Graphical Abstract

It is a given that ECM proteins function outside cells. We evaluate emerging data that implicate non-canonical roles in the cytoplasm or nucleus. We discuss the conceptual and experimental challenges that will need to be met to investigate this under-studied area of ECM biology more rigorously.
Insider Trading: Extracellular Matrix Proteins and their Non-Canonical Intracellular Roles

A “Problems and Paradigms” article

Andrew L. Hellewell and Josephine C. Adams+

School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK

+Corresponding author: Josephine C. Adams

Email: jo.adams@bristol.ac.uk

Key words: Extracellular Matrix, Protein targeting, ER Stress, Nucleus, Thrombospondin, SLRPs, Cancer

Abbreviations: ECM, Extracellular Matrix; ER, Endoplasmic Reticulum; ERAD, Endoplasmic reticulum-associated degradation, IGF, Insulin-Like Growth Factor; IGFBP, Insulin-Like Growth Factor Binding Protein; NLS, Nuclear Localisation Sequence; SLRP, Small Leucine Rich Proteoglycan; TSP, Thrombospondin.
Summary

In metazoans, the extracellular matrix (ECM) provides a dynamic, heterogeneous microenvironment that has important supportive and instructive roles. Although the primary site of action of ECM proteins is extracellular, evidence is emerging for non-canonical intracellular roles. Examples include osteopontin, thrombospondins, IGF-binding protein 3 and biglycan, and relate to roles in transcription, cell-stress responses, autophagy and cancer. These findings pose conceptual problems on how proteins signalled for secretion can be routed to the cytosol or nucleus, or can function in environments with diverse redox, pH and ionic conditions. We review evidence for intracellular locations and functions of ECM proteins, and current knowledge of the mechanisms by which they may enter intracellular compartments. We evaluate the experimental methods that are appropriate to obtain rigorous evidence for intracellular localisation and function. Better insight into this under-researched topic is needed to decipher the complete spectrum of physiological and pathological roles of ECM proteins.
Introduction

The extracellular matrix (ECM) of animals is a complex, structured, proteinaceous network that has vital roles in supporting multi-cellularity and the organisation of tissues. Over the last decades, our understanding of the ECM has expanded significantly such that the metazoan ECM is now recognised as a dynamic, heterogeneous environment that mediates complex interactions between constitutive ECM components, cell surfaces, growth factors, morphogens, cytokines and other signalling molecules [1]. From their extracellular locations, ECM proteins also have profound effects on intracellular signalling pathways – the so-called ‘outside-in’ signalling [2]. Whilst many diverse and tissue-specific roles exist for the ECM and its components, the primary site of function of ECM proteins is firmly established to be extracellular.

Papers in the recent literature challenge this paradigm by pointing to an intriguing and paradoxical phenomenon, that of intracellular locations and functions for ECM proteins themselves. Exciting new data propose functions within different intracellular compartments, including active roles within the secretory pathway, as well as localisation and specific functions in the cytoplasm or nucleus. Collectively, these studies implicate intracellular roles in transcription, ER stress, autophagy and cancer. These data are problematic in view of the fact that ECM proteins contain secretory signal peptides: the growing evidence for intracellular roles raises many basic questions about how ECM proteins can be routed to the cytosol or nucleus, either from the secretory pathway or by escape from the endo/lysosomal degradation pathway. The data also imply that ECM proteins must maintain protein folding and function in milieu with differing redox, pH and ionic conditions [3,4].

The data are also surprising with regard to current knowledge of the evolutionary origin of metazoan ECM proteins. The dogma is that ECM proteins evolved for extracellular roles; indeed, all the indications from molecular phylogeny are that these proteins evolved as secreted protein innovations and are specific to metazoans [5,6]. It appears unlikely that intracellular forms represent an ancestral state. Indeed, the examples of intracellular localisations and functions that we will discuss pertain to examples of ECM proteins that evolved either early or late in the metazoan phylogenetic tree.

In reviewing the literature on non-canonical intracellular locations and functions for ECM proteins, it was necessary to first establish our working definition of an ECM protein. Preparing a comprehensive list of ECM proteins is not straightforward. Because ECM proteins are typically large, multi-domain proteins with many repeated domains, database annotations can be inaccurate and current Gene Ontology categories under-represent ECM proteins. We began with Richard Hynes’ definition of the ‘Core Matrisome’, which includes the proteins that are considered to have a central role in the ECM, and is based on analysis of domain architectures [7]. Whilst there is strong experimental evidence for the majority of these proteins as bon fide ECM proteins, some are included on the basis of their domain structure and have not been rigorously proven to enter the ECM. The scope of the examples we will discuss excludes ‘matrisome affiliated’ proteins that only share limited architectural or biochemical similarities with core ECM proteins, and non-ECM ‘adhesome’ proteins.
(this term considers the complete biological system of the ECM, which depends on post-translational processing, cell surface receptors, extracellular proteases and cross-linking enzymes, as well as the ECM resident proteins [8]). We note that the ‘matrisome affiliated’ and ‘adhesome’ categories do include secreted proteins, such as lysyl oxidase (LOX) and matrix metalloproteinases (MMPs), that have also been reported to have intracellular roles [9,10].

