Blood and cerebellar abundance of *ATXN3* splice variants in spinocerebellar ataxia type 3/Machado-Joseph disease

Mafalda Raposo, PhD[1,2*], Jeannette Hübener-Schmid, PhD[3,4*], Rebecca Tagett, MSc[5], Ana F. Ferreira, MSc[2], Ana Rosa Vieira Melo, MSc[2], João Vasconcelos, MD[6], Paula Pires, MD[7], Teresa Kay, MD, PhD[8], Hector Garcia-Moreno, MD[9,10], Paola Giunti, MD, PhD[9,10], Magda M. Santana, PhD[11,12,13], Luis Pereira de Almeida, PhD[11,12,14], Jon Infante, MD, PhD[15], Bart P. van de Warrenburg, MD, PhD[16], Jeroen J. de Vries, MD[17], Jennifer Faber, MD[18,19], Thomas Klockgether, MD[18,19], Nicolas Casadei, PhD[3,20], Jakob Admard, MSc[3,20], Ludger Schöls, MD[21,22], Olaf Riess, MD[3,4,20], European Spinocerebellar ataxia type 3/Machado-Joseph disease Initiative (ESMI) study group, Maria do Carmo Costa, PhD[23*], Manuela Lima, PhD[2*]

1. Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal.
2. Faculdade de Ciências e Tecnologia, Universidade dos Açores, Ponta Delgada, Portugal.
3. Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany.
4. Centre for Rare Diseases, University of Tübingen, Tübingen, Germany.
5. Bioinformatics Core, Michigan Medicine, University of Michigan, Ann Arbor, MI, USA
6. Serviço de Neurologia, Hospital do Divino Espírito Santo, Ponta Delgada, Portugal.
7. Serviço de Neurologia, Hospital do Santo Espírito da Ilha Terceira, Angra do Heroísmo, Portugal.
8. Serviço de Genética Clínica, Hospital D. Estefânia, Lisboa, Portugal.
9. Ataxia Centre, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, UK.
10. Department of Neurogenetics, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK.
11. Center for Neuroscience and Cell Biology (CNC), University of Coimbra, Coimbra, 3000-075, Portugal.
12. Center for Innovative Biomedicine and Biotechnology (CIBB), University of Coimbra, Coimbra, Portugal.
13. Institute for Interdisciplinary Research, University of Coimbra (IIIUC), Coimbra, Portugal.
14. Faculty of Pharmacy, University of Coimbra (FFUC), Coimbra, Portugal
15. Neurology Service, University Hospital Marqués de Valdecilla-IDIVAL, Universidad de Cantabria, Centro de Investigación en Red de Enfermedades Neurodegenerativas (CIBERNED), Santander, Spain.
16. Radboud University Medical Centre, Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Nijmegen, The Netherlands.
17. Department of Neurology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
18. Department of Neurology, University Hospital Bonn, Bonn, Germany.
19. German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany.
20. NGS Competence Center Tübingen, Tübingen, Germany.
21. Department for Neurodegenerative Diseases, Hertie-Institute for Clinical Brain Research and Center for Neurology, University of Tübingen, Germany.
22. German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany.
23. Department of Neurology, Michigan Medicine, University of Michigan, Ann Arbor, MI, USA.

*Corresponding authors:

Mafalda Raposo

Instituto de Investigação e Inovação em Saúde (i3S)

Rua Alfredo Allen, 208

4200-135 Porto, Portugal

Phone: +351 226 074 900

E-mail address: msraposo@i3s.up.pt

Jeannette Hübener-Schmid

Institute of Medical Genetics and Applied Genomics

Nägelestraße 5

72074 Tübingen, Germany

Phone: +49 7071 29 72276

E-mail address: jeannette.huebener@med.uni-tuebingen.de

Maria do Carmo Costa, Ph.D.

Department of Neurology, University of Michigan

A. Alfred Taubman Biomedical Sciences Research Building, Room 4027

109 Zina Pitcher Place, Ann Arbor, MI 48109-2200, USA

Phone: +1 734-615-5632

E-mail address: mariadoc@med.umich.edu

Manuela Lima

Faculdade de Ciências e Tecnologia, Universidade dos Açores

Rua Mãe de Deus, N.13

9500-321 Ponta Delgada, Portugal

Phone: +351 296 650 118

E-mail address: maria.mm.lima@uac.pt

Word count: 3981 words

Running title: *ATXN3* splice variants in SCA3/MJD

Financial Disclosures/Conflict of Interest

TK is receiving research support from the Bundesministerium für Bildung und Forschung (BMBF), the National Institutes of Health (NIH) and Servier. Within the last 24 months, he has received consulting fees from Biogen, UCB and Vico Therapeutics. BvdW is supported by grants from ZonMw, NOW, Hersenstichting, Gossweiler Foundation, and Radboud university medical center; he has served on the scientific advisory boards or steering committees of uniQure, VICO Therapeutics, and Servier. LS is receiving research support from the European Commission, the Bundesministerium für Bildung und Forschung (BMBF) and the Bundesministerium für Gesundheit (BMG) as well as the Deutsche Forschungsgemeinschaft (DFG), Servier and Vigil Neuroscience. Within the last 24 months, he has received consulting fees from Vico Therapeutics.

The remaining authors declare that they have nothing to report.

Funding Sources

This work is an outcome of ESMI, an EU Joint Programme - Neurodegenerative Disease Research (JPND) project. The ESMI project was supported through the following funding organisations under the aegis of JPND: Germany, Federal Ministry of Education and Research (BMBF; funding codes 01ED1602A/B); Netherlands, The Netherlands Organisation for Health Research and

Development; Portugal, Fundação para a Ciência e a Tecnologia (FCT; funding codes JPCOFUND/0002/2015); United Kingdom, Medical Research Council. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 643417. MR is supported by FCT (CEECIND/03018/2018). AFF and ARVM received PhD fellowships from the FCT (SFRH/BD/121101/2016 and SFRH/BD/129547/2017, respectively). Fundo Regional para a Ciência e Tecnologia (FRCT, Governo Regional dos Açores) is currently supporting ESMI in the Azores, under the PRO-SCIENTIA program. NGS sequencing methods were performed with the support of the DFG-funded NGS Competence Center Tübingen (INST 37/1049-1). Bioinformatic analysis was funded by Fundo Regional para a Ciência e Tecnologia (FRCT, Governo Regional dos Açores), under the PRO-SCIENTIA program (ML) and by discretionary funds from the University of Michigan (MCC). MCC is supported by the Heeringa Ataxia Research Fund (University of Michigan). Several authors of this publication are members of the European Reference Network for Rare Neurological Diseases - Project ID No 739510.

ABSTRACT

Background: Spinocerebellar ataxia type 3 (SCA3)/Machado-Joseph disease (MJD) is an autosomal dominant polyglutamine disease. SCA3/MJD causative gene, *ATXN3*, is known to undergo alternative splicing (AS) and 54 transcripts are currently annotated. Differences in the toxicity of ataxin-3 protein isoforms, harbouring on its C-terminus two or three ubiquitin interacting motifs (UIMs), were previously uncovered, raising the hypothesis that specific *ATXN3* splice variants play key roles in promoting the selective toxicity displayed in SCA3/MJD.

Methods: Using RNA-seq datasets we identified and determined the abundance of annotated *ATXN3* transcripts in blood (n=60) and cerebellum (n=12) of SCA3/MJD subjects and controls.

Results: Globally, the number and the abundance of individual *ATXN3* transcripts were higher in the cerebellum than in the blood. While the most abundant transcript in the cerebellum was a protein with a coding sequence not defined of unknown function (*ATXN3*-208), the transcript with the highest abundance in blood was the reference transcript (*ATXN3*-251) which translates into an ataxin-3 isoform harboring three UIMs. Noteworthy, the abundance of *ATXN3*-251 and *ATXN3*-214, two out of the four transcripts that encode full-length ataxin-3 protein isoforms but differ in the C-terminus were strongly related with tissue expression specificity: *ATXN3-251* (3UIM) was expressed in blood 50-fold more than in cerebellum, whereas *ATXN3*-214 (2UIM) was expressed in the cerebellum 20-fold more than in blood.

Conclusions: These findings provide new insights into the elucidation of *ATXN3* AS in different tissues, contributing for a better understanding of SCA3/MJD pathogenesis and providing information for the development of future effective *ATXN3* mRNA-lowering therapies.

KEYWORDS: ataxin-3, polyQ diseases, neurodegenerative disease, RNA-seq, mRNA, alternative splicing

INTRODUCTION

Alternative splicing (AS) is a fundamental step in the regulation of eukaryotic gene expression through which introns are excised and exons are ligated together to form the messenger RNA (mRNA).[1] Multiple evidence indicates that genetic variation in *cis*-acting splicing sequence elements or in *trans*-acting splicing factors plays a relevant role in patients with neurodegenerative conditions (reviewed in [2,3]). Importantly, *post-mortem* brains of Huntington disease (HD) patients and of mouse models show aberrant splicing of the HD causative gene (huntingtin, *HTT*) itself, leading to the production of high levels of a splice variant only containing exon 1 that encodes a highly aggregation-prone protein and thus contributes to disease pathogenesis.[4–7] Currently, several pathologies triggered by splicing defects, including neurodegenerative diseases, can be treated using different types of RNA molecules, such as antisense oligonucleotide (ASO) or small interference RNA (siRNA), either by reducing the levels of the defective protein or by producing the corrected protein (reviewed in [8,9]). In fact, ASO-mediated therapies are currently available for some common and rare diseases, including neurological diseases such as spinal muscular atrophy and Duchenne muscular dystrophy (reviewed in [9]).

