Gene 509 (2012) 273–281

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Computational prediction of the PolyQ and CAG repeat spinocerebellar ataxia network based on sequence identity to untranslated regions

Jean L. Spence [a,*], Scott Wallihan [b]

[a] Omnitron Biosciences, USA
[b] UCSD Extension, San Diego, CA, USA

ARTICLE INFO

*Article history:*
Accepted 30 July 2012
Available online 6 August 2012

*Keywords:*
Spinocerebellar ataxia
UTR
Motif
Protein interaction networks
Bioinformatics

ABSTRACT

Computational prediction of biological networks would be a tremendous asset to systems biology and personalized medicine. In this paper, we use a moving window bioinformatic screen to identify transcripts with partial identity to the 5′ and 3′UTRs of the polyQ spinocerebellar ataxia (SCA) genes ATXN1, ATXN2, ATXN3, ATXN7, TBP and CACNA1A and the CAG repeat expansion gene PPP2R2B. We find that the bioinformatic screen enriches for transcripts that encode proteins that interact and that have functions relevant to polyQ SCA. Transcription control and RNA binding are the primary functional groups represented in the proteins from the combined screens. The insulin growth factor pathway, the WNT pathway, long term potentiation, melanogenesis and ATM mediated DNA repair pathways were identified as important pathways. UGUUU repeats were identified as an abundant motif in the SCA network and PAXIP1, CELF2, CREBBP, EBF1, PLEKHG4, SRSF4, C5orf42, NFIA, STK24, and YWHAG were identified as statistically significant proteins in the polyQ and PPP2R2B network.



## 1. Introduction

CAG repeat expansions have been identified as the cause of several inherited diseases including Huntington's disease, 8 forms of autosomal dominant spinocerebellar ataxia (SCA), spinal and bulbar muscular atrophy and Dentatorubral pallidoluysian atrophy (Zoghbi and Orr, 2000). The CAG repeats in ATXN1, ATXN2, ATXN3, CACNA1A, ATXN7, and TBP occur in the coding sequences resulting in an expanded poly glutamine (polyQ) region in SCA1, SCA2, SCA3, SCA6, SCA7 and SCA16, respectively (Durr, 2010; La Spada and Taylor, 2010). In SCA12, the CAG repeats occur in the 5′UTR/promoter of PPP2R2B, indicating that the DNA/RNA is also toxic (O'Hearn et al., 2001). SCA8 results from an expanded CAG repeat in ATXN8 which encodes primarily a polyglutamine repeat and expanded CTG repeats in the overlapping gene, ATXN8OS, which is transcribed into a non-coding RNA (Moseley et al., 2006). Both the coding and non-coding transcripts contribute to the disease through toxic gain of function by both RNA and protein (Moseley et al., 2006). Autosomal dominant spinocerebellar ataxia is an often fatal disease with no known cure (La Spada and Taylor, 2010; Zoghbi and Orr, 2000).

