IDENTIFICATION AND COPY NUMBER VARIANT ANALYSIS
OF ENHANCER REGIONS OF GENES CAUSING
SPINOCEREBELLAR ATAXIA

Fatemeh Ghorbani1, Eddy N. de Boer1, Michiel R. Fokkens1, Jelkje de Boer-Bergsma1, Corien C. Verschuuren-Bemelmans1, Elles Wierenga1, Hamidreza Kasaei2, Daan Noordermeer3, Dineke S. Verbeek1, Helga Westers1†, Cleo C. van Diemen1†*

1Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
2Department of Artificial Intelligence, University of Groningen, Groningen, the Netherlands
3Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France

* Correspondence:
Cleo C. van Diemen
c.c.van.diemen@umcg.nl
†These authors have contributed equally to this work.

Keywords: Spinocerebellar ataxia, Enhancers, 4C-seq, Luciferase assay, Genetic diagnosis, Copy number variants

Abstract
Currently, routine diagnostics for spinocerebellar ataxia (SCA) consist of analyses that look for polyQ repeat expansions and conventional variations affecting the proteins encoded by known SCA genes. Despite all the effort, ~40% of the patients still remain without a genetic diagnosis after routine tests. Increasing evidence suggests that variations in the enhancer regions of genes involved in neurodegenerative disorders can also be disease-causing. Since the enhancers of SCA genes are not yet known, it remains to be determined whether variations in these regions are a cause of SCA. In this pilot project, we aimed to identify the enhancers of the SCA genes ATXN1, ATXN3, TBP and ITPR1 in human cerebellum using 4C-seq experiments, publicly available datasets, reciprocal 4C-seq and luciferase assays. We then screened these enhancers for copy number variants (CNVs) in a cohort of genetically undiagnosed SCA patients. We identified two active enhancers for each of the four SCA genes in human cerebellar tissue. In addition, for TBP and ITPR1, we observed interactions between the putative enhancers of each gene. CNV analysis did not reveal any CNVs in the active enhancers of
the four SCA genes in any of the genetically undiagnosed SCA patients. However, in one patient, we noted a CNV deletion with unknown clinical significance near one of the active enhancers of \textit{ITPR1}. These results not only reveal elements involved in regulation of SCA genes, they can also lead to the discovery of novel SCA-causing genetic variants. As enhancer variations are increasingly recognized to cause brain disorders, screening the enhancers of \textit{ATXN1}, \textit{ATXN3}, \textit{TBP} and \textit{ITPR1} for variations other than CNVs and identifying and screening enhancers of other SCA genes might elucidate the genetic cause in genetically undiagnosed patients.

**Introduction**

The spinocerebellar ataxias (SCAs) are a heterogeneous group of autosomal dominantly transmitted neurodegenerative disorders that arise from progressive degeneration of the cerebellum and brainstem [1]. SCAs are clinically characterized by loss of balance and coordination, leading to severe disability. In the clinic, SCA patients are genetically diagnosed based on the presence of polyQ repeat expansions and conventional variations within protein-coding and splice regions of known SCA genes, but this still leaves a group of SCA patients without a genetic diagnosis. The presence of known variants within non-coding regulatory regions (5'UTR, promoter and 3'UTR) of SCA genes causing SCA8, SCA12 and SCA37 [2-4] demonstrate that genetic variations that affect gene regulation can also play an important role in SCA pathogenesis. Enhancers are another non-coding regulatory element that can activate the transcription of a gene in a tissue-specific manner [5], with increasing evidence showing that variations in enhancers, which can be located up to several 100s of kilobases from their target genes, can cause neurodevelopmental and neurodegenerative disorders such as intellectual disability, Alzheimer’s disease, Parkinson’s disease, Schizophrenia and Rett-like syndrome [5-9]. This led us to hypothesize that enhancers of SCA genes may also harbor variants that cause SCA. However, the enhancers of SCA genes and their role in disease pathogenesis have not yet been studied. Identifying gene enhancers is challenging due to tissue-specificity, the large distance between enhancers and their target genes and the existence of multiple enhancers for a single gene [10]. While the PubMed and enhancer databases include ~1.6 million putative enhancers specifically identified in the human brain, their target genes are mostly unknown due to the limited data on enhancer–promoter interactions [11 and 12]. Thus, it is still necessary to experimentally identify and validate gene-specific enhancers.

