Transcriptional promoter of the human α1(V) collagen gene (*COL5A1*)

Seungbok LEE* and Daniel S. GREENSPAN*†‡

*Department of Pathology and Laboratory Medicine and †Program in Cell and Molecular Biology, University of Wisconsin, 1300 University Avenue, Madison, WI 53706 U.S.A.

We have characterized the 5′ region of the human α1(V) collagen gene (*COL5A1*). The transcriptional promoter is shown to have a number of features characteristic of the promoters of 'housekeeping' and growth-control-related genes. It lacks obvious TATA and CAAT boxes, has multiple transcription start sites, has a high GC content, lies within a well-defined CpG island and has a number of consensus sites for the potential binding of transcription factor Sp1. This type of promoter structure, while unusual for a collagen gene, is consistent with the broad distribution of expression of *COL5A1* and is reminiscent of the promoter structures of the genes encoding type VI collagen, which has a similarly broad distribution of expression. Stepwise deletion of *COL5A1* 5′ sequences, placed upstream of a heterologous reporter gene, yielded a gradual decrease in promoter activity, indicating that the *COL5A1* promoter is composed of an array of *cis*-acting elements. A minimal promoter region contained within the 212 bp immediately upstream of the major transcription start site contained no consensus sequences for the binding of known transcription factors, but gel mobility shift assays showed this region to bind nuclear factors, including Sp1, at a number of sites. The major transcription start site is flanked by an upstream 34-bp oligopurine/oligopyrimidine stretch, or 'GAGA' box, and a downstream 56-bp GAGA box which contains a 10-bp mirror repeat and is sensitive to cleavage with S1 nuclease.

INTRODUCTION

Type V collagen is widely distributed throughout the tissues of vertebrates, where it is found at relatively low levels of abundance, compared with other components of the extracellular matrix (for a review, see [1]). Type V collagen is a fibrillar collagen usually composed of heterotrimeric monomers of the composition $\alpha1(V)_2\alpha2(V)$, although homotrimeric $\alpha1(V)_3$ [2] and heterotrimeric $\alpha1(V)\alpha2(V)\alpha3(V)$ [3] forms have also been reported. Although the function of type V collagen is largely unknown, it can combine with the much more abundant type I collagen, in various tissues, to form heterotypic fibres and may regulate the diameter of these fibres [4].

Type V collagen shares a number of characteristics with type XI collagen. Type XI collagen, characterized initially as a minor component of cartilage, is thought to combine and interact with the more abundant cartilage type II collagen in a manner analogous to the interaction of types V and I in other tissues [5,6]. Structurally, the α1(V) and α1(XI) chains show more similarity of structure than do any other two fibrillar collagen chains [7,8]. In fact, α1(V) chains can substitute for α1(XI) chains in type XI collagen in cartilage, in an age-dependent fashion [9]. It has become apparent that α1(XI) chains are found in a number of non-cartilagenous tissues where they are capable of substituting for α1(V) chains in the formation of $\alpha1(XI)_2\alpha2(V)$ heterotrimers [10–13]. In light of these results, the various type V and type XI chains may be considered to actually constitute a single collagen type.

It is not known what differences in functions the different combinations of type V and type XI chains may have. It seems probable, however, that the major determinant of which α1 chain, α1(V) or α1(XI), will complex with a given α2 chain is the co-expression of a given pair of genes within a single cell. Evidence that expression of type V collagen is up-regulated in a number of disease states, including inflammation [14], atherosclerosis [15] and some forms of cancer [16], provides added impetus to the study of mechanisms that control expression of the type V collagen genes.

In our efforts to understand the differential expression of type V α-chains, we have previously described the transcriptional promoter of the human α2(V) gene (*COL5A2*) [17]. Here we provide the first characterization of the promoter region of the α1(V) gene of any species: the human *COL5A1* gene. In contrast with *COL5A2*, the *COL5A1* promoter resembles promoters of housekeeping and growth-control-related genes.

MATERIALS AND METHODS

Cell culture, transfections and luciferase assays

AH1F normal neonatal foreskin fibroblasts have been described previously [18]. These fibroblasts, the HeLa line of epitheloid cervical carcinoma cells (A.T.C.C. CCL2) and the HT1080 fibrosarcoma line (A.T.C.C. CCL121) were all grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS). Transfections were performed in 6-cm dishes. Fibroblasts were transfected by calcium phosphate precipitation as described previously [18] using 15 μg of plasmid DNA per dish. HeLa cells were transfected using Transfectam (Promega). The DNA/Transfectam mixture was prepared in 2 ml of serum-free DMEM containing 5 μg of plasmid DNA and 25 μg of Transfectam. Cells, which were 60% confluent, were washed three times with serum-free DMEM and incubated for

Abbreviations used: DMEM, Dulbecco's modified Eagle's medium; FCS, fetal calf serum; SV40, simian virus 40; AP, activator protein; NF, nuclear factor.

‡ To whom correspondence should be addressed.

The nucleotide sequence reported in this paper has been submitted to the EMBL/GenBank/DDBJ Nucleotide Sequence Databases under accession number L38808.

18 h with DNA/Transfectam mixture. The cells were fed with fresh medium and harvested 30 h later for luciferase assays. HT1080 cells were transfected by the DEAE-dextran method [19] with some modifications. Subconfluent cells were incubated for 30 min with 2 μg of plasmid DNA in 1 ml of DMEM containing 500 μg of DEAE-dextran (Pharmacia). Cells were then treated for 2 h with 100 μM chloroquine in 4 ml of DMEM/4% FCS and then maintained for 46 h in DMEM/4% FCS before use in luciferase assays. For all transfections of HeLa and HT1080 cells, luciferase fusion plasmids were co-transfected with an equal amount of β-galactosidase plasmid PCH110 [20] for normalization of transfection efficiency. β-Galactosidase activity was determined as described previously [21] and luciferase activity was determined using a kit (Promega), following the manufacturer's instructions, and a Berthold Lumat LB9501 luminometer. A minimum of three transfection experiments was carried out for each construct in HeLa and HT1080 cell lines and seven transfection experiments were carried out for each construct in dermal fibroblasts.

DNA sequence analysis

Restriction fragments were subcloned into the vector pBluescript II KS+ (Stratagene) and sequences obtained from double-stranded templates by dideoxy chain termination, as described previously [18]. Ends of subclones were sequenced using T3 and T7 primers, with internal portions of the subclones made accessible to sequencing through introduction of appropriate deletions or use of oligonucleotide primers complementary to insert sequences. All reported sequences have been confirmed by sequencing of both strands.

Isolation of RNA

Poly(A)$^+$ RNA was isolated from cultured cells as described by Badley et al. [22].

