TAF1-dependent transcriptional dysregulation underlies multiple sclerosis

Claudia Rodríguez-López[1,2‡], Ivó H. Hernández[1,2‡], José Terrón-Bautista[3¶], Eneritz Agirre[4¶], David Lozano-Muñoz[1,2], Julia Pose-Utrilla[1,2], Inés García-Ortiz[1], Miriam Lucas-Santamaría[1,2], Irene Ruiz-Blas[1,2], María González-Bermejo[1,2], María C. Ortega[2,5], Juan C. Chara[2, 6], Zara Martínez[2,6], Fernando Pérez-Cerdá[2,6], Nina L. Fransen[7,8], Miriam Martínez-Jiménez[1], Beatriz Sancho-González[1,2], Alberto Pérez-Samartín[2,6], Mukund Kabbe[4], Chao Zheng[4], Marcos Casado-Barbero[1,2], María Santos-Galindo[1,2], Aldo Borroto[1], Balbino Alarcón[1], Diego Clemente[2,5], Claudio Toma[1,9,10], Carlos Matute[2,6], Felipe Cortés-Ledesma[3], Gonçalo Castelo-Branco[4] and José J. Lucas[1,2]*

[1] Center for Molecular Biology Severo Ochoa (CBM) CSIC/UAM, Madrid 28049, Spain

[2] Networking Research Center on Neurodegenerative Diseases (CIBERNED). Instituto de Salud Carlos III, Madrid 28031, Spain

[3] Topology and DNA breaks group, Spanish National Cancer Research Centre (CNIO), Madrid, 28029 Spain.

[4] Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.

[5]Neuroimmuno-Repair Group, Hospital Nacional de Parapléjicos (HNP), SESCAM-IDISCAM, Finca "La Peraleda" s/n, 45071, Toledo, Spain

[6]Achucarro Basque Center for Neuroscience and Department of Neuroscience, University of the Basque Country UPV/EHU, 48940, Leioa, Spain

[7]Department of Pathology, Amsterdam UMC location Vrije Universiteit Amsterdam, De Boelelaan 1117, Amsterdam, The Netherlands.

[8]Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands.

[9] Neuroscience Research Australia, Sydney, NSW, Australia.

[10] School of Clinical Medicine, Discipline of Psychiatry and Mental Health, University of New South Wales, Sydney, NSW, Australia.

‡ / ¶: These authors contributed equally

* Corresponding author: José J. Lucas
Center for Molecular Biology Severo Ochoa (CBM)
C/ Nicolás Cabrera,1. Campus UAM de Cantoblanco
28049 Madrid, Spain
e-mail: jjlucas@cbm.csic.es

Abstract

Multiple sclerosis (MS) is characterized by neuroinflammation and demyelination of the central nervous system (CNS), leading to disablility[1, 2]. Genetic variants that confer MS risk implicate genes involved in immune function[3], while variants related to severity of the disease are associated with genes preferentially expressed within the CNS[4]. Current MS therapies decrease relapse rates by preventing immune-mediated damage of myelin, but they ultimately fail to slow long-term disease progression, which apparently depends on CNS intrinsic processes. The molecular events that trigger progressive MS are still unknown. Here we report that the C-terminal region of TAF1 (the scaffolding subunit of the general transcription factor TFIID) is underrepresented in postmortem brain tissue from individuals with MS. Furthermore, we demonstrate *in vivo,* in genetically modified mice, that C-terminal alteration of TAF1 suffices to induce an RNA polymerase II (RNAPII)-elongation deficit that particularly affects oligodendroglial myelination-related genes and results in an MS-like brain transcriptomic signature, including increased expression of proinflammatory genes. This transcriptional profile is accompanied by CNS-resident inflammation, robust demyelination and MS-like motor phenotypes. We also identify numerous interactors of C-terminal TAF1 that participate in RNAPII-promoter escape, of which two show evidence for genetic association to MS[3, 4]. Our study reveals that TAF1 dysfunction converges with genetic susceptibility to cause transcriptional dysregulation in CNS cell types, such as oligodendrocytes, to ultimately trigger MS.

Introduction

MS is a disabling neurological disease characterized by neuroinflammation and demyelination affecting the brain, the spinal cord and the optic nerves, profoundly reducing quality of life for the majority of affected individuals, which is about 2.8 million worldwide[1, 2]. Most individuals debut with partially or fully reversible episodes of neurological deficits (relapses) in the clinical form known as relapsing–remitting MS (RRMS). A great proportion of these evolve to the secondary progressive MS (SPMS) clinical form, with the development of permanent neurological deficits and the continuous progression of clinical disability. Approximately 15% of individuals with MS have a progressive course from disease onset, which is referred to as primary progressive MS (PPMS)[5].

Although the precise etiology of MS remains unknown, epidemiology indicates that it is a complex disease influenced by both environmental and genetic factors[5]. Environmental factors include Epstein-Barr virus (EBV) infection in adolescence and early adulthood, smoking, lack of sun exposure and low vitamin D levels[5]. The genetic architecture of MS consists of hundreds of small-effect common variants that increase disease risk[3] or severity[4]. While risk variants implicate genes strongly enriched for immune relevance[3], MS severity is associated with variation in genes preferentially expressed within the CNS[4]. This suggests that processes related to severity leading to permanent disability originate from the CNS, explaining the lack of efficacy of the current available immunological therapies to prevent or halt the final progression of the disease independent of relapse activity.

In the search for new targets for therapeutic intervention, MS genome-wide association studies (GWAS) have provided clues about etiology-relevant molecular mechanisms, but the individual effect of each of the identified variants is minimal. Thus, a current priority is to detect specific biochemical alterations in the CNS of MS patients, related to environmental factors and/or to GWAS-identified molecular mechanisms, that significantly contribute to inducing MS progression.

The TATA-box binding protein associated factor 1 (TAF1) protein is the largest and scaffolding subunit of transcription factor IID (TFIID), the general transcription factor that makes contacts with core promoter DNA elements to aid to define the transcription start site (TSS), and

coordinates the formation of the RNAPII transcriptional preinitiation complex (PIC) on all protein-coding genes[6]. TAF1 also participates in promoter-proximal pausing of RNAPII[7]. In line with its inferred essential role in the initial steps of transcription, TAF1 is among the top loss-of-function intolerant genes (ranked 53rd of the top 1,003 constrained human genes)[8]. Importantly, *TAF1* mutations have been implicated in neural diseases: germ line point mutations lead to the MRXS33 X-linked intellectual disability (ID) syndrome[9] (OMIM: 300966) and an inherited retrotransposon insertion in intron 32 of *TAF1* causes X-linked dystonia-parkinsonism (XDP) syndrome[10, 11] (OMIM: 314250), the latter presumably through decreasing TAF1 expression. Furthermore, we have also recently reported decreased TAF1 protein levels in Huntington's disease attributable to TAF1 splicing alterations that affect its protein stability[12].

A cell type–specific interaction analysis of MS risk loci, used to determine the relevant pathogenic cell types for MS, found three MS associated loci to be pathogenic specifically in the oligodendrocyte lineage, of which two are involved in the release of RNAPII from promoter proximal pausing[13]. We thus wondered whether TAF1 could have a role in MS pathogenesis.

Here we show that the C-terminal end of TAF1 is underrepresented in the postmortem brain of individuals with MS. We then generated a mouse model (Taf1d38 mice) with deleted C-terminal end of TAF1, which exhibits an MS-like brain transcriptomic signature accompanied by CNS-resident inflammation, robust demyelination and MS-like motor phenotypes. Through proteomics analyses of differential interactors with full-length TAF1/TAF1d38, and ChIP-seq analyses of total and elongating RNAPII, we uncovered a regulatory role of C-terminal TAF1 in RNAPII promoter escape that particularly affects transcriptional elongation of oligodendroglial myelination–related genes.

Results

Decreased detection of the C-terminal end of TAF1 in brains from individuals with MS

To explore whether TAF1 is altered in CNS tissue where the neurodegenerative process is still not evident, we investigated its protein levels on postmortem brain cortical samples containing normal-appearing grey plus white matter (NAG+WM) from individuals with PPMS or SPMS (Extended Data Table 1 contains the demographic data of analysed cases and controls), with antibodies against different parts of the TAF1 protein. The human *TAF1* gene is located on chromosome X and encompasses 38 exons that encode the 1893–amino acid canonical form of the protein (Uniprot P21675-2) (Fig. 1a). We used three commercial antibodies which were raised against the following amino acid sequences: 103 to 123 (encoded by exon 3), 1771 to 1821 (encoded by exon 37), or 1821 to 1871 (encoded by exon 38) (Fig. 1a). Strikingly, we found a specifically reduced immunodetection at the C-terminal end of TAF1 with the antibody raised against amino acids encoded by the last exon (exon 38), in both PPMS and SPMS samples (Fig. 1b).

We then performed a bioinformatics analysis of the C-terminal sequence of TAF1. In agreement with a previous prediction using AlphaFold[14], the disorder prediction tool PrDOS depicts the secondary structure of the entire sequence after the tandem bromo domains (BrDs) as disordered (Extended Data Fig. 1a). We observed a high degree of conservation at the BrDs (encoded by exons 28–32), followed by a valley of low homology. Notably, conservation reappears at the particularly unstructured (disorder probability > 0.6) region downstream of mid-exon 36, suggesting the existence of functional domains at the C-terminal end of TAF1. Finally, the sequence encoded by exon 38 is particularly rich in conserved glutamic and aspartic residues (Extended Data Fig. 1b,c). Overall, our results show decreased detection of the C-terminal end of TAF1 (which is highly conserved, intrinsically disordered and acidic), with no changes in TAF1 total levels, in postmortem brain tissue from individuals with progressive MS.

We then performed immunohistochemistry on sections from snap-frozen brain tissue containing normal-appearing grey matter (NAGM), normal-appearing white matter (NAWM) and demyelinating lesions (Fig. 1c,d). We used the two Human Protein Atlas–validated TAF1 antibodies, raised against the C-terminal region (ex37+ex38 Ab) or the whole protein (whole TAF1 Ab). In control individuals, the ex37+ex38 Ab recognized the nuclei and (to a lesser extent)

the cytoplasm of both glial cells and neurons. The signal intensity of ex37+ex38 Ab showed a tendency to decrease in both NAGM and NAWM of PPMS samples. Notably, lower labelling intensity in NAGM significantly correlated with shorter disease duration (Fig. 1c). In contrast, the whole TAF1 Ab selectively stained the entire surface of the core of lesions (Fig. 1d). These results further support diminished detection of the C-terminal end of TAF1 in NAGM and NAWM of individuals with MS and suggest an even more pronounced TAF1 alteration associated to demyelinated lesions.

Progressive disability in mice lacking the C-terminal end of TAF1

To explore the function of this conserved sequence at the C-terminal end of TAF1 and its possible pathological relevance *in vivo*, we generated a genetically modified mouse model that mimics the decreased detection of this region in MS brains. For this, we emulated one of the human alternative splicing events that generates TAF1 isoforms lacking all or most of exon 38–encoded sequences (Extended Data Fig. 2a). More precisely, we mimicked the retention of intron 37, which results in truncation of the TAF1 protein after the exon 37–encoded section (Extended Data Fig. 2b), with the only change of the last amino acid (S1820R). We designed a CRISPR-Cas9 strategy to eliminate the reported splicing acceptor sites in exon38 and in the canonical 3'UTR. For this, we generated a guide RNA mapping to intron 37 and another mapping downstream of the canonical 3'UTR (the latter within an adjacent and longer cryptic alternative 3'UTR) (Fig. 2a). Simultaneous injection of both guide RNAs into single-cell mouse embryos led to a mouse line with a *Taf1d38* allele in which the 4,246 base-pair sequence flanked by both RNA guides sites was deleted, with no point mutations (Fig. 2a). RNA-seq profiles of hemizygous males or homozygous females for the *Taf1d38* allele showed that the remaining intron 37 sequence and the distal UTR formed a new 3'UTR immediately downstream of the canonical exon 37 (Fig. 2b). By Western blot, we verified that there was no signal with the exon38 antibody but that the total TAF1 levels remain unaltered (using the N-terminal [exon3] antibody) (Fig. 2c).

Mice carrying at least one *Taf1d38* allele were born at the expected frequencies and they looked normal until the age of 3 months, when hemizygous males and homozygous females began to show a visible progressive disability phenotype mainly affecting the hindlimbs, mimicking some of the symptomatology associated with mouse MS models such as the experimental autoimmune encephalomyelitis (EAE) model[15]. At 3 months, *Taf1d38* hemizygous and

homozygous mice start showing episodes of tail tremor and of lowered pelvis during walking. At 6 months, the action tremor of the tail became permanent and the episodes of lowered pelvis more frequent. Kyphosis was evident in some mice at 6 months and increased in the percentage of mice in the following months. At 10 months, the tail tremor was accompanied by lower body shivering that extended to the whole trunk at 12 months. The phenotype then progressed to paraplegia, which was frequent in 20-month-old mice (Fig. 2d and Extended Data Video). For this reason, we established 18 months as the humane endpoint for the entire colony of mice with the *Taf1del38* allele (including heterozygote females whose visible phenotype is subtle).

We decided to continue only with mice with full suppression of exon 38 (i.e., hemizygous males and homozygous females), hereby termed Taf1d38 mice. In line with their evident phenotype of progressive disability, Taf1d38 mice also showed decreased activity (both ambulatory and vertical) in the open field test (Fig. 2e), progressive hindlimb clasping (Fig. 2f) and motor coordination deficit in the rotarod test (Fig. 2g). Together, these results indicate that deletion of the C-terminal portion of TAF1, which is underdetected in human MS brains, results in a progressive disabling phenotype.

Massive demyelination in Taf1d38 mice

We observed a striking loss of the whitish appearance of spinal cords of 12-month-old Taf1d38 mice (Fig. 3a), indicating massive demyelination, which was further corroborated by FluoroMyelin staining (Fig. 3b). Indeed, demyelination was already observed in the spinal cord of 7-month-old Taf1d38 mice, as evidenced by transmission electron microscopy (with significant increases of the G-ratios in anterolateral, lateral and dorsal columns; Fig. 3c). In the brain, we observed an apparent decreased staining with BlackGold reagent in corpus callosum and superior colliculus at 6 months (Fig. 3d), and a significant decrease of FluoroMyelin staining at 12 months (Extended Data Fig. 3a).

We next explored whether demyelination in Taf1d38 mice was mirrored by a decrease in conduction velocity in white matter tracts, as reported for the EAE mouse model[16]. Electrophysiological recordings of the fornix of Taf1d38 mice showed a significant increase in the latency of the first peak, which corresponds to conduction through myelinated fibers, and an unaltered latency of the second peak (unmyelinated fibers) (Fig. 3e). This demonstrated a reduced conduction velocity in the fornix that specifically affected the myelinated axons.

White matter neuroinflammation in Taf1d38 mice is driven by resident glia

Immunohistochemistry of sagittal brain sections of Taf1d38 mice with the microglia/macrophage lineage markers CD68 and Iba1 revealed a profuse staining in most white matter tracts, such as those in the brainstem, the superior colliculus and the corpus callosum (Fig. 3d,f). Immunostainings for two additional markers of MS lesions—namely, GFAP and Spp1 (OPTN, osteopontin)—yielded patterns similar to those of CD68 and Iba1 (Fig. 3d,f and Extended Data Fig. 3b). The microglial activation in Taf1d38 mice did not appear to be accompanied by infiltration of T or B lymphocytes, as these were not detected by immunofluorescence using the CD3 and B220 markers, not even in the brain regions of Taf1d38 mice with profuse CD68 and Iba1 staining (Extended Data Fig. 3c). Together, these results demonstrate CNS resident neuroinflammation, demyelination and decreased conduction velocity affecting the brain and the spinal cord of Taf1d38 mice.

Taf1d38 causes transcriptomic alteration of neuroinflammation and demyelination pathways

To investigate the molecular mechanism underlying the MS-like phenotype of Taf1d38 mice, we first examined whether eliminating the conserved C-terminal domain in the scaffolding subunit of the general transcription factor TFIID would affect transcription. We performed bulk tissue RNA-seq analysis of brain samples from 8-week-old (before onset of apparent disability) Taf1d38 mice. We analysed the striatum, a grey and white matter structure that shows atrophy and demyelination in MS[17, 18]. Strikingly, only 1.8% of the 15,627 detected genes showed altered transcript levels in Taf1d38 mice with respect to wild-type littermates (Fig. 4a and Extended Data Table 2), with 163 genes downregulated and 131 upregulated (adjusted p-value <0.05). Gene ontology analysis (with Ingenuity Pathway Analysis, IPA) on the differentially expressed genes (DEG) with log2FC ≥ |0.5| revealed that the canonical pathways with the most significant enrichments [–log10(p-value) ≥ 3] were related to inflammation, pathogenesis of MS and vitamin D receptor/retinoid X receptor activation (Fig. 4b and Extended Data Table 3), thus coinciding with key pathways that have been implicated in MS neuropathology and as environmental risk factors[5, 19]. These results suggested that the observed progressive disability phenotype of Taf1d38 mice is due to a very selective alteration in the expression of a restrictive set of genes within MS related pathways.

Upregulated genes included some of the markers of the neuroinflammation associated to demyelination in MS plaques, such as *Spp1* (osteopontin, OPTN) and *GFAP*, as well as many other immunoinflammation-related genes, including *H2-D1*, *H2-K1*, *Ccl4*, *Cxcl10*, *Cd52*, *B2m* and *C4b* (Fig. 4c). In contrast, oligodendroglial-enriched genes[20], such as *Ptgds*, *Anln*, *Il33* and *Klk6*, were among the downregulated transcripts (Fig. 4c).

We next specifically analysed the similarities between the DEG transcriptomic signature of Taf1d38 mice and the signatures of NAG+WM or active lesions from individuals with MS. For NAG+WM, we performed bulk tissue RNA-seq analysis on samples from controls and individuals with PPMS (Fig. 4d, Extended Data Fig. 4a and Extended Data Table 4); for active lesions, we analysed results from a previous report[21]. Notably, the Taf1d38-DEG signature showed overrepresentation of both the NAW+GM and the active lesion-DEG signatures, and this was particularly remarkable for the top downregulated genes, with a 7-fold enrichment ($P<1.8 \times 10^{-5}$) for the NAW+GM signature, and a 9.6-fold enrichment ($P< 1.7 \times 10^{-4}$) for active lesions one (Fig. 4e).

Using the GWAS summary statistics of the largest meta-analysis thus far (of 47,429 individuals with MS and 68,374 controls) from the International Multiple Sclerosis Genetics Consortium (IMSGC)[3, 4] to perform a gene-based association analysis via MAGMA[22], we found that 21.7% of the Taf1d38 DEGs were nominally associated with MS (Extended Data Table 5). We then performed a gene set enrichment analysis (GSEA) of the Taf1d38 DEG dataset against different databases (Extended Data Table 6). In the context of the Human Phenotype Ontology (HPO) database we detected interesting enriched categories and, after validation *in silico* based on MS genetic associations via gene-set MAGMA analysis, we found *abnormal CNS myelination, abnormality of complement* and *abnormal serum IFN* among the top significant ones (Fig. 4f and Extended Data Table 6). Together, these results indicate that C-terminal TAF1 modification selectively affects the expression of a relatively small set of genes, including many potential MS susceptibility genes, resulting in an MS-like transcriptomic signature.