Given the high number of SCA genes and the heavy workload required to identify and validate gene-specific enhancers, we carried out a pilot study to identify enhancers of four SCA genes – \textit{ATXN1}, \textit{ATXN3}, \textit{TBP} and \textit{ITPR1} – in human adult cerebellum. These four genes are either causal for the most common SCA types or dosage-sensitive. Identifying cerebellar-specific enhancers of SCA genes can expand our knowledge of tissue-specific molecular mechanisms governing the regulation of SCA genes and provide new avenues for genetic diagnostics.
Materials and Methods

To identify enhancers of \( ATXN1 \), \( ATXN3 \), \( TBP \) and \( ITPR1 \) in the human cerebellum, we used the following workflow (Figure 1): (I) circularized chromosome conformation capture sequencing (4C-seq) on human cerebellum to capture genomic regions interacting with the promoters of the four SCA genes, (II) peak-calling to identify \( \text{cis} \)-genomic regions that interact significantly with gene promoters (4C-contact peaks), (III) annotation of these 4C-contact peaks using publicly available data sets to identify putative enhancers (PutEs), (IV) reciprocal 4C-seq for the PutEs to confirm interaction with the associated gene promoters (R-PutEs) and (V) analysis of the enhancer activity of R-PutEs using luciferase assays to identify active enhancers. Following identification of enhancers, we screened these regions for copy number variants (CNVs) in a cohort of genetically undiagnosed SCA patients.

![Workflow](https://example.com/workflow.png)

**Figure 1.** Workflow used to identify enhancers of \( ATXN1 \), \( ATXN3 \), \( TBP \) and \( ITPR1 \) in the human cerebellum.

4C-sequencing

**Promoter viewpoint selection.** We obtained the promoter regions and transcription start sites (TSSs) of \( ATXN1 \), \( ATXN3 \), \( TBP \) and \( ITPR1 \) from the Eukaryotic promoter database (https://epd.epfl.ch/index.php) and Database of TSS (https://dbtss.hgc.jp/), respectively. Promoter regions selected as viewpoints included the TSS (Supplementary Table 1).

**4C template preparation.** Cerebellar tissue from three healthy individuals was obtained from the Netherlands Brain Bank (http://www.brainbank.nl/) and stored at -80°C. For each 4C-template, 100 mg of frozen tissue was dissected at -20°C and homogenized on ice in pre-cooled 2 ml PBS with 10% Fetal Bovine Serum (FBS) using a cold 15 ml Dounce homogenizer. The cell suspensions obtained from the cerebellar tissue were immediately processed using the 4C protocol of Krijger et al. [13]. In brief, after cross-linking the cells with 2% formaldehyde, cells were lysed and DNA was digested with 400 U of \( NlaIII \) restriction enzyme (RE) (New England Biolabs, Ipswich, Massachusetts, United States) and ligated with 50 U of T4 DNA ligase (Merck, Darmstadt, Germany). Following de-crosslinking of the ligated DNA concatemers with proteinase K (Thermo Fisher Scientific, Waltham, Massachusetts, United States), the DNA was purified with NucleoMag P-beads (Macherey-Nagel, Düren, Germany). A second round of digestion was performed with 50 U of \( DpnII \) or \( Csp6I \) RE (Thermo Fisher Scientific), followed by ligation with 100 U T4 DNA ligase and a purification step with NucleoMag P-beads. Two 4C-templates were generated using the REs. For each 4C-template, three replicates were generated using human cerebellum from the three individuals. The 4C template used for each promoter viewpoint can be found in Supplementary Table 1.
**Preparation of 4C sequencing libraries.** Primer design and sequencing library preparation were performed as described in Krijger et al. [13], with some adjustments to the library preparation step. For each viewpoint of interest, specific reading and non-reading primers were designed for the first PCR to amplify genomic regions interacting with the viewpoints of interest (Supplementary Table 2). Per viewpoint, this PCR was performed in eight reactions of 25 µl (200 µl total) using 150 ng of 4C-template per reaction and the Expand Long Template PCR System (Roche, Basel, Switzerland) (Supplementary Tables 3 and 4). After pooling the PCR products, AMPure XP beads (Beckman Coulter, Brea, California, United States) were used to purify the products. A second round of PCR was performed on the initial PCR products (Supplementary Tables 5 and 6) with universal primers containing Illumina adaptor sequences and indexes (Supplementary Table 2). AmPure XP beads were used to purify the final PCR products. The PCR products were analyzed with the Agilent D5000 Screen Tape System (Agilent Technologies, Santa Clara, California, United States) to measure the fragment size and DNA concentrations in the range of 250–5000 bp.

**Sequencing.** We pooled 18 to 24 4C-seq libraries obtained from different viewpoints/replicates. The pools were loaded conservatively (0.9 pM), including 25% PhiX to ensure optimal diversity for base calling. Sequencing was performed on an Illumina Nextseq500 (Illumina, San Diego, California, United States) using the Mid-Output v2 kit with a single-read run and 75 bp read length.

**Data processing and 4C contact peak-calling.** Sequenced libraries were first demultiplexed based on the reading primers to separate the reads of each gene. We then used Illumina indexes introduced by the second PCR to separate the replicates of each gene. Demultiplexed data were analyzed using the 4C-seq pipeline of Krijger et al. [13]. In brief, reads were trimmed based on the first RE site in order to extract the captured sequence that also carries the RE motif. The trimmed reads were then mapped to the human reference genome (hg19) using Bowtie2. To count reads that map to RE fragment ends, the reference genome was *in silico* digested and the fragment ends were mapped against the genome to identify fragment ends that were unique to different RE combinations. After smoothing and normalizing the read counts, the data were further analyzed in R using the PeakC peak-calling algorithm to identify genomic regions with statistically significant interactions with the viewpoints of interest known as 4C-contact peaks [14]. Peak-calling was based on three biological replicates with parameters set on alphaFDR=0.05, qWr=1 and qWd=1.5.

