The evolution of metazoan extracellular matrix

Richard O. Hynes

Howard Hughes Medical Institute, Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139

The modular domain structure of extracellular matrix (ECM) proteins and their genes has allowed extensive exon/domain shuffling during evolution to generate hundreds of ECM proteins. Many of these arose early during metazoan evolution and have been highly conserved ever since. Others have undergone duplication and divergence during evolution, and novel combinations of domains have evolved to generate new ECM proteins, particularly in the vertebrate lineage. The recent sequencing of several genomes has revealed many details of this conservation and evolution of ECM proteins to serve diverse functions in metazoa.

Introduction

ECM proteins are typically composed of multiple protein domains, and their gene structures were some of the first recognized to have arisen by exon shuffling (Engel, 1996; Patthy, 1999; Hohenester and Engel, 2002). Biochemical analyses of ECM proteins began in vertebrates. However, as cDNA and genomic sequences became available, it became increasingly evident that many ECM genes such as collagens and laminins are very ancient, and, in the last decade, as genomic sequences were determined for many metazoa, it was recognized that many ECM-encoding genes originated early in metazoan evolution. In particular, genomes of diverse bilaterian organisms (mammals, flies, worms, sea urchins, and ascidians) revealed a common set of ECM proteins shared by all bilateria (Hynes and Zhao, 2000; Whittaker et al., 2006; Huxley-Jones et al., 2007), which is consistent with the presence of common ECM structures such as basement membranes in all these organisms. Most recently, genomes of nonbilaterian eumetazoa and basal metazoa (see Box 1 for a summary of metazoan phylogeny), as well as uncellular relatives of metazoa, have allowed investigation of the origins of this common set of ECM proteins. Furthermore, the increasing amount of genomic information has allowed investigation of the elaboration, diversification, and specialization of ECM proteins in different evolutionary lineages to subserve differing functional roles. In this brief review, I will summarize our current understanding of the diversity and evolution of ECM proteins and attempt to relate them to the evolution of multicellularity and the subsequent evolution of metazoa.

Major characteristics and categories of ECM proteins in metazoa

ECMs are, by definition, relatively or completely insoluble assemblies of proteins that form structures such as basement membranes, interstitial matrices, tendons, cartilage, bones, and teeth. The proteins that comprise these various ECMs are frequently large, with multiple characteristic domains specialized for protein interactions necessary for ECM assembly or for the recruitment of cells or other proteins (such as growth factors or cytokines) to the ECM (Hynes, 2009; Hynes and Naba, 2011; see Figs. 2 and 3 for illustration of domain structures). ECM proteins are frequently cross-linked by enzymatic and nonenzymatic reactions, further contributing to their insolubility. The large size, complexity, and insolubility of ECM proteins has made their analysis challenging, but the availability of complete genome sequences and their inferred complement of encoded proteins has made available reasonably reliable inventories of ECM proteins and allowed comparative analyses among species. These analyses have made clear that all bilaterian taxa share a common set of ECM proteins, with occasional examples of gene loss in certain lineages and many examples of taxon-specific elaborations based on the common set.

Basement membrane toolkit

Basement membranes are a characteristic feature of most metazoa, arguably an essential feature of tissue and epithelial organization, providing a locus for adhesion of epithelial cell layers and definition of basal-apical polarity of the cells (Fahey and Degnan, 2010). Studies, initially in vertebrates but more recently in invertebrates, have defined the major protein components of basement membranes (Fig. 2). All basement membranes are composed of a common set of interacting proteins (Yurchenco, 2011): a core network of cross-linked type IV collagen is associated with laminin (a trimer of related α, β, and γ subunits); nidogen, a laminin-binding glycoprotein; and perlecan, a very large and complex heparan sulfate proteoglycan. Strikingly, genes
Figure 1. Eukaryotic phylogeny and the appearance of ECM proteins. The figure diagrams the major taxonomic divisions of eukaryotes (Box 1), and is based on published phylogenetic analyses (e.g., Philippe et al., 2005, 2009; Shalchian-Tabrizi et al., 2008; Pick et al., 2010) as well as on analyses of whole genome sequences (see main text). Taxa for which whole genome sequences are available are marked with asterisks. The diagram represents topologies of relationships, and the branch lengths are not intended to reflect accurate evolutionary distances. The relationships of apusomonads and placozoa are not well defined; their placement in the diagram is influenced in part by evidence from analyses of ECM-related genes (see main text). Taxa with the complement of ECM proteins typical of all bilateria are within the light blue hexagon. Taxa highlighted with green or blue text have some homologues of bilaterian ECM proteins or their receptors (see main text), but lack a complete set. The complexity of the known ECM protein sets in these organisms increases from left to right. In contrast, the taxa within the gray trapezoid show no evidence of any creditable examples of metazoan ECM proteins. First known appearances of relevant genes/proteins or domains are marked in red. Note that taxa appearing to the right of (i.e., subsequent to) the origin of a given feature may have lost it; examples would be the absence of integrins in choanoflagellates or domains are marked in red. Note that taxa appearing to the right of (i.e., subsequent to) the origin of a given feature may have lost it; examples would be the absence of integrins in choanoflagellates sequenced to date and the absence of fibrillar collagens in Drosophila (in both cases, presumably by gene loss).

