Am. J. Hum. Genet. 49:555–565, 1991

Isolation and Characterization of the CRIPTO Autosomal Gene and Its X-linked Related Sequence

Rosanna Dono,* Nunzia Montuori,* Mariano Rocchi,† Liliana De Ponti-Zilli,* Alfredo Ciccodicola,* and M. Graziella Persico*,‡

*International Institute of Genetics and Biophysics, CNR, Naples; †Molecular Biology Laboratory, G. Gaslini Institute, Genoa; and ‡Laboratory of Molecular Biology, NIDDK, National Institutes of Health, Bethesda

Summary

We have previously reported on the identification of a cDNA clone encoding a novel human growth factor, named "CRIPTO," that is abundantly expressed in undifferentiated human NTERA-2 clone D1 (NT2/D1) and mouse (F9) teratocarcinoma cells. We now report the organization and nucleotide sequence of two related genomic sequences. One (CR-1) corresponds to the structural gene encoding the human CRIPTO protein expressed in the undifferentiated human teratocarcinoma cells, and the other (CR-3) corresponds to a complete copy of the mRNA containing seven base substitutions in the coding region representing both silent and replacement substitutions. The 440 bp 5′ to the CAP site of CR-1 are preserved in CR-3. CR-1 maps to chromosome 3, and CR-3 maps to Xq21-q22. Southern blot analysis reveals that multiple CRIPTO-related DNA sequences are present in the human as well as in the mouse genome.

Introduction

Polypeptide growth factors play a role in stimulating cell proliferation, and their genes are expressed in the developing embryo, in normal adult tissues, and in tumor cells (for review, see Devel 1987; Sporn and Roberts 1987; Whitman and Melton 1989). Characterization of these factors and sequencing of their genes have permitted their grouping into a relatively small number of families on the basis of sequence similarities (Mercola and Stiles 1988). One of these is the epidermal growth factor (EGF) family. EGF (Savage et al. 1972), transforming growth factor α (TGFα) (Derynck et al. 1984), and amphiregulin (AR) (Plowman et al. 1990) share structural similarities including the conservation of six cysteins of the "EGF motif," which in EGF are involved in three disulfide bonds defining the tertiary structure. The presence of the EGF motif also in developmental genes, such as *Notch* in *Drosophila* (Kidd et al. 1986) and lin-12 in *C. elegans* (Greenwald 1985), may imply a novel role for the growth factors of the "EGF family." It has been suggested that they may exert their action on the cell surface during development, to mediate cell-cell interactions by recognizing a complementary receptor on another cell.

Elsewhere we have described the isolation of a human cDNA, referred to as "CRIPTO" (Ciccodicola et al. 1989), encoding a protein of 188 amino acids. The central portion of this protein shares structural similarities with the human TGFα (Derynck et al. 1984), human AR (Plowman et al. 1990), and human EGF (Savage et al. 1972). Northern blot analysis of a wide variety of tumor and normal cell lines and tissues (e.g., choriocarcinoma, fibroblast, neuroblastoma, HeLa, placenta, and testis) has shown that CRIPTO transcripts are detected only in undifferentiated human NTERA-2 clone D1 (NT2/D1) and mouse (F9) teratocarcinoma cells and that these disappear after cell differentiation induced by retinoic acid treatment (Ciccodicola et al. 1989). In the present paper we describe the isolation and characterization of two human genomic CRIPTO-encoding sequences, one mapping to chromosome 3 and the other, possibly a functional pseudogene, mapping to the Xq21-22 region.

Received January 17, 1991; final revision received May 3, 1991.

Address for correspondence and reprints: M. Graziella Persico, International Institute of Genetics and Biophysics, CNR, Via Marconi 10, 80125 Naples, Italy.

0002-9297/91/4903-0008$02.00

Material and Methods

Southern Blot Analysis and Chromosomal Mapping Panels

All the hybrid cell lines were hamster × human, obtained by following a published protocol (Davidson 1976). The hybrid clones were characterized for their human chromosome content (Rocchi et al. 1986).

DNA preparation from human peripheral blood lymphocytes and cell lines, restriction-enzyme digestion, electrophoresis, and Southern blotting were performed using standard techniques (Maniatis et al. 1982). In general, 10 μg DNA was digested with 40 units enzyme. Electrophoresis of DNA digests was carried out in agarose gel (0.8%) in TEB buffer (89 mM Tris, 2 mM EDTA, 89 mM boric acid). DNAs were transferred by Southern capillary blot onto nylon membranes (ZETABIND; AMF Cuno, Meriden, CT), were fixed by UV cross-linking, and were hybridized to 10^7 dpm DNA probes labeled by nick translation (Rigby et al. 1977) to a specific activity of about 2×10^8 dpm/μg. Washing was carried out at 65°C in 2 × SSC, 0.2% SDS and subsequently in 0.2 × SSC, 0.2% SDS at 65°C.

Isolation of CRIPTO Genomic Clones

Genomic clones were isolated from two different human genomic libraries, one obtained by partial *Mbo*I digestion of genomic DNA cloned in the *Bam*HI site of the pcos2EMBL Cosmid Vector (Poustka et al. 1984) and the other obtained by partial *Mbo*I digestion of genomic DNA that had been flush ended and cloned into the flush-ended *Xho*I site of Lambda Fix Vector (Stratagene). Cosmids (5×10^5) and 10^6 phages were screened using the CRIPTO cDNA fragment 2B3 (see fig. 1*C*) by standard techniques (of Grunstein and Hogness [1975] and Benton and Davis [1977], respectively). The positive clones were analyzed by restriction mapping, and the genomic fragments hybridizing to the human cDNA were subcloned in pUC18 Vector (Yanisch-Perron et al. 1985) or in pGEM-1 Vector (Promega). DNA sequencing of the genomic subclones was carried out using the modified dideoxynucleotide chain-termination procedure (Hattori et al. 1985). An oligonucleotide walking strategy was performed using synthetic 17-mer oligonucleotides (Applied Biosystems) deduced from the genomic sequence previously determined.

S1 Nuclease Mapping

Total RNA from undifferentiated teratocarcinoma cells NT2/D1 (Andrews et al. 1984) was isolated by cell lysis in 4 M guanidine thiocyanate and sedimentation through 5.7 M CsCl (Chirgwin et al. 1979). Poly(A)+ RNA was selected by chromatography on oligo(dT) cellulose (Aviv and Leder 1972). Either 5 μg poly(A)+ RNA or 40 μg total RNA was hybridized with the 320-bp *Sau*96 fragment of CR-1-73-H (fig. 3), ^{32}P 5′-end labeled, in 20 μl 40 mM Pipes pH 7, 0.4 M NaCl, 1 mM EDTA pH 7, 80% formamide for 16 h at 50°C. Following hybridization, the reaction was diluted 10-fold with S1 nuclease buffer (0.28 M NaCl, 0.05 M sodium acetate pH 4.5, 4.5 mM $ZnSO_4$, 20 μg single-strand DNA/ml). S1 nuclease (1,200 units) was added, and the reaction mixture was incubated for 2 h at 37°C. The reaction was terminated by the addition of 44 μl termination buffer (2.5 M ammonium acetate, 50 mM EDTA); the DNA:RNA hybrids were extracted with phenol, precipitated with ethanol, resuspended in sequencing dye, heated to 90°C, and resolved on a 6% acrylamide, 7 M urea sequencing gel.

