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

69 Phone: +351 296 650 118

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

71

72 **Word count: 3981 words**

73

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

75

76 Financial Disclosures/Conflict of Interest

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

87 The remaining authors declare that they have nothing to report.

88 Funding Sources

89 This work is an outcome of ESMI, an EU Joint Programme - Neurodegenerative Disease Research
90 (JPND) project. The ESMI project was supported through the following funding organisations
91 under the aegis of JPND: Germany, Federal Ministry of Education and Research (BMBF; funding
92 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

130 **INTRODUCTION**

131 Alternative splicing (AS) is a fundamental step in the regulation of eukaryotic gene expression
132 through which introns are excised and exons are ligated together to form the messenger RNA
133 (mRNA) (Vuong *et al*, 2016). Multiple evidence indicates that genetic variation in *cis*-acting
134 splicing sequence elements or in *trans*-acting splicing factors plays a relevant role in patients
135 with neurodegenerative conditions (reviewed in (Daguenet *et al*, 2015; Nik & Bowman, 2019)).
136 Importantly, *post-mortem* brains of Huntington disease (HD) patients and of mouse models
137 show aberrant splicing of the HD causative gene (huntingtin, *HTT*) itself, leading to the
138 production of high levels of a splice variant only containing exon 1 that encodes a highly
139 aggregation-prone protein and thus contributes to disease pathogenesis (Mangiarini *et al*,
140 1996; Sathasivam *et al*, 2013; Barbaro *et al*, 2015; Neueder *et al*, 2017). Currently, several
141 pathologies triggered by splicing defects, including neurodegenerative diseases, can be treated
142 using different types of RNA molecules, such as antisense oligonucleotide (ASO) or small
143 interference RNA (siRNA), either by reducing the levels of the defective protein or by
144 producing the corrected protein (reviewed in (Dhuri *et al*, 2020; Zhu *et al*, 2022)). In fact, ASO-
145 mediated therapies are currently available for some common and rare diseases, including
146 neurological diseases such as spinal muscular atrophy and Duchenne muscular dystrophy
147 (reviewed in (Zhu *et al*, 2022)).

148 SCA3/MJD is an autosomal dominant neurodegenerative polyglutamine (polyQ) disorder and
149 the most common dominant ataxia worldwide (Sequeiros *et al*, 2012). SCA3/MJD is caused by
150 an abnormal polyQ-encoding CAG repeat motif, consensually containing more than 60
151 trinucleotides within exon 10 of the *ATXN3* gene (Takiyama *et al*, 1993; Kawaguchi *et al*, 1994;
152 Maciel *et al*, 2001). The *ATXN3* gene (ENSG00000066427) originates 54 transcripts (Martin *et*
153 *al*, 2023) in the GRCh38.p13 human genome assembly including the canonical transcript
154 (ENST00000644486/*ATXN3*-251) containing 11 exons and spanning a genomic region about 48
155 kb (Ichikawa *et al*, 2001), encoding the ataxin-3 canonical protein isoform (P54252-2, UniProt).

156 Ataxin-3 is a ubiquitously expressed deubiquitinating enzyme (*e.g.*, (Burnett *et al*, 2003))
157 harbouring on its N-terminus the globular Josephin domain (JD) that includes the catalytic
158 residues, and on its C-terminus a protein segment of undetermined structure containing the
159 polyQ tract and two or three ubiquitin interacting motifs (UIM) (Goto *et al*, 1997; Masino *et al*,
160 2003). Aggregation of the ataxin-3 protein is a hallmark of SCA3/MJD (reviewed in (Costa &
161 Paulson, 2012)) and occurs not only in brain regions with neuronal loss, such as the cerebellar
162 dentate nucleus and the motor and pontine nuclei of the brainstem, but also in brain areas
163 spared of neuronal destruction, like the cerebellar cortex (Koeppen, 2018).

164 Despite nearly 30 years of research following the identification of *ATXN3* as the causative gene
165 of SCA3/MJD (Takiyama *et al*, 1993; Kawaguchi *et al*, 1994), the characterization of the
166 transcriptional species of *ATXN3* remains very incomplete. While *ATXN3* is known to undergo
167 AS (Kawaguchi *et al*, 1994; Goto *et al*, 1997; Ichikawa *et al*, 2001; Harris *et al*, 2010;
168 Bettencourt *et al*, 2010), the extent and regulation of AS, the relative abundance of alternative
169 transcripts in health and disease conditions, the existence of transcript tissue/cell-specificity,
170 and the impact of AS in SCA3/MJD neurotoxicity remain poorly understood. To the best of our
171 knowledge, at least 24 *ATXN3* transcripts, likely to be translated into different isoforms of
172 ataxin-3, have been detected in the human brain and/or peripheral tissues including the blood
173 (Kawaguchi *et al*, 1994; Goto *et al*, 1997; Ichikawa *et al*, 2001; Harris *et al*, 2010; Bettencourt *et*
174 *al*, 2010), although validation of a large number of such transcripts has not been conducted. *In*
175 *silico* analysis performed to infer the functional impact of the several putative ataxin-3 protein
176 isoforms showed that some variants were predicted to be "protective" (lacking the CAG tract),
177 whereas others were foreseen to show increased toxicity (Bettencourt *et al*, 2010). Among the
178 *ATXN3* transcripts which are predicted to be translated, only two transcripts encoding ataxin-3
179 isoforms that differ on the C-terminal containing either two (2UIM) or three UIMs (3UIM) have
180 been mostly studied so far (Harris *et al*, 2010; Johnson *et al*, 2019; Weishäupl *et al*, 2019).
181 These studies using cellular, mouse or *Drosophila* models and *post-mortem* human brain

182 tissues from healthy individuals, though controversial, uncovered differences in the toxicity of
183 these two ataxin-3 protein isoforms by mainly revealing a faster degradation rate but higher
184 aggregation propensity of the 2UIM isoform compared to the 3UIM (Harris *et al*, 2010;
185 Johnson *et al*, 2019; Weishäupl *et al*, 2019). The fact that these two ataxin-3 isoforms seem to
186 differentially impact SCA3/MJD pathogenesis, raises the possibility that other species
187 generated by *ATXN3* AS play key roles in promoting the selective neuronal toxicity displayed by
188 SCA3/MJD patients.

189 Several phase I and II clinical trials are ongoing for polyQ disorders aiming to reduce the levels
190 of disease-related transcripts/proteins using ASO-based drugs, but none was yet been
191 approved for clinical use (reviewed in (Helm *et al*, 2022)). Non-allele-specific as well as allele-
192 specific ASO-mediated reduction of *ATXN3* have been successfully tested *in vivo* and *in vitro*
193 models of MJD (McLoughlin *et al*, 2018; Moore *et al*, 2019, 2017; Hauser *et al*, 2022).
194 Noteworthy, the pharmacokinetics and safety study of a non-selective ASO for SCA3/MJD is
195 currently ongoing (NCT05160558). The identification of specific *ATXN3* transcripts, whose
196 expression needs to be blocked is, therefore, paramount to unveil better strategies in future
197 ASO-based therapies of SCA3/MJD.

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

201

202 **MATERIAL AND METHODS**

203 Samples and RNA-seq datasets

204 Two RNA-seq datasets, generated from a total of 72 samples, namely 60 samples from whole
205 blood and 12 samples from *post-mortem* cerebellum (cerebellar lobules) were used in this
206 study: i) a RNA-seq dataset from whole blood samples of 40 SCA3/MJD subjects (30 patients

207 and 10 pre-ataxic) and 20 control individuals (Raposo *et al*, 2023), obtained in the scope of the
208 European spinocerebellar ataxia type 3/Machado-Joseph Initiative (ESMI); and ii) RNA-seq
209 dataset from *post-mortem* cerebellar samples of six SCA3/MJD patients and six controls
210 publicly available from the work of Haas and colleagues (Haas *et al*, 2022). Detailed
211 information about the blood and cerebellum samples from SCA3/MJD subjects (and
212 corresponding healthy controls), is provided in supplementary Table 1. Briefly, RNA extraction
213 and cDNA library preparation was performed according to manufacturer's instructions, as
214 described elsewhere (Haas *et al*, 2022; Raposo *et al*, 2023). Libraries from blood samples were
215 sequenced as paired end 100bp reads on an Illumina NovaSeq6000 (Illumina), whereas
216 libraries from cerebellum samples were sequenced as paired-end 68 bp reads on a HiSeq2000
217 (Illumina), both next-generation sequencing (NGS) platforms with limited read length (short-
218 reads).

