The *in-silico* characterization of the *Caenorhabditis elegans* matrisome and proposal of a novel collagen classification

Alina C. Teuscher[a], Elisabeth Jongsma[a], Martin N. Davis[b], Cyril Statzer[a], Jan M. Gebauer[c], Alexandra Naba[b,d] and Collin Y. Ewald[a]

***a* - Eidgenössische Technische Hochschule (ETH) Zürich,** *Department of Health Sciences and Technology, Institute of Translational Medicine, Schwerzenbach, Zürich, Switzerland*
***b* - Department of Physiology and Biophysics,** *University of Illinois at Chicago, Chicago, IL 60612, USA*
***c* - Institute of Biochemistry,** *University of Cologne, Cologne, Germany*
***d* - University of Illinois Cancer Center,** *Chicago, IL 60612, USA*

Correspondence to Jan M. Gebauer, Alexandra Naba and Collin Y. Ewald: *Co-corresponding authors. jan.gebauer@uni-koeln.de, anaba@uic.edu, collin-ewald@ethz.ch.*
https://doi.org/10.1016/j.mbplus.2018.11.001

Abstract

Proteins are the building blocks of life. While proteins and their localization within cells and sub-cellular compartments are well defined, the proteins predicted to be secreted to form the extracellular matrix - or matrisome - remain elusive in the model organism *C. elegans*. Here, we used a bioinformatic approach combining gene orthology and protein structure analysis and an extensive curation of the literature to define the *C. elegans* matrisome. Similar to the human genome, we found that 719 out of ~20,000 genes (~4%) of the *C. elegans* genome encodes matrisome proteins, including 181 collagens, 35 glycoproteins, 10 proteoglycans, and 493 matrisome-associated proteins. We report that 173 out of the 181 collagen genes are unique to nematodes and are predicted to encode cuticular collagens, which we are proposing to group into five clusters. To facilitate the use of our lists and classification by the scientific community, we developed an automated annotation tool to identify ECM components in large datasets. We also established a novel database of all *C. elegans* collagens (CeColDB). Last, we provide examples of how the newly defined *C. elegans* matrisome can be used for annotations and gene ontology analyses of transcriptomic, proteomic, and RNAi screening data. Because *C. elegans* is a widely used model organism for high throughput genetic and drug screens, and to study biological and pathological processes, the conserved matrisome genes may aid in identifying potential drug targets. In addition, the nematode-specific matrisome may be exploited for targeting parasitic infection of man and crops.

Introduction

Around one third of the human world population, including a majority of children, is infected by parasitic nematodes [1,2]. In addition, plant-parasitic nematodes are one of the most infectious species in agriculture with an impact on economic loss of about 100 billion dollars per year [3]. The major barriers for drugs to penetrate parasitic nematodes are its collagenous cuticle, an exoskeleton, and an extracellular matrix (ECM). The free-living nematode *C. elegans* has been widely used as a surrogate model organism for parasitic nematodes [4], as well as for host-pathogen interactions [5], and other fundamental biological processes [6]. *C. elegans* is also used as a pioneering in-vivo model for biomedical research because about 40% of *C. elegans* genes are conserved in the human genome [7], and vice versa between 60 and 80% of human genes have a corresponding orthologue in the *C. elegans* genome [8]. In addition, 40% of human genes associated with diseases are well conserved in *C. elegans* [9]. *C.*

elegans is genetically tractable for high throughput screens and is one of the best curated organisms for genetic, genomic, and phenotypic data. The vast array of openly shared molecular tools paved the way to gain molecular, functional, and mechanistic insights into gene and protein functions [8]. In particular, the two major extracellular matrices of *C. elegans*, the cuticle [10] and basement membrane [11–13], have recently become models to study cancer cell invasion [14] and aging [15]. However, a precise Gene Ontology term or a comprehensive compendium of genes predicted to form the *C. elegans* matrisome remains to be defined.

Using characteristic features of ECM proteins and a computational pipeline combining interrogation of protein and gene databases, we previously defined the matrisome as the ensemble of ECM and ECM-associated proteins [16–18]. In mammals, the matrisome represents 4% of the genome, or approximately 1000 genes. We further classified these genes into core matrisome components, consisting of collagens, proteoglycans, and glycoproteins (including laminins, fibronectins, etc.), and matrisome-associated components, including proteins that could incorporate into ECMs or are co-purified with ECM proteins. These components are further subdivided into ECM-affiliated proteins (*e.g.*, C-type lectins, galectins, annexins, semaphorins, syndecans, and glypicans), ECM regulators (*e.g.*, MMPs, ADAMs, and crosslinking enzymes), and secreted factors (*e.g.*, TGF-β, BMPs, FGFs, Wnt proteins, and chemokines) [16–18]. More recently, we employed a computational approach to predict the *in-silico* matrisome of the zebrafish [19]. Defining the matrisome of organisms has been instrumental to annotate transcriptomic and proteomic data and has permitted the identification of ECM signatures of biological processes [20] and of human diseases including cancers and fibrosis [21–25].

Here, we devised a novel bioinformatic pipeline combining gene orthology and *de-novo* identification to define the *C. elegans* matrisome. We report the identification of 719 genes potentially encoding ECM and ECM-associated proteins, including 181 collagens of which 173 are predicted to be components of the cuticle. Based on their collagen-domain organization, we propose to group these cuticular collagens into five novel clusters and further divide them in sub-clusters. In addition, we demonstrate that the newly defined *C. elegans* matrisome can be used to annotate data from high throughput RNAi screens, transcriptomic, and proteomic data, and can assist with the identification of ECM genes or signatures relevant in the context of various physiological and pathological processes.

Computational approach to define the *C. elegans* matrisome

The workflow and steps for defining the *C. elegans* matrisome are outlined in Fig. 1.

Identification of *C. elegans* orthologues of human matrisome genes

The orthologue list was created by comparing the human matrisome gene list downloaded from the Matrisome Project website (http://matrisome.org/)[26] with the *C. elegans* genome using the Greenwald Lab

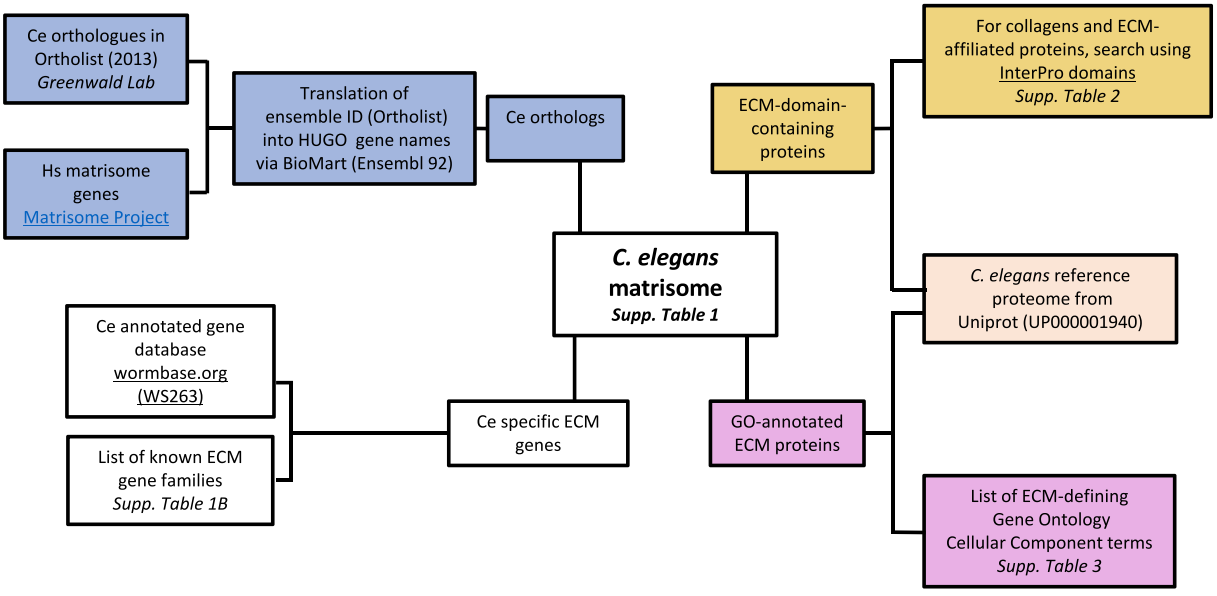

Fig. 1. Workflow of the pipeline devised to define the *in-silico C. elegans* matrisome.