Specific transcriptional disturbances of mature oligodendrocytes in Taf1d38 mice

To investigate the cell types underlying the transcriptional changes unveiled by bulk tissue RNA-seq, we performed single-nuclei RNA-seq (snRNA-seq) on striatal samples from 8-week-old wild-

type or Taf1d38 mice (n=3 in two batches). After quality control filtering, we identified 27 clusters (Extended Data Fig. 4b) that were annotated on the basis of the expression of lineage marker genes (Extended Data Table 7) and particularly those from a mouse striatal snRNA-seq study[23]. As expected, we could identify and analyze the proportion of eight major cell populations: medium spiny neurons, interneurons, microglia, astrocytes, endothelial cells, oligodendrocyte precursor cells (OPCs), newly formed oligodendrocytes (NFOLs) and mature oligodendrocytes (MOLs) (Fig. 4g and Extended Data Fig. 4c).

We then performed differential gene expression analyses and we could not detect at the single-cell level the increase of inflammation-related genes observed in the bulk RNA-seq. Thus, these inflammatory responses might be restricted to specific areas within the striatum, as the anterior commissure white matter, which might not be captured in the single cell analysis. On the other hand, we found that many downregulated genes identified in the bulk RNA-seq analysis, such as *Ninj2*, *Il33*, *Anln*, *Padi2* and *Ptgds*, were indeed found downregulated in snRNA-seq specifically in mature oligodendrocytes, together with other oligodendrocyte markers, such as Sox10 (Fig. 4h and Extended Data Table 8). Finally, to interrogate which cell clusters showed the highest transcriptional disturbance (and as we found no significant perturbations in any of the eight major cell populations shown in Fig. 4g), we performed perturbation-based analyses on the 27 initially identified subclusters. We found a small (240 cells) microglial subcluster (16, AUC 0.8055480) and a large (502 cells) mature oligodendrocytes subcluster (6, AUC 0.6917725) to be the most perturbed in Taf1d38 mice (Extended Data Fig. 4d and Extended Data Table 9). Together, these analyses detected oligodendrocyte populations among those most affected transcriptionally by the lack of exon 38 of TAF1, while also pointing to microglia affectation.

TAF1 C-terminal end interacts with MS-associated transcriptional release factors

We next explored the mechanism by which the lack of the C-terminal end of TAF1 results in a restricted MS-like transcriptional alteration with associated neuroinflammation and demyelination and leads to an MS-like progressive disability. For this, we used proteomics to identify the physiological interactors of TAF1 affected by the presence or absence of its exon 38–encoded C-terminal end. We first performed GFP-Trap immunoprecipitation from cells transfected with full-length TAF1 (TAF1FL) or TAF1d38, each with EGFP fused to the N-terminal, or only EGFP (three replicates each), followed by liquid chromatography–mass spectrometry and

differential analysis (Fig. 5a). We used two different cell lines, one oligodendroglial (Oli-neu) and the other neuronal (N2a neuroblastoma), and determined the interactors of TAF1FL or TAF1d38, and their relative efficiencies, for each cell type (Extended Data Table 10). Consistent with the restricted transcriptomic alteration in Taf1d38 mice, we detected that TAF7, which dimerizes with TAF1 in TFIID[24] to drive PIC assembly at core promoter sequences[14], interacted with TAF1FL as well as with TAF1d38, with the same efficiency, in both Oli-neu and N2a cells (Extended Data Table 10). This validated the efficacy of our approach at capturing physiological interactors of TAF1.

For interactors that bound more efficiently to TAF1FL than to TAF1d38 (TAF1FL>TAF1d38 interactors), gene ontology analysis detected *assembly of RNAPII complex* among top significant terms, as well as *nuclear receptor (androgen- and glucocorticoid-) signaling* (Fig. 5b and Extended Data Table 11). Among these differential interactors, we noticed multiple components of the general transcription factors TFIIH and TFIIB, which are also part of the PIC and are involved in RNAPII promoter escape[25-27]. Specifically, the detected subunits are Gtf2h1 (p62), Gtf2h2 (p44), Gtf2h4 (p52) and Ercc2 (XPD, p80) as components of TFIIH, and Gtf2b, which is the only component of TFIIB (Fig. 5c). Other transcription initiation and elongation factors detected as TAF1FL>TAF1d38 interactors included: i) Inst12 (in Oli-neu cells), which is a subunit of the Integrator complex that, among other functions, regulates transcriptional initiation and pause release following activation[28, 29]; and ii) EloB (in N2a cells), which is a subunit of the trimeric complex Elongin (SIII), an elongation factor that is thought to stimulate transcription by suppressing transient pausing of RNAPII[30]. Finally, other oligodendroglial differential interactors also involved in transcriptional regulation included Flii, Hmgn2, H2ac21 and Csnk2a2 (Fig. 5c).

The identification of these TAF1FL>TAF1d38 interactors indicates that the C-terminal end of TAF1 interplays with multiple transcriptional regulators involved in release of RNAPII from the promoter itself, from promoter proximal pausing and even from pausing events distant from the promoter. Strikingly, a similar mechanism regarding promoter proximal pausing had previously been associated to MS through a cell type–specific intralocus interaction analysis of MS GWAS data, which specifically involves loci pathogenic in the oligodendrocyte lineage[13]. In addition, upon gene-based MAGMA analysis of the GWAS data from the IMSGC[3, 4], we found that the oligodendroglial TAF1FL>TAF1d38 interactors Gtf2h4 and Gtf2b showed genetic nominal

association with MS risk and severity, respectively (Extended Data Table 12); the same applies to Flii (Flightless), an actin remodeling protein and transcriptional coactivator of hormone-activated nuclear receptors[31], for MS risk. In summary, these data provide a possible mechanism by which C-terminal alteration of TAF1 observed in brain tissue from individuals with MS synergizes with the MS-associated common variants to cause transcriptional, histological and functional alterations, leading to progressive disability.

RNAPII-promoter pausing and decreased elongation at myelination and MS related genes

We next tested whether the C-terminal of TAF1 regulates RNAPII promoter escape and proximal pausing release, selectively affecting myelination related genes. For this, we performed ChIP-seq analyses (two replicates) in brain tissue of 8-week-old Taf1d38 or wild-type mice and analysed the distribution of total and elongating (Ser2-P) RNAPII. Globally, we observed that Taf1d38, as compared to wild-type mice, had an increase of total RNAPII specifically at promoter regions (TSS +/– 1.5 kb) (Fig. 5d), but an unaltered occupancy profile along gene bodies and transcription termination sites (TTS). Together, this resulted in increased RNAPII pausing index in Taf1d38 mice (Fig. 5d and Extended Data Table 13). Gene ontology analysis of the genes displaying higher increased RNAPII levels at the promoter regions in Taf1d38 mice as compared to wild-type (Extended Data Table 14) showed MS signaling to be one of the top affected pathways, together with lipid metabolism (adipogenesis and triacylglycerol biosynthesis), IFN signaling and nucleotide excision pathway (Fig. 5e and Extended Data Table 15). In line with the increase in RNAPII promoter pausing, we observed decreased Ser2P RNAPII levels along gene bodies, indicating a general deficit in RNAPII elongation. Of note, this defect in productive RNAPII elongation was more pronounced in longer genes ($P<10^{-4}$ for genes >30 kb) (Fig. 5f). Remarkably, gene ontology analysis of top genes with decreased Ser2P RNAPII gene body occupancy in Taf1d38 as compared to wild-type mice (Extended Data Table 16) detected the myelination signaling pathway as the most significant term, followed by various protein kinase signaling pathways (Fig. 5g and Extended Data Table 17). These results further support decreased RNAPII release as the most plausible molecular mechanism leading to MS upon C-terminal TAF1 modification.

Discussion

The C-terminal end (encoded by exon 38) of TAF1 is rich in acidic amino acids and is predicted to be intrinsically disordered, with none of the functional or structural domains previously reported for TAF1. However, it has a degree of evolutionary conservation of sequence similar to that observed for known functional domains of TAF1, and its alternative splicing events are also conserved in multiple species to generate forms of TAF1 with or without exon 38–encoded sequences, suggesting a relevant and regulation-prone physiological role for the C-terminal end of TAF1. We found that the C-terminal end of TAF1 is under-detected in the brain of individuals with MS. Moreover, the Taf1d38 mouse model, which only expresses the splice variant of TAF1 lacking this C-terminal end, exhibits progressive motor disability, starting at the age of 3 months with tremor and walking abnormalities and leads to paraplegia at around 20 months. This strongly suggested that C-terminal TAF1 modification does contribute to MS etiology. Indeed, histological analyses of Taf1d38 mice confirmed CNS-resident neuroinflammation in white matter tracts together with progressive demyelination. We gained mechanistic insight through transcriptomic and ChIP-seq analyses of Taf1d38 mice, together with proteomics analyses of full length-TAF1 and TAF1d38 differential interactors, to reveal a regulatory role of the C-terminal TAF1 in RNAII-promoter escape, which especially affects elongation of oligodendroglial myelination–related genes. To the best of our knowledge, this study provides the first reported alteration of the transcription machinery in the brain of individuals with MS that, when placed into an animal model, suffices to induce a full, progressive MS-like phenotype. Further, by identifying TAF1 to be a key player in MS pathogenesis, this study has important etiological and translational implications for MS and provides a novel therapeutic target. Finally, our genetic mouse model of MS -with construct and face validity- can be used to preclinically test new therapies that have the potential to prevent or halt the progressive phase of the disease in individuals with PPMS or SPMS.

Inefficient transcriptional elongation particularly affecting oligodendrocytes, has already been proposed to contribute to the etiology of MS, based on genetic variants that affect expression of regulators of RNAPII promoter proximal pausing release in oligodendrocytes[13]. More precisely, BRD3, which facilitates transcriptional elongation, and HEXIM1/2, which inhibits RNAPII release upon promoter proximal pausing. Besides, MS lesions were found to have

decreased BRD3 RNA and increased HEXIM1 RNA and protein. Therefore, the seminal study by Corradin and coworkers[13] identified oligodendroglial RNAPII promoter-proximal pausing as a potential mechanism in the etiology of MS, involving two modifiers -BRD3 and HEXIM1/2-, affected by common genetic variants. Our study reveals that TAF1 C-terminal alteration, which is observed in MS brains and suffices to induce progressive demyelination and disability in mice, converges with genetic risk into the disruption of transcription elongation in CNS cell types as oligodendrocytes. Interestingly, the transcriptional regulators GTF2B and GTF2H4 that were detected in our study as C-terminal TAF1 interactors with genetic association to MS are also involved in RNAPII release, but at an earlier point in transcription, namely promoter escape[25, 26]. It therefore seems that oligodendroglial cells are particularly sensitive to transcriptional pausing, with this having a likely role in progression of the disease.

TAF1 alterations previously linked to the neurodegenerative diseases XDP[11] and Huntington's disease[12] essentially consist of decreased protein levels, and the prominent affectation of striatal medium spiny neurons in both diseases suggests that these cells might be particularly vulnerable to decreased TAF1 levels. Diminished TAF1 in Huntington's disease may be caused by a degron sequence in an alternative exon 5 due to decreased SREK1 levels[12]. In XDP, the causing haplotype contains multiple variants in noncoding regions within and around TAF1, and the retrotransposon in intron 32 of TAF1[11] is believed to be the etiologically relevant one, as its variable number of hexameric repeats correlates with decreased TAF1 transcript levels[32] and earlier age at disease onset[33]. However, the retrotransposon has also been reported to cause TAF1 splicing alterations, and specifically by increasing partial retention of intron 32[10], and diminished usage of exons downstream of intron 32, such as the 34' microexon[11], although the latter has recently been ruled out through Nanopore long-read sequencing of *TAF1* mRNAs from XDP and control brains tissues[34]. Reduced expression levels of TAF1 exons downstream of the retrotransposon have been reported in XDP-derived cell systems[10], possibly due to diminished passage of RNAP II through the SVA hexameric repeats. It would be therefore intriguing to know whether C-terminal TAF1 antibodies can detect any imbalance in XDP brain tissue similar to the one reported here regarding MS, which particularly affects oligodendroglial cells. It is worth noting that XDP neuropathology, apart from the mentioned loss of striatal medium spiny neurons, also includes overall changes of the structural integrity within the white matter[35].

Our study also opens numerous questions regarding the mechanism by which immunodetection of C-terminal TAF1 decreases in MS brains. We can think of at least three possibilities: i) alternative splicing favoring transcript isoforms lacking exon 38, ii) post-translational modifications that preclude recognition by antibodies raised against the exon 38–encoded native sequence and iii) endoproteolysis at C-terminal motifs. Regarding splicing, we analysed our bulk tissue RNA-seq data of NAG+WM (n=3) from controls and individuals with PPMS with three independent splicing analysis programs, and none of them detected significant differences in TAF1 C-terminal exons. Unfortunately, the previous bulk RNA-seq study with higher numbers of analysed individuals and brain structures[21] does not have enough sequencing depth for efficient interrogation of alternative splicing differences.

In summary, we now have identified that TAF1, a protein that has not previously been linked to MS, is altered in postmortem brains from individuals with MS, both in the NAG+WM and in the demyelinating plaques, and that TAF1 genetically modified mice develop a full, progressive, MS-like neuroinflammatory, demyelinating and disabling phenotype due to deficient transcriptional elongation that particularly affects oligodendrocytes. This study therefore has important etiological and translational implications for MS that may help to develop therapies for the progressive phase of the disease.

Main references

1. Thompson AJ, Baranzini SE, Geurts J, Hemmer B, Ciccarelli O. Multiple sclerosis. *Lancet* 2018; **391**(10130)**:** 1622-1636.

2. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA *et al.* Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult Scler* 2020; **26**(14)**:** 1816-1821.

3. International Multiple Sclerosis Genetics C. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* 2019; **365**(6460).

4. International Multiple Sclerosis Genetics C, Multiple MSC. Locus for severity implicates CNS resilience in progression of multiple sclerosis. *Nature* 2023; **619**(7969)**:** 323-331.

5. Filippi M, Bar-Or A, Piehl F, Preziosa P, Solari A, Vukusic S *et al.* Multiple sclerosis. *Nat Rev Dis Primers* 2018; **4**(1)**:** 43.

6. Bhuiyan T, Timmers HTM. Promoter Recognition: Putting TFIID on the Spot. *Trends Cell Biol* 2019; **29**(9)**:** 752-763.

7. Fant CB, Levandowski CB, Gupta K, Maas ZL, Moir J, Rubin JD *et al.* TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. *Mol Cell* 2020; **78**(4)**:** 785-793 e788.

8. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 2014; **46**(9)**:** 944-950.

9. O'Rawe JA, Wu Y, Dorfel MJ, Rope AF, Au PY, Parboosingh JS *et al.* TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations. *Am J Hum Genet* 2015; **97**(6)**:** 922-932.

10. Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* 2018; **172**(5)**:** 897-909 e821.

11. Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S *et al.* Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. *Am J Hum Genet* 2007; **80**(3)**:** 393-406.

12. Hernandez IH, Cabrera JR, Santos-Galindo M, Sanchez-Martin M, Dominguez V, Garcia-Escudero R *et al.* Pathogenic SREK1 decrease in Huntington's disease lowers TAF1 mimicking X-linked dystonia parkinsonism. *Brain* 2020; **143**(7)**:** 2207-2219.

13. Factor DC, Barbeau AM, Allan KC, Hu LR, Madhavan M, Hoang AT *et al.* Cell Type-Specific Intralocus Interactions Reveal Oligodendrocyte Mechanisms in MS. *Cell* 2020; **181**(2)**:** 382-395 e321.

14. Bernardini A, Mukherjee P, Scheer E, Kamenova I, Antonova S, Mendoza Sanchez PK *et al.* Hierarchical TAF1-dependent co-translational assembly of the basal transcription factor TFIID. *Nat Struct Mol Biol* 2023; **30**(8)**:** 1141-1152.

15. Kipp M, van der Star B, Vogel DY, Puentes F, van der Valk P, Baker D *et al.* Experimental in vivo and in vitro models of multiple sclerosis: EAE and beyond. *Mult Scler Relat Disord* 2012; **1**(1)**:** 15-28.

16. Hobom M, Storch MK, Weissert R, Maier K, Radhakrishnan A, Kramer B *et al.* Mechanisms and time course of neuronal degeneration in experimental autoimmune encephalomyelitis. *Brain Pathol* 2004; **14**(2)**:** 148-157.

17. Bermel RA, Innus MD, Tjoa CW, Bakshi R. Selective caudate atrophy in multiple sclerosis: a 3D MRI parcellation study. *Neuroreport* 2003; **14**(3)**:** 335-339.

18. Vercellino M, Masera S, Lorenzatti M, Condello C, Merola A, Mattioda A *et al.* Demyelination, inflammation, and neurodegeneration in multiple sclerosis deep gray matter. *J Neuropathol Exp Neurol* 2009; **68**(5)**:** 489-502.

19. Ascherio A, Munger KL, Simon KC. Vitamin D and multiple sclerosis. *Lancet Neurol* 2010; **9**(6)**:** 599-612.

20. Marques S, Zeisel A, Codeluppi S, van Bruggen D, Mendanha Falcao A, Xiao L *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* 2016; **352**(6291)**:** 1326-1329.

21. Elkjaer ML, Frisch T, Reynolds R, Kacprowski T, Burton M, Kruse TA *et al.* Molecular signature of different lesion types in the brain white matter of patients with progressive multiple sclerosis. *Acta Neuropathol Commun* 2019; **7**(1)**:** 205.

22. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11**(4)**:** e1004219.

23. Munoz-Manchado AB, Bengtsson Gonzales C, Zeisel A, Munguba H, Bekkouche B, Skene NG *et al.* Diversity of Interneurons in the Dorsal Striatum Revealed by Single-Cell RNA Sequencing and PatchSeq. *Cell Rep* 2018; **24**(8)**:** 2179-2190 e2177.

24. Wang H, Curran EC, Hinds TR, Wang EH, Zheng N. Crystal structure of a TAF1-TAF7 complex in human transcription factor IID reveals a promoter binding module. *Cell Res* 2014; **24**(12)**:** 1433-1444.

25. Cabart P, Ujvari A, Pal M, Luse DS. Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. *Proc Natl Acad Sci U S A* 2011; **108**(38)**:** 15786-15791.

26. Moreland RJ, Tirode F, Yan Q, Conaway JW, Egly JM, Conaway RC. A role for the TFIIH XPB DNA helicase in promoter escape by RNA polymerase II. *J Biol Chem* 1999; **274**(32)**:** 22127-22130.

27. Rimel JK, Taatjes DJ. The essential and multifunctional TFIIH complex. *Protein Sci* 2018; **27**(6)**:** 1018-1037.

28. Gardini A, Baillat D, Cesaroni M, Hu D, Marinis JM, Wagner EJ *et al.* Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell* 2014; **56**(1)**:** 128-139.