The paradigmatic lifecycle of an ECM protein: A Summary

Before considering the evidence for intracellular localisations and functions, we will outline the canonical synthesis and trafficking of ECM proteins (Figure 1, green arrows). ECM proteins contain secretory signal peptides and enter the secretory pathway cotranslationally via the translocon of the rough endoplasmic reticulum (rER). Within the ER lumen, protein folding proceeds with the assistance of chaperones. Many ECM protein molecules are oligomeric: these oligomers assemble upon translation and prior to secretion. As examples, members of the thrombospondin (TSP) family undergo oligomerisation into trimers or pentamers via coiled-coil domains [11]. Tenascin-C assembles into a 6-armed hexabrachion structure, composed of two trimers that each assemble as an α-helical coiled-coil by heptad repeats within the N-terminal region and then associate into a hexamer by N-terminal assembly domains and adjacent stabilizing disulphide bonds [12]. In the case of collagens, hydroxylation of lysine and proline residues is essential for thermal stability of triple helix assembly [13]. Many ECM proteins also undergo extensive post-translational carbohydrate modifications involving both N- and O-linked sugars [11,12].

Because a major role of ECM proteins is to form extracellular structural networks and fibrils, an important aspect of transit through the secretory pathway is the shielding of matrix assembly sites that would otherwise promote intermolecular interactions. For example, fibronectin is secreted as a globular dimer that only becomes extended and competent to assemble fibrils upon binding to cell-surface integrins [14]. The triple helices of procollagen molecules are assembled in the ER/Golgi, but globular N- and C-terminal domains inhibit the onset of fibril assembly; these domains are cleaved by specific proteases in coordination with secretion to enable higher order assembly [13]. For other ECM proteins, associated ER-resident chaperones may contribute to blockade of matrix assembly sites. As we will cover in the main part of this article, some ECM proteins co-traffic with extracellular interaction partners and thereby modulate their activity. In common with other secreted proteins, export of ECM proteins from the ER requires the general export factor COPII (coat protein II complex) [15-17].

Whilst many ECM proteins are secreted constitutively, for others secretion is regulated according to physiological conditions. Thus, von Willebrand Factor resides in cytoplasmic storage granules of endothelial cells until its release is activated by hypoxia or damage to the endothelium [18].

Once in the extracellular milieu, incorporation of ECM proteins into the ECM is ensured by specific protein-protein or carbohydrate-based interactions. At the external face of the plasma membrane, binding to receptors such as integrins activates intracellular signalling [2]. Dynamic turnover of the ECM is effected by extracellular proteases including the MMPs and A Disintegrin And Metalloproteinase with Thrombospondin Motifs (ADAMTS) proteinases and/or by cellular uptake of ECM proteins, or their proteolysed fragments, for intracellular degradation [19,20]. Endocytosis of ECM proteins often occurs by interaction.
with heparin and low density lipoprotein-receptor-related proteins (LRP-1) [20,21], and is followed by trafficking to endocytic vesicles and subsequent intracellular degradation in lysosomes (Figure 1). As this summary of the canonical lifecycle of an ECM protein makes clear, intracellular residency is limited to membrane-bound compartments of the secretory and endo/lysosomal systems.

How Can ECM Proteins Reside Intracellularly?

How might an ECM protein circumvent the above processes to function from within a cell? A number of possibilities are outlined here and in Figure 1 (blue arrows).