Table 1.
Comparison of the number of human to *C. elegans* matrisome genes. Corresponding genes for each category are found in Supplementary Table 1.

		Human	*C. elegans*
	Complete matrisome	1027	719
Core matrisome	ECM glycoproteins	195	35
	Collagens	44	181
	Proteoglycans	35	10
	Cuticlins	0	12
	Total	274	238
Matrisome-associated	ECM-affiliated proteins	171	301
	ECM regulators	238	128
	Secreted factors	344	52
	Total	753	481

OrthoList website (http://greenwaldlab.org/ortholist/; accessed 07.04.2017, [7]). The OrthoList uses four different orthology-prediction programs (Ensembl Compara, In Paranoid, Homologene, and OrthoMCL) to obtain the *C. elegans* orthologues from human Ensembl ID numbers. We included all genes that were found by at least one prediction program from the OrthoList. The human Ensemble IDs were then translated back into HUGO gene names using Ensembl BioMart [27]. This approach allows the identification of 348 *C. elegans* genes orthologous to human matrisome genes (Supplementary Table 1).

Domain-based definition and Gene Ontology annotations of matrisome proteins

We initially defined the mammalian matrisome by using the presence of characteristic protein domains commonly found in ECM proteins [17,18]. To verify that the orthology approach identified key components of the *C. elegans* matrisome, we focused on a specific category of matrisome proteins: the ECM-affiliated proteins, which are proteins that share structural and functional homologies with ECM components [17,18]. To do so, we retrieved the *C. elegans* reference proteome (UP000001940 downloaded August 14, 2017; Supplementary Table 2A) from the UniProt database [28] and identified proteins containing domains previously defined as characteristic of the 6 families of ECM-affiliated components (Fig. 1 and Supplementary Table 2B): the transmembrane proteoglycans syndecans and glypicans, the galectins, the plexins and semaphorins, and the annexins. Comparison of the list of proteins obtained using this approach (Supplementary Table 2C–H) and the list of genes identified by orthology revealed that all but 2 ECM-affiliated proteins were found by both approaches (Supplementary Table 1), suggesting that both approaches may be used to define the complete matrisome. However, the orthology-based approach does not permit to identify nematode-specific ECM proteins. One ECM structure specific to *C. elegans* is the cuticle. It is made of cuticular collagens and cuticlins [10,29,30]. To define the ensemble of proteins potentially contributing to cuticle ECM, we identified an InterPro domain termed "Nematode cuticle collagen, N-terminal" (InterPro domain IPR002486; [31]), this domain retrieved 171 UniProt entries in the *C. elegans* proteome, out of which 128 also contain the canonical collagen triple helix repeat (IPR008160) (Supplementary Table S2I and J). Close examination of the list of collagen genes obtained revealed that some might not have been found using the InterPro domain. We thus further sought to define *C. elegans* collagens using a more rigorous structure-based approach.

De-novo identification of *C. elegans* collagens

To identify in an unbiased manner and *de-novo* all the collagen proteins in *C. elegans*, we downloaded all reviewed and unreviewed protein entries from the UniProt database (release 2018_01; [28]). Using HMMER3 [32] and the standard HMM profile (PF01391) for collagens from the Pfam website [33] identified 219 sequences, which included several duplicated entries. However, this approach missed

Table 2.
Comparison of conserved versus nematode-specific matrisome genes. Corresponding genes for each category are found in Supplementary Table 1.

		Conserved matrisome	Nematode-specific matrisome
		Human to *C. elegans* orthologues [# found/total (percentage)]	Not found in mammals [# found/total (percentage)]
	Complete matrisome	467/1027 (45%)	252/719 (35%)
Core matrisome	ECM glycoproteins	35/195 (18%)	0/35 (0%)
	Collagens	4/44 (9%)	177/181 (98%)
	Proteoglycans	3/35 (9%)	7/10 (70%)
	Cuticlins	–	12/12 (100%)
	Total	42/274 (15%)	196/238 (82%)
Matrisome-associated	ECM-affiliated proteins	290/171 (170%)	11/301 (4%)
	ECM regulators	97/238 (41%)	31/128 (24%)
	Secreted factors	38/344 (11%)	14/52 (27%)
	Total	425/753 (56%)	56/481 (12%)

various bona fide collagens in *C. elegans*, such as *col-51* and *col-142*, probably due to their small collagenous domains interrupted by non-collagenous stretches. Collagen domains are characterized by their glycine-X-Y amino acid triplet repeats, whereby X and Y are frequently proline and 4-hydroxyproline residues, respectively. *In vitro*, 10 Gly-X-Y repeats are typically sufficient to form stable triple helices with melting temperatures, depending on the content of proline and hydroxyproline residues on the X and Y positions. Therefore, for a more sensitive approach, we generated a simple regular expression in Python matching at least 10 Gly-X-Y repeats (regex = r(G.){10,}) and used it against the above-mentioned dataset. In total, we found 243 entries in the UniProt database matching this pattern. After deleting duplicated entries and cross-referencing against WormBase (WS262) [34], we obtained a list of 201 unique entries. However, besides a repeating Gly-X-Y pattern, collagens also need frequent proline residues at the X and Y positions as this amino acid is important for the formation of single poly-proline II helices, which are the backbone of the collagen triple helix. In vertebrates, the percentage of proline residues at the X and Y position is approximately 30%. To avoid missing any potential sequences, we decided to use a cut-off of 10% proline at both positions, which still represents the double of the normal frequency of proline residues in the *C. elegans* proteome [35]. With this criterion, the list was narrowed to 190 potential collagen sequences, which were all curated manually for the likelihood of being a collagen. Five sequences (UniProt Q4W4Y5, G5EDS0, B3GWA1, O61209, Q9N3I0) were excluded, since they contain short glycine rich repeats, with only very few proline residues and no apparent collagen structures. These five proteins were also not recognized by the Pfam collagen profile. On the other hand, various sequences were not predicted to be collagen by the Pfam profile, but upon manual inspection have proper collagen domains (COL-51, COL-103, COL-161, COLl-142 and COL-183,). Finally, after manual curation, we identified 185 genes in total that encode collagen-domain containing proteins in *C. elegans* (Supplementary Table 4). Of these, 4 genes could be classified as gliomedins or collectins based on their small collagenous domain and the presence of further signature domains (*see below*). The remaining 181 proteins define the existent collagens in *C. elegans*.

The *C. elegans* matrisome consists of 719 genes

After combining the lists of genes and proteins identified above, we manually curated each entry and assigned them to matrisome divisions and categories. Last, in order to identify putative matrisome genes and proteins that have not been captured by the gene orthology approach or the structural domain-based approach, we searched both WormBase (http://www.wormbase.org/, release WS263; [34]) and the *C. elegans* reference proteome from UniProt to identify genes and proteins annotated as ECM genes by a selection of Gene Ontology – Cellular Component terms (Supplementary Table 3). This last step allowed us to identify an additional 11 genes that had not been identified otherwise and may be considered as matrisome components (Supplementary Table 1; see Column A, with the exception of *col-78*, which was identified earlier by the structural domain-based approach).

Altogether, we identified 719 *C. elegans* matrisome genes out of the total ~20,000 *C. elegans* protein-encoding genes, suggesting that 4% of the *C. elegans* genome is dedicated to ECM genes (Table 1; Supplementary Table 1). This is comparable to the 1027 human matrisome genes, which also represents about 4% the human genome [17,26]. We further classified these 719 genes into divisions and categories proposed to classify the mammalian matrisomes. We found 226 genes for the *C. elegans* core-matrisome (ECM glycoproteins, collagens, proteoglycans). 181 out of the 226 core-matrisome genes are collagen genes (Table 1), of which 173 are predicted to be nematode-specific cuticular collagens (*see below*). We found that the *C. elegans* genome encodes 35 ECM glycoproteins compared to 195 found in humans (Table 1). All 35 *C. elegans* ECM glycoproteins have mammalian orthologues and thus far no *C. elegans*-specific ECM glycoprotein was identified (Table 2). By contrast, 7 out of the 10 *C. elegans* proteoglycans are nematode-specific and several are sulfate-less-chondroitin-binding proteoglycans (*cpg-1-4*, *cpg-7-9*) [36]. The remaining three proteoglycans are similar to the heparan sulfate proteoglycan perlecan (*unc-52*; *Hspg2 in mammals*) [13], a SPOCK/Testican (*test-1*), and a leucine-rich proteoglycan nyctalopin (*lron-8*) (Supplementary Table 1).