219 Ataxin-3 protein levels in plasma (n=28) and in cerebrospinal fluid (CSF, n=8) from a subset of
220 the total 72 SCA3/MJD subjects were available from previous studies(Gonsior *et al*, 2021;
221 Hübener-Schmid *et al*, 2021). Briefly, total full-length ataxin-3 and/or mutant ataxin-3 were
222 measured using a time-resolved fluorescence energy transfer (TR-FRET)-based immunoassay
223 (Gonsior *et al*, 2021) and a single molecule counting (SMC) immunoassay (Hübener-Schmid *et*
224 *al*, 2021).

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

227

228 <u>Bioinformatic analyses of *ATXN3* splicing</u>

229 The reads were trimmed using Cutadapt v2.3. Next, reads were mapped to the reference
230 genome GRCh38 (Ensembl), using STAR v2.7.8a (Dobin *et al*, 2013) and assigned count
231 estimates to genes with RSEM v1.3.3(Li & Dewey, 2011). Alignment options followed ENCODE

232 standards for RNA-seq (Dobin *et al*, 2013). FastQC (v0.11.8) (Andrews, 2010) was run on .bam
233 files in a post-alignment step, including both aligned and un-aligned reads, to ensure data
234 quality. From the raw reads, TPM was calculated according to Zhao and colleagues (Zhao *et al*,
235 2020), considering the denominator as the sum over all samples (brain and blood).

236

237 Ensembl classification biotype

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

250

251 Statistical analyses

252 To increase the sample size and the power of statistical tests, *ATXN3* transcripts were labelled
253 according to its frequency by sample. Transcripts not detected in at least half of our set of
254 samples (i.e., 5 or less samples from cerebellum and 29 or less samples for blood) were not
255 included in statistical analyses (Suppl. Table 2, Fig. S1). Transcripts levels were compared

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

268

269 Availability of data and materials

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

272

273 **RESULTS**

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

276 Data from RNA-seq of whole blood (n=60) and *post-mortem* cerebellum (n=12) from
277 SCA3/MJD subjects and age- and sex-matched controls was used to identify the currently
278 annotated 54 *ATXN3* transcripts and quantify their abundance (Fig. 1). All annotated
279 transcripts were identified in our study. According to Ensembl annotation (Martin *et al*, 2023),

280 these 54 transcripts can be classified into four main biotypes: protein-coding (n=16), nonsense
281 mediated decay (NMD; n=19), protein CDS not defined (n=16) and retained intron (long
282 noncoding) (n=3) transcripts (Fig. 1). In the 16 protein-coding transcripts group, some
283 structural-related particularities are highlighted (Fig. 1): four transcripts encode full-length
284 ataxin-3 protein isoforms (*ATXN3*-214, *ATXN3*-215, *ATXN3*-227, *ATXN3*-251), including an
285 intact Josephin domain (JD), whereas the remaining 12 transcripts might encode proteins with
286 variations of the N-terminus (*ATXN3*-201, *ATXN3*-203, *ATXN3*-204, *ATXN3*-206, *ATXN3*-207,
287 *ATXN3*-209, *ATXN3*-213, *ATXN3*-219, *ATXN3*-228, *ATXN3*-231, *ATXN3*-235, *ATXN3*-246). Nine
288 transcripts lack exon 11 and, therefore, translate ataxin-3 isoforms only with 2UIM (*ATXN3*-
289 209, *ATXN3*-213, *ATXN3*-214, *ATXN3*-219, *ATXN3*-227, *ATXN3*-228, *ATXN3*-231, *ATXN3*-235,
290 *ATXN3*-246), while seven transcripts retain exon 11 (*ATXN3*-201, *ATXN3*-203, *ATXN3*-204,
291 *ATXN3*-206, *ATXN3*-207, *ATXN3*-215, *ATXN3*-251), and thus encode 3UIMs isoforms. Finally,
292 and although the annotation of the 3' untranslated (UTR) is incomplete, two *ATXN3* transcripts
293 seem to lack exon 10, which contains the polyQ-encoding tract (*ATXN3*-209, *ATXN3*-213) (Fig.
294 1).
295 To explore putative tissue-specific patterns of *ATXN3* AS regardless of the disease status, the
296 frequency of each annotated *ATXN3* transcript detected in samples from SCA3/MJD subjects
297 and controls was determined within each tissue (blood or cerebellum). Expression levels of
298 *ATXN3* transcripts ranged from 0.01 (blood) to 50.49 (cerebellum) transcripts per million
299 (TPM). Globally, cerebellum samples displayed higher expression levels than blood samples
300 (Fig. 2; Suppl. Table 2). The biotype with more abundant transcripts in blood and cerebellum is
301 the retained intron and the noncoding, respectively (Fig. 2). Thirty out of the 46 (74%)
302 transcripts identified in the cerebellum and 44 out of the 52 (85%) transcripts expressed in the
303 blood could only be detected in less than half of samples (Fig. S1; Suppl. Table 2); for example,
304 the protein-coding transcript *ATXN3*-204 was only expressed in one out of 12 cerebellum
305 samples (8%) and in six out of 60 blood samples (10%) (Suppl. Table 2).

306 The protein coding *ATXN3*-251 reference transcript and the protein CDS not defined *ATXN3*-
307 208 transcript showed, respectively, the highest expression in blood and in cerebellum.
308 Additionally, the transcripts more abundant in blood by biotype were the protein coding
309 *ATXN3*-251, the NMD *ATXN3*-202, the noncoding *ATXN3*-208 and the retained intron *ATXN3*-
310 210, while in cerebellum the more abundant were the protein coding *ATXN3*-214, the NMD
311 *ATXN3*-234, the noncoding *ATXN3*-208 and the retained intron *ATXN3*-205 (Fig. 2).
312 Eight of the overall 54 annotated transcripts were exclusively expressed in blood: (i) two
313 corresponded to protein-coding transcripts (*ATXN3*-215, which is one of the four transcripts
314 encoding the full-length ataxin-3 protein isoforms, and *ATXN3*-228); (ii) four were transcripts
315 targeted for NMD (*ATXN3*-212, *ATXN3*-218, *ATXN3*-221, *ATXN3*-236) and (iii) two were protein
316 CDS not defined (*ATXN3*-220, *ATXN3*-233). *ATXN3*-212 transcript showed the highest
317 expression among all transcripts detected in the blood; (Fig. 2, Suppl. Table 2, and Fig. S1). Two
318 NMD transcripts were exclusively expressed in the cerebellum (*ATXN3*-242, *ATXN3*-244)
319 (Suppl. Table 2, Fig. S1).
320
321 **Cerebellar abundance of *ATXN3* transcripts is similar in SCA3 patients and controls**
322 Analysis of the distribution of *ATXN3* transcripts by biological group (SCA3/MJD patients
323 and/or controls) in cerebellum samples revealed that several transcripts were detected only in
324 individual or in a limited number of samples (Suppl. Table 2).
325 To compare distributions of cerebellar *ATXN3* transcripts between SCA3/MJD patients and
326 controls, transcripts detected in more than 50% of all cerebellum samples (six or more of 12
327 samples) were selected. Following this criterium, 16 transcripts were identified in cerebellum
328 samples corresponding to seven protein-coding (the reference variant - *ATXN3*-251, plus
329 *ATXN3*-201, *ATXN3*-206, *ATXN3*-207, *ATXN3*-214, *ATXN3*-219, *ATXN3*-227), four targeted for
330 NMD (*ATXN3*-234, *ATXN3*-243, *ATXN3*-253, *ATXN3*-254), three protein CDS not defined
331 (*ATXN3*-208, *ATXN3*-217, *ATXN3*-226) and two retained intron (*ATXN3*-205, *ATXN3*-210; Fig. 2).

