The mRNA for α1(XIX) Collagen Chain, a New Member of FACITs, Contains a Long Unusual 3′ Untranslated Region and Displays Many Unique Splicing Variants[1]

Kazuhito Inoguchi, Hidekatsu Yoshioka, Mohammed Khaleduzzaman, and Yoshifumi Ninomiya[2]

Department of Molecular Biology and Biochemistry, Okayama University Medical School, Okayama, Okayama 700

Received for publication, August 10, 1994

We have isolated cDNAs and completed for the first time the primary structure for a novel collagenous chain that was partially characterized earlier and named α1(Y) chain [Yoshioka, H. *et al.* (1992) *Genomics* 13, 884-886]. The size of the coding region was unexpectedly small compared with the length of the mRNA (>10 kb), owing to the presence of a long 3′ untranslated region (>5 kb). The predicted polypeptide contained 1,142 amino acid residues with a 23-residue signal peptide consisting of 5 collagenous domains of 70-224 residues in length, interspersed and flanked with 6 noncollagenous (NC) domains. The primary structure is distinct from those of the 32 known collagen α-chains of types I through XVIII. Therefore, we designate this newly discovered collagen chain the α1 chain of type XIX collagen. Sequence analysis suggested that this chain belongs to the recently discovered group of collagens known as FACITs (fibril associated collagens with interrupted triple-helices). Northern blotting analysis demonstrated hybridization of the cDNA to a large mRNA species (>10 kb) extracted from a rhabdomyosarcoma cell line (CCL 136). We also isolated numerous truncated cDNA clones of which the 3′ parts were different from the "proto" type of the mRNA of >10-kb size. Sequence comparison between cDNAs and corresponding genomic DNA fragments indicated that unusual splicing events occurred through insufficient recognition at acceptor sites. Expression of the gene was extremely infrequent in the rhabdomyosarcoma cell line; it could be restricted to certain animal tissues both temporally and spatially during early development.

Key words: COL19A1, FACIT, mRNA splicing, rhabdomyosarcoma cell, type XIX collagen.

Collagen constitutes a large family of extracellular matrix proteins with distinct tissue distributions and functions. To date, 18 different types of collagen have been described with distinct primary structure and function (*1-5*). The collagen superfamily can be divided into several subgroups: the fibrillar collagens (types I, II, III, V, and XI), the FACIT (fibril-associated collagens with interrupted triple-helices: types IX, XII, XIV, and XVI), the short-chain collagens (types VIII and X), the basement membrane collagens (type IV), the multiplexins (proteins with multiple triple-helix domains and interruptions, XV and XVIII, *5*), and others (types VI, VII, XIII, and XVII). The FACIT are associated with the major fibrillar collagens (types I or II) and minor fibrillar collagens (types V and XI), which form fibrils staggered by 67 nm with different diameters in non-cartilaginous and cartilaginous tissues. These fibrils play important biological roles in interaction with other matrix components. Members of the FACIT share several common structural features (*1*). Namely, they have short stretches of collagenous (COL) domains interspersed by non-collagenous (NC) domains. The structure of these molecules can be divided into functional subdomains. One domain, which comprises a couple of collagenous regions, serves for interaction with and adhesion to the fibrils. A second domain, comprising another collagenous region, serves as a rigid arm that projects out of the fibril. The third domain, which is a non-collagenous domain, may serve for interaction with other matrix components or with cells. In contrast to the common features of the family, differences are seen in the size of their molecules and in their tissue-specific expression. For instance, type IX collagen, which is the best-characterized molecule of this group, is found in tissues containing type II collagen such as hyaline cartilage and the vitreous body of the eye. On the other hand, type XII collagen, which is another well-characterized member of this group, is found in dense connective tissues such as tendons and ligaments where type I collagen exists. The other two molecules of this group, type XIV (*6*) and XVI (*7*) collagens, were discovered recently and have only been characterized partially. Type XIV collagen is similar to type XII molecules in overall structure, having a large

[1] This work was supported by a Grant-in-Aid for Scientific Research (#03670178) from the Ministry of Education, Science and Culture of Japan. The nucleotide sequence reported in this paper has been submitted to the GenBank/EMBL Data Bank with accession numbers, D38163, D38164, D38460, D38461, D38462, D38463, D38464, D38465, and D38466.

[2] To whom correspondence should be addressed.

Abbreviations: bp, base pairs; COL, collagenous; FACIT, fibril-associated collagens with interrupted triple-helices; kb, kilobases; SSC, 0.15 M NaCl, 0.015 M sodium citrate (pH 6.8); NC, noncollagenous; nt, nucleotide(s); RACE, rapid amplification of cDNA ends; RT-PCR, reverse transcriptional polymerase chain reaction.

amino-terminal domain divided into multi subdomains with homologies to von Willebrand's factor A-domains, fibronectin type III repeats, and NC 4 of α1(IX) collagen chain. Type XIV collagen seems to be uniformly distributed in adult tissues. However, some differences in tissue distribution were noted in various developmental stages (*8*). Type XVI collagen, which is characterized by a cysteine-rich motif in its NC domains, is found in human placental tissues (*9*).

We (*10*) and Myers *et al.* (*11*) recently reported on the isolation and characterization of cDNAs encoding a novel collagenous polypeptide. Sequence analysis suggested that this polypeptide is a member of FACIT. Interestingly, the chromosomal location of the gene was fixed at position 6q12-q14, the same region to which the α1(IX) collagen gene (COL9A1, *12*) and α1(XII) gene (COL12A1, *13*) were assigned. Here we present the entire amino acid sequence of this peptide derived from the analysis of overlapping cDNAs. Our results have shown that the coding region is unexpectedly short relative to the length of the mRNA due to a long 3′ untranslated region. In addition to cDNA clones encoding a "proto type" of the chain, we also isolated many clones that seem to encode forms of the chain truncated at their carboxy-terminal regions. Comparison of the sequences of cDNA and genomic clones has revealed that these variant transcripts are produced through incomplete splicing.

MATERIALS AND METHODS

Isolation and Characterization of cDNA and Genomic DNA Clones—RNA was extracted from human rhabdomyosarcoma cell line (CCL 136) by the guanidium thiocyanate method (*14*). cDNAs were synthesized from random hexamers and constructed in λgt10 by use of a cDNA library synthesis kit (Amersham, RPN1256Y, RPN1712, and RPN1717). For the screening, we carried out prehybridization with 5×SSC [1×SSC, 0.15 M NaCl, 0.015 M sodium citrate (pH 7.0)], 1% *N*-laurylsarcosine, and 50 μg/ml salmon sperm DNA at 65°C, 1 h; hybridization with 5×SSC, 1% *N*-laurylsarcosine, 50 μg/ml salmon sperm DNA, and labeled probe at 65°C, overnight; and washing (3×SSC and 0.5% *N*-laurylsarcosine at 65°C, 15 min, twice, and 3×SSC at 65°C, 15 min, twice) (*3*). A human genomic DNA library in EMBL-3 (Clontech, HL1067j) was also used in screening for the genomic DNA fragments under the same conditions.

Northern Blotting Analysis and the 5′ and 3′ RACE—Poly(A)$^+$ RNA was fractionated from total rabdomyosarcoma cell RNA with an mRNA extraction kit (Japan Roche, Oligo-Tex). RNAs were electrophoresed on 0.8% agarose gel under denaturing conditions, blotted onto Hybond N (Amersham) nylon filters, and hybridized with specific probes under regular conditions (*14*).

In order to determine the structure of the 5′ end of the cDNA, we used 5′ RACE as described (*15*). Four primers were synthesized for this purpose: primer 1 = 5′-GTCATATGTCAGCTGATG-3′ (the other strand from nucleotide number 241 to 258 in Fig. 2); primer 2 = 5′-GTGCCTTGAAACCATGTG-3′ (the other strand from nucleotide number 99 to 116 in Fig. 2); hybrid primer = 5′-CTGAATTCTCGAGTCGAC(T)$_{17}$-3′; and adapter primer = 5′-CTGAATTCTCGAGTCGAC-3′. The first strand cDNA was synthesized by Moloney murine leukemia virus reverse transcriptase (U.S. Biochemical) from primer 1 with total RNA used as the template. A stretch of d(A) was added to the first strand cDNA at its 3′ end by terminal deoxynucleotidyl transferase (BIORAD). An aliquot of this material was used for PCR. The reaction was performed from primer 2 and hybrid/adapter primer under the following conditions: 94°C for 140 s, followed by 36 cycles of 94°C, 40 s; 55°C, 1 min; 70°C, 3 min. At the end of the last cycle, the sample was further incubated at 70°C for 15 min. The products were electrophoresed in 1.2% agarose gel, blotted onto Hybond N$^+$ (Amersham), and hybridized with the ^{32}P-labeled specific probe (nucleotide number from 10 to 82). For the 3′ RACE experiment, we used oligo d(T) for first strand synthesis and specific primers of a part of KI 50 and KI 51 for PCR. For cloning the PCR products, the TA cloning system (In Vitrogen) was used.

Nucleotide Sequencing Analysis—Nucleotide sequencing analysis was performed by the dideoxy chain termination technique (*16*) using double-stranded pBluescript II vectors (*17*). We used the ^{35}S-dATP labeling method, the fluorescence-labeled dye-terminator method, and an ABI automatic sequencer (373A) as well. For long fragments, a Kilo-Sequence deletion kit (Takara Biomedicals) using exonuclease III and mung bean nuclease and specific primers were utilized for sequencing. For analysis of the 5′ region of the gene, we used the MacVector program (version 4.1).

