The triple helix of collagens – an ancient protein structure that enabled animal multicellularity and tissue evolution

Aaron L. Fidler1,2,+, Sergei P. Boudko1,2,+, Antonis Rokas3 and Billy G. Hudson1,2,4,5,6,7,8,9,

ABSTRACT
The cellular microenvironment, characterized by an extracellular matrix (ECM), played an essential role in the transition from unicellularity to multicellularity in animals (metazoans), and in the subsequent evolution of diverse animal tissues and organs. A major ECM component and member of the collagen superfamily – comprising 28 types in vertebrates – that exist in diverse supramolecular assemblies ranging from networks to fibrils. Each assembly is characterized by a hallmark feature, a protein structure called a triple helix. A current gap in knowledge is understanding the mechanisms of how the triple helix encodes and utilizes information in building scaffolds on the outside of cells. Type IV collagen, recently revealed as the evolutionarily most ancient member of the collagen superfamily, serves as an archetype for understanding the mechanisms of how a triple helix encodes and utilizes information in building the extracellular matrix, which enabled animal multicellularity and tissue evolution.

KEY WORDS: Cell biology, Collagen, Evolution, Extracellular matrix, Multicellularity, Triple helix

Introduction
The extracellular matrix (ECM) played an essential role during the transition from unicellular organisms to multicellular animals (metazoans). The ECM comprises a basement membrane (BM) that underlies epithelia cells, and an interstitial matrix (IM) that is positioned between cells in the intercellular spaces and undergoes continuous controlled remodeling (Hynes, 2012; Bonnans et al., 2014; Nelson and Bissell, 2006; Inman et al., 2015). Yet, a major gap in cell biology is to understand how cells generate and interact with the ECM (Sherwood, 2015; Jayadev and Sherwood, 2017).

The collagen superfamily of proteins is a major component of ECMs, which – in vertebrates – comprises 28 types (I–XXVIII) that are derived from a total of 46 α-chains across the superfamilly (Fig. 1) (Ricard-Blum, 2011, Kadler et al., 2007; Ricard-Blum and Ruggiero, 2005). Invertebrates generally contain collagen IV, XV or XVIII, some fibrillar collagens, as well as some fibril-associated collagens with interrupted triple helices (FACITs) (Fidler et al., 2014, 2017; Fahey and Degnan, 2010; Meyer and Moussian, 2009; Boot-Handford and Tuckwell, 2003; Whittaker et al., 2006; Kadler et al., 2007). Among these collagens, type IV is the evolutionarily most ancient, based on recent studies of non-bilaterian animals (sponges, ctenophores, placozoans and cnidarians) and unicellular groups (Fidler et al., 2017; Grau-Bove et al., 2017) (Fig. 1).

Collagens are the most abundant protein in the human body (Kadler et al., 2007; Shoulders and Raines, 2009). They occur as diverse supramolecular assemblies, ranging from networks to fibrils, and broadly function in structural, mechanical and organizational roles that define tissue architecture and influence cellular behavior (Shoulders and Raines, 2009; Ricard-Blum, 2011; Ricard-Blum and Ruggiero, 2005). Defects in collagens underlie the cause of almost 40 human genetic diseases, affecting numerous organs and tissues in millions of people worldwide (summarized in Table 1).

Disease pathogenesis typically involves genetic alterations of the triple helix, a unique structure that is a hallmark feature common to all collagens. The triple helix bestows exceptional mechanical resistance to tensile forces and a capacity to bind a plethora of macromolecules. Yet, there is a gap in our current knowledge in understanding the mechanisms of how a triple helix encodes and utilizes information in building supramolecular assemblies on the outside of cells. Here, we present collagen IV, the most ancient of the collagen superfamily, and argue that it is ideally suited to serve as an archetype for investigating and describing core functions of a triple helix.

The triple helix – assembly and structural features that encode information
The chemical structure of the triple helix was determined through the seminal work of structural biologists and chemists over the last century (see Box 1 in supplementary material). Its unique structure bestows upon collagens an exceptional mechanical resistance to tensile forces and a plethora of organizing information for building an ECM (Fig. 2). The triple helix presents all residues, except glycine (Gly), on its surface, which is the most economical and robust way to encode binding motifs of any protein structure. Moreover, the triple helix exhibits extensive post-translational modifications (PTMs), such as hydroxylation, glycosylation and phosphorylation, adding – in tandem – a secondary layer of information in addition to its amino acid (aa) code (Yamauchi and Shiiba, 2008). These PTMs confer even more diversity with tissue-specific and disease-specific variations, even amongst identical types of collagen (Pokidysheva et al., 2013). Furthermore, additional collagen modifications are mediated by specific extracellular enzymes, such as peroxidasin and lylx oxidases-like
proteins (LOXLs) for crosslinking, and Goodpasture antigen-binding protein (COL4A3BP, hereafter referred to as GPBP) and other extracellular kinases for phosphorylation (Bhave et al., 2012; Añazco et al., 2016; Revert et al. 1995, Raya et al. 1999; Yalak and Olsen, 2015). Non-enzymatic modifications, such as glycation, oxidation or chlorination, add even more complexity (Brown et al., 2015). Together, these modifications may serve as regulatory mechanisms on the outside of cells that may instruct cell behavior and influence tissue architecture and stability (Yalak and Olsen, 2015; Pedchenko et al., 2010).

