Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest

Raymond P. Boot-Handford* and Danny S. Tuckwell

Summary
Fibril-forming (fibrillar) collagens are extracellular matrix proteins conserved in all multicellular animals. Vertebrate members of the fibrillar collagen family are essential for the formation of bone and teeth, tissues that characterise vertebrates. The potential role played by fibrillar collagens in vertebrate evolution has not been considered previously largely because the family has been around since the sponge and it was unclear precisely how and when those particular members now found in vertebrates first arose. We present evidence that the classical vertebrate fibrillar collagens share a single common ancestor that arose at the very dawn of the vertebrate world and prior to the associated genome duplication events. Furthermore, we present a model, 'molecular incest', that not only accounts for the characteristics of the modern day vertebrate fibrillar collagen family but demonstrates the specific effects genome or gene duplications may have on the evolution of multimeric proteins in general. *BioEssays* 25:142–151, 2003. © 2003 Wiley Periodicals, Inc.

What is a collagen?

Collagens are major components of the extracellular matrices of all metazoan life and play crucial roles in developmental processes and tissue homeostasis (e.g. Refs. 1,2). The conservation of basement membrane and fibril-forming collagens between the simplest (e.g. sponges, *Hydra*) and most-complex animals (vertebrates), illustrates the importance of these proteins and the extracellular matrices that they form. Inhibition of the synthesis of either of these classes of collagen in *Hydra*, a simple diploblastic organism, results in a failure of tissue regeneration after injury.(3,4) Mutations in similar genes in man are either incompatible with life or result in a number of severe diseases including osteogenesis imperfecta, aortic aneurysms, Alports syndrome and even some forms of cancer.(2,5) The 21 different types of collagen described in mammals to date form a variety of structures from networks to microfilaments to fibrils.(1,2,5) In each case, the collagen molecules provide strong and supportive extracellular scaffolds or assemblies as well as contributing other important properties such as defined interaction sites for cells and macromolecules.

The Wellcome Trust Centre for Cell-Matrix Research, School of Biological Sciences, University of Manchester.
Funding agencies: RBH is the recipient of a Wellcome Trust Research Leave Award. DST is a BBSRC Advanced Research Fellow.
*Correspondence to: Ray Boot-Handford, Wellcome Trust Centre for Cell-Matrix Research, 2,205 Stopford, School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK.
E-mail: Ray.Boot-Handford@man.ac.uk
DOI 10.1002/bies.10230
Published online in Wiley InterScience (www.interscience.wiley.com).

Abbreviations: ER, endoplasmic reticulum; Gly, glycine; vWF A (C), von Willebrand factor A (C) domain; WAP, partial whey acidic protein cysteine repeat; SURF, Sea URchin Fibrillar domain. FACIT, Fibril Associated Collagen with an Interrupted Triple helix.

Collagens are composed of three polypeptide chains (α chains) that fold together to form the characteristic triple helical collagenous domain (Fig. 1). Different types of collagen contain either three identical α chains (homotrimers) or a mixture of two or three genetically distinct chains (heterotrimers). The sequences required to form a collagenous domain are Gly-X-Y repeats in which the X and Y positions are frequently proline and hydroxyproline. Glycine is required every third residue as it is the only amino acid small enough to pack into the central core of the triple helix. Amino acids with larger side groups cannot fit into the available space in the core of the helix and hence disrupt or prevent the folding of the helix during trimer assembly in the ER (see Refs. 2,5). The hydroxyproline residues stabilise the triple helical domain by hydrogen bonding along the helix. Collagenous domains vary in length from as short as 30 to in excess of 1500 amino acid residues and are either uninterrupted or contain imperfections in the Gly-X-Y repeat that result in flexible regions within the relatively stiff, rod-like triple helix.(1)

How are collagens synthesised?

Following transcription, collagen synthesis (Fig. 1) starts with the translation of the α chain mRNAs on the rough ER. As the growing α chains are translocated into the lumen of the ER, the collagenous domains are post-translationally modified by hydroxylation of peptidyl prolyl and lysyl residues to form hydroxyproline and hydroxylysine respectively. In addition, some hydroxylysine residues are glycosylated with galactose and glucose residues. Once the full-length α chain has been

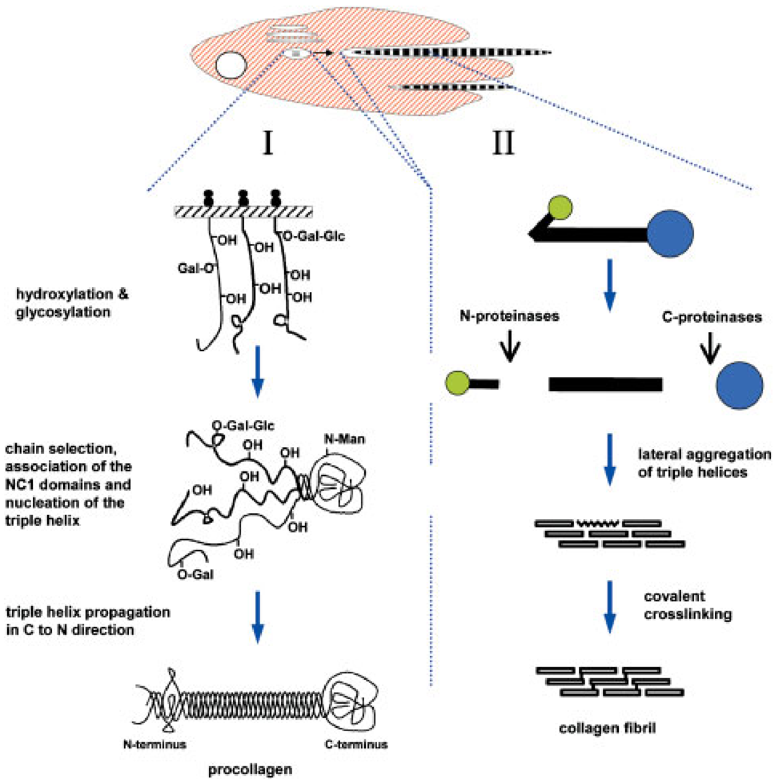


Figure 1. The biosynthesis of collagens. **I:** mRNA encoding the proα chains is translated by ribosomes on the ER and the nascent chains translocated into the lumen of the ER. The nascent polypeptide chains are post-translationally modified by hydroxylation and glycosylation reactions. Full-length proα chains select appropriate partners, trimerise at their C-terminal NC1 domains and fold the triple helical collagenous domain in the C- to N-terminal direction. **II:** Upon secretion, the carboxyl- and amino-terminal propeptides of fibrillar collagens are enzymatically removed by C- and N-proteinases, respectively. The collagenous domains laterally aggregate forming fibrils that are stabilised by covalent crosslinks.