RESULTS

Isolation of the Overlapping cDNAs Encoding the Entire Novel Collagen Chain—We previously reported the partial sequence and identification of the chromosomal location of a novel human collagen chain belonging to the FACIT group of collagen (*10*). In order to complete the primary structure of the chain, we constructed a human cDNA library from fetal rhabdomysarcoma cell line (CCL 136) RNA and screened it with a cDNA clone HY 67 as a probe. In the screening of 1×10^6 plaques, 19 positive clones were obtained. Three clones, KI 1, KI 6, and KI 18, which gave strong hybridization signals were characterized further. Nucleotide sequence analysis indicated that all three clones were overlapped with HY 67. One of the clones (KI 6) extended farther in the 5′ direction, as shown in Fig. 1 and contained an ATG codon (nucleotides 118-120 in Fig. 2) followed by a sequence encoding a 23-residue hydrophobic peptide (nucleotides 121-186). The other two clones, KI 1 and KI 18, which overlapped with HY 67 at its 3′ end, each contained a different sequence at its 3′ end. The question here is which 3′ end of the two really encodes the "proto" type of the chain. So we further screened the same cDNA library using KI 1 as probe, and isolated clone KI 40. Similarly KI 50 and KI 51 clones were obtained with KI 40 as the probe. These three overlapping cDNAs contained the coding region for the collagenous polypeptide. KI 50 encoded a collagenous domain of 70 amino acid residues and a noncollagenous domain (19 residues) containing two cysteinyl residues followed by a termination codon and a 3′ untranslated region. The relative position of the two cysteinyl residues is typical for polypeptides in the FACIT group. Northern blot analysis of the cDNAs of KI 1, 40, 50, and 51 revealed the coding region and noncoding region

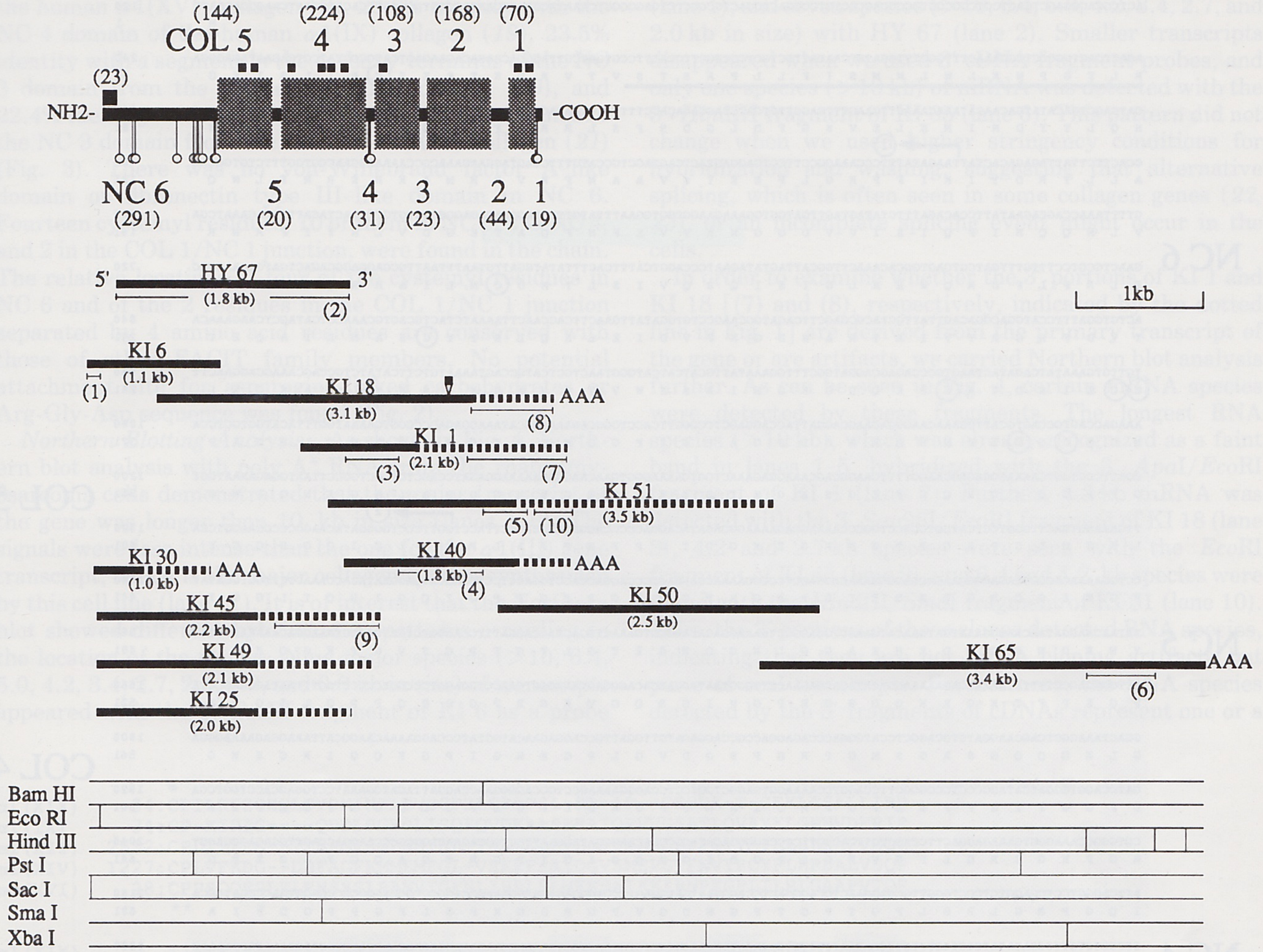
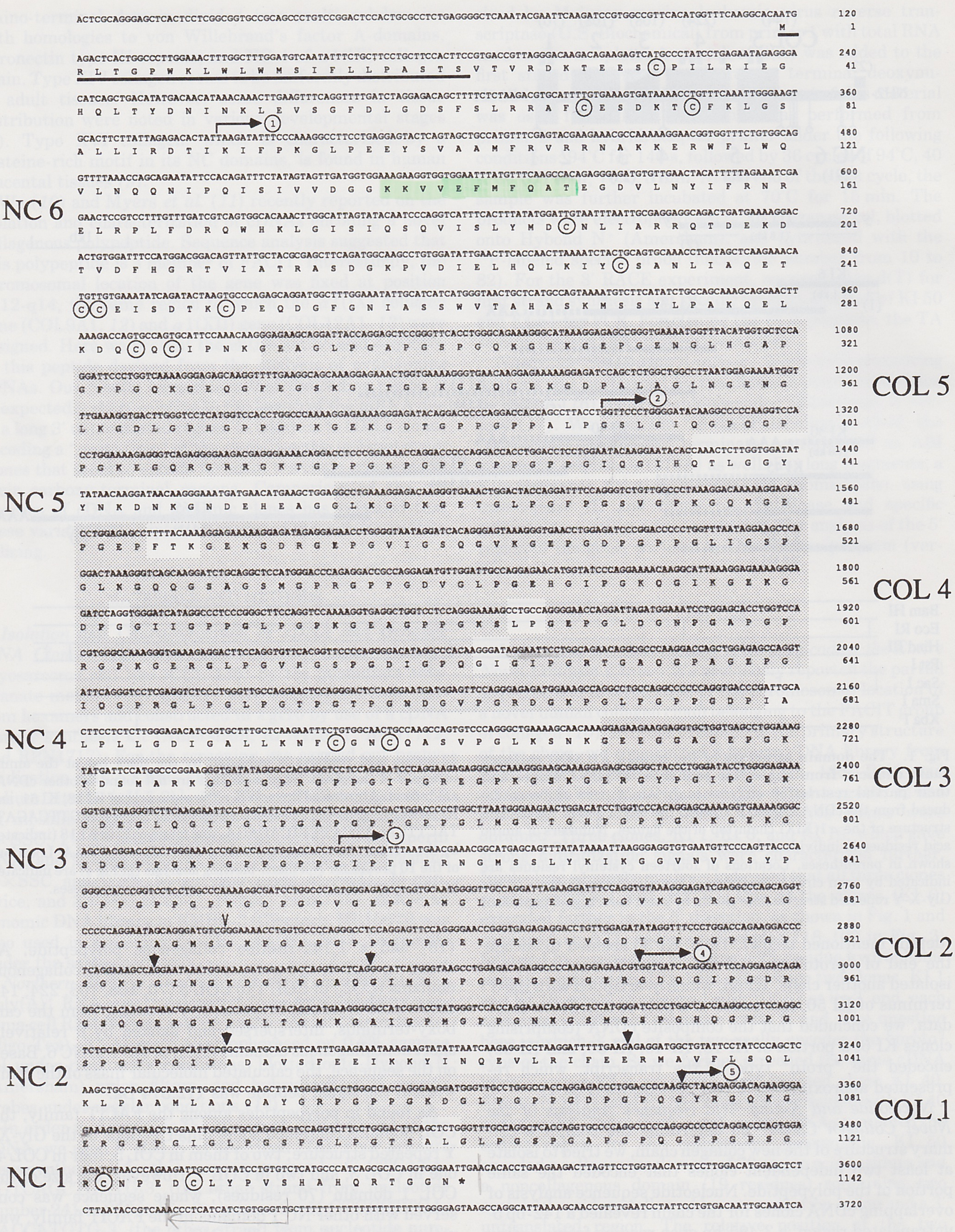


Fig. 1. **The domain structure of the human α1(XIX) collagen chain deduced from nucleotide sequences of cDNA clones and their partial restriction enzyme map.** Amino acid sequence deduced from the cDNA sequence allowed the prediction of the domain structure of the α1(XIX) chain(top of the figure). Numbers of amino acid residues in individual domains of COL 1 to 5 and NC 1 to 6 are shown in parentheses. Locations of fourteen cysteinyl residues are indicated by open circles. Positions of the nine imperfections of the Gly-X-Y repeated structure are shown by short bars. Signal peptide of 23 amino acid residues is indicated by the thick bar at the amino terminus of the chain. Sequences that are not found in other cDNAs are indicated by dotted lines. When the sequences of KI 18, KI 51, and KI 40 were compared, an extra sequence of 33 bp, GTGAGTTCAGATTACTTCACATCATTTTCACAGA, was found only in KI 18 (indicated by the wedge). Location of the poly A tail is shown by AAA. Locations of the 10 probes for Northern-blotting analysis (in Fig 4) are indicated by thin bars with the numbers 1 through 10 in parentheses.

