, suporte e paciência na concretização deste trabalho. Ao Professor Doutor Jorge Vieira, co-orientador, agradeço também toda a formação, suporte e disponibilidade demonstrado ao longo deste trabalho. À Doutora Sara Rocha, co-orientadora, agradeço também toda a formação, apoio e paciência na realização deste trabalho. A larga experiência profissional de todos contribuiu para ajudar-me a concretizar este trabalho. Agradeço a todos a oportunidade de trabalhar neste laboratório e das competências dadas que me ajudaram no meu desenvolvimento profissional e pessoal.

Não poderia deixar de agradecer a todos os meus colegas do laboratório, a quem considero amigos, no apoio e contribuição na realização deste trabalho.

Por fim, gostaria de demonstrar a minha gratidão à minha família, em especial, ao meu pai, à minha irmã e à minha mãe, que através do seu suporte e apoio, contribuíram para a minha formação e concretização deste trabalho. Agradeço também a todos os meus amigos pelo apoio, que me ajudou na realização deste trabalho.

Resumo

As doenças associadas com os tratos de poliglutamina (poliQ) são causadas pela expansão irregular destes tratos, que contêm repetições do aminoácido glutamina, em nove proteínas não relacionadas. Estas proteínas incluem o recetor de androgénio (AR; associado à atrofia muscular espinhal e bulbar ligada ao cromossoma X (SBMA), atropina-1 (ATN1; atrofia dentato-rubro-palido-luisiana (DRPLA)), Ataxina 1 (ATXN1; ataxia espinocerebelosa (SCA) tipo 1 (SCA1), Ataxina 2 (ATXN2; SCA2), Ataxina 3 (ATXN3; doença de Machado – Joseph (MJD ou SCA3)), Ataxina 7 (ATXN7; SCA7), a subunidade alfa1A do canal de cálcio (CACNA1A; SCA6), Huntingtin (HTT; doença de Huntington (HD)) e a proteína TATA (TBP: SCA17). A toxicidade da proteína poliQ expandida tem sido sugerida como associada a um ganho ou uma perda patológica da função da proteína. Esta patogenicidade pode ser definida por alterações irregulares na proteína, mais precisamente, na conformação estrutural da proteína poliQ na forma expandida, que implica diferentes locais de interação, incluindo locais em que ocorrem modificações pós-traducionais específicas, regiões de ligação ao RNA ou chaperones, necessárias para a atividade proteica normal. Abordagens *in silico* e *in vitro* foram, assim, usadas para perceber as possíveis modificações na interação proteica quando as expansões nas regiões poliQ ocorrem. Na abordagem *in silico*, defini uma metodologia baseada no número de resíduos envolvidos nas interações proteicas reportadas para a proteína ATXN1. Primeiramente, a inferência da estrutura 3D da proteína foi efetuada pelo I-TASSER. De seguida, a previsão da ancoragem molecular da proteína ATXN1 e os seus interactores foi efetuada usando o HADDOCK. Por último, a ferramenta PISAePDB foi usada para analisar e comparar as formas wt e expandida, ao nível residual, pra se verificar a estabilidade das interações proteicas. Tudo isto foi validado usando dados experimentais reportados. Os resultados obtidos para a proteína ATXN1 corroboram a hipótese de que a expansão do trato poliQ modifica a conformação proteica da proteína ATXN1, aumentando a afinidade da interação com os seus interactores. Este aumento do número de resíduos a interagir ocorre sobretudo na região N-terminal do trato poliQ na forma expandida. A associação deste padrão a outras proteínas poliQ, incluindo sete proteins relacionadas com doenças neurodegenerativas (ATXN1 inclusive) e 11 proteínas poliQ não relacionadas com esse tipo de doenças, foi também analisado. Os resultados obtidos demonstram, por comparação, um maior número de interações com maior afinidade na forma expandida, embora sem significância (53% versus 47%), indicando que o padrão observado para a

proteína ATXN1 não é generalizado para as outras proteínas poliQ. De facto, entre as proteínas relacionadas com doenças neurodegenerativas, a maior afinidade com a forma expandida não sucede em todas as proteínas. Os resultados apontam também que, contrariamente ao que é observado especificamente para a ATXN1, na generalidade das proteínas poliQ, a afinidade da interação nas formas wt e expandida está associada à região C-terminal do trato polyQ, sendo a região onde a variação é significativa quer no ganho ou quer na perda de resíduos a interagir. Na abordagem *in vitro* foi usada a técnica de co-imunoprecipitação com *Drosophila melanogaster* transgénica, codificando a proteína humana ATXN1, nas formas wt (30Q) e expandida (82Q) do trato poliQ. A maioria das proteínas identificadas interage com os interactores da proteína ATXN1 (Segundo nível de interação), tornando impossível analisar as diferenças na interação com as duas formas de ATXN1, wt e expandida. No entanto, esses dados não só validaram as interações reportadas, mas também foram usadas para identificar novas interações para a proteína ATXN1 humana. Novos putativos interactores para o homólogo putativo de *Drosophila* de ATXN1 (dAtx1) foram também identificados. Também foi usada uma abordagem evolutiva para analisar o número de genes poliQ e o tamanho do trato poliQ em genomas de espécies que pertencem ao reino Animalia. Os resultados demonstram que em grande parte dos grupos de animais, o número de genes e o tamanho do trato são relativamente constantes, embora com algumas exceções em alguns grupos, nomeadamente, em Insecta (com ênfase no grupo *Drosophila*), que demonstram elevadas variações no número de genes que codificam proteínas poliQ. Essas descobertas são discutidas também no contexto *in vitro* deste trabalho.

Palavras-chave

Neurodegeneração; Proteína; Interação; Evolução; Animalia; *In silico*; *in vitro*

Abstract

Polyglutamine (PolyQ) diseases are a group of nine hereditary neurodegenerative diseases, caused by the expansion of the repetitive polyQ tracts in nine unrelated proteins, namely androgen receptor (AR; associated with spinal and bulbar muscular atrophy X-linked 1 (SBMA), Atrophin-1 (ATN1; dentatorubral pallidoluysian atrophy (DRPLA)), Ataxin 1 (ATXN1; spinocerebellar ataxia (SCA) types 1 (SCA1), Ataxin 2 (ATXN2; SCA2), Ataxin 3 (ATXN3; Machado–Joseph disease (MJD or SCA3)), Ataxin 7 (ATXN7; SCA7), calcium voltage-gated channel subunit alpha1 A (CACNA1A; SCA6), Huntingtin (HTT; Huntington’s disease (HT)), and TATA-binding protein (TBP: SCA17). It is known that the polyQ tracts influence protein-protein-interactions (PPI). The toxicity of the expanded polyQ protein has been suggested to be both a gain of function and a loss of function of the normal protein. The pathology could be determined by alterations of the conformation of the expanded polyQ protein that imply different accessibility at specific post-translational modifications, RNA binding regions, or chaperone binding regions needed for the normal protein activity. To understand how PPI change when polyQ regions are expanded, I used both *in silico* and *in vitro* approaches. In the *in silico* approach I first establish an methodology, based on the binding surfaces of wt and expanded ATXN1 with the reported interactors (by using I-TASSER to infer the predicted 3D-structure, HADDOCK to infer docking of ATXN1 and their interactors, and then the PISAePDB web tool to infer the stability of the interactions and the regions of the expanded and wild-type polyQ proteins that are involved in the PPI), that has been validated using available experimental data. My results for ATXN1 supports that the polyQ expansion alters the ATXN1 conformation and that it enhances the strength of interaction with ATXN1 partners. Moreover, the difference in the interaction strength of the ATXN1 variants was due to an increase in the number of interactions at the N-terminal region, before the polyQ, for the expanded form. To address the generality of these findings I applied the same approach to seven polyQ diseases proteins (including ATXN1), and 11 human polyQ genes not involved in neurodegenerative diseases, using the 18 interactors that are common between two polyQ neurodegenerative diseases and one polyQ protein not involved in this type of disease. My results show that there is a no significant increase in the number of interactions in the expanded form (53 versus 47), implying that preference binding for expanded polyQ protein is not a general feature. Indeed, among neurodegenerative diseases, not all proteins show an increase in the number of interactions with the

expanded form. I observed that the interaction interface variances seen when comparing both forms is associated mostly with the C-terminal region of the polyQ tract. For the *in vitro* approach, I performed co-immunoprecipitation of transgenic *Drosophila melanogaster* expressing the human SCA1 gene (ATXN1) with either a normal and expanded polyQ tract. Most of the obtained proteins interact with proteins that interact with ATXN1 (Second level of interaction), making it impossible to address differences of interaction with the two forms of ATXN1. Nevertheless, this data not only validated the interactions reported but also was used to identify novel interactions of Human ATXN1. Putative novel interactors for the putative *Drosophila* homolog of ATXN1 (dAtx1) were also identified. I also used an evolutionary approach, to analyze polyQ number and size in the Animalia kingdom. I show that in some large animal groups the number of genes encoding polyQ proteins, as well as the average size of the polyQs remains fairly constant, while in some other groups such as the insects (with emphasis in the *Drosophila* group) there is large variations in the number of genes encoding polyQ proteins. Such findings are discussed in the context of the *in vitro* experiments.

Keywords

Neurodegeneration; Protein; Interaction; Evolution; Animalia; *In silico*; *in vitro*

Contents

Introduction	1
Objectives	5
Chapter 1 – ATXN1 N-terminal region explains the binding differences of wild-type and expanded forms	7
Introduction	7
Materials and Methods	11
1. Importance of N-terminal region in the ATXN1 protein on the binding variances of expanded and wild-type forms.	11
Results	16
Discussion	29
Conclusions	32
Chapter 2 – Consequences of polyglutamine tract expansion: are there general patterns?	33
Introduction	33
Materials and Methods	44
1. Selection of the 18 polyQ proteins to be studied	44
2. Selection of the polyQ proteins interactors	45
3. Analyzes of the 3D structure of the selected human polyQ proteins and their interactors	49
Results	51
Discussion	57
Conclusions	60
Chapter 3 – Identification of the interactors for human ATXN1 using <i>D. melanogaster</i>	61
Introduction	61
Materials and Methods	64

1. Flies expressing human ATXN1 wild-type and expanded forms	64
2. Protein extraction	64
3. Co-Immunoprecipitation	65
4. Mass Spectrometry screen and data analyzes	66
5. <i>In silico</i> prediction of the interacting residues.....	66
Results.....	68
Discussion	77
Conclusion	78
Chapter 4 – Evolution of the polyglutamine tract in the Animalia Kingdom.....	81
Introduction	81
Materials and Methods.....	84
1. Sequence data and annotation.....	84
2. Design of cladograms.....	85
3. Statistical analyzes	85
Results.....	86
Discussion	90
Conclusion	92
General Conclusions	93
Bibliography	95
Supplementary	I

Index of Figures

FIGURE 1. FLOWCHART DESCRIBING THE <i>IN SILICO</i> METHODOLOGY USED, WHICH SHOWED THE BEST AGREEMENT WITH SUTER <i>ET AL.</i> (2013) RESULTS.	13
FIGURE 2. GRAPHICAL REPRESENTATION OF THE COMPARISON OF THE NUMBER OF INTERACTORS BETWEEN PROTEINS CONTAINING POLYQ TRACTS IN BOTH <i>H. SAPIENS</i> AND <i>M. MUSCULUS</i> SPECIES (A) AND THE PROTEINS DISPLAYING A POLYQ TRACT IN HUMANS ONLY (B). FROM THIS GRAPHIC, IS POSSIBLE TO SEE A HIGHER NUMBER OF INTERACTORS FOR THE PROTEINS CONTIANING A POLYQ TRACT, COMPARING TO PROTEINS THAT NOT POSSESSES THEM.	17
FIGURE 3. REPRESENTATION OF THE 3D STRUCTURE OF THE POLYQ TRACT (IN CYAN) AND 20 AMINO ACIDS RESIDUES FLANKING THE POLYQ TRACT (IN YELLOW THOSE IN THE N-TERMINAL REGION, AND IN WHITE THOSE IN THE C-TERMINAL REGION) IN: A) WT ATXN1 AND B) EXPANDED ATXN1.	23
FIGURE 4. GRAPHICAL REPRESENTATION OF THE AXH DOMAIN, SHOWING HIGH PERCENTAGE OF THE NUMBER OF INTERFACE RESIDUES IN THIS DOMAIN FROM ALL THE PPI STUDIED, ESPECIALLY IN THREE REGIONS. THE PERCENTAGE OF THE NUMBER OF INTERFACE RESIDUES AT THE AXH DOMAIN WITH THE 43 INTERACTORS IN AGREEMENT WITH SUTER <i>ET AL.</i> (2013) FOR THE WT (IN BLUE) AND EXPANDED (IN RED) ATXN1.	24
FIGURE 5. COMPARISON OF THE WT AND EXPANDED ATXN1 DOCKING RESULTS OF THE 43 INTERACTORS ANALYZED A) INTERFACING RESIDUES OF THE ATXN1. B) PERCENTAGE OF INTERFACING RESIDUES OF THE ATXN1. C) NUMBER OF INTERFACING RESIDUES FOR THE 43 INTERACTORS. D) PERCENTAGE OF INTERFACING RESIDUES OF THE 43 INTERACTORS. E) INTERFACE OF THE SOLVENT ACCESSIBLE AREA (Å) OF THE ATXN1. F) PERCENTAGE OF THE INTERFACE OF THE SOLVENT ACCESSIBLE AREA OF THE ATXN1.	26
FIGURE 6. DIFFERENCE OF THE NUMBER OF INTERFACE RESIDUES OF THE 43 INTERACTORS ANALYZED. PERCENTAGE OF THE DIFFERENCE OF THE NUMBER OF INTERFACE RESIDUES OF THE 43 INTERACTORS IN AGREEMENT WITH SUTER <i>ET AL.</i> (2013), WITH WT AND EXPANDED ATXN1. POLYQ TRACT IS NOT REPRESENTED.	27
FIGURE 7. FLOW-CHART DESCRIBING THE PROCESS USED TO SELECT THE 18 POLYQ PROTEINS STUDIED.	45
FIGURE 8. FLOW-CHART USED TO SELECT THE INTERACTORS FOR THE 18 COMMON INTERACTORS FOR THE POLYQ PROTEINS HERE STUDIED.....	46
FIGURE 9. GRAPHICAL REPRESENTATION OF THE GENE ONTOLOGY ANALYZES MADE TO POLYQ PROTEINS (WHITE BARS) AND THEIR INTERACTORS (GREY BARS), WITH FOUR MAJOR FUNCTIONS, INCLUDING TRANSCRIPTION AND TRANSLATION REGULATION, BINDING AND CATALYTIC ACTIVITY.	49
FIGURE 10. MOLECULAR FUNCTION OF THE INTERACTORS (PANTHERDB) OF THE 391 POSSIBLE INTERACTORS OF THE HUMAN ATAXIN-1 PROTEIN OBTAINED BY CO-IMMUNOPRECIPITATION AND	

MASS SPECTROMETRY ASSAYS (ONLY ORTHOLOGUES EXAMINATION WITH BEST SCORE WERE USED).....	68
FIGURE 11. COMPARISON OF THE INTERACTORS DETECTED FROM THE MUTANT THE <i>DROSOPHILA</i> CO-IMMUNOPRECIPITATION AND MASS SPECTROMETRY ASSAYS AND THE INTERACTIONS REPORTED IN EvoPPI FOR ATXN1 THAT PRESENT A <i>DROSOPHILA</i> ORTHOLOGOUS.....	69
FIGURE 12. MOLECULAR FUNCTION (ACCORDING TO PANTHERDB) FOR THE 143 ATXN1 INTERACTIONS INFERRED BY ANALYSES OF THE PROTEINS OF THE SECOND LEVEL OF INTERACTION.	75
FIGURE 13. MOLECULAR FUNCTION (ACCORDING TO PANTHERDB) OF THE 229 POTENTIAL NEW INTERACTIONS INFERRED BY ANALYSES OF THE PROTEINS OF THE SECOND LEVEL OF INTERACTION.	76
FIGURE 14. GENERAL CLADOGRAM SHOWING A HIGH VARIANCE IN THE NUMBER AND COMPOSITION OF THE POLYQ GENES ACROSS THE ANIMALIA KINGDOM. IN THIS CLADOGRAM IS REPRESENTED THE NON-BILATERIA GROUP AND THE BILATERIAN GROUP, WITH PROTOSTOMIANS AND DEUTEROSTOMIANS SUBGROUPS. THE GROUPS INSECTA ¹ , ARACHNIDA ² , NEMATODA ³ , GASTROPODA ⁴ , ACTINOPTERI ⁵ , REPTILES ⁶ , AVES ⁷ AND MAMMALIA ⁸ DISPLAYED A HIGH VARIANCE IN THE GROUP, AND, THEREFORE, WHERE DEEPER ANALYZED. THE RELANTIOSHIPS AMONG THE DIFFERENT SPECIES WERE MADE ACCORDINGLY TO (TELFORD <i>ET AL.</i> , 2015) AND (HENRIQUES <i>ET AL.</i> , 2019).	86
FIGURE 15. ANIMALIA KINGDOM DISTRIBUTION OF THE FRACTION OF POLYQ GENES IN THE GENOME (A); NUMBER OF POLYQ TRACTS PER POLYQ GENE (B); AVERAGE SIZE OF THE POLYQ TRACTS (C); THE POLYQ TRACT WITH A MAXIMUM SIZE (D); AND THE POLYQ TRACT WITH THE LOWER SIZE (E).	89

Index of Tables

TABLE 1. THE POLYQ DISORDERS: CHARACTERISTICS OF WILD-TYPE AND EXPANDED POLYQ PROTEINS, INCLUDING GENE ID AND UNIPROT ID REFERENCES, POLYQ TRACT SIZE (NORMAL AND PATHOLOGIC), LOCATION, PROTEIN LENGTH AND POLYQ TRACT REFERENCE SIZE FROM UNIPROT DATABASE (ADAPTED FROM: (TODD AND LIM, 2013)).	2
TABLE 2. PPI WITH THE WT AND EXPANDED ATXN1 ACCORDING TO THE METHODOLOGIES USED. 1) THE MOST RELIABLE STRUCTURE ACCORDING TO HADDOCK, 2) DOCKING STRUCTURE THAT MAXIMIZES THE PROBABILITY OF HAVING A PPI.	19
TABLE 3. PREDICTED (APPROACH 2) AND DETERMINED (SUTER ET AL., 2013) BINDING PREFERENCES OF ATXN1 PARTNERS WITH THE WT AND EXPANDED ATXN1. IN BOLD ARE THOSE PROTEINS FOR WHICH THE PREDICTION IS DIFFERENT WHEN USING THE TWO IN-SILICO METHODOLOGIES.	21
TABLE 4. NUMBER OF INTERFACE RESIDUE (%) FOR THE PPI IN SIX REGIONS: BEFORE AND AFTER THE POLYQ TRACTS (INCLUDING THE 20 AMINO ACIDS RESIDUES FLANKING THE POLYQ TRACT), THE POLYQ TRACTS AND THE AXH DOMAIN.	27
TABLE 5. ATXN1 N-TERMINAL REGION EUKARYOTIC LINEAR MOTIFS (ELM.EU.ORG).	28
TABLE 6. THE POLYQ PROTEINS AND INTERACTORS HERE STUDIED. IN BRACKETS ARE THE PROTEIN NAME, GENEID, AND UNIPROTID. UNDERLINED ARE THE POLYQ PROTEINS ASSOCIATED WITH NEURODEGENERATIVE DISEASES.	46
TABLE 7. POLYQ PROTEINS INTERACTIONS WITH BINDING PREFERENCE FOR THE WT AND EXPANDED FORMS. UNDERLINED ARE POLYQ PROTEINS ASSOCIATED WITH NEURODEGENERATIVE DISEASES.	52
TABLE 8. PERCENTAGE OF INTERACTIONS SHOWING BINDING PREFERENCES FOR WT AND EXPANDED FOR THE NINE POLYQ PROTEINS WITH THREE OF MORE INTERACTIONS STUDIED. UNDERLINE ARE POLYQ PROTEINS ASSOCIATED WITH NEURODEGENERATIVE DISEASES	55
TABLE 9. IDENTIFIED GENES AND PATHWAYS BY LARGE-SCALE GENETIC SCREENS OF <i>D. MELANOGASTER</i> EXPRESSING FULL-LENGTH HUMAN ATXN1 GENE.	62
TABLE 10. <i>D. MELANOGASTER</i> STRAINS USED IN THIS STUDY. THE BLOOMINGTON <i>D. MELANOGASTER</i> STOCK CENTER NUMBER, THE TRANSGENE AND THEIR CHROMOSOME LOCALIZATION AS WELL AS THE GENDER USED IN THIS STUDY ARE REPORTED.	64
TABLE 11. ATXN1 INTERACTORS FOUND IN THE CO-IMMUNOPRECIPITATION AND MASS SPECTROMETRY ASSAYS COMMON IN <i>DROSOPHILA</i> AND HUMANS, REPORTED IN EVO PPI FOR THE HUMAN ATXN1 PROTEIN.	69
TABLE 12. THE LOCATION OF THE BINDING RESIDUES (N-TERMINAL, POLYQ, AND C-TERMINAL) BETWEEN THE HUMAN ATXN1 AND THE SEVEN <i>DROSOPHILA</i> PROTEINS AND THERE HUMAN ORTHOLOGOUS, INFERRED BY THE <i>IN SILICO</i> APPROACH.	70

TABLE 13. ATXN1 SECOND LEVEL INTERACTORS FOUND IN THE CO-IMMUNOPRECIPITATION AND MASS SPECTROMETRY ASSAYS, COMMON TO THE ATXN1 REPORTED SECOND LEVEL INTERACTION (EXCEPT ATXN1 AND UBC INTERACTORS).....	71
---	----

Abbreviations

ABCF1 - ATP-binding cassette sub-family F member 1
AES - Amino-terminal enhancer of split
AR – Androgen receptor
ATP1A3 - Sodium/potassium-transporting ATPase subunit alpha-3
ATXN1 - Ataxin-1 protein
ATXN1L - Ataxin-1-like
ATXN2 – Ataxin-2 protein
ATXN3 – Ataxin-3 protein
ATXN7 – Ataxin-7 protein
BMI1 - Polycomb complex protein BMI-1
CACNA1A - Calcium Channel Voltage-dependent P/Q Type 1 Alpha subunit
CAG – Cytosine-Adenine-Guanine
CASP3 - Caspase-3
CC – Coiled-coil
CCAR1 - cell cycle and apoptosis regulator 1
CDK9 - Cyclin-dependent kinase 9
CDS – Coding Sequence
CELF3 - CUGBP Elav-like family member 3
CIC - Capicua
CNS - central nerve system
DDX5 - Probable ATP-dependent RNA helicase DDX5
DJBP - DJ-1-binding protein
DLG4 - Disks large homolog 4
DRPLA - Dentatorubral-pallidoluysian atrophy disorder
ERCC3 - p53-excision repair cross complementing factor 3
EvoPPI - Evolution Protein-Protein Interaction web tool
EWSR1 - RNA-binding protein EWS
FAR1 - Fatty acyl-CoA reductase 1
GAPDH - Glyceraldehyde-3-phosphate dehydrogenase II
GST - glutathione-S-transferase
HADDOCK - High Ambiguity Driven protein-protein DOCKing
HD - Huntington's disease
HDAC1 - Histone deacetylase 1

HDAC3 - Histone deacetylase 3
HDAC4 - Histone deacetylase 4
HIP1 - Huntingtin interacting adaptor protein 1
HIST1H3D - Histone H3.1 subunit d
HNF4A - Hepatocyte nuclear factor 4-alpha
HOT - HOX transcript antisense RNA
HSPA8 - Heat Shock Protein Family
HTT - Huntington protein
IMMT - MICOS complex subunit MIC60
I-TASSER - Iterative Threading ASSEmbly Refinement
KAT2B - Histone acetyltransferase KAT2B
KAT7 - lysine acetyltransferase 7
LIM - Four and a half LIM domains protein 2
LPXN - leupaxin
MAGEA11 - melanoma-associated antigen 11
MAK - male germ cell-associated kinase
MEF2A - Myocyte-specific enhancer factor 2A
MJD - Machado– Joseph disease
MRF1 - modulator recognition factor 1
NCBI - National Center for Biotechnology Information database
NCOA - nuclear receptor coactivator
NES - nuclear export signal
NLS - Nuclear Localization Signal
NUMB - Numb-like protein
PDB - protein data bank file
PHC1 - Polyhomeotic-like protein 1
PISAePDB - Proteins, Interfaces, Structures and Assemblies
PML - Protein PML
PolyQ - polyglutamine
POU6F2 - POU domain, class 6, transcription factor 2
PPI - Protein-protein interaction
PSMC3 - 26S proteasome regulatory subunit 6A
PSMC5 - 26S proteasome regulatory subunit 8
Ran protein - Ras-related nuclear protein
RBFOX1 - RNA binding protein fox-1 homolog 1

RBM17 - RNA-binding motif protein 17
RBPJ - Recombining binding protein suppressor of hairless
RBPMS - RNA-binding protein with multiple splicing
RUNX2 - Runt-related transcription factor 2
SAR - self-association region
SATB1 - DNA binding protein SATB1
SBMA - Spinal-Bulbar Muscular Atrophy disorder
SCA1 - Spinocerebellar Ataxia 1
SCA17 - Spinocerebellar Ataxia 17
SCA2 - Spinocerebellar Ataxia 2
SCA3 - Spinocerebellar Ataxia 3
SCA6 - Spinocerebellar Ataxia 6
SCA7 - Spinocerebellar Ataxia 7
SEDA - SEquence DATaset builder
SPEDF - SAM pointed domain-containing Ets transcription factor
SUMO-1 - small ubiquitin-like modifier
TBP - TATA-binding protein
TFCP2 - Alpha-globin transcription factor CP2
TFIID - DNAbinding subunit transcription factor II D
TFTC - TATA-binding protein associated factor-containing complex
THAP11 - THAP domain containing protein 11
UBC – Ubiquitin C
VEZF1 - Vascular endothelial zinc finger 1
wt – wild-type
XRCC6 - X-ray repair cross-complementing protein 6
Y2H - Yeast two-hybrid method
YWHAE - 14-3-3 protein epsilon

Introduction

Neurodegeneration is characteristic of many debilitating, incurable diseases, and thus, there is an urgent need to understand the mechanism of pathogenesis of these devastating diseases, to find new therapeutic strategies (Gitler *et al.*, 2017). There is a wide variety of neurodegeneration disorders such as the repeat expansion disorders, that includes polyglutamines and polyalanine diseases (Heemels, 2016). Polyglutamine (polyQ) expansion diseases are a family of nine dominantly transmitted neurodegenerative disorders caused by a pathogenic expansion of CAG (Cytosine-Adenine-Guanine) trinucleotide repeats in the corresponding coding region of the gene associated with each of the nine neurodegenerative disorders. This family of polyQ diseases (Table 1) includes Huntington's disease (HD), spinal bulbar muscular atrophy X-linked type 1 (SMA1/SBMA) and dentatorubral pallidoluysian atrophy (DRPLA), as well as spinocerebellar ataxias (SCA types 1, 2, 6, 7 and 17) and Machado-Joseph disease (MJD/SCA3) (Gardian *et al.*, 2005). This particular group of diseases share an autosomal inheritance, except SBMA, which is linked to a mutation in the androgen receptor gene located on the X chromosome (Zoghbi and Orr, 2000). First description of these diseases goes back to 1991, with the identification of spinal bulbar muscular atrophy X-linked type 1 (SBMA/SMA1) disease, which is the only heterosomal polyQ disease. Although having a simple genetic basis, the origin and mechanism of pathogenesis of these diseases is far from clear (Xu *et al.*, 2015). Details of each polyQ disease are described in Table 1.

Table 1. The polyQ disorders: characteristics of wild-type and expanded polyQ proteins, including Gene ID and Uniprot ID references, polyQ tract size (normal and pathologic), location, protein length and polyQ tract reference size from Uniprot database (Adapted from: (Todd and Lim, 2013)).

Protein	Disease	Gene ID	Uniprot ID	Normal polyQ length	Pathologic polyQ length	Location of the polyQ tract (polyQ2.0 (Li et al,2016))	Length of canonical protein (Uniprot)	PolyQ length canonical protein (Uniprot)
AR (Androgen receptor)	Kennedy's Disease/spinal and bulbar muscular atrophy (SBMA)	367	P10275	9–36	38–62	Mid domain PolyQs	920	23
ATN1 (Atrophin-1)	Dentatorubralp allidoluysian atrophy (DRPLA)	1822	P54259	7–34	49–88	Mid domain PolyQs	1190	19
ATXN1 (Ataxin 1)	Spinocerebellar ataxias type1 (SCA1)	6310	P54253	6–44	39–82	N-terminal PolyQs	815	29
ATXN2 (Ataxin 2)	Spinocerebellar ataxias type2 (SCA2)	6311	Q99700	15–24	32–200	N-terminal PolyQs	1312	22
ATXN3 (Ataxin 3)	Machado-Joseph Disease (MJD/SCA3)	4287	P54252	13–36	61–84	Mid domain PolyQs	364	14
Ataxin 7 (ATXN7)	Spinocerebellar ataxias type7 (SCA7)	6314	O15265	4–35	37–306	N-terminal PolyQs	892	10
CACNA1A (Calcium voltage-gated channel subunit alpha1 A)	Spinocerebellar ataxias type6 (SCA6)	773	280	4–19	10–33	C-terminal PolyQs	2506	13
HTT (Huntingtin)	Huntington's Disease (HD)	3064	348	6–34	36–121	N-terminal PolyQs	3144	23
TBP (TATA-binding protein)	Spinocerebellar ataxias type17 (SCA17)	6908	P20226	25–42	47–63	Mid domain PolyQs	338	41

Although they are caused by unrelated proteins, polyQ disorders share similarities such as mutational mechanism, progressive neurodegeneration in specific neuronal populations, formation of protein aggregates, negative correlation between length of CAG repeats and age of onset, and positive correlation with the disease severity (Fan *et al.*, 2014). The phenomenon anticipation is also a prominent feature of all polyQ diseases, with the expanded CAG tract being unstable, which tends to expand further, leading to an earlier age of onset and a higher severity disease symptoms in late generations (Bauer and Nukina, 2009).

During protein synthesis, the expanded CAG repeats are translated into a series of uninterrupted glutamine residues, forming an abnormal polyQ tract and leading to an accumulation of deficient polyQ proteins, that in the case of the nine neurodegenerative diseases, causes neurotoxicity (Fan *et al.*, 2014). The neurotoxicity is the result of the aberrant protein interactions (Harjes and Wanker, 2003, MacDonald, 2003, Lam *et al.*, 2006, Lim *et al.*, 2006, Nedelsky *et al.*, 2010, Ratovitski *et al.*, 2012, Suter *et al.*, 2013, Hosp *et al.*, 2015). It is known that polyQ expansion mediates pathology mainly by modulating the native protein interactions of the disease protein, and not with novel complexes formed by misfolded or aggregating proteins (Ratovitski *et al.*, 2012, Suter *et al.*, 2013, Takeuchi and Nagai, 2017). Therefore, polyQ diseases are determined by transient and stable protein interactions, that change due to the conformational alterations of the expanded polyQ protein. Since these nine proteins are unrelated, it is not surprising, that when their interactors are compared, only 15 proteins are common to at least four of these polyQ disease proteins (Vázquez *et al.*, 2019). Furthermore, 40% of these 15 proteins are involved in ubiquitin protein ligase binding function. Therefore, the polyQ disease proteins are involved in different functional networks. PolyQ expansion causes a gain of abnormal interactions of the native wild-type (wt) protein network, that results in the identification of more proteins in quantitative PPI detection methodologies when comparing with the wt polyQ protein (for ATXN1 see (Lim *et al.*, 2006, Suter *et al.*, 2013, Hosp *et al.*, 2015); for HTT see (Ratovitski *et al.*, 2012, Hosp *et al.*, 2015); for AR see (Nedelsky *et al.*, 2010)). Nevertheless, loss of function can cause toxicity in these diseases (Kratte and Finkbeiner, 2010). Determining how polyQ expansions alters the protein interactions is thus, fundamental to understand polyQ diseases. The wt protein may present different conformations, and thus different functions, depending on post-translation modifications. The expanded polyQ may modify wt conformations causing loss of function. Furthermore, expanded polyQ proteins may disrupt the cell's protein folding, contributing to loss of function. At present, the suggestion is that polyQ modulates protein-protein interaction (PPI) by expanding a neighboring coiled-coil (CC) domain (that consists of parallel or antiparallel α -helices that further fold like the strands of a rope), upon interaction with a CC from a partner protein, since there is a statistically significant presence of predicted CC regions next to polyQ region (Schaefer *et al.*, 2012, Petrakis *et al.*, 2013). Nevertheless, regions other than the neighboring polyQ region could also affect the PPIs in expanded polyQ proteins. Indeed, for the ATXN1, phosphorylation at the serine 886 is necessary for toxicity (Lim *et al.*, 2008). Therefore, all evidence suggests that the polyQ action is to stabilize PPI and/or spacer elements

between individual folded domains in molecules that mediate PPI (Karlin and Burge, 1996, Schaefer *et al.*, 2012). The underlying mechanisms of nuclear accumulation of expanded polyQ proteins in these polyQ diseases include gain-of-toxic effects, affecting gene expression, or disrupting nuclear organization and function (Fan *et al.*, 2014). As a consequence, these aggregated polyQ proteins are found in the degenerated neurons, such as in the cerebellum, brainstem, and spinal tract (Zoghbi and Orr, 1999).

Besides the association with disease, polyQ repeats are the most common tandem repeats of single amino acids that are widely observed in eukaryote genomes (Golding, 1999, Huntley and Golding, 2000, Huntley and Clark, 2007). In human and metazoan proteomes, since a stretch of five consecutive glutamines is not a random feature, it can be considered a polyQ (Lobanov *et al.*, 2016). Nevertheless, to account for the functional and structural implications of a polyQ, a stretch longer than nine must be considered (Schaefer *et al.*, 2012, Totzeck *et al.*, 2017). PolyQs are not characteristic of certain gene families, and they can occur at N-terminal or C-terminal positions of the sequence. They may have a possible biological function. Supporting this hypothesis is the fact that, in untranslated regions, CAG repeats are found in low number. Among the 1825 protein complexes formed in the human cell, 149 are defined to have at least one polyQ protein, suggesting an involvement of these proteins in the protein interaction network. The polyQ tract modulates protein interactions since they form a flexible spacer between protein domains (Mitchell and Tjian, 1989, Karlin and Burge, 1996). PolyQ expansion also affects mRNA stability and structure, and their abnormal expansion in disease can influence splicing (Broda *et al.*, 2005, Neueder *et al.*, 2017).

Objectives

In the study, I have five major goals:

1- To develop an *in silico* approach, based on the inferred binding surfaces of wt and expanded ATXN1 with the reported interactors. To validate this methodology, I will use the available results of ATXN1 *in vitro* experiments (Lim *et al.*, 2006, Suter *et al.*, 2013, Hosp *et al.*, 2015) (Chapter 1)

2- Understand the implications of expansion of the polyglutamine tract of ATXN1 protein, namely, how it affects PPI. The aim is to identify essential regions such as the AXH domain, that have been described as essential for ATXN1 PPI, as well as other regions determining the strength of the PPI with the ATXN1 variants. Moreover, for these regions I will address ATXN1 post-transcriptional modifications (Chapter 1).

3- Verify the generality of the results obtained for ATXN1 by applying the same methodology to other human polyQ proteins. Therefore, I will address the role of the polyQ region, by analyzing differences in the binding affinity of the wt and expanded forms in proteins involved and not involved in neurodegenerative diseases (Chapter 2).

4- Identify putative new interactors for the human ATXN1, by performing co-immunoprecipitation in mutant flies with wt and expanded human ATXN1. This experiment will reveal the importance of the domain conservation, such as the AXH domain in the ATXN1 C-terminal region. It should be noted that the *Drosophila* protein orthologous of the human ATXN1 only has the C-terminal region. Since only three interactors are described for Atx-1, new interactors will be likely identified (Chapter 3).

5- To understand the selection pressure acting on polyQ tracts, I will analyze the evolution of the polyQ proteins in the Animalia kingdom (Chapter 4).

Chapter 1 – ATXN1 N-terminal region explains the binding differences of wild-type and expanded forms

The content of this chapter has been published as a scientific article with the following DOI: <https://doi.org/10.1186/s12920-019-0594-4> / PMID: 31655597

Research article | [Open Access](#) | [Open Peer Review](#) | Published: 26 October 2019

ATXN1 N-terminal region explains the binding differences of wild-type and expanded forms

[Sara Rocha](#), [Jorge Vieira](#), [Noé Vázquez](#), [Hugo López-Fernández](#), [Florentino Fdez-Riverola](#), [Miguel Reboiro-Jato](#), [André D. Sousa](#) & [Cristina P. Vieira](#) 

BMC Medical Genomics 12, Article number: 145 (2019) | [Cite this article](#)

Introduction

Ataxia Type 1 (SCA1) is caused by an expansion of a CAG trinucleotide repeat on the coding region of Ataxin-1, that encodes a protein which is mainly nuclear in neurons and cytoplasmic in peripheral cells, and is involved in transcriptional regulation and RNA metabolism. The Ataxin-1 gene is 10,660 bases in length and has nine exons. Only exons eight and nine encode the Ataxin-1 (ATXN1) coding region (the first seven exons encode the 5' untranslated region), and the polyQ stretch is encoded within exon eight. The resulting protein possesses a minimum of 816 amino acids, but varies according to the size of the polyQ tract. The polyQ tract in this protein usually contains 4 to 39 glutamine repeats and its expanded forms vary from 40 to 83 glutamines (Friedrich *et al.*, 2018). The polyQ tract is located in the N-terminal region of the protein (residues 1-566). Two C-terminal motifs, important for the interaction with other proteins have been identified: the AXH domain, amino acids 570–689, responsible for transcriptional repression and RNA-binding activity and the Nuclear Localization Signal (NLS) domain, located at 771–776, that is determinant for Ataxin-1 transport into the nucleus (Klement *et al.*, 1998). The interaction of these regions with certain proteins changes with the length of the polyQ region and/or phosphorylation of Ser-776. This residue is a site of phosphorylation in ATXN1 that, together with the expansion of the polyglutamine tract, increases the interaction of ATXN1 with other proteins, such as the RBM17 (RNA-binding motif protein 17) (Zoghbi and Orr, 2009). It should be noted that, ATXN1 forms at least

two distinct complexes, one with a transcription factor containing HMG-box domain called CIC (Capicua) and another with a RBM17 (RNA-binding motif protein 17, also known as SPF 45). CIC is involved in ErbB signaling and central nerve system (CNS) development (Lee *et al.*, 2002, Lee *et al.*, 2011). The CIC interacts with the AXH domain of ATXN1 and at the N-terminal region, and the reduction of CIC protein level partially rescues SCA1-like phenotype in a mouse model (Fryer *et al.*, 2011). RBM17, is a part of the spliceosome complex regulating mRNA splicing and abundant in Purkinje cell nucleus (Kim *et al.*, 2014). The phosphorylation of the Ser776 of ATXN1 serves a binding platform for RBM17 (Chen *et al.*, 2003, Kim *et al.*, 2014). The first complex is preferentially formed with the wt ATXN1 and the second with expanded ATXN1 (Lim *et al.*, 2008). Both complexes are, however, involved in the SCA1 disease (Orr, 2012). Therefore, protein phosphorylation plays a major role in the function and activity of ATXN1.

Wild-type ATXN1 function includes RNA binding (Yue *et al.*, 2001), to shuttle between the nucleus and cytoplasm (Irwin *et al.*, 2005), and to interact with transcription factors (Okazawa *et al.*, 2002, Tsai *et al.*, 2004, Tsuda *et al.*, 2005, Lam *et al.*, 2006, Serra *et al.*, 2006). The function of Ataxin-1 is also to regulate and modulate protein folding, through interactions with the chaperone and ubiquitin–proteasome pathways (Hands and Wyttenbach, 2010). Therefore, it is not surprising that 311 interactors are reported for ATXN1 in PPI datasets, that report the results of computational and experimental detection methods (Vázquez *et al.*, 2019). The expression of the expanded Ataxin-1 polyQ tract form in the nucleus is reported as being critical for the pathogenesis of SCA1 disease. The expanded protein gains toxic effects due to aggregation and accumulation in this localization resulting in variable degrees of neurodegeneration in the cerebellum, brainstem, and spinocerebellar tracts (Matilla-Duenas *et al.*, 2008). The observations in Neuro-2a cells of the polyQ expanded form of Ataxin-1 (85Q) interacting more with essential nuclear transporters and their cargoes, point to consequential disruptions in nuclear transport pathways. Other relevant interactors, other than CIC and RBM17, include the 14-3-3 family of proteins (that are highly conserved, ubiquitously expressed, that constitute 1% of all soluble protein in the brain and that are involved in neuronal development, signal transduction, vesicular transport, apoptosis, and cell cycle control) which prevents Ataxin-1 dephosphorylation and degradation (Steinacker *et al.*, 2011, Zhang *et al.*, 2018a). Thus, protein phosphorylation, ubiquitination, sumoylation and transglutamination play a major role in the function and activity of ATXN1. Functional

evidence for the role of ubiquitination in SCA1 aggregation came from the treatment with proteasome inhibitors of transfected cells with expanded polyQ ATXN1, that promotes the formation of ATXN1 aggregates. Several enzymes of the ubiquitination system, such as UBC6, CHIP, and UBE3A have been implicated in the modulation of the ATXN1 transcriptional activity, but the ATXN1 regions involved in such regulation are unknown. In sumoylation, a similar enzymatic cascade to ubiquitination is used, and SUMO-1 (small ubiquitin-like modifier) has been detected in affected brain regions in ATXN1 patients. Although 17 ATXN1 consensus sumoylation sites have been identified, only five (K¹⁶, K¹⁹⁴, K⁶¹⁰, K⁶⁹⁷, and K⁷⁴⁶) have been implicated in decreased sumoylation of ATXN1 (Ju *et al.*, 2014). Therefore, protein phosphorylation, ubiquitination, sumoylation and transglutamination play a major role in the function and activity of ATXN1. Indeed ATXN1 presents typical eukaryotic linear motifs for all these processes (de Chiara and Pastore, 2014).

By performing yeast two-hybrid (Y2H) screens for proteins involved in inherited ataxias, and by validating the results using randomly sampled interacting pairs using co-affinity purification co-AP GST (glutathione-S-transferase) pull-down assays in human HEK293T cells, 165 ATXN1 PPI have been identified (Lim *et al.*, 2006). Also, using human HEK293T cells and quantitative affinity purification and mass spectrometry, Hosp *et al.* (2015) identified 54 proteins that interact with ATXN1, that were not reported by Lim *et al.* (2006), and not included in the main EvoPPI databases. These authors observed that approximately 80% of the interaction partners were shared between the wt and the expanded ATXN1, suggesting that polyQ expansion does not dramatically change the interaction partners of ATXN1. Nevertheless, Suter *et al.* (2013) using ATXN1Q32 and ATXN1Q79 Y2H interaction screens and DNA microarrays for high-throughput quantitative PPI detection, reported only seven out of 81 proteins in common in the two experiments. Changes in PPI are observed in SCA1 (Faux *et al.*, 2005, Schaefer *et al.*, 2012, Pelassa and Fiumara, 2015), and thus, comparison of PPI networks in controls and patients have been used to get insights into the basis of this disease (Lim *et al.*, 2006, Suter *et al.*, 2013). In these studies, the ATXN1 interactors bind with higher affinity to the expanded than the wt form. Nevertheless, Hosp *et al.* (2015) were not able to validate this finding, raising the issue of whether the previous results can be explained based on different binding affinities. The human paralog of ATXN1, ATXN1L (Ataxin-1-like), does not encode a protein with a polyQ tract, shares with ATXN1 conserved domains (the NBA at the N-terminal region, the SAR and AXH),

several interactors, and their expression profiles are very similar (Mizutani *et al.*, 2005). Indeed, ATXN1 and ATXN1L seem to be functionally redundant, since, in flies and mice, increased ATXN1L levels induce the sequestration of expanded ATXN1 into nuclear inclusions, possibly by replacing ATXN1 from the endogenous complexes containing CIC (Lee *et al.*, 2011).

In this chapter I developed an *in silico* approach to address whether the expanded ATXN1 protein is predicted to bind with higher affinity to reported interactors. This methodology is based on the number and the percentage of interfacing residues of the expanded and wt protein with their interactors. It uses the 3D-protein structure (obtained using I-TASSER (Zhang, 2008)), the interfaces prediction using the 3D structure (with CPORT (de Vries and Bonvin, 2011)), the protein–protein docking predictions (PDP using HADDOCK (van Zundert *et al.*, 2016)) and PISAePDB (Krissinel and Henrick, 2007) to calculate the characteristics of the interaction-interface of the PDB results of the HADDOCK clusters. The predictions obtained using this approach can shed light on the ATXN1 regions that could be responsible for the higher binding affinity. I compared my results primarily with those of Suter *et al.* (2013), since these authors report the largest set of interactors using full-length proteins and not truncated protein segments/domains as baits. My results support the observation made by Suter *et al.* (2013) and Lim *et al.* (2006) that the expanded ATXN1 shows larger number of interactors than the wt form, that the number of interface residues is larger in the expanded ATXN1, as well as the importance of the AXH domain for protein binding. Thus, I suggest that the N-terminal region of ATXN1 is the one responsible for the binding differences of the wt and expanded ATXN1. Based on the amino acid motifs at this region, as well as the inferred RNA binding region, I discuss how interactions with expanded polyQ region can affect ATXN1 function.

Materials and Methods

1. Importance of N-terminal region in the ATXN1 protein on the binding variances of expanded and wild-type forms.

1.1. Common interactome of the ATXN1 protein

Interactomes of model organisms including *Homo sapiens*, *Mus musculus*, *Xenopus laevis*, *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans* of ATXN1 and ATXN1-Like proteins (accession numbers are listed in the Supplementary Table 1) have been retrieved using EvoPPI (evoppi.i3s.up.pt) that unifies and compare data from the public available protein-protein interaction main databases, including BioGRID, CCSB, Droid, FlyBase, HIPPIE, HitPredict, HomoMINT, Instruct, Interactome3D, mentha, MINT, and PINA. For the between species comparison of the interactomes of ATXN1 and ATXN1L used EvoPPI was also used, with the following parameters: 0.05 for the minimum expect value (e-value); 50 for the minimum length of alignment block; 40 for the minimum identity (%) and 1 for the number of descriptions (max_target_seqs) and the interaction level. Venny web tool (v.2.1.0: <http://bioinfogp.cnb.csic.es/tools/venny/>) was used to identify the common interactors.

1.2. *In silico* approaches to predict the wt and expanded ATXN1 interacting surfaces with 71 interacting partners

The 81 proteins reported in Suter *et al.* (2013) were obtained from UniProtKB. The wt form ATXN1Q29 used on this work was obtained from UniProtKB and ATXN1Q82 as expanded ATXN1. The 3D structure prediction of these proteins was obtained with I-TASSER, that uses the know structure of protein segments (LOMETS), assembling all of them to predict the structure of the entire protein. The models predicted are benchmarked by a scoring system (C-score) to obtain quantitative assessments and the confidence of the results. The models with higher confidence (C-score), therefore better scored, were then used. DOCK5 and IGF2R proteins had size longer than 1500 residues (the limit size for I-TASSER analyses), and therefore, were excluded. To evaluate the structure similarity of wt ATXN1 and expanded ATXN1, TM-align on-line was used, an algorithm for protein sequences comparisons, based on residue-to-to-residue alignment. The score varies from 0 to 1, corresponding the value 1 to a perfect match between two

structures. The structural images were obtained using the PyMOL molecular package (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.).

For the prediction of protein-protein interface residues, active and passive involved in the interaction, CPORT (de Vries and Bonvin, 2011) was used. C-PORT (de Vries and Bonvin, 2011) is a integrative application from the HADDOCK bioinformatics tool to predict Ambiguous interaction restraints (AIR), using five interface prediction methods into a consensus predictor, more precisely, PIER (Kufareva *et al.*, 2007), ProMate (Neuvirth *et al.*, 2004), cons-PPISP (Chen and Zhou, 2005), SPPIDER (Porollo and Meller, 2007), and PINUP (Liang *et al.*, 2006). CPORT(de Vries and Bonvin, 2011) predictions interfaces describes active and passive residues that can be used to predict the docking of macromolecular complexes in HADDOCK (van Zundert *et al.*, 2016). Two interactors, more specifically, the NPHP3 and the UHRF1BPL1L failed to predict the active and the passive residues, and, because of that, were not further analyzed.

The docking prediction of wt ATXN1 and expanded ATXN1 with the interacting partners was made further using HADDOCK (van Zundert *et al.*, 2016). HADDOCK is a web tool that accurately predicts and models the structure of biomolecular complexes, based on the integration of a variety of biochemical and biophysical interaction data such as NMR spectroscopy and X-ray crystallography interface restraints and other data, mutagenesis experiments and cryo-electron and anisotropy microscopy experiments. The clusters of macromolecular complexes that result from the possible dockings are then classified statistically by a Z-score, with the lowest values, being the best predictions (Dominguez *et al.*, 2003). CLCN2, EHMT1, FUBP3, LOXL1, SLC4A2, and ZXDC failed the docking with ATXN1Q29 and/or ATXN1Q82, and thus I analyzed 71 ATXN1 interactors.

