Widespread distribution of collagens and collagen-associated domains in eukaryotes

Tess A. Linden¹ and Nicole King^{1,2}

1. Howard Hughes Medical Institute and the Department of Molecular and Cell Biology,
University of California, Berkeley, CA, USA
2. Corresponding author. Email: nking@berkeley.edu

Abstract

The origin of collagen, the dominant structural component of metazoan extracellular matrix, has long been cited as a critical step in the evolution of metazoan multicellularity. While collagens were once thought to be found only in metazoans, scattered reports of collagen domains in Fungi, and more recently in close relatives of metazoans, have called into question whether collagens are truly unique to metazoans. Here, we take advantage of recently sequenced genomes and transcriptomes of diverse holozoans (the clade encompassing metazoans and their close relatives), as well as publicly available proteomes from diverse non-holozoan eukaryotes, to conduct a systematic search for collagen domains across eukaryotic diversity. We find that collagen domains are ubiquitous in choanoflagellates, the sister group of metazoans, and widespread across many other major eukaryotic taxa. Many predicted collagens in non-metazoans are comparable to metazoan collagens in length and proline content. Moreover, most are present in species that also encode putative prolyl 4-hydroxylase domains, suggesting that, like metazoan collagens, they may be stabilized through the hydroxylation of prolines. Fibrillar collagen and collagen IV appear to be unique to metazoans, and we posit that their ability to assemble into superstructures may have contributed to the origin of metazoan multicellularity.

Introduction

In metazoans, the extracellular matrix (ECM) is indispensable to development, form, and function. Collagen is the major structural protein of metazoan ECM (Frantz et al., 2010), and is the most abundant protein in metazoan bodies by mass, making up one third of the dry weight of an adult mammal (Shoulders and Raines, 2009) and over 70% of the weight of some sponges (Ehrlich et al., 2018). The diagnostic feature of collagens is the repeated triplet motif Gly-X-Y, in which X and Y can be any amino acid (Box 1). The repetition of the collagen repeat motif Gly-X-Y leads to the formation of a stable, rod-like triple helix, which in turn supplies tensile strength to metazoan tissues (Buehler, 2006; Shoulders and Raines, 2009; Wenger et al., 2007).

For most of the twentieth century, collagen was thought to be unique to metazoans and absent from other lineages with complex multicellularity (e.g., plants and fungi). Thus, because of its apparent phylogenetic restriction and its importance to metazoan ECM, collagen was repeatedly hypothesized to be one of the key innovations that allowed metazoans and their characteristic tissues to evolve (Erwin, 1993; Towe, 1970). The hypothesized connection between metazoan origins and the evolution of collagen was brought into question with the sequencing of genomes from three close relatives of metazoans—two choanoflagellates and one filasterean—each of which encode collagen domains (Fairclough et al., 2013; Grau-Bové et al., 2017; Hynes, 2012; King et al., 2008). These joined reports of putative collagens in *Dikarya* Fungi (Celerin et al., 1996; de Bruin et al., 2002; Wang and Leger, 2006), the malaria parasite *Plasmodium falciparum* (Rasmussen et al., 2003), and the amoebozoan *Dictyostelium discoideum* (Fidler et al., 2017) to create a growing catalog of putative collagens in non-metazoan eukaryotes. In addition, collagen domains have been detected in several species of bacteria (Bachert et al., 2015; Charalambous et al., 1988; Ghosh et al., 2012; Kananavičiūtė et al., 2020; Lukomski et al., 2017; Mohs et al., 2007; Rasmussen et al., 2003; Yu et al., 2014) and viruses (Bamford and Bamford, 1990; La Scola et al., 2008; Luther et al., 2011; Medveczky et al., 1993; Raoult et al., 2004; Tidona and Darai, 1997; van Hulten et al., 2001; Yau et al., 2011; Zhang et al., 2004), although due to their evolutionary distance, these were typically hypothesized to have evolved independently from metazoan collagens (Rasmussen et al., 2003) or through horizontal gene transfer from metazoans (Luther et al., 2011; Rasmussen et al., 2003).

Nonetheless, the growing number of reports has pointed to the possibility that collagens might be more ancient and/or more widespread than previously realized.

Therefore, to better understand the phylogenetic distribution and ancestry of collagen domains across the tree of eukaryotes, we analyzed recently sequenced genomes and transcriptomes of close relatives of metazoans, the choanoflagellates and other non-metazoan holozoans (Table S1), as well as those of diverse other eukaryotes representing many major clades. We found that collagen domains are common in the close relatives of metazoans, including choanoflagellates, many of which encode extended collagen domains comparable in length and proline content to metazoan collagens. Furthermore, we found that collagen domains are widely distributed across many eukaryotes that are only distantly related to metazoans, including Fungi, Archaeplastida, and members of the SAR clade. In contrast, canonical metazoan collagens such as fibrillar collagens and collagen IV are apparently restricted to metazoans and may help explain unique features of metazoan biology.

Box 1. Collagen terminology.

The criteria for what constitutes “a collagen” are not well defined (Garrone, 1999; Gay and Miller, 1983; Ricard-Blum, 2011). The major defining characteristic of a collagen is the collagen triple helix, but some authors restrict the term “a collagen” to mean a protein that plays a structural role in the ECM (Garrone, 1999; Gay and Miller, 1983), while others define “a collagen” as any protein containing a collagen triple helix, including intracellular proteins of the immune system (Casals et al., 2019; Fraser and Tenner, 2008) or transmembrane signaling proteins (Maertens et al., 2007). Here, we adopt the latter usage, for two reasons: first, many of the proteins we discuss have not been functionally characterized, and whether a given protein plays a structural role in the ECM is difficult to predict from sequence alone. Second, this usage acknowledges the possibility that collagens could evolve dynamically between structural and non-structural roles or between intracellular and extracellular localizations over evolutionary time. For brevity and simplicity, we use the term “collagen domain” to refer to any stretch of collagen triple helix repeats, i.e., (Gly-X-Y)_n.

Results

Collagen domains and conserved collagen-associated domains predate metazoan origins

To start, we focused our search on close relatives of metazoans, hypothesizing that they were most likely to illuminate the proximal ancestry of metazoan collagens (Fig. 1). From the 29 genomes and transcriptomes analyzed (Table S1), we found that all 22 choanoflagellates, three of four filastereans (*Ministeria vibrans*, *Pigoraptor vietnamica*, and *Pigoraptor chileana*) and one of two ichthyosporeans (*Sphaeroforma arctica*) encoded proteins with collagen domains (domains annotated by InterProScan as IPR008160). Like metazoan collagens, many of these proteins were predicted to carry transmembrane regions and/or N-terminal signal sequences (Fig. S1), suggesting that they are secreted into the ECM. Furthermore, all 29 species were found to encode putative prolyl 4-hydroxylase domains (IPR006620), the class of enzyme responsible for the post-translational modification of proline to 4-hydroxyproline, a step thought to be required for collagen triple helix formation in metazoans (Berg and Prockop, 1973; Canty and Kadler, 2005; Juva et al., 1966; Kao et al., 1979; Walmsley et al., 1999). (The strictness of this requirement has been challenged by observations of triple helix formation without hydroxyproline in some contexts, such as hymenopteran cocoons (Sutherland et al., 2013), recombinant non-metazoan cells (Olsen et al., 2001; Perret et al., 2001; Ruggiero et al., 2000), and bacteria (Yu et al., 2014).) Thus, 26 of the 29 non-metazoan holozoans that we analyzed appear to have the genetic machinery required to produce collagen triple helices that may be modified with 4-hydroxyproline.

