A new FACIT of the collagen family: COL21A1[1]

Jamie Fitzgerald*, John F. Bateman

Cell and Matrix Biology Research Unit, Department of Paediatrics, University of Melbourne and the Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, Vic. 3052, Australia

Received 24 July 2001; accepted 24 July 2001

First published online 29 August 2001

Edited by Veli-Pekka Lehto

Abstract Interrogation of the Human Genome data for sequences related to the von Willebrand factor A-domain module identified a previously unreported 4.1 kb full-length cDNA that is predicted to encode a new member of the collagen superfamily of extracellular matrix proteins, collagen XXI. The domain organization of collagen XXI comprised an N-terminal signal sequence, followed by single von Willebrand factor A-domain and thrombospondin domains, and an interrupted collagen triple helix. The organization of these motifs predict that collagen XXI is a new member of the FACIT collagen sub-family. Expression analysis indicated that COL21A1 mRNA is present in many tissues including heart, stomach, kidney, skeletal muscle and placenta, and radiation hybrid mapping localized the COL21A1 gene to 6p11-12. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Fibril-associated collagen with interrupted triple helices; von Willebrand factor A-domain; Thrombospondin repeat; Interrupted collagen triple helix; Extracellular matrix

1. Introduction

The extracellular matrix (ECM) of connective tissues is a highly regulated tissue-specific network of collagens, non-collagenous proteins and glycoproteins, and proteoglycans. Many of these ECM components are composed of different combinations of characterized protein domains or modules [1]. The completion of the first draft of the Human Genome Project makes it feasible to identify new modular ECM proteins by database homology searching using these conserved modules as probes. It is anticipated that this strategy will help reveal the full complement of modular ECM proteins and proteoglycans that so far have eluded conventional biochemical identification.

One module present in a number of proteins is the type-A domain, first described in von Willebrand factor (reviewed in [2]). ECM components that contain one or more von Willebrand factor A-domains (VA) include fibril-associated collagen with interrupted triple helices (FACIT) collagens XII [3], XIV [4], and XX [5], collagens VI [6,7] and VII [8], matrilins 1–4 (reviewed in [9]), cochlin [10], polydom [11] and nine transmembrane α integrin chains. The VA domain is an independently folding protein module that attains a classic α/β 'Rossman' fold consisting of a parallel β sheet surrounded by amphipathic α helices and a metal-ion-dependent adhesion site (MIDAS) at the C-terminal end of the β sheet [12–14]. Although the role of VA domains has not been precisely defined, they appear to play an important role in protein–protein interactions [15–20].

To further define the molecular and biochemical roles VA domains play in ECM architecture and function, we searched the Human Genome database for novel VA-domain-containing proteins. In this report we describe COL21A1, a new VA-domain-containing collagen with a domain structure that predicts it is a member of the FACIT collagen sub-family expressed in various tissues including heart, stomach, placenta, skeletal muscle, kidney and liver.

2. Materials and methods

2.1. COL21A1 mRNA analysis

A poly(A)$^+$ human multiple tissue expression (MTE) blot containing 76 tissues and cell lines (Clontech) was hybridized to a [^{32}P]dCTP random primer-labelled human COL21A1 polymerase chain reaction (PCR) product amplified from the COL21A1 cDNA clone obtained from the German cDNA Consortium (GenBank accession number NM_030820). The 642 bp probe was generated using COL21A1F1 (^{3247}5′-GCTCCTCAGTCATTTGGAGC-3′^{3266}) and COL21A1R1 (^{3889}5′-TGGACATGCACATASTAAGTG-3′^{3870}) primers designed to anneal within the 3′ untranslated region of the COL21A1 cDNA. The numbering of nucleotides is from the start of the clone including the 5′ untranslated region (see Fig. 1A). The PCR was performed in a 50 μl reaction volume with 100 ng of COL21A1 template cDNA, 1.5 mM $MgCl_2$ and 2.5 units of *Taq* polymerase (Perkin Elmer) at 94°C for 5 min, followed by 36 cycles of 94°C for 30 s, 58°C for 30 s, 72°C for 30 s, and a final extension at 72°C for 7 min. Following overnight hybridization at 65°C in ExpressHyb hybridization solution (Clontech), the blot was washed four times in 1×SSC/1% SDS (w/v) at 65°C, then twice in 1×SSC/0.5% SDS (w/v) at 55°C, exposed to a phosphor-screen for 2–4 days and scanned with a Storm phosphorimager (Molecular Dynamics). The blot was stripped and re-probed with a ubiquitin cDNA to confirm that approximately equal amounts of poly(A)$^+$ RNA were loaded on each dot.

A Northern blot containing poly(A)$^+$ RNA (2 μg per lane) from 12 human tissues (Clontech) was probed with the [^{32}P]dCTP random primer-labelled human COL21A1 PCR product in ExpressHyb hybridization solution. The blot was washed to a stringency of 0.1×SSC/0.1% SDS (w/v) at 55°C and exposed to a phosphor-screen as described above. The blot was stripped and re-probed with a human β-actin control cDNA to demonstrate that each lane contained approximately equal amounts of poly(A)$^+$ RNA.

*Corresponding author. Fax: (61)-3-9345 7997.
E-mail address: fitzgerj@cryptic.rch.unimelb.edu.au (J. Fitzgerald).

[1] The collagen 21 DNA (and protein) sequence has been deposited in GenBank (accession number AF414088).

Abbreviations: FACIT, fibril-associated collagen with interrupted triple helices; ECM, extracellular matrix; VA, von Willebrand factor A-domain; TN, thrombospondin domain; COL, collagenous domain; NC, non-collagenous domain; MTE, multiple tissue expression; PCR, polymerase chain reaction

0014-5793/01/$20.00 © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
PII: S0014-5793(01)02754-5

2.2. Chromosomal mapping of the human COL21A1 gene

The chromosomal location of the COL21A1 gene was determined by radiation hybrid mapping [21] using the Genebridge 4 radiation hybrid panel. Primers and amplification conditions were the same as those used to generate the PCR probe for MTE and Northern analysis. The results were submitted to the Sanger Centre (http://www.sanger.ac.uk/software/rhserver) for scoring using 2-pt RHMAP software [22].

