SHORT COMMUNICATION

Kavita Bhalla · Helen J. Eyre · Scott A. Whitmore
Grant R. Sutherland · David F. Callen

***C16orf5*, a novel proline-rich gene at 16p13.3, is highly expressed in the brain**

Received: April 12, 1999 / Accepted: June 4, 1999

Abstract A novel gene has been characterized, designated *C16orf5*, with an unusually high content of proline residues (40% over 104 residues) at the N-terminus of the protein. The C-terminus of the protein is also cysteine rich with 14 cysteine residues present. Analysis using Northern and dot blots showed that the highest expression of this gene is in the brain. The gene was located on chromosome 16 at band p13.3 by FISH to metaphase chromosomes. Southern blot analysis with a human–rodent somatic cell hybrid panel showed a location between the somatic hybrid breakpoints 23HA and CY196. This gene comprises at least four exons and an open reading frame of 786bp encoding a predicted protein of 261 amino acids. Analysis of this protein using PSORTII predicted a nuclear localization.

Key words Proline · Cysteine · Brain · Chromosome 16p13.3

Introduction

Balanced chromosome translocations occur in humans with a frequency of approximately 1 in 800 individuals and are usually associated with a normal clinical phenotype. However, de novo translocations are associated with an increased risk of clinical abnormalities (Warburton 1991), which can arise from the disruption of genes at the translocation breakpoints. An extensive somatic cell mouse/human hybrid panel has been constructed (Callen et al. 1995). Included in this panel is the hybrid CY196, which is derived from a de novo translocation ascertained in a patient with epilepsy and mental retardation.

An EST WI-16589 was mapped on the GB4 radiation hybrid panel at 16p13.3 (<http://www.ncbi.nlm.nih.gov/>

genemap98) and also has been located in the adjacent distal interval to the somatic hybrid breakpoint CY196 (Callen et al. 1989). Because this gene was highly expressed in the brain, this was a candidate for disruption by the CY196 breakpoint.

Materials and methods

Sequence of the cDNA was generated by amplification of cDNA clone inserts and sequencing with an ABI 377 sequencer using standard techniques. The 5'-end of the gene was confirmed by 5'-RACE (rapid amplification of cDNA end) on poly A⁺ mRNA from fetal brain tissue. The following nested gene-specific primers were used to generate 5'-RACE (Gibco BRL, Life Technologies, Gaithersburg, USA) products, which were then sequenced: GSP#1 (nt 346–362) 5'-CCT TTC TCT TCC AGA AG, GSP#2 (nt 290–308) 5'-CTC GCT GGA CAT CTT CGC TG, and GSP#3 (nt 277–296) 5'-TTC GCT GCT TCT CCT CAA AT.

To map the gene relative to the CY196 breakpoint, cDNA probes were used to screen two different PAC libraries (Genome Systems, St. Louis, Missouri, and PAC/BAC Resources, Buffalo, New York) each with a threefold coverage of the human genome. Positive clones were purchased and inserts sized and restriction mapped using pulsed-field gel electrophoresis (PFGE) and the infrequently cutting restriction enzymes *NotI* and *EagI*. This method allowed determination of clone overlaps. A further walk was undertaken from one of the furthest extending PACs by screening a BAC library (Genome Systems) with the cloned SP6 end fragment of the insert.

To determine the partial genomic structure of the gene DNA was prepared from a PAC, shown to contain the gene, using a plasmid kit (Qiagen, Hilden, Germany) and digested separately with the blunt end-cutting restriction enzymes *AluI*, *RsaI*, *ScaI*, *HaeIII*, and *PvuII* for ligation to bubble linkers (Gecz et al. 1999). Tentative exonic boundaries were generated by PCR using gene-specific primers

K. Bhalla (✉) · H.J. Eyre · S.A. Whitmore · G.R. Sutherland · D.F. Callen
Centre for Medical Genetics, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, 72 King William Road, Adelaide SA 5006, Australia
Tel. +61-8-8204-6270; Fax +61-8-8204-7342
e-mail: kbhalla@medicine.adelaide.edu.au

and bubble linker-specific primers. The resulting PCR products were purified and then directly sequenced using dye terminator fluorescence chemistry (ABI-Perkin Elmer, Warrington, UK) and the original PCR primers. Sequencing reactions were run on an ABI377 sequencer and analyzed using the Laser Gene DNA software package (DNASTar, Madison, USA). Further analysis utilized the PROSITE database (Hofmann et al. 1999) and PSORTII (Nakai and Horton 1999). Expression of the *Cl6orf5* transcript was evaluated using a Multiple Tissue Northern and RNA Master Blot (Clontech, Palo Alto, USA). The blots were probed with the insert of cDNA clone AA070715 as per manufacturer's protocol.

