Collagens, modifying enzymes and their mutations in humans, flies and worms

Johanna Myllyharju and Kari I. Kivirikko

Collagen Research Unit, Biocenter Oulu and Department of Medical Biochemistry and Molecular Biology, University of Oulu, FIN-90014 Oulu, Finland

Collagens and proteins with collagen-like domains form large superfamilies in various species, and the numbers of known family members are increasing constantly. Vertebrates have at least 27 collagen types with 42 distinct polypeptide chains, >20 additional proteins with collagen-like domains and ~20 isoenzymes of various collagen-modifying enzymes. Caenorhabditis elegans has ~175 cuticle collagen polypeptides and two basement membrane collagens. Drosophila melanogaster has far fewer collagens than many other species but has ~20 polypeptides similar to the catalytic subunits of prolyl 4-hydroxylase, the key enzyme of collagen synthesis. More than 1300 mutations have so far been characterized in 23 of the 42 human collagen genes in various diseases, and many mouse models and C. elegans mutants are also available to analyse the collagen gene family and their modifying enzymes.

The collagens are a family of extracellular matrix proteins that play a dominant role in maintaining the structure of various tissues and also have many other important functions. For example, collagens are involved in cell adhesion, chemotaxis and migration, and the dynamic interplay between cells and collagen regulates tissue remodeling during growth, differentiation, morphogenesis and wound healing, and in many pathologic states.

All collagen molecules consist of three polypeptide chains, called α chains (Box 1), and contain at least one domain composed of repeating Gly-X-Y sequences in each of the constituent chains. In some collagens all three α chains are identical, whereas in others the molecules contain two or even three different α chains. The three α chains are each coiled into a left-handed helix and are then wound around a common axis to form a triple helix with a shallow right-handed superhelical pitch, so that the final structure is a rope-like rod. The presence of glycine, the smallest amino acid, in every third position is essential for the packing of this coiled-coil structure. The X and Y positions can have any amino acid other than glycine, but proline is often found in the X position and 4-hydroxyproline in the Y position. The 4-hydroxyprolines play a particularly important role because these residues are essential for the stability of the triple helix.

Collagens are the most abundant proteins in the human body, constituting ~30% of its protein mass. The important roles of these proteins have been clearly demonstrated by the wide spectrum of diseases caused by a large number of mutations found in collagen genes. This article will review the collagen superfamilies and their mutations in vertebrates, Drosophila melanogaster and the nematode Caenorhabditis elegans. The genomes of these two model invertebrate species have been fully sequenced, and it is therefore possible to identify all of the collagen genes present in these species. Because of the extensive literature in these fields, this review will focus primarily on recent advances. More detailed accounts and more complete references can be found in previous reviews, for example Refs [1–6].

The collagen superfamily in vertebrates

Types of collagen

Vertebrates have at least 27 collagen types with 42 distinct α chains in total, and >20 additional proteins have collagen-like domains. All collagens also possess non-collagenous domains in addition to the actual collagen domains. Most collagens form supramolecular assemblies, such as fibrils and networks, and the superfamily can be

---

**Box 1. Human collagen nomenclature**

- **Collagen types I-XXVII.** Collagens are numbered with roman numerals in the order of their discovery.
- **Collagen polypeptide chains.** These are called α chains, each collagen molecule consisting of three of them. Depending on the collagen type, the three α chains can be either identical or the molecule can contain two or even three different α chains. The α chains of a specific collagen type are numbered with arabic numerals and the collagen type is given in parentheses. For example α1(I) and α2(I) are the α1 and α2 chains of type I collagen, and α1(III) is the α1 chain of type III collagen.
- **Procollagen.** The fibril-forming collagens (Figure 1a) are synthesized as procollagen molecules, which have propeptides at the N and C-terminal ends of their polypeptide chains, called proα chains.
- **Genes encoding the collagen chains.** These are named by the prefix COL followed by an arabic number for the collagen type, the letter ‘A’ stands for α chain and an arabic number for the chain. For example, the genes COL1A1 and COL1A2 encode the α1 and α2 chains of type I collagen, respectively, and the gene COL2A1 encodes the α1 chain of type II collagen.
divided into several subfamilies on the basis of these assemblies or other features (Figure 1). Collagen types I–XIX have been discussed in many previous reviews [1–4], whereas collagen types XX–XXVII (Table 1) have been reported only during the past three years [7–14]. Some collagens have a restricted tissue distribution: for example, types II, IX and XI, which are found almost exclusively in cartilage; type X, found only in hypertrophic cartilage; the family of type IV collagens in basement membranes; type VII in the anchoring fibrils for basement membranes; and type XVII in skin hemidesmosomes. By contrast some collagen types are found in most extracellular matrices. The highly heterogeneous group of proteins that contain collagen domains but have not been defined as collagens (Figure 1) includes: the subcomponent C1q of complement, a C1q-like factor, adiponectin, at least eight collectins and three ficolins (humoral lectins of the innate immune defence system), the tail structure of acetylcholinesterase, three macrophage receptors, ectodysplasin, two EMILINS (elastic fibre-associated glycoproteins) and a src-homologous-and-collagen protein [3,4,6,15].