1. Alternative splicing of mRNA, that results in omission of the region encoding the N-terminal secretory signal peptide, is a well-defined mechanism that can directly determine an intracellular localisation/function for an ECM protein. We will discuss the examples of adipocyte enhancer binding protein 1 (Aebp1) [22] and fibulin 1D' [23].
2. Alternative translation initiation within an mRNA transcript can also result in omission of the secretory signal peptide. We will discuss the example of osteopontin [24].
3. For a number of ECM proteins, distinctive intracellular roles have been identified that are enacted within the endoplasmic reticulum or subsequent compartments of the secretory pathway. We will discuss examples of chaperone-like activities and the modulation of ER stress.
4. Some ECM proteins can escape the endo/lysosomal system and enter the cytoplasm or nucleus. We will discuss the example of insulin-like growth factor binding protein-3 (IGFBP3) [25].
5. A number of ECM proteins contain a nuclear localisation sequence and, despite the presence of a secretory signal peptide, localise to the nucleus. The best evidence for this mechanism is for members of the small leucine-rich proteoglycan (SLRP) family, biglycan [26] and prolargin [27].
6. Conceptually, ECM proteins might enter the cytosol by delocalisation from the ER or other compartments of the secretory pathway, via the ER-associated degradation (ERAD) pathway [28,29]. Certain bacterial toxins exploit this pathway to dislocate from the ER and enter the cytosol [30,31], thus a mechanism of protection from ERAD is not inconceivable. There is no definite example of this pathway for ECM proteins at present.

Intracellular Localisation and Roles – Evidence of Mechanisms

Alternative Splicing

Alternative splicing of mRNA, that results in omission of the region encoding the N-terminal secretory signal peptide, will lead to protein translation within the cytosol and thus intracellular retention (Figure 1). This form of regulation has been established for a few ECM proteins and there is evidence of distinctive roles of the intracellular variants (Table 1). Adipocyte enhancer binding protein 1 (Aebp1) exists as two isoforms because of alternative splicing of its mRNA (Figure 2A) [22]. The smaller isoform, Aebp-1, encodes an intracellular, nuclear protein of 82 kDa that is expressed ubiquitously and influences adipogenesis and
cholesterol homeostasis by binding to DNA as a transcriptional repressor [32,33]. The longer protein, aortic carboxypeptidase-like protein (Aclp; 172 kDa), is targeted for secretion by inclusion of an N-terminal signal peptide [34]. The extra 380 bp in the Aclp open reading frame also encode an 11 amino acid lysine and proline rich repeat motif and a discoidin domain: the latter is thought to function in cell aggregation, cell adhesion or cell-cell recognition [34-36]. Aclp localises predominantly to the ECM of collagen-rich tissues in developing mouse embryos, notably the dermis, vasculature and skeleton [37]. Aclp+/Aclp- mice, generated by a strategy to knock-out expression of both variants, die early during development; those that survive to adulthood display non-healing skin wounds [38]. At least some of the actions of Aclp within the ECM are mediated by transforming growth factor-β (TGF-β)-dependent pathways, for example by binding to TGF-β receptor II [36].

A second example of alternative splicing that affects inclusion of the signal peptide occurs in Fibulin 1D [23]. Fibulins are ECM proteins with roles in elastic matrix fibre assembly and function [39]. Fibulin 1 is a major ECM glycoprotein, expressed predominantly in blood vessels [40], where it interacts with many ECM components, including fibronectin and laminin [41]. However, fibulin 1 was first isolated from placental extracts as a binding partner of a synthetic peptide corresponding to the cytoplasmic domain of integrin β1 subunit [42]. More recently, Twal et al. investigated the splice variants of fibulin 1 and identified a new variant, fibulin 1D', expressed by the A431 cell line. This new variant does not encode the N-terminal signal sequence or the anaphlatoxin region and is not detected in conditioned media of transfected CHO cells. Instead, fibulin 1D’ was found to be located intracellularly in CHO cells and in vivo in human placenta tissue (Table 1) [23]. A role for this new splice variant in cell adhesion and motility through binding to integrin β1 subunit has been proposed, but these functions remain to be investigated [23].

**Alternative Translation**

A second mechanism for omission of a secretory signal peptide from a polypeptide is alternative translation (Table 1): this often involves translational start from non-AUG codons, in particular CUG, UUG, GUG, ACG, AUA or AUU [43]. This form of regulation applies to a limited number of human genes, including osteopontin, a phosphorylated ECM glycoprotein which has multiple roles in tissue homeostasis and inflammation [44]. Osteopontin mRNA includes two translation initiation sites that allow for the production of osteopontin either with, or without, the N-terminal signal sequence (Figure 2B). The initiation of translation of intracellular osteopontin is predicted to occur because of leaky ribosome scanning or stable secondary structures that stall the ribosome and allow recognition of an alternative translation initiation site [24]. The first evidence for intracellular osteopontin was obtained by confocal microscopy of cultured rat calvarial (cranial bone) cells. Intracellular staining was identified in perinuclear regions, consistent with Golgi localisation, yet also in distinct intracellular patches adjacent to the plasma membrane of migratory cells [45]. Subsequently, a role for membrane-proximal osteopontin in cell motility of fibroblasts, dependent on co-localised CD44-ERM (ezrin-radixin-moesin) complex at the plasma membrane was identified [46,47]. Osteopontin has also been identified in the cytoplasm associated with Toll-like receptor 9 (TLR9), a transmembrane protein which recognises unmethylated CpG sequences in DNA common to bacterial and viral DNA [48] and in nuclear extracts in association with the transcription factor polo-like-kinase-1 (Plk-1). A role in mitosis was identified [49]. Intracellular osteopontin is also
implicated in suppression of TLR-mediated immune responses [50] and regulation of homeostasis and function of natural killer cells [51]. The diversity of cellular sites of action is especially intriguing given that osteopontin is present only in amniote vertebrates. Whether the occurrence of alternative splicing or translation is widespread in ECM proteins is an intriguing question for future research.

