

## Differential Expression of XAP5, a Candidate Disease Gene

Richard Mazzarella,<sup>\*,1</sup> Gina Pengue,<sup>\*</sup> Jaeyoung Yoon,<sup>\*,†</sup> Jonathan Jones,<sup>\*</sup> and David Schlessinger<sup>\*</sup><sup>\*</sup>Department of Microbiology and Center for Genetics in Medicine, Washington University School of Medicine, St. Louis, Missouri 63119; and <sup>†</sup>Department of Pharmacological and Physiological Sciences, St. Louis University School of Medicine, St. Louis, Missouri 63104

Received February 13, 1997; accepted July 17, 1997

**We have isolated a full-length cDNA corresponding to the XAP5 gene in Xq28. An unusual feature of the cDNA is that it contains runs of CCG repeats in the 5' untranslated region, typical of genes that exhibit anticipation. It has a striking pattern of differential expression and is greatly enhanced in various fetal tissues. This predicted protein encodes a unique 339-amino-acid polypeptide that contains a large percentage of highly charged residues and a possible nuclear localization signal. A comparison to genomic sequence shows that XAP-5 comprises 13 exons spanning 6.5 kb. An examination of the human population indicates that the longest CCG run is polymorphic and varies in length from 8 to 12 repeats.** © 1997 Academic Press

An increasing number of human disorders are based on expansion of the triplet repeats CAG and CCG and their complements (7, 11, 13). The diseases show the phenomenon of anticipation, in which age of onset becomes earlier and the severity of the disease increases in successive generations. This pattern can be explained by the gradual increase in size of the triplet upon passage of the affected chromosome to the offspring.

If the mechanism of repeat expansion leading to disease is a general one, depending primarily on initial repeat size, then the presence of triplet repeats in a gene is interesting as a clue to possible involvement in inherited disorders. There are more than 40 described genes that contain triplet-associated repeats (13). Here we report the recovery and analysis of the full cDNA sequence of one of the genes in a gene-rich portion of Xq28, which shows both a series of CCG repeats in its 5' untranslated region and an unusual pattern of tissue- and developmentally regulated expression.

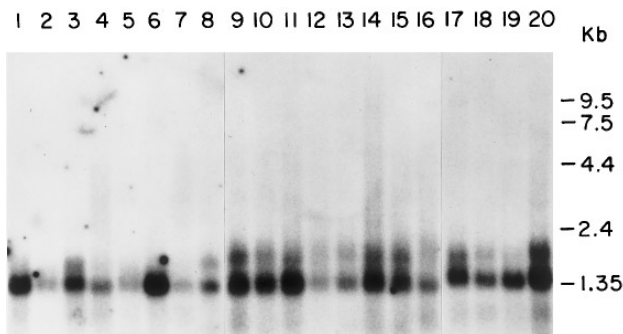
In a 1.5-Mb area of Xq28, high GC content indicated

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. AD001530.

<sup>1</sup>To whom correspondence should be addressed at Washington University School of Medicine, Department of Microbiology, Box 8230, 660 S. Euclid Avenue, St. Louis, MO 63110. Telephone: (314) 362-2191. Fax: (314) 362-3203. E-mail: rich@genetics.wustl.edu.

that the region is densely populated with genes (8, 9), and genetic linkage mapping information also placed a high concentration of disease genes in the region (6). Recently, we sequenced a 220-kb high-GC region of the X chromosome in Xq28 containing 13 known and 6 candidate genes between the RCP/GCP (color vision) locus and glucose 6-phosphate locus (2). Part of one of the genes, XAP5 (also called 9F; Genbank Accession Nos. X74611 and X87199), had previously been identified as an expressed sequence tag (EST) (1). The EST, near the center of the gene, includes the sequence of one exon and partial sequences of two neighboring exons. Another homologous segment of the gene was found cloned in the reverse orientation at the 5' end of a cDNA for the HEX2 gene. [HEX2 is localized to chromosome 2 (Z46376).] When the portion of the chimeric cDNA that was homologous to the genomic sequence was used as a guide, 80% of the cDNA/mRNA sequence of the gene was inferred. Further screening has now recovered a full-length cDNA for XAP5, revealing the 5' untranslated region of CCG repeats.

