Blood transcriptome sequencing identifies biomarkers able to track disease stages in spinocerebellar ataxia type 3

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

ABSTRACT

Transcriptional dysregulation has been described in spinocerebellar ataxia type 3/Machado-Joseph disease (SCA3/MJD), an autosomal dominant ataxia caused by a polyglutamine expansion in the ataxin-3 protein. As ataxin-3 is ubiquitously expressed, transcriptional alterations in blood may reflect early changes that start before clinical onset and might serve as peripheral biomarkers in clinical and research settings. Our goal was to describe enriched pathways and report dysregulated genes which can track disease onset, severity, or progression in carriers of the *ATXN3* mutation (pre-ataxic subjects and patients).

Global dysregulation patterns were identified by RNA sequencing of blood samples from 40 carriers of *ATXN3* mutation and 20 controls and further compared with transcriptomic data from *post-mortem* cerebellum samples of MJD patients and controls. Ten genes - *ABCA1, CEP72, PTGDS, SAFB2, SFSWAP, CCDC88C, SH2B1, LTBP4, MEG3 and TSPOAP1* - whose expression in blood was altered in the pre-ataxic stage and simultaneously, correlated with ataxia severity in the overt disease stage, were analysed by quantitative real-time PCR in blood samples from an independent set of 170 SCA3/MJD subjects and 57 controls.

Pathway enrichment analysis indicated the Gαi signalling and the oestrogen receptor signalling to be similarly affected in blood and cerebellum. *SAFB2*, *SFSWAP* and *LTBP4* were consistently dysregulated in pre-ataxic subjects compared to controls, displaying a combined discriminatory ability of 79%. In patients, ataxia severity was associated with higher levels of *MEG3* and *TSPOAP1*.

We propose expression levels of *SAFB2*, *SFSWAP* and *LTBP4* as well as *MEG3* and *TSPOAP1* as stratification markers of SCA3/MJD progression, deserving further validation in longitudinal studies and in independent cohorts.

Author affiliations

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 Serviço de Neurologia, Hospital do Divino Espírito Santo, Ponta Delgada, Portugal

6 Serviço de Neurologia, Hospital do Santo Espírito da Ilha Terceira, Angra do Heroísmo, Portugal

7 Serviço de Genética Clínica, Hospital D. Estefânia, Lisboa, Portugal

8 Ataxia Centre, Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, UK

9 Department of Neurogenetics, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK

10 Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra, 3000-075, Portugal

11 Neurology Service, University Hospital Marqués de Valdecilla-IDIVAL, Universidad de Cantabria, Centro de Investigación en Red de Enfermedades Neurodegenerativas (CIBERNED), Santander, Spain

12 Radboud University Medical Centre, Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Nijmegen, The Netherlands

13 Department of Neurology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

14 Department of Neurology, University Hospital Bonn, Bonn, Germany

15 German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

16 NGS Competence Center Tübingen, Tübingen, Germany

17 Department for Neurodegenerative Diseases, Hertie-Institute for Clinical Brain Research and Center for Neurology, University of Tübingen, Germany

18 German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany

Correspondence to: Mafalda Raposo

Instituto de Investigação e Inovação em Saúde (i3S)

Rua Alfredo Allen, 208

4200-135 Porto, Portugal

E-mail: msraposo@i3s.up.pt

Running title: Novel blood-based biomarkers of SCA3/MJD

Keywords: *ATXN3*; ataxin-3; polyQ diseases; neurodegenerative disease; RNA-seq

Abbreviations: AO = age at onset; ATXN3 = ataxin 3; ABCA1 = ATP binding cassette subfamily A member 1; CAG = cytosine-adenine-guanine; CCDC88C = coiled-coil domain containing 88C; CEP72 = centrosomal protein 72; DD = disease duration; DE = differentially expressed; ESMI = European spinocerebellar ataxia type 3/MJD Initiative; R = false discovery rate; GPCRs = G protein-coupled receptors; HD = Huntington disease; HSPB1 = heat shock protein family B (small) member 1; lncRNA = long non-coding RNA; LTBP4 = latent transforming growth factor beta binding protein 4; MAPT = microtubule associated protein tau; MEG3 = maternally expressed 3; NfL = Neurofilament light chain; PA = pre-ataxic subject; PTGDS = prostaglandin D2 synthase; qPCR = quantitative real-time PCR; RNA-seq = RNA sequencing; SARA = Scale for the assessment and rating of ataxia; SCA = spinocerebellar ataxia; SCA3/MJD = spinocerebellar ataxia type 3/Machado-Joseph disease; SAFB2 = scaffold attachment factor B2; SFSWAP = splicing factor SWAP; SH2B1 = SH2B adaptor protein 1; TGFβ = transforming growth factor beta 1; TSPOAP1 = TSPO associated protein 1; TP53 = tumor protein p53

INTRODUCTION

Spinocerebellar ataxia type 3 (SCA3)/Machado-Joseph disease (MJD) is an autosomal dominant neurodegenerative disorder characterized by selective dysfunction and degeneration of the cerebellum and brainstem[1,2]. Disease onset, occurring on average at midlife (~40 years), inversely correlates with the elongation of an exonic CAG motif at the *ATXN3* gene[3,4]. The presence of an expanded allele, harbouring consensually above 60 repeats[5], leads to a mutated form of the deubiquitinating enzyme ataxin-3[6]. Misfolding of mutant ataxin-3 and its subsequent aggregation, predominantly in the nucleus of affected cells, are the pathognomonic hallmarks of SCA3/MJD and are associated with the disruption of key cellular pathways[7,8], including transcriptional regulation[7–9]. Progressive gait and limb ataxia are the clinical hallmark of SCA3/MJD[10,11], whose severity is almost universally graded using the Scale for the Assessment and Rating of Ataxia (SARA;[12]).

Genetic diagnosis through predictive testing allows the identification of asymptomatic or pre-ataxic individuals[13] offering an unique opportunity to prevent or slow neuronal damage before clinical onset. However, interventional trials are currently hampered by the lack of sensitive markers for monitoring the disease in its early stages, and even more evident in its asymptomatic phase. To date, several biomarkers of SCA3/MJD have been investigated[14–17]. Among these, the mutated ataxin-3 and the neurofilament light chain (NfL) are highlighted, due to the explicit association to SCA3/MJD pathogenesis or the inherent neurodegenerative process, respectively[15,16,18,19]. Although the value of such biomarkers is acknowledged, the specific stratification of the pre-ataxic stage using molecular data is not yet guaranteed. Of note, it is unlikely that a single biomarker will be enough to monitor disease progression; more likely, a combination of biomarkers will be necessary, which is currently undiscovered.

Mutant ataxin-3 is known to be ubiquitously expressed across tissues[20] and increasing evidence suggests it exerts its effects also in easily available tissues, such as blood[14], a fact that provides the opportunity to find consistent peripheral alterations that correlate with clinical data. Upon identification, such peripheral biomarkers should be particularly suitable in the context of therapeutic strategies using compounds that can be taken systemically and delivered across the blood-brain barrier (such as small molecules, amongst others). Biomarkers may also serve to select patients for first therapeutic studies considering that it is unlikely that treatments will reverse progressed neurodegeneration in late-stage patients.

Cross-sectional, whole transcriptome microarray analyses have shown that there is global dysregulation in blood samples from SCA3/MJD subjects[9]. The extent to which such gene expression alterations reflect clinically meaningful dynamics (i.e., correlate with aspects of disease onset, progression and/or severity), however, remains elusive.

Profiting from a large and well-established cohort of European SCA3/MJD subjects, enrolled through the multicentric European Spinocerebellar Ataxia Type 3/Machado-Joseph Disease Initiative (ESMI), we performed next-generation sequencing-based transcriptome analysis in blood of SCA3/MJD mutation carriers (pre-ataxic subjects and patients). We describe the global dysregulation patterns found in blood and report transcriptional alterations than can track disease severity/progression, starting at the pre-ataxic stage. Moreover, we explore whether transcriptional changes seen in blood (both at the level of individual genes or enriched pathways) paralleled those from a previous RNA-sequencing (RNA-seq) study using *post-mortem* cerebellum samples from SCA3/MJD patients.

SUBJECTS AND METHODS

Cohort and sample collection

A total of 210 SCA3/MJD subjects and 77 controls, recruited between 2016 and 2019, were included in the present work. The ESMI (European Spinocerebellar Ataxia Type 3/Machado-Joseph Disease Initiative) cohort comprised subjects with confirmed SCA3/MJD and non-expanded *ATXN3* carriers without neurological disease (controls). The determination of the *ATXN3* genotype for all samples was performed centrally (University of Tubingen).

Clinical assessments and blood collection were performed at visit 1 for all sites, using a harmonized common protocol implemented in ESMI. For a subset of subjects (n=74), clinical data and blood samples were also available from a second annual visit, performed within 2 months around the specific timepoint (visit 2). SARA scores[12] were available for all SCA3/MJD subjects and were used to classify mutation carriers as either patients (SARA score $\geq$3, n=165) or pre-ataxic subjects (PA; SARA score <3, n=45)[13].

Age at visit was calculated as the difference between the year of birth and the year of the clinical evaluation/blood collection. Age at onset (AO) was defined as the age of the first gait disturbances, reported by the patient or a close relative/caregiver. Disease duration (DD) was calculated as the number of years elapsed between age at onset and age at visit. For pre-ataxic

carriers, time to onset was defined as the difference between age and predicted AO, which was determined according to Tezenas du Montcel and colleagues[21].