332 While all 16 transcripts showed similar abundance in the cerebellum of SCA3/MJD patients and
333 controls (p>0.05; Fig. 3), the protein CDS not defined transcript *ATXN3*-226 was expressed in all
334 samples from controls but was not expressed in four out of the six patients (Fisher's exact test,
335 p=0.06; Suppl. Table 2).
336
337 **Blood abundance of *ATXN3* transcripts is similar in SCA3/MJD subjects and controls.**
338 Analysis of the distribution of *ATXN3* transcripts in blood samples from SCA3/MJD subjects
339 (pre-ataxic and patients) and controls revealed that, similarly to the cerebellum, several
340 transcripts were only detected in individual or in a limited number of samples (Suppl. Table 2).
341 Following the same criterium used for the cerebellum, eight *ATXN3* transcripts detected in
342 more than 50% of all blood samples (SCA3/MJD subjects and controls) were selected for
343 comparative analyses of patients versus controls and pre-ataxic carriers versus controls. These
344 eight transcripts corresponded to two protein coding (*ATXN3*-214 and the reference transcript
345 - *ATXN3*-251), two transcripts targeted for NMD (*ATXN3*-202, *ATXN3*-254), one protein CDS not
346 defined (*ATXN3*-249), and three retained introns (*ATXN3*-210, *ATXN3*-205, *ATXN3*-248) (Fig. 2).
347 All eight transcripts showed similar levels between SCA3/MJD patients or pre-ataxic carriers
348 compared to their respective controls (Fig. 4).
349
350 **The amount of *ATXN3*-205 transcript in blood is directly associated with CSF levels of soluble**
351 **mutant ataxin-3.**
352 The relationship between *ATXN3* transcript levels from blood cells and soluble levels of ataxin-
353 3 protein (mutant and/or total) quantified in plasma, peripheral mononuclear blood cells
354 (PBMCs) and CSF samples from the same SCA3/MJD subjects was explored in this study (Fig.
355 S2a). Interestingly, levels of the retained intron transcript *ATXN3*-205 in blood cells directly
356 correlated with the abundance of soluble mutant ataxin-3 in the CSF of SCA3/MJD subjects
357 (p=0.031, Fig. S2c). Although not reaching statistical significance, high levels of *ATXN3*-248 and

358 of the reference transcript *ATXN3*-251 displayed a trend to subtly associate with low levels of
359 total ataxin-3 measured in PBMCs of SCA3/MJD individuals (Fig. S2b, d).
360
361 **Expression of the four transcripts encoding full-length ataxin-3 2UIM and 3UIM protein**
362 **isoforms are tissue-specific, irrespective of disease status.**
363 From the 16 protein-coding transcripts, only four encode a full-length ataxin-3 protein with an
364 intact JD and a preserved C-terminus (*ATXN3*-214, *ATXN3*-215, *ATXN3*-227 and *ATXN3*-251; Fig.
365 5a). *ATXN3*-214 and *ATXN3*-227 transcripts encode ataxin-3 2UIM isoforms, whereas *ATXN3*-
366 215 and *ATXN3*-251 transcripts encode for ataxin-3 3UIM isoforms (Fig. 5a). To elucidate
367 whether the individual expression levels of the four transcripts encoding full-length ataxin-3
368 2UIM and 3UIM isoforms contribute to the tissue-specific vulnerability observed in SCA3/MJD,
369 the abundance of these transcripts in blood and cerebellum was compared between
370 SCA3/MJD patients and controls (Fig. 5b).
371 Interestingly, two of these four transcripts were only expressed in some samples. The *ATXN3*-
372 215 transcript (3UIM) was only identified in one blood sample (Supp. Table 2); while the
373 *ATXN3*-227 transcript (2UIM) was detected at distinct frequencies in both tissues, it was only
374 measured in three out of 60 (5%) blood samples and in seven out of 12 (58%) in cerebellum
375 samples (Supp. Table 2; Fig. 3). Although being overall commonly detected in the cerebellum,
376 *ATXN3*-227 transcript showed similar levels and frequency in SCA3/MJD patients and controls
377 (Fig. 3).
378 The reference transcript, *ATXN3*-251 (3UIM), showed a ~50-fold increase of abundance in
379 blood samples over cerebellum samples, whereas the *ATXN3*-214 (2UIM) transcript was more
380 expressed in the cerebellum about 20-fold more than in blood (Fig. 5b). The magnitude of the
381 fold change values, however, could be partially reflecting technical differences due to different
382 RNA-seq experimental setups. Remarkably, in cerebellum samples, the *ATXN3*-214 (2UIM) was
383 expressed 265-fold more than the *ATXN3*-251 (3UIM) transcript, whereas in blood samples the

384 *ATXN3*-251 (3UIM) transcript was expressed 4-fold more than *ATXN3*-214 (2UIM). These
385 observed differences in transcript abundances between tissues were similar in SCA3/MJD
386 subjects and controls.

387 **DISCUSSION**

388 Here, we describe the diversity and abundance of annotated *ATXN3* transcripts detected in
389 blood and cerebellum samples from SCA3/MJD subjects and controls. Globally, the number of
390 different transcripts and the abundance of transcripts were higher in the cerebellum than in
391 blood, both in SCA3/MJD subjects and controls. Interestingly, in SCA3/MJD subjects and
392 controls pooled samples altogether, while the most abundant transcript in the cerebellum
393 (*ATXN3*-208) is categorized as a "protein coding with a CDS not defined of unknown function
394 transcript" by Ensembl, the transcript showing the highest abundance in the blood is the
395 protein-coding *ATXN3* reference transcript (*ATXN3*-251) that is translated into an ataxin-3
396 3UIM protein isoform. No differences in the abundance of the most frequent transcripts were
397 found in blood or cerebellum samples of SCA3/MJD subjects and controls. Noteworthy, the
398 abundance of *ATXN3*-251 and *ATXN3*-214 transcripts, two out of the four transcripts that
399 encode full-length ataxin-3 protein isoforms but differ in the C-terminus (2UIM *versus* 3UIM)
400 were strongly associated with tissue expression specificity: *ATXN3-251* (3UIM) transcript was
401 expressed in blood 50-fold more than in cerebellum, whereas *ATXN3*-214 (2UIM) transcript
402 was expressed in the cerebellum 20-fold more than in blood.

403 Studies using a *Drosophila* model of SCA3/MJD indicate that in addition to a gain-of-function of
404 the mutant ataxin-3 protein, *ATXN3* transcripts harbouring expanded CAG repeats show
405 independent and progressive cellular toxicity (Li *et al*, 2008). However, whether such toxic
406 transcript action occurs in human cells remains to be elucidated. Hence, the characterization
407 of *ATXN3* AS in mutation carriers of SCA3/MJD might provide fundamental knowledge to
408 unveil the contribution of such mechanism in SCA3/MJD pathogenesis.

409 Globally, AS shows its highest complexity (high diversity of transcripts), in the brain amongst
410 other tissues (Porter *et al*, 2018). As expected, *ATXN3* transcripts diversity and abundance (a
411 greater number of transcripts, which are expressed in a larger number of samples with the
412 highest levels of expression) was higher in the cerebellum than in the blood and might indicate
413 an important role of *ATXN3* AS and its regulation in the neurodegenerative process of
414 SCA3/MJD.

415 Twelve out of 16 protein-coding transcripts are predicted to encode proteins with truncations
416 in the JD. Such modifications imply that several ataxin-3 protein forms with a variable N-
417 terminus might exist, displaying for example different forms of the globular JD or other N-
418 terminal domains that show cysteine protease activity or even another yet unknown activity.
419 The fact that ataxin-3 interacts with so many proteins in cells (reviewed in (Costa & Paulson,
420 2012)) and is ubiquitously expressed (*e.g.,* (Paulson *et al*, 1997)) together with our data
421 indicates that several ataxin-3 isoforms might exist with distinct functions other than the
422 currently known deubiquitinating enzyme role. Interestingly, four out of these 12 transcripts
423 (*ATXN3*-201, *ATXN3*-206, *ATXN3*-207, *ATXN3*-219) are highly expressed in the cerebellum
424 compared to blood suggesting that these potential encoded proteins may play specific roles in
425 the cerebellum and that the correspondent mutant forms could potentially contribute for the
426 observed selective cerebellar toxicity observed in SCA3/MJD patients.