The *C. elegans* genome comprises 493 matrisome-associated genes (ECM-affiliated proteins, ECM regulators, and secreted factors) compared to the 753 human matrisome-associated genes. The majority of these 493 *C. elegans* matrisome-associated genes are C-type lectins (240 genes; Table 1 and Supplementary Table 1) [37].

Orthology relationship between human and *C. elegans* matrisome genes

Next, we determined the conserved versus nematode-specific matrisome genes for each matrisome category (Table 2). We compared the human matrisome genes to the *C. elegans* matrisome genes and vice-versa using OrthoList [7], or directly aligned them and examined the conservation of domains. In agreement with previous reports [10,38], most of the *C. elegans* collagens are predicted to be cuticular collagens that share no or little orthology to mammalian collagens (Table 2). However, other collagens and

ECM proteins that originated in basal metazoans are found to be well conserved in *C. elegans* (Fig. 2). These include basement membrane proteins (laminins, collagen type IV, nidogen, perlecan), transmembrane proteoglycans classified as ECM-affiliated proteins, syndecan and glypican, other collagens (type IX, XVIII, and XXV collagens), and axon guidance proteins (netrins, slits, agrin, fibrillin) (Fig. 2). Although thrombospondins are found in metazoans and we found many *C. elegans* proteins containing thrombospondin domains, we did not find a thrombospondin orthologue in agreement with previous reports [39]. Furthermore, ECM proteins that evolved during the vertebrate expansions, such as fibronectin, complex collagens, LINK proteins, and hyalectans (Fig. 2), were not identified in the *in-silico* searches in *C. elegans*, consistent with previous reports [16]. Last, some ECM proteins identified are shared between nematodes and humans, but not with other organisms like yeasts or *Drosophila*. These proteins include hemicentin (*him-4*) [40], SPARC/osteonectin (*ost-1*) [41,42], fibulin (*fbl-1*) [43], spondin (*spon-1*) [44], and olfactomedin (*unc-122*) [45].

Taken together, our survey of the *C. elegans* genome and proteome provides the first comprehensive compendium of the *C. elegans* matrisome. To facilitate the use of our lists of predicted genes encoding ECM and ECM-associated proteins in the *C. elegans* genome, we have deposited them on a dedicated page of the Matrisome Project website (http://matrisome.org) [26]. In addition, we have built an online tool, the *C. elegans* Matrisome Annotator, which, provided a list of genes, returns it annotated for matrisome components (http://ce-matrisome-annotator.permalink.cc/; tutorial provided as Supplementary Data).

Proposal of a novel classification of *C. elegans* collagens

In order to better classify and study the 185 collagen-domain-containing proteins in *C. elegans*, we propose to define a novel nomenclature based on their collagen-domain organization and the addition of other characteristic protein domains (*e.g.* C-type lectin; C4, the collagen IV NC1 domain; TSP; FNIII), similar to the mammalian collagen classification [46]. To do so, we clustered the 181 collagens and the 4 collagen-domain containing proteins into four major groups: (1) the vertebrate-like collagens (similar to mammalian type IV, XVIII, XXV), (2) the collagen-domain-containing proteins with mammalian orthologues (collectins and gliomedin), (3) the non-cuticular collagens with no clear orthology to mammalian collagens, and (4) the cuticular collagens. This last group contains the largest number of 173 collagens and which we further propose to subdivide into five main clusters (A to E). For detailed comparison and to facilitate the

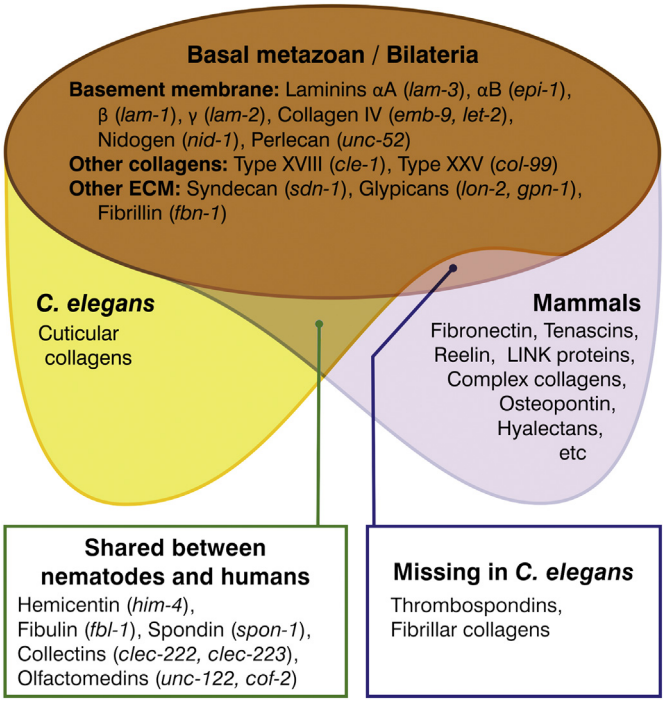

Fig. 2. Conserved *C. elegans* matrisome in the context of the evolution of ECM proteins. Figure is adapted from [16]. Corresponding *C. elegans* orthologues are italicized and indicate in parenthesis *e.g.* (*lam-1*). Individual genes and corresponding orthologues are found in Supplementary Table 1.

diffusion of this proposed classification, we constructed the *C. elegans* collagen database, CeColDB, available at: http://CeColDB.permalink.cc/.

Group 1: the conserved vertebrate-like collagens in *C. elegans*

Although fibrillar collagens are found in metazoans [47,48], our computational approach did not find any genes encoding fibrillar collagens in the *C. elegans* genome, which is in agreement with previous reports [49]. It has been hypothesized that *C. elegans* might have lost fibrillar collagens, since no evidence for an interstitial matrix is found in *C. elegans* [14,49,50]. However, the basement membrane type IV collagens are well conserved in *C. elegans* [13]. The *C. elegans* collagen-IV-like proteins are encoded by two genes *emb-9* and *let-2*, which both have collagenous domains of 1488 and 1487 amino acids, respectively, similar to their vertebrate (1398 amino acids) counterparts (Fig. 3A). The C-terminal domain (C4) is well conserved and the sequence identity of 52% and 69% among the nematode and human domains are similar to the variance in the 6 existing human genes. In phylogenetic analyses, the C-terminal domain of LET-2 consistently clusters with the even-numbered collagen alpha chains (α2 (IV), α4(IV) and α6(IV)), while EMB-9 groups with the odd-numbered collagen IV chains (α1(IV), α3(IV) and α5(IV)) (Fig. 3B). In humans, heterotrimers of collagen IV are formed by one even-numbered and two odd-numbered chains ($[\alpha1]_2[\alpha2]$, [α3][α4][α5], $[\alpha5]_2[\alpha6]$). Thus, we speculate that the collagen IV in *C. elegans* is an $[EMB\text{-}9]_2[LET\text{-}2]$ heterotrimer. In addition, *let-2* is alternatively spliced whereby one version is predominantly found in embryos and the other version in larval stages [51]. Both *emb-9* and *let-2* are essential genes and glycine mutations in the Gly-X-Y repeats result in retainment of this mutant collagen in the endoplasmic reticulum and arrest in embryonic development [52,53].

The *C. elegans* CLE-1 protein has similarities to collagen type XV [54] and XVIII [55]. CLE-1 is also found in basement membranes, but predominantly localized around neurons. Similar to the phenotype of the *Col18a1*-null mice [55], reduction of *C. elegans cle-1* function results in defects in the organization of the nervous system, but, in contrast, also results in a partially-penetrant embryonic lethality which may be due to failure of epidermal cell migration [56]. CLE-1 has one or two fibronectin-type-III-like domains, a laminin G-like domain, a very short interrupted collagenous domain, and an endostatin domain (Fig. 3C). Based on the last domains, CLE-1 is classified as an orthologue of collagen type XV [54] or XVIII [55], however, it is worth noting that the overall sequence identity is only 14% and 18% for collagen XV and XVIII, respectively, and the collagenous domain of CLE-1 is short in contrast to collagen XV or XVIII. The *C. elegans* collagen COL-99 is also not an essential protein, but is important for the organization of the nervous system [57]. COL-99 is a type II transmembrane-domain-containing protein with a smaller cytoplasmic region and a larger extracellular region containing 10 smaller collagenous domains. It therefore formally groups with the vertebrate Membrane-Associated Collagens with Interrupted Triple-helices (MACITs: collagen types XIII, XXIII, and XXV; Fig. 3D) [58].