3. Results and discussion

3.1. Identification of COL21A1 cDNA

To identify novel VA-domain ECM proteins, the non-redundant database at the National Center of Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/blast/) was interrogated with the N-terminal VA-domain protein sequence of human matrilin-1 using the blastp program (v2.0). One of the highest scoring hits, with an *E* value of 4×10^{-24}, was a cDNA sequence encoding a previously unreported collagen-like gene (GenBank accession numbers for the protein and nucleotide sequences are NP_110447 and NM_030820, respectively). This cDNA clone, which had been isolated and sequenced in full by the German cDNA consortium (http://www.rzpd.de/), was predicted to contain an uninterrupted protein coding region based on identification of a complete open reading frame and a poly(A) tail at the 3′ end of the clone. To test the accuracy of the reported NM_030820 nucleotide sequence, it was used to interrogate the Human Genome database at NCBI. Multiple hits were scored within a BAC (NT_007336) clone assigned to a working draft of chromosome 6 by the Sanger Centre. Inspection of the aligned NM_030820 cDNA and two Sanger Centre genomic clones (RP4-708F5 and RP1-181C24) revealed a nucleotide discrepancy (A to C change at nucleotide 587) between the cDNA and the two genomic clones, which is predicted to alter the amino acid sequence (Asp to Ala). Interrogation of the Human EST database revealed that in an EST sequence (BG699698) that covered this region, a C nucleotide was present at nucleotide 587, thus confirming the BAC sequences and suggesting that the nucleotide difference in the cDNA clone is either a cloning artefact or a polymorphism. Furthermore, since the alignment of the cDNA and genomic DNA provides information about the position and size of exons and introns, we can predict that this gene is approximately 190 kb in size and is composed of 30 exons.

Initial inspection of the predicted protein sequence revealed that in contains two significant repeating Gly-X-Y motifs of 339 and 112 amino acids in the C-terminus (Figs. 1A and 2A), identifying it as a new member of the collagen family of ECM proteins. We followed the traditional naming system for collagens and assigned it the next sequential number, XXI.

3.2. Features of COL21A1 DNA and predicted protein sequence

The COL21A1 cDNA is 4141 bp in size, exclusive of the poly(A) tail, with a predicted start methionine at nucleotides 203–205 and a TAG stop codon at nucleotides 3074/3076 (Fig. 1A). The deduced open reading frame is 2873 bp in size with 202 bp of 5′UTR and 1066 bp of 3′UTR.

The COL21A1 open reading frame encodes a 957 amino acid protein with a predicted molecular weight of 99 kDa. A 22 amino acid signal sequence with a cleavage site between Ala[22] and Glu[23] is predicted by SignalP signal sequence pre-

