Hypothesis

Is there an evolutionary relationship between WARP (von Willebrand factor A-domain-related protein) and the FACIT and FACIT-like collagens?

Jamie Fitzgerald*, John F. Bateman

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

Received 10 June 2003; revised 22 August 2003; accepted 25 August 2003

First published online 8 September 2003

Edited by Takashi Gojobori

Abstract We suggest that there is an evolutionary relationship between von Willebrand factor A-domain-related protein (WARP), and the fibril-associated collagen with interrupted triple helix (FACIT) and FACIT-like subfamilies of collagens. Data from a comparison of amino acid sequences, domain organisation and chromosomal location are consistent with the hypothesis that WARP and these collagens share a common collagen ancestor. In support of this is the observation that the WARP 3′ coding region is GC-rich suggesting that this may represent the remnant of a triple helix protein domain which WARP has 'lost' during evolution.

Key words: Von Willebrand factor A-domain; Fibril-associated collagen with interrupted triple helix; Collagen evolution; Genome duplication; Extracellular matrix

1. Introduction

The collagen family of proteins are important structural and regulatory components of the extracellular matrix (ECM) and are present in all multicellular eukaryotes of the animal kingdom [1–3]. The completion of the first draft of the human genome project has led to an explosion in the number of new collagen genes reported with 27 distinct collagen chains described, many of which can be further classified into distinct subfamilies based on size, domain organisation and function. The group with the largest number of members are the fibril-associated collagen with interrupted triple helix (FACIT) and FACIT-like collagen subfamilies with nine members although it should be noted that these collagens are classified according to domain organisation and there are little data to indicate whether these collagens are functionally conserved. The FACIT collagens include the three collagen IX chain genes (COL9A1, COL9A2, COL9A3), collagen XII (COL12A1), collagen XIV (COL14A1), and the recently described collagen XX (COL20A1) and collagen XXI (COL21A1) genes. The FACIT-like collagens, which share protein domain similarities with the FACIT collagens, are the collagen XVI (COL16A1) and collagen XIX (COL19A1) genes. All the members of these FACIT and FACIT-like collagen subfamilies contain an interrupted triple helix with a characteristic pair of conserved Cys residues at the C-terminal end of the interrupted triple helix (Fig. 1). In addition to these domains, collagens XII, XIV, XX, and XXI contain one or more copies of the von Willebrand factor A-domain (VWFA-domain). Collagens XII, XIV and XX also contain at least six copies of the fibronectin type III (FN3) repeat module and all collagens except the genes for COL9A2 and COL9A3 contain a thrombospondin domain. For more detailed information about the FACIT and FACIT-like collagens the reader is directed to a review [4] and original papers for the recently reported collagens XX [5] and XXI [6–8]. We recently identified WARP (for von Willebrand factor A-domain-related protein), a new member of the VWFA-domain-containing superfamily of ECM proteins, located on human chromosome 1p36 [9]. The WARP domain structure comprises an amino-terminal VWFA-domain followed by two FN3 domains and a short 21 amino acid proline/arginine-rich sequence [9].

2. Results and discussion

The first indication that WARP and the FACIT and FACIT-like collagens could be related is the observation that even though VWFA- and FN3 domains are widespread ECM proteins, they are only found adjacent to each other in the FACIT collagens XII, XIV, XX and XXI (see Fig. 1), collagen VII and in WARP. A direct comparison of VWFA-domain amino acid sequences from a range of ECM molecules reveals that WARP has highest amino acid similarity to the four VWFA-domains from collagen XII (58–60%) and the two collagen XIV VWFA-domains (60 and 61%) with less similarity to domains in collagens XX (57%) and XXI (53%) (not shown). The FN3 repeat motif is most similar to that of FN3 repeats in collagens XII and XIV [9]. However, since the possibility exists that the protein sequences are functionally constrained, we sought alternative lines of evidence that do not rely on direct amino acid comparisons.

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

Abbreviations: WARP, von Willebrand factor A-domain-related protein; FACIT, fibril-associated collagen with interrupted triple helix; VWFA, von Willebrand factor A; FN3, fibronectin type III; ECM, extracellular matrix