(details mentioned below). Thus, KI 50 seemed to encode the end of "proto" type of the chain. Furthermore, we isolated another clone, KI 65, which overlapped with the 3′ terminus of KI 50, and ended in a poly A tail. From these data, we concluded that the composite cDNA (comprising clones KI 6, a portion of KI 18, KI 40, KI 50, and KI 65) encoded the "proto" type of the transcript, which represented approximately 9,000 nucleotides (Fig. 1).

Nucleotide and Amino Acid Sequence Analysis of the Novel Collagen Chain—To determine the complete primary structure of the new collagen chain, we tried to isolate at least two independent clones that encoded the same portion of the polypeptide. Nucleotide sequence analysis of overlapping cDNA clones for the chain revealed a 142-bp 5′ untranslated region, a 3,495-bp open reading frame coding for 1,165 amino acid residues, and a 3′-untranslated region (>5 kb). The predicted protein contained a putative 23-residue hydrophobic signal peptide at the amino terminus (*18*) and a 1,142-residue collagenous polypeptide. As shown in Fig. 2, the polypeptide contained 5 collagenous domains (COL) of COL 1 to COL 5 (70, 168, 108, 224, 144 amino acid residues, respectively) counting from the carboxy-terminus, interspersed and flanked by 6 relatively short noncollagenous domains (NC), of NC 1 to NC 6. Based on the sequence, the calculated molecular mass of the chain without the signal peptide was 112,374 Da.

As found in polypeptides among the FACIT family, the COL domains contained some imperfections of the Gly-X-Y repeated structure; two of them in COL 5, four in COL 4, one in COL 3, and two in COL 1. However, the size of the COL 1 domain (70 residues), whose sequence was conserved with other NC 1 domains in the FACIT family, was rather small (Fig. 2). The NC 6 domain shared high sequence identity (22–29%) or sequence similarity (74–80%) with the amino-terminal NC domains of other FACIT family members: 29.4% identity with the NC 11 domain of



```
ACTCGCAGGGAGCTCACTCCTCGGCGGTGCCGCAGCCCTGTCCGGACTCCACTGCGCCTCTGAGGGGCTCAAATACGAATTCAAGATCCGTGGCCGTTCACATGGTTTCAAGGCACAATG  120
                                                                                                                      M   1

AGACTCACTGGCCCTTGGAAACTTTGGCTTTGGATGTCAATATTTCTGCTTCCTGCTTCCACTTCCGTGACCGTTAGGGACAAGACAGAAGAGTCATGCCCTATCCTGAGAATAGAGGGA  240
 R  L  T  G  P  W  K  L  W  L  W  M  S  I  F  L  L  P  A  S  T  S  V  T  V  R  D  K  T  E  E  S  C  P  I  L  R  I  E  G   41

CATCAGCTGACATATGACAACATAAACAAACTTGAAGTTTCAGGTTTTGATCTAGGAGACAGCTTTTCTCTAAGACGTGCATTTTGTGAAAGTGATAAAACCTGTTTCAAATTGGGAAGT  360
 H  Q  L  T  Y  D  N  I  N  K  L  E  V  S  G  F  D  L  G  D  S  F  S  L  R  R  A  F  C  E  S  D  K  T  C  F  K  L  G  S   81

GCACTTCTTATTAGAGACACTATTAAGATATTTCCCAAAGGCCTTCCTGAGGAGTACTCAGTAGCTGCCATGTTTCGAGTACGAAGAAACGCCAAAAAGGAACGGTGGTTTCTGTGGCAG  480
 A  L  L  I  R  D  T  I  K  I  F  P  K  G  L  P  E  E  Y  S  V  A  A  M  F  R  V  R  R  N  A  K  K  E  R  W  F  L  W  Q  121

GTTTTAAACCAGCAGAATATTCCACAGATTTCTATAGTAGTTGATGGTGGAAAGAAGGTGGTGGAATTTATGTTTCAAGCCACAGAGGGAGATGTGTTGAACTACATTTTTAGAAATCGA  600
 V  L  N  Q  Q  N  I  P  Q  I  S  I  V  V  D  G  G  K  K  V  V  E  F  M  F  Q  A  T  E  G  D  V  L  N  Y  I  F  R  N  R  161

GAACTCCGTCCTTTGTTTGATCGTCAGTGGCACAAACTTGGCATTAGTATACAATCCCAGGTCATTTCACTTTATATGGATTGTAATTTAATTGCGAGGAGGCAGACTGATGAAAAGGAC  720
 E  L  R  P  L  F  D  R  Q  W  H  K  L  G  I  S  I  Q  S  Q  V  I  S  L  Y  M  D  C  N  L  I  A  R  R  Q  T  D  E  K  D  201

ACTGTGGATTTCCATGGACGGACAGTTATTGCTACGCGAGCTTCAGATGGCAAGCCTGTGGATATTGAACTTCACCAACTTAAAATCTACTGCAGTGCAAACCTCATAGCTCAAGAAACA  840
 T  V  D  F  H  G  R  T  V  I  A  T  R  A  S  D  G  K  P  V  D  I  E  L  H  Q  L  K  I  Y  C  S  A  N  L  I  A  Q  E  T  241

TGTTGTGAAATATCAGATACTAAGTGCCCAGAGCAGGATGGCTTTGGAAATATTGCATCATCATGGGTAACTGCTCATGCCAGTAAAATGTCTTCATATCTGCCAGCAAAGCAGGAACTT  960
 C  C  E  I  S  D  T  K  C  P  E  Q  D  G  F  G  N  I  A  S  S  W  V  T  A  H  A  S  K  M  S  S  Y  L  P  A  K  Q  E  L  281

AAAGACCAGTGCCAGTGCATTCCAAACAAGGGAGAAGCAGGATTACCAGGAGCTCCGGGTTCACCTGGGCAGAAAGGGCATAAAGGAGAGCCGGGTGAAAATGGTTTACATGGTGCTCCA 1080
 K  D  Q  C  Q  C  I  P  N  K  G  E  A  G  L  P  G  A  P  G  S  P  G  Q  K  G  H  K  G  E  P  G  E  N  G  L  H  G  A  P  321

GGATTCCCTGGTCAAAAGGGAGAGCAAGGTTTTGAAGGCAGCAAAGGAGAAACTGGTGAAAAGGGTGAACAAGGAGAAAAAGGAGATCCAGCTCTGGCTGGCCTTAATGGAGAAAATGGT 1200
 G  F  P  G  Q  K  G  E  Q  G  F  E  G  S  K  G  E  T  G  E  K  G  E  Q  G  E  K  G  D  P  A  L  A  G  L  N  G  E  N  G  361

TTGAAAGGTGACTTGGGTCCTCATGGTCCACCTGGCCCAAAAGGAGAAAAGGGAGATACAGGACCCCCAGGACCACCAGCCTTACCTGGTTCCCTGGGGATACAAGGCCCCCAAGGTCCA 1320
 L  K  G  D  L  G  P  H  G  P  P  G  P  K  G  E  K  G  D  T  G  P  P  G  P  P  A  L  P  G  S  L  G  I  Q  G  P  Q  G  P  401

CCTGGAAAAGAGGGTCAGAGGGGAAGACGAGGGAAAACAGGACCTCCCGGAAAACCAGGACCCCCAGGACCACCTGGACCTCCTGGAATACAAGGAATACACCAAACTCTTGGTGGATAT 1440
 P  G  K  E  G  Q  R  G  R  R  G  K  T  G  P  P  G  K  P  G  P  P  G  P  P  G  P  P  G  I  Q  G  I  H  Q  T  L  G  G  Y  441

TATAACAAGGATAACAAGGGAAATGATGAACATGAAGCTGGAGGCCTGAAAGGAGACAAGGGTGAAACTGGACTACCAGGATTTCCAGGGTCTGTTGGCCCTAAAGGACAAAAAGGAGAA 1560
 Y  N  K  D  N  K  G  N  D  E  H  E  A  G  G  L  K  G  D  K  G  E  T  G  L  P  G  F  P  G  S  V  G  P  K  G  Q  K  G  E  481

CCTGGAGAGCCTTTTACAAAAGGAGAAAAAGGAGATAGAGGAGAACCTGGGGTAATAGGATCACAGGGAGTAAAGGGTGAACCTGGAGATCCCGGACCCCCTGGTTTAATAGGAAGCCCA 1680
 P  G  E  P  F  T  K  G  E  K  G  D  R  G  E  P  G  V  I  G  S  Q  G  V  K  G  E  P  G  D  P  G  P  P  G  L  I  G  S  P  521

GGACTAAAGGGTCAGCAAGGATCTGCAGGCTCCATGGGACCCAGAGGACCGCCAGGAGATGTTGGATTGCCAGGAGAACATGGTATCCCAGGAAAACAAGGCATTAAAGGAGAAAAGGGA 1800
 G  L  K  G  Q  Q  G  S  A  G  S  M  G  P  R  G  P  P  G  D  V  G  L  P  G  E  H  G  I  P  G  K  Q  G  I  K  G  E  K  G  561

GATCCAGGTGGGATCATAGGCCCTCCCGGGCTTCCAGGTCCAAAAGGTGAGGCTGGTCCTCCAGGGAAAAGCCTGCCAGGGGAACCAGGATTAGATGGAAATCCTGGAGCACCTGGTCCA 1920
 D  P  G  G  I  I  G  P  P  G  L  P  G  P  K  G  E  A  G  P  P  G  K  S  L  P  G  E  P  G  L  D  G  N  P  G  A  P  G  P  601

CGTGGGCCAAAGGGTGAAAGAGGACTTCCAGGTGTTCACGGTTCCCCAGGGGACATAGGCCCACAAGGGATAGGAATTCCTGGCAGAACAGGCGCCCAAGGACCAGCTGGAGAGCCAGGT 2040
 R  G  P  K  G  E  R  G  L  P  G  V  H  G  S  P  G  D  I  G  P  Q  G  I  G  I  P  G  R  T  G  A  Q  G  P  A  G  E  P  G  641

ATTCAGGGTCCTCGAGGTCTCCCTGGGTTGCCAGGAACTCCAGGGACTCCAGGGAATGATGGAGTTCCAGGGAGAGATGGAAAGCCAGGCCTGCCAGGCCCCCCAGGTGACCCGATTGCA 2160
 I  Q  G  P  R  G  L  P  G  L  P  G  T  P  G  T  P  G  N  D  G  V  P  G  R  D  G  K  P  G  L  P  G  P  P  G  D  P  I  A  681

CTTCCTCTCTTGGGAGACATCGGTGCTTTGCTCAAGAATTTCTGTGGCAACTGCCAAGCCAGTGTCCCAGGGCTGAAAAGCAACAAAGGAGAAGAAGGAGGTGCTGGTGAGCCTGGAAAG 2280
 L  P  L  L  G  D  I  G  A  L  L  K  N  F  C  G  N  C  Q  A  S  V  P  G  L  K  S  N  K  G  E  E  G  G  A  G  E  P  G  K  721

TATGATTCCATGGCCCGGAAGGGTGATATAGGGCCACGGGGTCCTCCAGGAATCCCAGGAAGAGAGGGACCAAAGGGAAGCAAAGGAGAGCGGGGCTACCCTGGGATACCTGGGGAGAAA 2400
 Y  D  S  M  A  R  K  G  D  I  G  P  R  G  P  P  G  I  P  G  R  E  G  P  K  G  S  K  G  E  R  G  Y  P  G  I  P  G  E  K  761

GGTGATGAGGGTCTTCAAGGAATTCCAGGCATTCCAGGTGCTCCAGGCCCGACTGGACCCCCTGGCTTAATGGGAAGAACTGGACATCCTGGTCCCACAGGAGCAAAAGGTGAAAAGGGC 2520
 G  D  E  G  L  Q  G  I  P  G  I  P  G  A  P  G  P  T  G  P  P  G  L  M  G  R  T  G  H  P  G  P  T  G  A  K  G  E  K  G  801

AGCGACGGACCCCCTGGGAAACCCGGACCACCTGGACCACCTGGTATTCCATTTAATGAACGAAACGGCATGAGCAGTTTATATAAAATTAAGGGAGGTGTGAATGTTCCCAGTTACCCA 2640
 S  D  G  P  P  G  K  P  G  P  P  G  P  P  G  I  P  F  N  E  R  N  G  M  S  S  L  Y  K  I  K  G  G  V  N  V  P  S  Y  P  841

GGGCCACCCGGTCCTCCTGGCCCAAAAGGCGATCCTGGCCCAGTGGGAGAGCCTGGTGCAATGGGGTTGCCAGGATTAGAAGGATTTCCAGGTGTAAAGGGAGATCGAGGCCCAGCAGGT 2760
 G  P  P  G  P  P  G  P  K  G  D  P  G  P  V  G  E  P  G  A  M  G  L  P  G  L  E  G  F  P  G  V  K  G  D  R  G  P  A  G  881

CCCCCAGGAATAGCAGGGATGTCGGGAAAACCTGGTGCCCCAGGGCCTCCAGGAGTTCCAGGGGAACCGGGTGAGAGAGGACCTGTTGGAGATATAGGTTTCCCTGGACCAGAAGGACCC 2880
 P  P  G  I  A  G  M  S  G  K  P  G  A  P  G  P  P  G  V  P  G  E  P  G  E  R  G  P  V  G  D  I  G  F  P  G  P  E  G  P  921

TCAGGAAAGCCAGGAATAAATGGAAAAGATGGAATACCAGGTGCTCAGGGCATCATGGGTAAGCCTGGAGACAGAGGCCCCAAAGGAGAACGTGGTGATCAGGGGATTCCAGGAGACAGA 3000
 S  G  K  P  G  I  N  G  K  D  G  I  P  G  A  Q  G  I  M  G  K  P  G  D  R  G  P  K  G  E  R  G  D  Q  G  I  P  G  D  R  961

GGCTCACAAGGTGAACGGGGAAAACCAGGCCTTACAGGCATGAAGGGGGCCATCGGTCCTATGGGTCCACCAGGAAACAAGGGCTCCATGGGATCCCCTGGCCACCAAGGCCCTCCAGGC 3120
 G  S  Q  G  E  R  G  K  P  G  L  T  G  M  K  G  A  I  G  P  M  G  P  P  G  N  K  G  S  M  G  S  P  G  H  Q  G  P  P  G 1001

TCTCCAGGCATCCCTGGCATTCCGGCTGATGCAGTTTCATTTGAAGAAATAAAGAAGTATATTAATCAAGAGGTCCTAAGGATTTTTGAAGAGAGGATGGCTGTATTCCTATCCCAGCTC 3240
 S  P  G  I  P  G  I  P  A  D  A  V  S  F  E  E  I  K  K  Y  I  N  Q  E  V  L  R  I  F  E  E  R  M  A  V  F  L  S  Q  L 1041

AAGCTGCCAGCAGCAATGTTGGCTGCCCAAGCTTATGGGAGACCTGGGCCACCAGGGAAGGATGGGTTGCCTGGGCCACCAGGAGACCCTGGACCCCAAGGCTACAGAGGACAGAAGGGA 3360
 K  L  P  A  A  M  L  A  A  Q  A  Y  G  R  P  G  P  P  G  K  D  G  L  P  G  P  P  G  D  P  G  P  Q  G  Y  R  G  Q  K  G 1081

GAAAGAGGTGAACCTGGAATTGGGCTGCCAGGGAGTCCAGGTCTTCCTGGGACTTCAGCTCTGGGTTTGCCAGGCTCACCAGGTGCCCCAGGCCCCCAGGGCCCCCCAGGACCCAGTGGA 3480
 E  R  G  E  P  G  I  G  L  P  G  S  P  G  L  P  G  T  S  A  L  G  L  P  G  S  P  G  A  P  G  P  Q  G  P  P  G  P  S  G 1121

AGATGTAACCCAGAAGATTGCCTCTATCCTGTGTCTCATGCCCATCAGCGCACAGGTGGAATTGAACACACCTGAAGAAGACTTAGTTCCTGGTAACATTTCCTTGGACATGGAGCTCT 3600
 R  C  N  P  E  D  C  L  Y  P  V  S  H  A  H  Q  R  T  G  G  N  *                                                       1142

CTTAATACCGTCAAACCCTCATCATCTGTGGGTTGCTTTTTTTTTTTTTTTTTTTTTGGGGAGTAAGCCAGGCATTAAAAGCAACCGTTTGAATCCTATTCCTGTGACAATTGCAAATCA 3720
```