translocated into the ER lumen, the carboxyl terminal non-collagenous (NC1) domain folds. Correctly folded NC1 domains select appropriate partners and trimerise bringing the three α chain collagenous domains together in register. The triple helix nucleates at its carboxyl terminal end allowing the complete collagenous triple helical domain to subsequently fold or 'zip up' in the C-to-N direction (Fig. 1). Upon secretion, collagen molecules assemble to form macromolecular aggregates. For fibril-forming or 'fibrillar' collagens, which are the focus of this article, the mechanisms controlling macromolecular assembly are relatively well understood (in vertebrate systems—Fig. 1). A full-length fibrillar collagen molecule, or 'procollagen', consists of the C-terminal NC1 domain (C-propeptide), a collagenous domain with approximately 337–340 Gly-X-Y repeats (major helix), and an N-propeptide consisting of a short triple helical domain (minor helix) and a non-collagenous NC2 domain. The N- and C-propeptides are highly soluble and keep the insoluble collagenous domain in solution. As the fibrillar procollagen is secreted from the cell, specific enzymes (C- and N-proteinases) cleave the propeptides from the collagenous domains. The dramatic drop in solubility causes the collagenous domains to laterally aggregate with a $\frac{1}{4}/\frac{1}{3}$ stagger to the overlap regions. Lysyl oxidase catalyses covalent cross linking between adjacent collagen molecules stabilising the fibril (Fig. 1). More detailed accounts of collagen synthesis and assembly are given in other recent reviews.[1,2,5]

Vertebrate fibrillar collagens

The first collagens to be purified and characterised biochemically were vertebrate and fibrillar (see Ref. 1), since these proteins are the most abundant in connective tissues such as skin, bone, tendon, ligament, blood vessel walls and cornea. Much of the initial interest in the biochemistry of these tissues arose due to their importance in the gelatin and leather industries. It is therefore not surprising that the specific characteristics and properties of the much-studied and -utilised vertebrate forms have dominated our perception of fibrillar collagens. The naming of vertebrate collagens (types I, II, III etc.) reflects the order in which they were discovered. The fibrillar forms comprise collagen types I, II, III, V and XI (Table 1). Most members of this family of collagens are crucial for skeletogenesis. Type I collagen is quantitatively the most abundant fibrillar collagen in the body and is the major structural protein of bone and teeth as well as many other tissues. Type II collagen is the major structural protein of cartilage. Types V and XI collagens, although quantitatively less abundant, are found in association with types I and II collagen in bone and cartilage respectively as well as in other tissues (Table 1). Specific mutations in type I collagen cause

Table 1. Fibrillar collagens in the animal kingdom

Organism	Name	Accession	Chain composition(s)	Characteristics*	Tissue distribution
Sponge	EMF1	P18856 CAA49472	Unknown	Contains major and minor helix; single interruption in major helix (one aa insertion) 244 residues from NC1; NC2 incomplete	Unknown
Hydra (Cnidaria)	Hcol 1	AF525468	Homotrimer	Contains major and minor helix; uninterrupted major helix	Central region of mesoglea
Hydra	Hcol 2	Unpublished	Unknown	Contains major and minor helix; uninterrupted major helix; has partial WAP in NC2 domain	Unknown
Hydra	Hcol 3	Unpublished	Unknown	Contains major and minor helix; uninterrupted but short (969 residues) major helix; NC2 incomplete but contains 2 partial WAP and 2vWF A domains	Unknown
Riftia (Annelid)	Fibrillar collagen	AAB24972 AAF80453	Unknown	Partial sequences, contains major helix of 1011 residues plus NC1 domain; interruption of helix 414 residues (Gly to Ala substitution) from NC1	Body wall
Arenicola (Annelid)	Fam 1 alpha	AAC47545	Unknown	Partial sequence, NC1 plus 415 residues of helical domain; interruption at 414 residues from NC1 domain	Body wall
Alvinella (Annelid)	FAp 1 alpha	AAC35289	Unknown	Partial sequence, NC1 plus 621 residues of helical domain; interruptions at 550 residues (one aa deletion) and 384 residues (Gly to Ala substitution) from NC1	Unknown
Abalone (Mollusc)	Hdcol 1 alpha	AB017600	Unknown	Contains only major helix; single interruption in major helical domain (Gly to Ser substitution) 417 residues from NC1 domain; NC2 contains vWF C domain	Foot and adductor muscles
Abalone	Hdcol 2 alpha	AB017601	Unknown	Contains major and minor helices; single interruption in major helical domain (Gly to Ala substitution) 417 residues from NC1 domain; NC2 contains partially conserved vWF C domain	Foot and adductor muscles
Sea urchin (Echinoderm)	alpha 1 alpha 2	M92040 M92041	$\alpha 1_2\alpha 2$	Both chains contains major and minor helices; uninterrupted major helices; alpha 2 chain NC2 contains vWF C and 12 SURF domains	Unknown
Homo sapien (Vertebrate)	Type I	α1: P02452 α2: P08123	$\alpha 1(\mathrm{I})_2\alpha 2(\mathrm{I})$	Both chains contain major and minor helix; uninterrupted major helices; α1(I) has vWF C in NC2 domain	Bone, skin, tendon, cornea
Homo sapien	Type II	P02458	$\alpha 1(\mathrm{II})_3$	Contains major and minor helix; uninterrupted major helix; has vWF C in NC2 domain	Cartilage, vitreous
Homo sapien	Type III	P02461	$\alpha 1(\mathrm{III})_3$	Contains major and minor helix; uninterrupted major helix of 1029 residues; has vWF C in NC2 domain	Skin, aorta, uterus, intestine
Homo sapien	Type V	α1: P20908 α2: P05997 α3: NP_056534	α1(V) α2(V) α3(V), $\alpha 1(\mathrm{V})\ {}_2\alpha 2(\mathrm{V})$, $\alpha 1(\mathrm{V})_3$	All chains contain major and minor helix; uninterrupted major helices; α1 & α3 (V) have PARP in NC2 domain; α2 (V) has vWF C in NC2 domain	Bone, skin, tendon, cornea, placenta
Homo sapien	Type XI	α1: P12107 α2: P13942	α1(XI) α2(XI) α3(XI)	All chains contain major and minor helix; uninterrupted major helices; α1 & α2 (XI) have PARP in NC2 domain. α3 (XI) is derived from the type II collagen gene	Cartilage, intervertebral disc

*Major helix 1011–1020 amino acid residues in length unless otherwise stated.

osteogenesis imperfecta (brittle bone disease) whereas mutations in types II and XI collagen can result in a variety of skeletal dysplasias and osteoarthrosis (see Refs. 2,5 for recent reviews). Since these collagens play specific and essential roles in the formation and growth of the cartilage and/or bone-based skeleton, as well as other complex tissues that characterise vertebrates, it is of interest to examine how and when in evolution these collagen types evolved.