Construction of *COL5A1* promoter–luciferase expression plasmids

For promoter analyses, various sequences from the 5′-portion of *COL5A1* were fused upstream of the firefly luciferase reporter gene in expression vector pGL2-Basic (Promega). Immediately downstream of the luciferase gene in this vector are the simian-virus-40 (SV40) small t splice site and early polyadenylation signal. To obtain construct pSBL10 (see Figure 5) a 2244-bp *COL5A1* fragment (nucleotides −2203 to +41; Figure 2) was excised with *Bst*BI and *Xho*I, blunt-ended with Klenow fragment and inserted into the *Hin*dIII site, which had also been blunt-ended with Klenow fragment, of pGL2-Basic. Subsequent cleavage of pSBL10 with *Nde*I, which cuts within the *COL5A1* sequence, and with *Xho*I, which cuts within the vector polylinker, followed by blunt-ending with Klenow fragment and re-ligation, gave construct pSBL12. Cleavage of pSBL10 with *Pvu*II, which cuts within the *COL5A1* sequence, and with *Xho*I, followed by blunt-ending with Klenow fragment and re-ligation, gave construct pSBL14. Cleavage of pSBL10 with *Sac*I, which cuts sites within both the *COL5A1* insert and the vector polylinker, followed by re-ligation, gave construct pSBL16. Cleavage of pSBL10 with *Xma*I, which cuts sites within both the *COL5A1* insert and the vector polylinker, followed by re-ligation, gave construct pSBL18. A 2359-bp *Bst*BI–*Eag*I *COL5A1* fragment (nucleotides −2203 to +156; Figure 2) was excised, blunt-ended with Klenow fragment and inserted into the blunt-ended *Bgl*II site of pGL2-Basic to give construct pSBL8. Cleavage of pSBL8 with either *Xho*I or *Xma*I, each of which cuts within both the *COL5A1* insert and the vector polylinker, and re-ligation gave construct pSBL20 or pSBL24 respectively. Partial digestion of pSBL10 with *Xma*I was followed by gel purification of a 2038-bp fragment and ligation of this fragment into the *Xma*I site of pGL2-Basic to give construct pSBL22. Remaining constructs pSBL26, 56, 57, 63, 79 and 93 were generated using standard cloning procedures and DNA fragments produced by PCR. For these constructions, the ends of oligonucleotide primers contained appropriate restriction sites for directional subcloning of DNA fragments into the polylinker of pGL2-Basic. All constructions employing PCR-generated fragments were confirmed by sequencing.

Assay for S1 nuclease sensitivity

Supercoiled plasmid containing a 1434-bp *Dra*I–*Sma*I *COL5A1* fragment (nucleotides −1008 to +426) was prepared by alkaline lysis and purified using a Qiagen-tip 100 column. Approx. 10 μg of DNA was incubated with 200 units of S1 nuclease (Pharmacia) in 200 μl of buffer containing 250 mM NaCl, 30 mM sodium acetate, pH 4.5, and 1 mM $ZnSO_4$ for 30 min at 37 °C. The reaction was terminated by addition of EDTA to a final concentration of 5 mM, followed by phenol extraction and ethanol precipitation. Portions of each sample were digested with *Hin*dIII for 2 h at 37 °C. Samples were electrophoresed on 0.8% agarose gels and bands revealed by staining with ethidium bromide. For precise mapping of the S1 site, portions of samples digested with both S1 nuclease and *Hin*dIII were blunt-ended with Klenow fragment and a resulting 3270-bp fragment, which contained insert sequences from the S1 site to nucleotide +426 and most of pBluescript, was recircularized, amplified in *Escherichia coli* and the S1 site sequenced as above.

Gel mobility shift assays

Probes for gel mobility shift assays were obtained by PCR using primers containing a 5′-extension for a unique restriction site. DNA fragments were digested with appropriate restriction enzymes and end-labelled with [α-^{32}P]dCTP using the Klenow fragment. For binding reactions, 5 μg of HeLa nuclear extract (Promega) was preincubated with 0.5 μg of the synthetic competitor poly(dI-dC) in 9 μl of reaction buffer containing 10 mM Tris/HCl, pH 7.5, 4% glycerol, 1 mM $MgCl_2$, 0.5 mM EDTA, 0.5 mM dithiothreitol and 50 mM NaCl. About 50000 c.p.m. of radiolabelled probe was added to the preincubation mixture. After 20 min incubation at room temperature, DNA–protein complexes were resolved on a non-denaturing 4% acrylamide (acrylamide/bisacrylamide, 80:1, w/w) gel. Electrophoresis was performed at room temperature at a constant 100 V in 0.5 × Tris/HCl/borate/EDTA buffer. Gels were dried under vacuum and exposed to X-ray film with intensifying screens at −70 °C. For competition experiments, molar excesses of unlabelled probes or synthetic double-stranded oligonucleotides (Sp1, 5′-GATCGATCGGGGCGGGGCGATC-3′; activator protein (AP) 1, 5′-CTAGTGATGAGTCAGCCGGATC-3′; AP2, 5′-GATCGAACTGACCGCCCGCGGCCCGT-3′; AP3, 5′-CTAGTGGGACTTTCCACAGATC-3′; nuclear factor (NF) 1/CTF, 5′-ATTTTGGCTTGAAGCCAATATG-3′) (Stratagene) were added to the binding reaction mixture.

RESULTS

Isolation of the 5′ portion of *COL5A1*

A 365-bp fragment, corresponding mostly to 5′-untranslated sequences and to sequences that encode the prepro-α1(V) signal peptide, was isolated from the 5′-end of the previously described human prepro-α1(V) cDNA clone CW334 [7] by restriction with

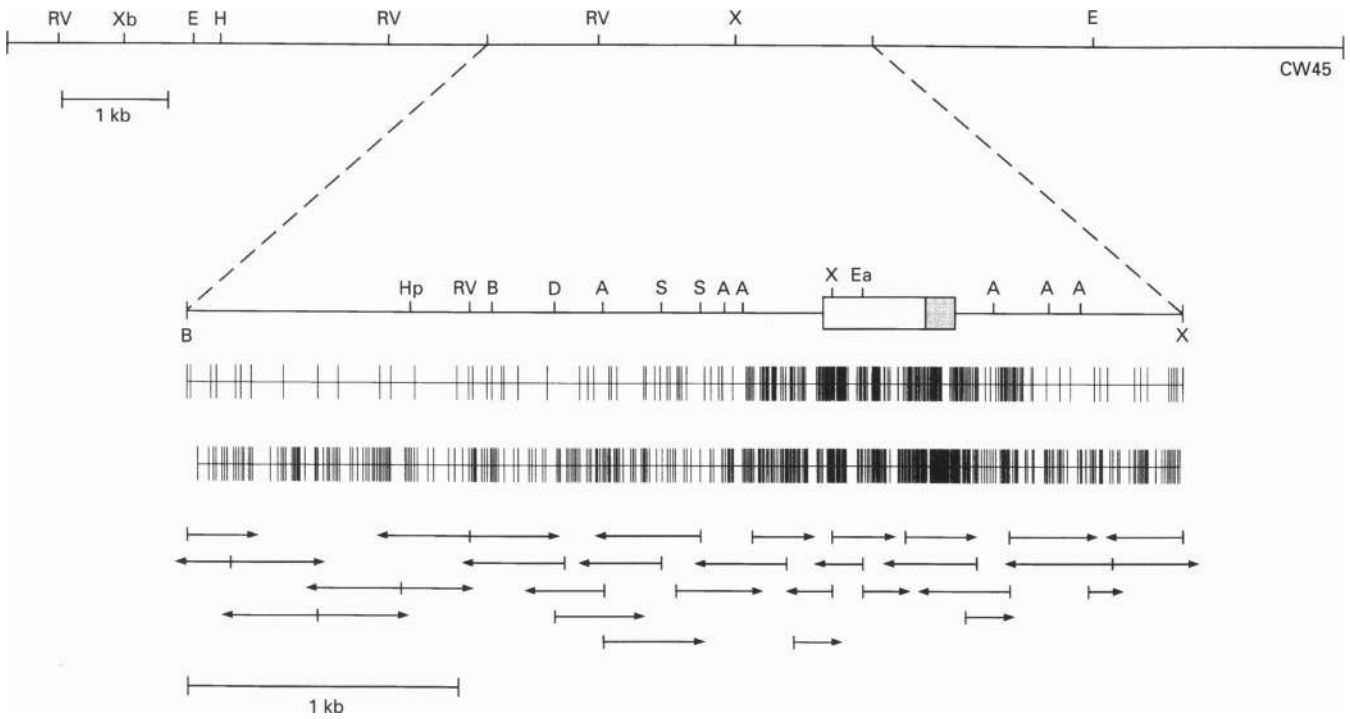

Figure 1 Partial restriction map, distribution of CpG dinucleotides and sequencing strategy of the 5′-end of the *COL5A1* gene

CW45 is a genomic clone of approx. 13 kb. A, *Apa*I; B, *Bsp*EI; D, *Dra*I; E, *Eco*RI; Ea, *Eag*I; H, *Hin*dIII; Hp, *Hpa*I; RV, *Eco*RV; S, *Sac*I; X, *Xho*I; Xb, *Xba*I. Open and filled boxes denote untranslated and coding regions, respectively, of the first exon. Each vertical line represents a CpG pair (upper line) or GpC pair (lower line). Arrows indicate region and direction of sequencing.

*Xba*I. This fragment was used as a probe to screen a human placental genomic DNA library, resulting in isolation of the approx. 13-kb genomic clone CW45 (Figure 1). CW45 was characterized by digestion with a number of restriction endonucleases and Southern-blot analysis using the 365-bp cDNA fragment as a probe (results not shown). Suitable restriction fragments were subcloned into the vector pBluescript and the nucleotide sequence of 3732 bp was determined (Figure 2). Comparison of the genomic sequence with the published cDNA sequence [7,8] allowed identification of the *COL5A1* first exon (Figures 1 and 2). We have previously predicted the first 36 amino acids of the prepro-α1(V) chain to correspond to the signal peptide, based on comparison with other secreted proteins [7]. If this prediction is correct, then *COL5A1* is similar to other characterized fibrillar collagen genes [17,23] in having the signal peptide encoded, in its entirety, by the first exon. Unlike the other characterized fibrillar collagen genes, however, the *COL5A1* first exon does not encode the first few amino acid residues of the N-propeptide, but instead contains only the first G residue of the codon for the alanine that begins the N-propeptide.

Determination of *COL5A1* transcription initiation sites

Several attempts to map the transcription start site of *COL5A1* by extension of various short synthetic oligonucleotide primers with reverse transcriptase failed, presumably due to the high GC content of the promoter region. Primer extension was successfully accomplished by using a 115-bp *Eag*I–*Xho*I restriction fragment (nucleotides +42 to +156; Figure 2) as primer. The major cDNA product obtained was 156 bases long (Figure 3a) and is consistent with a major transcription start site 382 bp upstream of the pro-α1(V) translation initiation codon. This major start site has been designated +1 in the *COL5A1* sequence (Figure 2). Consistent with the apparent absence of a *COL5A1* TATA box, a number of minor start sites were also detected both immediately adjacent to +1 and a short distance upstream of +1 (Figure 3a). RNase protection analysis produced a pattern of protected fragments (Figure 3b) that agreed with the pattern of cDNA products obtained through primer extension, thus confirming them as representing transcription start sites. A protected band approx. 37 bases shorter than the band corresponding to nucleotide +1 was also detected by both RNase protection (Figure 3b) and S1 nuclease mapping (results not shown). However, since the primer for reverse transcription used in this study hybridizes upstream of this potential site, its role in transcription remains to be confirmed by a method independent of nuclease protection.

Analysis of DNA sequences at the 5′-end of *COL5A1*

Analysis for the presence of previously described consensus sequences found potential binding sites for NF1 and AP2 in *COL5A1* 5′-flanking sequences and a cluster of 13 potential Sp1 binding sites in *COL5A*1 5′-flanking, first exon and first intron sequences (Figure 2). An additional AP2 site is found in the *COL5A1* first intron as is an enhancer 'core' element, similar to those found in viral enhancers [24] and in the first introns of other fibrillar [17,25,26] and non-fibrillar [27] collagen genes.

COL5A1 lacks a canonical CAAT box in a position typical for functional CAAT boxes [28]. The promoter region of *COL5A1* is also similar to that of a housekeeping gene [29] in that it lacks an apparent TATA box and is rich in GC nucleotides. In fact, as

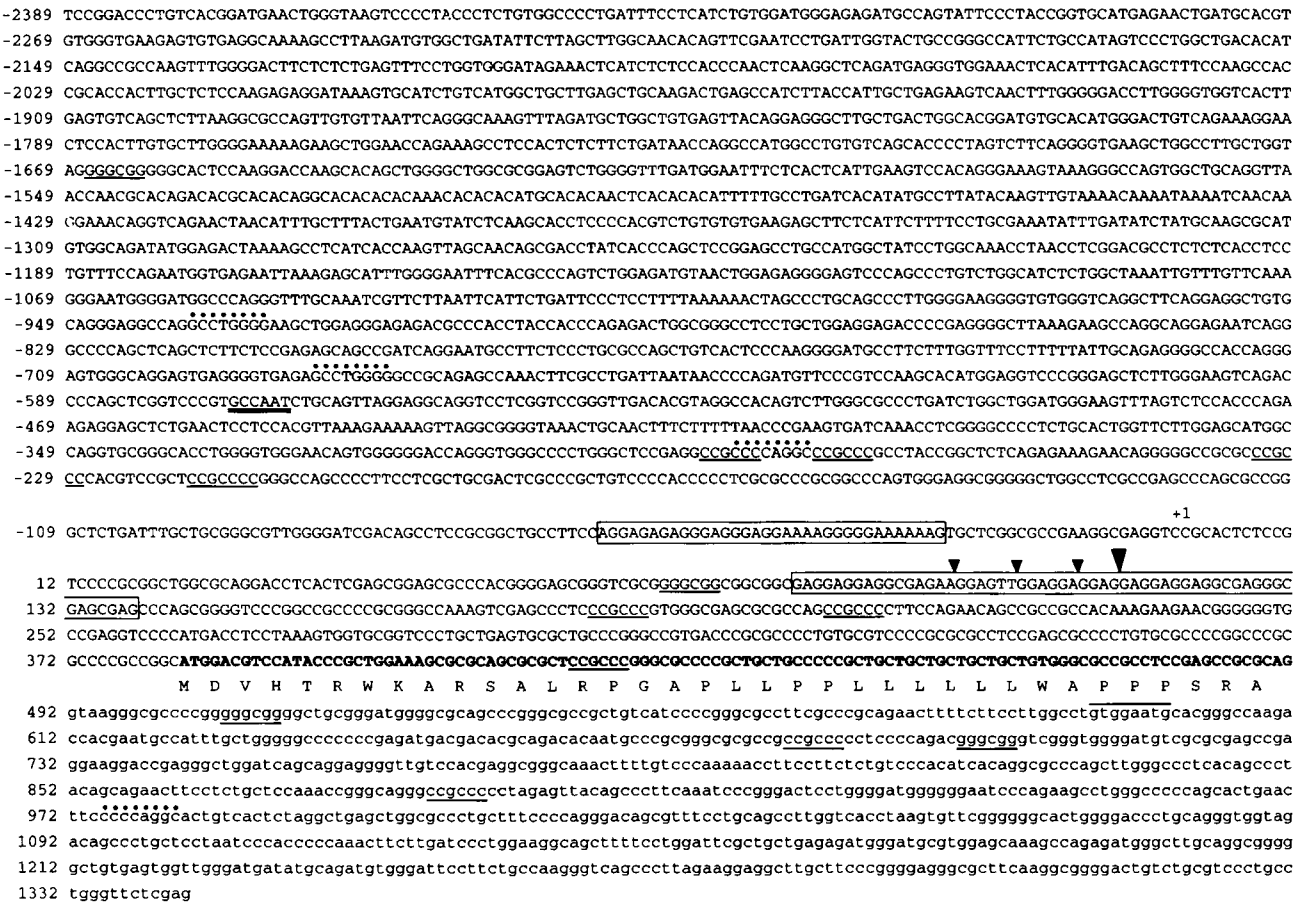

Figure 2 DNA sequences at the 5′-end of the *COL5A1* gene

Numbering to the left indicates nucleotide position relative to the major transcription start site, which is designated +1. Upper-case letters indicate 5′-flanking and first exon sequences. Lower-case letters indicate first intron sequences. Coding sequences are presented in **bold**, with the predicted amino acid sequence displayed below the corresponding nucleotide sequence. GAGA boxes are enclosed in boxes. Arrowheads represent S1-nuclease-cleavage sites. The large arrowhead represents an S1-cleavage site indicated by the endpoints of three separate clones (see the text). Underline, consensus sequence for Sp1 binding; double underline, consensus sequence for NF1 binding; dotted overline, consensus sequence for AP2 binding [42]; overline, viral-core-enhancer motif [24]. DNA sequences were analysed using programs of the Genetics Computer Group.

with housekeeping genes, the 5′-end of *COL5A1* lies within a CpG island (Figure 1). This CpG island, which is particularly rich in GC content (72%), begins upstream of the +1 transcription start site at about nucleotide −292 and extends into the first intron up to about nucleotide +730.

COL5A1 and the *COL5A2* gene, which encodes the α2(V) collagen chain, must be coexpressed in some tissues to produce the $\alpha1(V)_2\alpha2(V)$ heterotrimer that is believed to be the most widely distributed form of type V collagen [1]. However, comparison of genomic sequences upstream of the *COL5A1* translation initiation AUG with sequences from the 5′-untranslated and promoter regions of *COL5A2* [17], using the programs BEST and GAP (Genetics Computer Group), did not detect any extensive similarities.

Immediately upstream of the *COL5A1* major transcription start site is a 34-base-long (nucleotides −24 to −57) homopurine region on the sense strand. Downstream of the major transcription site is a 56-base-long purine-rich region (nucleotides +83 to +138), on the sense strand, interrupted by only six pyrimidine residues. Similar regions, composed predominantly of purines on one strand and pyrimidines on the other, have been found near the promoter regions of other genes [30–39] and have been referred to as GAGA boxes [30,31]. Some of these GAGA boxes have been implicated in affecting levels of transcription from their cognate promoters [30,32–36,39] and, in some cases, these sequences may adopt secondary structures which render them susceptible to cleavage with single-strand endonucleases such as S1 [30,32,35,36,38–40].

S1 nuclease sensitivity

To determine whether sequences surrounding the *COL5A1* major transcription start site might be susceptible to cleavage with S1 nuclease, a 1434-bp *Dra*I–*Sma*I fragment (nucleotides −1008 to +426) was subcloned into pBluescript vector and the supercoiled plasmid was digested with S1 nuclease (see the Materials and methods section). Incubation with S1 linearized the supercoiled plasmid, and subsequent digestion of linearized plasmid with *Hin*dIII generated two fragments of approx. 1100 and 3270 bp (Figure 4). The sizes of these fragments, and subsequent digestion with additional restriction endonucleases (results not shown), placed the S1-sensitive site(s) in the vicinity of the GAGA box downstream of transcription start site +1. To further localize the S1-sensitive site(s), the 3270-bp band, which contains most of

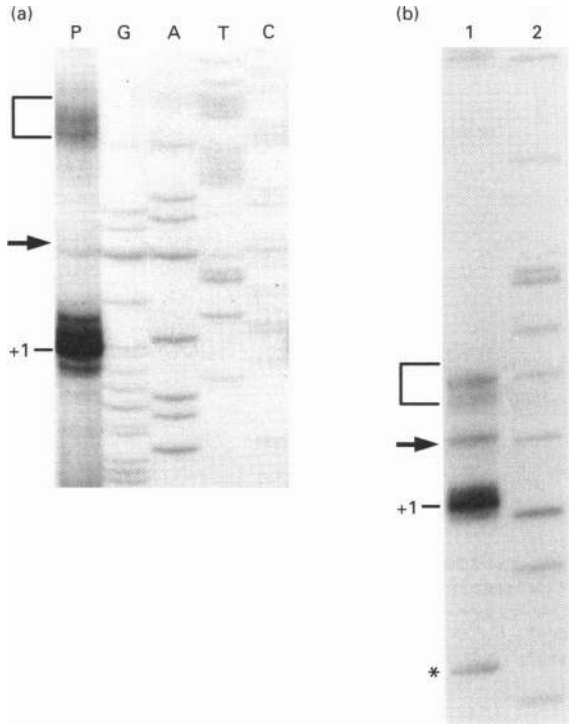

Figure 3 Determination of the transcription initiation site of the *COL5A1* gene

(**a**) Primer extension analysis. A 115-bp *Xho*I–*Eag*I fragment, corresponding to nucleotides +42 to +156 inclusive (see Figure 2), was ^{32}P-labelled by 3′-filling with Klenow fragment, hybridized at 50 °C in formamide buffer to 5 μg of dermal fibroblast poly(A)$^+$ RNA and extended with reverse transcriptase (lane P) as described previously [19]. A 25-base oligomer (nucleotides +132 to +156), with the same 5′-end as the *Xho*I–*Eag*I primer, was employed for dideoxy sequencing (lanes G, A, T and C, which show the antisense sequence) in which the template was a 1480-bp *Eco*RV–*Eag*I *COL5A1* fragment (nucleotides −1325 to +156; Figure 2). Autoradiograms are shown of the same gel exposed for 13 h without intensifying screen (lane P) or for 24 h with intensifying screen (lanes G, A, T and C). (**b**) RNase protection analysis. The 1480-bp *Eco*RV–*Eag*I fragment described above was subcloned between the *Eco*RV and *Eag*I sites of pBluescript II KS+ (Stratagene). The resultant plasmid was linearized at a *COL5A1 Sma*I site (nucleotide −210; Figure 2), and a uniformly ^{32}P-labelled 393-base riboprobe was generated by transcription with RNA polymerase T7. Conditions for annealing of the antisense probe to 2 μg of dermal fibroblast poly(A)$^+$ RNA (lane 1) and incubation with RNases A and T1 were as described previously [17]. Lane 2, ^{32}P-labelled markers of *Msp*I-digested pBR322. The major transcription start site is marked by +1. Minor start sites upstream of +1 are marked with brackets and an arrow. A potential start site downstream of +1 is marked with an asterisk. Extension products and nuclease-protected fragments were electrophoresed on denaturing 6% polyacrylamide gels.

the vector pBluescript, was recircularized with T4 ligase and six individual clones were isolated for sequencing of the S1 site(s). The ends of all of these clones were shown by sequencing to be within the GAGA box downstream of transcription start site +1 (see Figure 2). Moreover, three of the endpoints were found to be identical and five of the six endpoints were found associated with a mirror repeat containing six copies of the trinucleotide AGG (Figure 2).

Expression analysis of the *COL5A1* promoter region

A survey, by RNase protection analysis, determined that human dermal fibroblasts produce high levels of endogenous *COL5A1* mRNA and that HeLa cells produce only slightly (approx. 2-fold) lesser amounts of *COL5A1* mRNA and use the same *COL5A1* transcription start sites as the dermal fibroblasts (results not shown). HeLa cells have previously been shown to produce mRNA for the α2(V) collagen chain [41] and may, thus, produce $\alpha1(V)_2\alpha2(V)$ heterotrimers. The RNase protection survey also found that the HT1080 fibrosarcoma cell line did not produce any detectable endogenous *COL5A1* RNA. Thus dermal fibroblasts, HeLa cells and HT1080 cells were used in the transient expression assays, described below, to determine which *COL5A1* 5′ sequences may be involved in modulating expression in cell types which either express or do not express endogenous *COL5A1* genes. Interestingly, the RNase protection survey of cell lines found the A204 line of rhabdomyosarcoma cells, which has recently been shown to produce collagen composed primarily of $\alpha1(XI)_2\alpha2(V)$ heterotrimers [12], to produce a low but detectable level of *COL5A1* mRNA with transcription start sites similar to those found in fibroblasts and HeLa cells. This suggests that the A204 line may produce some α1(V) chains.

To gauge the importance of various sequences on *COL5A1* transcription, a series of stepwise deletions of *COL5A1* 5′ sequences were constructed, fused to the firefly luciferase reporter gene and transfected into HeLa and HT1080 cells and into dermal fibroblasts. For transfections into HeLa and HT1080 cells a β-galactosidase plasmid was included for normalization of transfection efficiencies (see the Materials and methods section). Activities of β-galactosidase were found to be too low, however, for use as controls in fibroblasts, presumably due to the lower transfection efficiency of these cells. Nevertheless, data obtained from fibroblasts, after seven independent transfection experiments performed to compensate for possible experimental variation, were found to be consistent and to agree with results from HeLa cells (Figure 5). Transcriptional activities of the various constructs in the different types of cells (Figure 5a) suggest that 5′-flanking sequences from −1469 to −770 contain positive *cis*-acting elements that can operate in a somewhat cell-type-specific manner. Thus constructs that contain these sequences (pSBL10 and 12) were expressed at markedly higher levels in HeLa cells and fibroblasts than in HT1080 cells, whereas removal of these sequences (pSBL14 and 16) decreased levels of expression in HeLa cells and fibroblasts to levels closer to those found in HT1080 cells. Further 5′-stepwise removal of the flanking sequences located between positions −464 and −212, to yield construct pSBL18, elevated transcriptional levels in HeLa cells and fibroblasts somewhat, but not in HT1080 cells, showing a cell-type-specific negative effect on transcription exerted by sequences in this region. Removal of the sequences between −212 and −75 reduced transcriptional activity to near basal levels, thus showing that sequences constituting a minimal *COL5A1* promoter are contained within the first 212 bp immediately upstream of the major transcription start site. Sequences upstream of −212 have no promoter activity in the absence of sequences from the minimal promoter region (Figure 5a; pSBL22).

COL5A1 minimal promoter region contains multiple binding sites for nuclear factors

To further explore the nature of the −212 to −75 region, three different probes corresponding to sequences from −212 to −171, from −170 to −119 and from −118 to −71 were prepared, incubated with HeLa cell nuclear extract and subjected to gel mobility shift assays. Each fragment was found to bind nuclear factors (Figure 6). The −118 to −71 region produced a single major protein–DNA complex whose protein factor(s) was not competed away using oligonucleotides containing binding sites for several known transcription factors (Figure 6a). The pattern of binding to the −212 to −171 region was more complex, with

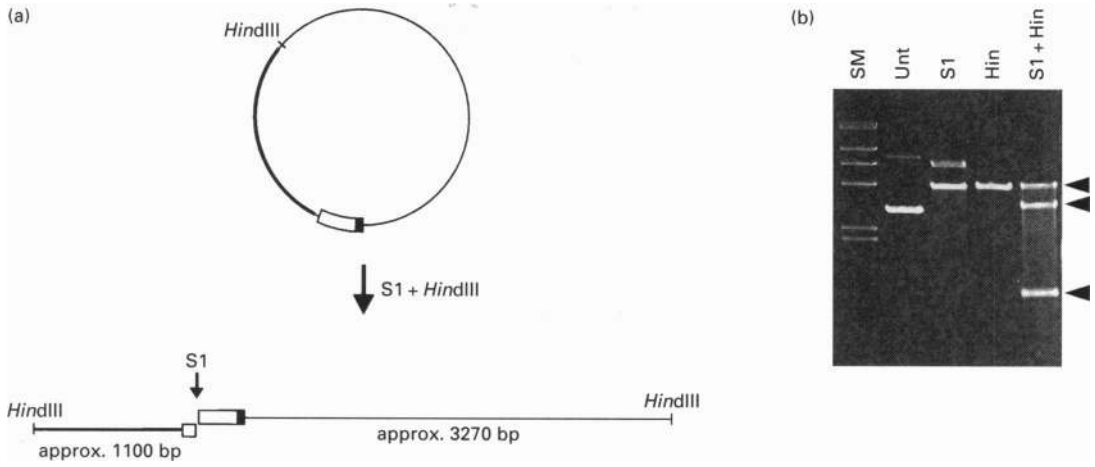

Figure 4 S1 nuclease sensitivity of the *COL5A1* promoter region

(**a**) A diagram is shown of a plasmid in which a 1434-bp *Dra*I–*Sma*I fragment of the *COL5A1* 5′ end was subcloned between the *Eco*RV and *Sma*I sites of pBluescript. Also shown, diagrammatically, is the generation of 1100- and 3270-bp fragments upon digestion of the plasmid with S1 and *Hin*dIII. The thin line represents pBluescript sequences, the heavy line represents *COL5A1* 5′-flanking sequences and the open and filled boxes represent untranslated and coding regions, respectively, of the *COL5A1* first exon. (**b**) A 0.8% agarose gel is shown with lanes containing size markers comprising a *Hin*dIII digest of lambda DNA (SM), untreated plasmid (Unt), plasmid digested with S1 nuclease (S1), plasmid digested with *Hin*dIII (Hin) and S1-digested plasmid redigested with *Hin*dIII (S1 + Hin). Arrowheads mark the positions of linearized plasmid and the 1100- and 3270-bp fragments.

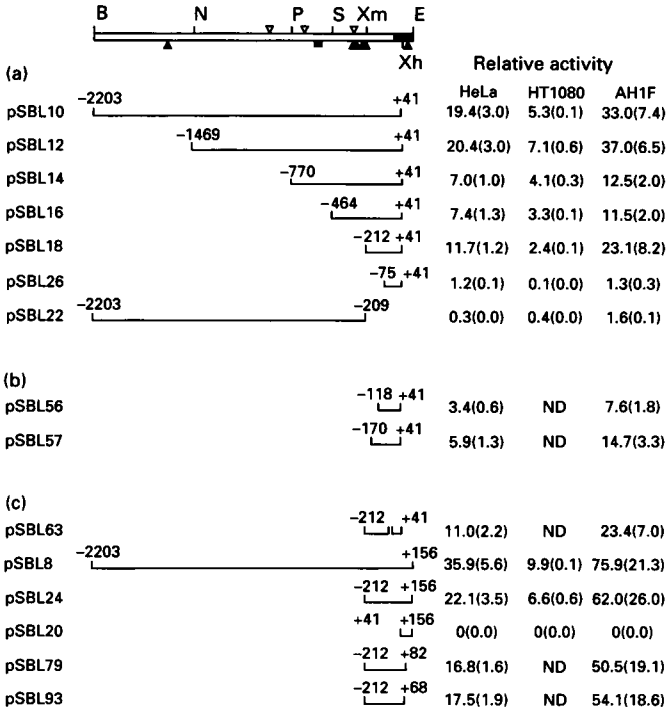

Figure 5 Promoter activities of sequences from the 5′-end of *COL5A1*

Lines beneath a partial restriction map of the 5′-portion of *COL5A1* denote the extent of *COL5A1* DNA fragments fused to the luciferase gene for transient expression analysis. Numbers at the ends of each line indicate the positions of the 5′- and 3′-ends of each fragment relative to the *COL5A1* major transcription start site. The luciferase activity of each construct is expressed relative to the level achieved by the vector pGL2-Control (100.0). Construct pGL2-Control (Promega) is similar to the pGL2-Basic vector from which the various *COL5A1* fusion plasmids were constructed. However, it also contains the SV40 early promoter and transcriptional enhancer located upstream and downstream, respectively, of the luciferase gene. Data represent means ± S.D. (in parentheses) of 3–7 independent transfection experiments. ND, not done; B, *Bst*BI; E, *Eag*I; N, *Nde*I; P, *Pvu*II; S, *Sac*I ; Xh, *Xho*I; Xm, *Xma*I. Open and closed triangles and closed boxes represent consensus sites for binding of AP2, Sp1 and NF-1 respectively.

one major and a number of minor complexes produced (Figure 6b). Competition experiments showed the major complex to contain the transcription factor Sp1 (Figure 6b). The −170 to −119 region produced two major complexes, and competition experiments showed the complex with least mobility to contain Sp1 (Figure 6c). The complex with greater mobility did not contain factors which were competed for by oligonucleotides containing the consensus sequences of several known transcription factors (Figure 6c). The binding sites for the two major complexes in the −170 to −119 region were further localized by competition of the labelled −170 to −119 probe with two unlabelled competitor DNA fragments corresponding to sequences from −170 to −134 (Figure 6d, competitor A) and from −149 to −119 (Figure 6d, competitor B). These experiments showed that binding sites for either complex were not located in the region where the two competitors overlap (−149 to −134). Instead, the binding site for the complex containing the unknown factor(s) was located to sequences between −170 and −149 and the binding site for the complex containing Sp1 was located to sequences somewhere between −134 and −119 (Figure 6d). Although the above experiments show Sp1 to bind sequences between −212 and −171 and between −134 and −119, neither region contains the reported Sp1 consensus binding site 5′-GGGCGG-3′ [42]. Both regions are, however, GC-rich (see Figure 2).

Two luciferase fusion constructs, containing *COL5A1* sequences either from −118 to +41 (Figure 5b; pSBL56) or from −170 to +41 (Figure 5c; pSBL57), were prepared to test which regions of the *COL5A1* 5′ sequence corresponding to the three mobility shift probes shown to contain nuclear-factor-binding sites (above) might be of functional significance to the *COL5A1* promoter. As can be seen (Figures 5a and 5b), stepwise addition of sequences from −118 to −75 (Figure 5b; pSBL56), from −170 to −118 (Figure 5b; pSBL57) and from −212 to −170 (Figure 5a; pSBL18) produces incremental increases in levels of *COL5A1* promoter activity. Thus, each region that we have shown to bind nuclear factors appears to be of functional significance to the operation of the *COL5A1* promoter.

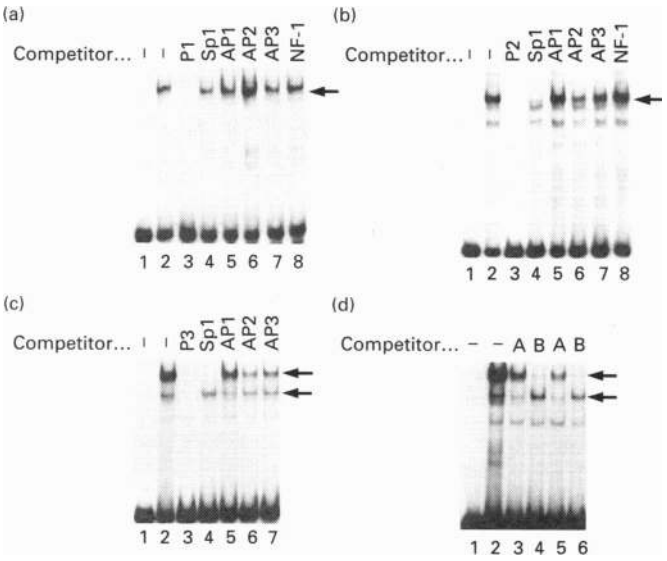

Figure 6 Binding of nuclear factors to the *COL5A1* minimal promoter region

Gel mobility shift assays were performed with labelled probes corresponding to *COL5A1* nucleotides −118 to −71 (P1; **a**), −212 to −171 (P2; **b**) and −170 to −119 (P3; **c** and **d**). Control assays were performed without nuclear extract or competitors (lane 1), or with nuclear extract, but without competitors (lane 2). In panels (**a**), (**b**) and (**c**), each labelled probe was competed against a 300-fold, unlabelled, molar excess either of itself (lane 3) or of oligonucleotides containing consensus binding sites for the indicated protein factors (lanes 4–8). Panel (**d**) shows competition for binding of nuclear factors between labelled probe P3 (−170 to −119) and 100-fold (lanes 3 and 4) or 200-fold (lanes 5 and 6) molar excesses of unlabelled competitor fragments A (−170 to −134) and B (−149 to −119). Major protein–DNA complexes are marked by arrows.

Testing for possible contributions of GAGA boxes to *COL5A1* promoter function

To test for possible contribution of the upstream GAGA box to *COL5A1* promoter activity, a 34-bp deletion was made in the −212 to +41 minimal promoter that exactly removed the upstream GAGA box. The resultant construct (Figure 5c; pSBL63) showed the same levels of expression in HeLa cells and fibroblasts as did an undeleted −212 to +41 construct (Figure 5a; pSBL18). Thus, the GAGA box in the *COL5A1* 5′-flanking sequence does not appear to contribute significantly to the transcription levels observed during transient expression in cultured cells.

To test for possible contributions of the downstream GAGA box, located at +83 to +138, to *COL5A1* promoter activity, two constructs were created in which 5′-untranslated sequences were extended to +156 (Figure 5c; pSBL8 and pSBL24) such that these constructs now contained the downstream GAGA box. The result was that expression levels were approximately doubled, in HeLa cells and fibroblasts, compared with constructs that contained similar 5′-flanking sequences but 5′-untranslated sequences extending only to +41 (Figure 5a; pSBL10 and 18). The sequences between +41 and +156 showed no intrinsic promoter activity (Figure 5c; pSBL20), indicating that these sequences had stimulated the expression levels of constructs in some other manner. To ascertain whether this effect was due to the presence of the downstream GAGA box, a construct was made which contained *COL5A1* sequences extending from −212 to +82 (Figure 5c; pSBL79). Transfection with this construct, however, yielded a level of expression intermediate between that of constructs containing *COL5A1* sequence extending from −212 to +41 (pSBL18) and from −212 to +156 (pSBL24). This suggests that although the downstream GAGA box may play some role in increasing levels of transient expression, other 5′-untranslated sequences, upstream of this GAGA box, are also involved in increasing expression levels. The 5′-untranslated sequences upstream of the GAGA box contain an Sp1 consensus binding site, at +70 to +75. However, further deletion of 5′-untranslated sequences to remove this site (Figure 5c; pSBL93) does not further reduce expression levels. This suggests that sequences between +41 and +68, which contain no known consensus site, also contribute to increased levels of transient expression. It remains to be determined whether various sequences of the *COL5A1* 5′-untranslated region, including the downstream GAGA box, increase expression levels by affecting transcription or by increasing the stability of RNA transcripts.

DISCUSSION

This report constitutes the first characterization of the promoter region of a type V α1 chain collagen gene from any species. The *COL5A1* promoter is shown to be similar to the promoters of housekeeping genes and the promoters of a number of growth-control-related genes in that it is GC-rich, lacks an identifiable TATA box, contains a number of potential, and demonstrated, binding sites for transcription factor Sp1 and has multiple transcription start sites. Thus, the nature of the *COL5A1* promoter region suggests a broad distribution of expression consistent with previous findings of *COL5A1* transcripts in a wide range of tissues [7]. It is also highly probable, however, that the α1(V) chain is not expressed in all cell types in the body, and we demonstrate that endogenous *COL5A1* genes are not expressed in the HT1080 cell line. Moreover, evidence is presented that *cis*-acting elements within the region bounded by nucleotides −1469 and −770 and between nucleotides −212 and −75, within the minimal promoter region, may be involved in conferring cell-type specificity upon the *COL5A1* promoter. Future studies, which will utilize and extend the systems established here, should provide further information regarding the nature of these elements and corresponding protein factors.

Stepwise deletion of *COL5A1* 5′ sequences yielded a gradual decrease in promoter activity, indicating that the *COL5A1* promoter is composed of an array of *cis*-acting elements. A number of these elements are shown to be contained between nucleotides −212 and −75, within the minimal promoter region. This region is shown to contain a number of binding sites for as yet unknown nuclear factors and for Sp1, which has previously been implicated as an important transcription factor in the operation of housekeeping-gene-like promoters [29]. Although the *COL5A1* 5′ region is shown to contain multiple canonical Sp1 binding sites, those areas from −212 to −75 that bind Sp1 must contain non-canonical sites.

The *COL5A1* minimal promoter region also contains a homopurine/homopyrimidine GAGA box, as does the 5′-untranslated region. GAGA boxes have now been found in the TATA-less promoter regions of a number of growth-related [32–34,36,39] and extracellular matrix [30,31,35,37,38] genes. We are not, however, aware of previous examples of GAGA boxes in 5′-untranslated regions. Such GAGA boxes may assume secondary structures, making them susceptible to cleavage with S1 nuclease. All *COL5A1* S1 breakpoints examined occurred within the downstream GAGA box and five of the six breakpoints were found associated with a mirror repeat containing six copies of the trinucleotide AGG, suggesting the possible formation of a

three-stranded triplex structure with a single-stranded loop (H-DNA) [40]. Transient expression analyses showed that removal of the upstream GAGA box, which is not S1 sensitive, had no effect on basic promoter function. Removal of the downstream GAGA box had a small negative effect on luciferase levels that, however, was not definitively ascribed to either transcriptional or post-transcriptional effects. Previously, S1-sensitive GAGA boxes have been shown to be important elements in the functioning of a number of TATA-less promoters [30,32–36,39]. An exception was an S1-sensitive GAGA box in the gene for the $\alpha 2$ chain of type VI collagen [38]. S1-sensitive sites that form structures with some single-strandedness may bind nucleosomes poorly and effect a more 'open' chromosome structure that may facilitate transcription *in vivo*. Indeed, many sequences shown to be S1 sensitive *in vitro* are also DNase I-sensitive sites in chromatin *in vivo* [43]. Thus, the downstream *COL5A1* GAGA box may influence transcription levels of *COL5A1 in vivo* (e.g. during development) even though transcriptional effects are not apparent in transient expression assays. Preliminary results with gel shift assays have found that both upstream and downstream GAGA boxes bind nuclear factors and, moreover, that the two GAGA boxes compete with each other in binding the same factors (results not shown). Therefore, it is possible that the two elements interact *in vivo* by binding shared nuclear factors in a single complex, thus further affecting chromatin structure at the 5′ end of *COL5A1*. Roles that the two GAGA boxes may play in determining chromatin structure can be probed in future experiments of DNase I sensitivity of these regions both in cells that express and in cells that do not express endogenous *COL5A1* genes.

This work was supported by National Institutes of Health Grant GM46846 (D.S.G.) and by a Cremer Scholarship Award for Research in the Basic Medical Sciences (S.L.).

REFERENCES

1 Fessler, J. H. and Fessler, L. I. (1987) in Structure and Function of Collagen Types (Mayne, R. and Burgeson, R. E., eds.), pp. 81–103, Academic Press, Orlando
2 Haralson, M. A., Mitchell, W. M., Rhodes, R. K., Kresina, T. F., Gay, R. and Miller, E. J. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 5206–5210
3 Niyibizi, C., Fietzek, P. P. and van der Rest, M. (1984) J. Biol. Chem. **259**, 14170–14174
4 Birk, D. E., Ficht, J. M., Babiarz, J. P. and Linsenmayer, T. G. (1988) J. Cell Biol. **106**, 999–1008
5 Morris, N. P. and Bachinger, H. P. (1987) J. Biol. Chem. **262**, 11345–11350
6 Mendler, M., Eich-Bender, S. G., Vaughn, L., Winterhalter, K. H. and Bruckner, P. (1989) J. Cell Biol. **108**, 191–197
7 Greenspan, D. S., Cheng, W. and Hoffman, G. G. (1991) J. Biol. Chem. **266**, 24727–24733
8 Takahara, K., Sato, Y., Okazawa, K., Okamoto, N., Noda, A., Yaoi, Y. and Kato, I. (1991) J. Biol. Chem. **266**, 13124–13129
9 Eyre, D. R. and Wu, J. J. (1987) in Structure and Function of Collagen Types (Mayne, R. and Burgeson, R. E., eds.), pp. 261–281, Academic Press, Orlando
10 Niyibizi, C. and Eyre, D. R. (1989) FEBS Lett. **242**, 314–318
11 Brown, K. E., Lawrence, R. and Sonenshein, G. E. (1991) J. Biol. Chem. **266**, 23268–23273
12 Kleman, J.-P., Hartmann, D. J., Ramirez, F. and van der Rest, M. (1992) Eur. J. Biochem. **210**, 329–335
13 Mayne, R., Brewton, R. G., Mayne, P. M. and Baker, J. R. (1993) J. Biol. Chem. **268**, 9381–9386
14 Narayanan, A. S., Engel, L. D. and Page, R. C. (1983) Collagen Relat. Res. **3**, 323–334
15 Ooshima, A. (1981) Science **213**, 666–668
16 Barsky, S. H., Rao, C. N., Grotendorst, G. R. and Liotta, L. A. (1982) Am. J. Pathol. **108**, 276–283
17 Greenspan, D. S., Lee, S.-T., Lee, B.-S. and Hoffman, G. G. (1991) Gene Expression **1**, 29–39
18 Lee, S.-T., Smith, B. D. and Greenspan, D. S. (1988) J. Biol. Chem. **263**, 13414–13418
19 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn., Cold Spring Harbor Laboratory, Cold Spring Harbor
20 Hall, C. V., Jacobs, P. E., Ringold, G. M. and Lee, F. (1983) J. Mol. Appl. Genet. **2**, 101–109
21 Shen, R.-F., Li, Y., Sifers, N., Wang, H., Hardick, C., Tsai, S. and Woo, S. (1987) Nucleic Acids Res. **15**, 8399–8415
22 Badley, J. E., Bishop, G. A., St. John, T. and Frelinger, J. A. (1988) BioTechniques **6**, 114–116
23 Sandell, L. J. and Boyd, C. D. (1990) in Extracellular Matrix Genes (Sandell, L. J. and Boyd, C. D., eds.), pp. 1–56, Academic Press, San Diego
24 Weiher, H., Konig, M. and Gruss, P. (1983) Science **219**, 626–631
25 Bornstein, P., McKay, J., Morishima, J. K., Devarayalu, S. and Gelinas, R. E. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 8869–8873
26 Rossi, P. and de Crombrugghe, B. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 5590–5594
27 Christiano, A. M., Hoffman, G. G., Chung-Honet, L. C., Lee, S., Cheng, W., Uitto, J. and Greenspan, D. S. (1994) Genomics **21**, 169–179
28 Maniatis, T., Goodbourn, S. and Fischer, J. A. (1987) Science **236**, 1237–1245
29 Dynan, W. S. (1986) Trends Genet. **6**, 196–197
30 Young, M. F., Findlay, D. M., Dominguez, P., Burbelo, P. D., McQuillan, C., Kopp, J. B., Robey, P. G. and Termine, J. D. (1989) J. Biol. Chem. **264**, 450–456
31 Nomura, S., Hashmi, S., McVey, J. H., Ham, J., Parker, M. and Hogan, B. L. M. (1989) J. Biol. Chem. **264**, 12201–12207
32 Hoffman, E. K., Trusko, S. P., Murphy, M. and George, D. L. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 2705–2709
33 Kennedy, G. C. and Rutter, W. J. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 11498–11502
34 Bossone, S., Asselin, C., Patel, A. J. and Marcu, K. B. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 7452–7456
35 Santra, M., Danielson, K. G. and Iozzo, R. V. (1994) J. Biol. Chem. **269**, 579–587
36 Mavrothalassitis, G. J., Watson, D. K. and Papas, T. S. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 1047–1051
37 Koller, E. and Trueb, B. (1992) Eur. J. Biochem. **208**, 769–774
38 Koller, E., Hayman, A. R. and Trueb, B. (1991) Nucleic Acids Res. **19**, 485–491
39 Johnson, A., Jinno, Y. and Merlino, G. T. (1988) Mol. Cell. Biol. **8**, 4174–4184
40 Wells, R. D., Collier, D. A., Hanvey, J. C., Shimizu, M. and Wohlrab, F. (1988) FASEB J. **2**, 2939–2949
41 Furth, J. J., Wroth, T. H. and Ackerman, S. (1991) Exp. Cell. Res. **192**, 118–121
42 Jones, N. C., Rigby, P. W. J. and Ziff, E. (1988) Genes Dev. **2**, 267–281
43 Weintraub, H. (1983) Cell **32**, 1191–1203

Received 6 January 1995/5 April 1995; accepted 20 April 1995