To fully appreciate the capacity and versatility of the triple helix, one should consider the building blocks (polypeptide chains) of this structure. These chains are simple ‘rope-like’ structures built from proline residues. In aqueous solutions, such a polypoline chain adopts a polyproline type II (PPII) helix in a left-handed conformation (Adzhubei et al., 2013) (Fig. 2A). A unique feature of this structure is that it is non-extensible (Okuyama et al., 1981). Such simple ‘ropes’ occur in the extracellular space of plants (Lamport, 1974) and have been identified as extensions of triple helices within mini-collagens of cnidarians (e.g. Hydra) (Holstein et al., 1994), where they perform structural and mechanical roles. To allow for the formation of a triple helix from this simple rope, every third proline residue of the PPII helix is replaced with a glycine residue – the smallest aa residue as it lacks a side chain. This endows the structure with flexibility and sufficient space to tightly pack three chains (Fig. 2A) that are each a left-handed helix, wind around each other with a shift in one residue, so that the smaller glycine residues are buried inside the triple helix, whereas only proline residues are exposed on the surface (Fig. 2A) (Ramachandran and Kartha, 1954; Rich and Crick, 1955; Okuyama et al., 1981). Moreover, as these three chains wind together forming a triple helix, each individual left-handed chain adopts a right-handed superhelix. The unique role of the glycine residues in the packing of the collagen triple helix explains the adverse effects of mutations at these positions, as any other residue in place of glycine will distort its tight packing (Bella et al., 1994).

Three main changes are achieved upon the transition from a single polyproline helix to a triple helix: (i) higher bending rigidity, (ii) a less accessible chain backbone that is, thus, less prone to proteolysis and, (iii) the ability to essentially accept any aa in place of the proline residues at position X and Y without any significant destabilization to the triple helix structure (Fig. 2B). Moreover, proline and other aa within the triple helix remain accessible to solvent and, thus, their numerous known PTMs are possible without disturbing the native helix structure. For example, probably the most important PTM is the 4(R)-hydroxylation of proline residues position Y, because it substantially increases the thermal stability of the collagen triple helix (Sakakibara et al., 1973). Although all proline and hydroxyproline residues are essential for the stability of a PPII helix, ~65% of aa at these positions can vary in the native triple helix, while still maintaining its stability. Thus, the triple helix, in contrast to a single PPII helix, confers additional capacity to specify binding partners.

Once assembled from three superhelical α-chains, the collagen triple helix provides several distinct sites to tether macromolecules through three modes of binding motifs (single, double and triple α-chain) with further three levels of variations to these modes (mode 1, 2 and 3) (Fig. 2B). The three binding modes operate by utilizing...
Table 1. Human genetic diseases/disorders due to mutation(s) in collagens and experimental collagen knockout studies