Fibrillar collagens in invertebrates

The presence of fibril-forming collagens in the invertebrates has been known for many years based on ultrastructural and biochemical experiments (e.g. Refs. 6–8). With the advent of molecular cloning and sequencing technologies, it became apparent that fibrillar collagens belong to a family of genes conserved at the sequence level throughout the animal kingdom from sponges to man (Table 1). The only parts of the multicellular animal kingdom that have lost fibrillar collagens during their evolution (based on genome sequencing) are the arthropods and nematodes. It is unfortunate that these two genetically most tractable invertebrate systems lack this class of protein which plays such important roles in the rest of the animal kingdom and, in particular, in vertebrates. Fibrillar collagen sequences from sea urchin[9,10] and sponge[11] have been known for around a decade and the recent additions of full-length sequences from abalone[12] and *Hydra*[3] have significantly increased our ‘invertebrate’ database (Table 1). The delineation of multiple sequences from several invertebrate phyla has enabled a more detailed comparison of fibrillar collagens across species than was previously possible based on the limited number and sometimes incomplete sequences available.

The structures of fibrillar collagens in the invertebrates seem somewhat less rigidly conserved than those in vertebrates. The fixed length of the major triple helical domain in the known vertebrate fibrillar collagens is clearly inherited from the earliest evolved invertebrates since this length of domain is present in sponge (339 Gly-X-Y repeats). *Hydra* also has α chains with similar sized collagenous domains (340 triplet repeats for both Hcol 1 & 2) but in addition has co-evolved a fibrillar collagen chain with a shorter major helix (Hcol 3—323 Gly-X-Y repeats). Several lineages including sponge, annelids and molluscs have evolved fibrillar collagens with interruptions in the helical domain (Table 1), a feature not exhibited by any of the classical vertebrate genes. The N-propeptides of fibrillar collagens represent the least conserved region of the molecule with several species exhibiting unique characteristics (e.g. vWF A and WAP domains in *Hydra* (Hcol 2 & 3), multiple SURF domains in sea urchin (α2), deletion of the minor helix in abalone (α1), Table 1). In contrast, some of the invertebrate proα chains have a vWF C domain located in their N-propeptide (abalone α1 and α2; sea urchin α2) in common with several of the vertebrate proα chains [α1(I), α1(II), α1(III), α2(V)].

Current models of fibrillar collagen evolution

Comparison of the invertebrate and vertebrate chain characteristics (Table 1) leads one to the inevitable conclusion that the basic template for those fibrillar collagens found in the vertebrates was laid down extremely early in metazoan evolution. There have been several attempts to describe how the family of genes evolved and the likely structure of the ancestral gene.[13–17] The discovery that the collagenous domains of vertebrate fibrillar collagen genes are encoded by exons with, or derived from, a 54 bp repeat (e.g. 108, 99, 54 or 45 bp coding for 12, 11, 6 or 5 Gly-X-Y amino acid repeats, respectively) was the first indication that not only amino acid sequence but also gene structure may be conserved.[13,18] Indeed, the subsequent demonstrations that sea urchin[9,10] and sponge[11,14] fibrillar collagen genes also have very similar exon organisations to those found in the vertebrate genes further emphasised the high level of conservation in this family of genes. Several attempts to define the mechanisms by which the family have evolved have been based on predicting how subtle variations in the exact pattern of exons between different phyla may have arisen.[14,15] Similarities in sequence and domain structure has led to the suggestion that the sea urchin α1 chain may be the invertebrate equivalent of the α2(I) vertebrate gene[19] and that *Hydra* Hcol 1 is most closely related to the type I/II vertebrate collagens.[3] Structural and biochemical studies on the collagens forming thin fibrils in primitive invertebrates such as sponge and jelly fish have revealed similarities with the vertebrate type V and XI collagens suggesting a close evolutionary relationship.[16] The natural tendency in these studies is to compare invertebrate fibrillar collagens with their vertebrate cousins on the assumption that there is a direct relationship between individual chains.

Phylogenetic analyses

Establishing the direct evolutionary relationships between invertebrate and vertebrate α chains is hampered by the fact that phylogenetic analyses have not provided clear-cut insights (e.g. see Fig. 2A and Ref. 17). The problem arises when analysing modern-day invertebrate sequences because the species in which these fibrillar collagen chains occur diverged from each other such a long time ago. The sequences of modern-day invertebrate α chains have had time to diverge sufficiently from their closest relatives that the sequence sites that can be mutated are saturated, and all the chains now share approximately the same level of sequence identity. For instance, two chains in *Hydra* (Hcol 2 and Hcol 3) contain WAP domains in their N-propeptides, presumably because they were derived by duplication from a common *Hydra*-specific ancestor. However, despite sharing this unique feature in the

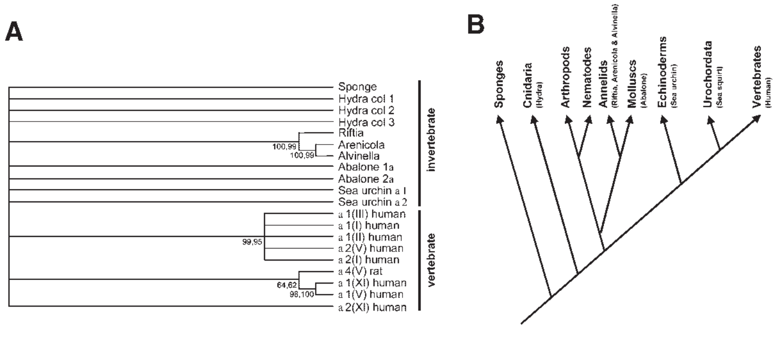


Figure 2. A: Phylogenetic analysis of invertebrate and vertebrate fibrillar collagens and **B:** the evolutionary relationships of the species from which they are derived. Distance matrix and parsimony analyses were carried out on the NC1 domains (C-propeptides) of fibrillar collagens and the consensus tree from the two methods is shown. Bootstrapping values are for consensus results where these are above 50%. In each case, the first value is from distance analysis, the second from parsimony. It should be noted that in the distance analysis, the α2(XI) clusters with α1(V), α1(XI) and α4(V) with 100% bootstrap value.

fibrillar collagen family, today the two *Hydra* chains are, upon phylogentic analysis, no more similar to each other than to any other chain in the fibrillar family, be it invertebrate or vertebrate in origin (Fig. 2A). Similarly, molluscs are the closest evolutionary relatives of annelids (Fig. 2B) and the fibrillar collagen chains in both phyla share a conserved interruption in the collagenous domain. However, despite this evidence suggesting these two phyla have derived their fibrillar collagen chains from a common ancestor specific to their two lineages, the mollusc and annelid chains have been evolving separately for too long a period for phylogenetic analysis to detect a closer relationship than that shared by all family members. This lack of clustering of sequences in the invertebrates holds true for all but the three sequences from annelids (*Riftia, Arenicola* and *Alvinella*) that form a distinct group or clade (Fig. 2A). Almost certainly, this is because the three sequences that cluster represent the same collagen gene in three closely related species that have evolved relatively recently from a common ancestor.