**Identification of PutEs**

To identify the PutEs of *ATXN1, ATXN3, TBP* and *ITPR1* in the human cerebellum, the 4C-seq contact peaks were intersected and annotated using publicly available datasets in ENCODE ([https://www.encodeproject.org/](https://www.encodeproject.org/)) and GEO ([https://www.ncbi.nlm.nih.gov/geo/](https://www.ncbi.nlm.nih.gov/geo/)). For the human cerebellar tissue, ATAC-seq (GSE101912) [15], ChIP-seq for H3K27ac (GSM1119154) and DNase-seq (GSM736538) data were used. Hi-C data (GSM2824367) from cerebellum were used for the analysis of topologically associated domains (TADs). The PyGenome track was used to analyze and visualize the genome tracks [16]. For gene annotation, we used the Genome Browser annotation track database to export data from: group: genes and gene predictions, track: UCSC genes and table: known genes ([https://genome.ucsc.edu/cgi-bin/hgTables](https://genome.ucsc.edu/cgi-bin/hgTables)). PutEs for each gene were identified as the section of the 4C-contact peak in which the chromatin was open (based on the ATAC-seq and DNase-seq data).
and carried the H3K27ac active enhancer marker. Hi-C data were used to identify TADs and study the location of the promoters and PutEs with respect to TADs. The data analysis and visualization procedures can be found at https://github.com/GhorbaniF/4CSeqEnhancerDetection.

**Reciprocal 4C-seq**

Reciprocal 4C-seq was performed for each PutE viewpoint (Supplementary Table 7) to identify the genomic regions interacting with the PutEs and confirm their interaction with the original gene promoters. The 4C-template used for the PutE viewpoints and the specific reading and non-reading primers can be found in Supplementary Tables 7 and 8. We used the same 4C-sequencing procedure described earlier to perform the reciprocal 4C-seq experiments. To identify regions significantly interacting with the PutE viewpoints (4C-contact peaks), we performed peak-calling with two biological replicates and alphaFDR=0.2, qWr=1 and qWd=1.5. When the original promoter was called as one of the 4C-contact peaks of the PutE, we considered the interaction of the promoter with the PutE to be confirmed, and these PutEs were designated Reciprocal PutEs (R-PutEs). The reciprocal 4C-seq data of the PutEs was also used to reveal interactions between the PutEs/R-PutEs identified for each gene.

**Luciferase assays**

Candidate regions were PCR-amplified using human genomic DNA (Roche, Basel, Switzerland) as template and cloned in pGL4.23 Firefly luciferase reporter vector (Promega, Wisconsin, Madison, United States). In addition, a previously reported validated enhancer region was PCR-amplified and cloned as a positive control [9]. Since it has been demonstrated that the sizes of transfected constructs are inversely correlated with luciferase activity [17], for the negative control, we cloned a 1479 bp stuffer DNA into the vector in order to ensure similar insert sizes for the test constructs (which ranged from 768 to 2690 bp) and the negative control construct (see primer sequences in Supplementary Tables 9, 10 and 11). Candidate regions >3 kb were amplified as two overlapping segments (size of the overlap was on average 1800 bp) and cloned into two separate vectors to facilitate the amplification and cloning procedures. * Constructs were transfected into SH-SY5Y cells because this cell line is a widely used model for neurodegenerative disorders [18]. The pGL4.23 constructs were co-transfected with a renilla internal control vector pRL-TK (Promega, Madison, Wisconsin, United States) in a 50:1 ratio in SH-SY5Y cells in 6-well plates with a seeding density of 1 million cells per well. Cells were grown in Dulbecco’s modified Eagle medium supplemented with 15% FBS and 1% penicillin-streptomycin in a 37°C incubator with 5% CO₂. Transfections were performed in triplicate using nucleofection (Lonza, Thermo Fisher Scientific, Waltham, Massachusetts, United States) according to the manufacturer’s instructions. Luciferase assays were performed 48h after transfection using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer’s instructions. The activity of the firefly and renilla for each sample were measured with a Promega GloMax® 96 Microplate Luminometer (Promega). Per well, the firefly luciferase activity was normalized by
dividing the firefly signal by the renilla signal. The fold change of the normalized firefly luciferase activity for each candidate region is expressed as:

\[
\text{Fold change} = \frac{\text{Normalized firefly luciferase activity of the candidate region}}{\text{Normalized firefly luciferase activity of the negative control}}
\]

Candidate enhancer regions with at least a two-fold increase in luciferase activity were regarded as active enhancers.

**CNV analysis in SCA patients using SNP array**

**Patient selection.** We studied 285 patients with cerebellar ataxia who were referred to the Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands for SCA genetic diagnostics. At time of inclusion in our study, all patients were without a genetic diagnosis and, for most, repeat expansions in the most frequent SCA genes (SCA1–3, 6, 7 and 17), conventional variants and CNV in the 36 SCA genes were excluded. Informed consent was not requested as this analysis was in line with the original request for SCA diagnostic testing.