Collagen fibrils often consist of more than one collagen type. For example, the type I collagen fibrils often contain small amounts of types III, V and XII, whereas the type II collagen fibrils of cartilage also contain types IX and XI. Collagen types V and XI can also form hybrid molecules (e.g. having an α1(XI) and an α2(V) chain in the same molecule). The six α chains of type IV form at least three types of molecule [α1(IV)α2(IV), α3(IV)α4(IV)α5(IV) and α6(IV)α7(IV)] [16]. Further heterogeneity within the superfamily is caused by alternative splicing of the transcripts of many of the genes and the use of alternative promoters in some genes. The large number of structures present in members of the superfamily implies that they are involved in numerous different biological functions [1–4].

The non-collagenous domains of many collagens also have important functions. Major interest has been focused on endostatin, a proteolytically derived 20 kDa C-terminal fragment of collagen XVIII, and restin, a corresponding fragment of collagen XV, which inhibit endothelial cell migration and angiogenesis and reduce tumour growth in animal models [3,4,17]. The C-terminal non-collagenous domain of collagen IV also inhibits angiogenesis and tumour growth [3,17], whereas other functions have been described for the non-collagenous domains of certain other collagens [3].

### Table 1. Recently identified collagen types XX–XXVII

<table>
<thead>
<tr>
<th>Type</th>
<th>Chain</th>
<th>Residues</th>
<th>Location</th>
<th>Group</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX</td>
<td>α1(XXI)</td>
<td>1473</td>
<td>Corneal epithelium, skin, cartilage and tendon</td>
<td>b</td>
<td>[7]</td>
</tr>
<tr>
<td>XXI</td>
<td>α1(XXII)</td>
<td>957</td>
<td>Many tissues</td>
<td>b</td>
<td>[8]</td>
</tr>
<tr>
<td>XXII^a</td>
<td>α1(XXII)</td>
<td>1616</td>
<td>Tissue junctions^a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>XXIII</td>
<td>α1(XXIII)</td>
<td>540</td>
<td>Metastatic tumour cells</td>
<td>g</td>
<td>[9]</td>
</tr>
<tr>
<td>XXIV</td>
<td>α1(XXIV)</td>
<td>1714</td>
<td>Developing bone and cornea</td>
<td>a</td>
<td>[10]</td>
</tr>
<tr>
<td>XXV</td>
<td>α1(XXV)</td>
<td>666</td>
<td>Neurons</td>
<td>g</td>
<td>[11]</td>
</tr>
<tr>
<td>XXVI</td>
<td>α1(XXVI)</td>
<td>438</td>
<td>Testis, ovary</td>
<td>b</td>
<td>[12]</td>
</tr>
<tr>
<td>XXVII</td>
<td>α1(XXVII)</td>
<td>1860</td>
<td>Cartilage, eye, ear and lung</td>
<td>a</td>
<td>[13,14]</td>
</tr>
</tbody>
</table>

^aGroup in Figure 1.
^bGenBank accession number AF406780.
^cM. Koch and L. Bruckner-Tuderman, unpublished.

---

**Biosynthesis and modifying enzymes**

Collagen synthesis involves many post-translational modifications that require three collagen hydroxylases [18,19], two collagen glycosyltransferases [1,3,4], two specific proteinases [20] to cleave the N and C propeptides from the procollagen molecules (family a in Figure 1) and one specific oxidase [21] to initiate crosslink formation (Figure 2). Other enzymes include peptidyl proline cis-trans isomerase and protein disulfide isomerase (PDI), which has at least three functions: (i) to catalyze the formation of intrachain and interchain disulfide bonds; (ii) to serve as the β subunit in collagen prolyl 4-hydroxylases; and (iii) to act as a chaperone that binds nascent collagen chains and prevents their aggregation [3,18,19]. Collagen synthesis also involves a specific chaperone, Hsp47. Homozygous knockout of this gene in mice is lethal at the embryonic stage, indicating that this protein is essential for normal development [22,23].

Collagen prolyl 4-hydroxylase, an α2β2 tetramer located within the lumen of the endoplasmic reticulum, plays a central role in collagen synthesis because 4-hydroxyproline residues are essential for the formation of triple-helical molecules in vivo. Collagen prolyl 4-hydroxylase has at least three isoenzymes in humans, with distinct α subunits but all have PDI as their β subunit [18,19,24,25]. A novel family of three cytoplasmic prolyl 4-hydroxylases has recently been shown to play a key role in the regulation of the hypoxia-inducible transcription factor HIF [19,26–28]. These enzymes have no PDI subunit, have different requirements with respect to the sequence flanking the prolines that are hydroxylated and have markedly higher \(K_m\) values for oxygen than the collagen prolyl 4-hydroxylases [26–29]. Lysyl hydroxylase also has at least three isoenzymes [3,4,30,31], whereas lysyl oxidase has at least five [3,21,32,33]. The two proteinases that cleave the N and C propeptides from procollagen molecules in large transport vesicles close to the plasma membrane [34] and in the extracellular matrix each have at least three isoenzymes [3,4,20,35]. The C proteinases also process various other precursor proteins of the extracellular matrix and belong to the tolloid family. The main C proteinase isozyme is identical to a protein previously called bone morphogenic protein-1 (BMP-1) [3,4,20]. The two specific collagen glycosyltransferases have not been cloned but one of the lysyl hydroxylase isozymes has also been shown to possess small amounts of these enzyme activities [36,37]. Work is
Members of the collagen superfamily and their known supramolecular assemblies. The collagen superfamily can be divided into nine families based on the supramolecular assemblies and other features of its members:

(a) fibril-forming collagens; 
(b) fibril-associated collagens with interrupted triple helices (FACITs) located on the surface of fibrils, and structurally related collagens; 
(c) collagens forming hexagonal networks; 
(d) the family of type IV collagens located in basement membranes; 
(e) type VI collagen, which forms beaded filaments; 
(f) type VII collagen, which forms anchoring fibrils for basement membranes; 
(g) collagens with transmembrane domains; and 
(h) the family of type XV and XVIII collagens. The supramolecular assemblies of families (g) and (h) are unknown and are therefore not shown in this figure.