29. Stadelmayer B, Micas G, Gamot A, Martin P, Malirat N, Koval S *et al.* Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun* 2014; **5:** 5531.

30. Elmendorf BJ, Shilatifard A, Yan Q, Conaway JW, Conaway RC. Transcription factors TFIIF, ELL, and Elongin negatively regulate SII-induced nascent transcript cleavage by non-arrested RNA polymerase II elongation intermediates. *J Biol Chem* 2001; **276**(25)**:** 23109-23114.

31. Lee YH, Campbell HD, Stallcup MR. Developmentally essential protein flightless I is a nuclear receptor coactivator with actin binding activity. *Mol Cell Biol* 2004; **24**(5)**:** 2103-2117.

32. Westenberger A, Reyes CJ, Saranza G, Dobricic V, Hanssen H, Domingo A *et al.* A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism. *Ann Neurol* 2019; **85**(6)**:** 812-822.

33. Bragg DC, Mangkalaphiban K, Vaine CA, Kulkarni NJ, Shin D, Yadav R *et al.* Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proc Natl Acad Sci U S A* 2017; **114**(51)**:** E11020-E11028.

34. Capponi S, Stoffler N, Penney EB, Grutz K, Nizamuddin S, Vermunt MW *et al.* Dissection of TAF1 neuronal splicing and implications for neurodegeneration in X-linked dystonia-parkinsonism. *Brain Commun* 2021; **3**(4)**:** fcab253.

35. Bruggemann N, Heldmann M, Klein C, Domingo A, Rasche D, Tronnier V *et al.* Neuroanatomical changes extend beyond striatal atrophy in X-linked dystonia parkinsonism. *Parkinsonism Relat Disord* 2016; **31:** 91-97.

Main figure legends

Fig. 1. Underdetection of C-terminal TAF1 in NAG+WM from individuals with MS, and staining of demyelinated lesions with anti-TAF1 antibody. a, Schematic view of human *TAF1* gene and protein (Uniprot P21675), with the following domains depicted in the protein: TAF1 N-terminal domain (TAND), DNA binding domain (DUF3591), zinc knuckle (ZnK) and the bromodomains (BrD1 and BrD2). Also indicated the positions of the protein sequences used as immunogens to generate the antibodies employed in the Western blot analyses, each mapping to a different exon (3, 37 and 38). **b**, TAF1 detection with antibodies against sequences encoded by exons 3, 37 or 38 in nuclear cortical extracts (NAG+WM) from controls (n = 10) and individuals with PPMS (n = 10) or SPMS (n = 8), quantified by normalizing respect to SMC1 levels (one-way ANOVA followed by Tukey's post hoc test). Graphs show mean ± SEM. n.s., non-significant. **c,** Immunohistochemistry in control and PPMS cortical sections with the TAF1ex37+38 antibody (raised against the sequence encoded by exons 37 and 38); top: control white matter (WM) and PPMS NAWM, bottom: control gray matter (GM) and PPMS NAGM (controls, n = 3; PPMS, n = 6; SPMS, n = 6). Histograms show mean optical density quantification (mean ± SEM, one-way ANOVA followed by Tukey's post hoc test) and graph shows simple linear regression (Spearman's rank correlation coefficient) with significant correlation in GM between decreased C-terminal TAF1 staining and lower disease duration of both PPMS and SPMS cases. Scale bars: 50 µm (top); 100 µm (bottom). **d,** Cortical sections (from two different individuals with SPMS) containing a WM inactive lesion (top) or a WM active–inactive lesion (bottom) as well as NAGM and NAWM. Left panels: lesions are delimited based on loss of eriochrome cyanine staining of myelin (blue) and increased HLA-DR immunoreactivity. Right panels: the same lesions immunostained, in sections proximal to those in left panels, with a TAF1 monoclonal antibody raised against the whole protein. Scale bar: 500 µm (top); 2 mm (bottom).

Fig. 2. Generation of Taf1d38 mice and behavioral characterization. a, Schematic view of the 3' region of *TAF1* gene comprising exon 37 to 3'UTR. The scissors show the targets of the guide RNA in a region inside intron 37 with low conservation and inside the alternative 3'UTR, leading to a deletion of 4,246 bp. **b**, Integrative Genomics Viewer (IGV) profiles of RNA-seq data from a wild-type (wt) and a hemizygous Taf1d38 mouse at the 3'region of *TAF1*. In the wt mouse, exons

37 and 38 and the canonical 3'UTR (blue) are present, minimal intron 37 retention (gray) is seen and the alternative 3'UTR (light blue) has few reads. In the Taf1d38 mouse, exon 37 (blue) is followed by the initial part of intron 37 (gray) fused to the distal part of the alternative 3'UTR (light blue) which provides a polyA sequence. **c**, Detection of TAF1 protein by Western blot in brain of 1.5-month-old wt or Taf1d38 mice with the antibodies against sequences encoded by exons 38 or 3. **d,** Time course of the evident motor phenotype of Taf1d38 mice from early phases (3 months) to end-stage disease (20 months). **e,** Ambulatory distance traveled and number of vertical activity counts in open field test by 2-month-old wt (n = 10) or Taf1d38 mice (n = 16). **f,** Limb clasping score evolution from early to symptomatic stages of wt (n = 10 for 2-8months and n= 9 after 8 months) or Taf1d38 mice (n = 16). **g,** Mean latency to fall in the rotarod test for 10-month-old wt (n = 9) or Taf1d38 mice (n = 16). **c, e-g,** Two-tailed unpaired *t*-test. Graphs show mean ± SEM.

Fig. 3. Robust demyelination and white matter neuroinflammation in Taf1d38 mice.

a, Photograph of the entire spinal cord of two 12-month-old mice, wt on the left and Taf1d38 on the right, showing severe atrophy and white matter loss in the latter. **b,** FluroMyelin staining of lumbar spinal cord of 12-month-old wt or Taf1d38 mice showing atrophy and profound white matter loss in both lateral and anterior columns. Insets (boxed areas) in lateral columns are shown enlarged on the right. Histogram shows quantification of the mean signal in the lateral columns of the lumbar spinal cord (n=2). Scale bars: 200 µm (left panels), 10 µm (right panels). **c,** Transmission electron microscopy of spinal cord white matter of 7-month-old mice showing marked myelin loss in Taf1d38 mice with respect to wt. G-ratio quantification is shown for the three main white matter tracts of the spinal cord at thoracolumbar level (left histogram). Axonal diameter vs G-ratio plot (right down) shows that increased G-ratio in Taf1d38 mice occurs independently of axonal size. Twenty microphotographs per tract and per animal (n = 3) were analyzed. Scale bar: 1 µm. **d,** Black Gold myelin staining (top) allowing discern myelin loss, and immunohistochemistry staining for CD68 (bottom left) and GFAP (bottom right) showing glial activation, in sagittal sections from 6-month-old Taf1d38 mice with as compared to wt ones. Scale bars: 200 µm. **e,** Electrophysiological recording of propagated compound action potentials (CAPs) in the fornix of 14-month-old wt (n = 4) or Taf1d38 mice (n = 4), showing increased latency in the conduction through myelinated fibers. Latencies are defined as the time between the

onset of the stimulus artifact and each of the two main negativities, N1 and N2, corresponding to conduction through myelinated and unmyelinated fibers respectively (see inset). **f,** Immunohistochemistry for GFAP (left) and Iba1 (right) in brainstem sagittal sections from 6-month-old wt or Taf1d38 mice, showing marked glial activation. Scale bars: 500 µm. **b, c, e** Two-tailed unpaired *t*-test. Graphs show mean ± SD.

Fig. 4. Transcriptomic alterations of Taf1d38 mice. a, Percentage of differentially expressed genes (DEG) identified through bulk tissue polyA+ RNA-seq analysis of striata from 2-month-old wt (n = 4) or Taf1d38 mice (n = 4). **b,** Gene ontology analysis, using Ingenuity Pathway Analysis (IPA) tool, of data in panel **a**. The most significant enriched canonical pathways for the top (log2FC ≥ |0.5|) DEG in Taf1d38 mice are shown. **c,** Volcano plot of Taf1d38 DEG in panel **a**. **d,** Bulk RNA-seq analyses of human MS NAG+WM (polyA+) or WM lesions (ribo-depleted) used in panel **e**. For our polyA+ RNA-seq of cortical NAG+WM from controls or individuals with PPMS, the principal component analysis of expression data of all detected genes is shown. **e,** Representation factors (RF) showing the DEG overlap between Taf1d38 mice and MS samples. Left: overlap between the polyA+ RNA-seq data of Taf1d38 mice and PPMS cortical NAG+WM. Right: overlap between ribo-depleted RNA-seq data of Taf1d38 mice and active lesions of individuals with progressive MS. The top DEGs (up or down) were defined for Taf1d38 mice as those with log2FC ≥ |0.25|, and for MS samples as those with log2FC ≥ |1.25|. **f,** Bubble plot of top enriched categories from the Human Phenotype Ontology (HPO) after gene-set enrichment analysis (GSEA) of bulk polyA+ RNA-seq data. The enriched categories that were validated for MS with a gene-set level analysis via MAGMA ($p<0.05$) are coloured and labeled in purple. CNS, central nervous system; ESR, erythrocyte sedimentation rate; IFN, interferon; MCV, mean corpuscular volume. **g,** UMAP plot of cell clusters sorted by cell population after single nuclei RNA-seq of the striatum (n = 10,351 nuclei from 3 wt and 3 Taf1d38 mice). AS, astrocytes; END, endothelium; IN, interneurons; MG, microglia; MOL, mature oligodendrocytes; MSN, medium spiny neurons; NFOL, newly formed oligodendrocytes; OPC, oligodendrocyte precursor cells. **h,** Top: violin plots of significantly dowregulated genes in the MOL population (n = 1372 nuclei) identified by Wilcoxon rank sum test, most of which were also identified in bulk RNA-seq. Bottom: IGV track for Ninj2 gene in MOL (merged tracks of 3 replicates are shown).

Fig. 5. Exon 38–dependent TAF1 interactors and altered RNAPII dynamics in Taf1d38 mice.

a, Proteomics and differential analysis experimental design. Oli-neu or N2a cells were transfected (in triplicates) with either EGFP, EGFP-TAF1 full length (FL) or EGFP-TAF1d38 plasmids and processed for GFP-Trap. The resulting co-immunoprecipitates were analysed by liquid chromatography–mass spectrometry (LC-MS/MS). EGFP and TAF1 structures were taken from from AlphaFold[36] and drawings were imported from Biorender. **b,** Top significant canonical pathways (from IPA gene ontology tool) for TAF1FL>TAF1d38 interactors found in Oli-neu cells (n =74) and N2a cells (n =18). **c,** TAF1FL>TAF1d38 interactors with transcriptional functions found in Oli-neu and N2a cells. Shown in purple the proteins encoded by genes associated to MS, according to gene-based analysis by MAGMA of the latest IMSGC GWAS data. **d,** Total RNAPII signal (RPGC, spike-in normalized) after bulk ChIP-seq performed on striata of 2-month-old wt or Taf1d38 mice. Focus on the promoter region is shown below. The box plot shows the distribution of log2(pausing index) based on two biological replicates (right). Bottom: normalized BigWig track of total RNAPII signal along *Npas2* gene showing increased promoter occupancy in Taf1d38. **e,** Top significant IPA canonical pathways for genes whose promoter occupancy by total RNAPII is higher in Taf1d38 mice than in controls. **f,** Ser2P RNAPII signal (RPGC, spike-in normalized) at genes larger than 30 kb after bulk ChIP-seq. Focus on the gene body is shown below. The box plot shows the distribution of normalized RPGCs on gene bodies (right). Bottom: normalized BigWig track of Ser2P RNAPII signal along the *Anln* gene showing decreased gene body occupancy in Taf1d38. **g,** Top significant IPA canonical pathways for genes whose gene body occupancy by Ser2P RNAPII is lower in Taf1d38 mice than in controls. **d, f,** Mann–Whitney *U* test. Box plots show interquartile range; center line represents the median value.

Extended data figure legends

Extended Data Fig. 1. Multiple sequence alignment of C-terminal TAF1

a, Disorder prediction of TAF1 determined with PrDOS webservice. Maximum false positive rate was stablished at 5% for the prediction, and a probability of 0.5 was taken as the threshold of disorder (red line). All amino acids in the C-terminal region (1740-1893 residues) have a disorder score higher than 0.6 (blue dashed line). **b,** Conservation according to JalView of the last 476 amino acids of TAF1 protein (1418-1893 residues) across 15 species (the ones listed in the sequence alignment plus the reptile *Notechis scutatus* and the bony fishes *Takifugu rubripes*, *Xiphophorus maculatus* and *Amphiprion ocellaris*). The multiple sequence alignment corresponds to the 1794-1893 sequence encoded by part of exon 37 and the entire 38, residues are coloured according to Clustal X colour scheme. The protein consensus sequence is shown below. **c,** Multiple sequence alignment also of the 1794-1893 sequence, but across 30 species (including *Saccharomyces cerevisiae, Drosophila melanogaster* and *Caenorhabditis elegans*) evidences that conservation extends beyond vertebrates. Note that interruptions in the alignment are caused by insertions in some sequences, which shift the rest in the alignment. The protein consensus sequence is shown below.

Extended Data Fig. 2. Splicing variants affecting exon 38 inclusion

a, Transcript structure (3' from exon 36) of canonical *TAF1* RNA and of reported isoforms lacking exon 38 sequences. The *Canonica*l form includes the whole exon 38 and the canonical 3' UTR. The *Short exon 38* isoform lacks the sequence encoding the last 52 amino acids encoded by exon 38 and is predicted to add 21 C-terminal amino acids encoded by part of the canonical 3' UTR. The *Intron 37 retention* isoform includes intron 37 and is predicted to result in a truncated protein lacking the whole exon 38, due to the highly conserved stop codon at the beginning of the intron (see **b**). The *Distal exon* isoform splices a large region comprising exon 38 and the 3' UTR and incorporate distal exons that are predicted to add only three additional amino acids (RYQ). The *Distal exon* isoform has only been reported in humans, but the rest are common to both human and mouse (h/m). **b,** Multiple sequence alignment of the genomic sequence comprising the last 20 nucleotides of exon 37 and the beginning of intron 37 across 25 species. A dashed yellow line marks the border between exonic and intronic sequence. The amino acid sequence for *Homo sapiens* is shown below. Note the high conservation of the stop codon (red) if intron 37 is retained.

Extended Data Fig. 3. Additional histopathological characterization of Taf1d38 mice brain

a, Examples of FluroMyelin staining of coronal brain slices of 12-month-old wt and Taf1d38 mice (left). The boxes in corpus callosum and subcortical white matter (cingulum bundle and medial external capsule) show the areas whose signal was quantified, as shown in the histograms (right). Quantifications of the mean intensity values were performed on 3 and 12-month-old wt and Taf1d38 mice (n = 2). Scale bar: 500 µm. **b,** Spp1 (osteopontin) immunofluorescence in corpus callosum (sagittal sections) from 6-month-old wt and Taf1d38 mice, exemplifying the upregulation in Taf1d38 white matter. Scale bar: 50 µm. **c,** CD3 or BC220 immunofluorescence (red) together with DAPI nuclear counterstaining (blue) in corpus callosum (coronal sections) from 3-month-old wt and Taf1d38 mice, evidencing the absence of lymphocyte infiltration. Scale bar: 20 µm.

Extended Data Fig. 4. Additional information regarding bulk tissue RNA-seq on human samples and snRNA-seq on mouse samples

a, Principal component analysis of all detected genes in the polyA+ RNA-seq experiment on human cortical samples, that was performed on 3 individuals with PPMS and 4 controls. The CT4 control sample was atypical as it mapped closer to PPMS samples than to the rest of control samples. Accordingly, we decided to exclude this sample from the final analysis. **b,** In the mouse snRNA-seq experiment, the UMAP plot of cell clusters in resolution 1 yield a total of 27 clusters. **c,** Proportion of nuclei corresponding to each genotype (wt or Taf1d38) for each of the eight major cell populations: AS, astrocytes; END, endothelium; IN, interneurons; MG, microglia; MOL, mature oligodendrocytes; MSN, medium spiny neurons; NFOL, newly formed oligodendrocytes; and OPC, oligodendrocyte precursor cells. **d,** UMAP plot highlighting in red the most perturbed clusters (6 and 16 in **b**), which correspond to MOL and MG, respectively.

Materials & Methods

Human brain tissue samples

Brain specimens from NAGM and NAWM cortex of individuals with PPMS or SPMS and control individuals used in immunoblotting, immunohistochemistry and RNA sequencing, were provided by the UK Multiple Sclerosis Tissue Bank at Imperial College, London. Written informed consent for brain removal was obtained from brain donors and/or next of kin. Procedures, information and consent forms have been approved by the Bioethics Subcommittee of Consejo Superior de Investigaciones Científicas (CSIC, Madrid, Spain). Donor metadata are summarized in Extended Data Table 1.

Animals

CRISPR mice lacking exon 38 of Taf1 (Taf1d38) were generated (for details, see 'Generation of Taf1d38 mice' below) for this study and used in a C57BL/6J background. All mice were housed in the CBM animal facility under *ad libitum* food and water availability, and maintained in a temperature-controlled environment on 12 h light–dark cycles. Animal housing and maintenance protocols followed local authority guidelines approved by the CBM Institutional Animal Care and Utilization Committee (Comité de Ética de Experimentación Animal del CBM, CEEA-CBM), and Comunidad de Madrid PROEX 293/15 and PROEX 247.1/20. Mice were euthanized with CO_2.

Generation and genotyping of Taf1d38 mice

Taf1d38 mice were generated in a 50% C57BL/6J and 50% CBA/CA background via CRISPR/Cas9-mediated gene editing, by the institutional core facility (CNB-CBM mouse transgenesis facility). We decided to emulate one of the alternative splicing events that in humans generate TAF1 isoforms lacking exon 38 encoded sequences (Extended Data Fig. 2a). More precisely, we mimicked the retention of intron 37 that happens in approximately 20% of TAF1 transcripts in human brain (https://vastdb.crg.eu/event/HsaINT1043632@hg38), which results in a TAF1 protein truncated at the end of exon 37 with the only substitution of its last amino acid (S1820R), due to the highly evolutionary conserved stop codon at the beginning of intron 37 (Extended Data Fig. 2b). We designed a CRISPR-Cas9 strategy to eliminate the reported splicing acceptor sites at exon 38 and the canonical 3'UTR. For this, we generated a guide RNA mapping to intron 37 (5': CCAAGTTGTGAAGGACGTAAATG) and another mapping downstream of the canonical 3'UTR (3': GGGCTACAAGCCCCTAACTTAGG), the latter within an adjacent and longer cryptic alternative 3'UTR (Fig. 2a). The CRISPR gRNAs were co-injected into fertilized oocytes with Cas9

protein. Confirmation of the correct 4,246 bp deletion without introducing any point mutations was performed through Sanger sequencing. RNA-seq analysis of both hemizygous males and homozygous females also showed the clean deletion and the constitution of a new 3'UTR downstream the canonical exon 37 (Fig. 2b). Taf1d38 mice were genotyped using primers that anneal in intron 37 before the deletion and in the distal 3' UTR after the deletion (GGCAATTCAGTCTTTGTGTATAGG; AGCTCTAGCTGACCTTAAACTCC).