**Array analysis.** Genome-wide SNP array was performed using the Infinium Global Screening Array-24 v3.0-EA-MD (Illumina), following the manufacturer’s protocol. Data processing was done with in-house pipelines ([https://github.com/molgenis/AGCT](https://github.com/molgenis/AGCT) and [https://github.com/molgenis/GAP](https://github.com/molgenis/GAP)). Processed data were uploaded into NxClinical version 5.0 software (BioDiscovery, El Segundo, California, United States) to perform CNV analysis and visualize CNVs in the active enhancers of *ATXN1*, *ATXN3*, *TBP* and *ITPR1*. To exclude artifacts, only CNVs comprising at least 10 subsequent SNPs were selected for further analysis. An in-house genome-wide CNV database comprising data from 3,280 healthy Dutch controls was used to filter out benign variants. Sanger sequencing was performed to confirm the break points of the identified CNVs (for primer sequences see Supplementary Table 12).

**Gene expression analysis of *ITPR1***

To assess the effect of a CNV deletion found near a potential *ITPR1* enhancer region in one patient, we collected blood from the patient in a PAXgene tube (PreAnalytiX, Hombrechtikon, Switzerland) for mRNA isolation. For controls, we obtained whole-blood RNA samples from three healthy unrelated individuals of different age (Amsbio, Cambridge, Massachusetts, United States). mRNA was reverse-transcribed using the RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific). To measure *ITPR1* expression levels in blood, quantitative RT-PCR analysis was performed using iTaq Universal SYBR® Green Supermix (Thermo Fisher Scientific) on a Thermo Fisher Quantstudio 7 flex Momentum machine according to the manufacturer’s instructions. The Geomean of two housekeeping genes, *GAPDH* and *Beta-actin*, was used for normalization of raw *ITPR1* expression values. Two sets of primers were used to measure *ITPR1* expression levels and one set was used for *GAPDH* and *Beta-actin* (Supplementary Tables 13). We used the delta-delta Ct method (2\(^{-\Delta\Delta Ct}\) method) to calculate the
relative fold change in gene expression. The relative gene expression for each individual is based on three independent replicates.

Results

Identifying enhancers of ATXN1, ATXN3, TBP and ITPR1 in human cerebellum

To identify the enhancers of the four SCA genes, we performed 4C-seq on human cerebellum to capture genomic regions that interact with the promoter viewpoints. The 4C-contact peaks (Supplementary Figure 1) were then overlapped with publicly available datasets to identify PutEs, and we performed reciprocal 4C-seq for the PutEs to confirm their interaction with the corresponding promoter (R-PutEs) (Supplementary Figures 2–5). Finally, we performed luciferase assays for the R-PutEs to identify the active enhancers of each gene. We present detailed descriptions of the results for each gene below.

ATXN1 enhancer identification. 4C-seq data analysis for ATXN1 promoter showed six significant 4C-contact peaks (Supplementary Figure 1A). Based on the overlap of the 4C-contact peaks with open chromatin and H3K27ac data, we identified two PutE regions for ATXN1: PutE1 and PutE2 (Figure 2A). The ATXN1 promoter and PutE2 were located in one TAD, whereas PutE1 was located in the neighboring TAD (Supplementary Table 15). An inter-TAD interaction between the two neighboring TADs [19] in which the ATXN1 promoter and PutE1 are located can explain the interaction between them. Both PutEs were confirmed to interact with the promoter through reciprocal 4C-seq (R-PutE1 and R-PutE2) (Supplementary Figure 2) and showed an increase in luciferase activity (2.4- and 7.2-fold, respectively) (Figure 3). R-PutE1 is located in the last intron of ATXN1. R-PutE2 is located in an intergenic region upstream of ATXN1. Based on the reciprocal 4C-seq results, no significant mutual interactions were observed between the R-PutEs and PutEs of ATXN1.

ATXN3 enhancer identification. 4C-seq data analysis for the promoter of ATXN3 identified eight significant 4C-contact peaks (Supplementary Figure 1B), from which five PutEs were identified: PutE1–PutE5 (Figure 2B). The ATXN3 promoter and PutE1–PutE5 were all located in one TAD (Supplementary Table 15). Reciprocal 4C-seq experiments for these PutEs confirmed that three of them, R-PutE1, R-PutE2 and R-PutE4, interacted with the promoter (Supplementary Figure 3), but only R-PutE1 and R-PutE2 showed an increase in luciferase activity (59.4- and 5.9-fold, respectively) (Figure 3). R-PutE1 is located in the first intron of PPP4R3A. R-PutE2 is located in the first exon and intron of TRIP11. Reciprocal 4C-seq experiments did not show mutual interactions between the R-PutEs and PutEs of ATXN3.