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00968-2

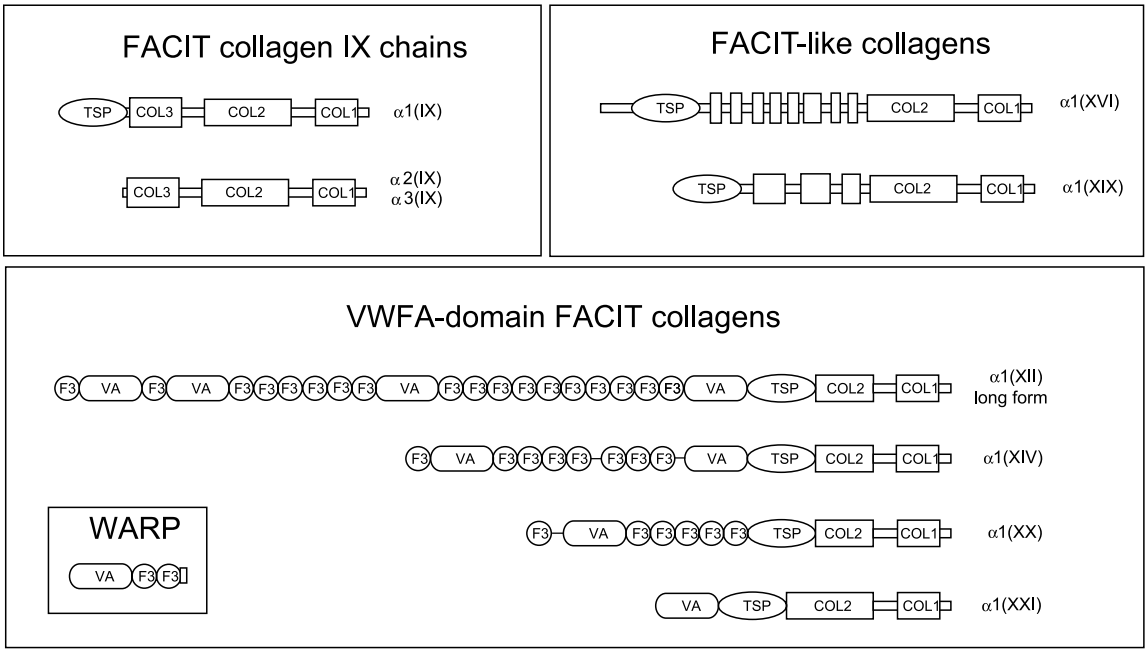

Fig. 1. Domain organisation of the FACIT, FACIT-like collagens and WARP. The broad family of FACIT collagens can be further divided based on domain organisation which places together the FACIT collagen IX genes (COL9A1, COL9A2 and COL9A3), and the FACIT-like collagen XVI (COL16A1) and collagen XIX (COL19A1) genes, and a group containing the VWFA-domain FACIT collagens (COL12A1, COL14A1, COL20A1, COL21A1) and WARP. The von Willebrand factor A-domain (VA), fibronectin type III (F3), thrombospondin (TSP) and interrupted triple helix domain modules (COL) are drawn using standard symbols [13].

It is clear that genomes evolve through a combination of single nucleotide mutation, segmental duplication, and chromosome rearrangement. In addition, one early observation was that vertebrate, but not invertebrate, genomes show extensive intra-genomic synteny [10] leading to the conclusion that two rounds of whole genome duplication (octoploidy) occurred early in vertebrate evolution (for reviews see [10,11]). In humans, the FACIT and FACIT-like collagen genes cluster on four chromosomal blocks on chromosomes 1, 6, 8 and 20, that show evidence of intra-genomic synteny [12] (Fig. 2). Intra-genomic synteny between these four chromosome blocks is supported by the finding that additional genes have three or four paralogues that reside on these blocks as well. All four matrilin (MATN1 to MATN4), syndecan (SDC1 to SDC4), eyes absent (EYA1 to EYA4) genes and three MYC family members (MYC, MYCL1, MYCN), have paralogues on 1p, 20q, 8q, and 6q/2p [12], near the FACIT collagen genes. Chromosome 6q13 contains the COL12A1, COL9A1 and COL19A1 genes, 8q23 the COL14A1 gene, 20q13 the COL20A1 and COL9A3 genes, and 1p36 the COL9A2 and COL16A1 genes. Intriguingly, the gene for WARP is located on one of these chromosomal segments, 1p36, in close proximity to the COL9A2 and COL16A1 genes. Furthermore, both 6q13 and 1p36 loci contain a collagen IX gene (COL9A1 and COL9A2) and a closely related FACIT-like collagen XVI/XIX gene (COL19A1 and COL16A1). 6q13 contains a VWFA-FACIT collagen gene (COL12A1) but since 1p36 lacks an obvious VWFA-FACIT collagen gene candidate, we propose that WARP is the VWFA-FACIT 'collagen' gene representative on 1p36.