Bioinformatic analysis of C-terminal TAF1

Genomic and protein sequences used in the alignments were downloaded from Ensembl and UniProt public databases. Sequence alignments were done with state-of-the-art aligner ClustalO[37]. JalView 2.11.0[38] was used for sequence alignment editing, visualization and analysis. Chosen conservation measure for protein, among JalView's repertoire, was Valdar's conservation[39]. Protein DisOrder prediction System (PrDOS[40]) was used to study the intrinsic disorder of TAF1. ProsperousPlus[41] was run to find putative protease cleavage sites in TAF1. Experimentally determined post-translational modifications were searched through PhosphoSitePlus[42], and predicted phosphorylation sites were found with NetPhos[43] and ELM[44].

Western Blot

Samples from human brain were stored at –80 °C and ground with a mortar in a frozen environment to prevent thawing of the samples, resulting in tissue powder. Mouse brains were quickly dissected on an ice-cold plate and the different structures stored at –80 °C.

Mouse extracts were prepared as previously described[45, 46] by homogenizing the brain areas in ice-cold extraction buffer (20 mM HEPES pH 7.4, 100 mM NaCl, 5 mM EDTA, 20 mM NaF, 1% Triton X-100, 1 mM sodium orthovanadate, 1 µM okadaic acid, 5 mM sodium pyrophosphate, 30 mM β-glycerophosphate and protease inhibitors [Complete, Roche]). Homogenates were centrifuged at 15000*g* for 15 min at 4 °C, the supernatant was collected, and protein content was determined by Quick Start Bradford kit assay (Bio-Rad, 500-0203).

Nuclear and cytoplasmic extracts from human tissue were prepared by homogenizing the tissue powder in ice-cold extraction buffer A (10 mM HEPES pH 7.9, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 1% NP40, 2.5 mM DTT, 1 mM sodium orthovanadate, 1 µM okadaic acid, 5 mM sodium pyrophosphate, 30 mM β-glycerophosphate and protease inhibitors [Complete, Roche]) and centrifuging at 4000*g* for 1 min at 4 °C. The supernatant constituted the cytoplasmic fraction, and the pellet was washed in extraction buffer A and resuspended in ice-cold extraction buffer C (20 mM HEPES pH 7.9, 25% glycerol, 400 mM NaCl, 1 mM EDTA, 1 mM EGTA, 2.5 mM DTT, 1 mM sodium orthovanadate, 1 µM okadaic acid, 5 mM sodium pyrophosphate, 30 mM β-

glycerophosphate and protease inhibitors [Complete, Roche]). After sonication (two pulses of 10 s and 100% amplitude, Labsonic M, Sartorius), 1% NP40 was added, and the extracts were centrifuged at 15000*g* for 15 min at 4 °C. The supernatant constituted the nuclear extract, and protein content was determined by Quick Start Bradford kit assay (Bio-Rad, 500-0203).

Between 10 and 30 µg of total protein was electrophoresed on 6% SDS–polyacrylamide gel, transferred to a nitrocellulose blotting membrane (Amersham Protran 0.45 µm, GE Healthcare Life Sciences, 10600002) and blocked in TBS-T (150 mM NaCl, 20 mM Tris–HCl, pH 7.5, 0.1% Tween 20) supplemented with 5% non-fat dry milk. Membranes were incubated overnight at 4 °C with either rabbit anti-TAF1 exon 3 (1:500; Abcam, ab188427), rabbit anti-TAF1 exon 37 (1:500; Bethyl, A303-504A), rabbit anti-TAF1 exon 38 (1:500; Bethyl, A303-505A), rabbit anti-SMC1 (1:2000; Bethyl, A300-055A), rabbit anti-vinculin (1:15,000; Abcam, ab129002) or mouse anti-β actin (1:50,000; Sigma, A2228) in TBS-T supplemented with 5% non-fat dry milk. After washings with TBS-T, the membranes were incubated with secondary HRP-conjugated anti-rabbit IgG (1:2,000, DAKO, P0448) or anti-mouse IgG (1:2,000, DAKO, P0447), and developed using the ECL detection kit (Perkin Elmer, NEL105001EA). Images were scanned with densitometer (Bio-Rad, GS-900) and quantified with ImageJ 1.52d software.

Immunohistochemical and myelin staining analyses

Human tissue. Postmortem control and MS cortical snap-frozen tissue was analysed (Extended Data Table 1). MS lesions were classified according to demyelination and cellular distribution as described[47-50].

Immunohistochemistry. The immunohistochemistry protocol was carried out as previously described[48]. Briefly, cryosections (10µm, Leica) were thawed and fixed in 4% paraformaldehyde (PFA) for 1 h at room temperature (RT). Antigen retrieval with citrate buffer 0.1M pH 6 at 90 °C for 10 min was performed for TAF1 labelling. Endogenous peroxidases were quenched with 10% methanol and 3% H_2O_2 and the slides were blocked with 5% normal goat serum and 0.2 % Triton X-100 in PBS. Slides were then incubated with rabbit anti-TAF1 exon 37+38 (1:100; Sigma, HPA001075), mouse anti-TAF1 whole protein (1:30; Santa Cruz, sc735) or anti-human leukocyte antigen (HLA)-DR (1:200 Agilent, M0746, TAL.1B5 clone) in blocking solution overnight. Anti-rabbit biotinylated (1:200; Vector Labs) or anti-mouse biotinylated (1:200; Vector Labs) secondary antibody was used and the reaction was amplified using the Vectastain Elite ABC reagent (Vector Labs). The peroxidase reaction product was developed with 0.05% 3,3´-diaminobenzidine (DAB, Sigma-Aldrich) and 0.003% H_2O_2 in 0.1 M Tris–HCl, pH 7.6. The reaction was monitored under the microscope and terminated by rinsing the slides with PB.

Eriochrome cyanine for myelin staining. To visualize myelin after HLA-DR immunostaining, eriochrome cyanine (EC) staining was carried out as described previously[47]. The sections were air-dried and placed in fresh acetone; then, they were stained in 0.5% EC for 1 h and differentiated in 5% iron alum and borax-ferricyanide for 10 and 5 min, respectively.

Quantification of TAF1 staining in human samples. To quantify the density of $TAF1^+$ cells and the intensity of TAF1 labelling, photomicrographs of the GM and WM from controls and MS samples were captured on BX61 microscope (Olympus) equipped with a color camera CX9000 (MBF). The quantification was performed using the QuPath software. A mean of 5 or 7 ROIs/sample were used respectively for GM or WM, and the regions were defined throughout all the cortical layers in GM (1.78 ± 0.18 mm^2/ROI) or WM (0.67 ± 0.05 mm^2/ROI). A threshold for TAF1 staining was assessed and the number of $TAF1^+$ cells (density: mean value of $TAF1^+$ cells/mm^2 per sample) and TAF1 intensity (mean value of the optical density obtained from all $TAF1^+$ cells/sample) were measured.

Mouse tissue. Mice were deeply anaesthetized with an intraperitoneal injection of Dolethal (2 mg/g) and perfused transcardially with 50 ml of NaCl followed by 4% PFA. Brains and spinal cords were fixed overnight in 4% PFA, cryopreserved for 72 h in 30% sucrose in PBS and added to optimum cutting temperature (OCT) compound (Tissue-Tek, Sakura Finetek Europe, ref. 4583), frozen and stored at -80 °C until use. 30 µm sagittal (brain) or 20 µm transversal (spinal cord) sections were cut on a cryostat (Thermo-Fisher Scientific), collected and stored free floating at -20 °C in glycol containing buffer (30% glycerol, 30% ethylene glycol in 0.02 M PB).

Immunohistochemistry. Mouse sections were washed in PBS, immersed in 0.1% H_2O_2 in PBS for 45 min to quench endogenous peroxidase activity and incubated for 1 h in blocking solution (PBS containing 0.5% fetal bovine serum, 0.3% Triton X-100 and 1% BSA). The sections were incubated overnight at 4 °C with either rat anti-CD68 (1:200; Abcam ab53444), rabbit anti-Iba1 (1:1000; Wako 019-19741) or rabbit anti-GFAP (1:20,000; Novus NB300-141) diluted in blocking solution. Sections were incubated with biotinylated anti-rat or anti-rabbit secondary antibody and then with avidin-biotin complex (Elite Vectastain kit, Vector Laboratories PK-6101, PK-6104). Chromogen reactions were carried with 3,3´-diaminobenzidine (SIGMAFAST DAB, Sigma-Aldrich, D4293) for 10 min. Finally, the sections were mounted on glass slides and coverslipped with Mowiol (Calbiochem, Cat. 475904). A DMC6200 CMOS camera was used to capture the images.

Immunofluorescence. Sections were blocked with 4% of serum of the species in which secondary antibody was obtained, diluted in 0.1 M PBS and 0.1% Triton X-100 for 1 h at RT. Sections were incubated overnight at 4 °C with either mouse anti-SPP1 (1:150; Biotechne MAB808), rat anti-CD3 (1:50; BioRad MCA1477) or rat-anti-B220 (1:200; BD Pharmingen clone RA3-6B2) in 1% serum and 0.1% Triton X-100. Afterwards, the sections were incubated with the secondary

fluorochrome-conjugated antibodies (donkey anti-mouse Alexa 488 1:500 [Thermo-Fisher Scientific, A-21202]; goat anti-rat Alexa 594 1:400 [Invitrogen]) for 1 h at RT, and nuclei were labelled with Hoechst 33258 (1.5ug/ml) for 10 min. The sections were mounted on glass slides using Glycergel mounting medium (Dako). A Leica Stellaris 5 confocal microscope was used to capture the images.

FluoroMyelin. Brain or spinal cord sections were mounted on glass slides and let dry. After hydration in PBS, staining in FluoroMyelin™ (Green Fluorescent Myelin Stain F34651, Molecular Probes) was performed for 20 min at RT according to the manufacturer's protocol. Sections were coverslipped with ProLong™ Gold Antifade Mountant (Invitrogen). A fluorescence microscope (Axiovision, Zeiss) was used to capture the images, and ImageJ software was used to measure the mean intensity gray value in each region. One level per animal was analysed.

Black gold. Mouse sections were mounted on SuperFrost slides (Thermo Fisher) and let dry overnight. The staining was performed according to manufacturer's instructions (Black Gold II Myelin Staining Kit, Avantor BSENTR-100-BG): after rehydration in distilled water, the slides were incubated with Black-Gold II reagent at 65 °C for 12 min in a water bath, and fixed with sodium thiosulfate solution for 3 min. The sections were coverslipped with DePeX (SERVA). A DMC6200 CMOS camera was used to capture the images.

Transmission electron microscopy

7-month-old mice (3 wt and 3 Taf1d38) were deeply anaesthetized with an intraperitoneal injection of Dolethal (2 mg/g) and perfused transcardially with 50 ml of NaCl followed by a solution of 4% PFA and 2 % glutaraldehyde in 0.1M PB pH 7.4. Spinal cords were dissected and 2 mm long fragments from the lumbar level were collected in the same solution for 2 h at RT and overnight at 4 °C. Fragments were postfixed in 1% osmium tetroxide in 0.1 M cacodylate buffer, dehydrated in ethanol and embedded in Epon-Araldite. Serial ultrathin transversal sections were collected on pioloform-coated, single-hole grids, and stained with uranyl acetate and lead citrate. The sections were analysed with a JEM-1010 transmission electron microscope (Jeol, Japan) equipped with a side-mounted CCD camera Mega View III from Olympus Soft Imaging System GmBH (Muenster, Germany). 20 photographs per white matter region (dorsal, lateral and anterior tracts) and per animal were taken at a magnification of 10,000X. Blind analysis of the g-ratio was performed using ImageJ as previously described[51], from circular areas equivalent to the measured areas of axons and myelin sheath including the axon.

Electrophysiology

Propagated compound action potentials (CAPs) were measured as reported[52] in the fornix of 14-month-old wt (n=4) and Taf1d38 (n=4) mice. Animals were anesthetized with isofluorane and brains were cut in 400 µm-thick horizontal sections (Pelco 3000 vibrating blade microtome, Ted Pella Inc., USA) in an ice-cold solution (215 mM sucrose, 2.5 mM KCl, 26 mM $NaHCO_3$, 1.6 mM NaH_2PO_4, 1 mM $CaCl_2$, 4 mM $MgCl_2$, 4 mM $MgSO_4$ and 20 mM glucose). Sections were incubated in artificial cerebrospinal fluid (aCFS) (124 mM NaCl, 2.5 mM KCl, 10 mM glucose, 25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4, 2.5 mM $CaCl_2$ and 1.3 mM $MgCl_2$) for 1 h. CAPs were recorded with a pulled borosilicate glass pipette (≈ 1 MΩ resistance) filled with aCFS by stimulation of the fornix with a bipolar electrode (CE2C55 FHC, USA) with up to 2 mA, 100 µs pulses (Master-8, AMPI, Israel). The distance between the recording and stimulating electrodes was 1 mm. Using pClamp 10.0 (Molecular Devices, USA), latencies between the onset of the stimulus artifact and the two negative waves of the CAP's (N1 and N2) were measured (ms).

Behavioural testing

Open field test. Locomotor activity was measured as previously described[46] in clear Plexiglas® boxes of 27.5cm x 27.5cm, outfitted with photo-beam detectors for monitoring horizontal and vertical activity. Mouse movements were recorded with a MED Associates' Activity Monitor and analysed with the MED Associates' Activity Monitor Data Analysis v.5.93.773 software. Mice were placed in the center of the open-field apparatus and left to move freely for 15 min. Total traveled ambulatory distance and number of vertical episodes when the mouse stands on its hindlimbs (vertical counts) were measured.

Limb clasping. Assessment of corticospinal function was performed by the limb clasping test[53]. Mice were held vertically by the tail for 10 s and recorded. Videos were scored blind to the genotype as follows: 0 – normal extension of both hindlimbs; 1 – unilateral hindlimb partial retraction during more than 50% of the recorded time; 2 – bilateral hindlimb partial retraction during more than 50% of the recorded time; 3 - bilateral hindlimb total retraction during more than 50% of the recorded time.

Rotarod. Motor coordination was assessed in an accelerating rotarod apparatus (Ugo Basile). The first time that the mice performed this test (2-month-old), they were trained at constant speed over two consecutive days: on the first day, 4 trials of 1 min with angular velocity fixed at 4 rpm were carried; and on the second day, 4 trials of 2 min (1 min at 4 rpm and 1 min at 8 rpm) were completed. The test was performed on the third consecutive day with the rotarod set to accelerate from 4 to 40 rpm over 5 min in 4 independent trials. For the following tests (6 and 10-month-old), the training was performed in a single session (2 trials of 1 min at 4 rpm and 2 trials of 2 min with acceleration from 4 to 8 rpm over 1 min and the second minute at 8 rpm).

The test was performed on the second day as described above. The latency to fall from the rod was measured as the mean of the four accelerating trials.

Bulk RNA sequencing

RNA was isolated from striatum of 2-month-old wild type (wt) mice (n=4) and Taf1d38 mice (n=4) and from normal appearing gray and white matter (NAG+WM) from cortex of control (CT) subjects (n = 4) and individuals with PPMS (n = 3) (Extended Data Table 1). Mice were subjected to daily manipulation for one week prior to sacrifice in their environment the day of the experiment, to avoid any stress that could modify the brain basal transcriptomic state.

Total RNA was extracted using the Maxwell 16 LEV simplyRNA Tissue Kit (Promega, AS1280). Quantification and quality determination of RNA was performed on a NanoDrop One spectrophotometer (Thermo Fisher) and an Agilent 2100 Bioanalyzer. For mouse samples a RNA integrity number (RIN) over 8.5 was required, while for human samples the RIN cut-off was set to 6.

Two independent libraries were prepared from the same mouse samples. The first (n = 4) was polyA+, had a read length of 1 x 50 bp, and was sequenced in HiSeq2500, generating >45 million single-end reads per sample. The second (n = 3) was ribo-depleted, had a read length of 2 x 150 bp and was sequenced in NovaSeq6000, generating >180 million paired-end reads per sample. For human samples, libraries were polyA+, with a read length of 2 x 150 bp and were sequenced in NovaSeq6000, generating >210 million paired-end reads per sample.

Bulk RNA-seq data processing and analysis

Quality analyses over the reads were performed using FastQC 0.11.9 software. Salmon 1.9.0 alignment mapping algorithm was used[54], employing indexes constructed with either *Mus musculus* reference genome (Ensembl GRCm39) or *Homo sapiens* reference genome (Ensembl GRCh38). Differential expression analysis at gene level was carried with DESeq2 1.40.2 [55]. Genes with less than 20 normalized counts on average in both wt and Taf1d38 groups or CT and PPMS groups were filtered out. Significant differentially expressed genes (DEGs) were defined for an adjusted p value < 0.05.

DEGs were validated *in silico* via gene-based association test via MAGMA 1.9[22], using the GWAS summary statistics from the International Multiple Sclerosis Genetics Consortium MS (47,429 MS cases; 68,374 controls)[3, 4]. For MAGMA analysis the model applied was 'multi=snp-wise', which aggregates the results of the 'snp-wise=top' and 'snp-wise=mean' analyses models, improving the power and accuracy of gene-based association tests especially when the genetic model of the disease under study is unknown. For gene-based analysis, gene boundaries were

retrieved from RefSeq (GRCh38.p14), lifted over to GRCh37 coordinates, and expanded to 5 kb flanking regions, upstream and downstream respectively, to encompass potential regulatory elements.

Hierarchical clustering based on sample distances and principal component analysis on human samples revealed that one of the controls (CT4) grouped together with the 3 PPMS samples (Extended Data Fig. 4a), and was removed it from the final analysis. Clustering of the remaining six samples resulted as expected (Fig. 4d) and differential expression analysis was performed in this subset (Extended Data Table 4). Alternative splicing was also using three complementary software: 1) vast-tools v. 2.5.1[56] using human junction libraries for the align and combine modules, and a |dPSI| (absolute difference in per cent spliced in) ≥ 10 in the compare module; 2) rMATS v.4.1.2[57] using STAR v. 2.7.10[58] to align fastq reads together with a custom splice-junction library, and a FDR < 5% for differential splicing analysis; and 3) Majiq v. 2.5[59] after alignment with STAR and default parameters and using deltapsi, with a probability of changing >95%.

We also reanalyzed raw data of ribo-depleted RNA-seq from Elkjaer et al.[21] to find DEGs between control white matter samples and white matter with active MS lesions. Fastq files corresponding to samples from the same individual and the same type (i.e all the samples taken from the same control individual, or all the active lesions from a given subject with MS) were concatenated. Then, Salmon alignment mapping, Salmon quantification and DESeq2 differential expression analysis were carried as described earlier. Finally, the aligned reads were also processed with samtools 1.9 and the profiles were visualized using the Integrative Genomics Viewer (IGV) 2.4.

