
Peer reviewed version

Link to published version (if available): 10.1042/EBC20180048

Link to publication record in Explore Bristol Research

PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Portland Press at https://portlandpress.com/essaysbiochem/article-lookup/doi/10.1042/EBC20180048. Please refer to any applicable terms of use of the publisher.

**University of Bristol - Explore Bristol Research**

**General rights**

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/user-guides/explore-bristol-research/ebr-terms/
Modelling the Early Evolution of Extracellular Matrix from Modern Ctenophores and Sponges

Graham Draper, Deborah K. Shoemark and Josephine C. Adams

School of Biochemistry, University of Bristol, Bristol BS8 1TD, U.K.

*Corresponding author. Email: jo.adams@bristol.ac.uk

Keywords: Ctenophora, Integrins, Metazoa, Multicellularity, Porifera, Protein Domains

Running Title: Early evolution of Metazoan ECM

Abbreviations: ADAMTS, A disintegrin and metalloproteinase with thrombospondin motifs; BLAST, basic local alignment search tool; BMP1; Bone morphogenetic protein 1; ECM, extracellular matrix; HIF, Hypoxia inducible factor; MMP, Matrix metalloproteinase; SPARC, Small protein acidic and rich in cysteine; TGM, Transglutaminase; TSA, Transcriptome shotgun assembly.
ABSTRACT
Animals (metazoans) include some of the most complex living organisms on Earth, with regard to their multicellularity, numbers of differentiated cell types and lifecycles. The metazoan extracellular matrix (ECM) is well-known to have major roles in the development of tissues during embryogenesis and in maintaining homeostasis throughout life, yet insight into the ECM proteins which may have contributed to the transition from unicellular eukaryotes to multicellular animals remains sparse. Recent phylogenetic studies place either ctenophores or poriferans as the closest modern relatives of the earliest-emerging metazoans. Here, we review the literature and representative genomic and transcriptomic databases for evidence of ECM and ECM-affiliated components known to be conserved in bilaterians, that are also present in ctenophores and/or poriferans. Whereas an extensive set of related proteins are identifiable in poriferans, there is a strikingly lack of conservation in ctenophores. From this perspective, much remains to be learnt about the composition of ctenophore mesoglea. The principal ECM-related proteins conserved between ctenophores, poriferans and bilaterians include collagen IV, laminin-like proteins, thrombospondin superfamily members, integrins, membrane-associated proteoglycans and tissue transglutaminase. These are candidates for a putative ancestral ECM that may have contributed to the emergence of the metazoans.

INTRODUCTION
Some of the most fascinating and mysterious steps in the evolution of life on earth involve the debut of multicellular organisms from single-celled ancestors. Modern multicellular lifeforms are present in both the bacterial and eukaryotic domains of life and there is evidence that multicellularity has emerged independently multiple times [1-5]. For eukaryotes, particularly the Metazoa (animals), the transition from unicellularity to multicellularity is complex to consider because ancestral forms are either not represented in the fossil record, or are extremely rare and difficult to identify. Genome and transcriptome sequencing projects offer new routes to consider these evolutionary transitions more systematically. For example, an analysis of
genomic data from many modern prokaryotes to identify commonalities in protein-coding sequences allowed inference of the possible repertoire of proteins in a last common ancestral cell [6].

In considering central attributes of multicellular organisms, the evolution of stable mechanisms for organised cell-to-cell attachments is a key requirement of multicellularity. The most complex modern multicellular organisms, including between 3 to about 122 different cell types, are found amongst the Metazoa [5]. The single-celled eukaryotes most closely related to animals are the choanoflagellates [7] and filastereans such as Capsaspora owczarzaki, [8,9]. From comparative studies of the transcriptomes and predicted proteomes of these protists, it has been deduced that the origin of metazoans probably involved the functional adaptation of pre-existing gene products, for example new or adapted roles of integrin and cadherin receptors, both of which have been identified in certain unicellular eukaryotes [10, 11], as well as genetic rearrangement events that led to the origin of new types of gene products with novel functional capacities for cell interactions and inter-cellular communications [12].

A central mediator of metazoan multicellularity is the extracellular matrix (ECM), a structured extracellular network of collagens, glycoproteins, proteoglycans and associated carbohydrates such as glycosaminoglycans. The secreted proteins that build the ECM appear to fall within the novel category of metazoan gene products, because many ECM proteins are highly-conserved throughout animals and yet are not represented in choanoflagellates or filastereans [13, 14]. Williams et al. [13] established that three species of these protists express distinct sets of predicted secreted proteins (identified by presence of a N-terminal secretory signal peptide and no transmembrane domain), none of which have a domain composition equivalent to a metazoan ECM protein, although individual domains common in ECM proteins are present. Specific examples are the separate domains of fibrillar collagens [15] or thrombospondins [16]. These data suggest that gene rearrangement and domain shuffling had an important role in the emergence of the large, multidomain, secreted proteins that characterise the metazoan
ECM. In modern metazoans, the secretion and extracellular assembly of structural proteins of the ECM depends on many ECM-affiliated proteins, both intracellular and extracellular: for example, to effect post-translational modifications, proteolytic processing, or interactions with non-structural and matricellular proteins within the ECM [17, 18]. Thus, consideration of the phylogeny of these affiliated proteins is also relevant to constructing models for the evolution of metazoan ECM.