A total of 361 blood samples, collected in PAXGene Blood RNA tubes (Cat ID: 762165, BD) according to the manufacturer's instructions, were used to perform:

(i) *RNA-seq analysis:* Samples from 10 pre-ataxic carriers, 30 patients, and 20 controls were used. Patients were selected according to their SARA score to represent a wide range of the disease severity: 10 mild (score ≥3 and <10), 10 moderate (score ≥15 and <25), and 10 severe (score ≥25). Controls were matched by age (similar range) and sex (similar proportion) to *ATXN3* carriers.

(ii) qPCR analysis: Samples from 35 pre-ataxic carriers, 135 patients, and 57 controls (visit 1) were used. For a subset of SCA3/MJD subjects (12 pre-ataxic carriers and 62 patients), samples from visit 2 (1-year interval) were also analysed.

The study was approved by local ethics committees and all subjects provided written informed consent.

Post –mortem brain tissues of six SCA3/MJD patients and six control individuals (average age at death of 67 years for patients and 64 years for controls) were available from a previous study[22].

The workflow of the study is shown at Figure 1; briefly, data from a RNA-seq experiment using whole blood from SCA3/MJD carriers and controls was used: (**A**) to identify common enriched pathways in blood and cerebellum (cross-sectional design) by overlapping our data with RNA-seq datasets from *post-mortem* cerebellum samples; and (**B**) to select expression alterations that correlate with disease onset (biomarker study), severity, or progression (including the pre-ataxic stage).

Total RNA isolation and cDNA synthetization

For RNA-seq or qPCR analysis, total RNA was isolated from blood cells using the Qiasymphony PAXGene Blood RNA kit (Cat ID 762635, Qiagen), following the automated protocol V5 or the MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit, compatible with PAXgene™ Blood RNA Tubes (Cat ID: 4451894, Invitrogen), respectively. The RNA concentration, RNA purity and RNA Integrity Number were evaluated using the Qubit RNA BR Assay Kit (ThermoFisher Scientific), the NanoDrop ND-1000 Spectrophotometer (PEQLAB), and the Bioanalyzer 2100 (RNA 6000 Nano Kit, Agilent), respectively. For library

preparation, total RNA libraries were prepared using the TruSeq Stranded Total RNA with Ribo-Zero Globin (Illumina), according to the manufacturer's instructions. The libraries were denaturated, diluted to 270 pM and sequenced as paired end 100bp reads on an Illumina NovaSeq6000 (Illumina) with a sequencing depth of approximately 60 million clusters, in average, per sample.

For qPCR analysis, 500 nanograms of total RNA was used to synthesize complementary DNA (cDNA), using the High-Capacity cDNA Reverse Transcription Kit with RNase inhibitor (Cat ID: 4374966, Applied Biosystems).

RNA sequencing analysis

Read quality of RNA-seq data (fastq files) was assessed using ngs-bits (v.2019_04), to identify sequencing cycles with low average quality, adaptor contamination, or repetitive sequences from PCR amplification. Reads were aligned to the GRCh37 using STAR v2.7.0f[23] and alignment quality was analysed using ngs-bits. Normalized read counts for all genes were obtained using Subread (v1.6.4) and edgeR (v.3.26.4). Raw expression values were available for 60,790 genes in the 60 samples. Raw gene expression data was filtered by demanding a minimum expression value of 1 cpm (counts per million) in at least 8 samples. Filtered data contains expression values for 16,888 genes.

Global differential expression (DE) analysis in blood samples

To identify the blood-based global transcriptional profile of SCA3/MJD, differential expression (DE) analysis between SCA3/MJD carriers (pre-ataxic subjects and patients) compared to controls were performed using expression data from 16,888 genes, by fitting a negative binomial distribution using a generalized linear model (GzLM) conducted at edgeR version 3.18.1. For each gene, expression fold change values (log2 fold change) were calculated, and statistical significance was given as nominal p-value and/or q-value (FDR, obtained by Benjamini-Hochberg procedure).

Global differential expression analysis in cerebellum samples

To assess the global transcriptional profile of SCA3/MJD in the cerebellum, DE analysis using expression values from six *post-mortem* cerebellum samples of SCA3/MJD patients, and six controls were performed according to Haas and colleagues[22]; expression data was available for 17,543 genes.

Pathway enrichment analysis

Pathway analyses were performed with the Ingenuity Pathway Analysis software[24], using as input data the dysregulated genes at p-value<0.05 from global DE analysis using blood; for cerebellum samples a q-value<0.05 was used. Pathways with a -log (Benjamini-Hochberg p-value)>1.3 were considered significantly enriched. A z-score, which is a measure of the predicted direction of the pathway activity, was calculated; pathways with a z-score > 2.0 or < -2.0 were significantly activated or inhibited, respectively. Enriched pathways from blood were intersected (Venn diagram, http://bioinformatics.psb.ugent.be/webtools/Venn/) with those from cerebellum analysis to uncover pathways common to both tissues.

Selection of candidate biomarkers

To select expression alterations correlatable with disease onset, severity, or progression (including the pre-ataxic stage), RNA-seq data was analysed to: (i) compare gene expression levels between PA subjects and controls (analysis of covariance with age as covariate and log2 transforming all variables prior to the test); and (ii) correlate gene expression levels and SARA score in patients (partial Spearman rank correlation). The potential effects of age, number of CAG repeats in the expanded allele and disease duration were statistically removed in the partial Spearman rank correlations. Statistical analyses were run at R version 3.6.2 and a significance level of 5% were considered. To further identify alterations which could simultaneously, distinguish PA from controls and correlate with SARA scores in patients, DE genes from (i) were intersected with DE genes from (ii) (Suppl. Table 1), which resulted in a set of 62 common genes.

Quantitative real-time PCR analysis

cDNA amplification by qPCR was performed using TaqMan Gene Expression Assays (IDs are described in Supp. Table 1) and TaqMan Fast Advanced Master Mix (Applied Biosystems), according to the supplier's instructions. qPCR experiments were performed in the Bio-Rad CFX384 system (Bio-Rad). For each gene, samples were run in triplicate alongside the reference gene – *TRAP1*[25]. Furthermore, to minimize possible batch effects, each plate always contained samples from one control, one pre-ataxic subject (visit 1 and visit 2) and one patient (visit 1 and visit 2) from each research center. Relative expression values were calculated by the $2^{-\Delta Ct}$ method[26] through the CFX Maestro 1.1 Software, version 4.1.2433.1219 (Bio-Rad). Amplification curves from 29 pre-ataxic carriers, 129 patients, and 51 controls were successfully obtained and further used in statistical analysis.

Statistical procedure

The ROUT method (Q=1%) was used to exclude outliers from qPCR data previously to statistical analyses. A chi-square test of independence was used to compare the proportion of subjects by sex and biological groups (PA subjects, patients, and controls). Differences between biological groups on age, number of CAG repeats in the expanded *ATXN3* allele, AO, DD, and SARA score were determined by Mann-Whitney U or Kruskal-Wallis tests. Using the controls dataset, the relationship between gene expression levels and age, total RNA concentration, and RNA Purity was assessed by Spearman rank order correlation. Differences between groups of categorical variables (sex, research center, country of origin, time of blood collection, fasting and blood storage time) on gene expression levels were tested by Mann-Whitney U or Kruskal-Wallis tests. Expression data for the 10 candidate biomarkers was used to perform comparisons between biological groups and to establish associations with clinical and genetic data. To account for age as a potential cofounder, two sub-sets of controls were formed: controls matched to pre-ataxic carriers (CTRL-PA, n=24) and controls matched to patients (CTRL-P, n=27). Differences in expression levels between biological groups were determined by the Kruskal-Wallis test. To analyse the ability of expression levels of the 10 genes in discriminating PA from matched controls ROC analysis was performed. To explore the direction and strength of the relationship between expression levels of the ten genes and demographic (age), clinic (time to predicted onset, AO, DD, SARA score) and genetic data (number of CAG repeats in the expanded *ATXN3* allele), Spearman correlation coefficients (rho) were calculated; to account for the influence (i) of the number of CAG repeats in expanded allele on AO and (ii) of age, the number of CAG repeats in expanded allele and disease duration on expression levels, partial Spearman correlation coefficients (rho*) were also computed. For follow-up analyses, differences on expression levels between the two timepoints (visit 1 and visit 2) were compared using the Wilcoxon signed rank test.

Statistical analyses were performed in IBM SPSS Statistics for windows version 25.0 (IBM Corp. Released 2017) and GraphPad Prism 8.0.1. The significance level of all tests was set to 5%. To control for Type I errors, *post hoc* analyses using the Dunn's multiple comparisons tests were performed. Graphic bars are shown as median ± 95% CI (confidence interval).

Data availability

Most data are available in this manuscript and in supplementary material. Raw transcriptomic data will be made available, upon request, to the corresponding author.

RESULTS

Demographic, genetic, and clinical characterization of the study participants is detailed in Table 1 and Suppl. Table 3. Age, sex, as well as several technical variables related with RNA-seq experiments were shown not to be confounders of gene expression levels (Suppl. Table 4).

Similar patterns of affectation of Gαi and oestrogen receptor signalling pathways in blood and cerebellum

Using RNA-seq data, blood expression levels of PA and patients were compared with those of controls as well as levels of patients were compared with PA subjects (genes at a q-value <0.05 are shown in suppl. Table 5). Furthermore, global DE analyses identified a total of 1467 dysregulated genes (785 downregulated and 682 upregulated) significantly associated with the SCA3/MJD carrier (PA and patients) status, at a nominal p-value significance (p<0.05). Using expression levels of these 1469 genes as input data, a total of 51 pathways were found to be significantly enriched (-log p-value>1.30) (Suppl. Table 6). Noteworthy, the two pathways with both the highest overlap and -log p-value were the interferon signalling (overlap=22%, -log p-value=3.61) and the inflammasome pathway (overlap=25%, -log p-value=2.67; Suppl. Table 6), which were both activated (z-score>2).