Primer Extension

For primer extension analysis, a 30-bp synthetic oligonucleotide, ol GP2 (3′-CCCGGTAGAAGGACGTCAGGTATCGAAATTGTTAA-5′) corresponding to basepairs −9 to +21 of the first exon was end labeled using T4 polynucleotide kinase to a specific activity of 10^8 cpm/μg. End-labeled primer (10^6 cpm) was annealed with 10 μg poly(A)+ mRNA from NT2/D1 cells (Andrews et al. 1984) in a 40-μl vol containing 10 mM Pipes pH 6.4, 0.4 M NaCl, 1 mM EDTA, by heating the reaction mixture for 3 min at 90°C and for 2 min at 75°C and by gradual cooling to 42°C. After 14 h at 42°C the resulting DNA:RNA hybrids were ethanol precipitated and dissolved in reverse transcription buffer (50 mM Tris HCl pH 8, 0.1 M KCl, 10 mM $MgCl_2$) in the presence of 500 μM deoxynucleotide and 20 units reverse transcriptase. After 1 h at 42°C the DNA:RNA hybrids were phenol extracted, ethanol precipitated, dissolved in sequencing dye, heated to 90°C, and resolved on a 6% acrylamide, 7 M urea sequencing gel.

Results

Genomic Complexity of CRIPTO Gene-Related Sequences in Human Chromosomes

The 2020-bp-long CRIPTO cDNA, described elsewhere (Ciccodicola et al. 1989), contains an open reading frame of 564 bp, a 245-bp-long 5′ untranslated region, and a 1,209-bp-long 3′ untranslated re-

gion that includes an Alu sequence element. As a first approach to characterize the genomic organization of the gene encoding the CRIPTO protein, Southern blot analyses were carried out. The two cDNA fragments, 2B3 and G2 (fig. 1C), used as probes, hybridized to several genomic restriction fragments (fig. 1*A*). The 2B3 probe, used to analyze by Southern blot the genomic DNA of mouse and chicken, hybridized to several bands in the lanes containing mouse DNA (fig. 1*B*, first three lanes), whereas no hybridization was seen with chicken DNA (fig. 1*B*, fourth and fifth lanes).

Isolation and Characterization of CRIPTO Human Genomic Clones

To better understand the nature of the CRIPTO gene-related sequences, a human genomic library (Poustka et al. 1984) was screened (see Material and Methods) using CRIPTO fragment 2B3 as a probe, and 34 positive cosmid clones were isolated. *Eco*RI restriction analysis of 10 of the isolated clones revealed only three different restriction patterns in the inserts.

The isolated clones were hybridized to a synthetic oligonucleotide (G1) corresponding to nucleotides −91 to −110 of the 5′ noncoding region of CRIPTO cDNA (Ciccodicola et al. 1989; and fig. 2), with the intention of isolating the complete gene and discarding possible incomplete pseudogenes. A positive 800-bp *Pst*I/*Eco*RI fragment (CR-1-P7) was identified in the CR-1 cosmid clone (fig. 3, *top*).

CRIPTO Gene

DNA sequencing analysis revealed that clone CR-1 includes an intact structural gene encoding the entire human CRIPTO protein. The CRIPTO coding se-

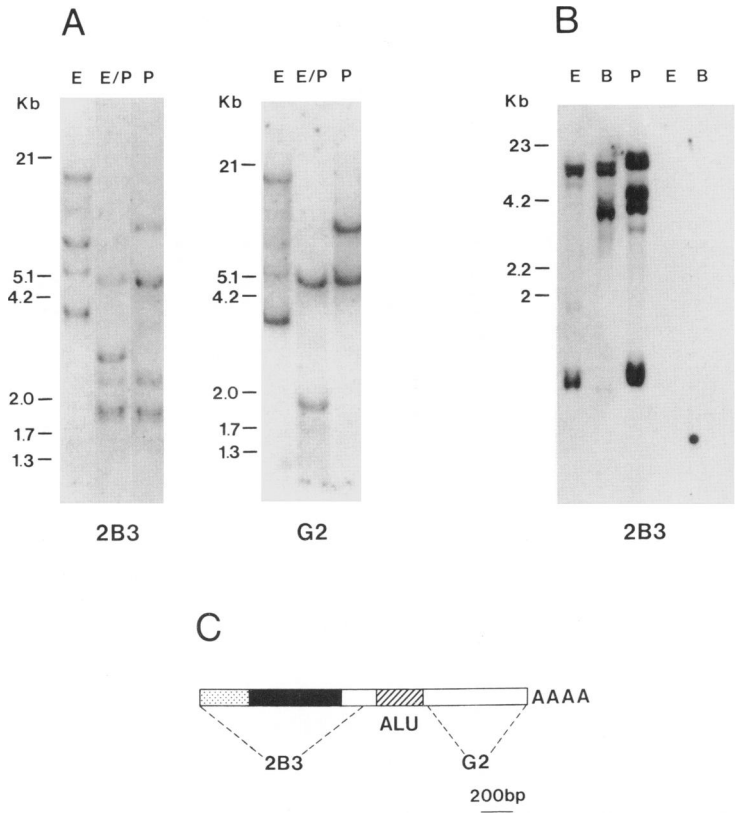
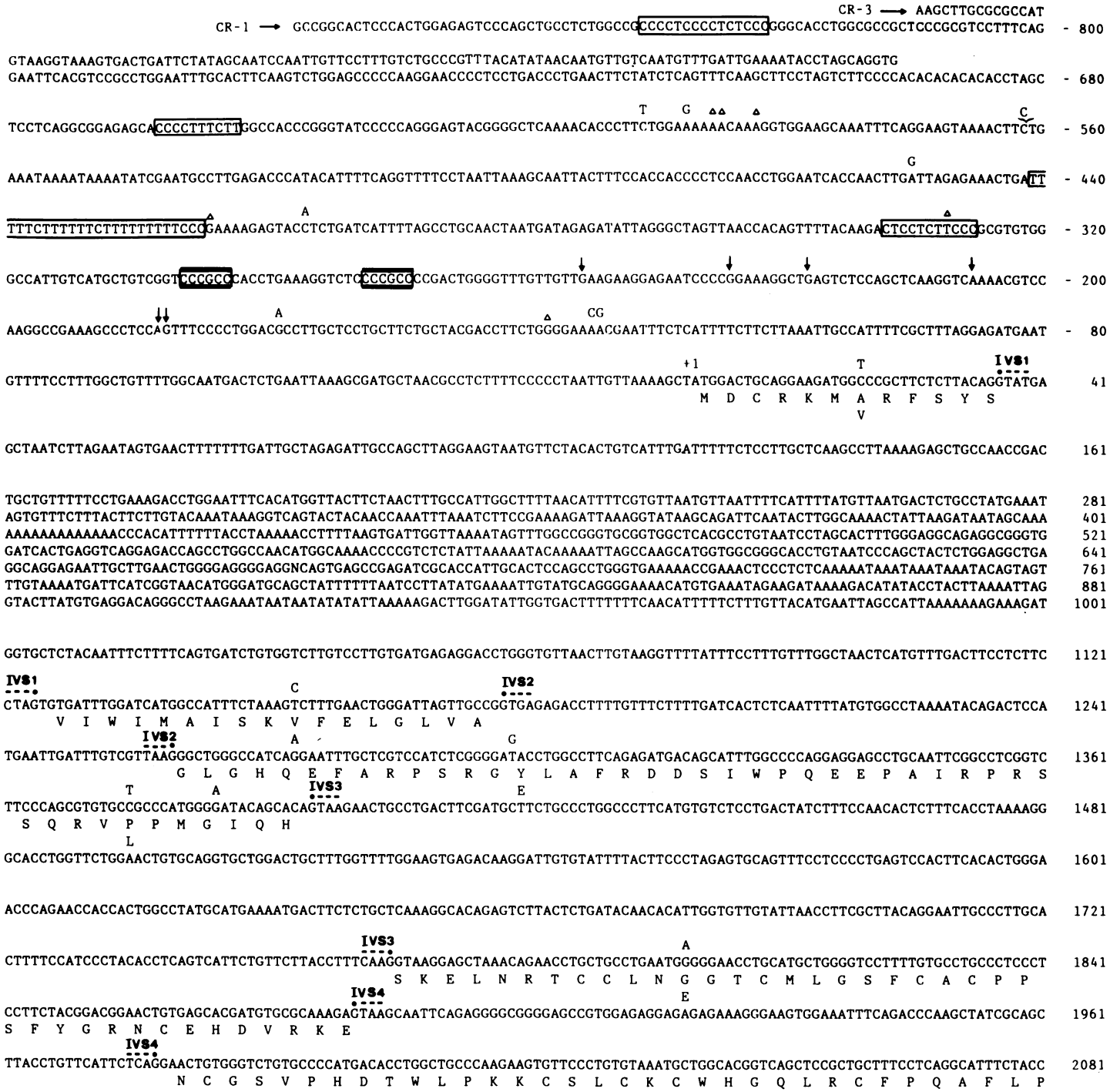