**Modulation of ER Stress**

Evidence is accumulating that numerous ECM proteins do not traverse the ER lumen and secretory pathway as simple cargoes (Figure 1), but undergo functionally significant interactions. For example, latent TGF-β-binding protein-2 and -3 (LTBP-2, -3) bind proprotein convertase 5/6A (PCS/6A) in the ER before being co-secreted and sequestered in the ECM as a complex in which PCS activity is blocked [52].

In other cases, ER-localisation is associated with functions distinct from those within the ECM. Insulin-like growth factor binding proteins (IGFBPs) are important regulators and carrier molecules for insulin-like growth factor (IGF) [53]. The family member IGFBP3 binds and regulates the activity of IGF through heparin-binding motifs, which anchor IGFBP3 to the ECM [53]. IGFBP3 influences ER function by binding the ER luminal resident protein, 78 kDa glucose-regulated protein/binding immunoglobulin protein (Grp78/BiP), a master regulator of the unfolded protein response. This was demonstrated by multiple methods (Table 1). In breast cancer cells experiencing nutrient starvation and hypoxia, this interaction promotes the autophagic recycling of cytoplasmic components to increase cell survival [54].

Thrombospondins (TSPs) are a large family of evolutionarily ancient, calcium-binding glycoproteins that are classed as matricellular or adhesion-modulating ECM proteins because they interact extracellularly with both cells and the ECM, and do not form structural fibrils [11]. The complex extracellular roles of TSPs involve binding to ECM components, cell surface receptors or growth factors [11]. There are five TSP genes in mammals and several family members have been implicated in ER-located, stress-related, or chaperone-like activities. The best studied is the pentameric family member, TSP4. TSP4 is present in the myocardium, blood vessel walls, skeletal muscle and tendons, and its gene knockout in mice results in altered ECM deposition and composition in the heart, tendon and muscle, and reduced grip muscle strength [55-58].

In-depth study in a transgenic mouse engineered to over-express TSP4 specifically in the heart, established that TSP4 over-expression correlated with elevation of TSP4 within the ER, and with upregulation of the ER chaperone proteins BiP, Sdf211, Creld2, Calreticulin, Armet, Hyou1, Mthfd2 and Pdi [59]. ER stress response proteins, including the PKR-like ER kinase (PERK) pathway component eukaryotic translation initiation factor 2 α (eIF2α), activating transcription factor-4 (Atf4), and Atf6α transcription factors, become activated. The ER and secretory vesicles become enlarged, consistent with the activation of Atf6α, which acts to increase the secretory capacity of cells during injury [59]. A similar, but weaker, response was apparent in mice with cardiac over-expression of TSP1 [59]. Atf6α is a transmembrane protein of the ER from which a cytosolic fragment is cleaved by proteolysis in response to initial ER stress and enters the nucleus to function as a transcription factor. In *vitro*, in mouse cardiac extracts, recombinant TSP4 was identified to bind, through its TSP...
type 3 repeats, to the ER luminal domain of Atf6α, and overexpression of TSP4 in cells correlated with protease-dependent nuclear accumulation of Atf6α. The Atf6α response was lost in Thbs4-/− mice, and expression of Atf6α could not be induced by addition of exogenous TSP-4 protein to cultured neonatal cardiomyocytes [59]. Therefore, the authors proposed that the conserved type 3 repeats of TSPs have a critical intracellular role in ER stress response by driving Atf6α-dependent enhancement of protein secretion, thereby promoting reconstruction of the ECM after injury. More recently, correlated upregulated expression of Thbs4 and Atf6-target genes has been identified in vivo, in mesenteric arteries of the spontaneously hypertensive rat [60].