At least two classes of collagens may have been present in the Urmetazoan (the last common ancestor of all metazoans) based on their broad phylogenetic distribution within metazoans. The first is fibrillar collagen, which is characterized by a long, uninterrupted collagen domain and a C-terminal COLF1 domain (Exposito et al., 2010). The second is collagen IV, which is characterized by a triple helix with many interruptions and a C-terminal NC1 domain (Boute et al., 1997; Exposito et al., 2002; Fidler et al., 2018, 2017). Included in this second class is the collagen IV variant spongin, which is distinguished by a truncated collagen domain and the lack of a conserved HSQ motif in its NC1 domain (Aouacheria et al., 2006; Fidler et al., 2017). To determine whether these key metazoan collagens may have predated the Urmetazoan, we searched our data set for proteins with collagen domains and COLF1 or NC1 domains. Prior

investigations found that *S. rosetta* and *M. brevicollis* encode COLF1 domains homologous to the C-terminal domains of metazoan fibrillar collagens, but that these domains are not present in the same proteins as collagen domains (Hynes, 2012; King et al., 2008). We found that this pattern holds true across choanoflagellate diversity: COLF1 was detected in 12 of 22 choanoflagellate proteomes (Fig. 1), but none of the choanoflagellate COLF1 domains were present in proteins with collagen domains. COLF1 was not detected in any other holozoans. Thus, the COLF1 domain likely evolved on the Choanozoan stem lineage, but canonical fibrillar collagens appear to have evolved, presumably through domain shuffling, after the split between the choanoflagellate and metazoan lineages.

Next, we analyzed whether any non-metazoan holozoans encoded NC1 domains. Consistent with previous findings in *S. rosetta* and *M. brevicollis* (Fidler et al., 2017), we did not detect NC1 domains in any of the 22 choanoflagellate proteomes. Outside choanozoans, we detected an NC1 domain only in one protein encoded by the filasterean *Ministeria vibrans* (Fig. 1; Fig. S2C), as previously reported by Grau-Bové et al. (2017). Like metazoan collagen IV, this *M. vibrans* protein (hereafter “MvCN”, for *Ministeria vibrans* Collagen + NC1) contains collagen domains and a C-terminal NC1 domain (Fig. S2C). This raises the question of whether MvCN might be homologous to collagen IV, and thus that collagen IV might be an ancient protein that predates the last common ancestor of filastereans and metazoans. To investigate this possibility more closely, we examined each of the domains of MvCN individually. By aligning the MvCN NC1 domain with metazoan collagen IVs, we found that it contains the conserved HSQ motif found in the NC1 domains of metazoan collagen IVs but not in those of spongins (Fig. S3B) (Fidler et al., 2017). Next, we examined the MvCN collagenous region. In contrast with metazoan collagen IV, the MvCN collagenous region is much shorter: it contains a total of 81 Gly-X-Y repeats, as opposed to the hundreds of Gly-X-Y repeats typical of metazoan collagen IV: e.g., 443 repeats in *Mus musculus* (UniProt ID P02463), 264 repeats in the mollusk *Lottia gigantea* (UniProt ID V4A1B9), or 473 repeats in the cnidarian *Nematostella vectensis* (UniProt ID V9GW22). Finally, we examined its non-collagenous domain. Unlike metazoan collagen IV, MvCN contains a long, N-terminal non-collagenous domain of unknown function. Through BLAST search, we found that this domain shares sequence similarity with uncharacterized bacterial proteins (Fig. S3A), indicating that MvCN may be a eukaryotic-bacterial fusion protein. Thus, the NC1 domain appears to predate the divergence between

filastereans and choanozoans, but whether MvCN shares a common ancestry with metazoan collagen IV/spongins remains unclear. It is possible, for example, that an NC1 domain was present in a protein without collagen domains in a stem holozoan and became fused to collagen domains independently in the filasterean and metazoan lineages.

Choanoflagellates encode collagens with diverse non-collagenous domains

Because choanoflagellates are the closest living relatives of metazoans and provide a unique window into the ancestry of metazoan genes, we examined the protein domain architecture of their collagens in more detail. The non-collagenous protein domain found most frequently in choanoflagellate collagens was the von Willebrand Factor A (VWFA) domain, a domain commonly found in metazoan ECM proteins such as collagen VI and FACIT collagens (Fig. S2D) (Ricard-Blum, 2011). VWFA domains often mediate protein-protein interactions, facilitating the binding of ECM proteins to each other (Hynes and Naba, 2012).

Interestingly, two choanoflagellate species encode collagens with domains that are also present in metazoan collagenous immune proteins. The choanoflagellate *Salpingoeca kjevrii* encodes a collagen with an SRCR domain, a combination found in some metazoan scavenger receptors that play a role in phagocytosis and innate immunity (Fig. S2D) (Gowen et al., 2001; Kodama et al., 1996; Neubauer et al., 2016; Yamada et al., 1998). *Microstomoeca roanoka* encodes a collagenous protein with a gamma fibrinogen domain, making it similar to ficolins, a group of collagens that function in innate immune system activation in chordates (Endo et al., 2015; Matsushita, 2010). Because both species nest deeply within choanoflagellates, we posit that these choanoflagellate proteins evolved independently from metazoan scavenger receptors and ficolins. Nonetheless, the shared combination of protein domains raises the possibility that these collagens might play analogous roles in bacterial recognition, phagocytosis, or immunity in choanoflagellates in a manner similar to the metazoan defense collagens.

Because the choanoflagellate clade contains as much genetic diversity as its sister group, the Metazoa (Richter et al., 2018), we were interested in whether our data set could reveal patterns of collagen evolution within the choanoflagellate lineage. The diversity of collagen repertoires across choanoflagellates and the high degree of variability in collagen number and

domain structure between closely related species suggested that collagens are rapidly evolving within choanoflagellates. For example, *Choanoeca flexa* was found to encode six collagens, whereas its closely related sister species *C. perplexa* encodes 16 collagens, and their next-closest relative, *M. brevicollis*, encodes only two (Fig. S1); many of these proteins have unique domain structures without clear orthologs in the other species. By contrast, some collagens appear to be conserved across choanoflagellate evolution. Mirroring the conserved collagens of metazoans (collagen IV and fibrillar collagens), we identified a possibly ancient collagen that is distributed widely within choanoflagellates. This protein, which we refer to as “ChCL” (for Choanoflagellate Collagenous Lectin) has a characteristic architecture of a ~20-amino acid N-terminal transmembrane domain followed by an extracellular region with a ~50-amino acid collagen domain, a ~700-amino acid non-collagenous domain, and a ~200-amino acid C-terminal Concanavalin A-like lectin/glucanase domain (IPR013320) (Fig. S4A). The non-collagenous domain of ChCL has sequence similarity with non-collagenous metazoan proteins of unknown function (Fig. S4B), while the C-terminal domain closely resembles genes of bacterial origin (such as MSX41286.1, an Actinobacterium protein of unknown function, or WP_007139689.1, a putative cell wall-associated protein in a Flavobacterium) (Fig. S4C). Based on its distribution and on the most recent understanding of choanoflagellate phylogenetic relationships, ChCL appears to have been present in the last common ancestor of all choanoflagellates (Fig. 1).

Collagen domains are widespread across eukaryotes

The finding that collagenous proteins are nearly ubiquitous in the closest living relatives of metazoans inspired us to broaden our search to include more distant lineages. To investigate the distribution of collagen domains across eukaryotic diversity, we took advantage of the proteomes available in the UniProtKB database (UniProt Consortium, 2021), which includes representatives of 288 families of non-metazoan eukaryotes across many, but not all, of the major eukaryotic groups (Fig. 2A). We detected proteins with collagen domains in nearly all major eukaryotic groups analyzed (Fig. 2B). For example, out of 169 fungal families in our analysis, 73 families encode at least one protein with a collagen domain (IPR008160) (Fig. 2C). Families encoding collagens were concentrated within Ascomycota, but we also detected

collagens sparsely distributed in Basidiomycota, Zoopagomycota, Mucoromycota, Cryptomycota, and Microsporidia (Fig. S5B). Within Archaeplastida, 17 out of 71 families encode collagens (Fig. 2D). Nine of these 17 families are within land plants (Embryophyta), revealing that collagens are present in at least two lineages with complex multicellularity. In the SAR clade, 19 out of 31 families were found to encode collagens (Fig. 2E). The only major eukaryotic group in which we detected no collagen domains was Haptista (Fig. 2F), which is represented by only two proteomes in our data set.