<table>
<thead>
<tr>
<th>Collagen type</th>
<th>Human genetic diseases/disorders due to mutation(s) in collagens</th>
<th>Loss of protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (G)</td>
<td>Osteogenesis imperfecta (Wenstrup et al., 1988; Vogel et al., 1987; Peltonen et al., 1988; Lamande et al., 1989; Labhard et al., 1988; Cohn et al., 1986; Baldwin et al., 1989)</td>
<td>Embryonic lethality during organogenesis before E15, embryonic growth arrest at E11.5–E12 (Schnieke et al., 1983; Harbers et al., 1984)</td>
</tr>
<tr>
<td>(N) Ehlers–Danlos syndrome (Wirtz et al., 1987; Malfait et al., 2006; Nuytinck et al., 2000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffey disease (Gensure et al., 2005) (for all, see <a href="http://www.le.ac.uk/ge/collagen/">http://www.le.ac.uk/ge/collagen/</a>)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II (G)</td>
<td>Spondyloepiphyseal dysplasias (Lee et al., 1989; Tiller et al., 1990; Chan et al., 1993; Jurgens et al., 2015; Barat-Houari et al., 2016)</td>
<td>Perinatal lethality, abnormal skeleton, abnormal cranium, abnormal respiratory system (Li et al., 1995a). IIA isoform-specific knockout: prenatal lethality, holoprosencephal</td>
</tr>
<tr>
<td>(G) Achondrogenesis 2 (Vissing et al., 1989; Mortier et al., 1995, 2000; Körkö et al., 2000; Chan et al., 1995; Barat-Houari et al., 2016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Legg–Calvé–Perthes disease (Miyamoto et al., 2007)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Knies dysplasia (Wilkin et al., 1994; Winterpacht et al., 1996)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Avascular necrosis of femoral head (Liu et al., 2005; Kannu et al., 2011)</td>
<td></td>
</tr>
<tr>
<td>Platsypodylic lethal skeletal dysplasia, Torrance type (Nishimura et al., 2004; Zankl et al., 2005)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spondyloperipheral dysplasia (Zankl et al., 2005; Nishimura et al., 2004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Stickler syndrome 1 (Richards et al., 2000, 2010; McAlinden et al., 2008; Körkö et al., 1993; Barat-Houari et al., 2016)</td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>Osteoarthritis with mild chondro dysplasia (Ala-Kokko et al., 1990)</td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>Czech dysplasia (Williams et al., 1993; Tzschach et al., 2008; Matsui et al., 2009; Barat-Houari et al., 2016)</td>
<td></td>
</tr>
<tr>
<td>III (G)</td>
<td>Ehlers–Danlos syndrome, vascular types (Narcisi et al., 1994; Tromp et al., 1989a,b; Pepin et al., 1993) (<a href="http://www.le.ac.uk/ge/collagen/">http://www.le.ac.uk/ge/collagen/</a>)</td>
<td>Embryonic lethality before E9.5, pre- to mid-gestation lethality (Smith et al., 2011)</td>
</tr>
<tr>
<td>(G)</td>
<td>Cerebral small vessel disease (Gould et al., 2006; Shah et al., 2012; Sibon et al., 2007)</td>
<td>Embryonic lethality at E10.5–E11.5 (Poschl et al., 2004)</td>
</tr>
<tr>
<td>(G)</td>
<td>Hereditary angiopathy with nephropathy aneurysms and muscle cramps (HANAC) (Plaisier et al., 2007, 2010)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Porencephaly 1 (Breedveld et al., 2006; de Vries et al., 2009; Gould et al., 2005; Yoneda et al., 2013)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Porencephaly 2 (Yoneda et al., 2012)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Tortuosity of retinal arteries (Zenteno et al., 2014)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Schizencephaly (Yoneda et al., 2013)</td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>Intracerebral hemorrhage (Jeanne et al., 2012; Weng et al., 2012)</td>
<td></td>
</tr>
<tr>
<td>IV (α1β1γ2)</td>
<td>Alport syndrome (Barker et al., 1990; Boye et al., 1998; Heidet et al., 2001; Kashtan, 2017; Knebelmann et al., 1992; Mehta and Jim, 2017; Mochizuki et al., 1994; Nagel et al., 2005; Renieri et al., 1992; Zhou et al., 1991; Van Der Loop et al., 2000; Zhou et al., 1992) (<a href="http://databases.lovd.nl/shared/genes/COL4A5">http://databases.lovd.nl/shared/genes/COL4A5</a>)</td>
<td>Decreased survival rate, premature death, proteinuria, glomerulosclerosis, glomerulonephritis, renal interstitial fibrosis, hearing loss (Cosgrove et al., 1996, 1998; Rheault et al., 2004; LeBlue et al., 2010)</td>
</tr>
<tr>
<td>(G)</td>
<td>Thin basement membrane nephropathy (TBMD) (Badenas et al., 2002; Buzza et al., 2003; Lemmink et al., 1996; Kashtan, 2004)</td>
<td></td>
</tr>
<tr>
<td>IV (α5γ5α6)</td>
<td>X-linked deafness 6 (Rost et al., 2014)</td>
<td>No abnormal phenotype detected (Fox et al., 2007)</td>
</tr>
<tr>
<td>V (G)</td>
<td>Ehlers–Danlos syndrome (Watanabe et al., 2016; Symoens et al., 2009; Richards et al., 1998; Michalickova et al., 1998; Malfait et al., 2005; Giunta and Steinmann, 2000; Giunta et al., 2002; De Paepe et al., 1997) (<a href="http://www.le.ac.uk/ge/collagen/">http://www.le.ac.uk/ge/collagen/</a>)</td>
<td>Embryonic lethality at E10–E12 (Wenstrup et al., 2004; Park et al., 2015)</td>
</tr>
<tr>
<td>(G)</td>
<td>Bethlem myopathy 1 (Scacheri et al., 2002; Pepe et al., 1999; Pan et al., 1998; Lucidi et al., 2005; Lampe et al., 2006; Jöbsis et al., 1996; Baker et al., 2007)</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>Ullrich congenital muscular dystrophy 1 (Lampe et al., 2005; Kawahara et al., 2007; Giusti et al., 2005; Baker et al., 2005) Dystonia 27 (Zech et al., 2015)</td>
<td></td>
</tr>
<tr>
<td>VII (G)</td>
<td>Epidermolysis bullosa dystrophica, multiple types (Shinkuma, 2015; Fine et al., 1989; Christiano et al., 1993, 1994, 1995, 1996; Woodley et al., 2004)</td>
<td>Postnatal lethality within the first week from complications of blistering (Heinonen et al., 1999)</td>
</tr>
<tr>
<td>VIII (N)</td>
<td>Corneal dystrophies (Biswas et al., 2001)</td>
<td>Dysgenesis of the anterior segment of the eye (Hopfer et al., 2005)</td>
</tr>
<tr>
<td>IX (G)</td>
<td>Multiple epiphyseal dysplasia 2, 3 and 6 (Paassilta et al., 1999; Jeong et al., 2014; Holden et al., 1999; Czarny-Ratajczak et al., 2001) Stickler Syndrome 4 and 5 (Van Camp et al., 2006; Baker et al., 2011)</td>
<td>Osteoarthrits, abnormalities in morphology of bone and cartilage, abnormal bone healing, impaired hearing (Fassler et al., 1994; Asamura et al., 2005; Ting et al., 1999)</td>
</tr>
</tbody>
</table>