Fig. 2

the human α1(XVI) collagen (*7*), 28.6% identity with the NC 4 domain of the human α1(IX) collagen (*19*), 23.5% identity with a segment in the carboxy-terminus of the NC 3 domain from the chicken α1(XII) collagen (*20*), and 22.4% identity with a segment in the carboxy-terminus of the NC 3 domain from the chicken α1(XIV) collagen (*21*) (Fig. 3). There was no von-Willebrand factor A-like domain or fibronectin type III-like domain in NC 6. Fourteen cysteinyl residues, 10 of them in NC 6, 2 in NC 4, and 2 in the COL 1/NC 1 junction, were found in the chain. The relative locations of four of the cysteinyl residues in NC 6 and of the 2 residues in the COL 1/NC 1 junction separated by 4 amino acid residues are conserved with those of other FACIT family members. No potential attachment site for asparagine-linked carbohydrates or Arg-Gly-Asp sequence was found (Fig. 2).

Northern Blotting Analysis—As shown in Fig. 4, Northern blot analysis with poly A^+ RNA from the rhabdomyosarcoma cells demonstrated that the major transcript of the gene was longer than 10 kb in size (lanes 1-6). The signals were less intense than the one for the α1(III) gene transcript, which is the major collagen species synthesized by this cell line (lane 11). It is of interest that the Northern blot showed different hybridization patterns according to the location of the probes. Nine major species (>10, 6.4, 5.0, 4.2, 3.4, 2.7, 2.0, 1.3, and 0.9 kb in size) of transcripts appeared with the 5′ *Eco*RI fragment of KI 6 as a probe (lane 1), and seven species (>10, 6.4, 5.0, 4.2, 3.4, 2.7, and 2.0 kb in size) with HY 67 (lane 2). Smaller transcripts disappeared when we used 3′ cDNA fragment probes, and only one species (>10 kb) of mRNA was detected with the 3′-*Hin*dIII fragment of KI 65 (lane 6). This pattern did not change when we used higher stringency conditions for hybridization and washing, suggesting that alternative splicing, which is often seen in some collagen genes (*22, 23*), or an incomplete splicing event might occur in the cells.