Group 2: collagen-domain-containing proteins with mammalian orthologues

We identified four collagens with additional non-collagenous domains, which, based on their domain organization, resemble mammalian gliomedins and collectins. The group of gliomedin-like proteins consists of *cof-2* and *unc-122*. Both have a predicted N-terminal transmembrane domain, followed by a collagenous domain of 15 triplets and a C-terminal olfactomedin-like domain (Fig. 4A). However, despite their similar domain organization both proteins only share approximately 26% sequence identity with each other. Mutations in *unc-122* cause an uncoordinated locomotory behavior, the so-called Unc phenotype [59]. The group of collectins harbours two genes (*clec-222* and *clec-223*) which are oriented in a head-to-tail fashion on chromosome V. Both have very short collagenous domains (10 triplets), which might still permit trimerization and 1 or 3 C-type lectin domains (Fig. 4B).

Group 3: the *C. elegans* non-cuticular collagens with no clear orthology to mammalian collagens

This group consists of four collagens that cluster neither with basement membrane or vertebrate-like collagens nor with cuticular collagens. We can speculate that they might have specialized functions or localize to other ECMs than the basement membrane or the cuticle. MEC-5 is a collagen of medium size with a short collagen domain, similar to the N-terminal pro-helices found in fibrillar collagens, followed by a major uninterrupted collagenous domain. There is no further domain predicted and no obvious similarities to vertebrate collagens (Fig. 4C). The MEC-5 collagen is produced and secreted from hypodermal cells to anchor the ion channel/degenerin complex (MEC-4/10) that is expressed from touch receptor neurons to the ECM and thus MEC-5 is essential for the mechanosensory response to gentle touch [60,61]. COL-55 and ROL-8 are similar to the cuticular collagens discussed below, but are missing certain features, like the N-Pro-helix or the characteristic cysteine knots. COL-55 and ROL-8 are predicted to have a transmembrane domain, which overlaps with the predicted N-cuticular domain (Fig. 4C). Mutations in *rol-8* cause a left-handed rolling phenotype (a helically twisted

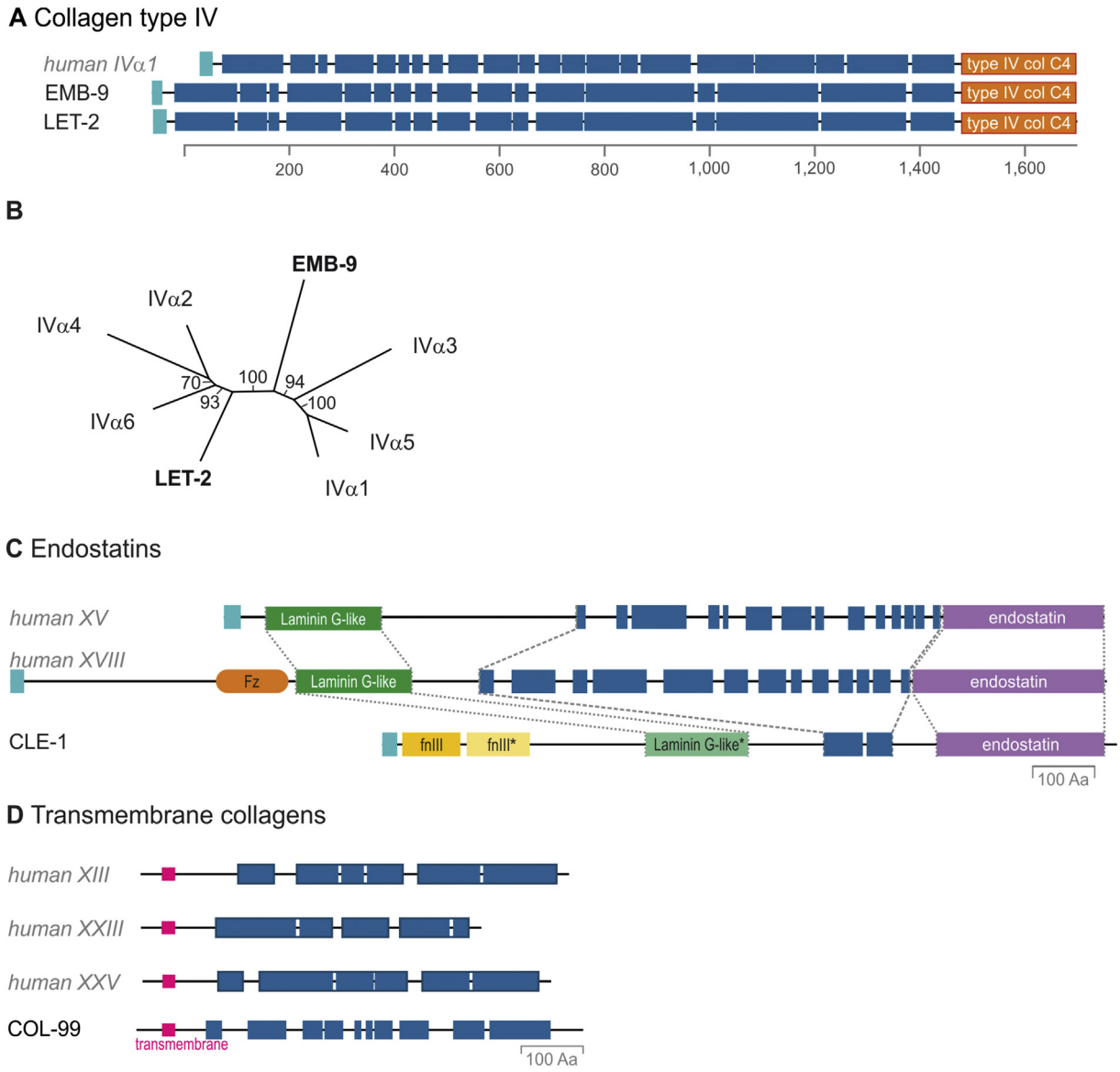

Fig. 3. Group 1: *C. elegans* collagens with orthologues in vertebrates. (A) Collagen type IV. Alignment of collagen type IV α1 from *Homo sapiens* with EMB-9 and LET-2 from *Caenorhabditis elegans*. (B) Phylogenetic analysis of collagen type IV. The C-terminal collagen type IV domains (C4) of humans and *C. elegans* were analyzed using ClustalO [70] followed by Neighbor Joining. The numbers indicate the bootstrap values of 100 replicates. The nematode-specific sequences are indicated in bold (EMB-9 and LET-2). (C) Endostatins. Comparison of CLE-1 from *C. elegans* with endostatin-containing collagens type XV and XVIII from *H. sapiens*. Asterisk (*) indicate domain predictions with weak significance. (D) Transmembrane collagens. Comparison of *C. elegans* COL-99, the only non-cuticular transmembrane collagen, with its human orthologues (collagens type XIII, XXIII, XXV). For human proteins, dark outlines group collagenous stretches recognized as collagen domains in earlier publications. All panels are drawn to scale. Colour codes are as follow: light blue: signal peptides; pink: transmembrane region; orange: frizzled domain or collagen C4 domain; yellow: Laminin G-like domain; purple: endostatin domain; blue: collagenous Gly-X-Y repeats.

body), suggesting its importance in cuticle assembly and/or chirality [62]. COL-135 has been predicted to be a collagen but has a very particular composition. Its sequence contains a signal peptide, three short collagen domains, and a rather large domain of Gly-X-Y repeats. Proline residues are under-represented compared to other *C. elegans* collagens especially at the Y position (17.0% and 3.6% for X and Y in COL-135, 43.3% and 23.9% in all collagens), but lysine and aspartate residues are over-represented