Figure 1 CRIPTO-related sequences in human and mouse DNA. *A*, Genomic DNA (10 μg) digested with *Eco*RI (lanes E), *Pst*I (lanes P), and *Eco*RI + *Pst*I (lanes E/P) and size fractionated by agarose gel electrophoresis. Hybridization probes are ^{32}P-nick-translated 2B3 and G2 segments. The molecular-weight markers included are *Hin*dIII/*Eco*RI-digested lambda DNA. *B*, DNA (10 μg) of mouse (first three lanes from left) and chicken (fourth and fifth lanes), digested with *Pst*I (P), *Bam*HI (B), and *Eco*RI (E). The hybridization probe is a ^{32}P-nick-translated 2B3 segment. Electrophoresis, transfer, and hybridization were as described in Material and Methods, except for washing conditions (2 × SSC at 60°C). *C*, Schematic representation of human CRIPTO cDNA. The coding region is indicated by the blackened box; AAAA = poly(A) tail. cDNA regions corresponding to 2B3 and G2 probes are indicated.


CR-3 → AAGCTTGCGCGCCAT
CR-1 → GCCGGCACTCCCACTGGAGAGTCCCAGCTGCCTCTGGCCGCCCCTCCCCTCTCCCGGGCACCTGGCGCCGCTCCCGCGTCCTTTCAG - 800
GTAAGGTAAAGTGACTGATTCTATAGCAATCCAATTGTTCCTTTGTCTGCCCGTTTACATATAACAATGTTGTCAATGTTTGATTGAAAATACCTAGCAGGTG
GAATTCACGTCCGCCTGGAATTTGCACTTCAAGTCTGGAGCCCCCAAGGAACCCCTCCTGACCCTGAACTTCTATCTCAGTTTCAAGCTTCCTAGTCTTCCCCACACACACACCCTAGC - 680
TCCTCAGGCGGAGAGCACCCCCTTTCTTGGCCACCCGGGTATCCCCCAGGGAGTACGGGGCTCAAAACACCCTTCTGGAAAAAACAAAGGTGGAAGCAAATTTCAGGAAGTAAAACTTCTG - 560
AAATAAAATAAAATATCGAATGCCTTGAGACCCATACATTTTCAGGTTTTCCTAATTAAAGCAATTACTTTCCACCACCCCTCCAACCTGGAATCACCAACTTGATTAGAGAAACTGATT - 440
TTTCTTTTTTCTTTTTTTTTCCGGAAAAGAGTACCTCTGATCATTTTAGCCTGCAACTAATGATAGAGATATTAGGGCTAGTTAACCACAGTTTTACAAGACTCCTCTTCCCGCGTGTGG - 320
GCCATTGTCATGCTGTCGGTCCCGCCCACCTGAAAGGTCTCCCCGCCCCGACTGGGGTTTGTTGTTGAAGAAGGAGAATCCCCGGAAAGGCTGAGTCTCCAGCTCAAGGTCAAAACGTCC - 200
AAGGCCGAAAGCCCTCCAGTTTCCCCTGGACGCCTTGCTCCTGCTTCTGCTACGACCTTCTGGGGAAAACGAATTTCTCATTTTCTTCTTAAATTGCCATTTTCGCTTTAGGAGATGAAT - 80
GTTTTCCTTTGGCTGTTTTGGCAATGACTCTGAATTAAAGCGATGCTAACGCCTCTTTTCCCCCTAATTGTTAAAAGCTATGGACTGCAGGAAGATGGCCCGCTTCTCTTACAGGTATGA 41
M D C R K M A R F S Y S
V
GCTAATCTTAGAATAGTGAACTTTTTTTGATTGCTAGAGATTGCCAGCTTAGGAAGTAATGTTCTACACTGTCATTTGATTTTTCTCCTTGCTCAAGCCTTAAAAGAGCTGCCAACCGAC 161
TGCTGTTTTTCCTGAAAGACCTGGAATTTCACATGGTTACTTCTAACTTTGCCATTGGCTTTTAACATTTTCGTGTTAATGTTAATTTTCATTTTATGTTAATGACTCTGCCTATGAAAT 281
AGTGTTTCTTTACTTCTTGTACAAATAAAGGTCAGTACTACAACCAAATTTAAATCTTCCGAAAAGATTAAAGGTATAAGCAGATTCAATACTTGGCAAAACTATTAAGATAATAGCAAA 401
AAAAAAAAAAAAACCCACATTTTTTACCTAAAAACCTTTTAAGTGATTGGTTAAAATAGTTTGGCCGGGTGCGGTGGCTCACGCCTGTAATCCTAGCACTTTGGGAGGCAGAGGCGGGTG 521
GATCACTGAGGTCAGGAGACCAGCCTGGCCAACATGGCAAAACCCCGTCTCTATTAAAAATACAAAAATTAGCCAAGCATGGTGGCGGGCACCTGTAATCCCAGCTACTCTGGAGGCTGA 641
GGCAGGAGAATTGCTTGAACTGGGGAGGGGAGGNCAGTGAGCCGAGATCGCACCATTGCACTCCAGCCTGGGTGAAAAACCGAAACTCCCTCTCAAAAATAAATAAATAAATACAGTAGT 761
TTGTAAAATGATTCATCGGTAACATGGGATGCAGCTATTTTTTAATCCTTATATGAAAATTGTATGCAGGGGAAAACATGTGAAATAGAAGATAAAAGACATATACCTACTTAAAATTAG 881
GTACTTATGTGAGGACAGGGCCTAAGAAATAATAATATATTAAAAAGACTTGGATATTGGTGACTTTTTTTCAACATTTTTCTTTGTTACATGAATTAGCCATTAAAAAAAGAAAGAT 1001
GGTGCTCTACAATTTCTTTTCAGTGATCTGTGGTCTTGTCCTTGTGATGAGAGGACCTGGGTGTTAACTTGTAAGGTTTTATTTCCTTTGTTTGGCTAACTCATGTTTGACTTCCTCTTC 1121
IVS1
CTAGTGTGATTTGGATCATGGCCATTTCTAAAGTCTTTGAACTGGGATTAGTTGCCGGTGAGAGACCTTTTGTTTCTTTTGATCACTCTCAATTTTATGTGGCCTAAAATACAGACTCCA 1241
V I W I M A I S K V F E L G L V A
IVS2
TGAATTGATTTGTCGTTAAGGGCTGGGCCATCAGGAATTTGCTCGTCCATCTCGGGGATACCTGGCCTTCAGAGATGACAGCATTTGGCCCCAGGAGGAGCCTGCAATTCGGCCTCGGTC 1361
G L G H Q E F A R P S R G Y L A F R D D S I W P Q E E P A I R P R S
IVS3
TTCCCAGCGTGTGCCGCCCATGGGGATACAGCACAGTAAGAACTGCCTGACTTCGATGCTTCTGCCCTGGCCCTTCATGTGTCTCCTGACTATCTTTCCAACACTCTTTCACCTAAAAGG 1481
S Q R V P P M G I Q H
L
GCACCTGGTTCTGGAACTGTGCAGGTGCTGGACTGCTTTGGTTTTGGAAGTGAGACAAGGATTGTGTATTTTACTTCCCTAGAGTGCAGTTTCCTCCCCTGAGTCCACTTCACACTGGGA 1601
ACCCAGAACCACCACTGGCCTATGCATGAAAATGACTTCTCTGCTCAAAGGCACAGAGTCTTACTCTGATACAACACATTGGTGTTGTATTAACCTTCGCTTACAGGAATTGCCCTTGCA 1721
IVS3
CTTTTCCATCCCTACACCTCAGTCATTCTGTTCTTACCTTTCAAGGTAAGGAGCTAAACAGAACCTGCTGCCTGAATGGGGGAACCTGCATGCTGGGGTCCTTTTGTGCCTGCCCTCCCT 1841
S K E L N R T C C L N G G T C M L G S F C A C P P
IVS4
CCTTCTACGGACGGAACTGTGAGCACGATGTGCGCAAAGAGTAAGCAATTCAGAGGGGCGGGGAGCCGTGGAGAGGAGAGAGAAAGGGAAGTGGAAATTTCAGACCCAAGCTATCGCAGC 1961
S F Y G R N C E H D V R K E
IVS4
TTACCTGTTCATTCTCAGGAACTGTGGGTCTGTGCCCCATGACACCTGGCTGCCCAAGAAGTGTTCCCTGTGTAAATGCTGGCACGGTCAGCTCCGCTGCTTTCCTCAGGCATTTCTACC 2081
N C G S V P H D T W L P K K C S L C K C W H G Q L R C F P Q A F L P