Continued
Table 1. Continued

<table>
<thead>
<tr>
<th>Collagen type</th>
<th>Human genetic diseases/disorders due to mutation(s) in collagens</th>
<th>Loss of protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Schmid metaphyseal chondrodysplasia (Wallis et al., 1994; McIntosh et al., 1994; Bateman et al., 2005)</td>
<td>Abnormal spleen morphology, abnormal hematopoietic system (Rosati et al., 1994)</td>
</tr>
<tr>
<td>XI</td>
<td>(G) Stickler syndrome 2 (Annunen et al., 1999a; Richards et al., 1996, 2010), Marshall syndrome (Annunen et al., 1999a) (G) Fibrochondrogenesis 1 (Tompson et al., 2010) (G) Otospondylocephalohypoplasia dysplasia (OSMED) (Melkonieni et al., 2000; Pihlajamaa et al., 1998; Sirko-Osadca et al., 1998; Vikkula et al., 1995) (G) Deafness (Chakchouk et al., 2015; Chen et al., 2005; McGirt et al., 1999) Fibrochondrogenesis 2 (Tompson et al., 2012)</td>
<td>Neonatal lethality; abnormalities in cartilage of limbs, ribs, mandible and trachea (Li et al., 1999b)</td>
</tr>
<tr>
<td>XII</td>
<td>Ullrich congenital muscular dystrophy 2 (Zou et al., 2014) (G) Bethlem myopathy 2 (Zou et al., 2014; Hicks et al., 2014)</td>
<td>Increased perinatal lethality, skeletal abnormalities, decreased body weight, abnormal skeletal muscle morphology and physiology (Izu et al., 2011; Zou et al., 2014)</td>
</tr>
<tr>
<td>XIII*</td>
<td>Congenital myasthenic syndrome type 19 (Logan et al., 2015)</td>
<td>Deletion of cytosolic and transmembrane regions, abnormal skeletal muscle morphology, muscle degeneration, myopathy, impaired exercise endurance, increased susceptibility to injury (Kivist et al., 2001)</td>
</tr>
<tr>
<td>XIV</td>
<td>unknown</td>
<td>Decreased tendon stiffness, decreased skin tensile strength (Anorge et al., 2009)</td>
</tr>
<tr>
<td>XV</td>
<td>unknown</td>
<td>Abnormal muscle morphology, increased muscle regeneration, myopathy (Eklund et al., 2001)</td>
</tr>
<tr>
<td>XVI</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>XVII</td>
<td>(G) Generalized atrophic benign epidermolysis bullosa (McGrath et al., 1996; Schumann et al., 1997; Tasanen et al., 2000a,b; Wu et al., 2002) Epithelial recurrent erosion dystrophy (Jonsson et al., 2015)</td>
<td>Significant postnatal lethality, sterility, postnatal growth retardation, blistering, abnormal renal glomerulus morphology (Hurskainen et al., 2012; Nishie et al., 2007)</td>
</tr>
<tr>
<td>XVIII</td>
<td>Knobloch syndrome (Aldahmesh et al., 2013; Sertie et al., 2000)</td>
<td>Eye abnormalities, intracranial hemorrhage, abnormalities in morphology of brain, cranium and renal glomerulus; abnormal lipid metabolism (Bishop et al., 2010; Fukai et al., 2002; Utirainen et al., 2004)</td>
</tr>
<tr>
<td>XIX</td>
<td>unknown</td>
<td>~95% postnatal lethality, dysphagia, abnormal esophagus morphology; weight loss (Sumiyoshi et al., 2004)</td>
</tr>
<tr>
<td>XX to XXIV</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>XXV</td>
<td>(G) Congenital cranial dysinnervation disorder (Shinwari et al., 2015)</td>
<td>Neonatal lethality; abnormalities in morphology of motor neurons, skeletal muscles; respiratory failure; cyanosis (Tanaka et al., 2014)</td>
</tr>
<tr>
<td>XXVI</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>XXVII*</td>
<td>(G) Steel syndrome (Gonzaga-Jauregui et al., 2015)</td>
<td>87-residue deletion of collagenous domain; neonatal lethality; chondrodysplasia; respiratory failure at birth; abnormalities in morphology of lung and skeleton; decreased fetal size (Plumb et al., 2011)</td>
</tr>
<tr>
<td>XXVIII</td>
<td>unknown</td>
<td>unknown</td>
</tr>
</tbody>
</table>