In contrast to the invertebrates, fibrillar collagen sequences from vertebrates cluster into two clades[17,18] (Fig. 2) indicating that chains in a clade are more similar to each other than to chains in either the other vertebrate clade or invertebrates. The type A clade comprises α1(I), α2(I), α1(II), α1(III) and α2(V) whereas the type B clade contains α1(V), α3(V), α1(XI) and α2(XI).[17] The α3(XI) is derived from the gene for α1(II) and so belongs to the type A clade. The clustering of these two sets of sequences immediately suggests that the multiple members of each clade have arisen much latter in evolution than their invertebrate cousins. However, the timing of the evolution of the founder genes of the two vertebrate clades is not clear from phylogenetic analyses, although it has been proposed to have occurred at the same time or after the duplication event giving rise to the two sea urchin α chains.[17]

Genome duplication and vertebrate evolution

Ohno[21] was the first to propose that major transitions such as the evolution of vertebrates may have required genome duplication events. Ohno's hypothesis, based on comparative data on genome size and chromosome number, was that genome duplication events provide large pools of 'redundant' genes that can rapidly mutate and develop the novel functions required to drive the evolutionary transition from, for instance, invertebrate to vertebrate. Studies on the evolution of the developmentally important Hox gene cluster that patterns the anteroposterior axis of most animal embryos has been particularly informative in this regard.[22] Amphioxus, an invertebrate chordate and closest living relative of the vertebrates, has a single Hox gene cluster[23] whereas mammals have four. The amplification of this developmentally crucial gene cluster and subsequent gain-of-function changes in the expression patterns of the amplified clusters is proposed to have facilitated the evolution of vertebrate-specific features such as control of neural crest cell fate, hindbrain differentiation and otic morphogenesis.[24] The exact mechanism by which the genome was expanded remains controversial,[25] but it seems logical that the resulting increased numbers of genes was a significant factor enabling vertebrate evolution.

Genome duplications and vertebrate fibrillar collagen evolution

The studies describing the amplification of the Hox gene cluster by way of partial or complete genome duplications early during vertebrate evolution to produce the four clusters seen in modern day mammals is of particular relevance to the vertebrate fibrillar collagen family for the following reason. The genes encoding each of the vertebrate fibrillar collagen type A clade members are physically linked to the Hox gene clusters in mammals (*COL1A2* to HOXA; *COL1A1* to HOXB; *COL2A1* to HOXC; *COL5A2* & *COL3A1* to HOXD) and are believed to have arisen through the same duplication mechanism.[26] Indeed, one of the most detailed phylogenetic analyses on vertebrate fibrillar collagens was in fact conducted to establish the evolutionary relationships between the linked mammalian Hox gene clusters which are not themselves particularly useful for such analyses because of the limited lengths of conserved sequences (183 bp for each Hox box).[20] From these studies, it is clear that the radiation of the type A and B clades occurred early during vertebrate evolution as a result of genome, or at least, large-scale duplication events. Therefore, since the multiple members of the two clades of vertebrate fibrillar collagens arose early in vertebrate evolution due to duplication events, the major question remaining to be resolved is, 'precisely when did the founder genes for the type A and B clades first evolve'?

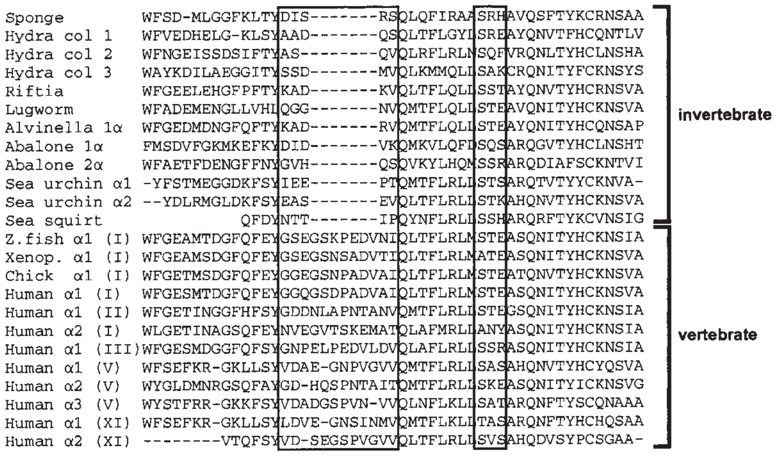
A rare genomic event characterises all vertebrate fibrillar collagens

As described above, phylogenetic analyses based on sequence comparisons have not given any precise indication as to when the type A and B clade founder genes first evolved. However, direct sequence comparisons are not the only way of determining molecular phylogeny. Rare genomic events, mutational changes that have occurred in and distinguish the genomes of particular clades, can also give clear insights into phylogeny.[27,28] Indeed, it is just such a rare genomic event described below, apparent in all the characterised vertebrate fibrillar collagens but none of the invertebrate chains, that provides the strongest indication to date of precisely when and how the two vertebrate clades evolved.

During collagen synthesis, three nascent α chains must first trimerise at their NC1 domains prior to nucleating and folding the triple helix (Fig. 1). In order to trimerise, the chains must be capable of recognising their correct partners. This is not a trivial exercise since in some cells many different α chains are being synthesised at the same time [e.g. α1(I), α2(I), α1(III), α1(V), α2(V) and α3(V)] and yet these chains always associate to form the correct trimers for type I, III and V collagen as shown in Table 1. Lees et al.[29] reasoned that the regions of the NC1 domain with little or no sequence conservation probably account for the 'collagen-type'-specific trimerisation of NC1 domains. Accordingly, sequence swapping was used to identify the discontinuous 15 amino acid residue sequence required for chain selectivity during fibrillar collagen trimer assembly in vertebrates.[29] During detailed analyses of novel *Hydra* fibrillar collagen NC1 domain sequences (RBH, unpublished), it became apparent that each was missing 7 or 8 amino acid residues from the core of their chain selectivity sequences (Fig. 3). Indeed, further comparisons revealed that all known invertebrate NC1 domains, including one from an invertebrate chordate (the closest relatives of the vertebrates) are also missing 7 amino acid residues from the core of the chain selection sequence (Fig. 3). In contrast, all the characterised vertebrate fibrillar NC1 domains, irrespective of collagen type or species, contain the full-length sequence originally identified in the human. By far the most likely way in which this rare genomic change distinguishing invertebrate and vertebrate fibrillar collagens could have arisen is by a single founder gene for the vertebrate lineage acquiring the sequence as an insertion, followed by multiple rounds of gene and/or genome duplication coupled with numerous mutation events, to produce firstly the founders for the type A and B clades and, subsequently, their multiple members. One inevitable outcome of this assertion is that the founder vertebrate collagen gene formed a homotrimer. Since the insertional mutation is not found in invertebrates yet occurred prior to the early vertebrate genome duplications and the radiation of the vertebrate collagen A and B clades, it is evidently one of the most ancient, or evolutionarily earliest, detectable genetic events specific to the vertebrate lineage.