Global DE analyses using data from a previous study of *post-mortem* cerebellum samples[22] identified a total of 1058 dysregulated genes (732 downregulated and 326 upregulated) in patients compared to controls (q-value<0.05). Pathway enrichment analysis (using the 1058 DE genes) revealed 52 enriched pathways at a -log B-H p-value>1.30. The pathway with both the highest overlap and statistical significance was the glutamate receptor signalling (overlap=20%, -log B-H p-value=3.74; Suppl. Table 7), which was predicted to be inhibited (z-score< 2).

We further intersected enriched pathways identified from blood with those from cerebellum analysis. Five pathways were commonly enriched (-log B-H p-value>1.30) in both tissues (Suppl. Fig.1a): from these, the Gαi signalling, and the oestrogen receptor signalling showed a consistent predicted direction of activity in both tissues (activated and inhibited, respectively), although this prediction failed to reach significance (Suppl. Fig.1b, Suppl. Table 6 and 7).

Promising RNA-seq based candidate biomarkers of SCA3/MJD

Aiming to identify gene expression alterations that would be detectable already in the pre-ataxic phase of the disease and that, simultaneously, could be correlated with ataxia severity in the overt disease stage, we intersected genes whose expression levels showed significant differences between PA subjects and controls (n= 1002; p-value<0.05) with those which, in patients, correlated with the SARA score (n=962; p value<0.05). Sixty-two genes were identified (Suppl. Table 1); from these, *ABCA1*, *CEP72*, *PTGDS*, *SAFB2*, *SFSWAP*, *CCDC88C*, *SH2B1*, *LTBP4*, *MEG3 and TSPOAP1* were prioritized (prioritization criteria described in Suppl. Table 2) and were further analysed by qPCR in an independent set of 28 pre-ataxic carriers, 124 patients, and 47 controls.

Analysis of qPCR data revealed several significant disease-related expression patterns for five out of 10 genes analysed: *SAFB2*, *SFSWAP*, *LTBP4*, *MEG3* and *TSPOAP1*; furthermore, expression patterns of these five genes were specific for the disease stage: levels of *SAFB2*, *SFSWAP* and *LTBP4* were associated with the pre-ataxic stage, whereas levels of *MEG3* and *TSPOAP1* were correlated with ataxia severity. None of the 10 genes was able to simultaneously distinguish PA from matched-controls and correlate with SARA scores in patients, as previously observed in RNA-seq analysis.

SAFB2 levels are increased in the pre-ataxic stage and show an increase with disease progression

Levels of *SAFB2*, encoding for the scaffold attachment factor B2, a transcriptional regulator, were confirmed to be significantly increased in pre-ataxic carriers compared to matched controls (Fig. 2a). Levels of *SAFB2* discriminated PA from controls with an accuracy of 0.71 (p-value=0.0059, Fig.2b). The correlation of expression levels of *SAFB2* with SARA score, previously identified in RNA-seq, was not significant in the independent set of patients (Suppl. Table 8; Suppl. Fig.3). Noteworthy, expression levels of *SAFB2* were increased in patients with an earlier age at onset (rho=-0.271, p=0.004; Fig.2c). Also, in patients, an increase of *SAFB2* levels was further observed when analysing follow-up data, with levels from the second visit being, in average, significantly higher than those from visit 1 (p=0.023, Fig.2d). This trend, however, was not observed in the pre-ataxic stage (Suppl. Fig.4).

SAFB2, SFSWAP and _LTBP4_ display a high combined ability to classify the pre-ataxic stage

Transcript levels of the splicing factor SWAP gene – *SFSWAP* - were significantly increased in pre-ataxic carriers compared to matched controls (Fig. 3a), whereas no significant correlation was found between expression levels of *SFSWAP* and SARA score (Suppl. Table 8, Suppl. Fig 3). Transcript levels of the latent transforming growth factor beta binding protein 4 - *LTBP4*- were significantly lower in pre-ataxic carriers than in matched controls (Fig. 3b). Again, the correlation between *LTBP4* levels and SARA score observed in RNA-seq experiments was lost in the independent set of SCA3/MJD patients. Similar levels of *SFSWAP* and *LTBP4* between visit 1 and visit 2 were observed in pre-ataxic carriers and patients (Suppl. Fig 4).

Since the levels of *SAFB2, SFSWAP* and *LTBP4* were significantly dysregulated in pre-ataxic subjects, we analysed the joint discriminative ability of the three genes. Combined expression levels of these three genes are expected to be able to distinguish PA from controls, with a 79% chance (p=0.002, Fig. 3c).

Levels of _MEG3_ and _TSPOAP1_ are increased in more severe cases of the SCA3/MJD

For the *MEG3* (maternally expressed 3 gene), a long non-coding RNA gene, as well as for the TSPO associated protein 1 gene (*TSPOAP1*), differences in expression between PA and controls were not replicated in the larger cohort. Noteworthy, for these two genes in the patient's group, the correlation between expression levels and SARA score was maintained. Thus, patients showing higher SARA scores consistently presented higher levels of *MEG3* (rho*=0.346, p-value=0.003, Fig.4a) and *TSPOAP1* (rho*=0.222, p-value=0.030, Fig.4b) after the adjustment of confounders (age, number of CAG in the expanded allele, and disease duration; Suppl. Table 8).

In our large independent set of SCA3/MJD subjects, expression levels of *ABCA1*, *CCDC88C*, *CEP72, PTGDS,* and *SH2B1* failed to distinguish PA carriers from controls and/or to correlate with the respective SARA score in patients (Suppl. Fig. 2, Suppl. Table 8).

Using expression data from the present study and from a previous study with *post-mortem* cerebellum samples[22], we analysed the consistency of gene dysregulation patterns of *SAFB2, SFSWAP, LTBP4, MEG3 and TSPOAP1* in blood and cerebellum (Suppl. Fig. 5). In blood samples, patients, in comparison with controls, presented similar levels of the five genes (p>0.05), whereas in cerebellum samples, levels of *SAFB2, SFSWAP, LTBP4, and TSPOAP1*

were significantly dysregulated ($p<0.05$); *SAFB2, SFSWAP and TSPOAP1* were increased in patients whereas levels of *LTBP4* were decreased (Suppl. Fig. 5). It appears that dysregulation patterns of *SAFB2, SFSWAP*, and *LTBP4* levels in cerebellum samples are more similar to the dysregulation observed in blood samples from pre-ataxic carriers than to what is observed in patients.

DISCUSSION

In this study we confirmed the presence of peripheral transcriptional dysregulation in SCA3/MJD through performing next-generation sequencing-based transcriptome analysis of whole blood samples from SCA3/MJD mutation carriers (pre-ataxic and patients) and controls. To assess the transcriptional signature of SCA3/MJD in a highly affected tissue, the cerebellum, we also analysed data from a previous RNA-seq study using *post-mortem* samples from SCA3/MJD patients and controls[22]. Although brain samples can be biased towards the end-stage of the disease, comparison with blood datasets allowed insights on the similarity/differences between the periphery and a highly affected region.

In *post-mortem* cerebellum, downregulated genes represented 69% of all total dysregulated genes, a finding according to the recruitment of transcription factors into aggregates by mutated ataxin-3[27]. This pattern was not seen in blood, where the proportion of downregulated (54%) *versus* upregulated genes did not evidence a trend towards a decrease in transcription, similarly to what has been observed in previous microarray-based transcriptomic studies[9]. Although transcription dysregulation in blood of SCA3/MJD subjects was confirmed, the magnitude of the differential expression was limited, with all differences involving nominal p-values. The limited magnitude of the differences found between expression levels of controls and SCA3/MJD subjects in the present study contrasts with the high number of dysregulated genes identified after controlling for multiple comparisons (FDR) in two previous microarray-based expression studies[9] (Ana F. Ferreira, personal communication). As the frequency of false-positive signals in microarray analyses is known to be much higher than in RNA-seq, especially in transcripts with low expression levels[28], we can postulate that dysregulation levels provided from array data are overestimated.

Intersection of blood and cerebellum RNA-seq datasets allowed the identification of two commonly enriched pathways, the Gαi signalling and the oestrogen receptor signalling, with an expected direction of activity which is consistent in both tissues. The identification of

enriched pathways common to both SCA3/MJD blood and brain supports the use of blood cells to investigate features of disease biology, highlighting new pathogenic signatures to be explored in further studies. The Gαi signalling is predicted to be activated in blood as well as in cerebellum of SCA3/MJD subjects. Heterotrimeric guanine nucleotide-binding (G) proteins are transducers of G protein-coupled receptors (GPCRs), which translate signals from extracellular ligands into intracellular responses[29]. Gαi is one of the four types of Gα subunits which undergo a conformational change when coupled with GPCRs (previously activated by a ligand). Several receptors (e.g., dopamine, serotonin and glutamate) are amongst the Gαi-coupled GPCRs highly abundant in brain, whose activity is generally related to the inhibition of the adenylate cyclase enzyme, leading ultimately to reduced neuronal excitability[29]. Remarkably, evidence of impaired neurotransmission in SCA3/MJD by defects in acetylcholine, glutamatergic, dopaminergic and serotonergic signalling has been previously described[7]. Pathway analysis further indicated that the oestrogen receptor signalling pathway is predicted to be inhibited in blood and in cerebellum of SCA3/MJD subjects. Oestrogens are cholesterol-derived sex hormones playing an essential role in sex but also in non-sex specific physiological processes, including neuroprotective actions under basal and pathologic conditions[30]. Two previous studies pointed to the existence of sex differences in SCA3/MJD but its effect on disease onset and progression was not elucidated[31,32]; more recently, and using also data from the ESMI cohort, mean deterioration rate in SARA total score or appendicular sub-score was two and five-fold increased, respectively, in men compared to women[33]. Although we could hypothesize that neuroprotection mediated by oestrogens might be impaired in SCA3/MJD, which such neuroprotection would be more evident in men, further studies, specifically designed to address this issue, need to be conducted.