For wt and expanded ATXN1, HADDOCK docking structures were selected based on 1) most reliable structure according to HADDOCK (the structure with the lowest Z-score), and 2) the structure presenting the highest probability of having a protein-protein interaction between wt ATXN1 and expanded ATXN1. The number and percentage of interface residues (including those that make hydrogen bonds, and salt bridges), the solvent-accessible area (Å²), and the percentage of the solvent-accessible area of the wt ATXN1 and the expanded ATXN1 with the interacting protein partners were determined using PISAePDB (Krissinel and Henrick, 2007). PISAePDB (Krissinel and Henrick, 2007) is an interactive tool for the exploration of macromolecular interfaces. It calculates results interactively for structures including structural and chemical

properties of macromolecular surfaces and interfaces and probable quaternary structures (assemblies), their structural and chemical properties and probable dissociation pattern. If these values are greater for the docking with the expanded ATXN1 than for the wt ATXN1 I consider that there is a higher binding affinity with the former, and vice versa. I assume no preference if the value is identical for the wt ATXN1 and expanded ATXN1. A flowchart showing methodology 2 (the one showing the best agreement with Suter *et al.* (2013; see Results) is presented as Figure 1. The PDB files describing the docking of these proteins are available as Supplementary Table 2.

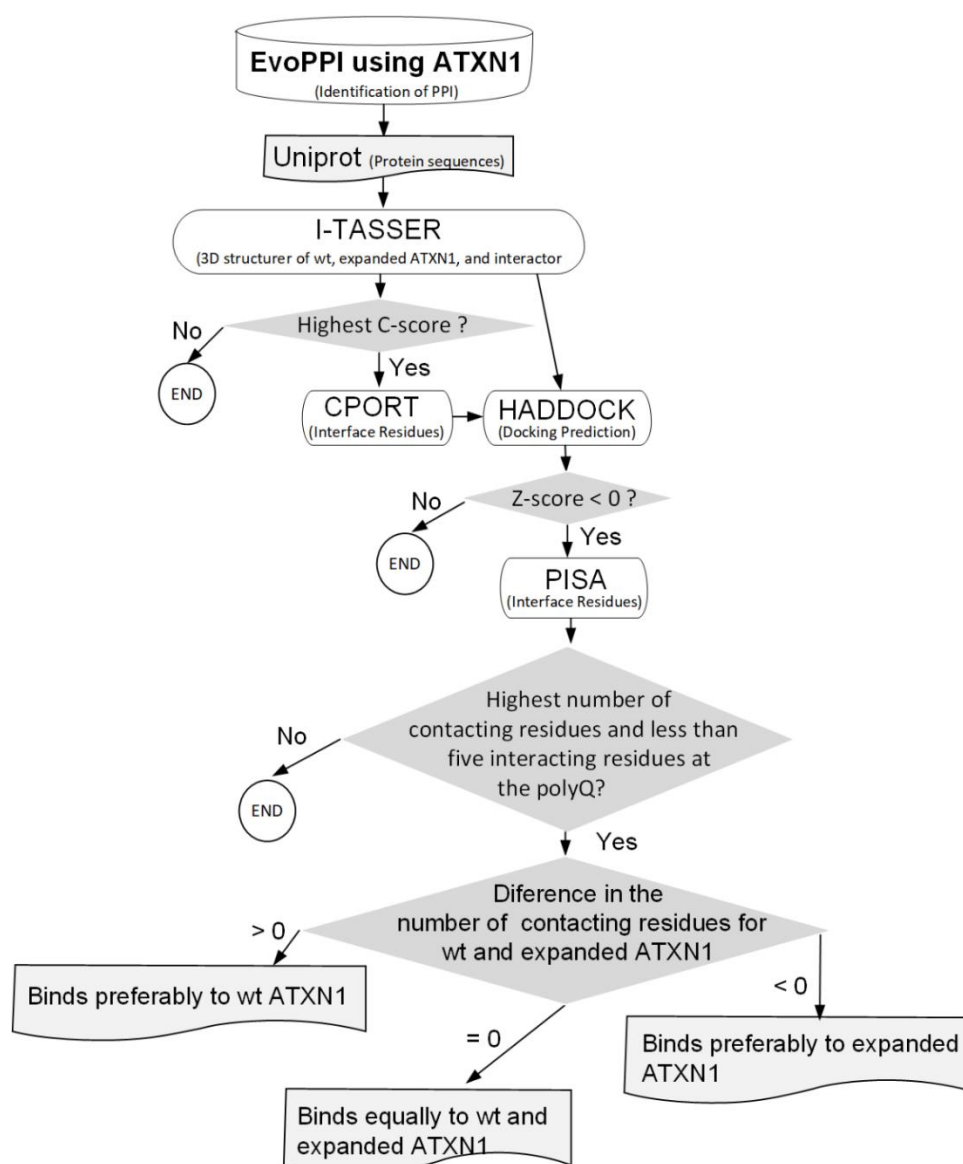


Figure 1. Flowchart describing the *in silico* methodology used, which showed the best agreement with Suter *et al.* (2013) results.

The interface residues of the 19 model structures of the AXH domain in complex with CIC (PDB ID: 2M41, (Chiara *et al.*, 2013) PDB ID: 4J2L (Kim *et al.*, 2013)) were also determined using PISAePDB (Krissinel and Henrick, 2007).

1.3. Gene ontology analysis

Proteins were divided into two groups according to whether the *in silico* prediction is in agreement or not with the wet bench results of Suter *et al.* (2013). Then, I used PANTHER webtool (Mi *et al.*, 2013), <http://www.pantherdb.org/>, that classify proteins to test for possible functional enrichment in each group, including molecular function, biological process or pathways.

1.4. Coiled-coil regions prediction

Marcoil1.0, a tool MARKov model-based program that predicts presence and position of potential coiled-coil domains in protein sequences, was used to predict the presence of coiled-coil regions in the protein partners of wt and expanded ATXN1 reported in Suter *et al.* (2013). As parameters, I used the 9FAM as Coiled-Coil Emission P. Matrix and the precomputed (MARCOIL-H) as HMM Transition P. Matrix. I only considered as proteins that form coiled-coils regions, those that present coiled-coil predicted domains with a minimum threshold of 90%.

1.5. Statistical analyses

To address whether the presence of a polyQ influences the observed (according to EvoPPI) number of interactor proteins, regression analyses of the number of ATXN1 interactors in humans and mouse were performed for 65 human proteins that have a polyQ, divided into two groups according to whether in mouse they also have a polyQ. 1000 random permutations were used to compare the slopes and the intercept of the regressions in humans and mouse. Regression analyses were also performed to address whether in humans, the number of interactors is influenced by protein size.

In order to address possible technical biases, non-parametric Mann-Whitney U tests were performed to look for possible associations between sequence length, the I-TASSER C-score, the HADDOCK Z-score, and being in agreement or not with the results of Suter *et al.* (2013).

To test whether the docking predictions that are in agreement with Suter *et al.* (2013) are enriched in coiled-coil domains when compared to the results that are not in agreement with Suter *et al.* (2013), Marcoil1.0 (Delorenzi and Speed, 2002) was used. Fisher exact test was used to compare the frequency of cases in the two groups. Moreover, in order to test whether the docking predictions that are in agreement/disagreement with Suter *et al.* (2013) are enriched in a particular gene ontology category, according to PANTHERdb (Mi *et al.*, 2013).

To infer binding preference between wt and expanded ATXN1 forms and the interactors, non-parametric Sign tests were also used to compare the number of residues in the interface of the interacting partners, percentage of number of interface residues, solvent-accessible area (Å), and percentage of the solvent-accessible area.

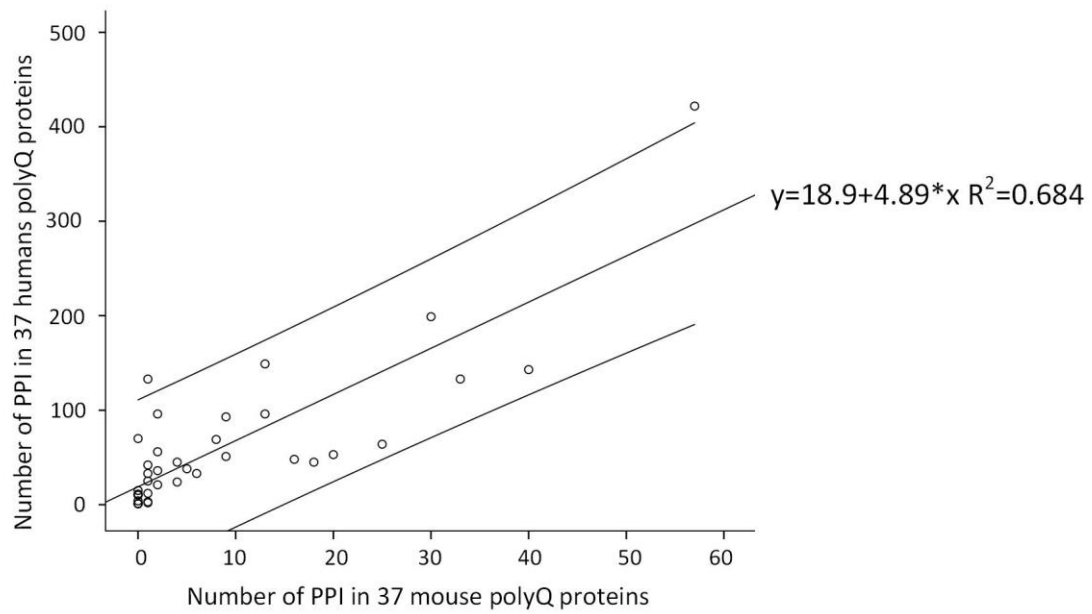
In order to identify the regions responsible for the higher binding affinity I also compared, using non-parametric sign tests the number of interface residues before and after the polyQ tracts, and the polyQ tracts and 20 amino acids residues flanking the polyQ tract for the wt and expanded ATXN1 forms.

All analyses were performed using SPSS software (<https://www.ibm.com/analytics/spss-statistics-software>).

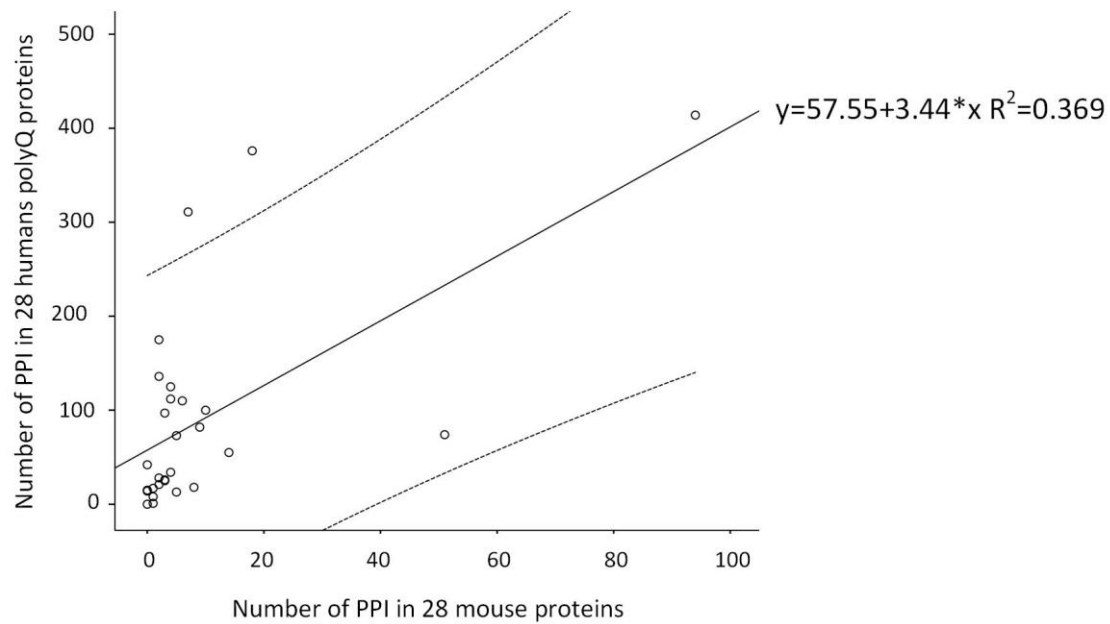
Results

Human polyQ proteins have more interactors than non-polyQ proteins

In humans, polyQ proteins tend to have more interactors than proteins lacking a polyQ tract (Schaefer *et al.*, 2012; Vazquez *et al.*, 2019b) which supports the view that the presence of a polyQ region increases the number of partner proteins. Nevertheless, in humans, more than 73% of the 65 genes encoding polyQ proteins are associated with diseases (Supplementary table 3), and thus it is unlikely that the reported interactions are a random sample of the interactome. Therefore, I retrieved from EvoPPI (Vázquez *et al.*, 2019a; <http://evoppi.i3s.up.pt>), the number of interactors for the 65 human proteins that have a polyQ (a stretch with a minimum of 10 consecutive glutamines) as well as for the proteins encoded by the *Mus musculus* orthologues. Then, I compare the number of interactors for the proteins presenting a polyQ in both species, and those presenting a polyQ in humans only (Figure 2A-B, respectively).



A



B

Figure 2. Graphical representation of the comparison of the number of interactors between proteins containing polyQ tracts in both *H. sapiens* and *M. musculus* species (A) and the proteins displaying a polyQ tract in humans only (B). From this graphic, is possible to see a higher number of interactors for the proteins containing a polyQ tract, comparing to proteins that not possess them.

For proteins having polyQ in both humans and mouse (37), the regression of the number of interactions in the two species explains 68.4% of the variability, while for proteins that have polyQ in humans but not in mouse explains 36.9% of the variability. The slope of the former (4.89) is higher than for the latter (3.44). Nevertheless, when I pull all data together and get 1000 random samples of size 37, a slope as high or higher than 4.89 is obtained 11.4% of the times. Moreover, when I obtain 1000 random samples of size 28, a slope as low or lower than 3.44 is obtained 27.0% of the times. Therefore, the observed slopes for the regression of the number of interactions of proteins having polyQ in both humans and mouse and proteins that have polyQ in humans but not in mouse, could have been generated by taking two random samples of size 37 and 28 from a single population, meaning that the two regression lines are likely parallel. The slope of these regressions indicate the difference in the amount of data available for humans and mouse (in between 3.44 and 4.89 more data for humans than for mouse). The interception of the y axis is much lower (18.9) for the regression obtained when using proteins having polyQ in both humans and mouse than when using proteins that have polyQ in humans but not in mouse (57.55). When I pull all data together and get 1000 random samples of size 37, an interception as low or lower than 18.9 is obtained in only 0.1% of the cases. Moreover, when I obtain 1000 random samples of size 28 an interception as high or higher than 57.55 is obtained only 2.9% of the times. It is thus highly unlikely that the two datasets represent two random samples. Therefore, proteins that have polyQ in humans but not in mouse have many more interactions in humans than what is expected considering the number of reported interactions in mouse and the regression obtained for proteins having polyQ in both humans and mouse. In humans, the number of interactors is not related to the size of the proteins (size according to UniProtKB; linear regression analyzes $P=0.084$; Supplementary Figure 1), in agreement with Schaefer *et al.* (2012).

The interactomes of the *Homo sapiens* ATXN1 and ATXN1L retrieved from EvoPPI (Vázquez *et al.*, 2019), revealed a minimum of 311 and 70 interactors, respectively. Despite the conservation between these proteins (Mizutani *et al.*, 2005) only 35 interact with both ATXN1 and ATXN1L (Supplementary Table 4). Seven of these proteins (RBF2, SREBF1, NCOR1, GMEB2, PHPT1, ANKHD1, and STAC2) are reported as interacting with expanded ATXN1 (Lim *et al.*, 2006, Suter *et al.*, 2013), despite ATXN1L not having a polyQ tract. Moreover, of the different *M. musculus* ATXN1 interactors (YWHAE, RBM17, CIC, VCP, KAT5, and Nr1f1) YWHAE and RBM17 are

reported as interacting in humans with expanded ATXN1 (Suter *et al.*, 2013), although the mouse ATXN1 does not have a polyQ tract. Moreover, none of these interactors are predicted to have coiled-coil regions, that have been implicated in the interaction with polyQ regions (Fiumara *et al.*, 2010, Schaefer *et al.*, 2012, Petrakis *et al.*, 2013). For the three *Drosophila melanogaster* interactions reported for the orthologous ATXN1, none is in common with the human ATXN1. For *X. tropicalis*, *D. rerio*, and *C. elegans* no information is available for the orthologous ATXN1, as well as for the orthologous ATXN1L for *M. musculus*, *X. laevis*, *D. rerio*, and *C. elegans*.

Validation of the *in silico* method used for predicting protein binding surfaces

Although two methodologies were used for the identification of the protein interfaces, namely: 1) the most reliable complex structure according to HADDOCK, and 2) the complex structure that maximizes the probability of having a PPI, for 73.2 % (60 out 71) of the proteins reported by Suter *et al.* (2013) as interacting with wt ATXN1(Q29) (7; Table 2) and/or expanded ATXN1(Q82) (53; Table 2) the results were the same (Table 2).

Table 2. PPI with the wt and expanded ATXN1 according to the methodologies used. 1) the most reliable structure according to HADDOCK, 2) docking structure that maximizes the probability of having a PPI.

ATXN1		binding		Proteins names (GeneID; UniProtKB)
preference				
wt	(in	both		CRY2 (1408; Q49AN0); EIF1B (10289; O60739);
	approaches)			GGA2 (23062; Q9UJY4); LITAF (9516; Q99732); RBM26 (64062; Q5T8P6); SEMA4G (57715; Q9NTN9); TOMM20 (9804; Q15388)
Expanded	(in	both		ADD3 (120; Q9UEY8); ARID5A (10865; Q03989);
	approaches)			ASNS (440; P08243); BAALC (79870; Q8WXS3); BASP1 (10409; P80723); C16orf5 (29965; Q9H305); C2orf27B (408029; Q580R0); CAMK2B (816; Q13554); CHRNA7 (1139; P36544); CRK (1398; P46108); DHRSX (207063; Q8N5I4); DHX37 (57647; Q8IY37); DIXDC1 (85458; Q155Q3); EIF3F (8665; O00303); ESRRA (2101; P11474); ETV4 (2118; P43268); FAM46B (115572; Q96A09); FAR1

	(84188; Q8WVX9); FOSL1 (8061; P15407); GATAD1 (57798; Q8WUU5); HEY2 (23493; Q9UBP5); HEYL (26508; Q9NQ87); HNRPLL (92906; Q8WVV9); ILVBL (10994; A1L0T0); IMMT (10989; Q16891); KCTD15 (79047; Q96SI1); KIF22 (3835; Q14807); LASP1 (3927; Q14847); LPAR2 (9170; Q9HBW0); MAGEB18 (286514; Q96M61); MAGEB2 (4113; O15479); MAGEB6 (158809; Q8N7X4); MCART1 (92014; Q9H1U9); MLST8 (64223; Q9BVC4); NCAM1 (4684; P13591); OTX2 (5015; P32243); PIAS1 (8554; O75925); PPAT (5471; Q06203); QKI (9444; Q96PU8); RAPGEF1 (2889; Q13905); RBFOX2 (23543; O43251); SF1 (7536; Q15637); SLC6A13 (6540; Q9NSD5); STAM2 (10254; O75886); TMX2 (51075; Q9Y320); TRIM38 (10475; O00635); TSC1 (7248; Q92574); TTRAP (51567; O95551); UHRF2 (115426; Q96PU4); WBSCR16 (81554; Q96I51); YWHAE (7531; P62258); ZC3H10 (84872; Q96K80); ZSCAN1 (284312; Q8NBB4)
wt in 1/ Expanded in 2	CREM (1390; Q03060); CRIP2(1397; P52943); MSX2 (4488; P35548); PLEKHB1 (58473; Q9UF11); PSPH (5723; P78330); SV2A (9900; Q7L0J3)
wt in 2/ Expanded in 1	CXorf27 (25763; O75409); RAI2 (10742; Q9Y5P3); TCTA (6988; P57738); TP53I11 (9537; O14683)
wt in 1/ Expanded in 1 and 2	FAM46A (55603; Q96IP4)

For the remaining cases, to address the accuracy of the two approaches, the predictions generated by them regarding binding strength towards the wt and expanded ATXN1, are compared with the wet bench results of Suter et al. (2013). When I used approach 1), 38 out of the 71 PPI analyzed were in agreement in both studies (53.5%) and when I use approach 2) 43 out of the 71 PPI analyzed were in agreement in both studies (60.6%) (Table 3). Therefore, the second approach is more reliable in predicting the interaction of proteins with wt ATXN1 (three of the 16 proteins (18.7%)) and expanded ATXN1 (40 of the 48 (83.3%)).

Table 3. Predicted (approach 2) and determined (Suter et al., 2013) binding preferences of ATXN1 partners with the wt and expanded ATXN1. In bold are those proteins for which the prediction is different when using the two in-silico methodologies.

This work	Suter <i>et al.</i>	Proteins names (GeneID; UniProtKB)
wt	Wt	CXorf27 (25763; O75409) ; GGA2 (23062; Q9UJY4); TOMM20 (9804; Q15388)
expanded	Expanded	BAALC (79870; Q8WXS3); BASP1 (10409; P80723); C16orf5 (29965; Q9H305); C2orf27B (408029; Q580R0); CREM (1390; Q03060) ; CRIP2 (1397; P52943) ; CRK (1398; P46108); DHRSX (207063; Q8N5I4); ESRRA (2101; P11474); FAM46A (55603; Q96IP4) ; FAM46B (115572; Q96A09); FAR1 (84188; Q8WVX9); GATAD1 (57798; Q8WUU5); HEY2 (23493; Q9UBP5); HNRPLL (92906; Q8WVV9); IMMT (10989; Q16891); KCTD15 (79047; Q96SI1); KIF22 (3835; Q14807); LPAR2 (9170; Q9HBW0); MCART1 (92014; Q9H1U9); MSX2 (4488 (P35548)) ; NCAM1 (4684; P13591); OTX2 (5015; P32243); PIAS1 (8554; O75925); PLEKHB1 (58473; Q9UF11) ; PPAT (5471; Q06203); PSPH (5723; P78330) ; QKI (9444; Q96PU8); RAPGEF1 (2889; Q13905); RBM9 (23543; O43251); SF1 (7536; Q15637); SLC6A13 (6540; Q9NSD5); SV2A (9900; Q7L0J3) ; TMX2 (51075; Q9Y320); TRIM38 (10475; O00635); TTRAP (51567; O95551); UHRF2 (115426; Q96PU4); YWHAE (7531 (P62258); ZC3H10 (84872; Q96K80); ZSCAN1 (284312; Q8NBB4)
wt	Expanded	CRY2 (1408; Q49AN0); EIF1B (10289; O60739); LITAF (9516; Q99732); RAI2 (10742; Q9Y5P3) ; RBM26 (64062; Q5T8P6); SEMA4G (57715; Q9NTN9); TCTA (6988; P57738) ; TP53I11 (9537; O14683)
expanded	wt	ADD3 (120; Q9UEY8); ASNS (440; P08243); CAMK2B (816; Q13554); DHX37 (57647; Q8IY37); DIXDC1 (85458; Q155Q3); EIF3F (8665; O00303); ETV4 (2118; P43268); HEYL (26508; Q9NQ87); ILVBL (10994; A1L0T0); MAGEB18 (286514; Q96M61); MAGEB2

		(4113; O15479); MAGEB6 (158809; Q8N7X4); WBSCR16 (81554; Q96I51)
expanded	no preference	ARID5A (10865; Q03989); CHRNA7 (1139; P36544); FOSL1 (8061; P15407); LASP1 (3927; Q14847); MLST8 (64223; Q9BVC4); STAM2 (10254; O75886); TSC1 (7248; Q92574)

Moreover, the use of the same methodology but different parameters such as, the number of residues in the interface of the interacting partners, percentage of number of interface residues, solvent-accessible area (Å), and percentage of the solvent-accessible area of wt ATXN1, expanded ATXN1 and interacting partners, does not result in a better agreement with Suter *et al.* (2013) results (Supplementary Table 5). The sequence length of the interacting proteins (P=0.312), the I-TASSER C-score of the interacting proteins (that estimates the quality of predicted models by I-TASSER, calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations; P=0.892), and the Haddock Z-score (that indicates how many standard deviations from the average the cluster is located in terms of score) of both forms of ATXN1 (wt, P=0.911; expanded, P=0.608) are not statistically significantly associated with the ability of making the correct prediction (Mann-Whitney U test). The main difference in the I-TASSER predicted structures of the wt ATXN1 (Q29) and expanded ATXN1 (Q82) presenting the most similar structures (TM-score= 0.97361; if normalized by length of wt ATXN1; and the highest C-score), is that the wt polyQ tract is predicted as random coil (absence of regular secondary structures), while in the expanded polyQ this region is predicted as three α -helices connected by loops (Figure 3).

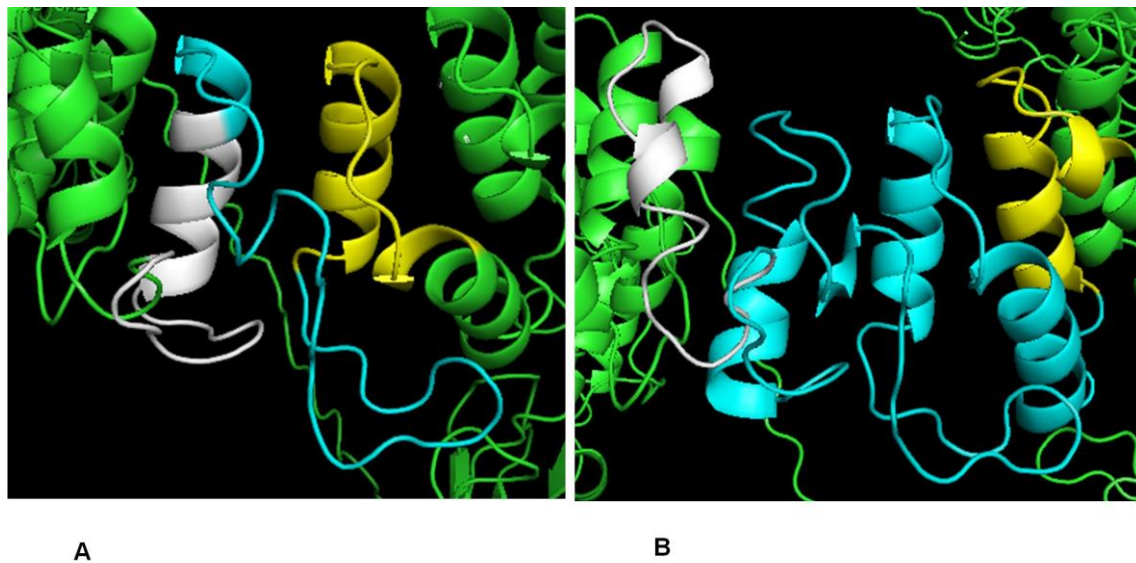


Figure 3. Representation of the 3D structure of the polyQ tract (in cyan) and 20 amino acids residues flanking the polyQ tract (in yellow those in the N-terminal region, and in white those in the C-terminal region) in: A) wt ATXN1 and B) expanded ATXN1.

It should be noted that I-TASSER modeling of the 3D structures starts from the structure templates from the PDB library and uses the templates of the highest significance in the threading alignments (Yang *et al.*, 2015), and the polyQ region has fewer amino acids used as threading templates (Supplementary Figure 2), making the prediction of these regions less reliable. Therefore, the *in silico* methodology is likely more robust when no interactions are predicted with the polyQ tract.

Since alpha-helical coiled-coil domains have been suggested as being critical for the spontaneous aggregation of Q/N-rich yeast prions and polyQ disease proteins (Fiumara *et al.*, 2010, Schaefer *et al.*, 2012, Petrakis *et al.*, 2013), I have addressed if the proteins that show similar results in the *in silico* and Suter *et al.* (2013) approaches are enriched in such domains. Nevertheless, using a Fisher exact test, the presence of predicted coiled-coil domains (using Marcoils prediction threshold of 90%) is not statistically different ($P=0.347$) in the two datasets.

I also address if when our results are/are not in agreement with those of Suter *et al.* (2013), the interacting proteins are enriched in a particular function. According to PANTHER (Mi *et al.*, 2013) there is no statistically significant enrichment for the two datasets (for all gene ontology classes $P>0.05$).

The AXH domain is important for PPI

The AXH domain region, relevant to PPI, presents a similar pattern of interaction for both wt and expanded ATXN1 types. In Lim *et al.* (2006), 72.5% of the ATXN1 protein partners bind the AXH domain. It should be noted that the AXH domain can mediate neurodegeneration through its interaction with other proteins (de Chiara *et al.*, 2005, Kang *et al.*, 2015). Here I show that most of the proteins studied interact with the AXH domain at three regions (Figure 4).

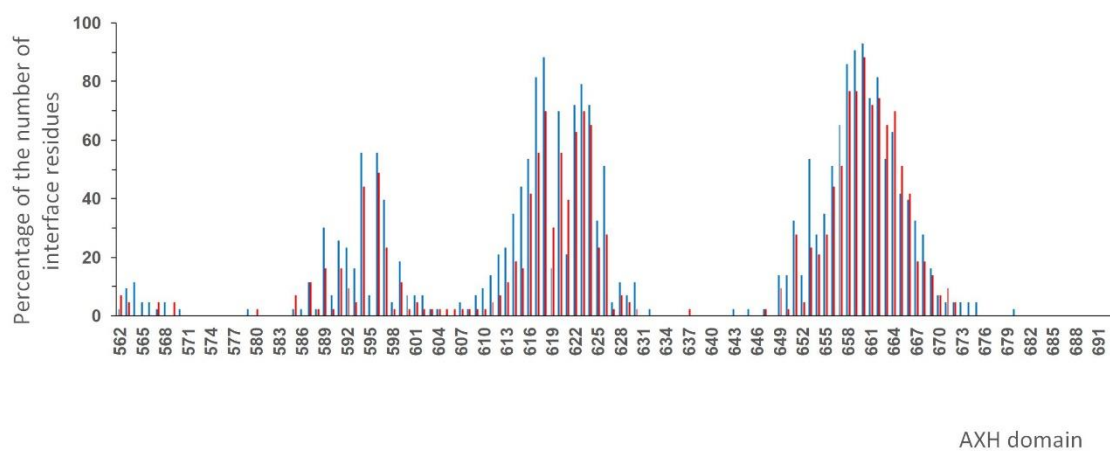


Figure 4. Graphical representation of the AXH domain, showing high percentage of the number of interface residues in this domain from all the PPI studied, especially in three regions. The percentage of the number of interface residues at the AXH domain with the 43 interactors in agreement with Suter *et al.* (2013) for the wt (in blue) and expanded (in red) ATXN1.

An overlap is observed at seven residues (ATXN1 sites 588, 591, 594, 599, 602, 609, and 649; see Supplementary Figure 3) between our predictions and those identified when using the 19 available crystal structures for the complex ATXN1/AXH domain and CIC (de Chiara *et al.*, 2013a. b; Kim *et al.*, 2013). Therefore, these regions are fundamental for the PPI of ATXN1.

The expanded ATXN1 shows more residues at the binding interface than the wt form

Comparison of the docking results of the expanded and wt ATXN1 with the 43 proteins for which the *in silico* results are in agreement with those of Suter *et al.* (2013), revealed a significant increase in the interfacing residues and in the percentage of

interfacing residues of the expanded ATXN1 compared with wt ATXN1 (Sign test; $P=0.000$ for both; Fig. 5A-B). Moreover, the number and the percentage of interfacing residues of the 43 interacting proteins with the expanded ATXN1 is increased compared with wt ATXN1 ($P=0.015$ for both; Fig. 5C-D). The buried solvent accessible area ($\text{\AA}$) at the predicted binding interface of the expanded ATXN1 ($P=0.002$; Fig. 5E) also increases compared with the wt ATXN1. Moreover, the percentage of the solvent accessible area buried in the docking complexes of the expanded mutated ATXN1 ($P=0.025$) increases compared with the wt ATXN1 (Fig. 5F). It should be noted that the analysis of the docking results of the expanded and wt ATXN1 when using all the 71 proteins revealed similar results (Supplementary Figure 4) except for the percentage of the buried solvent accessible area of the expanded ATXN1 ($P=0.104$) that does not increase compared with the wt ATXN1.

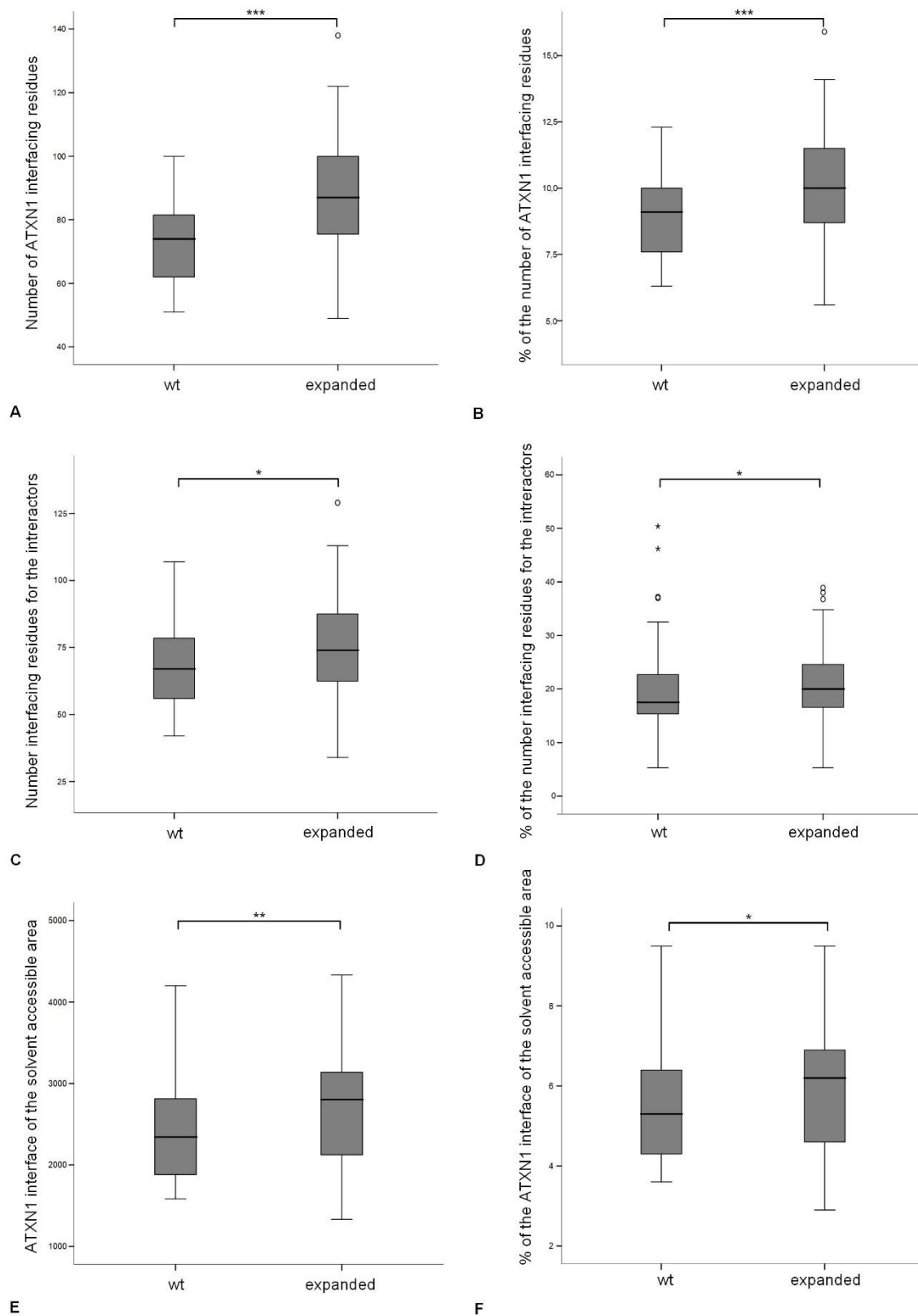


Figure 5. Comparison of the wt and expanded ATXN1 docking results of the 43 interactors analyzed A) interfacing residues of the ATXN1. B) percentage of interfacing residues of the ATXN1. C) number of interfacing residues for the 43 interactors. D) percentage of interfacing residues of the 43 interactors. E) interface of the solvent accessible area (Å) of the ATXN1. F) percentage of the interface of the solvent accessible area of the ATXN1.

The N-terminal region of the expanded ATXN1 shows more residues at the predicted binding interface and an increase in the number of hydrogen bonds and salt bridges when compared to the wt

To identify the regions of the ATXN1 that are responsible for the overall significant increase in the number of interfacing residues of the expanded ATXN1 when compared with the wt ATXN1, I analyzed separately the sum of the number of interface residues in six regions: before and after the polyQ tracts (including the 20 amino acids residues flanking the polyQ tract), the polyQ tracts, the 20 amino acid residues before and after the polyQ tracts and the AXH domain (Table 4).

Table 4. Number of interface residue (%) for the PPI in six regions: before and after the polyQ tracts (including the 20 amino acids residues flanking the polyQ tract), the polyQ tracts and the AXH domain.

ATXN1	Before the polyQ	PolyQ tract	After the polyQ	20 amino acid before polyQ	20 amino acid after polyQ	AXH domain
wt	12.2	0.2	87.6	0.7	0.0	31.1
Expanded	23.6	1.0	75.3	0.5	0.1	20.2

Using the docking results of the expanded ATXN1 and wt ATXN1 with the 43 proteins that are in agreement with Suter *et al.* (2013), a significant increase is only observed in the number of residues before the polyQ tract (Non-parametric Sign test; for both $P=0.000$; Figure 6).

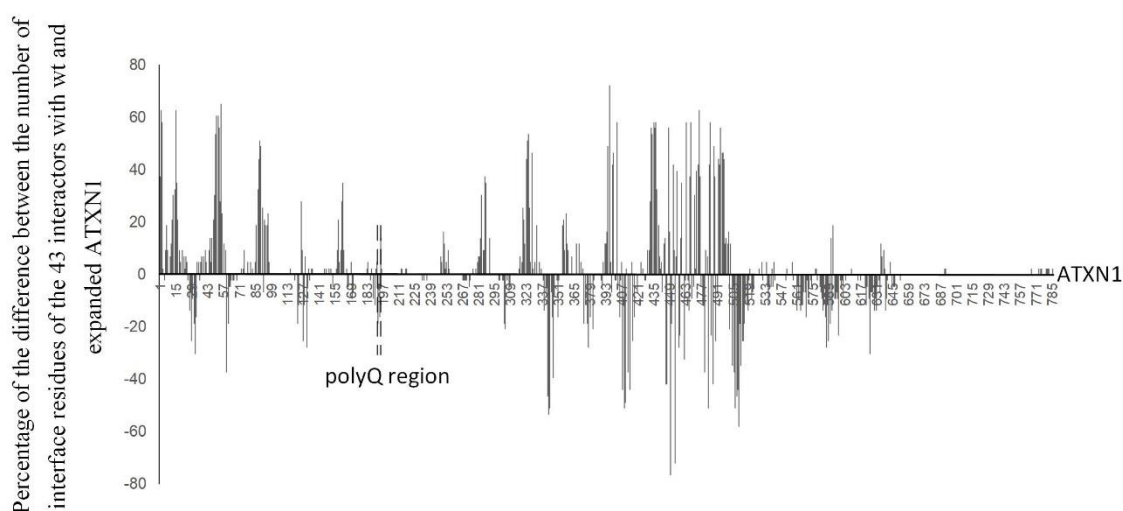


Figure 6. Difference of the number of interface residues of the 43 interactors analyzed. Percentage of the difference of the number of interface residues of the 43 interactors in agreement with Suter *et al.* (2013), with wt and expanded ATXN1. PolyQ tract is not represented.

The percentage of the difference of the number of predicted interface residues of the 43 proteins when they interact with expanded and wt ATXN1 revealed that the largest changes occurred in the N-terminal region of ATXN1 (Figure 6). Moreover, the number of predicted hydrogen bonds as well as salt bridges is higher at the N-terminal region of the expanded form than in the wt (Non-parametric Sign test; for both $P=0.000$). This is in agreement with Lim *et al.* (2006), which using an Y2H assay, also describe a predominance of proteins interacting with the N-terminal region of the expanded ATXN1 (55.7%) when compared with wt ATXN1 (15.2%). Therefore, it seems that the presence of a polyQ region affects mostly the interaction in the region before the polyQ tract. In this region, motifs related with transcription, cell communication, phosphorylation, and sumoylation (Table 5) were identified, using the eukaryotic linear motifs (elm.eu.org) prediction tool.

Table 5. ATXN1 N-terminal region eukaryotic linear motifs (elm.eu.org).

Pathway	Sequence motif	Location	Proteins that bind in the region	Description	Pattern	Probability
Transcription	YSPPSAP	87-93	32	Motif recognized by class I SH3 domains	[RKY]..P..P	1.237e-03
	AWLPGNP	44-50	21		...[PV]..P	1.317e-02
Phosphorylation	SPPSAPR	88-94	31	CK1 recognition site	S..([ST])...	1.704e-02
	SPPSAPRS	88-95	32	GSK3 recognition site	...([ST])...[ST]	2.679e-02
	PVTSAVAS	154-161	21			
Cell communication	PSAP	90-93	27	Glycosaminoglycan attachment site	[ED].(S)[GA]	1.792e-02
	ASAA	160-163	24			
Sumoylation	TFQFI	126-130	21	Glycosaminoglycan attachment site	[ED].(S)[GA]	1.792e-02

Moreover, using RNApred (a webserver for prediction of RNA-binding proteins), a sliding window of 50 residues and an increment of 25, four putative RNA binding regions are inferred, namely one before the polyQ region and three after (Supplementary Figure 5). One of them corresponds to the region described in the literature.

Discussion

There is only a partial agreement between our predictions and the experimental data retrieved from Suter *et al.* (2013). Therefore, in what follows I discuss the possible sources of error. First, I must consider the wrong prediction of the protein structure for some interactors. Nevertheless, the quality of the prediction is usually evaluated by the C-score value, and I found no association between being or not in agreement with Suter *et al.* (2013) (independently of considering the difference in the number ($P=0.892$), percentage of interfacing residues ($P=0.325$), the buried solvent accessible area ($P=0.2$), or the percentage of the solvent accessible area buried in the docking complexes ($P=0.2$) and the C-score value (non-parametric Mann-Whitney tests).

The second possible source of error is the quality of the prediction of the docking structures, evaluated by the associated Z-score. For both the wt and expanded ATXN1, when I use the docking structure presenting the lowest Z-score or the docking structure with the highest probability of having a protein-protein interaction, I found no association between the Z-score and being or not in agreement (independently of considering the difference in the number of interacting residues for ATXN1 (wt, $P=0.911$; expanded, $P=0.608$) or for the interactors (wt, $P=0.086$; expanded, $P=0.848$), percentage of interfacing residues for ATXN1 (wt, $P=0.339$; expanded, $P=0.784$) or for the interactors (wt, $P=0.086$; expanded, $P=0.848$), the buried solvent accessible area for ATXN1 (wt, $P=0.751$; expanded, $P=0.708$) or for the interactors (wt, $P=0.470$; expanded, $P=0.777$), or the percentage of the solvent accessible area buried in the docking complexes for ATXN1 (wt, $P=0.372$; expanded, $P=0.854$) or for the interactors (wt, $P=0.548$; expanded, $P=0.894$)) with Suter *et al.* (2013); non-parametric Mann-Whitney tests). Another possible source of error is the wrong inference of the binding surfaces. The prediction that the majority of the proteins bind at the AXH region, suggests that this is not a major source of error either. Nevertheless, the number of interacting residues at the polyQ region is significantly associated with the wrong outcome (Mann-Whitney test; $P=0.036$), but this association is no longer significant when interactors showing six or more predicted interactions with the ATXN1 polyQ region are removed ($P=0.804$). It should also be noted, that ATXN1 likely participates in a multi-protein complex but our predictions are based on the binding of ATXN1 to a single interactor at a time. Although ATXN1 forms dimers mediated by the AXH domain, it has been reported that it binds with other proteins as a monomer (de Chiara *et al.*, 2013). Lastly, the different quantitative methods that have been developed to identify the strength of the binding are

not error-free, and thus, for a few cases, the outcome reported in Suter *et al.* (2013) may be wrong. Indeed, the comparison of the results using different methodologies, as for instance with Lim *et al.* (2006), revealed that six of the seven proteins that were detected by Suter *et al.* (2013) as partners of expanded ATXN1 were detected as partners of both forms of ATXN1 by Lim *et al.* (2006) (CRK; FAM46A; FAM46B; OTX2; RBFOX2; ZC3H10) and ARID5A detected by Suter *et al.* (2013) as an interacting partner of both forms of ATXN1, was detected by Lim *et al.* (2006) as interacting with the expanded ATXN1 only. The *in silico* approach here used has a better performance for proteins that interact preferably with expanded ATXN1 (83.33%) compared with the wt ATXN1 (18.75%). It is unclear whether this is a bias of the *in silico* method, or whether the number of interactors that interact preferably with the expanded ATXN1 is underestimated by Suter *et al.* (2013). Our results agree with the hypothesis that the expansion of the polyQ tract that occurs in SCA1 alters the conformation of ATXN1, leading to abnormal strength interactions and conferring a toxic gain of function (Tsuda *et al.*, 2005, Lim *et al.*, 2008, Nalavade *et al.*, 2013, de Chiara and Pastore, 2014, Paulson *et al.*, 2017). It should be noted that I obtain similar results when using the 43 proteins only that are in agreement with Suter *et al.* (2013) or when using all 71 proteins.

An expansion of the polyQ tract of ATXN1 has been identified as the cause of the neurodegenerative disease SCA1 (Orr *et al.*, 1993). The overall function of wt polyQ stretch is likely to stabilize PPI and/or spacer elements between individual folded domains in molecules that mediate PPI (Karlin and Burge, 1996, Schaefer *et al.*, 2012). Evolutionary analyses, however, suggest that the polyQ tract does not seem to be necessary for protein function (Schaefer *et al.*, 2012). Indeed, neither the ATXN1L paralog, nor the ATXN1 orthologous genes of *M. musculus*, *X. tropicalis*, *X. laevis*, *D. rerio*, *D. melanogaster* and *C. elegans* show a polyQ tract. All these proteins present the AXH domain, and as here shown, all human proteins here studied interact at the C-terminal region (where the SAR and AXH domains are located), and not with the polyQ region. The AXH domain is essential for binding with ATXN1, in agreement with previous studies (Tsai *et al.*, 2004, Tsuda *et al.*, 2005, Lam *et al.*, 2006, Serra *et al.*, 2006, Goold *et al.*, 2007). When the polyQ region is expanded, more interactions are established at the N-terminal region. This result is in agreement with the observation that a larger number of interactions are established with the N-terminal region of the expanded than with the wt ATXN1 (Lim *et al.*, 2006). Although I did not find evidence supporting the importance of the polyQ surrounding sequences on the PPI modulation (Petrakis *et al.*,

2013), this may be due to difficulties in determining the structure of polyQ regions (Masino, 2004, de Chiara and Pastore, 2011, Totzeck *et al.*, 2017). In the literature, the polyQ tract of the two ATXN1 forms has been proposed as being largely the same (Klein *et al.*, 2007), slightly different in the overall secondary structure content (Davies *et al.*, 2008), or even as a sharp change from an extended monomeric conformation to a collapsed state (Walters and Murphy, 2009). Here I used the models of both forms of ATXN1 with higher confidence scoring and with the most similar secondary structure.

For the N-terminal ATXN1 region, I identify motifs related with transcription, cell communication, phosphorylation, and sumoylation. The latter two are the major posttranslational modifications described for ATXN1 (see Background). 80% of the interactors that bind with expanded ATXN1 interact in these motif regions, thus showing the importance of these processes in SCA1. Moreover, it is conceivable that there is also a RNA binding domain in this region, and that the proteins that interact with expanded ATXN1 at the N-terminal region block the access of other proteins or RNA. Phosphorylation of serine residue 776 by Protein Kinase A (PKA) has been reported, but there are three other putative residues that could be phosphorylated (a Threonine at position 236 and a Serine at residues 239 and 254). The 776 residue is part of a canonical Arg-containing phospho-motif responsible for a strong interaction with 14-3-3 proteins, such as YWHAЕ. In mouse ATXN1, seven phosphorylation sites have been identified, and two are in the N-terminal region of the protein (Huttlin *et al.*, 2010). Kang and Hong (2010) showed that the SUMO-1 protein interacts with mutant ATXN1, but not with wt ATXN1, suggesting the involvement of the SUMO-1 system in the pathogenesis of SCA1 disease.