The low information-density of the collagen domain (Box 1) makes it difficult to infer whether the collagen domains found in diverse eukaryotic clades are homologous with metazoan collagens or whether they evolved convergently. However, we can ask if non-metazoan collagen domains resemble metazoan collagens in length and proline content, and whether the collagen-encoding species also encode a putative prolyl 4-hydroxylase. We found that most non-metazoan eukaryotic proteomes in our data set encode at least one protein with a putative prolyl 4-hydroxylase domain (Fig. S6); specifically, 84% of the 2,011 proteomes that encode collagens also encode a putative prolyl 4-hydroxylase. Whether an enzyme is capable of hydroxylating proline in a collagen-like substrate is difficult to predict from sequence alone, but such enzymatic activity outside metazoans would not be unprecedented: endogenous enzymes with collagen prolyl 4-hydroxylase activity have been reported in plants (Hieta and Myllyharju, 2002) and Fungi (de Bruin et al., 2002). Next, we compared the collagen repeat length across all collagens detected in metazoans ($n = 57,104$ collagens), choanoflagellates ($n = 258$ collagens), Archaeplastida ($n = 75$ collagens), Fungi ($n = 508$ collagens), and SAR ($n = 176$ collagens) (Fig. 3A; Fig. S7A). Metazoan collagens are the longest, with a median collagen repeat length of 192 amino acids (i.e., a median of 64 Gly-X-Y repeats per protein); choanoflagellate collagens have the second-longest, with a median collagen repeat length of 145.5 amino acids; and SAR collagens have the shortest, with a median of 85.5 amino acids. We then quantified proline content across the same data set (Fig. 3B; Fig. S7B) and found that metazoan collagen domains contain the most proline at the X and Y amino acid positions, with a median of 29.9% proline. Choanoflagellate collagens were the second-most proline-rich at 15.6%, while SAR collagens were the least proline-rich at a median of 0%. While the median metazoan collagen may have longer collagen domains and more proline than the median non-metazoan collagen, it is notable that the distributions of collagen repeat length and proline content overlap significantly across all

these eukaryotic groups (Fig. 3). Together, these data suggest that some non-metazoans likely express collagen triple helices with similar structural qualities to metazoan collagens, which could therefore play similar roles in their biology.

Discussion

Extracellular proline-rich glycoproteins have long been known to exist in non-metazoan eukaryotes (Davies et al., 1997; Hallmann, 2006; Lamport, 1977; Perfect et al., 1998; Showalter, 1993; Sumper and Hallmann, 1998). However, collagen domains have rarely been reported in distant eukaryotic relatives of Metazoa. The furthest examples, to our knowledge, have been reports of collagen repeats in the genomes of the apicomplexan (SAR) parasite *Plasmodium falciparum* (Rasmussen et al., 2003) and the amoebozoan *Dictyostelium discoideum* (Fidler et al., 2017), and of cell-surface proteins with collagen repeats in a handful of species of Dikarya fungi (Celerin et al., 1996; de Bruin et al., 2002; Wang and Leger, 2006). Our findings suggest that collagen domains may be a near-ubiquitous feature of holozoans (Fig. 1, Fig. S5C) and may also be widespread across eukaryotic diversity (Fig. 2). The collagen domain is characterized by a repeated triplet motif (Box 1) whose simplicity makes it a reasonable candidate for repeated convergent evolution (Chen et al., 1997; Stern, 2013). With the exception of metazoans (Fig. S5A), choanoflagellates (Fig. S5C), and perhaps ascomycete Fungi (Fig. S5B), the distribution of collagen domains across all of the eukaryotic groups in our data set is notably scattered: many pairs of sister families differ from each other in whether they encode collagen domains (Fig. 2C-F). Such a pattern could indicate either rampant convergent evolution or rampant loss. Due to the inherent difficulty of aligning highly repetitive domains, it remains unclear how many times collagen domains may have evolved in the history of eukaryotes. Even if the collagen domains outside Metazoa evolved independently from those within Metazoa or Holozoa, it is nonetheless possible that proteins with collagen-like properties might be important to the biology of many organisms across eukaryotic diversity, not just metazoans.

The evolution of collagen has often been cited as a key prerequisite for metazoan complex multicellularity (Erwin, 1993; Fidler et al., 2018; Reinhard et al., 2016; Towe, 1970). Here, we found that while collagen domains are encoded by diverse eukaryotes, several features distinguish metazoan collagen repertoires from those of non-metazoans. Metazoan collagen

domains tend to be slightly longer and more proline-rich than those of any other eukaryotic clade in our data set (Fig. 3). Furthermore, several classes of collagens are restricted to metazoans in our data set: fibrillar collagen, a major component of metazoan connective tissues; and collagen IV *sensu stricto* (i.e. excluding MvCN, which has many fewer triple helix repeats than metazoan collagen IV), the major component of metazoan basement membranes. The C-terminal domains of these collagens allow them to form highly oligomeric, cross-linked superstructures—fibrils and networks, respectively—that are key to their roles as the structural scaffold of metazoan ECM (Ricard-Blum, 2011). In addition to proline hydroxylation, metazoan collagens also undergo diverse, functionally important post-translational modifications, such as lysine hydroxylation and cross-linking (Eyre and Wu, 2005; Fidler et al., 2014; Rodriguez-Pascual and Slatter, 2016), proteolytic cleavage (Canty and Kadler, 2005; Ricard-Blum, 2011), and glycosylation (Hennet, 2019), and it remains to be seen whether non-metazoan collagens might undergo similar modifications. Thus, while collagen domains are not unique to Metazoa, it may be the case that innovations related to collagen, such as the origin of fibrillar collagens/collagen IV and their associated superstructures, increased modification with 4-hydroxyproline, and/or innovations in collagen-interacting and collagen-modifying proteins not addressed by our data set, facilitated the evolution of complex multicellularity during metazoan origins.

The ubiquity of collagen domains in choanoflagellates raises the question of what role collagens might be playing in single-celled and colonial relatives of metazoans. In a possible parallel to fibrillar collagen and collagen IV in metazoans, we detected an ancient collagenous lectin, ChCL, that is widespread across choanoflagellates and was likely present in the last common ancestor of all choanoflagellates. Choanoflagellates secrete extracellular matrices that are dynamically regulated across multiple life history stages and play functional roles in substrate adhesion (Dayel and King, 2014; Leadbeater, 2008), mating regulation (Woznica et al., 2017), and multicellular colony formation (Dayel et al., 2011; Larson et al., 2020; Levin et al., 2014). Furthermore, choanoflagellates interact closely with bacteria as predators (Dayel and King, 2014), as signaling partners (Alegado et al., 2012; Ireland et al., 2020; Woznica et al., 2017, 2016), and possibly as symbionts (Hake et al., 2021). We found that some choanoflagellates encode collagens with domains characteristic of metazoan defense collagens, such as scavenger receptors and ficolins, and that the choanoflagellate collagen ChCL contains a domain similar to putative bacterial cell wall proteins, suggestive of a potential role for these

collagens in mediating choanoflagellate-bacteria interactions. In general, while non-ECM collagens—i.e., the defense collagens (Casals et al., 2019; Fraser and Tenner, 2008), transmembrane collagens (Franzke et al., 2005; Loria et al., 2004; Maertens et al., 2007), and soluble collagenous signaling proteins (Leclère et al., 2020; Mikkola and Thesleff, 2003; Pyagay et al., 2005)—are relatively well characterized in vertebrates and in some invertebrate model organisms, little is known about non-ECM collagens in basal metazoans. This leaves a gap in our understanding of any ancient non-ECM collagens that may have been present in early metazoan evolution. Future work to characterize the function and localization of collagens in basal metazoans and close metazoan relatives, including choanoflagellates, may shed light on the early roles of collagens both within and outside of the ECM.