Given the existence of transcriptional dysregulation in SCA3/MJD blood cells, expression levels of specific genes could constitute suitable peripheral biomarkers. In fact, previous attempts to identify transcriptional biomarkers were based only on the establishment of differences relative to controls, whereas the link between abnormal expression levels and clinical rating measures was missing[9]. Attempting to solve this major drawback, we have selected candidate transcriptional biomarkers grounded on the rationale of ideally detecting alterations which are already present in the pre-ataxic stage and, additionally, when evaluated in patients, show a correlation with ataxia worsening, as measured by the SARA score. Using this strategy, we identified a set of 62 genes and prioritized *ABCA1, CEP72, PTGDS, SAFB2, SFSWAP, CCDC88C, SH2B1, LTBP4, MEG3, and TSPOAP1* to be tested by qPCR in a large

and independent set of SCA3/MJD subjects and controls. As clinical biomarkers are devoid of utility in the pre-ataxic stage of the disease, the identification of molecular biomarkers for this specific phase is urgent. We were able to identify three genes - *SAFB2, SFSWAP,* and *LTBP4* - that show a distinct expression behaviour in the pre-ataxic stage of SCA3/MJD. The discriminatory ability of the combined expression levels of the three genes to distinguish pre-ataxic carriers from controls was 79%, which is similar to levels of mutant ataxin-3 (78%) and NfL (84%)[15,34]. Levels of *SAFB2*, which were found to be increased in pre-ataxic subjects (compared to controls), further increased in most patients with a one-year follow up visit; thus, *SAFB2* is a promising candidate biomarker for disease progression, whose behaviour deserves further investigation in a longitudinal setup. SAFB2 is part of the SAFB family, formed by DNA–RNA-binding proteins which are involved in regulation of transcription and mRNA processing, DNA repair and cellular response to stress[35]. Although SAFB proteins are widely expressed, SAFB1 and SAFB2 show high expression levels in the central nervous and immune systems[36]. Interestingly, repressor activity of SAFBs on oestrogen receptor signalling has been described[36]; we could thus hypothesize that upregulation of *SAFB2* in blood and *post-mortem* cerebellum samples of SCA3/MJD subjects can, at least in part, be associated with inhibition of the oestrogen receptor signalling pathway, predicted for both tissues. Moreover, SAFBs are also regulators of the promoter activity of HSPB1 (also known as HSP27)[37], a heat-shock protein whose downregulation was observed in lymphoblastoid cells from SCA3/MJD patients and in two cell models of SCA3/MJD[38–40]. An association between SAFB1 expression and spinocerebellar ataxia (SCA) as well as with Huntington disease (HD) has been recently reported[41]; SAFB1 cytoplasmic immunopositivity was more frequent in cerebellar Purkinje cells from SCA patients than in controls ($p<0.05$), whereas in cerebellar dentate nucleus neurons SAFB1 expression was increased in the nucleus and cytoplasm[41]. Using a cell model of SCA1, Buckner and colleagues also have shown that SAFB1 bound significantly more to the pathogenic (ATXN85Q) mRNA[41]. Of note, SAFB1 and SAFB2 are homologous proteins, presenting high similarity and highly conserved functional domains and although they can show unique properties, they might function in a similar manner[36]. Evidence of increased expression of SAFB1 protein in Purkinje cells and dentate nucleus neurons of SCA patients is in accordance with our results for *SAFB2* mRNA levels (higher expression in patient's cerebellum as well as in blood of PA subjects compared to controls). A genome-wide study revealed a link between variants in DNA repair genes and earlier age at onset in a large cohort of polyglutamine disease patients', including SCA3/MJD[42]; authors suggested that DNA repair is compromised (by genetic variation) which can cause somatic expansions and therefore

modify age at onset[42]. Exploring the role of SAFB family as potential modifiers of DNA repair, we hypothesized that the upregulation pattern of *SAFB2* observed in MJD (higher levels in pre-ataxic carriers, higher levels in patients with earlier onset and higher levels in one year follow-up) could be associated with an inhibition of DNA repair, implying an increase of somatic expansion in blood cells (and probably also in cerebellum). The investigation of somatic mosaicism in blood and other tissues measured over time in SCA3/MJD will elucidate this hypothesis. Nevertheless, has been recently described that somatic instability in blood increased with age in blood samples of Huntington disease carriers[43], and the same observation can be expectable in SCA3/MJD.

Altered levels of *SFSWAP* and *LTBP4* were also observed in the pre-ataxic stage of SCA3/MJD, although their individual discriminative power is below clinical usefulness and no evidence of associations with disease measures in the symptomatic stage were found. SFSWAP is an RS-domain containing (SR-Like) protein, belonging to a family of proteins which participates in the regulation of RNA processing, including splicing and transcript elongation[44]. SFSWAP regulates splicing of itself and several other genes[44], including the *MAPT* gene (which encodes the Tau protein[45]). *LTBP4*, whose transcript levels were downregulated in pre-ataxic carriers, is a latent TGFβ binding protein (LTBP; LTBPs are extracellular matrix proteins, which bind and sequester TGFβ in the extracellular matrix to modulate its availability to the TGFβ receptor[46]. TGFβ1, amongst other processes, contributes to maintain neuronal survival and integrity of the central nervous system and is involved in immune functions[47]. Plasma levels of TGFβ1 were significantly reduced in asymptomatic HD subjects, whereas in patients, at different stages, levels were similar to controls[48]. Due to the modulatory link between LTBP4 and TGFβ we speculate that if *LTBP4* is lower, the availability of TGFβ will be also lower, implying that the neuroprotective role of this cytokine is compromised in SCA3/MJD.

Concerning the overt disease stage, we found a positive correlation between expression levels of *MEG3* and *TSPOAP1* with the SARA score, hence with disease severity. MEG3 is a long noncoding RNA (lncRNA), maternally expressed, with antiproliferative and TP53-stimulating functions[49]. Analyses of lncRNAs, using microarray data of caudate nucleus samples from 44 HD patients and 36 controls, revealed that *MEG3* was downregulated in HD brain[50]. However, this result failed to be confirmed in two different models of the disease[51]; *MEG3* levels were increased in the cortex region of early (6 weeks) and late (8 weeks) disease stages of R6/2 mice compared to age-matched wild-type mice. The same up-regulation tendency was observed in

mouse immortalized striatal cells expressing the full-length huntingtin gene with 111 glutamine repeats[51]. Moreover, a significant decrease of mutant huntingtin aggregates and downregulation of the endogenous TP53 protein levels in two cell lines transfected with HTT-83Q-DsRed and treated with siRNAs against *MEG3* were observed[51]. In turn, TP53 has been previously identified as a novel substrate of ataxin-3; mutated ataxin-3 abnormally interacts with TP53, leading to its upregulation and to increased TP53-dependent neuronal cell death[52]. Along with the potential role of *MEG3* as a biomarker of SCA3/MJD severity, its potential as a therapeutic target deserves further investigation.

RIMBP1 (Rab3-interacting molecule, RIM–binding protein 1), the protein encoded by *TSPOAP1*, whose expression levels we found to be correlated with SARA score, is one of the main elements of the presynaptic active zone, which in turn is a cytomatrix responsible for precise neurotransmitter release and synaptic transmission[53]. Mutations on this gene are causative of an autosomal recessive form of dystonia[54]. Motor abnormalities suggestive of dystonia were further observed in mice whose *TSPOAP1* was knocked-out, as well as alterations in the biochemical composition and morphology of dendritic arbors of Purkinje cells[54].

No transcriptional dysregulation of *SAFB2, SFSWAP, LTBP4, and TSPOAP1* in blood of SCA3/MJD patients was observed, whereas in brain the expression levels of these genes were different between patients and controls. This observation suggests that dysregulation of *SAFB2, SFSWAP, LTBP4 and TSPOAP1* seems to be tissue-specific in the overt ataxic stage; thus, our results are consistent with previous studies that showed a weak correlation at transcript level between blood and brain samples[55] (GTEx Portal on 28.01.22). Noteworthy, the dysregulation of *SAFB2*, *SFSWAP*, and *LTBP4* levels in blood samples from pre-ataxic carriers' mirrors in a better way the dysregulation observed in brain; such observation seems to indicate that blood of pre-ataxic carriers reflects more accurately transcriptional alterations of brain cells in which degenerative processes occurs. This behaviour was also described for some markers in HD, such as the case of TGFβ[48].