The distinct domain architectures of secreted proteins of choanoflagellates and metazoans suggest that additional insights into metazoan ECM evolution could be ascertained from careful comparative analysis of ECM and affiliated proteins encoded in modern species from the earliest-diverging metazoan phyla. By analogy with the analysis of a prokaryotic ancestral cell [6], proteins in common between extant species in the earliest-diverging phyla would be candidates for membership of “ancestral ECM”. Of the early-diverging metazoan phyla (Ctenophora (comb jellies), Porifera (sponges), Placozoa and Cnidaria), cnidarians are, to date, by far the most-studied with regard to their ECM and cell-adhesion mechanisms. This relates to the lengthy history of Hydra as an experimental model [18], the phylogenetic position of cnidarians as the sister group to bilaterians, and the presence of a morphologically well-defined ECM, the mesoglea, that, in Hydra, can be isolated away from the cell layers of the body wall as an acellular structure and is thus suitable for biochemical study [19, 20].

Knowledge of the ECM of the other early-diverging phyla is much more sparse. Placozoa comprise an enigmatic phylum that to date includes only a few species and will not be considered further here [21]. A mesoglea between the epithelial cell layers is apparent in ctenophores, but information on its molecular composition is very limited (discussed further below). Sponges have bio-mineralised extracellular structures (spicules) embedded in a fibrillar meshwork, but most classes lack ECM as recognised morphologically in bilaterians. A limitation for the study of ctenophores and sponges until recent years has been a lack of laboratory model species or cell culture [22, 23].
The limited knowledge is significant because both Porifera and Ctenophora are considered to be of earlier evolutionary origin than Cnidaria, as evidenced by the fossil record and molecular phylogeny reconstructions [24, 25]. However, there remains considerable controversy over whether the sponges or the ctenophores are of earliest evolutionary origin, i.e., which phylum represents the sister group to all other animals. Historically, sponges were placed at the base of the animal evolutionary tree due to their simple morphological organisation, limited number of cell types, and the absence of recognisable nerve or muscle structures: both of the latter are present in ctenophores and cnidarians [26] (Fig. 1A). With the expansion of molecular phylogenetics, several studies have surprisingly placed ctenophores as the sister group to all other animals [e.g., 27] (Fig. 1B), whereas others continue to support the traditional “poriferans-sister” model [e.g, 28, 29]. Genome sequencing of two species of ctenophores indicated major differences in the categories of encoded proteins in comparison to all other metazoans, with many proteins of bilaterians noted to be absent [30, 31, 32]. Thus the “ctenophore first” hypothesis remains under active discussion and investigation [33, 34].

The advent of genomic and transcriptomic sequencing projects for an increasing number of poriferan and ctenophore species has revolutionised the possibility to gain insight into ECM content in sponges and ctenophores, through analysis of the predicted proteomes of individual species. This article will review the published literature and discuss our findings from a recent detailed survey of public genomic and transcriptomic databases for ECM proteins in species representing three classes of sponges and three species of ctenophores.

**Known Components of ECM in Ctenophores and Poriferans**

*Ctenophores.* The unique anatomy and ultrastructure of ctenophores has been studied by light and electron microscopy, with major interest in developmental processes, nerve and muscle tissues, and the specialised rows of locomotary ciliated combs (Fig. 1C, 1D) [35, 36]. Prey capture is carried out by specific colloblast cells on extensible tentacles (most species)
or by direct engulfment (*Beroe* species, that lack tentacles) [27, 37]. The mesoglea is typically described as transparent and jelly-like. From transmission electron microscopy studies of the tentacles of *Euplokamis*, the mesoglea was observed to contain networks of striated fibrils, interpreted as collagen fibrils, as well as muscle fibres, mesenchymal cells and a network of nerve cells. Curious box-like, acellular, extracellular structures were also observed [38]. Later immunofluorescent staining studies of *Pleurobrachia* species or *Beroe abyssicola* also identified many cell types within the bodywall mesoglea, including networks of nerve cells, muscle and other cell types [39-41]. Transmission electron microscopy has also provided views of a basement membrane-like layer that underlies ectodermal cells in *Pleurobrachia bachei* and *B. ovata*, but is not visible in *Mnemiopsis leidyi* [14]. There is very little direct knowledge of the composition of ctenophore mesoglea, but a phylogenetic study of the basement membrane proteoglycan, perlecan, concluded that perlecan is absent from *M. leidyi* [42]. Fidler et al. [14] detected collagen IV by immunohistochemistry as diffuse arrays in proximity to ectodermal cells in *M. leidyi*, and with more appearance of linear elements in *Beroe* and *Pleurobrachia*. Genomic and transcriptomic analyses identified many collagen IV paralogues, whereas small collagenous proteins (designated spongins from their initial identification in *Porifera*) were absent and a unique type of secreted protein, containing only a non-collagenous (NC) domain, was identified and designated NC1 protein [14].