Conclusions

In conclusion, the *in silico* method here used can provide useful insight into which interactors bind preferably to the expanded ATXN1, predict binding surfaces, and possible functional consequences of the expansion of the polyQ tract. Our results are 84.5% in agreement with the largest data set available using as baits ATXN1 full-length proteins (Suter *et al.*, 2013). The major limitations of the *in silico* methodology are the size of the proteins considered (the prediction programs do not infer the structure of proteins larger than 1500 amino acids), as well as considering binary interactions only and not larger protein complexes. Such methodology can be used to address differences in the binding surfaces caused by the expansion of the polyQ in other diseases. Using this approach, I revealed the importance of the N-terminal region, before the expanded polyQ region of the ATXN1, where an increase in the number of interactions is predicted. In this region, I identify motifs related with transcription, cell communication, phosphorylation, sumoylation, and RNA binding.

Chapter 2 – Consequences of polyglutamine tract expansion: are there general patterns?

Introduction

Proteins containing a polyglutamine tract (defined as minimum of 10 consecutive Qs; Schaefer *et al.* (2012)) are a group of 65 proteins in the human proteome, involved in widespread functions and pathways. Based on their distribution in the human genome, it seems that they have a function at the protein and also at transcription level (Schaefer *et al.*, 2012). The glutamine tract (polyQ) that is present in the wt protein, in healthy individuals usually has less than fifty repeats (Cohen-Carmon and Meshorer, 2012). Depending on the disease etiology, the polyQ tract reaches above fifty and often over hundred glutamines (Paulson *et al.*, 2000). Expansion of the CAG repeat coding for a polyglutamine stretch in specific target proteins are associated to severe neurodegenerative disorders (Silva *et al.*, 2018). Huntington's disease (HD), spinobulbar muscular atrophy (SBMA) and various spinocerebellar ataxias (SCAs) are the known polyglutamine disorders that cause neurodegenerative diseases (Paulson *et al.*, 2000). The pathologies are dominant, mostly autosomal (except heterossomal dominant disease SBMA), with the polyQ tract having a role in the aggregation propensity, that leads in late stages to nuclear dysfunction and consequently neural cell death of Purkinje cells (PCs), brain-stem nuclei and the spinocerebellar tracts. The specific role of the polyQ in these diseases is, however, not clear. For instance, aggregation has been suggested as having a protective role rather than the case of de disease (Arrasate *et al.*, 2004).

Ataxia type 1 (SCA1) is an autosomal dominate ataxia caused by an expansion on the coding sequence of the N-terminal polyglutamine tract region of *Ataxin-1* gene. As described in Chapter 1, in this protein there are various domains located in the C-terminal region that play a role in PPI, such as the AXH domain, conserved in various animal models, including *D. melanogaster* and *C. elegans*, that play an important role in the recognition and interaction with ATXN1 protein partners. Also located in the C-terminal region, the self-association region (SAR), and the RNA-binding regions, have also been described as important for ATXN1 self-interaction and aggregation of the mutant ATXN1. ATXN1 is involved in the formation and regulation of key multimeric

protein complexes within the nucleus, such as the one involving the transcription factor CIC (Capicua) and RBM17 (a splicing factor), involved in major developmental processes including the nervous system development. The 14-3-3 protein family, proteins containing phospho-serine motifs, are also reported interactors of ATXN1 protein. Thus, protein phosphorylation plays a major role in the function and activity of ATXN1, but other processes such as ubiquitination, sumoylation and transglutamination are also important for the ATXN1 activity (see references in Chapter 1). Using an *in silico* approach to predict the wt and expanded ATXN1 interacting surfaces with 71 interacting partners, I observed that the expanded ATXN1 shows larger number of interactions than the wt form. Furthermore, I show the importance of the AXH domain for protein binding, and the role of the N-terminal region of ATXN1 in the binding differences of the wt and expanded ATXN1. In this region, I identified putative amino acid motifs related with transcription, cell communication, phosphorylation, sumoylation, and RNA binding. The generality of these findings in other polyQ proteins is here addressed. Since there are 65 polyQ proteins in the human genome, and each of these proteins interact with a varying number of proteins (that can be as many as 422), I have selected a set of 18 (the ATXN1 and six polyQ proteins determining neurodegenerative disorders that present a size compatible with the *in silico* methodology (smaller than 1500 amino acids), plus 11 proteins that share the largest number of interactors with the neurodegenerative disease set; see Material and Methods). Bellow I present a summary description of each of the proteins studied.

ATXN2 is a cytosolic polyQ protein, 1312 amino acids long, encoded by 25 exons. The protein contains an N-terminal polyQ, encoded by the exon 1, in which an expansion of the polyQ to 32, or more repeats, causes Ataxia type 2 (SCA2; Table 1). This protein is widely expressed in neuronal and non-neuronal tissues, and presents high conservation at nucleotide and amino acid level (Jiménez-López and Guzmán, 2014). ATXN2 is involved in endocytosis, modulation of mTOR signals, ribosomal translation, and mitochondrial functions. Three domains have been described, the Sm_like and LsmAD domain that are located in the N-terminus, and the PAM2 domain that is located in the C-terminus. The specific function of ATXN-2 is largely unknown, with reports of a cytoplasmic protein with Golgi association, which may play a role in RNA splicing (Shibata *et al.*, 2000). The involvement of ATXN-2 in RNA stress granules, was suggested because of the interactions with PABPC1 and GW182 (proteins involved in mRNA processing and translation), promoting the degradation of silenced mRNAs

(Kozlov *et al.*, 2010). The assembly of these proteins with ATXN2 is made by the N-terminal LsmAD domain and a PAM2 motif (Satterfield and Pallanck, 2006). Yeast two-hybrid analyzes using human Ataxin-2 as bait, revealed partners such the brain proteins endophilin A1 and A3, that belong to the endocytosis endophilin family, involved in synaptic vesicle endocytosis. GST pull-down assays experiments showed that the interaction is made through binding proline-rich domains in Ataxin-2, with the SH3 domain of endophilin proteins, at the endoplasmic reticulum. Interactions with other endocytic proteins such as the adaptor protein CIN85 and the ubiquitin ligase c-Cbl have been also reported. Co-immunoprecipitation experiments showed that ATXN2 and TDP-43 associate in a complex that depends on RNA (Elden *et al.*, 2010). ATXN2 has a role in translation regulation, stress granules in the endocytosis complex and to affects EGF receptor trafficking (Nonhoff *et al.*, 2007, Jiménez-López and Guzmán, 2014). According to EvoPPI (Vázquez *et al.*, 2019), 82 proteins have been reported to interact with Ataxin-2.

ATXN3 protein has 370 amino acids and is constituted by a N-terminal Josephin Domain (JD/JOS1), mainly involved in protein-protein interactions and a C-terminal unstructured tail with a polyQ tract (Deriu *et al.*, 2014). When the polyQ tract is elongated over a threshold of 40 glutamines, this protein is associated with Ataxia type 3/Machado-Joseph disease (SCA3), a rare disorder also known as Azorean disease of nervous system, due his high prevalence in Azores population (Raposo *et al.*, 2017). It has a putative nuclear localization signal (NLS), that is highly conserved in evolution. Two ubiquitin interaction motifs (UIMs) are also known, where a broad range of ubiquitinated proteins interact, that accumulate when the proteasome is inhibited. The expression of a catalytically inactive ATXN3 (normal or expanded) causes ubiquitinated proteins to accumulate in cells, even in the absence of proteasome inhibitor, leading to a formation of intranuclear inclusions (Berke *et al.*, 2005). ATXN3 interacts with UV excision repair protein Rad23 and VCP (ATPase valosin-containing protein), forming a protein complex, related to protein degradation machinery (Doss-Pepe *et al.*, 2003). Thus, function of ATXN-3 is mainly as a de-ubiquitinating enzyme. Immunoprecipitation studies results revealed that HAP (a component of the stigmoid body (STB) involved in neuroprotection of proteins such as HTT, AR and TFII), interacts with the normal and expanded form of ATXN3, through the Josephin domain, possibly modifying their physiological function (Takeshita *et al.*, 2011). Interaction with the Chk1 protein, which is known to be linked to DNA repair, has also been reported. When the repair is done, Chk1 dissociates from

ATX3, binding E3 ligase, which causes Chk1 proteasomal degradation, thus suggesting linkage of ATXN3 to the ubiquitination-proteasome system (Tu *et al.*, 2017). The interaction with E3 ligase and E2 ligase, an deubiquitinating enzymes, and indirectly with parkin self-ubiquitination protein was also reported (Durcan *et al.*, 2012). ATXN3 was also described to be a partner of transcriptional regulators and factors, regulating the transcription of certain target genes, as revealed by the interaction with FOXO4 transcription factor, activating the transcription of the manganese superoxide dismutase gene (Araujo *et al.*, 2011). Furthermore, interaction between ATXN3 and the histone acetyltransferases cAMP-response-element binding protein, inhibits transcription by these activators. Interaction of wt and expanded polyQ ATXN3 with more than fifty proteins of the quality control system and mitochondria (such as UBB, STUB1, COA7, HSPH1, HSPA and CDK4) have been reported (Kristensen *et al.*, 2018). Ataxin-3 has been associated with transcription repression via interaction of the ubiquitin-interacting motifs with histone deacetylation, through recruitment of histone deacetylase 3 and the nuclear receptor corepressor that binds to the target DNA sequences in specific chromatin regions of the GATA-2 gene promoter (Evert *et al.*, 2006). ATXN3 also interacts at the Josephin domain with Bcl-XL protein and indirectly with Bax forming a protein complex that has a protective role against cellular oxidative stress induced by H₂O₂ (Zhou *et al.*, 2013). Using EvoPPI (Vázquez *et al.*, 2019), a total of 100 proteins partners are found to interact with ATXN3, in agreement with Lim *et al.* (2006).

ATXN7 protein is widely expressed in the cytoplasm and nucleus of neural and non-neural tissues, is 892 amino acids long, encoded by 13 exons, with CAG repeat located in exon 3. Expansions above 34 of the N-terminal polyQ tract causes SCA7 (Ataxia type 7; Table 1; (Fan *et al.*, 2014). This protein contains three nuclear localization signals (NLS) and one nuclear export signal (NES) (Duncan *et al.*, 2013). ATXN7 also contains a domain characterized by an atypical zinc- finger, known to be involved in protein-DNA interaction (McCullough *et al.*, 2012). Association of ATXN7 with Cbl-associated proteins have been reported, and thus ATXN7 plays a role in neuronal intranuclear inclusions in expanded ATXN7 (Lebre *et al.*, 2001). Co-immunoprecipitation analysis revealed the ATXN7 interaction with CRX protein (cone-rod homeobox protein) at the NLS domain of the ATXN7 protein, but the polyQ also plays a role in this interaction. The expanded ATXN7 causes a drastic suppression of CRX transactivation and binding activity, triggering cone-rod dystrophy disorder. ATXN7 is also a subunit of the GCN5 (general control of amino acid synthesis protein 5), histone acetyltransferase-

containing coactivator multicomplexes interacting with TFTC (a (TATA)-binding protein TBP-associated factor-containing complex), the transcriptional adaptor 1, TBP-associated factor 9 and the STAGA complex (GCN5 acetyltransferase complex). ATXN7, ATXN7L3, USP22, and ENY2 are crucial cofactors for the transcriptional activity by nuclear receptors that, together with STAGA complex, has a role in histone deubiquitinase function (Zhao *et al.*, 2008). The molecular pathogenesis may be caused by polyQ-expanded Ataxin-7 deregulated TFTC/STAGA, suggesting a “transcriptionopathy” (Helmlinger *et al.*, 2004, Fan *et al.*, 2014). ATXN7 is also associated to microtubules, by interacting with proteins such as ACTN1 that stabilizes the cytoskeletal of the cell (Nakamura *et al.*, 2012). Using EvoPPI (Vázquez *et al.*, 2019), 97 proteins are reported to interact with Ataxin-7.

TBP has a length of 339 amino acids, and is an important general transcription initiation factor and the DNA-binding subunit of RNA polymerase II TFIID (transcription factor II D, the multi-subunit complex crucial for the expression of most genes) (Koide *et al.*, 1999). Expansion of the polyQ tract above 42 glutamines, located in the N-terminal, is associated with ataxia 17 (SCA17; (Fan *et al.*, 2014, Tsagalioti *et al.*, 2018). Initiation of the transcription of several genes by RNA polymerase II requires the TFIID to assemble several initiation factors (such as BRF1, GTF2A/B/F1/E1, TAF1/1C/1D, TAF2, TAF4, TAF5, TAF6, TAF7, TAF8, TAF9, TAF10 and TAF13, ZNF76) to form a preinitiation multicomplex. This macromolecular complex requires also the interaction with cleavage-polyadenylation specificity factors types one, two and three (CPSFs) (Dantonel *et al.*, 1997, Lim *et al.*, 2006). In a SCA17 representative PC12 cell model binding of TBP to DNA is reduced (Friedman *et al.*, 2009). Several target specific cofactors have been also reported to bind TBP protein, such as the MBF-1 (multiprotein bridging factor) that modulate the activity of fushi tarazu factor 1, a nuclear receptor originally implicated in *Drosophila* development (Brendel *et al.*, 2002, Wang *et al.*, 2008). In vivo, co-immunoprecipitation assays showed that the phosphoprotein NS5A (HCV-encoded nonstructural protein 5A) forms a heteromeric complex with TBP and tumor suppressor protein p53 inhibiting their DNA binding. Interactions with p53 and ERCC3 (p53-excision repair cross complementing factor 3) were also showed to be inhibited, and thus, TBP plays a role in control of a cellular growth and proliferation pathway (Qadri *et al.*, 2002). Using EvoPPI (Vázquez *et al.*, 2019), it can be seen that TBP interacts with 199 proteins.

ATN1 (also known as DRPLA) is a transcriptional corepressor localized in the nucleus and cytoplasm of neurons in the human nervous central system (Yamada, 2010, Xu *et al.*, 2015). The polyQ expansion is located in exon 5 of the ATN1 gene, and this repeat is associated with DRPLA, also known as Haw River syndrome and Naito–Oyanag (Table 1). ATN1 contains a nuclear localizing signal (NLS) in the N-terminal and a nuclear export signal (NES) in the C-terminal domain. Mutation of the NES region changed the ATN1 localization, causing its accumulation in the nucleus. The functional domains of this protein includes BAH (bromo adjacent homology), ELM2 (EGL-27 and MTA1 homology 2), SANT (SWI3/ADA2/N-CoR/TFIIIB), RE-repeat (arginine-glutamic acid dipeptide-repeat), and ZnF-GATA. These domains are linked to a diversity of functions, including protein interaction mediation, transcriptional regulation and cell signaling (Wang and Tsai, 2008). In disease, there is abnormal aggregation of ATN1 with other proteins including transcription factors (such as TBP, TAF, CREB, and CBP), inducing transcriptional abnormalities and triggering of progressive neuronal degeneration (Yamada *et al.*, 2001). Mutation may inhibit the ability of ATN1 to associate with activator complex component, leading to transcription repression (Shen *et al.*, 2007). A report showed also an interaction with transcription repressor ETO8 to recruit and repress histone acetylase CBP transcription (Wang and Tsai, 2008). Five Atrophin-1 interactors (AIP), (1/2/3/4/5) that belong to two WW family domain proteins, bind also to ATN1, suggesting a common function between Huntington disease and DRPLA (Wood *et al.*, 1998). TLX protein (Orphan nuclear receptor) has been also showed to interact with ATN1, forming a complex that can promote tissue-specific proliferation and differentiation, to prevent retinal malformation and degeneration (Zhang *et al.*, 2006). DRPLA protein is also classified as a phosphoprotein. RBPMS (RNA-binding protein with multiple splicing), a member of RNA-binding proteins family, is a known ATXN1 genetic modifier. A total of over 94 proteins were identified to interact with ATN1, with a wide range of function, including transcription regulation factors (NOC2L and CBFA2T2), transport (SLC4A4, SLCO6A1 and PLSCR1), signaling (LTBP1 and LAIR1/2), T-cell activation (NCK2) and Androgen biosynthesis (HSD17B2) (Lim *et al.*, 2006). Using EvoPPI web tool (Vázquez *et al.*, 2019) I retrieved 136 interactions for the human ATN1 protein.

AR protein is a nuclear receptor that belongs to the steroid/thyroid hormone receptor family. After binding ligand, either testosterone (T) or dihydrotestosterone (DHT), AR translocates to the nucleus and regulates the expression of target genes. It

is encoded by 8 exons, and exon 1 encodes an amino-terminal transactivation domain (NTD) that contains the CAG trinucleotide repeat associated with spinal and bulbar muscular atrophy (SBMA), also known as Kennedy's disease, when composed by more than 37 repeats (Fan *et al.*, 2014). Similar to the other polyQ diseases, the expanded protein has abnormal interactions, causing aggregation with other proteins and possibly triggering the formation of nuclear inclusions in tissues where it is expressed. The loss of its function caused by this expansion leads also to a transcriptional dysregulation and other not well-defined mechanisms. This loss-of-function explains the symptoms manifested, such as muscle atrophy (Breza and Koutsis, 2018). This protein is part of the family of nuclear steroid receptors, that includes glucocorticoid receptor and the progesterone receptor, which are all involved in transactivation of specific genes. These receptors share a common function due to the high homology of a conserved characteristic LBD and DBD domains (Mangelsdorf *et al.*, 1995). A more variable NTD and hinges domain (more precisely the (629)RKLLKK(633) motif) is also present, being involved in mediation of Androgen receptor stability, the nuclear localization, the DNA binding, the transactivation potential and the intranuclear mobility (Tanner *et al.*, 2010). Therefore, this protein is known to be a nuclear hormone receptor that is responsive to G-protein coupled receptor signaling, binding to DNA and regulating gene expression via modulation of chromatin complexes. Thus, AR is a regulatory protein, functioning in many different physiological processes and many reproductive functions. It is reported also that AR protein plays a role in the assembly of signaling complexes at the synapse by interacting with membrane-associated guanylate kinases (Hands and Wyttenbach, 2010). AR interaction partners are involved in the transcriptional regulation of different genes, including other nuclear receptors, ubiquitination, cell development and signaling pathways. An interaction between the MRF1 (modulator recognition factor 1) and the AR was reported, that functions as co-repressor of the transcription, that regulates cell growth and development (Georgescu *et al.*, 2005). Cryptochromes 1, a circadian clock component, interacts also with the AR as a co-repressor, by a ligand-dependent interaction, to regulate gene transcription, that guarantee homeostasis (Lamia *et al.*, 2011). The association of the C-terminal region of AR with CCAR1 (cell cycle and apoptosis regulator 1), function as a co-activator to enhance transcriptional activity of AR, has been reported as well (Seo *et al.*, 2013). It has also been described the role of DAPK3 (ZIP kinase) as coactivator with AR in the regulation of cell development and transcription (Leister *et al.*, 2008). CDK9 (Cyclin-dependent kinase 9) has been demonstrated to interact with AR, through phosphorylation on S81, regulating AR

transcriptional activity and prostate cancer cell growth (Gordon *et al.*, 2010). Immunoprecipitation screens identified PIAS protein (protein inhibitor of activated STAT) to interact with AR, repressing its transcription activity. Contrary, DJBP (DJ-1-binding protein) restored transcription activity, positively regulating AR transcription via binding of DJBP to the DNA-binding domain (Niki *et al.*, 2003). The involvement of AR in a nuclear protein complex with FHL2 (four-and-a-half Lim domains protein 2) and filamin, has been shown to function as a co-activator of AR transcription, linked to cell migration and oncogenesis (McGrath *et al.*, 2013). The coiled coil domain of HIP1 (Huntingtin interacting adaptor protein 1), by association with AR, induces cellular transformation through transcription enhancement (Mills *et al.*, 2005). SPEDF (SAM pointed domain-containing Ets transcription factor), that binds to the DNA-binding domain of AR, is involved in cellular differentiation. Interaction of ZIMP7 and ZIMP10 with AR, at the NTD domain, plays a role in enhancing AR sumoylation, DNA synthesis and chromatin modification (Beliakoff and Sun, 2006). Several other transcription regulators are described to interact with the AR, such as the HOT (HOX transcript antisense RNA) or MAGEA11 (melanoma-associated antigen 11) that binds to NTD domain of AR, LPXN (leupaxin) that binds to LBD domain, MAK (male germ cell-associated kinase) that binds to DBD domain and the KAT7 (lysine acetyltransferase 7) that binds to both LBD and DBD domains (Iizuka and Stillman, 1999, Bai *et al.*, 2005, Zhang *et al.*, 2015, Shi *et al.*, 2016). AR also interacts with nuclear receptors, including NCOA1 (nuclear receptor coactivator 1), NCOA2, NCOA3, NCOA4 or SHP (small heterodimer partner), by binding to the LBD domain of AR, activating transcription of the target genes (Estebanez-Perpina *et al.*, 2005, Tetel and Acharya, 2013). The involvement of AR in ubiquitination transcription regulation was identified through interaction with several proteins of this system, including the E3 ubiquitin-protein ligase RNF14 and RNF16 and the tripartite motif-containing protein 24 and 68 that associates with C-terminal LBD domain and the ubiquitin-specific protease 26 (USP26) (Dirac and Bernards, 2010, Li *et al.*, 2014). The involvement of AR in signaling transcription regulation was also described by the association with many proteins including the Ran protein (Ras-related nuclear protein) that binds to the AR DBD and NTD domains. A total of 136 interactions have been described, according to EvoPPI (Vázquez *et al.*, 2019).

For the proteins non-related to the neurodegenerative diseases, the PPI, especially when looking to the expanded forms, is poorly studied. One protein non-associated with neurodegenerative disease is TFCEP2 (alpha-globin transcription factor

CP2 protein). This protein has 450 amino acids, containing several important domains for the interaction. These include the C-terminal polyQ domain and the N-terminal CP2 domain (Li *et al.*, 2016). Indeed, this protein belongs to the family of CP2 DNA-binding transcription factors that are ubiquitous activators of transcription (Rodda *et al.*, 2001). The Sterile alpha motif (SAM domain) is also present in the C-terminal region of the protein. This domain, widespread across the eukaryotes proteome, has been described as having a key role in the RNA binding function (Kim and Bowie, 2003). The transcription gene regulation of this protein has been linked to several key roles in the brain, including the regulation of acid-sensing ion channel 2a, glycogen synthase kinase 3 β and cAMP responsive element binding protein 1 (Lau *et al.*, 1999, Zhang *et al.*, 2017). This protein dysregulation has been linked to a gain-of-function in several diseases, including Alzheimer and major depressive disorders (Taracha *et al.*, 2018). In EvoPPI (Vázquez *et al.*, 2019) 125 interactions are reported.

THAP11 (THAP domain-containing protein 11) is a polyQ protein non-related with neurodegenerative diseases that possesses 314 aminoacids, being an essential factor involved in ES cell pluripotency and cell growth (Lian *et al.*, 2012). This protein contains several essential domains, an N-terminal polyQ tract and a characteristic THAP domain (Li *et al.*, 2016). This domain has been associated with DNA-binding function, presenting similarities with the *Drosophila* DNA-binding domain (DBD) of the P-element transposase (Roussigne *et al.*, 2003). This domain is zinc-dependent, recognizing a consensus distinctive DNA target sequence (Clouaire *et al.*, 2005). Another domain in the N-terminal region, is the Glucocorticoid receptor-like domain (Li *et al.*, 2016). This protein has been described as a possible candidate to a polyglutamine disorder. Indeed, *In silico* and polymorphism analyzes, showed that this protein exhibit features characteristic of genes associated with repeat disorders. The conformational change in the regions that are "natively unfolded" happened over a threshold of 40 glutamines (Pandey *et al.*, 2004). THAP11 mediated gene regulation and its has been found to be associated with the transcriptional regulator host cell factor 1 (HCF-1) (Parker *et al.*, 2012). The role of THAP11 in transcription regulation is also supported by its binding to PCBP1 (Lian *et al.*, 2012). 70 interactions are reported in EvoPPI(Vázquez *et al.*, 2019).

MEF2A (Myocyte-specific enhancer factor 2A) is also a polyQ protein transcription factor non-associated with neurodegenerative disorders, which possesses 366 aminoacids. This protein contains a C-terminal polyQ stretch, usually with ten glutamines (Li *et al.*, 2016). A serum response factor like (SRF) domain is also found in

this protein, which is linked to DNA-binding (Maeda *et al.*, 2002). This protein also contains an N-terminal Holliday junction regulator protein repeat (HJURP) of unknown function (Sanchez-Pulido *et al.*, 2009). In Evo PPI (Vázquez *et al.*, 2019) 73 interactions are reported for this protein.

CELF3 (CUGBP Elav-like family member 3) protein, with 388 aminoacids, is RNA-binding protein involved in the regulation of pre-mRNA alternative splicing (Ladd *et al.*, 2001). It contains a Interdomain polyQ tract and over three RNA recognition motifs. These motifs are located wider on the protein that play a main role in translation regulation and post-transcriptional processing of pre-mRNA (Glisovic *et al.*, 2008). In Evo PPI (Vázquez *et al.*, 2019) 12 interactions are reported for this protein.

NUMBL (Numb-like protein) is a polyQ protein crucial in neuronal development, with 568 aminoacids. This protein possesses a N-terminal phosphotyrosine binding (PTB) domain and a C-terminal polyQ tract. The PTB domain is known to be involved in modulation of signaling complexes involved in important functions, including neuronal development (Uhlik *et al.*, 2005). There is also a putative NUMB domain, generally found in the NUMB protein family adjacent to the PTB domain. In neurogenesis, through interaction with TAB2 and TRAF6 proteins, NUMBL protein maintains neural progenitor cells (Ma *et al.*, 2008, Zhou *et al.*, 2010). 38 interactions are reported in Evo PPI (Vázquez *et al.*, 2019).

PHC1 (Polyhomeotic-like protein 1) is a major component of a Polycomb group multiprotein PRC1-like complex that repress the transcription of several genes throughout development. This protein possesses 946 aminoacids, an N-terminal polyQ tract and a characteristic C-terminal SAM domain. The SAM domain is generally involved in RNA binding (Kim and Bowie, 2003). In Evo PPI (Vázquez *et al.*, 2019) 45 interactions are reported.

SATB1 (DNA-binding protein SATB1), with 691 aminoacids, is a polyQ protein linked also to transcription regulation, being a crucial factor that contributes to the initiation of X inactivation that occurs during embryogenesis. For that function the DNA-binding domain, Lambda repressor-like is crucial (Wang *et al.*, 2009). Other domains of this protein include an Interdomain polyQ tract and a Sec23/Sec24 beta-sandwich domain. The Sec23/Sec24 beta-sandwich domain function is largely unknown, although it may be involved in the transport of the protein from the endoplasmic reticulum to the

Golgi complex (Lederkremer *et al.*, 2001). In EvoPPI (Vázquez *et al.*, 2019) 56 interactions are reported.

RUNX2 (Runt-related transcription factor 2), with 485 aminoacids, is a main transcription factor involved in skeletal morphogenesis and brain development (Jeong *et al.*, 2008). This protein contains a N-terminal polyQ tract and a runt domain that confers DNA binding ability to a specific sequence. The beta-sandwich domain of Sec23/24 is also found in the N-terminal region of this protein. There are 64 interactions reported in EvoPPI (Vázquez *et al.*, 2019)

POU6F2 (POU domain, class 6, transcription factor 2), with 655 aminoacids, is a putative transcription factor likely involved in the differentiation of amacrine and ganglion cells (Zhou *et al.*, 1996). The N-terminal polyQ domain and the POU DNA binding domain, are fundamental for the transcription function (Andersen and Rosenfeld, 2001). In EvoPPI (Vázquez *et al.*, 2019) 133 interactions are reported.

ABCF1(ATP-binding cassette sub-family F member 1), with 807 aminoacids, is a polyQ protein involved in the efficient mRNA translation initiation (Paytubi *et al.*, 2009). The N-terminal polyQ tract and a P-loop containing nucleoside triphosphate hydrolase domain, are important in nucleotide-binding protein folding (Leipe *et al.*, 2004). In EvoPPI (Vázquez *et al.*, 2019) 74 interactions are reported.

VEZF1 (Vascular endothelial zinc finger 1) is a putative transcription factor with 512 amino acids, that specifically binds to a CT/GC-rich region. This protein contains a C-terminal polyQ tract and over four zinc-binding repeats domains. These domains are fundamental for protein-protein and protein-nucleic acid binding (Koyano-Nakagawa *et al.*, 1994). In EvoPPI (Vázquez *et al.*, 2019) 51 interactions are reported.

Materials and Methods

1. Selection of the 18 polyQ proteins to be studied

For the 65 polyQ proteins in the human proteome (see Chapter 1), I have recorded the size of the protein, the size of the polyQ, and its location in the protein (according to PolyQ 2.0; (Li *et al.*, 2016). Since CACNA1A protein having a polyQ tract of ten glutamines is already associated with a neurodegenerative disease (Schaefer *et al.*, 2012), I considered, as described in Chapter 1, only polyQ proteins that have a polyQ tract with the minimum of 10 consecutive uninterrupted glutamines. Moreover, since the polyQ tract of ATXN3 is defined by consecutive glutamines with an interruption of a methionine (Kawaguchi *et al.*, 1994), when recording the size of the polyQ, I considered the possibility of the polyQ tract to have at least one interruption. Since the maximum size of the wt polyQ in the nine neurodegenerative polyQ proteins is 44 (Table 1), I considered a size of 50 uninterrupted glutamines the pathological minimum polyQ length. Indeed, this value is above the minimum size reported in the literature as being pathological for all polyQ diseases, excepted for SCA3.

Because the *in silico* methodology can be applied only to proteins smaller than 1500 amino acids, 13 proteins were discarded (Figure 7). Furthermore, one protein that presents no interaction in EvoPPI was also removed from the analyses. Seven proteins with more than one polyQ tract were excluded from further analyses, since it would be difficult to infer which polyQ tract influences the interactions and the binding. To address expression in the brain I used the expression atlas database (Petryszak *et al.*, 2016). To identify possible diseases associated with polyQ proteins, I used OMIM database (Amberger *et al.*, 2015). In this study, we selected seven of the polyQ degenerative diseases (ATXN1, ATXN2, ATXN3, ATXN7, TBP, ATN1, and AR). From the remaining 37 polyQ proteins, 17 present one interactor in common with two of the polyQ degenerative diseases, allowing to make comparisons of the same interaction with different polyQ proteins. Nevertheless, for four out of the 17 selected proteins, CPORT failed to predict the active and the passive residues, and thus, these proteins have been also excluded from further analyses. Also, for two proteins I failed to predict the docking, using HADDOCK, and, consequently, these also have also been excluded. In conclusion, I have selected 18 polyQ proteins for further study (Figure 7).

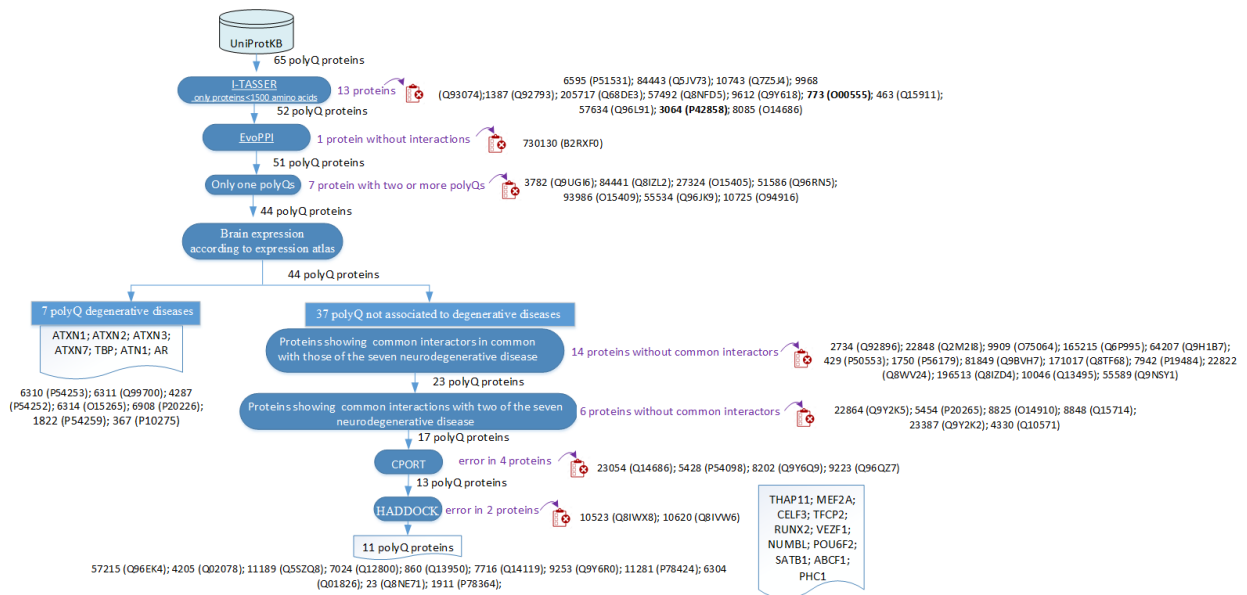


Figure 7. Flow-chart describing the process used to select the 18 polyQ proteins studied.

2. Selection of the polyQ proteins interactors

The interactome of the 18 polyQ proteins here studied was obtained using EvoPPI (Vázquez *et al.*, 2019). Although 561 interactors were obtained, only 44 show interactions with two polyQ neurodegenerative disease proteins plus one polyQ protein not associated to these diseases (Figure 8). Of these, four proteins are larger than 1500 amino acids, and thus were excluded from this study. I also excluded from this study nine ubiquitin related proteins, that show interactions with large number of proteins. The three proteins that present a polyQ were also excluded from the analyses. I also removed the non-interacting histones, since only histone H3.1 subunit is reported as interacting with polyQ proteins (Chou *et al.*, 2014). Self-interactions between proteins were also excluded from analyzes. Therefore, 20 common interactors were selected for further analyses. Of these, two were not studied since one produces an error in CPORT and the other in HADDOCK.

An evolutionary approach to understand how interactomes change when expansions of the polyglutamine (polyQ) tracts occur: implications for the human polyQ diseases

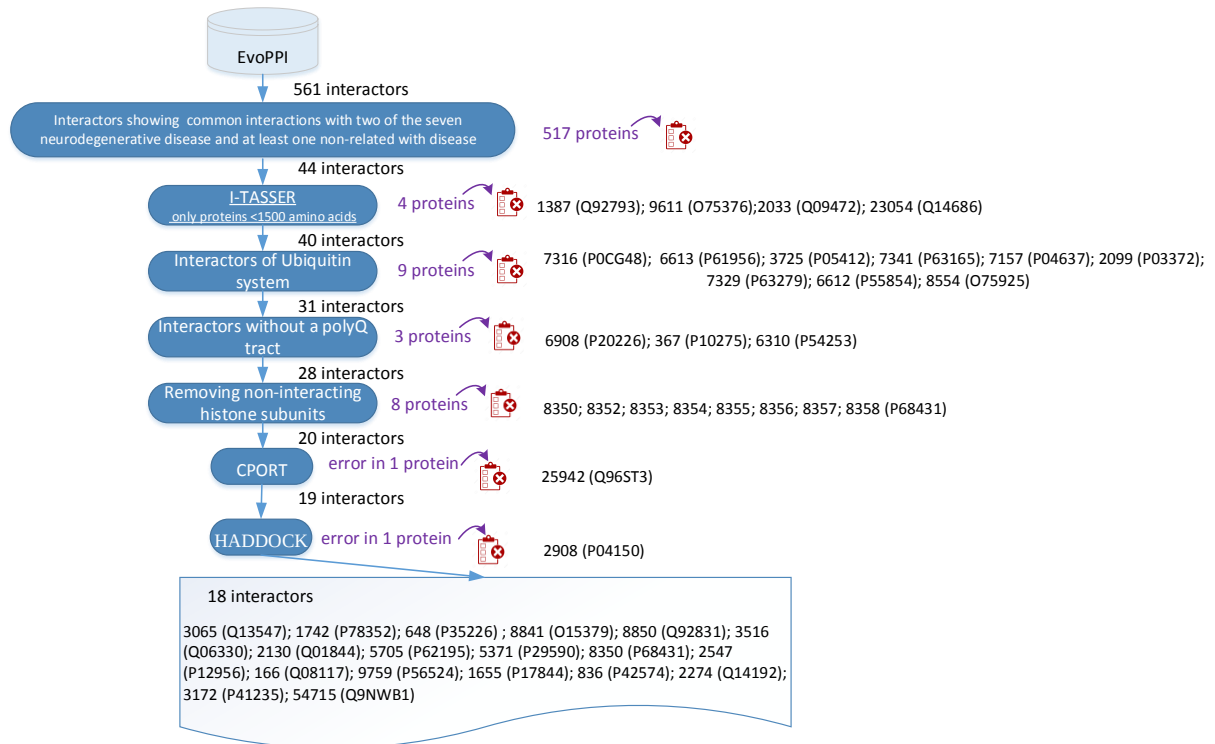


Figure 8. Flow-chart used to select the interactors for the 18 common interactors for the polyQ proteins here studied

A total of 118 interactions were analyzed in this study (Table 6). According to PANTHERdb the polyQ proteins and interactors here studied are mainly involved in binding, catalytic activity, and transcription regulation activity (Figure 9).

Table 6. The polyQ proteins and interactors here studied. In brackets are the protein name, GeneID, and UniProtID. Underlined are the polyQ proteins associated with neurodegenerative diseases.

PolyQ proteins	Interactors
<u>AR (Androgen receptor; 367; P10275)</u>	HDAC1 (Histone deacetylase 1; 3065; Q13547) XRCC6 (X-ray repair cross-complementing protein 6; 2547; P12956) HDAC4 (Histone deacetylase 4; 9759; P56524) LIM (Four and a half LIM domains protein 2; 2274; Q14192) CASP3 (Caspase-3; 836; P42574) KAT2B (Histone acetyltransferase KAT2B; 8850; Q92831) AES (Amino-terminal enhancer of split; 166; Q08117) HNF4A (Hepatocyte nuclear factor 4-alpha; 3172; P41235) HDAC3 (Histone deacetylase 3; 8841; O15379)

<u>ATXN1 (Ataxin-1; 6310; P54253)</u>	RBFOX1 (RNA binding protein fox-1 homolog 1; 54715; Q9NWB1) RBPJ (Recombining binding protein suppressor of hairless; 3516; Q06330) LIM (Four and a half LIM domains protein 2; 2274; Q14192) HDAC4 (Histone deacetylase 4; 9759; P56524) HIST1H3D (Histone H3.1 subunit d; 8351; P68431) PML (Protein PML; 5371; P29590)
<u>ATXN2 (Ataxin-2; 6311; Q99700)</u>	DLG4 (Disks large homolog 4; 1742; P78352) BMI1 (Polycomb complex protein BMI-1; 648; P35226) DDX5 (Probable ATP-dependent RNA helicase DDX5; 1655; P17844)
<u>ATXN3 (Ataxin-3; 4287; P54252)</u>	HDAC3 (Histone deacetylase 3; 8841 ; O15379) KAT2B (Histone acetyltransferase KAT2B; 8850; Q92831) XRCC6 (X-ray repair cross-complementing protein 6; 2547; P12956) PSMC5 (26S proteasome regulatory subunit 8; 5705; P62195) PML (Protein PML; 5371; P29590) EWSR1 (RNA-binding protein EWS; 2130; Q01844)
<u>ATXN7 (Ataxin-7; 6314; O15265)</u>	PSMC5 (26S proteasome regulatory subunit 8; 5705; P62195) HIST1H3D (Histone H3.1 subunit d; 8351; P68431)
<u>ATN1 (Atrophin-1; 1822; P54259)</u>	DLG4 (Disks large homolog 4; 1742; P78352) RBFOX1 (RNA binding protein fox-1 homolog 1; 54715; Q9NWB1) EWSR1 (RNA-binding protein EWS; 5705; Q01844) AES (Amino-terminal enhancer of split; 166; Q08117) CASP3 (Caspase-3; 836; P42574)
<u>TBP (TATA-box-binding protein; 6908; P20226)</u>	BMI1 (Polycomb complex protein BMI-1; 648; P35226) HDAC4 (Histone deacetylase 4; 9759; P56524) DDX5 (Probable ATP-dependent RNA helicase DDX5; 1655; P17844) RBPJ (Recombining binding protein suppressor of hairless; 3516; Q06330) HDAC1 (Histone deacetylase 1; 3065; Q13547) HNF4A (Hepatocyte nuclear factor 4-alpha; 3172; P41235)
TFCP2 (Alpha-globin transcription factor CP2; 7024; Q12800)	DDX5 (Probable ATP-dependent RNA helicase DDX5; 1655; P17844) HDAC1 (Histone deacetylase 1; 3065; Q13547)

MEF2A (Myocyte-specific enhancer factor 2A; 4205; Q02078)	CASP3 (Caspase-3; 836; P42574) HDAC3 (Histone deacetylase 3; 8841; O15379) DLG4 (Disks large homolog 4; 1742; P78352) HDAC4 (Histone deacetylase 4; 9759; P56524)
SATB1 (DNA binding protein SATB1; 6304; Q01826)	RBPJ (Recombining binding protein suppressor of hairless; 3516; Q06330) KAT2B (Histone acetyltransferase KAT2B; 8850; Q92831)
ABCF1 (ATP-binding cassette sub-family F member 1; 23; Q8NE71)	PSMC5 (26S proteasome regulatory subunit 8; 5705; P62195)
CELF3 (CUGBP Elav-like family member 3; 11189; Q5SZQ8)	RBFOX1 (RNA binding protein fox-1 homolog 1; 54715; Q9NWB1)
NUMB (Numb-like protein; 9253; Q9Y6R0)	RBFOX1 (RNA binding protein fox-1 homolog 1; 54715; Q9NWB1)
PHC1 (Polyhomeotic-like protein 1; 1911; P78364)	BMI1 (Polycomb complex protein BMI-1; 648; P35226)
POU6F2 (POU domain, class 6, transcription factor 2; 11281; P78424)	LIM (Four and a half LIM domains protein 2; 2274; Q14192)
RUNX2 (Runt-related transcription factor 2; 860; Q13950)	HDAC4 (Histone deacetylase 4; 9759; P56524) LIM (Four and a half LIM domains protein 2; 2274; Q14192) KAT2B (Histone acetyltransferase KAT2B; 8850; Q92831) RBPJ (Recombining binding protein suppressor of hairless; 3516; Q06330)
THAP11 (THAP domain containing protein 11 ; 57215; Q96EK4)	PSMC5 (26S proteasome regulatory subunit 8; 5705; P62195) CASP3 (Caspase-3; 836; P42574) HDAC3 (Histone deacetylase 3; 8841; O15379) HDAC1 (Histone deacetylase 1; 3065; Q13547)
VEZF1 (Vascular endothelial zinc finger 1; 7716; Q14119)	AES (Amino-terminal enhancer of split; 166; Q08117)

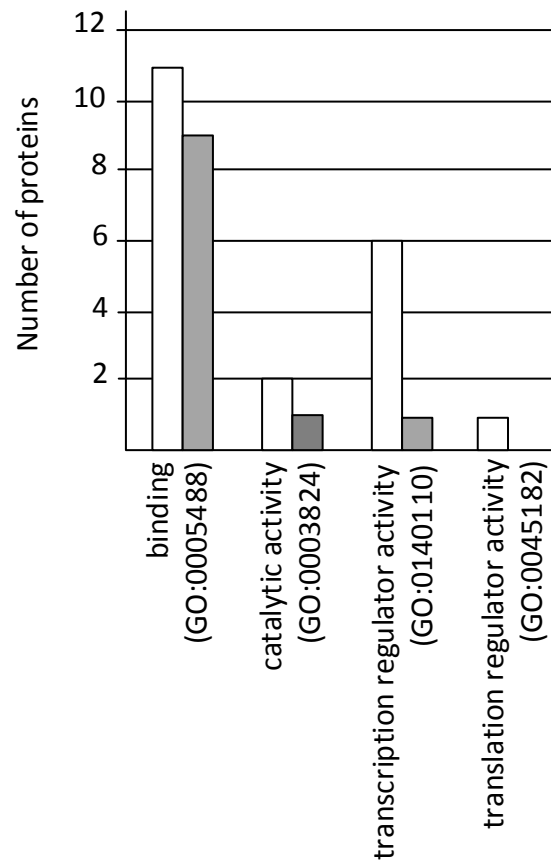


Figure 9. Graphical representation of the gene ontology analyzes made to polyQ proteins (white bars) and their interactors (grey bars), with four major functions, including transcription and translation regulation, binding and catalytic activity.

3. Analyzes of the 3D structure of the selected human polyQ proteins and their interactors

3.1. Protein structures prediction and identification of the interaction residues

To predict the 3D-structure of the proteins, I used the methodology described in the previous chapter in the Figure 1. In summary, I used I-TASSER (Zhang, 2008), to predict the structure of each protein (18 wt and 18 expanded polyQ proteins, plus 18 interactors), and the C-score to obtain a quantitative assessment of the confidence of the I-TASSER models. The structures with higher C-score (higher confidence) were then used for the identification of the interface residues, using CPORT (de Vries and Bonvin, 2011).

3.2. Protein-protein docking

The docking of the polyQ proteins with the known partners was made using HADDOCK2.2 (High Ambiguity Driven protein-protein DOCKing (van Zundert *et al.*, 2016). The clusters of macromolecular complexes that result from the possible dockings are then classified statistically by a Z-score, with the lowest values, being the best predictions (Dominguez *et al.*, 2003).

3.3. Wild-type vs. expanded cluster analyses

To analyze in detail the docking results PISAePDB (Krissinel and Henrick, 2007) was used. PISAePDB take into consideration the structural and chemical properties of molecular interfaces and protein assemblies and the likely dissociation pattern based on physical-chemical models of macromolecular interactions and chemical thermodynamics. The accessible surface area (ASA), that calculates the surface area of a biomolecule that is accessible to a solvent, and buried surface area (BSA) that estimates the size of the interface in a protein interaction, were retrieved, giving insights on the number and which residues are involved on the specific protein interaction with a given interactor (Chakravarty *et al.*, 2013). PISAePDB also identified the number and percentage of interface residues (including those that make hydrogen bonds, and salt bridges), the solvent-accessible area (Å²), and the percentage of the solvent-accessible area of the wt and the expanded form of the different polyQ proteins. From all HADDOCK docking structures, I always selected for further study the structure presenting the highest probability of having a protein-protein interaction. I also considered that if the number of interface residues is greater for the docking with the expanded polyQ protein than for the wild-type polyQ, that there is a higher binding affinity with the former, and vice versa. I assumed no preference if the value is identical for the wild-type polyQ and expanded polyQ docking structures.

Results

No technical bias is observed in the *in silico* method used in this approach.

In order to test for possible technical biases, I compared the I-TASSER C-Score of the 18 polyQ proteins of the structure for wt and expanded forms. The average I-TASSER C-Score is -1,345 (standard deviation is 1,04287) for the wt form and -1,49 (standard deviation is 1,31651) for the expanded form. The C-Score usually ranges from -5 to 2, higher values signifying models with higher confidence. There is no tendency for the I-TASSER C-Scores of wt and expanded forms being different (non-parametric Sign test; $P=1,000$). The non-parametric Spearman's correlation coefficient between I-TASSER C-Scores for wt and expanded forms is 0,628 ($P=0,005$), thus, as expected the confidences of the wt and expanded structures are correlated. The average HADDOCK Z-Score is -1,147 (standard deviation is 0,6942; $N=59$) for those interactions involving wt forms, and -1,200 (standard deviation is 0,7719; $N=59$) for those involving expanded forms. The HADDOCK Z-Score is not statistically different between the two groups (non-parametric Sign test; $Z=-0,408$; $P=0,683$). The non-parametric Spearman's correlation coefficient between HADDOCK Z-Score for wt and expanded forms interactions is 0,013 ($P=0,925$), showing that the quality of the docking is not related in wt and expanded forms, although there is no tendency in the quality of the inferred structures (see above).

Interestingly, protein size is positively correlated with the I-TASSER C-Score, showing that large proteins tend to be predicted with better confidence in wt form (Non-parametric Spearman's correlation coefficient: 0,481; $P=0,043$; $N=18$) and in the expanded form (Non-parametric Spearman's correlation coefficient: 0,553; $P=0,017$; $N=18$). Nevertheless, the protein size is not correlated with the HADDOCK Z-Score, for the interactions with the wt form (Non-parametric Spearman's correlation coefficient: -0,146; $P=0,269$; $N=59$) and with the expanded form (Non-parametric Spearman's correlation coefficient: -0,055; $P=0,677$; $N=59$). The I-TASSER C-Score of the wt polyQ proteins showed no correlation with the HADDOCK Z-Score for the interactions involving these proteins (Non-parametric Spearman's correlation coefficient: -0,189; $P=0,152$; $N=59$). The I-TASSER C-Score of the expanded forms of the polyQ proteins showed no correlation with the HADDOCK Z-Score for the interactions involving these proteins (Non-parametric Spearman's correlation coefficient: -0,196; $P=0,136$; $N=59$).