None of the genes identified in this RNA-seq study has been reported in the two previous transcriptional studies of blood samples from SCA3/MJD subjects, which were both conducted using an array-based approach in the discovery stage[9]. Constraints in replicating results from transcriptional biomarkers have been widely acknowledged for other polyglutamine diseases, such as HD[56]. These difficulties are usually attributed to the insufficient sample size as well as the lack of standardization in sample collection and storage[57]; however, both issues were

accounted for in our study. Cellular heterogeneity of blood, namely fluctuations of cell counts[58,59] as well as specific gene expression profiles of cell subpopulations[60] or different treatment regimens[61] could be the primary source to explain the non-replication of transcriptional biomarkers between different studies. Finally, the pleiotropic nature of SCA3/MJD, as the disease shows itself through a variety of clinical signs/symptoms and progression rates, could not be rolled out as well.

To better molecularly assess SCA3/MJD, a battery of different biomarkers should be further trained and optimized depending on the disease stage. We propose the expression levels of *SAFB2*, *SFSWAP*, *LTBP4*, *MEG3* and *TSPOAP1* as stratification markers of pre-ataxic or symptomatic disease stages, deserving further validation in longitudinal studies and in independent cohorts.

ACKNOWLEDGEMENTS

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

FUNDING

This work is an outcome of ESMI, an EU Joint Programme - Neurodegenerative Disease Research (JPND) project (see www.jpnd.eu). 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); 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 (SFRH/BD/121101/2016) and ARVM (SFRH/BD/129547/2017) has received a PhD fellowship from the FCT. Fundo Regional para a Ciência e Tecnologia (FRCT, Governo Regional dos Açores) is currently supporting ESMI in 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). Several authors of this publication are

members of the European Reference Network for Rare Neurological Diseases - Project ID No 739510.

COMPETING INTERESTS

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, Hersenstichting, Gossweiler Foundation, and Radboud university medical center; he has served on the scientific advisory board of uniQure.

The remaining authors report no competing interests.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Brain* online.

APPENDIX 1

The European Spinocerebellar Ataxia type 3/Machado-Joseph Initiative (ESMI) Study Group:

Janna Krahe, Kathrin Reetz, José González, Carlos Gonzalez, Carlos Baptista, João Lemos, Ilaria Giordano, Marcus Grobe-Einsler, Demet Önder, Patrick Silva, Cristina Januário, Joana Ribeiro, Inês Cunha, João Lemos, Maria M Pinto, Dagmar Timmann, Katharina M. Steiner, Andreas Thieme, Thomas M. Ernst, Heike Jacobi, Nita Solanky, Cristina Gonzalez-Robles, Judith Van Gaalen, Ana Lara Pelayo-Negro, Leire Manrique, Holger Hengel, Matthis Synofzik, Winfried Ilg.

REFERENCES

1. Seidel K, Siswanto S, Brunt ERP, den Dunnen W, Korf H-W, Rüb U. Brain pathology of spinocerebellar ataxias. *Acta Neuropathol*. 2012;124(1):1-21. doi:10.1007/s00401-012-1000-x

2. Riess O, Rüb U, Pastore A, Bauer P, Schöls L. SCA3: Neurological features, pathogenesis and animal models. *The Cerebellum*. 2008;(March):1-13.

doi:10.1007/s12311-008-0013-4

3. Takiyama Y, Nishizawa M, Tanaka H, et al. The gene for Machado-Joseph disease maps to human chromosome 14q. *Nat Genet*. 1993;4(3):300-304. doi:10.1038/ng0793-300

4. Kawaguchi Y, Okamoto T, Taniwaki M, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet*. 1994;8(3):221-228. doi:10.1038/ng1194-221

5. Maciel P, Costa MC, Ferro A, et al. Improvement in the molecular diagnosis of Machado-Joseph disease. *Arch Neurol*. 2001;58(11):1821-1827. doi:10.1001/archneur.58.11.1821

6. Burnett B, Li F, Pittman RN. The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity. *Hum Mol Genet*. 2003;12(23):3195-3205. doi:10.1093/hmg/ddg344

7. Da Silva JD, Teixeira-Castro A, Maciel P. From Pathogenesis to Novel Therapeutics for Spinocerebellar Ataxia Type 3: Evading Potholes on the Way to Translation. *Neurotherapeutics*. 2019;16(4):1009-1031. doi:10.1007/s13311-019-00798-1

8. Costa M do C, Paulson HL. Toward understanding Machado-Joseph disease. *Prog Neurobiol*. 2012;97(2):239-257. doi:10.1016/j.pneurobio.2011.11.006

9. Raposo M, Bettencourt C, Maciel P, et al. Novel candidate blood-based transcriptional biomarkers of machado-joseph disease. *Mov Disord*. 2015;30(7):968-975. doi:10.1002/mds.26238

10. Lima L, Coutinho P. Clinical criteria for diagnosis of Machado-Joseph disease: report of a non-Azorena Portuguese family. *Neurology*. 1980;30(3):319-322. Accessed March 7, 2013. http://www.ncbi.nlm.nih.gov/pubmed/7189034

11. Sequeiros J, Coutinho P. Epidemiology and clinical aspects of Machado-Joseph disease. *Adv Neurol*. 1993;61:139-153. Accessed January 28, 2013. http://www.ncbi.nlm.nih.gov/pubmed/8421964

12. Schmitz-Hübsch T, du Montcel ST, Baliko L, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology*. 2006;66(11):1717-1720. doi:10.1212/01.wnl.0000219042.60538.92

13. Maas RPPWM, van Gaalen J, Klockgether T, van de Warrenburg BPC. The preclinical stage of spinocerebellar ataxias. *Neurology*. 2015;85(1):96-103. doi:10.1212/WNL.0000000000001711

14. Lima M, Raposo M. *Towards the Identification of Molecular Biomarkers of Spinocerebellar Ataxia Type 3 (SCA3)/Machado-Joseph Disease (MJD)*. Vol 1049.; 2018. doi:10.1007/978-3-319-71779-1_16

15. Wilke C, Haas E, Reetz K, et al. Neurofilaments as blood biomarkers at the preataxic and ataxic stage of spinocerebellar ataxia type 3: A cross-species analysis in humans and mice. *EMBO Mol Med*. Published online January 1, 2019:19011882. doi:10.1101/19011882

16. Hübener-Schmid J, Kuhlbrodt K, Peladan J, et al. Polyglutamine-Expanded Ataxin-3: A Target Engagement Marker for Spinocerebellar Ataxia Type 3 in Peripheral Blood. *Mov Disord*. Published online August 16, 2021. doi:10.1002/mds.28749

17. Raposo M, Ramos A, Santos C, et al. Accumulation of Mitochondrial DNA Common Deletion Since The Preataxic Stage of Machado-Joseph Disease. *Mol Neurobiol*. 2019;56(1). doi:10.1007/s12035-018-1069-x

18. Li Q-F, Dong Y, Yang L, et al. Neurofilament light chain is a promising serum biomarker in spinocerebellar ataxia type 3. *Mol Neurodegener*. 2019;14(1):39. doi:10.1186/s13024-019-0338-0

19. Garcia-Moreno, Hector; Prudencio, Mercedes; Thomas-Black, Gilbert; Solanky, Nita; Jansen-West, Karen R; Hanna Al Shaikh, Rana; Heslegrave, Amanda; Zetterberg, Henrik; Santana, Magda M; Pereira de Almeida, Luis; Ferreira, Ana Cristina; Januário, Cristina; P. TAU AND NEUROFILAMENT LIGHT-CHAIN AS FLUID BIOMARKERS IN SPINOCEREBELLAR ATAXIA TYPE 3. *Submitted*.

20. Ichikawa Y, Goto J, Hattori M, et al. The genomic structure and expression of MJD, the Machado-Joseph disease gene. *J Hum Genet*. 2001;46(7):413-422. doi:10.1007/s100380170060

21. Tezenas du Montcel S, Durr A, Bauer P, et al. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain*. 2014;137(Pt 9):2444-2455. doi:10.1093/brain/awu174

22. Haas E, Incebacak RD, Hentrich T, 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*. Published online October 30, 2021. doi:10.1007/s12035-021-02610-8

23. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

24. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*. 2014;30(4):523-530. doi:10.1093/bioinformatics/btt703

25. Ferreira AF, Raposo M, Vasconcelos J, Costa MC, Lima M. Selection of Reference Genes for Normalization of Gene Expression Data in Blood of Machado-Joseph Disease/Spinocerebellar Ataxia Type 3 (MJD/SCA3) Subjects. *J Mol Neurosci*. 2019;69(3). doi:10.1007/s12031-019-01374-0

26. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc*. 2008;3(6):1101-1108. Accessed June 25, 2014. http://www.ncbi.nlm.nih.gov/pubmed/18546601

27. Evers MM, Toonen LJA, van Roon-Mom WMC. Ataxin-3 protein and RNA toxicity in spinocerebellar ataxia type 3: current insights and emerging therapeutic strategies. *Mol Neurobiol*. 2014;49(3):1513-1531. doi:10.1007/s12035-013-8596-2

28. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. Zhang S-D, ed. *PLoS One*. 2014;9(1):e78644. doi:10.1371/journal.pone.0078644

29. de Oliveira PG, Ramos MLS, Amaro AJ, Dias RA, Vieira SI. Gi/o-Protein Coupled Receptors in the Aging Brain. *Front Aging Neurosci*. 2019;11. doi:10.3389/fnagi.2019.00089

30. Bustamante-Barrientos FA, Méndez-Ruette M, Ortloff A, et al. The Impact of Estrogen and Estrogen-Like Molecules in Neurogenesis and Neurodegeneration: Beneficial or Harmful? *Front Cell Neurosci*. 2021;15. doi:10.3389/fncel.2021.636176

31. Klockgether T, Lüdtke R, Kramer B, et al. The natural history of degenerative ataxia: a retrospective study in 466 patients. *Brain*. 1998;121 (Pt 4:589-600. http://www.ncbi.nlm.nih.gov/pubmed/9577387