Chromosomal assignment of the human *Cl6orf5* was performed by fluorescence in situ hybridization (FISH). The cDNAs AA070715 and H07909 were nick-translated with biotin-14-ATP and hybridized in situ at a final concentration of 20 ng/μl to metaphase chromosomes from a normal male. The FISH method was modified from that previously described (Callen et al. 1990). Chromosomes were stained before analysis with both propidium iodide (as counterstain) and DAPI (for chromosome identification). Images of metaphase preparation were captured by a cooled CCD camera using the CytoVision Ultra Image collection and enhancement system (Applied Imaging, New Castle, UK). Twenty metaphases from a normal male were examined for the fluorescence signal.

Results and discussion

A BLAST search of the dbEST database with the sequence of W1-16589 identified two overlapping cDNA clones, AA070715 and H07909 (Fig. 2B). End sequences for both these clones were available, which enabled design of primers to generate the complete cDNA sequence. This sequence, together with a further 193 bp generated by 5'-RACE, is presented in Fig. 1B. The gene *Cl6orf5* (GenBank accession number, AF131218) consists of an open reading frame of 786 bp, encoding a putative protein of 261 amino acids and a large 1517-bp 3'-untranslated region. The putative translation initiation site begins with an ATG at the nucleotide 298, and the sequence surrounding the site conforms to Kozak's rule (Kozak 1991, 1996). Two polyadenylation sites are present, one at 2144 bp and the second at 2756 bp.

The Northern blot showed a single band approximately 2.4 kb in size with the strongest signal in brain but also expression in heart, skeletal muscle, kidney, pancreas, and liver (Fig. 1A). This is likely to correspond to transcripts that are generated using the polyadenylation site starting at 2,144 bp. Although some transcripts in the NCBI database dBEST were generated using the second polyadenylation site starting at 2,756 bp, there was no clear evidence of transcripts of this size from the Northern blot. Also probed was a RNA dot blot that contained poly A⁺ RNAs from 50 different human tissues with the quantity of RNA normalized with respect to several housekeeping genes. This pro-

cedure showed that there was expression in all the brain samples (whole brain, amygdala, caudate nucleus, cerebellum, cerebral cortex, frontal lobe, hippocampus, medulla oblongata, occipital lobe, putamen, substantia nigra, temporal lobe, thalamus, subthalamic nucleus, and spinal cord), while heart and all other tissues showed very weak or absent expression (data not shown). The expression patterns determined from the normalized RNA of the dot blot therefore show a brain-restricted expression, suggesting that the *Cl6orf5* gene has some brain-specific function.

Database searching using the entire *Cl6orf5* cDNA sequence showed no significant homology to known proteins and domains. Searching of the PROSITE database with the *Cl6orf5* protein sequence predicted a single N-glycosylation site, three N-myristoylation sites, and two protein kinase C phosphorylation sites, which suggested that the protein might encode a secreted or plasma membrane protein (Habuchi et al. 1998). The notable features of the peptide include its proline richness (40% over 104 residues) at the N-terminus of the protein. Proline richness is a characteristic of DNA binding proteins (Bennet and Reed 1993). Proline-rich activation domains are associated with transcription factors like CTF/NF-1, AP-2, and steroid receptors for progesterone and estrogen (Mermoud et al. 1989). The predicted *Cl6orf5* protein is also cysteine rich in the C-terminal half of the gene, with 14 cysteine residues present that may participate in intra- and/or interchain disulfide bonding (Marquardt et al. 1998). Computer prediction of the subcellular localization using PSORT II indicated that *Cl6orf5* would most probably localize to the nucleus.