Subsequent analysis of the genomic sequence revealed a 1-kb Merck project EST (yb99a04) from a



**FIG. 1.** Northern analysis of the XAP5 gene. The tissues analyzed are lane 1, heart; lane 2, brain; lane 3, placenta; lane 4, lung; lane 5, liver; lane 6, skeletal muscle; lane 7, kidney; lane 8, pancreas; lane 9, spleen; lane 10, thymus; lane 11, prostate; lane 12, testis; lane 13, ovary; lane 14, small intestine; lane 15, colon; lane 16, leukocytes; lane 17, fetal brain; lane 18, fetal lung; lane 19, fetal liver; and lane 20, fetal kidney. No significant differences in the mRNA levels between the various tissues were detected with an actin probe.

```

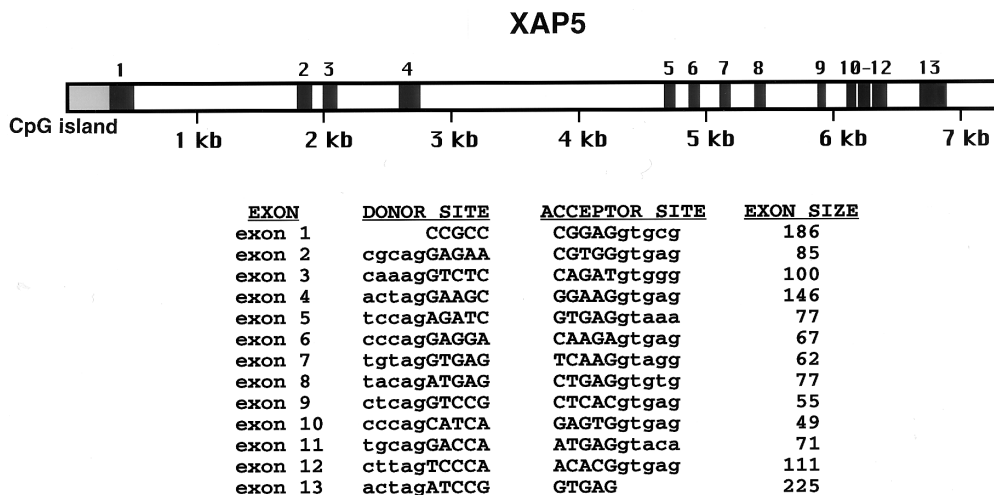
CCGCCGCTGCCGCTGCCGCTGTCGCTGTCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCTGCC 75
      10                                20
ATGGCTCAATACAAGGGCGCCGCGAGCGAGGCGGCCGCCGCATGCACCTGATGAAGAAGCGGGAGAAGCAGCGC 150
M A Q Y K G A A S E A G R A M H L M K K R E K Q R
      30                                40                                50
GAGCAGATGGAGCAGATGAAGCAGCGCATCGCGGAGGAGAACATCATGAAATCCAACATTGACAAGAAGTTCTCT 225
E Q M E Q M K Q R I A E E N I M K S N I D K K F S
      60                                70
GCGCACTACGACGCGGTGGAGGCAGAGCTCAAGTCCAGCACCGTGGGTCTCGTGACCCTGAATGACATGAAGGCC 300
A H Y D A V E A E L K S S T V G L V T L N D M K A
      80                                90                                100
AAGCAGGAGGCTCTGGTGAAGGAGCGGGAGAAGCAGCTGGCCAAGAAGGAGCAGTCCAAGGAGCTGCAGATGAAG 375
K Q E A L V K E R E K Q L A K K E Q S K E L Q M K
      110                                120
CTGGAGAAGCTTCGAGAGAAGGAGCGTAAAGAAGGAAGCCAAGCGGAAGATCTCCAGCCTGTCTTCACCCCTGGAG 450
L E K L R E K E R K K E A K R K I S S L S F T L E
      130                                140                                150