Box 1. Outline of metazoan phylogeny

Taxon: Any phylogenetic group, such as a phylum, class, genus, or species.

Clade: A group of organisms that are related by evolution and divergence.

Metazoa: Multicellular animals.

Eumetazoa: All metazoan animals apart from Porifera (sponges), Placozoa, and a few other obscure taxa. Within the eumetazoa there are two well defined clades of bilaterally symmetric animals: protostomes and deuterostomes, which are grouped together as Bilateria. Protostomes have two subdivisions: ecdysozoa, which include arthropods and nematodes; and lophotrochozoa, which include mollusks, annelids, flatworms, and others. Deuterostomes include echinoderms, hemichordates, and chordates. Eumetazoa also include two additional clades: ctenophores (comb jellies) and cnidaria (Hydrazoa, sea anemones, jelly fish, and the like), which are traditionally viewed as radially symmetrical (but see Martindale et al., 2002; Ball et al., 2004). They are sometimes classed together as Radiata or Coelenterata; however, the phylogenetic relationships between ctenophores and cnidaria are not certain and there are, as yet, no complete genomic sequences available for ctenophores. The major eumetazoan phylogenetic divisions are outlined in Fig. 1; they all arose before the Cambrian era, >540 million years ago.

All eumetazoan have epithelial layers showing apical-basal polarity and underlain by basement membranes. Bilateria have three germ layers (ectoderm, endoderm, and mesoderm), whereas Radiata have two epithelial layers with limited interstitial cells among and between them. Two metazoan phyla (sometimes grouped as Parazoa) are basal to the eumetazoa and lack any obvious axes of symmetry: the Placozoa, which are flat, bilayered organisms with a very limited number of cell types (approximately four) and no obvious basement membranes or ECM; and Porifera, or sponges, in which most cells lack epithelial organization. Most sponges lack basement membranes, but interstitial ECM is present. The exact evolutionary relationships between the Parazoa and Eumetazoa remain incompletely defined. Most phylogenetic analyses place Placozoa closer than Porifera to the Eumetazoa, as shown in Fig. 1 (but see Schierwater et al., 2009), and, as we will see, analyses of the complement of ECM proteins conform with this conclusion.

encoding this characteristic set of proteins, long-defined in vertebrates, were found in the genomes of two model protostomes, Caenorhabditis elegans (Hutter et al., 2000) and Drosophila melanogaster (Hynes and Zhao, 2000), when they were sequenced a little over a decade ago. The two homologous minor collagens XV and XVIII had also been observed to be associated with vertebrate basement membranes, although their functions were, and are, less clear. Genes encoding a collagen XV/XVIII orthologue were also found in both the fly and worm genomes. This set of 9–10 genes (2 laminin α, 1 laminin β, 1 laminin γ, 2 type IV collagen subunits, nidogen, perlecan, and 1–2 collagen XV/XVIII homologues; Fig. 2) has subsequently shown up in essentially every bilaterian genome sequenced, and we have called it the “basement membrane toolkit” (Whittaker et al., 2006). As is typical of most ECM proteins, the core constituents of basement membrane proteins are built from a set of well-defined protein domains (Fig. 2; Engel, 1996; Hohenester and Engel, 2002). This highly conserved set of genes has persisted in bilaterian genomes for well over half a billion years. This conservation indicates the essential nature of both this toolkit and the individual domains of its constituent proteins.