The location of the *Cl6orf5* gene was confirmed by FISH analysis of metaphase chromosomes. Twenty metaphases all showed signal on one or both chromatids of chromosome 16 at 16p13.3 for the cDNA probes H07907 and AA070715. 3 and 5 nonspecific dots were observed in these 20 metaphases respectively. The gene was further localized by Southern analysis of a human/rodent somatic cell hybrid panel (Callen et al. 1989). *Pst*I- and *Hind*III- digested somatic cell hybrid DNA was blotted and hybridized with the inserts of cDNA clones (data not shown). Results of FISH and Southern analysis localized the gene to chromosome 16p13.3 between the somatic cell hybrid breakpoints 23HA and CY196 (Fig. 2A, B).

Fig. 1. A Detection of *Cl6orf5* transcript by Northern blot analysis. The filter with 2 μg of poly A⁺ RNA from different tissues in each lane was probed with 1.5-kb insert of cDNA AA070715. The position of the site markers is indicated. **B** Sequence of *Cl6orf5*. Nucleotide sequence of the *Cl6orf5* cDNA and the deduced amino acid sequence of its predicted protein are shown. Asterisks denote the start and termination codons. Bold capital letters in the sequence flanking the start codon, (GCGAAGATGT), conform to the optimum Kozak consensus. The first 193 nucleotides were obtained by 5'-RACE. The proposed polyadenylation signals (Sheets et al. 1990; Wahle and Keller 1992), and an in-frame stop codon before the initiation codon are underlined; proline residues are boxed and cysteine residues are shown in bold-face. The predicted 1N-glycosylation (ASN), 3N-myristoylation (MYRISTYL), and two protein kinase C (PKC) phosphorylation sites are indicated. Each exon boundary is shown by >> above the nucleotide sequence. The nucleotide sequence reported here will appear in GenBank under the accession number AF131218

