Time-lapsed proteomics reveals a role for the novel protein, SNED1, in modulating ECM composition and protein folding

Fred Lee[1#], Xinhao Shao[2#], Yu (Tom) Gao [2, 3, *], and Alexandra Naba [1, 3, *]

[1] Department of Physiology and Biophysics, University of Illinois at Chicago, Chicago, IL 60612, U.S.A

[2] College of Pharmacy, University of Illinois at Chicago, Chicago, IL 60612, U.S.A

[3] University of Illinois at Chicago Cancer Center, Chicago, IL 60612, U.S.A

[#] These authors contributed equally to the study

[*] **Correspondence:**
Dr. Alexandra Naba
Department of Physiology and Biophysics - University of Illinois at Chicago
835 S. Wolcott Avenue
Room MSB-E202
Chicago, IL 60612
U.S.A
Phone: +1 (312) 355-5417
Email: anaba@uic.edu

Dr. Yu (Tom) Gao
College of Pharmacy - University of Illinois at Chicago
833 S. Wood St,
Room - Pharm 505
Chicago, IL 60612
U.S.A
Phone: +1 (312) 996-8087
Email: yugao@uic.edu

ABSTRACT

The extracellular matrix (ECM) is a complex and dynamic meshwork of proteins providing structural support to cells. It also provides biochemical signals governing cellular processes including proliferation and migration. Alterations of ECM structure and/or composition has been shown to lead to, or accompany, many pathological processes including cancer and fibrosis. To understand how the ECM contributes to diseases, we first need to obtain a comprehensive characterization of the ECM of tissues and of its changes during disease progression. Over the past decade, mass-spectrometry-based proteomics has become the state-of-the-art method to profile the protein composition of ECMs. However, existing methods do not fully capture the broad dynamic range of protein abundance in the ECM, nor do they permit to achieve the high coverage needed to gain finer biochemical information, including the presence of isoforms or post-translational modifications. In addition, broadly adopted proteomic methods relying on extended trypsin digestion do not provide structural information on ECM proteins, yet, gaining insights into ECM protein structure is critical to better understanding protein functions. Here, we present the optimization of a time-lapsed proteomic method using limited proteolysis of partially denatured samples and the sequential release of peptides to achieve superior sequence coverage as compared to standard ECM proteomic workflow. Exploiting the spatio-temporal resolution of this method, we further demonstrate how 3-dimensional time-lapsed peptide mapping can identify protein regions differentially susceptible to trypsin and can thus identify sites of post-translational modifications, including protein-protein interactions. We further illustrate how this approach can be leveraged to gain insight on the role of the novel ECM protein SNED1 in ECM homeostasis. We found that the expression of SNED1 by mouse embryonic fibroblasts results in the alteration of overall ECM composition and the sequence coverage of certain ECM proteins, raising the possibility that SNED1 could modify accessibility to trypsin by engaging in protein-protein interactions.

KEYWORDS: Extracellular matrix, matrisome, limited proteolysis, mass spectrometry, protein folding, peptide mapping

INTRODUCTION

The extracellular matrix (ECM) is a remarkably complex assembly of proteins that plays a critical architectural role in multicellular organisms by governing cell polarization, organizing cells into tissues, conferring mechanical properties to tissues and organs, and contributing to morphogenesis [1–4]. The ECM also conveys biochemical and biomechanical signals that tightly control all cellular processes from cell proliferation and survival, to migration, to differentiation [1,5]. Alterations of ECM structure and/or composition arising from mutations in ECM genes, or from an imbalance in ECM homeostasis, *e.g.*, excessive accumulation or degradation, lead to, or accompany, many pathological processes including skeletal diseases, cardiovascular diseases, fibrosis, and cancer [6–12]. To understand the underlying molecular bases of ECM involvement in diseases, we need methods and tools capable of probing the complexity of this compartment.