A

```
GGGGGCCCGCTGCAGGGAGAACGGACTCCGGGCGGAGGGCAGCCAATCCGTTTCAGCGCA 60
GGTCTTGCTCGGGTTGGGCTTGCCACTGCCTGGAACATACCTGTCCCCCTGGCGCAACAC 120
TCAGCTGGCTGCGACCGCAACCCCGAGCCTGGACACTGCGCCAGGAATCCTAAAACCAAA 180
ATATTAGAACGAAAACAGAAACATGGCTCACTATATTACATTTCTCTGCATGGTTTTGGT 240
                  M  A  H  Y  I  T  F  L  C  M  V  L  V  13
GCTGCTTCTTCAGAATTCTGTGTTAGCTGAAGATGGGGAAGTAAGATCAAGTTGTCGTAC 300
 L  L  L  Q  N  S  V  L  A  E  D  G  E  V  R  S  S  C  R  T  33
TGCTCCGACAGATTTAGTTTTCATCTTAGATGGCTCTTATAGTGTTGGCCCAGAAAACTT 360
 A  P  T  D  L  V  F  I  L  D  G  S  Y  S  V  G  P  E  N  F  53
TGAAATAGTGAAAAAGTGGCTTGTCAATATCACAAAAAACTTTGACATAGGGCCGAAGTT 420
 E  I  V  K  K  W  L  V  N  I  T  K  N  F  D  I  G  P  K  F  73
TATTCAAGTTGGAGTGGTTCAATATAGTGACTACCCTGTGCTGGAGATTCCTCTCGGAAG 480
 I  Q  V  G  V  V  Q  Y  S  D  Y  P  V  L  E  I  P  L  G  S  93
CTATGATTCAGGAGAACATTTGACGGCAGCAGTGGAATCCATACTCTACTTAGGAGGAAA 540
 Y  D  S  G  E  H  L  T  A  A  V  E  S  I  L  Y  L  G  G  N  113
CACAAAGACAGGGAAGGCCATCCAGTTTGCGCTCGATTACCTTTTTGCCAAGTCCTCACG 600
 T  K  T  G  K  A  I  Q  F  A  L  D  Y  L  F  A  K  S  S  R  133
ATTTCTGACTAAGATAGCAGTGGTACTTACGGATGGCAAGTCCCAAGATGACGTCAAGGA 660
 F  L  T  K  I  A  V  V  L  T  D  G  K  S  Q  D  D  V  K  D  153
TGCAGCTCAAGCAGCAAGAGATAGTAAGATAACATTATTTGCTATTGGTGTTGGTTCAGA 720
 A  A  Q  A  A  R  D  S  K  I  T  L  F  A  I  G  V  G  S  E  173
AACAGAAGATGCCGAACTTAGAGCTATTGCCAACAAGCCTTCGTCTACTTATGTGTTTTA 780
 T  E  D  A  E  L  R  A  I  A  N  K  P  S  S  T  Y  V  F  Y  193
TGTGGAAGACTATATTGCAATATCCAAAATAAGGGAAGTGATGAAGCAGAAACTTTGTGA 840
 V  E  D  Y  I  A  I  S  K  I  R  E  V  M  K  Q  K  L  C  E  213
AGAATCTGTCTGTCCAACACGAATTCCAGTGGCAGCTCGTGATGAAAGGGGATTTGATAT 900
 E  S  V  C  P  T  R  I  P  V  A  A  R  D  E  R  G  F  D  I  233
TCTTTTGGGTTTAGATGTAAATAAAAAGGTTAAGAAAAGAATACAGCTTTCACCAAAAAA 960
 L  L  G  L  D  V  N  K  K  V  K  K  R  I  Q  L  S  P  K  K  253
GATAAAAGGATATGAAGTAACATCAAAAGTTGATTTATCAGAACTCACAAGCAATGTTTT 1020
 I  K  G  Y  E  V  T  S  K  V  D  L  S  E  L  T  S  N  V  F  273
CCCAGAAGGTCTTCCTCCATCATATGTATTTGTGTCTACTCAAAGATTTAAAGTCAAGAA 1080
 P  E  G  L  P  P  S  Y  V  F  V  S  T  Q  R  F  K  V  K  K  293
AATTTGGGATTTATGGAGAATATTAACTATTGATGGAAGGCCACAAATAGCAGTTACCTT 1140
 I  W  D  L  W  R  I  L  T  I  D  G  R  P  Q  I  A  V  T  L  313
AAATGGTGTGGACAAAATCTTATTATTTACAACAACCAGCGTAATTAATGGCTCACAAGT 1200
 N  G  V  D  K  I  L  L  F  T  T  T  S  V  I  N  G  S  Q  V  333
GGTTACCTTTGCTAACCCTCAAGTTAAGACGTTGTTTGATGAAGGCTGGCACCAAATTCG 1260
 V  T  F  A  N  P  Q  V  K  T  L  F  D  E  G  W  H  Q  I  R  353
TCTCTTAGTAACAGAACAAGATGTGACTTTGTATATTGATGACCAACAAATTGAAAACAA 1320
 L  L  V  T  E  Q  D  V  T  L  Y  I  D  D  Q  Q  I  E  N  K  373
GCCCTTACATCCAGTTTTAGGGATCTTGATCAATGGGCAAACCCAAATTGGAAAATATTC 1380
 P  L  H  P  V  L  G  I  L  I  N  G  Q  T  Q  I  G  K  Y  S  393
TGGAAAAGAAGAAACTGTTCAGTTTGATGTCCAAAAGTTGCGAATCTACTGTGACCCAGA 1440
 G  K  E  E  T  V  Q  F  D  V  Q  K  L  R  I  Y  C  D  P  E  413
ACAGAACAACCGGGAGACAGCATGTCAGATTCCTGGATTTAATGGAGAGTGCCTTAATGG 1500
 Q  N  N  R  E  T  A  C  E  I  P  G  F  N  G  E  C  L  N  G  433
TCCCAGTGATGTAGGTTCAACTCCAGCTCCCTGTATTTGTCCTCCGGGAAAACCAGGACT 1560
 P  S  D  V  G  S  T  P  A  P  C  I  C  P  P  G  K  P  G  L  453
TCAAGGCCCCAAAGGTGACCCTGGACTGCCTGGGAACCCTGGCTACCCTGGACAACCTGG 1620
 Q  G  P  K  G  D  P  G  L  P  G  N  P  G  Y  P  G  Q  P  G  473
TCAAGATGGTAAGCCTGGATATCAGGGAATTGCAGGGACACCAGGTGTTCCAGGATCTCC 1680
 Q  D  G  K  P  G  Y  Q  G  I  A  G  T  P  G  V  P  G  S  P  493
AGGAATACAAGGAGCTCGAGGACTACCAGGTTACAAAGGAGAACCAGGGCGAGATGGTGA 1740
 G  I  Q  G  A  R  G  L  P  G  Y  K  G  E  P  G  R  D  G  D  513
CAAGGGTGATCGTGGACTTCCTGGTTTTCCTGGGCTTCATGGCATGCCAGGATCAAAGGG 1800
 K  G  D  R  G  L  P  G  F  P  G  L  H  G  M  P  G  S  K  G  533
TGAAATGGGTGCCAAAGGAGACAAAGGATCACCTGGATTTTATGGCAAAAAGGGTGCAAA 1860
 E  M  G  A  K  G  D  K  G  S  P  G  F  Y  G  K  K  G  A  K  553