427 Four out of 54 transcripts, *ATXN3*-214, *ATXN3*-215, *ATXN3*-227 and *ATXN3*-251, encode full-
428 length ataxin-3 protein isoforms, which differ only in the C-terminus. In the present study, the
429 *ATXN3*-214 and the *ATXN3*-227 transcripts, both encoding 2UIM isoforms, are preferentially
430 abundant in cerebellum samples, whereas the *ATXN3*-251 transcript (3UIM) is more expressed
431 in blood. Altogether, these findings suggest that ataxin-3 isoforms with 2UIM might have a
432 unique role in the cerebellum and consequently in neurodegeneration. The correlation
433 between transcript and protein abundance is still not known and future studies exploring the

434 role of the *ATXN3*-214 and *ATXN3*-227 transcripts and respective protein isoforms in brain
435 tissues could contribute for the elucidation of mechanisms involved in cellular vulnerability to
436 mutant *ATXN3* gene products in SCA3/MJD.

437 Of note, a significant direct relationship between the abundance of the retained intron *ATXN3*-
438 205 transcript in blood and soluble mutant ataxin-3 in CSF samples was observed in a small
439 sub-set of SCA3/MJD subjects. Because *ATXN3*-205 is classified by Ensembl as a long non-
440 coding RNA, such finding suggests that this transcript might have a regulatory function in the
441 transcription and/or translation of protein coding transcripts in the nervous system and,
442 therefore, contribute to the observed correlation with the soluble mutant ataxin-3 protein in
443 the CSF. More comprehensive studies about the function of these type of transcripts are
444 needed as well as larger number of CSF samples should be further analyzed to better explore
445 this preliminary but important finding.

446 In this study, transcripts more frequently found in blood (n=8) and in cerebellum (cerebellar
447 lobules, n=16) showed similar levels in SCA3/MJD patients and controls. Mort and colleagues
448 also described similar amount of two *HTT* transcripts in cerebellum samples of HD patients and
449 controls (Mort *et al*, 2015). In SCA3/MJD, neuronal loss in the cerebellar lobules is not as
450 severe as, for example, in the dentate nucleus or pons (Koeppen, 2018). Whether our findings
451 are specific of this brain area is unknown and, thus, characterization of *ATXN3* transcripts in
452 other SCA3/MJD patients brain tissues showing more severe lesions would be important to
453 conduct in the future.

454 In this work, the diversity and abundance of annotated *ATXN3* transcripts in SCA3/MJD
455 patients and controls is provided. The methodology used in this study did not allow to
456 specifically analyze the *ATXN3* transcripts with CAG repeat expansions. Future studies using
457 other NGS approaches, including novel bioinformatic pipelines to correctly detect CAG
458 expansions are needed to elucidate the contribution of the expanded CAG repeat itself in

459 *ATXN3* AS. Moreover, functional studies to define the role of the most promising transcripts
460 found in the cerebellum, for example for its most abundant known non-coding *ATXN3*-208
461 transcript, would be important to conduct in the future. In fact, the majority of annotated
462 *ATXN3* transcripts might have distinct functions or regulatory properties and its
463 comprehension will reveal novel mechanisms of SCA3/MJD pathogenesis.

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

470

471 **ACKNOWLEDGEMENTS**

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

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

479

480 **AUTHORS' ROLES**

481 Design and conceptualization of the study: MR, JH-S, RT, MCC, ML; Subject

482 recruitment/Acquisition of participants biomaterials/data: MR, AFF, ARVM, JV, PP, TK, HG-M,

483 PG, MMS, LPA, JI, BPW, JV, JF, TK, NC, JA, LS, OR, ML; Bioinformatic/Statistical analysis of data:

484 MR, RT; Drafting of the manuscript: MR, JH-S, MCC, ML; Revision of the Manuscript: MR, JH-S,

485 RT, AFF, ARVM, JV, PP, TK, HG-M, PG, MMS, LPA, JI, BPW, JV, JF, TK, NC, JA, LS, OR, MCC, ML.

486 All authors read and approved the final manuscript.

487

488 **REFERENCES**

489 Andrews S (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].
490 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [PREPRINT]

491 Barbaro BA, Lukacsovich T, Agrawal N, Burke J, Bornemann DJ, Purcell JM, Worthge SA,
492 Caricasole A, Weiss A, Song W, *et al* (2015) Comparative study of naturally occurring
493 huntingtin fragments in Drosophila points to exon 1 as the most pathogenic species in
494 Huntington's disease. *Hum Mol Genet* 24: 913–925

495 Bettencourt C, Santos C, Montiel R, Costa M do C, Cruz-Morales P, Santos LR, Simões N, Kay T,
496 Vasconcelos J, Maciel P, *et al* (2010) Increased transcript diversity: novel splicing variants
497 of Machado–Joseph Disease gene (ATXN3). *Neurogenetics* 11: 193–202

498 Burnett B, Li F & Pittman RN (2003) The polyglutamine neurodegenerative protein ataxin-3
499 binds polyubiquitylated proteins and has ubiquitin protease activity. *Hum Mol Genet* 12:
500 3195–205

501 Costa M do C & Paulson HL (2012) Toward understanding Machado-Joseph disease. *Prog*
502 *Neurobiol* 97: 239–257 doi:10.1016/j.pneurobio.2011.11.006 [PREPRINT]

503 Daguenet E, Dujardin G & Valcárcel J (2015) The pathogenicity of splicing defects: mechanistic
504 insights into pre- <scp>mRNA</scp> processing inform novel therapeutic approaches.
505 *EMBO Rep* 16: 1640–1655

506 Dhuri K, Bechtold C, Quijano E, Pham H, Gupta A, Vikram A & Bahal R (2020) Antisense
507 Oligonucleotides: An Emerging Area in Drug Discovery and Development. *J Clin Med* 9:
508 2004

509 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M & Gingeras
510 TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21

511 Gonsior K, Kaucher GA, Pelz P, Schumann D, Gansel M, Kuhs S, Klockgether T, Forlani S, Durr A,
512 Hauser S, *et al* (2021) PolyQ-expanded ataxin-3 protein levels in peripheral blood
513 mononuclear cells correlate with clinical parameters in SCA3: a pilot study. *J Neurol* 268:
514 1304–1315

515 Goto J, Watanabe M, Ichikawa Y, Yee S-B, Ihara N, Endo K, Igarashi S, Takiyama Y, Gaspar C,

516 Maciel P, *et al* (1997) Machado–Joseph disease gene products carrying different carboxyl
517 termini. *Neurosci Res* 28: 373–377

518 Haas E, Incebacak RD, Hentrich T, Huridou C, Schmidt T, Casadei N, Maringer Y, Bahl C,
519 Zimmermann F, Mills JD, *et al* (2022) A Novel SCA3 Knock-in Mouse Model Mimics the
520 Human SCA3 Disease Phenotype Including Neuropathological, Behavioral, and
521 Transcriptional Abnormalities Especially in Oligodendrocytes. *Mol Neurobiol* 59: 495–522

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

525 Hauser S, Helm J, Kraft M, Korneck M, Hübener-Schmid J & Schöls L (2022) Allele-specific
526 targeting of mutant ataxin-3 by antisense oligonucleotides in SCA3-iPSC-derived neurons.
527 *Mol Ther - Nucleic Acids* 27: 99–108

528 Helm J, Schöls L & Hauser S (2022) Towards Personalized Allele-Specific Antisense
529 Oligonucleotide Therapies for Toxic Gain-of-Function Neurodegenerative Diseases.
530 *Pharmaceutics* 14: 1708

531 Hübener-Schmid J, Kuhlbrodt K, Peladan J, Faber J, Santana MM, Hengel H, Jacobi H, Reetz K,
532 Garcia-Moreno H, Raposo M, *et al* (2021) Polyglutamine-Expanded Ataxin-3: A Target
533 Engagement Marker for Spinocerebellar Ataxia Type 3 in Peripheral Blood. *Mov Disord*
534 36: 2675–2681