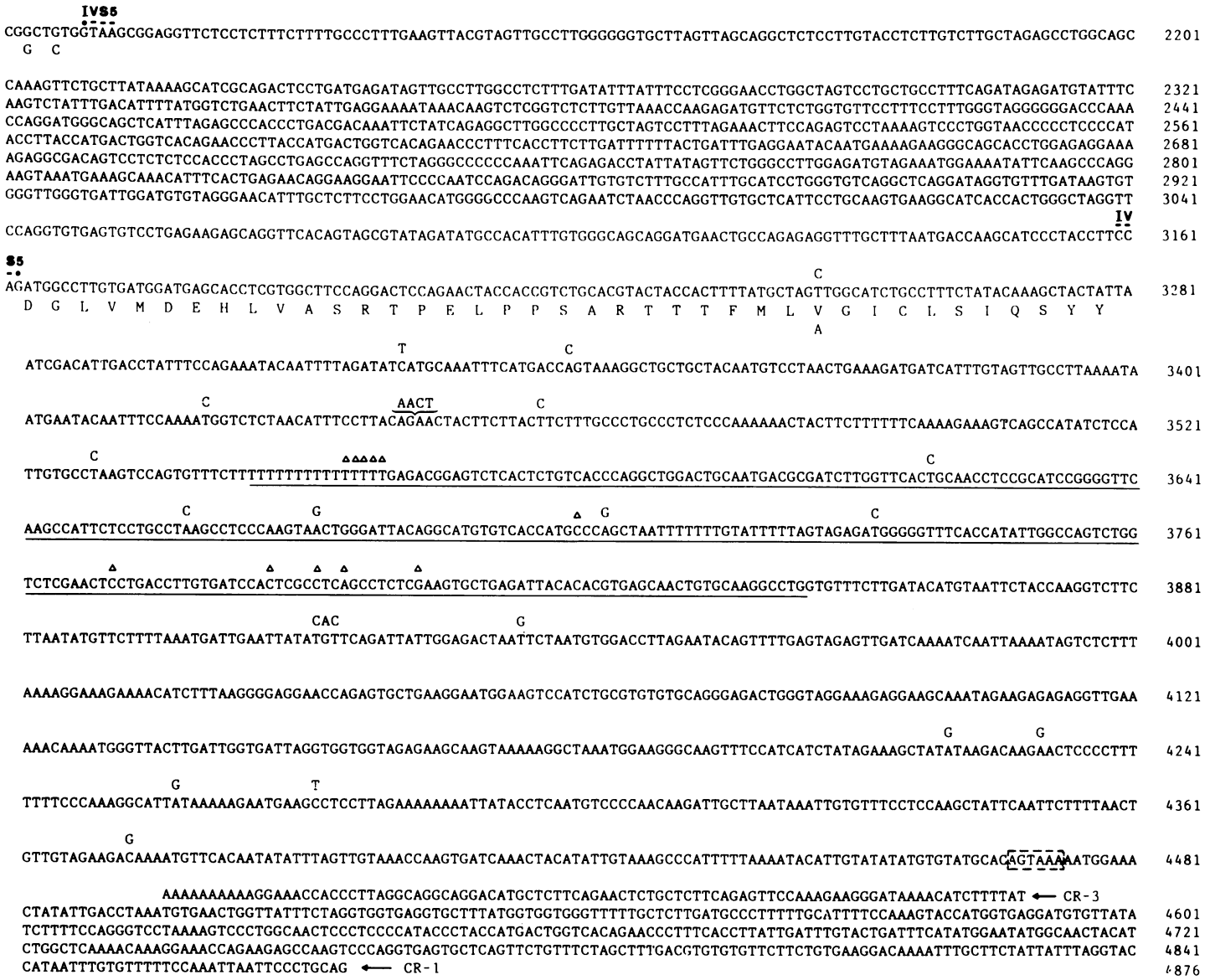


Figure 2 Nucleotide sequence of CR-1 and CR-3 genomic DNAs. The sequence of 5,763 nucleotides of the CR-1 gene is shown. The nucleotides are numbered from the start codon, and the amino acids for CR-1 are shown below. The nucleotide sequence of CR-3 is shown on top of CR-1. Nucleotide changes and deletions (Δ) in the CR-3 sequence are indicated above the CR-1 sequence. The six amino acid changes are indicated below the CR-1 protein sequence. It is to be noted that all the introns of CR-1 are absent in the CR-3 sequence. The boxed motifs are Sp1 binding sites (solid-line boxes), pyrimidine stretches (thin-line boxes), and polyadenylation signals (broken-line box). The vertical arrows indicate the multiple transcription starts. The Alu sequence present in the mRNA is underlined.