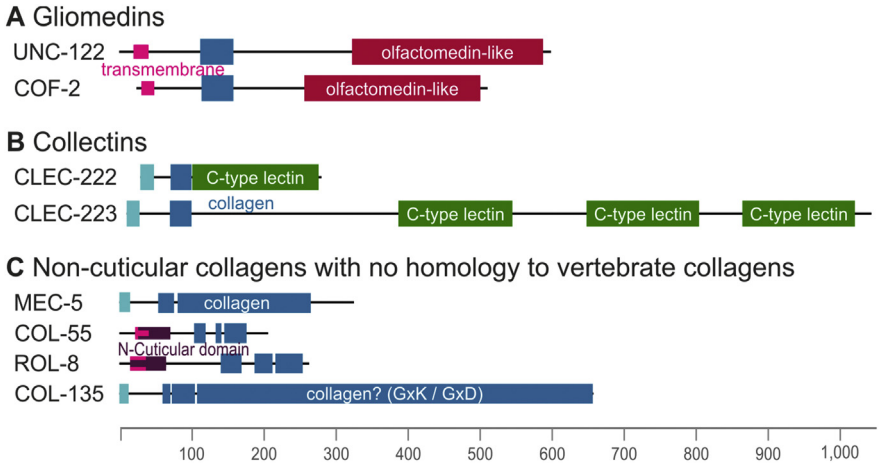

Fig. 4. Domain organization of *C. elegans* collagens from Group 2 and 3. (A-B) Group 2: Collagen-domain containing proteins with mammalian orthologues. (A) Gliomedin-like collagens are characterized by an olfactomedin-like domain. (B) Collectins carry C-type lectin domains and only small collagenous domains. (C) Group 3: The *C. elegans* non-cuticular collagens with no clear orthology to mammalian collagens. MEC-5 is a neuron-specific collagen. For COL-55 and ROL-8 the hypothetical Nematode cuticle collagen N-terminal domain (PF01484) overlaps with the predicted transmembrane domain. COL-135 is high in GxK and GxD repeats (Supplementary Fig. 1). All panels are drawn to a common scale. Colour codes are as follow: light blue: signal peptides; pink: transmembrane region; dark violet: N-cuticular domain (PF01484); green: C-type lectin; red: olfactomedin-like domain; blue: collagenous Gly-X-Y repeats.

(28.6% and 22.2%, respectively in COL-135, compared to 4.3% and 8.1% in all collagens) (Supplementary Fig. 1A). Furthermore, 109 out of the 198 triplets are Gly-X-Lys repeats (Supplementary Fig. 1B). Although COL-135 meets the criteria stated above for being recognized as a collagen, it is uncertain whether it is able to form a bona fide collagen triple helix.

Group 4: the cuticular collagens

Approximately 80% of the cuticle is made of collagenous proteins [29]. Previously, cuticular collagens were grouped according to their cysteine knots into 6 groups based on 20 collagen sequences known at that time [63]. Furthermore, these 20 known collagen sequences showed four shared amino-acid-sequence motifs (termed "homology blocks") in the N-terminal region before the Gly-X-Y domains [63]. Expanding the analysis from 20 to the 173 cuticular collagen genes identified in our study, we did not identify these shared homology blocks. We found three conserved features which occur in various combinations in many but not all cuticular collagens: a serine (in position 21 in BLI-6 and position 78 in the alignment; conserved in 75% of all cuticular collagens), a potential furin cleavage site with an RxxR consensus (in position 71–74 and 195–198, respectively; 93% conservation), and a tyrosine (in

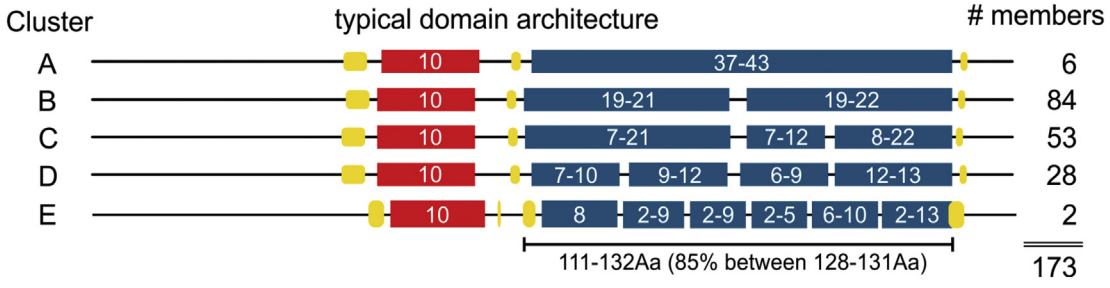

Fig. 5. Organization of the cuticular collagen clusters. The 173 cuticular collagens were grouped into 5 clusters based on the interruptions in their main collagenous domain (blue boxes). All cuticular collagens contain a shorter N-terminal helical domain (red box) and typically three cysteine-rich regions (yellow) flanking the collagenous domains. Numbers in the boxes correspond to the number of Gly-X-Y repeats.

position 78 and 273, respectively; approx. 50% conservation) that may be important for tyrosine-tyrosine crosslinking (Supplementary Fig. 2).

Here, we defined the cuticular collagens based on their characteristic collagenous domains consisting of 37 to 43 Gly-X-Y triplets, which are flanked by N- and C-terminal cysteine knots (Fig. 5; Supplementary Table 4). As in fibrillar collagens found in vertebrates, there is an additional N-Pro-helix of usually 10 G-X-Y repeats long located between 12 and 31 residues (in 97% of the cases between 13 and 23) N-terminally of the collagenous domain. This N-Pro-helix is often stabilized by an additional cysteine knot (Fig. 5; Supplementary Table 4). Based on the interruptions in their main collagenous domain, we grouped the cuticular collagens into five main clusters (A to E), having either 0, 1, 2, 3, or >3 interruptions (Fig. 5; Supplementary Table 4). We further classified these five main clusters based on the length of their collagenous domains, the positions of interruptions, and their prediction of being transmembrane or secreted (Supplementary Figs. 3–7; Supplementary Table 4). These sub-clusters are numbered based on the length of their uninterrupted collagenous stretches, counting from the C-terminus in an ascending manner (Supplementary Table 4). Members of a sub-cluster often have the same cysteine knot (Supplementary Table 4) although the same cysteine knot might also occur in different clusters (Supplementary Table 4). One type of cysteine knot, reported to be important for tyrosine-crosslinking [63], can be found in 35 collagens (Supplementary Table 4/CysKnot C-Col domain, red-marked Y). Many, but not all, of the predicted transmembrane collagens have predicted furin cleavage sites potentially enabling the shedding of these collagens (Supplementary Table 4).

Cluster A

Cluster A comprises six members and is subdivided into four sub-clusters (A1–4; Supplementary Fig. 3; Supplementary Table 4). Members of sub-clusters A2 and A4 are predicted to be transmembrane proteins. Interestingly, sub-clusters A3 and A4 have the same length of their collagenous domain (Supplementary Fig. 3; Supplementary Table 4).

Cluster B

Cluster B comprises 84 members divided into 18 sub-clusters (Supplementary Fig. 4; Supplementary Table 4). The two collagenous domains are typically 20 or 21 triplets long and the non-collagenous interruption only differs by a few amino acids.

Cluster C

Cluster C comprises 53 members divided into 27 sub-clusters. 19 out of the 27 C sub-clusters consist only of one cuticular collagen gene (Supplementary Fig. 5; Supplementary Table 4).

Cluster D

Cluster D comprises 28 members forming 17 sub-clusters (Supplementary Fig. 6; Supplementary Table 4). Of note, COL-51 is predicted to be a multispan-membrane protein with a total of four transmembrane (TM) regions, with the collagenous domain between the second and the third TM. If this prediction is correct, it would be interesting to know how the collagenous domain of COL-51 forms.

Cluster E

Cluster E comprises two members, with five interruptions in their collagenous domain (Supplementary Fig. 7; Supplementary Table 4).

Taken together, the newly proposed classification of cuticular collagens is based on the structural similarity of the collagenous domain. This will help identify similar collagens with similar, redundant or compensatory functions. Furthermore, it is likely that heterotrimers exist within the cuticular collagen family. We hope that our system will help to identify potential candidates for heterotrimerization.

Low amino acid sequence similarity among cuticular collagens

Among the 173 cuticular collagens, approximately 30 genes form 9 similarity groups that share high sequence similarity to each other (>90%, Supplementary Fig. 8, pink/white).

Sequences with a similarity of over 90% normally belong to the same sub-cluster, with the minor exception of COL-146 and COL-147, which belong to C21a and C21c, respectively. However, not all members of a sub-cluster group by sequence similarity. For example, although many members of the sub-cluster B9 group together based on the sequence alignment (Supplemental Fig. 8, green bars), some members, like COL-152 or COL-123, are separated and only show weak sequence similarity to the sub-cluster (approx. 35%, Supplementary Fig. 8 or visit http://CeColDB.permalink.cc/ website and use “recursive on” with cluster B9). Additionally, some collagens show sequence similarity with members of B9, but group differently based on their collagen domain organization (*e.g.* COL-148 and COL-150). A similar pattern can be observed in the sub-cluster B14 (Supplemental Fig. 8, green bars). Overall, the cuticular collagens only show a relatively low sequence similarity (81% with

<40% identity; Supplementary Fig. 8), with the exception of COL-126 and COL-127. These two proteins are identical at the amino acid level, but also at the nucleotide level (in both exons and introns), which raises the question of whether this is one gene misannotated as two collagens. We confirmed by PCR that *col-126* and *col-127* are indeed two distinct genes located next to each other in inverse direction (Supplementary Fig. 9), suggesting a very recent gene duplication event.