535 Ichikawa Y, Goto J, Hattori M, Toyoda A, Ishii K, Jeong SY, Hashida H, Masuda N, Ogata K, Kasai
536 F, *et al* (2001) The genomic structure and expression of MJD, the Machado-Joseph
537 disease gene. *J Hum Genet* 46: 413–22

538 Johnson SL, Blount JR, Libohova K, Ranxhi B, Paulson HL, Tsou W-L & Todi S V. (2019)
539 Differential toxicity of ataxin-3 isoforms in Drosophila models of Spinocerebellar Ataxia
540 Type 3. *Neurobiol Dis* 132: 104535

541 Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H,
542 Nakamura S, Nishimura M & Akiguchi I (1994) CAG expansions in a novel gene for
543 Machado-Joseph disease at chromosome 14q32.1. *Nat Genet* 8: 221–8

544 Koeppen AH (2018) The Neuropathology of Spinocerebellar Ataxia Type 3/Machado-Joseph
545 Disease. In pp 233–241.

546 Li B & Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or
547 without a reference genome. *BMC Bioinformatics* 12: 323

548 Li L-B, Yu Z, Teng X & Bonini NM (2008) RNA toxicity is a component of ataxin-3 degeneration
549 in Drosophila. *Nature* 453: 1107–1111

550 Maciel P, Costa MC, Ferro A, Rousseau M, Santos CS, Gaspar C, Barros J, Rouleau GA, Coutinho
551 P & Sequeiros J (2001) Improvement in the molecular diagnosis of Machado-Joseph
552 disease. *Arch Neurol* 58: 1821–1827

553 Mangiarini L, Sathasivam K, Seller M, Cozens B, Harper A, Hetherington C, Lawton M, Trottier
554 Y, Lehrach H, Davies SW, *et al* (1996) Exon 1 of the HD Gene with an Expanded CAG
555 Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice.
556 *Cell* 87: 493–506

557 Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R,
558 Berry A, Bhai J, *et al* (2023) Ensembl 2023. *Nucleic Acids Res* 51: D933–D941

559 Masino L, Musi V, Menon RP, Fusi P, Kelly G, Frenkiel TA, Trottier Y & Pastore A (2003) Domain
560 architecture of the polyglutamine protein ataxin-3: a globular domain followed by a
561 flexible tail. *FEBS Lett* 549: 21–5

562 McLoughlin HS, Moore LR, Chopra R, Komlo R, McKenzie M, Blumenstein KG, Zhao H,
563 Kordasiewicz HB, Shakkottai VG & Paulson HL (2018) Oligonucleotide therapy mitigates
564 disease in spinocerebellar ataxia type 3 mice. *Ann Neurol* 84: 64–77

565 Moore LR, Keller L, Bushart DD, Delatorre RG, Li D, McLoughlin HS, do Carmo Costa M,
566 Shakkottai VG, Smith GD & Paulson HL (2019) Antisense oligonucleotide therapy rescues
567 aggresome formation in a novel spinocerebellar ataxia type 3 human embryonic stem cell
568 line. *Stem Cell Res* 39: 101504

569 Moore LR, Rajpal G, Dillingham IT, Qutob M, Blumenstein KG, Gattis D, Hung G, Kordasiewicz
570 HB, Paulson HL & McLoughlin HS (2017) Evaluation of Antisense Oligonucleotides
571 Targeting ATXN3 in SCA3 Mouse Models. *Mol Ther - Nucleic Acids* 7: 200–210

572 Mort M, Carlisle FA, Waite AJ, Elliston L, Allen ND, Jones L & Hughes AC (2015) Huntingtin
573 Exists as Multiple Splice Forms in Human Brain. *J Huntingtons Dis* 4: 161–171

574 Neueder A, Landles C, Ghosh R, Howland D, Myers RH, Faull RLM, Tabrizi SJ & Bates GP (2017)
575 The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's
576 disease patients. *Sci Rep* 7: 1307

577 Nik S & Bowman T V. (2019) Splicing and neurodegeneration: Insights and mechanisms. *WIREs*
578 *RNA* 10

579 Paulson HL, Das SS, Crino PB, Perez MK, Patel SC, Gotsdiner D, Fischbeck KH & Pittman RN
580 (1997) Machado-Joseph disease gene product is a cytoplasmic protein widely expressed
581 in brain. *Ann Neurol* 41: 453–62

582 Porter RS, Jaamour F & Iwase S (2018) Neuron-specific alternative splicing of transcriptional
583 machineries: Implications for neurodevelopmental disorders. *Mol Cell Neurosci* 87: 35–45

584 Raposo M, Hübener-Schmid J, Ferreira AF, Vieira Melo AR, Vasconcelos J, Pires P, Kay T, Garcia-
585 Moreno H, Giunti P, Santana MM, *et al* (2023) Blood transcriptome sequencing identifies
586 biomarkers able to track disease stages in spinocerebellar ataxia type 3. *Brain* accepted

587 Sathasivam K, Neueder A, Gipson TA, Landles C, Benjamin AC, Bondulich MK, Smith DL, Faull
588 RLM, Roos RAC, Howland D, *et al* (2013) Aberrant splicing of HTT generates the
589 pathogenic exon 1 protein in Huntington disease. *Proc Natl Acad Sci* 110: 2366–2370

590 Sequeiros J, Martins S & Silveira I (2012) Epidemiology and population genetics of degenerative
591 ataxias. In pp 227–251.

592 Takiyama Y, Nishizawa M, Tanaka H, Kawashima S, Sakamoto H, Karube Y, Shimazaki H,
593 Soutome M, Endo K & Ohta S (1993) The gene for Machado-Joseph disease maps to
594 human chromosome 14q. *Nat Genet* 4: 300–4

595 Vuong CK, Black DL & Zheng S (2016) The neurogenetics of alternative splicing. *Nat Rev*
596 *Neurosci* 17: 265–281

597 Weishäupl D, Schneider J, Peixoto Pinheiro B, Ruess C, Dold SM, von Zweydorf F, Gloeckner CJ,
598 Schmidt J, Riess O & Schmidt T (2019) Physiological and pathophysiological characteristics
599 of ataxin-3 isoforms. *J Biol Chem* 294: 644–661

600 Zhao S, Ye Z & Stanton R (2020) Misuse of RPKM or TPM normalization when comparing across
601 samples and sequencing protocols. *RNA* 26: 903–909

602 Zhu Y, Zhu L, Wang X & Jin H (2022) RNA-based therapeutics: an overview and prospectus. *Cell*
603 *Death Dis* 13: 644