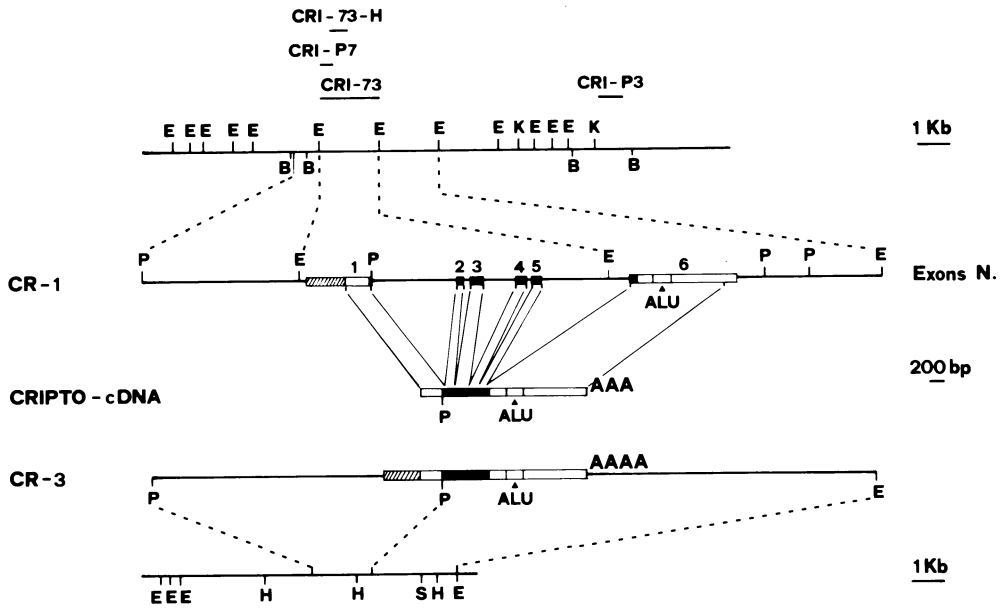


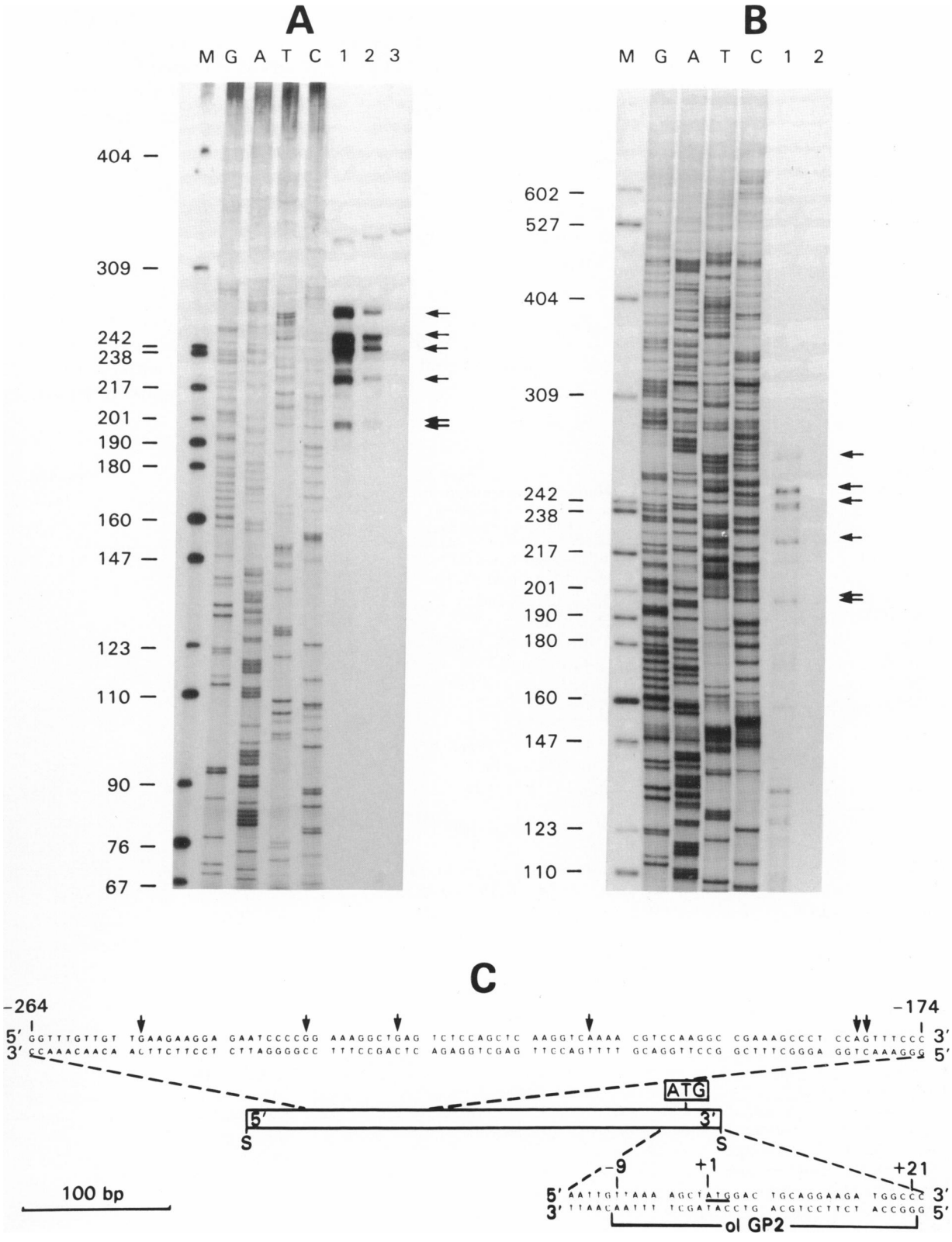
Figure 3 Maps summarizing information obtained from DNA sequencing and restriction mapping of isolated CRIPTO homologous recombinant clones. *Top,* Physical map of CR-1. Numbered exons are indicated by blackened boxes, for coding region, and by unblackened boxes, for noncoding regions. The hatched box represents the 440 bp upstream of the most common transcription start, present also in CR-3. Restriction sites indicated are *Eco*RI (E), *Bam*HI (B), and *Pst*I (P). Thick lines above the map denote genomic subclones used as probes. *Bottom,* Physical map of CR-3. Below are represented the genomic region isolated, as well as *Eco*RI (E), *Hin*dIII (H), and *Pst*I (P) restriction sites. CR-3 contains all the exons and a polyA tail (AAAA).

quence is encoded by six exons spanning a 4.8-kb-long DNA interval (fig. 3, *top*) The nucleotide sequences at the exon-intron boundaries were established by DNA sequence comparison of cDNA and genomic subclones. The 5′ donor and 3′ acceptor splice sites in each of the five introns conform to the GT . . . . AG rule and agree with the consensus sequence compiled for the exon-intron boundaries (Mount 1982), except for the acceptor sequence of the second and third introns (fig. 2). Exon 1 is 281 bp in length and contains the initiator methionine. The other exons range in size from 52 to 1,329 bp. The 3′-most exon, 1,329 bp in length, contains 118 bp of coding sequence and all of the 3′ untranslated region (3′ UT), which is 1,209 nucleotides long (fig. 2). The EGF-like domain exhibited by the CRIPTO protein (Ciccodicola et al. 1989) is encoded by exon 4.