AGGTGAAAAGGGAATGCTGGCTTCCCTGGCCTCCCTGGACCTGCTGGAGAACCAGGAAG 1920
 G  E  K  G  N  A  G  F  P  G  L  P  G  P  A  G  E  P  G  R  573
ACATGGAAAGGATGGATTAATGGGTAGTCCCGGTTTCAAGGGAGAAGCAGGATCCCCTGG 1980
 H  G  K  D  G  L  M  G  S  P  G  F  K  G  E  A  G  S  P  G  593
TGCTCCGGGGCAGGATGGAACACGGGGAGAGCCTGGAATCCCAGGATTTCCTGGAAACCG 2040
 A  P  G  Q  D  G  T  R  G  E  P  G  I  P  G  F  P  G  N  R  613
AGGATTAATGGGCCAAAAGGGAGAAATTGGGCCTCCAGGACAGCAAGGAAAAAAAGGAGC 2100
 G  L  M  G  Q  K  G  E  I  G  P  P  G  Q  Q  G  K  K  G  A  633
CCCAGGGATGCCTGGTTTAATGGGAAGCAATGGCTCACCAGGCCAGCCTGGAACACCGGG 2160
 P  G  M  P  G  L  M  G  S  N  G  S  P  G  Q  P  G  T  P  G  653
ATCTAAGGGAAGCAAAGGTGAACCTGGAATTCAAGGGATGCCTGGGGCTTCAGGGCTCAA 2220
 S  K  G  S  K  G  E  P  G  I  Q  G  M  P  G  A  S  G  L  K  673
GGGAGAACCAGGAGCAACGGGTTCCCCAGGAGAACCAGGATACATGGGTTACCCGGGAT 2280
 G  E  P  G  A  T  G  S  P  G  E  P  G  Y  M  G  L  P  G  I  693
TCAAGGAAAAAAGGGGACAAAGGAAATCAAGGTGAAAAAGGTATTCAGGGTCAAAAGGG 2340
 Q  G  K  K  G  D  K  G  N  Q  G  E  K  G  I  Q  G  Q  K  G  713
AGAAAATGGAAGACAGGGAATTCCAGGGCAACAGGGAATTCAAGGCCATCATGGTGCAAA 2400
 E  N  G  R  Q  G  I  P  G  Q  Q  G  I  Q  G  H  H  G  A  K  733
AGGAGAGAGAGGTGAAAAGGGAGAACCTGGTGTCCGAGGTGCCATTGGATCAAAAGGAGA 2460
 G  E  R  G  E  K  G  E  P  G  V  R  G  A  I  G  S  K  G  E  753
ATCTGGGGTGGATGGCTTGATGGGGCCCGCAGGTCCTAAGGGGCAACCTGGGGATCCAGG 2520
 S  G  V  D  G  L  M  G  P  A  G  P  K  G  Q  P  G  D  P  G  773
TCCTCAGGGACCCCCAGGTTTGGATGGGAAGCCCGGAAGAGAGTTTTCAGAACAATTTAT 2580
 P  Q  G  P  P  G  L  D  G  K  P  G  R  E  F  S  E  Q  F  I  793
TCGACAAGTTTGCACAGATGTAATAAGAGCCCAGCTACCAGTCTTACTTCAGAGTGGAAG 2640
 R  Q  V  C  T  D  V  I  R  A  Q  L  P  V  L  L  Q  S  G  R  813
AATTAGAAATTGTGATCATTGCCTGTCCCAACATGGCTCCCCGGGTATTCCTGGGCCACC 2700
 I  R  N  C  D  H  C  L  S  Q  H  G  S  P  G  I  P  G  P  P  833
TGGTCCGATAGGCCCAGAGGGTCCCAGAGGATTACCTGGTTTGCCAGGAAGAGATGGTGT 2760
 G  P  I  G  P  E  G  P  R  G  L  P  G  L  P  G  R  D  G  V  853
TCCTGGATTAGTGGGTGTCCCTGGACGTCCAGGTGTCAGAGGATTAAAAGGCCTACCAGG 2820
 P  G  L  V  G  V  P  G  R  P  G  V  R  G  L  K  G  L  P  G  873
AAGAAATGGGGAAAAAGGGAGCCAAGGGTTGGGTATCCTGGAGAACAAGGTCCTCCTGG 2880
 R  N  G  E  K  G  S  Q  G  F  G  Y  P  G  E  Q  G  P  P  G  893
TCCCCCAGGTCCAGAGGGCCCTCCTGGAATAAGCAAAGAAGGTCCTCCAGGAGACCCAGG 2940
 P  P  G  P  E  G  P  P  G  I  S  K  E  G  P  P  G  D  P  G  913
TCTCCCTGGCAAAGATGGAGACCATGGAAAACCTGGAATCCAAGGGCAACCAGGCCCCCC 3000
 L  P  G  K  D  G  D  H  G  K  P  G  I  Q  G  Q  P  G  P  P  933
AGGCATCTGCGACCCATCACTATGTTTTAGTGTAATTGCCAGAAGAGATCCGTTCAGAAA 3060
 G  I  C  D  P  S  L  C  F  S  V  I  A  R  R  D  P  F  R  K  953
AGGACCAAACTATTAGTGTCTGATGCCTCATTCAGCAGCCTAGGCATGGTGCTTTTTCTG 3120
 G  P  N  Y  *                                                 957
TGGTCTTTTGCATCTCAGGAAGATAACCAACAGTATCCCTTGAAAAGAAACTTAAGTACC 3180
TCGGTGTTTTATTTTTTTTTTCTTATGGAAAAAAATATAAAAGATCACATATACTGATT 3240
TTAAAGGCTCCCTCAGTCATTTGGAGCCCTTGGATTAGCAGCATTAATTAAATCTCAAGGG 3300
TTTCTTGTAAAGTCCATTTATGTTAATCAAAGTTGAATATAAAAATCCACCATTGCCTGT 3360
TAGCCAGTCAGTTTTAGTCACTGTGAAATATTTCACATTCAGCCTCCATGCAGTAGAGAT 3420
TTGAGTTTAATTTCATGTCCATGTGACTTTCATGTTTCCTATCTCATAGCTCATGCTACT 3480
ACATAAGCCAAAACATGTATCTCATCATTGGAAGTAAGATCAGGGCTGATATTCACCTGG 3540
GATAGACAGTATTGGTGAACTACTCATTTACTACAGTGTCTCAGCCTTGATAAAGGGCAG 3600
TGGATTGCCTGTTGTTCGGTGTTGTGAATAGCACCTCTGAATAAGATTAGAGTGTTTCTT 3660
AATTCATTTCAAACTCTAAAATTAGATTAATGGTGGTGCTAAGAAAGAGTATTAATTACT 3720
TTGGGAATGGTCAAAATTAACATTAAAAACATTTTAGACAAAAAGTTTCATTGTACATTC 3780
AAAGAAAATGTAAGTTTGGAAGTACTAAAAGACTATTTTATACTTGTTGATTAATCGGAA 3840
TGTTTGTTGTATGCCTTCATTTTCCATTTCACTTATATGTGCATGTCCATATATGTTAAT 3900
TTTCATTGTAGCAAAGCTAATGGAAATAAAGCTAATGCTCTAGTTGAAAGAAAAGGAAAA 3960
CTCCTGAAATCCTAGAATGTCTTGTTATTTTTAGCTGACTGTAAAATTATTATGAACAGTC 4020
TTTGTGTATTGTGCTTAATGCTTTTGTAAGAAACAGAATTTGAAATATTTCATCCTTGTC 4080
ATGCTCAAAATTTTGTTACATGCTTGTTATTCAGAGTATAATAAAGTTTTGTACAGGCCT 4140
GAAAAAAAAAAAAAAAAAAA - 4160
```