GAGGAAGAAGAGGGGAGGCGAGGAGGAAGAGGAGGCGGCCATGTATGAGGAGGAGATGGAAAGGGAAGAGATCACC 525
E E E E G G E E E E E A A M Y E E E M E R E E I T
      160                                170
ACGAAGAAGAGAAAAGTGGGGAAAGAACCCAGACGTTGACACAAGCTTCTTGCCTGATCGAGACCGTGAGGAGGAG 600
T K K R K L G K N P D V D T S F L P D R D R E E E
      180                                190                                200
GAGAATCGGCTTCGGGAAGAGCTGCGGCAGGAGTGGGAAGCCAAGCAGGAGAAGATCAAGAGTGAGGAGATCGAG 675
E N R L R E E L R Q E W E A K Q E K I K S E E I E
      210                                220
ATCACCTTACGCTACTGGGATGGCTCTGGGCACCGGCGGACAGTCAAGATGAGAAAGGGCAACACCATGCAGCAG 750
I T F S Y W D G S G H R R T V K M R K G N T M Q Q
      230                                240                                250
TTCCTGCAGAAGGCGCTCGAGATCCTTCGGAAAGACTTCAGTGAGCTGAGGTCCGCAGGGGTGGAGCAGCTCATG 825
F L Q K A L E I L R K D F S E L R S A G V E Q L M
      260                                270
TACATCAAGGAGGACTTGATCATCCCTCACCATCACAGCTTCTACGACTTCATCGTCACCAAGGCACGGGGGAAG 900
Y I K E D L I I P H H H S F Y D F I V T K A R G K
      280                                290                                300
AGTGGACCACTCTTCAACTTTGATGTTTCATGACGATGTGCGGTTGCTCAGTGACGCCACTGTGGAGAAGGATGAG 975
S G P L F N F D V H D D V R L L S D A T V E K D E
      310                                320
TCCCATGCAGGCAAGGTGGTGTCTGAGGAGCTGGTACGAGAAGAACAAGCACATCTTCCCGCCAGCCGCTGGGAA 1050
S H A G K V V L R S W Y E K N K H I F P A S R W E
      330
CCCTACGACCCTGAAAAGAAGTGGGACAAGTACAGATCCGCTGAGCATCCAGGAGGCTGCGCGGCCCCCGCTCC 1125
P Y D P E K K W D K Y T I R *
      TCAGCTCCCTCAGTGTGCCCGTGGTGTACCCGGGACTCCAGGCACCCGCTCCCTGCGACCATGCCAGGCACGC 1200
      TGGAGGAGGACGGCAGCTGCTGCTGCTGCCCTGCCACATCAGTGAAGTCTTATTCTTTCCAATAAAGAA 1275
      GTGCACGTGTCAGAGCTGGAGCGCCTGCATTGTGAG 1311

