Isolation and Structure of the COL4A6 Gene Encoding the Human α6(IV) Collagen Chain and Comparison with Other Type IV Collagen Genes

(Received for publication, April 25, 1995, and in revised form, August 7, 1995)

Toshitaka Oohashi, Yasuyoshi Ueki, Manabu Sugimoto, and Yoshifumi Ninomiya

From the Department of Molecular Biology and Biochemistry, and the Department of Ophthalmology, Okayama University Medical School, 2-5-1, Shikata-cho, Okayama 700, J apan

The genes COL4A5 and COL4A6, coding for the basement membrane collagen chains, α5(IV) and α6(IV), respectively, are located head-to-head in close proximity on human chromosome Xq22, and COL4A6 is transcribed from two alternative promoters in a tissue-specific fashion (Sugimoto M., Oohashi T., and Ninomiya Y. (1994) Proc. Natl. Acad. Sci. U. S. A. 91, 11679–11683). Immuno-fluorescence studies using α chain-specific antibodies demonstrated that the two genes are expressed in a tissue-specific manner (Ninomiya, Y., Kagawa, M., Iyama, K., Naito, I., Kishiro, Y., Seyer, J. M., Sugimoto, M., Oohashi, T., and Sado, Y. (1995) J. Cell Biol. 130, 1219–1229). We report here for the first time the isolation and the structural organization of the human COL4A6 gene. The entire gene presumably exceeds 200 kilobase pairs and contains 46 exons. Exons 1’ and 1 encode the two different 5’-UTRs and the two aminотerminal parts of the signal peptide. The carboxy part of the signal peptide and the 7 S domain are coded for by the following 6 different exons, 2–7, whereas the exons 7–42 encode the central COL 1 domain, which contains the Gly-X-Y repeats. The last three exons, 43–45, encode the carboxy-terminal NC1 domain. Sizes of more than a half of the exons of the gene are the same as those of COL4A2 but quite different from those of COL4A5. Within the COL4A6 gene we found three CA repeat markers that can be used for allele detection. The detailed structure of the COL4A6 gene and the high heterozygosity microsatellite markers located within the gene will be useful for linkage analysis and familial diagnosis of diseases caused by mutations of this gene.

Epithelial cells present a sheet-like extracellular structure, basement membrane (BM),1 at their basal surfaces. This structure appears as a ~100-nm-thick layer, consisting of type IV collagen, laminin, heparan sulfate proteoglycan, and some other glycoproteins (1). Since BMs are attached to the underlying extracellular matrix by anchoring fibrils formed by type VII collagen and microfibrils, there may be functional interactions between matrix molecules. On the other hand, BMs also play a crucial role in adhesion of epithelial and other types of cells, differentiation of a variety of cells, and tissue repair (2). BMs are known to be affected in certain disease states. For example, the epitope for the circulating antibody for Goodpasture syndrome that causes glomerulonephritis and pulmonary hemorrhage has been shown to be on the carboxy-terminal part of the NC1 domain of the α3(IV) collagen chain (3). Also, a hereditary Alport syndrome characterized by sensorineural hearing loss, progressive glomerulonephritis, and, occasionally, ocular defects has been shown to be caused by mutations of the gene COL4A5, coding for the α5(IV) collagen chain (4).

Type IV collagen was initially thought to be a heterotrimeric molecule of two α1(IV) chains and one α2(IV) chain. But recent biochemical investigation and the molecular biology approach have made possible the identification of additional four different α(IV) chains, α3, α4, α5, and α6 (5–7). However, almost nothing is known about how these chains are involved in the assembly of the individual molecules. The newest chain, α6(IV), was found by characterizing the region upstream of the neighboring gene, COL4A5 (8), and by cross-hybridization with the homologous gene, COL4A4 (7). Of interest is not only that the two most recently discovered genes, COL4A5 and COL4A6, are arranged in a head-to-head fashion on chromosome Xq22 but also that the other set of “new” genes for collagen IV, COL4A3 and COL4A4, are located on chromosome 2 (9, 10) in a presumably similar fashion as the former two (11). Furthermore, α1(IV) and α2(IV) chains are known to exist in all BMs, and α3(IV) and α4(IV) chains appear to co-localize in certain BMs but not all; whereas the α5(IV) and α6(IV) chains are not necessarily distributed together in some tissues (12). Differential expression of the latter two genes could be explained by the transcription from the two alternative promoters in the COL4A6 gene (13). Precise characterization of the COL4A6 gene is essential for studies on pathogenesis of certain diseases such as Alport’s syndrome associated with diffuse esophageal leiomyomatosis (14). In the present study, we describe the isolation and characterization of the COL4A6. The gene harbors 46 exons, spanning at least 200 kb in size. We also determined the sequences of the exon/intron junctions of the entire gene.