32. Jacobi H, du Montcel ST, Bauer P, et al. Long-term disease progression in

spinocerebellar ataxia types 1, 2, 3, and 6: a longitudinal cohort study. *Lancet Neurol*. 2015;14(11):1101-1108. doi:10.1016/S1474-4422(15)00202-1

33. Roderick P.P.W.M. Maas, Steven Teerenstra, Manuela Lima, Luis Pereira de Almeida, Judith van Gaalen, Dagmar Timmann, Jon Infante, Khalaf Bushara, Chiadi Onyike, Heike Jacobi, Kathrin Reetz, Magda Santana, Jeanette Hübener-Schmid, Jeroen de Vries, Ludger S BPC van de W. Differential temporal dynamics of axial and appendicular ataxia in SCA3. *Accepted*. doi:10.1002/mds.29135

34. Li Q-F, Dong Y, Yang L, et al. Neurofilament light chain is a promising serum biomarker in spinocerebellar ataxia type 3. *Mol Neurodegener*. 2019;14(1):39. doi:10.1186/s13024-019-0338-0

35. Norman M, Rivers C, Lee Y-B, Idris J, Uney J. The increasing diversity of functions attributed to the SAFB family of RNA-/DNA-binding proteins. *Biochem J*. 2016;473(23):4271-4288. doi:10.1042/BCJ20160649

36. Townson SM, Dobrzycka KM, Lee A V., et al. SAFB2, a New Scaffold Attachment Factor Homolog and Estrogen Receptor Corepressor. *J Biol Chem*. 2003;278(22):20059-20068. doi:10.1074/jbc.M212988200

37. Oesterreich S. Scaffold attachment factors SAFB1 and SAFB2: Innocent bystanders or critical players in breast tumorigenesis? *J Cell Biochem*. 2003;90(4):653-661. doi:10.1002/jcb.10685

38. Wen F-C, Li Y-H, Tsai H-F, et al. Down-regulation of heat shock protein 27 in neuronal cells and non-neuronal cells expressing mutant ataxin-3. *FEBS Lett*. 2003;546(2-3):307-314. doi:10.1016/S0014-5793(03)00605-7

39. Chang WH, Cemal CK, Hsu YH, et al. Dynamic expression of Hsp27 in the presence of mutant ataxin-3. *Biochem Biophys Res Commun*. 2005;336(1):258-267. doi:10.1016/j.bbrc.2005.08.065

40. Evert BO, Vogt IR, Vieira-Saecker AM, et al. Gene expression profiling in ataxin-3 expressing cell lines reveals distinct effects of normal and mutant ataxin-3. *J Neuropathol Exp Neurol*. 2003;62(10):1006-1018. Accessed March 6, 2014. http://www.ncbi.nlm.nih.gov/pubmed/14575237

41. Buckner N, Kemp KC, Scott HL, et al. Abnormal scaffold attachment factor 1 expression and localization in spinocerebellar ataxias and Huntington's chorea. *Brain*

Pathol. 2020;30(6):1041-1055. doi:10.1111/bpa.12872

42. Bettencourt C, Hensman-Moss D, Flower M, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann Neurol*. 2016;79(6). doi:10.1002/ana.24656

43. Kacher R, Lejeune F-X, Noël S, et al. Propensity for somatic expansion increases over the course of life in Huntington disease. *Elife*. 2021;10. doi:10.7554/eLife.64674

44. Twyffels L, Gueydan C, Kruys V. Shuttling SR proteins: more than splicing factors. *FEBS J*. 2011;278(18):3246-3255. doi:10.1111/j.1742-4658.2011.08274.x

45. Gao QS, Memmott J, Lafyatis R, Stamm S, Screaton G, Andreadis A. Complex regulation of tau exon 10, whose missplicing causes frontotemporal dementia. *J Neurochem*. 2000;74(2):490-500. doi:10.1046/j.1471-4159.2000.740490.x

46. Su C-T, Urban Z. LTBP4 in Health and Disease. *Genes (Basel)*. 2021;12(6). doi:10.3390/genes12060795

47. Meyers EA, Kessler JA. TGF-β Family Signaling in Neural and Neuronal Differentiation, Development, and Function. *Cold Spring Harb Perspect Biol*. 2017;9(8):a022244. doi:10.1101/cshperspect.a022244

48. Battaglia G, Cannella M, Riozzi B, et al. Early defect of transforming growth factor β1 formation in Huntington's disease. *J Cell Mol Med*. 2011;15(3):555-571. doi:10.1111/j.1582-4934.2010.01011.x

49. Zhang X, Rice K, Wang Y, et al. Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: isoform structure, expression, and functions. *Endocrinology*. 2010;151(3):939-947. doi:10.1210/en.2009-0657

50. Johnson R. Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiol Dis*. 2012;46(2):245-254. doi:10.1016/j.nbd.2011.12.006

51. Chanda K, Das S, Chakraborty J, et al. Altered Levels of Long NcRNAs Meg3 and Neat1 in Cell And Animal Models Of Huntington's Disease. *RNA Biol*. 2018;15(10):1348-1363. doi:10.1080/15476286.2018.1534524

52. Liu H, Li X, Ning G, et al. The Machado-Joseph Disease Deubiquitinase Ataxin-3 Regulates the Stability and Apoptotic Function of p53. *PLoS Biol*. 2016;14(11):e2000733. doi:10.1371/journal.pbio.2000733

53. Liu KSY, Siebert M, Mertel S, et al. RIM-Binding Protein, a Central Part of the Active Zone, Is Essential for Neurotransmitter Release. *Science (80-).* 2011;334(6062):1565-1569. doi:10.1126/science.1212991

54. Mencacci NE, Brockmann MM, Dai J, et al. Biallelic variants in TSPOAP1, encoding the active-zone protein RIMBP1, cause autosomal recessive dystonia. *J Clin Invest*. 2021;131(7). doi:10.1172/JCI140625

55. Cai C, Langfelder P, Fuller TF, et al. Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics*. 2010;11(1):589. doi:10.1186/1471-2164-11-589

56. Hensman Moss DJ, Flower MD, Lo KK, et al. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. *Sci Rep*. 2017;7(1):44849. doi:10.1038/srep44849

57. Mastrokolias A, Ariyurek Y, Goeman JJ, et al. Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *Eur J Hum Genet*. 2015;23(10):1349-1356. doi:10.1038/ejhg.2014.281

58. Whitney AR, Diehn M, Popper SJ, et al. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci*. 2003;100(4):1896-1901. doi:10.1073/pnas.252784499

59. Di Pardo A, Alberti S, Maglione V, et al. Changes of peripheral TGF-β1 depend on monocytes-derived macrophages in Huntington disease. *Mol Brain*. 2013;6(1):55. doi:10.1186/1756-6606-6-55

60. Xie X, Liu M, Zhang Y, et al. Single-cell transcriptomic landscape of human blood cells. *Natl Sci Rev*. 2021;8(3). doi:10.1093/nsr/nwaa180

61. Liu X, Zeng P, Cui Q, Zhou Y. Comparative analysis of genes frequently regulated by drugs based on connectivity map transcriptome data. Zou Q, ed. *PLoS One*. 2017;12(6):e0179037. doi:10.1371/journal.pone.0179037

FIGURE LEGENDS

Figure 1 Workflow of the study. We performed a cross-sectional RNA-seq experiment using whole blood from pre-ataxic subjects (PA), patients (P) and controls. To identified common enriched pathways (**A**) of both tissues, we overlapped our data with RNA-seq datasets from *post-mortem* cerebellum samples of six MJD patients and six controls previously obtained [22]. To select expression alterations (biomarker study - **B**) that correlate with disease onset, severity, or progression (including the pre-ataxic stage), RNA-seq data was used to: (**i**) compare gene expression levels between PA subjects and controls (analysis of covariance with age as covariate and log2 transforming all variables prior to the test); and (**ii**) correlate gene expression levels and SARA scores in patients (partial Spearman rank correlation); the potential effects of age, number of CAG repeats in the expanded allele and disease duration were statistically removed in the partial Spearman rank correlations (statistical analyses were run at R version 3.6.2 and a significance level of 5% were considered). To further identify alterations which could simultaneously distinguish PA from controls and correlate with SARA scores in patients, DE genes from (i) were intersected with DE genes from (ii) which resulted in a set of 62 common genes (Suppl. Table 1). Ten candidate genes (prioritization criteria are provided in Suppl. Table 2) were selected to be further tested by qPCR.

Figure 2 *SAFB2* expression levels in SCA3/MJD. (a) *SAFB2* levels were significantly increased in pre-ataxic carriers compared to age-matched controls; (b) levels of *SAFB2* allowed to significantly distinguish pre-ataxic carriers and age-matched controls with an accuracy of 0.71; (c) SCA3/MJD patients with an earlier age at onset presented higher levels of *SAFB2*; (d) in patients, levels of *SAFB2* from the second visit (median=1.41) were, in average, significantly higher than those from visit 1 (median=1.06); the difference of expression values (range) between visits for each pair of patients is also shown.

Figure 3 *SFSWAP* and *LTBP4* expression levels in SCA3/MJD. (a) *SFSWAP* levels were significantly increased in pre-ataxic carriers compared to age-matched controls. (b) Levels of *LTBP4* were significantly decreased in pre-ataxic carriers (PA) compared to age-matched controls (CTRL-PA). (c) combined levels of *SAFB2*, *SFSWAP* and *LTBP4* allowed to significantly distinguish pre-ataxic carriers and age-matched controls with an accuracy of 0.79;

individual ROC curves of *SAFB2* (Fig.1b), *SFSWAP* (AUC=0.65, 95%CI [0.519-0.782], p=0.034) and *LTBP4* (AUC=0.65, 95%CI [0.504-0.796], p=0.047) are also shown.