Over the past decade, bottom-up mass-spectrometry (MS)-based proteomics, has become a method of choice to profile the global protein composition of ECMs – or matrisomes – of tissues, organs, or produced by cells *in vitro* [13–17]. This method infers protein presence based on the identification of the most abundant peptides in any given sample. Proteomics has revealed that the ECM of any given tissue is made of well over 100 distinct proteins and has allowed the identification of ECM protein signatures characteristic of physiological processes (*e.g.*, development [18], aging [19–22]) and pathological states (*e.g.*, cardiovascular diseases [23], cancer [24], fibrosis [25–27]). However, existing proteomic approaches present limitations. First, the size and relative abundance of proteins found in the ECM span a very broad dynamic range, from very large and hyper-abundant collagens to small and low-abundance ECM remodeling enzymes or ECM-bound growth factors. As such, and despite improvements in instrumentation and sample preparation, including protein and peptide fractionation, proteins present in lower abundance are often eclipsed [14]. In addition, trypsin digestion remains the gold standard to generate peptides in a bottom-up MS workflow [28] but is not without limitations. For example, some proteins can be intrinsically resistant to tryptic digestion and, as such, may not be detected. ECM proteins are often highly cross-linked and cannot be easily solubilized, thus limiting trypsin accessibility. To overcome this, we and others previously proposed that hard-to-digest ECM proteins would benefit from sequential multi-proteases digestion (in our case LysC + trypsin)

3

[17,29]. Last, lower abundance proteins found in the ECM (proteins we have previously grouped under the term "matrisome-associated and including ECM-affiliated proteins, ECM regulators, and ECM-bound secreted factors [17,30]) tend also to be smaller, and so generate less peptides and, consequently, are less likely to be identified. Yet, these proteins are critical regulators of ECM homeostasis and functions. We and others have contributed to the enhancement of proteomic methods, however, until now, the focus has remained on improving protein identification and increasing the number of proteins identified [14,31,32].

Moreover, like all other proteins, ECM proteins exist in multiple proteoforms arising from alternative splicing for isoforms, from single-nucleotide variants, or from post-translational modifications [33,34] and these proteoforms can play important roles in health and disease. While proving the presence of a protein in a given sample only needs a few unique peptides, capturing a specific proteoform requires the detection of unique peptide sequences that cover all the variation and modification sites of a given protein, which often necessitates higher sequence coverage. Of note, the combined average sequence coverage of ECM proteins in MatrisomeDB, a database compiling ECM proteomic datasets [35], is 36.85% with a median sequence coverage <30%. The addition of new datasets to MatrisomeDB contributes little to improving overall sequence coverage, as the same trypsin-based proteomics protocol simply repeats previous observations. This implies that conventional "identification-oriented" proteomic workflows cannot meet the need for studies of ECM proteoforms. Instead, we propose to steer the development towards "coverage-oriented" approaches to increase the sequence coverage of ECM proteins and thus obtain more information on ECM proteoforms.

Last, certain critical aspects of ECM biochemistry remain largely understudied, including folding and ECM protein-protein interactions, in the context of the assembled and insoluble ECM meshwork, as opposed to studying individual ECM proteins or fragments in soluble forms. Yet, we know that ECM protein folding and the nature of the protein complexes formed in the ECM are of paramount importance to achieve proper ECM function.

Limited proteolysis coupled to proteomics has emerged as a powerful method to elucidate protein structures and conformational changes, and to map sites of protein-protein interactions [36,37]. However, until now, this approach has mainly been used to study soluble proteins, as they provide omnidirectional access for proteases. Here, we used limited proteolysis to devise a time-

4

lapsed digestion pipeline to study the ECM with two goals: 1) to release peptides *sequentially* over time and thus generate peptide samples of lower complexity, with the goal of enhancing protein identification and protein coverage, 2) since ECM proteins are highly insoluble and can only be partially solubilized prior to tryptic digestion [17,38] (*see also* Figure 1A), a time-lapsed tryptic digestion will result in the release of peptides from more exposed or accessible regions of proteins at earlier timepoints, and of peptides from buried protein domains as the digestion progresses. As a result, the mapping of peptides obtained via time-lapsed tryptic digestion can provide important information on protein folding and possible sites of protein-protein interactions.