MATERIALS AND METHODS

Isolation of Genomic Clones—Total EMBL3-genomic library (HL1111J) from human female leukocytes was purchased from Clontech Laboratories Inc. The library was screened with the previously isolated cDNAs encoding the human α6(IV) collagen chain (7). The
screening of the library was performed by standard procedures (15). 13 purified genomic clones were characterized by genomic mapping, and necessary restriction fragments were subcloned into pBluescript II vectors for further characterization. Genomic DNA preparation was performed by the standard procedures (15).

DNA Sequencing—DNA sequencing was carried out by the dye-terminator method (Perkin Elmer). Universal primers were used for the sequencing reaction. We used a 373S automatic sequencer (Perkin Elmer) for further characterization. Genomic DNA preparation was performed by the standard procedures (15).

RESULTS AND DISCUSSION

Isolation of Genomic Clones—We have previously described (13) that COL4A5 and COL4A6 have a unique arrangement in that they are co-localized on chromosome Xq22 in a head-to-head fashion and appear to share a common bidirectional promoter. In addition we reported a novel observation that the COL4A6 gene is transcribed from two alternative promoters in a tissue-specific manner (13). In this present study, we analyzed the structure of the COL4A6 gene by isolation and characterization of phage clones. We used previously isolated cDNAs, TMS1, 46, 30, and 29, and screened the genomic library. We isolated 50 clones all together, but only 13 of them, AF 2, 4, 17, 18, 27, 30, 31, 33, 39, 40, 47, 48, and YU2 (Fig. 1), that cover exons were characterized further.

### Table I

| Name   | Locus   | CA strand primer | Repeat sequence | GT strand primer |
|--------|---------|------------------|-----------------|------------------|
| A6YU1  | intron 25 | ACA CCA GAG AGC AAA CTT TT | (CA)15AT(CA)15 | GTG TCT CCT CAC CTT GCT CT |
| A6YU2  | intron 31 | ACA GAA CCC TGT GCC TTC TT | (CA)18 | GTA GTC AGC ACT ATC CAG TC |
| A6YU3  | intron 45 | GAT GCA GCT GCT AAA GCT CC | (CA)1AT(CA)15 | CTT AGT CAT CTA ATG GGA TG |

**Fig. 1. Structure of the human α6(IV) gene, COL4A6 and relative location of genomic DNA clones.** The COL4A6 gene contains 46 exons: exon 1 (one white box) and exons 1 to 45 (black boxes). The COL4A5 gene is located next to the COL4A6 gene as indicated by arrows. Analysis of cDNAs revealed the presence of the two transcripts: exon 1 is spliced to exon 2 in one transcript and exon 1 to 2 in the other (13). Top part of the figure shows the only one of the two transcripts, and the relative location of the coding regions is indicated along with the individual exons. The 7S domain is encoded by the first 7 exons. The central COL1 domain containing 25 imperfections is coded for exons 7–43, whereas the last three exons code for the NC1 domain. The gene spans a length of >200 kb. Relative location of the 13 phage clones is shown by short bars. Numbers in parentheses under the bars indicate the length of the individual phage clones. Four ovoid-shaped marks within the gene indicate the locations in the gene that are not covered by the 13 isolated phage clones. Sizes of the EcoRI fragments containing individual exons are shown by numbers under each exon. EcoRI fragments that contain multiple exons are shown by brackets above the numbers. Four exons 4, 12, 29, and 39 contain EcoRI sites, whose locations are marked by E. The area covering exons 38–41 is enlarged in a square to show small EcoRI fragments.
Structure of the COL4A6 Gene—The COL4A6 gene contains 46 exons including exon 1’ at the 5’ end. Since exon 1 and 1’ are alternatively transcribed presumably from alternative promoters (13), we named the following exons from exons 2–45. Fig. 1 shows the relative location of the 46 exons. Since the five 5′ genomic clones did not overlap each other, there were four intronic gaps between the clones; however the fragment sizes for the entire coding region were all isolated. The 13 clones span approximately 150 kb of the structural gene. Several intron regions (introns 2, 3, 5, and 7) are not covered by these 13 clones. Considering the previous data on introns 2 and 3 characterized by pulse-field gel electrophoresis (14) together with the results presented here, we estimate the total length of the COL4A6 gene to be not shorter than 200 kb. The neighboring gene, COL4A5, also represents a long gene, at least 140 kb (20). The IV collagen genes (COL4A1, 100 kb (21); COL4A2, 100 kb (22); COL4A5, >140 kb (20)) seem to be in general larger than fibrillar collagen genes, i.e. COL1A1, 39 kb; COL1A2, 38 kb (23); COL2A1, 35 kb; and COL3A1, 35 kb (24). Some other nonfibrillar collagen genes in other species have been demonstrated to be large, i.e. chicken a(II)X gene, >100 kb (25) and human COL1A1, 140 kb (26).