Fibrillar and other collagens

The most prevalent and earliest described collagens of vertebrates are those with long uninterrupted series of collagen repeats, typically ~1,000 amino acids long. They comprise multiple repeats of the tripeptide unit Gly-X-Y, where X is frequently proline and Y is often hydroxyproline. This repeating amino acid structure allows collagen
proteoglycans, around three dozen in mammals. Many of these
in bilateria and a few more examples are shown in Fig. 3 A. However, some of them are widespread
taxa, as we will discuss. However, some of them are widespread
taxa, as we will discuss. However, some of them are widespread
and vertebrate ECM proteins are restricted to later-evolving
different types (Figs. 2 and 3). Like the collagen repeats, these
domains are typically encoded by single exons or groups of
exons that have allowed shuffling during evolution of the ex-
onic units encoding these domains to build a large variety of
ECM proteins. Although the same domains can occur in many
different proteins, including both ECM and non-ECM proteins,
the domain composition, order, and number are characteristic
of individual ECM proteins; that is, they are defined by their
domain architectures. This is illustrated in Fig. 2, where the
laminin subunits are clearly related to each other and share
domains with nidogen and perlecan. Many of the mammalian
and vertebrate ECM proteins bind to the collagen network and to
nidogen. Perlecan, which is a complex heparan sulfate proteoglycan, is also incorporated into the basement membrane. Two other collagens, types XV and XVIII, are also associated with vertebrate
basement membranes, and an orthologue is present in all bilaterian clades. The high degree of conservation of this “toolkit” over more than half a billion
years testifies to the essentiality of basement membranes.

Figure 2. The basement membrane toolkit.
The figure shows the domain structures of the core proteins of metazoan basement
membranes, based on domain predictions largely using SMART and Pfam domain definitions.
All bilaterian clades encode this set of nine
proteins in their genomes. There are typically
two distinct laminin α subunits and one each of the laminin β and γ subunits. Each subunit has a characteristic domain organization. A laminin protomer is an αβγ trimer (shown at the top left) associated through coiled-coil domains in each subunit (red) and disulfide bonds [not depicted]. Type IV collagen is a
trimer of two homologous subunits, α1 and α2, usually adjacent in the genome in a head-
to-head arrangement with a single promoter between the two genes. Signature pairs of C4
domains lie at the C termini of all type IV col-
lagens, and the collagen segment (fuscha) is
interrupted, allowing flexibility. Type IV colla-
gen protomers associate through their N and
terminus and through disulfide bonding to
form a “chicken wire” network that provides
structural strength to the basement membrane.
Lamins bind to the collagen network and to
nidogen. Perlecan, which is a complex heparan sulfate proteoglycan, is also incorporated into the basement membrane. Two other collagens, types XV and XVIII, are also associated with vertebrate
basement membranes, and an orthologue is present in all bilaterian clades. The high degree of conservation of this “toolkit” over more than half a billion
years testifies to the essentiality of basement membranes.

subunits to assemble into triple-helical protomers. A primordial
exon (54 bp) encoding exactly six repeating Gly-X-Y tripeptides
underwent duplications and modifications (such as deletions and
fusions), always retaining the same phasing of introns, so that
exons encoding collagen repeat units can be assembled in varying
numbers and with other domains. In vertebrates, there are >40
collagen genes encoding diverse collagens (Ricard-Blum, 2011).
Mammalian fibrillar collagens (11 genes) have collagen repeats
flanked by characteristic noncollagenous domains at the N termi-

As mentioned previously, type IV collagens have a pre-
Cambrian origin. The same is true for fibrillar collagens. The fibril-
lar collagens assemble into the characteristic striated collagen
fibris of interstitial connective tissue matrices and provide structural
strength to those ECMs. As such, they play crucial roles in the
integration of multicellular organisms. Fibrillar collagens are
found in sponges, the most primitive metazoan phylum (Box 1
and Fig. 1). Three fibrillar collagen subclades (A, B, and C) arose
before the eumetazoan radiation and are widespread, although
not universal, in bilateria (Exposito et al., 2008, 2010; Heino
et al., 2009). For example, Drosophila lacks any fibrillar collagens,
which indicates the loss of the relevant genes in that lineage.

Proteoglycans
In addition to perlecan, vertebrate genomes encode many other
proteoglycans, around three dozen in mammals. Many of these
fall into two families (Merline et al., 2009; Schaefer and Schaefer
2010); one built of LRR domains and one, known as hyalectans,
containing N-terminal IgV and LINK domains and C-terminal
EGF-CLEC-CCP domain units, flanking a central section bearing
attached glycosaminoglycans. In addition, a small family of
glycoproteins named SPOCKs or testicans are related to the ECM
glycoproteins SPARC/osteonectin. The testicans, LRR repeat
proteoglycans, and hyalectans have been reported only in chor-
dates, and will be discussed later. Two membrane-bound fami-
lies of proteoglycans—syndecans and glypicans (Couchman,
2010)—like perlecan, are found throughout bilateria (Ozbek
et al., 2010).