The most widely-studied activity of a thrombospondin within the ER relates to a pathological context. TSP5, also known as cartilage oligomeric matrix protein, (COMP), is expressed highly in cartilage and tendon. Point or deletion mutations in the type 3 repeats or L-lectin-like domain of TSP5 are causal for the skeletal dysplasia, pseudoachondroplasia (PSACH), and certain forms of multiple epiphyseal dysplasia (MED/EDM1) [61-63]. The mutations cause misfolding and thereby disrupt secretion of TSP5 by chondrocytes. Because this is a constitutive effect, degradation of mis-folded protein via ERAD cannot clear the ER to maintain normal ER function, and the chondrocytes develop swollen, giant ER that are full of TSP5 [64]. An important feature of the pathology is that the accumulation of TSP5 within ER is accompanied by the co-retention of other ECM proteins that are extracellular binding partners of TSP5, notably type IX collagen and matrilin-3, and the onset of chronic ER stress [65,66]. It has not been explored whether this stress response depends on Atf6α. Overall, TSP5 retention leads to aberrant ‘intracellular matrix assembly’ and compromised cellular function, leading to premature chondrocyte death, impaired secretion of multiple ECM proteins and compromised quantity and stability of cartilage ECM [64,66]. Overall, the ER is emerging as an organelle of importance for specific intracellular roles of ECM proteins.

**Functions Subsequent to Endocytosis**

It is recognised that endocytosis and entry into the endo/lysosomal system is a normal stage in the lifecycle of an ECM protein (Figure 1 green arrows), and is a step that can also involve specific functions. For example, TSP1 inhibits vascular endothelial growth factor (VEGF) signalling by binding VEGF for co-internalisation via LRP1 that leads to lysosomal degradation [67]. Some evidence supports the concept that endocytosis can be the prelude to distinct intracellular roles. Micutkova et al. demonstrated by multiple methods that IGFBP3 can undergo dynamin-dependent endocytosis, followed by association with the nuclear envelope and entry of IGFBP3 into the nucleus [25] (Table 1). A nuclear localisation sequence (NLS) was identified in the basic C-terminal region of IGFBP3 (amino acids 215-232) (Figure 3A) and nuclear import was found to depend on importin-β [68,69]. A similar NLS motif is present in IGFBP5 and is conserved across vertebrate species (Figure 3A, 3B) [68]. However, nuclear import of IGFBP5 depends on the nucleolar protein, nucleolin [70]. The active NLS may explain why several nuclear interaction partners have been reported for IGFBP3. These include the transcription factor, retinoid X receptor and the epidermal growth factor (EGFR)-DNA-dependent kinase (DNA-PKcs) complex. Through these interactions, IGFBP3 influences apoptosis and is implicated in the DNA damage response [71,72]. For example, in oestrogen receptor-negative breast cancer cell lines, responses to DNA-damaging drugs include increased association between IGFBP3 and EGFR-DNA-PKcs (Table 1) [71].
In other cases, it is not yet clear whether a cytoplasmic location of an ECM protein is achieved after endocytosis or by another route, and the mechanisms by which an ECM protein might escape from lysosomes or autophagosomes without degradation remains perplexing. An interaction of TSP1 and phosphorylated extracellular-signal-regulated kinase (pERK) in the cytosol has been demonstrated by co-Immunoprecipitation (co-IP) and the Duolink antibody method; the latter produces a signal only for epitopes within 40 nm of each other [73]. The interaction promoted a level of mitogen-activated protein kinase (MAPK) signalling that resulted in stabilised p53 and subsequent cell senescence. Addition of TSP1 antibody to cell cultures failed to prevent the co-IP of TSP1 and pERK, favouring the notion that only intracellular TSP1 might be needed for this function [74]. The mechanism by which TSP1 enters the cytosol to interact with pERK is unknown. Given the mechanism implicated for IGFB3, the escape of endocytosed ECM proteins into the cytosol and nucleus warrants more general investigation.

**SLRPs: Nuclear Localisation and Interactions with Transcription Factors**

A prominent group of ECM proteins that undergo nuclear localisation are the SLRPs (small leucine-rich proteoglycans). SLRPs are matricellular proteoglycans that consist of a protein core of leucine-rich repeat motifs modified post-translationally by addition of glycosaminoglycan (GAG) chains [75-77]. SLRPs undergo an astonishing range of protein-protein interactions, including binding to growth factors, cell surface receptors and collagens [75,76]. They influence innate immunity, inflammation, cell proliferation and cell differentiation [77]. Several reports describe nuclear localisation and NLSs in SLRPs [27,78,79]. For example, a bipartite NLS is predicted in the C-terminal region of decorin (Figure 3A). Decorin was localised to the nucleus in dysplastic and malignant oral epithelial cells by multiple methods (Table 1). Decorin knockdown by siRNA in these cells resulted in significantly reduced cell migration and invasion [79].