*Porifera*. Adult sponges are sessile, vase-shaped animals with pores that filter water into the body cavity for food uptake by specialised choanocytes (Fig. 1E) [43, 44]. The body wall consists of an epithelial bilayer supported by mineralised spicules and a meshwork of extracellular fibrils termed the mesohyl (Fig. 1F). Unlike ctenophores, overt cell-cell junctions are apparent between epithelial cells [45]. There are four extant classes of sponges (Fig. 1G) and different classes have different processes of biomineralisation. In calcereous sponges, calcium carbonate-based spicules are assembled extracellularly through carbonic anhydrase activity and possibly in association with acidic extracellular proteins [46-48]. In siliceous sponges, silicon dioxide spicules are templated through intracellular and extracellular processes
involving (in many species) the polymerising enzyme, silicatein, and frequently with templating onto collagen fibres [49-51]. Silicateins are related to the cathepsin family of intracellular processing and degrading enzymes and are thought to have arisen in the sponge lineage through ancestral gene duplication and point mutation of cathepsin L [52]. Chitin has also been identified as a spicule-associated, possible template in demosponges and a glass sponge [53, 54].

By electron microscopy, class Homoscleromorpha is distinguished by the presence of a basement membrane structure [55, 56]. Indeed, a collagen IV cloned from a homoscleromorph sponge was shown to have a basement membrane-like localisation [55]. In addition, sponges (along with various other invertebrates) encode short-chain spongins that contribute to 10nm microfibrils within the mesohyl. Spongins comprise around 79-100 Gly-Xaa-Yaa triplets and 3 noncollagenous regions, with the C-terminal non-collagenous regions having homology and a proposed shared evolutionary origin with the NC1 domain of collagen IV [57, 58]. Collagen fibril structures have been identified by ultrastructural criteria in several sponge species [59-61] and molecular cloning led the recognition of a diversity of molecular forms of collagens of the fibrillar or interrupted-triple-helix types in addition to spongins ([62-64] and reviewed by [65]).

Species-specific, carbohydrate-based cell-to-cell adhesion mechanisms have also been characterized functionally and biochemically in sponges. Sulphated polysaccharides have major roles in species-specific cell-to-cell adhesion, in conjunction with proteins termed aggregation factor complex, spongican or glyconectins [66-71]. These proteins can adopt unusual, ring-like conformations and appear to be specific to the demosponge lineage [72, 73]. Other mechanisms may involve C-type lectins and a calcium-dependent lectin, clathrilectin [74, 75], as well as self-association of carbohydrates [76]. Chemically, the sulphated polysaccharides appear very varied, with varying amounts of sulphated galactose, fucose, arabinose or hexuronic acid identified across species [77]. Examination of the structures of the acid-labile carbohydrates of glyconectins from several species identified these to
include heterogeneous sulphated oligosaccharides with variable amounts of fucose, arabinose, or py(4,6)Galactose residues, and thus distinct from the repeated disaccharide units of bilaterian glycosaminoglycans [78].

With regards to other mechanisms of cell-ECM associations, integrin subunits have been cloned from several sponge species [79, 80, 81]. Integrin(s) have been implicated functionally in autograft fusion in Geodia cydonium [82] and the response of Microciona prolifera cells to depletion of extracellular sulphate [83]. The identification of vinculin in Oscarella pearsei and the localisation of this protein to cell-cell and cell-ECM contact sites further supports that poriferan integrins are likely to function in adhesion and cell signaling, as in bilaterians [84]. Dystroglycan-like proteins with possible laminin-binding capacity have been recognised in several sponges in addition to the dystroglycans of cnidarians and bilaterians [85].

**Insights from Genomics and Transcriptomics: A Structured Survey of ECM Proteins Encoded in Poriferans and Ctenophores**

The sequencing of the genome of the demosponge Amphimedon queenslandica expanded the view on candidate ECM proteins of sponges. Analysis of the predicted proteins indicated that, even in the absence of overt basement membrane-like structures, laminin-like proteins are encoded [86, 87]. Since 2010, transcriptomes for sponges of other classes and genomes and transcriptomes for several species of ctenophores have been published [31, 32, 88, 89].