An unpublished collagen gene, designated COL22A1 (GenBank accession number AF406780), is located in close proximity to the COL14A1 gene on human chromosome 8 (mouse chromosome 15). The predicted protein contains a VWFA-domain followed by a thrombospondin domain and a C-terminal FACIT collagen-like interrupted triple helix, suggesting that this molecule is a new VWFA-domain FACIT collagen. The location of this gene is consistent with our observations

Table 1
The FACIT and FACIT-like collagen gene clusters are syntenic in the human and mouse genomes

Human gene	FACIT collagen subfamily	human chromosome	Mouse gene	mouse chromosome
WARP	VWFA-FACIT	1 (0.9 M)	Warp	4 (151.2 M)
COL16A1	FACIT-like	1 (31.9 M)	Col16a1	4 (127.6 M)
COL9A2	collagen IX FACIT	1 (40.8 M)	Col9a2	4 (118.9 M)
COL12A1	VWFA-FACIT	6 (75.7 M)	Col12a1	9 (80.4 M)
COL19A1	FACIT-like	6 (70.5 M)	Col19a1	1 (24.7 M)
COL9A1	collagen IX FACIT	6 (70.9 M)	Col9a1	1 (24.6 M)
COL14A1	VWFA-FACIT	8 (120.2 M)	Col14a1	15 (72.4 M)
COL20A1	VWFA-FACIT	20 (61.8 M)	Col20a1	2 (179.3 M)
COL9A3	collagen IX FACIT	20 (61.3 M)	Col9a3	2 (178.9 M)

The mouse homologues are located on chromosomal regions that are syntenic with the human genes. Numbers in parentheses refer to approximate location of gene from the top (p arm) of each chromosome. The chromosomal locations of the human (June 2002 release) and mouse (February 2002 release) genes for WARP, collagens XII, XIV, XVI, XIX, and XX, and the three collagen IX chains were obtained from the UCSC Genome Bioinformatics (http://genome.ucsc.edu/) [14–16].

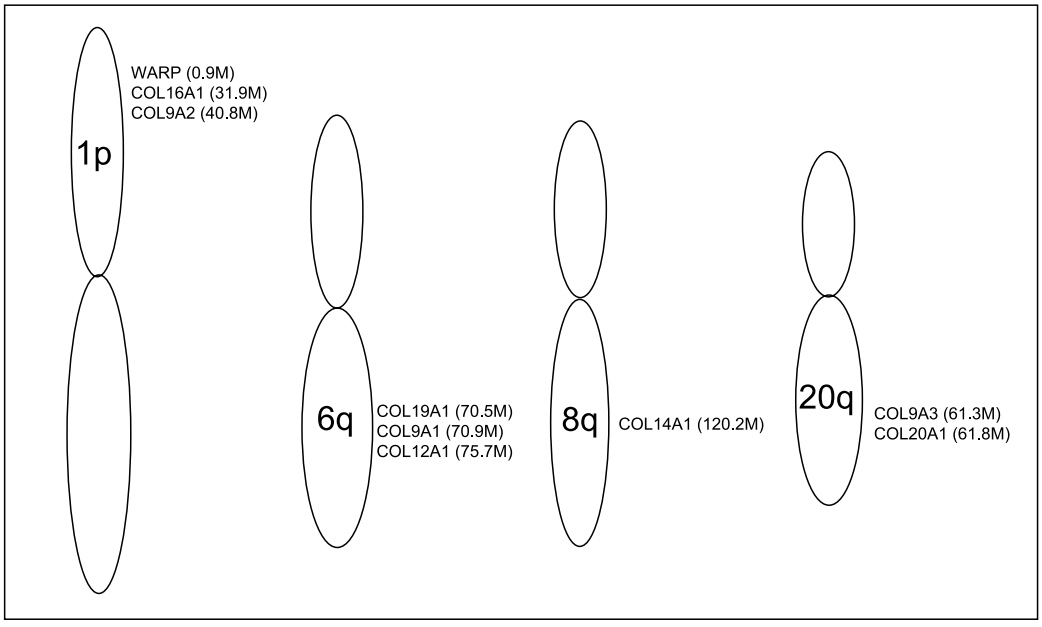

Fig. 2. Chromosomal locations of human WARP, and FACIT and FACIT-like collagen genes.

that the FACIT and FACIT-like collagen genes are clustered on four chromosomal blocks.

In the mouse genome, the WARP and FACIT and FACIT-like collagen gene homologues are clustered on discrete chromosomal blocks that show synteny with the corresponding locations in the human genome (Table 1). For the cluster that includes the WARP gene (Chr. 1 in humans and chr. 4 in mice) the gene order is also conserved between the two species with the COL16A1/Col16a1 genes located between the WARP/warp and COL9A2/col9a2 genes. This pattern of human/mouse synteny for the broad FACIT collagen family indicates that the genome dispersal of these collagens occurred prior to the mammalian radiation and that the intra-genomic synteny is probably conserved in all mammals.