Overlap analyses. Overlap between MS and Taf1d38 DEG signatures was analysed using http://nemates.org/MA/progs/overlap_stats.html., which runs an hypergeometric test and provides a representation factor (RF), which is the number of overlapping genes divided by the expected number of overlapping genes drawn from two independent groups. PolyA+ datasets were compared to extract common DEG signature between Taf1d38 mice and PPMS NAG+WM; whereas ribo-depleted datasets were compared to find common DEG signature between Taf1d38 mice and MS active lesions.

Single nuclei RNA sequencing

2-month-old wt (n = 3) and Taf1d38 (n = 3) mice were manipulated as described for bulk RNA-seq experiment, and the striata were quickly dissected on an ice-cold plate and stored at -80 °C. The six samples were processed in two batches (first batch n = 2, second batch n = 1). The whole striatum was homogenized in nuclei isolation buffer (10 mM Tris-HCl pH 8.0, 0.25 M sucrose, 5 mM $MgCl_2$, 25 mM KCl, 0.1% Triton X-100, 0.1 mM dithiothreitol (DTT) and 0.2 U/µl RNase

inhibitors [RNAsin Plus 40 U/µl]). The suspension was sequentially filtered through a 70-µm strainer followed by a 30-µm strainer and centrifuged for 10 min at 900*g* at 4 °C. The pellet was resuspended in 150 µl of nuclei isolation buffer, and was mixed with 250 µl of 50% iodixanol for debris removal. 500 µl of 29% iodixanol were overlaid with this solution, and centrifuged for 30 min at 13,500*g* at 4 °C. The supernatant was removed and the pellet was resuspended in 100 µl of phosphate-buffered saline (PBS) with 1% BSA with 0.2 U/µl RNase inhibitors.

The nuclei were stained with Trypan Blue 0.4% and counted in a haemocytometer. A total of 10,000 estimated nuclei for each sample were loaded on the Chromium Next GEM Chip G, although a much lower number of nuclei was recovered after sequencing. cDNA libraries were prepared following the Chromium Next GEM Single Cell 3' Reagents Kits v3.1 User Guide and sequenced in NovaSeq6000.

Single nuclei RNA-seq data processing and analysis

The 6 samples were aligned with Cellranger v.7.0.1 with reference genome GRCm38-mm10. Filtered count matrices and barcodes were uploaded into R 4.3.2 to be processed following Seurat v4[60]. As samples were sequenced in two batches (replicates 1 and 2 run in the first and replicate 3 in the second for both conditions), they were processed separately for single initial QC to avoid differences related to sequencing batches. Cells were filtered for number UMIs per nuclei, number of genes per nuclei and mitochondrial percentage (percent_mito < 2 & nFeature_RNA > 50 & nFeature_RNA < 5000 & nCount_RNA > 5 & nCount_RNA < 10000). The final selected nuclei had the following features: median number of UMI, median number of genes and median mitochondrial percentage for WT samples (8056, 3376, 0.1804958) and for KO samples (8404, 3454, 0.1160056) with a total final number of 10,351 nuclei.

Log-normalization, feature selection and SCT normalization regressing mitochondrial percentage, was performed in the whole object with the 6 samples. The first 30 PCs were selected using variable features determined by SCTransform (2000 genes). Sample integration was performed with Harmony[61] (RunHarmony), on the selected PCs using as grouping variables the replicates and the batch information. Dimension reduction was performed using runUMAP. KNN graph was constructed using Findneighbors on the harmony reduction, with annoy method, 50 trees and a K of 20. Clusters were identified by running Findclusters to different resolutions from 0.2 to 1.2.

First exploratory view of the clusters showed expression of a specific set of hormonal genes ("Prl", "Gh", "Pou1f1", "Tshb") from nuclei corresponding only to one of the replicates. These genes are not expected to be expressed in the striatum, and probably came from contamination from the hypothalamus or the bed nucleus of the stria terminalis in this specific sample. An

"Hormonal_gene_score" was calculated using AddModuleScore with 4 controls and 50 bins for all the nuclei, and we filtered out the possibly contaminated nuclei by setting a "Hormonal_gene_score">0. All the processing starting from the QC to cluster identification was repeated without these spurious cells, using the parameters and procedures as explained before.

Clusters were annotated using FindAllMarkers in resolution 1 and 0.8. Top markers per cluster showed classical markers for main cell types in the striatum allowing the annotation of broad cell types, which was complemented with label transfer using as reference the single cell striatum dataset from Muñoz-Manchado et al.[23]. The GSE97478 annotated expression matrix was downloaded from GEO repository, and common anchors between query and reference were calculated (FindTransferAnchors, with parameters cca reduction and SCT normalized query and reference). Annotations were transferred with TransferData on the first 30 dimensions of the cca reduction to the query dataset. The predicted cell types corresponded with our broad cell types and identified specific known striatum cells types, such as MSN cell types. Differential expression analysis was performed splitting the object according to cell type and running FindAllMarkers between wt and Taf1d38 cells in each cluster using the default Wilcoxon Rank Sum test. The dataset can be explored at:

https://castelobranco.shinyapps.io/snrnaseq_ct_vs_taf1ex38del/.

Identification of perturbed cell types

AUGUR R package[62] (https://github.com/neurorestore/Augur) was used to identify cell types that showed changes in Taf1d38 compared to wt, defined as perturbed. AUGUR trains a machine-learning model specific to the cell types to predict their condition (wt or Taf1d38) and, as a result, for each cell type AUGUR reports and AUC score of the accuracy of the model on that cell type. AUC score > 0.5 was considered as a significant perturbation. Input data was loaded after normalization and batch correction, as explained in the section above. AUGUR scores are not affected by cluster size; however, different iterations were run by downsampling to the same number of nuclei per comparison.

Functional enrichment analysis

The Ingenuity Pathway Analysis (IPA) software (QIAGEN Inc., version 107193442, http://www.ingenuity.com) was used for analysis of enriched canonical pathways within Taf1d38 DEG signatures, TAF1FL>TAF1d38 interactors (after eliminating potential unspecific interactors such as cytoskeletal proteins), genes with higher total RNAPII occupancy at the promoter and genes with lower Ser2P RNAPII occupancy in Taf1d38 mice.

DEGs were also examined for enriched categories via gene-set enrichment analysis (GSEA) using clusterProfiler v.4.8.1 package in R[63], using a modified Kolmogorov-Smirnov statistics to assess significant categories. Categories found enriched (adjusted p-value <0.05) were additionally validated for MS with a gene-set level analysis via MAGMA using Z scores from gene p-values.

Plasmid generation

Human TAF1 cDNA (TAF1FL, ENST00000373790.4 transcript) was cloned into a pcDNA3 vector plasmid and EGFP coding sequence was added N-terminally by restriction free cloning[64] (RFC) to generate the pcDNA3-EGFP-TAF1FL plasmid. Exon 38 sequence was deleted and substituted by intron 37 retention sequence until the stop codon (Extended Data Fig. 2b) by RFC to generate the pcDNA3-EGFP-TAF1d38 plasmid. pEGFP-C1 plasmid was used as an EGFP only expressing plasmid.

Cell culture and plasmid transfection

N2a cells (mouse neuroblastoma cell line) were grown in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) (Gibco), 2 mM glutamine and penicillin–streptomycin (Sigma) at 37 °C with 5% CO_2. Oli-neu cells were grown in DMEM supplemented with 1% horse serum (Gibco), 10 µg/mL insulin (Sigma I9278), 10 µg/mL apotransferrin (Sigma T4382), 100 µM putrescine dihydrochloride (Sigma P5780), 200 nM progesterone (Sigma P0130), 220 nM sodium selenite (Sigma S5261) and penicillin–streptomycin (Sigma) at 37 °C with 10% CO_2. N2a cells were transfected with either pEGFP-C1, pcDNA3-EGFP-TAF1FL or pcDNA3-EGFP-TAF1d38 using Lipofectamine™ 2000 (Invitrogen; 4 µL/µg of plasmid), while jetOPTIMUS transfection agent (Polyplus; 1.5 µL/µg of plasmid) was used for Oli-neu cells. Technical triplicates were carried for each condition.

GFP-Trap

24 h after transfection, cells were harvested in PBS with protease inhibitors (1mM PMSF, Phosphatase Inhibitor Cocktail 3 [Sigma-Aldrich, P004] and Complete [Roche]) and centrifuged for 3 min at 500*g* at 4°C. After washing the pellet in PBS with protease inhibitors, the samples were incubated in lysis buffer (50 mM NaCl, 20 mM TrisHCl pH 8, 0.5% Triton-X100, 2 mM $MgCl_2$) with 0.1 U/µL of benzonase for 30 min in ice. 250 mM NaCl, 0.5 mM EDTA and 0.1 mM DTT were added to the lysed sample and after 5 min centrifugation at 16,000*g* at 4°C, the supernatant was collected and 10% of the volume was saved as the input. GFP-Trap dynabeads (Chromotek) were washed with IP buffer (150 mM NaCl, 0.5 mM EDTA, 20 mM TrisHCl pH 7.5) supplemented with

protease inhibitors, and each sample was incubated with 25 µL of beads in IP buffer for 1 h at 4 °C. After 3 washes with IP buffer, beads were stored at -20 °C until usage.

Mass spectrometry and proteomic data processing

N2a experiment. Proteins were eluted twice in 100 µL of 8 M urea in 100 mM Tris-HCl pH 8, and then reduced, alkylated (15 mM TCEP, 50 mM CAA, 30 min in the dark at RT) and digested with 200 ng of Lys-C (Wako) and 200 ng of trypsin (Promega) for 16 h at 37 °C. Peptides were desalted (C18 stage-tips, speed-vac dried) and re-dissolved in 21 µl of 0.5% formic acid (FA).

Liquid chromatography–mass spectrometry (LC-MS/MS) was performed by coupling an UltiMate 3000 RSLCnano LC system to an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific). 5 µL of peptides were loaded into a trap column (Acclaim™ PepMap™ 100 C18 LC Columns 5 µm, 20 mm length) for 3 min at a flow rate of 10 µl/min in 0.1% FA. Then, peptides were transferred to an EASY-Spray PepMap RSLC C18 column (Thermo) (2 µm, 75 µm x 50 cm) operated at 45°C and separated using a 60 min effective gradient (buffer A: 0.1% FA; buffer B: 100% acetonitrile [ACN], 0.1% FA) at a flow rate of 250 nL/min (4% to 6% B the first 2 min, 6% to 33% B the following 58 minutes, plus 10 additional minutes at 98% B). Peptides were sprayed at 1.5 kV into the mass spectrometer via the EASY-Spray source, with a capillary temperature of 300°C. The mass spectrometer was operated in a data-dependent mode, with an automatic switch between MS and MS/MS scans using a top 18 method (intensity threshold ≥ 3.6e5, dynamic exclusion of 20 s and excluding charges, +1 and > +6). MS spectra were acquired from 350 to 1500 mass-to-charge ratio (m/z) with a resolution of 60,000 FWHM (200 m/z). Ion peptides were isolated using a 1.0 Th window and fragmented using higher-energy collisional dissociation (HCD) with a normalized collision energy of 27. MS/MS spectra resolution was set to 15,000 (200 m/z). The ion target values were 3e6 for MS (maximum IT of 25 ms) and 1e5 for MS/MS (maximum IT set to auto).

Raw files were processed with MaxQuant (v 2.1.2.0) against a mouse protein database (UniProtKB/Swiss-Prot, 21,990 sequences) supplemented with contaminants, with the following settings: carbamidomethylation of cysteines as a fixed modification; oxidation of methionines and protein N-term acetylation as variable modifications; minimal peptide length = 7 amino acids; maximum of two tryptic missed-cleavages. Results were filtered at 0.01 FDR and loaded in Prostar (v1.8.0)[65] using the intensity values for further statistical analysis. Proteins with less than 4 measured valid values in at least one condition (n=6) were filtered out. A global normalization of log2-transformed intensities across samples was performed using the LOESS function. Partial observed missing values (POV) were imputed using the SLSA algorithm, whereas missing in entire condition values (MEC) were imputed with the DETquantile option. Differential

analysis was performed using the empirical Bayes statistics Limma. Proteins with an adjusted p-value<0.05 and a log2(fold-change)>|3| respect to EGFP were defined as interactors. Proteins with an adjusted p-value<0.05 and a log2(fold-change)>|0.6| in the comparison between TAF1FL and vs TAF1d38 were defined as differential interactors. The adjusted p-value was estimated to be below 5% by Benjamini-Hochberg.

Oli-neu experiment. Identical volumes of each immunoprecipitate were loaded on STRAP columns (PROTIFI, Farmingdale, NY) and individually digested with trypsin (250 ng) in reducing and alkylating (chloroacetamide) conditions. Tryptic peptides were desalted using C18 tips and subjected to independent LC-MS/MS analysis using a nano liquid chromatography system (Ultimate 3000 nano HPLC system, Thermo Fisher Scientific) coupled to an Orbitrap Exploris 240 mass spectrometer (Thermo Fisher Scientific). 5 µL of each sample were injected on a C18 PepMap trap column (5 µm, 300 µm I.D. x 2 cm, Thermo Scientific) at 20 µL/min, in 0.1% FA, and the trap column was switched on-line to a C18 PepMap Easy-spray analytical column (2 µm, 100 Å, 75 µm I.D. x 50 cm, Thermo Scientific). Equilibration was done in mobile phase A (0.1% FA), and peptide elution was achieved in a 90 min gradient from 4% - 50% B (0.1% FA in 80% acetonitrile) at 250 nL/min. Data acquisition was performed in DDA (Data dependent acquisition) mode, in full scan positive mode. Survey MS1 scans were acquired at a resolution of 60,000 at m/z 200, scan range was 375-1200 m/z, with Normalized Automatic Gain Control (AGC) target of 300% and a maximum injection time of 45 ms. The top 20 most intense ions from each MS1 scan were selected and fragmented by Higher-energy collisional dissociation (HCD) of 30%. Resolution for HCD spectra was set to 15,000 at m/z 200, with AGC target of 100 % and maximum ion injection time of 80 ms. Precursor ions with single, unassigned, or more than six charge states from fragmentation selection were excluded.

Raw files were processed using Proteome Discoverer v2.5 (Thermo Fisher Scientific). MS2 spectra were searched using four Mascot v2.7 as search engine, against a target/decoy database built from sequences corresponding to the *Mus musculus* reference proteome downloaded from Uniprot Knowledge database including contaminants. Search parameters considered fixed carbamidomethyl modification of cysteine, and the following variable modifications: methionine oxidation, possible pyroglutamic acid from glutamine at the peptide N-terminus and acetylation of the protein N-terminus. The peptide precursor mass tolerance was set to 10 ppm and MS/MS tolerance was 0.02 Da, allowing for up to two missed tryptic cleavage sites. Spectra recovered after FDR <= 0.01 filter were selected for quantitative analysis. Label-free protein quantification used unique + razor peptides. The same criteria as described in the N2a experiment was adopted to define the interactors. MAGMA gene-based analysis was also performed for genes encoding both N2a and Oli-neu the TAF1FL>TAF1d38 interactors. In this regard, it should be noted Gtf2h4

is located on the complement region on Chr6 and, therefore, its genetic association might be not as robust as those found for Gtf2b and Flii.

Bulk ChIP sequencing

Mice were manipulated as described for bulk RNA-seq experiment, and the striata were quickly dissected on an ice-cold plate. A pool of 4 striata were used for each sample, and a total of 2 wt and 2 Taf1d38 pooled samples were processed in 2 batches.

Chromatin immunoprecipitation was performed as previously described[66]. After homogenizing the tissue in ice-cold PBS with protease inhibitors (1mM PMSF; Phosphatase Inhibitor Cocktail 3, Sigma-Aldrich; Complete, Roche), the extracts were crosslinked with 1% formaldehyde for 10 minutes at 37 °C, and the reaction was quenched with 125mM glycine for 5 minutes. After centrifugation at 5000*g* for 5 minutes at 4°C and washing in ice-cold PBS with protease inhibitors, the pellets were resuspended in 5 mL lysis buffer A (5 mM Pipes pH 8.0, 85 mM KCl, 0.5% NP40) supplemented with protease inhibitors and incubated for 10 minutes on ice. Cell nuclei were pelleted by centrifugation at 4000*g* for 5 minutes at 4°C, and resuspended in 1.5 mL of lysis buffer B (50 mM Tris HCl pH 8.1, 1% SDS, 10 mM EDTA) supplemented with protease inhibitors. Chromatin was sonicated for 25 minutes using a Covaris ultrasonicator (E220 Evolution) with the following parameters: 200 cycles/burst, 5% duty factor and 140 W of peak incident power. After a 10 min centrifugation at 20,000*g* the supernatant was collected. 50 µL of sonicated chromatin was reverse-crosslinked using Proteinase K in lysis buffer B at 65°C overnight and, after phenol chloroform extraction, DNA fragmentation was checked (intended fragment size 200-600 bp). For each immunoprecipitation (IP), 40 µg of chromatin spiked-in with 5% of human chromatin from MCF7 cell line, and 4 µg of the specific antibody (rabbit anti-Rpb1, Cell Signaling, Cat.14958; rabbit anti-RNA Pol II phospho S2, Abcam, ab5095) was incubated in IP buffer (0,1% SDS, 1% TX-100, 2mM EDTA, 20 mM TrisHCl pH8, 150 mM NaCl) at 4 °C overnight. Inputs consisting of 10% of the mix of chromatin and IP buffer were stored at 4 °C. Dynabeads protein A and G (ThermoFisher) were blocked overnight with IP buffer supplemented with 1 mg/ml BSA, and each IP was incubated with 25 µL of beads for 4 hours at 4 °C. Beads were sequentially washed with wash buffer 1 (20 mM Tris HCl pH 8, 0.1% SDS, 2 mM EDTA, 150 mM NaCl), wash buffer 2 (20 mM Tris HCl pH 8, 0.1% SDS, 2 mM EDTA, 500 mM NaCl), wash buffer 3 (20 mM Tris HCl pH 8, 1% NP40, 1% sodium deoxycholate, 1 mM EDTA, 250 mM LiCl) and T buffer (10 mM Tris HCl pH8), all supplemented with protease inhibitors. ChIPmentation on beads was performed as previously described[67], using Tn5A enzyme provided by the Proteomics Service of CABD (Centro Andaluz de Biología del Desarrollo) in tagmentation buffer (10 mM Tris HCl pH 8, 10% dimethylformamide, 5 mM $MgCl_2$) for 10 min at 37°C. After serial washes in wash

buffer 3 and TE buffer (10 mM Tris HCl pH8, 1mM EDTA), tagmented DNA was eluted with 1% SDS and 100 mM $NaHCO_3$ in TE buffer at 50°C for 30 minutes. Overnight decrosslinking was performed using Proteinase K in 175 mM NaCl at 37 °C. After incubation for 10 min with RNase A (ThermoFisher), DNA was purified (QIAGEN, 28106) and inputs were also tagmented as described above and purified. Libraries were amplified for the optimum cycles determined by qPCR reaction using NEBNext High-Fidelity Polymerase (New England Biolabs, M0541). Libraries were purified with Sera-Mag Select Beads (GE Healthcare, 29343052) and sequenced after Qubit quantification and size profile analysis with Agilent 2100 Bioanalyzer, using Illumina NextSeq 500 and single-end configuration with a read length of 75 bp.