SCA3/MJD is an autosomal dominant neurodegenerative polyglutamine (polyQ) disorder and the most common dominant ataxia worldwide.[10] SCA3/MJD is caused by an abnormal polyQ-encoding CAG repeat motif, consensually containing more than 60 trinucleotides within exon 10 of the *ATXN3* gene.[11–13] The *ATXN3* gene (ENSG00000066427) originates 54 transcripts[14] in the GRCh38.p13 human genome assembly including the canonical transcript (ENST00000644486/*ATXN3*-251) containing 11 exons and spanning a genomic region about 48 kb[15], encoding the ataxin-3 canonical protein isoform (P54252-2, UniProt). Ataxin-3 is a ubiquitously expressed deubiquitinating enzyme (*e.g.*, [16]) harbouring on its N-terminus the globular Josephin domain (JD) that includes the catalytic residues, and on its C-terminus a protein segment of undetermined structure containing the polyQ tract and two or three

ubiquitin interacting motifs (UIM).[17,18] Aggregation of the ataxin-3 protein is a hallmark of SCA3/MJD (reviewed in [19]) and occurs not only in brain regions with neuronal loss, such as the cerebellar dentate nucleus and the motor and pontine nuclei of the brainstem, but also in brain areas spared of neuronal destruction, like the cerebellar cortex.[20]

Despite nearly 30 years of research following the identification of *ATXN3* as the causative gene of SCA3/MJD[11,12], the characterization of the transcriptional species of *ATXN3* remains very incomplete. While *ATXN3* is known to undergo AS[12,15,17,21,22], the extent and regulation of AS, the relative abundance of alternative transcripts in health and disease conditions, the existence of transcript tissue/cell-specificity, and the impact of AS in SCA3/MJD neurotoxicity remain poorly understood. To the best of our knowledge, at least 24 *ATXN3* transcripts, likely to be translated into different isoforms of ataxin-3, have been detected in the human brain and/or peripheral tissues including the blood[12,15,17,21,22], although validation of a large number of such transcripts has not been conducted. *In silico* analysis performed to infer the functional impact of the several putative ataxin-3 protein isoforms showed that some variants were predicted to be "protective" (lacking the CAG tract), whereas others were foreseen to show increased toxicity.[22] Among the *ATXN3* transcripts which are predicted to be translated, only two transcripts encoding ataxin-3 isoforms that differ on the C-terminal containing either two (2UIM) or three UIMs (3UIM) have been mostly studied so far.[21,23,24] These studies using cellular, mouse or *Drosophila* models and *post-mortem* human brain tissues from healthy individuals, though controversial, uncovered differences in the toxicity of these two ataxin-3 protein isoforms by mainly revealing a faster degradation rate but higher aggregation propensity of the 2UIM isoform compared to the 3UIM.[21,23,24] The fact that these two ataxin-3 isoforms seem to differentially impact SCA3/MJD pathogenesis, raises the possibility that other species generated by *ATXN3* AS play key roles in promoting the selective neuronal toxicity displayed by SCA3/MJD patients.

Several phase I and II clinical trials are ongoing for polyQ disorders aiming to reduce the levels of disease-related transcripts/proteins using ASO-based drugs, but none was yet been approved

for clinical use (reviewed in [25]). Non-allele-specific as well as allele-specific ASO-mediated reduction of *ATXN3* have been successfully tested *in vivo* and *in vitro* models of MJD.[26–29] Noteworthy, the pharmacokinetics and safety study of a non-selective ASO for SCA3/MJD is currently ongoing (NCT05160558). The identification of specific *ATXN3* transcripts, whose expression needs to be blocked is, therefore, paramount to unveil better strategies in future ASO-based therapies of SCA3/MJD.

Hence, the goal of this study was to profile annotated *ATXN3* AS transcripts in blood and cerebellum of mutation carriers of SCA3/MJD and healthy individuals using RNA sequencing (RNA-seq) data.

MATERIAL AND METHODS

Samples and RNA-seq datasets

Two RNA-seq datasets, generated from a total of 72 samples, namely 60 samples from whole blood and 12 samples from *post-mortem* cerebellum (cerebellar lobules) were used in this study: i) a RNA-seq dataset from whole blood samples of 40 SCA3/MJD subjects (30 patients and 10 pre-ataxic) and 20 control individuals[30], obtained in the scope of the European spinocerebellar ataxia type 3/Machado-Joseph Initiative (ESMI); and ii) RNA-seq dataset from *post-mortem* cerebellar samples of six SCA3/MJD patients and six controls publicly available from the work of Haas and colleagues.[31] Detailed information about the blood and cerebellum samples from SCA3/MJD subjects (and corresponding healthy controls), is provided in supplementary Table 1. Briefly, RNA extraction and cDNA library preparation was performed according to manufacturer's instructions, as described elsewhere.[30,31] Libraries from blood samples were sequenced as paired end 100bp reads on an Illumina NovaSeq6000 (Illumina), whereas libraries from cerebellum samples were sequenced as paired-end 68 bp reads on a HiSeq2000 (Illumina), both next-generation sequencing (NGS) platforms with limited read length (short-reads).

Ataxin-3 protein levels in plasma (n=28) and in cerebrospinal fluid (CSF, n=8) from a subset of the total 72 SCA3/MJD subjects were available from previous studies.[32,33] Briefly, total full-length ataxin-3 and/or mutant ataxin-3 were measured using a time-resolved fluorescence energy transfer (TR-FRET)-based immunoassay[32] and a single molecule counting (SMC) immunoassay.[33]

The study was approved by local ethics committees of all participating centers; all subjects provided written informed consent.

Bioinformatic analyses of *ATXN3* splicing

The raw bulk RNA-seq paired-end data was first trimmed of adapter and poor-quality sequence using Cutadapt (v. 2.3) and we used STAR[34] as the mapper. The trimmed fasta files were then filtered to include only reads in the *ATXN3* locus. To do this, we leveraged the NCBI standalone BLAST (Basic Local Alignment Search Tool) suite of tools (ncbi-blast-2.11.0), with *ATXN3* as query and the fasta file of reads from each paired-end sample as subject. To avoid spurious matches, the query sequence, Ensembl's *ATXN3* reference transcript ENST00000644486.2 (GRCh38; MANE Select v0.95), was masked for low-complexity sequence, using the "dustmasker" tool in the BLAST suite. The blast command was designed leniently to avoid missing any *ATXN3* reads, especially *ATXN3* fragments on the ends (blastn arguments: "-word_size 20 -evalue 0.1 -outfmt 6 -max_target_seqs 10000"). Setting the word-size to 20 identified reads with a perfect match of at least 20 nt. The results were parsed and curated to remove the (very few) false positives and to ensure one match per read. When only one of a pair of reads matched *ATXN3*, both pairs were kept for downstream analysis. For each set of *ATXN3* reads per sample, de-novo assembling was performed with Trinity (v.2.14.0)[35], a method which integrates the orientation of the paired reads into its analysis. The output is a fasta file with the resulting contigs. There were typically ~20 contigs for a given sample. Trinity provides supplementary scripts which are wrappers for a variety of established methods to quantify the number of reads used to create

the contigs; we used 'bowtie' for alignment and 'RSEM' for quantitation of reads per contig. Transcript levels are shown in transcripts per million (TPM).

Ensembl classification biotype

ATXN3 transcripts were clustered according to the Ensembl classification biotype (Ensembl release 108 – Jan 2023[14]) as: (i) protein coding (a transcript that contains an open reading frame), (ii) nonsense mediated decay (NMD; a transcript with a premature stop codon, predicted as most likely to be subjected to targeted degradation because the in-frame termination codon is found more than 50 bp upstream of the final splice junction), (iii) protein coding sequence (CDS) not defined (a transcript of a protein coding gene for which a CDS has not yet been defined), and (iv) retained intron (a transcript that lacks an open reading frame believed to contain intronic sequence relative to other, coding, transcripts of the same gene) which is included in the major category of long non-coding RNA transcripts. Additionally, using the Ensembl annotation for the *ATXN3* gene[14], *ATXN3* transcripts were clustered according to the presence or absence of the CAG tract, inferred from the presence or absence of exon 10 where the CAG repeat is located.

Statistical analyses

To increase the sample size and the power of statistical tests, *ATXN3* transcripts were labelled according to its frequency by sample. Transcripts not detected in at least half of our set of samples (i.e., 5 or less samples from cerebellum and 29 or less samples for blood) were not included in statistical analyses (Suppl. Table 2, Fig. S1). Transcripts levels were compared between biological groups (pre-ataxic *versus* controls and patients *versus* controls) and between tissues (cerebellum *versus* blood), and statistical differences were determined by the Mann-

Whitney U test. Among the blood samples from controls (n=20), a subset (n=10) was paired-matched by age and sex to pre-ataxic (CTRL-PA) and used for group comparisons (pre-ataxic *versus* CTRL-PA). The significance of the association between absence/presence of transcripts within each tissue by biological group (contingency table) was tested using a Fisher's exact test. In SCA3/MJD subjects' group, the relationship between transcript and protein levels was determined using a Spearman rank correlation test. The ROUT method (Q=1%) was used to exclude outliers previously to correlation test. Statistical analyses were performed in GraphPad Prism version 8.0.1 for Windows (GraphPad Software, San Diego, California USA). The significance level of all tests was set to 5%. Graphic bars are shown as median ± 95% CI (confidence interval).

Availability of data and materials

Data supporting the findings of this study are available in this manuscript and in supplementary material.

RESULTS

The abundance and the frequency of annotated *ATXN3* transcripts are higher in cerebellum than in blood regardless of the disease status

Data from RNA-seq of whole blood (n=60) and *post-mortem* cerebellum (n=12) from SCA3/MJD subjects and age- and sex-matched controls was used to identify the currently annotated 54 *ATXN3* transcripts and quantify their abundance (Fig. 1). All annotated transcripts were identified in our study. According to Ensembl annotation[14], these 54 transcripts can be classified into four main biotypes: protein-coding (n=16), nonsense mediated decay (NMD; n=19), protein CDS not defined (n=16) and retained intron (long noncoding) (n=3) transcripts (Fig. 1). In the 16

protein-coding transcripts group, some structural-related particularities are highlighted (Fig. 1): four transcripts encode full-length ataxin-3 protein isoforms (*ATXN3*-214, *ATXN3*-215, *ATXN3*-227, *ATXN3*-251), including an intact Josephin domain (JD), whereas the remaining 12 transcripts might encode proteins with variations of the N-terminus (*ATXN3*-201, *ATXN3*-203, *ATXN3*-204, *ATXN3*-206, *ATXN3*-207, *ATXN3*-209, *ATXN3*-213, *ATXN3*-219, *ATXN3*-228, *ATXN3*-231, *ATXN3*-235, *ATXN3*-246). Nine transcripts lack exon 11 and, therefore, translate ataxin-3 isoforms only with 2UIM (*ATXN3*-209, *ATXN3*-213, *ATXN3*-214, *ATXN3*-219, *ATXN3*-227, *ATXN3*-228, *ATXN3*-231, *ATXN3*-235, *ATXN3*-246), while seven transcripts retain exon 11 (*ATXN3*-201, *ATXN3*-203, *ATXN3*-204, *ATXN3*-206, *ATXN3*-207, *ATXN3*-215, *ATXN3*-251), and thus encode 3UIMs isoforms. Finally, and although the annotation of the 3' untranslated (UTR) is incomplete, two *ATXN3* transcripts seem to lack exon 10, which contains the polyQ-encoding tract (*ATXN3*-209, *ATXN3*-213) (Fig. 1).