Regarding the interactors ($N=18$), the average I-TASSER C-Score is -0,87 (standard deviation is 1,41901). In contrast to the observation made for polyQ proteins,

the protein size of the interactors is not correlated with their I-TASSER C-Score (Non-parametric Spearman's correlation coefficient: -0,028; P=0,913; N=18). The interactor protein size is not correlated with the HADDOCK Z-Score, when involving both the wt form (Non-parametric Spearman's correlation coefficient: -0,012; P=0,927; N=59) and the expanded form (Non-parametric Spearman's correlation coefficient: 0,016; P=0,903; N=59). Therefore, there are no major technical biases in the approach here used.

The binding preference for the expanded form is not common to all the polyQ proteins.

In Chapter 1, I show that the expanded ATXN1 shows more residues at the binding interface than the wt form. To address the generality of this pattern, I used a set of 59 interactions involving 18 polyQ proteins (seven of them involved in disease) and 18 interactors. Only 31 (52%) of interactions studied showed a higher interaction in the expanded form, 27 (46%) interactions showed higher interaction in the wild-type form, and one interaction (2%) had similar interaction in both forms (Table 7).

Table 7. PolyQ proteins interactions with binding preference for the wt and expanded forms. Underlined are polyQ proteins associated with neurodegenerative diseases.

PolyQ (Gene ID; Uniprot ID)	Interactor (Gene ID; Uniprot ID)	Binding preference
<u>AR</u> (367; P10275)	HDAC1 (3065; Q13547)	expanded
	XRCC6 (2547; P12956)	wt
	HDAC4 (9759; P56524)	wt
	LIM (2274; Q14192)	wt
	CASP3 (836; P42574)	expanded
	KAT2B (8850; Q92831)	expanded
	AES (166; Q08117)	expanded
	HNF4A (3172; P41235)	expanded
	HDAC3 (8841; O15379)	expanded
<u>ATXN1</u> (6310; P54253)	RBFOX1 (54715; Q9NWB1)	expanded
	RBPJ (3516; Q06330)	wt
	LIM (2274; Q14192)	wt
	HDAC4 (9759; P56524)	expanded
	HIST1H3D (8351; P68431)	expanded
	PML (5371; P29590)	expanded
<u>ATXN2</u> (6311; Q99700)	DLG4 (1742; P78352)	wt
	BMI1 (648; P35226)	wt

<u>ATXN3</u> (4287; P54252)	DDX5 (1655; P17844)	expanded
	HDAC3 (8841; O15379)	wt
	KAT2B (8850; Q92831)	wt
	XRCC6 (2547; P12956)	expanded
	PSMC5 (5705; P62195)	-
<u>ATXN7</u> (6314; O15265)	PML (5371; P29590)	expanded
	EWSR1 (2130; Q01844)	expanded
	PSMC5 (5705; P62195)	wt
	HIST1H3D (8351; P68431)	expanded
	DLG4 (1742; P78352)	wt
<u>ATN1</u> (1822; P54259)	RBFOX1 (54715; Q9NWB1)	expanded
	EWSR1 (2130; Q01844)	expanded
	AES (166; Q08117)	expanded
	CASP3 (836; P42574)	expanded
	BMI1 (648; P35226)	wt
<u>TBP</u> (6908; P20226)	HDAC4 (9759; P56524)	wt
	DDX5 (1655; P17844)	wt
	RBPJ (3516; Q06330)	wt
	HDAC1 (3065; Q13547)	wt
	HNF4A (3172; P41235)	expanded
TFCP2 (7024; Q12800)	DDX5 (1655; P17844)	wt
	HDAC1 (3065; Q13547)	expanded
ABCF1 (23; Q8NE71)	PSMC5 (5705; P62195)	expanded
CELF3 (11189; Q5SZQ8)	RBFOX1 (54715; Q9NWB1)	expanded
SATB1 (6304; Q01826)	RBPJ (3516; Q06330)	expanded
	KAT2B (8850; Q92831)	expanded
MEF2A (4205; Q02078)	CASP3 (836; P42574)	wt
	HDAC3 (8841; O15379)	wt
	DLG4 (1742; P78352)	expanded
	HDAC4 (9759; P56524)	expanded
	RBFOX1 (54715; Q9NWB1)	wt
PHC1 (1911; P78364)	BMI1 (648; P35226)	wt
POU6F2 (11281; P78424)	LIM (2274; Q14192)	wt
RUNX2 (860; Q13950)	HDAC4 (9759; P56524)	wt
	LIM (2274; Q14192)	expanded
	KAT2B (8850; Q92831)	wt
	RBPJ (3516; Q06330)	wt
	PSMC5 (5705; P62195)	wt
THAP11 (57215; Q96EK4)	CASP3 (836; P42574)	expanded
	HDAC3 (8841; O15379)	expanded
	HDAC1 (3065; Q13547)	expanded
<u>VEZF1</u> (7716; Q14119)	AES (166; Q08117)	wt

The number of interaction residues involving wt and expanded polyQ proteins is similar (non-parametric Sign test; $Z = -0,394$; $P = 0,694$; $N = 59$). This result holds when we analyze separately the regions before the polyQ tract (non-parametric Sign test; $Z = -0,560$; $P = 0,575$; $N = 59$), at the polyQ region (non-parametric Sign test; $Z = -1,925$; $P = 0,054$; $N = 59$) and after the polyQ tract (non-parametric Sign test; $Z = -0,539$; $P = 0,590$; $N = 59$). A simple classification regarding the position of the polyQ tract such as located in the first or second half of the protein gives the same results. The proteins with a polyQ tract located in the first half of the protein showed a similar binding preference between wt and expanded forms (non-parametric Sign test; $Z = -0,154$; $P = 0,877$; $N = 42$). In the proteins with a polyQ tract located in the second half of the protein, a similar binding preference between wt and expanded forms is also seen (non-parametric Sign test; $P = 0,804$; $N = 17$). Then, I refine the location of the polyQ tract, according to Li *et al.* (2016), in N-terminal (polyQ appears before all Pfam domains), Mid Domain (the polyQ repeat appears in the middle of a Pfam domain or overlaps with a Pfam domain), C-terminal (the polyQ appears after all Pfam domains), and Interdomain (the polyQ appear between the first Pfam and last Pfam domain). When all the interactions are analyzed, the Interdomain and Mid Domain polyQs tend to have a higher binding preference for the expanded form than the polyQ tracts located in the N-terminal and C-terminal regions (Kruskal-Wallis test; $H = 8,668$; $P = 0,034$; $N = 59$). Thus, it is not surprising that no statistically significant difference in binding preferences involving wt and expanded polyQ proteins is observed for the proteins having a N-terminal (non-parametric Sign test; $P = 0,815$; $N = 18$), and a C-terminal polyQ tract (non-parametric Sign test; $P = 0,727$; $N = 8$). For the Mid Domain no statistically significant difference in binding preferences involving wt and expanded polyQ proteins is observed (non-parametric Sign test; $Z = -0,743$; $P = 0,458$; $N = 30$), and only the proteins having a Interdomain polyQ showed a higher binding preference for expanded polyQ proteins (non-parametric Sign test; $P = 0,250$; $N = 3$), as observed for ATXN1. Nevertheless, the latter group has three interactions only.

PolyQ proteins involved in neurodegenerative diseases do not show binding preferences different from that observed for polyQ proteins not involved in these diseases (Mann-Whitney test; $Z = -0,612$; $P = 0,541$; $N = 59$). The composition of the polyQ tract (the interruptions in the polyQ tract with the minimum of 10 consecutive glutamines) is not associated with the binding differences between both forms (Mann-Whitney test; $Z = -0,855$; $P = 0,392$; $N = 59$).

For the proteins with three or more interactors I infer binding preferences, as the difference in number of interface residues in expanded and wt forms for each interaction (Table 8). It should be noted that for ATXN1 although only six proteins have been here studied, the binding preferences are not significantly different from those obtained in Chapter 1 (Fisher's exact test; $P=0.056$). Therefore, I considered binding preferences for percentages higher than 66.0, as observed for ATXN1. Binding preference for expanded form is inferred for ATN1, AR, and THAP11. Binding preference for wt is inferred for TBP, ATXN2, and RUNX2.

Table 8. Percentage of interactions showing binding preferences for wt and expanded for the nine polyQ proteins with three or more interactions studied. Underline are polyQ proteins associated with neurodegenerative diseases

PolyQ protein	Percentage (%) of interactions	
	wt	expanded
<u>AR</u>	33.3	66.6
<u>ATXN1</u>	33.3	66.6
<u>ATXN2</u>	66.6	33.3
<u>ATXN3</u>	33.3	50.0
<u>ATN1</u>	20.0	80.0
<u>TBP</u>	83.3	16.7
MEF2A	50.0	50.0
RUNX2	75.0	25.0
THAP11	25.0	75.0

The C-terminal region explains binding preferences

The number of interface residues in the N-terminal and C-terminal regions according to the location of the polyQ tract in wt and expanded forms are different. The C-terminal region had more interaction residues in the wt form (2185) and the expanded form (2039) when compared to the N-terminal region in wt (1251) and expanded form (1241), but the proportions are similar (Fisher exact test; $P=0.235$). In the wt form, there is a tendency to have a higher number of interactions in the C-terminal region (non-parametric Sign test; $Z=-2.083$; $P=0.037$; $N=59$). In the expanded form, there is no tendency (non-parametric Sign test; $Z=-1.823$; $P=0.068$; $N=59$), although it is a borderline result. When, we compare the interacting residues in the N-terminal and C-terminal regions between both forms, there is no tendency for a higher number of

interaction residues in either form for both the N-terminal (non-parametric Sign test; $Z=-0,280$; $P=0,779$; $N=59$) and the C-terminal (non-parametric Sign test; $Z=-0,270$; $P=0,787$; $N=59$) regions. Nevertheless, when I consider the subset of proteins that show binding preference for the expanded form only in the N-terminal polyQ region (non-parametric Sign test; $Z=0$; $P=1.000$; $N=31$) there is no binding differences between wt and expanded forms, but for the C-terminal polyQ region (non-parametric Sign test; $Z=-2,008$; $P=0,045$; $N=31$) there is an increase in the number of interaction residues in the expanded form. When similar analyses were performed using the subset of proteins that show binding preference for the wt form, the N-terminal polyQ region (non-parametric Sign test; $P=0,523$; $N=27$) shows also no binding differences between wt and expanded forms but the C-terminal polyQ region (non-parametric Sign test; $P=0,007$; $N=27$) shows a decrease in the number of interaction residues in the expanded form. Therefore, loss and gain of interaction in the expanded form is mainly due to differences in the C-terminal region, in contrast with ATXN1 findings

When I divide the results into polyQ proteins related and non-related with neurodegenerative diseases, there is a higher binding preference for the expanded form in the latter group for the C-terminal region (Mann-Whitney test; $Z=-2,181$; $P=0,026$; $N=59$), but not for the N-terminal region (Mann-Whitney test; $Z=-1,358$; $P=0,174$; $N=59$).

Discussion

The nine polyQ proteins involved in neurodegenerative diseases are a group of unrelated proteins with expanded polyQ repeats, with similar mechanisms of neuronal toxicity, involving protein conformational changes that result in the formation and deposition of aggregates (Fan *et al.*, 2014). Therefore, the polyQ tract plays a key role in PPI. In the previous chapter, to understand how the expanded polyQ affects PPI, the predicted protein structures for wt and expanded ATXN1, and for 71 ATXN1 interactors, were used to infer the binding surfaces of wt and expanded ATXN1. The expanded ATXN1 form presented more residues (the sum of interface residues for the 71 proteins is 12.63) at the binding interface than the wt form. Here, I show that this pattern is not observed for all polyQ proteins, neither is associated with proteins being involved in neurodegenerative diseases, or the position of the polyQ in the protein. This result is independent of the *in silico* method used since there is no bias regarding the quality of the structures, or the docking predictions of the different polyQ forms. The protein size and the interactor size do not affect the results gathered. Therefore, it can be hypothesized that each protein may have a set of unique features that determines the binding preference for wt or expanded form, explaining the results gathered.

Among the seven polyQ proteins associated with neurodegenerative diseases, AR, ATN1 and ATXN1 are proteins where the expanded form show more binding interface residues than the wt (previous Table 8). This observation is in agreement with the findings in chapter 1 for the ATXN1, using a large number of interactors. Therefore, it is conceivable that the effect of the polyQ expansion depends of the specific biological function associated to the PPI. According to the literature, in neurodegenerative diseases, polyQ expansion could lead to a gain or loss of function (van Roon-Mom *et al.*, 2005, La Spada and Taylor, 2010), but the relative importance of each is still debatable (Housman, 1995, Yang *et al.*, 2002, Schaffar *et al.*, 2004, van Roon-Mom *et al.*, 2005). For ATXN1, the increase in the interfacing residues in the expanded form seems to be associated with a gain-of-function. For instance, Lim *et al.* (2006) showed that the pathogenesis of the expanded form of ATXN1 with RBM17 is caused by a gain-of-function. Also the interaction between HDAC4 and the expanded ATXN1, lead to aberrant functions or inappropriate manifestations of normal functions (Bolger *et al.*, 2007). Nevertheless, loss-of function for expanded ATXN1 is also reported for interactions with the CIC (Chen *et al.*, 2003, Steinacker *et al.*, 2011). For expanded ATN1, at least one report shows that interaction with Caspase-3 leads to abnormal gain-

of-function in the cleavage of ATN1, that lately leads to an accumulation of ATN1 protein fragments in the cell, which can contribute for the neuronal dysfunction (Wellington *et al.*, 1998, Wood *et al.*, 2000, Suzuki and Yazawa, 2011). For AR extended polyQ tract is also reported as causing a pathologic gain-of-function (Piccioni *et al.*, 2001), although the basis and mechanism linked to the anomalous gain of function impaired by the polyQ expansion remains unknown. Genetic screens identified AF-2-based coregulators as responsible for the toxic interaction with expanded AR, which leads to a gain-of-function with dysregulated transcription overexpression, that lately causes neuronal degeneration (Nedelsky *et al.*, 2010). The abnormal gain-of-function function has been also reported for histone deacetylase (Kang *et al.*, 2004, Katsuno *et al.*, 2006, Xiang *et al.*, 2018). Nevertheless, how expansion of the polyQ affects protein function is largely unknown. THAP domaincontaining protein 11 has been described as a potential candidate for causing a polyQ disorder, mainly due to the similar characteristics to the other polyQ proteins (Pandey *et al.*, 2004). In conclusion, it is conceivable that increase binding preference in the expanded polyQ form could be associated with gain-of-function.

ATXN2, and TBP showed a binding preference for the wt form, and it is possible that expanded form of these proteins could be associated with loss of function. The role of the expanded ATXN2 in the formation of aggregates and cell degeneration, is described as either gain or loss of function (Ostrowski *et al.*, 2017). Expansion of the ATXN2 polyQ tract induces cell death through a loss of some wild-type function, especially in the Golgi complex (Huynh *et al.*, 2003, Ostrowski *et al.*, 2017). Indeed, the expression of the expanded ATXN2 disrupted the interaction with Golgi markers, disrupting the stability and morphology of the Golgi complex (Huynh *et al.*, 2003). TBP expansion of the polyQ tract causes a conformational change that leads to an enhancement in the interaction with the general transcription factor IIB (TFIIB), and in loss-of-function, since there is a reduction in the expression transcription activity (Friedman *et al.*, 2007, Friedman *et al.*, 2008, Mohan *et al.*, 2014a). Nevertheless, the interaction of the expanded form of TBP with the nuclear factor κ B (NF- κ B) signaling protein showed a gain-of-function with an increased hyperactivated signaling pathway (Yang *et al.*, 2017). RUNX2 polyQ proteins not related with neurodegenerative diseases showed also a higher binding affinity for the wt form. The runt-related transcription factor 2 protein interacts with histone acetyltransferase KAT2B, which could cause a loss of function in the acetyltransferase activity (Mohan *et al.*, 2014b). Therefore, it is

conceivable that expansion of the polyQ in proteins that show higher number of binding residues with the wt form, could cause a loss of function of the protein complex.

In contrast with ATXN1, where I observed a large number of interface residues at the N-terminal region in the expanded form, here I observed that it is the C-terminal region (the region after the polyQ) that presents binding affinity differences in wt and expanded forms, and determines binding strength as a whole. Detail analyses concerning motifs associated with processes such as ubiquitination, sumoylation and transglutamination in this region are needed, in order to understand the importance of this finding.

Conclusions

The pattern observed for ATXN1 regarding binding preference for the expanded form, is not common to all the polyQ proteins. No association was found with the localization of the polyQ tract, as well as for polyQ proteins being or not involved in neurodegenerative diseases. Here I rise the hypothesis that gain and loss of function, can be related to the binding preference for either the wt or the expanded form, since polyQ proteins that showed a binding preference for the expanded form are associated with gain-of-function and polyQ proteins that showed a binding preference for the wt form, seem more associated with the potential loss-of-function.

Chapter 3 – Identification of the interactors for human ATXN1 using *D. melanogaster*

Introduction

Ataxin-1 is one of the nine polyglutamine proteins which expansion of the coding glutamine repeat tract causes neurodegenerative disease (Zoghbi and Orr, 2009). *D. melanogaster* expressing full-length human ATXN1, is a useful model to elucidate and dissect the genetic pathways of SCA1 pathogenesis (Fernandez-Funez *et al.*, 2000, Nolo *et al.*, 2000, Chen *et al.*, 2003, Lam *et al.*, 2006, Ju *et al.*, 2013, Park *et al.*, 2013, Bondar *et al.*, 2018). The identification of 18 genes involved in protein folding/heat-shock response, ubiquitin-proteolytic pathway, RNA metabolism and transcription regulation, have been identified in mutant flies with 30Q (a wild-type human isoform) or 82Q (an expanded isoform) human ATXN1, by performing genetic screens (Fernandez-Funez *et al.*, 2000).

Flies with wt and expanded human ATXN1 genes with a *gmr*-GAL4 driver (Fernandez-Funez *et al.*, 2000) are commercially available. The existence of a GAL4-UAS system for conditional gene expression makes possible to target expression of the desired transgene in a particular tissue, such as the eye (*gmr*-GAL4 driver (Glass-Multiple-Repeats promoter) (Duffy, 2002), although it is also expressed in additional tissues, including specific neuronal cells in larval and pupal ventral and cerebral ganglia. This system allows to obtain a phenotype that can be easily examined under a dissecting microscope (both the eye color and external morphology). This is relevant since the eye is not essential for survival of the fly (does not affects viability or fertility), and its structure has been well characterized (Ray and Lakhotia, 2015). Large-scale genetic screens of these flies allowed the identification of the genes (and pathways) involved in pathogenesis (Table 9).

Table 9. Identified genes and pathways by large-scale genetic screens of *D. melanogaster* expressing full-length human ATXN1 gene.

Human interactor	Pathway	Methodology used	References
PAK1 (Serine/Threonine-protein kinase;	Ubiquitination	<i>In-vivo</i> Knock-down by RNAi (showed decreased levels of ATXN1)	(Bondar <i>et al.</i> , 2018)
MSK1 (ribosomal protein S6 kinase alpha-5;	Phosphorylation	<i>In-vivo</i> siRNA and Kinase assay	(Park <i>et al.</i> , 2013)
NLK (nemo serine-kinase;	Phosphorylation	Co-immunoprecipitation Luciferase assay	(Ju <i>et al.</i> , 2013)
AKT1 (rac-alpha serine/threonine-protein kinase;	Transcription	Y2H – Yeast two-hybrid method	(Chen <i>et al.</i> , 2003)
CIC (CAPICUA;	Transcription	Co-immunoprecipitation Luciferase assay	(Lam <i>et al.</i> , 2006)
GFI1 (growth factor independent 1;	Transcription	Co-immunoprecipitation	(Nolo <i>et al.</i> , 2000)
RpA1 and its essential partners BRCA1/2	DNA damage repair systems	<i>In vivo</i> screen of all available <i>D. melanogaster</i> homolog DNA repair genes	(Barclay <i>et al.</i> , 2014)
TERA/VCP/P97	DNA double strand break repair	Immunohistochemistry	(Fujita <i>et al.</i> , 2013)

Generally, this genetic screen studies support a toxic gain-of-function on the cell, frequently affecting transcription and gene regulation pathways, mostly caused by abnormal overexpression with transcription factor partners (Matilla-Duenas *et al.*, 2008). The involvement of these interactions and their contribution to the disease pathogenesis is not yet fully understood.

Overexpression of the expanded form of ATXN1 as showed to cause neurodegeneration in flies (Fernandez-Funez *et al.*, 2000). It should be noted that, in *D. melanogaster* the expression of the human wild-type ATXN1 shows a phenotype similar but less pronounced than the expanded human ATXN1 (Fernandez-Funez *et al.*, 2000, Chen *et al.*, 2003). Mutation at the serine (S776) was showed to be critical for neuronal degeneration, as seen in mice expressing a ATXN1 expanded form in which S776 is replaced by alanine (Emamian *et al.*, 2003). Interestingly, overexpression of the wt form of human ATXN1 was associated to cause neurodegeneration when expressed at

considerable high levels in *D. melanogaster*, suggesting that protein domains besides the expanded polyQ tract might contribute to pathogenesis of Ataxin-1 (Fernandez-Funez *et al.*, 2000).

The *D. melanogaster* orthologous ATXN1 (Atx1; GeneID31624) is a small protein, without the N-terminal region, that shares homology with the human protein at the AXH domain only. Since this domain is one of the regions of interaction of ATXN1 with most of the interactors (Chapter 1), a large fraction of common interactors in the two species is expected. For *Drosophila*, only three interactions have been reported for Atx1 (Nmnat (AAL90149, GeneID42987); MTA1-like (CG2244, 40693); senseless (CG10714, 45328). The latter interaction (*D. melanogaster* zinc-finger transcription factor senseless) is also described in EvoPPI for the human ATXN1 (the human homolog growth factor independence-1, Gfi-1; (Nolo *et al.*, 2000)). Nevertheless, interactions with *Drosophila* orthologous binding partners of the human ATXN1 have been reported (Chen *et al.*, 2003, Tsuda *et al.*, 2005, Lam *et al.*, 2006, Ju *et al.*, 2013, Park *et al.*, 2013, Bondar *et al.*, 2018). Here, using co-immunoprecipitation and mass spectrometry of the head of the mutant flies, I identify seven such proteins that have been reported to interact with human ATXN1. FAR1, YWHAЕ, and IMMT, analyzed in Chapter 1, show interaction with ATXN1 both at the N-terminal and the C-terminal region. For the GAPDH (Koshy *et al.*, 1996) PSMC3 (Lim *et al.*, 2006), ATP1A3 (Lim *et al.*, 2006), and HSPA8 (Choi *et al.*, 2007), the N-terminal region is critical for ATXN1 interaction. I also found indirect evidence for 143 human interactors of ATXN1, supporting the evidence for a large conservation of the human ATXN1 network in *Drosophila*, despite conservation in the C-terminal region only. Moreover, there are 229 proteins that are not reported as ATXN1 interactors. Supporting such hypotheses is the fact that 67 % of these proteins are ubiquitin interactors.

Materials and Methods

1. Flies expressing human ATXN1 wild-type and expanded forms

Mutant flies with wt (two (39738) and 30 (30Q; 39739) consecutive glutamines) and expanded (with 80 (33818) and 82Q (39740)) human ATXN1 (Chen *et al.*, 2003) were obtained from the Bloomington Center (Table 10).

Table 10. *D. melanogaster* strains used in this study. The Bloomington *D. melanogaster* Stock Center number, the transgene and their chromosome localization as well as the gender used in this study are reported.

<i>D. melanogaster</i> strains	
Expanded ATXN1 (82Q)	33818 (w[1118]; P{w[+mC]=UAS-Hsap\ATX1. 82Q } selected M6 (Chromosome 3))
	39740 (P{w[+mC]=UAS-Hsap\ATX1. 82Q }F7, y[1] w[*] (Chromosome X))
wt ATXN1 (30Q)	39738 (P{w[+mC]=UAS-Hsap\ATX1. 2Q }F5, y[1] w[*] (Chromosome X))
	39739 (y[1] w[*]; P{w[+mC]=GAL4-ninaE.gmr}12/CyO; P{w[+mC]=UAS-Hsap\ATX1. 30Q }F6 (Chromosome 2; Chromosome 3))
Drive	GMR-GAL4 (Chromosome X)
Negative control	5905 – white

These flies also hold the *Cyo* gene, expressing the curly wings mutant, lethal when homozygote, being widely used for selection of heterozygotes. Flies with *gmr-GAL4*, known to target expression exclusively in the developing eye cells by expression of the GAL4 protein (Ray and Lakhota, 2015) were used as females. The males used in the crosses expressed the human transgene ATXN1 with UAS site, a GAL4 protein-binding site, which linked with the respective protein, activates the expression of the transgene. As negative control, I used the flies line white, displaying the characteristic phenotype white eyes, possessing the transposable element white (*w(*)*) used to insert the *ATXN1* of the other lines.

2. Protein extraction

I used flies heads for protein extraction using the protocol of Emery (2007). For each sample, 400 flies were frozen in liquid nitrogen and then were decapitated by

vigorously vortexing twice for 15s. The heads of the flies were grounded after adding of 1,25 µl of extraction buffer per fly (500 µl) (extraction buffer: 50 mM TRIS; 150 mM NaCl, pH 7.5 (TBS); 5% glycerol; 10 mM EDTA; 0.1% Triton; 1 mM dithiothreitol (DTT); 0.5 mM phenylmethylsulfonyl fluoride; 20 mg/mL aprotinin, 5 mg/mL leupeptin, 5 mg/mL pepstatin A (Protease inhibitor cocktail; Complete™, Mini, EDTA-free Protease Inhibitor Cocktail, Roche). After a centrifugation step for 10 minutes, at 12 000g at 4°C, I collected the supernatant and disposed the pellet. The concentration of the proteins was measured by a NanoDrop Microvolume Spectrophotometer and the proteins were used to perform co-immunoprecipitation followed by mass spectrometry to analyze the interactors of wild-type and expanded form of ATXN1.

3. Co-Immunoprecipitation

For the co-immunoprecipitation assay, I used the classic protocol of the Protein G Mag Sepharose (GE HealthCare) and the antibody for the N-terminal region of the polyQ tract, characteristic of the human protein (11NQ anti-ATXN1; NeuroMab (Park *et al.*, 2013). The co-immunoprecipitation experiment involves three essential steps: the protein extraction (made previously), coupling of the antibody to the beads and isolation of the protein complexes. To perform the classical protocol I first prepared the magnetic beads, then wash the beads (calibration) with the binding buffer (TBS (50 mM TRIS, 150 mM NaCl pH 7,5)) and then bind the antibody to the beads. The antibody was prepared with the binding buffer (1:500) and was added to the beads, incubated and resuspended in a benchtop shaker at 4°C for two hours. After a wash of the beads, the protein extracted from the flies (made previously) was added to the beads and a new cycle of incubation and resuspension on a benchtop shaker at 4°C for two hours was performed. After, the unbound fraction liquid was removed. Three washing steps were then made with washing buffer. Before the last wash process, the solution with the beads was transferred to a new microtube, since some unbound fraction could be restrained. Subsequently, the elution process, or isolation of the proteins complex was made with two different solutions. First I used, 50 µL of the elution buffer (5% acetic acid) for two minutes and collect the solution with the eluted proteins (E1). Then, I made a second step of elution to ensure the complete elution of the proteins binding to the beads. In the second step, I used 50 µL of the elution buffer (2% SDS) and the samples are heated at 95°C for 5 minutes (E2).

4. Mass Spectrometry screen and data analyzes

The method used to identify the isolated proteins present in the macromolecular complexes with ATXN1 was mass spectrometry. This method consists on the ionization of the solutions that are further transported by electric fields to the mass analyzer(that is based on their mass-to-charge ratio), retrieving the isotopic signature of the samples (spectrums). These spectrums are analyzed using Proteome discovery software and BLAST (Boratyn *et al.*, 2013). The elution samples prepared in the co-immunoprecipitation step were analyzed at I3S. Since, two elution steps were made to ensure the desirable complete elution (E1 with acetic acid and E2 with SDS), I considered the results from E2 samples, as these had a higher elution and consequently, a higher degree of confidence to be interactors of ATXN1. For each protein identified, I retrieve the GeneID using Uniprot Retrieve ID tool (<https://www.uniprot.org/uploadlists/>) (UniProt, 2019) and the Flybase ID converter tool (<https://flybase.org/convert/id>) (Thurmond *et al.*, 2019). I also identified the human orthologues using DIOPT - DRSC Integrative Ortholog Prediction Tool (<https://fgr.hms.harvard.edu/diopt>) (Hu *et al.*, 2011). To address expression in brain I used DAVID (<https://david.ncifcrf.gov/>) (Huang *et al.*, 2007) that integrates data from Uniprot expression database (UniProt, 2019) and the Human Protein Atlas Database (<https://www.proteinatlas.org/>) (Uhlen *et al.*, 2015). Only the human orthologues of interactors founded in *D. melanogaster* with expression in the brain were selected for further analyzes. The molecular function, the biological process and pathways of the interactors were gathered using PANTHER (<http://www.pantherdb.org/>) (Mi *et al.*, 2013). The HGNC:ID of each protein needed to use this tool was retrieved using Uniprot Retrieve ID tool with a batch download (UniProt, 2019). To validate the ATXN1 reported interactors we compared our results with those obtained using EvoPPI (Vázquez *et al.*, 2019), using Venny 2.1.0 (<http://bioinfogp.cnb.csic.es/tools/Venny/>).

5. *In silico* prediction of the interacting residues

In this case, the human polyQ protein was the ATXN1 protein with the common found interactors for the wt and expanded ATXN1 forms in both *D. melanogaster* and *H. sapiens*. Firstly, the I-TASSER tool (Zhang, 2008) was used to predict the structure of the entire protein. Moreover, the prediction of the protein-protein interface residues of each protein was made using CPORT (de Vries and Bonvin, 2011). The docking of the ATXN1 protein with the resulting interactors was made using HADDOCK2.2 (High Ambiguity Driven protein-protein DOCKing (van Zundert *et al.*, 2016), that predicts the

structure PPI of biomolecular complexes. The analyze of the docking results was made using PISAePDB (Proteins, Interfaces, Structures and Assemblies - Protein Data Bank) web tool (Krissinel and Henrick, 2007) to identify the properties of molecular interfaces at a residue level.

Results

No interactions were obtained for the white flies, here used as negative control. This supports the specificity of the antibody used for the human N-terminal ATXN1 region. In order to obtain a larger network of ATXN1 PPI, the proteins identified from the different experiments were grouped. After removing duplicates, I identify 408 *D. melanogaster* proteins (274 unique using mutant flies with wt ATXN1, 17 unique using mutant flies with expanded ATXN1, and 117 common to both experiments) as interacting with the human ATXN1. Only 391 proteins present human orthologous. These proteins are assigned to diverse functions, with emphasis on catalytic activity, binding, cytoskeletal proteins, transport activity and structural molecular activity (Figure 10).

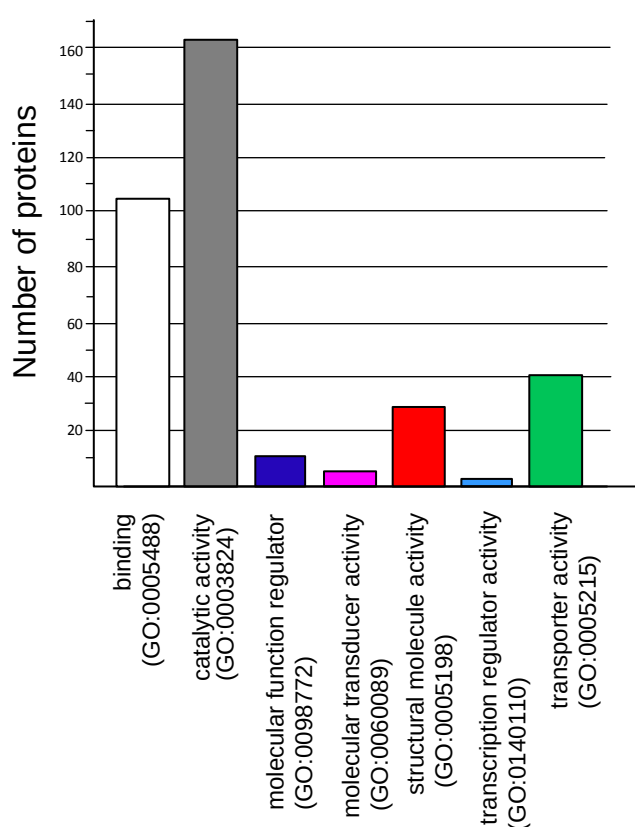


Figure 10. Molecular function of the interactors (PantherDB) of the 391 possible interactors of the human Ataxin-1 protein obtained by co-immunoprecipitation and mass spectrometry assays (only orthologues examination with best score were used).

According to EvoPPI, for the human ATXN1 there are 311 interactors reported, but only for 253 I found a *Drosophila* orthologous. By comparing the two data sets seven proteins are common (Figure 11; Table 11).

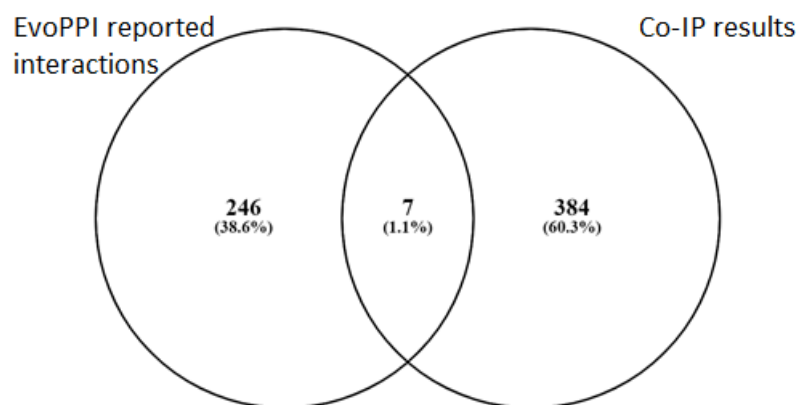


Figure 11. Comparison of the interactors detected from the mutant the *Drosophila* co-immunoprecipitation and mass spectrometry assays and the interactions reported in EvoPPI for ATXN1 that present a *Drosophila* orthologous

Table 11. ATXN1 interactors found in the co-immunoprecipitation and mass spectrometry assays common in *Drosophila* and humans, reported in EvoPPI for the human ATXN1 protein.

Human	Size (aa)	<i>D. melanogaster</i>	Size (aa)
ATP1A3 (Sodium/potassium-transporting ATPase subunit alpha-3; 478; P13637)	1041	Atpalpha (48971; E1JIR4)	1013
GAPDH (Glyceraldehyde-3-phosphate dehydrogenase II; 2597; P04406)	332	GAPDH2 (32545; P07487)	335
PSMC3 (26S proteasome regulatory subunit 6A; 5702; P17980)	439	Rpt5 (26S proteasome regulatory subunit p50; 42805; Q9V3V6)	428
YWHAE (14-3-3 protein epsilon; 7531; P62258)	255	14-3-3epsilon (42186; P92177)	262
IMMT (MICOS complex subunit MIC60; 10989; Q16891)	758	Mitofilin (42587; P91928)	735

FAR1 (Fatty acyl-CoA reductase 1; 84188, Q8WVX9)	515	CG8306 (36857; Q960W6)	516
HSPA8 (Heat Shock Protein Family; 3312; P11142)	646	Hsc70-4 (41840; P11147)	651

Since in *Drosophila* the orthologous of the human ATXN1 (Atx1) does not have the N-terminal region, I expect that in the binding interface would be with the AXH domain. Indeed, *in silico* analyzes using the methodology in chapter 1 revealed a higher number of interaction residues in the C-terminal region, especially in the AXH domain. Interaction in the ATXN1 N-terminal region are, however, inferred for six human proteins (Table 12).

Table 12. The location of the binding residues (N-terminal, polyQ, and C-terminal) between the human ATXN1 and the seven *Drosophila* proteins and there human orthologous, inferred by the *in silico* approach.

ATXN1 interactions	Total	N-terminal	polyQ	C-terminal
Gapdh2 (<i>Drosophila</i>)	65	18	1	46
GAPDH (human)	59	15	0	44
Rpt5 (<i>Drosophila</i>)	56	0	0	56
PSMC3 (human)	68	12	0	56
CG8306 (<i>Drosophila</i>)	66	13	0	53
FAR1 (human)	67	11	0	56
14-3-3epsilon (<i>Drosophila</i>)	68	24	0	44
YWHAE (human)	62	2	0	60
Atpalpha (<i>Drosophila</i>)	56	18	0	38
ATP1A3 (human)	61	9	0	52
Hsc70-4 (<i>Drosophila</i>)	57	0	0	57
HSPA8 (human)	51	0	0	51
Mitofilin (<i>Drosophila</i>)	52	0	0	52
IMMT (human)	77	28	1	48

For the remaining interactors (384) we address if these proteins have been described as proteins that interact with those proteins that interact with ATXN1 (second level of interaction at EvoPPI). As ATXN1 is degraded by the ubiquitin –proteasome pathway, I have excluded UBC (7316) interactors from the following analyses. By comparing the reported interactions in EvoPPI of the 309 (UBC and ATXN1 have been

removed) ATXN1 interactors (the second level of the interaction), I found 143 proteins common to the two data sets, presented in the following Table 13.

Table 13. ATXN1 second level interactors found in the co-immunoprecipitation and mass spectrometry assays, common to the ATXN1 reported second level interaction (except ATXN1 and UBC interactors).

Protein Symbol	Gene ID	Gene Name
ACACA	31	Acetyl-CoA carboxylase 1
AKT1	207	RAC-alpha serine/threonine-protein kinase
ASNS	440	Asparagine synthetase
PRDM1	639	PR domain zinc finger protein 1
CAMK2B	816	Calcium/calmodulin-dependent protein kinase type II subunit beta
CFL1	1072	Cofilin-1
CRIP2	1397	Cysteine-rich protein 2
CRK	1398	Proto-oncogene c-Crk
CRY2	1408	Cryptochrome-2
CST3	1471	Cystatin-3
ESRRA	2101	Steroid hormone receptor ERR1
FDPS	2224	Farnesyl pyrophosphate synthase
FHL2	2274	Four and a half LIM domains protein 2
FYN	2534	Tyrosine-protein kinase Fyn
GSPT1	2935	Eukaryotic peptide chain release factor GTP-binding subunit ERF3A
HIVEP1	3096	Zinc finger protein 40
NR4A1	3164	Nuclear receptor subfamily 4 group A member 1
HSPA1A	3303	Heat shock 70 kDa protein 1A
HSPA1B	3304	Heat shock 70 kDa protein 1B
HSPA4	3308	Heat shock 70 kDa protein 4
ITGB4	3691	Integrin beta-4
KIF22	3835	Kinesin-like protein KIF22
LASP1	3927	LIM and SH3 domain protein 1
LOXL1	4016	Lysyl oxidase homolog 1
MAGEB2	4113	Melanoma-associated antigen B2
MATN2	4147	Matrilin-2
MBP	4155	Myelin basic protein
NARS	4677	Asparagine--tRNA ligase
NFYA	4800	Nuclear transcription factor Y subunit alpha
OPCML	4978	Opioid-binding protein/cell adhesion molecule
PEPD	5184	Xaa-Pro dipeptidase
PML	5371	Protein PML
PROP1	5626	Homeobox protein prophet of Pit-1
PSPH	5723	Phosphoserine phosphatase
RCN1	5954	Reticulocalbin-1
REL	5966	Proto-oncogene c-Rel
ATXN2	6311	Ataxin-2

SMARCC2	6601	SWI/SNF complex subunit SMARCC2
SUMO3	6612	Small ubiquitin-related modifier 3
SUMO2	6613	Small ubiquitin-related modifier 2
SREBF2	6721	Sterol regulatory element-binding protein 2
TRAF2	7186	TNF receptor-associated factor 2
TRIP6	7205	Thyroid receptor-interacting protein 6
TSC1	7248	Tuberous sclerosis 1 protein
UBE2I	7329	SUMO-conjugating enzyme UBC9
SUMO1	7341	Small ubiquitin-related modifier 1
VCP	7415	Transitional endoplasmic reticulum ATPase
VSINL1	7447	Visinin-like protein 1
YWHAZ	7534	14-3-3 protein zeta/delta
SF1	7536	Splicing factor 1
ZYX	7791	Zyxin
USP7	7874	Ubiquitin carboxyl-terminal hydrolase 7
PRRC2A	7916	Proline-rich and coiled-coil-containing protein 2A
KAT6A	7994	Histone acetyltransferase KAT6A
ANP32A	8125	Acidic leucine-rich nuclear phosphoprotein 32 family member A
COIL	8161	Coilin
PICALM	8301	Phosphatidylinositol-binding clathrin assembly protein
HIST1H3A	8350	Histone H3.1 subunit A
HIST1H3D	8351	Histone H3.1 subunit D
HIST1H3C	8352	Histone H3.1 subunit C
HIST1H3E	8353	Histone H3.1 subunit E
HIST1H3I	8354	Histone H3.1 subunit I
HIST1H3G	8355	Histone H3.1 subunit G
HIST1H3J	8356	Histone H3.1 subunit J
HIST1H3H	8357	Histone H3.1 subunit H
HIST1H3B	8358	Histone H3.1 subunit B
PLA2G10	8399	Group 10 secretory phospholipase A2
CAMK1	8536	Calcium/calmodulin-dependent protein kinase type 1
BHLHE40	8553	Class E basic helix-loop-helix protein 40
PIAS1	8554	E3 SUMO-protein ligase PIAS1
TCAP	8557	Telethonin
EIF3F	8665	Eukaryotic translation initiation factor 3 subunit F
CCNK	8812	Cyclin-K
FUBP3	8939	Far upstream element-binding protein 3
HIST1H3F	8968	Histone H3.1
EIF1AY	9086	Eukaryotic translation initiation factor 1A, Y-chromosomal
DNAJA3	9093	DnaJ homolog subfamily A member 3, mitochondrial
ATP6V0D1	9114	V-type proton ATPase subunit d 1
LITAF	9516	Lipopolysaccharide-induced tumor necrosis factor-alpha factor
NCOR1	9611	Nuclear receptor corepressor 1
NCOR2	9612	Nuclear receptor corepressor 2

DZIP3	9666	E3 ubiquitin-protein ligase DZIP3
PPP6R2	9701	Serine/threonine-protein phosphatase 6 regulatory subunit 2
ARHGAP32	9743	Rho GTPase-activating protein 32
KMT2B	9757	Histone-lysine N-methyltransferase 2B
HDAC4	9759	Histone deacetylase 4
IST1	9798	IST1 homolog
TOMM20	9804	Mitochondrial import receptor subunit TOM20 homolog
TOX4	9878	TOX high mobility group box family member 4
POM121	9883	Nuclear envelope pore membrane protein POM 121
UBAP2L	9898	Ubiquitin-associated protein 2-like
PQBP1	10084	Polyglutamine-binding protein 1
SUGP2	10147	SURP and G-patch domain-containing protein 2
STUB1	10273	E3 ubiquitin-protein ligase CHIP
EIF1B	10289	Eukaryotic translation initiation factor 1b
BASP1	10409	Brain acid soluble protein 1
TRIM38	10475	E3 ubiquitin-protein ligase TRIM38
KAT5	10524	Histone acetyltransferase KAT5
IPO8	10526	Importin-8
ILVBL	10994	Acetolactate synthase-like protein
CLASRP	11129	CLK4-associating serine/arginine rich protein
ZHX1	11244	Zinc fingers and homeoboxes protein 1
ATXN2L	11273	Ataxin-2-like protein
ACOT7	11332	Cytosolic acyl coenzyme A thioester hydrolase
U2AF2	11338	Splicing factor U2AF 65 kDa subunit
ATF5	22809	Cyclic AMP-dependent transcription factor ATF-5
EPN2	22905	Epsin-2
TRIM32	22954	E3 ubiquitin-protein ligase TRIM32
GGA2	23062	ADP-ribosylation factor-binding protein GGA2
POGZ	23126	Pogo transposable element with ZNF domain
GPATCH8	23131	G patch domain-containing protein 8
RBFOX2	23543	RNA binding protein fox-1 homolog 2
PHPT1	29085	14 kDa phosphohistidine phosphatase
EFEMP2	30008	EGF-containing fibulin-like extracellular matrix protein 2
TMX2	51075	Thioredoxin-related transmembrane protein 2
POMP	51371	Proteasome maturation protein
TDP2	51567	Tyrosyl-DNA phosphodiesterase 2
MED15	51586	Mediator of RNA polymerase II transcription subunit 15
NLK	51701	Serine/threonine-protein kinase NLK
TOLLIP	54472	Toll-interacting protein
PLEKHA5	54477	Pleckstrin homology domain-containing family A member 5
RBFOX1	54715	RNA binding protein fox-1 homolog 1
TTC19	54902	Tetratricopeptide repeat protein 19, mitochondrial
RNF31	55072	E3 ubiquitin-protein ligase RNF31
FAM46A	55603	Terminal nucleotidyltransferase 5A

SYBU	55638	Syntabulin
EIF4ENIF1	56478	Eukaryotic translation initiation factor 4E transporter
SAR1A	56681	GTP-binding protein SAR1a
UBQLN4	56893	Ubiquilin-4
DHX37	57647	Probable ATP-dependent RNA helicase DHX37
METTL17	64745	Methyltransferase-like protein 17, mitochondrial
CCDC136	64753	Coiled-coil domain-containing protein 136
WNK1	65125	Serine/threonine-protein kinase WNK1
EHMT1	79813	Histone-lysine N-methyltransferase EHMT1
CPSF7	79869	Cleavage and polyadenylation specificity factor subunit 7
DOCK5	80005	Dedicator of cytokinesis protein 5
C1orf94	84970	Uncharacterized protein C1orf94
UNK	85451	RING finger protein unkempt homolog
HNRNPLL	92906	Heterogeneous nuclear ribonucleoprotein L-like
FAM46B	115572	Terminal nucleotidyltransferase 5B
USP54	159195	Inactive ubiquitin carboxyl-terminal hydrolase 54
ATXN1L	342371	Ataxin-1-like
MYO18A	399687	Molecule associated with JAK3 N-terminus

Therefore, at maximum I found evidence for 46.3 % of the human ATXN1 interactors. Nevertheless, the seven protein that interact with ATXN1, presented in Table 11, that I found in the experiments can already account for 56 out of the 143 proteins, and thus 19.7 % of the human ATXN1 interactors. These proteins are not bias according to molecular function (Figure 12), and they are assigned in the same seven molecular functions groups as described before, plus the translation regulator activity (GO:0045182) group.

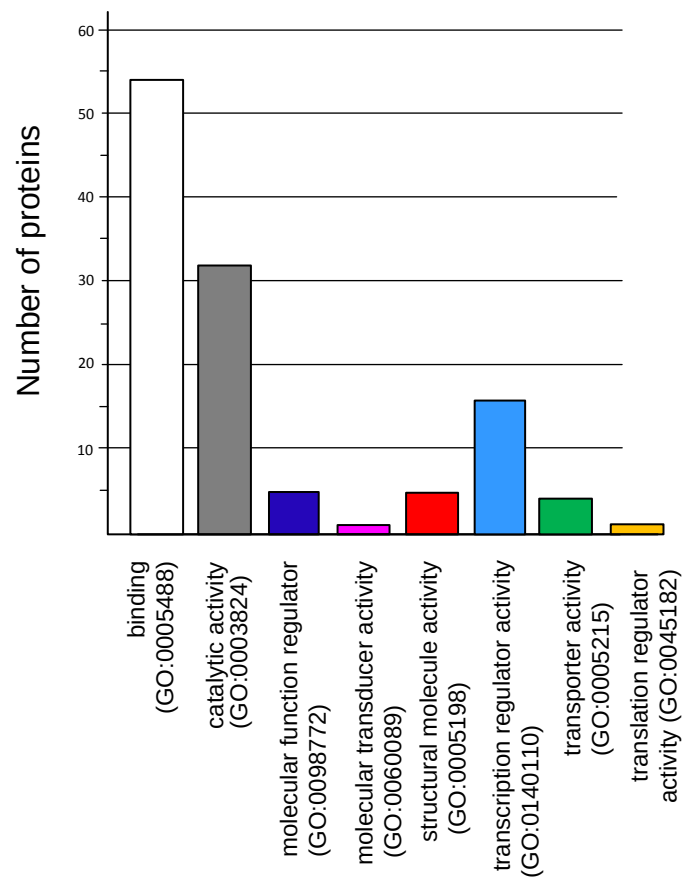


Figure 12. Molecular function (according to PANTHERdb) for the 143 ATXN1 interactions inferred by analyses of the proteins of the second level of interaction.

The remaining 229 proteins obtained could represent new interactions for ATXN1. Accordingly, the molecular function of these proteins are similar to the 143 ATXN1 proteins, supporting there putative involvement in ATXN1 network (Figure 13).