604

605 **LEGENDS OF FIGURES**

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

622

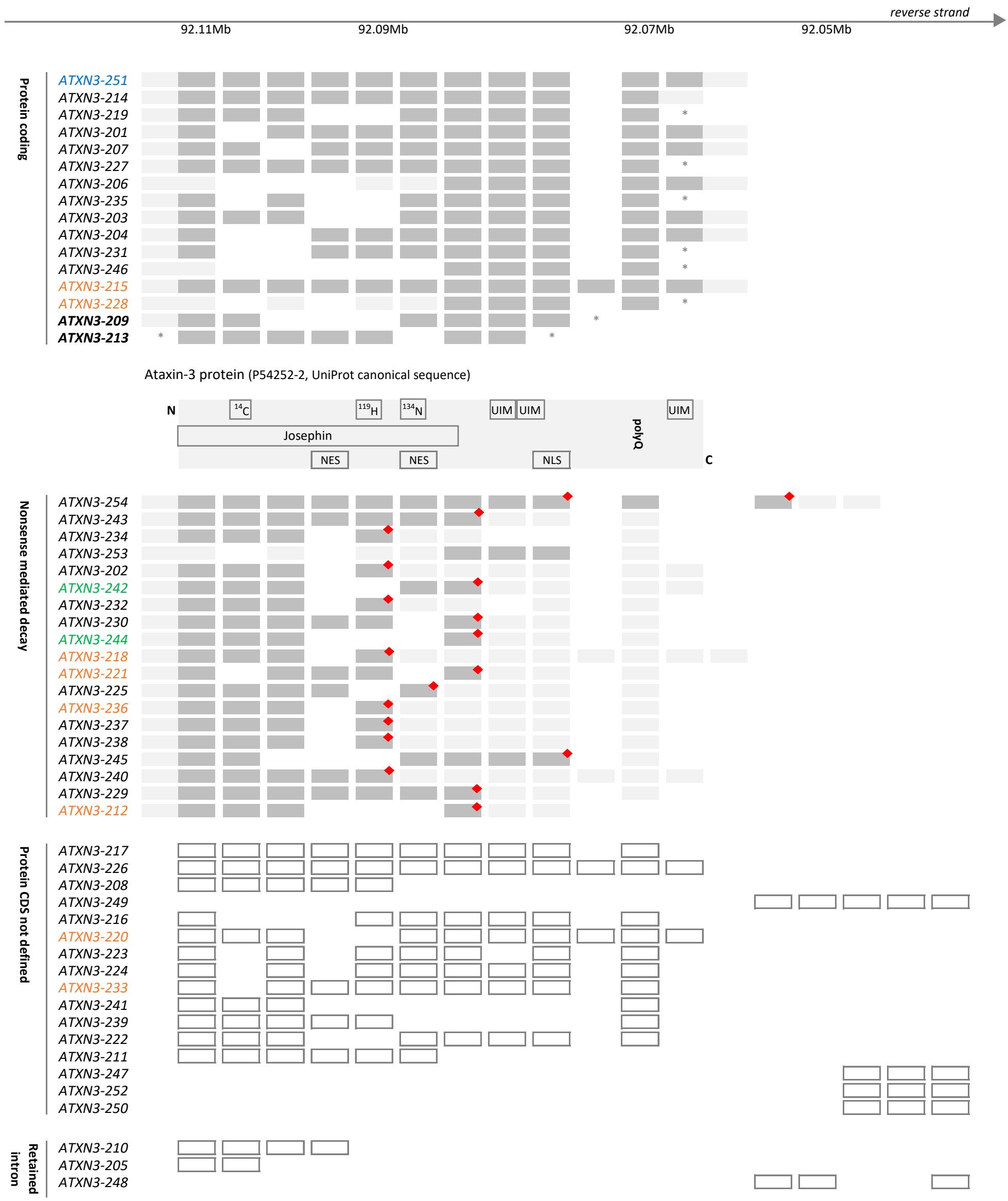
623 **Figure 2.** Heatmap of the 54 annotated *ATXN3* transcripts detected in cerebellum and blood
624 samples of SCA3/MJD subjects and healthy individuals. Transcripts are clustered by biotype
625 (protein coding, nonsense mediated decay, protein sequence (CDS) not defined, and retained
626 intron)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

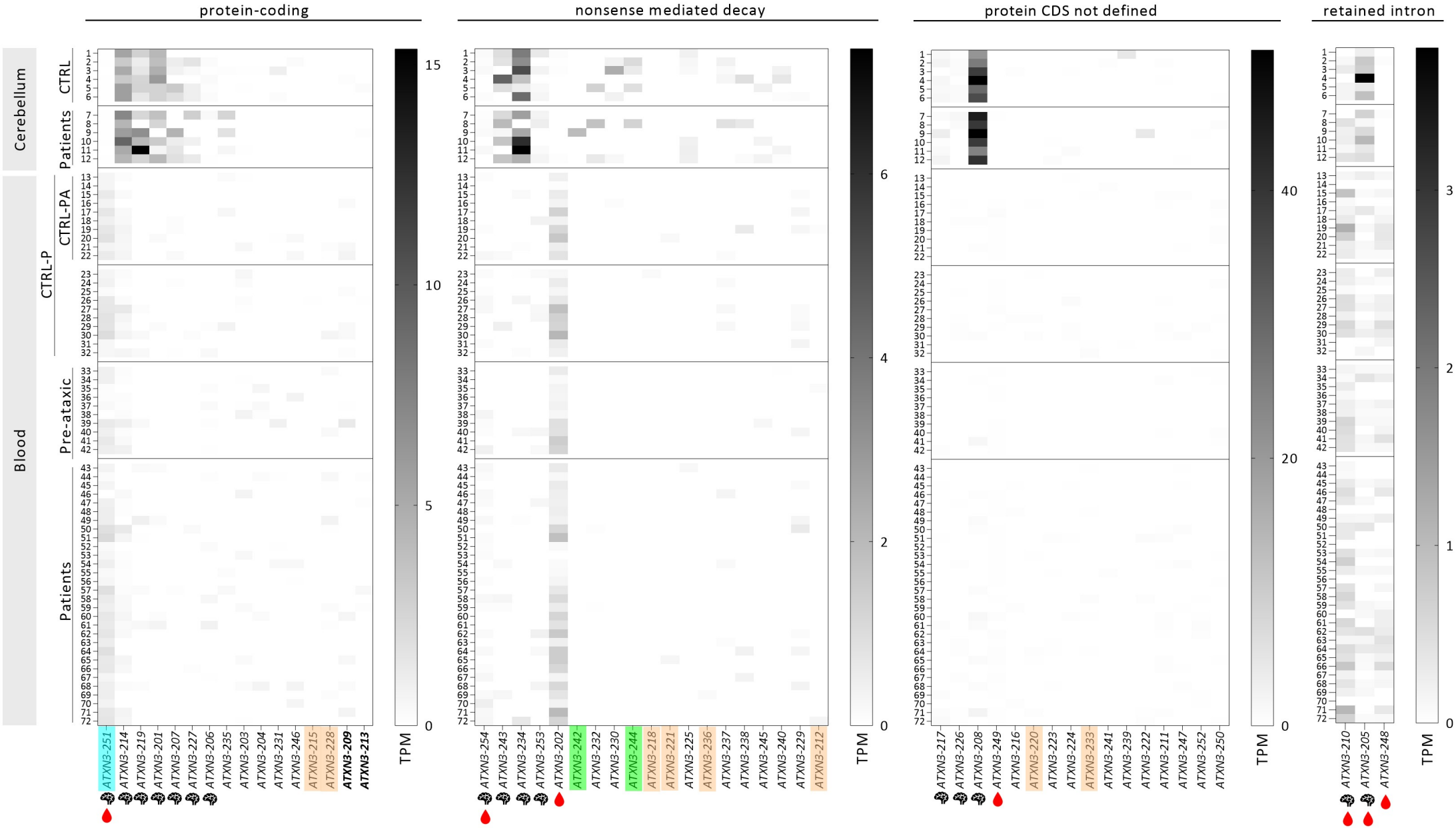
631 detected in more than 50% of all samples in each tissue. *ATXN3* transcripts with exclusive
632 expression in blood and transcripts with exclusive expression in cerebellum are highlighted in
633 orange and green, respectively. The Ensembl's *ATXN3* reference transcript
634 ENST00000644486.2|*ATXN3*-251 (MANE Select v0.95) is highlighted in blue.
635
636 **Figure 3.** Levels of the 16 *ATXN3* transcripts expressed in more than half of all cerebellum
637 samples from SCA3/MJD patients (n=6) and controls (n=6). The transcripts are clustered
638 according to the type of transcript, as defined by Ensembl classification biotype (Ensembl release
639 108 - Jan 2023, REF). Transcript expression levels are shown in transcript per million (TPM). Bars
640 in the graphs represent the median ± 95% CI (confidence interval, error bars).
641
642 **Figure 4.** Expression levels of eight *ATXN3* transcripts in blood samples from pre-ataxic (PA)
643 carriers (n=10), SCA3/MJD patients (n=30) and respective controls (CTRL-P, n=20 and CTRL-PA,
644 n=10). CTRL-PA was defined as a subset of age-matched controls with pre-ataxic carriers within
645 the CTRL-P group. The transcripts are clustered according to biotype, as defined by Ensembl
646 classification biotype (Ensembl release 108 - Jan 2023(Martin *et al*, 2023)). Expression levels are
647 shown in transcript per millions (TPM). Bars in the graphs represent the median ± 95% CI
648 (confidence interval, error bars).
649
650 **Figure 5.** (a) Sequence alignment of the four full-length ataxin-3 protein isoforms harbouring an
651 intact Josephin domain (JD) and a preserved C-terminus. Ataxin-3 protein isoforms, namely the
652 P54252-2, F5H211, P54252-1 and G3V3R7 are encoded by the *ATXN3*-251, *ATXN3*-215, *ATXN3*-
653 214 and *ATXN3*-227 transcripts, respectively. Multiple sequence alignment was performed using
654 Clustal Omega program (Sievers F. et al., 2011). JD (green line), the three catalytic sites (orange
655 dot) and the three UIMs (blue boxes) are highlighted. (b) Levels of *ATXN3*-251 (3UIM) and of the
656 *ATXN3*-214 as well as *ATXN3*-227 (both 2UIM) transcripts in the blood and cerebellum samples