ECM glycoproteins
Mammalian genomes encode around 200 further ECM glyco-
proteins distinct from collagens and proteoglycans (Hynes
and Naba, 2011; Naba et al., 2011). These ECM glycoproteins are
also built from characteristic arrays of domains of >50 dif-
f erent types (Figs. 2 and 3). Like the collagen repeats, these
domains are typically encoded by single exons or groups of
exons that have allowed shuffling during evolution of the ex-
onic units encoding these domains to build a large variety of
ECM proteins. Although the same domains can occur in many
different proteins, including both ECM and non-ECM proteins,
the domain composition, order, and number are characteristic
of individual ECM proteins; that is, they are defined by their
domain architectures. This is illustrated in Fig. 2, where the
laminin subunits are clearly related to each other and share
domains with nidogen and perlecan. Many of the mammalian
and vertebrate ECM proteins are restricted to later-evolving
taxa, as we will discuss. However, some of them are widespread
in bilateria and a few more examples are shown in Fig. 3 A.
These ancient ECM glycoproteins, like those of the basement
membrane toolkit (Fig. 2), have been subject to strong selection since the divergence of bilateria >600 million years ago and must have fundamental functions.

Challenges of ECM phylogeny
Analyses of the evolution of ECM proteins present some challenges. As discussed, ECM proteins are large and complex, with multiple domains, which they share both among themselves and with many other proteins. Domains such as EGF, LRR, FN3, and Ig are widespread in many proteins encoded by metazoan genomes and do not themselves define ECM proteins. Therefore, simple Basic Local Alignment Search Tool (BLAST) or domain searches yield multiple partial homologues for most ECM proteins and can be misleading if not supplemented by analyses of domain composition. It is the patterns or arrangements of domains that are diagnostic of specific ECM proteins. However, because the genes are large, with many exons, they are frequently incomplete or interrupted in current databases of genomes, ESTs, cDNAs, and inferred proteins. Therefore, gene predictions for ECM proteins are significantly harder than for many other genes. Thorough analyses require high-quality genomic or cDNA sequences and, often, further annotation to yield complete and reliable ECM protein predictions. This has only become possible fairly recently for many taxa, but there has been an explosion of genomic information in recent years that has shed light on the origins of ECM proteins and, indeed, of ECM itself. These data have allowed extension of the comparative genomics of ECM beyond bilateria.

The genomes of Nemastomella vectensis (starlet sea anemone; Putnam et al., 2007) and Hydra magnipapillata (Chapman et al., 2010) reveal that cnidaria share many but not the entire core set of ECM proteins found in bilateria. Some of these proteins had been described previously, based on cDNA cloning, but the completed genomes allow conclusions about what is absent as well as what is present (subject to the qualifications mentioned in the previous paragraph). The comparative analyses by Fahey and Degnan (2010) are particularly informative. They show clearly that N. vectensis encodes good homologues of most of the basement membrane toolkit: laminin (1α, 1β, and 1γ), nidogen, perlecans, and collagen (IV and XV/XVIII). H. magnipapillata also encodes these proteins, and cnidaria encode examples of all three fibrillar collagen clades. They also encode homologues of fibrillins and thrombospondins (Fig. 3 A), as well as some other ECM proteins. Also conserved across eumetazoa are cellular receptors for ECM proteins: integrins, which bind many ECM proteins; dystroglycan, which binds laminin and agrin; and CD36, which binds thrombospondins, as well as the membrane proteoglycans, syndecan, and glypicans (Hynes and Zhao, 2000; Huhtala et al., 2005; Ewan et al., 2005; Whittaker et al., 2006; Knack et al., 2008; Ozbek et al., 2010). Therefore, it appears that all eumetazoan genomes encode a common set of ECM proteins, although data for cnidarians are sparse. Individual taxa may lack some of this set but it is clear that the common ancestor of eumetazoa had a reasonably complex repertoire of ECM proteins that has been largely conserved throughout subsequent evolution.