To obtain a wider view of ECM and ECM-associated proteins in ctenophores and poriferans, we surveyed sponge and ctenophore genome- and/or transcriptome-predicted proteins for selected ECM and ECM-affiliated proteins. The ECM proteins chosen for study are highly conserved in invertebrate and vertebrate bilaterians [90, 91] and have known functional roles in the fibrils and meshworks of the ECM. A range of collagens were included as search tools to assist identification of possible disparate forms. Major glycoprotein and proteoglycan receptors that tether ECM proteins at cell surfaces and extracellular proteases important for ECM dynamics in
bilaterians were also included (Fig. 2A), along with intracellular proteins that are important for procollagen assembly, processing and collagen fibril formation (Fig. 2B) [92], or for the post-translational assembly of the core tetrasaccharide linker for glycosaminoglycan substitution on proteoglycan core proteins [93] (Fig. 2C). Spongin and silicatein are not present in vertebrates but were included because of their known importance in poriferan ECM. The suite of 41 proteins analysed is listed in Supplementary Table 1. The sponges species included represent the classes Demospongiae (the most abundant in terms of species), Homoscleromorpha and Calcarea and the three species of ctenophores have sequenced genomes and/or a large transcriptome dataset.

From initial BLASTP or TBLASTN searches, protein sequences of highest homology (determined by evalue <1e-30 and extensive % coverage of the query sequence) were compiled and validated by best-reciprocal BLAST search. Sequences passing this step were subjected to additional quality control analysis for secretory signal peptides, transmembrane domains, overall domain architecture and, where relevant, enzyme active site. The results obtained are summarised in Tables 1 and 2, respectively. The underlying data are presented in Supplementary Tables 2-8.

Collectively, the data demonstrate dramatic differences in the profile of conserved proteins in ctenophores versus poriferans (Fig. 3, Table 1). Many more proteins in common with bilaterian ECM are encoded in poriferans than in ctenophores. Nevertheless, the ctenophore list does include a repertoire for a basic cell-ECM adhesion system: cell-surface receptors, ECM proteins, a cross-linking enzyme and a potential ECM-proteolytic enzyme (Fig. 3). The proteins identified in ctenophores were for the most part present in all three species, with the exception of syndecan, identified only in P. bachei, and potential matrix metalloproteases, identified in M. leidyi and H. californensis. In agreement with Fidler et al. [14], many collagen IV-like paralogues were identified. Post-translational modifications of proline and lysine residues contribute to the stability of collagen triple helices in vertebrates, however only prolyl-4-hydroxylase and not pro-collagen lysine dioxygenase (lysine
hydroxylase) was identified in these ctenophores (Table 1, Fig. 3). The encoding of multiple integrin alpha and beta subunits (Table 1) indicates potential for diverse specificities of integrin-mediated cell adhesion, perhaps in line with the relatively large number of cell types now documented in ctenophores [94]. Silicatein-like proteins were identified in all three species; however, given the recognised general divergence of protein sequences in ctenophores [31, 32], in-depth studies will be needed to determine the relationship of these to the cathepsin family (Table 1).

We examined the laminin-like proteins in more detail in view of the early evolution of collagen IV [14] and the central role of laminin in basement membrane assembly in bilaterians [95]. With the caveat that some of the identified sequences are incomplete, the laminin subunits identified present a complex picture. Although all are large proteins with many of the characteristic domains of laminins, many variations in domain organisation are apparent, including atypical domains such as thrombospondin or fibronectin III domains. Overall, the laminin proteins of sponges are more similar to those of bilaterians, yet distinctions between beta and gamma subunits are blurred at both the sequence and domain levels. Notably, the laminin N-domain is lacking from (apparently full-length) ctenophore proteins and two alpha-like subunits but no beta- or gamma-like subunits were identified in P. bachei (Fig. 4). The numbers of alpha-like and beta/gamma-like subunits varied between species and the alpha-like subunits included at most three laminin-G domains. Biochemical experiments will be needed to determine if these proteins are capable of forming stable heterotrimers and undergoing extracellular polymerisation or integrin-binding. For the ctenophore proteins, it is of interest whether heteromers including two different alpha subunits can be assembled.

We also examined the tissue transglutaminase 2- (TGM2) like proteins. In vertebrates, TGM2 has major roles in cross-linking fibrillin or fibronectin molecules into extracellular fibrils via redox-dependent processes [96, 97]. The TGM-like proteins identified in the ctenophores and A. queenslandica sponge each include all the major domains of TGM2 but have only 30%-35% sequence identity to human TGM2. However, at the active site of TTG2, the
identity is 65%-80% and the cysteine residue is completely conserved (Fig 5A). Molecular models of the active site region for four of the sequences, constructed against secondary structure alignments of four TGM2 structures from Protein database (4KTY_B, 1G0D_A, 1LM9_A and 2Q3Z_A), are presented overlaid with the structure of human TGM2 (4PYG) that was not used for modelling (Fig. 5B, 5C). The models demonstrate that residues within 60nm of the active site cysteine align very well with the known structure (Fig. 5B), as do the highly-conserved residues at the active site (Fig. 5C). We predict that the ctenophore and sponge proteins should be active transglutaminases.