Exon Sizes and Exon/Intron Boundaries—The exon 1’ encodes the 5’-untranslated region (UTR) of the gene and the 5’ part of the signal peptide (5 amino acid residues). Similarly exon 1 codes for another 5’-untranslated region and another 5’ part of the signal peptide (4 residues), resulting in two different signal peptides (13). The following exons of 2–7 code for the 7S part of the central 1 domain. The carboxyl NC1 domain is encoded by the last three exons, 43, 44, and 45. We did not determine whether there are more exons farther in the 3’ direction. The exons are distributed sparsely at the 5’-half of the gene but rather densely at the 3’-half, as seen in Fig. 1.

The exon sizes vary between 36 and ~980 bp, but if the 5’- and 3’-untranslated region and the NC1 domain are excluded, they vary from 36 to 222 bp (Fig. 2). Exon sizes of the COL4A6 gene are summarized and compared with those of other type IV collagen genes, COL4A5 (20), Col4a2 (27), COL4A2 (28), and COL4A4 (29) (Fig. 3). Exon sizes of the gene are rather small; most of them are smaller than 150 bp. Especially exons coding for only Gly-X-Y repeats, characteristic for collagen, are small. Not many exons represent 54 or 45 bp in size as seen in fibrillar collagen genes (23).

Exon sizes of the COL4A6 gene were compared with those of the other type IV collagen genes. Exons 4–42 all encode Gly-X-Y repeats. Many of them code for imperfect Gly-X-Y repeats. As shown in Fig. 3, sizes of many exons coding for the odd-numbered a(IV) chains, a1, a3, and a5 (29). The NC1 domain of the a6(IV) chain is encoded by three different exons, 43, 44, and 45. This pattern is common for isolated and characterized even-numbered collagen genes, COL4A2, Col4a2, and COL4A4, whereas COL4A1 and COL4A5 encode the NC1 domain by means of five different exons.

In the previous study on the cDNAs we did not notice a very small EcoRI fragment. However, when we compared the nucleotide sequences between cDNA and genomic DNA, we identified a 36-bp EcoRI fragment located in the exon 39 (see Fig. 1). It contains AATTCTGGACCTAAAGGGCCTAAGG. Structure of a Gene Encoding Human α(6)(IV) Collagen Chain 26865
the cDNA sequence in the previous paper (12). This makes amino acid sequence (G)IPGPKGPKGDQG(I) (amino acid residues in parentheses indicate that the 36 bp contained a part of coding nucleotides).

All of the exon/intron boundary sequences are shown in Fig. 2. As highlighted, dinucleotides at the beginning (gt) and ending (ag) of introns are all conserved. The first 15 exons, 3–17, of the COL4A6 gene begin with an intact glycine codon, whereas all of the last 38 exons start with a two-thirds intact glycine codon except for the three exons (22, 23, and 32), which start with the intact glycine codon. Split glycine codons were also found in exons of other collagen genes (30, 31). It is intriguing that the genes all code for collagens with imperfections in the Gly-X-Y repeats and that the glycine codon is almost always split after the first G. This type of split codon is also found in noncollagenous genes that harbor Gly-X-Y repeat sequences with imperfections, such as those encoding mannos-binding protein (32), lung surfactant apoprotein (33), acetylcholinestrase (34), complement C1q B chain (35).