Figure 4 Expression behaviour of *MEG3* and *TSPOAP1* in SCA3/MJD. SCA3/MJD patients with higher SARA scores had higher levels of (a) *MEG3* and (b) *TSPOAP1*.

Table 1 Characterization of the participants (controls, pre-ataxic subjects, and patients) used in this study

	Controls	Pre-ataxic subjects	Patients	
RNA-seq experiments (*n* = 60)				
Sample size, *n*	20	10	30	
Gender (Female:Male)	10:10	5:5	15:15	ns
Age, years	49 [33.3–62.3]	36 [30–40]	51.5 [42.8–61.5]	C ≠ PA; PA ≠ P
CAG_n allele 1	14.5 [14–23]	18.5 [14–23]	23 [20–26.3]	C ≠ P
CAG_n allele 2	23 [23–27]	69 [64–71]	70.5 [66.5–72.3]	ns[a]
Age at onset (AO), years	na	[b]	38.5 [26.3–46.8][c]	na
SARA score	0 [0–0.5]	1 [0–1.1]	17.8 [5.9–28.5]	PA ≠ P; C ≠ P
qPCR analyses (*n* = 290)				
Sample size, n				
Visit 1	51 (all) 24 (CTRL-PA)[d] 27 (CTRL-P)[d]	29	129	na
Visit 2	na	12	62	na
Gender (Female:Male)				
Visit 1	29:22 (all) 14:10 (CTRL-PA) 15:12 (CTRL-P)	19:10	65:64	ns
Visit 2	na	9:3	32:30	na
Age, years				
Visit 1	42 [33–56] (all) 32.5 [29–40] (CTRL-PA) 56 [46–61] (CTRL-P)	35 [29–39.5]	52 [44–59]	ns
Visit 2	na	33.5 [25–39]	50.5 [44–58]	na
CAG_n allele 1				
Visit 1	22 [14-23]	22.5 [20–26.3][c]	23 [17.8–24][c]	ns
Visit 2	na	21 [17–24]	23 [14–25][c]	na
CAG_n allele 2				
Visit 1	24 [23–27]	69 [66–71][c]	69 [66–71][c]	ns
Visit 2	na	69 [67–71]	70 [68–72][c]	na
Time to preAO, years				
Visit 1	na	−8 [−12 to −6]	na	na
Visit 2		−11 [−12 to −7]	na	na
Age at onset (AO), years				
Visit 1	na	[e]	38 [33–46][c]	na
Visit 2	na	[e]	37 [32.5–44.5][c]	na
Disease duration (DD), years				
Visit 1	na	[e]	11 [7–16][c]	na
Visit 2	na	[e]	11 [7–16.5][c]	na
SARA score				
Visit 1	*n* = 44 0 [0–0.88]	1 [0.25–2] subcohort 1 [0–2]	12.5 [9–22] subcohort 13 [9–22]	PA ≠ P; C ≠ P
Visit 2	na	1 [0–2]	15 [10–23]	
		ns	≠	

Continuous variables are shown as median [Interquartile range: 1stQ–3rdQ]. A chi-square test of independence was used to compare the proportion of subjects by gender and biological groups (pre-ataxic subjects, patients, and controls). Differences between biological groups on age, the number of CAG repeats in *ATXN3*, AO, and SARA score, were determined by Mann-Whitney U or Kruskal-Wallis tests. Differences between visit 1 and visit 2 for SARA score were calculated by Wilcoxon matched pairs signed rank test. Significant differences were lower than 0.05 (≠); ns= not statistically significant; na= not applicable. Sub-cohort= the number of subjects whose data and blood samples were also available at a second annual visit (visit 2).

[a]Differences were only assessed between pre-ataxic subjects and patients.

[b]Age at disease onset was reported by four pre-ataxic carriers.

[c]This variable was missing at a proportion between 3–6% of total sample size.

[d]To account for age and gender (potential cofounders), two sub-sets of controls were formed: controls matched to pre-ataxic carriers (CTRL-PA) and controls matched to patients (CTRL-P).

[e]Age at disease onset was reported by three pre-ataxic carriers.

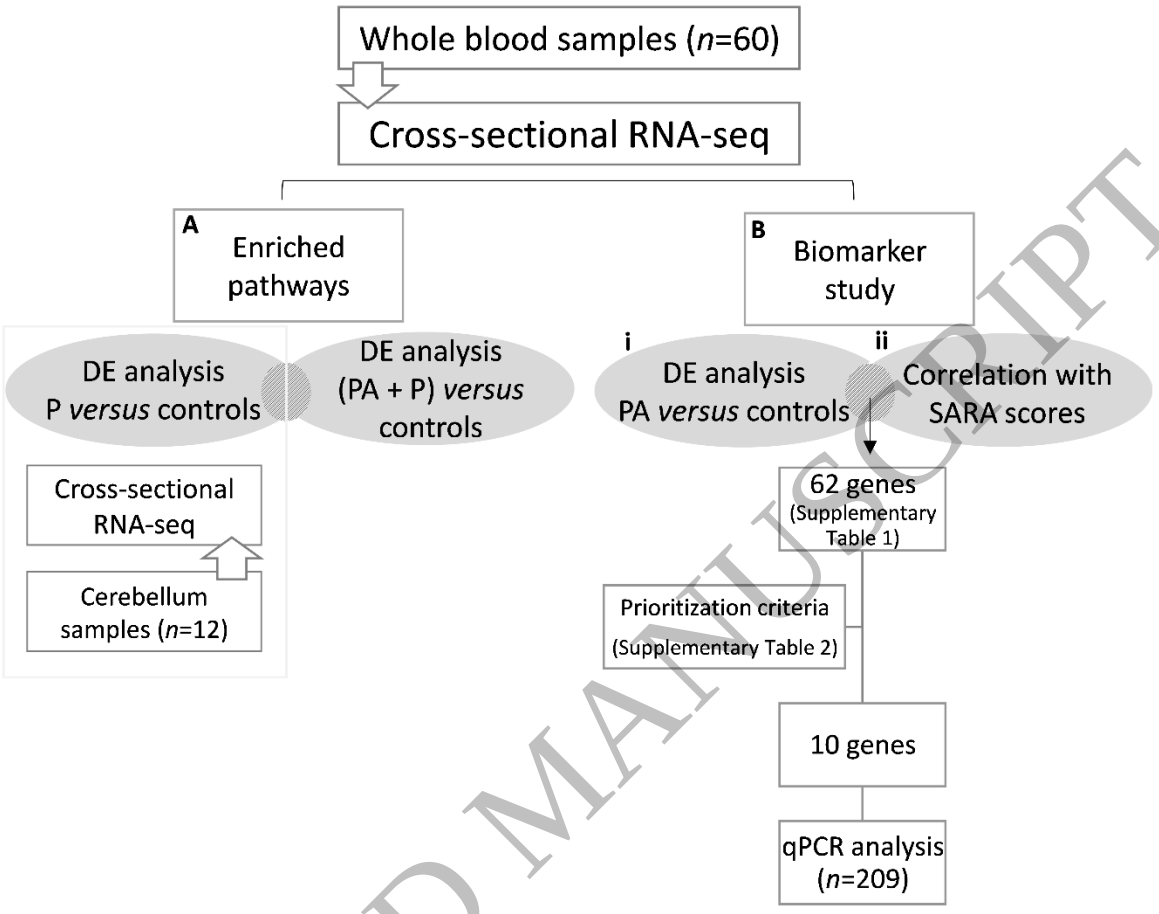


Figure 1
159x126 mm (x DPI)