The polypeptide chains found in the 27 collagen types are coded by 42 genes in total (shown in blue), each molecule consisting of three polypeptide chains that can be either identical or different. An additional highly heterogeneous group (i) within the superfamily comprises proteins that possess collagenous domains but have not been defined as collagens. Some of the group (i) proteins could also be defined as collagens, although some of the collagens might also belong to this group because there are no distinct criteria for distinguishing between a collagen and a protein containing a collagen domain(s) [1]. The collagen domains are shown in purple, the N and C-terminal non-collagenous domains are in dark pink, and the non-collagenous domains interrupting the triple helix in light blue, short interruptions of a few amino acids are not shown. For acetylcholinesterase, the catalytic domain (shown in green) and the tail structure are products of separate genes. Modified and updated from Refs [1,3].

Abbreviation: PM, plasma membrane.
in progress to elucidate differences in the expression patterns and functions of the various isoenzymes of the collagen modifying enzymes.

**Mutations in human collagens and their modifying enzymes and corresponding mouse models**

About 1100 mutations have been reported in just six of the 42 collagen genes, COL1A1, COL1A2, COL2A1, COL3A1, COL4A5 and COL7A1, which encode the two kinds of polypeptides of type I collagen, the polypeptides of types II, III and VII and the α5 chain of type IV [1,3,38–44]. The number of identified mutations will probably increase because many of those found recently might not have been regarded as worth publishing separately. About 200 other mutations have been reported in 17 additional genes [1,3,38,39,41–47], whereas no data are available on mutations in 19 of the 42 genes. The mutations reported by mid-2000 have been reviewed previously [3]. The main development since then is that several mutations in one additional gene, COL8A2, have been reported in two forms of corneal endothelial dystrophy, one of which is among the most common indications for corneal transplantation [46]. In addition, new mutations have been reported in the 22 genes in which mutations were known in 2000 [38–45,47].

Most of the mutations that have been reported are single-base substitutions that convert the codon of the obligate glycine to that of a bulkier residue and either prevent the folding of the triple helix beyond this point or cause an interruption in the helix. Because the triple helix of most collagens is propagated from the C-terminus to the N-terminus (Figure 2), there is a tendency for a glycine substitution that is closer to the C-terminal end of the triple-helical domain to produce a more severe phenotype than a similar substitution that is closer to the N-terminal end [1,3]. There are numerous exceptions to this rule, however, which are probably explained by the fact that the triple helix has regions of high and low stability as determined by the amino acids present in the X and Y positions [48]. Mutations in different regions, therefore, have different effects. A continuous triple helix seems to be particularly important for the fibril-forming collagens, whereas some collagens that normally contain several interruptions in their triple helices can tolerate an additional interruption with mild or no consequences. Other mutations change the codon of an X or Y-position residue to that of another amino acid or to a translation
stop codon. Further mutations lead to abnormal RNA splicing or are gene deletions, insertions, duplications or complex rearrangements. Most of the amino acid substitutions in the X and Y positions produce milder phenotypes than mutations of the obligate glycines, and some X and Y-position substitutions are probably non-pathogenic polymorphisms [3].

Mutations that lead to the production of a structurally altered polypeptide chain that is still able to associate with other chains usually cause more severe phenotypes than mutations that prevent trimer formation and heterozygous null alleles. Trimer formation from a mutant chain and normal chains can interfere with either the folding of the triple helix or the formation of the supramolecular assemblies. In the former case, the unfolded trimers containing both mutant and normal chains will first accumulate within cells and subsequently be degraded. If the chains form triple-helical molecules, the mutant molecules can have kinks or other abnormalities, which can reduce or delay the formation of supramolecular assemblies or alter their structure and function [1,3]. Most of the diseases caused by collagen mutations are dominantly inherited but there are also many examples of recessive inheritance [38–47].

The vast majority of the known collagen mutations have been identified in relatively rare heritable diseases (Table 2), including: (i) osteogenesis imperfecta, which is characterized by bone fragility, but also involves other tissues that are rich in type I collagen; (ii) various subtypes of the Ehlers-Danlos syndrome (EDS), a heterogeneous group of diseases characterized by joint hypermobility, skin changes, occasional skeletal deformities and rupture of the hollow organs; (iii) various chondrodysplasias, varying in severity from perinatal lethality to a very mild disease and non-syndromic hearing loss; (iv) autosomally inherited and X-linked forms of Alport syndrome, a disease characterized by haematuria and progressing to end-stage renal failure; (v) Behçet myopathy and Ulrich muscular dystrophy; (vi) two forms of epidermolysis bullosa, a blistering skin disease; (vii) two forms of cutaneous endothelial dystrophy; and (viii) Knobloch syndrome, a disease characterized by high myopia, vitreoretinal detachment and occipital encephalocele [1,3,38–47].