Repetitive Sequences in COL4A6—To find repetitive CA sequences, we hybridized the 5 phage clones, AF4, 18, 17, and 30, and YU2, that cover 50 kb from the 3′ end of COL4A6 (Fig. 1) with a (CA), probe. Three clones, AF18, 17, and 30, were positive. Further analysis revealed three repetitive sequences, A6YU1, A6YU2, and A6YU3 (Table I), which are located in introns 25, 31, and 45, respectively. To determine polymorphic variations of the three microsatellites in the general population, we analyzed genomic DNAs isolated from 50 unrelated Japanese females, since the COL4A6 gene is located on chromosome X. Polymerase chain reaction products from the flanking primers were analyzed on sequencing gels. As indicated in Table II, one of the three repeats, A6YU2, displayed five alleles with 0.76 heterozygosity, and 0.26 PIC (polymorphism information content), whereas the other two, A6YU1 and A6YU3, showed 2 alleles with 0.34 heterozygosity, 0.65 PIC and 4 alleles with 0.42 heterozygosity, and 0.36 PIC, respectively. The combined heterozygosity was 0.82. These results indicate that the three markers are useful to distinguish alleles in relation to the COL4A6 gene.

More than 60 mutations in the COL4A5 gene have been identified in X-linked Alport syndrome patients (36). Recently, a new report showed mutations in COL4A3 and COL4A4 genes from autosomal recessive-type patients (37). Intriguingly, deletions of the 5′ part of both COL4A5 and COL4A6 genes have
also been revealed in seven patients with Alport's syndrome associated with diffuse leiomyomatosis (14); however, how the COL4A6 gene is involved in the pathogenesis of diffuse leiomyomatosis has not yet been clarified. Therefore the high heterozygosity microsatellite markers located within the COL4A6 gene will be a useful tool for linkage analysis and familial diagnosis of the diseases caused by COL4A6 gene mutations.

In conclusion we reported the isolation and structure of the COL4A6 gene, which is aligned together with COL4A5 in head-to-head fashion on chromosome X. The detailed structure of the COL4A6 gene determined in this study will be important for finding mutations in patients with X-linked Alport's syndrome and/or leiomyomatosis and in those with other diseases caused by the mutated gene.

Acknowledgments—We thank Drs. Hidekatsu Yoshikawa and Nobuhiko Matsuo for constructive discussion and continuous support for the project and Aiko Fukutomi, Tsugio Seki, and Yan Ling Wu for technical assistance in isolating and characterizing the genomic fragments.