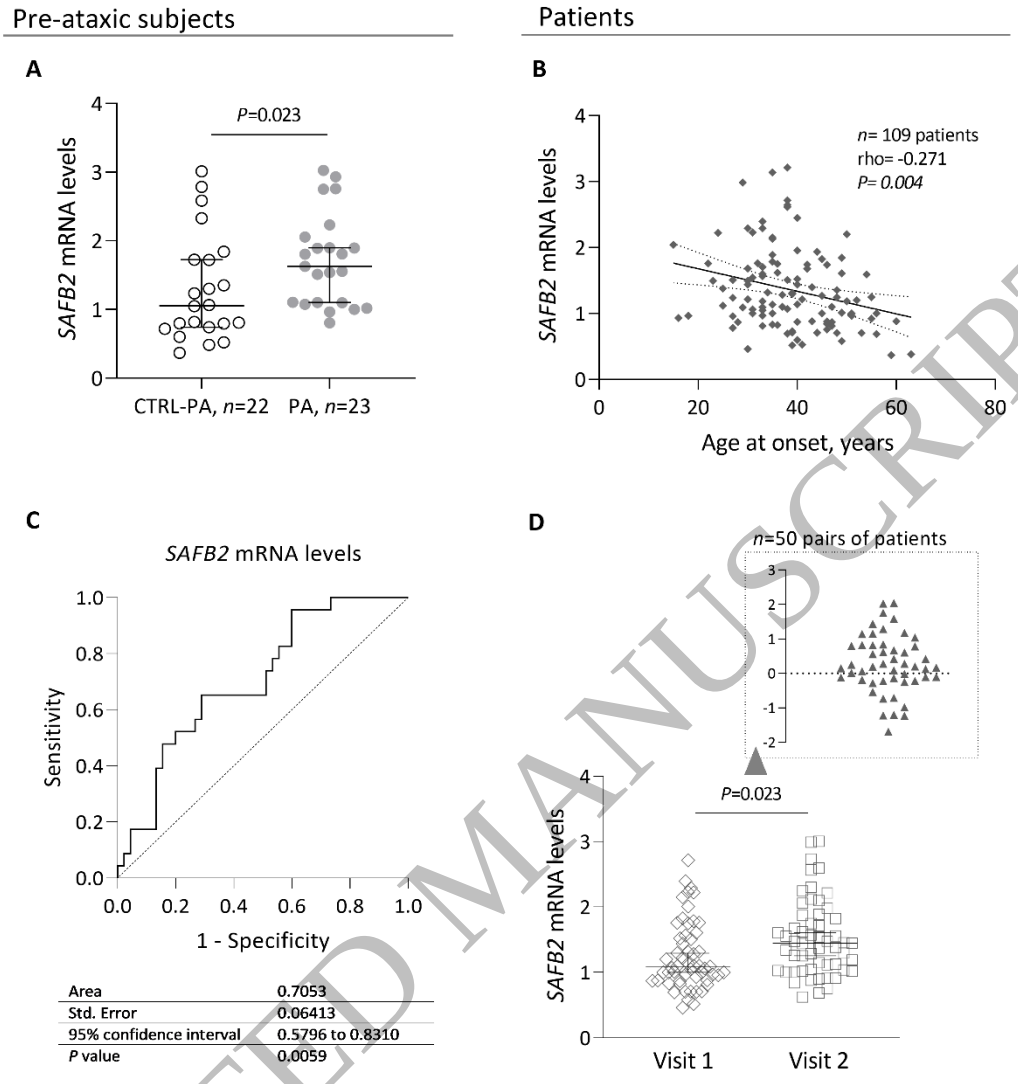
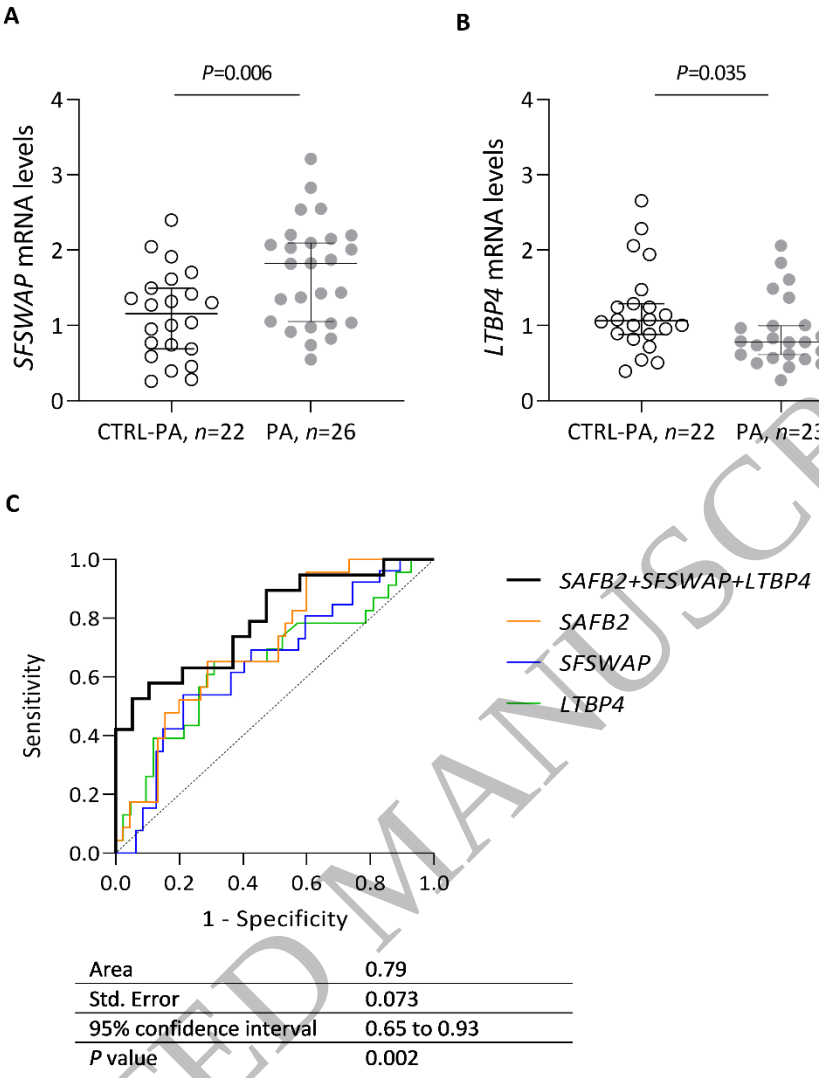


Figure 2
140x150 mm (x DPI)



Area	0.79
Std. Error	0.073
95% confidence interval	0.65 to 0.93
P value	0.002

Figure 3
125x150 mm (x DPI)

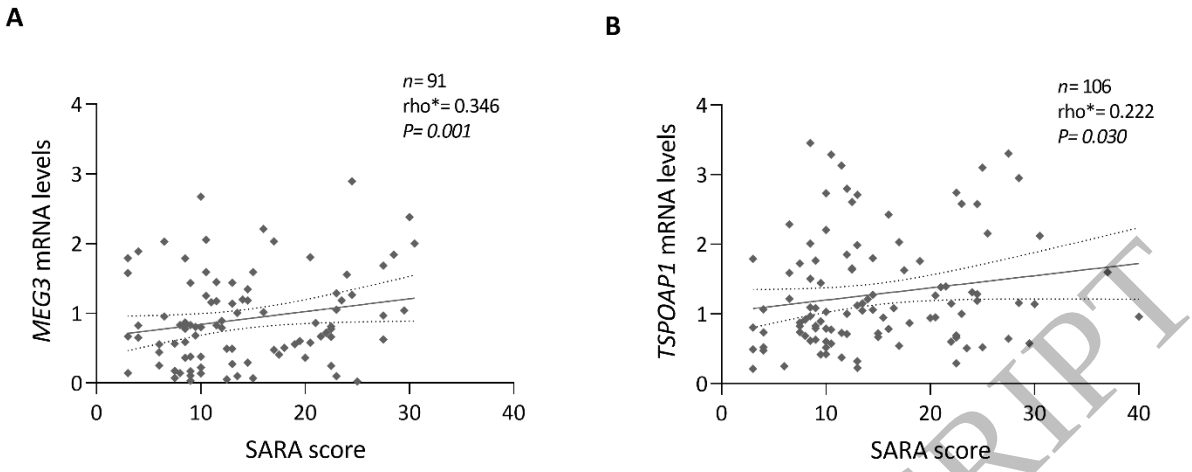


Figure 4
159x66 mm (x DPI)

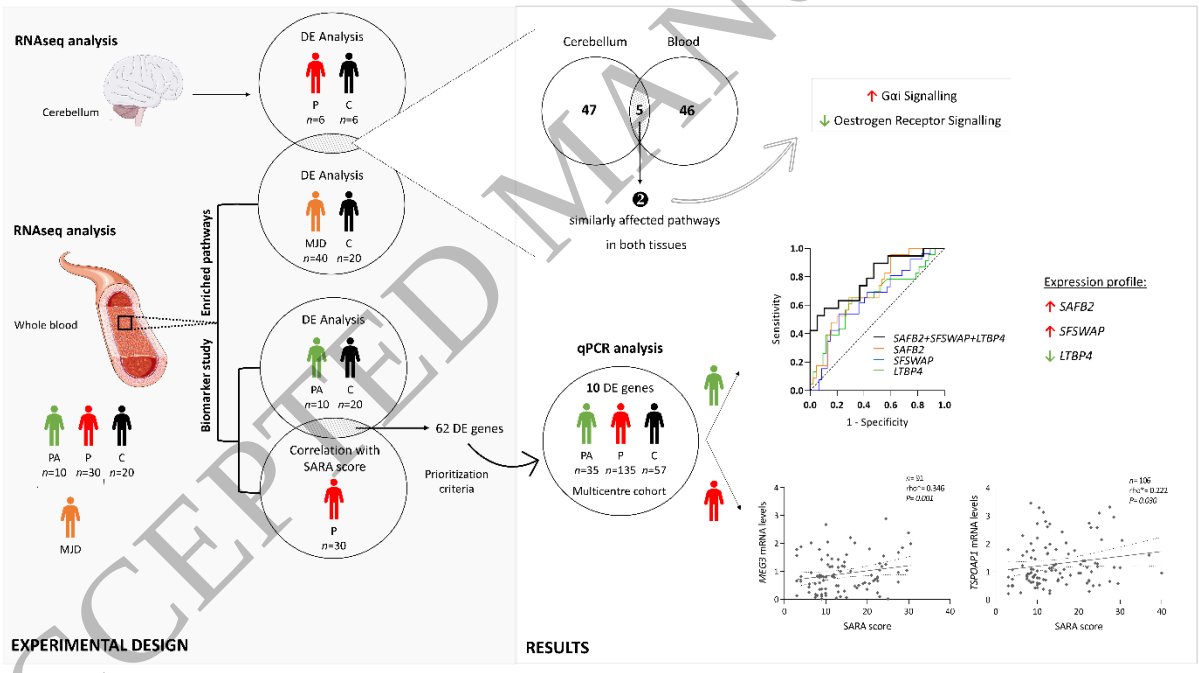


Figure 5
159x88 mm (x DPI)