Methods

Identification of collagen domains and collagen-associated domains in non-metazoan holozoans

To catalogue putative prolyl 4-hydroxylase and collagen domains in holozoans, we obtained the predicted proteomes of 22 choanoflagellates, four filastereans, two ichthyosporeans, and one corallochytrid (Table S1). We performed InterProScan annotation of each proteome using the AgBase InterProScan Singularity container (running InterProScan version 5.41-78.0; available at <https://hub.docker.com/r/agbase/interproscan>) on the University of California Berkeley Savio computational cluster to obtain Gene3D, Panther, Pfam, Phobius, PrositePatterns, SMART, and SuperFamily annotations of every protein. The Phobius data were used for transmembrane region predictions and signal sequence predictions. Any protein region annotated as IPR008160 (“Collagen triple helix repeat”; <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR008160/>) was considered a collagen domain, and any region annotated as IPR006620 (“Prolyl 4-hydroxylase, alpha subunit”; <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR006620/>) was considered a putative prolyl 4-hydroxylase domain. Domains annotated as IPR001442 (“Collagen IV, non-collagenous”; <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR001442/>) or IPR036954 (“Collagen IV, non-collagenous domain superfamily”; <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR036954/>) were considered NC1 domains, domains annotated as IPR000885 (“Fibrillar collagen, C-terminal”; <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR000885/>) were considered COLF1 domains.

Identification of collagen domains and prolyl 4-hydroxylases in metazoans and non-holozoan eukaryotes

To characterize the distribution of putative prolyl 4-hydroxylase domains and collagen domains across non-holozoan eukaryotes and metazoans, we took advantage of the publicly available proteomes in the UniProtKB database. We first assembled a list of all eukaryotic proteomes available in the UniProtKB database as of June 30th, 2021 (UniProt release 2021_03; <https://www.uniprot.org/>) by searching the UniProtKB Proteomes database (<https://www.uniprot.org/proteomes/>) with the query

taxonomy:"Eukaryota [2759]"

and downloading the results. We filtered out low-quality proteomes using the following criteria: first, we discarded any proteomes with a protein count less than 200; second, we discarded proteomes classified by UniProt's Complete Proteome Detector (CPD) as "Outlier (low value)"; and third, we discarded proteomes with BUSCO scores lower than 50% (or, for proteomes with no BUSCO score provided, we discarded those with a protein count less than 20,000). These criteria were chosen in order to minimize inclusion of low-quality proteomes while retaining at least one proteome from each major eukaryotic group.

After filtering the proteomes, the "Taxonomic lineage" provided by UniProt for each proteome was then used as the basis for the topologies of all phylogenetic trees presented in this paper, including assignment of each organism to a taxonomic family. At this step, the following modifications were made: (1) We rearranged the base of the eukaryotic phylogeny to match the most up-to-date eukaryotic tree of life based on Burki et al. (2020), and the base of the metazoan phylogeny to reflect known relationships between phyla. (2) We discarded proteomes from organisms labeled "fungal sp. no. X" (where X is a species number) due to unclear phylogenetic placement. (3) For 40 species whose UniProt taxonomic lineage did not include a family name, a family name and taxonomic lineage were assigned from the Global Biodiversity Information Facility (GBIF) database (<https://www.gbif.org/>); in the 11 out of 40 cases where the GBIF database did not list a family name due to phylogenetic uncertainty, the organism and its proteome was discarded.

Having assembled our eukaryotic phylogeny and list of proteomes for analysis, we next searched for collagen domains in metazoans and non-holozoan eukaryotes by searching the UniProtKB protein database (<https://www.uniprot.org/uniprot/>) with the following queries:

1. NOT taxonomy:"Opisthokonta [33154]" IPR008160 taxonomy:"Eukaryota [2759]"
2. taxonomy:"Fungi [4751]" IPR008160
3. taxonomy:"Metazoa [33208]" IPR008160

We downloaded all hits for these three queries, then discarded any hits whose UniProt Organism ID was not present in our list of filtered proteomes. To catalogue putative prolyl 4-hydroxylase domains in metazoans and non-holozoan eukaryotes, we repeated the above procedure replacing "IPR008160" with "IPR006620".

Quantification of collagen repeat length and proline content

To compare the characteristics of collagens across eukaryotic diversity, we analyzed the collagen repeat length and proline content of collagens in different eukaryotic clades. Collagen repeat length was assessed by quantifying, for each protein, the total number of triplet repeats matching the pattern (Gly-X-Y)_n, where $n \geq 2$ and X and Y can be any amino acid. The resulting value was either expressed in units of repeats or multiplied by 3 when expressed in units of amino acids. Proline content was assessed by dividing the total number of prolines found in the X and Y positions of those triplet repeats by twice the total number of repeats. For example, a protein with the sequence MGPPGNNGPNNNNPGNPNNNNNGPN* has a “collagen repeat length” of 5 repeats or 15 amino acids, and its “proportion of proline in X and Y positions” is 50%.

Analysis of MvCN domains.

We investigated the non-collagenous domains of *Ministeria vibrans* MvCN through a combination of BLAST search and amino acid alignment. First, we conducted a search of the NCBI BLAST database (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) using the MvCN N-terminal non-collagenous domain (amino acids 1-731) as a query. This resulted in three hits, NBV83445.1, NTW51976.1, and WP_161718411.1, all of which are bacterial proteins. Alignment of the MvCN N-terminal domain against these hits (Fig. S3) was performed using the Geneious alignment algorithm in Geneious Prime 2021.1.1 (<https://www.geneious.com>). We aligned the MvCN NC1 domain (amino acids 938-1122) against the NC1 domains of collagen IV from the ctenophore *Mnemiopsis leidyi* (protein ML18198a-PA) and the sponge *Oscarella carmela* (protein Ocar_m.306941) using the same alignment procedure (Fig. S3).

Analysis of ChCL domains.

To investigate the origins of the conserved choanoflagellate collagen ChCL, we first performed a search of the NCBI BLAST database using the entire amino acid sequence of the representative *Salpingoeca urceolata* ChCL protein (Surc_m.480486) as a query. This yielded as hits many metazoan proteins that aligned to the middle non-collagenous domain of the protein (amino acids ~200-900). We aligned *S. urceolata* ChCL to the top hit, XP_013383323.1, as described above (Fig. S4). We then performed a second search with only the C-terminal domain of *S. urceolata* ChCL (amino acids 920-1136) as a query, which resulted in predominantly bacterial hits, and we

aligned the *S. urceolata* ChCL C-terminal domain to the top hit, MSX41286.1, as described above (Fig. S4).

Acknowledgements

We thank O. Muellerklein of Berkeley Research Computing and F. McCarthy and S. Saha of the University of Arizona for advice on InterProScan implementation. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). T.A.L. was supported by an NSF GRFP Fellowship and the Berkeley Fellowship for Graduate Study.