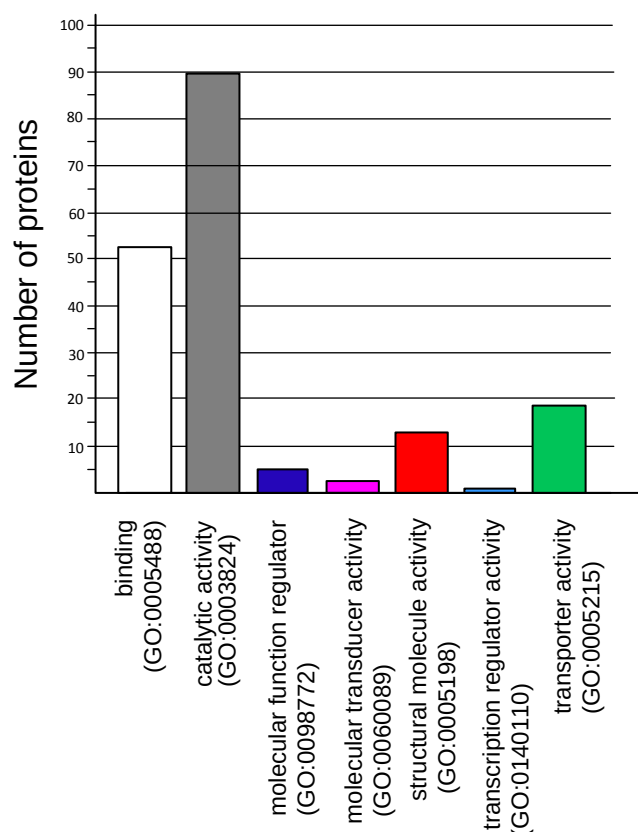


Figure 13. Molecular function (according to PantherDB) of the 229 potential new interactions inferred by analyses of the proteins of the second level of interaction.

To get further support for the involvement of these proteins as ATXN1 interactors, I compared these proteins with interactors of the two proteins, ATXN1-like, and HBP1 (HMG-box transcription factor 1; 26959), that also present an AXH domain. None of these proteins is reported as level one interactors of these proteins.

Discussion

The well-known spinocerebellar ataxia type 1 (SCA1) is autosomal progressive disease is caused by an expansion in the polyQ tract of the ATXN-1 protein, which affects the PPI, with an abnormal higher affinity, that in lately cases, can lead to the formation of aggregates that could cause degeneration of the cell. The *D. melanogaster* ATXN1 orthologous, the Atx1, does not have a N-terminal region, but the C-terminal region is conserved and in both proteins have an AXH domain (Tsuda *et al.*, 2005). In *Drosophila* only three proteins are known to interact with Atx1, and one (Senseless (Sens)) is also described in EvoPPI for the human ATXN1 (Gfi-1; (Nolo *et al.*, 2000)). The humans ATXN1 protein contains a unique extensive N-terminal region, which contains the polyglutamine tract. Co-immunoprecipitation and mass spectrometry assays of mutant flies with human ATXN1, revealed seven *Drosophila* interactors for the human ATXN1. Comparison of the docking results, using an *in silico* approach, of the human ATXN1 with the *Drosophila* and human orthologous proteins revealed that the AXH domain is essential for binding with ATXN1. In human these proteins can also interact in the N-terminal ATXN1 region. These proteins appear in both wt and expanded ATXN1 fly mutants, making impossible to address differences in binding preferences. Considering the 143 proteins (presented in the Table 13) that represent interactors level 2 of the ATXN1, there is a high conservation between the genetic pathways in humans and *Drosophila*, making *Drosophila* a suitable organism for detail analyses to identify genes that overexpressed or suppressed SCA1-induced neurodegeneration. For the putative new interactors of humans Ataxin-1, further *in silico* analyses are needed, to predict interface residues between the two proteins. These possible interactors of human ATXN1 could give new insights in SCA1 pathogenesis.

Since only seven proteins were obtained as level 1 in the ATXN1 interactome, this implies that the methodology used or/and the number of runs per sample is not appropriated to address binding differences between wt and expanded ATXN1 forms. For each sample, in the near future, I will perform two more assays.

Conclusion

The *in vitro* assays done in this study with transgenic flies coding the human Ataxin-1 protein showed the importance of the AXH domain in the PPI. Seven interactors reported for ataxin-1 were confirmed with this assay, including the proteins ATP1A3 (Sodium/potassium-transporting ATPase subunit alpha-3), GAPDH (Glyceraldehyde-3-phosphate dehydrogenase), PSMC3 (26S proteasome regulatory subunit 6A), YWHAE (14-3-3 protein epsilon), IMMT (MICOS complex subunit MIC60), FAR1 (Fatty acyl-CoA reductase 1) and HSPA8 (Heat Shock Protein Family). The *in silico* analyzes of these proteins, confirmed that the influence of the polyQ tract in the PPI, especially in the binding interface of the N-terminal region of the polyQ tract. As expected the C-terminal AXH domain showed the higher number of interaction residues in all the interactions analyzed, supporting the key role of this domain in the recognition and binding involving the PPI of Ataxin-1 protein. Considering proteins of the second level of interactions of the ATXN1, a maximum of 143 proteins were also gathered, supporting the conservation of the ATXN1 network in these distantly related organisms. I also identify new putative ATXN1 interactors, but for these proteins *in silico* analyses must be performed to support the interaction with ATXN1.

Chapter 4 – Evolution of the polyglutamine tract in the Animalia Kingdom

Introduction

Homorepeats stretches are widespread in the coding sequences across the animals, with orthologues being conserved across distantly related taxa, especially those involved in development (Fondon and Garner, 2004). In the eukaryotic proteome the amino acid glutamine is the most common residue in stretches with over 16 residues (Djian, 1998). Proteins with polyQ tracts are present in lower and higher eukaryotes, and their number varies among species, especially for distant related ones (Schaefer *et al.*, 2012). This observation suggests that the number of polyQ proteins is not random and depends on variable properties among species (Schaefer *et al.*, 2012). Analyses of polyQs in 14 eukaryotes reveals that the species with higher number of polyQ proteins, as well as the longest ones, are those of *Drosophila* (Ramazzotti *et al.*, 2012, Mier and Andrade-Navarro, 2018). Generally, the percentage of the number of polyQs in the genome is similar in taxonomically related species, although there are a few exceptions. For example, when looking at the *Drosophilae* species, the percentage of number of polyQs in the genome varies widely from a range of 2,7% in *D. simulans* to 8.9% in *D. grimshawi* (Schaefer *et al.*, 2012). In mammals, related species displays a similar number of polyQs. In primates, the length of orthologous polyQ tracts is similar, in particular for the shorter ones, in contrast with those proteins related to polyQ-associated diseases that are longer in human than in evolutionarily closer non-human primates (Mier and Andrade-Navarro, 2018). The hypothesized evolutionary model for polyQ in primates suggested that shorter polyQ tract are positively selected, since a short polyQ region is probably long enough to be functional, and thus their growth is no longer selected. The longer stretch is then actively counter selected to avoid instability, or is neutral (Mier and Andrade-Navarro, 2018). The longer polyQ regions that “escape” this selection tend to grow uncontrollably, being involved in the development of polyQ-related diseases (Mier and Andrade-Navarro, 2018). This selection is hypothesized not to be random, with trend towards the addition CAG repeats in primates, more particularly in humans (Djian, 1998).

An example that shows that the selection pressure of the polyQ tract is independent in different groups is the HTT protein family. The N-terminal polyQ tract in chordates show different length, including in humans. The *Drosophila* group shows HTT orthologue with more than one C-terminal polyQ tract, validating that there are independent events of insertion of polyQ tracts. This would suggest that the huntingtin orthologues are under selecting pressure to the inclusion of the polyQ tract, but this pressure not act on a precise position in the sequence (Schaefer *et al.*, 2012).

The observation that species with high polyQ protein content have a higher number of proteins related to phosphatidylinositol signaling and ubiquitin degradation (important mechanisms in the clearance of polyQ aggregates) suggest differential selection (Schaefer *et al.*, 2012). Organisms that can select polyQ co-evolve the appropriate machinery to clear polyQ protein aggregates (Petrakis *et al.*, 2013). In humans, there are only 65 proteins with a polyQ larger than 10 repeats. Such a small number can be explained either by a low rate of origin of polyQ tracts or strong purifying selection against polyQ tracts and/or tract size (Rocha *et al.*, 2019).

Since in humans, polyQ containing proteins tend to be transcription factors, rather than randomly found across different functional categories, they likely have an important functional role, and therefore, in those proteins, they likely confer some selective advantage when the tract is below a given size (Faux *et al.*, 2005). Indeed, this bias is also seen in Bovines species, since gene ontology analyzes revealed that the encoded proteins with polyQ tracts were enriched for functions associated with transcriptional regulation in nuclear complexes (Whan *et al.*, 2010). Such an advantage is likely related to the number of binding partners or binding strength, since in humans polyQ containing proteins have on average a higher number of interactors than proteins without polyQ tracts (Rocha *et al.*, 2019). It should be noted that, on average, transcription factors that do not have a polyQ tract do not have more interactors than non-transcription factor proteins without a polyQ tract. *Mus musculus* polyQ containing proteins also have, on average, a higher number of interactors than proteins without polyQ tracts, but transcription factors without a polyQ tract already have a larger number of interactors than proteins without a polyQ that are not transcription factors. In this species polyQ containing proteins are, as in humans, also have a tendency to be transcription factors (Rocha *et al.*, 2019). Therefore, the reason for the higher number of interactors in *M. musculus* is not so clear. In *Rattus norvegicus*, although polyQ proteins are significantly enriched for the transcription factor category, the average number of interactors is similar

in polyQ and non-polyQ proteins as well as in non-polyQ transcription factors and non-polyQ non-transcription factor, although it could be argued that there is less data for *R. norvegicus* than for *M. musculus* and *H. sapiens*. For *D. melanogaster* and *C. elegans* there is, however, a reasonably large amount of data on interactions available. While in *Drosophila* the findings are similar to those obtained in *M. musculus*, in *C. elegans* proteins with a polyQ tract do not have a tendency to be transcription factors, and polyQ containing proteins do not have on average a higher number of interactors. Moreover, the number of genes that encode proteins with a polyQ tract varies widely from under 20 in *C. elegans* to more than 240 in *D. melanogaster* (Rocha *et al.*, 2019). Therefore, during evolution the rate of origin of polyQ tracts, as well as the selective forces acting on polyQ tracts and tract length likely varies. In order to make a detailed characterization of the evolution of polyQ tracts in the Animalia kingdom, here I estimate the fraction of genes with polyQ tracts, the number of polyQ tracts per gene, as well as the minimum, average, and maximum polyQ tract size in 374 animal species.

Materials and Methods

1. Sequence data and annotation

Reference sequences from 376 species sequenced from Kingdom Animalia were retrieved from NCBI RefSeq. Because this work is centralized in proteins, and more specifically the polyQ tract, I obtained only the coding sequences (CDS) from these genomes. From all obtained sequences, I selected only sequences coding polyQ tracts, by using the pattern filter tool on SEDA software (Lopez-Fernandez *et al.*, 2018). Since most software applications do not function with degenerative bases I removed all the sequences with the known IUPAC degenerative bases (R (A or G); Y (C or T); S (G or C); W (A or T); K (G or T); M (A or C); B (C or G or T); D (A or G or T); H (A or C or T); V (A or C or G); N (all bases)) using SEDA.

As described in chapter 2, since the focus of this work is on polyQ proteins, those sequences were filtered to select the coding sequences of proteins with polyQ tracts using Pattern filter of SEDA software (Lopez-Fernandez *et al.*, 2018), filtering for protein sequences with at least 10 consecutive glutamines, which is defined as a polyQ protein and accordingly to (Schaefer *et al.*, 2012). Since the polyQ tract of Ataxin-3 is, according to Kawaguchi *et al.*, (1994), defined by consecutive glutamines with interruption of a methionine, I considered the possibility of the polyQ tract to have at least one inconsecutive interruption on the characterization of the tract. I then selected reference sequences for each protein, removing isoforms and identical sequences, using SEDA (Lopez-Fernandez *et al.*, 2018). I also identify the class and family for each sequence, by taking into account the NCBI taxonomy, to make it possible to cluster results by taxonomy groups, and remove possible contaminants like virus and bacteria that were sequenced and included in the genome data. Two contaminates genomes were removed. I also renamed the sequence files using SEDA (Lopez-Fernandez *et al.*, 2018), with taxonomy classification (the class and family).

To characterize the polyQ tract in terms of minimum, maximum and average size, the number of polyQ tracts, the number of genes showing a polyQ, and finally the ratio of polyQ tracts per gene, I used a BASH script.

2. Design of cladograms

By using the NCBI taxonomy, I could retrieve the relations among all species for the sequences previously obtained. In Microsoft Office Visio I draw a cladogram describing the relation among classes and families of species, from non bilaterian to mammalian. The classes that showed a high variety in the number of polyQ sequences were further analyzed. Since *Drosophila* family presents the higher number of these sequences, further analyzes were made specific to this family. I then refine the cladograms using the literature available about those groups and the reference database related to taxonomy, Tree-of-life web project (Maddison, 2007). In order to predict and evaluate the evolution pattern of the polyQ proteins, I represented the previous data obtained with BASH script, through bar charts embedded, all species in the cladograms with those factors.

3. Statistical analyzes

The statistical analyzes was made using SPSS (<https://www.ibm.com/analytics/spss-statistics-software>) with Box-plot analyzes and representations, in order to see distinctive groups/ species for each category (minimum, maximum and average size, the percentage of polyQ genes in the genomes, and the ratio of polyQ tracts per gene), that could explain a possible selective pressure in a certain group.

Results

As can be seen in Figure 14, the fraction of genes encoding polyQ proteins varies widely in the different animal groups.

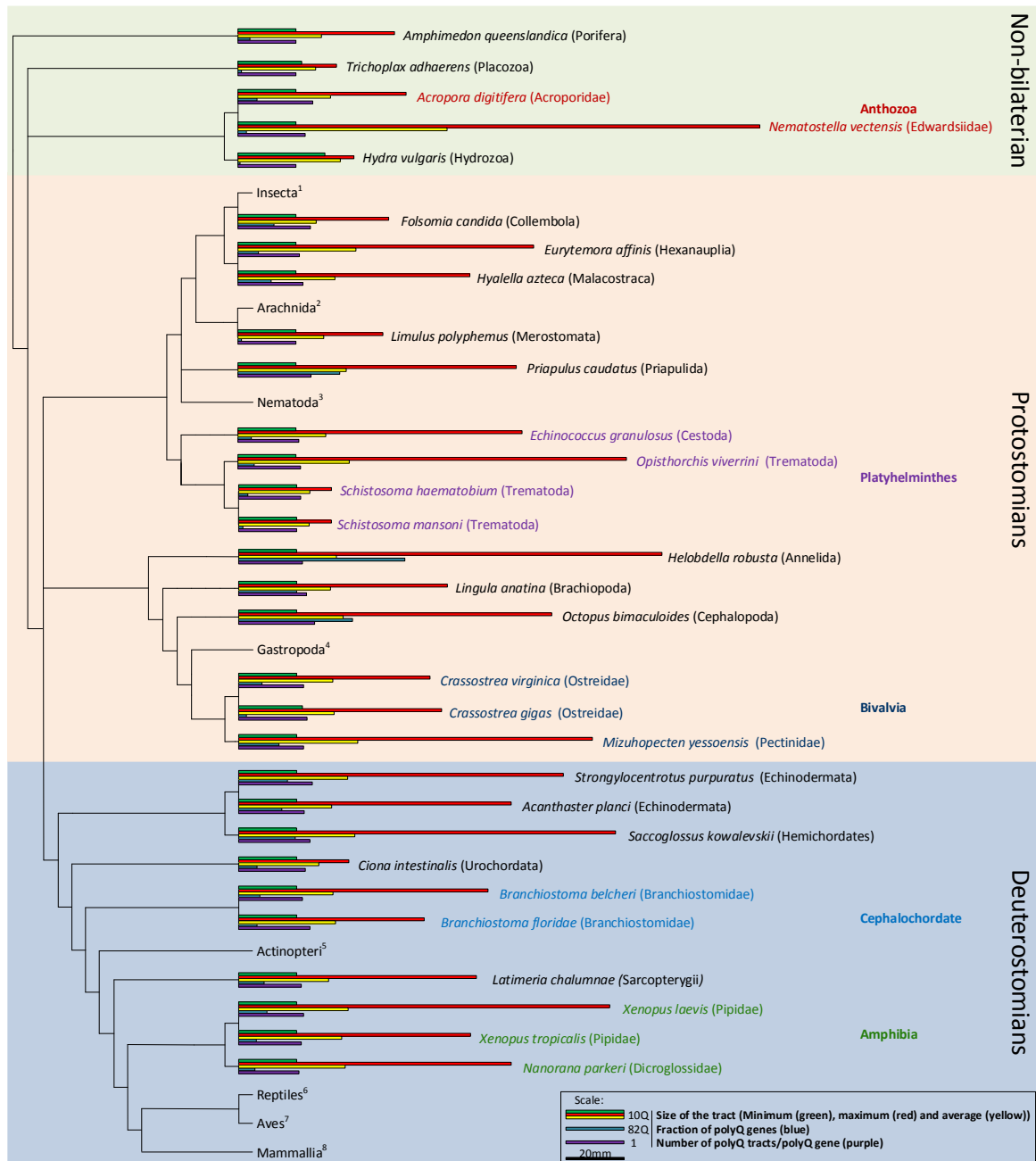


Figure 14. General cladogram showing a high variance in the number and composition of the polyQ genes across the Animalia Kingdom. In this cladogram is represented the Non-bilaterian group and the bilaterian group, with Protostomians and Deuterostomians subgroups. The groups Insecta¹, Arachnida², Nematoda³, Gastropoda⁴, Actinopteri⁵, Reptiles⁶, Aves⁷ and Mammalia⁸ displayed a high variance in the group, and, therefore, where deeper analyzed. The relationships among the different species were made accordingly to (Telford *et al.*, 2015) and (Henriques *et al.*, 2019).

In non-Bilateria only a small fraction of genes encodes proteins with a polyQ, and thus we infer that the ancestral state in animals is to have few such genes. The specie *Nematostella vectensis* shows a distinctive high average tract size, possible due the longest polyQ tract in the low number of polyQ genes (Figure 15-D). Bilateria, showed a higher variance in the number of polyQ genes, especially in the Protostomes group. Indeed, the Insecta (Supplementary figure 6, Supplementary figure 7, Supplementary figure 8, Supplementary figure 9), the Arachnida (Supplementary figure 10), the Nematoda (Supplementary figure 11) and the Gastropods (Supplementary figure 12) classes displayed the most variations, and therefore, were deeper analyzed. The *Drosophila* family presented the distinctive higher fraction of polyQ genes, and therefore, was also deeper analyzed, as presented in the Supplementary figure 9. In the Deuterostomes group the variations in the fraction of polyQ genes is also seen, however, with a lower threshold. The classes Actinopteri (Supplementary figure 13, Supplementary figure 14), Reptiles (Supplementary figure 15), Aves (Supplementary figure 16, Supplementary figure 17, Supplementary figure 18) and Mammalia (Supplementary figure 19, Supplementary figure 20, Supplementary figure 21, Supplementary figure 22) showed the highest variation, being also deeper analyzed. Preliminary observations of these cladograms seem not to disclose a clear evolution pattern for the inclusion of the polyQ tract and the expansion of these repeats.

In general, protostomes have a larger fraction of genes encoding proteins with a polyQ than deuterostomes, but there is ample variation in Protostomes, as presented in the Figure 15-A. The Insecta group display the highest number of genes encoding polyQ proteins, especially in the *Drosophila* group. In *Drosophila* this value is highest, followed by the group of non-*Drosophila* Hymenoptera, the Phthiroptera/Blattodea/Isoptera group and the Lepidoptera. Therefore, there is a decrease in the fraction of genes encoding polyQ proteins as one moves away from *Drosophila* along the phylogeny. Nevertheless, in Coleoptera and Hemiptera there are significantly more genes encoding polyQ proteins than in Lepidoptera, and thus it seems that the fraction of genes encoding polyQ proteins started to increased independently at least twice.

In Deuterostomes, Primates is the one with the highest fraction of genes encoding polyQ proteins. The number of polyQ tracts per polyQ protein follows the pattern observed for the fraction of genes encoding polyQ proteins (when using the average for each group, non-parametric Spearman's rho = 0.488; P=0.047; N=17).

For the average size of the polyQ tract (Figure 15-C), I infer that the ancestral state is to have small tracts, although *N. vectensis* (non-bilateria) shows the highest value for the average tract size, as well as for the minimum tract size. Interestingly, the Lepidoptera group, that showed the smallest fraction of genes encoding polyQ proteins, is the one showing the highest value for average tract size. As one approaches the *Drosophila* group this value decreases, although for insects, as a whole the average size of the polyQ tract and the number of polyQ tracts per gene (Figure 15-B) is not significantly correlated (when using the average for each group, non-parametric Spearman's $\rho = -0.143$; $P=0.787$; $N=6$). Moreover, the putative inverse relationship is not observed for other protostomes (Arachnida vs Nematodes vs Gastropods), and for deuterostomes there is a significant direct correlation between the average size of the polyQ tract and the number of polyQ tracts per gene (when using the average for each group, non-parametric Spearman's $\rho = 0.821$; $P=0.023$; $N=7$).

More than 95% of the species analyzed show at least one polyQ tract with more than 19 repeats, and 36.1% of the species show a polyQ tract with more than 49 repeats. Therefore, at least some genes can tolerate very large polyQ tracts. The average and maximum tract size increases in Deuterostomes as one approaches the Primates species group. For the vast majority of species, the minimum polyQ tract size is 10, the limit imposed by us, as the lower limit for a polyQ tract (Figure 15-E). The detailed distribution per species of the measured quantities can be found in Figures 15(A-E).

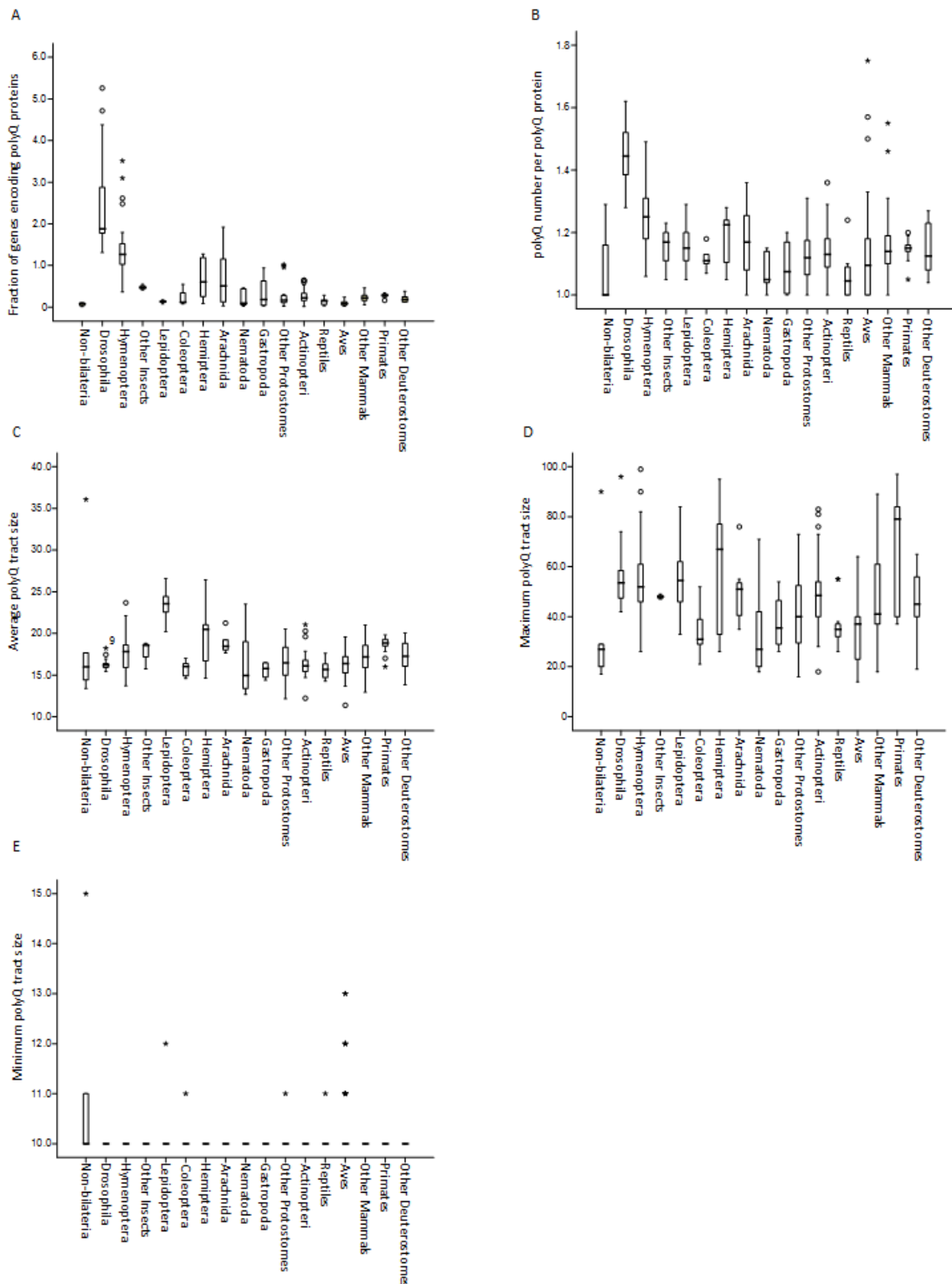


Figure 15. Animalia Kingdom distribution of the fraction of polyQ genes in the genome (A); number of polyQ tracts per polyQ gene (B); average size of the polyQ tracts (C); the polyQ tract with a maximum size (D); and the polyQ tract with the lower size (E).

Discussion

Four interesting observations emerged from the analysis of genes encoding polyQ proteins in animals: 1) the ancestral state seems to be having few genes encoding polyQ proteins, and about a single polyQ per gene of average size 19-20; 2) the number of polyQ tracts per polyQ protein follows the pattern observed for the fraction of genes encoding polyQ proteins; 3) in insects the average size of the polyQ tract and the number of polyQ tracts per gene may be weakly negatively correlated; and 4) in Deuterostomes the average and maximum tract size seems to increase as one approaches the Primates species group.

The observation that in animals the number of polyQ tracts per polyQ protein follows the pattern observed for the fraction of genes encoding polyQ proteins, may be taken as an indication that species groups that tolerate a large fraction of genes encoding polyQ proteins also tolerate a large number of polyQ tracts per polyQ protein. The positive correlation between the average size of the polyQ tract and the number of polyQ tracts per gene also lends support to this interpretation, since species that tolerate well a large number of genes encoding polyQ proteins may also tolerate better larger polyQ tracts. If true, when compared to non-primate mammals, birds, reptiles and fish, Primates seem to be a species group that tolerates especially well polyQ tracts. Indeed, this confirms the reports suggesting that length of orthologous polyQ regions is mostly stable in primates, with a tendency trend polyQ repeat additions in primates (Djian, 1998, Mier and Andrade-Navarro, 2018).

Nevertheless, for insects it is conceivable that species groups such as *Drosophila* show a large fraction of genes encoding polyQ proteins because polyQ tracts may be often interrupted, which could attenuate the deleterious effects of having very long polyQ tracts as well as avoid polyQ tract expansion. This would explain why the species showing the largest fraction of genes encoding polyQ proteins have the highest number of polyQ tracts per gene as well, since the polyQ tract is by definition flanked by two consecutive non-glutamine amino acids. Therefore, a highly degenerated long polyQ tract may be counted as two or more tracts, giving rise to a negative correlation between average tract size and number of tracts per polyQ gene, which is not observed for Deuterostomes. In conclusion, it is likely that the selective and mutational forces acting on different animal groups are different, and that there is no single explanation to account the evolution of polyQ tracts across animals. Nevertheless, as more and more animal

genomes are sequenced, soon it will be possible to study in detail the evolution of polyQ in great detail and infer with certainty the forces shaping polyQ evolution.

Conclusion

This evolutionary approach gave new insights knowledge about the inclusion of the polyQ tract in the genes and their size. The ancestral species seems to have fewer genes encoding polyQ proteins, suggesting that the ancestral state tend to have fewer polyQ genes. Generally, this study suggested there is not a general evolution pattern across the Animalia kingdom for the inclusion of the polyQ tract and the size of the tract in the genes. However, when looking at specific groups, there is certain selection pressures for the inclusion of polyQ tracts or inclusion of glutamines in the tract. These inclusions seem to be independent. The number of polyQ tracts per polyQ seems correlated with the number of polyQ genes. There is a tendency for inclusion of polyQ tracts, specially, in the Insecta group. This Protostome group shows a high number of the number of polyQ tracts per gene, although the average size of the tract being negatively correlated. Therefore, this specific clade seems to have a tolerance for the inclusion of tracts in their genes, even more than one per gene. In the Deuterostomes group, especially in Primates, the average and maximum tract size seems to be higher. Indeed, Primates seem to be a species group that tolerates especially well longer size polyQ tracts. Therefore, this shows that the selective pressure acting on the different animal groups are independent, unveiling that the evolution of polyQ tracts in the Animalia Kingdom is specific for each group analyzed.

General Conclusions

In this study, *in silico*, *in vitro*, and evolutionary analyzes of polyQ proteins were performed to make inferences that could give a better understanding how PPI are affected by polyQ region, as well as the evolution of this type of proteins.

A four step *in silico* approach was developed: prediction of the 3D structures of the proteins using I-TASSER, prediction of the interface residues using C-PORT, docking of the proteins using HADDOCK, and PISAePDB for the analyzes of the PPI at the residues level. This approach was validated using the ATNX1 protein and the 71 interactions, for which *in vitro* data was available, in order to compare the binding interface of wt and expanded ATXN1 forms. Although this methodology has flows, as mentioned in chapter 1, it is useful to infer binding preferences, and can be applied easily to many proteins.

Analyzes of reported ATXN1 PPI confirmed the importance of the AXH domain, with emphasis in three regions of interaction. Furthermore, by comparing the number of interaction residues between ATXN1 and 71 interactors we addressed the effect of N-terminal region, before the expanded polyQ region of the ATXN1, where an increase in the number of interactions is predicted in the expanded form. In this region we identify motifs related with transcription, cell communication, phosphorylation and sumoylation. In-vivo experiments are, however needed to address the importance of these motifs.

To address the importance of the N-terminal region in binding differences between wt and expanded polyQ, I used the *in silico* methodology to analyze the common interactions of seven polyQ proteins related with neurodegenerative diseases and eleven polyQ proteins that are not associated with this type of disease. The results obtained from these analyses revealed that the binding preference for the expanded form is not always due to the N-terminal region. Furthermore, not all polyQ proteins in the expanded form show more binding interactions than wt form. Here I hypothesize that bind preferences could be related with loss / gain of function resulting from the PPI. Contrarily to what is seen specifically for ATXN1, changes in the binding interface between wt and expanded forms occurs mainly in the C-terminal region of the polyQ tract. Further work is needed to address the presence, in the C-terminal region, of motifs related with processes common for polyQ proteins such as ubiquitination, sumoylation and transglutamination.

Using transgenic *Drosophila* with the ATXN1 protein, I validate the importance of the AXH domain in ATXN1 PPI. Furthermore, I found evidence for a large conservation of the human ATXN1 network in *Drosophila*, despite the long evolutionary time of separation of these species. Novel putative interactors were also identified, although further *in silico* analyses are needed to confirm that interaction at the AXH domain is observed.

In order to understand the mechanisms behind the inclusion of the polyQ tract in the genes and the evolution of polyQ size, evolutionary analyzes were also performed. The results showed that the inclusion of the polyQ tract and the size of the tract in the genes is generally similar in the Animalia kingdom. Nevertheless, when looking at specific groups, there is a tendency for inclusion of polyQ tracts, specially, in the Insecta group. Despite this observation, the size of the tract is similar to the average in the Animalia Kingdom. The Primates group showed a higher size of tract relative to what is seen in the Animalia Kingdom. The number of polyQ tracts in the genomes is, however, similar to the average in the Animalia Kingdom. Therefore, this shows that the selective pressure acting on the different animal groups are independent, unveiling that the evolution of polyQ tracts in the Animalia Kingdom is complex.

Future Approaches

The knowledge gathered related to the effects of the expansion of the polyQ tract of ATXN1, especially in the N-Terminal, could give new insights to better understand the associated disease SCA1. In the future, it will be interesting to identify the contribution of the interactors of the protein ATXN1 for the pathogenesis of the disease by gene silencing in *Drosophila* RNAi mutants.

The fact that the pattern found for ATXN1 is not similar when looking to polyQ proteins in general shows that the effects of the expansion should be looked in a particular way, when talking about contribution to the pathogenesis of the disease. In fact, taking in account what is reported, with expansion being associated with gain or loss-of-function effects accordingly to the function of the interaction, the effect of the polyQ expansion on the PPI analyzed could also depend on the associated specific biological function. Therefore, this could be a factor to take in account in future analyzes. Still, a pattern was generally seen, with the binding difference between both forms seen in each interaction explained mostly by the differences in the C-terminal region of the polyQ tract. Therefore, it could be important to address the presence, in this region, of motifs/domains related with pathways associated to processes common for all the polyQ proteins, such as ubiquitination, sumoylation and transglutamination that could be affected by the expansion. This could explain the common phenotype factors/symptoms seen in these diseases, although these proteins are not related.

The validation of the interactions for ATXN1 found together with the conservation of the interactome network for *Drosophila*, confirms the importance of this species for studying polyQ proteins related with neurodegenerative disease. Therefore, it will be useful, address the same in vitro methodology made for the other available lines of other polyQ proteins, to see the effects of the expansion in the interactome of them. The new interactions found could give new insights to better understand the pathogenesis of SCA1. Confirming the novel interactions that were identified, could help understand their contribution for the pathogenesis of SCA1. Further in silico analyses, to compare if the pattern observed for the AXH domain (Overlap in three regions) is similar in these potential new interactions, could support the possibility of being novel interactors of ATXN1.

Identification of the orthologues in all species to identify potential gain/loss of tract in the genes, could help giving new insights to the mechanism of inclusion of this tract.

Bibliography

- AGNARSSON, I., ZAMBRANA-TORRELIO, C. M., FLORES-SALDANA, N. P. & MAY-COLLADO, L. J. 2011. A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). *PLoS Curr*, 3, RRN1212.
- AMBERGER, J. S., BOCCHINI, C. A., SCHIETTECATTE, F., SCOTT, A. F. & HAMOSH, A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*, 43, D789-98.
- ANDERSEN, B. & ROSENFELD, M. G. 2001. POU domain factors in the neuroendocrine system: lessons from developmental biology provide insights into human disease. *Endocr Rev*, 22, 2-35.
- ARAUJO, J., BREUER, P., DIERINGER, S., KRAUSS, S., DORN, S., ZIMMERMANN, K., PFEIFER, A., KLOCKGETHER, T., WUELLNER, U. & EVERT, B. O. 2011. FOXO4-dependent upregulation of superoxide dismutase-2 in response to oxidative stress is impaired in spinocerebellar ataxia type 3. *Hum Mol Genet*, 20, 2928-41.
- ARRASATE, M., MITRA, S., SCHWEITZER, E. S., SEGAL, M. R. & FINKBEINER, S. 2004. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature*, 431, 805-10.
- BAI, S., HE, B. & WILSON, E. M. 2005. Melanoma antigen gene protein MAGE-11 regulates androgen receptor function by modulating the interdomain interaction. *Mol Cell Biol*, 25, 1238-57.
- BARCLAY, S. S., TAMURA, T., ITO, H., FUJITA, K., TAGAWA, K., SHIMAMURA, T., KATSUTA, A., SHIWAKU, H., SONE, M., IMOTO, S., MIYANO, S. & OKAZAWA, H. 2014. Systems biology analysis of Drosophila in vivo screen data elucidates core networks for DNA damage repair in SCA1. *Hum Mol Genet*, 23, 1345-64.
- BAUER, P. O. & NUKINA, N. 2009. The pathogenic mechanisms of polyglutamine diseases and current therapeutic strategies. *J Neurochem*, 110, 1737-65.
- BELIAKOFF, J. & SUN, Z. 2006. Zimp7 and Zimp10, two novel PIAS-like proteins, function as androgen receptor coregulators. *Nucl Recept Signal*, 4, e017.
- BERKE, S. J., CHAI, Y., MARRS, G. L., WEN, H. & PAULSON, H. L. 2005. Defining the role of ubiquitin-interacting motifs in the polyglutamine disease protein, ataxin-3. *J Biol Chem*, 280, 32026-34.
- BERTONE, M. A., COURTNEY, G. W. & WIEGMANN, B. M. 2008. Phylogenetics and temporal diversification of the earliest true flies (Insecta: Diptera) based on multiple nuclear genes. *Systematic Entomology*, 33, 668-687.
- BETANCUR, R. R., BROUGHTON, R. E., WILEY, E. O., CARPENTER, K., LOPEZ, J. A., LI, C., HOLCROFT, N. I., ARCILA, D., SANCIANGCO, M., CURETON II, J. C., ZHANG, F., BUSER, T., CAMPBELL, M. A., BALLESTEROS, J. A., ROAVARON, A., WILLIS, S., BORDEN, W. C., ROWLEY, T., RENEAU, P. C., HOUGH, D. J., LU, G., GRANDE, T., ARRATIA, G. & ORTI, G. 2013. The tree of life and a new classification of bony fishes. *PLoS Curr*, 5.
- BOLGER, T. A., ZHAO, X., COHEN, T. J., TSAI, C. C. & YAO, T. P. 2007. The neurodegenerative disease protein ataxin-1 antagonizes the neuronal survival function of myocyte enhancer factor-2. *J Biol Chem*, 282, 29186-92.
- BONDAR, V. V., ADAMSKI, C. J., ONUR, T. S., TAN, Q., WANG, L., DIAZ-GARCIA, J., PARK, J., ORR, H. T., BOTAS, J. & ZOGHBI, H. Y. 2018. PAK1 regulates ATXN1 levels providing an opportunity to modify its toxicity in spinocerebellar ataxia type 1. *Hum Mol Genet*, 27, 2863-2873.
- BORATYN, G. M., CAMACHO, C., COOPER, P. S., COULOURIS, G., FONG, A., MA, N., MADDEN, T. L., MATTEN, W. T., MCGINNIS, S. D., MERZHUH, Y.,

- RAYTSELIS, Y., SAYERS, E. W., TAO, T., YE, J. & ZARETSKAYA, I. 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res*, 41, W29-33.
- BRENDEL, C., GELMAN, L. & AUWERX, J. 2002. Multiprotein bridging factor-1 (MBF-1) is a cofactor for nuclear receptors that regulate lipid metabolism. *Mol Endocrinol*, 16, 1367-77.
- BREZA, M. & KOUTSIS, G. 2018. Kennedy's disease (spinal and bulbar muscular atrophy): a clinically oriented review of a rare disease. *Journal of Neurology*.
- BRODA, M., KIERZEK, E., GDANIEC, Z., KULINSKI, T. & KIERZEK, R. 2005. Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry*, 44, 10873-82.
- CHAKRAVARTY, D., GUHARROY, M., ROBERT, C. H., CHAKRABARTI, P. & JANIN, J. 2013. Reassessing buried surface areas in protein-protein complexes. *Protein Sci*, 22, 1453-7.
- CHEN, H. K., FERNANDEZ-FUNEZ, P., ACEVEDO, S. F., LAM, Y. C., KAYTOR, M. D., FERNANDEZ, M. H., AITKEN, A., SKOULAKIS, E. M., ORR, H. T., BOTAS, J. & ZOGHBI, H. Y. 2003. Interaction of Akt-phosphorylated ataxin-1 with 14-3-3 mediates neurodegeneration in spinocerebellar ataxia type 1. *Cell*, 113, 457-68.
- CHOI, D. H., HA, J. S., LEE, W. H., SONG, J. K., KIM, G. Y., PARK, J. H., CHA, H. J., LEE, B. J. & PARK, J. W. 2007. Heat shock protein 27 is associated with irinotecan resistance in human colorectal cancer cells. *FEBS Lett*, 581, 1649-56.
- CHOU, A. H., CHEN, Y. L., HU, S. H., CHANG, Y. M. & WANG, H. L. 2014. Polyglutamine-expanded ataxin-3 impairs long-term depression in Purkinje neurons of SCA3 transgenic mouse by inhibiting HAT and impairing histone acetylation. *Brain Res*, 1583, 220-9.
- CLOUAIRE, T., ROUSSIGNE, M., ECOCHARD, V., MATHE, C., AMALRIC, F. & GIRARD, J. P. 2005. The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. *Proc Natl Acad Sci U S A*, 102, 6907-12.
- COHEN-CARMON, D. & MESHORER, E. 2012. Polyglutamine (polyQ) disorders: the chromatin connection. *Nucleus*, 3, 433-41.
- DANTONEL, J. C., MURTHY, K. G., MANLEY, J. L. & TORA, L. 1997. Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. *Nature*, 389, 399-402.
- DAVIES, P., WATT, K., KELLY, S. M., CLARK, C., PRICE, N. C. & MCEWAN, I. J. 2008. Consequences of poly-glutamine repeat length for the conformation and folding of the androgen receptor amino-terminal domain. *J Mol Endocrinol*, 41, 301-14.
- DE CHIARA, C., MENON, R. P., DAL PIAZ, F., CALDER, L. & PASTORE, A. 2005. Polyglutamine is not all: the functional role of the AXH domain in the ataxin-1 protein. *J Mol Biol*, 354, 883-93.
- DE CHIARA, C. & PASTORE, A. 2011. Prediction and Experimental Detection of Structural and Functional Motifs in Intrinsically Unfolded Proteins.
- DE CHIARA, C. & PASTORE, A. 2014. Kaleidoscopic protein-protein interactions in the life and death of ataxin-1: new strategies against protein aggregation. *Trends in neurosciences*, 37, 211-218.
- DE CHIARA, C., REES, M., MENON, R. P., PAUWELS, K., LAWRENCE, C., KONAREV, P. V., SVERGUN, D. I., MARTIN, S. R., CHEN, Y. W. & PASTORE, A. 2013. Self-assembly and conformational heterogeneity of the AXH domain of ataxin-1: an unusual example of a chameleon fold. *Biophys J*, 104, 1304-13.
- DE VRIES, S. J. & BONVIN, A. M. 2011. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One*, 6, e17695.

- DELORENZI, M. & SPEED, T. 2002. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, 18, 617-25.
- DERIU, M. A., GRASSO, G., LICANDRO, G., DANANI, A., GALLO, D., TUSZYNSKI, J. A. & MORBIDUCCI, U. 2014. Investigation of the Josephin Domain protein-protein interaction by molecular dynamics. *PLoS One*, 9, e108677.
- DIRAC, A. M. & BERNARDS, R. 2010. The deubiquitinating enzyme USP26 is a regulator of androgen receptor signaling. *Mol Cancer Res*, 8, 844-54.
- DJIAN, P. 1998. Evolution of Simple Repeats in DNA and Their Relation to Human Disease. *Cell*, 94, 155-160.
- DOMINGUEZ, C., BOELEN, R. & BONVIN, A. M. 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125, 1731-7.
- DOSS-PEPE, E. W., STENROOS, E. S., JOHNSON, W. G. & MADURA, K. 2003. Ataxin-3 interactions with rad23 and valosin-containing protein and its associations with ubiquitin chains and the proteasome are consistent with a role in ubiquitin-mediated proteolysis. *Mol Cell Biol*, 23, 6469-83.
- DUFFY, J. B. 2002. GAL4 system in Drosophila: a fly geneticist's Swiss army knife. *Genesis*, 34, 1-15.
- DUNCAN, C. E., AN, M. C., PAPANIKOLAOU, T., RUGANI, C., VITELLI, C. & ELLERBY, L. M. 2013. Histone deacetylase-3 interacts with ataxin-7 and is altered in a spinocerebellar ataxia type 7 mouse model. *Molecular Neurodegeneration*, 8, 42.
- DURCAN, T. M., KONTOGIANNEA, M., BEDARD, N., WING, S. S. & FON, E. A. 2012. Ataxin-3 deubiquitination is coupled to Parkin ubiquitination via E2 ubiquitin-conjugating enzyme. *J Biol Chem*, 287, 531-41.
- ELDEN, A. C., KIM, H. J., HART, M. P., CHEN-PLOTKIN, A. S., JOHNSON, B. S., FANG, X., ARMAKOLA, M., GESER, F., GREENE, R., LU, M. M., PADMANABHAN, A., CLAY-FALCONE, D., MCCLUSKEY, L., ELMAN, L., JUHR, D., GRUBER, P. J., RUB, U., AUBURGER, G., TROJANOWSKI, J. Q., LEE, V. M., VAN DEERLIN, V. M., BONINI, N. M. & GITLER, A. D. 2010. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, 466, 1069-75.
- EMAMIAN, E. S., KAYTOR, M. D., DUVICK, L. A., ZU, T., TOUSEY, S. K., ZOGHBI, H. Y., CLARK, H. B. & ORR, H. T. 2003. Serine 776 of ataxin-1 is critical for polyglutamine-induced disease in SCA1 transgenic mice. *Neuron*, 38, 375-87.
- EMERY, P. 2007. Protein extraction from Drosophila heads. *Methods Mol Biol*, 362, 375-7.
- ENGEL, M. S. 2001. Monophyly and extensive extinction of advanced eusocial bees: insights from an unexpected Eocene diversity. *Proc Natl Acad Sci U S A*, 98, 1661-4.
- ESTEBANEZ-PERPINA, E., MOORE, J. M., MAR, E., DELGADO-RODRIGUES, E., NGUYEN, P., BAXTER, J. D., BUEHRER, B. M., WEBB, P., FLETTERICK, R. J. & GUY, R. K. 2005. The molecular mechanisms of coactivator utilization in ligand-dependent transactivation by the androgen receptor. *J Biol Chem*, 280, 8060-8.
- EVERT, B. O., ARAUJO, J., VIEIRA-SAECKER, A. M., DE VOS, R. A., HARENDZA, S., KLOCKGETHER, T. & WULLNER, U. 2006. Ataxin-3 represses transcription via chromatin binding, interaction with histone deacetylase 3, and histone deacetylation. *J Neurosci*, 26, 11474-86.
- FAN, H. C., HO, L. I., CHI, C. S., CHEN, S. J., PENG, G. S., CHAN, T. M., LIN, S. Z. & HARN, H. J. 2014. Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transplant*, 23, 441-58.
- FAUX, N. G., BOTTOMLEY, S. P., LESK, A. M., IRVING, J. A., MORRISON, J. R., DE LA BANDA, M. G. & WHISSTOCK, J. C. 2005. Functional insights from the

- distribution and role of homopeptide repeat-containing proteins. *Genome Res*, 15, 537-51.
- FERNANDEZ-FUNEZ, P., NINO-ROSALES, M. L., DE GOUYON, B., SHE, W. C., LUCHAK, J. M., MARTINEZ, P., TURIEGANO, E., BENITO, J., CAPOVILLA, M., SKINNER, P. J., MCCALL, A., CANAL, I., ORR, H. T., ZOGHBI, H. Y. & BOTAS, J. 2000. Identification of genes that modify ataxin-1-induced neurodegeneration. *Nature*, 408, 101-6.
- FIGUEROA, A., MCKELVY, A. D., GRISMER, L. L., BELL, C. D. & LAILVAUX, S. P. 2016. A Species-Level Phylogeny of Extant Snakes with Description of a New Colubrid Subfamily and Genus. *PLoS One*, 11, e0161070.
- FIUMARA, F., FIORITI, L., KANDEL, E. R. & HENDRICKSON, W. A. 2010. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell*, 143, 1121-35.
- FONDON, J. W., 3RD & GARNER, H. R. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*, 101, 18058-63.
- FRIEDMAN, M. J., LI, S. & LI, X.-J. 2009. Activation of Gene Transcription by Heat Shock Protein 27 May Contribute to Its Neuronal Protection. *Journal of Biological Chemistry*, 284, 27944-27951.
- FRIEDMAN, M. J., SHAH, A. G., FANG, Z. H., WARD, E. G., WARREN, S. T., LI, S. & LI, X. J. 2007. Polyglutamine domain modulates the TBP-TFIIB interaction: implications for its normal function and neurodegeneration. *Nat Neurosci*, 10, 1519-28.
- FRIEDMAN, M. J., WANG, C. E., LI, X. J. & LI, S. 2008. Polyglutamine expansion reduces the association of TATA-binding protein with DNA and induces DNA binding-independent neurotoxicity. *J Biol Chem*, 283, 8283-90.
- FRIEDRICH, J., KORDASIEWICZ, H. B., O'CALLAGHAN, B., HANDLER, H. P., WAGENER, C., DUVICK, L., SWAYZE, E. E., RAINWATER, O., HOFSTRA, B., BENNEYWORTH, M., NICHOLS-MEADE, T., YANG, P., CHEN, Z., ORTIZ, J. P., CLARK, H. B., OZ, G., LARSON, S., ZOGHBI, H. Y., HENZLER, C. & ORR, H. T. 2018. Antisense oligonucleotide-mediated ataxin-1 reduction prolongs survival in SCA1 mice and reveals disease-associated transcriptome profiles. *JCI Insight*, 3.
- FRYER, J. D., YU, P., KANG, H., MANDEL-BREHM, C., CARTER, A. N., CRESPO-BARRETO, J., GAO, Y., FLORA, A., SHAW, C., ORR, H. T. & ZOGHBI, H. Y. 2011. Exercise and genetic rescue of SCA1 via the transcriptional repressor Capicua. *Science*, 334, 690-3.
- FUJITA, K., NAKAMURA, Y., OKA, T., ITO, H., TAMURA, T., TAGAWA, K., SASABE, T., KATSUTA, A., MOTOKI, K., SHIWAKU, H., SONE, M., YOSHIDA, C., KATSUNO, M., EISHI, Y., MURATA, M., TAYLOR, J. P., WANKER, E. E., KONO, K., TASHIRO, S., SOBUE, G., LA SPADA, A. R. & OKAZAWA, H. 2013. A functional deficiency of TERA/VCP/p97 contributes to impaired DNA repair in multiple polyglutamine diseases. *Nat Commun*, 4, 1816.
- GARDIAN, G., BROWNE, S. E., CHOI, D. K., KLIVENYI, P., GREGORIO, J., KUBILUS, J. K., RYU, H., LANGLEY, B., RATAN, R. R., FERRANTE, R. J. & BEAL, M. F. 2005. Neuroprotective effects of phenylbutyrate in the N171-82Q transgenic mouse model of Huntington's disease. *J Biol Chem*, 280, 556-63.
- GEORGESCU, S. P., LI, J. H., LU, Q., KARAS, R. H., BROWN, M. & MENDELSON, M. E. 2005. Modulator recognition factor 1, an AT-rich interaction domain family member, is a novel corepressor for estrogen receptor alpha. *Mol Endocrinol*, 19, 2491-501.
- GITLER, A. D., DHILLON, P. & SHORTER, J. 2017. Neurodegenerative disease: models, mechanisms, and a new hope. *Dis Model Mech*, 10, 499-502.