ECM-related proteins not found in the ctenophores raise other intriguing questions about the biochemical nature of ctenophore mesoglea. It was previously noted that core proteins of secreted proteoglycans of bilaterians are not conserved in early-diverging metazoans [19, 90]. However, carbohydrate has been reported as <1% of dry weight of ctenophores [98]. With the exception of glucuronyl-transferase, homologues of the bilaterian enzymes for addition of the core O-linked saccharides (Fig. 2C), are not identifiable (Table 1, Fig 3). Whether this pathway mechanism evolved later or has been lost through lineage-specific gene losses in ctenophores is unclear.

With regard to ECM structure, no fibrillar-like collagens were identified and the collagen cross-linking enzyme lysyl oxidase was also absent, raising questions over the nature of observed striated fibrils in ctenophore mesoglea [38]. However, bone morphogenetic protein 1 (BMP1), which cleaves the C-propeptide of fibrillar procollagen [99] (Fig. 2B), was present. In bilaterians, BMP1 has many other substrates including a laminin gamma chain [99] and it may be expected that the ctenophore protein can target other substrates. In agreement with [14], spongin was not identified.

In contrast, poriferans were confirmed to encode a wider repertoire of ECM proteins including fibrillar-like collagens of various domain architectures and fibrillin, as well as SPARC and one or more thrombospondin superfamily members (see [16] for details of the thrombospondin superfamily). Collagen IV proteins were apparent in the calcerous sponges as well as (as expected)
in the homoscleromorph sponge. The encoding of lysyl oxidase is in agreement with the detection of striated collagen fibrils in sponges [59-61]. Nonwithstanding the unusual carbohydrate structures reported in sponges (see section above), the suite of carbohydrate-addition enzymes encoded indicate potential for addition of the O-linked core tetrasaccharide of glycosaminoglycans (Table 2, Fig. 3). Other aspects of ECM in common with bilaterian ECM include the encoding of A Disintegrin and Metalloproteinase with Thrombospondin motifs (ADAMTS) proteases and a wider repertoire of cell-ECM attachment receptor types, including membrane-bound proteoglycan core proteins, dystroglycan and integrin subunits. Spongins were identified in A. queenslandica and O. carmela, but not in the calcerous sponges. Silecateins were identified as expected in the demosponge A. queenslandica and also in the other species examined (Table 3).

Perspective

Current laboratory experiments and analyses of genome-predicted proteins indicate that ctenophore ECM has a very different protein composition to other metazoans. This cannot be interpreted as a result of the early phylogenetic emergence of this phylum because poriferans, traditionally considered the earliest-diverging metazoans, are found to have an array of ECM and ECM-affiliated proteins that is clearly closer to the conserved repertoire of cnidarians and bilaterians. The difference in ECM might be considered an indication that ctenophores evolved prior to poriferans, in which case the limited set of proteins conserved between ctenophores, poriferans and bilaterians can be taken to represent a prototypic “toolkit” for a minimal metazoan ECM [100]. The combination of collagen IV, laminin-like proteins and thrombospondin superfamily members is of great interest, as the concept of coordinated function of these three proteins within ECM has received little consideration.

However, other factors also need to be taken into consideration. There are estimated to be around 5,000 species of extant poriferans, yet only about 150 known species of ctenophores. This may reflect that the deep-sea lifestyle of many ctenophores makes it difficult to identify the true number of species, or
alternatively could indicate very different evolutionary histories of poriferans and ctenophores. Sponges and ctenophores evolved when oxygen levels on Earth were far lower than at present [101]. Indeed, members of both phyla lack hypoxia-inducible factor α (HIF) indicating that oxygen availability does not drive gene expression through the HIF pathway as in cnidarians and bilaterians [102]. In a “poriferan-first” evolutionary scenario, the limited repertoire of known ECM proteins in ctenophores would represent secondary gene losses, leading to a proposal of a relatively complex ECM in the metazoan ancestor. Many ctenophores live in the deep sea and the environment of low oxygen and sparse food sources [101, 103], and high hydrostatic pressure may have driven selection for a unique form of ECM. Clearly, the anatomy of ctenophores does include a mesoglea and, to date, to our knowledge, an unbiased study of ctenophore mesoglea by proteomic methods has not been carried out. Only through this type of approach will a clear view of ctenophore mesoglea composition be gained. A limitation of focusing on the ECM proteins conserved with bilaterians is that possible ctenophore- or poriferan-specific ECM proteins remain undisclosed. As discussed by others, it is very likely that a considerable “hidden biology” of ctenophores and poriferans remains to be discovered [104]. Nevertheless, the positive identification of certain ECM proteins conserved between these early-emerging phyla and other metazoans increases the precision of models for an ancestral metazoan ECM.

**SUMMARY POINTS**

1. Ctenophores and poriferans have distinct sets of ECM-related proteins in comparison to the most highly-conserved ECM and ECM-affiliated proteins of bilaterians.
2. In particular, ctenophores lack many of the structural ECM proteins and enzymes for addition of the core O-linked tetrasaccharide that is characteristic of bilaterian glycosaminoglycan substitutions.
3. Proteomic studies are needed to gain more comprehensive views of the composition of ECM in ctenophores and poriferans.
4. The very limited set of conserved ECM and ECM-affiliated proteins in ctenophores may represent a minimal toolkit that reflects the earliest emerging form of metazoan ECM. Given the current controversies over the relative phylogenetic placements of ctenophores and poriferans it cannot be excluded that this apparent simplicity may result from
secondary gene losses that are specific to the evolution of the ctenophore lineage.