To explore putative tissue-specific patterns of *ATXN3* AS regardless of the disease status, the frequency of each annotated *ATXN3* transcript detected in samples from SCA3/MJD subjects and controls was determined within each tissue (blood or cerebellum). Expression levels of *ATXN3* transcripts ranged from 0.01 (blood) to 50.49 (cerebellum) transcripts per million (TPM). Globally, cerebellum samples displayed higher expression levels than blood samples (Fig. 2; Suppl. Table 2). The biotype with more abundant transcripts in blood and cerebellum is the retained intron and the noncoding, respectively (Fig. 2). Thirty out of the 46 (74%) transcripts identified in the cerebellum and 44 out of the 52 (85%) transcripts expressed in the blood could only be detected in less than half of samples (Fig. S1; Suppl. Table 2); for example, the protein-coding transcript *ATXN3*-204 was only expressed in one out of 12 cerebellum samples (8%) and in six out of 60 blood samples (10%) (Suppl. Table 2).

The protein coding *ATXN3*-251 reference transcript and the protein CDS not defined *ATXN3*-208 transcript showed, respectively, the highest expression in blood and in cerebellum. Additionally, the transcripts more abundant in blood by biotype were the protein coding *ATXN3*-251, the

NMD *ATXN3*-202, the noncoding *ATXN3*-208 and the retained intron *ATXN3*-210, while in cerebellum the more abundant were the protein coding *ATXN3*-214, the NMD *ATXN3*-234, the noncoding *ATXN3*-208 and the retained intron *ATXN3*-205 (Fig. 2).

Eight of the overall 54 annotated transcripts were exclusively expressed in blood: (i) two corresponded to protein-coding transcripts (*ATXN3*-215, which is one of the four transcripts encoding the full-length ataxin-3 protein isoforms, and *ATXN3*-228); (ii) four were transcripts targeted for NMD (*ATXN3*-212, *ATXN3*-218, *ATXN3*-221, *ATXN3*-236) and (iii) two were protein CDS not defined (*ATXN3*-220, *ATXN3*-233). *ATXN3*-212 transcript showed the highest expression among all transcripts detected in the blood; (Fig. 2, Suppl. Table 2, and Fig. S1). Two NMD transcripts were exclusively expressed in the cerebellum (*ATXN3*-242, *ATXN3*-244) (Suppl. Table 2, Fig. S1).

Cerebellar abundance of *ATXN3* transcripts is similar in SCA3 patients and controls

Analysis of the distribution of *ATXN3* transcripts by biological group (SCA3/MJD patients and/or controls) in cerebellum samples revealed that several transcripts were detected only in individual or in a limited number of samples (Suppl. Table 2).

To compare distributions of cerebellar *ATXN3* transcripts between SCA3/MJD patients and controls, transcripts detected in more than 50% of all cerebellum samples (six or more of 12 samples) were selected. Following this criterium, 16 transcripts were identified in cerebellum samples corresponding to seven protein-coding (the reference variant - *ATXN3*-251, plus *ATXN3*-201, *ATXN3*-206, *ATXN3*-207, *ATXN3*-214, *ATXN3*-219, *ATXN3*-227), four targeted for NMD (*ATXN3*-234, *ATXN3*-243, *ATXN3*-253, *ATXN3*-254), three protein CDS not defined (*ATXN3*-208, *ATXN3*-217, *ATXN3*-226) and two retained intron (*ATXN3*-205, *ATXN3*-210; Fig. 2). While all 16 transcripts showed similar abundance in the cerebellum of SCA3/MJD patients and controls ($p>0.05$; Fig. 3), the protein CDS not defined transcript *ATXN3*-226 was expressed in all samples

from controls but was not expressed in four out of the six patients (Fisher's exact test, p=0.06; Suppl. Table 2).

Blood abundance of *ATXN3* transcripts is similar in SCA3/MJD subjects and controls.

Analysis of the distribution of *ATXN3* transcripts in blood samples from SCA3/MJD subjects (pre-ataxic and patients) and controls revealed that, similarly to the cerebellum, several transcripts were only detected in individual or in a limited number of samples (Suppl. Table 2).

Following the same criterium used for the cerebellum, eight *ATXN3* transcripts detected in more than 50% of all blood samples (SCA3/MJD subjects and controls) were selected for comparative analyses of patients versus controls and pre-ataxic carriers versus controls. These eight transcripts corresponded to two protein coding (*ATXN3*-214 and the reference transcript - *ATXN3*-251), two transcripts targeted for NMD (*ATXN3*-202, *ATXN3*-254), one protein CDS not defined (*ATXN3*-249), and three retained introns (*ATXN3*-210, *ATXN3*-205, *ATXN3*-248) (Fig. 2). All eight transcripts showed similar levels between SCA3/MJD patients or pre-ataxic carriers compared to their respective controls (Fig. 4).

The amount of *ATXN3*-205 transcript in blood is directly associated with CSF levels of soluble mutant ataxin-3.

The relationship between *ATXN3* transcript levels from blood cells and soluble levels of ataxin-3 protein (mutant and/or total) quantified in plasma, peripheral mononuclear blood cells (PBMCs) and CSF samples from the same SCA3/MJD subjects was explored in this study (Fig. S2a). Interestingly, levels of the retained intron transcript *ATXN3*-205 in blood cells directly correlated with the abundance of soluble mutant ataxin-3 in the CSF of SCA3/MJD subjects (p=0.031, Fig. S2c). Although not reaching statistical significance, high levels of *ATXN3*-248 and of the reference transcript *ATXN3*-251 displayed a trend to subtly associate with low levels of total ataxin-3 measured in PBMCs of SCA3/MJD individuals (Fig. S2b, d).

Expression of the four transcripts encoding full-length ataxin-3 2UIM and 3UIM protein isoforms are tissue-specific, irrespective of disease status.

From the 16 protein-coding transcripts, only four encode a full-length ataxin-3 protein with an intact JD and a preserved C-terminus (*ATXN3*-214, *ATXN3*-215, *ATXN3*-227 and *ATXN3*-251; Fig. 5a). *ATXN3*-214 and *ATXN3*-227 transcripts encode ataxin-3 2UIM isoforms, whereas *ATXN3*-215 and *ATXN3*-251 transcripts encode for ataxin-3 3UIM isoforms (Fig. 5a). To elucidate whether the individual expression levels of the four transcripts encoding full-length ataxin-3 2UIM and 3UIM isoforms contribute to the tissue-specific vulnerability observed in SCA3/MJD, the abundance of these transcripts in blood and cerebellum was compared between SCA3/MJD patients and controls (Fig. 5b).

Interestingly, two of these four transcripts were only expressed in some samples. The *ATXN3*-215 transcript (3UIM) was only identified in one blood sample (Supp. Table 2); while the *ATXN3*-227 transcript (2UIM) was detected at distinct frequencies in both tissues, it was only measured in three out of 60 (5%) blood samples and in seven out of 12 (58%) in cerebellum samples (Supp. Table 2; Fig. 3). Although being overall commonly detected in the cerebellum, *ATXN3*-227 transcript showed similar levels and frequency in SCA3/MJD patients and controls (Fig. 3).

The reference transcript, *ATXN3*-251 (3UIM), showed a ~50-fold increase of abundance in blood samples over cerebellum samples, whereas the *ATXN3*-214 (2UIM) transcript was more expressed in the cerebellum about 20-fold more than in blood (Fig. 5b). The magnitude of the fold change values, however, could be partially reflecting technical differences due to different RNA-seq experimental setups. Remarkably, in cerebellum samples, the *ATXN3*-214 (2UIM) was expressed 265-fold more than the *ATXN3*-251 (3UIM) transcript, whereas in blood samples the *ATXN3*-251 (3UIM) transcript was expressed 4-fold more than *ATXN3*-214 (2UIM). These observed differences in transcript abundances between tissues were similar in SCA3/MJD subjects and controls.

DISCUSSION

Here, we describe the diversity and abundance of annotated *ATXN3* transcripts detected in blood and cerebellum samples from SCA3/MJD subjects and controls. Globally, the number of different transcripts and the abundance of transcripts were higher in the cerebellum than in blood, both in SCA3/MJD subjects and controls. Interestingly, in SCA3/MJD subjects and controls pooled samples altogether, while the most abundant transcript in the cerebellum (*ATXN3*-208) is categorized as a "protein coding with a CDS not defined of unknown function transcript" by Ensembl, the transcript showing the highest abundance in the blood is the protein-coding *ATXN3* reference transcript (*ATXN3*-251) that is translated into an ataxin-3 3UIM protein isoform. No differences in the abundance of the most frequent transcripts were found in blood or cerebellum samples of SCA3/MJD subjects and controls. Noteworthy, the abundance of *ATXN3*-251 and *ATXN3*-214 transcripts, two out of the four transcripts that encode full-length ataxin-3 protein isoforms but differ in the C-terminus (2UIM *versus* 3UIM) were strongly associated with tissue expression specificity: *ATXN3-251* (3UIM) transcript was expressed in blood 50-fold more than in cerebellum, whereas *ATXN3*-214 (2UIM) transcript was expressed in the cerebellum 20-fold more than in blood.