Although the VWFA- and FN3 domains are clearly recognisable as conserved modules, the C-terminal domain of WARP does not resemble any other domain as determined by extensive BLAST P searching. In human WARP, this region is 24 amino acids in size (21 amino acids in mouse) and strikingly, 14 of these amino acids are either proline or arginine residues coded for by CCX and GCX (or AGpu) codons respectively which make the encoding nucleotide sequence 88% GC nucleotide-rich. Collagen triple helix domains are composed of repeating glycine-X-Y triplets where X and Y are often proline residues and since glycine is coded for by GGX codons the FACIT collagen triple helix is also GC-rich (average of 61%). The skewed GC:AT ratio raises the possibility that the WARP C-terminal region may be the remnant of a collagen triple helix. In support of this is the observation that although the amino acid sequence of the VWFA-domain and FN3 repeats have 91% similarity between the human and mouse, the C-terminal domains have only 52% similarity and differ in size by three amino acids suggesting that because this region cannot function as a triple helix it has experienced a loss of selection pressure and accumulated nucleotide changes.

In summary, we propose that WARP and the FACIT and FACIT-like collagens share a common ancestor and that at one time WARP may have contained a C-terminal triple helical domain. Our observations which are consistent with this are: (1) WARP contains a VWFA-domain and an FN3 repeat which are protein modules only found in tandem in the VWFA-domain FACIT collagens; (2) the amino acid sequences of the WARP VWFA and FN3 modules have highest identity with VWFA and FN3 modules in the FACIT collagens rather than other ECM proteins; (3) the human WARP gene is located on 1p36 close to the COL9A2 and COL16A1 genes, on one of four chromosomal regions known to form an intra-genomic syntenic block which have a common evolutionary history; (4) the C-terminal domain of WARP is GC-rich suggesting that at one time it may have encoded a triple helix domain.

A greater understanding of the evolution of the FACIT and FACIT-like collagen genes will be provided in time by the sequence analysis of additional vertebrate genomes that separated from the mammalian lineage prior to the rodent–human split. Since it is postulated that large scale genome duplication occurred early in vertebrate evolution we would predict that the invertebrate genomes such as *Drosophila melanogaster* have few or no FACIT-like collagen genes and probably do not contain a WARP homologue. Homology searching the available invertebrate genome data fails to reveal a WARP-like molecule. In vertebrates, analysis is complicated by the realisation that additional rounds of large genome DNA duplication have occurred in many fish and amphibian species, although one genome that may illuminate FACIT collagen evolution is that of *Xenopus tropicalis*. The *X. tropicalis* genome is genetically diploid and is currently the subject of a large scale sequencing effort due to be completed in 2005. It will be interesting to ask whether the FACIT and FACIT-like collagen sequences are clustered on discrete chromosomal segments in this species and whether a WARP homologue exists and if so, whether it contains a collagenous domain.

Acknowledgements: This work was supported by grants from the National Health and Medical Research Council of Australia and the Murdoch Childrens Research Institute.

References

[1] Bateman, J.F., Lamande, S.R. and Ramshaw, J.A.M. (1996) in: Extracellular Matrix (Comper, W.D., Ed.), pp. 22–67, Harwood Academic Publishers, Amsterdam.
[2] Kielty, C.M., Hopkinson, I. and Grant, M.E. (1993) in: Connective Tissue and its Heritable Disorders. Molecular, Genetic, and Medical Aspects (Royce, P.M. and Steinmann, B., Eds.), pp. 103–147, Wiley-Liss, New York.
[3] van der Rest, M. and Garrone, R. (1991) FASEB J. 5, 2814–2823.
[4] Ricard-Blum, S., Dublet, B. and van der Rest, M. (2000) in: Protein Profile, Oxford University Press, Oxford.
[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] Fitzgerald, J. and Bateman, J.F. (2001) FEBS Lett. 505, 275–280.
[7] Tuckwell, D. (2002) Matrix Biol. 21, 63–66.
[8] Chou, M.Y. and Li, H.C. (2002) Genomics 79, 395–401.
[9] Fitzgerald, J., Ting, S.T. and Bateman, J.F. (2002) FEBS Lett. 517, 61–66.
[10] Ohno, S. (1970) Evolution by gene duplication, Springer, New York.
[11] Furlong, R.F. and Holland, P.W. (2002) Philos. Trans. R. Soc. Lond. B. Biol. Sci. 357, 531–544.
[12] Gibson, T.J. and Spring, J. (2000) Biochem. Soc. Trans. 28, 259–264.
[13] Bork, P. and Bairoch, A. (1995) Trends Biochem. Sci. 20, poster C02.
[14] Karolchik, D. et al. (2003) Nucleic Acids Res. 31, 51–54.
[15] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) Genome Res. 12, 996–1006.
[16] Kent, W.J. (2002) Genome Res. 12, 656–664.