As the structural similarity of cuticular collagens is striking, it is very likely that they originate from a common evolutionary ancestor. The low sequence identity further suggests that upon gene multiplication cuticular collagens diversified to fulfill various important functions in *C. elegans*. However, as the prerequisites for collagen helices are relatively low, there is also the possibility that evolutionary pressure was mostly directed towards domain organization and less to the primary sequence.

Functions of cuticular collagens

Twenty one out of the 173 cuticular collagens have been isolated in genetic mutagenesis screens [10] (*bli-1*, *bli-2*, *bli-6*, *dpy-2*, *dpy-3*, *dpy-4*, *dpy-5*, *dpy-7*, *dpy-8*, *dpy-9*, *dpy-10*, *dpy-13*, *dpy-14*, *dpy-17*, *lon-3*, *ram-2*, *rol-1*, *rol-6*, *sqt-1*, *sqt-2*, *sqt-3*; Supplementary Table 4). Mutations in these cuticular collagens affect the synthesis or assembly of the cuticle and thereby alter body morphology. These cuticular collagens are named based on their phenotype: long (*lon-#*) are about 1.5 times the length of wild type, dumpy (*dpy-#*) are short and fat-looking, roller (*rol-#*) roll around their helical axis instead of the sinusoid-curve crawling of wild type, blister (*bli-#*) show detachment of cuticular layer forming blisters along the body, Ray abnormal (*ram-#*) affects the morphology of the male tail, and squat (*sqt-#*) can lengthen, shorten, or helically twist *C. elegans* [10] (Supplementary Table 4). For instance, *sqt-1*(*e1350*) mutation leads to a R69C substitution altering the predicted furin **R**VR**R** cleavage site to **R**VR**C** before the collagen domains. These *sqt-1* (*e1350*) mutant *C. elegans* show stage-specific phenotypes: larval stage L1–2 are wild-type, L3 or dauer are rolling, and L4 are dumpy [64]. By contrast, *sqt-1* null mutants are wild type [64], suggesting that the absence of SQT-1 collagen has no effect on the cuticular structure and there is redundancy among the cuticular collagens to compensate for the absence of SQT-1. However, the genotype to phenotype interpretation of how these collagens interact to from and integrate into their ECM is complex [65]. For instance, *sqt-1* null mutations suppress the rolling phenotype of *rol-6* mutants, suggesting that collagen ROL-6 gene product depends on the presence of collagen SQT-1 [66], whereas both null mutations of *sqt-1* or *rol-6* suppress *lon-3* mutant phenotype [67]. Taken together, with the complete matrisome list, it becomes now possible to start dissecting out the complex genetics underlying the formation of ECM structures *in vivo*.

Utilizing the *in-silico C. elegans* matrisome to annotate large datasets

RNA sequencing and proteomics are standard techniques used by many *C. elegans* research laboratories to elucidate physiological and pathological processes. In addition, genome-wide RNA interference (RNAi) screens are commonly used to identify the mechanism(s) underlying phenotypes of interest.

To demonstrate the applicability and power of our matrisome definition and classification, we used the Matrisome Annotator we developed here (http://ce-matrisome-annotator.permalink.cc/) to re-annotate existing datasets. We first re-analyzed our previously published study using transcriptomics to identify genes involved in longevity [15]. We found 79 matrisome genes out of the total 426 transcriptionally upregulated genes when comparing long-lived *C. elegans* under reduced Insulin/IGF-1 conditions with short-lived *C. elegans* that lack the oxidative stress transcription factor SKN-1/Nrf1,2,3 (Supplementary Table 5) [15]. Although, we previously recognized the upregulated collagens and potentially secreted proteases [15], the re-annotation of this data set paints a more complete picture to envision a remodeling of the ECM in long-lived *C. elegans*. Our list can also be used to annotate proteomic datasets. Here, we re-annotated a proteomic dataset from a recently published study aimed at studying longevity in *C. elegans* [68]. In contrast to the 11 collagens highlighted in their study, we found 25 matrisome proteins out of the 177 total upregulated proteins when comparing long-lived germ stem cell mutant *glp-1* with wild-type *C. elegans* (Supplementary Table 6). The additionally identified matrisome proteins includes laminin A and B (EPI-1 and LAM-1), prolyl 4-hydroxylase (DPY-18), and secreted proteases (Supplementary Table 6). Together with the 11 previously identified collagens [68], this suggests a potential remodeling of the ECM in long-lived *C. elegans*, consistent with the findings from the mRNA expression profile [15]. Last, we set out to re-annotate data from a whole-genome RNAi screen aimed at identifying antifungal innate immunity genes [69], since this would help to identify functional importance of matrisome genes. We found that 18 out of the 297 gene hits that regulate antimicrobial peptide gene expression are matrisome genes (Supplementary Table 7) [69]. These 18 matrisome genes include six cuticular collagens, three secreted proteases, and one collagen cross-linking enzyme (Supplementary Table 7), suggesting a potential role for strengthening or stiffening of the ECM to form a protective barrier against fungal infections.

By using the *C. elegans* Matrisome Annotator tool, we found substantial enrichment for matrisome genes in these data sets. Thus, re-analyzing -omic datasets with the *C. elegans* Matrisome Annotator tool may be useful to generate novel hypotheses about the role of the *C. elegans* matrisome for various biological processes.

Conclusions

Defining proteins in cellular compartments has helped understand their functions and implication in various processes. The ECM has been implicated in many biological processes. Components of the ECM have essential roles for *C. elegans* development, cell migration, and aging. In this study, we defined the *C. elegans* matrisome, an ensemble of ECM proteins and associated factors. We identified conserved and nematode-specific components, which informs biomedical research and provides potential targets to fight pathogenic nematodes. The categorization and clustering of *C. elegans* collagens lays the foundation to experimentally test, for example, whether cuticular collagens might form heterotrimers. Using the *C. elegans* Matrisome Annotator tool, we found enrichment of ECM genes at the mRNA, protein, and phenotypic level. This will assist researchers in delineating genotype-to-phenotype relationships for ECM genes. Modern science is hypothesis-driven. We hope that our contribution in defining the *C. elegans* matrisome and providing tools to analyze -omic data will aid generating novel hypotheses to propel science forward.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mbplus.2018.11.001.

Acknowledgement

We thank Gary Williams from WormBase and Paolo Bazzicalupo for helpful discussions about *col-126/-127* and cuticlin-like genes, respectively, and members of the Naba and Ewald labs for discussion and comments on the manuscript.

Funding sources

This work was supported by Swiss National Science Foundation [163898] to ACT, CS, and CYE, the work of AN and MND was supported by a start-up fund from the Department of Physiology and Biophysics at the University of Illinois at Chicago. JMG was supported by the Deutsche Forschungsgemeinschaft SFB829/B11.

Author contributions

All authors participated in analyzing and interpreting the data.

AN, ACT, MND, EJ, and CYE defined the *in-vitro* matrisome and analyzed expression data. CS developed the online matrisome annotation script.

JMG identified and established the *C. elegans* collagen classification.

JMG, AN, and CYE wrote the manuscript in consultation with the other authors.

Author information

The authors have no competing interests to declare.

Received 19 October 2018;
Received in revised form 26 November 2018;
Accepted 26 November 2018
Available online 21 February 2019

Keywords:
Nematode;
Extracellular matrix;
Collagen;
Cuticle;
Basement membrane

References

[1] G. Stepek, D.J. Buttle, I.R. Duce, J.M. Behnke, Human gastrointestinal nematode infections: are new control methods required? Int. J. Exp. Pathol. 87 (2006) 325–341, https://doi.org/10.1111/j.1365-2613.2006.00495.x.

[2] Deworming for health and development: report of the third global meeting of the partners for parasite control, Partners for Parasite Control Meeting 3rd 2004 Geneva, World Health Organization 2005, pp. 1–64http://www.who.int/iris/handle/10665/69005.

[3] G.C. Bernard, M. Egnin, C. Bonsi, The impact of plant-parasitic nematodes on agriculture and methods of control, Nematology-Concepts, Diagnosis and Control, InTech 2017, pp. 1–33, https://doi.org/10.5772/intechopen.68958.