Why should an elongated form of the chain selection sequence be conserved?

A discontinuous selection sequence chain of 8 amino acids appears completely adequate for promoting collagen proα chain homo- or hetero-trimerisation in all invertebrates (Fig. 3; Table 1). Why, therefore, should a selection sequence of



```
Sponge           WFSD-MLGGFKLTYDIS-------RSQLQFIRAASRHAVQSFTYKCRNSAA
Hydra col 1      WFVEDHELG-KLSYAAD-------QSQLTFLGYLSREAYQNVTFHCQNTLV
Hydra col 2      WFNGEISSDSIFTYAS--------QVQLRFLRLNSQFVRQNLTYHCLNSHA
Hydra col 3      WAYKDILAEGGITYSSD-------MVQLKMMQLLSAKCRQNITYFCKNSYS
Riftia           WFGEELEHGFPFTYKAD-------KVQLTFLQLLSSTAYQNVTYHCRNSVA
Lugworm          WFADEMENGLLVHLQGG-------NVQMTFLQLLSTEAVQNITYHCKNSVA
Alvinella 1α     WFGEDMDNGFQFTYKAD-------RVQMTFLQLLSTEAYQNITYHCQNSAP
Abalone 1α        FMSDVFGKMKEFKYDID-------VKQMKVLQFDSQSARQGVTYHCLNSHT
Abalone 2α       WFAETFDENGFFNYGVH-------QSQVKYLHQMSSRARQDIAFSCKNTVI
Sea urchin α1   -YFSTMEGGDKFSYIEE-------PTQMTFLRLLSTSARQTVTYYCKNVA-
Sea urchin α2   --YDLRMGLDKFSYEAS-------EVQLTFLRLLSTKAHQNVTYHCKNSVA
Sea squirt                 QFDYNTT-------IPQYNFLRLLSSHARQRFTYKCVNSIG
Z.fish α1 (I)   WFGEAMTDGFQFEYGSEGSKPEDVNIQLTFLRLMSTEASQNITYHCKNSIA
Xenop. α1 (I)   WFGEAMSDGFQFEYGSEGSNSADVTIQLTFLRLMATEASQNITYHCKNSVA
Chick  α1 (I)   WFGETMSDGFQFEYGGEGSNPADVAIQLTFLRLMSTEATQNVTYHCKNSVA
Human α1 (I)    WFGESMTDGFQFEYGGQGSDPADVAIQLTFLRLMSTEASQNITYHCKNSVA
Human α1 (II)   WFGETINGGFHFSYGDDNLAPNTANVQMTFLRLLSTEGSQNITYHCKNSIA
Human α2 (I)    WLGETINAGSQFEYNVEGVTSKEMATQLAFMRLLANYASQNITYHCKNSIA
Human α1 (III)WFGESMDGGFQFSYGNPELPEDVLDVQLAFLRLLSSRASQNITYHCKNSIA
Human α1 (V)    WFSEFKR-GKLLSYVDAE-GNPVGVVQMTFLRLLSASAHQNVTYHCYQSVA
Human α2 (V)    WYGLDMNRGSQFAYGD-HQSPNTAITQMTFLRLLSKEASQNITYICKNSVG
Human α3 (V)    WYSTFRR-GKKFSYVDADGSPVN-VVQLNFLKLLSATARQNFTYSCQNAAA
Human α1 (XI)   WFSEFKR-GKLLSYLDVE-GNSINMVQMTFLKLLTASARQNFTYHCHQSAA
Human α2 (XI)   ---------VTQFSYVD-SEGSPVGVVQLTFLRLLSVSAHQDVSYPCSGAA-
```

Figure 3. Invertebrate NC1 domains lack amino acid residues at the core of the vertebrate fibrillar collagen chain selection sequence. Alignment of the segment of representative NC1 domains containing the discontinuous 15 residue chain selection sequence[29] (boxed). Homologous regions from Sponge, *Hydra* (Cnidaria), *Alvinella* (Annelid), Abalone (Mollusc), Sea urchin (Echinoderm) Sea squirt (Urochordata: *Ciona intestinalis* genome project, http://www.jgi.doe.gov/programs/ciona.htm, LQW104235)], Zebra fish (CAC51021), *Xenopus laevis* (BAA94972), chicken (P02457), and human NC1 domains are aligned. Unless otherwise stated, accession numbers are presented in Table 1.

almost twice the length be so conserved in vertebrate fibrillar collagens? Firstly, this conservation indicates functional importance—if the acquired extra residues were non-essential, one would expect most if not all the chains to have mutated back to the presumably minimal sequence length required for chain selection as seen in the invertebrate genes. Secondly, we wish to advance the hypothesis that the elongated chain selection sequence of the founder vertebrate collagen gene acquired functional significance in the context of gene and genome duplications. For multimeric proteins such as collagen, gene duplications offer the possibility of rapidly evolving not just one novel protein but several by what we term a mechanism of 'molecular incest'. The vertebrate-specific elongated chain selection sequence may have been essential to allow the rapid development of the large number of 'incestuous' interactions required to promote the evolution of the complex family of vertebrate fibrillar collagens needed to form tissues that characterise vertebrates, such as bones and teeth.