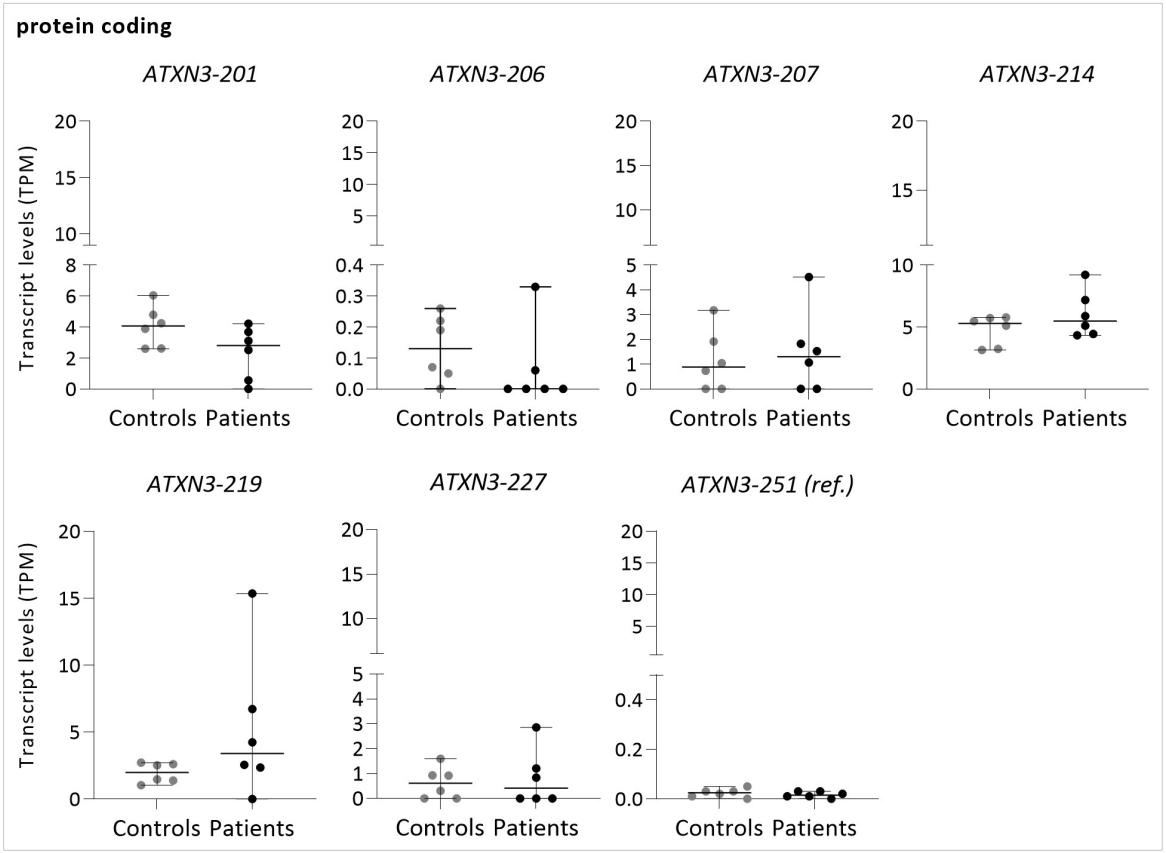
657 from SCA3/MJD subjects and controls. Transcript levels are shown in transcripts per million

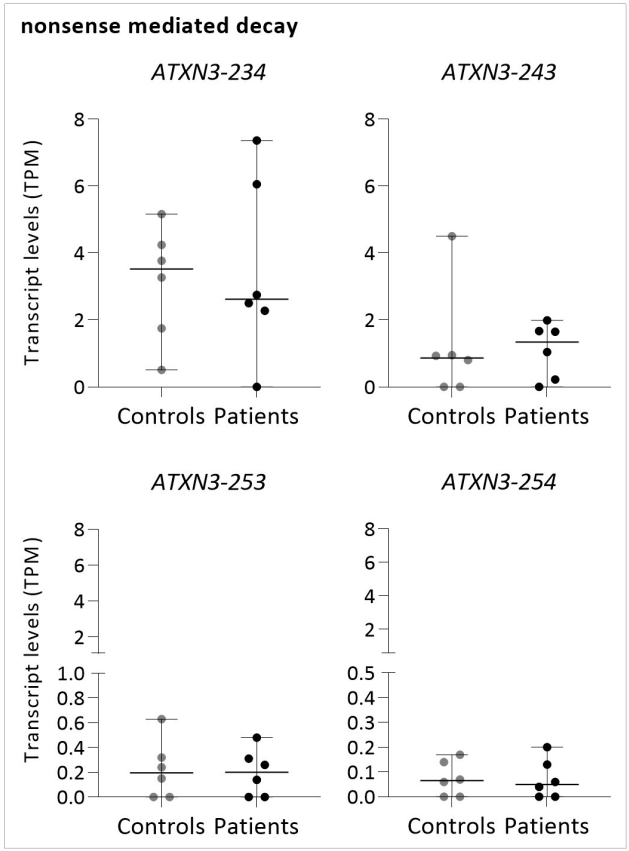
658 (TPM). Bars in the graphs represent the median ± 95% CI (confidence interval, error bars).

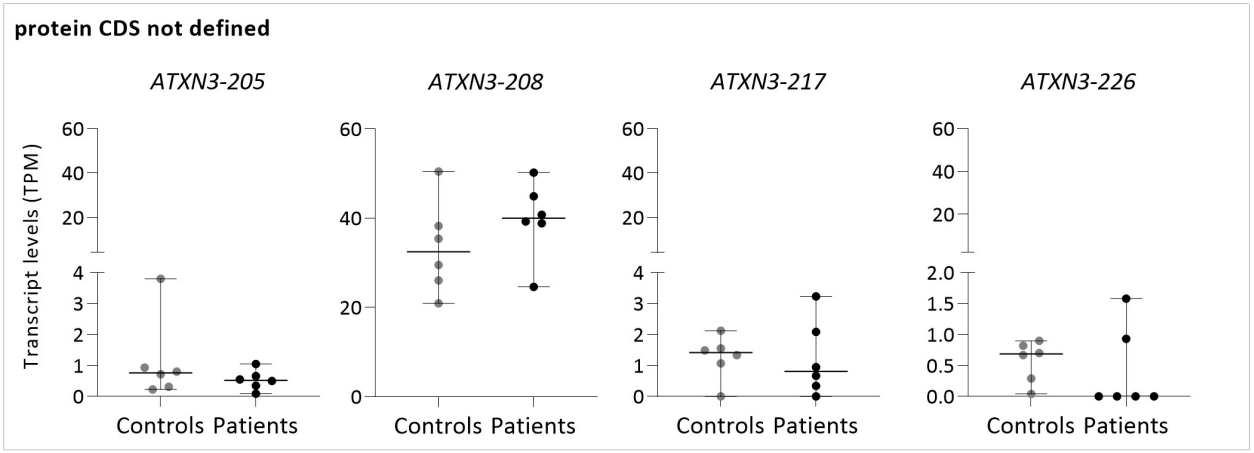
659

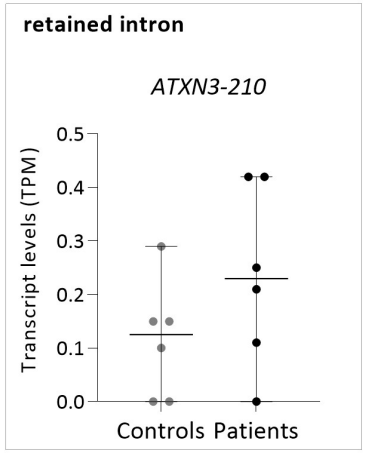

reverse strand
92.11Mb
92.09Mb
92.07Mb
92.05Mb
Protein coding
ATXN3-251
ATXN3-214
ATXN3-219
ATXN3-201
ATXN3-207
ATXN3-227
ATXN3-206
ATXN3-235
ATXN3-203
ATXN3-204
ATXN3-231
ATXN3-246
ATXN3-215
ATXN3-228
ATXN3-209
ATXN3-213
Ataxin-3 protein (P54252-2, UniProt canonical sequence)
N
14C
119H
134N
UIM
UIM
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