We further demonstrate here that this novel approach can shed light on the functions of under-studied ECM proteins by using as a case study the novel ECM protein SNED1 (Sushi, Nidogen and EGF Like Domains 1). We previously identified SNED1 in a proteomic screen comparing the ECM of poorly and highly metastatic mammary tumors [39]. Functionally we showed that SNED1 acts as a metastasis promoter, since knocking down SNED1's expression significantly decreased metastatic dissemination. We also showed that the organization of fibrillar collagens within the tumor ECM was altered upon SNED1 knockdown [39], which led us to postulate that SNED1 may modulate ECM architecture. However, the precise mechanisms by which it could do so remain unknown. Using immunofluorescence microscopy, we have recently reported that SNED1 forms fibrils within the ECM [40]. Using, molecular modeling, we have predicted the interactome of SNED1 and found that SNED1 could potentially interact with 40 ECM components, including fibronectin, nidogens 1 and 2, and this may serve as the basis for SNED1's incorporation in the ECM [40]. However, none of these interactions have been validated experimentally yet, and we still do not know the mechanisms underlying SNED1's assembly in the ECM. Here, by applying time-lapsed proteomics to profile ECM produced *in-vitro* by cells overexpressing SNED1, we show that SNED1's expression and presence in the ECM modulates ECM protein composition. Interestingly, we also show that SNED1 increases the accessibility of some ECM proteins to trypsin resulting in an increase sequence coverage for these proteins (*e.g.*, thrombospondins 1 and 2, Efemp1, Efemp2, fibrillin 2) but also decreased accessibility of other ECM proteins to trypsin, resulting in a decrease sequence coverage for these proteins (*e.g.*, nidogen 2, Tinagl1, Tgm2). While the mechanisms by which SNED1 modulates ECM composition and

ECM protein folding remains to be discovered, this proof-of-concept study highlights the ability of time-lapsed proteomics to shed light on novel or unsuspected roles for ECM proteins.

RESULTS

Time-limited proteolysis increases protein identification and coverage

To test the impact of modifying trypsin digestion duration on protein identification, we chose to work with ECMs produced by mouse embryonic fibroblasts *in vitro* (Figure 1A). The generation of cell-derived matrices (CDMs) after decellularization of fibroblast cultures is highly reproducible [41] and thus offers a robust system to test and optimize ECM proteomic workflows. Once CDMs are obtained, they are partially solubilized and pre-digested with Lys-C (Figure 1A). CDMs are then either digested into peptides using a standard 18-hour tryptic digestion protocol [17,38] or using time-limited digestion (30 minutes, 2 hours, or 4 hours; Figure 1B). Peptides obtained were then characterized using LC-MS/MS (Supplementary Table 1, Supplementary Table 2). Overall, we observed that, time-limited digestions resulted in the highest absolute number of overall and matrisome spectra, while the proportion of matrisome spectra remained equivalent (Supplementary Tables 1 and 2).

In terms of protein identification, we detected, 77, 79, 88, and 76 ECM proteins across two biological replicates using 30-min, 2-hour, 4-hour, and 18-hour digestions, respectively (Figures 1C-D, Supplementary Tables 1 and 2). Across all digestions, about 2/3 of the proteins belong to the core matrisome (ECM glycoproteins, collagens, proteoglycans) while 1/3 are matrisome-associated proteins (ECM-affiliated proteins, ECM remodeling enzymes, or secreted factors). Comparison of the proteins identified at the different timepoints reveals that 68 proteins were identified across all timepoints (Figure 1D, Supplementary Table 3). In addition, 2 proteins (Col4a3 and Fgf2) were uniquely identified by the 30-minute digestion, while 9 proteins were uniquely identified by the longer 2-hour and 4-hour digestions (Mfge8, Vwa1, Thbs2, Lum, Anxa4, Anxa5, Loxl3, Egfl7 and S100a13), and an additional 6 only using the 4-hour digestion protocol. While the standard 18h digestion protocol uniquely identified 2 proteins (annexin 5 and S100a13), it resulted in an overall lower number of proteins identified (76 vs. 87 with the 4-hour digestion).

6

Beyond protein identification, we next sought to evaluate the impact limited proteolysis on sequence coverage. We found that limited digestions resulted in an increase in the percentage of sequence coverage for all categories of matrisome proteins as compared with a standard 18hour digestion (Table 1, Supplementary Table 2).

Sequential tryptic digestion as a mean to fractionate peptides and further improve protein identification and sequence coverage

We next sought to perform a time-lapsed digestion where peptides are collected sequentially, at different timepoints, over an 18-hour period (Figure 2A). To compare the two digestion protocols (single timepoints vs time-lapsed), we first evaluated their output in terms of protein identification. Using the time-lapse digestion protocol, we identified 76, 75, 67, and 41 ECM proteins in two biological replicates, at the 30-min, 2-hour, 4-hour, and 18-hour timepoints, respectively (Figure 2B-C). We also observed that the absolute number of spectra and peptides decreased at later timepoints of the time-lapse digestion (Supplementary Table 1). This is expected, since easy-to-digest peptides are depleted from the samples at the earlier timepoints of the sequential digestion.