Figure 1. Collagen domains and collagen-associated domains are conserved in close

relatives of Metazoa. We detected collagen domains (IPR008160) and putative prolyl 4-hydroxylase domains (IPR006620) in the predicted proteomes of all metazoans and choanoflagellates and in most other holozoans analyzed, and found that the distributions of other collagen-associated domains such as COLF1 (IPR000885) and NC1 (IPR001442/IPR036954) are also consistent with pre-metazoan origins. Holozoans include choanoflagellates, filastereans, corallochytreans (“Co.”) and ichthyosporeans (“Ic.”). Collagen length in choanoflagellates rivals that of metazoans, as shown here in the “Max. collagen repeat length” column, in which the maximum collagen repeat length of any protein detected in the species’ predicted proteome is indicated by degree of shading (see key and Methods). **(A – E)** Mapping the distribution of collagen domains and key collagen-associated domains helps resolve the pre-metazoan evolution of collagens. **(A)** Putative prolyl 4-hydroxylase domains, which in metazoans perform a post-translational modification step important for collagen triple helix formation, were detected in all 29 non-metazoan holozoans. This domain is also present in many non-holozoan eukaryotes (Fig. S6), and thus might predate the divergence between Fungi and Holozoa. **(B)** The NC1 domain characteristic of metazoan collagen IV was detected in a collagen domain-containing protein (MvCN) of the filasterean *Ministeria vibrans* as previously reported (Grau-Bové et al., 2017), but not in any other holozoans, including choanoflagellates. Thus, the NC1 domain was likely present in the common ancestor of filastereans and choanozoans. **(C)** COLF1 domains, which are diagnostic of metazoan fibrillar collagens, are broadly distributed across choanoflagellates and metazoans, indicating that they predate the Urchoanozoan. However, in choanoflagellates, the COLF1 domain was not detected in any proteins that also contained collagen domains. **(D)** ChCL, a choanoflagellate-specific transmembrane protein consisting of a collagen domain and a lectin/glucanase domain, is broadly distributed across the choanoflagellate phylogeny, suggesting that it was likely present in the last common ancestor of choanoflagellates. **(E)** Fibrillar collagens (those containing COLF1 and collagen domains) appear to be restricted to metazoans. Likewise, canonical collagen IVs (proteins with >100 Gly-X-Y repeats and a C-terminal NC1 domain with a conserved HSQ motif) and spongins (truncated collagen IV variants lacking the conserved HSQ motif) are also restricted to metazoans. *: The MvCN protein was previously described as a collagen IV (Grau-Bové et al., 2017), but given the differences between MvCN and metazoan collagen IV, it is currently unclear whether this filasterean protein

is homologous to collagen IV (see main text). Figure format is adapted with permission from Figure 1 of Brunet et al. (2017). Consensus phylogeny of choanoflagellate relationships is based on Richter et al. (2018).

Figure 2. Collagen domains are detected throughout eukaryotic diversity. Having found that collagens are widespread in holozoans, we broadened our search to investigate whether collagen domains are found in other eukaryotic taxa. Our analysis included 2,598 eukaryotic proteomes present in the UniProtKB database that met quality criteria based on BUSCO score and protein count (see Methods). **(A)** The current consensus eukaryotic tree. The UniProtKB database contains proteomes from many branches of the eukaryotic tree of life, but not all. Clades shown in black are represented by at least one proteome in UniProtKB, while clades in gray are unrepresented. Dashed lines represent uncertainty about the monophyly of the indicated groups. **(B)** For each major eukaryotic group, the number of analyzed families in which a collagen domain was detected (black bar) or not detected (white bar) is shown (y-axis). For the Amorphea supergroup, which contains the Holozoa, we only show data for non-holozoans. Collagen domains were detected in at least one family from all eukaryotic groups analyzed, with the exception of Haptista. Even when omitting Holozoa, proteomes from Amorphea (particularly Fungi; see panel C) are heavily overrepresented in the UniProtKB database. Meanwhile, several major eukaryotic groups are represented by <10 families each (inset). **(C to F)** Collagen domains are broadly distributed not just across, but within, all major eukaryotic groups. In these trees, each stem represents one taxonomic family for which at least one proteome meeting our quality criteria is present in the UniProtKB database. The presence of a circle indicates that at least one species in that family encodes a collagen domain (IPR008160). The size of the circle represents the maximum collagen repeat length of any collagen in the family, while the shading of the circle represents the maximum proline content of any collagen in the family (see Methods). Note that the maximum collagen repeat length and maximum proline content may be from different proteins. **(C)** Distribution of collagen domains across Amorphea families. Holozoans were excluded in order to highlight that collagen domains are broadly distributed even in distant relatives of metazoans. **(D)** Distribution of collagen domains across Archaeplastida families. The chlorophyte *Coccomyxa subellipsoidea* encodes the longest collagen domain detected in our entire data set, with a collagen repeat length of >4000 amino acids. **(E)** Distribution of collagen domains across SAR families. Collagen domains were detected in Stramenopiles and Alveolates, but not Rhizarians. **(F)** Distribution of collagen domains across the families of the other major groups of eukaryotes: Haptista, Cryptista, Metamonada, and Discoba. Haptista, which is represented by only two families in our analysis, is the only eukaryotic group for which no collagen domains were detected. Panel A was adapted with permission from Figure 1 of Burki et al. (2020).

Figure 3. Metazoans encode collagens with more collagen repeats and higher proline content than other eukaryotes. To compare the characteristics of collagen domains across eukaryotic taxa, we examined collagen repeat length and proline content (see Methods) across all collagens detected in all analyzed species of choanoflagellates ($n = 258$ collagens), metazoans ($n = 57,104$ collagens), Archaeplastida ($n = 75$ collagens); fungi ($n = 508$ collagens); and SAR ($n = 176$ collagens). **(A)** Metazoans have the highest total collagen repeat length per protein (median = 192 amino acids), choanoflagellates the second-highest (median = 145.5 amino acids), and SAR the lowest (median = 85.5 amino acids). Note that the y-axis is on a log-scale. **(B)** Metazoan collagens have the highest proline content (median = 29.9%), choanoflagellates have the second-highest (median = 15.6%), and SAR the lowest (median = 0%). Thus, metazoan collagens are characterized by a greater number of repeats and higher proline content compared to those of other eukaryotic taxa; however, it is notable that the distributions of these characteristics across all five taxa are highly overlapping. Red bars represent medians. *: to more clearly visualize the distribution of the metazoan collagen data, a randomly selected 1% of metazoan collagens are shown here. For comparison of collagen repeat length and proline content across all eukaryotic groups analyzed, see Figure S7.