*Abbreviations:* SCA, spinocerebellar ataxia; PolyQ, poly glutamine; LTP, long term potentiation; Nt, nucleotide; UTR, untranslated region; DAVID, Database for Annotation, Visualization and Integrated Discovery; CAG, cytosine–adenosine–guanosine; RefSeq, NCBI Reference Sequence; ATXN1, ataxin 1; ATXN2, ataxin 2; ATXN3, ataxin 3; ATXN7, ataxin 7; CACNA1A, calcium channel, voltage-dependent, P/Q type, alpha 1A subunit; PPP2R2B, protein phosphatase 2 regulatory subunit B, beta isoform; TBP, TATA box binding protein; AR, androgen receptor; ATM, ataxia telangiectasia mutated; ATN1, atrophin 1; ATXN8, ataxin 8; ATXN8OS, ATXN8 opposite strand; C5orf42, chromosome 5 open reading frame 42; CACNB4, calcium channel, voltage-dependent, beta 4 subunit; CAMK2A, calcium/calmodulin-dependent protein kinase II alpha; CELF2, CUG triplet repeat, RNA binding protein 2; CFTR, cystic fibrosis transmembrane conductance regulator; CHUK, conserved helix–loop–helix ubiquitous kinase; CLSTN1, calsyntenin 1; CREBBP, CREB binding protein; CSNK2A1, casein kinase 2, alpha 1 polypeptide; CTBP1, C-terminal binding protein 1; DCLRE1C, DNA cross-link repair 1C; DUSP10, dual specificity phosphatase 10; EBF1, early B-cell factor 1; ELMO1, engulfment and cell motility 1; EPHA4, EPH receptor A4; FRS2, fibroblast growth factor receptor substrate 2; GABRA1, gamma-aminobutyric acid A receptor, alpha 1; GBA, glucosidase, beta, acid; GNAQ, guanine nucleotide binding protein, q polypeptide; GRIN1, glutamate receptor, ionotropic, N-methyl D-aspartate 1; GRIN2A, glutamate receptor, ionotropic, N-methyl D-aspartate 2A; HPRD, Human Protein Reference Database; IGF1, insulin-like growth factor 1; JPH3, junctophilin 3; KIF5B, kinesin family member 5B; KLC1, kinesin light chain 1; LRRK2, leucine-rich repeat kinase 2; MAF, v-maf musculoaponeurotic fibrosarcoma oncogene homolog; MCM9, minichromosome maintenance complex component 9; MECP2, methyl CpG binding protein 2; MIR9-2, microRNA 9-2; MLL2, myeloid/lymphoid or mixed-lineage leukemia 2; MRE11A, MRE11 meiotic recombination 11 homolog A; MSH2, mutS homolog 2; MYT1L, myelin transcription factor 1-like; NFIA, nuclear factor I/A; PABPN1, poly(A) binding protein, nuclear 1; PARK2, Parkinson disease 2 (parkin); PAXIP1, PAX interacting protein 1; PLCB1, phospholipase C, beta 1; PLEKHG4, pleckstrin homology domain containing, family G member 4; PPP3CA, protein phosphatase 3, catalytic subunit, alpha isoform; PRKCG, protein kinase C, gamma; PSEN1, presenilin 1; PSMD2, proteasome 26 S subunit, non-ATPase, 2; PTPN11, protein tyrosine phosphatase, non-receptor type 11; QKI, quaking homolog; RAI1, retinoic acid induced 1; RBM17, RNA binding motif protein 17; RDH12, retinol dehydrogenase 12; SCAMP5, secretory carrier membrane protein 5; SIAH1, seven in absentia homolog 1; SPRY1, sprouty homolog 1, antagonist of FGF signaling; SRSF4, serine/arginine-rich splicing factor 4; STK24, serine/threonine kinase 24; SYT11, synaptotagmin XI; TNFRSF1A, tumor necrosis factor receptor superfamily, member 1A; TRPM7, transient receptor potential cation channel, subfamily M, member 7; UTR, Untranslated Region; WRN, Werner syndrome, RecQ helicase-like; YWHAG, tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide.

* Corresponding author at: 3090 Admiral Avenue, San Diego, CA 92123, USA. Tel.: +1 858 576 0877.
*E-mail address:* jlspence@san.rr.com (J.L. Spence).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.07.068

The hallmark of the disease is uncoordinated limb movement that can escalate to uncontrolled breathing and swallowing resulting in death (La Spada and Taylor, 2010; Moseley et al., 2006).

There is a correlation between the length of the CAG repeat and the time of onset and severity of the disease in SCA1, 2, 3, 6, 7 and 17 (Reetz et al., 2011; Stevanin et al., 2000). However, the method by which the expanded CAG repeats result in disease is under debate and is likely to be multifactoral. It has been proposed that the polyQ repeats result in either toxic gain or loss of function of the protein or CAG repeats result in toxicity of the RNA or possible toxicity at the DNA through recruitment of cellular factors leading to cell death (Lin and Wilson, 2011). Although the CAG and polyQ repeats are toxic, the proteins containing the repeats play a critical role in the etiology of the disease. Some of the evidence for this lies in the fact that, although the polyQ containing proteins occur in multiple tissues, the deleterious effects of the polyQ expanded protein target specific cells. For example, polyQ expansions in Huntington's disease target striatal and cortical neurons, whereas polyQ expansions in ataxin 1 affect Purkinje cells (La Spada and Taylor, 2010). The local environment of the protein is critical as demonstrated by TBP, a housekeeping protein involved in the initiation of transcription by RNA polymerase II. Despite the universality of transcription by RNA polymerase II, polyQ expansions in this gene cause a well-defined disease that affects specific regions of the brain. The polyQ expansions may alter the function of the protein by causing misfolding leading to aggregation or by affecting the activity of the protein by altering its modification by phosphorylation, sumoylation, or acetylation (La Spada and Taylor, 2010). Experimentally, the pathology of the polyQ in ATXN1 has been attributed to both gain of function and loss of function (Lim et al., 2008).