ChIP-seq data processing and analysis

Initial filtering and trimming was performed using FASTQ Toolkit 1.0.0. Second trimming was performed using Trimmomatic 0.36 (20:4 sliding window, 2:leading, 2:trailing, 35:minlen). The reads reads from the IP and input libraries aligned to the mouse and human reference genomes using bowtie 2.4.5, and signal tracks were obtained using the bamCoverage tool (deepTools 3.5.1[68]).

To correct technical biases and global differences, spike-in normalization was applied using human chromatin as control. Signal tracks were obtained without normalization and with a bin size of 50 bp. To measure the enrichment ratio in human and mouse signal tracks, the ratio between the IP and input samples was calculated to avoid local biases and background noise, providing a more accurate representation of the occupancy ratio (OR). OR was obtained by dividing the enrichment ratio in mouse and human reads. For the genome tracks, the OR was used to normalize the signal tracks using bamCoverage. The normalization was performed using the RPGC (Reads Per Genomic Content) method with a bin size of 50 bps, a read extension of 300 bps, and the OR as a scale factor.

Spike-in quantifications were used to generate definitive signal profiles and the normalized RPGC counts tables. Signal profiles were generated by Seqplots[69]: signal values were shown by mean in bin tracks of 100 bps, using 15,000 bp as default gene length with 10,000 bp as upstream and downstream regions.

For signal distribution analysis using normalized RPGCs, two regions were defined: the transcription start site (TSS) region, which comprised the sequence 1.5 kb up and downstream of the TSS; and the gene body region, which was defined between the TSS and the transcriptional termination site (TTS). With such definition of the TSS region, only genes larger than 4.5 kb were plotted in total RNAPII metaplots, as shorter genes may introduce noise in the promoter/gene body definition. Log2FC of Taf1d38 vs wt counts for each IP and each experiment were

calculated. The top 15% genes with higher log2FC for total RNAPII or lower log2FC for Ser2P RNAPII in a given experiment were extracted. Then, the overlapping genes which resulted from crossing the 15% top lists obtained in batches 1 and 2, were taken as the top affected genes. Thus, we defined the 1285 genes with highest total RNAPII promoter region occupancy in Taf1d38 mice (Extended Data Table 14) and the 2397 genes with lowest Ser2P RNAPII gene body occupancy (Extended Data Table 16) in Taf1d38 mice.

To estimate the level of paused RNAPII, we used the pausing index (PI), which is defined as the ratio of total RNAPII enrichment within the promoter to that in the gene body. The PI was estimated[66] as the ratio between the occupation of RNAPII within the promoter (computed as the sum of reads in 400 bp surrounding the TSS), and the occupation of RNAPII within the gene (calculated as the average number of reads in 400 bp windows throughout the gene body, starting 200 bp downstream the TSS). Again, only genes larger than 4.5 kb were considered for PI computation (Extended Data Table 13).

Statistical analysis

Statistics were performed with SPSS 21.0 (SPSS Statistic IBM). The normality of the data was checked by Shapiro-Wilk and Kolmogorov-Smirnov tests. For two-group comparison, two-tailed unpaired Student's t-test or Mann-Whitney U test were performed. For multiple comparisons, data were analysed by one-way ANOVA test followed by a Tukey's post hoc test. Pearson correlation test was used to test the strength of correlation analysis between continuous variables. A critical value for significance of $p\text{-value} < 0.05$ was used throughout the study. Benjamini-Hochberg correction was applied for multiple testing for DEG. Unless otherwise stated in the corresponding figure legend and methods, all measurements were taken from distinct samples.

Data availability

The datasets generated during the current study will be available in the GEO repository upon publication of the peer-reviewed version of this manuscript.

Methods references

36. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021; **596**(7873)**:** 583-589.

37. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019; **47**(W1)**:** W636-W641.

38. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009; **25**(9)**:** 1189-1191.

39. Valdar WS. Scoring residue conservation. *Proteins* 2002; **48**(2)**:** 227-241.

40. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 2007; **35**(Web Server issue)**:** W460-464.

41. Li F, Wang C, Guo X, Akutsu T, Webb GI, Coin LJM *et al.* ProsperousPlus: a one-stop and comprehensive platform for accurate protease-specific substrate cleavage prediction and machine-learning model construction. *Brief Bioinform* 2023; **24**(6).

42. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 2015; **43**(Database issue)**:** D512-520.

43. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 1999; **294**(5)**:** 1351-1362.

44. Kumar M, Michael S, Alvarado-Valverde J, Zeke A, Lazar T, Glavina J *et al.* ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res* 2024; **52**(D1)**:** D442-D455.

45. Olla I, Pardinas AF, Parras A, Hernandez IH, Santos-Galindo M, Pico S *et al.* Pathogenic Mis-splicing of CPEB4 in Schizophrenia. *Biol Psychiatry* 2023; **94**(4)**:** 341-351.

46. Picó S, Parras A, Santos-Galindo M, Pose-Utrilla J, Castro M, Fraga E *et al.* CPEB alteration and aberrant transcriptome-polyadenylation lead to a treatable SLC19A3 deficiency in Huntington's disease. *Sci Transl Med* 2021; **13**(613)**:** eabe7104.

47. Clemente D, Ortega MC, Arenzana FJ, de Castro F. FGF-2 and Anosmin-1 are selectively expressed in different types of multiple sclerosis lesions. *J Neurosci* 2011; **31**(42)**:** 14899-14909.

48. Ortega MC, Lebron-Galan R, Machin-Diaz I, Naughton M, Perez-Molina I, Garcia-Arocha J *et al.* Central and peripheral myeloid-derived suppressor cell-like cells are closely related to the clinical severity of multiple sclerosis. *Acta Neuropathol* 2023; **146**(2)**:** 263-282.

49. Kuhlmann T, Ludwin S, Prat A, Antel J, Bruck W, Lassmann H. An updated histological classification system for multiple sclerosis lesions. *Acta Neuropathol* 2017; **133**(1)**:** 13-24.

50. Luchetti S, Fransen NL, van Eden CG, Ramaglia V, Mason M, Huitinga I. Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: a retrospective autopsy cohort analysis. *Acta Neuropathol* 2018; **135**(4)**:** 511-528.

51. Soria FN, Zabala A, Pampliega O, Palomino A, Miguelez C, Ugedo L *et al.* Cystine/glutamate antiporter blockage induces myelin degeneration. *Glia* 2016; **64**(8)**:** 1381-1395.

52. Palma A, Chara JC, Montilla A, Otxoa-de-Amezaga A, Ruiz-Jaen F, Planas AM *et al.* Clemastine Induces an Impairment in Developmental Myelination. *Front Cell Dev Biol* 2022; **10:** 841548.

53. Lalonde R, Strazielle C. Brain regions and genes affecting limb-clasping responses. *Brain Res Rev* 2011; **67**(1-2)**:** 252-259.

54. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017; **14**(4)**:** 417-419.

55. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**(12)**:** 550.

56. Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res* 2017; **27**(10)**:** 1759-1768.

57. Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* 2012; **40**(8)**:** e61.

58. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**(1)**:** 15-21.

59. Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-Vallinas J, Lahens NF, Hogenesch JB *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 2016; **5:** e11752.

60. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021; **184**(13)**:** 3573-3587 e3529.

61. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019; **16**(12)**:** 1289-1296.

62. Squair JW, Skinnider MA, Gautier M, Foster LJ, Courtine G. Prioritization of cell types responsive to biological perturbations in single-cell data with Augur. *Nat Protoc* 2021; **16**(8)**:** 3836-3873.

63. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021; **2**(3)**:** 100141.

64. van den Ent F, Lowe J. RF cloning: a restriction-free method for inserting target genes into plasmids. *J Biochem Biophys Methods* 2006; **67**(1)**:** 67-74.

65. Wieczorek S, Combes F, Lazar C, Giai Gianetto Q, Gatto L, Dorffer A *et al.* DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* 2017; **33**(1)**:** 135-136.

66. Herrero-Ruiz A, Martinez-Garcia PM, Terron-Bautista J, Millan-Zambrano G, Lieberman JA, Jimeno-Gonzalez S *et al.* Topoisomerase IIalpha represses transcription by enforcing promoter-proximal pausing. *Cell Rep* 2021; **35**(2)**:** 108977.

67. Schmidl C, Rendeiro AF, Sheffield NC, Bock C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* 2015; **12**(10)**:** 963-965.

68. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 2014; **42**(Web Server issue)**:** W187-191.

69. Stempor P, Ahringer J. SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome Open Res* 2016; **1:** 14.

Acknowledgements

This work was supported by grants: to J.J.L. from Networking Research Center on Neurodegenerative Diseases (CIBERNED) - Health Institute Carlos III (ISCIII) (Collaborative Grant No. PI2018/06-1), Spanish Ministry of Economy and Competitiveness (MINECO) (SAF2015-65371-R), Spanish Ministry of Science, Innovation and Universities (MICIU) (RTI2018-096322-B-I00) and Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033/FEDER,UE Grant No. PID2021-123141OB-I00); to G.C-B. from Swedish Research Council (grant 2019-01360), the Swedish Brain Foundation (FO2018- 0162, FO2023-0032), Knut and Alice Wallenberg Foundation (grants 2019-0107), the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, the Swedish Society for Medical Research (SSMF, grant JUB2019) and Karolinska Institutet; to F. C.-L. from Spanish Ministry of Sciences (PID2020-119570RB-I00); C.M. from: Spanish Ministry of Sciences (PID2022-143020OB-I00), CIBERNED (CB06/05/0076) and Basque Government (IT1551-22); to C.T. from MCIU PID2020-114996RB-I00 (MCIU/AEI/FEDER,UE), RyC2018-024106-I, and CNS2022-135318; to D.C. from Instituto de Salud Carlos III (PI21/00302, co-funded by the European Union, and CB22/05/00016); and to B.A. from AEI/10.13039/501100011033 Grant PID2022-136745OB-I00 and European Union NextGeneration EU/PRTR.

C.R.-L and I.G.-O were awarded fellowships from Tatiana Guzmán El Bueno Foundation. C.Z. was funded by the Chinese Scholarship Council (CSC).

Tissue specimens used in this research were obtained from the UK Multiple Sclerosis Tissue Bank at Imperial College (London). We thank the patients and families who participated in the tissue donation programs.

We thank Dr. Fernando de Castro for advice at different stages of the study and for critical reading of the manuscript, Dr. Manuel Comabella for advice on genetics and clinical features of progressive MS, and Dr. Jorge R. Cabrera for advice on the role of TAF1 in brain diseases. We also thank Cristina Martos-Polo (Group of Neuro Immuno-Repair of the HNP) and Dr Javier Mazarío (Microscopy Service of the HNP) for their assistance with the histological quantifications in the human CNS and Nuria de la Torre (CBM) for administrative assistance.

Computational analyses were run on the Center for Scientific Computing provided by the Universidad Autónoma de Madrid (CCC-UAM), the High-Performance Computing cluster provided by the Centro Informático Científico de Andalucía (CICA) and the Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for Advanced

Computational Science, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

We thank the following core facilities: CBM-CNB Mouse Transgenesis, CBM-Animal Facility, CBM-Genomics & Massive Sequencing, CBM-Bioinformatics, CBM-Advanced Light Microscopy Facility, CBM-Electron Microscopy, CNIO-Genomics Core Unit, CNIO-Proteomics Core Unit, CNB-Proteomics, Science for Life Laboratory-National Genomics Infrastructure in Stockholm and CRG-Genomics Unit in Barcelona.

CBM is a Severo Ochoa Center of Excellence (MICIN, Award CEX2021-001154-S).

Author contributions

C.R.-L. contributed to study design and was involved in all assays and data collection. I.H.H contributed to study design and was involved in Taf1d38 mouse model generation and characterization. J.T-.B contributed to ChIP-seq and GFP-Trap experiments and analyses. E.A. performed single nuclei RNA-seq analyses. D.L.-M. and J.P.-U. performed bioinformatics and immunohistochemistry analyses. I.G.-O. and M.M.-J. performed GSEA and gene-based analyses from GWAS data. M.M.-S. performed behavioral tests on mice. I.R.-B., M.G.-B., B.S.-G. and M.S.-G. performed immunohistochemistry and Western blot. M.C.O. and N.L.F. performed immunohistochemistry on human samples. J.C.C, Z.M. and F.P.-C. performed electron microscopy, immunofluorescence and fluoromyelin analyses. A.P.-S. performed electrophysiological recordings. M.K and C.Z. contributed to single cell RNA-seq experiments. M.C.-B. performed bioinformatics analyses of C-terminal TAF1 motifs and conservation. A.B., B.A., D.C., C.T., C.M., F.C.-L. and G.C.-B. made intellectual contributions to experimental design and discussion. C.M., F.C.-L., G.C.-B and J.J.L. directed the study and designed experiments. J.J.L. wrote the paper with input from all authors.

Declaration of interests

The authors declare no competing interests.

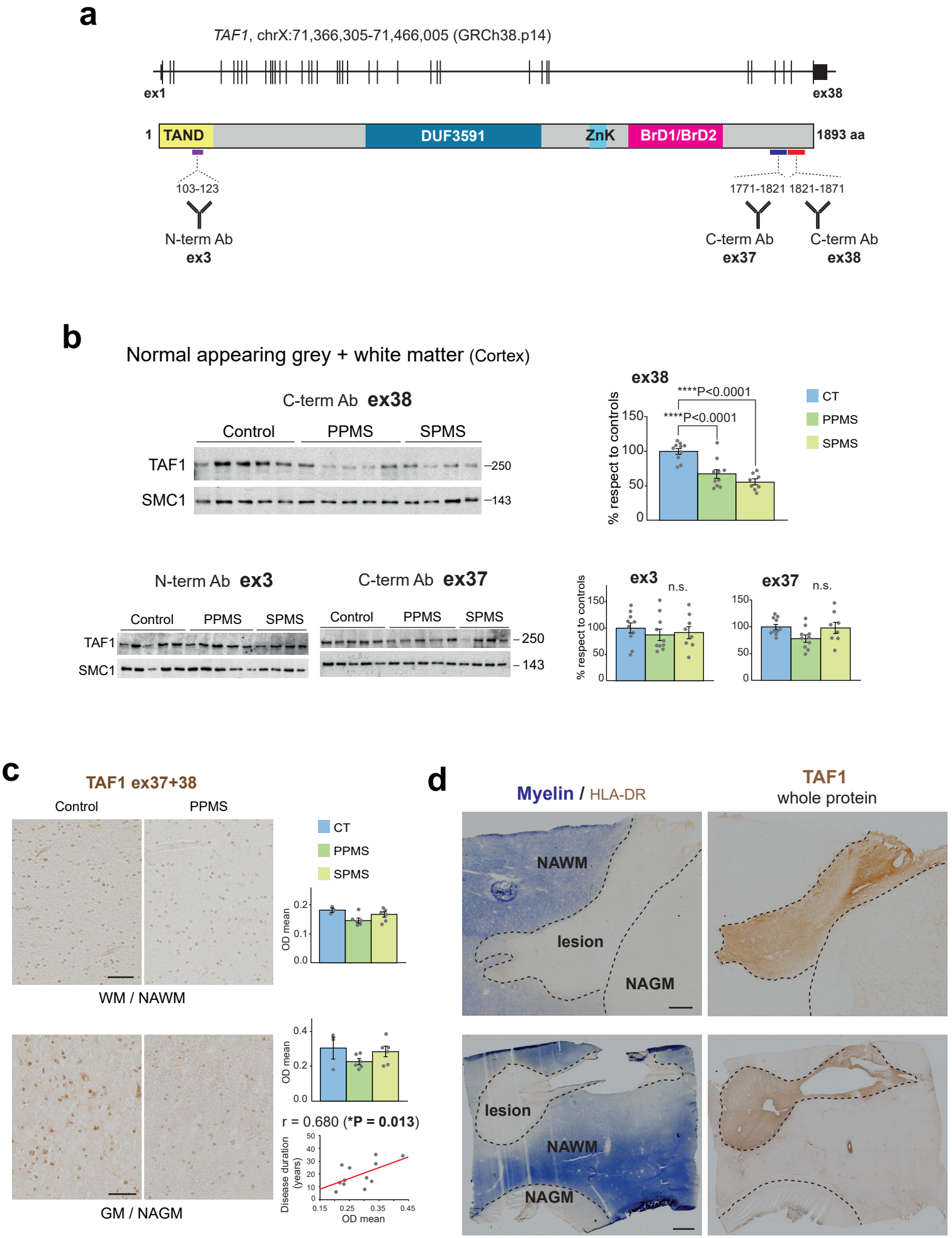

a
TAF1, chrX:71,366,305-71,466,005 (GRCh38.p14)
ex1
ex38
1
TAND
DUF3591
ZnK
BrD1/BrD2
1893 aa
103-123
N-term Ab
ex3
1771-1821
1821-1871
C-term Ab
ex37
C-term Ab
ex38
b
Normal appearing grey + white matter (Cortex)
C-term Ab ex38
Control
PPMS
SPMS
TAF1
SMC1
250
143
ex38
****P<0.0001
****P<0.0001
% respect to controls
CT
PPMS
SPMS
N-term Ab ex3
C-term Ab ex37
ex3
n.s.
ex37
n.s.
c
TAF1 ex37+38
Control
PPMS
WM / NAWM
GM / NAGM
OD mean
r = 0.680 (*P = 0.013)
Disease duration (years)
OD mean
d
Myelin / HLA-DR
TAF1
whole protein
NAWM
lesion
NAGM