Studies using a *Drosophila* model of SCA3/MJD indicate that in addition to a gain-of-function of the mutant ataxin-3 protein, *ATXN3* transcripts harbouring expanded CAG repeats show independent and progressive cellular toxicity.[36] However, whether such toxic transcript action occurs in human cells remains to be elucidated. Hence, the characterization of *ATXN3* AS in mutation carriers of SCA3/MJD might provide fundamental knowledge to unveil the contribution of such mechanism in SCA3/MJD pathogenesis.

Globally, AS shows its highest complexity (high diversity of transcripts), in the brain amongst other tissues.[37] As expected, *ATXN3* transcripts diversity and abundance (a greater number of transcripts, which are expressed in a larger number of samples with the highest levels of

expression) was higher in the cerebellum than in the blood and might indicate an important role of *ATXN3* AS and its regulation in the neurodegenerative process of SCA3/MJD.

Twelve out of 16 protein-coding transcripts are predicted to encode proteins with truncations in the JD. Such modifications imply that several ataxin-3 protein forms with a variable N-terminus might exist, displaying for example different forms of the globular JD or other N-terminal domains that show cysteine protease activity or even another yet unknown activity. The fact that ataxin-3 interacts with so many proteins in cells (reviewed in [19]) and is ubiquitously expressed (*e.g.,* [38]) together with our data indicates that several ataxin-3 isoforms might exist with distinct functions other than the currently known deubiquitinating enzyme role. Interestingly, four out of these 12 transcripts (*ATXN3*-201, *ATXN3*-206, *ATXN3*-207, *ATXN3*-219) are highly expressed in the cerebellum compared to blood suggesting that these potential encoded proteins may play specific roles in the cerebellum and that the correspondent mutant forms could potentially contribute for the observed selective cerebellar toxicity observed in SCA3/MJD patients.

Four out of 54 transcripts, *ATXN3*-214, *ATXN3*-215, *ATXN3*-227 and *ATXN3*-251, encode full-length ataxin-3 protein isoforms, which differ only in the C-terminus. In the present study, the *ATXN3*-214 and the *ATXN3*-227 transcripts, both encoding 2UIM isoforms, are preferentially abundant in cerebellum samples, whereas the *ATXN3*-251 transcript (3UIM) is more expressed in blood. Altogether, these findings suggest that ataxin-3 isoforms with 2UIM might have a unique role in the cerebellum and consequently in neurodegeneration. The correlation between transcript and protein abundance is still not known and future studies exploring the role of the *ATXN3*-214 and *ATXN3*-227 transcripts and respective protein isoforms in brain tissues could contribute for the elucidation of mechanisms involved in cellular vulnerability to mutant *ATXN3* gene products in SCA3/MJD.

Of note, a significant direct relationship between the abundance of the retained intron *ATXN3*-205 transcript in blood and soluble mutant ataxin-3 in CSF samples was observed in a small sub-set of SCA3/MJD subjects. Because *ATXN3*-205 is classified by Ensembl as a long non-coding RNA, such finding suggests that this transcript might have a regulatory function in the transcription and/or translation of protein coding transcripts in the nervous system and, therefore, contribute to the observed correlation with the soluble mutant ataxin-3 protein in the CSF. More comprehensive studies about the function of these type of transcripts are needed as well as larger number of CSF samples should be further analyzed to better explore this preliminary but important finding.

In this study, transcripts more frequently found in blood (n=8) and in cerebellum (cerebellar lobules, n=16) showed similar levels in SCA3/MJD patients and controls. Mort and colleagues also described similar amount of two *HTT* transcripts in cerebellum samples of HD patients and controls.[39] In SCA3/MJD, neuronal loss in the cerebellar lobules is not as severe as, for example, in the dentate nucleus or pons.[20] Whether our findings are specific of this brain area is unknown and, thus, characterization of *ATXN3* transcripts in other SCA3/MJD patients brain tissues showing more severe lesions would be important to conduct in the future.

In this work, the diversity and abundance of annotated *ATXN3* transcripts in SCA3/MJD patients and controls is provided. The methodology used in this study did not allow to specifically analyze the *ATXN3* transcripts with CAG repeat expansions. Future studies using other NGS approaches, including novel bioinformatic pipelines to correctly detect CAG expansions are needed to elucidate the contribution of the expanded CAG repeat itself in *ATXN3* AS. Moreover, functional studies to define the role of the most promising transcripts found in the cerebellum, for example for its most abundant known non-coding *ATXN3*-208 transcript, would be important to conduct in the future. In fact, the majority of annotated *ATXN3* transcripts might have distinct functions

or regulatory properties and its comprehension will reveal novel mechanisms of SCA3/MJD pathogenesis.

We show that *ATXN3* AS is complex and, thus, its elucidation in the SCA3/MJD-related tissues should become an integral part of future design of *ATXN3* transcripts-targeted therapeutics to overcome potential limitations of efficacy of current RNA-based therapies targeting a limited number of *ATXN3* transcripts. Our findings further aid to the clarification of the specific tissue vulnerability observed in SCA3/MJD and, therefore, may help unravelling novel mechanisms involved in SCA3/MJD pathogenesis.

ACKNOWLEDGEMENTS

The ESMI consortium would like to thank Nina Roy for coordination and managing of the project. We gratefully thank Dr. Aires Raposo for the collaboration on blood collection in Azores islands.

Additional members of the European Spinocerebellar Ataxia type 3/Machado-Joseph Initiative (ESMI) Study Group are listed in Supplementary Material (Table S3) and participated in subject recruitment and acquisition of participants data as well as read and approved the final manuscript.

AUTHORS' ROLES

Design and conceptualization of the study: MR, JH-S, RT, MCC, ML; Subject recruitment/Acquisition of participants biomaterials/data: MR, AFF, ARVM, JV, PP, TK, HG-M, PG, MMS, LPA, JI, BPW, JV, JF, TK, NC, JA, LS, OR, ML; Bioinformatic/Statistical analysis of data: MR, RT; Drafting of the manuscript: MR, JH-S, MCC, ML; Revision of the Manuscript: MR, JH-S,

RT, AFF, ARVM, JV, PP, TK, HG-M, PG, MMS, LPA, JI, BPW, JV, JF, TK, NC, JA, LS, OR, MCC, ML. All authors read and approved the final manuscript.

REFERENCES

1. Vuong CK, Black DL, Zheng S. The neurogenetics of alternative splicing. Nat Rev Neurosci [Internet]. 2016 May 20;17(5):265–81. Available from: https://www.nature.com/articles/nrn.2016.27

2. Daguenet E, Dujardin G, Valcárcel J. The pathogenicity of splicing defects: mechanistic insights into pre- <scp>mRNA</scp> processing inform novel therapeutic approaches. EMBO Rep [Internet]. 2015 Dec 13;16(12):1640–55. Available from: https://onlinelibrary.wiley.com/doi/10.15252/embr.201541116

3. Nik S, Bowman T V. Splicing and neurodegeneration: Insights and mechanisms. WIREs RNA [Internet]. 2019 Jul 20;10(4). Available from: https://onlinelibrary.wiley.com/doi/10.1002/wrna.1532

4. Mangiarini L, Sathasivam K, Seller M, Cozens B, Harper A, Hetherington C, et al. Exon 1 of the HD Gene with an Expanded CAG Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice. Cell [Internet]. 1996 Nov;87(3):493–506. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867400813690

5. Sathasivam K, Neueder A, Gipson TA, Landles C, Benjamin AC, Bondulich MK, et al. Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. Proc Natl Acad Sci [Internet]. 2013 Feb 5;110(6):2366–70. Available from: https://pnas.org/doi/full/10.1073/pnas.1221891110

6. Barbaro BA, Lukacsovich T, Agrawal N, Burke J, Bornemann DJ, Purcell JM, et al. Comparative study of naturally occurring huntingtin fragments in Drosophila points to exon 1 as the most pathogenic species in Huntington's disease. Hum Mol Genet [Internet]. 2015 Feb 15;24(4):913–25. Available from: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddu504

7. Neueder A, Landles C, Ghosh R, Howland D, Myers RH, Faull RLM, et al. The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. Sci Rep [Internet]. 2017 May 2;7(1):1307. Available from: https://www.nature.com/articles/s41598-017-01510-z

8. Dhuri K, Bechtold C, Quijano E, Pham H, Gupta A, Vikram A, et al. Antisense Oligonucleotides: An Emerging Area in Drug Discovery and Development. J Clin Med [Internet]. 2020 Jun 26;9(6):2004. Available from: https://www.mdpi.com/2077-0383/9/6/2004

9. Zhu Y, Zhu L, Wang X, Jin H. RNA-based therapeutics: an overview and prospectus. Cell Death Dis [Internet]. 2022 Jul 23;13(7):644. Available from: https://www.nature.com/articles/s41419-022-05075-2

10. Sequeiros J, Martins S, Silveira I. Epidemiology and population genetics of degenerative

ataxias. In 2012. p. 227–51. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780444518927000140

11. Takiyama Y, Nishizawa M, Tanaka H, Kawashima S, Sakamoto H, Karube Y, et al. The gene for Machado-Joseph disease maps to human chromosome 14q. Nat Genet [Internet]. 1993 Jul [cited 2012 Jun 4];4(3):300–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8358439

12. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet [Internet]. 1994 Nov;8(3):221–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7874163

13. Maciel P, Costa MC, Ferro A, Rousseau M, Santos CS, Gaspar C, et al. Improvement in the molecular diagnosis of Machado-Joseph disease. Arch Neurol. 2001;58(11):1821–7.

14. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res [Internet]. 2023 Jan 6;51(D1):D933–41. Available from: https://academic.oup.com/nar/article/51/D1/D933/6786199

15. Ichikawa Y, Goto J, Hattori M, Toyoda A, Ishii K, Jeong SY, et al. The genomic structure and expression of MJD, the Machado-Joseph disease gene. J Hum Genet [Internet]. 2001 Jan;46(7):413–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11450850

16. Burnett B, Li F, Pittman RN. The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity. Hum Mol Genet [Internet]. 2003 Dec 1;12(23):3195–205. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14559776

17. Goto J, Watanabe M, Ichikawa Y, Yee SB, Ihara N, Endo K, et al. Machado–Joseph disease gene products carrying different carboxyl termini. Neurosci Res [Internet]. 1997 Aug;28(4):373–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168010297000564

18. Masino L, Musi V, Menon RP, Fusi P, Kelly G, Frenkiel TA, et al. Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail. FEBS Lett [Internet]. 2003 Aug 14;549(1–3):21–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12914917

19. Costa M do C, Paulson HL. Toward understanding Machado-Joseph disease. Vol. 97, Progress in Neurobiology. 2012. p. 239–57.