Evolution of ECM in basal metazoa
Given this strong conservation of a core set of ECM proteins in all eumetazoa, it is of obvious interest to ask when the genes encoding these proteins arose during evolution and to attempt to correlate their emergence with the acquisition of novel
morphological and developmental features. The taxa closest to Metazoa are the Placozoa and the Porifera (sponges). Genomes from these two phyla have recently been completed: the placozoan Trichoplax adhaerens (Srivastava et al., 2008) and the demosponge Amphinom da queenslandica (Srivastava et al., 2010). These genomes have proven quite informative concerning the origins of ECM proteins (see also Fahey and Degnan, 2010 and Ozbek et al., 2010). As mentioned earlier, neither organism has any true basement membranes. However, the T. adhaerens genome encodes reasonably good orthologues of type IV collagen (two subunits); laminin α, β and γ subunits; and nidogen and perlecan—essentially the entire basement membrane toolkit apart from type XV/XVIII collagen. This is a surprising result given the reported absence of basement membranes in T. adhaerens and it suggests that T. adhaerens has the ingredients to make a basement membrane. Perhaps there are stages in the T. adhaerens life cycle where basement membranes are assembled or perhaps some other protein is needed for coassembly or as a cell-surface receptor. T. adhaerens does encode potential laminin receptors, including dystroglycan as well as an integrin, although the homology of the latter with subclasses of bilaterian integrins has not yet been explored. It will be of interest to determine the biosynthetic patterns and distributions of the basement membrane proteins and these potential receptors in T. adhaerens.

In contrast, the A. queenslandica genome encodes homologues of all three laminin subunits, albeit with imperfect matches in domain composition (Fahey and Degnan, 2010), but does not encode any of the other proteins of the basement membrane toolkit, which is consistent with the absence of basement membranes in demosponges. The more complete set of basement membrane proteins encoded by T. adhaerens as compared with A. queenslandica is consistent with a closer evolutionary relationship of Placozoa with eumetazoa, as shown in Fig. 1. However, it should be noted that sponges are diverse, with four distinguishable clades (Gazave et al., 2010), one of which, homoscleromorphs, has been reported to have basement membranes. Indeed, type IV collagen cDNA has been isolated from Pseudocorticium jarrei, a homoscleromorph sponge (Boute et al., 1996). Thus, it remains plausible that some sponges may express the basement membrane toolkit and assemble basement membranes, an obvious topic for future investigations.

The T. adhaerens genome also encodes many other candidate ECM glycoproteins, including a homologue of B-type thrombospondins (although in the current genome assembly, the gene may be fused with another) and a partial match with agrin. The genome includes many genes with known ECM domains in unusual combinations not seen in eumetazoa. Some of these inferred proteins include predicted transmembrane domains and may, in fact, be surface glycoproteins rather than true ECM proteins. In contrast with sponges, there is little evidence for collagens other than type IV in T. adhaerens. However, it is clear that this simple organism with only four known cell types has elaborated large numbers of genes encoding multiple ECM domains. The elaboration of ECM proteins appears further developed in Placozoa than in the sponge species analyzed to date. Further comparative analyses of the T. adhaerens genome and those of sponges should shed further light on the evolution of diverse combinations of extracellular domains in these simple metazoan animals.

Hints of earlier evolution of ECM domains in unicellular organisms

There is widespread agreement that choanoflagellates are the closest unicellular relatives of metazoa (King et al., 2003, 2008). Their characteristic cellular organization, with a collar of actin-based filopodia surrounding a single apical flagellum, is similar to that of choanocytes, the feeding cells of sponges. The complete genome of Monosiga brevicollis (King et al., 2008) and the partial one of Salpingoeca rosetta (Broad Institute Origins of Multicellularity Initiative; http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html) have revealed that these two choanoflagellates encode several proteins previously considered to be specific to metazoa. These include homologues of the cell–cell adhesion receptor cadherins. The presence of some integrin domains in choanoflagellates might also suggest a role in ECM-mediated adhesion, but there are no true integrins. There are a few genes encoding α integrin repeats, but none of them looks like a fully developed integrin subunit, and there is no evidence for any β subunits (King et al., 2008). Furthermore, neither genome encodes any of the proteins of the basement membrane toolkit. Although there are several proteins that include one or more laminin domains, only one approaches eumetazoan (or Placozoan or sponge) laminin subunits in the complexity of domain organization. However, it lacks some domains and is not a true orthologue, and there is no evidence for laminin αβγ heterotrimers. Also, collagen IV, nidogen, and perlecan all appear to be absent (King et al., 2008; unpublished data). Both choanoflagellate species encode several proteins with collagen repeats and others with COLFI domains, but so far never in the same protein, which indicates that they lack true fibrillar collagens. Both choanoflagellate species do encode a protein with multiple collagen repeats and VWA domains. This is superficially reminiscent of certain vertebrate collagens, but the matches in domain architecture are not at all good (unpublished data).