TBP enhancer identification. 4C-seq data analysis for the TBP promoter identified three significant 4C-contact peaks (Supplementary Figure 1C), from which two PutE regions were identified: PutE1 and PutE2. Both are located in the same TAD as the TBP promoter (Figure 2C and Supplementary Table 15). Reciprocal 4C-seq experiments confirmed that both PutEs interacted with the TBP promoter (R-PutE1 and R-PutE2) (Supplementary Figure 4). Both R-PutE1 and R-PutE2 showed an increase in
luciferase activity (6.7- and 27.1-fold, respectively) (Figure 3). R-PutE1 encompasses DYNLT2, while R-PutE2 contains two long intergenic non-coding RNAs: LINC00242 and LINC00574. Reciprocal 4C-seq also revealed a mutual interaction between R-PutE1 and R-PutE2 (Supplementary Figure 5).

**ITPR1 enhancer identification.** 4C-seq data analysis for the ITPR1 promoter showed 11 significant 4C-contact peaks (Supplementary Figure 1D), of which seven were considered PutEs (Figure 2D). Three PutEs (PutE1, PutE2 and PutE3) and the ITPR1 promoter are within one TAD. The other four PutEs (PutE4–PutE7) are located in a neighboring TAD (Supplementary Table 15). However, as seen in figure 2D, a looping across the two neighboring TADs (an inter-TAD) [19] can explain the interaction between the promoter and PutE4–PutE7. Reciprocal 4C-seq experiments confirmed only the interaction of PutE2 and PutE7 with the promoter (R-PutE2 and R-PutE7) (Supplementary Figure 6). R-PutE2 showed a 23.9-fold increase in luciferase activity. For cloning purposes, R-PutE7 was analyzed as two fragments (A and B) (Supplementary Table 9). While fragment A showed a 7.3-fold increase in luciferase activity, fragment B showed no increase (Figure 3). R-PutE2 encompasses intron 8 of SUMF1 and exon 1 and intron 1 of SETMAR. Fragment A of R-PutE7 is located in intron 1 of the ncRNA BHLHE40-AS1.

Reciprocal 4C-seq analysis showed significant mutual interactions between PutEs and R-PutEs of ITPR1 (Supplementary Figure 7) including: PutE1 with R-PutE2, PutE4 with R-PutE7, PutE4 with PutE6, PutE5 with PutE6, PutE5 with R-PutE7 and PutE6 with R-PutE7 (Figure 4A and B). Based on this, we decided to check the enhancer activity for PutE1, PutE4, PutE5 and PutE6 (Figure 4C). Due to their size, PutE1 and PutE5 were both analyzed as two separate fragments (A and B). Both PutE1 fragments showed an increase (9.5- and 4.6-fold, respectively) in luciferase activity, but only fragment B of PutE5 showed an increase (6.5-fold). PutE4 did not show any enhanced luciferase activity, whereas PutE6 showed an increase (6.9-fold) (Figure 4C). Although PutE1, PutE5 and PutE6 are active enhancer regions, they cannot be considered as enhancers for ITPR1 because we could not confirm their interaction with the ITPR1 promoter.
Figure 2. Identification of enhancers of (A) ATXN1, (B) ATXN3, (C) TBP and (D) ITPRI in the human cerebellum. The red track shows the 4C-seq data generated for the promoter of the corresponding gene. Black boxes show genomic regions significantly interacting with the promoters (4C-contact peaks). Blue, green and black tracks show publicly available human cerebellum datasets for ATAC-seq, DNase-seq and ChIP-seq for H3K27ac, respectively. Red, blue and green dashed boxes represent, respectively, putative enhancers (PutEs), PutEs confirmed to interact with their associated promoter through reciprocal 4C-seq (R-PutEs), and R-PutEs confirmed to have an enhancer activity. The bottom track shows the genomic coordinates and the genes (UCSC gene track; hg19). TADs obtained from the human cerebellum HiC data are depicted in the bottom plot.
Figure 3. Fold change of luciferase activity for R-PutEs of ATXN1, ATXN3, TBP and ITPR1. Fold change is expressed as the ratio of the normalized luciferase activity (firefly/renilla) of the candidate region to that of the negative control (pGL4.23 Firefly luciferase reporter vector with a 1479 bp stuffer DNA). A previously reported validated enhancer was used as a positive control [9]. Fold change values are based on the means of three independent experiments. Standard deviations are shown as error bars.
Figure 4. Analysis of the PutE–R-PutE interactions for ITPRI. (A) 4C-seq and reciprocal 4C-seq data for the ITPRI promoter and PutEs revealed mutual interactions between PutEs and R-PutEs. Red bars: PutEs. Green bars: R-PutEs. Dashed bar: ITPRI promoter. Green arrows: mutual interaction between the ITPRI promoter and R-PutEs. Black arrows: mutual interactions between PutEs and R-PutEs. Chromosomal locations and gene annotations are indicated underneath the figure. (B) Schematic representation of the hypothetical interactions between ITPRI PutEs and R-PutEs. Only R-PutE2 and R-PutE7 mutually interact with the ITPRI promoter. Asterisks indicate PutEs and R-PutEs which showed an increase in luciferase activity. Mutual interactions between PutEs and R-PutEs are shown as gray hatching. (C) Fold change of luciferase activity for PutE1, R-PutE2, PutE4, PutE5, PutE6 and R-PutE7. Fold change is expressed as the ratio of normalized luciferase activity (firefly/renilla) of the candidate region to that of the negative control (pGL4.23 Firefly luciferase reporter vector with a 1479 bp stuffer DNA). A previously reported validated enhancer was used as a positive control. Fold change values are based on the means of three independent experiments. Standard deviations are shown as error bars.
We next screened the enhancers of *ATXN1*, *ATXN3*, *TBP* and *ITPR1* for CNVs in a cohort of genetically undiagnosed SCA patients. CNV analysis did not reveal deletions or duplications in these regions in any of the patients. However, in one patient, we noted a CNV deletion in one of the 4C-contact peaks of the *ITPR1* promoter that did not overlap with the delineated active enhancer in this region (PutE1) (Figure 5). This CNV deletion had a minimum size of 40,590 bp (20 SNP probes, GRCh37; chr3:3,799,881-3,840,471). Examination of the SNPs flanking the deleted region allowed us to delimit the deletion to 3,836 bp on the centromeric side (between rs711544 and rs139693033) and 294 bp on the telomeric side (between rs711582 and rs11705993). To identify the breakpoints of the deletion within the delimited regions, we designed several sets of primers at different locations and performed Sanger sequencing on the PCR products. The deleted region was 41,104 bp (UCSC; GRCh37; chr3:3,797,792-3,838,896). It encompasses part of intron 11 of the old version of the annotated *SUMF1* transcript (NM_182760) on the telomeric side and an intergenic region upstream of this transcript on the centromeric side, but it does not encompass the currently used *SUMF1* transcript (NM_182760.4).