In order to examine whether the 3′ portions of KI 1 and KI 18 [(7) and (8), respectively, indicated by the dotted line in Fig. 1] are derived from the primary transcript of the gene or are artifacts, we carried Northern blot analysis further. As can be seen in Fig. 4, certain mRNA species were detected by these fragments. The longest RNA species (>10 kb), which was already recognized as a faint band in lanes 1-5, hybridized with the 3′ *Apa*I/*Eco*RI fragment of KI 1 (lane 7). Further, 4.2 kb mRNA was detected with the 3′ *Sau*96I/*Eco*RI fragment of KI 18 (lane 8), 4.2 and 2.7 kb species were seen with the *Eco*RI fragment of KI 45 (lane 9), and 6.4 and 4.2 kb species were found with the *Hin*dIII/*Sma*I fragment of KI 51 (lane 10). Thus, the 3′ portions of these clones detected RNA species, indicating that they are not cDNA cloning artifacts but parts of real transcripts. Furthermore, the RNA species detected by the 3′ fragments of cDNAs represent one or a

```
                 ●
α1(XIX)       34:CPILRIEGHQLTYDNINKLEVSGFDLGDSFSL-RR-AF--CESDK-T-CFKLGSALLIRD
α1(IX)        34:CP-KIRIG----QDDLPGFDLISQFQVDKAASRRAIQRVVGSATLQVAYKLGNNVDFRIP
α1(XII)     2512:CPLIYLEG--YTSPGFKMLESYNL-TEKHFASVQGVSLESGSFPSYVAYRLHKNAFVSQP
α1(XIV)     1227:CPLVFKDG--DNFAGFKMMEMFGL-VEKEFSAIDGVSMEPGTFNVYPCYRLHRDALVSQP
α1(XVI)       28:CPPSQQEGLKLEHSSSLPANVTGFNLIHRLSLMKKSAIKKIRNPKGPLILRLGAAPVTQP
                 **     *

α1(XIX)          TIKIFPKGLPEEYSVAAMFRVRRNAKKERWFLWQVLNQQNIPQISIVVDGGKKVVEFMFQ
α1(IX)           TRNLYPSGLPEEYSFLTTFRMTGSTLKKNWNIWQIQDSSGKEQVGIKINGQTQSVVFSYK
α1(XII)          IREIHPEGLPQAYTIIMLFRLLPESPSEPFAIWQITDRDYKPQVGVVLDPGSKVLSFFNK
α1(XIV)          TKYLHPEGLPSDYTITFLFRILPDTPQEPFALWEILNEQYEPLVGVILDNGGKTLTFFNY
α1(XVI)          TRRVFPRGLPEEFALVLTLLLKKHTHQKTWYLFQVTDANGYPQMSLEVNSQERSLELA-Q
                      * ***
                                                          ●
α1(XIX)          ATEGDVLNYIFRNRELRPLFDRQWHKLGISIQSQVISLYMDCNLIARRQTDEKDTVDFHG
α1(IX)           GLDGSLQTAAF--SNLSSLFDSQWHKIMIGVERSSATLFVDCNRIESLPIKPRGPIDIDG
α1(XII)          DTRGEVQTVTFDNDEVKKIFYGSFHKVHIVVTSSNVKIYIDCSEILEKPIKEAGNITTDG
α1(XIV)          DYKGDFQTVTFEGPEIRKIFYGSFHKLHVVISKTTAKIIIDCKEAGEKTINAAGNISSDG
α1(XVI)          GQDGDFVSCIF--P-VPQLFDLRWHKLMLSVAGRVASVHVDCSSASSQPLGPRRPMPV-G
                    *      *        *    **              **                 *
                                             ●         ●
α1(XIX)          RTVIATRA-SDG-KPVDIEL-HQ-LKIYCSANLIAQETCCEISDTKCPEQDGFGNIASSW
α1(IX)           ---FAVLGKLADNPQVSVPFELQWMLIHCDPLRPRRETCHELPARITPSQTT
α1(XII)          YEILGKLL--KGDR-RSATLEIQNFDIVCSPVWTSRDRCCDLPSMR-DEAKC
α1(XIV)          IEVLGRMVRSRGPRDNSAPLQLQMFDIVCATSWANRDKCCELPGLR-DEENC
α1(XVI)          HVFLGLDA-EQG-KPVSFDL-QQ-VHIYCDPELVLEEGCCEILPAGCP-PET-SK-A-RR
                                       *   * *         *

α1(XIX)          VTAHASKMSSYLP-AKQELKDQCQC
α1(IX)
α1(XII)
α1(XIV)
α1(XVI)          -DTQSNELIEINPQSEGKVYTRCFC
```

Fig. 3. **Alignment of amino acid sequences of the NC 6 domain from the α1(XIX) collagen chain (position 1) with the NC 4-like domains of the other FACIT family.** Positions 2, 3, 4, and 5, are the human α1(IX) NC 4 (*19*), the chicken α1(XII) NC 3 (*20*), the chicken α1(XIV) NC 3 (*21*), and the human α1(XVI)NC 11 (*7*) domains, respectively. Gaps (indicated by -) are introduced to obtain the highest homology among the sequences. The asterisks indicate the conserved amino acids in all 5 chains. Location of the four cysteinyl residues conserved in position in all 5 chains is indicated by closed circles. Predicted β-strand areas are shaded (*32*).

Fig. 2. **Nucleotide sequence of cDNA and deduced amino acid sequence of α1(XIX) collagen chain.** The overlapping cDNAs, as shown in Fig. 1, correspond to a 3,720-bp mRNA that contains 117-bp 5′ and 177-bp 3′ untranslated regions and a 3,495-bp protein-coding region. The putative signal peptide (amino acid number 1 to 23) is indicated by an underline. Fourteen cysteinyl residues, the place of insertion of KI 18, and the termination codon TGA are indicated by circled Cs, a wedge, and a small star, respectively. Collagenous Gly-X-Y repeating sequence is shaded. The beginnings of the sequences which differ from that of "proto" type in KI 30, KI 25/KI 45/KI 49, KI 1, KI 18, and KI 40/KI 51 are indicated by arrows with the circled numbers 1, 2, 3, 4, and 5, respectively (see Figs. 1 and 5). When cDNA sequence and genomic DNA sequences were compared, several exon boundaries (indicated by arrowheads) were identified.

few of the multiple RNA transcripts seen in lanes 1-3 in Fig. 4.

Identification of Genomic Clones and Partial Characterization of Exon/Intron Structure—We thus confirmed that the 3′ fragments of these clones just mentioned above were really present in mature mRNA by Northern blotting analysis. To define the exon/intron structure in the gene and to determine whether there is some relationship between the unique cDNAs and intron sequences, we isolated several human genomic clones. For this purpose, we used 3 cDNA clones (KI 18, KI 40, and KI 51) as probes.

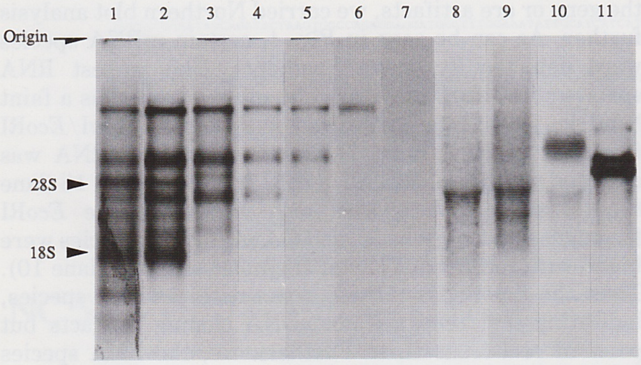


Fig. 4. **Northern blot analysis of the α1(XIX) mRNA.** Samples (3 μg/lane in lanes 1-10, 1 μg/lane in lane 11) of mRNA isolated from human rhabdomyosarcoma cells were electrophoresed on 0.8% agarose gels, blotted onto nylon filters, and hybridized with cDNA probes (lanes 1-10). 5′ *Eco*RI fragment of KI 6 [indicated by "(1)" in Fig. 1], whole HY 67 fragment ["(2)" in Fig. 1], *Eco*RI fragment of KI 1/KI 51 ["(3)"], *Eco*RI/*Bam*HI fragment of KI 40/KI 51 ["(4)"], *Bam*HI fragment of KI 51 ["(5)"], and 3′ *Hind*III fragment of KI 65 ["(6)"], were used as probes for hybridization in lanes 1, 2, 3, 4, 5, and 6, respectively (see Fig. 1). The 3′ fragments cDNAs whose sequences differed from the "proto" type in KI 1 [shown by "(7)" in Fig. 1], KI 18 ["(8)" in Fig. 1], KI 45"(9)", and KI 40/KI 51"(10)" were used as probes as well for lanes 7, 8, 9, and 10, respectively. As a control, cDNA encoding α1(III) was used as a probe for hybridization with the same RNA (1.0 μg, lane 11). Note that the largest transcript is >10 kb in lanes 1-6 and that the number of bands decreases as the probes go in the 3′ direction, disappearing from smaller RNA species.

Three overlapping genomic clones (KIG 12, KIG 23, and KIG 13) that span more than 25 kb of the human gene were isolated with the 3′ fragments of KI 18 and KI 40. The genomic clones were characterized by restriction mapping and Southern blot analysis. Relevant segments of the genomic clones were subcloned and sequenced to determine the exon/intron structure. Five consecutive exons covering 448 bp in the positions of the carboxy-terminal portion of COL 2, NC 2, and amino-terminal portion of COL 1 were sequenced (Fig. 5). The size of the 5 exons, named +1 to +5 from the 5′ end, varied from 36 bp to 171 bp, which feature is distinct from that of fibrillar collagen genes, but similar to that of the FACIT family. Of interest was that the sequence of the 3′ portion (640 bp) of KI 18 was found between exons +2 and +3 (indicated exon 3′ in Fig. 5). The region is a portion of the intron when spliced to form the "proto" type of the transcript. Within exon +3′ there is a polyadenylation signal, of AATAAA (indicated by the short bar in Fig. 5), which might be utilized for poly A tailing. Similar splicing events were observed in KI 40 and KI 51. Namely, the sequences of 3′ fragments of KI 40 and KI 51 were found downstream of exon +5 (+6′ and +6′α in Fig. 5). The segment of exon +6′ contained a polyadenylation signal, which was utilized to add a poly A tail to KI 40. The nucleotide sequence of KI 51 indicated that the presence of three polyadenylation signals, but they are not utilized, and the cDNA extended to farther in the 3′ direction. The exon/intron boundaries were analyzed from the sequence of genomic fragments as shown in Table I. The GT-AG rules were conserved in exons +1 to +5, but the donor sites of "exon" +3′, and +6′ (or +6′α) were changed (Table I).