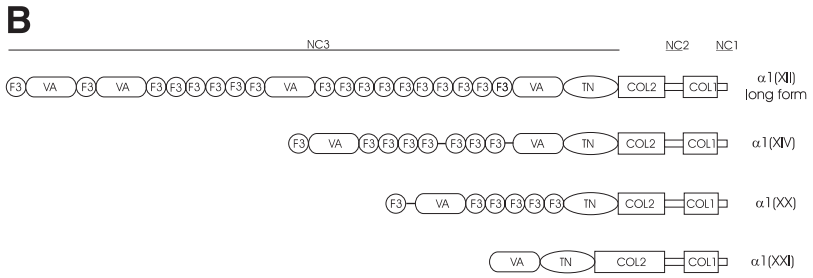

Fig. 1. Sequence and modular organization of collagen. A: Nucleotide and deduced amino acid sequence of COL21A1 cDNA. The predicted signal sequence is in italic type, the stop codon is marked with an asterisk and a potential polyadenylation site in the 3′ untranslated region is underlined. Putative *N*- (Asn^{66} and Asn^{329}) and *O*-linked (Ser^{439}, Thr^{440}, Thr^{486}, Ser^{645}, Thr^{651}, Thr^{679}, Ser^{826}, Ser^{904}) glycosylation sites are underlined and C-terminal cysteine residues (Cys^{936} and Cys^{941}), conserved in all FACIT collagens, are in bold type. B: Modular structure of FACIT collagen XII (long form), XIV, XX and XXI chains showing positions of VA, fibronectin type III repeat (F3), TN (TSP) and interrupted collagen triple helix domains (drawn using standard symbols [35]).

diction software (http://genome.cbs.dtu.dk/services/signalp-2.0) [23] (data not shown). The signal sequence is followed by a VA domain of approximately 170 amino acids from amino acids 35–194, which contains a putative MIDAS [13], and adjacent to the VA domain is a thrombospondin (TN) domain spanning amino acids 229–412. At the C-terminal end of the molecule distal to amino acid 449 are two collagen triple helical domains (COL1 and COL2) separated by short non-collagenous interruption (NC2). At the C-terminus is another short non-collagenous domain (NC1) (Fig. 2A). There are putative *N*-linked glycosylation sites at Asn^{66} and Asn^{329} that fit the consensus sequence NxS/T and potential *O*-linked sites at Ser^{439}, Thr^{440}, Thr^{486}, Ser^{645}, Thr^{651}, Thr^{679}, Ser^{826} and Ser^{904}, as predicted by NetOGlyc v2.0 software (http://genome.cbs.dtu.dk/services/netoglyc) [24] (Fig. 1A).

3.3. Similarity of collagen XXI to the FACIT collagen sub-family members

The modular organization and amino acid sequence identity strongly suggest that collagen XXI belongs to the FACIT sub-family of collagens, which include as members collagens XII, XIV, XVI, XIX (reviewed in [25]) and XX [5]. The key structural feature which defines the FACIT collagens is a pair of highly conserved Cys residues separated by four amino acids at the NC1–COL1 boundary (Fig. 2A). A second sequence feature common to all FACIT collagens in the presence of two imperfections in the Gly-X-Y repeat structure within the COL2 domain [26]. Collagen XXI contains both the conserved Cys residues at Cys^{936} and Cys^{941}, and the two imperfections at amino acids 882–885 and 902–906. Another notable similarity between collagen XXI and the other FACIT collagens is the size of the NC1, COL1, NC2 and COL2 domains of collagen XXI, which are broadly similar to those in the other FACIT collagens (Table 1).

The NC3 domain of collagen XXI is composed of single copies of VA and TN domains. All FACIT collagens contain a TN domain immediately N-terminal to the NC2 domain, but only collagens XII, XIV and XX contain one or more VA domains. The TN domain, which is predicted to attain an anti-parallel β-sheet structure, is composed of nine β strands and is believed to be involved in molecular recognition [27]. The TN repeat of collagen XXI has the most identity to those present in collagens XII, XIV and XX and is less related to the sub-group of FACIT collagens that include IX, XVI and XIX (not shown). The other feature of collagen XXI is the presence of a VA domain, which is only found in FACIT collagens XII, XIV and the recently described collagen XX (Fig. 2B). All amino acids within the MIDAS motif which have been found to be critical for ion-binding, Asp^{40}, Ser^{42}, Ser^{44} and Thr^{113}, are conserved in collagen XXI, although further studies are required to directly demonstrate a functional MIDAS motif. In addition, the overall arrangement of α helices and β sheets that form the important secondary structure framework that is shared between all VA-like domains is conserved in collagen XXI. Thus, we conclude that collagen XXI is member of the FACIT collagen sub-family and is most closely related to the VA-domain-containing FACIT collagens XII, XIV and XX. The currently accepted classification of collagens as members of the FACIT family has been based on homology to the prototypical FACIT collagen IX (reviewed in [25]). The original defining functional feature of the FACIT collagens, association with interstitial collagen

Table 1
Amino acid length of interrupted collagen triple helix domains in FACIT collagens

Collagen chain	NC1 domain	COL1 domain	NC2 domain	COL2 domain
FACIT collagens containing VA domains				
Human α1(XII)	19	103	43	152
Chicken α1(XIV)	119/88[a]	106	43	152
Chicken α1(XX)	15	102	45	154
Human α1(XXI)	21	112	39	339
FACIT collagens lacking VA domains				
Human α1(IX)	30	115	30	339
Human α2(IX)	25	115	30	339
Human α3(IX)	22	112	31	339
Human α1(XVI)	26	106	39	422
Human α1(XIX)	19	70	44	168

[a]Collagen XIV has an alternatively spliced NC1 domain [25].