Another SLRP, prolargin, has been proposed to act as an inhibitor of osteoclastogenesis through interaction with the transcription factor, nuclear factor kappa- B (NF-κB) [27]. A bipartite NLS is predicted within the N-terminal heparin-binding region of prolargin (hbdPRELP) (Figure 3A) and active NF-κB is a potential interaction partner for hbdPRELP in the nucleus. Indeed, upon incubation with primary osteoclasts, a synthetic peptide corresponding to hbdPRELP became internalised and associated with endosomes and, after 20 minutes, accumulated in the nucleus. Introduction of hbdPRELP into RAW 264.7 cells significantly reduced NF-κB transcriptional activity [27]. It is important to note that these results are based on a synthetic fragment of prolargin: whether this fragment is generated in vivo or whether full length prolargin has intracellular roles requires further investigation.

**Suspected Intracellular Localisations of Additional ECM Proteins**

Further examples of ECM proteins with suspected intracellular localisations are outlined in Table 2. These studies rely, for the most part, on immunohistochemistry, which is insufficient alone for reliable determination of a specific intracellular localisation. One example is laminin, a molecule composed of heterotrimeric of α, β and γ chains that is essential for basement membrane assembly [80]. Immunohistochemistry for laminin-332 in gastric cancer cell lines revealed an unusual intracellular accumulation of β3 and γ2 laminin chains, which correlated with altered cell motility and cancer cell invasiveness [81]. The accumulation of β3 and γ2 laminin chains may result from suppression of α3 chain
translation and thereby inhibition of laminin-332 heterotrimer assembly. However, it is not yet defined whether β3 and γ2 subunits accumulate within the ER or in other intracellular locations, or whether there is a causal link between their accumulation and effects on cell motility.

Lactadherin/Mfge8 is a secreted glycoprotein of milk now known to have roles in cell-ECM interactions in multiple tissues [82]. By immunohistochemistry of cells from breast cancer biopsies, lactadherin is located in the cytoplasm or nucleus of tumour cells, in contrast with its association with ductal lumens in normal mammary glands (Figure 4) [83]. Intracellular lactadherin is proposed to enhance the tumorigenic potential of mammary epithelial cells through regulation of cyclins D1 and D3, but the mechanism for intracellular localisation is unknown, and experimental investigations are needed [83].

Other cases in Table 2 include SPARC (secreted protein acidic and rich in cysteine) and lumican. SPARC, also known as osteonectin, is a matricellular and collagen-binding protein with multiple roles in the modulation of cell-matrix interactions [84,85]. Several studies have reported intracellular SPARC [86-90] and an interaction with axonemal tubulin in ciliated epithelial cells has been identified [91]. Emerging evidence suggests the possibility of further intracellular roles for SPARC, some of which implicate it in tumor progression. For example, upregulation of intracellular SPARC protein has been reported in chronic myelogenous leukemia cells, resulting in resistance to the tumor inhibitor, Imatinib [87]. In lung cancers, the SLRP lumican is expressed aberrantly intracellularly [92], and its cytoplasmic localisation has been correlated with the aggressiveness of lung squamous cell carcinomas and adenocarcinomas [93]. In both cases the functions of the intracellular proteins remain to be identified.

Several ECM proteins have NLSs proven to target them to the nucleus, and nuclear localisation by unknown mechanisms has been demonstrated for others. Examples include biglycan, optincin, dentin matrix protein-1 (DMP-1) and CYR61/CTGF/NOV (CCN) proteins (Table 2). In most cases, their functions in the nucleus remain unclear. The SLRP biglycan and DMP-1, an acidic ECM protein from teeth, both contain functional nuclear localisation signals and localise to the nuclei of cultured cells [26,94,95]. Optincin, another SLRP, is present in the nucleus and ER of patient-derived chronic lymphocytic leukaemia cells [96]. CCN proteins are matricellular proteins which, extracellularly, bind to integrins and various growth factors to influence wound repair, cancer and skeletal development. Nuclear locations have been reported for multiple family members, but the mechanisms are unknown [97-102].