Although the deletion did not overlap with PutE1, as the array data from the in-house control population did not show deletions affecting this locus, we still decided to perform follow-up studies for this deletion. Luciferase assay for the deleted region of the 4C-contact peak showed a 3-fold increase in luciferase activity (data not shown).

**Figure 5.** Schematic representation of the deleted region in one patient and its overlap with the 4C-contact peak of *ITPR1*. Upper plot shows the significant 4C-contact peaks (red bars) of the *ITPR1* promoter. Lower plot shows the deleted region in the patient (GRCh37: chr3:3,797,792-3,838,896; blue dotted box), which partly overlaps (dashed box) the broad 4C-contact peak (GRCh37: chr3:3,836,985-3,845,608; green box) of the *ITPR1* promoter. PutE1 (red line in the lower plot), the active enhancer region delineated in the 4C-contact peak, does not overlap with the deleted region in the patient. Gene annotations are indicated underneath the figure. The deletion encompasses part of intron 11 of the old version of the *SUMF1* transcript (NM_182760) but does not encompass the updated transcript (NM_182760.4).

To assess whether the deletion in the 4C-contact peak affected the expression levels of *ITPR1*, we measured *ITPR1* expression in blood from the patient carrying the deletion and three unrelated, apparently healthy individuals. The relative *ITPR1* expression levels varied markedly between the three
healthy individuals. While the patient showed $ITPR1$ expression similar to control 1, they showed a 4–5-fold decrease in expression compared to controls 2 and 3 (Figure 6 and Supplementary Figure 8).

**Figure 6.** Expression of $ITPR1$ in blood of a SCA patient carrying a CNV deletion in a 4C-contact peak of the $ITPR1$ promoter compared to expression in three healthy individuals. The delta-delta Ct was used to determine the relative fold change in expression levels. Expression values are normalized to control 1. The relative fold change in expression for each individual was based on the means of three independent replicates. Standard deviations are shown as error bars.

**Discussion**

This is the first study to experimentally identify enhancers of four SCA genes, $ATXN1$, $ATXN3$, $TBP$ and $ITPR1$, in the human cerebellum in order to broaden our knowledge of molecular mechanisms regulating the expression of these genes and assist the discovery of novel genetic variants causing SCA. Using a comprehensive strategy involving 4C-seq on post-mortem human cerebellum tissue, annotation and filtering using publicly available datasets of the human cerebellum, reciprocal 4C-seq and functional validation using luciferase assays, we identified two specific enhancers for each of the four SCA genes in the human cerebellum. In our undiagnosed patient cohort, we did not identify a CNV in any of the active enhancers.