All the mutations in the two type I collagen genes that cause EDS type VII (Table 2) prevent cleavage of the N propeptide, EDS type VIIA results from mutations in the COLIAI gene and tends to be phenotypically more severe than type VIIb, which results from mutations in COLIA2 [3,41]. One COLIA1 mutation (not shown in Table 2) that replaces the codon of an X-position arginine with one of cysteine has also been reported in EDS I/II, thus indicating an overlap of phenotypes caused by mutations in different collagen types [3]. Overlaps are also seen in the case of cartilage collagens, where mutations in the main cartilage collagen, type II, usually produce more severe phenotypes than corresponding mutations in the minor cartilage collagens (i.e. types IX and XI) [3,42]. In the case of type VI collagen, Ulrich dystrophy is caused in most cases by recessive mutations whereas the milder Bethlem myopathy is caused by dominant mutations [3,45].

A few collagen mutations have also been identified in common diseases, such as osteoporosis, arterial aneurysms and the two most common musculoskeletal diseases, osteoarthrosis and intervertebral disc disease (Table 2). Of particular interest are two recently characterized mutations in the COL9A2 and COL9A3 genes that cause a codon of an X-position glutamine or Y-position arginine, respectively, to a tryptophan codon [49,50]. The mutation in the COL9A2 gene causes intervertebral disc disease in many families in a dominantly inherited fashion whereas the COL9A3 mutation is a strong predisposing factor for this disease. It is probable that many additional collagen mutations will be found that either cause or act as predisposing factors for these and other common diseases.

All the 42 collagen genes identified in humans are probably present in the mouse genome and many mouse lines (Table 2) are now available that harbour a mutation in a collagen gene [1,3,51–56]. Such mouse models are particularly useful for defining the significance and function of proteins such as collagens, which are large, insoluble and difficult to study functionally. In addition, mouse models are useful for analyzing the consequences of mutations in various genes of the superfamily, for identifying additional diseases caused by these mutations and even for testing potential therapies. Indeed, many mouse models have reproduced the phenotypes of various human diseases (Table 2) and in several cases have even led to the identification of additional human diseases that are caused by mutations in collagen genes [1,3].

Homozygous mutations have been reported in only three out of ~20 genes encoding various isoenzymes of the human collagen-modifying enzymes (Table 2), namely in the genes for lysyl hydroxylase-1 and a procollagen N proteinase (also known as ADAMTS-2), involved in two subtypes of EDS [1,3,18,41], and for lysyl hydroxylase-2 in Bruck syndrome, a disease characterized by osteoporosis, joint contractures, fragile bones and short stature [57]. The lysyl hydroxylase mutations prevent the formation of stable hydroxylysine-derived crosslinks, whereas the N proteinase mutations lead to an accumulation of partially processed molecules containing the N propeptide.

In addition to these human enzyme mutations, homozygous knockout mouse models have been generated [1,3,58–61] for genes encoding the catalytic subunit of two of the three known collagen prolyl 4-hydroxylase isoenzymes, one lysyl hydroxylase, one procollagen N proteinase, two procollagen C proteinases and the first described isoenzyme of lysyl oxidase (Table 2). Homozygous inactivation of many of these genes causes embryonic or perinatal lethality, demonstrating a crucial role for these enzymes in collagen synthesis and development.

Collagens and their modifying enzymes in Drosophila

Drosophila melanogaster has three conserved genes encoding basement membrane collagens, two α chains of type IV collagen and a homologue of type XV and XVIII collagens [62]. An additional gene encodes pericardin, a protein in which the collagen domain shows some
similarity to type IV and which is involved in the morphogenesis and maintenance of the heart epithelium during dorsal ectoderm closure [63], but Drosophila has no fibril-forming collagen with a long triple helix [62].

One surprising aspect of the Drosophila collagen superfamily is the presence of ~20 genes encoding polypeptides of 480–550 residues with a similarity to the catalytic α subunits of the vertebrate collagen prolyl 4-hydroxylases [19,64]. Many of these show tissue-specific embryonic expression (e.g. in the salivary gland, mouth-part precursor, proventriculus or epidermis) [64]. Only one of the encoded polypeptides has been characterized in detail [65]. It is expressed only in larvae and the embryonic mouth-part precursor, but not in adults [64,65], and combines with PDI to form an active α1β2 tetramer with properties similar to those of the vertebrate collagen prolyl 4-hydroxylases [65]. An additional gene encodes a HIF prolyl 4-hydroxylase indicating that Drosophila has a hypoxic response pathway similar to that in vertebrates [27]. The Drosophila collagen prolyl 4-hydroxylase family appears to be markedly larger than the corresponding vertebrate or nematode families, even though the number of collagen genes in Drosophila is much smaller. It is therefore possible that the functions of many of the Drosophila prolyl 4-hydroxylases might be to hydroxylate proline residues in proteins other than the collagens, and