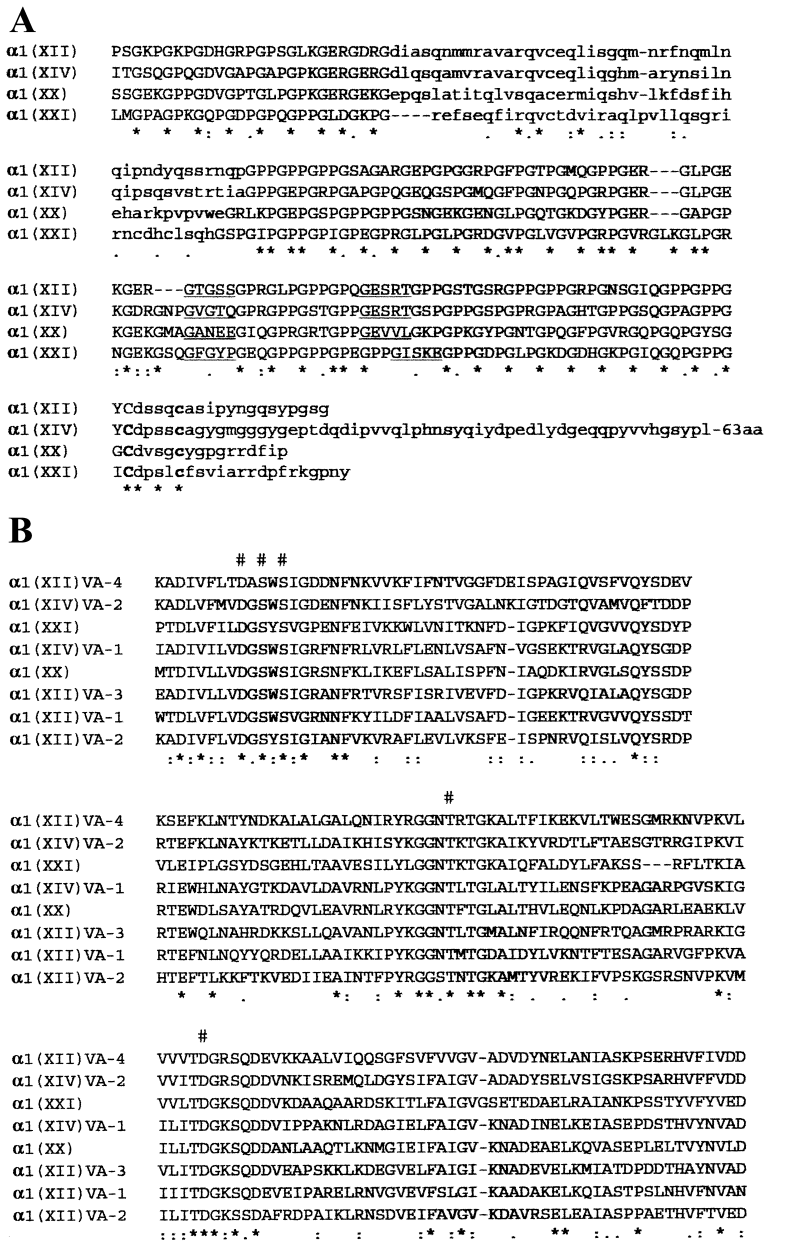

Fig. 2. Alignment of the interrupted collagen triple helix and VA-domains of collagen XXI with homologous domains in other FACIT collagens. Sites where amino acids are identical in all sequences are marked with an asterisk, conserved substitutions are marked with a colon and semi-conserved substitutions with a full-stop. Alignments were performed using CLUSTALW (http://www.ch.embnet.org/software/clustalw.html) [36] with some changes performed visually. A: Alignment of C-terminal interrupted collagenous region containing NC1, COL1, NC2 and COL2 domains from FACIT collagens XII, XIV, XX and XXI. Sequences are human α1(XII) (GenBank accession number Q99715, amino acids 2870–3063) and α1(XXI) (759–957), and chicken α1(XIV) (CAA50064, 1592–1825) and α1(XX) (AAK58847, 1281–1472) chains. The complete COL1 and C-terminal end of the COL2 domains are shown in upper case, and the NC1 and NC2 are in lower case. Conserved Cys residues are in bold type and conserved imperfections in the NC1 domain are underlined. B: Alignment of VA domains from FACIT collagens XII, XIV, XX and XXI. The sequences are human α1(XII) VA domain-1 (amino acids 138–300), -2 (438–599), -3 (1197–1357) and -4 (2321–2483 and α1 (XXI) (35–194), and chicken α1(XIV) VA domain-2 (1040–1199) and α1(XX) (239–402). Conserved amino acids important for the MIDAS [13] are indicated (#).

fibrils, has been demonstrated for collagens IX, XII and XIV, but not for collagens XVI, XIX and XX [5] (reviewed in [25,28]). Likewise, we have yet to demonstrate that collagen XXI can interact with other collagen types, but based on its homology with collagens XII and XIV, we suggest that this interaction is likely to occur.

3.4. Expression of COL21A1 mRNA

To determine the pattern of COL21A1 mRNA expression, a MTE blot containing poly(A)$^+$ RNA from 76 human and fetal tissues and cell lines was hybridized to a COL21A1-specific PCR product (Fig. 3A). COL21A1 mRNA was present in a variety of tissues, but expression was highest in adult and fetal heart, placenta, stomach, jejunum, skeletal muscle, kidney, lung, pancreas and lymph node. Interestingly, there was an asymmetric distribution of COL21A1 mRNA within the heart, with the right ventricle and atrium expressing higher levels compared to the left ventricle and atrium. The apex