Because the pathology of SCA involves both the polyQ expansion and its effect on the individual protein in specific locations in the brain, it is important to identify the networks for the individual proteins as well as the cellular response to the CAG expansions in the DNA and RNA and the polyQ expansions in the protein. Transcripts that encode proteins that function in common pathways should contain common regulatory sequences. Some of these are microRNA binding sites that are relatively short and difficult to define by bioinformatic analysis. Bioinformatic analysis of transcription factor binding sites has been previously used to define biological networks (Marbach et al., 2012; Zhao et al., 2012; Zhou et al., 2010). Transcription factor binding sites tend to be short and poorly conserved resulting in high error rates. Previously, we constructed a biological network for cystic fibrosis based on genes that shared partial identity to the cystic fibrosis 5′ and 3′ UTRs over 40 nt (Spence, 2009). In this manuscript, we use a bioinformatic screen to identify transcripts that have partial identity to the polyQ SCA and PPP2R2B UTRs over 40 to 60 nucleotides. We then construct protein–protein interaction maps with the proteins encoded by these transcripts. We present protein interaction maps that are universal and that are specific to the developing mouse cerebellum and Purkinje cell layer. These protein–protein interaction networks correlate with the literature and provide insight into the etiology of the disease. Moreover, transcripts that were identified in multiple UTR screens were found to be differentially regulated in the developing mouse cerebellum.

## 2. Materials and methods

### 2.1. Databases

The human RNA library, human.rna.fna.gz, was downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The protein interactions from the Human Protein Reference Database were downloaded from http://www.hprd.org/ and the protein interactions from BioGRID were obtained from http://thebiogrid.org/. These were integrated and redundancies were removed. The proteins in the mouse cerebellum and Purkinje cell layer were obtained from the Cerebellum Development Transcriptome database at www.cdtdb.neuroinf.jp.

### 2.2. Sequences

Some of the polyQ SCA genes had transcript variants. The following transcripts were used in the screen and the sequences with partial identity to these transcripts are recorded in Supplemental Table 1.

ATXN1: Transcript variant 1 (GenBank ID: NM_000332) and transcript variant 2 (GenBank ID: NM_001128164).

ATXN2: GenBank ID: NM_002973.3.

ATXN3: GenBank ID: NM_004993.5 (reference isoform).

ATXN7: Transcript variant b was not used because of the limited public record. Transcript variants SCA7a (GenBank ID: NM_000333) and SCA7c (GenBank ID: NM_001128149) were used.

CACNA1A: Transcript variant 2 (GenBank ID: NM_023035) and variant 4 (GenBank ID: NM_001127222) were used. Transcript variants 1, 3 and 5 were not used because the CAG repeats that are expanded in polyQ SCA were not included in the coding sequences.

PPP2R2B: Transcript variant 3 (GenBank ID: NM_181675) and transcript variant 7 (GenBank ID: NM_001127381) were used as these have the CAG repeats that are amplified in SCA12.

TBP: Transcript variant 1 (GenBank ID: NM_003194) and transcript variant 2 (GenBank ID: NM_001172085).

ATXN1, ATXN7, TBP and PPP2R2B transcript variants have alternative 5′UTRS. The results from these variants are recorded in Supplemental Table 1.

### 2.3. Motif identification

The moving window sequence alignment search is used to identify networks of related sequences by iteratively searching a database for sequences that meet a predetermined level of identity when compared to a search sequence. This screen uses local sequence alignment to iteratively compare predefined lengths of the search sequence to other sequences in the database. For each candidate sequence in the database, the search sequence assumes all possible overlapping alignments with the candidate sequence. For each alignment pair, arbitrary length subwindows of base pairs are compared for similarity. For each subwindow scoring a similarity count greater than a chosen threshold, a counter is incremented. Following the scoring of all sub-windows, the percentage of similarity is reported. Pairings that reported a similarity percentage greater than or equal to an arbitrarily chosen threshold are then identified as related to the search sequence. Sequences with highly repetitive elements or low complexity sequence were identified by RepeatMasker (http://www.repeatmasker.org/). Pattern matches to find additional sequences used regular expressions in a Perl script. In some cases, additional matches were found by BLAST (Basic Local Alignment Search Tool; blast.ncbi.nlm.nih.gov) as indicated in the text.

### 2.4. Random network simulation

The combined protein interactions from HPRD and BioGRID were converted into an integer array in R (http://cran.r-project.org/). A random selection of the same number of genes from RefSeq was obtained by using the R sample function and scanned for interacting proteins in the array. Self-interacting proteins were eliminated by deleting matches on the diagonal of the matrix. These results were compared to the results from the same analysis with the genes from the combined screens from the bioinformatic screens with the 5′ and 3′UTRs from the SCA genes which included miscellaneous RNAs, non-coding RNAs and mRNAs that encoded proteins with no known interactions.