$ITPR1$ is a dosage-sensitive gene for which haploinsufficiency is known to cause SCA15 [20]. This means that there may be unidentified genetic variants within the enhancers of $ITPR1$ that cause lower expression levels, resulting in SCA15. We found two active enhancers for $ITPR1$, R-PutE2 and R-PutE7, and the latter is located in the enhancer-associated non-coding RNA BHLHE40-AS1 [21]. Reciprocal 4C-seq analysis of the other PutEs of $ITPR1$ (PutE1, PutE3, PutE4, PutE5 and PutE6) did not confirm their direct interaction with the $ITPR1$ promoter but did reveal a cluster of mutually interacting PutEs and R-PutEs. While enhanced luciferase activity demonstrated that PutE1, PutE5 and PutE6 are active enhancers, the significance of their interactions with the promoter of $ITPR1$ was not confirmed by our reciprocal 4C-seq experiments. However, reciprocal 4C-seq for PutE1, PutE5 and PutE6 showed significant 4C-contact peaks close to the $ITPR1$ promoter (Supplementary Figure 6), suggesting that these PutEs might be engaged in a more transient interaction with the target promoter [22 and 23]. The CNV deletion we identified in one patient overlapped with a 4C-contact peak of $ITPR1$, but not with PutE1, which was delineated in this peak based on public datasets. However, the...
absence of the deleted region in our in-house control population and potential enhancer activity confirmed with the luciferase assay led us to study the functional consequence of the deletion in the patient. A study performed by Di Gregorio et al showed that the reduced expression of ITTP1 in blood can be used as a marker to identify SCA15 patients [24]. In our study, the ITTP1 expression level in the patient was similar to that in one of the healthy individuals but much lower than that in two other healthy individuals. The inconsistency in the ITTP1 expression levels among the healthy individuals makes it difficult to draw a conclusion about the causality of this deletion. Future studies that include more healthy individuals in the control cohort, as well as familial segregation studies, might help reveal the significance of this deletion. Furthermore, we suggest future studies to screen the active enhancers of ITTP1, as well as ATXN1, ATXN3 and TBP, for other types of variants such as smaller deletions, insertions and single nucleotide variants. This would be particularly interesting for ATXN3 since there is evidence that certain single nucleotide variants in the 3′ UTR are associated with early onset SCA3 [25].

It has already been shown that, in addition to the expanded polyglutamine tract in ATXN1 protein that causes SCA1, slight increases in wild-type ATXN1 levels can also cause SCA1 in mice [26]. Studies suggest targeting the expression levels of ATXN1 at the RNA and protein level for therapeutic purposes in SCA1 patients [27, 28]. While mouse studies have already shown that the 3′ UTR and 5′ UTR of ATXN1 are crucial for ATXN1 regulation via miRNAs [26, 29], enhancers of ATXN1 have not yet been studied. Other studies similarly support that reducing the levels of ATXN3 mutant transcript/protein is a promising preventive treatment strategy [30, 31]. An expanded polyQ tract in ATXN3 causes SCA3, the most common type of SCA, a devastating disease for which there is currently no preventive treatment [31]. Studies have shown that ATXN3 is a non-essential protein and that reducing the expression level of both the normal and disease-causing allele will probably have tolerable side effects [30]. In mice, it has already been shown that inducing artificial miRNAs expressing RNA-interfering sequences that target the ATXN3 3′UTR can suppress the expression of ATXN3 transcripts, an approach that can be used as a preventive treatment for SCA3 [30]. In this study, we identified two active enhancers for ATXN1 (R-PutE1 and R-PutE2) and two for ATXN3 (R-PutE1 and R-PutE2). Of all the R-PutEIs tested in this study, ATXN3 R-PutE1 showed the highest fold increase in the luciferase activity (a 59.4-fold increase), making it the strongest active enhancer detected in our study. Thus, the active enhancers identified for ATXN1 and ATXN3 may also be used as potential targets to reduce the expression of these genes, providing novel treatment strategies for SCA1 and SCA3 patients. A new class of drugs, bromo-and extra-terminal (BET) inhibitors, that target enhancers to inhibit gene expression are currently under investigation for treatment of disorders linked to impaired enhancer function [32].

The general assumption is that most enhancer-promoter pairs reside in the same TAD [33], and most of the active enhancers we identified for ATXN1, ATXN3, TBP and ITTP1 are indeed in the same TAD as the corresponding promoter. However, we also identified active enhancers not located in the same TAD as the promoter, including ATXN1 R-PutE1 and ATXN1 promoter that were located in two neighboring TADs and ITTP1 R-PutE7 and ITTP1 promoter that were also located in two different TADs. While the majority of regulatory loops are a result of intra-TAD looping, regulatory looping across TADs (inter-TAD looping) has also been reported [19]. The boundary of the two TADs
in which R-PutE1 and the ATXN1 promoter are located appears to be quite weak, and the interactions between the two TADs (inter-TAD) can explain the contact between R-PutE1 and the ATXN1 promoter. While the boundary of the two TADs containing R-PutE7 and the ITPR1 promoter appears to be strong based on the Hi-C data, resulting in two separate TADs, our 4C-seq data suggests a very strong interaction between the two TADs.