FIGURES AND TABLES

Figure 1. Ctenophores and Poriferans. A, B, Schematic evolutionary trees of the poriferan-first (A), or ctenophore-first (B), models of metazoan evolution. Choanoflagellates are included as the closest outgroup. C, Morphology of a ctenophore, Pleurobrachia pileus. Internal organs are visible through the transparent mesoglea (asterisk). Key: Mouth (mo); aboral neuro-sensory complex (nsc); comb rows (cr); tentacles (tcl); lateral branches of tentacles (tentillae), (tt); tentacle root (tr), and pharynx (ph). Scale bar: 1cm. (reproduced from [105] under CC-BY). D, Schematic diagram of a ctenophore. Mesoglea is present within the tentacles and body wall (reproduced from [106] under CC-BY). E, Morphology of a demosponge, Echinoclathria gibbosa (reproduced from [107] under CC-BY). F, Diagram of the body wall and cell types of a sponge, (reproduced from [108]; permission in progress). G, Tree diagram of the four classes of extant sponges.

Figure 2. The cell-ECM adhesion proteins included in our genomic and transcriptomic survey. A, Schematic diagram of molecular processes involved in cell-ECM adhesion; B, The major enzymes involved in procollagen processing and collagen fibril assembly; C, The enzymes involved in O-linkage of the core tetrasaccharide for glycosaminoglycan addition.

Figure 3. ECM and ECM-affiliated proteins of poriferans and ctenophores that are conserved with bilaterians, as identified by our survey. The Venn diagram includes all categories of proteins that were identified in at least one of the species studied. See Tables 1 and 2 and supplementary Tables for details of the protein identifications.

Figure 4. Domain architectures of the laminin-like proteins identified in the species studied. Domains were identified in InterProScan 5. Representative examples of human laminin α, β, and γ chains are shown at the top. The number of amino acids is given in small font at the C-terminus of each protein. Dashed lines indicate incomplete sequence. Accession numbers are given on the righthand side, see Tables 1 and 2 for details. The IV-B domain (in coral)
is specific to the β subunit. Key: SP, signal peptide; N, laminin N-terminal domain; EGF, epidermal growth factor-like domain, with numbers referring to the number of repeated domains; IV, laminin domain IV (in blue); G, laminin G domain; FNIII, fibronectin type III domain; GFR, growth factor-cysteine-rich domain; TSR, thrombospondin type 1 domain. The G domain is structurally related to the Concanavalin A domain (Con A), which was identified in some sequences. Not to scale.

Figure 5. Molecular modelling of the active sites of the predicted tissue transglutaminases identified in ctenophores and a sponge. A, CLUSTAL sequence alignment of the active site of human TGM2 (Hs) with the homologous regions of the identified ctenophore (Ml, *M. leidyi*; Pb, *P. bachei*, and poriferan (Aq, *A. queenslandica*) TGMs. See Tables 1 and 2 for full accession numbers. B, Structure of the active site of 4PYG.pdb (human TGM2) [96]. The helix is highlighted in yellow as orientation for the overlay models in Fig. 5C. C, Models were prepared by HHpred and MODELLER [110] and are shown as Aq (pink), ML03126a (green), ML25826a (salmon) and Pb3462531 (silver) overlaid with the crystal structure of human TGM2 from 4PYG.pdb, (black, with the catalytic cysteine (C277) labelled). The overlays show high conservation of the sidechains of residues within a 6 Angstrom radius around the catalytic cysteine.

**TABLES**

*Table 1.* The ECM-related proteins conserved with bilaterians identified from the ctenophore species studied.

*Table 2.* The ECM-related proteins conserved with bilaterians identified from the sponge species studied.

**SUPPLEMENTARY DATA**

*Supplementary Table 1.* The ECM and ECM-affiliated proteins of bilaterians included in our survey.

*Supplementary Tables 2-8.* Summary of BLAST top hits and validations for the 7 species surveyed.
REFERENCES