REFERENCES

1. J. Unoura, L. C., Carneiro, J., Kelley, R. O. (1992) Basic Histology, 7th Ed., Appleton & Lange, East Norwalk, CT
2. Rohrbach, D. H., and Timpl, R. (1993) Molecular and Cellular Aspects of Basement Membranes, Academic Press, New York
3. Kalluri, R., Gattone, V. H., II, Nadkarni, M. E., and Hudson, B. G. (1994) Proc. Natl. Acad. Sci. U. S. A. 91, 6201–6205
4. Barker, D. F., Hostikka, S. L., Zhou, J., Chow, L. T., Oliphant, A. R., Gerken, S., C., Gregory, M. C., Skolnick, M. H., Atkin, C. L., and Tryggvason, K. (1990) Science 248, 1224–1227
5. Hudson, B. G., Reinders, S. T., and Tryggvason, K. (1993) J. Biol. Chem. 268, 26033–26036
6. Hostikka, S. L., Eddy, R. D., Byers, M. G., Hoyt, M. A., Show, T. B., and Tryggvason, K. (1990) Proc. Natl. Acad. Sci. U. S. A. 87, 5972–5976
7. Fujisawa, T., Sugimoto, M., Nishino, Y., and Ninomiya, Y. (1995) J. Biol. Chem. 270, 5720–5726
8. Zhang, J., Mochizuki, T., Smeets, H., Antignac, C., Laurila, P., de Paepe, A., Tryggvason, K., and Reinders, S. T. (1994) Science 261, 1167–1169
9. Mariyama, M., Zheng, K., Yang-Feng, T. L., and Reinders, S. (1992) Genomics 13, 809–813
10. Kamagata, Y., Tanaka, M., and Ninomiya, Y. (1993) J. Biol. Chem. 268, 23735–23738
11. Hlaing, T., Funabiki, K., Togawa, M., Chapo, J. A., and Killen, D. P. (1994) Am. J. Hum. Genet. 54, 125–131
12. Ninomiya, Y., Kagawa, M., Iyama, K., Naito, I., Kishiro, Y., Seyer, J. M., Sugimoto, M., Oohashi, T., and Sado, Y. (1995) J. Cell Biol. 130, 1219–1229
13. Sugimoto, M., Oohashi, T., and Ninomiya, Y. (1994) Proc. Natl. Acad. Sci. U. S. A. 91, 11679–11683
14. Heidet, L., Dahan, K., Zhou, J., Xu, Z., Cochot, P., Gould, J. D. M., Leppig, K. A., Proesman, W., Guyot, C., Guillet, M., Roussel, B., Tryggvason, K., Galder, J.-P., Gubler, M.-C., and Antignac, C. (1995) Hum. Mol. Genet. 4, 99–108
15. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
16. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463–5467
17. Hattori, M., and Sakaki, Y. (1986) Anal. Biochem. 152, 232–238
18. Warman, M. L., Tiller, G. E., Polumbo, P. A., Seldin, M., Frisell, J. M., Knoll, J. H., Cheng, S.-D., and Olsen, B. R. (1993) Genetics 137, 694–698
19. Watanabe, K., and May, P. E. (1989) Am. J. Hum. Genet. 44, 179–190
20. Zhou, J., Leinonen, A., and Tryggvason, K. (1994) J. Biol. Chem. 269, 6608–6614
21. Sorenson, R., Huysman, M., Ganguly, A., Prokop, D. J., and Tryggvason, K. (1989) J. Biol. Chem. 264, 13563–13571
22. Boyd, C. D., Toth-Fejel, S., Gadi, I., Condon, M. R., Kolbe, M., Hagen, I. K., Kurkinen, M., Mackenzie, M. W., and Magenis, E. R. (1988) Am. J. Hum. Genet. 42, 309–314
23. Yamada, Y., Liu, M., Madry, M., Obici, S., and deCroonbrugghe, B. (1984) Nature 310, 333–337
24. Sorenson, R., Huysman, M., Ganguly, A., Prokop, D. J., and Tryggvason, K. (1989) J. Biol. Chem. 264, 13563–13571
25. Hattori, M., and Sakaki, Y. (1986) Anal. Biochem. 152, 232–238
26. Tikka, L., Elomaa, O., Pihlajaniemi, T., and Tryggvason, K. (1991) J. Biol. Chem. 266, 17713–17719
27. Buttice, G., Kayte, P., D’Arminio, J., Vogel, G., and Kurkinen, M. (1990) J. Biol. Chem. 265, 122–128
28. Hostikka, S. L., and Tryggvason, K. (1987) FEBS Lett. 224, 297–305
29. Sugimoto, M., Oohashi, T., Yoshikawa, H., Matsuo, N., and Ninomiya, Y. (1993) FEBS Lett. 330, 122–128
30. Kramer, J. M., Cox, G. N., and Hirsh, D. (1982) Cell 30, 509–606
31. Lozano, G., Ninomiya, Y., Thompson, H., and Olsen, B. R. (1984) Proc. Natl. Acad. Sci. U. S. A. 82, 4050–4054
32. Drickamer, K., and McCrea, V. (1987) J. Biol. Chem. 262, 2582–2589
33. White, R. T., Damm, D., Miller, J., Spratt, K., Shilling, J., Hagwood, S., Benson, B., and Cordell, B. (1985) Nature 317, 361–363
34. Sorensen, J. K., Zunin-Sonkin, D., Avni, A., Hall, L. M., and Spierer, P. (1985) Proc. Natl. Acad. Sci. U. S. A. 82, 1827–1831
35. Reid, K. B. M. (1985) Biochem. J. 233, 729–735
36. Lemmink, H. H., Mochizuki, T., van den Heuvel, L. P. J. W., Schroder, C. H., Barrientos, A., Mennens, L. A. H., van Oost, B. A., Brunner, H. G., Reinders, S. T., and Smeets, H. J. M. (1994) Hum. Mol. Genet. 3, 1269–1273
37. Mochizuki, T., Lemmink, H., Mariyama, M., Antignac, C., Gubler, M.-C., Pirson, Y., Verellen-Dumoulin, C., van den Heuvel, L. P. J. W., Schroder, C. H., Smeets, H. J. M., and Reinders, S. T. (1994) Nature Genet. 8, 77–82
38. Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980) Am. J. Hum. Genet. 32, 314–331