The vertebrate fibrillar collagen family evolved by 'molecular incest' resulting from gene duplications

Let us consider the situation of the founder vertebrate collagen gene immediately following the first gene (or genome) duplication event (Fig. 4 panel B). There are now two identical copies of the gene which, when expressed and translated, produce identical α chains that trimerise to form the same homotrimer produced by the parental gene. However, the sequences of the two identical genes inevitably start to diverge. Bear in mind that evolution is selecting for α chains that maintain the ability to trimerise and subsequently polymerise to form fibrils. There are two possible outcomes. Firstly, one of the duplicated genes becomes inactivated and deleted from the genome. Secondly, as depicted in Fig. 4 panel C, combinations of the evolving α chains derived from the now non-identical but instead 'sibling' genes continue to trimerise in what is now an incestuous molecular interaction between sibling α chains. Should these new forms of collagen trimers be

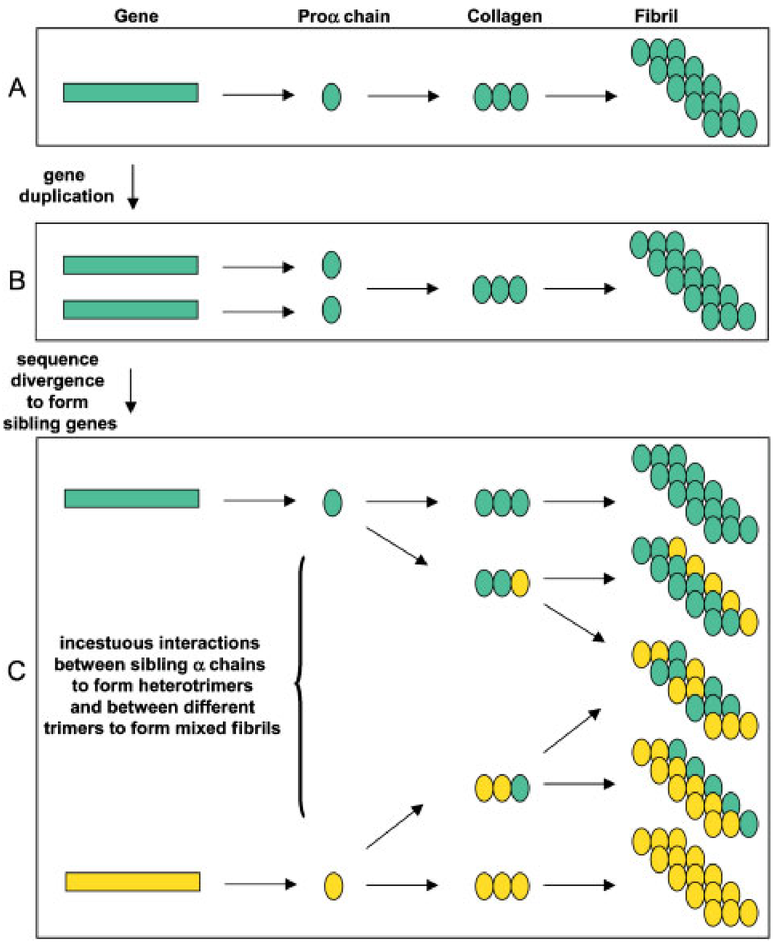

Gene
Proα chain
Collagen
Fibril
A
gene duplication
B
sequence divergence to form sibling genes
C
incestuous interactions between sibling α chains to form heterotrimers and between different trimers to form mixed fibrils

Figure 4. Vertebrate collagens evolved by a mechanism of gene duplication and molecular incest. **A:** The parental collagen gene encodes a proα chain that forms a homotrimeric collagen molecule that assembles to form a fibril. **B:** Immediately following gene (or genome) duplication, the two identical duplicate genes make an unchanged collagen that assembles to form an identical fibril to that formed before the duplication event. **C:** With time, the sequences of the duplicated genes drift so that they are no longer identical but instead become 'sibling' genes (yellow and green genes). The genes no longer encode identical proteins but the sibling proα chains may retain their ability to trimerise together in what is now an incestuous relationship. Novel, heterotrimers can be formed which may then assemble to form a number of new fibrils with differing compositions.

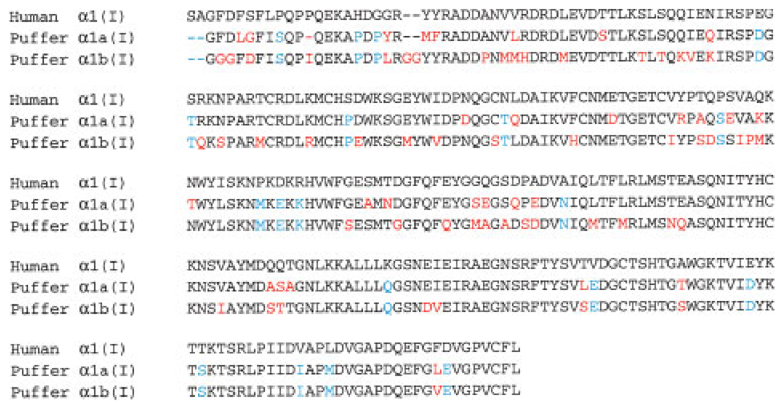
capable of polymerising to form fibrils with improved or advantageous properties, the duplicated but diverging genes would become conserved in the genome. As illustrated in Fig. 4, one round of gene duplication for a homotrimeric collagen produces three possible new trimer combinations and even more potential combinations in terms of polymerisation into fibrils. Clearly, we only see the successful and useful products of this selective process since unsuccessful combinations are not conserved. This mechanism of multimeric protein evolution, driven by gene duplication and subsequent molecular incest of the resulting sibling proteins, can readily account for the structure, complex set of α chain trimerization patterns, and patterns of co-polymerisation, seen in the vertebrate fibrillar collagen family. Indeed, when viewed in this light, the vertebrate family of fibrillar collagens is highly incestuous with all the classical fibrillar collagen genes of both clades being very close relatives.

The incestuous nature of interactions extends throughout and across both clades of the vertebrate family. For instance, types V and XI collagen molecules are heterotrimers composed of α chains from both clades (Table 1). These chains must not only be able to select appropriate partners for trimerisation through their chain selectivity sequences, but also maintain similar lengths of collagenous domain in order to correctly fold their triple helices. Furthermore, the co-polymerisation of type I collagen with types III or V collagen and type II with XI collagen generates further extensive and common evolutionary constraints on the collagenous domains of genes in both clades. Accordingly, the vertebrate fibrillar family of collagens represents a complex set of closely related and interdependent α chains. Consequently, a mutation in one α chain can potentially have profound effects throughout the fibrillar collagen system and the chains are therefore evolutionarily tied together in an interdependent manner. This is a major reason why the classical vertebrate collagen chains appear so similar (e.g. uninterrupted helix of same length and highly conserved NC1 domain with same length of chain selection sequence).