<table>
<thead>
<tr>
<th>ECM/ECM-affiliated Protein</th>
<th>Accession numbers of homologues identified in ctenophore species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M. leidyi</td>
</tr>
<tr>
<td>ADAMTS</td>
<td>-</td>
</tr>
<tr>
<td>Agrin</td>
<td>-</td>
</tr>
<tr>
<td>beta-1,4-galactosyltransferase 1</td>
<td>-</td>
</tr>
<tr>
<td>beta-1,4-galactosyltransferase 2</td>
<td>-</td>
</tr>
<tr>
<td>Bone morphogenetic protein 1</td>
<td>ML31401a</td>
</tr>
<tr>
<td>Bifunctional heparan sulphate N-deacetylase/N-sulphotransferase</td>
<td>-</td>
</tr>
<tr>
<td>Collagen I</td>
<td>-</td>
</tr>
<tr>
<td>Collagen IV</td>
<td>ML034334a</td>
</tr>
<tr>
<td></td>
<td>ML18175a</td>
</tr>
<tr>
<td></td>
<td>ML17501a</td>
</tr>
<tr>
<td></td>
<td>ML18363a</td>
</tr>
<tr>
<td></td>
<td>ML18198a</td>
</tr>
<tr>
<td></td>
<td>ML16441a</td>
</tr>
<tr>
<td></td>
<td>ML25069a</td>
</tr>
<tr>
<td></td>
<td>ML18197a</td>
</tr>
<tr>
<td></td>
<td>ML18176a</td>
</tr>
<tr>
<td></td>
<td>ML17504a</td>
</tr>
<tr>
<td></td>
<td>ML17502a</td>
</tr>
<tr>
<td></td>
<td>ML034335a</td>
</tr>
<tr>
<td>Collagen VI</td>
<td>-</td>
</tr>
<tr>
<td>Collagen X</td>
<td>-</td>
</tr>
<tr>
<td>Collagen XIV</td>
<td>-</td>
</tr>
<tr>
<td>Collagen XVII</td>
<td>-</td>
</tr>
<tr>
<td>Collagen XXIII</td>
<td>-</td>
</tr>
<tr>
<td>Collagen XXV</td>
<td>-</td>
</tr>
<tr>
<td>Dystroglycan</td>
<td>-</td>
</tr>
<tr>
<td>Fibrillin</td>
<td>-</td>
</tr>
<tr>
<td>Fibulin</td>
<td>-</td>
</tr>
<tr>
<td>Glucuronyl-transferase</td>
<td>ML271519a</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Heparan sulphate 2-O-sulphotransferase</td>
<td>-</td>
</tr>
<tr>
<td>Hsp47 (SERPINH1)</td>
<td>-</td>
</tr>
<tr>
<td>Integrin alpha</td>
<td>ML09951a</td>
</tr>
<tr>
<td></td>
<td>ML030811a</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Integrin beta</td>
<td>ML068314a</td>
</tr>
<tr>
<td></td>
<td>ML04042a</td>
</tr>
<tr>
<td></td>
<td>ML05951a</td>
</tr>
<tr>
<td>Laminin alpha</td>
<td>ML097515a</td>
</tr>
<tr>
<td>Laminin beta</td>
<td>-</td>
</tr>
<tr>
<td>Laminin gamma</td>
<td>ML05951a</td>
</tr>
<tr>
<td>Lysyl hydroxylase</td>
<td>-</td>
</tr>
<tr>
<td>Lysyl Oxidase</td>
<td>-</td>
</tr>
<tr>
<td>Matrix metalloproteinase</td>
<td>ML030215a</td>
</tr>
<tr>
<td>Nidogen</td>
<td>-</td>
</tr>
<tr>
<td>Perlecan</td>
<td>-</td>
</tr>
<tr>
<td>Procollagen C-proteinase enhancer</td>
<td>-</td>
</tr>
<tr>
<td>Prolyl 4-hydroxylase</td>
<td>GFA01094412.1, ML257632</td>
</tr>
<tr>
<td>Silicatein</td>
<td>ML017310a</td>
</tr>
<tr>
<td></td>
<td>ML10109a</td>
</tr>
<tr>
<td></td>
<td>ML007332a</td>
</tr>
<tr>
<td></td>
<td>PB3465652</td>
</tr>
<tr>
<td>SPARC</td>
<td>-</td>
</tr>
<tr>
<td>Spongion</td>
<td>-</td>
</tr>
<tr>
<td>Syndecan</td>
<td>-</td>
</tr>
<tr>
<td>Thrombospondin superfamily</td>
<td>ML34222a</td>
</tr>
</tbody>
</table>
Table 1. Homologues of ECM and affiliated proteins of bilaterians identified in three species of ctenophores. Accession numbers are from *M. leidyi* genome browser [31], *P. bachei* transcriptome at Neurobase [32], and GenBank TSA for the *H. californiensis* transcriptome [109]. Underlying BLAST results are reported in Supplementary Tables 2-4. Key. p = partial sequence. + BLASTP for silicatein in *H.californiensis* yielded >100 hits with many redundant sequences. Four top hits are included here.