```

**FIG. 2.** Nucleotide and deduced protein sequence of the XAP5 cDNA clone. The CCG repeats in the 5' untranslated region are underlined and an asterisk indicates the termination codon.

lung cDNA library. It overlapped the 169-bp XAP5 EST and extended an additional 350 bp in the 5' direction and 475 bp in the 3' direction. The 450-bp 5' *EcoRI*-*XhoI* fragment from the Merck EST was used as a probe in Northern blot analyses of adult and fetal tissues (Fig. 1). A major band of 1.4 kb was detected in all tissues examined, with an additional 1.8-kb band

detectable in some tissues. The expression pattern showed abundant fetal expression in brain (lane 17), liver (lane 19), and especially kidney (lane 20), with little signal from the equivalent adult organs. Instead, the gene shows a differential pattern of tissue-specific expression in the adult, with a high level observed in heart, skeletal muscle, spleen, thymus,



**FIG. 3.** The genomic structure of the XAP-5 gene. A schematic representation of the genomic intron-exon structure of XAP5 is shown at the top. Exons are depicted as shaded boxes. A table displaying the donor and acceptor splice junctions and the size of each exon is at the bottom.

prostate, and small intestine (lanes 1, 6, 9–11, and 14, respectively).

This EST fragment probe was used to isolate a 1311-bp cDNA from a teratocarcinoma library (10). The cDNA encodes a putative protein of 339 amino acids with a molecular weight of 40,241 (Accession No. GSDB:S:1236293) (Fig. 2). A striking feature of the sequence is a group of 17 CCG repeats that occur in 75 bp of the 5' untranslated region; the longest run of 9 CCGs is separated from an additional 4 CCGs by a single C nucleotide. The initiator methionine codon is an excellent start site according to Kozak's contextual rules (5). A consensus polyadenylation signal is also observed about 40 bases upstream of the 3' end, but it is unclear whether this signal is used, since neither this clone, nor three additional isolates from the teratocarcinoma library, nor the original clone from the lung library contained a poly(A) tract. The predicted protein sequence shows a distribution of highly charged residues that precludes likely signal peptide or transmembrane regions. Rather, further analysis reveals features suggestive of nuclear localization, since the peptide contains more than 23% basic residues and an SV40 large T antigen nuclear localization signal

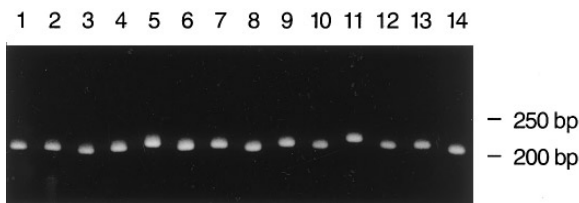
(KKRK; Ref. 4) at amino acid positions 152–155. The skewed amino acid distribution of the protein also includes 21% acidic residues and no cysteine.

A BLAST (3) search of the SwissPro database finds no significant protein sequences when low-complexity regions containing runs of glutamic acid residues are discounted. Thus, it appears that the XAP5 gene encodes a novel soluble protein that may be located in the nucleus.

A comparison of the cDNA with the genomic sequence shows that XAP5 comprises 13 exons spanning about 6.5 kb (Fig. 3). All of the exon-intron boundaries have consensus AG-GT splice junctions. The gene starts at a CpG island, coincident with the first exon and extending about another 300 bases upstream. The cDNA sequence exactly matches the genomic sequence of the region that we determined previously (2).

Another report of sequence from the XAP5 gene has appeared during the preparation of this article (12). That report, however, included no Northern data to compare the predicted and observed size of the mRNA, and the sequence lacks both the N-terminal 14 amino acids and the 5' UTR containing the CCG repeats. We have found in a number of experiments that the method used in that study to locate a possible 5' end, RT-PCR, often fails to cross the GC-rich 5' zone, which probably explains both the incompleteness of the reported sequence and the rarity with which a full-length cDNA is recovered. That Fig. 2 shows the full sequence of XAP5 is more likely, since there are no ATG codons or appropriate splice junctions for several hundred base-pairs upstream of the 5' end, and the size of the cDNA coincides with that observed in Northern analyses.

To examine possible triplet polymorphisms in the human population in the 5' untranslated region of the XAP5 gene, a PCR assay across the CCG repeat region was developed. Reactions were performed in a final



**FIG. 4.** PCR amplification across the CCG repeats in the XAP5 5' untranslated region. PCR products from 14 unrelated males containing 8 (lane 3), 9 (lanes 1, 2, 4, 6, 8, and 14), 10 (lanes 5, 7, 9, 10, 12, and 13) and 12 (lane 11) CCG repeats were fractionated on a 4% agarose gel and stained with ethidium bromide.