References

- Alegado, R.A., Brown, L.W., Cao, S., Dermenjian, R.K., Zuzow, R., Fairclough, S.R., Clardy, J., King, N., 2012. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *eLife* 2012. <https://doi.org/10.7554/eLife.00013>
- Aouacheria, A., Geourjon, C., Aghajari, N., Navratil, V., Deléage, G., Lethias, C., Exposito, J.-Y., 2006. Insights into early extracellular matrix evolution: spongin short chain collagen-related proteins are homologous to basement membrane type IV collagens and form a novel family widely distributed in invertebrates. *Mol. Biol. Evol.* 23, 2288–2302. <https://doi.org/10.1093/molbev/msl100>
- Bachert, B.A., Choi, S.J., Snyder, A.K., Rio, R.V.M., Durney, B.C., Holland, L.A., Amemiya, K., Welkos, S.L., Bozue, J.A., Cote, C.K., Berisio, R., Lukomski, S., 2015. A Unique Set of the Burkholderia Collagen-Like Proteins Provides Insight into Pathogenesis, Genome Evolution and Niche Adaptation, and Infection Detection. *PLOS ONE* 10, e0137578. <https://doi.org/10.1371/journal.pone.0137578>
- Bamford, D.H., Bamford, J.K.H., 1990. Collagenous proteins multiply. *Nature* 344, 497–497. <https://doi.org/10.1038/344497b0>
- Berg, R.A., Prockop, D.J., 1973. The thermal transition of a non-hydroxylated form of collagen. Evidence for a role for hydroxyproline in stabilizing the triple-helix of collagen. *Biochem. Biophys. Res. Commun.* 52, 115–120. [https://doi.org/10.1016/0006-291x\(73\)90961-3](https://doi.org/10.1016/0006-291x(73)90961-3)
- Boute, N., Exposito, J.Y., Boury-Esnault, N., Vacelet, J., Noro, N., Miyazaki, K., Yoshizato, K., Garrone, R., 1997. Type IV collagen in sponges, the missing link in basement membrane ubiquity. *Biol. Cell* 88, 37–44. [https://doi.org/10.1016/S0248-4900\(97\)86829-3](https://doi.org/10.1016/S0248-4900(97)86829-3)
- Brunet, T., King, N., 2017. The Origin of Animal Multicellularity and Cell Differentiation. *Dev. Cell* 43, 124–140. <https://doi.org/10.1016/j.devcel.2017.09.016>
- Buehler, M.J., 2006. Nature designs tough collagen: explaining the nanostructure of collagen fibrils. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12285–12290. <https://doi.org/10.1073/pnas.0603216103>
- Burki, F., Roger, A.J., Brown, M.W., Simpson, A.G.B., 2020. The New Tree of Eukaryotes. *Trends Ecol. Evol.* 35, 43–55. <https://doi.org/10.1016/j.tree.2019.08.008>
- Canty, E.G., Kadler, K.E., 2005. Procollagen trafficking, processing and fibrillogenesis. *J. Cell Sci.* 118, 1341–1353. <https://doi.org/10.1242/jcs.01731>
- Casals, C., García-Fojeda, B., Minutti, C.M., 2019. Soluble defense collagens: Sweeping up immune threats. *Mol. Immunol.* 112, 291–304.
- Celerin, M., Ray, J.M., Schisler, N.J., Day, A.W., Stetler-Stevenson, W.G., Laudénbach, D.E., 1996. Fungal fimbriae are composed of collagen. *EMBO J.* 15, 4445–4453.
- Charalambous, B.M., Keen, J.N., McPherson, M.J., 1988. Collagen-like sequences stabilize homotrimers of a bacterial hydrolase. *EMBO J.* 7, 2903–2909. <https://doi.org/10.1002/j.1460-2075.1988.tb03148.x>
- Chen, L., DeVries, A.L., Cheng, C.-H.C., 1997. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl. Acad. Sci.* 94, 3817–3822. <https://doi.org/10.1073/pnas.94.8.3817>
- Davies, H.A., Findlay, K., Daniels, M.J., Dow, J.M., 1997. A novel proline-rich glycoprotein associated with the extracellular matrix of vascular bundles of Brassica petioles. *Planta* 202, 28–35.

- Dayel, M.J., Alegado, R.A., Fairclough, S.R., Levin, T.C., Nichols, S.A., McDonald, K., King, N., 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate *Salpingoeca rosetta*. *Dev. Biol.* 357, 73–82. <https://doi.org/10.1016/j.ydbio.2011.06.003>
- Dayel, M.J., King, N., 2014. Prey capture and phagocytosis in the choanoflagellate *Salpingoeca rosetta*. *PLoS ONE* 9. <https://doi.org/10.1371/journal.pone.0095577>
- de Bruin, E.C., Werten, M.W.T., Laane, C., de Wolf, F.A., 2002. Endogenous prolyl 4-hydroxylation in *Hansenula polymorpha* and its use for the production of hydroxylated recombinant gelatin. *FEMS Yeast Res.* 1, 291–298. <https://doi.org/10.1111/j.1567-1364.2002.tb00047.x>
- Ehrlich, H., Wysokowski, M., \Zóltowska-Aksamitowska, S., Petrenko, I., Jesionowski, T., 2018. Collagens of Poriferan Origin, *Marine Drugs*. <https://doi.org/10.3390/md16030079>
- Endo, Y., Matsushita, M., Fujita, T., 2015. Chapter Two - New Insights into the Role of Ficolins in the Lectin Pathway of Innate Immunity, in: Jeon, K.W. (Ed.), *International Review of Cell and Molecular Biology*. Academic Press, pp. 49–110. <https://doi.org/10.1016/bs.ircmb.2015.01.003>
- Erwin, D.H., 1993. The origin of metazoan development: a palaeobiological perspective. *Biol. J. Linn. Soc.* 50, 255–274. <https://doi.org/10.1111/j.1095-8312.1993.tb00931.x>
- Exposito, J.-Y., Cluzel, C., Garrone, R., Lethias, C., 2002. Evolution of collagens. *Anat. Rec.* 268, 302–316. <https://doi.org/10.1002/ar.10162>
- Exposito, J.Y., Valcourt, U., Cluzel, C., Lethias, C., 2010. The fibrillar collagen family, *International Journal of Molecular Sciences*. <https://doi.org/10.3390/ijms11020407>
- Eyre, D.R., Wu, J.-J., 2005. Collagen Cross-Links, in: Brinckmann, J., Notbohm, H., Müller, P.K. (Eds.), *Collagen: Primer in Structure, Processing and Assembly, Topics in Current Chemistry*. Springer, Berlin, Heidelberg, pp. 207–229. <https://doi.org/10.1007/b103828>
- Fairclough, S.R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H.M., Begovic, E., Richter, D.J., Russ, C., Westbrook, M.J., Manning, G., Lang, B.F., Haas, B., Nusbaum, C., King, N., 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol.* 14, 1–15. <https://doi.org/10.1186/gb-2013-14-2-r15>
- Fidler, A.L., Boudko, S.P., Rokas, A., Hudson, B.G., 2018. The triple helix of collagens—an ancient protein structure that enabled animal multicellularity and tissue evolution. *J. Cell Sci.* 131.
- Fidler, A.L., Darris, C.E., Chetyrkin, S.V., Pedchenko, V.K., Boudko, S.P., Brown, K.L., Gray Jerome, W., Hudson, J.K., Rokas, A., Hudson, B.G., 2017. Collagen iv and basement membrane at the evolutionary dawn of metazoan tissues. *eLife* 6. <https://doi.org/10.7554/eLife.24176.001>
- Fidler, A.L., Vanacore, R.M., Chetyrkin, S.V., Pedchenko, V.K., Bhav, G., Yin, V.P., Stothers, C.L., Rose, K.L., McDonald, W.H., Clark, T.A., Borza, D.-B., Steele, R.E., Ivy, M.T., Aspirnauts, T., Hudson, J.K., Hudson, B.G., 2014. A unique covalent bond in basement membrane is a primordial innovation for tissue evolution. *Proc. Natl. Acad. Sci.* 111, 331–336. <https://doi.org/10.1073/pnas.1318499111>
- Frantz, C., Stewart, K.M., Weaver, V.M., 2010. The extracellular matrix at a glance. *J. Cell Sci.* 123, 4195–4200. <https://doi.org/10.1242/jcs.023820>
- Franzke, C.-W., Bruckner, P., Bruckner-Tuderman, L., 2005. Collagenous transmembrane proteins: recent insights into biology and pathology. *J. Biol. Chem.* 280, 4005–4008. <https://doi.org/10.1074/jbc.R400034200>