A **B**

aggaagcagaccctgcgacgggaccgaggaatggaggggtggcgggtgtgtggatggccagagtgcacagcttagggagaagccggatgtggaaggggaggaactgtgtgg 110
 tatccccagggccctgggcccagccctggcttggcagcccttgcctccctgctgagggcccctgcctttccctccctgaacacag > exon 1 220
 CATCCAGGCCAGCAGCTGAAGGAGCCTACCTGCCTCCCTTCTC IGAGTAGCAGGATTGTAGGAGAAGCA GCGAAGATGTCCAGCGAGCCTCCCTCCCTTATCCTGGG 330
 M S S E P P P P Y P G
 > exon 2 440
 GGCCCCACAGCCCACTTCTGGAAGAGAAAAGTGGAGCCCCGCCACCCAGGCCGTTCTCCCAAGCTGTGATGCAGCCCCCTCCAGGCATGCCACTGCCCTTCCGGA
 G P T A P L L E E K S G A P P T P G R S S P A V M Q P P P G M P L P P A D
 CATTGGCCCCCACCCTATGAGCGCGGGTCACCCAATGCCAGCCTGGCTTCATCCACCACATGAGTGCAGATGGCACCTACATGCCTCCGGGTTTCTACCTCT 550
 I G P P P Y E P P G H P M P Q P G F I P P H M S A D G T Y M P P G F Y P P
 CTCCAGGCCCCACCCACCATGGGCTACTACCCCCAGGGCCCTACACGCCAGGGCCCTACCTGGCCCTGGGGGCCACACAGCCACAGTCTGGTCCCTTCCAGGAGCT 660
 P G P H P P M G Y Y P P G P Y T P G P Y P G P G G H T A T V L V P S G A
 MYRISTYL MYRISTYL
 GCCACCACGGTGACAGTGTGCAGGGAGAGATCTTTGAGGGAGCGCTGTGCAGACGGTGTGTCCCACTGCCAGCAGGCCATCACCAAGATCTCTACGAGATTGG 770
 A T T V T V L Q G E I F E G A P V Q T V C P H C Q Q A I T T K I S Y E I G ASN
 CTGTGATGAATTTCTGCTGGGTTTCTTCTGTTGCTTCATGGGATGTGATCTGGGCTGCTGCCTGATCCCTGCCTCATCAATGACTTCAAGGATGTGACGCACACATGCC 880
 L M N F V L G F F C C F M G C D L G C C L I P C L I N D F K D V T H T C P
 MYRISTYL
 CCAGCTGCAAAAGCTACATCTACAGTACAAGCGCTGTGCTAACGGAGCTGGGACTCCCGCCGCTGTCACTGAGCCCTGTGCTTTGCTCCCTGCCTCAG 990
 S C K A T S T R T S A C A N G A G T R D S P A C Q S G P L C F A P C A Q
 PKC PKC
 TGGTCACTTTCCGCTCCCACTTGGGGCTGGGAGCCGTGCCACCATCCCTAGAAGTCCTGTCTCTTCACTCCCTGCCCTACCTGAGCCGC TGACTCTTCTGGCAAAAAT 1110
 W S L S R S H L G L G A V P P S P R S P V L F T L P Y L S R * > exon 4
 CTGTTGGGATTTAAGGCCAAGGGTCACTGGGTGGCAGGGGGCTGGCAATGAGCTTGTGTGTTGTTGGTCTGCTTGGTGTGTGATCGGGAAGATAAGCTGGGAGGGGTC 1210
 TCCTGCTGGGGTCTGATGCCTCTGTTTCCAAACAAGGTACAGGTTTCACTCAGCTCTTTCCCTCGGACCAACAGCAGCCAGAGCAGTTAGCCAGTTAGTCCCCAGG 1320
 CTTGTGGCACAGGCCTTCTGACCTGCTGGGCCGAGAAATGGGTAAAGTTGCTGGAGTCAGGTTGGGCCACAGTAGGACAGGGTCACAAAGCTGGGTTTGTCTTCTGGGTAC 1430
 TTTGCGCCTCTGGGGTCTAGAGGTGGGGCATGGTGGCTGGAAGTAAACTGCCAATCTGGCCCTCAGAACTCTCAGGTATAGAAGCCAGGATGTCTAATACCTGTCT 1540
 CCAGTGGCCGAGAGCTGCCCTGGTGTCAAGTGAAGAGGACACGTACCTGGGTGAATGATCAGACCTGGTAGCTAAGAAGGAACCTGTCCCTTTGAGTCAAGTGTGACAG 1660
 CCCCCTTTCAGGCCATGCGTCTGTGAACCTGTATTGCTGGGGCCGGAAGGAGCCCGTGAAGCTAGCCCTTCCCGCTGCCCTGTGCTCTCACTGCGTGTGGGTATGACC 1760
 CTGCGCTGGTGGCTGGTGTATCCCAACTGGGGAAGAGATGGCAGAGGGTCCCCCTTGTGGGTGCGCTGGATGTGCAGAGCCTTCTCCATGGATTTTCTTCCCTGTAAGT 1870
 GCCGGGCCCTCACCCAGCTGACAGGCTGTTGCTGTGCTGCTACACCTGCTCCTGCAAGGCACACTGGGCTAGGGACGAGGAAGGAGCAGCCACAAGTGGTAGAACTG 1980
 CCTTGGTGGACACAGCCTCGCCCTGCTCTTATTTCTGAAATGGTTTGTGAACCTTGCTCACTTGGACCATGTATCTTCTGCTGCTGCTGCTGCTGCTGCTGCTGCT 2090
 GCATGGCCTCCTGTCACTGTGAATCGTGGCCAGTCTCAGTTTGTAGTTTCTC ATTAAATTGGCCCTTTTCACTCCCTTGGCCCTTTTCACTCCCCGCGCTGGGCTCT 2220
 GCTCTCTTGGCTGGCTTCTTCTTTTGGAGGAAAGAGGGTGGGGCTGCAGGCAGTCTACTGGCAGGACGGGAGGCTGAGTCTCAGGGTCTCACACCTCAGTGTGA 2310
 TGCCATGCCAATCTGCTGGGACAAACCAACACCTAAGGACCTAATTAAACCAACACAGAGTGGGTGTAGACCAAGCCTGGGATTTCAGCTGTGACTGGGCCAGGGCA 2420
 CAGCTTGGTCTCGGCAGTGGCTGTAAGGTCACCTTCTTCTGATGCTGGTTCAACCATCTATATGGCATCCAGCATGGGATCTGCAAGCTGGAGCCCTCTAC 2530
 CCGCAGGCTTGAACACAGCATCCAGCCCTGGGGAGGCGCACCTTAAGCAAGACAGAGTGGGCTGAGGCTGACAGGCAGAAAGACTAACAGAGCGCAGCTGCACACG 2640
 CAGGTTCTGGGCGACCTCTGGCCTGGCCATCTCTGCACTAATCATGTAATTATGAAGGTGGCAGTCTTGGTCAAGTAGTTTAAAGAGTTTCCCTACTTTTAAACCTT 2750
 TTTGAATAAAACTTTTACAGT 2773

Localization of *Cl6orf5* relative to the CY196 breakpoint was achieved by constructing a PAC and BAC contig containing the entire gene (Fig. 2B). Genomic sequencing enabled the determination of four exon–intron boundaries (Fig. 1B; sequence available on request). Hence, the gene *Cl6orf5* consists of at least four exons. The genomic size of *Cl6orf5* was estimated to be 10kb by hybridizing Southern blots of *Pst*I-digested genomic DNA with cDNA clones AA070715 and H07909.