G, Gly and non-Gly mutations reported in triple helix; N, only non-Gly mutations reported in triple helix; *deletion mutants for collagens XIII and XXVII

one, two or all three distinct chains of the triple helix in different combinations to directly bind a molecule (Fig. 2B; modes 1–3). The three levels of variation to these modes add additional diversity to binding specificity on top of involving either one, two or three chains. The first level stems from the variation that is provided by the ≤20 possible aa residues that can occupy the variable positions X and Y of each tripeptide; this results in extensive structural variability across the three chains, as well as within each Gly-X-Y tripeptide of the same chain (Fig. 2B; level 1). The second level confers even more diversity because of the proclivity of collagens to become post-translationally modified and change their structure (Fig. 2B; level 2). The third level utilizes chain staggering that occurs during triple helix assembly between either two or three chains and enables new combinations of chains that can give rise to additional binding motifs (Fig. 2B; level 3).

In addition to these binding specificities, the triple helix also possesses other structural features that underlie its biological function. It is non-stretchable along its longitudinal axis, providing great tensile strength to withstand all physiological mechanical loads and stresses in our body. It also confers resistance to proteases – making collagens some of the most long-lived proteins – as well as to a wide range of pH values in order to withstand adverse conditions (Steven, 1965; Hafer and Hörmann, 1963; Grant and Alburn, 1960; Drake et al., 1966; Uzawa et al., 1998; Pokidysheva et al., 2013; Hudson et al., 2017; Eyre et al., 2011). Furthermore, the triple helix exhibits variable lengths among collagen suprastructures, such as networks, fibrils and filaments (Ricard-Blum, 2011). For example, the triple helix length of fibrillar collagens results from multiple duplications of exons that are either 54 or 45 base pairs in length and encode Gly-X-Y (Yamada et al., 1980; Exposito et al., 2010). Moreover, the terminal and internal incorporation of non-triple helical sequences (e.g. interruptions) in all collagen types extends their functional capacity (Ricard-Blum, 2011; Khoshnoodi et al., 2006). Collectively, the binding motifs (see above) together with these features of a collagen triple helix underlie the diversity of...
biological activities of the diverse supramolecular assemblies of collagens.

**Fig. 2. Building a collagen triple helix and encoding information.** (A) The building blocks (polypeptide chains) of the triple helix are left-handed polyproline type-II helices, i.e. non-extensible structures that have structural and mechanical roles in the ECM of plants and some animals. Replacement of every third proline (Pro) residue with a glycine (Gly) residue results in increased flexibility. (B) Three left-handed superhelices wind together and pack tightly owing to these Gly residues, thereby forming a right-handed triple helix. Once associated, the non-extensibility of the structure is restored, and the combination of three chains results in three binding modes (one, two or three chains) with three levels of variation to these modes, i.e. level 1: 20 variable aa residues, level 2: post-translational modifications (PTMs) and, level 3: chain stagger. Together, these specify the numerous possible binding motifs that are positioned along the length of the triple helix of any of the 28 collagen types to bind various macromolecules.