- Fraser, D.A., Tenner, A.J., 2008. Directing an appropriate immune response: the role of defense collagens and other soluble pattern recognition molecules. *Curr. Drug Targets* 9, 113–122. <https://doi.org/10.2174/138945008783502476>
- Garrone, R., 1999. Collagen, a common thread in extracellular matrix evolution. *Proc. Indian Acad. Sci.-Chem. Sci.* 111, 51–56.
- Gay, S., Miller, E.J., 1983. What is collagen, what is not. *Ultrastruct. Pathol.* 4, 365–377.
- Ghosh, N., McKillop, T.J., Jowitt, T.A., Howard, M., Davies, H., Holmes, D.F., Roberts, I.S., Bella, J., 2012. Collagen-Like Proteins in Pathogenic *E. coli* Strains. *PLOS ONE* 7, e37872. <https://doi.org/10.1371/journal.pone.0037872>
- Gowen, B.B., Borg, T.K., Ghaffar, A., Mayer, E.P., 2001. The collagenous domain of class A scavenger receptors is involved in macrophage adhesion to collagens. *J. Leukoc. Biol.* 69, 575–582. <https://doi.org/10.1189/jlb.69.4.575>
- Grau-Bové, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T.A., Ruiz-Trillo, I., 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* 6, e26036. <https://doi.org/10.7554/eLife.26036>
- Hake, K.H., West, P.T., McDonald, K., Laundon, D., Feng, C., Burkhardt, P., Richter, D., Banfield, J.F., King, N., 2021. Colonial choanoflagellate isolated from Mono Lake harbors a microbiome. <https://doi.org/10.1101/2021.03.30.437421>
- Hallmann, A., 2006. The pherophorins: Common, versatile building blocks in the evolution of extracellular matrix architecture in Volvocales. *Plant J.* 45, 292–307. <https://doi.org/10.1111/j.1365-313X.2005.02627.x>
- Hennet, T., 2019. Collagen glycosylation. *Curr. Opin. Struct. Biol.* 56, 131–138. <https://doi.org/10.1016/j.sbi.2019.01.015>
- Hieta, R., Myllyharju, J., 2002. Cloning and characterization of a low molecular weight prolyl 4-hydroxylase from *Arabidopsis thaliana*. Effective hydroxylation of proline-rich, collagen-like, and hypoxia-inducible transcription factor alpha-like peptides. *J. Biol. Chem.* 277, 23965–23971. <https://doi.org/10.1074/jbc.M201865200>
- Hynes, R.O., 2012. The evolution of metazoan extracellular matrix. *J. Cell Biol.* 196, 671–679. <https://doi.org/10.1083/jcb.201109041>
- Hynes, R.O., Naba, A., 2012. Overview of the matrisome-An inventory of extracellular matrix constituents and functions. *Cold Spring Harb. Perspect. Biol.* 4. <https://doi.org/10.1101/cshperspect.a004903>
- Ireland, E.V., Woznica, A., King, N., 2020. Synergistic Cues from Diverse Bacteria Enhance Multicellular Development in a Choanoflagellate. *Appl. Environ. Microbiol.* 86, e02920-19. <https://doi.org/10.1128/AEM.02920-19>
- Juva, K., Prockop, D.J., Cooper, G.W., Lash, J.W., 1966. Hydroxylation of Proline and the Intracellular Accumulation of a Polypeptide Precursor of Collagen. *Science* 152, 92–94.
- Kananavičiūtė, R., Kvederavičiūtė, K., Dabkevičienė, D., Mackevičius, G., Kuisienė, N., 2020. Collagen-like sequences encoded by extremophilic and extremotolerant bacteria. *Genomics* 112, 2271–2281. <https://doi.org/10.1016/j.ygeno.2019.12.023>
- Kao, W.W., Prockop, D.J., Berg, R.A., 1979. Kinetics for the secretion of nonhelical procollagen by freshly isolated tendon cells. *J. Biol. Chem.* 254, 2234–2243.
- King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K.J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J.B., Morris, A., Nichols, S., Richter, D.J., Salamov, A., Bork, P., Lim, W.A., Manning, G.,

- Miller, W.T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I.V., Rokhsar, D., 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451, 783–788. <https://doi.org/10.1038/nature06617>
- Kodama, T., Doi, T., Suzuki, H., Takahashi, K., Wada, Y., Gordon, S., 1996. Collagenous macrophage scavenger receptors. *Curr. Opin. Lipidol.* 7, 287–291. <https://doi.org/10.1097/00041433-199610000-00005>
- La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., Raoult, D., 2008. The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100–104. <https://doi.org/10.1038/nature07218>
- Lampert, D.T.A., 1977. Structure, Biosynthesis and Significance of Cell Wall Glycoproteins BT - The Structure, Biosynthesis, and Degradation of Wood, in: Loewus, F.A., Runeckles, V.C. (Eds.), . Springer US, Boston, MA, pp. 79–115. https://doi.org/10.1007/978-1-4615-8873-3_3
- Larson, B.T., Ruiz-Herrero, T., Lee, S., Kumar, S., Mahadevan, L., King, N., 2020. Biophysical principles of choanoflagellate self-organization. *Proc. Natl. Acad. Sci.* 117, 1303–1311. <https://doi.org/10.1073/pnas.1909447117>
- Leadbeater, B.S.C., 2008. Choanoflagellate evolution: the morphological perspective. *Protistology* 5, 256–267.
- Leclère, L., Nir, T.S., Bazarsky, M., Braitbard, M., Schneidman-Duhovny, D., Gat, U., 2020. Dynamic Evolution of the *Cthrc1* Genes, a Newly Defined Collagen-Like Family. *Genome Biol. Evol.* 12, 3957–3970. <https://doi.org/10.1093/gbe/evaa020>
- Levin, T.C., Greaney, A.J., Wetzel, L., King, N., 2014. The Rosetteless gene controls development in the choanoflagellate *S. rosetta*. *eLife* 3. <https://doi.org/10.7554/eLife.04070>
- Loria, P.M., Hodgkin, J., Hobert, O., 2004. A Conserved Postsynaptic Transmembrane Protein Affecting Neuromuscular Signaling in *Caenorhabditis elegans*. *J. Neurosci.* 24, 2191–2201. <https://doi.org/10.1523/JNEUROSCI.5462-03.2004>
- Lukowski, S., Bachert, B.A., Squeglia, F., Berisio, R., 2017. Collagen-like proteins of pathogenic streptococci. *Mol. Microbiol.* 103, 919–930. <https://doi.org/10.1111/mmi.13604>
- Luther, K.B., Hülsmeier, A.J., Schegg, B., Deuber, S.A., Raoult, D., Henet, T., 2011. Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme. *J. Biol. Chem.* 286, 43701–43709. <https://doi.org/10.1074/jbc.M111.309096>
- Maertens, B., Hopkins, D., Franzke, C.-W., Keene, D.R., Bruckner-Tuderman, L., Greenspan, D.S., Koch, M., 2007. Cleavage and Oligomerization of Gliomedin, a Transmembrane Collagen Required for Node of Ranvier Formation*. *J. Biol. Chem.* 282, 10647–10659. <https://doi.org/10.1074/jbc.M611339200>
- Matsushita, M., 2010. Ficolins: Complement-Activating Lectins Involved in Innate Immunity. *J. Innate Immun.* 2, 24–32. <https://doi.org/10.1159/000228160>
- Medveczky, M.M., Geck, P., Vassallo, R., Medveczky, P.G., 1993. Expression of the collagen-like putative oncoprotein of Herpesvirus saimiri in transformed T cells. *Virus Genes* 7, 349–365. <https://doi.org/10.1007/BF01703391>
- Mikkola, M.L., Thesleff, I., 2003. Ectodysplasin signaling in development. *Cytokine Growth Factor Rev.* 14, 211–224. [https://doi.org/10.1016/S1359-6101\(03\)00020-0](https://doi.org/10.1016/S1359-6101(03)00020-0)