### Table 2. Mutations in human collagens, their modifying enzymes and the corresponding mouse models

<table>
<thead>
<tr>
<th>Name</th>
<th>Human Gene</th>
<th>Human disease</th>
<th>Name</th>
<th>Mouse Gene</th>
<th>Mouse model</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL1A1, COL1A2</td>
<td>COL1A1, COL1A2</td>
<td>OI; osteoporosis; EDS type VI A and EDS type VII B [40,41]</td>
<td>Co1a1</td>
<td>TG, T</td>
<td></td>
</tr>
<tr>
<td>COL2A1</td>
<td>COL2A1</td>
<td>Several chondrodysplasias; osteoarthrosis [42]</td>
<td>Co2a1</td>
<td>TG, KO, N</td>
<td></td>
</tr>
<tr>
<td>COL3A1</td>
<td>COL3A1</td>
<td>EDS type IV; arterial aneurysms [41]</td>
<td>Co3a1</td>
<td>KO</td>
<td></td>
</tr>
<tr>
<td>COL4A3, COL4A4</td>
<td>COL4A3</td>
<td>Autosomal forms of Alport syndrome [43]</td>
<td>Co4a3</td>
<td>KO</td>
<td></td>
</tr>
<tr>
<td>COL4A5</td>
<td>COL4A5</td>
<td>X-linked forms of Alport syndrome [43]</td>
<td>Co4a5</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>COL5A1, COL5A2</td>
<td>COL5A2</td>
<td>Alport syndrome with diffuse oesophageal leiomyomatosis</td>
<td></td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>COL6A1, COL6A2, COL6A3</td>
<td>COL6A1, COL6A2, COL6A3</td>
<td>Bethlem myopathy; Ulrich muscular dystrophy [45]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL7A1</td>
<td>COL7A1</td>
<td>Dystrophic forms of EB [44]</td>
<td>Col7a1</td>
<td>KO</td>
<td></td>
</tr>
<tr>
<td>COL8A2</td>
<td>COL8A2, COL9A3</td>
<td>Two forms of corneal endothelial dystrophy [46]</td>
<td>Col9a1</td>
<td>TG, KO</td>
<td></td>
</tr>
<tr>
<td>COL10A1</td>
<td>COL10A1</td>
<td>Schmid metaphyseal chondrodysplasia [42]</td>
<td>Col10a1</td>
<td>TG, KO</td>
<td></td>
</tr>
<tr>
<td>COL11A1, COL11A2</td>
<td>COL11A1, COL11A2</td>
<td>Several mild chondrodysplasias; nonsyndromic hearing loss;</td>
<td>Col11a1</td>
<td>KO, T</td>
<td></td>
</tr>
<tr>
<td>COL12A1</td>
<td>COL12A1</td>
<td>Not identified; disruption of matrix structure of periodontal ligaments and skin in mouse</td>
<td>Col12a1</td>
<td>TG [52]</td>
<td></td>
</tr>
<tr>
<td>COL13A1</td>
<td>COL13A1</td>
<td>Not identified; embryonic lethality or progressive muscular atrophy in mouse</td>
<td>Col13a1</td>
<td>TG, T [53,54]</td>
<td></td>
</tr>
<tr>
<td>COL15A1</td>
<td>COL15A1</td>
<td>Not identified; skeletal myopathy and cardiovascular defects in mouse</td>
<td>Col15a1</td>
<td>KO [55]</td>
<td></td>
</tr>
<tr>
<td>COL17A1</td>
<td>COL17A1</td>
<td>Two forms of EB [44]</td>
<td>Col17a1</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>COL18A1</td>
<td>COL18A1</td>
<td>Knobloch and pigment dispersion syndromes [47]</td>
<td>Col18a1</td>
<td>KO [56]</td>
<td></td>
</tr>
<tr>
<td>COL19A1</td>
<td>COL19A1</td>
<td>Not identified; abnormal muscle layer in the oesophagus in mouse</td>
<td>Col19a1</td>
<td>KO</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modifying enzyme</th>
<th>Name</th>
<th>Human disease</th>
<th>Mouse model</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4H-α(Ⅰ)</td>
<td></td>
<td>Not identified; embryonic lethality in mouse</td>
<td>KO</td>
</tr>
<tr>
<td>P4H-α(Ⅱ)</td>
<td></td>
<td>Not identified; KO mice are viable</td>
<td>KO</td>
</tr>
<tr>
<td>BH-1</td>
<td></td>
<td>Type VI EDS [41]</td>
<td>KO</td>
</tr>
<tr>
<td>BH-2</td>
<td></td>
<td>Bruck syndrome [57]</td>
<td>None</td>
</tr>
<tr>
<td>BH-3</td>
<td></td>
<td>Not identified; embryonic lethality and lack of type IV collagen in basement membranes in mouse</td>
<td>KO</td>
</tr>
<tr>
<td>ADAMTS-2</td>
<td></td>
<td>type VIE EDS</td>
<td>KO [58]</td>
</tr>
<tr>
<td>BMP-1</td>
<td></td>
<td>Not identified; perinatal lethality and failure of ventral body wall closure in mouse</td>
<td>KO</td>
</tr>
<tr>
<td>Tolloid-like-1</td>
<td></td>
<td>Not identified; embryonic lethality and cardiac failure in mouse</td>
<td>KO [59]</td>
</tr>
<tr>
<td>Lox</td>
<td></td>
<td>Not identified; perinatal lethality and aortic aneurysms and cardiovascular dysfunction in mouse</td>
<td>KO [60,61]</td>
</tr>
</tbody>
</table>

*Abbreviations: OI, osteogenesis imperfecta; EDS, Ehlers-Danlos syndrome; EB, epidermolysis bullosa; P4H, collagen prolyl 4-hydroxylase; LH, lysyl hydroxylase; ADAMTS-2, N-proteinase isozyme; BMP-1, C-proteinase isozyme; Tolloid-like-1, C-proteinase isozyme; Lox, lysyl oxidase-1; T, transgenic (i.e. expression of a mutant polypeptide); KO, knock-out; T, other targeted, (i.e. knock-in); N, naturally occurring mutations.