Analysis of 5′ Region of mRNA and 5′ Flanking Region of the Gene—To determine the 5′ end structure of the mRNA, we performed a 5′ RACE experiment as mentioned under "MATERIALS AND METHODS." The PCR products were subcloned and the inserts were sequenced. As shown in Fig. 6, the sequence of five inserts out of seven started 25 bp downstream of the 5′ end of KI 6, one from 22 bp downstream, and one from 41 bp downstream. Thus, the tentative start site of the major transcripts is indicated as +1 in Fig. 6. Furthermore, to obtain the 5′ flanking region of the gene, we screened a total genomic DNA library

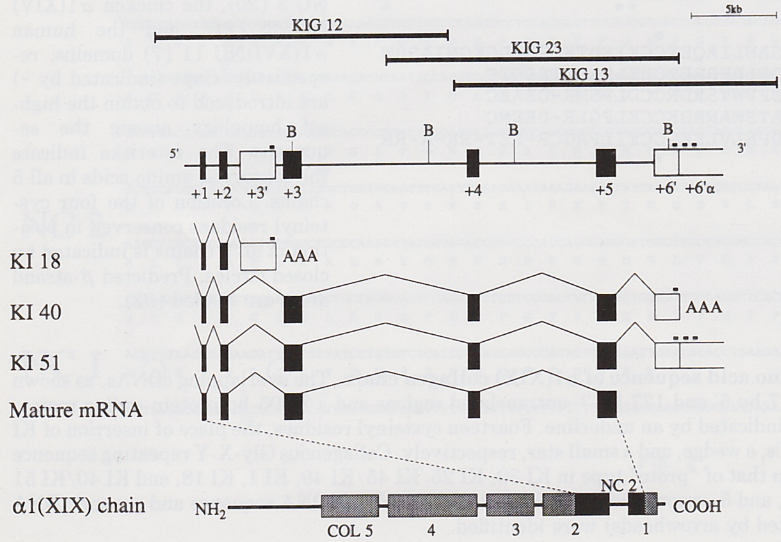


Fig. 5. **Partial characterization of alternative splicing pattern of α1(XIX) mRNA.** Top: Three genomic clones, KIG 12, KIG 23, and KIG 13, spanning more than 25 kb, are shown with a kb scale. Middle: Schematic pattern of exon/intron organization of the gene. Location of *Bam*HI sites is indicated by Bs. The five exons (tentatively named +1, +2, +3, +4, and +5, counting from 5′ to 3′), and two alternative "exons" (+3′ and +6′ or +6′α, revealed by cDNAs KI 18, KI 40, and KI 51) exist in the region. Exons are indicated by closed boxes; and alternative exons, by open boxes. Note that four different mRNAs are generated by alternative splicing. Bottom: Schematic representation of 5 exons encoding a carboxyl part of the COL 2 domain, the entire NC 2 domain, and the amino part of the COL 1 domain of α1(XIX) chain.

TABLE I. **Exon-intron boundaries of the five exons and two alternative exons of the human α1(XIX) gene.** When RNA transcripts are spliced as the cDNAs, KI 18 and KI 40 (in Fig. 5), the translation products will stop immediately at the beginning of the exons, +3′ and +6′, respectively. Location of the stop codon (TAG)s is indicated by underlines.

exon No. (nt)	exon-intron boundaries		
+1 (36)	tttatag	2893 GGAATAAATGGAAAAGATGGAATACCAGGTGCTCAG 2928	gtatggg
+2 (45)	ttttcag	2929 GGCATCATGGGTAAGCCTGGAGACAGAGGCCCCAAAGGAGAACGT 2973	gtatgta
+3′ (640)	ccactag	ATGCAGTAGCGCCACCATCCCTCCAATTGTGACAACCAAAAGTGTCTCCA GACATCACCAAATGTCTCCTGGGAAGCAAAATCACTGCAATTGAGAACTA CTGACATGCAGAAAAACTAAGTAGAGTAAAGAAGGTGGTGTGAGGGGTTA GAGAAATGGAAAGTGTAACTGAAAGTATAGTGAAGTTACTAGTAATGCCA ATTATGTTATATAGTTTTGTGTTGTTTGGAGGGTTAATGTCATTATTGCA ATTTCTCTTTATTTGGTTTAATGAACCTCAACATTTTGTAGTAGAATTGA TTTCTACTAGAATTAACTTCTACTAGAATTAATTTCCACTAGAATTAATA GAATTAATTTCTAGTAGAGTCATGCCACTGCAAGAAGTTACTCAAACATT GTTACTCAGTGTAATTCTAGTAGAGTCATGCCACTGCAAGAAGTTACTCA AACATTGTTTACTCAGTGTAATTCTATTAAGGGTTGCTATGTATGTGAGT GTGGGCACATCTGTTTGTATACTGGTGTCTAGTTTGCTAACTTTTATTTA TCATTATACAAGTAATCTATAATAACTGAAGGAAATGTGAAAATACAGAT GATGTAAATTTAAAAGGAAATAAATCTACCTTACCCCACT	acacaga
+3 (171)	ttaacag	2974 GGTGATCAGGGGATTCCAGGAGACAGAGGCTCACAAGGTGAACGGGGAAA ACCAGGCCTTACAGGCATGAAGGGGGCCATCGGTCCTATGGGTCCACCAG GAAACAAGGGCTCCATGGGATCCCCTGGCCACCAAGGCCCTCCAGGCTCT CCAGGCATCCCTGGCATTCCG 3144	gtaagta
+4 (67)	tcttcag	3145 GCTGATGCAGTTTCATTTGAAGAAATAAAGAAGTATATTAATCAAGAGGT CCTAAGGATTTTTGAAG 3211	gttagat
+5 (129)	catgaag	3212 AGAGGATGGCTGTATTCCTATCCCAGCTCAAGCTGCCAGCAGCAATGTTG GCTGCCCAAGCTTATGGGAGACCTGGGCCACCAGGGAAGGATGGGTTGCC TGGGCCACCAGGAGACCCTGGACCCCAAG 3340	gtaagtc
+6′ (436)	tgagaag	CCGGAGACGTATAGAAAGCAGTGAGAAAATCGACGTCAGACTGTGAGAGG TACACAATTCATATTGAAAAAAGGAAGAGCCGCTCCTACCAAGATTCTGA TCATGAAAGCTTGAGGATCCGCCTTCAGCACCCTGACCTTCTTAAAGTGA AAAGAAAGACCTTTTGGATTCTGAGTTGCATCACTTCATTTGCAGTTTTT CTAATGTTTCCAAAGGGGTGATTGTCGATGTCTTTTTTAAATTGAAAACA TTTCCAGACAGCTTTTAGATTTTAGCCGTCCCATGGGGTTCTGAGTATTG GCTTCCTGCTGTGCCTTGAGAGAACTTTCCTTGGTGTCCATTCCCACCTG GGGCTTGCTTGCAGCACAGGGCCTCATGGCTGGTTCTCCCGGGTATGAGC TAAAATGTGCCACATCAGAATAAAAGAAACCCCAAG	aagcatg

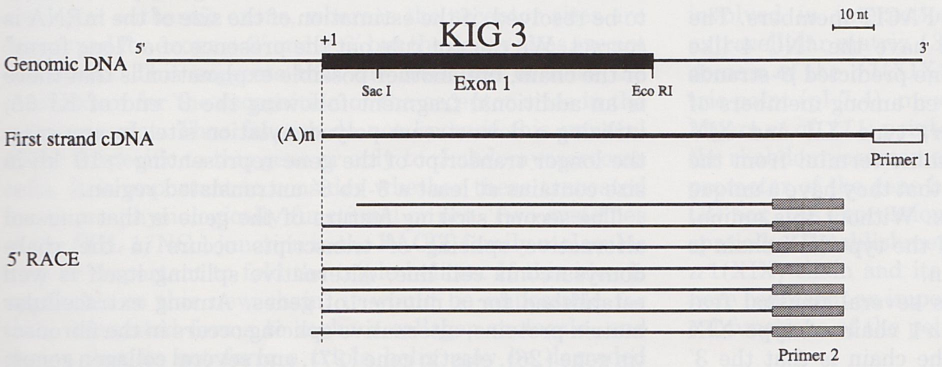


Fig. 6. **Analysis of the 5′ region of α1(XIX) mRNA.** Relative location of 5′ RACE products to the corresponding region of genomic DNA, KIG 3. First strand cDNA was synthesized from primer 1 and (dA)n was added to the 3′ end. Aliquots of the sample were used for amplification between hybrid/acceptor primer and primer 2. RACE products were cloned and sequenced. Relative location of 5′ RACE products is shown in the figure in relation to genomic DNA, KIG 3.