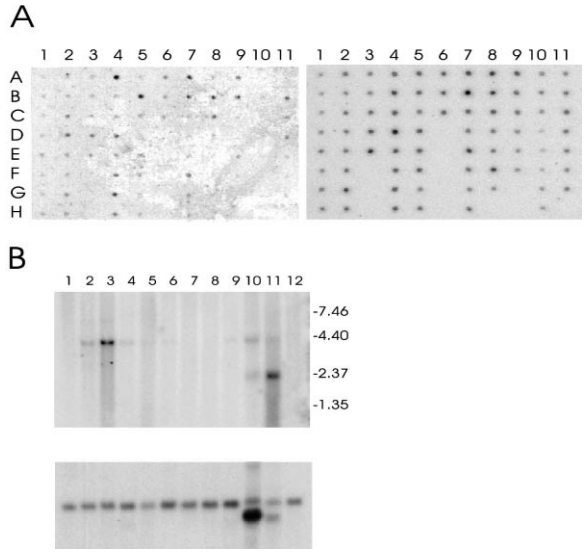

Fig. 3. Expression pattern of COL21A1 mRNA in human adult and fetal tissues and cell lines. A: MTE analysis. Probes for COL21A1 (left) and a ubiquitin control (right) were hybridized to a human MTE array of poly(A)$^+$ RNA from whole brain (A1), cerebral cortex (B1), frontal lobe (C1), parietal lobe (D1), occipital lobe (E1), temporal lobe (F1), paracentral gyrus (G1), pons (H1), left cerebellum (A2), right cerebellum (B2), corpus callosum (C2), amygdala (D2), caudate nucleus (E2), hippocampus (F2), medulla oblongata (G2), putamen (H2), substantia nigra (A3), nucleus accumbens (B3), thalamus (C3), pituitary gland (D3), spinal cord (E3), heart (A4), aorta (B4), left atrium (C4), right atrium (D4), left ventricle (E4), right ventricle (F4), inter-ventricular septum (G4), apex (H4), esophagus (A5), stomach (B5), duodenum (C5), jejunum (D5), ileum (E5), ilocecum (F5), appendix (G5), ascending colon (H5), transverse colon (A6), descending colon (B6), rectum (C6), kidney (A7), skeletal muscle (B7), spleen (C7), thymus (D7), peripheral blood leukocytes (E7), lymph node (F7), bone marrow (G7), trachea (H7), lung (A8), placenta (B8), bladder (C8), uterus (D8), prostate (E8), testis (F8), ovary (G8), liver (A9), pancreas (B9), adrenal gland (C9), thyroid gland (D9), salivary gland (E9), mammary gland (F9), leukemia HL-60 cells (A10), HeLa S3 cells (B10), leukemia K-562 (C10), leukemia MOLT-4 cells (D10), Burkitt's lymphoma (Raji) (E10), Burkitt's lymphoma (Daudi) (F10), SW480 colorectal adenocarcinoma (G10), A549 lung carcinoma (H10), fetal brain (A11), fetal heart (B11), fetal kidney (C11), fetal liver (D11), fetal spleen (E11), fetal thymus (F11), fetal lung (G11). B: Northern blot analysis. Probes for COL21A1 (top panel) and a β-actin control (bottom panel) were hybridized to a multiple tissue Northern blot containing approximately 2 μg of poly(A)$^+$ RNA from peripheral blood leukocytes (1), lung (2), placenta (3), small intestine (4), liver (5), kidney (6), spleen (7), thymus (8), colon (9), skeletal muscle (10), heart (11) and brain (12). RNA markers (kb) are indicated on the right. In most tissues, β-actin is present as a 2.0 kb transcript, except in skeletal muscle and heart where an additional 1.8 kb transcript is present.

and septum of the heart also expressed high mRNA levels. Re-probing the blot with a ubiquitin cDNA confirmed that approximately equal amounts of mRNA were present on each dot. This pattern of COL21A1 mRNA expression closely corresponds to that of the three chains of collagen V, COL5A1, COL5A2 and COL5A3 [29]. Since the MTE analysis does not provide information about transcript size, we also probed a Northern blot containing mRNA from some of the human tissues that were positive for COL21A1 on the MTE blot (Fig. 3B). COL21A1 mRNA was detected in heart, placenta, jejunum, skeletal muscle, colon, kidney, liver, lung and absent or present at low levels in brain, spleen, thymus and peripheral leukocytes, confirming the pattern of expression determined by MTE analysis. In most of these tissues, a 4.2 kb transcript was present, which is in good agreement with the size of the 4142 bp NM_030820 cDNA clone, confirming that it represents the full-length sequence. Interestingly, in heart and skeletal muscle, an additional 2.4 kb band was present that probably represents a splicing variant of the COL21A1 gene. Densitometric estimation using the phospho-imager indicated that in skeletal muscle, the two splicing variants were present in approximately equal amounts, but in heart, the 2.4 kb variant was 30-fold more abundant than the 4.2 kb version. When the Northern blot was re-probed with β-actin cDNA, a 2.0 kb transcript was detected in all tissues, confirming that approximately equal amounts of RNA were present in each lane.

In summary, analysis of the predicted amino acid sequence and the domain structure indicates that collagen XXI belongs to the FACIT sub-family of collagens and is the smallest member of this group reported to date. The mRNA expression data demonstrated that collagen XXI is present in tissues containing abundant ECM and in particular, in tissues expressing a muscle phenotype such as heart, skeletal muscle, and smooth muscle including stomach and jejunum, and in placenta, which has a large blood vessel network. These tissue are also enriched for collagen I and at least two members of the FACIT collagen family, collagens XII and XIV, have been shown to co-localize with collagen I [30,31] although a direct interaction has yet to be demonstrated [32]. This data raise the exciting possibility that co-expression of collagen XXI with collagen I in muscle and other tissues may by analogy with some of the other FACIT collagens, playing a role in the organization of interstitial collagen fibrils.

3.5. Chromosomal assignment of human COL21A1 gene

COL21A1 gene location was determined by radiation hybrid mapping [21], using PCR analysis of the Genebridge 4 radiation hybrid panel. COL21A1 mapped to marker AFM205yc7 with a maximum LOD score of 9.6, which places the gene on chromosome 6p11-12. While no characterized connective tissue disorder maps to 6p11-12, the importance of FACIT collagens in ECM structure and function has been demonstrated by the characterization of mutations in collagen IX (COL9A2 and COL9A3) in multiple epiphyseal dysplasia [33,34]. The marker, AFM205yc7, is a CA-repeat polymorphism with a maximum heterozygosity of 0.8298, and will be useful in screening families for linkage between COL21A1 and potential disease phenotypes.

Acknowledgements: This work was supported by grants from the National Health and Medical Research Council of Australia and the Murdoch Childrens Research Institute. We thank the German cDNA Consortium for generously providing cDNA clone NM_030820.