- GLISOVIC, T., BACHORIK, J. L., YONG, J. & DREYFUSS, G. 2008. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett*, 582, 1977-86.
- GOLDING, G. B. 1999. Simple sequence is abundant in eukaryotic proteins. *Protein Sci*, 8, 1358-61.
- GOOLD, R., HUBANK, M., HUNT, A., HOLTON, J., MENON, R. P., REVESZ, T., PANDOLFO, M. & MATILLA-DUENAS, A. 2007. Down-regulation of the dopamine receptor D2 in mice lacking ataxin 1. *Hum Mol Genet*, 16, 2122-34.
- GORDON, V., BHADEL, S., WUNDERLICH, W., ZHANG, J., FICARRO, S. B., MOLLAH, S. A., SHABANOWITZ, J., HUNT, D. F., XENARIOS, I., HAHN, W. C., CONAWAY, M., CAREY, M. F. & GIOELI, D. 2010. CDK9 regulates AR promoter selectivity and cell growth through serine 81 phosphorylation. *Mol Endocrinol*, 24, 2267-80.
- HANDS, S. L. & WYTTEBACH, A. 2010. Neurotoxic protein oligomerisation associated with polyglutamine diseases. *Acta Neuropathologica*, 120, 419-437.
- HARJES, P. & WANKER, E. E. 2003. The hunt for huntingtin function: interaction partners tell many different stories. *Trends Biochem Sci*, 28, 425-33.
- HEEMELS, M. T. 2016. Neurodegenerative diseases. *Nature*, 539, 179.
- HELGEN, K. M. 2011. Evolution. The mammal family tree. *Science*, 334, 458-9.
- HEMLINGER, D., HARDY, S., SASORITH, S., KLEIN, F., ROBERT, F., WEBER, C., MIGUET, L., POTIER, N., VAN-DORSSELAER, A., WURTZ, J. M., MANDEL, J. L., TORA, L. & DEVYS, D. 2004. Ataxin-7 is a subunit of GCN5 histone acetyltransferase-containing complexes. *Hum Mol Genet*, 13, 1257-65.
- HELMSTETTER, A. J., PAPADOPULOS, A. S., IGEA, J., VAN DOOREN, T. J., LEROI, A. M. & SAVOLAINEN, V. 2016. Viviparity stimulates diversification in an order of fish. *Nat Commun*, 7, 11271.
- HENRIQUES, S. F., DUQUE, P., LOPEZ-FERNANDEZ, H., VAZQUEZ, N., FDEZ-RIVEROLA, F., REBOIRO-JATO, M., VIEIRA, C. P. & VIEIRA, J. 2019. Multiple independent L-gulonolactone oxidase (GULO) gene losses and vitamin C synthesis reacquisition events in non-Deuterostomian animal species. *BMC Evol Biol*, 19, 126.
- HOSP, F., VOSSFELDT, H., HEINIG, M., VASILJEVIC, D., ARUMUGHAN, A., WYLER, E., GENETIC, ENVIRONMENTAL RISK FOR ALZHEIMER'S DISEASE, G. C., LANDTHALER, M., HUBNER, N., WANKER, E. E., LANNFELT, L., INGELSSON, M., LALOWSKI, M., VOIGT, A. & SELBACH, M. 2015. Quantitative interaction proteomics of neurodegenerative disease proteins. *Cell Rep*, 11, 1134-46.
- HOUSMAN, D. 1995. Gain of glutamines, gain of function? *Nat Genet*, 10, 3-4.
- HU, Y., FLOCKHART, I., VINAYAGAM, A., BERGWITZ, C., BERGER, B., PERRIMON, N. & MOHR, S. E. 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, 12, 357.
- HUANG, D. W., SHERMAN, B. T., TAN, Q., KIR, J., LIU, D., BRYANT, D., GUO, Y., STEPHENS, R., BASELER, M. W., LANE, H. C. & LEMPICKI, R. A. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*, 35, W169-75.
- HUGHES, L. C., ORTI, G., HUANG, Y., SUN, Y., BALDWIN, C. C., THOMPSON, A. W., ARCILA, D., BETANCUR, R. R., LI, C., BECKER, L., BELLORA, N., ZHAO, X., LI, X., WANG, M., FANG, C., XIE, B., ZHOU, Z., HUANG, H., CHEN, S., VENKATESH, B. & SHI, Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc Natl Acad Sci U S A*, 115, 6249-6254.
- HUNTLEY, M. & GOLDING, G. B. 2000. Evolution of simple sequence in proteins. *J Mol Evol*, 51, 131-40.

- HUNTLEY, M. A. & CLARK, A. G. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol*, 24, 2598-609.
- HUTTLIN, E. L., JEDRYCHOWSKI, M. P., ELIAS, J. E., GOSWAMI, T., RAD, R., BEAUSOLEIL, S. A., VILLEN, J., HAAS, W., SOWA, M. E. & GYGI, S. P. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143, 1174-89.
- HUYNH, D. P., YANG, H. T., VAKHARIA, H., NGUYEN, D. & PULST, S. M. 2003. Expansion of the polyQ repeat in ataxin-2 alters its Golgi localization, disrupts the Golgi complex and causes cell death. *Hum Mol Genet*, 12, 1485-96.
- IIZUKA, M. & STILLMAN, B. 1999. Histone acetyltransferase HBO1 interacts with the ORC1 subunit of the human initiator protein. *J Biol Chem*, 274, 23027-34.
- IRWIN, S., VANDELFT, M., PINCHEV, D., HOWELL, J. L., GRACZYK, J., ORR, H. T. & TRUANT, R. 2005. RNA association and nucleocytoplasmic shuttling by ataxin-1. *J Cell Sci*, 118, 233-42.
- JEONG, J. H., JIN, J. S., KIM, H. N., KANG, S. M., LIU, J. C., LENGNER, C. J., OTTO, F., MUNDLOS, S., STEIN, J. L., VAN WIJNEN, A. J., LIAN, J. B., STEIN, G. S. & CHOI, J. Y. 2008. Expression of Runx2 transcription factor in non-skeletal tissues, sperm and brain. *J Cell Physiol*, 217, 511-7.
- JIMÉNEZ-LÓPEZ, D. & GUZMÁN, P. 2014. Insights into the evolution and domain structure of ataxin-2 proteins across eukaryotes. *BMC Research Notes*, 7, 453.
- JONES, M. E., WERNEBURG, I., CURTIS, N., PENROSE, R., O'HIGGINS, P., FAGAN, M. J. & EVANS, S. E. 2012. The head and neck anatomy of sea turtles (Cryptodira: Chelonioidae) and skull shape in Testudines. *PLoS One*, 7, e47852.
- JU, H., KOKUBU, H. & LIM, J. 2014. Beyond the glutamine expansion: influence of posttranslational modifications of ataxin-1 in the pathogenesis of spinocerebellar ataxia type 1. *Mol Neurobiol*, 50, 866-874.
- JU, H., KOKUBU, H., TODD, T. W., KAHLE, J. J., KIM, S., RICHMAN, R., CHIRALA, K., ORR, H. T., ZOGHBI, H. Y. & LIM, J. 2013. Polyglutamine disease toxicity is regulated by Nemo-like kinase in spinocerebellar ataxia type 1. *J Neurosci*, 33, 9328-36.
- KANG, A. R., PARK, S. H., LEE, S., CHOI, D. Y., KIM, K. P., SONG, H. K., HONG, S. & KANG, S. 2015. A key lysine residue in the AXH domain of ataxin-1 is essential for its ubiquitylation. *Biochim Biophys Acta*, 1854, 356-64.
- KANG, S. & HONG, S. 2010. SUMO-1 interacts with mutant ataxin-1 and colocalizes to its aggregates in Purkinje cells of SCA1 transgenic mice. *Arch Ital Biol*, 148, 351-63.
- KANG, Z., JANNE, O. A. & PALVIMO, J. J. 2004. Coregulator recruitment and histone modifications in transcriptional regulation by the androgen receptor. *Mol Endocrinol*, 18, 2633-48.
- KARLIN, S. & BURGE, C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proceedings of the National Academy of Sciences*, 93, 1560-1565.
- KATSUNO, M., ADACHI, H., WAZA, M., BANNO, H., SUZUKI, K., TANAKA, F., DOYU, M. & SOBUE, G. 2006. Pathogenesis, animal models and therapeutics in spinal and bulbar muscular atrophy (SBMA). *Exp Neurol*, 200, 8-18.
- KAWAGUCHI, Y., OKAMOTO, T., TANIWAKI, M., AIZAWA, M., INOUE, M., KATAYAMA, S., KAWAKAMI, H., NAKAMURA, S., NISHIMURA, M., AKIGUCHI, I. & ET AL. 1994. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet*, 8, 221-8.
- KAWAHARA, A. Y. & BREINHOLT, J. W. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proc Biol Sci*, 281, 20140970.

- KIM, C. A. & BOWIE, J. U. 2003. SAM domains: uniform structure, diversity of function. *Trends Biochem Sci*, 28, 625-8.
- KIM, E., LEE, Y., CHOI, S. & SONG, J. J. 2014. Structural basis of the phosphorylation dependent complex formation of neurodegenerative disease protein Ataxin-1 and RBM17. *Biochem Biophys Res Commun*, 449, 399-404.
- KIM, H., LEE, S. & JANG, Y. 2011. Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. *PLoS One*, 6, e24749.
- KLEIN, F. A., PASTORE, A., MASINO, L., ZEDER-LUTZ, G., NIERENGARTEN, H., OULAD-ABDELGHANI, M., ALTSCHUH, D., MANDEL, J. L. & TROTTIER, Y. 2007. Pathogenic and non-pathogenic polyglutamine tracts have similar structural properties: towards a length-dependent toxicity gradient. *J Mol Biol*, 371, 235-44.
- KLEMENT, I. A., SKINNER, P. J., KAYTOR, M. D., YI, H., HERSCH, S. M., CLARK, H. B., ZOGHBI, H. Y. & ORR, H. T. 1998. Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in SCA1 transgenic mice. *Cell*, 95, 41-53.
- KOIDE, R., KOBAYASHI, S., SHIMOHATA, T., IKEUCHI, T., MARUYAMA, M., SAITO, M., YAMADA, M., TAKAHASHI, H. & TSUJI, S. 1999. A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? *Hum Mol Genet*, 8, 2047-53.
- KOSHY, B., MATILLA, T., BURRIGHT, E. N., MERRY, D. E., FISCHBECK, K. H., ORR, H. T. & ZOGHBI, H. Y. 1996. Spinocerebellar ataxia type-1 and spinobulbar muscular atrophy gene products interact with glyceraldehyde-3-phosphate dehydrogenase. *Hum Mol Genet*, 5, 1311-8.
- KOYANO-NAKAGAWA, N., NISHIDA, J., BALDWIN, D., ARAI, K. & YOKOTA, T. 1994. Molecular cloning of a novel human cDNA encoding a zinc finger protein that binds to the interleukin-3 promoter. *Mol Cell Biol*, 14, 5099-107.
- KOZLOV, G., SAFAR, N., ROSENAUER, A. & GEHRING, K. 2010. Structural basis of binding of P-body-associated proteins GW182 and ataxin-2 by the Mlle domain of poly(A)-binding protein. *J Biol Chem*, 285, 13599-606.
- KRATTER, I. H. & FINKBEINER, S. 2010. PolyQ disease: too many Qs, too much function? *Neuron*, 67, 897-9.
- KRISSINEL, E. & HENRICK, K. 2007. Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, 372, 774-97.
- KRISTENSEN, L. V., OPPERMANN, F. S., RAUEN, M. J., FOG, K., SCHMIDT, T., SCHMIDT, J., HARMUTH, T., HARTMANN-PETERSEN, R. & THIRSTRUP, K. 2018. Mass spectrometry analyses of normal and polyglutamine expanded ataxin-3 reveal novel interaction partners involved in mitochondrial function. *Neurochem Int*, 112, 5-17.
- KUFAREVA, I., BUDAGYAN, L., RAUSH, E., TOTROV, M. & ABAGYAN, R. 2007. PIER: protein interface recognition for structural proteomics. *Proteins*, 67, 400-17.
- LA SPADA, A. R. & TAYLOR, J. P. 2010. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat Rev Genet*, 11, 247-58.
- LADD, A. N., CHARLET, N. & COOPER, T. A. 2001. The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. *Mol Cell Biol*, 21, 1285-96.
- LAM, Y. C., BOWMAN, A. B., JAFAR-NEJAD, P., LIM, J., RICHMAN, R., FRYER, J. D., HYUN, E. D., DUVICK, L. A., ORR, H. T., BOTAS, J. & ZOGHBI, H. Y. 2006. ATAXIN-1 interacts with the repressor Capicua in its native complex to cause SCA1 neuropathology. *Cell*, 127, 1335-47.

- LAMIA, K. A., PAPP, S. J., YU, R. T., BARISH, G. D., UHLENHAUT, N. H., JONKER, J. W., DOWNES, M. & EVANS, R. M. 2011. Cryptochromes mediate rhythmic repression of the glucocorticoid receptor. *Nature*, 480, 552-6.
- LAU, K. F., MILLER, C. C., ANDERTON, B. H. & SHAW, P. C. 1999. Molecular cloning and characterization of the human glycogen synthase kinase-3 β promoter. *Genomics*, 60, 121-8.
- LEBRE, A.-S., JAMOT, L., TAKAHASHI, J., SPASSKY, N., LEPRINCE, C., RAVISÉ, N., ZANDER, C., FUJIGASAKI, H., KUSSEL-ANDERMANN, P., DUYCKAERTS, C., CAMONIS, J. H. & BRICE, A. 2001. Ataxin-7 interacts with a Cbl-associated protein that it recruits into neuronal intranuclear inclusions. *Human Molecular Genetics*, 10, 1201-1213.
- LEDERKREMER, G. Z., CHENG, Y., PETRE, B. M., VOGAN, E., SPRINGER, S., SCHEKMAN, R., WALZ, T. & KIRCHHAUSEN, T. 2001. Structure of the Sec23p/24p and Sec13p/31p complexes of COPII. *Proc Natl Acad Sci U S A*, 98, 10704-9.
- LEE, C. J., CHAN, W. I., CHEUNG, M., CHENG, Y. C., APPLEBY, V. J., ORME, A. T. & SCOTTING, P. J. 2002. CIC, a member of a novel subfamily of the HMG-box superfamily, is transiently expressed in developing granule neurons. *Brain Res Mol Brain Res*, 106, 151-6.
- LEE, Y., FRYER, J. D., KANG, H., CRESPO-BARRETO, J., BOWMAN, A. B., GAO, Y., KAHLE, J. J., HONG, J. S., KHERADMAND, F., ORR, H. T., FINEGOLD, M. J. & ZOGHBI, H. Y. 2011. ATXN1 protein family and CIC regulate extracellular matrix remodeling and lung alveolarization. *Dev Cell*, 21, 746-57.
- LEIPE, D. D., KOONIN, E. V. & ARAVIND, L. 2004. STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol*, 343, 1-28.
- LEISTER, P., FELTEN, A., CHASAN, A. I. & SCHEIDTMANN, K. H. 2008. ZIP kinase plays a crucial role in androgen receptor-mediated transcription. *Oncogene*, 27, 3292-300.
- LI, B., LU, W. & CHEN, Z. 2014. Regulation of Androgen Receptor by E3 Ubiquitin Ligases: for More or Less. *Receptors Clin Investig*, 1.
- LI, C., NAGEL, J., ANDROULAKIS, S., SONG, J. & BUCKLE, A. M. 2016. PolyQ 2.0: an improved version of PolyQ, a database of human polyglutamine proteins. *Database (Oxford)*, 2016.
- LIAN, W. X., YIN, R. H., KONG, X. Z., ZHANG, T., HUANG, X. H., ZHENG, W. W., YANG, Y., ZHAN, Y. Q., XU, W. X., YU, M., GE, C. H., GUO, J. T., LI, C. Y. & YANG, X. M. 2012. THAP11, a novel binding protein of PCBP1, negatively regulates CD44 alternative splicing and cell invasion in a human hepatoma cell line. *FEBS Lett*, 586, 1431-8.
- LIM, J., CRESPO-BARRETO, J., JAFAR-NEJAD, P., BOWMAN, A. B., RICHMAN, R., HILL, D. E., ORR, H. T. & ZOGHBI, H. Y. 2008. Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1. *Nature*, 452, 713-8.
- LIM, J., HAO, T., SHAW, C., PATEL, A. J., SZABÓ, G., RUAL, J.-F., FISK, C. J., LI, N., SMOLYAR, A., HILL, D. E., BARABÁSI, A.-L., VIDAL, M. & ZOGHBI, H. Y. 2006. A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. *Cell*, 125, 801-814.
- LOBANOV, M. Y., KLUS, P., SOKOLOVSKY, I. V., TARTAGLIA, G. G. & GALZITSKAYA, O. V. 2016. Non-random distribution of homo-repeats: links with biological functions and human diseases. *Sci Rep*, 6, 26941.
- LOPEZ-FERNANDEZ, H., DUQUE, P., HENRIQUES, S., VAZQUEZ, N., FDEZ-RIVEROLA, F., VIEIRA, C. P., REBOIRO-JATO, M. & VIEIRA, J. 2018.

- Bioinformatics Protocols for Quickly Obtaining Large-Scale Data Sets for Phylogenetic Inferences. *Interdiscip Sci*.
- MA, Q., ZHOU, L., SHI, H. & HUO, K. 2008. NUMBL interacts with TAB2 and inhibits TNFalpha and IL-1beta-induced NF-kappaB activation. *Cell Signal*, 20, 1044-51.
- MACDONALD, M. E. 2003. Huntingtin: Alive and Well and Working in Middle Management. *Science's STKE*, 2003, pe48.
- MADDISON, D. R. A. K. S. S. E. 2007. *The Tree of Life Web* [Online]. Available: <http://tolweb.org> [Accessed].
- MAEDA, T., GUPTA, M. P. & STEWART, A. F. 2002. TEF-1 and MEF2 transcription factors interact to regulate muscle-specific promoters. *Biochem Biophys Res Commun*, 294, 791-7.
- MANGELSDORF, D., THUMMEL, C., BEATO, M., HERRLICH, P., SCHÜTZ, G., UMESONO, K., BLUMBERG, B., KASTNER, P., MARK, M. & CHAMBON, P. 1995. *Evans, R.M.*
- MARKOW, T. A. 2015. The secret lives of Drosophila flies. *Elife*, 4.
- MASINO, L. 2004. Polyglutamine and neurodegeneration: structural aspects. *Protein Pept Lett*, 11, 239-48.
- MATILLA-DUENAS, A., GOOLD, R. & GIUNTI, P. 2008. Clinical, genetic, molecular, and pathophysiological insights into spinocerebellar ataxia type 1. *Cerebellum*, 7, 106-14.
- MCCULLOUGH, S. D., XU, X., DENT, S. Y. R., BEKIRANOV, S., ROEDER, R. G. & GRANT, P. A. 2012. Reelin is a target of polyglutamine expanded ataxin-7 in human spinocerebellar ataxia type 7 (SCA7) astrocytes. *Proceedings of the National Academy of Sciences*, 109, 21319-21324.
- MCGRATH, M. J., BINGE, L. C., SRIRATANA, A., WANG, H., ROBINSON, P. A., POOK, D., FEDELE, C. G., BROWN, S., DYSON, J. M., COTTLE, D. L., COWLING, B. S., NIRANJAN, B., RISBRIDGER, G. P. & MITCHELL, C. A. 2013. Regulation of the transcriptional coactivator FHL2 licenses activation of the androgen receptor in castrate-resistant prostate cancer. *Cancer Res*, 73, 5066-79.
- MERCHANT, M., HALE, A., BRUEGGEN, J., HARBSMEIER, C. & ADAMS, C. 2018. Crocodiles Alter Skin Color in Response to Environmental Color Conditions. *Sci Rep*, 8, 6174.
- MI, H., MURUGANUJAN, A. & THOMAS, P. D. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*, 41, D377-86.
- MIER, P. & ANDRADE-NAVARRO, M. A. 2018. Glutamine Codon Usage and polyQ Evolution in Primates Depend on the Q Stretch Length. *Genome biology and evolution*, 10, 816-825.
- MILLS, I. G., GAUGHAN, L., ROBSON, C., ROSS, T., MCCracken, S., KELLY, J. & NEAL, D. E. 2005. Huntingtin interacting protein 1 modulates the transcriptional activity of nuclear hormone receptors. *J Cell Biol*, 170, 191-200.
- MITCHELL, P. J. & TJIAN, R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245, 371-8.
- MIZUTANI, A., WANG, L., RAJAN, H., VIG, P. J., ALAYNICK, W. A., THALER, J. P. & TSAI, C. C. 2005. Boat, an AXH domain protein, suppresses the cytotoxicity of mutant ataxin-1. *EMBO J*, 24, 3339-51.
- MOHAN, R. D., ABMAYR, S. M. & WORKMAN, J. L. 2014a. The expanding role for chromatin and transcription in polyglutamine disease. *Curr Opin Genet Dev*, 26, 96-104.
- MOHAN, R. D., DIALYNAS, G., WEAKE, V. M., LIU, J., MARTIN-BROWN, S., FLORENS, L., WASHBURN, M. P., WORKMAN, J. L. & ABMAYR, S. M. 2014b.

- Loss of *Drosophila* Ataxin-7, a SAGA subunit, reduces H2B ubiquitination and leads to neural and retinal degeneration. *Genes Dev*, 28, 259-72.
- MOREAU, C. S., BELL, C. D., VILA, R., ARCHIBALD, S. B. & PIERCE, N. E. 2006. Phylogeny of the ants: diversification in the age of angiosperms. *Science*, 312, 101-4.
- NAKAMURA, Y., TAGAWA, K., OKA, T., SASABE, T., ITO, H., SHIWAKU, H., LA SPADA, A. R. & OKAZAWA, H. 2012. Ataxin-7 associates with microtubules and stabilizes the cytoskeletal network. *Hum Mol Genet*, 21, 1099-110.
- NALAVADE, R., GRIESCHE, N., RYAN, D. P., HILDEBRAND, S. & KRAUSS, S. 2013. Mechanisms of RNA-induced toxicity in CAG repeat disorders. *Cell Death Dis*, 4, e752.
- NEDELSKY, N. B., PENNUTO, M., SMITH, R. B., PALAZZOLO, I., MOORE, J., NIE, Z., NEALE, G. & TAYLOR, J. P. 2010. Native functions of the androgen receptor are essential to pathogenesis in a *Drosophila* model of spinobulbar muscular atrophy. *Neuron*, 67, 936-52.
- NEUEDER, A., LANGLES, C., GHOSH, R., HOWLAND, D., MYERS, R. H., FAULL, R. L. M., TABRIZI, S. J. & BATES, G. P. 2017. The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. *Sci Rep*, 7, 1307.
- NIKI, T., TAKAHASHI-NIKI, K., TAIRA, T., IGUCHI-ARIGA, S. M. & ARIGA, H. 2003. DJBP: a novel DJ-1-binding protein, negatively regulates the androgen receptor by recruiting histone deacetylase complex, and DJ-1 antagonizes this inhibition by abrogation of this complex. *Mol Cancer Res*, 1, 247-61.
- NOLO, R., ABBOTT, L. A. & BELLEN, H. J. 2000. Senseless, a Zn finger transcription factor, is necessary and sufficient for sensory organ development in *Drosophila*. *Cell*, 102, 349-62.
- NONHOFF, U., RALSER, M., WELZEL, F., PICCINI, I., BALZEREIT, D., YASPO, M. L., LEHRACH, H. & KROBITSCH, S. 2007. Ataxin-2 interacts with the DEAD/H-box RNA helicase DDX6 and interferes with P-bodies and stress granules. *Mol Biol Cell*, 18, 1385-96.
- OKAZAWA, H., RICH, T., CHANG, A., LIN, X., WARAGAI, M., KAJIKAWA, M., ENOKIDO, Y., KOMURO, A., KATO, S., SHIBATA, M., HATANAKA, H., MOURADIAN, M. M., SUDOL, M. & KANAZAWA, I. 2002. Interaction between mutant ataxin-1 and PQBP-1 affects transcription and cell death. *Neuron*, 34, 701-13.
- ORR, H. T. 2012. SCA1-phosphorylation, a regulator of Ataxin-1 function and pathogenesis. *Prog Neurobiol*, 99, 179-85.
- ORR, H. T., CHUNG, M. Y., BANFI, S., KWIATKOWSKI, T. J., JR., SERVADIO, A., BEAUDET, A. L., MCCALL, A. E., DUVICK, L. A., RANUM, L. P. & ZOGHBI, H. Y. 1993. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat Genet*, 4, 221-6.
- OSTROWSKI, L. A., HALL, A. C. & MEKHAIL, K. 2017. Ataxin-2: From RNA Control to Human Health and Disease. *Genes (Basel)*, 8.
- PANDEY, N., MITTAL, U., SRIVASTAVA, A. K. & MUKERJI, M. 2004. SMARCA2 and THAP11: potential candidates for polyglutamine disorders as evidenced from polymorphism and protein-folding simulation studies. *J Hum Genet*, 49, 596-602.
- PARK, J., AL-RAMAH, I., TAN, Q., MOLLEMA, N., DIAZ-GARCIA, J. R., GALLEGOFLORES, T., LU, H. C., LAGALWAR, S., DUVICK, L., KANG, H., LEE, Y., JAFAR-NEJAD, P., SAYEGH, L. S., RICHMAN, R., LIU, X., GAO, Y., SHAW, C. A., ARTHUR, J. S. C., ORR, H. T., WESTBROOK, T. F., BOTAS, J. & ZOGHBI, H. Y. 2013. RAS-MAPK-MSK1 pathway modulates ataxin 1 protein levels and toxicity in SCA1. *Nature*, 498, 325-331.

- PARKER, J. B., PALCHAUDHURI, S., YIN, H., WEI, J. & CHAKRAVARTI, D. 2012. A transcriptional regulatory role of the THAP11-HCF-1 complex in colon cancer cell function. *Mol Cell Biol*, 32, 1654-70.
- PAULSON, H. L., BONINI, N. M. & ROTH, K. A. 2000. Polyglutamine disease and neuronal cell death. *Proc Natl Acad Sci U S A*, 97, 12957-8.
- PAULSON, H. L., SHAKKOTTAI, V. G., CLARK, H. B. & ORR, H. T. 2017. Polyglutamine spinocerebellar ataxias - from genes to potential treatments. *Nat Rev Neurosci*, 18, 613-626.
- PAYTUBI, S., WANG, X., LAM, Y. W., IZQUIERDO, L., HUNTER, M. J., JAN, E., HUNDAL, H. S. & PROUD, C. G. 2009. ABC50 promotes translation initiation in mammalian cells. *J Biol Chem*, 284, 24061-73.
- PELASSA, I. & FIUMARA, F. 2015. Differential Occurrence of Interactions and Interaction Domains in Proteins Containing Homopolymeric Amino Acid Repeats. *Front Genet*, 6, 345.
- PERELMAN, P., JOHNSON, W. E., ROOS, C., SEUANEZ, H. N., HORVATH, J. E., MOREIRA, M. A., KESSING, B., PONTIUS, J., ROELKE, M., RUMPLER, Y., SCHNEIDER, M. P., SILVA, A., O'BRIEN, S. J. & PECON-SLATTERY, J. 2011. A molecular phylogeny of living primates. *PLoS Genet*, 7, e1001342.
- PETRAKIS, S., SCHAEFER, M. H., WANKER, E. E. & ANDRADE-NAVARRO, M. A. 2013. Aggregation of polyQ-extended proteins is promoted by interaction with their natural coiled-coil partners. *Bioessays*, 35, 503-7.
- PETRYSZAK, R., KEAYS, M., TANG, Y. A., FONSECA, N. A., BARRERA, E., BURDETT, T., FULLGRABE, A., FUENTES, A. M., JUPP, S., KOSKINEN, S., MANNION, O., HUERTA, L., MEGY, K., SNOW, C., WILLIAMS, E., BARZINE, M., HASTINGS, E., WEISSER, H., WRIGHT, J., JAISWAL, P., HUBER, W., CHOUDHARY, J., PARKINSON, H. E. & BRAZMA, A. 2016. Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res*, 44, D746-52.
- PICCIONI, F., SIMEONI, S., ANDRIOLA, I., ARMATURA, E., BASSANINI, S., POZZI, P. & POLETTI, A. 2001. Polyglutamine tract expansion of the androgen receptor in a motoneuronal model of spinal and bulbar muscular atrophy. *Brain Res Bull*, 56, 215-20.
- PRUM, R. O., BERV, J. S., DORNBURG, A., FIELD, D. J., TOWNSEND, J. P., LEMMON, E. M. & LEMMON, A. R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, 526, 569-73.
- QADRI, I., IWAHASHI, M. & SIMON, F. 2002. Hepatitis C virus NS5A protein binds TBP and p53, inhibiting their DNA binding and p53 interactions with TBP and ERCC3. *Biochim Biophys Acta*, 1592, 193-204.
- RAMAZZOTTI, M., MONSELLIER, E., KAMOUN, C., DEGL'INNOCENTI, D. & MELKI, R. 2012. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. *PLoS One*, 7, e30824.
- RAPOSO, M., BETTENCOURT, C., RAMOS, A., KAZACHKOVA, N., VASCONCELOS, J., KAY, T., BRUGES-ARMAS, J. & LIMA, M. 2017. Promoter Variation and Expression Levels of Inflammatory Genes IL1A, IL1B, IL6 and TNF in Blood of Spinocerebellar Ataxia Type 3 (SCA3) Patients. *Neuromolecular Med*, 19, 41-45.
- RATOVITSKI, T., CHIGHLADZE, E., ARBEZ, N., BORONINA, T., HERBRICH, S., COLE, R. N. & ROSS, C. A. 2012. Huntingtin protein interactions altered by polyglutamine expansion as determined by quantitative proteomic analysis. *Cell Cycle*, 11, 2006-2021.
- RAY, M. & LAKHOTIA, S. C. 2015. The commonly used eye-specific sev-GAL4 and GMR-GAL4 drivers in *Drosophila melanogaster* are expressed in tissues other than eyes also. *J Genet*, 94, 407-16.

- REEDER, T. W., TOWNSEND, T. M., MULCAHY, D. G., NOONAN, B. P., WOOD, P. L., JR., SITES, J. W., JR. & WIENS, J. J. 2015. Integrated analyses resolve conflicts over squamate reptile phylogeny and reveal unexpected placements for fossil taxa. *PLoS One*, 10, e0118199.
- REIS, M., SILVA, A. C., VIEIRA, C. P. & VIEIRA, J. 2016. The *Drosophila melanogaster* Muc68E Mucin Gene Influences Adult Size, Starvation Tolerance, and Cold Recovery. *G3 (Bethesda)*, 6, 1841-51.
- RODDA, S., SHARMA, S., SCHERER, M., CHAPMAN, G. & RATHJEN, P. 2001. CRTR-1, a developmentally regulated transcriptional repressor related to the CP2 family of transcription factors. *J Biol Chem*, 276, 3324-32.
- ROUSSIGNE, M., KOSSIDA, S., LAVIGNE, A. C., CLOUAIRE, T., ECOCHARD, V., GLORIES, A., AMALRIC, F. & GIRARD, J. P. 2003. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem Sci*, 28, 66-9.
- SANCHEZ-PULIDO, L., PIDOUX, A. L., PONTING, C. P. & ALLSHIRE, R. C. 2009. Common ancestry of the CENP-A chaperones Scm3 and HJURP. *Cell*, 137, 1173-4.
- SANCIANGCO, M. D., CARPENTER, K. E. & BETANCUR, R. R. 2016. Phylogenetic placement of enigmatic percomorph families (Teleostei: Percomorphaceae). *Mol Phylogenet Evol*, 94, 565-576.
- SATTERFIELD, T. F. & PALLANCK, L. J. 2006. Ataxin-2 and its *Drosophila* homolog, ATX2, physically assemble with polyribosomes. *Hum Mol Genet*, 15, 2523-32.
- SCHAEFER, M. H., WANKER, E. E. & ANDRADE-NAVARRO, M. A. 2012. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. *Nucleic Acids Research*, 40, 4273-4287.
- SCHAFFAR, G., BREUER, P., BOTEVA, R., BEHREND, C., TZVETKOV, N., STRIPPEL, N., SAKAHIRA, H., SIEGERS, K., HAYER-HARTL, M. & HARTL, F. U. 2004. Cellular toxicity of polyglutamine expansion proteins: mechanism of transcription factor deactivation. *Mol Cell*, 15, 95-105.
- SEO, W. Y., JEONG, B. C., YU, E. J., KIM, H. J., KIM, S. H., LIM, J. E., KWON, G. Y., LEE, H. M. & KIM, J. H. 2013. CCAR1 promotes chromatin loading of androgen receptor (AR) transcription complex by stabilizing the association between AR and GATA2. *Nucleic Acids Res*, 41, 8526-36.
- SERRA, H. G., DUVICK, L., ZU, T., CARLSON, K., STEVENS, S., JORGENSEN, N., LYSCHOLM, A., BURRIGHT, E., ZOGHBI, H. Y., CLARK, H. B., ANDRESEN, J. M. & ORR, H. T. 2006. RORalpha-mediated Purkinje cell development determines disease severity in adult SCA1 mice. *Cell*, 127, 697-708.
- SHEN, Y., LEE, G., CHOE, Y., ZOLTEWICZ, J. S. & PETERSON, A. S. 2007. Functional Architecture of Atrophins. *Journal of Biological Chemistry*, 282, 5037-5044.
- SHI, J., WU, W. J., HU, G., YU, X., YU, G. S., LU, H., YANG, M. L., LIU, B. & WU, Z. X. 2016. Regulation of beta-catenin transcription activity by leupaxin in hepatocellular carcinoma. *Tumour Biol*, 37, 2313-20.
- SHIBATA, H., HUYNH, D. P. & PULST, S. M. 2000. A novel protein with RNA-binding motifs interacts with ataxin-2. *Hum Mol Genet*, 9, 1303-13.
- SHIMIZU, K., IJIMA, M., SETIAMARGA, D. H., SARASHINA, I., KUDOH, T., ASAMI, T., GITTEBERGER, E. & ENDO, K. 2013. Left-right asymmetric expression of dpp in the mantle of gastropods correlates with asymmetric shell coiling. *Evodevo*, 4, 15.
- SILVA, A., DE ALMEIDA, A. V. & MACEDO-RIBEIRO, S. 2018. Polyglutamine expansion diseases: More than simple repeats. *J Struct Biol*, 201, 139-154.
- STEINACKER, P., AITKEN, A. & OTTO, M. 2011. 14-3-3 proteins in neurodegeneration. *Semin Cell Dev Biol*, 22, 696-704.

- SUTER, B., FONTAINE, J. F., YILDIRIMMAN, R., RASKO, T., SCHAEFER, M. H., RASCHE, A., PORRAS, P., VAZQUEZ-ALVAREZ, B. M., RUSS, J., RAU, K., FOULLE, R., ZENKNER, M., SAAR, K., HERWIG, R., ANDRADE-NAVARRO, M. A. & WANKER, E. E. 2013. Development and application of a DNA microarray-based yeast two-hybrid system. *Nucleic Acids Res*, 41, 1496-507.
- SUZUKI, Y. & YAZAWA, I. 2011. Pathological accumulation of atrophin-1 in dentatorubralpallidoluysian atrophy. *Int J Clin Exp Pathol*, 4, 378-84.
- TAKESHITA, Y., FUJINAGA, R., KOKUBU, K., ISLAM, M. N., JAHAN, M. R., YANAI, A., KAKIZUKA, A. & SHINODA, K. 2011. Interaction of ataxin-3 with huntingtin-associated protein 1 through Josephin domain. *Neuroreport*, 22, 232-8.
- TAKEUCHI, T. & NAGAI, Y. 2017. Protein Misfolding and Aggregation as a Therapeutic Target for Polyglutamine Diseases. *Brain Sci*, 7.
- TANNER, T. M., DENAYER, S., GEVERTS, B., VAN TILBORGH, N., KERKHOF, S., HELSEN, C., SPANS, L., DUBOIS, V., HOUTSMULLER, A. B., CLAESSENS, F. & HAELENS, A. 2010. A 629RKLKK633 motif in the hinge region controls the androgen receptor at multiple levels. *Cell Mol Life Sci*, 67, 1919-27.
- TARACHA, A., KOTARBA, G. & WILANOWSKI, T. 2018. Neglected Functions of TFCEP2/TFCEP2L1/UBP1 Transcription Factors May Offer Valuable Insights into Their Mechanisms of Action. *Int J Mol Sci*, 19.
- TELFORD, M. J., BUDD, G. E. & PHILIPPE, H. 2015. Phylogenomic Insights into Animal Evolution. *Curr Biol*, 25, R876-87.
- TETEL, M. J. & ACHARYA, K. D. 2013. Nuclear receptor coactivators: regulators of steroid action in brain and behaviour. *J Neuroendocrinol*, 25, 1209-18.
- THURMOND, J., GOODMAN, J. L., STRELETS, V. B., ATTRILL, H., GRAMATES, L. S., MARYGOLD, S. J., MATTHEWS, B. B., MILLBURN, G., ANTONAZZO, G., TROVISCO, V., KAUFMAN, T. C., CALVI, B. R. & FLYBASE, C. 2019. FlyBase 2.0: the next generation. *Nucleic Acids Res*, 47, D759-D765.
- TOTZECK, F., ANDRADE-NAVARRO, M. A. & MIER, P. 2017. The Protein Structure Context of PolyQ Regions. *PLoS One*, 12, e0170801.
- TSAGALIOTI, E., TRIFONOS, C., MORARI, A., VADIKOLIAS, K. & GIAGINIS, C. 2018. Clinical value of nutritional status in neurodegenerative diseases: What is its impact and how it affects disease progression and management? *Nutr Neurosci*, 21, 162-175.
- TSAI, C. C., KAO, H. Y., MITZUTANI, A., BANAYO, E., RAJAN, H., MCKEOWN, M. & EVANS, R. M. 2004. Ataxin 1, a SCA1 neurodegenerative disorder protein, is functionally linked to the silencing mediator of retinoid and thyroid hormone receptors. *Proc Natl Acad Sci U S A*, 101, 4047-52.
- TSUDA, H., JAFAR-NEJAD, H., PATEL, A. J., SUN, Y., CHEN, H. K., ROSE, M. F., VENKEN, K. J., BOTAS, J., ORR, H. T., BELLEN, H. J. & ZOGHBI, H. Y. 2005. The AXH domain of Ataxin-1 mediates neurodegeneration through its interaction with Gfi-1/Senseless proteins. *Cell*, 122, 633-44.
- TU, Y., LIU, H., ZHU, X., SHEN, H., MA, X., WANG, F., HUANG, M., GONG, J., LI, X., WANG, Y., GUO, C. & TANG, T. S. 2017. Ataxin-3 promotes genome integrity by stabilizing Chk1. *Nucleic Acids Res*, 45, 4532-4549.
- UHLEN, M., FAGERBERG, L., HALLSTROM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, A., KAMPF, C., SJOSTEDT, E., ASPLUND, A., OLSSON, I., EDLUND, K., LUNDBERG, E., NAVANI, S., SZIGYARTO, C. A., ODEBERG, J., DJUREINOVIC, D., TAKANEN, J. O., HOBER, S., ALM, T., EDQVIST, P. H., BERLING, H., TEGEL, H., MULDER, J., ROCKBERG, J., NILSSON, P., SCHWENK, J. M., HAMSTEN, M., VON FEILITZEN, K., FORSBERG, M., PERSSON, L., JOHANSSON, F., ZWAHLN, M., VON

- HEIJNE, G., NIELSEN, J. & PONTEN, F. 2015. Proteomics. Tissue-based map of the human proteome. *Science*, 347, 1260419.
- UHLIK, M. T., TEMPLE, B., BENCHARIT, S., KIMPLE, A. J., SIDEROVSKI, D. P. & JOHNSON, G. L. 2005. Structural and evolutionary division of phosphotyrosine binding (PTB) domains. *J Mol Biol*, 345, 1-20.
- UNIPROT, C. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47, D506-D515.
- VAN ROON-MOM, W. M., REID, S. J., FAULL, R. L. & SNELL, R. G. 2005. TATA-binding protein in neurodegenerative disease. *Neuroscience*, 133, 863-72.
- VAN ZUNDELT, G. C. P., RODRIGUES, J., TRELLET, M., SCHMITZ, C., KASTRITIS, P. L., KARACA, E., MELQUIOND, A. S. J., VAN DIJK, M., DE VRIES, S. J. & BONVIN, A. 2016. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol*, 428, 720-725.
- VÁZQUEZ, N., ROCHA, S., LÓPEZ-FERNÁNDEZ, H., TORRES, A., CAMACHO, R., FDEZ-RIVEROLA, F., VIEIRA, J., VIEIRA, C. P. & REBOIRO-JATO, M. EvoPPI: A Web Application to Compare Protein-Protein Interactions (PPIs) from Different Databases and Species. 2019 Cham. Springer International Publishing, 149-156.
- WALTERS, R. H. & MURPHY, R. M. 2009. Examining polyglutamine peptide length: a connection between collapsed conformations and increased aggregation. *J Mol Biol*, 393, 978-92.
- WANG, L., DI, L. J., LV, X., ZHENG, W., XUE, Z., GUO, Z. C., LIU, D. P. & LIANG, C. C. 2009. Inter-MAR association contributes to transcriptionally active looping events in human beta-globin gene cluster. *PLoS One*, 4, e4629.
- WANG, L. & TSAI, C. C. 2008. Atrophin proteins: an overview of a new class of nuclear receptor corepressors. *Nucl Recept Signal*, 6, e009.
- WANG, Y. L., FAIOLA, F., XU, M., PAN, S. & MARTINEZ, E. 2008. Human ATAC Is a GCN5/PCAF-containing acetylase complex with a novel NC2-like histone fold module that interacts with the TATA-binding protein. *J Biol Chem*, 283, 33808-15.
- WELLINGTON, C. L., ELLERBY, L. M., HACKAM, A. S., MARGOLIS, R. L., TRIFIRO, M. A., SINGARAJA, R., MCCUTCHEON, K., SALVESEN, G. S., PROPP, S. S., BROMM, M., ROWLAND, K. J., ZHANG, T., RASPER, D., ROY, S., THORNBERRY, N., PINSKY, L., KAKIZUKA, A., ROSS, C. A., NICHOLSON, D. W., BREDESEN, D. E. & HAYDEN, M. R. 1998. Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J Biol Chem*, 273, 9158-67.
- WHAN, V., HOBBS, M., MCWILLIAM, S., LYNN, D. J., LUTZOW, Y. S., KHATKAR, M., BARENDSE, W., RAADSMA, H. & TELLAM, R. L. 2010. Bovine proteins containing poly-glutamine repeats are often polymorphic and enriched for components of transcriptional regulatory complexes. *BMC Genomics*, 11, 654.
- WOOD, J. D., NUCIFORA, F. C., JR., DUAN, K., ZHANG, C., WANG, J., KIM, Y., SCHILLING, G., SACCHI, N., LIU, J. M. & ROSS, C. A. 2000. Atrophin-1, the dentato-rubral and pallido-lusian atrophy gene product, interacts with ETO/MTG8 in the nuclear matrix and represses transcription. *J Cell Biol*, 150, 939-48.
- WOOD, J. D., YUAN, J., MARGOLIS, R. L., COLOMER, V., DUAN, K., KUSHI, J., KAMINSKY, Z., KLEIDERLEIN, J. J., SHARP, A. H. & ROSS, C. A. 1998. Atrophin-1, the DRPLA gene product, interacts with two families of WW domain-containing proteins. *Mol Cell Neurosci*, 11, 149-60.
- XIANG, C., ZHANG, S., DONG, X., MA, S. & CONG, S. 2018. Transcriptional Dysregulation and Post-translational Modifications in Polyglutamine Diseases:

- From Pathogenesis to Potential Therapeutic Strategies. *Front Mol Neurosci*, 11, 153.
- XU, Z., TITO, A. J., RUI, Y. N. & ZHANG, S. 2015. Studying polyglutamine diseases in *Drosophila*. *Exp Neurol*, 274, 25-41.
- YAMADA, M. 2010. Dentatorubral-pallidoluysian atrophy (DRPLA): The 50th Anniversary of Japanese Society of Neuropathology. *Neuropathology : official journal of the Japanese Society of Neuropathology*, 30, 453-7.
- YAMADA, M., WOOD, J. D., SHIMOHATA, T., HAYASHI, S., TSUJI, S., ROSS, C. A. & TAKAHASHI, H. 2001. Widespread occurrence of intranuclear atrophin-1 accumulation in the central nervous system neurons of patients with dentatorubral-pallidoluysian atrophy. *Ann Neurol*, 49, 14-23.
- YANG, J., YAN, R., ROY, A., XU, D., POISSON, J. & ZHANG, Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nat Methods*, 12, 7-8.
- YANG, W., DUNLAP, J. R., ANDREWS, R. B. & WETZEL, R. 2002. Aggregated polyglutamine peptides delivered to nuclei are toxic to mammalian cells. *Hum Mol Genet*, 11, 2905-17.
- YANG, Y., YANG, S., GUO, J., CUI, Y., TANG, B., LI, X. J. & LI, S. 2017. Synergistic Toxicity of Polyglutamine-Expanded TATA-Binding Protein in Glia and Neuronal Cells: Therapeutic Implications for Spinocerebellar Ataxia 17. *J Neurosci*, 37, 9101-9115.
- YUE, S., SERRA, H. G., ZOGHBI, H. Y. & ORR, H. T. 2001. The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract. *Hum Mol Genet*, 10, 25-30.
- YURI, T., KIMBALL, R. T., HARSHMAN, J., BOWIE, R. C., BRAUN, M. J., CHOJNOWSKI, J. L., HAN, K. L., HACKETT, S. J., HUDDLESTON, C. J., MOORE, W. S., REDDY, S., SHELDON, F. H., STEADMAN, D. W., WITT, C. C. & BRAUN, E. L. 2013. Parsimony and model-based analyses of indels in avian nuclear genes reveal congruent and incongruent phylogenetic signals. *Biology (Basel)*, 2, 419-44.
- ZHANG, A., ZHAO, J. C., KIM, J., FONG, K. W., YANG, Y. A., CHAKRAVARTI, D., MO, Y. Y. & YU, J. 2015. LncRNA HOTAIR Enhances the Androgen-Receptor-Mediated Transcriptional Program and Drives Castration-Resistant Prostate Cancer. *Cell Rep*, 13, 209-221.
- ZHANG, C. L., ZOU, Y., YU, R. T., GAGE, F. H. & EVANS, R. M. 2006. Nuclear receptor TLX prevents retinal dystrophy and recruits the corepressor atrophin1. *Genes Dev*, 20, 1308-20.
- ZHANG, H., GAO, G., ZHANG, Y., SUN, Y., LI, H., DONG, S., MA, W., LIU, B., WANG, W., WU, H. & ZHANG, H. 2017. Glucose Deficiency Elevates Acid-Sensing Ion Channel 2a Expression and Increases Seizure Susceptibility in Temporal Lobe Epilepsy. *Sci Rep*, 7, 5870.
- ZHANG, S., WILLIAMSON, N. A. & BOGOYEVITCH, M. A. 2018a. Complementary proteomics strategies capture an ataxin-1 interactome in Neuro-2a cells. *Sci Data*, 5, 180262.
- ZHANG, S. Q., CHE, L. H., LI, Y., DAN, L., PANG, H., SLIPINSKI, A. & ZHANG, P. 2018b. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nat Commun*, 9, 205.
- ZHANG, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9, 40.
- ZHAO, Y., LANG, G., ITO, S., BONNET, J., METZGER, E., SAWATSUBASHI, S., SUZUKI, E., LE GUEZENNEC, X., STUNNENBERG, H. G., KRASNOV, A., GEORGIEVA, S. G., SCHULE, R., TAKEYAMA, K., KATO, S., TORA, L. & DEVYS, D. 2008. A TFTC/STAGA module mediates histone H2A and H2B

- deubiquitination, coactivates nuclear receptors, and counteracts heterochromatin silencing. *Mol Cell*, 29, 92-101.
- ZHOU, H., YOSHIOKA, T. & NATHANS, J. 1996. Retina-derived POU-domain factor-1: a complex POU-domain gene implicated in the development of retinal ganglion and amacrine cells. *J Neurosci*, 16, 2261-74.
- ZHOU, L., MA, Q., SHI, H. & HUO, K. 2010. NUMBL interacts with TRAF6 and promotes the degradation of TRAF6. *Biochem Biophys Res Commun*, 392, 409-14.
- ZHOU, L., WANG, H., WANG, P., REN, H., CHEN, D., YING, Z. & WANG, G. 2013. Ataxin-3 protects cells against H₂O₂-induced oxidative stress by enhancing the interaction between Bcl-X(L) and Bax. *Neuroscience*, 243, 14-21.
- ZOGHBI, H. Y. & ORR, H. T. 1999. Polyglutamine diseases: protein cleavage and aggregation. *Curr Opin Neurobiol*, 9, 566-70.
- ZOGHBI, H. Y. & ORR, H. T. 2000. Glutamine repeats and neurodegeneration. *Annu Rev Neurosci*, 23, 217-47.
- ZOGHBI, H. Y. & ORR, H. T. 2009. Pathogenic mechanisms of a polyglutamine-mediated neurodegenerative disease, spinocerebellar ataxia type 1. *J Biol Chem*, 284, 7425-9.