[4] L. Holden-Dye, R.J. Walker, How relevant is *Caenorhabditis elegans* as a model for the analysis of parasitic nematode biology? Parasitic Helminths, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany 2012, pp. 23–41, https://doi.org/10.1002/9783527652969.ch2.

[5] C.Y. Ewald, Redox signaling of NADPH oxidases regulates oxidative stress responses, immunity and aging, Antioxidants 7 (2018) 130–136, https://doi.org/10.3390/antiox7100130.

[6] A.K. Corsi, B. Wightman, M. Chalfie, A transparent window into biology: a primer on *Caenorhabditis elegans*, Genetics 200 (2015) 387–407, https://doi.org/10.1534/genetics.115.176099.

[7] D.D. Shaye, I. Greenwald, OrthoList: a compendium of *C. elegans* genes with human orthologues, PLoS ONE 6 (2011), e20085. https://doi.org/10.1371/journal.pone.0020085.

[8] T. Kaletta, M.O. Hengartner, Finding function in novel targets: *C. elegans* as a model organism, Nat. Rev. Drug Discov. 5 (2006) 387–398, https://doi.org/10.1038/nrd2031.
[9] E. Culetto, D.B. Sattelle, A role for *Caenorhabditis elegans* in understanding the function and interactions of human disease genes, Hum. Mol. Genet. 9 (2000) 869–877.
[10] A.P. Page, I.L. Johnstone, The cuticle, WormBook: the online review of *C. elegans*, Biology (2007) 1–15, https://doi.org/10.1895/wormbook.1.138.1.
[11] D.P. Keeley, D.R. Sherwood, Tissue linkage through adjoining basement membranes: the long and the short term of it, Matrix Biol. (2018)https://doi.org/10.1016/j.matbio.2018.05.009.
[12] J.C. Adams, Matricellular proteins: functional insights from non-mammalian animal models, Curr. Top. Dev. Biol. 130 (2018) 39–105, https://doi.org/10.1016/bs.ctdb.2018.02.003.
[13] J.M. Kramer, Basement membranes, WormBook: the online review of *C. elegans*, Biology (2005) 1–15, https://doi.org/10.1895/wormbook.1.16.1.
[14] D.R. Sherwood, J. Plastino, Invading, leading and navigating cells in *Caenorhabditis elegans*: insights into cell movement in vivo, Genetics 208 (2018) 53–78, https://doi.org/10.1534/genetics.117.300082.
[15] C.Y. Ewald, J.N. Landis, J. Porter Abate, C.T. Murphy, T.K. Blackwell, Dauer-independent insulin/IGF-1-signalling implicates collagen remodelling in longevity, Nature 519 (2015) 97–101, https://doi.org/10.1038/nature14021.
[16] R.O. Hynes, A. Naba, Overview of the matrisome—an inventory of extracellular matrix constituents and functions, Cold Spring Harb. Perspect. Biol. 4 (2012) a004903, https://doi.org/10.1101/cshperspect.a004903.
[17] A. Naba, K.R. Clauser, S. Hoersch, H. Liu, S.A. Carr, R.O. Hynes, The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices, Mol. Cell. Proteomics 11 (2012), M111.014647. https://doi.org/10.1074/mcp.M111.014647.
[18] A. Naba, S. Hoersch, R.O. Hynes, Towards definition of an ECM parts list: an advance on GO categories, Matrix Biol. 31 (2012) 371–372, https://doi.org/10.1016/j.matbio.2012.11.008.
[19] P. Nauroy, S. Hughes, A. Naba, F. Ruggiero, The *in-silico* zebrafish matrisome: a new tool to study extracellular matrix gene and protein functions, Matrix Biol. 65 (2018) 5–13, https://doi.org/10.1016/j.matbio.2017.07.001.
[20] P. Nauroy, A. Guiraud, J. Chlasta, M. Malbouyres, B. Gillet, S. Hughes, et al., Gene profile of zebrafish fin regeneration offers clues to kinetics, organization and biomechanics of basement membrane, Matrix Biol. (2018)https://doi.org/10.1016/j.matbio.2018.07.005.
[21] A. Naba, K.R. Clauser, J.M. Lamar, S.A. Carr, R.O. Hynes, Extracellular matrix signatures of human mammary carcinoma identify novel metastasis promoters, elife 3 (2014), e01308. https://doi.org/10.7554/eLife.01308.
[22] A.M. Socovich, A. Naba, The cancer matrisome: from comprehensive characterization to biomarker discovery, Semin. Cell Dev. Biol. (2018)https://doi.org/10.1016/j.semcdb.2018.06.005.
[23] Y. Zhou, J.C. Horowitz, A. Naba, N. Ambalavanan, K. Atabai, J. Balestrini, et al., Extracellular matrix in lung development, homeostasis and disease, Matrix Biol. (2018)https://doi.org/10.1016/j.matbio.2018.03.005.
[24] V.L. Massey, C.E. Dolin, L.G. Poole, S.V. Hudson, D.L. Siow, G.N. Brock, et al., The hepatic "matrisome" responds dynamically to injury: characterization of transitional changes to the extracellular matrix in mice, Hepatology 65 (2017) 969–982, https://doi.org/10.1002/hep.28918.
[25] M.C. Staiculescu, J. Kim, R.P. Mecham, J.E. Wagenseil, Mechanical behavior and matrisome gene expression in the aneurysm-prone thoracic aorta of newborn lysyl oxidase knockout mice, Am. J. Physiol. Heart Circ. Physiol. 313 (2017) H446–H456, https://doi.org/10.1152/ajpheart.00712.2016.
[26] A. Naba, K.R. Clauser, H. Ding, C.A. Whittaker, S.A. Carr, R.O. Hynes, The extracellular matrix: tools and insights for the "omics" era, Matrix Biol. 49 (2016) 10–24, https://doi.org/10.1016/j.matbio.2015.06.003.
[27] R.J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, et al., Ensembl BioMarts: a hub for data retrieval across taxonomic space, Database (Oxford) (2011) https://doi.org/10.1093/database/bar030.
[28] T. UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 46 (2018) 2699, https://doi.org/10.1093/nar/gky092.
[29] G.N. Cox, M. Kusch, R.S. Edgar, Cuticle of *Caenorhabditis elegans*: its isolation and partial characterization, J. Cell Biol. 90 (1981) 7–17.
[30] M. Sebastiano, F. Lassandro, P. Bazzicalupo, cut-1 a *Caenorhabditis elegans* gene coding for a dauer-specific noncollagenous component of the cuticle, Dev. Biol. 146 (1991) 519–530.
[31] R.D. Finn, T.K. Attwood, P.C. Babbitt, A. Bateman, P. Bork, A.J. Bridge, et al., InterPro in 2017-beyond protein family and domain annotations, Nucleic Acids Res. 45 (2017) D190–D199, https://doi.org/10.1093/nar/gkw1107.
[32] S.R. Eddy, Accelerated profile HMM searches, PLoS Comput. Biol. 7 (2011), e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
[33] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A. L. Mitchell, et al., The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res. 44 (2016) D279–D285, https://doi.org/10.1093/nar/gkv1344.
[34] R.Y.N. Lee, K.L. Howe, T.W. Harris, V. Arnaboldi, S. Cain, J. Chan, et al., WormBase 2017: molting into a new stage, Nucleic Acids Res. 46 (2018) D869–D874, https://doi.org/10.1093/nar/gkx998.
[35] Y. Zhuang, F. Ma, J. Li-Ling, X. Xu, Y. Li, Comparative analysis of amino acid usage and protein length distribution between alternatively and non-alternatively spliced genes across six eukaryotic genomes, Mol. Biol. Evol. 20 (2003) 1978–1985, https://doi.org/10.1093/molbev/msg203.
[36] S.K. Olson, J.R. Bishop, J.R. Yates, K. Oegema, J.D. Esko, Identification of novel chondroitin proteoglycans in *Caenorhabditis elegans*: embryonic cell division depends on CPG-1 and CPG-2, J. Cell Biol. 173 (2006) 985–994, https://doi.org/10.1083/jcb.200603003.
[37] K. Drickamer, R.B. Dodd, C-type lectin-like domains in *Caenorhabditis elegans*: predictions from the complete genome sequence, Glycobiology 9 (1999) 1357–1369.
[38] I.L. Johnstone, Cuticle collagen genes. Expression in *Caenorhabditis elegans*, Trends Genet. 16 (2000) 21–27.
[39] A.A. Bentley, J.C. Adams, The evolution of thrombospondins and their ligand-binding activities, Mol. Biol. Evol. 27 (2010) 2187–2197, https://doi.org/10.1093/molbev/msq107.
[40] C.A. Whittaker, R.O. Hynes, Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere, Mol. Biol. Cell 13 (2002) 3369–3387, https://doi.org/10.1091/mbc.e02-05-0259.