20. Koeppen AH. The Neuropathology of Spinocerebellar Ataxia Type 3/Machado-Joseph Disease. In 2018. p. 233–41. Available from: http://link.springer.com/10.1007/978-3-319-71779-1_11

21. Harris GM, Dodelzon K, Gong L, Gonzalez-Alegre P, Paulson HL. Splice isoforms of the polyglutamine disease protein ataxin-3 exhibit similar enzymatic yet different aggregation properties. PLoS One. 2010;5(10).

22. Bettencourt C, Santos C, Montiel R, Costa M do C, Cruz-Morales P, Santos LR, et al. Increased transcript diversity: novel splicing variants of Machado–Joseph Disease gene (ATXN3). Neurogenetics [Internet]. 2010 May 28;11(2):193–202. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19714377

23. Johnson SL, Blount JR, Libohova K, Ranxhi B, Paulson HL, Tsou WL, et al. Differential toxicity of ataxin-3 isoforms in Drosophila models of Spinocerebellar Ataxia Type 3. Neurobiol Dis [Internet]. 2019 Dec;132:104535. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0969996119302037

24. Weishäupl D, Schneider J, Peixoto Pinheiro B, Ruess C, Dold SM, von Zweydorf F, et al. Physiological and pathophysiological characteristics of ataxin-3 isoforms. J Biol Chem [Internet]. 2019 Jan;294(2):644–61. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925820403783

25. Helm J, Schöls L, Hauser S. Towards Personalized Allele-Specific Antisense Oligonucleotide Therapies for Toxic Gain-of-Function Neurodegenerative Diseases. Pharmaceutics [Internet]. 2022 Aug 16;14(8):1708. Available from: https://www.mdpi.com/1999-4923/14/8/1708

26. McLoughlin HS, Moore LR, Chopra R, Komlo R, McKenzie M, Blumenstein KG, et al. Oligonucleotide therapy mitigates disease in spinocerebellar ataxia type 3 mice. Ann Neurol [Internet]. 2018 Jul;84(1):64–77. Available from: https://onlinelibrary.wiley.com/doi/10.1002/ana.25264

27. Moore LR, Keller L, Bushart DD, Delatorre RG, Li D, McLoughlin HS, et al. Antisense oligonucleotide therapy rescues aggresome formation in a novel spinocerebellar ataxia type 3 human embryonic stem cell line. Stem Cell Res [Internet]. 2019 Aug;39:101504. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1873506119301345

28. Moore LR, Rajpal G, Dillingham IT, Qutob M, Blumenstein KG, Gattis D, et al. Evaluation of Antisense Oligonucleotides Targeting ATXN3 in SCA3 Mouse Models. Mol Ther - Nucleic Acids [Internet]. 2017 Jun;7:200–10. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2162253117301555

29. Hauser S, Helm J, Kraft M, Korneck M, Hübener-Schmid J, Schöls L. Allele-specific targeting of mutant ataxin-3 by antisense oligonucleotides in SCA3-iPSC-derived neurons. Mol Ther - Nucleic Acids [Internet]. 2022 Mar;27:99–108. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2162253121002900

30. Mafalda Raposo, Jeannette Hübener-Schmid, Ana F. Ferreira, Ana Rosa Vieira Melo, João Vasconcelos, Paula Pires, Teresa Kay, Hector Garcia-Moreno, Paola Giunti, Magda M. Santana, Luis Pereira de Almeida, Jon Infante, Bart P. van de Warrenburg et al. Blood transcriptome sequencing identifies biomarkers able to track disease stages in spinocerebellar ataxia type 3. Brain. *Accepted*

31. Haas E, Incebacak RD, Hentrich T, Huridou C, Schmidt T, Casadei N, et al. A Novel SCA3 Knock-in Mouse Model Mimics the Human SCA3 Disease Phenotype Including Neuropathological, Behavioral, and Transcriptional Abnormalities Especially in Oligodendrocytes. Mol Neurobiol [Internet]. 2022 Jan 30;59(1):495–522. Available from: https://link.springer.com/10.1007/s12035-021-02610-8

32. Gonsior K, Kaucher GA, Pelz P, Schumann D, Gansel M, Kuhs S, et al. PolyQ-expanded ataxin-3 protein levels in peripheral blood mononuclear cells correlate with clinical parameters in SCA3: a pilot study. J Neurol [Internet]. 2021 Apr 26;268(4):1304–15. Available from: https://link.springer.com/10.1007/s00415-020-10274-y

33. Hübener-Schmid J, Kuhlbrodt K, Peladan J, Faber J, Santana MM, Hengel H, et al. Polyglutamine-Expanded Ataxin-3: A Target Engagement Marker for Spinocerebellar

Ataxia Type 3 in Peripheral Blood. Mov Disord [Internet]. 2021 Nov 16;36(11):2675–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/34397117

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [Internet]. 2013 Jan;29(1):15–21. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635

35. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc [Internet]. 2013 Aug 11;8(8):1494–512. Available from: https://www.nature.com/articles/nprot.2013.084

36. Li LB, Yu Z, Teng X, Bonini NM. RNA toxicity is a component of ataxin-3 degeneration in Drosophila. Nature [Internet]. 2008 Jun 30;453(7198):1107–11. Available from: http://www.nature.com/articles/nature06909

37. Porter RS, Jaamour F, Iwase S. Neuron-specific alternative splicing of transcriptional machineries: Implications for neurodevelopmental disorders. Mol Cell Neurosci [Internet]. 2018 Mar;87:35–45. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1044743117302373

38. Paulson HL, Das SS, Crino PB, Perez MK, Patel SC, Gotsdiner D, et al. Machado-Joseph disease gene product is a cytoplasmic protein widely expressed in brain. Ann Neurol [Internet]. 1997 Apr;41(4):453–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9124802

39. Mort M, Carlisle FA, Waite AJ, Elliston L, Allen ND, Jones L, et al. Huntingtin Exists as Multiple Splice Forms in Human Brain. J Huntingtons Dis [Internet]. 2015 Jul 1;4(2):161–71. Available from: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JHD-150151

LEGENDS OF FIGURES

Figure 1. Schematic representation of the annotated 54 *ATXN3* transcripts (adapted from Ensembl release 108 – Jan 2023) detected in blood and cerebellum of pooled SCA3/MJD patients and controls. *ATXN3* transcripts in both SCA3/MJD subjects and controls are clustered according to Ensembl-related biotype - protein coding, nonsense mediated decay, protein coding sequence (CDS) not defined, and retained intron. Moreover, transcripts are clustered by descending order of frequency within tissue calculated in the present study as well as transcripts with exclusive expression in blood and transcripts with exclusive expression in cerebellum are highlighted in orange and green, respectively. The Ensembl's *ATXN3* reference transcript ENST00000644486.2|*ATXN3*-251 (MANE Select v0.95) is highlighted in blue. Protein-coding transcripts not containing exon 10 (which contains the CAG tract) are highlighted in bold. *The annotation of the transcript is incomplete at 5' untranslated (UTR), 3'UTR or both. The catalytic amino acids cysteine (C) 14, histidine (H) 119 and asparagine (N) 134 are shown in the respective exons; red diamonds represent a premature termination codon (PTC). Grey box = coding exon; light grey box = UTR; white box = non-coding exon; UIM = ubiquitin interacting motif; polyQ = polyglutamine tract; NES = nuclear export signal; NLS = nuclear localization signal. The size of exons/introns/UTR are not drawn to scale.

Figure 2. Heatmap of the 54 annotated *ATXN3* transcripts detected in cerebellum and blood samples of SCA3/MJD subjects and healthy individuals. Transcripts are clustered by biotype (protein coding, nonsense mediated decay, protein sequence (CDS) not defined, and retained intron) and disease/health status. Expression values, shown in transcripts per million (TPM), were obtained for 72 samples (12 from cerebellum - lines one to 12) and 60 from whole blood (lines 13 to 72). The magnitude of the expression values was set independently for each transcript type cluster. Protein-coding transcripts not containing exon 10 (which contains the CAG tract) are highlighted in bold. The brain (🧠) and/or blood (🩸) symbols indicate transcripts