*References to most human mutations are found in [1,3,38,39] and to most mouse mutations in [1,3,51]. References given in the Table therefore indicate only some additional reviews or recent original articles.

*The mouse gene is shown only when a mouse model is available.

*H. Sumiyoshi et al., unpublished, see Ref [3].

*T. Holster et al., unpublished.

The phenotype has not yet been analysed. O. Pakkanen et al., unpublished.

*K. Rautavaara et al., unpublished.
thus detailed studies on these enzymes might help researchers to identify additional functions for the human prolyl 4-hydroxylases.

The collagen families and their modifying enzymes in Caenorhabditis elegans

Two major collagen families are present in Caenorhabditis elegans – the cuticle collagens and the basement membrane collagens. The C. elegans cuticle is an exoskeleton that is synthesized by the underlying hypodermis. The major proteins of the cuticle are small collagen-like polypeptides of ~30 kDa encoded by a multigene family of ~175 members [66–68] (J. Kramer, unpublished). These polypeptides typically have two collagen domains, a smaller N-terminal domain and a larger C-terminal domain, with 8–10 and 40–42 Gly-X-Y repeats, respectively. The repeats in the C-terminal domain usually have 1–4 small interruptions. These two domains are separated and flanked by three cysteine-containing, non-collagenous domains, and the cuticle collagen family can be divided into four main groups and several additional small groups on the basis of conserved cysteine patterns. The N-terminal non-collagenous domain varies in size and contains a cleavage site for a putative subtilisin-like protease that processes a procollagen precursor to a mature polypeptide [66–68]. The cuticle is synthesized five times during the life cycle, once in the embryo before hatching and subsequently at the end of each of the four larval stages before moult. The cuticle collagen genes are expressed in a distinct temporal fashion, the pattern of which is repeated during each cuticle synthesis but not all the genes are expressed at the same time; some are early, some are intermediate and some are late with respect to the secretion of the new cuticle [67,68]. Sets of collagens that are temporally coexpressed have been shown in genetic studies to interact and to be capable of forming functionally distinct structures [69,70]. The collagen collagens contain many interchain disulfide bonds that can be either intramolecular or intermolecular. These collagens also contain tyrosine and putative γ carboxyl glutamine-derived crosslinks that are not found in vertebrate collagens, the tyrosine crosslinks probably being in the form of di and trityrosines [66–68].

The C. elegans basement membrane collagens consist of homologues of vertebrate type IV, a heterotrimer of α1(IV) and α2(IV) chains, and a type XVIII homotrimer, [α1(XVIII)]3 [66,71]. The catalytic α subunits of collagen prolyl 4-hydroxylases are encoded by four genes, phy-1–phy-4 [18,19,72–76]. PDI has two isoforms in C. elegans, PDI-1 and PDI-2, both have disulfide isomerase activity [68] and PDI-2 also serves as the β subunit in the prolyl 4-hydroxylase forms that are involved in the synthesis of cuticle collagens (Figure 3). PHY-1 and PHY-2 combine with PDI-2 to form a unique mixed tetramer (Figure 3) that catalyzes the synthesis of the cuticle collagens. Both PHY subunits also form a dimer with PDI-2 (Figure 3), although PHY-2 does this very ineffectively [75]. Phy-3 is expressed in embryos, late larval stages and adult nematodes but expression in adults is restricted to the spermatheca (a specialized region of the gonad where oocytes are fertilized) [76]. PHY-3 forms an active enzyme only with PDI-1, but its molecular composition is unknown, whereas PHY-4 has not yet been characterized. A further gene, egl-9 encodes a cytoplasmic HIF prolyl 4-hydroxylase, which is involved in the regulation of the response to hypoxia [26].

Lysoyl hydroxylase has only one isoenzyme in C. elegans [30,77], which is essential for the synthesis of type IV collagen [77]. Additional enzymes include: subtilisin-like proteases that are required in the processing of the cuticle procollagens [68]; a thioredoxin-like enzyme needed for proper crosslinking [78]; a homologue of the vertebrate PDI-like protein ERp60 with both disulfide isomerase and transglutaminase-like crosslinking activity [79]; and two multidomain dual oxidases with oxidase and peroxidase activity that catalyze the formation of the tyrosine-derived crosslinks [80].

Mutations in C. elegans collagens and their modifying enzymes

Many mutations in the cuticle collagen genes affect the body morphology (Table 3), and most of the mutations that have been characterized so far result from a glycine codon being altered to that of another amino acid. Insertion of a mutant collagen into the cuticle often causes a more severe phenotype than null alleles or mutations that lead to a loss or a reduction in the amount of the collagen in the matrix [66–70]. Different mutations in a given gene can produce different phenotypes (Table 3). Homozygous null alleles of the sqt-1 or rol-6 gene that result in complete loss of the SQP-1 or ROL-6 collagen, for example, have mild consequences, leading to a defective tail structure or a mild dumpy phenotype, respectively [81]. Mutations in either of these two genes that remove a conserved cysteine in the C-terminal non-collagenous domain eliminate a disulfide bond that is necessary for the formation of a tyrosine-derived cross-link and cause a recessive left-handed roller phenotype, whereas mutations that affect the N propeptide cleavage site and lead to an accumulation of molecules that retain this propeptide produce a recessive dumpy and dominant right-handed roller phenotype [81]. Interestingly, a corresponding mutation that affects the N propeptide cleavage site in the dpy-10 gene causes a dominant left-handed roller [81]. These mutations are similar to those found in the human COL1A1 and COL1A2 genes discussed previously, which prevent cleavage of the N propeptide from type I procollagen and lead to EDS types VIIA and VIIB, respectively. The data currently available indicate a high degree of complexity and redundancy between the C. elegans cuticle collagens, their interacting partners and the higher-order structures that they form [69].