using the 5′ *Eco*RI fragment of KI 6 as the probe. A positive clone, KIG 3, was isolated; and a 643- bp *Pst*I/*Eco*RI fragment that hybridized with the 5′ fragment of KI 6 was subcloned and sequenced (Fig. 7). The sequence of the genomic fragment was consistent with that of the 5′ end of KI 6 and of the RACE products. The nucleotide sequence immediately upstream of the 5′ end of the major transcripts from RACE revealed a TATA-like element (TAATAA) located at −31, where the TATA box is expected to be. One CCAAT box and a potential AP1 site were found farther upstream at −293 and −378, respectively (Fig. 7).

```
-561                                   CTGCAGGAAAATGACACTTTC   -541
-540   ACATATTCAAAGGGATCCCAAAAGAGTTCAAAAGCCAAGACTGGAAGCACCAACCCATGG   -481
-480   AAAGATGTATTTAAAATCACAAAATACTTCTGTGGACACTGAGAAAAACCAAGCCACACA   -421
-420   AACGTTTACATATGAGCGTGATTAGAAGGTGGGCAGACTGCCTTAATCAACTTTATATAA   -361
-360   ACTCTTGACATAAAAGTGAAAAAATGAAACATAAAGATATATTTTCTTAGCTCAACCCGT   -301
-300   TAAAAAGCCAATTACCTACTCGTGAACGTTTGCCACATACAATAATACTTTGAAGACACT   -241
-240   TCATGATCAGAAGTGAGAAACTCTAGGCATTTGCATTCAACCGTCATTTTATCTGTGAAA   -181
-180   CTGCTAGTAAAACATAAACCTGAAAGGCATTACTCCCAGCTCTAAGGCGACATCGCTGTC   -121
-120   ATTAAAGAAAACCTGTGCCTGAGTTGGCTGAGTACCTTCTTGCAGAGTCCTCTGTCCTCG    -61
 -60   CGGAGTGGGAGGGTCACACTGGGAGAGACTAATAAGGGCAGAGATGCGTCCCCCTTCCCC     -1
   1   ACTCGCAGGGAGCTCACTCCTCGGCGGTGCCGCAGCCCTGTCCGGACTCCACTGCGCCTC     60
  61   TGAGGGGCTCAAATACGAATTC                                            82
```

Fig. 7. **Nucleotide sequence of the 5′ region of COL19A1.** A genomic DNA fragment of KIG 3 *Pst*I/*Eco*RI was sequenced. Based on the results of the 5′ RACE experiment (Fig. 6), the major transcriptional start site (+1) was identified. A TATA-like element, a CAT box, and a potential AP 1 site are indicated by a small box, double underline, and single underline, respectively.

DISCUSSION

The selective cDNA clones reported here (KI 6, 5′ fragments of KI 18 and KI 40, and KI 50) and the HY 67 clone encode a "proto" type of new collagen chain beginning with a methionyl residue followed by a hydrophobic signal peptide, suggesting that the predicted protein is secreted into the extracellular matrix. The primary structure indicated the presence of five COL domains interspersed by four rather short NC domains, with short imperfections in the COL domains, and four and two conserved cysteines in NC 6 and at the COL 1/NC 1 junction, respectively. These are the characteristic features of the FACIT family (*10*). The structural characteristics of the putative chain are significantly different from those of the 18 distinct collagen types that contain collagen triple helices. Therefore we designated this newly discovered chain derived from cDNAs α1(XIX) collagen chain. As shown in Fig. 8, this chain is rather small in size in the FACIT family. One of the characteristic features of the FACIT family is the COL 1 domain and the two cysteinyl residues at the junction of COL 1/NC 1. The type XIX collagen chain contains a similar arrangement, although the COL 1 domain is somewhat shorter than that of the other FACIT members. The second feature is that all of them have the "NC 4-like domain." As shown in Fig. 3, the nine predicted β-strands of the NC-module are well-conserved among members of the FACIT family (*32*). However, type XII and XIV collagen chains contain different amino-termini from the other FACIT molecules, suggesting that they have a unique function in the extracellular matrix. Without this amino-terminal sequence, the structure of the type XIX chain is similar to that of the α1(XVI) chain.

The present study demonstrates several unusual features of the human gene for the α1 chain of type XIX collagen. One striking feature of the chain is that the 3′ untranslated region of the transcript is extremely long. Myers *et al.* suggested that the chain might be uncommonly large, as estimated from the size of mRNAs from rhabdomysarcoma cells and fibroblasts (*11*). Contrary to their suggestion, the size of the chain is similar to that of type IX chain, and very much shorter than that of the type XII. There are two possibilities to explain the discrepancy in length between the predicted peptide and mRNA. One possibility is the existence of a "long form" of the chain. In the α1 chain of type XII, there are two forms that share the same signal peptide and the same carboxy-terminus, but a fragment of the carboxy-terminus of the signal peptide, approximately one thousand amino acids is missing in the short form (*24, 25*). We tried to detect a "long form" of the mRNA. However, we could not find a clone encoding a "long form." Another possibility is the existence of a long 5′ or 3′ untranslated region in the transcript of the gene. To determine the 5′ or 3′ end, we employed the RACE procedure. Five of seven 5′ RACE products stopped 25 bp downstream from the 5′ end of KI 6, where we considered the major transcriptional start site to lie. On the other hand, we failed to obtain 3′ RACE products. Meanwhile, in screening of the cDNA library using a 3′ fragment of KI 50 as the probe, we isolated an overlapping clone, KI 65, extending in the 3′ direction with a poly A tail. Northern blotting analysis using a 3′ fragment of KI 65 demonstrated a single RNA band of longer than 10 kb. These data suggested that the transcript of the gene contained a long untranslated region at its 3′ end, spanning more than 5 kb. Some size difference between the length (>9 kb) shown by overlapping cDNA clones here and the size (>10 kb) of mRNA estimated from Northern blotting analysis remains to be resolved, if the estimation of the size of the mRNA is correct. We can not rule out the presence of a "long form" of the chain, but another possible explanation is that there is an additional fragment following the 3′ end of KI 65, utilizing a downstream polyadenylation site. In any case, the longer transcript of the gene representing >10 kb in size contains at least a 5-kb 3′ untranslated region.

The second striking feature of the gene is that unusual alternative splicing of transcripts occurs in the rhabdomysarcoma cell line. Alternative splicing itself is well established for a number of genes. Among extracellular matrix proteins, alternative splicing occurs in the fibronectin gene (*26*), elastin gene (*27*), and several collagen genes: type VI and XIII (*22, 23*). The alternative splicing events in the genes of the extracellular matrix components could be involved in changing specific functions such as cell-cell or cell-matrix binding. Usually, variant transcripts are made by skipping some exons (*22, 23*) or utilizing alternative promoters (*28-30*). In this gene, the manner of alternative splicing is different. The variant forms with the

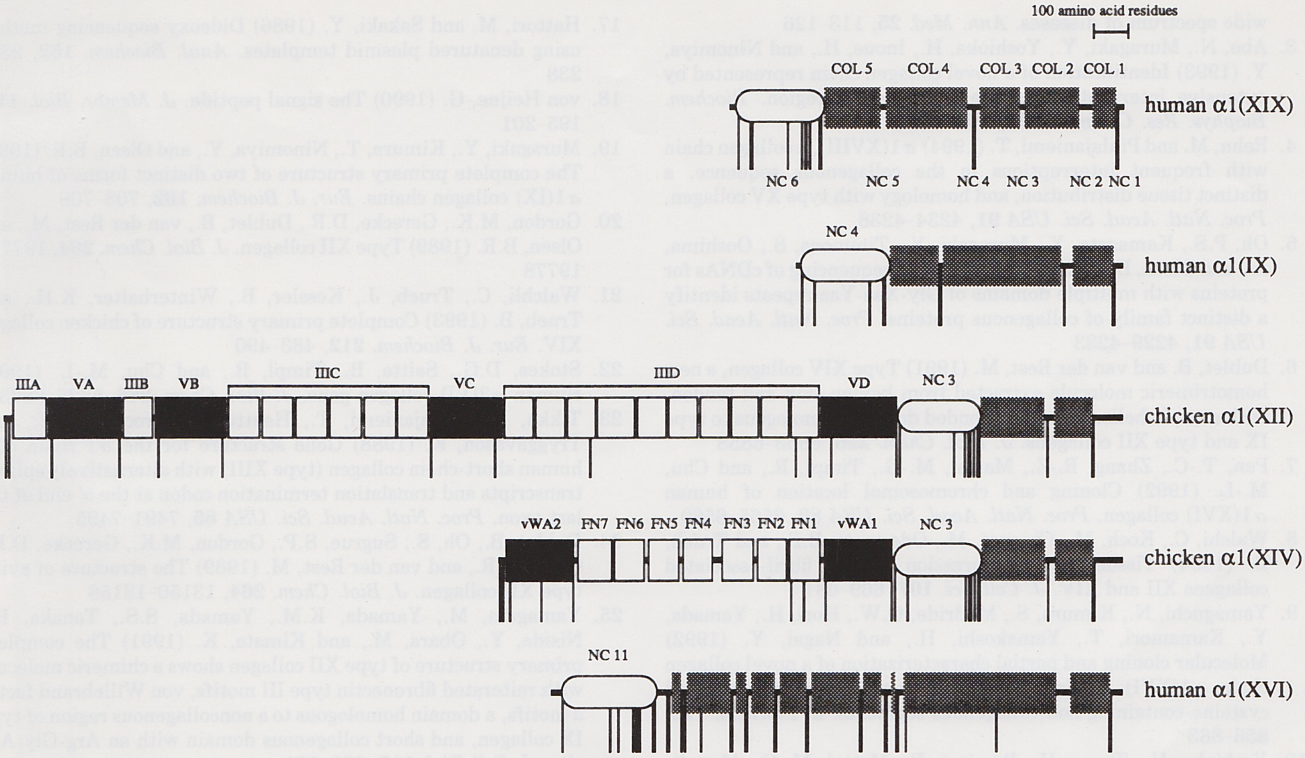


Fig. 8. **Schematic comparison of polypeptide structures of α1(XIX) collagen chain and other FACIT family.** The schematics are drawn to scale, and the sequences of the other FACIT family are taken from recent reports (*7, 19-21*). Collagenous domains, noncollagenous domains, NC 4-like domains [thrombospondin-like domains (*31, 32*)], von Willebrand factor A-like domain, fibronectin type III-like domains, and cysteinyl residues are indicated by shaded squares, horizontal bars, ellipses, black boxes, white boxes, and vertical bars, respectively.

different sizes showed common 5′ sides but different 3′ ends. Partial characterization of genomic clones revealed the mechanism of the occurrence of the variant forms. Of the variant forms we examined (KI 18, KI 40, and KI 51), the +3′ and +6′, +6′α exons were within an intron that is usually skipped in the "proto" type of the splicing. These exons have stop codons near the 5′ ends, which should result in truncated forms of the peptide, although we have not detected such polypeptides. The consensus sequence of the donor or acceptor site of the exon/intron junction is conserved at authentic exons, whereas the acceptor sites are conserved for exons +3′ and +6′ but the donor sites are not conserved. This suggests that the cells lose or miss the mechanism for the recognition of the acceptor sites in the splicing event. Thus far, we do not know if this unusual event is specific to this gene or only to rhabdomyosarcoma cells. It is important to consider whether these truncated forms are physiologically functional or not. Two cysteines at the COL 1/NC 1 junction in the FACIT family molecules may be important to fold the triple helices. If this idea is applicable to this new chain, it would be impossible for these truncated forms to make a collagen molecule, and chains that do not form a triple helix may be degraded within the cell. Alternatively, once truncated forms of the peptide successfully form a molecule, it may have a different function from the "proto" type peptide.