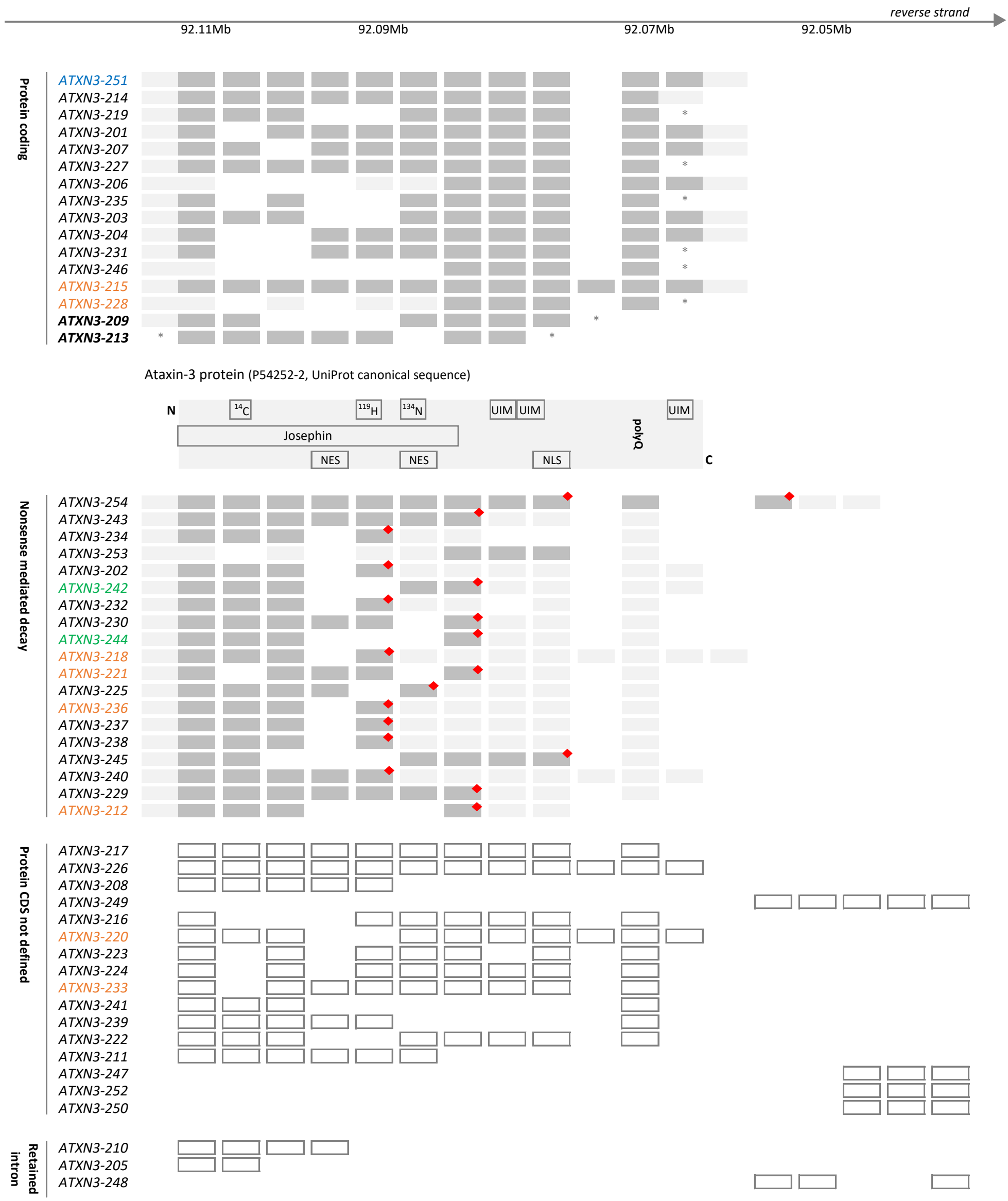
detected in more than 50% of all samples in each tissue. *ATXN3* transcripts with exclusive expression in blood and transcripts with exclusive expression in cerebellum are highlighted in orange and green, respectively. The Ensembl's *ATXN3* reference transcript ENST00000644486.2|*ATXN3*-251 (MANE Select v0.95) is highlighted in blue.

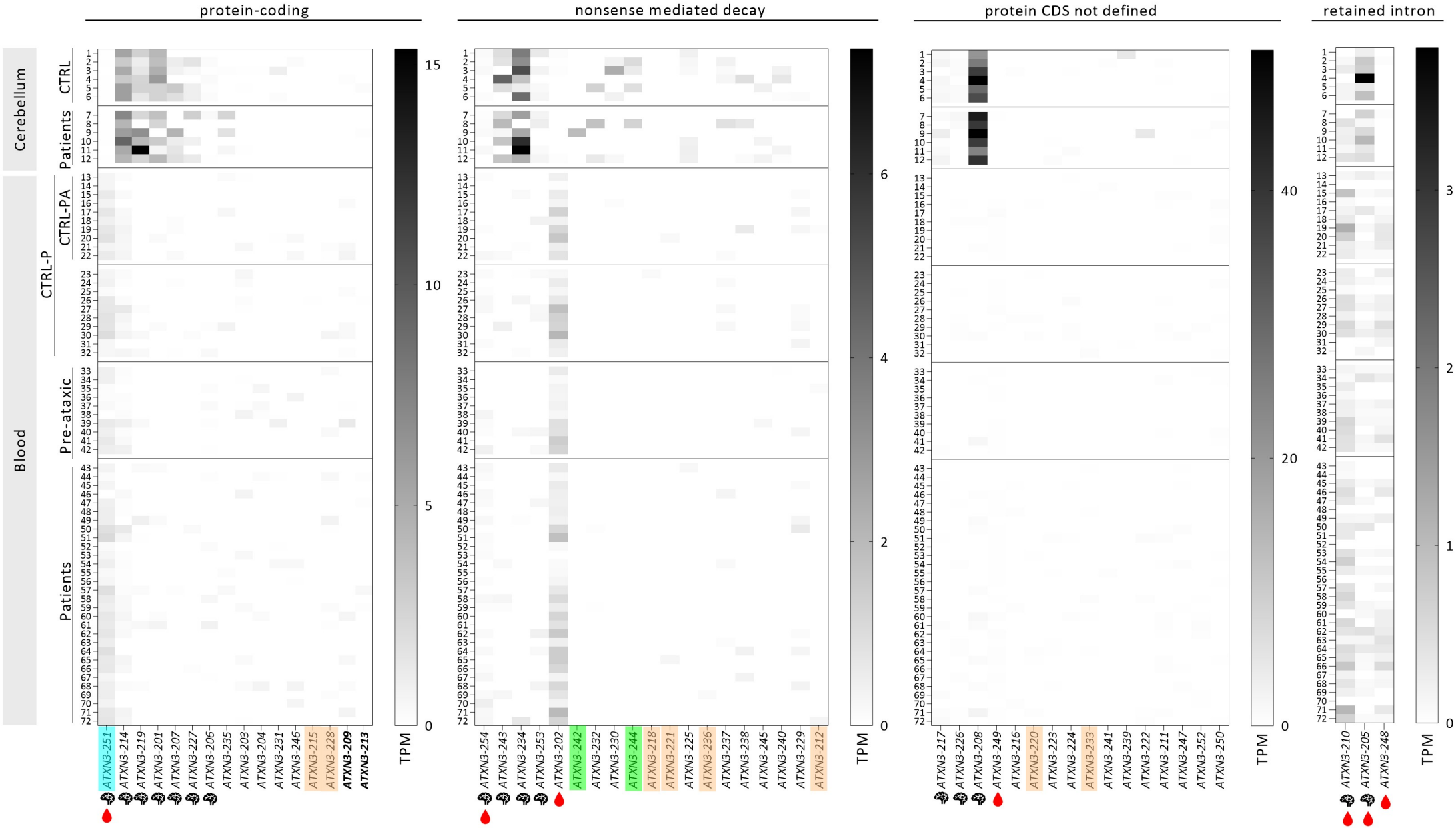
Figure 3. Levels of the 16 *ATXN3* transcripts expressed in more than half of all cerebellum samples from SCA3/MJD patients (n=6) and controls (n=6). The transcripts are clustered according to the type of transcript, as defined by Ensembl classification biotype (Ensembl release 108 - Jan 2023, REF). Transcript expression levels are shown in transcript per million (TPM). Bars in the graphs represent the median ± 95% CI (confidence interval, error bars).

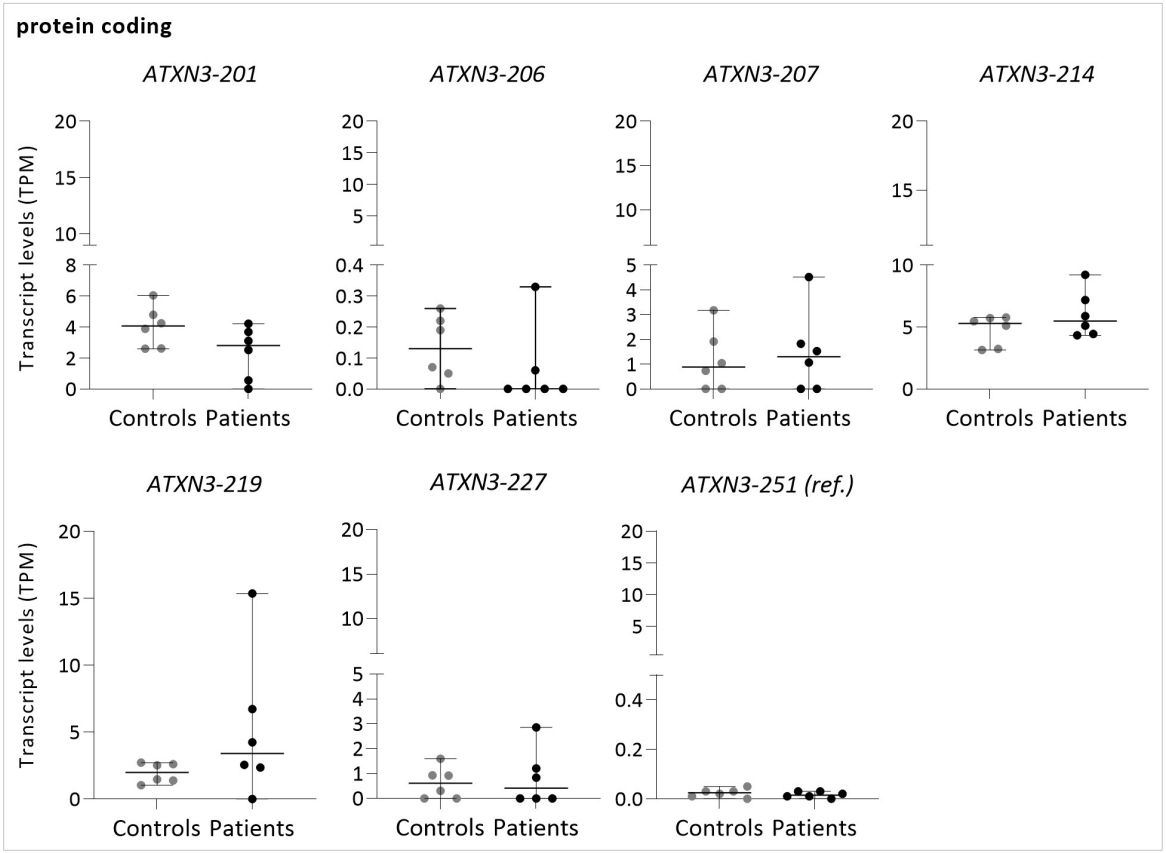
Figure 4. Expression levels of eight *ATXN3* transcripts in blood samples from pre-ataxic (PA) carriers (n=10), SCA3/MJD patients (n=30) and respective controls (CTRL-P, n=20 and CTRL-PA, n=10). CTRL-PA was defined as a subset of age-matched controls with pre-ataxic carriers within the CTRL-P group. The transcripts are clustered according to biotype, as defined by Ensembl classification biotype (Ensembl release 108 - Jan 2023[14]). Expression levels are shown in transcript per millions (TPM). Bars in the graphs represent the median ± 95% CI (confidence interval, error bars).