Supplementary

Supplementary Table 1. Uniprot and gene ID, gene name(s), and protein size for ATXN1 and ATXN1L in the species used.

Protein	Model Organism	UniProt	GeneID	Gene name(s)	Protein size (aa)
ATXN1	<i>H. sapiens</i>	P54253	6310	<i>ATXN1, ATX1, SCA1</i>	815
	<i>M. musculus</i>	P54254	20238	<i>Atxn1, Sca1</i>	791
	<i>X. tropicalis</i>	F6PW06	100492711	<i>atxn1</i>	775
	<i>D. rerio</i> (Ataxin 1a)	Q1L966	557340	<i>atxn1a, atxn1b</i>	781
	<i>D. rerio</i> (Ataxin 1b)	E7F6Z5	565841	<i>atxn1b</i>	827
	<i>D. melanogaster</i>	Q9W3V7	31624	<i>Atx-1, CG454RA, dAtx-1, CG4547,</i>	230
	<i>C.elegans</i>	O44771	187002	<i>CELE_K04F10.1, K04F10.1</i>	299
ATXN1-like	<i>H. sapiens</i>	P0C7T5	342371	<i>ATXN1L, BOAT, BOAT1</i>	689
	<i>M. musculus</i>	P0C7T6	52335	<i>Atxn1l, Boat</i>	687
	<i>X. laevis</i>	B1WBC1	100158423	<i>atxn1l, LOC100158423, XELAEV_18024898mg</i>	691
	<i>D. rerio</i>	E7FB74	793927	<i>atxn1l</i>	765

Supplementary Table 2. List of PDB files describing the interaction of wt (ATXN1Q29) and expanded (ATXN1Q82) ATXN1 with 71 interactors. Files can be downloaded from Zenodo DOI: 10.5281/zenodo.3416591

wt ATXN1- interactor	Zenodo's download link	expanded ATXN1- interactor	Zenodo's download link
ATXN1Q29- ADD3.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ADD3.pdb?download=1	ATXN1Q82- ADD3.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ADD3.pdb?download=1
ATXN1Q29- ARID5A.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ARID5A.pdb?download=1	ATXN1Q82- ARID5A.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ARID5A.pdb?download=1
ATXN1Q29- ASNS.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ASNS.pdb?download=1	ATXN1Q82- ASNS.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ASNS.pdb?download=1
ATXN1Q29- BAALC.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-BAALC.pdb?download=1	ATXN1Q82- BAALC.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-BAALC.pdb?download=1
ATXN1Q29- BASP1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-BASP1.pdb?download=1	ATXN1Q82- BASP1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-BASP1.pdb?download=1
ATXN1Q29- C16orf5.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-C16orf5.pdb?download=1	ATXN1Q82- C16orf5.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-C16orf5.pdb?download=1
ATXN1Q29- C2orf27B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-C2orf27B.pdb?download=1	ATXN1Q82- C2orf27B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-C2orf27B.pdb?download=1
ATXN1Q29- CAMK2B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CAMK2B.pdb?download=1	ATXN1Q82- CAMK2B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CAMK2B.pdb?download=1
ATXN1Q29- CHRNA7.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CHRNA7.pdb?download=1	ATXN1Q82- CHRNA7.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CHRNA7.pdb?download=1
ATXN1Q29- CREM.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CREM.pdb?download=1	ATXN1Q82- CREM.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CREM.pdb?download=1
ATXN1Q29- CRIP2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CRIP2.pdb?download=1	ATXN1Q82- CRIP2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CRIP2.pdb?download=1
ATXN1Q29- CRK.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CRK.pdb?download=1	ATXN1Q82- CRK.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CRK.pdb?download=1
ATXN1Q29- CRY2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CRY2.pdb?download=1	ATXN1Q82- CRY2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CRY2.pdb?download=1
ATXN1Q29- CXorf27.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-CXorf27.pdb?download=1	ATXN1Q82- CXorf27.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-CXorf27.pdb?download=1

ATXN1Q29-DHRX.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-DHRX.pdb?download=1	ATXN1Q82-DHRX.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-DHRX.pdb?download=1
ATXN1Q29-DHX37.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-DHX37.pdb?download=1	ATXN1Q82-DHX37.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-DHX37.pdb?download=1
ATXN1Q29-DIXDC1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-DIXDC1.pdb?download=1	ATXN1Q82-DIXDC1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-DIXDC1.pdb?download=1
ATXN1Q29-EIF1B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-EIF1B.pdb?download=1	ATXN1Q82-EIF1B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-EIF1B.pdb?download=1
ATXN1Q29-EIF3F.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-EIF3F.pdb?download=1	ATXN1Q82-EIF3F.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-EIF3F.pdb?download=1
ATXN1Q29-ESRRA.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ESRRA.pdb?download=1	ATXN1Q82-ESRRA.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ESRRA.pdb?download=1
ATXN1Q29-ETV4.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ETV4.pdb?download=1	ATXN1Q82-ETV4.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ETV4.pdb?download=1
ATXN1Q29-FAM46A.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-FAM46A.pdb?download=1	ATXN1Q82-FAM46A.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-FAM46A.pdb?download=1
ATXN1Q29-FAM46B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-FAM46B.pdb?download=1	ATXN1Q82-FAM46B.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-FAM46B.pdb?download=1
ATXN1Q29-FAR1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-FAR1.pdb?download=1	ATXN1Q82-FAR1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-FAR1.pdb?download=1
ATXN1Q29-FOSL1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-FOSL1.pdb?download=1	ATXN1Q82-FOSL1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-FOSL1.pdb?download=1
ATXN1Q29-GATAD1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-GATAD1.pdb?download=1	ATXN1Q82-GATAD1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-GATAD1.pdb?download=1
ATXN1Q29-GGA2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-GGA2.pdb?download=1	ATXN1Q82-GGA2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-GGA2.pdb?download=1
ATXN1Q29-HEY2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-HEY2.pdb?download=1	ATXN1Q82-HEY2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-HEY2.pdb?download=1
ATXN1Q29-HEYL.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-HEYL.pdb?download=1	ATXN1Q82-HEYL.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-HEYL.pdb?download=1
ATXN1Q29-HNRPLL.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-HNRPLL.pdb?download=1	ATXN1Q82-HNRPLL.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-HNRPLL.pdb?download=1

ATXN1Q29-ILVBL.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ILVBL.pdb?download=1	ATXN1Q82-ILVBL.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ILVBL.pdb?download=1
ATXN1Q29-IMMT.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-IMMT.pdb?download=1	ATXN1Q82-IMMT.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-IMMT.pdb?download=1
ATXN1Q29-KCTD15.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-KCTD15.pdb?download=1	ATXN1Q82-KCTD15.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-KCTD15.pdb?download=1
ATXN1Q29-KIF22.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-KIF22.pdb?download=1	ATXN1Q82-KIF22.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-KIF22.pdb?download=1
ATXN1Q29-LASP1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-LASP1.pdb?download=1	ATXN1Q82-LASP1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-LASP1.pdb?download=1
ATXN1Q29-LITAF.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-LITAF.pdb?download=1	ATXN1Q82-LITAF.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-LITAF.pdb?download=1
ATXN1Q29-LPAR2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-LPAR2.pdb?download=1	ATXN1Q82-LPAR2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-LPAR2.pdb?download=1
ATXN1Q29-MAGEB18.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-MAGEB18.pdb?download=1	ATXN1Q82-MAGEB18.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-MAGEB18.pdb?download=1
ATXN1Q29-MAGEB2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-MAGEB2.pdb?download=1	ATXN1Q82-MAGEB2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-MAGEB2.pdb?download=1
ATXN1Q29-MAGEB6.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-MAGEB6.pdb?download=1	ATXN1Q82-MAGEB6.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-MAGEB6.pdb?download=1
ATXN1Q29-MCART1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-MCART1.pdb?download=1	ATXN1Q82-MCART1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-MCART1.pdb?download=1
ATXN1Q29-MLST8.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-MLST8.pdb?download=1	ATXN1Q82-MLST8.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-MLST8.pdb?download=1
ATXN1Q29-MSX2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-MSX2.pdb?download=1	ATXN1Q82-MSX2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-MSX2.pdb?download=1
ATXN1Q29-NCAM1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-NCAM1.pdb?download=1	ATXN1Q82-NCAM1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-NCAM1.pdb?download=1
ATXN1Q29-OTX2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-OTX2.pdb?download=1	ATXN1Q82-OTX2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-OTX2.pdb?download=1
ATXN1Q29-PIAS1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-PIAS1.pdb?download=1	ATXN1Q82-PIAS1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-PIAS1.pdb?download=1

ATXN1Q29- PLEKHB1.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
PLEKHB1.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- PLEKHB1.pdb?download=1	ATXN1Q82- PLEKHB1.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
PLEKHB1.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- PLEKHB1.pdb?download=1
ATXN1Q29- PPAT.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
PPAT.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- PPAT.pdb?download=1	ATXN1Q82- PPAT.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
PPAT.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- PPAT.pdb?download=1
ATXN1Q29- PSPH.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
PSPH.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- PSPH.pdb?download=1	ATXN1Q82- PSPH.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
PSPH.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- PSPH.pdb?download=1
ATXN1Q29- QKI.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
QKI.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- QKI.pdb?download=1	ATXN1Q82- QKI.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
QKI.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- QKI.pdb?download=1
ATXN1Q29- RAI2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
RAI2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- RAI2.pdb?download=1	ATXN1Q82- RAI2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
RAI2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- RAI2.pdb?download=1
ATXN1Q29- RAPGEF1.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
RAPGEF1.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- RAPGEF1.pdb?download=1	ATXN1Q82- RAPGEF1.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
RAPGEF1.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- RAPGEF1.pdb?download=1
ATXN1Q29- RBFOX2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
RBFOX2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- RBFOX2.pdb?download=1	ATXN1Q82- RBFOX2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
RBFOX2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- RBFOX2.pdb?download=1
ATXN1Q29- RBM26.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
RBM26.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- RBM26.pdb?download=1	ATXN1Q82- RBM26.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
RBM26.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- RBM26.pdb?download=1
ATXN1Q29- SEMA4G.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
SEMA4G.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- SEMA4G.pdb?download=1	ATXN1Q82- SEMA4G.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
SEMA4G.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- SEMA4G.pdb?download=1
ATXN1Q29- SF1.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
SF1.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- SF1.pdb?download=1	ATXN1Q82- SF1.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
SF1.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- SF1.pdb?download=1
ATXN1Q29- SLC6A13.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
SLC6A13.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- SLC6A13.pdb?download=1	ATXN1Q82- SLC6A13.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
SLC6A13.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- SLC6A13.pdb?download=1
ATXN1Q29- STAM2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
STAM2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- STAM2.pdb?download=1	ATXN1Q82- STAM2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
STAM2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- STAM2.pdb?download=1
ATXN1Q29- SV2A.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
SV2A.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- SV2A.pdb?download=1	ATXN1Q82- SV2A.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
SV2A.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- SV2A.pdb?download=1
ATXN1Q29- TCTA.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
TCTA.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- TCTA.pdb?download=1	ATXN1Q82- TCTA.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
TCTA.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- TCTA.pdb?download=1
ATXN1Q29- TMX2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
TMX2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- TMX2.pdb?download=1	ATXN1Q82- TMX2.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
TMX2.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- TMX2.pdb?download=1
ATXN1Q29- TOMM20.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q29-
TOMM20.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q29- TOMM20.pdb?download=1	ATXN1Q82- TOMM20.pdb	<a href="https://zenodo.org/record/3416591/files/ATXN1Q82-
TOMM20.pdb?download=1">https://zenodo.org/record/3416591/files/ATXN1Q82- TOMM20.pdb?download=1

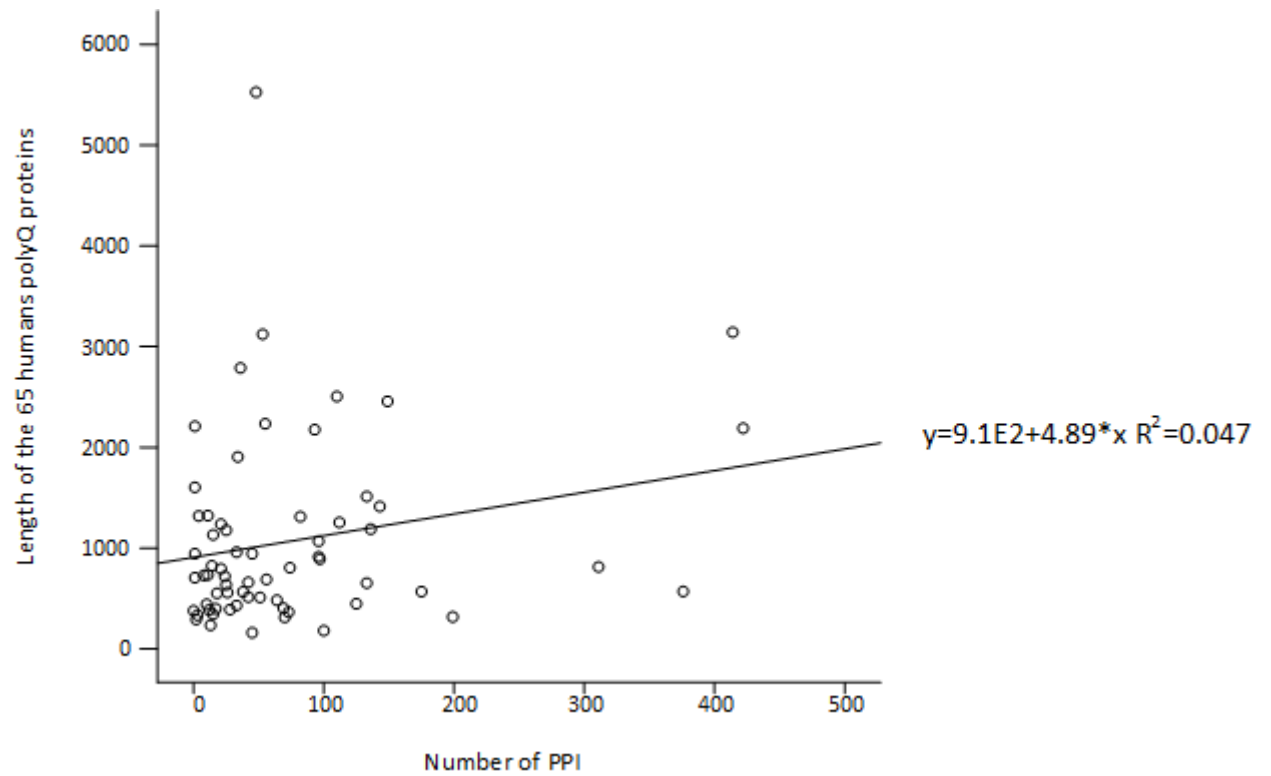
ATXN1Q29-TP53I11.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-TP53I11.pdb?download=1	ATXN1Q82-TP53I11.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-TP53I11.pdb?download=1
ATXN1Q29-TRIM38.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-TRIM38.pdb?download=1	ATXN1Q82-TRIM38.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-TRIM38.pdb?download=1
ATXN1Q29-TSC1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-TSC1.pdb?download=1	ATXN1Q82-TSC1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-TSC1.pdb?download=1
ATXN1Q29-TTRAP.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-TTRAP.pdb?download=1	ATXN1Q82-TTRAP.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-TTRAP.pdb?download=1
ATXN1Q29-UHRF2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-UHRF2.pdb?download=1	ATXN1Q82-UHRF2.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-UHRF2.pdb?download=1
ATXN1Q29-WBSCR16.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-WBSCR16.pdb?download=1	ATXN1Q82-WBSCR16.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-WBSCR16.pdb?download=1
ATXN1Q29-YWHAE.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-YWHAE.pdb?download=1	ATXN1Q82-YWHAE.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-YWHAE.pdb?download=1
ATXN1Q29-ZC3H10.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ZC3H10.pdb?download=1	ATXN1Q82-ZC3H10.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ZC3H10.pdb?download=1
ATXN1Q29-ZSCAN1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q29-ZSCAN1.pdb?download=1	ATXN1Q82-ZSCAN1.pdb	https://zenodo.org/record/3416591/files/ATXN1Q82-ZSCAN1.pdb?download=1

Supplementary Table 3. The involvement of the 65 human proteins with polyQ in disease.

zGene ID	Protein name	UniProtKB	Involvement in disease (according to NCBI-Gene;UniProt;and OMIM)
6310	ATXN1, ATX1, SCA1	P54253	Spinocerebellar ataxia 1 (SCA1)
367	AR, DHTR, NR3C4	P10275	Testicular feminization syndrome (TFM), and in Kennedy spinal and bulbar muscular atrophy (SBMA)
1822	ATN1, D12S755E, DRPLA	P54259	Dentatorubro-pallidoluysian atrophy (DRPLA)
6311	ATXN2, ATX2, SCA2, TNRC13	Q99700	Spinocerebellar ataxia 2 (SCA2); Amyotrophic lateral sclerosis 13 (ALS13); Parkinson disease, late-onset, susceptibility to
4287	ATXN3, ATX3, MJD, MJD1, SCA3	P54252	Machado-Joseph disease (MJD)
6314	ATXN7, SCA7	O15265	Spinocerebellar ataxia 7 (SCA7)
773	CACNA1A, CACH4, CACN3, CACNL1A4	O00555	Epileptic encephalopathy, early infantile; Episodic ataxia, type 2; Migraine, familial hemiplegic, 1; Migraine, familial hemiplegic, 1, with progressive cerebellar ataxia; Spinocerebellar ataxia 6 (SCA6)
3064	HTT, HD, IT15	P42858	Huntington disease (HD); Lopes-Maciel-Rodan syndrome (LOMARS)
6908	TBP, GTF2D1, TF2D, TFIID	P20226	Spinocerebellar ataxia 17 (SCA17); susceptibility to Parkinson disease
22848	AAK1, KIAA1048	Q2M2I8	Not Described
23	ABCF1, ABC50	Q8NE71	Play a role in enhancement of protein synthesis and the inflammation process
57492	ARID1B, BAF250B, DAN15, KIAA1235, OSA2	Q8NFD5	Coffin-Siris syndrome 1
10620	ARID3B, BDP, DRIL2	Q8IWW6	Present in K-562 erythrocytic leukemia cell line
429	ASCL1, ASH1, BHLHA46, HASH1	P50553	Highly expressed in 2 neuroendocrine cancers, medullary thyroid cancer (MTC; 155240) and small cell lung cancer (SCLC). Central hypoventilation syndrome, congenital, Haddad syndrome.
55589	BMP2K, BIKE, HRIHFB2017	Q9NSY1	Not Described
11189	CELF3, BRUNOL1, CAGH4, ERDA4, TNRC4	Q5SZQ8	Not Described
10523	CHERP, DAN26, SCAF6	Q8IWX8	Not Described
1387	CREBBP, CBP	Q92793	Rubinstein-Taybi syndrome 1; acute myeloid leukemia
196513	DCP1B	Q8IZD4	Not Described
9909	DENND4B, KIAA0476	O75064	Not Described
1750	DLX6	P56179	Not Described
57634	EP400, CAGH32, KIAA1498, KIAA1818, TNRC12	Q96L91	Not Described
165215	FAM171B, KIAA1946, NPD019	Q6P995	Not Described

93986	FOXP2, CAGH44, TNRC10	O15409	Speech-language disorder-1
84443	FRMPD3, KIAA1817	Q5JV73	Not Described
2734	GLG1, CFR1, ESL1, MG160	Q92896	Not Described
64207	IRF2BPL, C14orf4, EAP1, KIAA1865, My03	Q9H1B7	Neurodevelopmental disorder with regression, abnormal movements, loss of speech, and seizures
3782 8085	KCNN3, K3 KMT2D, ALR, MLL2, MLL4	Q9UGI6 KMT2D	Possible Association with Schizophrenia Kabuki syndrome 1
8825	LIN7A, MALS1, VELI1	O14910	LIN7A is a major determinant of cell-polarity defects in breast carcinomas
9223	MAGI1, AIP3, BAIAP1, BAP1, TNRC19	Q96QZ7	Not Described
84441	MAML2, KIAA1819	Q8IZL2	Mucoepidermoid salivary gland carcinoma
55534	MAML3, KIAA1816	Q96JK9	Not Described
10046	MAMLD1, CG1, CXorf6	Q13495	Hypospadias 2, X-linked
9968	MED12, ARC240, CAGH45, HOPA, KIAA0192, TNRC11	Q93074	Lujan-Fryns syndrome; Ohdo syndrome, X-linked; Opitz-Kaveggia syndrome
51586	MED15, ARC105, CTG7A, PCQAP, TIG1, TNRC7	Q96RN5	DiGeorge syndrome
4205	MEF2A, MEF2	Q02078	Autosomal dominant coronary artery disease 1 with myocardial infarction (ADCAD1)
4330	MN1	Q10571	Meningioma
8202	NCOA3, AIB1, BHLHE42, RAC3, TRAM1	Q9Y6Q9	Breast and ovarian cancers
23054	NCOA6, AIB3, KIAA0181, RAP250, TRBP	Q14686	Colon, breast and lung cancers.
9612	NCOR2, CTG26	Q9Y618	Aberrant expression of this gene is associated with certain cancers. Alt
10725	NFAT5, KIAA0827, TONEBP	O94916	Invasive human ductal breast carcinomas.
9253	NUMBL	Q9Y6R0	Numbl like regulates proliferation, apoptosis, and invasion of lung cancer cell.
1911	PHC1, EDR1, PH1	P78364	Microcephaly 11, primary, autosomal recessive (MCPH11)
22822	PHLDA1, PHRIP, TDAG51	Q8WV24	Myotonic dystrophy
5428	POLG, MDP1, POLG1, POLGA	P54098	Mitochondrial DNA depletion syndrome 4A (Alpers type); Mitochondrial DNA depletion syndrome 4B (MNGIE type); Mitochondrial recessive ataxia syndrome (includes SANDO and SCAE); Progressive external ophthalmoplegia, autosomal dominant 1; Progressive external ophthalmoplegia, autosomal recessive 1
5454	POU3F2, BRN2, OCT7, OTF7	P20265	Not Described
11281	POU6F2, RPF1	P78424	Wilms tumor susceptibility-5

22864	R3HDM2, KIAA1002	Q9Y2K5	Not Described
10743 860	RAI1, KIAA1820 RUNX2, AML3, CBFA1, OSF2, PEBP2A	Q7Z5J4 Q13950	Smith-Magenis syndrome (SMS) Cleidocranial dysplasia; Cleidocranial dysplasia, forme fruste, dental anomalies only; Cleidocranial dysplasia, forme fruste, with brachydactyly; Metaphyseal dysplasia with maxillary hypoplasia with or without brachydactyly
6304	SATB1	Q01826	Delineates specific epigenetic modifications at target gene loci, directly up-regulating metastasis-associated genes while down-regulating tumor-suppressor genes. Reprograms chromatin organization and the transcription profiles of breast tumors to promote growth and metastasis.
23387	SIK3, KIAA0999, QSK, L19	Q9Y2K2	Somatic mutation in a breast cancer sample; expression elevated in ovarian cancers,
6595	SMARCA2, BAF190B, BRM, SNF2A, SNF2L2	P51531	Nicolaides-Baraitser syndrome
81849	ST6GALNAC5, SIAT7E	Q9BVH7	Not Described
7024	TFCP2, LSF, SEF	Q12800	May be involved in the pathogenesis of Alzheimer's disease
7942	TFEB, BHLHE35	P19484	Renal carcinoma
57215	THAP11, HRIHFB2206	Q96EK4	Not Described
730130	TMEM229A	B2RXF0	Not Described
27324	TOX3, CAGF9, TNRC9	O15405	A minor allele of this gene has been implicated in an elevated risk of breast cancer.
8848	TSC22D1, KIAA1994, TGFB1I4, TSC22, hucep-2	Q15714	May play a critical role in tumor suppression through the induction of cancer cell apoptosis, and a single nucleotide polymorphism in the promoter of this gene has been associated with diabetic nephropathy.
205717	USF3, KIAA2018	Q68DE3	A deletion in the polyglutamine region was associated with risk for thyroid carcinoma.
7716	VEZF1, DB1, ZNF161	Q14119	Involved in both normal and abnormal cellular proliferation and differentiation.
463	ZFH3, ATBF1	Q15911	Prostate cancer, susceptibility to, somatic
171017	ZNF384, CAGH1, CIZ, NMP4, TNRC1	Q8TF68	Acute leukemia



Supplementary Figure 1. Linear regression analyzes showing that number of interactors is not related with the size of the polyQ proteins (Linear regression test; $P=0.084$).

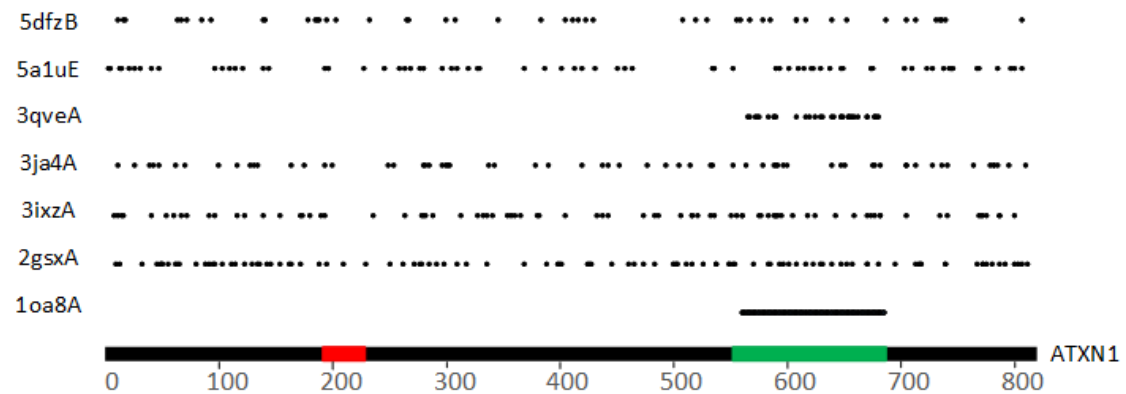
Supplementary Table 4. Proteins reported to interact with human ATXN1L. The cells marked in black represent common presence in ATXN1L and ATXN1. In bold are the proteins that preferentially bind to expanded ATXN1.

GeneID	Gene Name	In common with ATXN1
50	ACO2	
134	ADORA1	
166	AES	
322	APBB1	
1173	AP2M1	
1487	CTBP1	
1488	CTBP2	
1601	DAB2	
2037	EPB41L2	
2521	FUS	
3096	HIVEP1	
3516	RBPJ	
4077	NBR1	
4147	MATN2	
4152	MBD1	
4223	MEOX2	
4817	NIT1	
4914	NTRK1	
6310	ATXN1	
6601	SMARCC2	
6720	SREBF1	
7070	THY1	
7916	PRRC2A	
8161	COIL	
8522	GAS7	
8906	AP1G2	
9139	CBFA2T2	
9146	HGS	
9169	SCAF11	
9321	TRIP11	
9415	FADS2	
9463	PICK1	
9555	H2AFY	
9611	NCOR1	
9612	NCOR2	
9802	DAZAP2	
10147	SUGP2	
10472	ZBTB18	
11030	RBPMS	
11129	CLASRP	
11332	ACOT7	
23053	ZSWIM8	

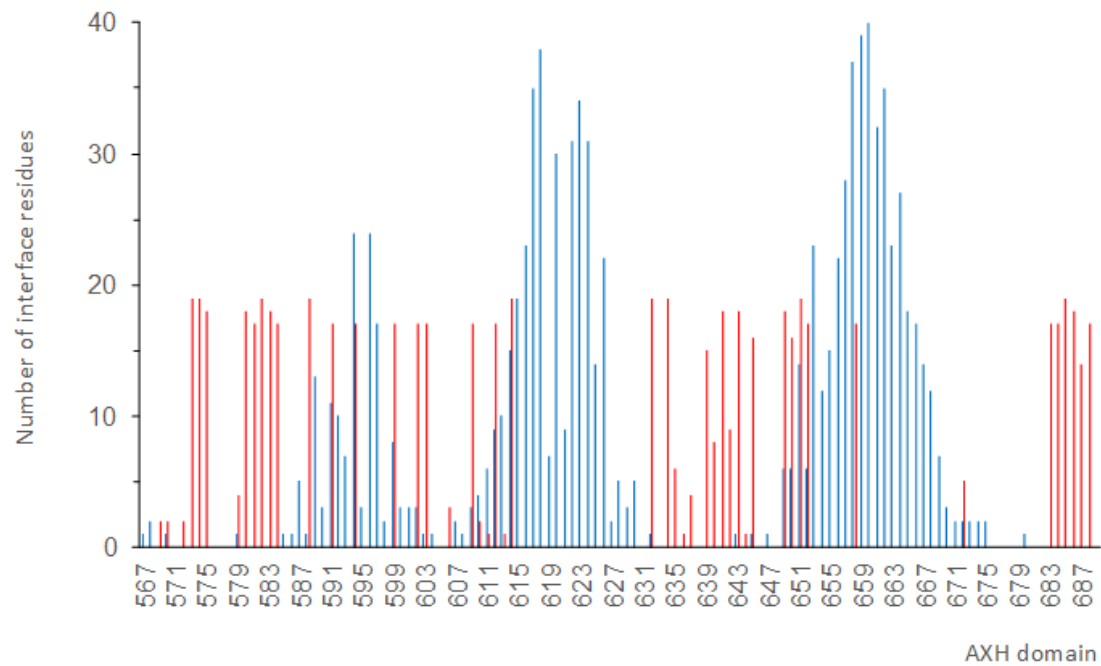
23131	GPATCH8	
23132	RAD54L2	
23136	EPB41L3	
23264	ZC3H7B	
23271	CAMSAP2	
23543	RBFOX2	
23587	ELP5	
26003	GORASP2	
26205	GMEB2	
29085	PHPT1	
30815	ST6GALNAC6	
51754	TMEM8B	
54882	ANKHD1	
55249	YY1AP1	
64121	RRAGC	
64745	METTLL17	
79868	ALG13	
84181	CHD6	
84632	AFAP1L2	
84726	PRRC2B	
84866	TMEM25	
114785	MBD6	
159195	USP54	
342371	ATXN1L	
342667	STAC2	
404734	ANKHD1EIF4EBP3	
100130086	HSFX2	
100506164	HSFX1	

Supplementary Table 5. Prediction agreement for the number of interactors with wt and expanded ATXN1 when using the *in silico* and Suter *et al.* (2013) methodologies. In brackets are the number of cases that show agreement versus the total number of interactors analyzed.

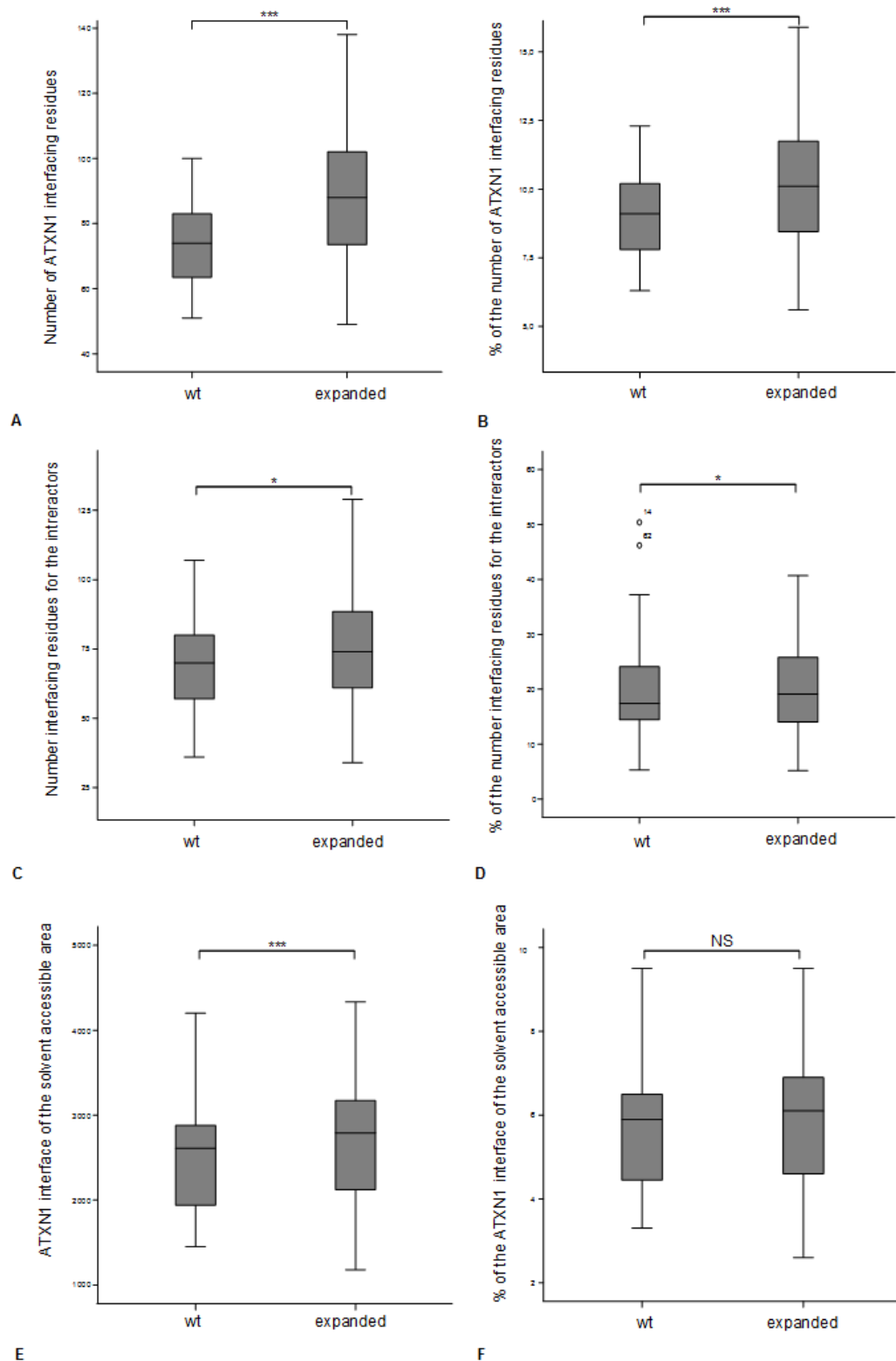
Parameters	Prediction agreement (%)	
NRI of the ATXN1	wt ATXN1	27.3 (3/11)
	expanded ATXN1	66.7 (40/60)
% NRI of the ATXN1	wt ATXN1	29.4 (5/17)
	expanded ATXN1	66.7 (36/54)
NRI of the interactor	wt ATXN1	28.0 (7/25)
	expanded ATXN1	71.7 (33/46)
% NRI of the interactor	wt ATXN1	28.0 (7/25)
	expanded ATXN1	71.7 (33/46)
SAA of the ATXN1	wt ATXN1	20.0 (4/20)
	expanded ATXN1	64.7 (33/51)
% SAA of the ATXN1	wt ATXN1	24.0 (6/25)
	expanded ATXN1	71.8 (28/39)
SAA of the interactor	wt ATXN1	41.2 (3/17)
	expanded ATXN1	64.8 (35/54)
% SAA of the interactor	wt ATXN1	18.8 (3/16)
	expanded ATXN1	64.2 (34/53)



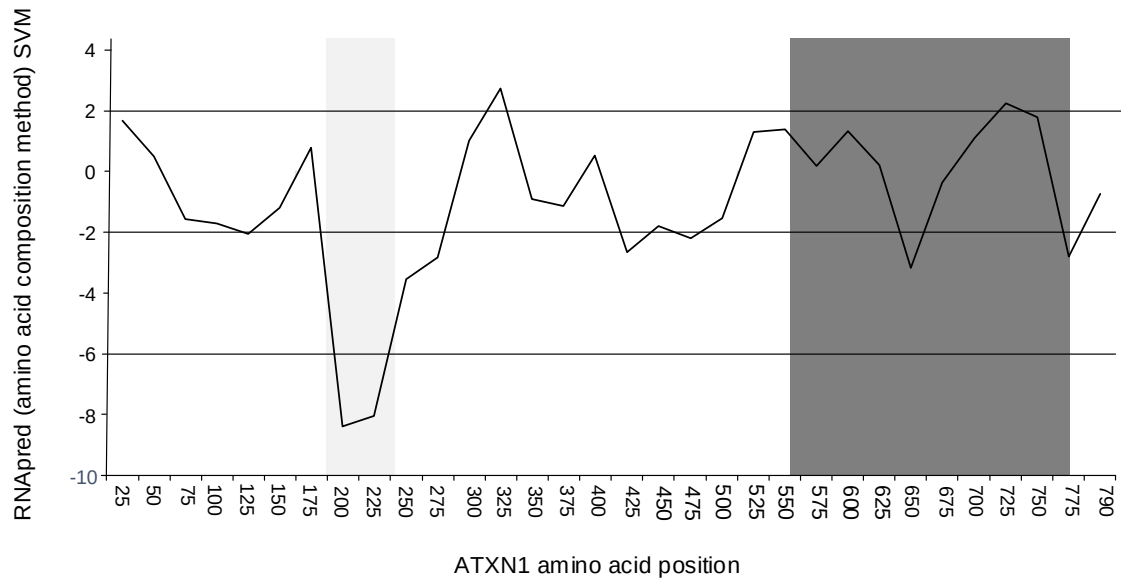
Supplementary Figure 2. Analyzes showing that the polyQ region (red) has fewer amino acids used as threading templates making the prediction of these regions less reliable. The AXH domain (green) shows a more reliable prediction with a higher amino acids used as threading templates.



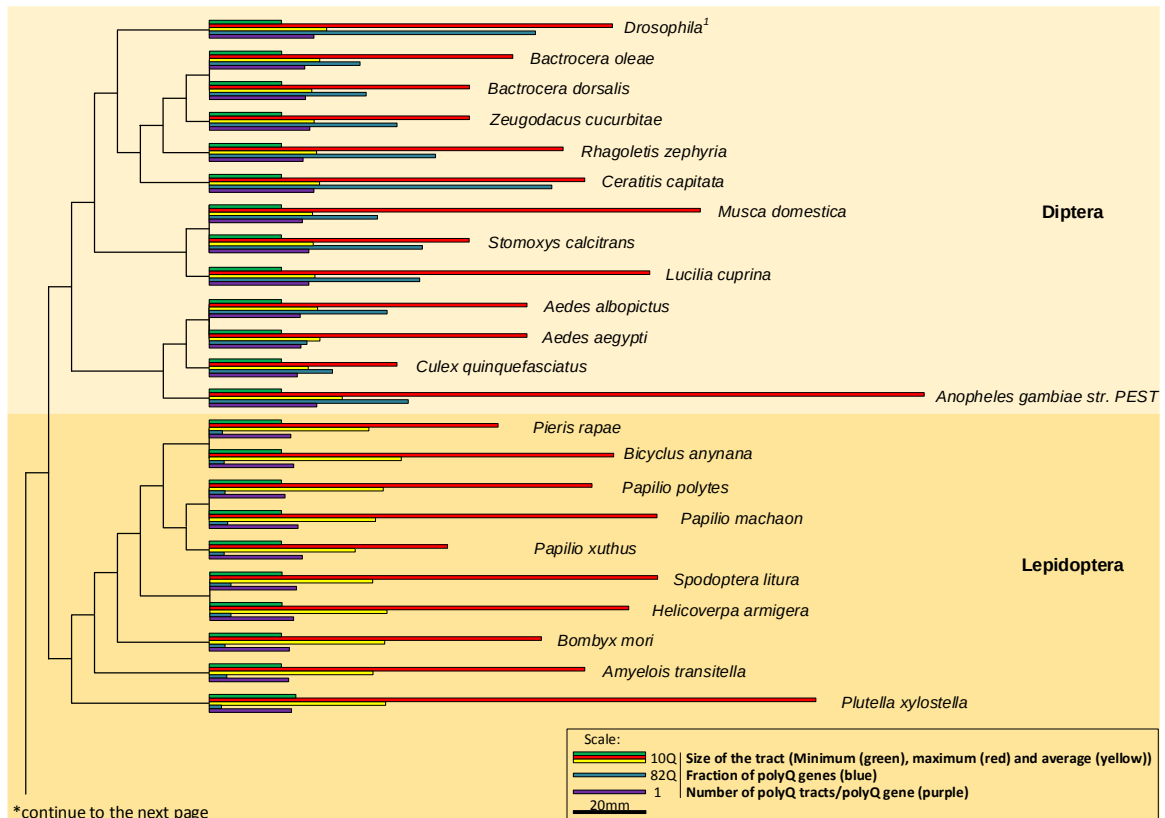
Supplementary Figure 3. Analyzes showing an overlap is seven interaction residues (ATXN1 sites 588, 591, 594, 599, 602, 609, and 649) between the predictions and those identified when using the 19 available crystal structures for the complex ATXN1/AXH domain and CIC (de Chiara *et al.*, 2013a. b; Kim *et al.*, 2013). Therefore, these regions are fundamental for the PPI of ATXN1. The percentage of the number of interface residues at the AXH domain for the wt (in blue) and expanded (in red) ATXN1.



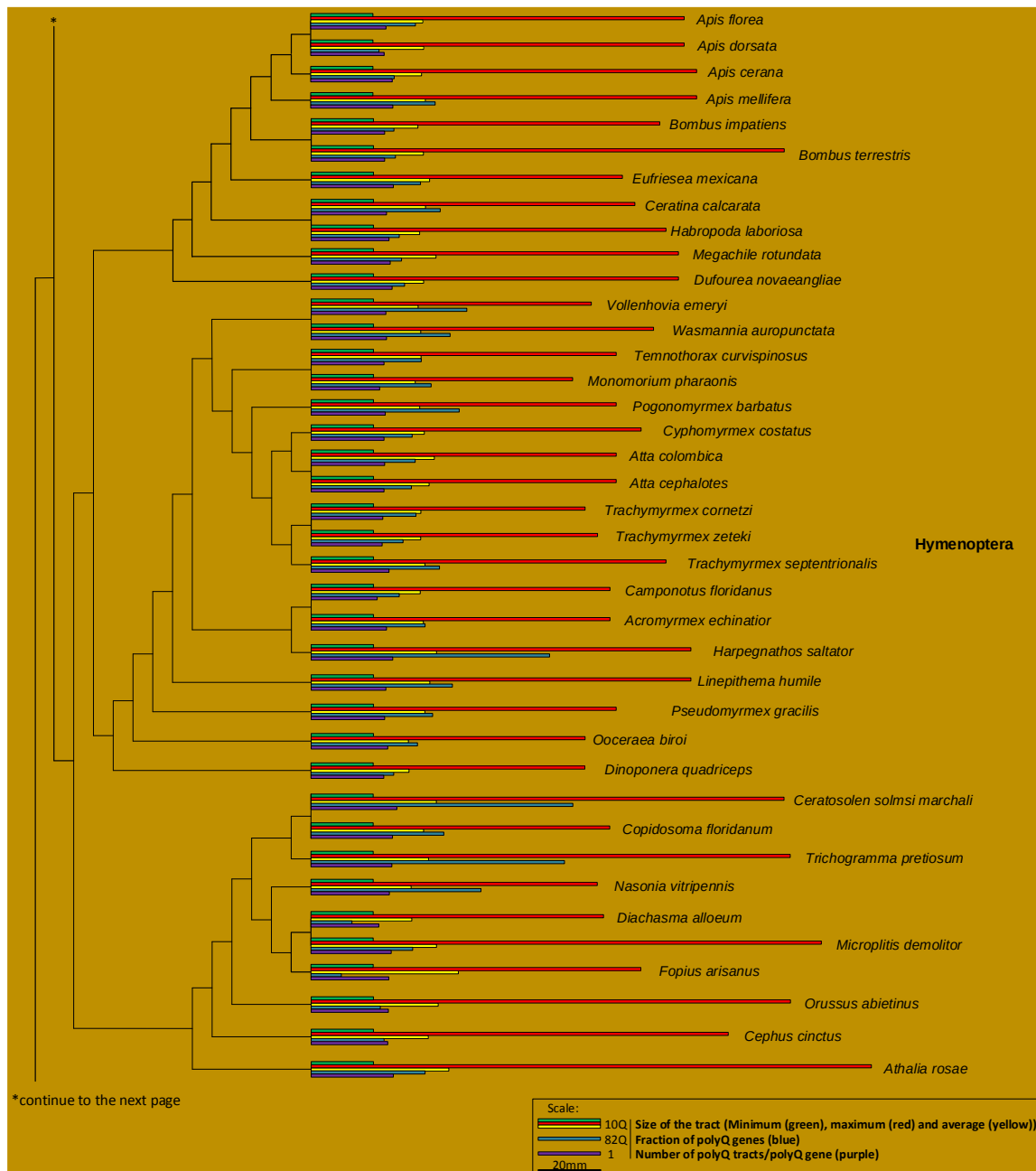
Supplementary Figure 4. Analysis of the docking results of the expanded and wt ATXN1 when using all the 71 proteins revealed similar results (A - Number of interface residues for ATXN1; B - % of the number of interface residues for ATXN1; C - Number of interface residues for the interactors; D - % of the number of interface residues for the interactors; E - Solvent-accessible area for ATXN1; F - % of the Solvent-accessible area for ATXN1).



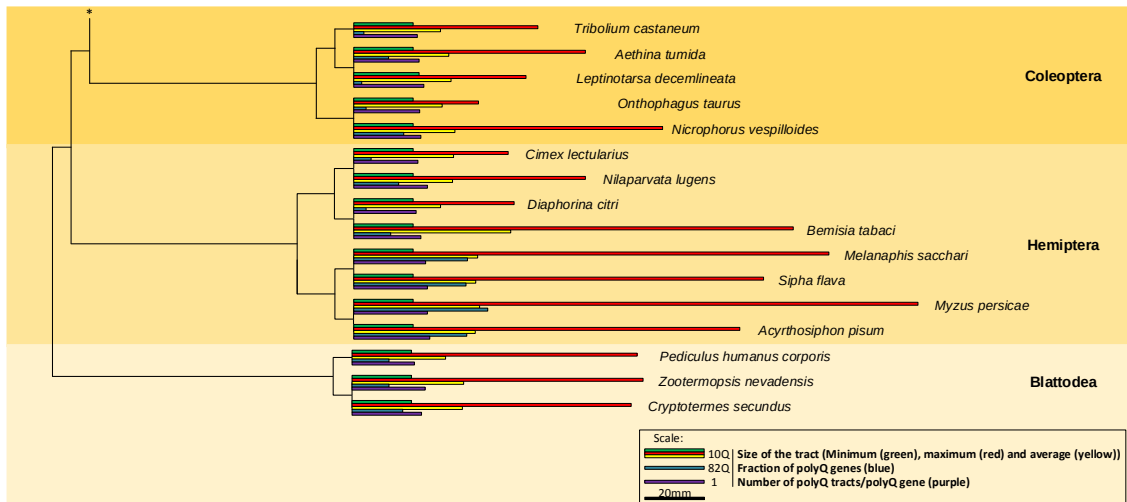
Supplementary Figure 5. Graphical representation of RNAPred results with a sliding window of 50 residues and an increment of 25 residues, suggesting four putative RNA binding regions, one before the polyQ region and three after.