Figure 1

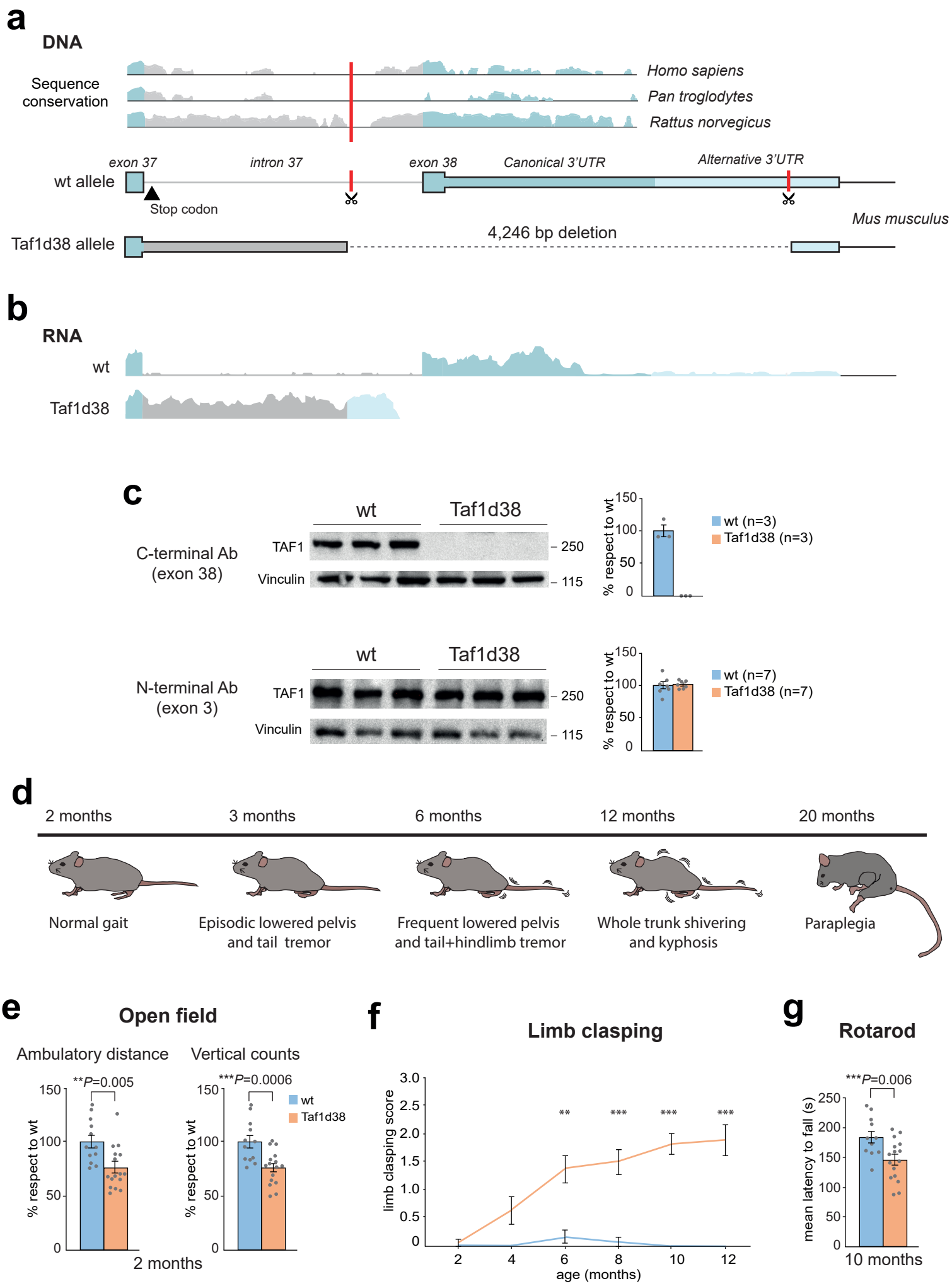

a
DNA
Sequence conservation
Homo sapiens
Pan troglodytes
Rattus norvegicus
exon 37
intron 37
exon 38
Canonical 3'UTR
Alternative 3'UTR
wt allele
Stop codon
Mus musculus
Taf1d38 allele
4,246 bp deletion
b
RNA
wt
Taf1d38
c
wt
Taf1d38
C-terminal Ab (exon 38)
TAF1
Vinculin
250
115
% respect to wt
150
100
50
0
wt (n=3)
Taf1d38 (n=3)
N-terminal Ab (exon 3)
wt (n=7)
Taf1d38 (n=7)
d
2 months
3 months
6 months
12 months
20 months
Normal gait
Episodic lowered pelvis and tail tremor
Frequent lowered pelvis and tail+hindlimb tremor
Whole trunk shivering and kyphosis
Paraplegia
e
Open field
Ambulatory distance
Vertical counts
**P=0.005
***P=0.0006
wt
Taf1d38
2 months
f
Limb clasping
limb clasping score
3.0
2.5
2.0
1.5
1.0
0.5
0
**
***
2
4
6
8
10
12
age (months)
g
Rotarod
***P=0.006
mean latency to fall (s)
250
200
150
100
50
0
10 months

Figure 2

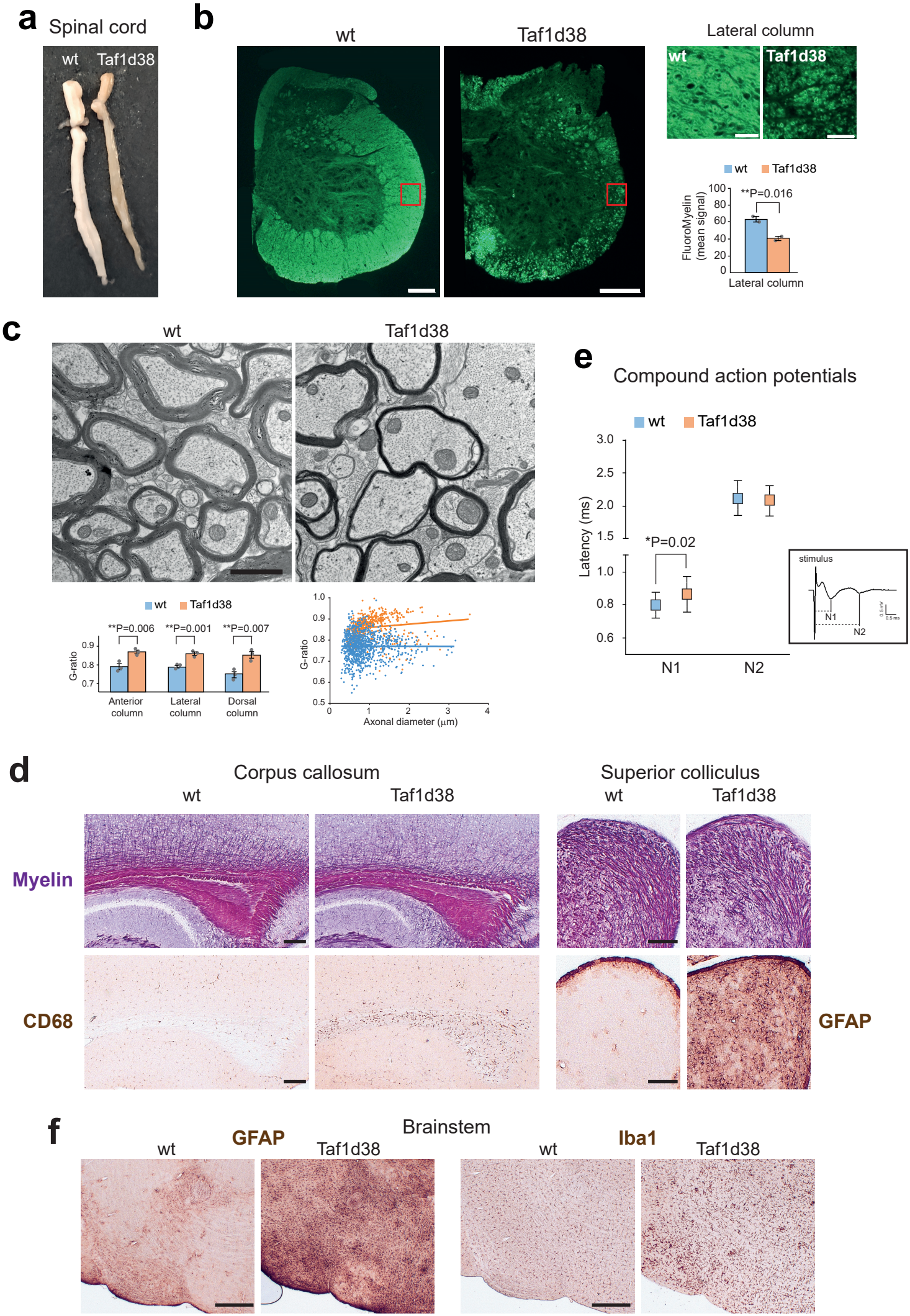

a
Spinal cord
wt
Taf1d38
b
wt
Taf1d38
Lateral column
wt
Taf1d38
wt
Taf1d38
**P=0.016
FluoroMyelin (mean signal)
100
80
60
40
20
0
Lateral column
c
wt
Taf1d38
wt
Taf1d38
**P=0.006
**P=0.001
**P=0.007
G-ratio
0.9
0.8
0.7
Anterior column
Lateral column
Dorsal column
G-ratio
1.0
0.9
0.8
0.7
0.6
0.5
0
1
2
3
4
Axonal diameter (µm)
e
Compound action potentials
wt
Taf1d38
Latency (ms)
3.0
2.5
2.0
1.5
1.0
0.8
0.6
*P=0.02
N1
N2
stimulus
N1
N2
d
Corpus callosum
wt
Taf1d38
Superior colliculus
wt
Taf1d38
Myelin
CD68
GFAP
f
Brainstem
GFAP
wt
Taf1d38
Iba1
wt
Taf1d38

Figure 3

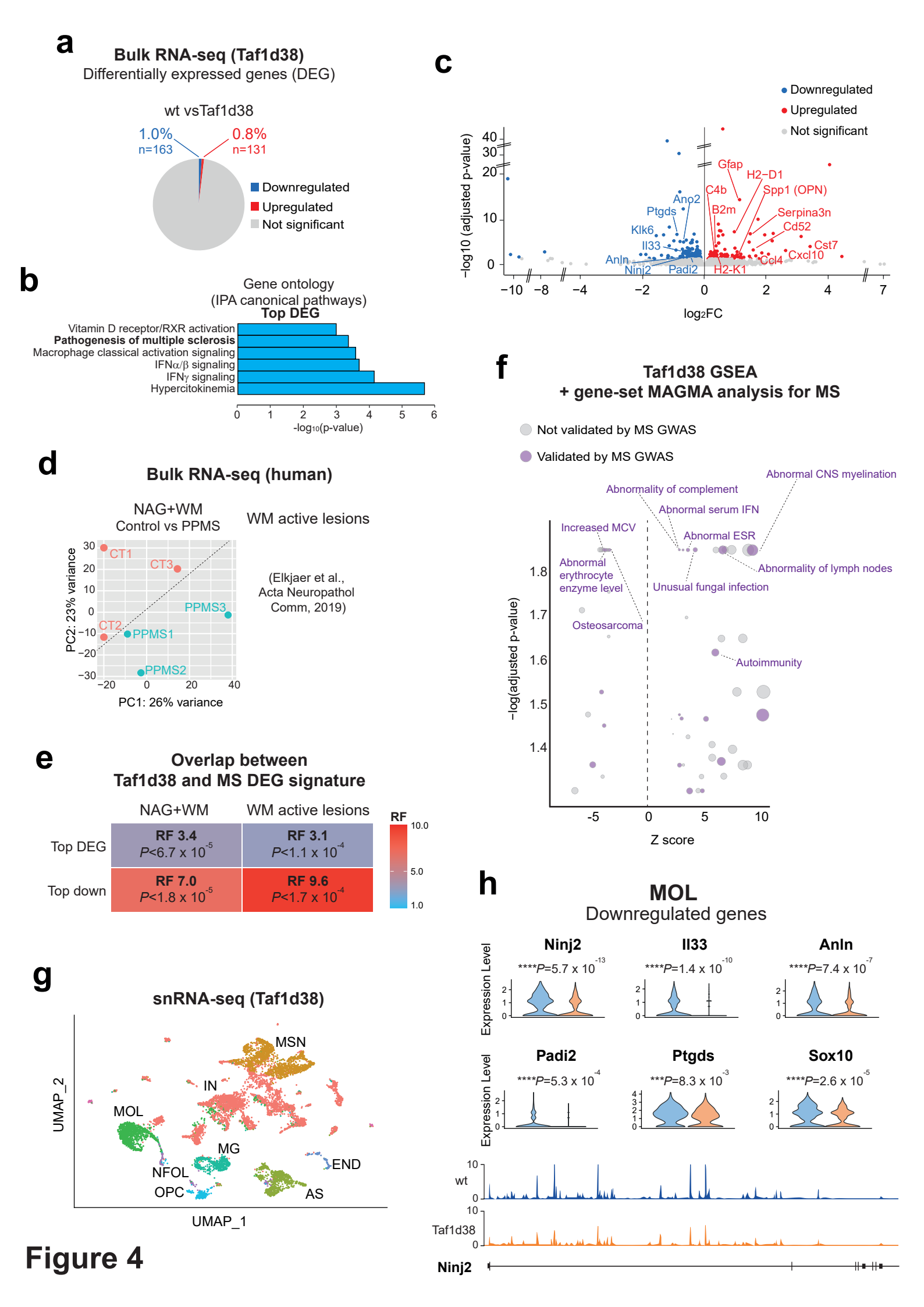

a
Bulk RNA-seq (Taf1d38)
Differentially expressed genes (DEG)
wt vsTaf1d38
1.0%
n=163
0.8%
n=131
Downregulated
Upregulated
Not significant
b
Gene ontology
(IPA canonical pathways)
Top DEG
Vitamin D receptor/RXR activation
Pathogenesis of multiple sclerosis
Macrophage classical activation signaling
IFNα/β signaling
IFNγ signaling
Hypercitokinemia
-log10(p-value)
c
Downregulated
Upregulated
Not significant
−log10 (adjusted p-value)
log2FC
Gfap
H2−D1
C4b
Spp1 (OPN)
Ano2
B2m
Ptgds
Serpina3n
Klk6
Cd52
Il33
Cst7
Anln
Ninj2
Padi2
H2-K1
Ccl4
Cxcl10
d
Bulk RNA-seq (human)
NAG+WM
Control vs PPMS
WM active lesions
CT1
CT3
CT2
PPMS3
PPMS1
PPMS2
PC2: 23% variance
PC1: 26% variance
(Elkjaer et al.,
Acta Neuropathol
Comm, 2019)
e
Overlap between
Taf1d38 and MS DEG signature
NAG+WM
WM active lesions
Top DEG
Top down
RF 3.4
P<6.7 x 10-5
RF 3.1
P<1.1 x 10-4
RF 7.0
P<1.8 x 10-5
RF 9.6
P<1.7 x 10-4
RF
f
Taf1d38 GSEA
+ gene-set MAGMA analysis for MS
Not validated by MS GWAS
Validated by MS GWAS
Abnormal CNS myelination
Abnormality of complement
Abnormal serum IFN
Increased MCV
Abnormal ESR
Abnormal erythrocyte enzyme level
Abnormality of lymph nodes
Unusual fungal infection
Osteosarcoma
Autoimmunity
−log(adjusted p-value)
Z score
g
snRNA-seq (Taf1d38)
MSN
IN
MOL
MG
NFOL
OPC
END
AS
UMAP_2
UMAP_1
h
MOL
Downregulated genes
Ninj2
Il33
Anln
Padi2
Ptgds
Sox10
Expression Level
wt
Taf1d38
Ninj2

Figure 4

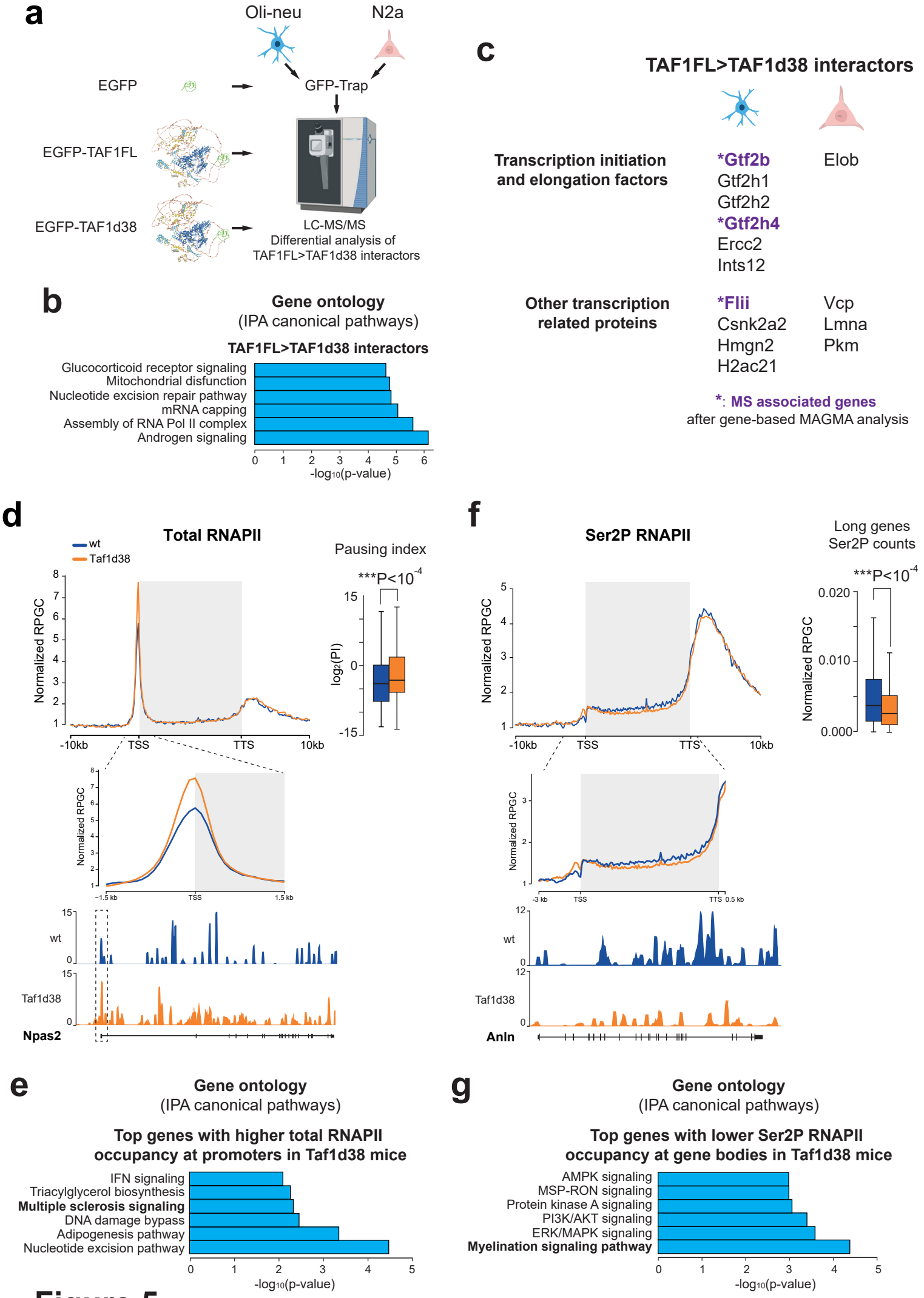
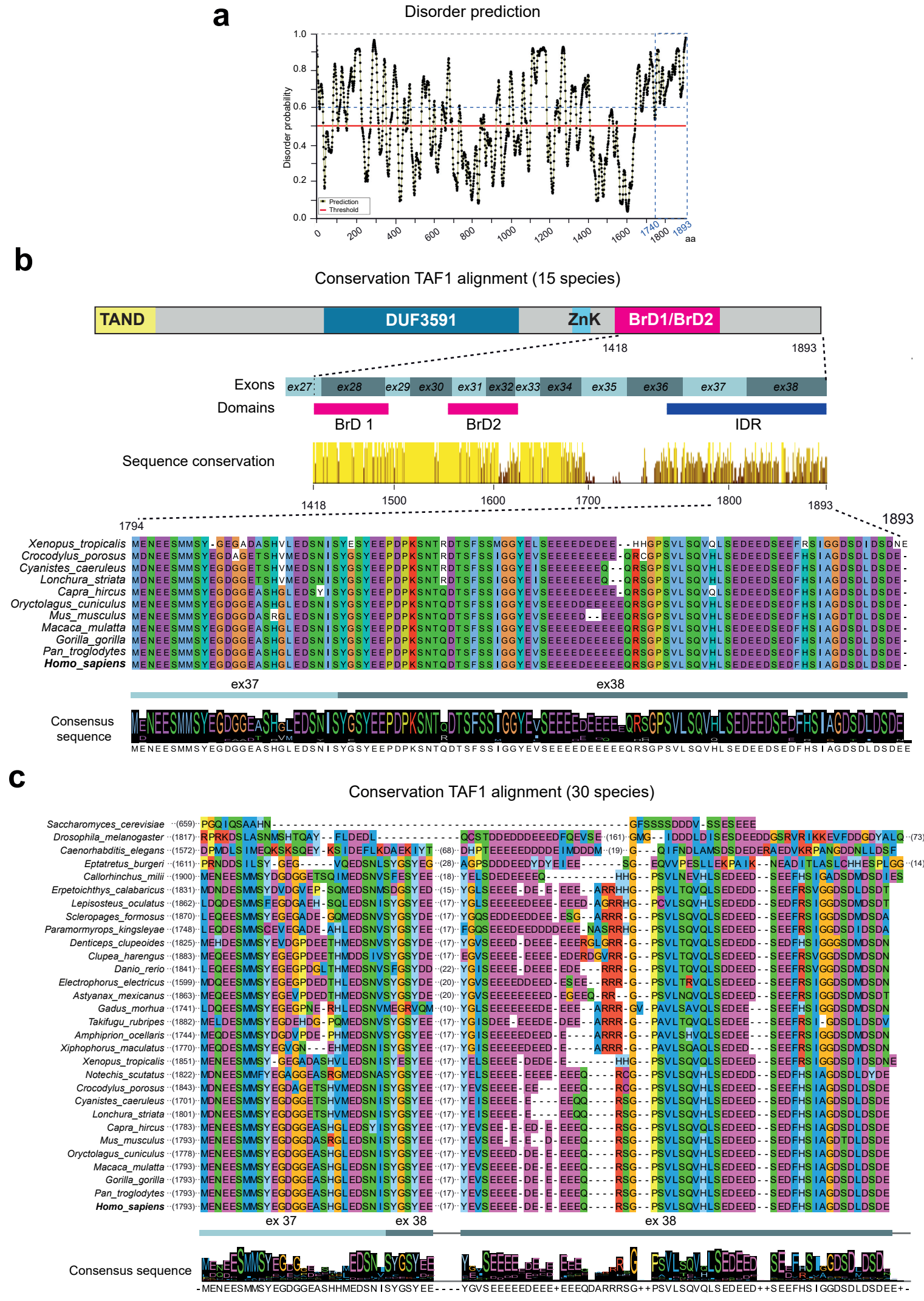
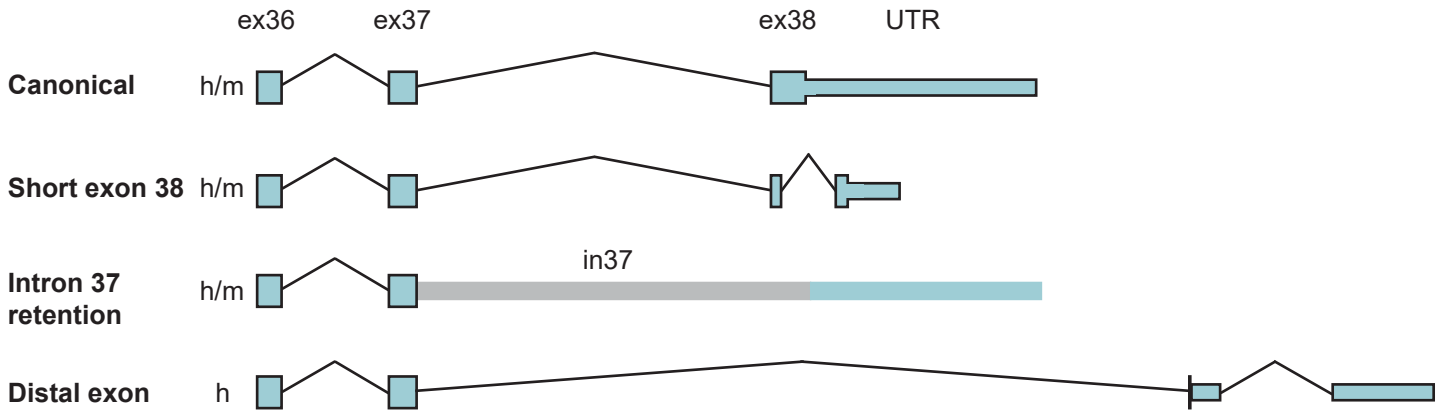