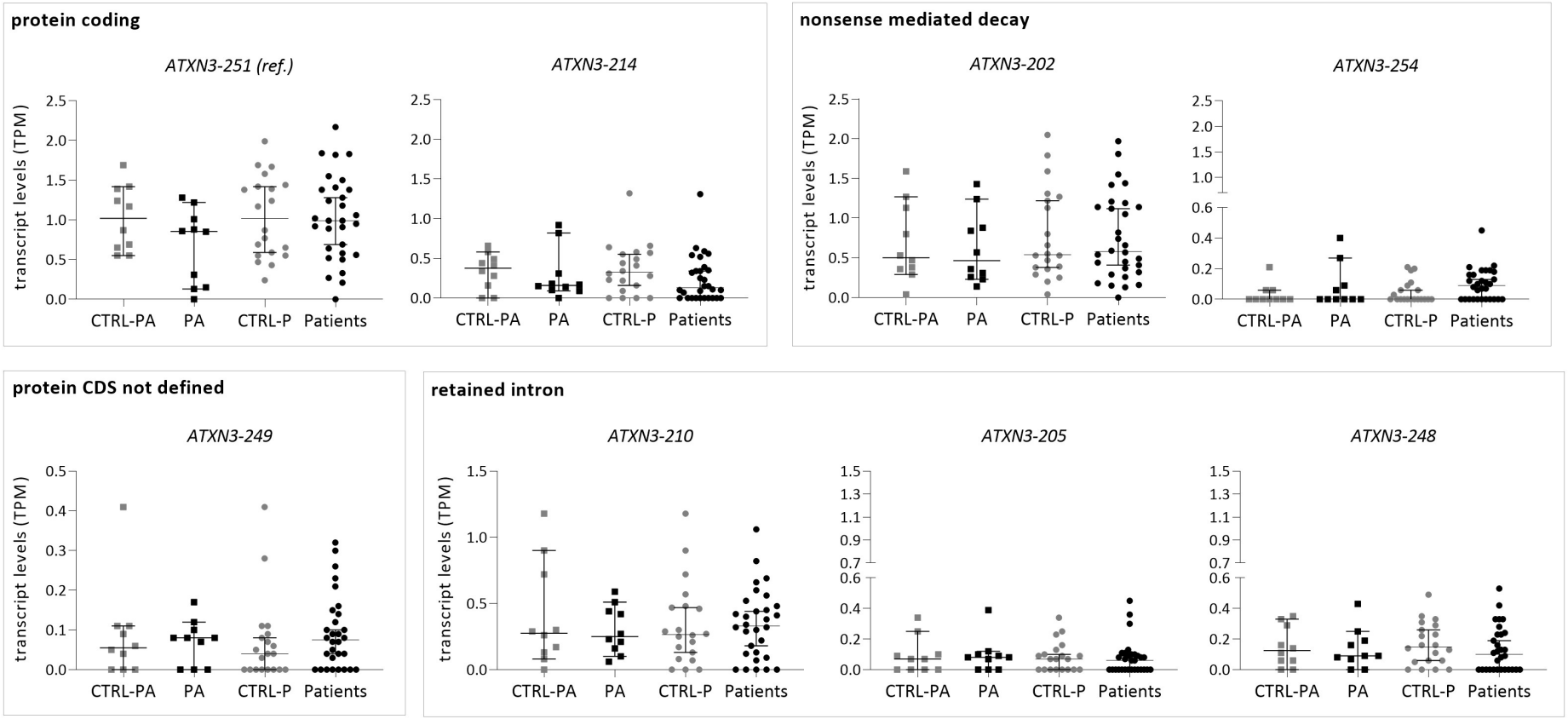

protein-coding
nonsense mediated decay
protein CDS not defined
retained intron
Cerebellum
Blood
CTRL
Patients
CTRL-P
CTRL-PA
Pre-ataxic
Patients
TPM
ATXN3-251
ATXN3-214
ATXN3-219
ATXN3-201
ATXN3-207
ATXN3-227
ATXN3-206
ATXN3-235
ATXN3-203
ATXN3-204
ATXN3-231
ATXN3-246
ATXN3-215
ATXN3-228
ATXN3-209
ATXN3-213
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

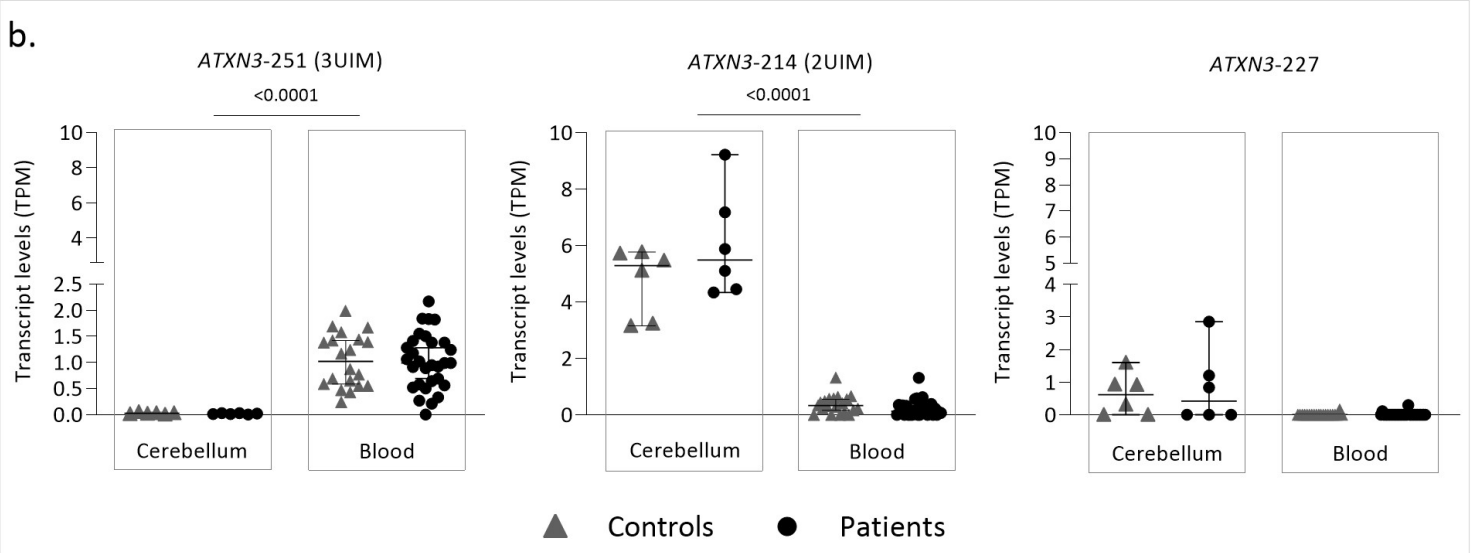

nonsense mediated decay
ATXN3-234
ATXN3-243
ATXN3-253
ATXN3-254
Transcript levels (TPM)
Controls
Patients


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.