CR-1 Gene Transcripts

A combination of S1 nuclease mapping and primer extension analyses was used to characterize the CR-1 transcription products. Since we found the CRIPTO gene to be expressed in an undifferentiated human teratocarcinoma cell line (NT2/D1) (Ciccodicola et al. 1989), we used poly(A)+ RNA isolated from cultured NT2/D1 cells. The probe used for S1 nuclease mapping was a double-strand DNA fragment encompassing nucleotides −302 to +18 of the genomic sequence and was labeled with ^{32}P at the 5′ end (fig. 4*C*). Five major S1 nuclease–protected fragments (fig. 4*A*)

Figure 4 Determination of CR-1 transcription initiation sites. *A,* S1 nuclease mapping analysis. Lanes 1 and 2, Products of S1 nuclease digestion in presence of 5 μg NT2/D1 poly(A)+ RNA and 40 μg NT2/D1 total RNA, respectively. Protected fragments are indicated on the right by horizontal arrows. Total yeast RNA (40 μg; lane 3) was used as a negative control. Lane M, Molecular-weight marker. The sizes of the fragments (in bp) are indicated on the left. Lanes G, to A, T, and C, Dideoxynucleotide-derived sequencing ladders obtained by priming recombinant pUC18 subclone of CR-1, CR-1-73 (fig. 3, *top*), with ol GP2 primer (described in fig. 4*C*). *B,* Reverse transcriptase analysis. Lanes M, G, A, T, C are as in panel *A*. Lanes 1 and 2, Results of primer extension in presence of 10 μg NT2/D1 poly(A)+ mRNA and total yeast RNA (as control), respectively, and reverse transcriptase (Boehringer). Extended fragments are indicated on the right by horizontal arrows. *C,* Schematic representation of probes used. The genomic DNA fragment *Sau96* (S)–*Sau96* (S), used for S1 analysis, and the oligonucleotide primer ol GP2, utilized for the primer extension reaction, are indicated. Arrows indicate major transcription initiation sites.


A
M G A T C 1 2 3
404
309
242
238
217
201
190
180
160
147
123
110
90
76
67
B
M G A T C 1 2
602
527
404
309
242
238
217
201
190
180
160
147
123
110
C
-264
-174
5′ GGTTTGTTGT TGAAGAAGGA GAATCCCCGG AAAGGCTGAG TCTCCAGCTC AAGGTCAAAA CGTCCAAGGC CGAAAGCCCT CCAGTTTCCC 3′
3′ CCAAACAACA ACTTCTTCCT CTTAGGGGCC TTTCCGACTC AGAGGTCGAG TTCCAGTTTT GCAGGTTCCG GCTTTCGGGA GGTCAAAGGG 5′
ATG
5′
3′
S
S
-9
+1
+21
5′ AATTGTTAAA AGCTATGGAC TGCAGGAAGA TGGCCC 3′
3′ TTAACAATTT TCGATACCTG ACGTCCTTCT ACCGGG 5′
ol GP2
100 bp

mapping between positions −180 to −253 of the genomic sequence were observed (fig. 4*C*).

The primer extension assay, performed with ol GP2 (see Material and Methods and fig. 4*C*), confirmed the five major products corresponding in length to the transcripts predicted by S1 analysis (fig. 4*B*). It should be noted that other bands are seen in primer extension experiments, probably because of both minor RNA species and early termination of the reverse transcriptase reaction. No further investigation was undertaken by us to study the origin of these minor bands.

Chromosome Mapping with Somatic Cell Hybrid Panel

A chromosome mapping panel was used to assign the CR-1 gene to human chromosomes. A 1.5-kb-long *Pst*I fragment derived from CR-1 (CR-1-P3; fig. 3, *top*) was used to probe a Southern blot of *Taq*I-digested genomic DNAs prepared from 23 hamster human somatic cell hybrids (table 1). Under conditions of high stringency one human-specific genomic fragment of 4.5 kb hybridized to the probe. The presence of the 4.5-kb fragment could be clearly distinguished in the DNA of the hybrid cell lines containing chromosome 3 (table 1).

When the *Eco*RI-*Pst*I fragment (CR-1-P7) containing 800 bp upstream of the translation initiation (see fig. 3, *top*) was used to probe the same Southern blot described previously and shown in figure 1, hybridization to two fragments was seen (e.g., in the lane containing human DNA digested with *Eco*RI and *Pst*I [fig. 5*A*], the 0.8-kb band corresponds to the genomic

Table 1

Segregation of CRIPTO-related Sequences in Human/Hamster Hybrids

Cell Line	Chromosome(s) Present	CR-1[a]	CR-3[a]
Hy.19.16T3D	Xq-, 10, 12–15, 18, and 20	−	−
HY.22AZA1	t(X;X),[b] 5, 12, 14, and 17–19	−	+
HY.31.24E	X, 5, 8, 11, 12, 14, and 21	−	+
HY.36.1	X, 8, 11, and 19	−	+
HY.60A	X, 5, 6, 8, 13, 14, 18, and 20	−	+
HY.70B1A	t(X;21),[c] 6, 15, and 17	−	−
HY.70B2	t(X;21),[c] 6, 13, 15, and 16	−	−
HY.75E1	X, 5, 9, and 12	−	+
HY.94A	X, 6–8, 16, and 22	−	+
HY.94BT1	t(X;Y),[d] 4, 7, 9, 11, 12, and 20	−	+
HY.95A1	X, 3, 5, 10, 11, and 14	+	+
HY.95B	X, 4, 6, 7, 14, 18, and 22	−	+
HY.95S	X, 2, 3, 13, and 21	+	+
HY.112F7	t(X;11),[e] 3, 4, 8, 10, and 20	+	+
RJ.369.1T2	13, 22	−	−
Y.173.5CT3	Xi, 1, 3, 4, 6, 8, 11, 12, 14, 15, 18, 21, and 22	+	+
YC2T1	X, 1, 11, 12, 14, 18, 19, and 20	−	+
HY.136C	X	−	+
Y.X6.8B2	t(X;6),[f] 1, 3, 5, 12–15, 17, 21, and 22	+	+
Y.162AZA	t(X;hamster)[g]	−	+
HY.87Z4	t(X;11),[h] 1, 2, 4–6, 12, 15, and 20	−	−
HY.85D30T2	t(X;1),[i] 2, 3, 8, 11, 13, 18, and 21	+	+
HY.84T2	Y	−	−
Chromosome assignment		3	Xq21-Xq22
Total no. of concordant +		6	17
Total no. of concordant −		17	6
Total no. discordant		0	0

[a] Symbols + and − indicate presence and absence, respectively.
[b] Xqter→Xq21::Xp22.3→Xqter
[c] Xqter→Xq22::21p13→21qter
[d] Xqter→Xp22.3::Yp→Yqter
[e] Xqter→Xq11.1::11p11.2→11q11
[f] Xqter→Xq21.3::6q27→6pter
[g] Xqter→Xq27.3::hamster
[h] Xqter→Xq26::11q23→11pter
[i] Xpter→Xq26::1q12→1qter

sequence CR-1). This indicated that the 5′ region of the CR-1 gene was present in two copies in the human genome. When the CR-1-P7 fragment was used to probe the above-mentioned hamster-human somatic cell hybrid panel, it was possible to obtain the segregation of the two sequences (fig. 5*B*). Because of both the hybridization pattern summarized in table 1 and, in particular, the pattern obtained using the hybrid cell lines containing portions of the X chromosome already described (Rocchi et al. 1986), we can assign the second genomic copy to the Xq21-22 region.