Support for the 'molecular incest' model of fibrillar collagen evolution is derived from modern day fish (teleosts). There is considerable evidence accumulating that fish have undergone a further round of genome duplication after they diverged, approx. 450 million years ago, from the lineage leading to amphibians, reptiles, birds and mammals. Indeed, some species of fish are thought to have up to seven Hox gene clusters (see Ref. 30 and references therein). So what about the fibrillar collagen genes? The puffer fish genome is currently being sequenced and the first draft was published in the autumn of 2001.[31] A initial simple search reveals that at least four of the vertebrate fibrillar collagen genes found in man are duplicated in the puffer fish [α1(I), α1(II), α2(V) and α1(XI)]. These duplicated genes are 'siblings' in that their sequences are not identical and appear to be slowly diverging both from each other and from the mammalian equivalent (Fig. 5). In puffer fish, the α1b(I) chain, as compared to α1a(I) chain, has accumulated almost twice as many amino acid changes. This difference is intriguing because it suggests different levels of evolutionary constraint on the two chains.

Does gene duplication and molecule incest give genes encoding multimeric proteins an evolutionary advantage?

The mechanisms of gene duplication and molecular incest illustrated in Fig. 4 are not only of relevance for collagen molecules but hold for all multimeric proteins, be they homomeric or heteromeric, including other matrix families such as laminins, thrombospondins and integrins. The opportunities for developing potentially useful 'novel' properties following a genome duplication are far greater for multimeric than monomeric proteins. In essence, following a duplication, monomeric proteins drift by acquiring new mutations and eventually one of the duplicates achieves a new



```
Human  α1(I)     SAGFDFSFLPQPPQEKAHDGGR--YYRADDANVVRDRDLEVDTTLKSLSQQIENIRSPEG
Puffer α1a(I)    --GFDLGFISQP-QEKAPDPYR--MFRADDANVLRDRDLEVDSTLKSLSQQIEQIRSPDG
Puffer α1b(I)    --GGGFDFISQPIQEKAPDPLRGGYYRADDPNMMHDRDMEVDTTLKTLTQKVEKIRSPDG

Human  α1(I)     SRKNPARTCRDLKMCHSDWKSGEYWIDPNQGCNLDAIKVFCNMETGETCVYPTQPSVAQK
Puffer α1a(I)    TRKNPARTCRDLKMCHPDWKSGEYWIDPDQGCTQDAIKVFCNMDTGETCVRPAQSEVAKK
Puffer α1b(I)    TQKSPARMCRDLRMCHPEWKSGMYWVDPNQGSTLDAIKVHCNMETGETCIYPSDSSIPMK

Human  α1(I)     NWYISKNPKDKRHVWFGESMTDGFQFEYGGQGSDPADVAIQLTFLRLMSTEASQNITYHC
Puffer α1a(I)    TWYLSKNMKEKKHVWFGEAMNDGFQFEYGSEGSQPEDVNIQLTFLRLMSTEASQNITYHC
Puffer α1b(I)    NWYLSKNMKEKKHVWFSESMTGGFQFQYGMAGADSDDVNIQMTFMRLMSNQASQNITYHC

Human  α1(I)     KNSVAYMDQQTGNLKKALLLKGSNEIEIRAEGNSRFTYSVTVDGCTSHTGAWGKTVIEYK
Puffer α1a(I)    KNSVAYMDASAGNLKKALLLQGSNEIEIRAEGNSRFTYSVLEDGCTSHTGTWGKTVIDYK
Puffer α1b(I)    KNSIAYMDSTTGNLKKALLLQGSNDVEIRAEGNSRFTYSVSEDGCTSHTGSWGKTVIDYK

Human  α1(I)     TTKTSRLPIIDVAPLDVGAPDQEFGFDVGPVCFL
Puffer α1a(I)    TSKTSRLPIIDIAPMDVGAPDQEFGLEVGPVCFL
Puffer α1b(I)    TSKTSRLPIIDIAPMDVGAPDQEFGVEVGPVCFL
```

Figure 5. Alignment of the NC1 domain of human α1(I) collagen with the duplicated α1(I) sequences found in Puffer fish. Amino acid changes most likely occurring after the fish-specific genome duplication event and therefore unique to either of the puffer fish sequences are highlighted in red. Sequence changes common to both puffer fish sequences and therefore probably occurring after the divergence of the lineages leading to fish and man, but prior to the fish genome duplication event are shown in blue.

function or becomes functionally inactivated and deleted from the genome. After the duplication of genes forming a multimeric protein, the same accumulation of mutations occurs but, what may be a neutral event in the context of, for instance, a homotrimer, may in fact be advantageous within the context of a heterotrimer. In this manner, multimeric proteins have a larger 'context' within which to test the effects of a specific mutation and would therefore seem more likely to achieve a novel function and their genes become conserved.

Concluding comments and questions for the future

In this review, we have taken a number of relatively recent findings and some completely new data to develop a novel and radical view of how the fibrillar collagen family and, in particular, the vertebrate members, evolved. We have presented evidence of a rare genomic event affecting the chain selectivity sequence of fibrillar collagens that strongly suggests that the classical fibrillar collagens share a single common ancestor that arose at the very dawn of the vertebrate world and prior to the genome duplication events. Furthermore, we have presented a model based on gene duplication and subsequent molecular incest that accounts for all aspects of the modern day vertebrate fibrillar collagen family.

It is clear that the emergence of vertebrates was the consequence of numerous factors acting in concert and Ohno's hypothesis that genome duplications played a part in driving these major transitions in evolution is an extremely plausible explanation. The fibrillar collagens have not previously been considered or suggested to play any particular role in vertebrate evolution since this class of collagens has been around since the sponge first evolved. The 'tongue-in-cheek' title to this review suggesting that fibrillar collagens may have played the key role in vertebrate evolution was created to draw attention to the fact that we now believe this family of vertebrate-specific genes played an active role in this process, albeit alongside numerous other gene families. Included amongst these are 15 other types of collagen (VI–X, XII–XXI) that have an uncertain evolutionary history in that, to date, they have only been described in vertebrates.[1] Most of these collagens, which presumably first evolved in early vertebrates and radiated alongside their more ancient fibrillar and basement membrane collagen relatives, also play important roles in the highly specialised and complex tissues that characterise vertebrates.[1,2,5]

Many fascinating questions remain to be answered including: are the duplicated genes encoding fibrillar collagens in puffer fish still expressed in the same cells permitting the development of new collagens by way of 'molecular incest' or are the sibling genes now expressed in different cells or tissues? What is the structure of the fibrillar collagen genes of hagfish and lampreys, living descendants of the earliest vertebrates that are thought to have undergone only one round of genome duplication? Did the FACIT (Fibril Associated Collagen with an Interrupted Triple helix) collagen family, which encodes collagens found in intimate association with fibrils in vertebrates, evolve in the invertebrates or vertebrates? Do multimeric proteins have a demonstrable evolutionary advantage over monomer proteins? Genome database searches reveal the presence of some novel collagens that have yet to be biologically characterised—how will these new genes influence our perception of collagen evolution?