reaction volume of 25  $\mu$ l using thin-walled GeneAmp reaction tubes (Perkin-Elmer Cetus) containing 1 $\times$  cloned *Pfu* DNA polymerase reaction buffer [20 mM Tris-HCl, pH 8.8, 10 mM KCl, 10 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 2 mM MgSO<sub>4</sub>, 0.1% Triton X-100, 100 mg/ml BSA], 0.125  $\mu$ M primer pair (5'-GCGGGGCATCCGTGCGTCTCC-TGGTGGCTG-3' and 5'-TGCTTCTCCCGCTTCTTCA-TCAGGTGCATG-3'), 250  $\mu$ M dNTPs, 250 ng of genomic DNA, and 1.25 units of Exo(-) *Pfu* DNA polymerase. Each reaction mixture was incubated in a DNA thermal cycler 480 (Perkin-Elmer Cetus) with an initial denaturation cycle of 5 min at 98°C, followed by 35 cycles of 1 min at 98°C, 1 min at 65°C, and 2 min at 75°C. A final extension cycle was performed for 10 min at 75°C.

An examination of 48 unrelated males showed that this locus is polymorphic in the human population. The most frequent allele size contained 9 triplet repeats in the longest CCG run (83.3%), while less frequent allele sizes of 8 repeats (2.1%), 10 repeats (12.5%), and 12 repeats (2.1%) were also observed. The polymorphic nature of this region is illustrated in Fig. 4.

Whether the variable number of CCG repeats is correlated with any disease state remains to be determined. If so, the disorder might be severe, since the expression pattern of the gene indicates tight control of the distribution of the putative protein, and the possible cellular location based on sorting predictions suggests that XAP5 may be a DNA-binding protein or transcriptional factor.

## REFERENCES

- Bione, S., Tamanini, F., Maestrini, E., Triboli, C., Poustka, A., Torri, G., Rivella, S., and Toniolo, D. (1993). Transcriptional organization of a 450-kb region of the human X chromosome in Xq28. *Proc. Natl. Acad. Sci. USA* **90**: 10977-10981.
- Chen, E. Y., Zollo, M., Mazzarella, R., Ciccodicola, A., Chen, C.-N., Zuo, L., Heiner, C., Burough, F., Repetto, M., Schlessinger, D., and D'Urso, M. (1996). Long-range sequence analysis in Xq28: Thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. *Hum. Mol. Genet.* **5**: 659-668.
- Gish, W., and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Genet.* **3**: 266-272.
- Kalderon, D., Roberts, B. L., Richardson, W. D., and Smith, A. E. (1984). A short amino acid sequence able to specify nuclear location. *Cell* **39**: 499-509.
- Kozak, M. (1996). Interpreting cDNA sequences: Some insights from studies of translation. *Mamm. Genome* **7**: 563-574.
- Mandel, J. L., Monaco, A. P., Nelson, D., Schlessinger, D., and Willard, H. (1992). Genome analysis and the human X chromosome. *Science* **258**: 103-109.
- Panzer, S., Kuhl, D. P. A., and Caskey, C. T. (1995). Unstable triplet repeat sequences: A source of cancer mutations. *Stem Cells* **13**: 146-157.
- Pilia, G., Little, R. D., Aissani, B., Bernardi, G., and Schlessinger, D. (1993). Isochores and CpG islands in YAC contigs in human Xq26.1-qter. *Genomics* **17**: 456-562.
- Saccone, S., DeSario, A., Wiegant, J. R., Raap, A., Della Valle, G., and Bernardi, G. (1993). Correlation between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA* **90**: 11929-11933.
- Skowronski, J., Fanning, T. G., and Singer, M. F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8**: 1385-1397.
- Sutherland, G. R., and Richards, R. I. (1995). Simple tandem DNA repeats and human genetic disease. *Proc. Natl. Acad. Sci. USA* **92**: 3636-3641.
- Toyoda, A., Sakai, T., Sugiyama, Y., Kusada, J., Hashimoto, K., and Maeda, H. (1996). Isolation and analysis on a novel gene, HXC-26, adjacent to the GDP dissociation inhibitor gene located at human chromosome Xq28 region. *DNA Res.* **3**: 337-340.
- Wells, R. D. (1996). Molecular basis of genetic instability of triplet repeats. *J. Biol. Chem.* **271**: 2875-2878.