Mutations in either of the two type IV basement membrane collagen genes are embryonically lethal, indicating that this collagen is essential for C. elegans embryogenesis [82], whereas mutations in the type XVIII collagen gene cause cell and axon migration defects and affect the organization and function of neuromuscular junctions [71,83]. Mutations in the lysoyl hydroxylase gene lead to intracellular accumulation of type IV collagen, which is then absent from basement membranes [77], and are therefore also lethal at the embryonic stage (Table 3).
C. elegans are coded by two conserved genes, as observed by RNA interference experiments. Abbreviations: Duox, dual oxidase; LH, lysyl hydroxylase; PDI, protein disulfide isomerase; P4H, prolyl 4-hydroxylase.

Recessive dumpy/dominant or recessive roller.

Definitions: Dumpy, shortening in the length and thickening of the nematode; Blister, blistering of the cuticle; roller, helical twisting of the nematode body; Long, elongation of the nematode.

Figure 3. Schematic representation of the forms of collagen prolyl 4-hydroxylase characterized in vertebrates and C. elegans, and phenotypes resulting from inactivation of the Caenorhabditis elegans phy-1 and phy-2 genes. (a) The three vertebrate isoenzymes have unique catalytic α subunits and the same β subunits [i.e. the protein disulfide isomerase (PDI) polypeptide] Refs [18,19,24,25]. (b) The catalytic α subunits of the prolyl 4-hydroxylase forms that catalyze the synthesis of the cuticle collagens in C. elegans are coded by two conserved genes, phy-1 and phy-2 [18,19,72–74]. The processed PHY-1 and PHY-2 polypeptides consist of 543 and 523 residues, respectively, PHY-1 being slightly longer than the human α subunits, which are 514–525 residues. PHY-1 and PHY-2 combine with a single β subunit, PDI-2, both in recombinant expression systems and in vivo to form a unique mixed tetramer PHY-1–PHY-2–(PDI-2)2, whereas neither forms a tetramer in the absence of the other [76]. Both also form a PHY–PDI-2 dimer, although the PHY-2–PDI-2 dimer is formed only in small, almost non-detectable amounts [75]. Homozygous inactivation of either the phy-1 or phy-2 prevents formation of the mixed tetramer [75]. The null mutant nematodes can in part compensate for the lack of the tetramer by increasing the formation of the corresponding PHY–PDI-2 dimer but the phy-1 mutants do this ineffectively owing to the very small amount of the PHY-2–PDI-2 dimer that is formed [76] and, therefore, have a dumpy phenotype (d), whereas the phy-2 null mutants (e) are of the wild-type (f) [72–75]. The homozygous phy-1,phy-2 double null [72,74] and the pdi-2 null mutants (not shown) [74] lack all the forms of prolyl 4-hydroxylase involved in the synthesis of the cuticle collagens [75] and therefore either have a severe dumpy phenotype or are embryonically lethal. Antony Page is gratefully acknowledged for panels c–f. Panel f has been reproduced with permission from Ref. [74].

Table 3. Mutations in Caenorhabditis elegans collagens and their modifying enzymes

<table>
<thead>
<tr>
<th>Polypeptide collagen</th>
<th>Gene</th>
<th>Typical phenotype*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuticle collagen</td>
<td>dpy-2, dpy-3, dpy-5, dpy-7, dpy-8, dpy-10, dpy-13</td>
<td>Dumpy</td>
</tr>
<tr>
<td></td>
<td>bli-1, bli-2</td>
<td>Blister</td>
</tr>
<tr>
<td></td>
<td>rol-6</td>
<td>Dpy-2 null mutants</td>
</tr>
<tr>
<td></td>
<td>sqt-1, sqt-3</td>
<td>Dpy-3 null mutants</td>
</tr>
<tr>
<td></td>
<td>lon-3</td>
<td>Long</td>
</tr>
<tr>
<td></td>
<td>emb-9, let-2</td>
<td>Embryonically lethal</td>
</tr>
<tr>
<td>Collagen IV: α1(IV); α2(IV)</td>
<td>cle-1</td>
<td>Defects in cell and axon migration and neuromuscular synapse function</td>
</tr>
<tr>
<td>Collagen XVIII</td>
<td>cle-1</td>
<td>Dumpy</td>
</tr>
<tr>
<td>P4H*, PHY-1</td>
<td>phy-1 (also known as dpy-18)</td>
<td>Wild-type</td>
</tr>
<tr>
<td>P4H*, PHY-2</td>
<td>phy-2</td>
<td>Wild-type</td>
</tr>
<tr>
<td>P4H*, PHY-1 and PHY-2</td>
<td>phy-1 and phy-2</td>
<td>Severe dumpy or embryonically lethal*</td>
</tr>
<tr>
<td>P4H*, PHY-3</td>
<td>phy-3</td>
<td>Wild-type</td>
</tr>
<tr>
<td>PDI-2</td>
<td>pdi-2</td>
<td>Severe dumpy or embryonically lethal*</td>
</tr>
<tr>
<td>LH</td>
<td>let-288</td>
<td>Embryonically lethal</td>
</tr>
<tr>
<td>Subtilisin-like protease</td>
<td>bli-4</td>
<td>Embryonically lethal or blister</td>
</tr>
<tr>
<td>Thiopeptidase</td>
<td>dpy-11</td>
<td>Dumpy</td>
</tr>
<tr>
<td>ER60</td>
<td>pdi-3</td>
<td>Mild disruption of cuticle collagen localization</td>
</tr>
<tr>
<td>Duox 1; duox 2</td>
<td>F56C11.1; F53G12.3</td>
<td>Dumpy and blister</td>
</tr>
</tbody>
</table>