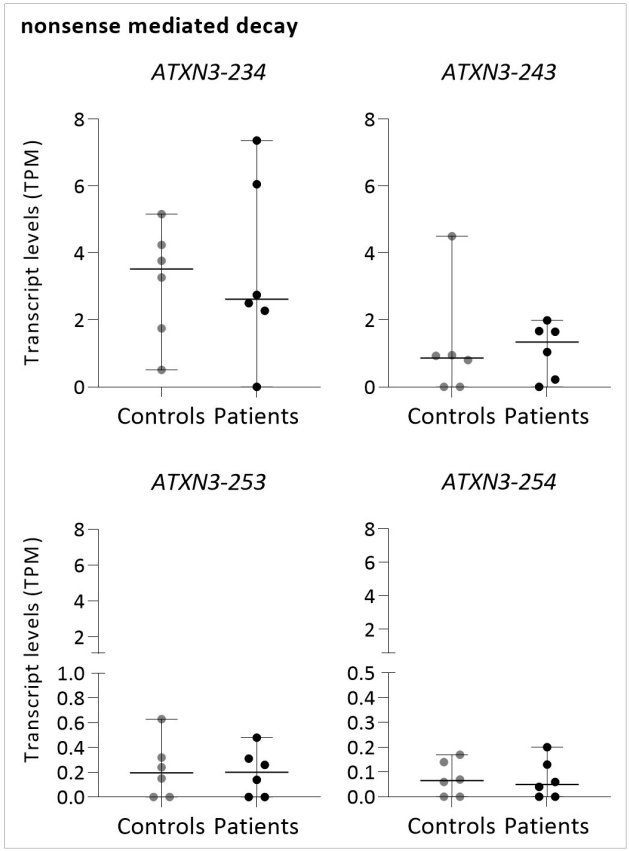
Figure 5. (a) Sequence alignment of the four full-length ataxin-3 protein isoforms harbouring an intact Josephin domain (JD) and a preserved C-terminus. Ataxin-3 protein isoforms, namely the P54252-2, F5H211, P54252-1 and G3V3R7 are encoded by the *ATXN3*-251, *ATXN3*-215, *ATXN3*-214 and *ATXN3*-227 transcripts, respectively. Multiple sequence alignment was performed using Clustal Omega program (Sievers F. et al., 2011). JD (green line), the three catalytic sites (orange dot) and the three UIMs (blue boxes) are highlighted. (b) Levels of *ATXN3*-251 (3UIM) and of the *ATXN3*-214 as well as *ATXN3*-227 (both 2UIM) transcripts in the blood and cerebellum samples

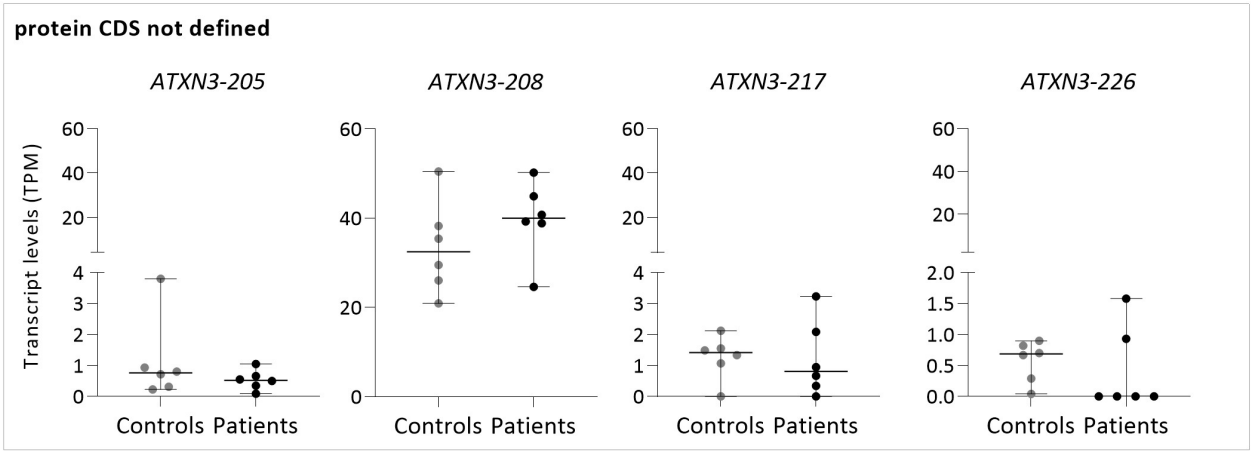
from SCA3/MJD subjects and controls. Transcript levels are shown in transcripts per million (TPM). Bars in the graphs represent the median ± 95% CI (confidence interval, error bars).

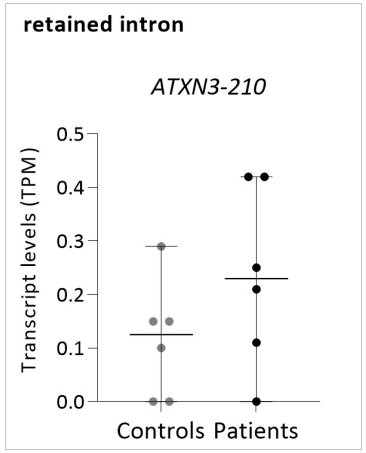

reverse strand
92.11Mb
92.09Mb
92.07Mb
92.05Mb
Protein coding
ATXN3-251
ATXN3-214
ATXN3-219
ATXN3-201
ATXN3-207
ATXN3-227
ATXN3-206
ATXN3-235
ATXN3-203
ATXN3-204
ATXN3-231
ATXN3-246
ATXN3-215
ATXN3-228
ATXN3-209
ATXN3-213
Ataxin-3 protein (P54252-2, UniProt canonical sequence)
N
14C
119H
134N
UIM
UIM
UIM
Josephin
polyQ
NES
NES
NLS
C
Nonsense mediated decay
ATXN3-254
ATXN3-243
ATXN3-234
ATXN3-253
ATXN3-202
ATXN3-242
ATXN3-232
ATXN3-230
ATXN3-244
ATXN3-218
ATXN3-221
ATXN3-225
ATXN3-236
ATXN3-237
ATXN3-238
ATXN3-245
ATXN3-240
ATXN3-229
ATXN3-212
Protein CDS not defined
ATXN3-217
ATXN3-226
ATXN3-208
ATXN3-249
ATXN3-216
ATXN3-220
ATXN3-223
ATXN3-224
ATXN3-233
ATXN3-241
ATXN3-239
ATXN3-222
ATXN3-211
ATXN3-247
ATXN3-252
ATXN3-250
Retained intron
ATXN3-210
ATXN3-205
ATXN3-248