The contig containing the *Cl6orf5* transcript was orientated by hybridizing to a YAC spanning this region (data not shown). This result demonstrated that the *Cl6orf5* gene has its 3'-end more distally located and is entirely contained within a region of this contig, which maps distal to the somatic cell hybrid breakpoint CY196. Therefore, the translocation breakpoint represented in CY196 does not disrupt this gene. Although the function of the *Cl6orf5* protein is unknown, its proline richness, presence of cysteine residues, predicted nuclear localization, and abundant brain-specific expression suggest that it is likely to have an important role in the development and function of neuronal cells.

Acknowledgments K.B. was supported by a fellowship from AusAID. This work was supported by a Program grant from the Australian National Health and Medical Research Council. We thank Dr. Jozef Gecz for assistance with the bubble PCR. We also greatly appreciate help from Ingrid L. Lensink for 5'-RACE work.

References

- Bennett M, Reed R (1993) Correspondence between a mammalian spliceosome component and an essential yeast splicing factor. *Science* 262:105–108
- Callen DF, Baker E, Eyre HJ, Chernos JE, Bell JA, Sutherland GR (1990) Reassessment of two apparent deletions of chromosome 16p to an ins(11;16) and a t(1;16) by chromosome painting. *Ann Genet* 33:219–221
- Callen DF, Hyland VJ, Baker EG, Fratini A, Gedeon AK, Mulley JC, Fernandez KEW, Breuning MH, Sutherland GR (1989) Mapping the short arm of human chromosome 16. *Genomics* 4:348–354
- Callen DF, Lane SA, Kozman H, Kremmidiotis G, Whitmore SA, Lowenstein M, Doggett NA, Kenmochi N, Page DC, Maglott DR, Nierman WC, Murakawa K, Berry R, Sikela JM, Houlgatte R, Auffray C, Sutherland GR (1995) Integration of transcript and genetic maps of chromosome 16 at near-1-Mb resolution: demonstration of a “hot-spot” for recombination at 16p12. *Genomics* 29:503–511
- Gecz J, Baker E, Donnelly A, Ming JE, McDonald-McGinn DM, Spinner NB, Zackai EH, Sutherland GR, Mulley JC (1999) Fibroblast growth factor homologous factor 2 (FHF2): gene structure, expression and mapping to the Börjeson-Forssman-Lehmann syndrome region in Xq26 delineated by a duplication breakpoint in a BFLS-like patient. *Hum Genet* 104:56–63
- Habuchi T, Luscombe M, Elder PA, Knowles MA (1998) Structure and methylation-based silencing of a gene (DBCCR1) within a candidate bladder cancer tumor suppressor region at 9q32–q33. *Genomics* 48:277–288
- Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res* 27 (1):215–219
- Kozak M (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 266(30):19867–19870
- Kozak M (1996) Interpreting cDNA sequences: some insights from studies on translation. *Mamm Genome* 7:563–574
- Marquardt A, Stöhr H, Passmore LA, Krämer F, Rivera A, Weber BHF (1998) Mutations in a novel gene, VMD2, encoding a protein of unknown properties cause juvenile-onset vitelliform macular dystrophy (Best's disease). *Hum Mol Genet* 7(9):1517–1525
- Mermod N, O'Neill EA, Kelly TJ, Tjian R (1989) The proline-rich transcriptional activator of CTF/NF-1 is distinct from the replication and DNA binding domain. *Cell* 58:741–753
- Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* 24 (1):34–36
- Sheets MD, Ogg SC, Wickens MP (1990) Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* 18(19):5799–5805
- Wahle E, Keller W (1992) The biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors. *Annu Rev Biochem* 61:419–440
- Warburton D (1991) De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am J Hum Genet* 49 (5):995–1013