Thus, it appears that choanoflagellates encode several characteristic ECM domains, but, to date, no true matches with bilaterian ECM proteins have been found (King et al., 2008; Ozbek et al., 2010; unpublished data). The unusual VWA collagen may represent an early ECM protein, and it has been suggested that there is a putative fibrillin-like protein encoded in each genome (Ozbek et al., 2010). However, these proposed fibrillin-like proteins consist solely of EGF repeats, lack the TGF-β-binding TB domains of fibrillins, and have transmembrane domains, so their homology with fibrillins is not at all close (unpublished data). Fibrillins and the homologous latent transforming growth factor β-binding proteins (LTBPs) are involved in binding and regulating TGF-β family members but, to date, appear to be eumetazoan in origin (Robertson et al., 2011); placozoa, sponges, and choanoflagellates do not have the TB domain. In fact, M. brevicollis does not actually encode very many ECM-type proteins, and many known ECM domains, which play important roles in conserved bilaterian ECM proteins (compare Figs. 2 and 3),
appear to be absent from the genome. There are also very few Ig family domains and only one or two copies of several other ECM domains, all of which are, in contrast, extremely prevalent in the *T. adhaerens* genome (unpublished data).

In conclusion, at this point it is clear that choanoflagellate genomes contain some domains typical of ECM proteins (LamNT, LamG, FN3, VWA, EGF, COLFI, and collagen repeats) but do not appear to have assembled them into the characteristic arrangements of domains seen in metazoan ECM proteins. They also lack many other ECM domains. Most choanoflagellates are unicellular, although *S. rosetta* does have a colonial phase. The transition to multicellularity therefore seems to have involved both considerable shuffling of preexisting domains (King et al., 2008) as well as evolution of many new ones.

The taxon that contains metazoa and choanoflagellates as well as fungi and several other unicellular relatives is called the opisthokonts. Although fungi contain no credible homologues of ECM proteins (or integrins), several of the other opisthokonts do encode some integrin subunits (Shalchian-Tabrizi et al., 2008; Sebédő-Pédrös et al., 2010), but so far there have been no reports of ECM proteins. One additional unicellular organism that encodes an integrin β subunit but, so far, no α subunits (*Thecamonas trahens* formerly known as *Amastrigomonas sp.*), is an apusomonad (Sebédő-Pédrös et al., 2010). This group is of uncertain phylogenetic position, but the shared integrin subunit suggests a relationship with the other unicellular organisms discussed here (compare Fig. 1). The presence of integrin homologues of unknown function in these unicellular opisthokonts suggests that integrins may have been lost in the choanoflagellate lineage. Why these unicellular organisms encode integrins is unclear. One possibility is that the integrins function in phagocytosis, as has been suggested for the cadherins in choanoflagellates (King et al., 2008). It will be of considerable interest to see the entire genomes of representatives of these unicellular taxa and to investigate the expression and functions of their integrins and whether or not there are any ECM ligands.

**Taxon-specific elaborations**

As for most other categories of genes and proteins, there is a steady increase in the complexity of the “matrisome,” the set of proteins contributing to the ECM, as one ascends the tree of life. This increase comprises several different processes. There are notable examples of taxon-specific elaborations of the matrisome, both by duplication and divergence of existing genes as well as by the addition of new domains, including domains not associated with taxon-specific divergence to suit particular purposes. Although the three clades of fibrillar collagens have an ancient origin before the divergence of eumetazoa (Exposito et al., 2008, 2010; Heino et al., 2009), individual lineages have expanded the set in different ways. Again, vertebrates provide some prime examples. Each of the three clades has expanded (to give a total of 11 fibrillar collagen genes), and individual members of each clade have become specialized for different functions; one from each clade of collagens is expressed selectively in notochord, cartilage, and bone (Wada et al., 2006). Vertebrate genomes also encode complex collagens with additional ECM domains, such as VWA and FN3. These are not newly developed domains; both are widespread and found in many other genes (Fig. 3; Whittaker and Hynes, 2002), and VWA domains do occur in collagen genes of unknown function in *H. magnipapillata* (Zhang et al., 2007) and, as mentioned, in choanoflagellates. There are several specialized vertebrate collagens incorporating VWA and FN3 domains. These include FACIT collagens, which form side branches on collagen fibrils; and collagens VI and VII, which assemble into short fibrils connecting basement membranes to underlying interstitial ECM in locations such as the skin (for review see Ricard-Blum, 2011). The inclusion of these extra domains confers additional interaction capabilities on these collagens, allowing assembly of higher-order structures important for the organisms.