- Mohs, A., Silva, T., Yoshida, T., Amin, R., Lukomski, S., Inouye, M., Brodsky, B., 2007. Mechanism of Stabilization of a Bacterial Collagen Triple Helix in the Absence of Hydroxyproline *. *J. Biol. Chem.* 282, 29757–29765. <https://doi.org/10.1074/jbc.M703991200>
- Neubauer, E.F., Poole, A.Z., Weis, V.M., Davy, S.K., 2016. The scavenger receptor repertoire in six cnidarian species and its putative role in cnidarian-dinoflagellate symbiosis. *PeerJ* 4, e2692. <https://doi.org/10.7717/peerj.2692>
- Olsen, D.R., Leigh, S.D., Chang, R., McMullin, H., Ong, W., Tai, E., Chisholm, G., Birk, D.E., Berg, R.A., Hitzeman, R.A., Toman, P.D., 2001. Production of Human Type I Collagen in Yeast Reveals Unexpected New Insights into the Molecular Assembly of Collagen Trimers *. *J. Biol. Chem.* 276, 24038–24043. <https://doi.org/10.1074/jbc.M101613200>
- Perfect, S.E., O’Connell, R.J., Green, E.F., Doering-Saad, C., Green, J.R., 1998. Expression cloning of a fungal proline-rich glycoprotein specific to the biotrophic interface formed in the *Colletotrichum*–bean interaction. *Plant J.* 15, 273–279. <https://doi.org/10.1046/j.1365-313X.1998.00196.x>
- Perret, S., Merle, C., Bernocco, S., Berland, P., Garrone, R., Hulmes, D.J.S., Theisen, M., Ruggiero, F., 2001. Unhydroxylated Triple Helical Collagen I Produced in Transgenic Plants Provides New Clues on the Role of Hydroxyproline in Collagen Folding and Fibril Formation *. *J. Biol. Chem.* 276, 43693–43698. <https://doi.org/10.1074/jbc.M105507200>
- Pyagay, P., Heroult, M., Wang, Q., Lehnert, W., Belden, J., Liaw, L., Friesel, R.E., Lindner, V., 2005. Collagen triple helix repeat containing 1, a novel secreted protein in injured and diseased arteries, inhibits collagen expression and promotes cell migration. *Circ. Res.* 96, 261–268. <https://doi.org/10.1161/01.RES.0000154262.07264.12>
- Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.-M., 2004. The 1.2-Megabase Genome Sequence of Mimivirus. *Science* 306, 1344–1350. <https://doi.org/10.1126/science.1101485>
- Rasmussen, M., Jacobsson, M., Björck, L., 2003. Genome-based Identification and Analysis of Collagen-related Structural Motifs in Bacterial and Viral Proteins *. *J. Biol. Chem.* 278, 32313–32316. <https://doi.org/10.1074/jbc.M304709200>
- Reinhard, C.T., Planavsky, N.J., Olson, S.L., Lyons, T.W., Erwin, D.H., 2016. Earth’s oxygen cycle and the evolution of animal life. *Proc. Natl. Acad. Sci. U. S. A.* 113, 8933–8938. <https://doi.org/10.1073/pnas.1521544113>
- Ricard-Blum, S., 2011. The Collagen Family. *Cold Spring Harb. Perspect. Biol.* 3, 1–19. <https://doi.org/10.1101/cshperspect.a004978>
- Richter, D.J., Fozouni, P., Eisen, M.B., King, N., 2018. Gene family innovation, conservation and loss on the animal stem lineage. *eLife* 7, e34226. <https://doi.org/10.7554/eLife.34226>
- Rodriguez-Pascual, F., Slatter, D.A., 2016. Collagen cross-linking: insights on the evolution of metazoan extracellular matrix. *Sci. Rep.* 6, 37374. <https://doi.org/10.1038/srep37374>
- Ruggiero, F., Exposito, J.-Y., Bournat, P., Gruber, V., Perret, S., Comte, J., Olgarnier, B., Garrone, R., Theisen, M., 2000. Triple helix assembly and processing of human collagen produced in transgenic tobacco plants. *FEBS Lett.* 469, 132–136. [https://doi.org/10.1016/S0014-5793\(00\)01259-X](https://doi.org/10.1016/S0014-5793(00)01259-X)
- Shoulders, M.D., Raines, R.T., 2009. Collagen Structure and Stability. *Annu. Rev. Biochem.* 78, 929–958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833>
- Showalter, A.M., 1993. Structure and function of plant cell wall proteins. *Plant Cell* 5, 9.

- Stern, D.L., 2013. The genetic causes of convergent evolution. *Nat. Rev. Genet.* 14, 751–764.
<https://doi.org/10.1038/nrg3483>
- Sumper, M., Hallmann, A., 1998. Biochemistry of the Extracellular Matrix of Volvox. *Int. Rev. Cytol.* 180, 51–85. [https://doi.org/10.1016/S0074-7696\(08\)61770-2](https://doi.org/10.1016/S0074-7696(08)61770-2)
- Sutherland, T.D., Peng, Y.Y., Trueman, H.E., Weisman, S., Okada, S., Walker, A.A., Sriskantha, A., White, J.F., Huson, M.G., Werkmeister, J.A., Glattauer, V., Stoichevska, V., Mudie, S.T., Haritos, V.S., Ramshaw, J.A.M., 2013. A new class of animal collagen masquerading as an insect silk. *Sci. Rep.* 3, 2864. <https://doi.org/10.1038/srep02864>
- Tidona, C.A., Darai, G., 1997. The complete DNA sequence of lymphocystis disease virus. *Virology* 230, 207–216. <https://doi.org/10.1006/viro.1997.8456>
- Towe, K.M., 1970. Oxygen-Collagen Priority and the Early Metazoan Fossil Record. *Proc. Natl. Acad. Sci.* 65, 781–788. <https://doi.org/10.1073/pnas.65.4.781>
- UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- van Hulten, M.C.W., Witteveldt, J., Peters, S., Kloosterboer, N., Tarchini, R., Fiers, M., Sandbrink, H., Lankhorst, R.K., Vlak, J.M., 2001. The White Spot Syndrome Virus DNA Genome Sequence. *Virology* 286, 7–22. <https://doi.org/10.1006/viro.2001.1002>
- Walmsley, A.R., Batten, M.R., Lad, U., Bulleid, N.J., 1999. Intracellular retention of procollagen within the endoplasmic reticulum is mediated by prolyl 4-hydroxylase. *J. Biol. Chem.* 274, 14884–14892. <https://doi.org/10.1074/jbc.274.21.14884>
- Wang, C., Leger, R.J.S., 2006. A collagenous protective coat enables *Metarhizium anisopliae* to evade insect immune responses. *Proc. Natl. Acad. Sci.* 103, 6647–6652.
<https://doi.org/10.1073/pnas.0601951103>
- Wenger, M.P.E., Bozec, L., Horton, M.A., Mesquida, P., 2007. Mechanical properties of collagen fibrils. *Biophys. J.* 93, 1255–1263. <https://doi.org/10.1529/biophysj.106.103192>
- Woznica, A., Cantley, A.M., Beemelmans, C., Freinkman, E., Clardy, J., King, N., 2016. Bacterial lipids activate, synergize, and inhibit a developmental switch in choanoflagellates. *Proc. Natl. Acad. Sci. U. S. A.* 113, 7894–7899.
<https://doi.org/10.1073/pnas.1605015113>
- Woznica, A., Gerdt, J.P., Hulett, R.E., Clardy, J., King, N., 2017. Mating in the Closest Living Relatives of Animals Is Induced by a Bacterial Chondroitinase. *Cell* 170, 1175–1183.e11.
<https://doi.org/10.1016/j.cell.2017.08.005>
- Yamada, Y., Doi, T., Hamakubo, T., Kodama, T., 1998. Scavenger receptor family proteins: roles for atherosclerosis, host defence and disorders of the central nervous system. *Cell. Mol. Life Sci. CMLS* 54, 628–640. <https://doi.org/10.1007/s000180050191>
- Yau, S., Lauro, F.M., DeMaere, M.Z., Brown, M.V., Thomas, T., Raftery, M.J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J.M., Gibson, J.A., Cavicchioli, R., 2011. Virophage control of antarctic algal host–virus dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 108, 6163–6168. <https://doi.org/10.1073/pnas.1018221108>
- Yu, Z., An, B., Ramshaw, J.A.M., Brodsky, B., 2014. Bacterial collagen-like proteins that form triple-helical structures. *J. Struct. Biol.* 186, 451–461.
<https://doi.org/10.1016/j.jsb.2014.01.003>
- Zhang, Q.-Y., Xiao, F., Xie, J., Li, Z.-Q., Gui, J.-F., 2004. Complete Genome Sequence of Lymphocystis Disease Virus Isolated from China. *J. Virol.* 78, 6982–6994.
<https://doi.org/10.1128/JVI.78.13.6982-6994.2004>