**The roles of the ECM and collagen IV in the transition of the Urmetazoan to multicellular animals**

The last common ancestor to animals, the Urmetazoan, almost certainly reproduced by gametogenesis, underwent gastrulation during early development, had the ability for cells to differentiate both during development and as stem cells, and comprised an epithelial layer of cells forming the body of the animal – features that are still fundamental to extant animals (Richter and King, 2013; King and Rokas, 2017). Importantly, these cellular activities ultimately required the invention of an ECM to provide a substrate for attachment and signaling cues to regulate cell behavior and function in tissue genesis and homeostasis (Abedin and King, 2010). The appearance of a specialized form of ECM, the BM, coincided with the transition to multicellularity. The BM functions in several cellular activities, including migration, adhesion, delineation of apical–basal polarity and modulation of differentiation during development (Petersen et al., 1992; Lukashev and Werb, 1998; Daley and Yamada, 2013; Hynes, 2009, 2012; Yurchenco, 2011; Ozbek et al., 2010; Henry and Campbell, 1998). Importantly, understanding the makeup of BMs between different animals sheds light on the functions of proteins in the evolution of animal multicellularity and tissues. BMs are composed of numerous proteins, vary between animals. The BM of bilaterian animals (e.g. human, fly, C. elegans, sea urchin) is composed of several proteins, including collagen IV, laminin, perlecan, nidogen, fibronectin, proteoglycans, peroxidase, GPBP, and collagens XV and XVIII (Hynes, 2012; Yurchenco, 2011; Jayadev and Sherwood, 2017). Potentially, there are many more components of the BM (Chew and Lennon, 2018). Among these, collagen IV is a major...
component that is conserved among animal phyla (Fidler et al., 2014). In a recent study, we described collagen IV at the origins of animal multicellularity and in tissue evolution, as revealed by close examination of sponges, ctenophores and other non-bilaterally symmetrical animals (Fidler et al., 2017). Ctenophores and sponges have been established as the two most likely candidates to be the sister-groups to the rest of animals, based on phylogenetic analyses of genomic and transcriptomic data and cell-type evolution (Ryan et al., 2013; Moroz et al., 2014; Pisani et al., 2015; Whelan et al., 2015; Telford et al., 2016; King and Rokas, 2017; Feuda et al., 2017). Our genomic analysis of the extracellular matrix components within the ctenophores and Homoscleromorph sponges revealed a BM ‘toolkit’ consisting of just collagen IV and laminin (Fidler et al., 2017). However, the demosponges, a sponge class, lack both laminin and classic collagen IV but do contain spongins, which are short collagen IV variants (Exposito et al., 1991; Aouacheria et al., 2006; Fidler et al., 2017). The order in which these variants and collagen IV first appeared is still unknown. Despite containing fewer BM proteins than bilaterians, many sponges and ctenophores form classic BMs (Fidler et al., 2017; Boute et al., 1996; Leys et al., 2009). The closest relatives of animals, no complete ECM proteins exist; yet, domains that are characteristic of laminins and short collagenous Gly-X-Y repeats are present (King et al., 2008; Fahey and Degnan, 2012; Fidler et al., 2017). Interestingly, a recent study reported the discovery of a collagen IV-like gene in the filasterean, Ministeria vibrans, a unicellular lineage that diverged prior to choanoflagellates and animals (Grau-Bove et al., 2017). This finding indicates that collagen IV has a premetazoan ancestry and a function for single cells. Collectively, these findings suggest that collagen IV played a role in the transition from unicellular organisms to multicellular animals (Grau-Bove et al., 2017; Fidler et al., 2017) (Fig. 1). Therefore, we consider collagen IV as an archetype of collagens to describe the fundamental features of a triple helix that underlie biological functions.

The triple helix of collagen IV scaffolds
Collagen IV forms a network that functions as a scaffold within BMs (Fig. 3). These scaffolds provide tensile strength, connect adjacent cells and organize supramolecular protein assemblies that are able to influence cell behavior (Wang et al., 2008; Emsley et al., 2000; Parkin et al., 2011; Cummings et al., 2016; Vanacore et al., 2009). Network assembly begins with the intracellular formation of triple-helical protomers comprising three α-chains (Brown et al., 2017).