a
Oli-neu
N2a
EGFP
GFP-Trap
EGFP-TAF1FL
EGFP-TAF1d38
LC-MS/MS
Differential analysis of
TAF1FL>TAF1d38 interactors
b
Gene ontology
(IPA canonical pathways)
TAF1FL>TAF1d38 interactors
Glucocorticoid receptor signaling
Mitochondrial disfunction
Nucleotide excision repair pathway
mRNA capping
Assembly of RNA Pol II complex
Androgen signaling
-log10(p-value)
c
TAF1FL>TAF1d38 interactors
Transcription initiation and elongation factors
*Gtf2b
Gtf2h1
Gtf2h2
*Gtf2h4
Ercc2
Ints12
Elob
Other transcription related proteins
*Flii
Csnk2a2
Hmgn2
H2ac21
Vcp
Lmna
Pkm
*: MS associated genes
after gene-based MAGMA analysis
d
Total RNAPII
wt
Taf1d38
Normalized RPGC
-10kb
TSS
TTS
10kb
Pausing index
***P<10-4
log2(PI)
-1.5 kb
1.5 kb
Npas2
f
Ser2P RNAPII
Long genes
Ser2P counts
-3 kb
0.5 kb
Anln
e
Gene ontology
(IPA canonical pathways)
Top genes with higher total RNAPII occupancy at promoters in Taf1d38 mice
IFN signaling
Triacylglycerol biosynthesis
Multiple sclerosis signaling
DNA damage bypass
Adipogenesis pathway
Nucleotide excision pathway
-log10(p-value)
g
Gene ontology
(IPA canonical pathways)
Top genes with lower Ser2P RNAPII occupancy at gene bodies in Taf1d38 mice
AMPK signaling
MSP-RON signaling
Protein kinase A signaling
PI3K/AKT signaling
ERK/MAPK signaling
Myelination signaling pathway
-log10(p-value)

Figure 5

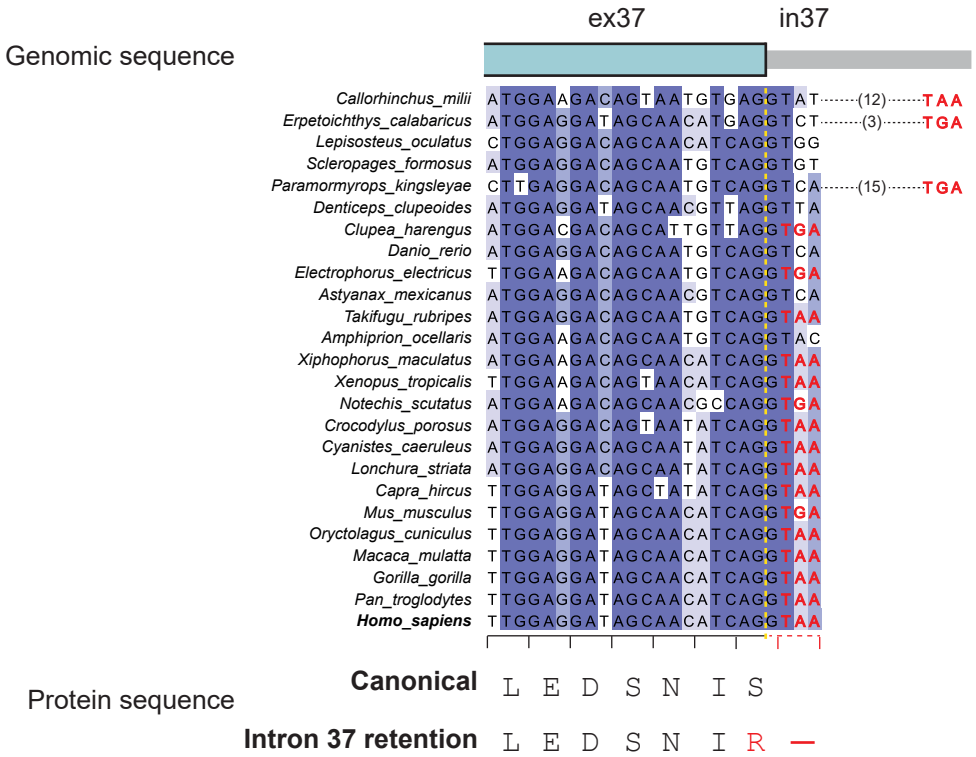


Extended Data Figure 1

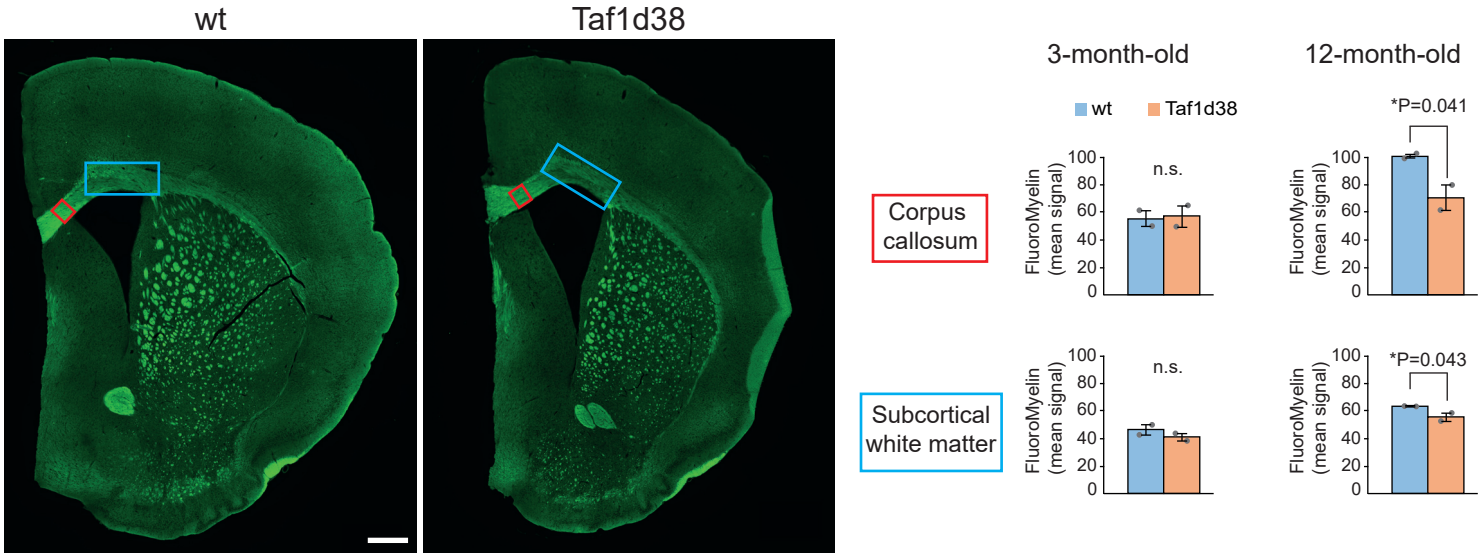
a

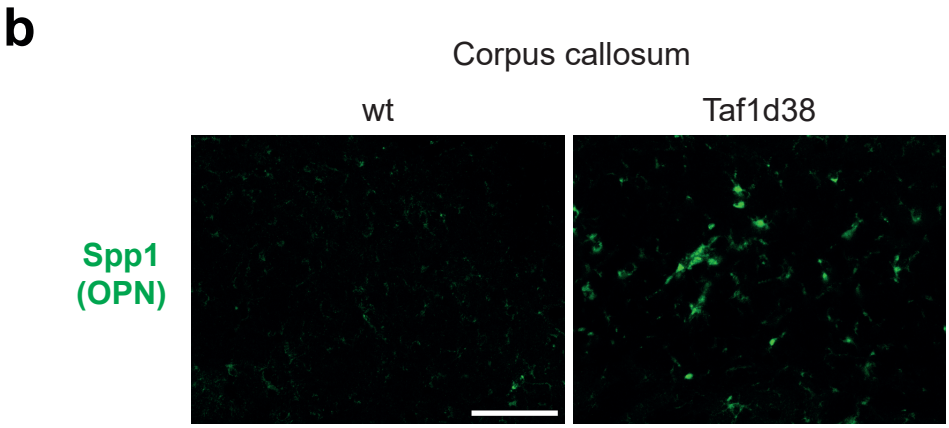


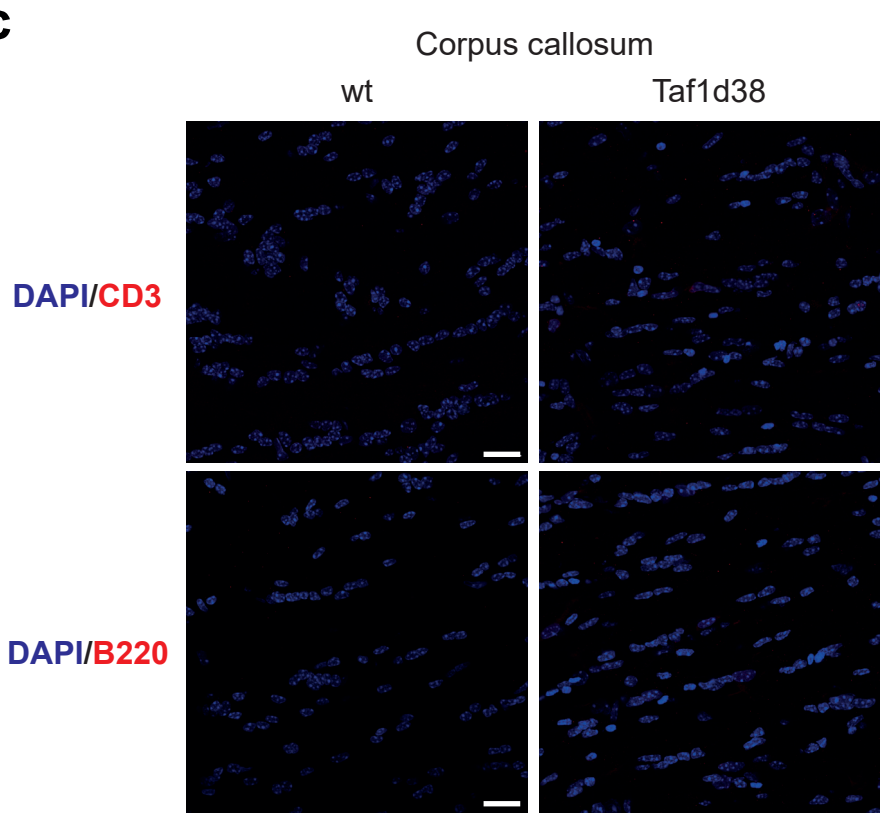
b



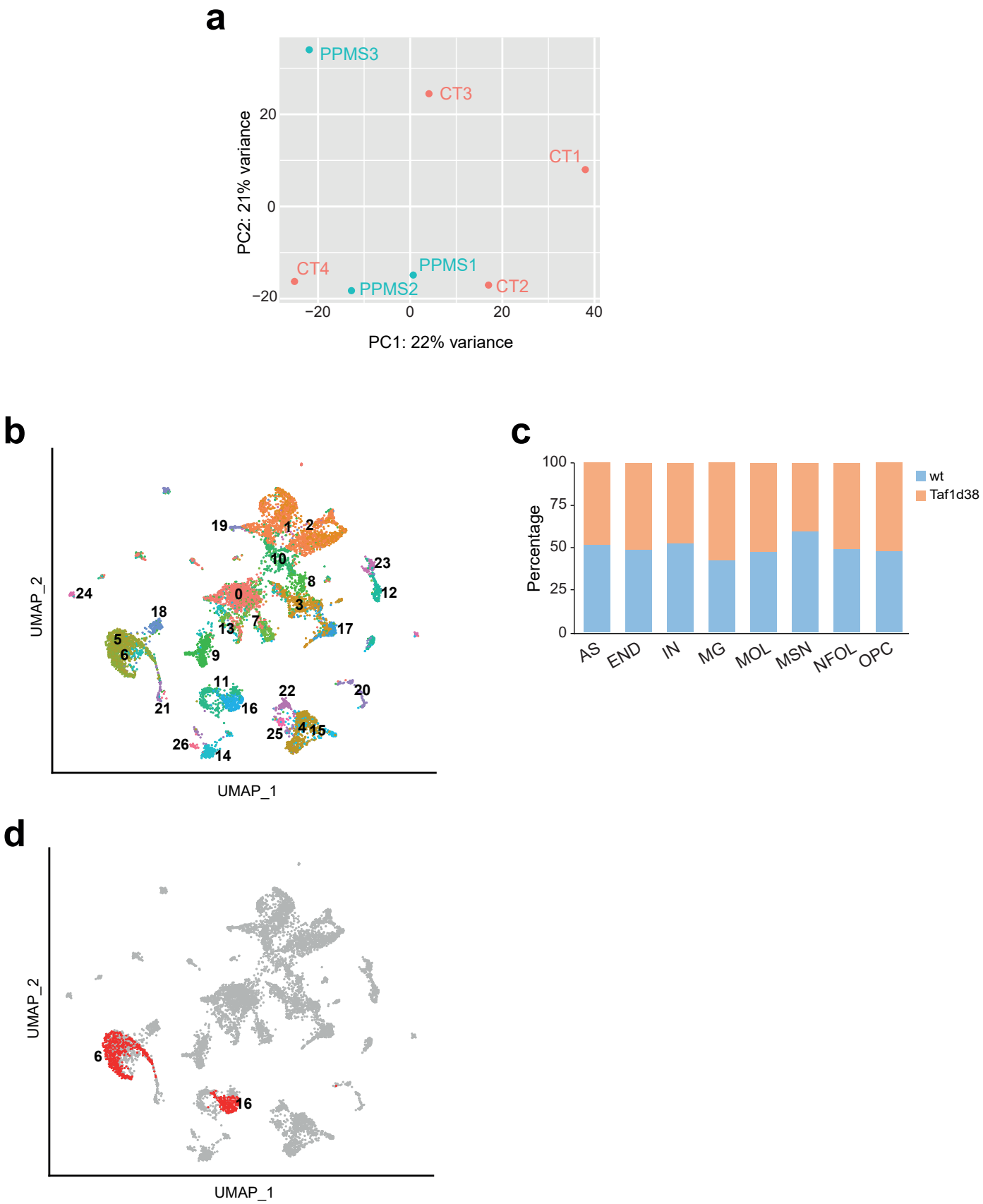
Extended Data Figure 2







Extended Data Figure 3



Extended Data Figure 4