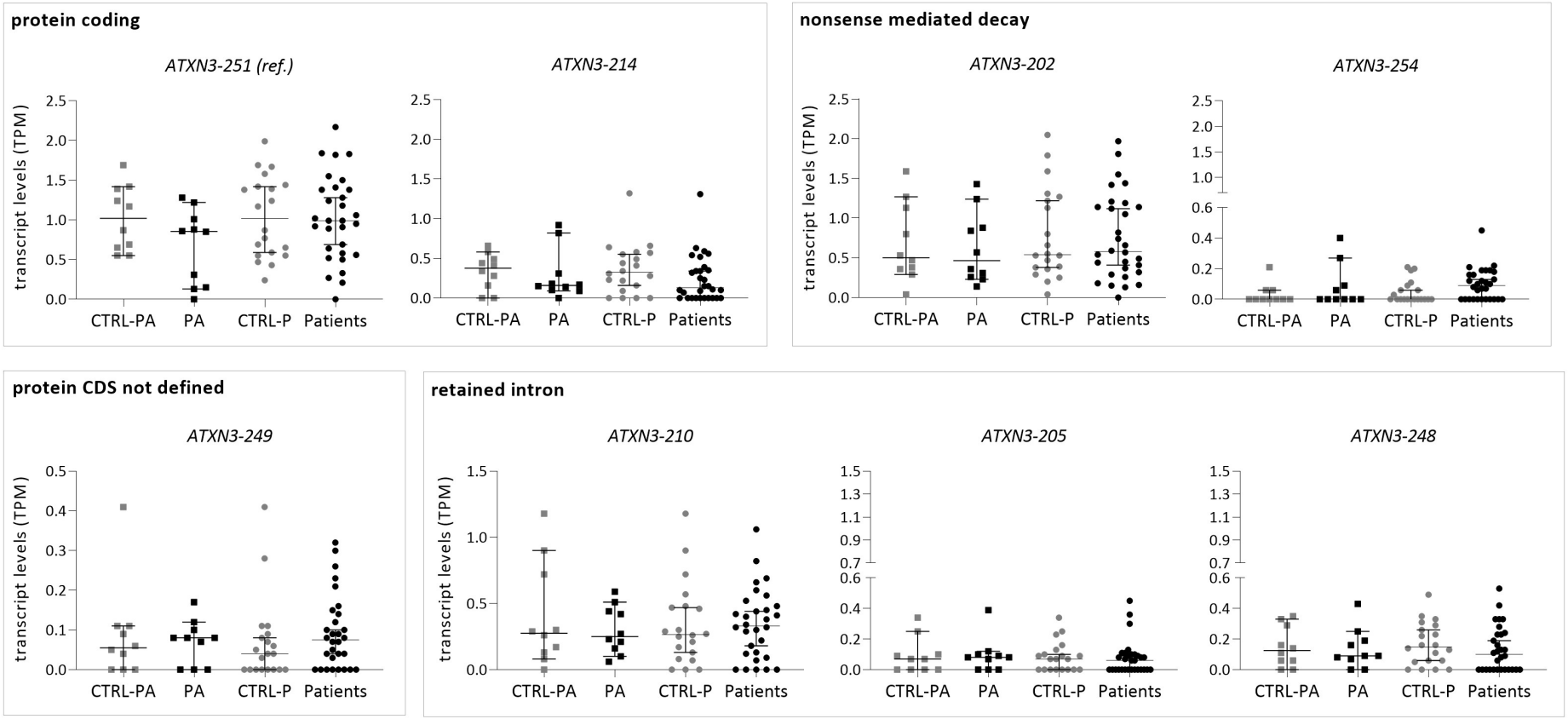

protein-coding
nonsense mediated decay
protein CDS not defined
retained intron
Cerebellum
Blood
CTRL
Patients
CTRL-P
CTRL-PA
Pre-ataxic
Patients
TPM
ATXN3-251
ATXN3-214
ATXN3-219
ATXN3-201
ATXN3-207
ATXN3-227
ATXN3-206
ATXN3-235
ATXN3-203
ATXN3-204
ATXN3-231
ATXN3-246
ATXN3-215
ATXN3-228
ATXN3-209
ATXN3-213
ATXN3-254
ATXN3-243
ATXN3-234
ATXN3-253
ATXN3-202
ATXN3-242
ATXN3-232
ATXN3-230
ATXN3-244
ATXN3-218
ATXN3-221
ATXN3-225
ATXN3-236
ATXN3-237
ATXN3-238
ATXN3-245
ATXN3-240
ATXN3-229
ATXN3-212
ATXN3-217
ATXN3-226
ATXN3-208
ATXN3-249
ATXN3-216
ATXN3-220
ATXN3-223
ATXN3-224
ATXN3-233
ATXN3-241
ATXN3-239
ATXN3-222
ATXN3-211
ATXN3-247
ATXN3-252
ATXN3-250
ATXN3-210
ATXN3-205
ATXN3-248


protein coding
ATXN3-201
ATXN3-206
ATXN3-207
ATXN3-214
ATXN3-219
ATXN3-227
ATXN3-251 (ref.)
Transcript levels (TPM)
Controls
Patients

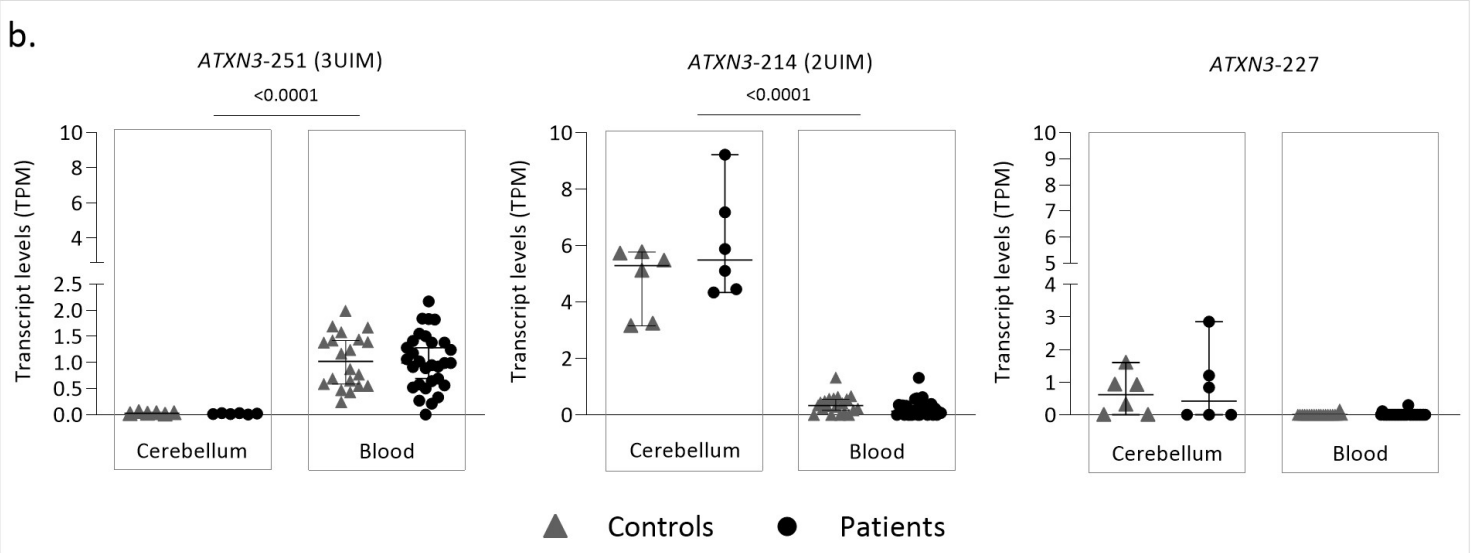

nonsense mediated decay
ATXN3-234
ATXN3-243
ATXN3-253
ATXN3-254
Transcript levels (TPM)
Controls
Patients


protein CDS not defined
ATXN3-205
ATXN3-208
ATXN3-217
ATXN3-226
Transcript levels (TPM)
Controls
Patients


retained intron
ATXN3-210
Transcript levels (TPM)
Controls
Patients


protein coding
ATXN3-251 (ref.)
ATXN3-214
nonsense mediated decay
ATXN3-202
ATXN3-254
protein CDS not defined
ATXN3-249
retained intron
ATXN3-210
ATXN3-205
ATXN3-248
transcript levels (TPM)
CTRL-PA
PA
CTRL-P
Patients

a.

```
P54252-2|ATXN3-251      MESIFHEKQEGSLCAQHCLNNLLQGEYFSPVELSSIAHQLDEEERMRMAEGGVTSEDYRT    60
F5H211|ATXN3-215        MESIFHEKQEGSLCAQHCLNNLLQGEYFSPVELSSIAHQLDEEERMRMAEGGVTSEDYRT    60
P54252-1|ATXN3-214      MESIFHEKQEGSLCAQHCLNNLLQGEYFSPVELSSIAHQLDEEERMRMAEGGVTSEDYRT    60
G3V3R7|ATXN3-227        MESIFHEKQEGSLCAQHCLNNLLQGEYFSPVELSSIAHQLDEEERMRMAEGGVTSEDYRT    60
                        ************************************************************

P54252-2|ATXN3-251      FLQQPSGNMDDSGFFSIQVISNALKVWGLELILFNSPEYQRLRIDPINERSFICNYKEHW    120
F5H211|ATXN3-215        FLQQPSGNMDDSGFFSIQVISNALKVWGLELILFNSPEYQRLRIDPINERSFICNYKEHW    120
P54252-1|ATXN3-214      FLQQPSGNMDDSGFFSIQVISNALKVWGLELILFNSPEYQRLRIDPINERSFICNYKEHW    120
G3V3R7|ATXN3-227        FL-QPSGNMDDSGFFSIQVISNALKVWGLELILFNSPEYQRLRIDPINERSFICNYKEHW    119
                        ** *********************************************************

P54252-2|ATXN3-251      FTVRKLGKQWFNLNSLLTGPELISDTYLALFLAQLQQEGYSIFVVKGDLPDCEADQLLQM    180
F5H211|ATXN3-215        FTVRKLGKQWFNLNSLLTGPELISDTYLALFLAQLQQEGYSIFVVKGDLPDCEADQLLQM    180
P54252-1|ATXN3-214      FTVRKLGKQWFNLNSLLTGPELISDTYLALFLAQLQQEGYSIFVVKGDLPDCEADQLLQM    180
G3V3R7|ATXN3-227        FTVRKLGKQWFNLNSLLTGPELISDTYLALFLAQLQQEGYSIFVVKGDLPDCEADQLLQM    179
                        ************************************************************

P54252-2|ATXN3-251      IRVQQMHRPKLIGEELAQLKEQRVHKTDLERVLEANDGSGMLDEDEEDLQRALALSRQEI    240
F5H211|ATXN3-215        IRVQQMHRPKLIGEELAQLKEQRVHKTDLERVLEANDGSGMLDEDEEDLQRALALSRQEI    240
P54252-1|ATXN3-214      IRVQQMHRPKLIGEELAQLKEQRVHKTDLERVLEANDGSGMLDEDEEDLQRALALSRQEI    240
G3V3R7|ATXN3-227        IRVQQMHRPKLIGEELAQLKEQRVHKTDLERVLEANDGSGMLDEDEEDLQRALALSRQEI    239
                        ************************************************************

P54252-2|ATXN3-251      DMEDEEADLRRAIQLSMQGSSRNISQDMTQTSGTNLTSEELRKRREAYFEKQ--------    292
F5H211|ATXN3-215        DMEDEEADLRRAIQLSMQGSSRNISQDMTQTSGTNLTSEELRKRREAYFEKVRLRLQKYI    300
P54252-1|ATXN3-214      DMEDEEADLRRAIQLSMQGSSRNISQDMTQTSGTNLTSEELRKRREAYFEKQ--------    292
G3V3R7|ATXN3-227        DMEDEEADLRRAIQLSMQGSSRNISQDMTQTSGTNLTSEELRKRREAYFEKQ--------    291
                        ************************************************************

P54252-2|ATXN3-251      -QQKQQQQQQQQQGDLSGQSSHPCERPATSSGALGSDLGDAMSEEDMLQAAVTMSLETV    351
F5H211|ATXN3-215        QQQKQQQQQQQQQGDLSGQSSHPCERPATSSGALGSDLGDAMSEEDMLQAAVTMSLETV    360
P54252-1|ATXN3-214      -QQKQQQQQQQQQGDLSGQSSHPCERPATSSGALGSDLGKACSPFIMFATFTLYLTYEL    351
G3V3R7|ATXN3-227        -QQKQQQQQQQQQGDLSGQSSHPCERPATSSGALGSDL---------------------    329
                         ************************************

P54252-2|ATXN3-251      RND--LKTEGKK-    361
F5H211|ATXN3-215        RND--LKTEGKK-    370
P54252-1|ATXN3-214      HVIFALHYSSFPL    364
G3V3R7|ATXN3-227        -------------    329
```

b.