References

[1] Engel, J., Efimov, V.P., and Maurer, P. (1994) Dev. Suppl., 35–42.
[2] Colombatti, A., Bonaldo, P. and Doliana, R. (1993) Matrix 13, 297–306.
[3] Yamagata, M., Yamada, K.M., Yamada, S.S., Shinomura, T., Tanaka, H., Nishida, Y., Obara, M. and Kimata, K. (1991) J. Cell Biol. 115, 209–221.
[4] Trueb, J. and Trueb, B. (1992) Eur. J. Biochem. 207, 549–557.
[5] Koch, M., Foley, J.E., Hahn, R., Zhou, P., Burgeson, R.E., Gerecke, D.R. and Gordon, M.K. (2001) J. Biol. Chem. 276, 23120–23126.

[6] Hayman, A.R., Koppel, J., Winterhalter, K.H. and Trueb, B. (1990) J. Biol. Chem. 265, 9864–9868.
[7] Chu, M.L., Zhang, R.Z., Pan, T.C., Stokes, D., Conway, D., Kuo, H.J., Glanville, R., Mayer, U., Mann, K., Deutzmann, R. and Timpl, R. (1990) EMBO J. 9, 385–393.
[8] Parente, M.G., Chung, L.C., Ryyanen, M., Woodley, D.T., Wynn, K.C., Bauer, E.A., Mattei, M.-G., Chu, M.-L. and Uitto, J. (1991) Proc. Natl. Acad. Sci. USA. 88, 6931–6935.
[9] Deak, F., Wagener, R., Kiss, I. and Paulsson, M. (1999) Matrix Biol. 18, 55–64.
[10] Robertson, N.G., Skvorak, A.B., Yin, Y., Weremowicz, S., Johnson, K.R., Kovatch, K.A., Battey, J.F., Bieber, F.R. and Morton, C.C. (1997) Genomics 46, 345–354.
[11] Gilges, D., Vinit, M.A., Callebaut, I., Coulombel, L., Cacheux, V., Romeo, P. and Vigon, I. (2000) Biochem. J. 352, 49–59.
[12] Emsley, J., Cruz, M., Handin, R. and Liddington, R. (1998) J. Biol. Chem. 273, 10396–10401.
[13] Lee, J.-O., Rieu, P., Arnaout, M.A. and Liddington, R. (1995) Cell 80, 631–638.
[14] Qu, A. and Leahy, D.J. (1995) Proc. Natl. Acad. Sci. USA 92, 10277–10281.
[15] Pareti, F.I., Niiya, K., McPherson, J.M. and Ruggeri, Z.M. (1987) J. Biol. Chem. 262, 13835–13841.
[16] Hoylaerts, M.F., Yamamoto, Y., Nuyts, K., Vreys, I., Deckmyn, H. and Vermylen, J. (1997) Biochem. J. 324, 185–191.
[17] Kamata, T., Liddington, R.C. and Takada, Y. (1999) J. Biol. Chem. 274, 32108–32111.
[18] Specks, U., Mayer, U., Nischt, R., Spissinger, T., Mann, K., Timpl, R., Engel, J. and Chu, M.-L. (1992) EMBO J. 11, 4281–4290.
[19] Fitzgerald, J., Morgelin, M., Selan, C., Wiberg, C., Keene, D.R., Lamande, S.R. and Bateman, J.F. (2001) J. Biol. Chem. 276, 187–193.
[20] Chen, Q., Zhang, Y., Johnson, D.M. and Goetinck, P.F. (1999) Mol. Biol. Cell 10, 2149–2162.
[21] Walter, M.A., Spillet, D.J., Thomas, P., Weissenbach, J. and Goodfellow, P.N. (1994) Nat. Genet. 7, 22–28.
[22] Boehnke, M., Lange, K. and Cox, D.R. (1991) Am. Hum. Genet. 49, 1174–1188.
[23] Nielson, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Protein Eng. 10, 1–6.
[24] Hansen, J.E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J.O., Hansen, J.-E.S. and Brunak, S. (1995) Biochem. J. 308, 801–813.
[25] Ricard-Blum, S., Dublet, B. and van der Rest, M. (2000) Unconventional Collagens: Types VI, VII, VIII, IX, X, XII, XIV, XVI and XIX, 1st edn., Oxford University Press, Oxford.
[26] Shaw, L.M. and Olsen, B.R. (1991) Trends Biochem. Sci. 16, 191–194.
[27] Moradi-Ameli, M., Deleage, G., Geourjon, C. and van der Rest, M. (1994) Matrix Biol. 14, 233–239.
[28] Shaw, L.M. and Olsen, B.R. (1991) Trends Biochem. Sci. 16, 191–194.
[29] Imamura, Y., Scott, I.C. and Greenspan, D.S. (2000) J. Biol. Chem. 275, 8749–8759.
[30] Schuppan, D., Cantaluppi, M.C., Becker, J., Veit, A., Bunte, T., Troyer, D., Schuppan, F., Schmid, M., Ackermann, R. and Hahn, E.G. (1990) J. Biol. Chem. 265, 8823–8832.
[31] Keene, D.R., Lunstrum, G.P., Morris, N.P., Stoddard, D.W. and Burgeson, R.E. (1991) J. Cell Biol. 113, 971–978.
[32] Brown, J.C., Mann, K., Wiedemann, H. and Timpl, R. (1993) J. Cell Biol. 120, 557–567.
[33] Muragaki, Y., Mariman, E.C., van Beersum, S.E., Perala, M., van Mourik, J.B., Warman, M.L., Olsen, B.R. and Hamel, B.C. (1996) Nat. Genet. 12, 103–105.
[34] Paassilta, P., Lohiniva, J., Annunen, S., Bonaventure, J., Le Merrer, M., Pai, L. and Ala-Kokko, L. (1999) Am. J. Hum. Genet. 64, 1036–1044.
[35] Bork, P. and Bairoch, A. (1995) Trends Biochem. Sci. 20, poster C02.
[36] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.