Fig. 3. The triple helix of collagen IV scaffolds encodes information for the tethering of macromolecules. (A) Assembly begins with collagen IV monomers (α-chains) transcribed and assembled into protomers inside of many cell types. Protomers are composed of three intertwining α-chains that form a triple helix. On the outside of cells, the NC1- and 7S-domains direct the assembly of protomers into network structures of higher order crosslinked by sulfilimine and lysyl oxidase-like protein 2 (LOXL2). The higher Cl− concentration of the ECM induces NC1-domain-directed oligomerization of protomers that form ‘smart’ scaffolds. (B) Once assembled, collagen IV networks function as smart scaffolds, bestowing BMs that underlie and surround cells of several capabilities. Vast amounts of structural information is encoded in motifs located at specific sites along the surface of the triple helix to tether macromolecules. The mode of variation is based on 20 variable aa residues on a single chain, or the combination of one, two or three chains, post-translational modifications (PTMs) and chain stagger (see Fig. 2). The tethers at specific sites spatially organizes molecules along the triple helix, resulting in a populated scaffold within the BM that provides tensile strength to tissues, and influences cell behavior, adhesion and migration during tissue development and regeneration. Partly modified, with permission (Protein Science) from Brown et al., 2017.
2017). Protoner assembly is directed and regulated by the non-collagenous (NC) I recognition modules, which are located at the C-terminus of each α-chain; this is followed by the twisting together of the collagenous domains – the Gly-X-Y repeats – into a triple helix. In mammals, three distinct protomers (α1I, α3I and α5I) are formed from six genetically distinct α-chains (α1–α6), thereby forming three distinct networks (Khoshnoodi et al., 2008). Once secreted into the extracellular space, protomers adjoin via their NC I domains and the N-terminal 7S domains (Fig. 3A). NC I-domain association is mediated by extracellular C1', which activates a molecular switch that enables adjacent protomers to adjoin two NC I domain trimers (Fig. 3A) (Cummings et al., 2016). 7S domains assemble into a complex of four independent protomer 7S domains (Añazco et al., 2016). Upon association mediated through NC I- and 7S-domains, the collagen IV networks are reinforced through covalent crosslinks at both the NC I- and 7S-domain interfaces. NC I-domain hexamers are stabilized through sulfilimine (-S=N-) double bonds, crosslinks that are induced by peroxidasin (PXDN) – an animal heme peroxidase embedded in the BM – and by Br− (Vanacore et al., 2009; Bhave et al., 2012; McCall et al., 2014). Concurrently, 7S dodecamers are crosslinked by lysyl oxidase-like protein 2 (LOXL2) (Añazco et al., 2016). Collagen IV networks function as smart scaffolds, bestowing BMs with several capabilities (Fig. 3B). Via the triple helix, scaffolds tether different extracellular molecules, i.e. laminins, proteoglycans, perlecans, nidogens, growth factors and extracellular scaffolds tether different extracellular molecules, i.e. laminins, proteoglycans, perlecans, nidogens, growth factors and extracellular enzymes (such as peroxidasin and lysyl oxidase-like protein 2) (Parkin et al., 2011; Hynes, 2012; Bhave et al., 2012; Añazco et al., 2016). Glycine mutations in triple helix

![Diseases and disorders caused by glycine mutations in triple helix](image)

**Fig. 4. Mutations in the triple helix can result in genetic diseases.** (A) A single missense mutation of a glycine to another residue results in disruption of the triple helix’s function. (B) Glycine mutations are responsible for a number of genetic diseases involving the kidney, teeth, muscle, joints, cartilage, brain, skin, vasculature, bone, inner ear and eye in humans (see Table 1). (B) Diseases and disorders due to replacement of glycine residue(s) in the triple helix of collagens. The type of collagen is indicated in bold Roman numerals. AVN, avascular necrosis of femoral head; CCDD, congenital cranial dysinnervation disorder; CSVVD, cerebral small-vessel disease; HANAC, hereditary angiopathy with nephropathy, aneurysms and muscle cramps; UMD, Ullrich congenital muscular dystrophy; LCPD, Legg–Calvé–Perthes disease; OSMED, otospondyloepiphyseal dysplasia; syn, syndrome.
Collagen properties of the collagen triple helix outside the cell (left) evokes an analogy to the DNA double helix inside the cell (right). (B) The triple helix protein

Table 1). Glycine residues are crucial for the structural integrity of the triple helix (see Figs 2 and 4A); therefore, mutated collagen molecules can assemble into faulty fibrils, networks, and other assemblies can cause tissue dysfunction.

As examples for such glycine mutations, osteogenesis imperfecta (OI; also known as brittle bone disease), and Alport syndrome are two collagen-dependent genetic disorders that are well-studied (Fig. 4C, Table 1). For OI, over 800 mutations in collagen I have been described (Marini et al., 2007; Forlino et al., 2011; Forlino and Marini, 2016). Approximately 80% are glycine mutations that occur in the triple helix (Forlino et al., 2011). However, such mutations, depending on the nature of substitution as well as its location, lead to different degrees of post-translational modifications and structural destabilization of the triple helix. For example, substitutions in the first 200 residues of collagen I are non-lethal, whereas there are two regions (helix positions 691–823 and 910–964), in which substitutions can cause lethality because they align with main ligand-binding sites for integrins, matrix metalloproteinases, fibronectin and cartilage oligomeric matrix protein (Marini et al., 2007). In Alport syndrome, the collagen IV scaffold is mutated, which leads to progressive organ failure in kidney, ear and eye (Williamson, 1961; Hudson, 2004; Hudson et al., 2003; Cosgrove and Liu, 2017). Over 1700 mutations have been found to occur in the collagen IV scaffold that is composed of α3, α4 and α5 chains. Of these, ~85% are glycine substitutions that are located in the triple helix (Hertz et al., 2012) (Table 1).