It has been shown that type IX collagen is localized on the surface of collagen fibrils in cartilage and may serve as molecular bridges binding the extracellular components (*1*). The structural similarities between the XIX collagen and the FACIT family raise the possibility that the XIX collagen may have similar functions. The NC 6 domain in the α1(XIX) collagen chain has sequence homology with the NC 4-like domains of the other FACIT members. The homologous region has been called the thrombospondin 1 (tsp 1) module by Bork (*31*). It contains a common antiparallel β-sheet structure composed of nine consensus β-strands, which could be involved in molecular recognition (*32*). The NC 4 domain of the α1(IX) collagen chain has a basic isoelectric point pI of ~10 and is thus thought to be involved in interaction with acidic components in the extracellular matrix (*33*). However, the homologous NC 6 domain of the α1(XIX) chain, which has a neutral isoelectric point (pI 7.4), may function in a different manner. As Myers *et al.* (*11*) pointed out, the expression of the gene in the rhabdomyosarcoma cell line is extremely low. Since the promoter of the gene has a TATA-like box at −31 and no GC box, the expression could be restricted temporally and spatially. Establishment of the primary structure of the α1(XIX) chain and its unique gene expression presented here represent an important step in elucidating the biological functions of the polypeptide.

The authors thank A. Fukutomi for excellent technical assistance, and M. Kohmoto for typing the manuscript.

REFERENCES

1. van der Rest, M. and Garrone, R. (1991) Collagen family of proteins. *FASEB J.* **5**, 2814-2823
2. Kivirikko, K.I. (1993) Collagens and their abnormalities in a

wide spectrum of diseases. *Ann. Med.* **25**, 113-126

3. Abe, N., Muragaki, Y., Yoshioka, H., Inoue, H., and Ninomiya, Y. (1993) Identification of a novel collagen chain represented by extensive interruptions in the triple-helical region. *Biochem. Biophys. Res. Commun.* **196**, 576-582
4. Rehn, M. and Pihlajaniemi, T. (1994) α1(XVIII), a collagen chain with frequent interruptions in the collagenous sequence, a distinct tissue distribution, and homology with type XV collagen. *Proc. Natl. Acad. Sci. USA* **91**, 4234-4238
5. Oh, P.S., Kamagata, Y., Muragaki, Y., Timmons, S., Ooshima, A., and Olsen, B.R. (1994) Isolation and sequencing of cDNAs for proteins with multiple domains of Gly-Xaa-Yaa repeats identify a distinct family of collagenous proteins. *Proc. Natl. Acad. Sci. USA* **91**, 4229-4233
6. Dublet, B. and van der Rest, M. (1991) Type XIV collagen, a new homotrimeric molecule extracted from bovine skin and tendon, with a triple helical disulfide-bonded domain homologous to type IX and type XII collagens. *J. Biol. Chem.* **226**, 6853-6858
7. Pan, T.-C., Zhang, R.-Z., Mattei, M.-G., Timpl, R., and Chu, M.-L. (1992) Cloning and chromosomal location of human α1(XVI) collagen. *Proc. Natl. Acad. Sci. USA* **89**, 6565-6569
8. Walchi, C., Koch, M., Chiquet, M., Odermatt, B.F., and Trueb, B. (1994) Tissue-specific expression of the fibril-associated collagens XII and XIV. *J. Cell Sci.* **107**, 669-681
9. Yamaguchi, N., Kimura, S., McBride, O.W., Hori, H., Yamada, Y., Kamamori, T., Yamakoshi, H., and Nagai, Y. (1992) Molecular cloning and partial characterization of a novel collagen chain, α1(XVI), consisting of repetitive collagenous domains and cysteine-containing non-collagenous segments. *J. Biochem.* **112**, 856-863
10. Yoshioka, H., Zhang, H., Ramirez, F., Mattei, M.-G., Moradi-Ameli, M., van der Rest, M., and Gordon, M.K. (1992) Synteny between the loci for a novel FACIT-like collagen locus (D6S228E) and α1(IX) collagen (COL9A1) on 6q12-q14 in humans. *Genomics* **13**, 884-886
11. Myers, J.C., Sun, M.J., D'Ippolito, J.A., Jabs, E.W., Neilson, E.G., and Dion, A.S. (1993) Human cDNA clones transcribed from an unusually high-molecular-weight RNA encode a new collagen chain. *Gene* (*Amst.*) **123**, 211-217
12. Kimura, T., Mattei, M.-G., Stevens, J.W., Goldring, M.B., Ninomiya, Y., and Olsen, B.R. (1989) Molecular cloning of rat and human type IX collagen cDNA and localization of the α1(IX) gene on the human chromosome 6. *Eur. J. Biochem.* **179**, 71-78
13. Oh, S.P., Taylor, R.W., Gerecke, D.R., Rochelle, J.M., Seldin, M.F., and Olsen, B.R. (1992) The mouse α1(XII) and human α1(XII)-like collagen genes are localized on mouse chromosome 9 and human chromosome 6. *Genomics* **14**, 225-231
14. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1989) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
15. Frohman, M.A., Dush, M.K., and Martin, G.R. (1988) Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85**, 8998-9002
16. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467
17. Hattori, M. and Sakaki, Y. (1986) Dideoxy sequencing method using denatured plasmid templates. *Anal. Biochem.* **152**, 232-238
18. von Heijne, G. (1990) The signal peptide. *J. Membr. Biol.* **115**, 195-201
19. Muragaki, Y., Kimura, T., Ninomiya, Y., and Olsen, B.R. (1990) The complete primary structure of two distinct forms of human α1(IX) collagen chains. *Eur. J. Biochem.* **192**, 703-708
20. Gordon, M.K., Gerecke, D.R., Dublet, B., van der Rest, M., and Olsen, B.R. (1989) Type XII collagen. *J. Biol. Chem.* **264**, 19772-19778
21. Walchli, C., Trueb, J., Kessler, B., Winterhalter, K.H., and Trueb, B. (1993) Complete primary structure of chicken collagen XIV. *Eur. J. Biochem.* **212**, 483-490
22. Stokes, D.G., Saitta, B., Timpl, R., and Chu, M.-L. (1991) Human α3(VI) collagen gene. *J. Biol. Chem.* **266**, 8626-8633
23. Tikka, L., Pihlajaniemi, T., Henttu, P., Prockop, D.J., and Tryggavason, K. (1988) Gene structure for the α1 chain of a human short-chain collagen (type XIII) with alternatively spliced transcripts and translation termination codon at the 5′ end of the last exon. *Proc. Natl. Acad. Sci. USA* **85**, 7491-7495
24. Dublet, B., Oh, S., Sugrue, S.P., Gordon, M.K., Gerecke, D.R., Olsen, B.R., and van der Rest, M. (1989) The structure of avian type XII collagen. *J. Biol. Chem.* **264**, 13150-13156
25. Yamagata, M., Yamada, K.M., Yamada, S.S., Tanaka, H., Nisida, Y., Obara, M., and Kimata, K. (1991) The complete primary structure of type XII collagen shows a chimeric molecule with reiterated fibronectin type III motifs, von Willebrand factor a motifs, a domain homologous to a noncollagenous region of type IX collagen, and short collagenous domain with an Arg-Gly-Asp site. *J. Cell Biol.* **115**, 209-221
26. Schwarzbauer, J.E., Tamkun, J.W., Lemischka, I.R., and Hynes, R.O. (1983) Three different fibronectin mRNAs arise by alternative splicing within the coding region. *Cell* **35**, 421-431
27. Indik, Z., Yeh, H., Ornstein-Goldstein, N., Sheppard, P., Anderson, N., Rosenbloom, J.C., Peltonen, L., and Rosenbloom, J. (1987) Alternative splicing of human elastin mRNA indicated by sequence analysis of cloned genomic and complementary DNA. *Proc. Natl. Acad. Sci. USA* **84**, 5680-5684
28. Bennett, V.D. and Adams, S.L. (1990) Identification of a cartilage-specific promoter within intron 2 of the chick α2(I) collagen gene. *J. Biol. Chem.* **265**, 2223-2230
29. Nah, H.D., Niu, Z.L., and Adams, S.L. (1994) An alternative transcript of the chick type III collagen gene that does not encode type III collagen. *J. Biol. Chem.* **269**, 16443-16448
30. Nishimura, I., Muragaki, Y., and Olsen, B.R. (1989) Tissue-specific forms of type IX collagen-proteoglycan arise from the use of two widely separated promoters. *J. Biol. Chem.* **264**, 20033-20041
31. Bork, P. (1992) The molecular architecture of vertebrate collagens. *FEBS Lett.* **307**, 49-54
32. Moradi-Ameli, M., Deleage, G., Geourjon, C., and van der Rest, M. (1994) Common topology within a non-collagenous domain of several different collagen types. *Matrix Biol.* **14**, 233-239
33. Shaw, L.M. and Olsen, B.R. (1991) FACIT collagens: Diverse molecular bridges in extracellular matrices. *Trends Biochem. Sci.* **16**, 191-194