Many of these indications of unusual intracellular locations of ECM proteins relate to cancer cells. Multiple aspects of normal cell physiology are altered in cancer cells as a consequence of genetic or epigenetic changes and these might favour the aberrant localisation of proteins. For example, in oral squamous cell carcinomas, desmosomal and hemidesmosomal proteins localise to the cytoplasm, instead of the plasma membrane [103]. Cancer cells can undergo cell-cell-fusion events or cell-in-cell invasion (entosis) which might also result in cytoplasmic mislocalisation of ECM proteins [104-106]. More research will be needed to determine whether ECM proteins in the cytoplasm of cancer cells have functional roles.
Identifying Intracellular Localisation of ECM Proteins: The Need for Appropriate Methodology

The indication that there are intracellular-located roles for ECM proteins is counter-intuitive: therefore it is vital that the experimental evidence presented to justify these activities is robust. In reporting an intracellular localisation of an ECM protein that is beyond the expected ER/Golgi pattern, the use of multiple, well-validated, antibodies is an essential first step. Antibodies need to have been robustly validated for specificity, both by immunoblot and also by immunohistochemistry in gene knockout or knockdown models, to ensure that any intracellular-detected staining is credible. Other controls to consider are the use of class-matched, non-immune immunoglobulins, or blocking peptides for antibody preabsorption. Independent antibody quality report websites can provide helpful indications of the properties of individual antibodies [107]. Immunohistochemistry and immunofluorescence of cultured cells are used frequently to identify locations of ECM proteins (Table 2) [81,83,87], however, these methods alone are insufficient to define a specific intracellular localisation. In this article, we have focussed, to the extent possible, on research that included multiple experimental approaches to establish intracellular localisation (Table 1). The value of imaging studies can be increased if immunofluorescence based on confocal microscopy is complemented by examination of co-localisation with well-established markers for intracellular compartments. More convincing evidence for intracellular localisation can be obtained if subcellular fractionation methods are used in combination with imaging techniques. Any cell or organelle fractionation methods also need to include appropriate controls in the form of established markers for specific intracellular fractions. Ideally, cell culture studies that identify intracellular roles would include analysis in a 3-dimensional ECM environment, such as that observed in tissues, to more closely mimic the physiological conditions.

If potential intracellular interaction partners are known, these should also be investigated in microscopy-based co-localisation studies. The Duolink antibody method is of value for identification of intracellular interactions between ECM proteins and constitutive intracellular proteins [73]. Förster resonance energy transfer (FRET), which determines the efficiency of energy transfer between two chromophores within 10nm of each other, is another definitive method to demonstrate proximity of a tagged ECM protein with a tagged intracellular resident protein. Again, co-localisation evidence needs to be supplemented with biochemical or biophysical evidence for a physical association between an ECM protein and a constitutive intracellular protein. In vitro assays to test direct binding of purified molecules and demonstrations of co-fractionation or association in cell extracts are all of value. Identification of interactions via proximity ligation with a biotin ligase-tagged ECM protein can be used to validate protein partners, or as an unbiased method to search for novel intracellular associated proteins [108].

The use of mouse gene knockout, transgenic, or clustered regularly interspaced short palindromic repeats (CRISPR) models to investigate specific roles of intracellular ECM proteins has to date been relatively limited, yet these are all important approaches that enable manipulation of either intracellular or extracellular isoforms in the in vivo context. With the exception of tissue immunohistochemistry, the majority of data on intracellular localisations and functions of ECM proteins to date have come from cell culture experiments. Progress to 3-dimensional cultures and in vivo models is needed to gain a
better understanding of the physiological occurrence, regulation, relevance and roles of the intracellular forms discussed in this article.

**Conclusion and Future Perspectives**

Whilst there are robust, convincing data for the intracellular localisation of certain ECM proteins, much remains unknown and speculative. Clear evidence for mechanisms and intracellular roles is lacking in many cases. The wider use of RNA interference, CRISPR, or gene knockout models, alongside thorough, well-designed controls, would help to uncover intracellular roles and place them in a physiological context. In many cases, the mechanisms by which an ECM protein becomes localised and functional inside the cell remain unknown or poorly characterised. A new research area would be the clear identification of pathways that lead to cytoplasmic or nuclear localisation. A good example is the study of Micutkova and colleagues in which the endocytic mechanisms involved in IGFBP3 uptake and penetration of the nuclear envelope were analysed systematically [25]. Two other fascinating mechanisms are alternative splicing [22,23] and alternative translation [24], which provide demonstrated routes by which intracellular forms of ECM proteins can be generated. The question of whether these levels of regulation occur for a wider repertoire of ECM proteins than the examples discussed here warrants investigation. How these mechanisms might be regulated in cells to adjust the balance between extracellular and intracellular location of an ECM protein is an open question. With regard to the perplexing question of cytoplasmic sites of function of ECM proteins with signal peptides, the possibility of interactions between ECM proteins and components of the ER dislocation process, Sec61 and Derlins, is of interest [29]. Whilst an enormous amount of research has furthered the consideration of the ECM as a multifaceted environment capable of complex signalling and regulation of cellular behavior, the emerging topic of intracellular roles for ECM proteins presents a paradigmatic challenge with potential to open up exciting new avenues of research in the future.