Isolation and Characterization of a Second Genomic CRIPTO-related Sequence

A genomic library was screened (see Material and Methods) to isolate the genomic clones containing the 5′ cDNA noncoding region, using as probe the labeled CR-1-P7 DNA fragment (fig. 3). Only two different classes of recombinant phages exhibiting the restriction pattern expected from the Southern blot were found (fig. 5).

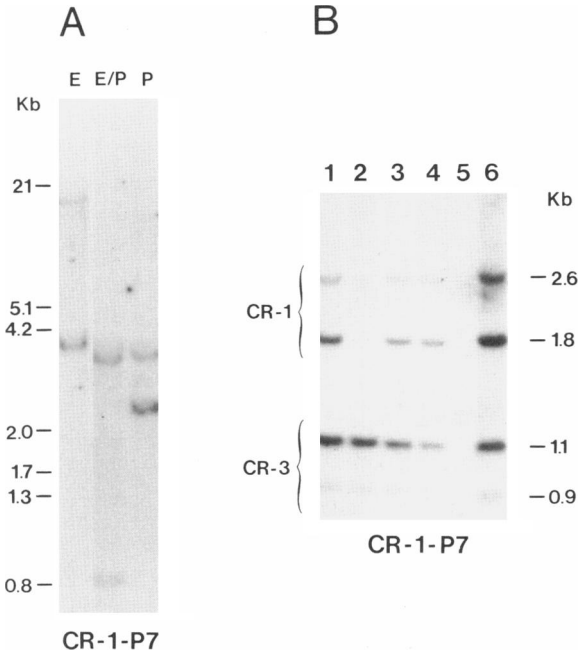


Figure 5 Southern blot analysis of human DNA by CR-1-P7 probe. *A*, Genomic DNA (10 μg) digested with *Eco*RI (E), *Pst*I (P), *Eco*RI + *Pst*I (E/P) and size-fractionated by agarose gel electrophoresis. On transfer of DNA to the Z-bind membrane, hybridization analysis was performed with the ^{32}P-nick-translated CR-1-P7 probe (fig. 3, *top*). Molecular-weight markers included are *Hin*dIII + *Eco*RI-digested lambda DNA. *B*, DNA (10 μg) derived from hybrids HY.95A1 (lane 1), HY.95B (lane 2), HY.95S (lane 3), HY.112F7 (lane 4), Chinese hamster (lane 5), and human (lane 6), digested with *Taq*I, and size-fractionated and hybridized as described above. The sizes (in kb) of the bands are indicated on the right-hand side. Refer to table 1 for chromosome content of the hybrid cell lines.

The restriction map of clone CR-3 is shown at the bottom of figure 3. To investigate whether the CRIPTO-related genomic sequences from recombinant lambda CR-3 clones encode a complete CRIPTO protein, we determined the nucleotide sequence of a 2,688-bp fragment hybridizing to 2B3 and G2 and compared this sequence with that of the cDNA (fig. 2 and fig. 3, *bottom*).

Analysis of the nucleotide sequence of CR-3 revealed that this clone includes a complete CRIPTO cDNA lacking introns and containing a poly(A) tract at the 3′ end. Seven single-basepair substitutions are observed in the coding region (fig. 2), and six of these give rise to amino acid changes. The 3′ noncoding sequence is less similar (97% identical) to CR-1. Most of the base changes, deletions, and insertions fall within the inverted Alu sequence. The unusual polyA addition site AGTAAA present in the CR-1 gene is conserved also in CR-3. The similarity between CR-1 and CR-3 extends for 976 nucleotides upstream of the initiator AUG; seven single-basepair substitutions and six nucleotide deletions are present in this DNA region.

Discussion

We have reported the isolation and characterization of two genomic clones—CR-1 and CR-3—encoding polypeptides related to the CRIPTO polypeptide previously described. We hypothesize that CR-1 represents the gene expressed in human teratocarcinoma cells, because of the identity between its exonic regions and the cDNA isolated elsewhere (Ciccodicola et al. 1989).

S1 mapping and primer extension experiments indicate that in NT2/D1 many CR-1 gene transcripts are present, initiating at multiple closely spaced sites located upstream of the translation start AUG, and that no intronic sequences interrupt the 5′ untranslated region of the CR-1 gene. Analysis of the genomic DNA region upstream of these mRNAs revealed (*a*) the absence of TATA and CAAT boxes (Breathnach et al. 1981; Benoist and Chambon 1981), (*b*) the presence of a 25-pyrimidine stretch from nucleotide −416 to nucleotide −391 (Evans et al. 1984; Yu and Manley 1986), and (*c*) two potential Sp1 binding sites (fig. 2). Only the Sp1 binding-site sequence located at nucleo-

tide −298 conforms to the consensus sequence 5′-G/TGGGCGGG/AG/AC/T3′ (Kadonaga et al. 1986).

The CR-1 gene is organized into six exons, with exon 4 containing the entire cysteine-rich EGF-like segment. The human AR and TGFα genes are both organized into six exons. The cysteine-rich EGF-like sequence is interrupted by an intron in the AR, TGFα, and EGF genes (Plowman et al. 1990) but is contained within a single exon in the eight EGF-like repeats of the EGF precursor gene, in the three repeats in the low-density-lipoprotein receptor gene, and in the blood-clotting-system genes (factors IX, X, and XII and protein C) (Sudhof et al. 1985*a*, 1985*b*). We can hypothesize that, from a common ancestral gene containing the cysteine-rich segment coded by only one exon, two groups of genes have been derived: one group, like the CR-1 gene, contains the motif in a single exon, whereas the other group, like the genes coding from the growth factors of the EGF family, contains this motif interrupted by an intron.

The CR-3 sequence has all the characteristics of a retroposon, including the poly(A) tail and the lack of introns (fig. 2 and fig. 3, *bottom*). A few mutations distinguish the CR-3 and cDNA sequences, and six amino acid changes are present in the CRIPTO-3 protein.

The transposition of genetic material from one locus to another through an RNA intermediate is a documented evolutionary mechanism for sequence duplication, dispersion, and rearrangement, which contribute to the fluidity of eukaryotic genomes. Some 5%–10% of the mammalian genome may consist of sequences that arose by retroposition of cellular RNAs (see review by Weiner et al. 1986). These retroposons, like the Alu and L1 sequences, generally possess an internal promoter, no introns, and a remnant of a 3′ polyadenylated sequence and are often flanked by short direct repeats. Deletions at the 5′ end and accumulation of mutations contribute to transform these retroposons into nonfunctional genes.

There are a few functional genes with characteristics of retroposons (Soares et al. 1985; Clark et al. 1986; Boer et al. 1987), suggesting that retroposition may be a mechanism for gene duplication. In these cases a new promoter drives the expression of the gene accounting, e.g., for the testis-specific expression of the phosphoglycerate kinase gene, pgK-2 (Boer et al. 1987; McCarrey and Thomas 1987).