It is an intriguing possibility that the elongated chain selection sequence acquired by the founder vertebrate fibrillar collagen gene produced a more sophisticated chain selection site that was better suited and perhaps even essential to support the complexity of incestuous interactions evident in the vertebrate family of fibrillar collagens. Indeed, it is conceivable that without the elongated chain selection sequence, the fibrillar family may have developed with fewer members or types causing the emergence of vertebrates from the invertebrate world to take a different evolutionary course, perhaps not involving the development of bone as we know it. Although fibrillar collagens are clearly not the key factor in vertebrate evolution, they play a key role in putting the vertebrae in vertebrates.

Acknowledgments

David Garrod is thanked for promoting the writing of this article and Checco Ramirez for critical reading and helpful comments. Karl Kadler and Gillian Wallis are thanked for many valuable discussions.

References

1. Kielty CM, Grant ME. The collagen family: structure, assembly and organisation in the extracellular matrix. In: Royce PM, Steinmann B, editors. Connective Tissue and its Heritable Disorders. Wiley-liss, Inc. 2002. p 159–221.
2. Prockop DJ, Kivirikko KI. Collagens: molecular biology, diseases and potentials for therapy. Ann Rev Biochem 1995;64:403–434.
3. Deutzmann R, Fowler S, Zhang X, Boone K, Dexter S, Boot-Handford RP, Rachel R, Sarras MP Jr. Molecular, biochemical, and functional analysis of a novel and developmentally important fibrillar collagen (Hcol-I) in *Hydra*. Development 2000;127:4669–4680.
4. Fowler SJ, Jose S, Zhang X, Deutzmann R, Sarras MP Jr, Boot-Handford RP. Characterisation of *Hydra* type IV collagen: Type IV collagen is essential for head regeneration and its expression is up-regulated upon exposure to glucose. J Biol Chem 2000;275:39589–39599.
5. Myllyharju J, Kivirikko KI. Collagens and collagen-related diseases. Ann Med 2001;33:7–21.
6. Adams E. Invertebrate collagens. Marked differences from vertebrate collagens appear in only a few invertebrate groups. Science 1978;202: 591–598.
7. Lenhoff HM, Mascatine L, Davis LV. Coelenterate biology: experimental research. Science 1968;160:1141–1146.
8. Baccetti B. Collagen and animal phylogeny. In: Bairati A, Garrone R, editors. Biology of invertebrate and lower vertebrate collagens. New York & London: Plenum Press. 1985. p 29–47.
9. Saitta B, Buttice G, Gambino R. Isolation of a putative collagen-like gene from the sea urchin *Paracentrotus lividus*. Biochem Biophys Res Commun 1989;158:633–669.

10. D'Alessio M, Ramirez F, Suzuki HR, Solursh M, Gambino R. Structure and developmental expression of a sea urchin fibrillar collagen gene. Proc Natl Acad Sci USA 1989;86:9303–9307.
11. Exposito JY, Garrone R. Characterization of a fibrillar collagen gene in sponges reveals the early evolutionary appearance of two collagen gene families. Proc Natl Acad Sci USA 1990;87:6669–6673.
12. Yoneda C, Hirayama Y, Nakaya M, Matsubara Y, Irie S, Hatae K, Watabe S. The occurrence of two types of collagen proα-chain in the abalone *Haliotis* muscle. Eur J Biochem 1999;261:714–721.
13. Yamada Y, Avvedimento VE, Mudryi M, Ohkubo H, Vogeli G, Irani M, Pastan I, de Crombrugge B. The collagen gene: Evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. Cell 1980;22:887–892.
14. Exposito J-Y, van der Rest M, Garrone R. The complete intron/exon structure of *Ephydatia mulleri* fibrillar collagen gene suggests a mechanism for the evolution of an ancestral gene module. J Mol Evol 1993;37:254–259.
15. Exposito J-Y, Cluzel C, Lethias C, Garrone R. Tracing the evolution of vertebrate fibrillar collagens from an ancestral α chain. Matrix Biology 2000;19:275–279.
16. Tillet E, Franc JM, Franc S, Garrone R. The evolution of fibrillar collagens: A sea-pen collagen shares common features with vertebrate type V collagen. Comp Biochem Physiol 1996;133B:239–246.
17. Sicot F-X, Exposito J-Y, Masselot M, Garrone R, Deutsch J, Gaill F. Cloning of an annelid fibrillar-collagen gene and phylogenetic analysis of vertebrate and invertebrate collagens. Eur J Biochem 1997;246: 50–58.
18. Chu M-L, Prockop DJ. Collagen: Gene structure. In: Royce PM, Steinmann B, editors. Connective Tissue and its Heritable Disorders. Wiley-liss, Inc. 2002. p 223–248.
19. Exposito JY, D'Alessio M, Solursh M, Ramirez F. Sea urchin collagen is evolutionarily homologous to vertebrate pro-alpha 2(I) collagen. J Biol Chem 1992;267:15559–15562.
20. Bailey WJ, Kim J, Wagner GP, Ruddle FH. Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol Biol Evol 1997;14:843–853.
21. Ohno S. Evolution by gene duplication. Springer-Verlag, Berlin. 1970.
22. Holland PWH. Gene duplication: Past, present and future. Seminars Cell Devel Biol 1999;10:541–547.
23. Garcia-Fernandez J, Holland PWH. Archetypal organisation of the amphioxus Hox gene cluster. Nature 1994;370:563–566.
24. Holland PWH. Homeobox genes in vertebrate evolution. Bioessays 1992; 14:267–273.
25. Smith NGC, Knight R, Hurst LD. Vertebrate genome evolution: a slow shuffle or a big bang? Bioessays 1999;21:697–703.
26. Ruddle FH, Bentley KL, Murtha MT, Risch N. Gene loss and gain in the evolution of vertebrates. Dev 1994;Suppl:155–161.
27. Rokas A, Holland PWH. Rare genomic changes as a tool for phylogenetics. Trend Ecol Evol 2000;15:454–459.
28. Venkatesh B, Erdmann MV, Brenner S. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proc Natl Acad Sci USA 2001;98:11382–11387.
29. Lees JF, Tasab M, Bulleid NJ. Identification of the molecular recognition sequence which determines the type-specific assembly of procollagen. EMBO J 1997;16:908–916.
30. Taylor JS, de Peer YV, Braasch I, Meyer A. Comparative genomics provides evidence for an ancient genome duplication event in fish. Phil Trans R Soc Lond B 2001;356:1661–1679.
31. see http://Fugu.jgi-psf.org, http://Fugu.jgi-pst.org, or http://fugu.hgmp.mrc.ac.uk.