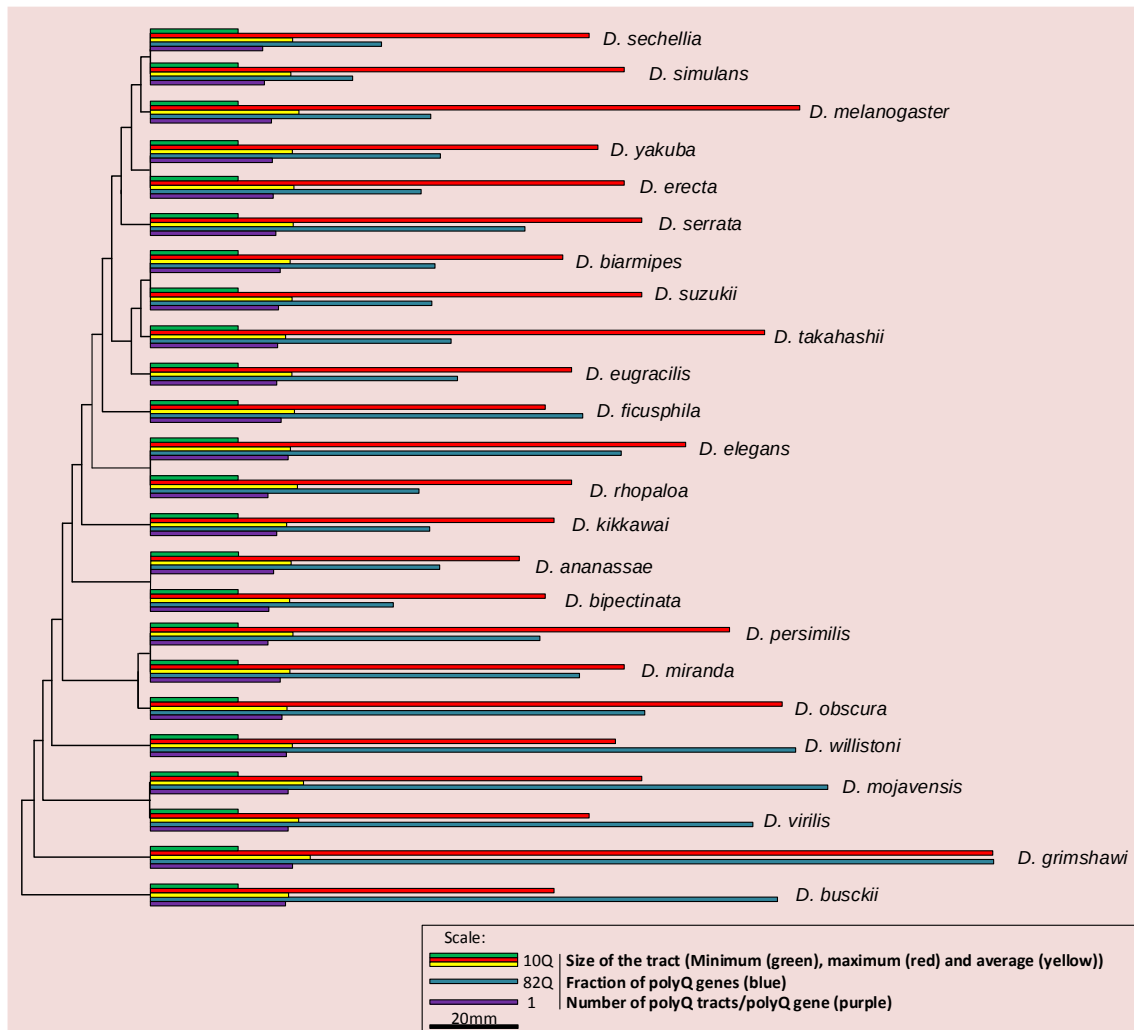
Supplementary Figure 6. Cladogram representation of the Insecta class (Protostomians), with the Diptera and Lepidoptera groups. The relationships were designed based and accordingly to NCBI taxonomy, Bertone *et al.* (2008) and Kawahara and Breinholt (2014). The fraction of polyQ genes and their composition varies widely in these groups, without displaying a general pattern of evolution. The Lepidoptera group shows a distinctive relative higher average size of the polyQ tract. *Drosophila*¹ displays a distinctive higher number of polyQ genes and, therefore, was deeper analyzed (Supplementary Figure 9).



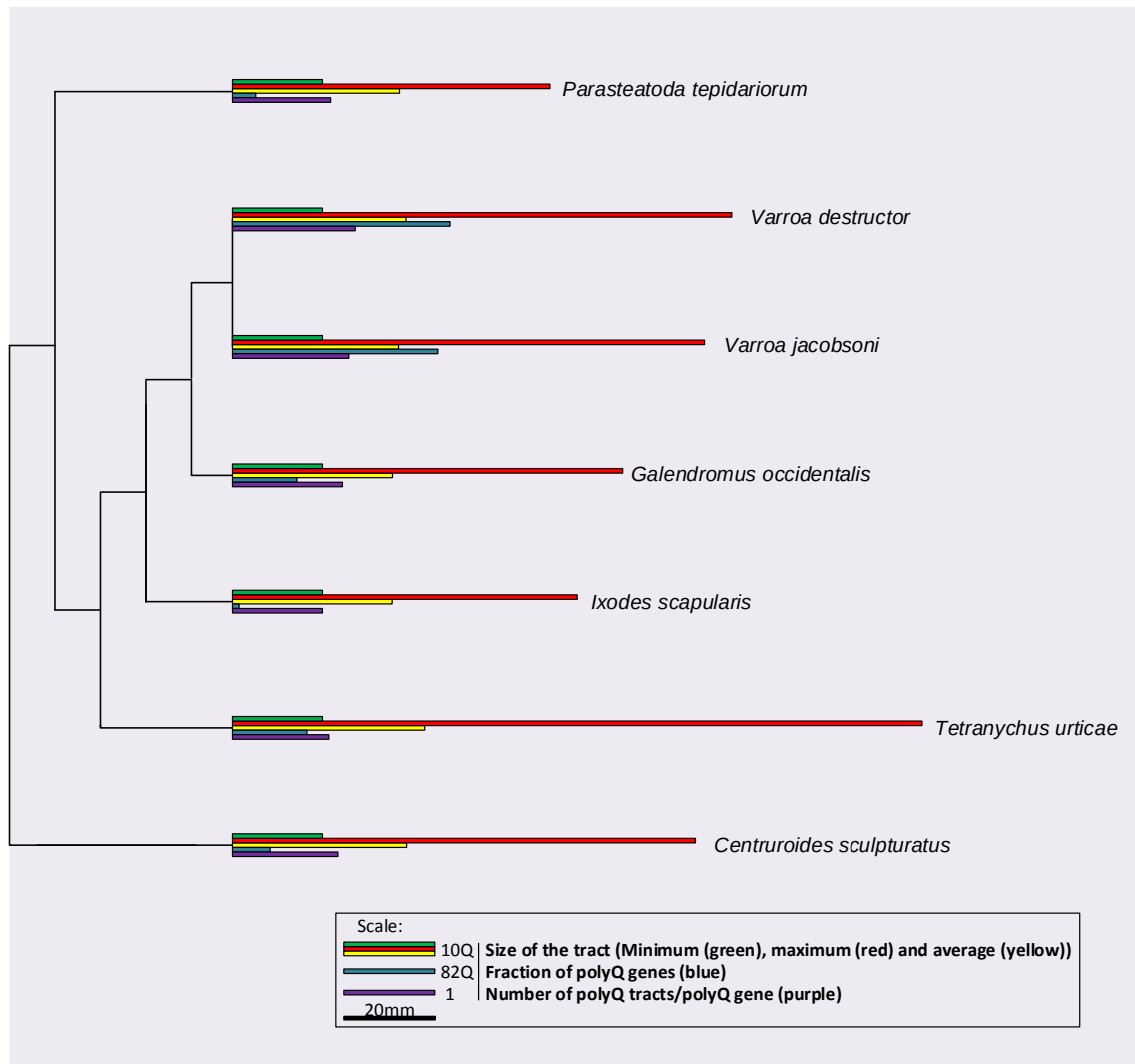
Supplementary Figure 7. Cladogram representation of the Insecta class (Protostomians), with the Hymenoptera group. The relationships were designed based and accordingly to NCBI taxonomy, Tree-of-life database (<http://tolweb.org> (Maddison, 2007)), Engel (2001) and Moreau *et al.* (2006). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution.



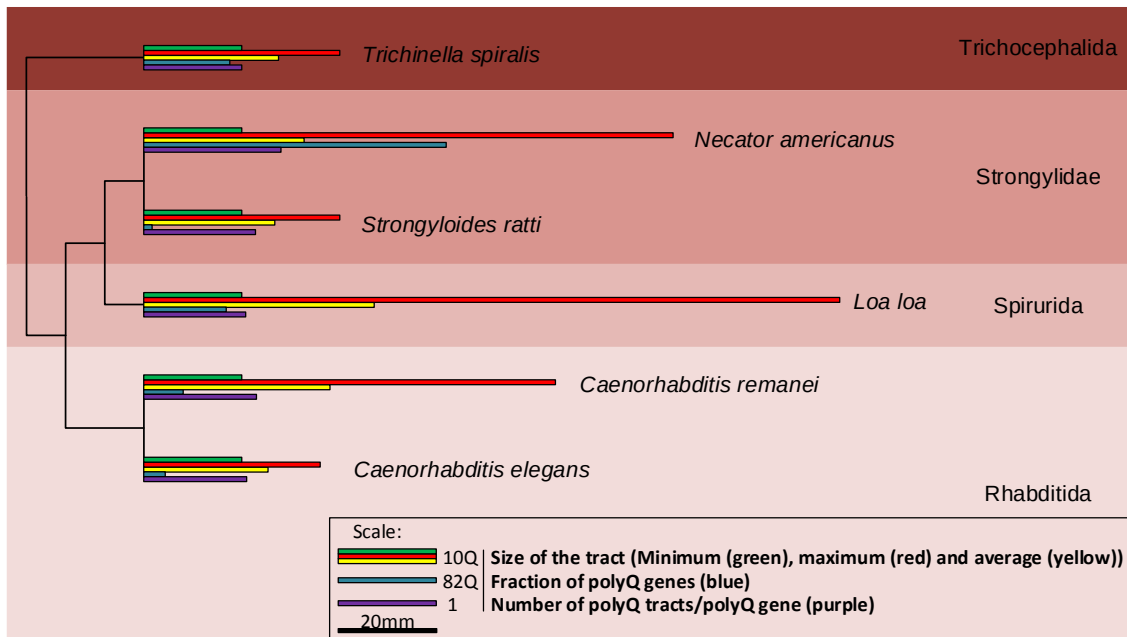
Supplementary Figure 8. Cladogram representation of the Insecta class (Protostomians), with the Coleoptera, the Hemiptera and Blattodea groups. The relationships were designed based and accordingly to NCBI taxonomy, Tree-of-life database (<http://tolweb.org>, (Maddison, 2007)), Zhang *et al.* (2018b) and Kim *et al.* (2011). The fraction of polyQ genes and their composition varies also widely in these groups, without displaying a general pattern of evolution.



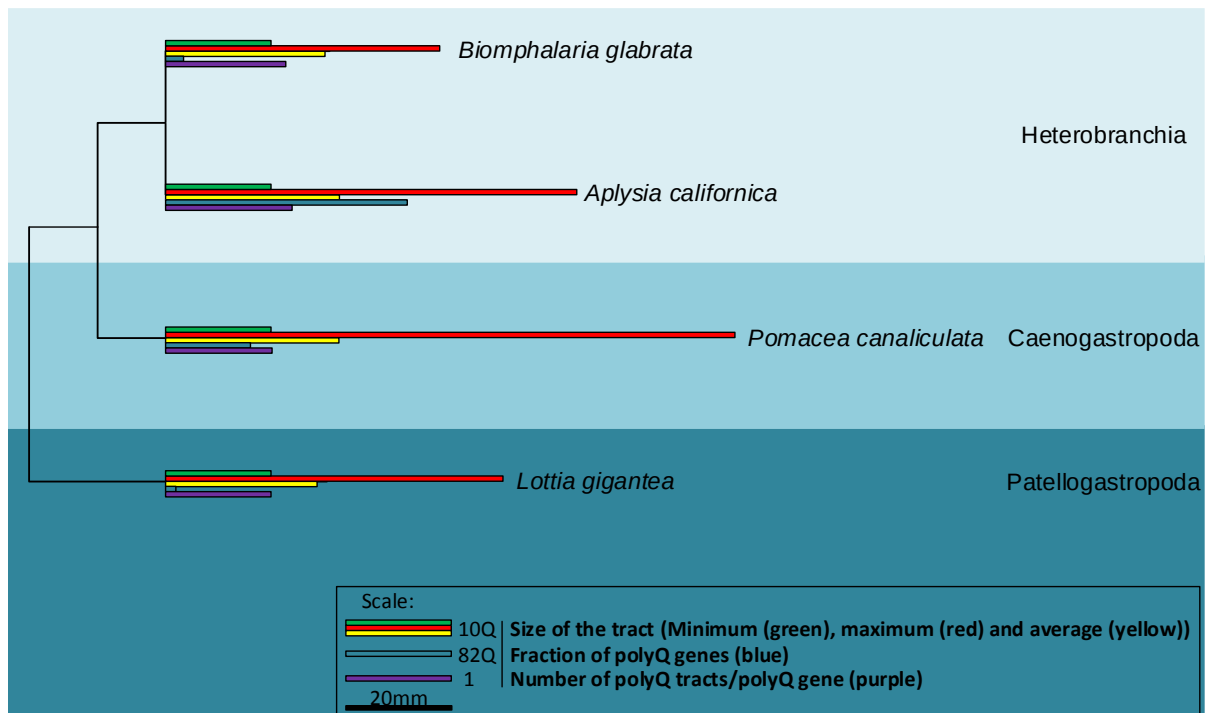
Supplementary Figure 9. Cladogram representation of the *Drosophila* group (Diptera) of the Insecta class (Protostomians). The relationships in this family were designed based and accordingly to NCBI taxonomy, Markow (2015) and Reis *et al.* (2016). This particular family displays a distinctive higher number of polyQ genes and number of tracts per gene, when compared to other species from the Animalia Kingdom.



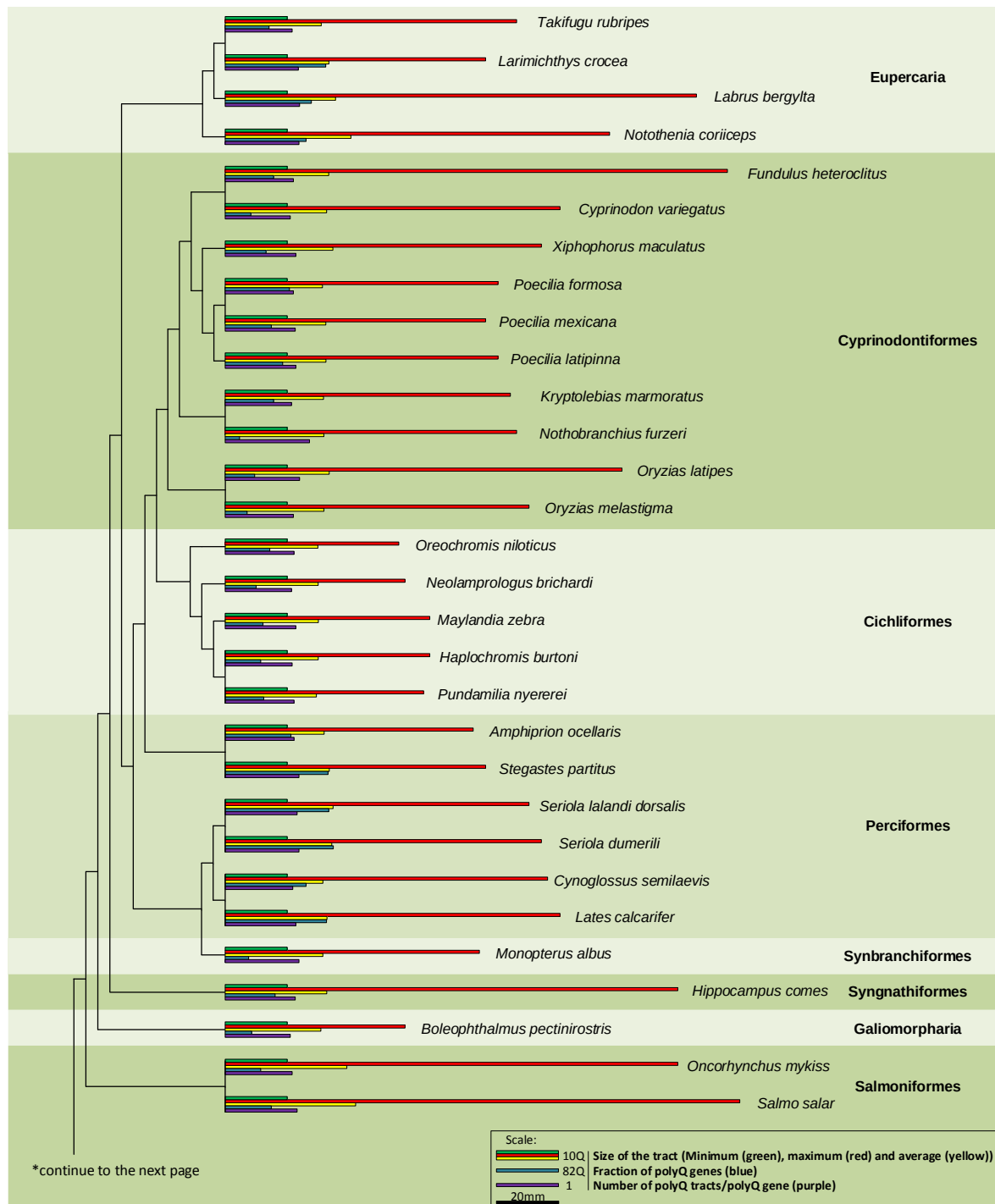
Supplementary Figure 10. Cladogram representation of the Arachnida class (Protostomians). The relationships in this group were designed based and accordingly to NCBI taxonomy, Henriques *et al.* (2019), Figueroa *et al.* (2016), Reeder *et al.* (2015), (Merchant *et al.*, 2018) and (Jones *et al.*, 2012). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution. The *Varroa* gender displays the highest number of polyQ genes in this particular class.



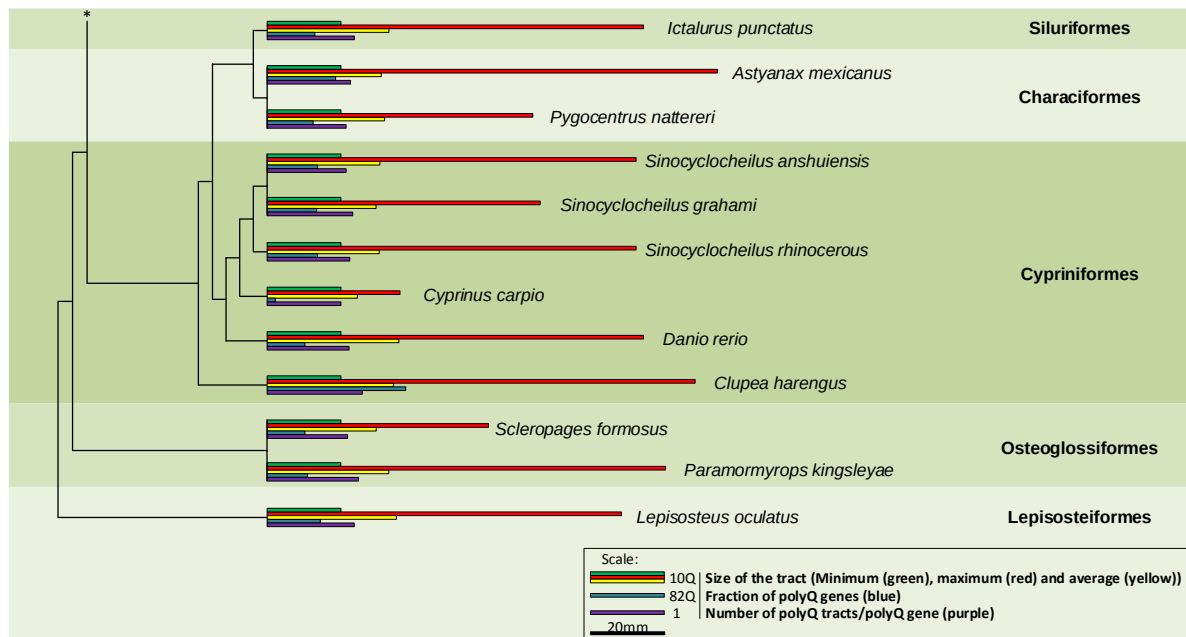
Supplementary Figure 11. Cladogram representation of the Nematoda class (Protostomians). The relationships in this group were designed based and accordingly to NCBI taxonomy and Henriques *et al.* (2019). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution. The *Loa loa* specie displays a distinctive higher polyQ tract size and the *Necator americanus* displays the highest number of polyQ genes.



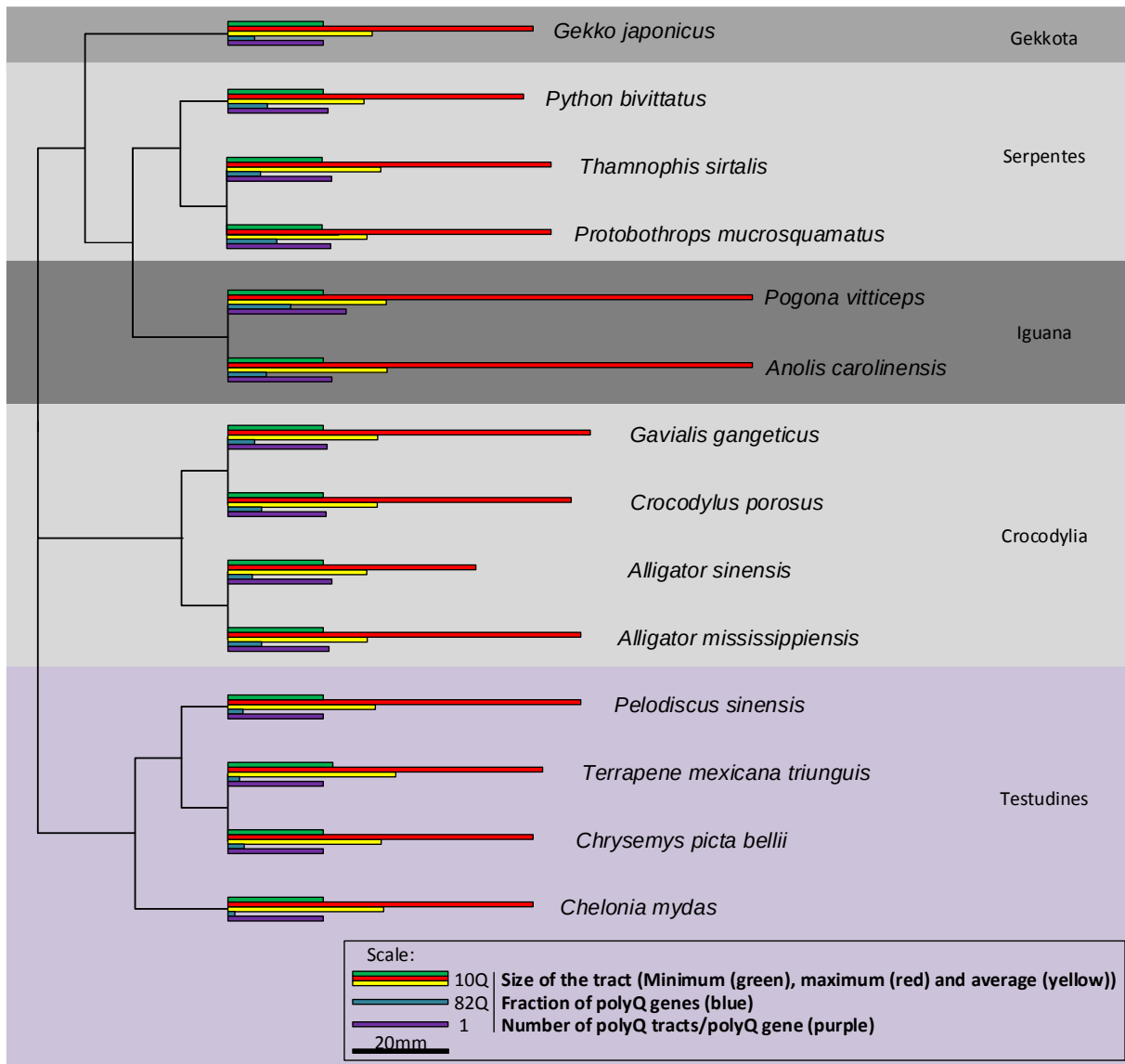
Supplementary Figure 12. Cladogram representation of the Gastropoda class (Protostomians). The relationships in this group were designed based and accordingly to NCBI taxonomy and Shimizu *et al.* (2013). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution. *Aplysia californica* displays the highest number of polyQ genes and *Pomacea canaliculata* displays the highest polyQ tract size.



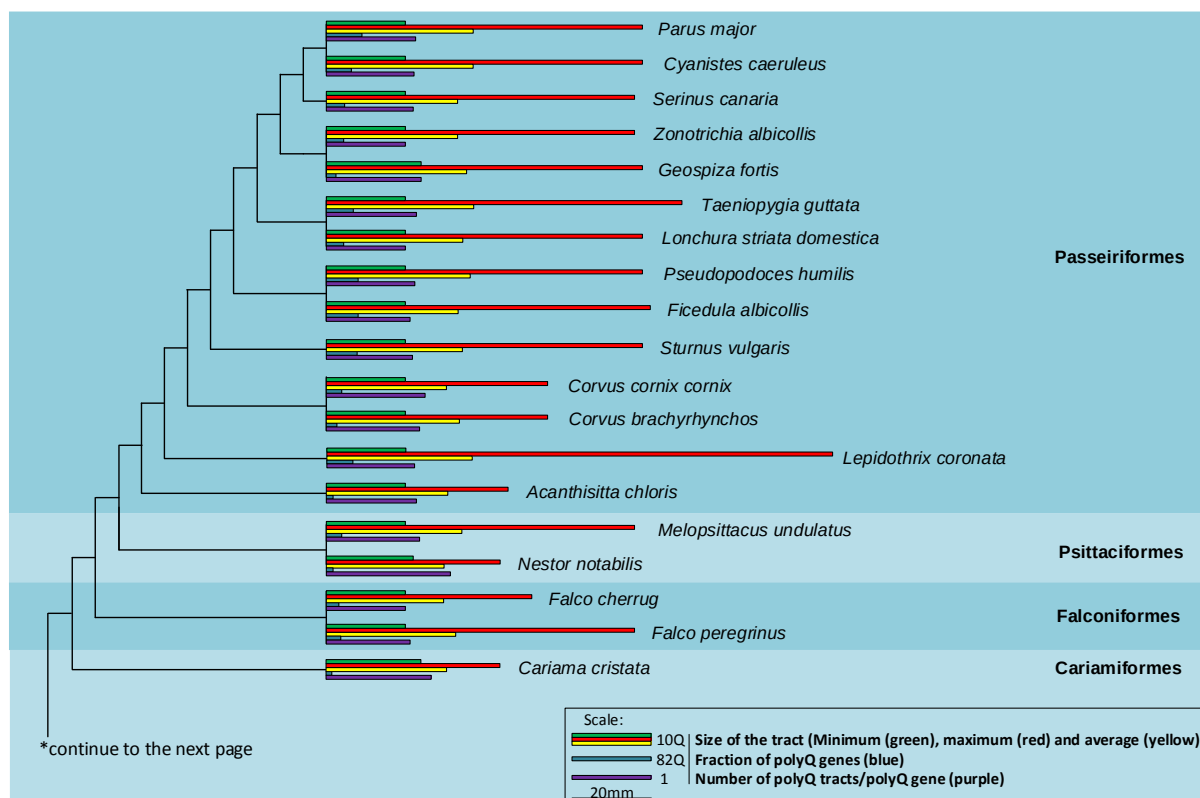
Supplementary Figure 13. Cladogram representation of the Actinopteri class (Deuterostomians), with the Eupercaria, Cyprinodontiformes, Cichliformes, Perciformes, Synbranchiformes, Syngnathiformes, Galiomorpha and Salmoniformes groups. The relationships in these groups were designed based and accordingly to NCBI taxonomy and Betancur *et al.* (2013), Sanciangco *et al.* (2016) and Helmstetter *et al.* (2016). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution.



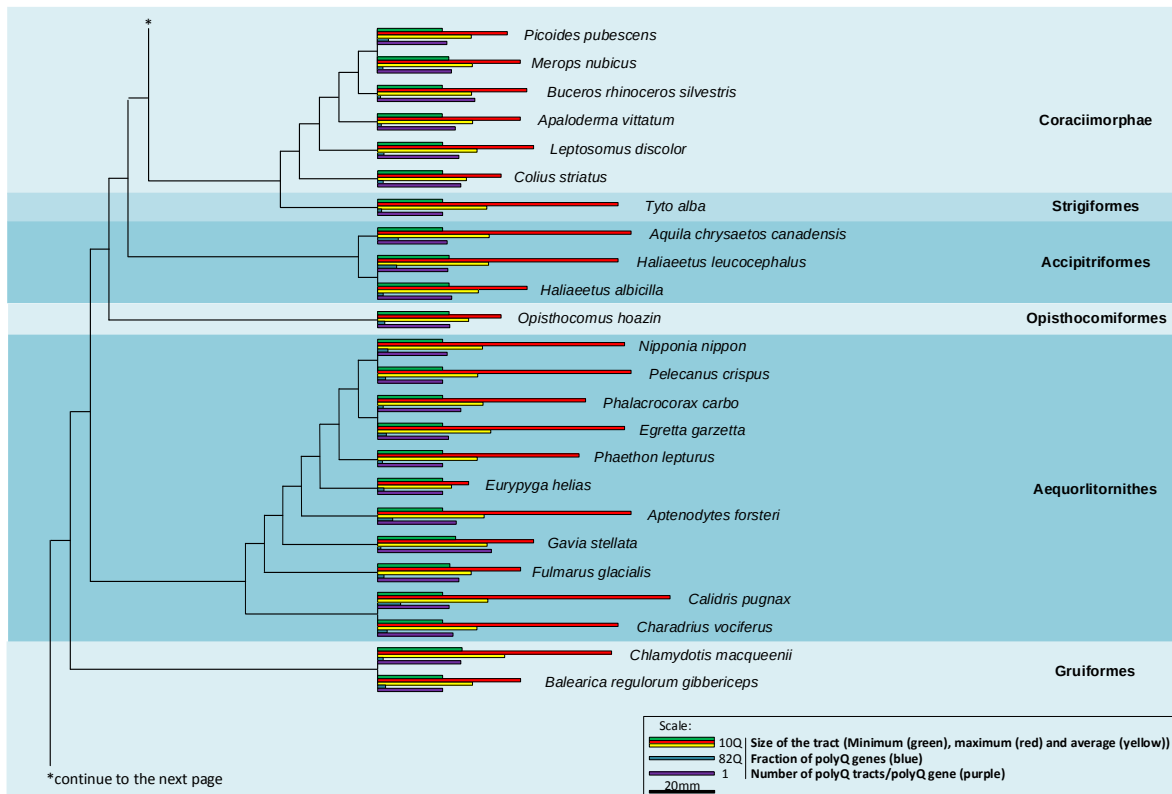
Supplementary Figure 14. Cladogram representation of the Actinopteri class (Deuterostomians), with Siluriformes, Characiformes, Cypriniformes, Osteoglossiformes and Lepisosteiformes. The relationships in these groups were designed based and accordingly, Betancur *et al.* (2013) and Hughes *et al.* (2018). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution.



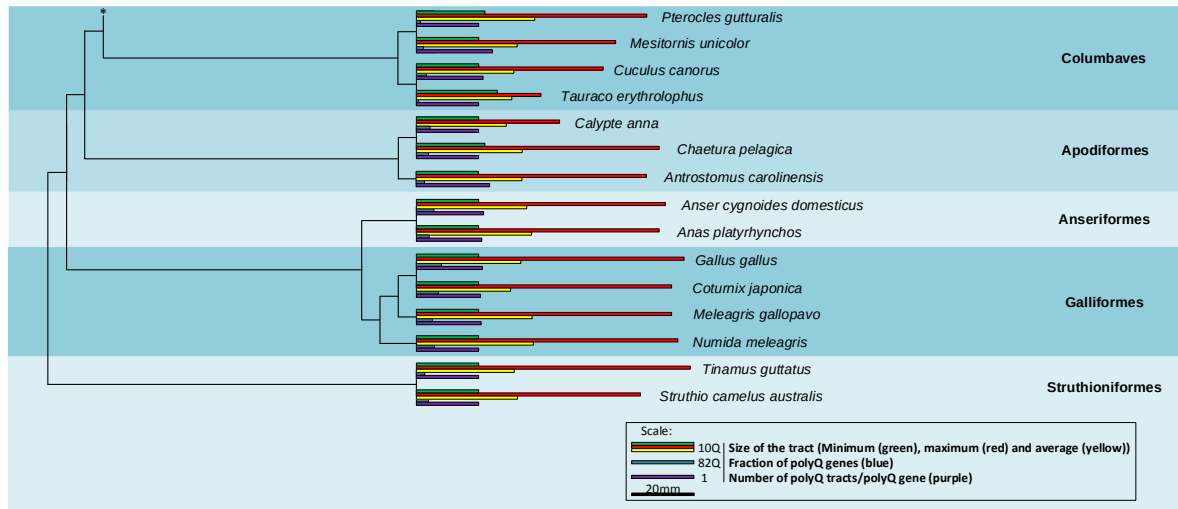
Supplementary Figure 15. Cladogram representation of the Reptilia group (Deuterostomians). The relationships in these groups were designed based and accordingly to NCBI taxonomy, Jones *et al.* (2012), Figueroa *et al.* (2016), Reeder *et al.* (2015) and Merchant *et al.* (2018). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution.



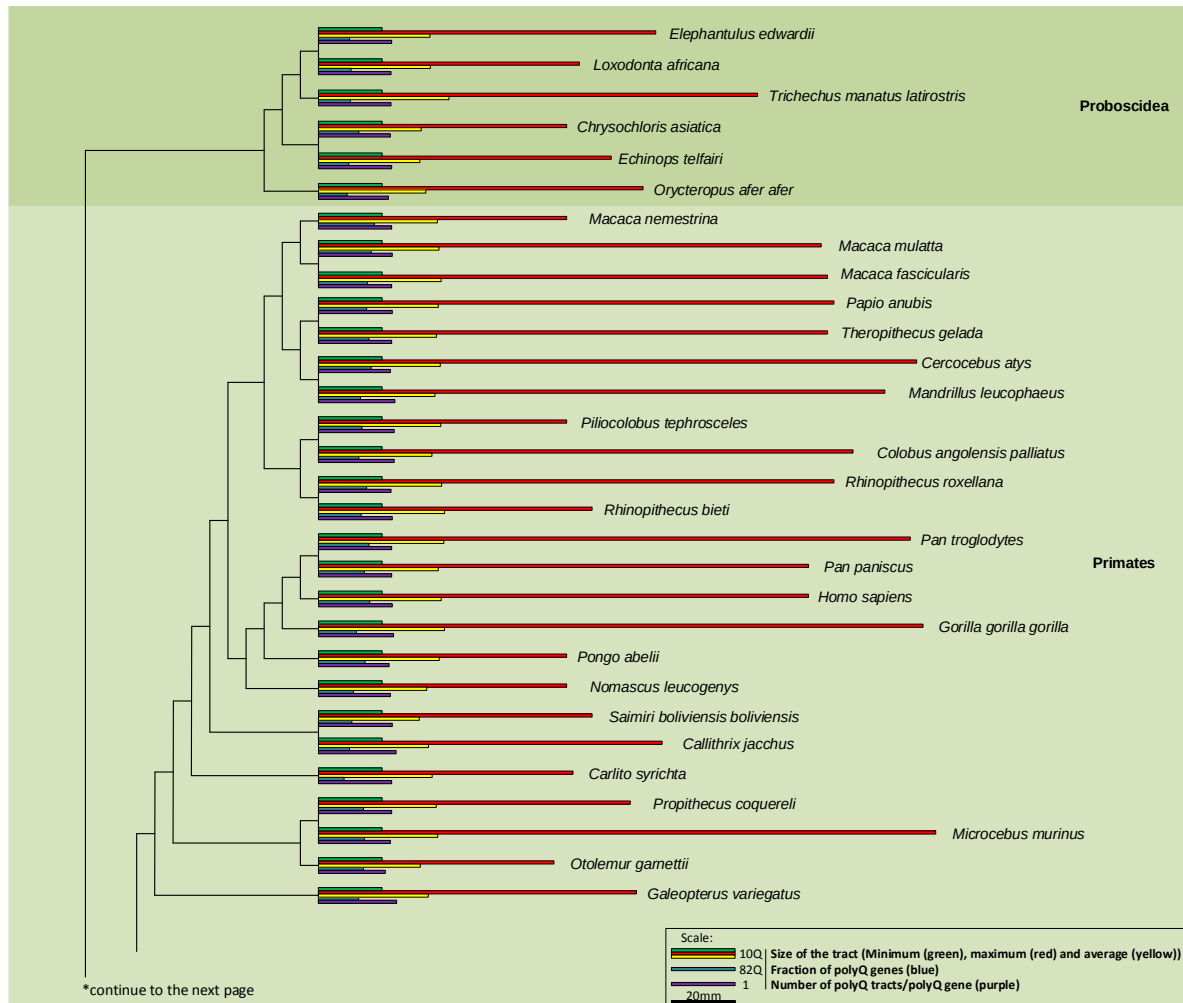
Supplementary Figure 16. Cladogram representation of the Aves class (Deuterostomians), with Passeriformes, Psittaciformes, Falconiformes and Cariamiformes. The relationships in these groups were designed based and accordingly to NCBI taxonomy and Prum *et al.* (2015). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution



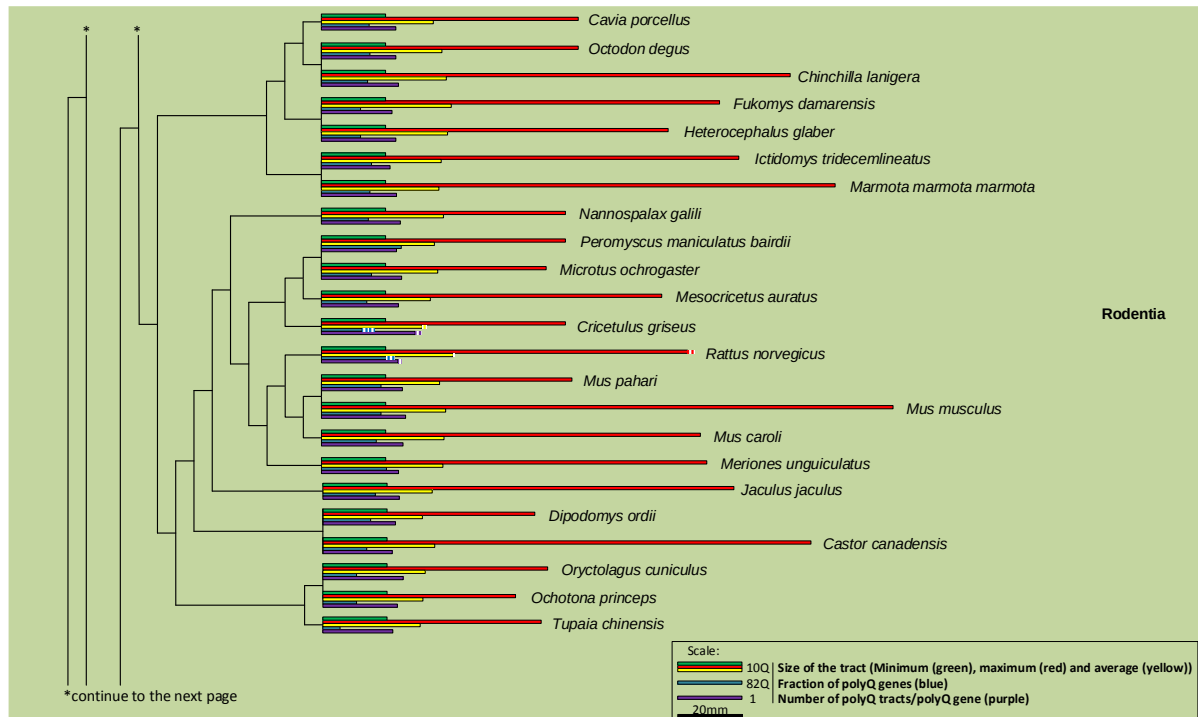
Supplementary Figure 17. Cladogram representation of the Aves class (Deuterostomians), Coraciimorphae, Strigiformes, Accipitriformes, Opisthocomiformes, Aequorlornithes and Gruiformes. The relationships in these groups were designed based and accordingly to NCBI taxonomy and Prum *et al.* (2015). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution



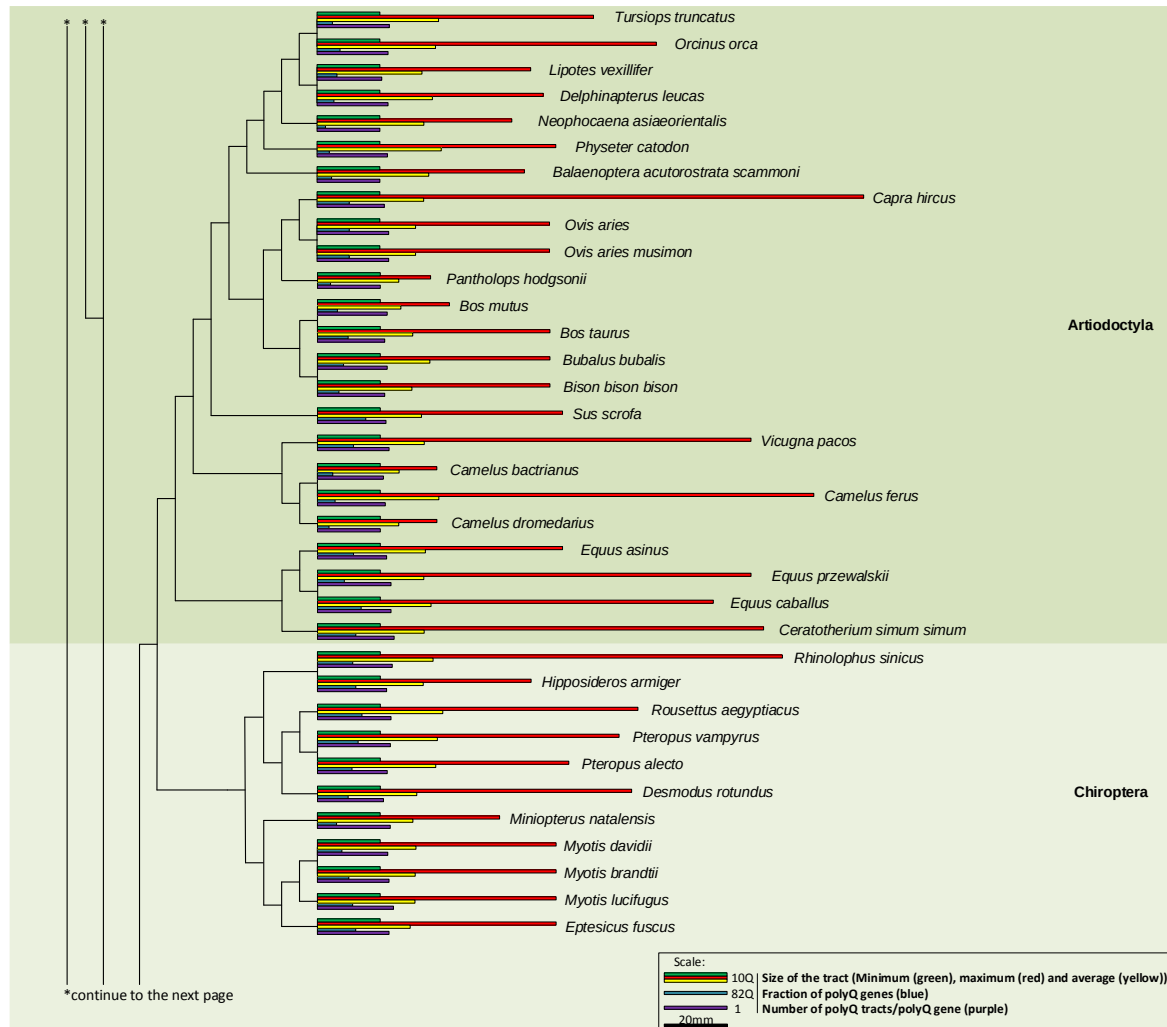
Supplementary Figure 18. Cladogram representation of the Aves class (Deuterostomians), with Columbaves, Apodiformes, Anseriformes, Galliformes and Struthioniformes. The relationships in these groups were designed based and accordingly to NCBI taxonomy, Prum *et al.* (2015) and (Yuri *et al.*, 2013). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution



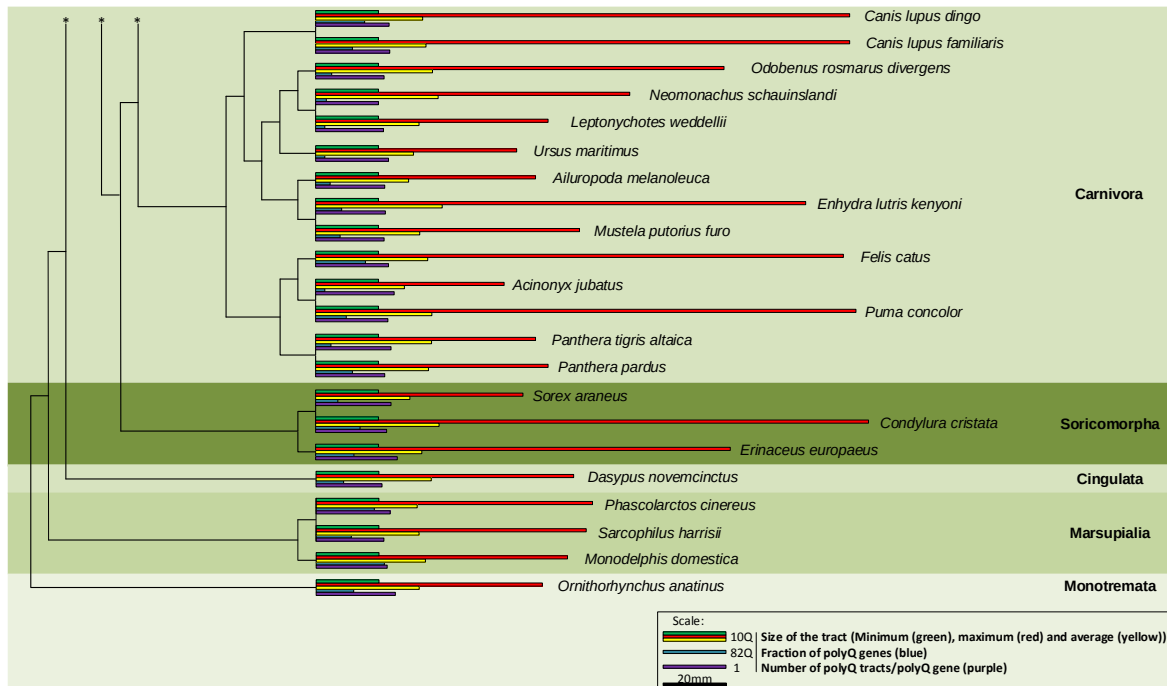
Supplementary Figure 19. Cladogram representation of the Mammalia class (Deuterostomians), with the Proboscidea and the Primates groups. The relationships in these groups were designed based and accordingly to NCBI taxonomy, Helgen (2011) and (Perelman *et al.*, 2011). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution. The primates group displays a distinctive higher size of the polyQ tract, when compared to other groups of the Animalia Kingdom.



Supplementary Figure 20. Cladogram representation of the Mammalia class (Deuterostomians), with the Rodentia group. The relationships in this group were designed based and accordingly to NCBI taxonomy and Helgen (2011). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution.



Supplementary Figure 21. Cladogram representation of the Mammalia class (Deuterostomians), with the Artiodactyla and Chiroptera groups. The relationships in these groups were designed based and accordingly to NCBI taxonomy, Helgen (2011) and Agnarsson *et al.* (2011). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution.



Supplementary Figure 22. Cladogram representation of the Mammalia class (Deuterostomians), with the Carnivora, Soricomorpha, Cingulata, Marsupialia and Monotremata groups. The relationships in these groups were designed based and accordingly to NCBI taxonomy, Helgen (2011) and Tree-of-life database (<http://tolweb.org> (Maddison, 2007)). The fraction of polyQ genes and their composition varies widely among the species of this group, without displaying a general pattern of evolution