The novel feature of CR-3 is the presence of a 440-bp sequence, upstream of the first start of transcription, that is 97% homologous to the CR-1 gene. The presence of this long 5′ upstream region of CR-1 in CR-3 can be explained by hypothesizing retroposition of a transcript starting either 440 bp upstream of the most common start of transcription or, possibly, as a *Pol*III transcript starting farther upstream. Alternatively, a gene conversion mechanism is bound to be involved to justify the similarity between these two regions. Martin et al. (1983) have proposed that function was restored to the human δ-globin gene, by a gene conversion mechanism, from part of the promoter region of the β-globin gene.

The analysis of the nucleotide sequence of clone CR-3 upstream of the region of similarity with CR-1 shows the presence of CAAT and TATA boxes, indicating that the retroposition event has probably located CR-3 downstream of a promoter region. At present we cannot describe the mechanism that has resulted in the same 5′ region in front of the CR-3 sequence, as we have no data demonstrating that this member of the family is expressed in certain cell types.

It is worth emphasizing that very few mutations have accumulated in the CR-3 sequence. It is conceivable that these two genes have related or identical functions and may be expressed at different times or in different cell types, the differences residing in the few mutations present in both noncoding and coding regions. This hypothesis is supported by preliminary data obtained from the analysis of the mouse genome. Among the mouse CRIPTO sequences analyzed, one presents, like the human CR-3, a few sense mutations (data not shown).

Acknowledgments

We thank Dr. Robert G. Martin (NIH) for useful and stimulating discussions and for his generous hospitality to M.G.P. in his laboratory. We thank Dr. J. Nickols (NIH) for providing the chicken DNA. We are grateful to Drs. Paolo Di Nocera, Edoardo Boncinelli, and Bruce Paterson for critical review of the manuscript. We thank Maria Terracciano for excellent technical assistance. This work was supported in part by grants (to M.G.P.) from both the Associatione Italiana per la Ricerca sul Cancro (AIRC) and the P.F. Ingegneria Genetica, CNR. R.D. and N.M. are AIRC Fellows.

References

Andrews PW, Damjanov I, Simon D, Banting G, Carlin C, Dracopolo NC, Fogh J (1984) Pluripotent embryonal carcinoma clones derived from the human teratocarcinoma cell line Tera-2. Lab Invest 50:147–162

Aviv H, Leder P (1972) Purification of biologically active globin messenger RNA by chromatography on oligothy-

midylic acid-cellulose. Proc Natl Acad Sci USA 69:1408–1412

Benoist C, Chambon P (1981) In vivo sequence requirements of the SV40 early promoter region. Nature 290:304–310

Benton W, Davis R (1977) Screening λ gt recombinant clones by hybridization to single plaque in situ. Science 196:180–192

Boer PH, Adra CN, Lau YF, McBurney MW (1987) The testis-specific phosphoglycerate kinase genes pgk-2 is a recruited retroposon. Mol Cell Biol 7:3107–3112

Breathnach R, Chambon P (1981) Organization and expression of eukaryotic split genes coding for proteins. Annu Rev Biochem 50:349–383

Chirgwin JM, Pzybyla AE, Macdonald RJ, Rutter WJ (1979) Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 18: 5294–5304

Ciccodicola A, Dono R, Obici S, Simeone A, Zollo M, Persico MG (1989) Molecular characterization of a gene of the "EGF family" expressed in undifferentiated human NTERA2 teratocarcinoma cells. EMBO J 8:1987–1991

Clark BD, Collins KL, Gandy MS, Webb AC, Auron PE (1986) Genomic sequences for human prointerleukin I beta: possible evolution from a reverse transcribed prointerleukin I alpha gene. Nucleic Acids Res 14:7897–7914

Davidson RD (1976) Improved technique for the induction of mammalian cell hybridization by polyethyleneglycol. Somatic Cell Genet 2:165–176

Derynck R, Roberts AB, Winkler ME, Cheng EY, Goeddel DV (1984) Human transforming growth factor-α: precursor structure and expression in E. coli. Cell 38:287–297

Devel TF (1987) Polypeptide growth factors: roles in normal and abnormal cell growth. Annu Rev Cell Biol 3:443–492

Evans T, Schon E, Gora-Maslak G, Patterson J, Efstratiadis A (1984) S1 hypersensitive sites in eukaryotic promoter region. Nucleic Acids Res 12:8043–8058

Greenwald I (1985) lin-12, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. Cell 43:583–590

Grunstein M, Hogness D (1975) Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci USA 72:3961–3965

Hattori M, Hidaka S, Sakaki Y (1985) Sequence analysis of a KpnI family member near the 3′ end of human β-globin gene. Nucleic Acids Res 13:7813–7827

Kadonaga JT, Jones KA, Tjian R (1986) Promoter-specific activation of RNA polymerase II transcription by SP1. Trends Biosci 11:20–23

Kidd S, Kelley MR, Young MW (1986) Sequence of the Notch locus of *Drosophila melanogaster:* relationship of the encoded protein to mammalian clotting and growth factors. Mol Cell Biol 6:3094–3108

McCarrey JR, Thomas K (1987) Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature 326:501–505

Maniatis T, Fritsch CF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Martin SL, Karen A, Vincent A, Wilson AC (1983) Rise and fall of the delta globin gene. J Mol Biol 164:513–528

Mercola M, Stiles CD (1988) Growth factor superfamilies and mammalian embryogenesis. Development 102:451–460

Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10:459–472

Plowman GD, Green JM, McDonald VL, Neubauer MG, Disteche CM, Todaro GJ, Shoyab M (1990) The amphiregulin gene encodes a novel epidermal growth factor-related protein with tumor-inhibitory activity. Mol Cell Biol 10:1969–1981

Poustka A, Rackwitz HR, Frischaug A, Hohn B, Lehrach H (1984) Selective isolation of cosmid clones by homologous recombination in Escherichia coli. Proc Natl Acad Sci USA 81:4129–4133

Rigby PWJ, Dieckmann M, Rhodes C, Berg P (1977) Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J Mol Biol 113:237–251

Rocchi M, Roncuzzi L, Santamaria R, Archidiacono N, Dente L, Romeo G (1986) Mapping through somatic cell hybrid and cDNA probes of protein C to chromosome 2, factor X to chromosome 13, and α1-acid glycoprotein to chromosome 9. Hum Gen 74:30–33

Savage CR, Inagami T, Cohen S (1972) The primary structure of the epidermal growth factor. J Biol Chem 247: 7612–7621

Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, et al (1985) RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol 5:2090–2103

Sporn MB, Roberts AB (1987) Peptide growth factors are multifunctional. Nature 332:217–219

Sudhof TC, Goldstein JL, Brown MS, Russel D (1985*a*) The LDL receptor gene: a mosaic of exons shared with different proteins. Science 228:815–822

Sudhof TC, Russel DW, Goldstein JL, Brown MS, Sanchez-Pescador R, Bell GI (1985*b*) Cassette of eight exons shared by genes for LDL receptor and EGF precursor. Science 228:893–897

Weiner AM, Deininger PL, Efstratiadis EA (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661

Whitman M, Melton DA (1989) Growth factors in early embryogenesis. Annu Rev Cell Biol 5:93–117

Yanisch-Perron C, Vieira J, Messing J (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of M13mp18 and pUC19 vectors. Gene 33: 103–109

Yu YT, Manley JL (1986) Structure and function of the S1 nuclease-sensitive site in the adenovirus late promoter. Cell 45:743–751