[41] J.E. Schwarzbauer, C.S. Spencer, The *Caenorhabditis elegans* homologue of the extracellular calcium binding protein SPARC/osteonectin affects nematode body morphology and mobility, Mol. Biol. Cell 4 (1993) 941–952.
[42] M.A. Morrissey, R. Jayadev, G.R. Miley, C.A. Blebea, Q. Chi, S. Ihara, et al., SPARC promotes cell invasion in vivo by decreasing type IV collagen levels in the basement membrane, PLoS Genet. 12 (2016), e1005905. https://doi.org/10.1371/journal.pgen.1005905.
[43] F. Segade, Molecular evolution of the fibulins: implications on the functionality of the elastic fibulins, Gene 464 (2010) 17–31, https://doi.org/10.1016/j.gene.2010.05.003.
[44] W.-M. Woo, E. Berry, M.L. Hudson, R.E. Swale, A. Goncharov, A.D. Chisholm, The *C. elegans* F-spondin family protein SPON-1 maintains cell adhesion in neural and non-neural tissues, Development 135 (2008) 2747–2756, https://doi.org/10.1242/dev.015289.
[45] C.A. Karavanich, R.R. Anholt, Molecular evolution of olfactomedin, Mol. Biol. Evol. 15 (1998) 718–726, https://doi.org/10.1093/oxfordjournals.molbev.a025975.
[46] S. Ricard-Blum, The collagen family, Cold Spring Harb. Perspect. Biol. 3 (2011) a004978, https://doi.org/10.1101/cshperspect.a004978.
[47] A.L. Fidler, C.E. Darris, S.V. Chetyrkin, V.K. Pedchenko, S.P. Boudko, K.L. Brown, et al., Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues, elife 6 (2017), e15040. https://doi.org/10.7554/eLife.24176.
[48] S. Ozbek, P.G. Balasubramanian, R. Chiquet-Ehrismann, R. P. Tucker, J.C. Adams, The evolution of extracellular matrix, Mol. Biol. Cell 21 (2010) 4300–4305, https://doi.org/10.1091/mbc.E10-03-0251.
[49] H. Hutter, B.E. Vogel, J.D. Plenefisch, C.R. Norris, R.B. Proenca, J. Spieth, et al., Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes, Science 287 (2000) 989–994.
[50] R.P. Boot-Handford, D.S. Tuckwell, Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest, Bioessays 25 (2003) 142–151, https://doi.org/10.1002/bies.10230.
[51] M.H. Sibley, J.J. Johnson, C.C. Mello, J.M. Kramer, Genetic identification, sequence, and alternative splicing of the Caenorhabditis elegans alpha 2(IV) collagen gene, J. Cell Biol. 123 (1993) 255–264.
[52] X.D. Guo, J.J. Johnson, J.M. Kramer, Embryonic lethality caused by mutations in basement membrane collagen of *C. elegans*, Nature 349 (1991) 707–709, https://doi.org/10.1038/349707a0.
[53] M.H. Sibley, P.L. Graham, N. von Mende, J.M. Kramer, Mutations in the alpha 2(IV) basement membrane collagen gene of *Caenorhabditis elegans* produce phenotypes of differing severities, EMBO J. 13 (1994) 3278–3285.
[54] R. Ramchandran, M. Dhanabal, R. Volk, M.J. Waterman, M. Segal, H. Lu, et al., Antiangiogenic activity of restin, NC10 domain of human collagen XV: comparison to endostatin, Biochem. Biophys. Res. Commun. 255 (1999) 735–739, https://doi.org/10.1006/bbrc.1999.0248.
[55] R. Heljasvaara, M. Aikio, H. Ruotsalainen, T. Pihlajaniemi, Collagen XVIII in tissue homeostasis and dysregulation - lessons learned from model organisms and human patients, Matrix Biol. 57–58 (2017) 55–75, https://doi.org/10.1016/j.matbio.2016.10.002.
[56] B.D. Ackley, J.R. Crew, H. Elamaa, T. Pihlajaniemi, C.J. Kuo, J.M. Kramer, The NC1/endostatin domain of *Caenorhabditis elegans* type XVIII collagen affects cell migration and axon guidance, J. Cell Biol. 152 (2001) 1219–1232.
[57] J. Taylor, T. Unsoeld, H. Hutter, The transmembrane collagen COL-99 guides longitudinally extending axons in *C. elegans*, Mol. Cell. Neurosci. 89 (2018) 9–19, https://doi.org/10.1016/j.mcn.2018.03.003.
[58] H. Tu, P. Huhtala, H.-M. Lee, J.C. Adams, T. Pihlajaniemi, Membrane-associated collagens with interrupted triple-helices (MACITs): evolution from a bilaterian common ancestor and functional conservation in *C. elegans*, BMC Evol. Biol. 15 (2015) 281, https://doi.org/10.1186/s12862-015-0554-3.
[59] P.M. Loria, J. Hodgkin, O. Hobert, A conserved postsynaptic transmembrane protein affecting neuromuscular signaling in *Caenorhabditis elegans*, J. Neurosci. 24 (2004) 2191–2201, https://doi.org/10.1523/JNEUROSCI.5462-03.2004.
[60] L. Emtage, G. Gu, E. Hartwieg, M. Chalfie, Extracellular proteins organize the mechanosensory channel complex in *C. elegans* touch receptor neurons, Neuron 44 (2004) 795–807, https://doi.org/10.1016/j.neuron.2004.11.010.
[61] J.G. Cueva, A. Mulholland, M.B. Goodman, Nanoscale organization of the MEC-4 DEG/ENaC sensory mechanotransduction channel in *Caenorhabditis elegans* touch receptor neurons, J. Neurosci. 27 (2007) 14089–14098, https://doi.org/10.1523/JNEUROSCI.4179-07.2007.
[62] D.C. Bergmann, J.R. Crew, J.M. Kramer, W.B. Wood, Cuticle chirality and body handedness in *Caenorhabditis elegans*, Dev. Genet. 23 (1998) 164–174, https://doi.org/10.1002/(SICI)1520-6408(1998)23:3<164::AID-DVG2>3.0.CO;2-C.
[63] J.M. Kramer, Structures and functions of collagens in *Caenorhabditis elegans*, FASEB J. 8 (1994) 329–336.
[64] J.M. Kramer, J.J. Johnson, R.S. Edgar, C. Basch, S. Roberts, The sqt-1 gene of *C. elegans* encodes a collagen critical for organismal morphogenesis, Cell 55 (1988) 555–565.
[65] A.P. Page, A.D. Winter, Enzymes involved in the biogenesis of the nematode cuticle, Adv. Parasitol. 53 (2003) 85–148.
[66] J.M. Kramer, J.J. Johnson, Analysis of mutations in the sqt-1 and rol-6 collagen genes of *Caenorhabditis elegans*, Genetics 135 (1993) 1035–1045.
[67] J. Nyström, Z.-Z. Shen, M. Aili, A.J. Flemming, A. Leroi, S. Tuck, Increased or decreased levels of *Caenorhabditis elegans* lon-3, a gene encoding a collagen, cause reciprocal changes in body length, Genetics 161 (2002) 83–97.
[68] Y.-Z. Pu, Q.-L. Wan, A.-J. Ding, H.-R. Luo, G.-S. Wu, Quantitative proteomics analysis of *Caenorhabditis elegans* upon germ cell loss, J Proteomics. 156 (2017) 85–93, https://doi.org/10.1016/j.jprot.2017.01.011.
[69] O. Zugasti, N. Thakur, J. Belougne, B. Squiban, C.L. Kurz, J. Soulé, et al., A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes, BMC Biol. 14 (2016) 35, https://doi.org/10.1186/s12915-016-0256-3.
[70] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539, https://doi.org/10.1038/msb.2011.75.
[71] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190, https://doi.org/10.1101/gr.849004.
[72] S.M. Reynolds, L. Käll, M.E. Riffle, J.A. Bilmes, W.S. Noble, Transmembrane topology and signal peptide prediction using dynamic bayesian networks, PLoS Comput. Biol. 4 (2008), e1000213. https://doi.org/10.1371/journal.pcbi.1000213.