In summary, the many collagen diseases that involve glycine mutations directly demonstrate the essentiality of the triple helix for tissue architecture and function. Its essentiality is further supported experimentally by collagen knockout studies in mice (Table 1). The respective mouse phenotypes include developmental lethality upon knockout of collagen types I, II, III, IV, V, VII, XII, XVII, XIX, XXVI and XXVII, muscle deformities in that of types VI, XII, XIII, XV and XXV, and bone and cartilage deformities in that of types IX and XI (see references in Table 1).

Conclusions and perspectives

The triple helix is unique among all other protein structures – globular or fibrous – in its capacity to encode vast amounts of information that is available on its surface for utilization on the outside of cells (Fig. 5A). The triple helix, arranged in various patterns forming diverse supramolecular scaffolds, tethers and

Fig. 5. The fundamental importance of a triple helix in enabling animal multicellularity and tissue evolution. (A) The unique structure and vast encoding properties of the collagen triple helix outside the cell (left) evokes an analogy to the DNA double helix inside the cell (right). (B) The triple helix protein structure was present in uncellular organisms and was co-opted in the form of collagen IV, enabling the transition to multicellular animals. The triple helix was also adapted as a key feature of all members of the diverse collagen superfamily in the ECM, enabling tissue evolution. LCA, last common ancestor.
cell biology, disease pathogenesis and evolution of animals. Knowledge may provide insights into unknown roles of collagens in the transition from unicellular organisms to multicellular animals, and the triple helix was also adapted to give rise to all the other members of the diverse collagen superfamily, thereby enabling the evolution of tissues and organs (Fig. 5B). Thus, the triple helix represents a fundamental protein structure that nature adapted for building an extracellular matrix.

There are many provocative and unanswered questions regarding the function of triple helices in suprastructures and their dysfunction in diseases. These include: (i) What are the unknown sites encoded in the triple helix for binding partners as exemplified by those in collagen I fibrils and collagen IV networks (Di Lullo et al., 2002; Parkin et al., 2011)? (ii) How is information in the triple helix used to assemble suprastructures (Orgel et al., 2011)? (iii) What are the mechanisms of how triple helices within suprastructures influence cell function? (iv) What is the impact of PTMs on the structure and function of the triple helix? (v) How do genetic mutations in the triple helix cause tissue dysfunction? (vi) How do genetic backgrounds affect phenotype variations? (vii) What are the mechanisms for the function of the triple helix in the transition from unicellular organisms to multicellular animals.

To answer such fundamental questions, collagen IV is an ideal archetype because it is the most ancient of the collagens, and it is present in unicellular organisms and non-bilaterian animals (Fig. 5B). Furthermore, recent studies have revealed that, in some organisms, collagen IV also occurs in the absence of a BM, such as in certain ctenophores and sponges, placozoans and the unicellular flagellate *Ministeria vibrans* (Fidler et al., 2017; Grau-Bove et al., 2017; Schierwater et al., 2009). Moreover, a recent study of *Drosophila* development provided evidence, that in the absence of BM, collagen IV has a role in intercellular adhesion and pro-growth signaling (Dai et al., 2017; Zajac and Horne-Badovinač, 2017). Together, these recent studies clearly indicate that collagen IV can have a direct role in influencing cell behavior, outside of the BM; thus, the core functions of the collagen triple helix, as well as its role in the transition from unicellular organisms to multicellular animals, can be addressed by comparative studies in these organisms. Such knowledge may provide insights into unknown roles of collagens in cell biology, disease pathogenesis and evolution of animals.

**Acknowledgements**

This article is the culmination of a ten-year Aspirnaut expedition to the dawn of the animal kingdom in search of the evolutionary origin of collagen IV. The expedition involved over 100 middle school, high school, undergraduate and graduate students in disadvantaged backgrounds, and was championed by Aaron Fidler, Julie Hudson and Billy Hudson. We are grateful for the contributions of all of these students in the discovery that collagen IV occurred in all animals, and the fundamental importance of the triple helical protein structure in animal evolution as presented in this capstone article.

**Competing interests**

The authors declare no competing or financial interests.

**Funding**

This work was supported by National Institutes of Health DK18381 to B.G.H., National Science Foundation DEB-1442113 to Antonis Rokas, March of Dimes Foundation March of Dimes Prematurity Research Center Ohio Collaborative to Antonis Rokas, and the Aspirnaut Program to Julie K. Hudson and B.G.H. Deposited in PMC for release after 12 months.

**Supplementary information**

Supplementary information available online at http://jcs.biologists.org/lookup/doi/10.1242/jcs.203950.supplemental

**References**


mutations in COL1A2, the gene encoding the alpha2(III) chain of type XI collagen. J. Pediatr. 132, 368-371.