<table>
<thead>
<tr>
<th></th>
<th>Accession Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue Transglutaminase</td>
<td>ML03126a</td>
</tr>
<tr>
<td></td>
<td>PB 3462531</td>
</tr>
<tr>
<td></td>
<td>GGLO01036061</td>
</tr>
<tr>
<td>Xylosyltransferase</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>ECM / ECM-affiliated protein</td>
<td>Accession numbers of top homologues identified in poriferan species</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td><strong>O. carmella</strong></td>
</tr>
<tr>
<td></td>
<td>g302.1.t, g3943.11, g5893.1</td>
</tr>
<tr>
<td></td>
<td>g1015.11</td>
</tr>
<tr>
<td></td>
<td>g1015.11</td>
</tr>
<tr>
<td></td>
<td>XP_00383478.3</td>
</tr>
<tr>
<td></td>
<td>g11091</td>
</tr>
<tr>
<td>Bone morphogenetic protein 1</td>
<td>g3280.1.t</td>
</tr>
<tr>
<td></td>
<td>XP_019854532.1</td>
</tr>
<tr>
<td></td>
<td>g9342.11</td>
</tr>
<tr>
<td></td>
<td>g11085.11</td>
</tr>
<tr>
<td></td>
<td>g426.11, g10272.11</td>
</tr>
<tr>
<td>Collagen IV</td>
<td>g1079.11</td>
</tr>
<tr>
<td></td>
<td>g1097.11</td>
</tr>
<tr>
<td></td>
<td>XP_019850772.1</td>
</tr>
<tr>
<td></td>
<td>g2667.11</td>
</tr>
<tr>
<td>Fibrillin</td>
<td>g3633.11</td>
</tr>
<tr>
<td>Fibulin</td>
<td>-</td>
</tr>
<tr>
<td>Glypican</td>
<td>g1352.1</td>
</tr>
<tr>
<td>Glucuronyl-transferase</td>
<td>g2277.11</td>
</tr>
<tr>
<td>Heparan sulphate 2-O- sulphotransferase</td>
<td>g2593.1</td>
</tr>
<tr>
<td>Integrin alpha</td>
<td>g1989.11</td>
</tr>
<tr>
<td></td>
<td>g9424.11</td>
</tr>
<tr>
<td></td>
<td>g551.11</td>
</tr>
<tr>
<td></td>
<td>XP_01986402.1</td>
</tr>
<tr>
<td></td>
<td>XP_01986540.1</td>
</tr>
<tr>
<td></td>
<td>XP_0198654508.1</td>
</tr>
<tr>
<td></td>
<td>XP_0198654533.1</td>
</tr>
<tr>
<td></td>
<td>XP_019865507.1</td>
</tr>
<tr>
<td></td>
<td>g9310.1</td>
</tr>
<tr>
<td></td>
<td>g9748.1</td>
</tr>
<tr>
<td>Laminin alpha</td>
<td>g7319.1</td>
</tr>
<tr>
<td>Laminin beta</td>
<td>-</td>
</tr>
<tr>
<td>Laminin gamma</td>
<td>g7968.1</td>
</tr>
<tr>
<td>Lysyl hydroxylase</td>
<td>g1142.1</td>
</tr>
</tbody>
</table>
Table 2. Homologues of ECM and affiliated proteins of bilaterians identified in four species of poriferans. Accession numbers are from genomes and transcriptome-predicted proteins of *O. carmela*, *S. ciliatum* and *L. complicata* [88] at Compagen [89], and GenBank entries of the *A. queenslandica* genome and transcriptome projects [86, 87]. Apart from collagen IV-like collagens and spongins, the various collagens identified have been grouped as fibrillar-like collagens. Underlying BLAST results and validations are reported in Supplementary Tables 5-8.
Figure 1

A
Choanoflagellata  Porifera  Ctenophora  Placozoa  Cnidaria  Bilateria  Metazoa

B
Choanoflagellata  Ctenophora  Porifera  Placozoa  Cnidaria  Bilateria  Metazoa

C

D

E
Hexactinellida  Demospongiae  Calcarea  Homoscleromorpha

F

G

Figure 3

Porifera

Ctenophora

MMP

- BMP1
- Collagen IV
- Glypican
- Glucuronil-transferase
- Integrin α, β
- Laminin chains
- Prolyl-4-hydroxylase
- Silicatein
- Syndecan
- Thrombospondin superfamily
- Tissue transglutaminase

ADAMTS
- β1,4 galactosyltransferase
- Collagens (Fibrillar)
- Collagen XXIII-like
- Dystroglycan
- Fibrillan
- HS-deacetylase
- HS 2-O-sulphotransferase
- Lysyl hydroxylase
- Lysyl oxidase
- SPARC
- Spong
- Spong
- Xylosyl-transferase
Figure 4

Human

- **N**: EGF x 4
- **IV**: EGF x 9
- **EGF x 4**: K1k2

Ctenophores

- **H. californiensis**
  - EGF x 4
  - EGF x 5
  - GFR x 2

- **M. leidy**
  - EGF x 2
  - EGF x 4
  - EGF x 5
  - EGF x 7

- **P. bachei**
  - EGF x 3
  - EGF x 5
  - EGF x 2

Poriferans

- **A. queenslandica** (Demospongia)
  - EGF x 11
  - FNIII

- **L. complicata** (Calcarea)
  - EGF x 18

- **S. ciliatum** (Calcarea)
  - EGF x 5
  - IV-B

- **O. carmela** (Homoscleromorpha)
  - EGF x 5