Another example comes from sponges. They encode a family of short-chain collagens (~120 Gly-X-Y repeats) called spongins, which form exoskeletons (familiar in the form of bath sponges). Spongins have a C-terminal domain distantly related to that of type IV collagens, and appear to have diverged from those basement membrane collagens before the parazoan/metazoa split (Aouacheria et al., 2006). Relatives of spongins are found in other invertebrates, although not in ecdysozoa or vertebrates, the spongins presumably having been lost in those lineages. The nematode, *C. elegans*, is one such ecdysozoan. The genome of this worm instead encodes a large number (>160) of collagen genes (Hutter et al., 2000; Myllyharju and Kivirikko, 2004). These encode short collagen chains (~50 collagen Gly-X-Y repeats), which form the cuticle of the worm, a structure that undergoes remodeling at each larval molt. Different sets of cuticle collagen genes are expressed at different times. This is therefore a nematode-specific expansion of this family of specialized collagens for a taxon-specific ECM function, the cuticle. In contrast, flies
Evolution of ECM proteins • Hynes

Among the ECM proteins missing from sea urchins as well as other invertebrates are tenascin, fibronectin, and von Willebrand factor (VWF; Whittaker et al., 2006). All three proteins comprise novel assemblages of domains in combinations not found in other ECM proteins (Fig. 3B), and they serve to illustrate some issues common to the many other vertebrate-specific ECM proteins. Tenascins include multiple EGF and FN3 domains and a single C-terminal FBG domain. All of these domains are ancient in origin, but the combination is only found in deuterostomes. The sea urchin genome does not encode a tenascin, but those of *Branchiostoma floridae* (amphioxus, lancelet, cephalochordate; Putnam et al., 2008), *Ciona intestinalis*, and *Ciona savignyi* (sea squirts, ascidians, tunicates, urochordates; Dehal et al., 2002) all do, and all vertebrates encode multiple tenascins (Tucker and Chiquet-Ehrismann, 2009; Chiquet-Ehrismann and Tucker, 2011). The different vertebrate tenascins are differentially expressed in various ECMs, including those in the central nervous system (CNS) and during inflammatory and carcinogenic processes, and, given their association with disease states, clearly play important roles in vertebrates (Chiquet-Ehrismann and Tucker, 2011). Fibronectin appeared even later in the deuterostome lineage. In contrast with tenascins, fibronectin does include novel domains; although FN3 domains are ancient in origin, FN2 and FN1 domains are much more recent developments largely confined to chordates. The structure of vertebrate fibronectin is highly conserved in the entire vertebrate phylum—once assembled, this gene appears to have been under strong selection—and it is essential for life in every species tested. Ascidians do encode a fibronectin-related gene (Tucker and Chiquet-Ehrismann, 2009) with all three fibronectin domains (FN1, -2, and -3), but it lacks key features (domains and motifs) of fibronectin structure and function,

Figure 4. Deuterostome phylogeny and elaboration of ECM proteins. The figure diagrams deuterostome evolution according to the same principles outlined in Fig. 1. The core set of ECM proteins encoded in the genomes of all bilateria is boxed in black. Many of these proteins are also found in cnidaria [see main text]. The main taxa of the deuterostome lineage and their relationships as currently understood are indicated with representative animals noted. The first reported appearances of particular proteins are marked in red.
has additional domains not found in vertebrate fibronectins, and is best viewed as a proto-fibronectin (unpublished data). VWF is the final vertebrate ECM protein we will discuss in this context. This gene appears conserved in mammals, birds, amphibians, and fish (and presumably other vertebrates). As for fibronectin, there appears to be a proto-VWF in ascidians with similar domains but differentially arranged and including additional domains (unpublished data). VWF is a key protein in hemo-
stasis, being responsible for platelet adhesion under high-shear conditions such as those in arterioles (Sadler, 2009; Bergmeier and Hynes, 2012). So, its function would appear to be necessary only in vertebrates. Its domain structure reveals that it is related to mucins, which are found in many invertebrates; the key in-
novation is the inclusion of a set of three VWA domains that are involved in binding to collagen (as in certain integrins) and to the cell-surface receptor, GPIb/V/IX, on platelets.