**Acknowledgements**

ECM research in JCA’s laboratory is supported by MRC K018043. The authors have no conflicts of interest to declare.
Figure Legends

**Figure 1. Schematic diagram of the conventional trafficking of ECM proteins and possible routes to intracellular localisation.** Green arrows indicate conventional secretion and endocytosis pathways for ECM proteins. Solid blue arrows indicate routes for intracellular localisation for which there is experimental evidence. Dotted blue arrows show potential routes for intracellular localisation with limited or no supporting evidence.

**Figure 2. Mechanisms of intracellular localisation. A)** Alternative splicing of AEBP1 transcripts results in a shorter mRNA which does not encode the signal peptide. Mouse Aebp1 was drawn in FancyGene [115]. The figure is adapted from [22]. **B)** Alternative translation initiation within osteopontin mRNA results in an intracellular protein which lacks the signal peptide. Opn-FL refers to the sequence that includes the signal peptide. Opn-i refers to the intracellular isoform of osteopontin. Δ1-39 and Δ1-48 were constructs engineered to identify the sequence important for alternative translation initiation between nucleotides 39 and 48 of the mRNA [24]. The figure is adapted from [24] and osteopontin mRNA has been shortened for illustrative purposes.

**Figure 3. Nuclear localisation sequences in ECM proteins. A)** Published and predicted examples of bipartite NLSs in SLRPS and IGFBPs. NLS in decorin and prolargin were predicted with cNLS mapper [116]. A cNLS Mapper score of 3, 4 or 5 indicates predicted localisation to both the nucleus and cytoplasm. NLSs in IGFBP3 and -5 are as published [69]. **B)** The bipartite NLS of IGFBP3 is highly conserved between vertebrate species. The sequence alignment was prepared in Muscle 3.8 [117] and is presented in Boxshade. Black shading indicates conserved residues and grey shading indicates related residues.

**Figure 4. Immunohistochemical demonstration of intracellular lactadherin.** Lactadherin is present in the cytoplasm (A) (arrow) and nucleus (B) (arrowhead) of invasive breast carcinoma specimens. **C)** Negative control without primary antibody. **D)** Distribution of lactadherin in normal mammary gland (asterisk indicates mammary duct lumen). Scale bar: 50 μm. Reprinted by permission from Macmillan Publishers Ltd: [Oncogene] [83], copyright (2011).

**Table 1. ECM proteins identified to have intracellular localisation and function.**

Key. EM, Electron microscopy; GST, Glutathione-S-Transferase; IF, Immunofluorescence; IHC, Immunohistochemistry; IP, Immunoprecipitation.
References


Figure 2
Figure 1

1. Alternative splicing

2. Alternative translation

3. Active roles in rER

4. Exit from endo / lysosomal system

5. Trafficking with nuclear localisation signal

6. Exit from ER, Golgi or vesicles

Conventional secretion

Internalisation from extracellular space

Extracellular roles

Plasma Membrane

Intracellular

Golgi apparatus

Rough Endoplasmic reticulum

mRNA

Gene

Nucleus

Lysosome Degradation
A)  

<table>
<thead>
<tr>
<th>Protein</th>
<th>Position</th>
<th>Sequence</th>
<th>cNLS Mapper Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decorin</td>
<td>138</td>
<td>NTKKASYSGVSLFSNPVQYWEIQPSTFCVY</td>
<td>4.5</td>
</tr>
<tr>
<td>Prolargin</td>
<td>2</td>
<td>RSPLCWLLPLLASVAQGQPTRRPRP</td>
<td>4.5</td>
</tr>
<tr>
<td>IGFBP3</td>
<td>215</td>
<td>KKGFYKKQCRPSKGRKR</td>
<td>Published [69]</td>
</tr>
<tr>
<td>IGFBP5</td>
<td>201</td>
<td>RKGFYKRQCKPSRGRKR</td>
<td>Published [69]</td>
</tr>
</tbody>
</table>

B)  

<table>
<thead>
<tr>
<th>Species</th>
<th>Position</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danio rerio</td>
<td>245</td>
<td>KGFYKKQCPSKGRKRGCWCV</td>
</tr>
<tr>
<td>Callorhinchus mili</td>
<td>240</td>
<td>KGFYKKQCPSKGRKRGCWCV</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>233</td>
<td>KGFYKKQCPSKGRKRGCWCV</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>248</td>
<td>KGFYKKQCPSKGRKRGCWCV</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>242</td>
<td>KGFYKKQCPSKGRKRGCWCV</td>
</tr>
</tbody>
</table>

Figure 3
Figure 4