The disease mechanism of SCA17 is sequestration of wild-type TBP in polyQ-pathological TBP aggregates in a dominant-negative manner, causing the loss of wild-type TBP. TBP is an important general transcription factor that also plays a role in the mechanism of other polyQ disorders (SCA1, SCA2, SCA3, Huntington’s disease and dentatorubral-pallidoluysian atrophy) by sequestering in the polyQ-pathological aggregates of those diseases [34, 35]. The loss of TBP has a serious impact on the highly regulated process of DNA transcription where any minor disruption could lead to major consequences for the cells [35, 36]. The two active enhancers of TBP that we identified, R-PutE1 and R-PutE2, encompass DYNLT2 and LINCO0242/LINCO0574, respectively. During transcriptional activation, long intergenic non-protein-coding RNAs (LINCRNAs) are transcribed from their genes and can act as enhancer-RNAs (eRNAs) [37, 38]. Chromatin looping between the genes coding for LINCRNAs and the target promoter brings the transcribed LINCRNAs eRNAs near the target gene promoter to allow gene activation [37]. Reciprocal 4C-seq data of R-PutE1 (encompassing DYNLT2) and R-PutE2 (encompassing LINCO0574) revealed a mutual interaction between these two regions (Supplementary Figure 5), suggesting that LINCO0574 has a regulatory effect on DYNLT2. Surprisingly, in a recent study, LINCO0574 was suggested to regulate the expression of the nearby gene DYNLT2 (TCTE3) [39]. Thus, our results not only show that LINCO0574 regulates expression of TBP, they also support LINCO0574’s regulatory role for DYNLT2. The enhancers of TBP we identified will provide insight into the molecular mechanisms governing the expression of this tightly regulated gene in polyQ-mediated neurodegenerative disorders.

Our study has several limitations. First, it is known that enhancers can be more than a Mb away from their target promoters [40]. However, the signal of the 4C-seq technique is highest and most reproducible within ~500 kb of the viewpoint, limiting 4C-seq’s ability to detect weaker/more transient interactions at larger distances [41]. We could therefore have missed relevant enhancers in these more distant regions. Such enhancers would be detectable using extremely deep-sequenced Hi-C or Micro-C, two techniques that can identify interactions with increased signal over background at any length-scale [42 and 43]. Second, the activity of the enhancers was assessed with luciferase assays, which have their own drawbacks. For example, transcription factors present in the correct cell type and genomic context necessary for the proper functioning of the enhancers might be missing in this in vitro system, although we did try to mimic environment by using the most appropriate cell line available. Moreover, all the enhancers were tested with an identical vector carrying a minimal promoter, but this can be a problem for enhancers that are only functional with their own target promoters. Third, we cannot conclude with complete certainty that the identified active enhancers are also functional for their target genes in the cerebellum, as enhancers can be active at different stages and timepoints. Fourth, the SNP array that we used to detect CNVs has a low genotyping resolution, and thus we may have missed smaller CNVs in the active enhancers.
A more direct way to study the impact of the identified active enhancers on \textit{ATXN1}, \textit{ATXN3}, \textit{TBP} and \textit{ITPR1} expression would be to knock-out the active enhancers with CRISPR/Cas9 in a proper cell line (e.g. SHSY5Y cells) and assess the target gene expression. Moreover, testing the functionality of the active enhancers in zebra fish can provide further insight into tissue-specificity and \textit{in vivo} activity patterns during development [9, 44]. Finally, as we only searched for CNVs in the active enhancers, we strongly suggest screening these regions for other types of variations in genetically undiagnosed patients in order to discover novel genetic variants.

In conclusion, we identified several active enhancers for \textit{ATXN1}, \textit{ATXN3}, \textit{TBP} and \textit{ITPR1}. The output of this study provides insights into the regulation of the genes in the human cerebellum, which can lead to the discovery of novel genetic variants causing SCA and to new therapeutic options for SCA patients. Although no CNVs were identified in the active enhancers in our SCA cohort, it may be worthwhile to look for CNVs in these elements in other SCA cohorts. Moreover, the identification of these elements would provide a starting point to study other types of variations in these regions. Since it is increasingly recognized that genetic variations that disturb enhancer functionality can cause brain disorders, we strongly suggest identifying and screening the enhancers of other SCA genes, in particular for those genes with known variants within regulatory regions (\textit{DAB1}, \textit{PPP2R2B} and \textit{ATXN8OS}), to explain a fraction of missing genetic diagnoses for SCA patients.

\textbf{Conflict of interest}

The authors report no conflict of interest.

\textbf{Authors Contribution}

FG, CC-D, HW and DS-V designed the work. FG, EN-B, MR-F, JB-B and EW performed the experiments. FG and HK analyzed the data. CC-D, HW, DN and DS-V supervised the study. CV-B collected the patient sample. FG took the lead in writing the manuscript. CC-D and HW revised the work for intellectual content. All authors provided critical feedback and helped shape the research, analysis and manuscript.

\textbf{Funding}

Funding for this study came from a Rosalind Franklin Fellowship from University Medical Center of Groningen (DSV) and a PhD grant from the Graduate School of Medical Science of the University of Groningen (FG).
Acknowledgments

We sincerely thank all the patients for their participation in this study and the Netherlands Brain Bank for providing human cerebellum samples. We thank Kim de Lange and Marloes Benjamins-Stok for demultiplexing the sequencing data and Kate Mc Intyre for proofreading the manuscript.

References:


