Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 11679–11683, November 1994
Genetics

The genes *COL4A5* and *COL4A6*, coding for basement membrane collagen chains α5(IV) and α6(IV), are located head-to-head in close proximity on human chromosome Xq22 and *COL4A6* is transcribed from two alternative promoters

(extracellular matrix/collagen IV/bidirectional promoter/alternative promoter/synteny)

MANABU SUGIMOTO*†, TOSHITAKA OOHASHI*, AND YOSHIFUMI NINOMIYA*‡

Departments of *Molecular Biology and Biochemistry, and †Ophthalmology, Okayama University Medical School, 2-chome 5-1, Shikata-cho, Okayama 700, Japan

Communicated by Darwin J. Prockop, June 24, 1994 (received for review January 28, 1994)

ABSTRACT **The genes for the α5(IV) and α6(IV) chains of human basement membrane collagen type IV have been found together on chromosome X at segment q22 and have been reported to be arranged in a head-to-head fashion. Here we report the 5′ flanking sequences of *COL4A5* and *COL4A6* and that *COL4A6* is transcribed from two alternative promoters in a tissue-specific fashion. Analysis of the sequence immediately upstream of the transcription start sites revealed some features of housekeeping genes—i.e., the lack of a TATA motif and the presence of CCAAT and CTC boxes. Further analysis revealed that *COL4A6* contains two alternative promoters that control the generation of two different transcripts. One transcription start site (from exon 1′) is 442 bp away from the transcription start site of *COL4A5*, while an alternative transcription start site (from exon 1) is located 1050 bp from the first one and drives the expression of a second transcript that encodes an α6(IV) chain with a different signal peptide. Reverse transcription–PCR experiments revealed that the transcript from exon 1′ is abundant in placenta, whereas the transcript from exon 1 is more frequently found in kidney and lung. These results provide additional clues to answering the general question of what mechanisms are used to generate unique basement membrane structures in different tissues.**

The collagen gene superfamily is composed of >30 distinct genes with products that form 18 or more heterotrimeric or homotrimeric molecules (1, 2). Many of the collagen genes are dispersed throughout the human genome. For instance *COL1A1* and *COL1A2*, the genes encoding the α1(I) and α2(I) chains of collagen type I, are located on chromosomes 17 and 7, respectively. To date, 28 collagen genes have been assigned on individual chromosomes in humans. However, for the genes that code for collagen IV the situation is different (1, 2). Not only were the genes for the α1(IV) and α2(IV) chains (*COL4A1* and *COL4A2*) initially mapped to the same segment of the long (q) arm of human chromosome 13 (3), but the two genes were found to be arranged in a head-to-head fashion and separated by only 127 bp (4, 5).

In addition to the two major subunit components of collagen IV, α1(IV) and α2(IV), four minor components, α3(IV), α4(IV), α5(IV), and α6(IV), have been identified and characterized. The genes for the α3(IV) and α4(IV) chains are colocalized on chromosome 2q35–q37 (6, 7). Based on amino acid sequence homology, domain structure, and gene structure, these six chains can be classified into two groups: group A, consisting of α1, α3, and α5 chains, and group B, with α2, α4, and α6 chains (8). As mentioned above, *COL4A1* and *COL4A2* are colocalized on chromosome 13 and *COL4A3* and *COL4A4* are on chromosome 2. Based on the existence of these two gene pairs, each encoding one member of the A and B groups, we predicted, even before the α6(IV) chain had been identified, the existence of an additional chain in the group B as the counterpart of the α5(IV) chain. We hypothesized that the genes encoding α5(IV) and this unknown chain would represent a third type IV collagen gene pair on the X chromosome, where *COL4A5* had been localized. This prediction and hypothesis have now been confirmed through the cloning of α6(IV) cDNAs (8) and through the identification of the α6(IV) gene in the region upstream of the α5(IV) gene promoter (9). In this paper we describe the precise relationship between the 5′ flanking regions of *COL4A6* and *COL4A5* and we report that *COL4A6* is transcribed from different transcription start sites in different tissues.§

MATERIALS AND METHODS

Screening of Genomic DNA. A restriction fragment of cDNA, TM51 (1550 bp, ref. 8), encoding the human α6(IV) chain was used for isolation of the related gene fragment. A total genomic library in λ phage EMBL3 was purchased from Clontech (HL1006d) and screened. Two positive clones, AF2 and AF3, were isolated from 6×10^5 phage plaques. One of them, AF2, was characterized further, because Southern blotting analysis revealed that AF2 encoded more of the 5′ part of the gene. Fragments were labeled by the random primer method (10). Hybridization and washing were as described (11).

Total human genomic DNA was isolated (12) from white blood cells from a healthy young man. Approximately 30 μg of the DNA was digested with *Pst* I, electrophoresed in a 0.8% agarose gel, blotted onto a nylon filter (Hybond-N^+; Amersham), and hybridized with two probes: TM51-5 (probe 1, 398 bp), representing the 5′ end of α6(IV) cDNA, and a *Pst* I–*Hin*cII restriction fragment (probe 2, 506 bp, indicated by an asterisk in Fig. 1) of the genomic DNA clone AF2, covering exon 1 of the α5(IV) gene.

Rapid Amplification of cDNA 5′ Ends (5′ RACE) and Reverse Transcription–Polymerase Chain Reaction (RT–PCR). RNA was extracted from frozen kidney, lung, and placenta and cultured keratinocytes by the guanidinium thiocyanate method (12). To determine the structure of the 5′ end of the cDNA, we used 5′ RACE (13). Four primers were synthesized for this purpose. Primer 1 was 5′-TTCCTCGGTCAGGCACAA-3′ [complementary to nt 268–285 from the 5′ end of cDNA for α6(IV) chain (8)]; primer 2 was 5′-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: RACE, rapid amplification of cDNA ends.

‡To whom reprint requests should be addressed.

§The sequence reported in this paper has been deposited in the GenBank data base (accession no. D28116).

AACCAGGAGCAGCCACAA-3′ (complementary to nt 247–264). Also, hybrid primer 5′-CTGAATTCTCGAGTCGAC-$(T)_{17}$-3′ and adaptor primer 5′-CTGAATTCTCGAGTCGAC-3′ were prepared. The first-strand cDNA was synthesized by Moloney murine leukemia virus reverse transcriptase (United States Biochemical) from primer 1 and total RNA extracted from placenta, kidney, lung, and keratinocytes. A stretch of dA residues was added to the first-strand cDNA by terminal deoxynucleotidyltransferase. Using an aliquot of the material, we performed PCR with primer 2 and the hybrid primer under the following conditions: 92°C for 40 sec, followed by 40 cycles of 94°C, 40 sec; 58°C, 1 min; 70°C, 1.5 min. At the end of the last cycle, the sample was further incubated at 70°C for 15 min.

The products were electrophoresed in 1.5% agarose gels, blotted onto Hybond-N+ (Amersham), and hybridized with ^{32}P-labeled specific probes: cDNA fragment TM51-5 (probe 1, 398 bp) and AF2/Eco1.4/Hinc Eco 0.8 fragment (probe 3, 774 bp), which covers the entire exon 1′ of the α6(IV) gene. The TA cloning system (Invitrogen) was used to clone the PCR product.

Nucleotide Sequence Analysis. Nucleotide sequence was determined by the dideoxy chain-termination method (14) on double-stranded pBluescript II vectors (15). We used the fluorescence labeled dye-terminator method on an Applied Biosystems model 373A automatic sequencer. For long fragments, a Kilo-Sequence deletion kit (Takara Shuzo, Kyoto) with exonuclease III, mung bean nuclease, and specific primers was utilized for sequencing. Nucleotide sequence was determined from both directions. Promoter sequence was analyzed by the computer program MacVector (International Biotechnologies).

RESULTS AND DISCUSSION

Isolation of Genomic Fragments That Contain 5′ Exons for the α6(IV) Chain. Two α6(IV) genomic clones, AF2 and AF3, were isolated by screening a genomic library with cDNA TM51 (8) as probe. This cDNA (1550 bp) has been sequenced and shown to encode the 5′ untranslated region, the signal peptide, the 7S domain, and the amino-terminal portion of the COL1 domain of the human α6(IV) chain (8). Southern blotting analysis revealed that AF2 encoded more of the 5′ part of the gene than AF3, and AF2 was therefore characterized further. Digestion of AF2 (20 kb) with *Pst* I and *Eco*RI and Southern blotting with TM51-5 (the most 5′ part of TM51, probe 1) showed hybridization to two restriction fragments, *Pst* I–*Pst* I (0.8 kb, see Fig. 1) and *Pst* I–*Eco*RI (0.8 kb). Their nucleotide sequences were determined. By comparing the nucleotide sequence obtained from the genomic DNA with that of the cDNA, we identified two exons, exons 1 and 2 (Fig. 2), 246 and 50 bp. We also determined the nucleotide sequence of the *Eco*RI–*Eco*RI fragment (1.4 kb, see Figs. 1 and 2) which is located upstream of the *Eco*RI–*Pst* I fragment (0.8 kb). Interestingly, we found exon 1 of *COL4A5* (16) within the sequence of this upstream fragment. The coding sequences of exons 1 and 2 of *COL4A6* and exon 1 of *COL4A5* were located on opposite strands. The distance between the two exons 1 was 1492 bp. This was an indication that the two genes were closely associated: this association was more precisely analyzed in subsequent experiments.

To demonstrate that the isolated genomic clone did not represent a cloning artifact, we compared the restriction map of clone AF2 with that seen by genomic Southern blot analysis. Human genomic DNA was digested with restriction enzyme *Pst* I and hybridized with probe 1 [a portion of the cDNA for the human α6(IV) chain, TM51-5; ref. 8] and probe 2 [a portion of genomic DNA for the human α5(IV) chain, indicated by an asterisk in Fig. 1). One of the *Pst* I restriction fragments (2.1 kb) hybridized with these two probes (data not shown). This result confirmed the restriction map of the upstream region of the two genes as shown in Fig. 1.

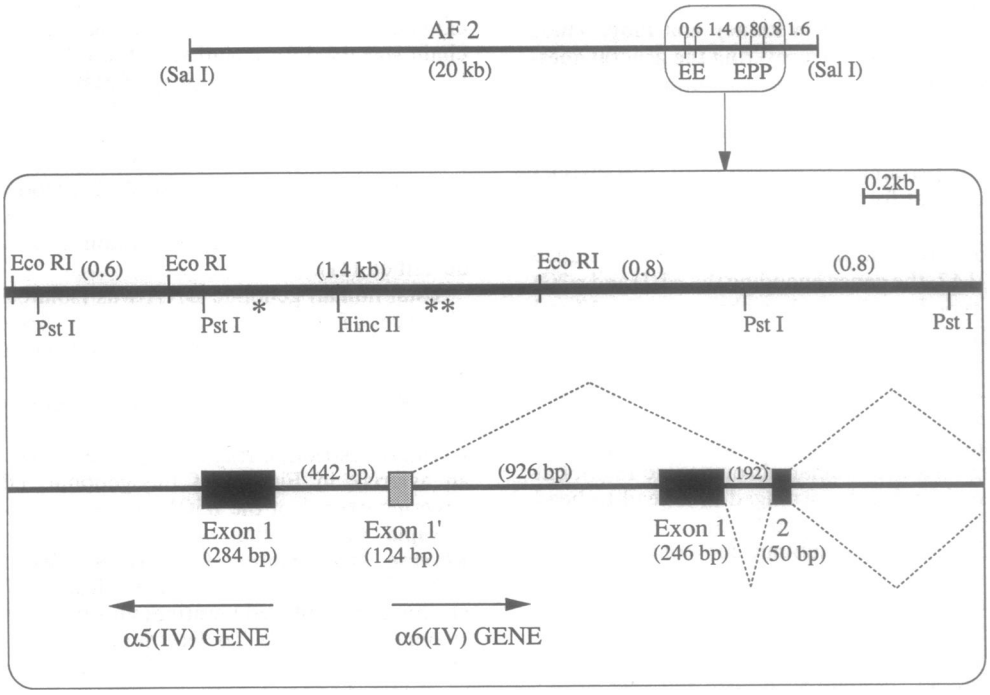

FIG. 1. Schematic representation of the genomic fragment AF2 and the head-to-head arrangement of the human α6(IV) and α5(IV) genes. The top line depicts the isolated genomic fragment AF2 (20 kb), harboring both α6(IV) and α5(IV) genes. A part of the fragment is highlighted. Genomic organization of the 5′ ends of the two genes, *COL4A5* and *COL4A6*, are arranged in opposite directions on opposite strands. Representative restriction sites and relative locations of exon 1 for *COL4A5* and exons 1′, 1, and 2 for *COL4A6* and introns are shown. The *Hinc*II site is indicated within the 1.4-kb *Eco*RI–*Eco*RI fragment. The locations of probes 2 and 3 are indicated by an asterisk and double asterisks, respectively. Numbers in parentheses indicate the lengths of the restriction fragments and exons and introns. As indicated by the dotted lines, exon 1′ is spliced to exon 2 and exon 1 is spliced to exon 2.

FIG. 2. Sequence of the 1.4-kb *Eco*RI–*Eco*RI and 0.8-kb *Eco*RI–*Pst* I fragments of genomic fragment AF2. Exon 1 for *COL4A5* and exons 1′ and 1 for *COL4A6* are indicated by boxed-in areas. Directions of transcription are shown by arrows. Amino acid sequences of the coding region for both α5(IV) and α6(IV) chains are shown by single-letter symbols. Only the amino-terminal part of the signal peptide is shown for the α6(IV) chain. Potential CCAAT boxes, GC boxes, CTC box, and AP1 and AP2 sites are marked by shaded boxes. The numbering of the sequence starts with 1 at the *Eco*RI site within the first intron of *COL4A5* and ends with 2229 at the *Pst* I site, and the nucleotide sequences of both strands are shown.

5′ RACE Indicates the Presence of the Alternative Exon. To find out whether the 5′ end sequence was the same among all *COL4A6* mRNAs, we performed 5′ RACE using RNAs from various tissues and cells. The primers were hybrid primer with T_{17} (as the forward primer) and primer 2 (as the reverse primer). The 5′ RACE products were hybridized with two probes. One of the probes (probe 1 in Fig. 3*B*) was the cDNA fragment TM51-5. This cDNA fragment encodes the most amino-terminal portion of the α6(IV) collagen chain (8). It showed one major band for placenta, kidney, lung, and keratinocytes. The other probe (probe 3) used for hybridization with the same Southern blot (Fig. 3*C*) was a genomic *Hin*cII–*Eco*RI fragment of AF2 shown in Figs. 1 and 2 (nt 631–1405). The PCR products from placenta, kidney, and

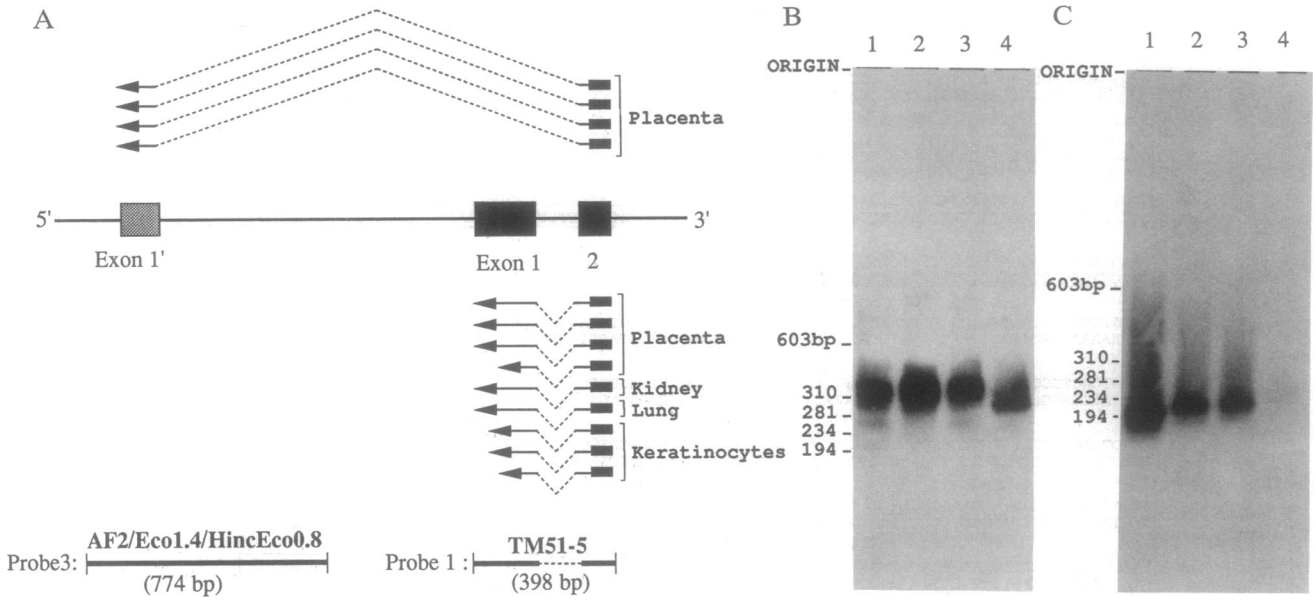

FIG. 3. 5′ RACE using RNAs from placenta, kidney, lung, and keratinocytes. First-strand cDNA was synthesized from RNAs from placenta (lanes 1 in *B* and *C*), kidney (lane 2), lung (lane 3), and keratinocytes (lane 4) as template from primer 1, and poly(A) was added at the 5′ end. The product was used as PCR template. PCR was performed with the hybrid primer and primer 2. Southern blot analysis showed that the PCR products hybridized with TM51-5 (probe 1, 398 bp), the most 5′ end of the cDNA for the human α6(IV) chain (*B*). The same sample was hybridized with probe 3, AF2/Eco 1.4/Hinc Eco 0.8 (774 bp), a *Hin*cII–*Eco*RI fragment of the genomic fragment AF2 (*C*). The exposure time for the film shown in *C* was five times longer than that for the film in *B*. Note that both forms of transcript are present in placenta, kidney, and lung, whereas only one form of transcript is present in keratinocytes (*C*). PCR products were subcloned and complete nucleotide sequences were determined. The nucleotide sequences were compared with those of the genomic fragment shown in Fig. 2. Some of the PCR products were shown to contain sequences of exons 1 and 2, whereas the other products contained sequences of exons 1′ and 2 (*A*). The expected size of PCR products when exon 1′ joins exon 2 is 194 bp (as seen in *C*) and when exon 1 joins exon 2 is 296 bp (*B*). Some clones were fully extended to the 5′ side, whereas others were not. Clones from keratinocyte RNA were significantly shorter than those isolated from placenta, kidney, and lung.

lung RNAs hybridized with probe 3 (Fig. 3*C*), and, interestingly, the product size was significantly smaller than that seen with probe 1 (Fig. 3*B*). This suggested that two kinds of RNAs were expressed in these tissues. The relative amounts of PCR products differed for the various tissues.

To find out how these two RNAs are different, the PCR products were cloned by the TA cloning method (Invitrogen) and sequenced. As expected, all 9 clones (4 from placenta, 1 from kidney, 1 from lung, and 3 from keratinocyte) screened by probe 1 and sequenced represented exons 1 and 2. Relative locations of the clones identified by nucleotide sequence analysis are drawn in Fig. 3*A*. However, nucleotide sequence analyses of the four clones from placenta screened by probe 3 demonstrated that they were the same and that they all contained a 5′ sequence of 124 bp different from the other clones, whereas the 3′ part of the sequence was the same as exon 2. Interestingly, the 5′ sequence was found in the genomic DNA, nt 838–961 in Fig. 2. We refer to this exon as exon 1′ in this report. The transcription initiation site is slightly different from that determined by cDNA analysis (9). We confirmed this initiation site by cloning and sequencing two different clones. Thus, from comparison of the nucleotide sequence between the genomic DNA and the two different kinds of cDNA encoding the α6(IV) chain, we conclude that there are two kinds of transcripts. One starts from exon 1, and exon 1 is spliced to the sequence of exon 2 in the usual manner. The alternative exon sequence, that of exon 1′, is spliced to the sequence of exon 2. These results indicate that the presence of the two different mRNAs is due to the use of the alternative transcription start sites. Alternative promoters are utilized in transcription of α1(IX) genes (17, 18). In the α1(IX) case, an alternative promoter located in intron 6 of the α1(IX) gene transcribes RNA from alternative exon in intron 6 and the alternative exon is spliced to exon 8. This transcription form results in an α1(IX) chain missing almost the entire NC4 noncollagenous domain, when it is translated. But the alternative promoter described in this paper is not far from the other promoter. These closely located alternative promoters result in different signal peptides at its amino terminus (Fig. 4).

Structure of the Promoter Region for α5(IV) and α6(IV) Genes. The complete nucleotide sequence of the upstream region of both *COL4A5* and *COL4A6* is shown in Fig. 2. Comparison between this genomic sequence and the 5′-terminal cDNA sequences of both genes confirmed that AF2 indeed contains the 5′ region of the α5(IV) gene as well as that of the α6(IV) gene. The coding sequences are aligned on opposite strands, indicating that the two genes are transcribed in opposite directions and thus are oriented in a head-to-head fashion. Further, cloning and characterization of the two kinds of cDNAs of different 5′ regions, as mentioned above, indicated that the presence of the two mRNAs is due to the use of the alternative transcription start site as shown in Fig. 1. The transcription initiation site for *COL4A5* and the start sites for the two transcripts generated from *COL4A6* are separated by 442 bp and 1492 bp. This suggests the presence of a common promoter region.

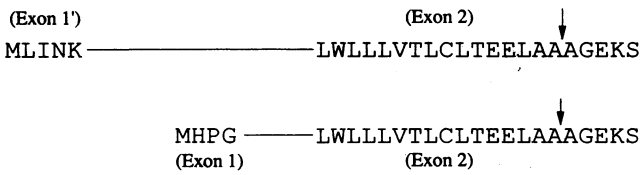

FIG. 4. Two signal peptides starting from exon 1′ and exon 1. Two different signal peptides for the α6(IV) chain are encoded by two different exons controlled by alternative promoters. Starting from the two alternative promoters, either exon 1′ or exon 1 is spliced to exon 2, resulting in two different signal peptides. The two vertical arrows indicate the putative signal peptidase cleavage sites.

The sequence of the promoter region is rich in G and C, and no typical TATA box or related sequences can be found at a location compatible with a TATA-box-type function. However, some other typical sequence motifs are found within the promoter sequence. CCAAT boxes are located about 90 bases and 120 bases upstream of the start site of exon 1′ of the *COL4A6* gene, and a GC box, a potential interaction site with transcription factor Sp1, is located 40 bases upstream of exon 1 of *COL4A6*. A GC box is located in the middle of the bidirectional promoter in *COL4A1* and *COL4A2* (4), but this is not the case for the promoter in *COL4A5* and *COL4A6*. Intriguingly, a recently described CTC box that has been suggested to bind CTC-binding factor (CTCBF) for efficient transcription of *COL4A1* and *COL4A2* (19) was found in the middle of the promoter region (nt 606–614) between exon 1 of *COL4A5* and exon 1′ of *COL4A6*. Both genes could contain elements in their first introns that would seem to be essential for efficient initiation of collagen gene transcription (20).

Exon 1′ of the α6(IV) gene includes 110 bp of the 5′ untranslated region and 14 bp coding for 4⅔ amino acid residues of the signal peptide. On the other hand, exon 1 contains a longer 5′ untranslated region (234 bp) and 11 bp coding for 3⅔ amino acid residues of the different amino terminus of the signal peptide. The two signal peptides are compared in Fig. 4. Exon 2 of the α6(IV) gene contains 50 bp coding for 16⅔ amino acid residues of the remaining signal peptide, indicating that the following exon, no. 3, should start with the coding region of the 7S domain.

It would be interesting to find out whether the genes for α3(IV) and α4(IV) collagen chains have the same regulatory arrangements of their promoter region on chromosome 2q36. We do not know whether expression of the homo- or heterotrimeric collagen IV including the α6(IV) chain is essential for the structural integrity and functional properties of the basement membrane, but we do know that the genes, *COL4A5* and *COL4A6*, coding for their respective subunits are located close to each other on chromosome X and are transcribed from a common bidirectional promoter element. We predict that the common promoter region of collagen IV does not represent an equally functional bidirectional element but may be better understood as two overlapping promoters with shared elements.

We thank Dr. Hidekatsu Yoshioka for constructive discussion and Ms. Aiko Fukutomi for technical assistance. We are very much indebted to Dr. Bjorn Reino Olsen for his continuous interest in the project and critical reading of the manuscript. This work was supported in part by Scientific Research Grant 06454250 from the Ministry of Education, Science, and Culture of Japan; Grant EY07334 from the U.S. National Eye Institute, and a Grant for the Promotion of Science from the CIBA-Geigy Foundation (Japan).

1. Kivirikko, K. I. (1993) *Ann. Med.* **25**, 113–126.
2. Chu, M.-L. & Prockop, D. J. (1992) in *Connective Tissue and Its Heritable Disorders: Molecular, Genetic and Medical Aspects*, eds. Royce, P. M. & Steinmann, B. (Wiley–Liss, New York), pp. 149–166.
3. Griffin, C. A., Emmanuel, B. S., Hansen, J. R., Cavenee, W. K. & Myers, J. C. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 512–516.
4. Poschl, E., Pollner, R. & Kuhn, K. (1988) *EMBO J.* **7**, 2687–2695.
5. Soininen, R., Houtari, M., Hostikka, S. L., Prockop, D. J. & Tryggvason, K. (1988) *J. Biol. Chem.* **263**, 17217–17220.
6. Kamagata, Y., Mattei, M.-G. & Ninomiya Y. (1992) *J. Biol. Chem.* **267**, 23753–23758.
7. Morrison, K. E., Mariyama, M., Yang-Feng, T. L. & Reeders, S. T. (1991) *Am. J. Hum. Genet.* **49**, 545–554.
8. Oohashi, T., Sugimoto, M., Mattei, M.-G. & Ninomiya Y. (1994) *J. Biol. Chem.* **269**, 7520–7526.
9. Zhou, J., Mochizuki, T., Smeets, H., Antignac, C., Laurila, P., de Paepe, A., Tryggvason, K. & Reeders, S. T. (1993) *Science* **261**, 1167–1169.
10. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
11. Overbeck, P. A., Mertino, G. T., Peters, N. K., Cohen, V. H., Moore, G. P. & Kleinsmith, L. J. (1981) *Biochim. Biophys. Acta* **656**, 195–205.
12. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
13. Frohman, M. A., Dush, M. K. & Martin, G. R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8998–9002.
14. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
15. Hattori, M. & Sakaki, Y. (1986) *Anal. Biochem.* **152**, 232–238.
16. Zhou, J., Hertz, J. M., Leinonen, A. & Tryggvason, K. (1992) *J. Biol. Chem.* **267**, 12475–12481.
17. Nishimura, I., Muragaki, Y. & Olsen, B. R. (1989) *J. Biol. Chem.* **264**, 20033–20041.
18. Kong, R. Y., Kwan, K. M., Lau, E. T., Thomas, J. T., Boot-Handford, R. P., Grant, M. E. & Cheah, K. S. (1993) *Eur. J. Biochem.* **213**, 99–111.
19. Schmidt, C., Fischer, G., Kadner, H., Genersch, E., Kuhn, K. & Poschl, E. (1993) *Biochim. Biophys. Acta* **1174**, 1–10.
20. Pollner, R., Fischer, G., Poschl, E. & Kuhn, K. (1990) *Ann. N.Y. Acad. Sci.* **580**, 44–54.