*Definitions: Dumpy, shortening in the length and thickening of the nematode; Blister, blistering of the cuticle; roller, helical twisting of the nematode body; Long, elongation of the nematode.

*Rescessive dumpy/dominant or recessive roller.

*Abbreviations: Duox, dual oxidase; LH, lysyl hydroxylase; PDI, protein disulfide isomerase; P4H, prolyl 4-hydroxylase.

*As observed by RNA interference experiments.
homologous inactivation of genes for most of the other collagen modifying enzymes also results in severe phenotypes (Table 3), indicating a crucial role for these enzymes in the synthesis of the various C. elegans collagens.

Conclusions
It is now well established that collagens and proteins with collagen domains form large superfamilies in many species. The number of family members is constantly growing but research to elucidate the specific properties and functions of the different members has a long way to go. The collagen-modifying enzymes also form large families with multiple isoenzymes, although research into their expression patterns and functions is still in its early stages. Collagen prolyl 4-hydroxylases play a key role in the synthesis of all collagens. An interesting new development has been the identification of a second prolyl 4-hydroxylase family that plays a key role in the regulation of the hypoxia-inducible transcription factor HIF, and it seems likely that other proteins might be found in which 4-hydroxyproline residues play a crucial role. An interesting aspect of the Drosophila collagen family is the presence of ~20 putative prolyl 4-hydroxylase isoenzymes, even though the number of collagens in Drosophila is much less than in vertebrates and C. elegans. Many of these enzymes might therefore be involved in the hydroxylation of proline residues in proteins other than the collagens, and studies on these enzymes might help identify additional functions for vertebrate prolyl 4-hydroxylases. Numerous human collagen mutations have been characterized, but these probably represent only a small fraction of all the existing mutations. It will be important to learn how critical are the roles of collagen mutations as direct causes or predisposing factors in common diseases. The existing mouse and C. elegans models for mutations in collagens and their modifying enzymes, and several additional mutants that are likely to be generated or identified, will provide important information on the functions of various members of the superfamily and the effects of mutations in them.

References
12 Sato, K. et al. (2002) Type XXVI collagen, a new member of the collagen family, is specifically expressed in the testis and ovary. J. Biol. Chem. 277, 37678–37684
20 Prockop, D.J. et al. (1998) Two unusual metalloproteinases that are essential for procollagen processing probably have important roles in development and cell signaling. Matrix Biol. 16, 399–408
26 Epstein, A.C.R. et al. (2001) C. elegans EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. Cell 107, 43–54
29 Hirasil, M. et al. (2003) Characterization of the human prolyl 4-hydroxylases that modify the hypoxia-inducible factor HIF. J. Biol. Chem. 278, 30772–30780
59 Clark, T.G. et al. (1999) The mammalian Tolloid-like 1 gene, Tll1, is necessary for normal septation and positioning of the heart. Development 126, 2631–2642
80 Edens, W.A. et al. (2001) Tyrosine cross-linking of extracellular matrix is catalyzed by Dux, a multidomain oxidase/peroxidase with homology to the phagocyte oxidase subunit gp91phox. J. Cell Biol. 154, 879–891

Articles of interest in Trends and Current Opinion journals

Progress in functional genomics approaches to antifungal drug target discovery
Marianne D. De Backer and Patrick Van Dijck
Trends in Microbiology 11, 470–478

Huntington’s disease: a synaptopathy?
Jia-Yi Li, Markus Plomann and Patrik Brundin
Trends in Molecular Medicine 9, 414–420

Inflammation, degeneration and regeneration in the injured spinal cord: insights from DNA microarrays
Florence M. Bareyre and Martin E. Schwab
Trends in Neurosciences 26, 555–563

Mining genome databases for therapeutic gold: SIM2 is a novel target for treatment of solid tumors
Rajiv R. Ratan
Trends in Pharmacological Sciences 24, 508–510

Turning germ cells into stem cells
Peter J. Donovan and Maria P. de Miguel
Trends in Biotechnology 21, 428–432

Peptidylarginine deiminase type 4: identification of a rheumatoid arthritis-susceptible gene
Ryo Yamada, Akari Suzuki, Xiotian Chang and Kazuhiko Yamamoto
Trends in Molecular Medicine 9, 503–508

Slipping while sleeping? Trinucleotide repeat expansions in germ cells
Christopher E. Pearson
Trends in Molecular Medicine 9, 490–495

Gene repression by Polycomb group protein complexes: a distinct complex for every occasion?
Arie P. Otte and Ted H.J. Kwaks
Current Opinion in Genetics and Development 13, 448–454

http://tigs.trends.com