These three proteins, as well as the collagens, exemplify the role of domain shuffling and the addition of novel domains to ECM proteins to confer novel functions. For VWF, it is plau-
sible to infer the novel functions from our knowledge of its he-
mostatic role in mammals. What could be the novel functions that selected for the evolution of tenascins and fibroactin in vertebrates? It could be that they are necessary for the develop-
ment of vertebrate-specific structural ECMs like cartilage (as for some collagens and proteoglycans), but tenasin and fibro-
nectin do not have obvious roles in such ECMs. Another pos-
sibility is neural crest migration, a key feature of vertebrate development; both tenasin C and fibronectin are strongly ex-
pressed in neural crest, and fibronectin has been functionally implicated in this migration as well as in condensation of somites (Hynes 1990), another vertebrate synapomorphy. Development and function of an endothelium-lined vasculature and high-
pressure circulation is also a specialization of vertebrates. Fibro-
nectin clearly plays a role there, and tenascins are expressed in the vertebrate CNS (Chiquet-Ehrismann and Tucker, 2011), as are many other ECM proteins, including both the pan-
eumetazoan proteins, laminins, netrins, slit, and agrin, as well as later-evolving proteins (e.g., reelin and thrombospondin-1) and vertebrate-specific proteins such as proteoglycans and SCO-spondin (Barros et al., 2011).

Concluding remarks

The recent completion of genome sequences for cnidaria has shown that the common set of ECM proteins already known to be shared by all bilaterian taxa originated before the eumeta-
zoan radiation >600 million years ago, and that many of these proteins have been conserved ever since, which indicates the importance of ECM for metazoan life. Furthermore, genome sequences of basal metazoan have shown that placoazans have many of the same proteins, most notably including the base-
ment membrane toolkit. Sponges have a somewhat simpler repertoire of ECM proteins, and the demosponge for which genomic information is available lacks the basement membrane toolkit. Thus, it appears that, with respect to ECM content, placoza are closer to eumetazoas than are demosponges, al-
though information on other sponge clades will be of future in-
terest. With the core of eumetazoan ECM proteins as a point of reference, one can ask when these proteins arose in premeta-
zoan organisms and how the repertoire has been expanded in higher-order taxa.

Genomes of the closest unicellular relatives of metazoa—
choanoflagellates—encode some domains characteristic of ECM proteins but appear not to have organized them in the combina-
tions and patterns typical of metazoan ECM proteins. Choano-
flagellates also lack ECM receptors such as integrins. However, some other unicellular opisthokonts do encode integrins, al-
though no metazoan-type ECM proteins have yet been detected. Therefore, assembly of complex domain structures in ECM pro-
teins seem to have accompanied the acquisition of multicellular-
ity, with placozoa showing extensive elaboration of novel ECM proteins with domain combinations not reported elsewhere. The core set of ECM proteins has shown multiple taxon-specific ex-
pansions to meet particular needs. This is particularly evident in the deuterostome lineage leading to chordates and vertebrates. These taxa have greatly expanded the repertoire of ECM pro-
teins both by gene duplication and divergence, and by the evolu-
tion of novel ECM proteins incorporating novel arrangements of old domains as well as the occasional addition of new ones. Evolution of this diverse set of ECM proteins has been enabled by their modular protein structures, with individual domains en-
coded as exonic units allowing shuffling during evolution.

I would like to thank Charlie Whittaker and Sebastian Haersch of the Swanson Biotechnology Bioinformatics Facility in the Koch Institute and Aleixandra Naba of my own laboratory for their collaborations on annotation of ECM proteins.

I would like to thank the Howard Hughes Medical Institute for finan-
cial support.

Submitted: 8 September 2011
Accepted: 23 February 2012

References

Aouacheria, A., C. Goeurjon, N. Aghajari, V. Navratil, G. Deléage, C. Lethias, and J.-Y. Exposito. 2006. Insights into early extracellular matrix evolution: spongion short chain collagen-related proteins are homologous to base-
ment membrane type IV collagens and form a novel family widely dis-
org/10.1093/molbev/msl100
Bergmeier, W., and R.O. Hynes. 2012. Extracellular matrix proteins in hemosta-

Downloaded from jcb.rupress.org on April 18, 2012