Aggregation of the data from the sequential digestions identified an average of 87 ECM proteins (Figure 2B, *Aggregated*), as compared to the 76 proteins identified using the standard 18-hour digestion. The comparison of the proteins identified in two biological replicates showed a smaller overlap of 37 proteins across all timepoints (Figure 2C) compared to that of the single timepoint digestions (Figure 1D). This overlap consists of 17 ECM glycoproteins, nine collagens, two proteoglycans, five ECM-affiliated proteins, and 4 ECM regulators (Figure 2C, Supplementary Table 3). 24 matrisome proteins were identified in samples from the first three timepoints, ten matrisome proteins in samples from the first two timepoints. We also found that the first timepoint of the time-lapse digestion (30 minutes) contributed the most to the aggregate list of ECM proteins identified as an additional five matrisome proteins (Grem1, Ctsz, P4ha2, Itih1, and Angptl2) were uniquely identified at this timepoint (Figure 2C, Figure 2F).

Of note, proteins that tended to vary more between protocols, timepoints, and replicates belonged to the matrisome-associated classification [17]. These proteins are expected to be present in lower abundance in CDMs than the more structural core matrisome proteins. For this reason, a short digestion cannot singlehandedly capture the complexity of CDM composition as it misses

7

proteins present in lower abundance. This is further demonstrated by the five-way comparison of datasets obtained with the four timepoints of the time-lapsed digestion or the standard 18-hour digestion (Figure 2D, Supplementary Table 3). The comparison yielded 36 matrisome proteins identified in all samples. These consisted of 17 ECM glycoproteins, 9 collagens, 2 proteoglycans, 4 ECM-affiliated proteins, and 4 ECM regulators. It also showed how later timepoints of the time-lapse digestion are contributing complementary datapoints to the 30-minute digestion (Figure 2E, Supplementary Table 3). The other timepoints captured four matrisome proteins identified using the standard digestion protocol that otherwise would have been missed, and further add the identification of four matrisome proteins that the standard digestion did not identify. We therefore defined the aggregate of identified proteins from the sequential digestion as the union of four timepoints, which creates a list greater than that of the classical digestion, with 13 uniquely identified proteins (Figure 2E).

To further refine the characterization of the matrisome produced by MEFs *in vitro*, we compared the relative abundances of the proteins identified using normalized spectral abundance factor (NSAF), a semi-quantitative method based on spectral counts normalized by protein size [42,43]. The most abundant matrisome proteins found in MEF CDMs are fibronectin (Fn1), fibrillin 1 (Fbn1), fibulin 2 (Fbln2) and the elastin microfibril interface-located protein 1 (Emilin 1) for the glycoproteins and type I and type VI collagens (Figure 2F, Supplementary Table 7). Interestingly, we also detected basement membrane proteins including nidogens 1 and 2 (Nid1 and Nid2), perlecan (Hspg2) and type IV collagens.

As anticipated, we observed that the abundances of most proteins decreased with time during the time-lapsed digestion (Figure 2B, Supplementary Table 1, Supplementary Table 2). We also observed that the relative abundance of some proteins increased with longer digestions (Tgfbi, Tgm2, Col1a1, Col1a2, Col18a1, Lgals1, Anxa5, Anxa4). This may suggest that these proteins require more extensive denaturation and digestion to release peptides.

Time-lapsed digestion increases matrisome protein sequence coverage

After evaluating the impact of the two digestion protocols on protein identification, we sought to compare the methods in terms of protein sequence coverage, in other word, the percentage of a given protein sequence matched (or "covered") by peptides detected by mass spectrometry. We found that proteins showed higher coverage at earlier digestion timepoints

(Supplementary Figure 1A). In particular, average coverage of proteins with peptides detected at the 2-hour timepoint of the sequential digestion was the highest of the four timepoints and surpassed that of the standard 18-hour digestion (Supplementary Figure 1B). However, by combining the identified peptide sequences of earlier timepoints with those of the next timepoint (aggregation), we observed that the coverage was significantly higher over time (Figures 3A, 3B, Supplementary Figure 1B). This increase in coverage over time for most proteins suggests that some of the peptide sequences identified are unique to the later timepoints of the time-lapsed digestion, thus increasing the cumulative coverage with time (Figure 3A).

To evaluate whether a particular matrisome category of proteins benefitted more from this approach, we compared the protein coverages obtained using the standard 18-hour digestion or after aggregating coverage data of the time-lapsed protocol. We observed a shift in the distribution of coverages particularly for ECM glycoproteins ($p < 0.001$), ECM-affiliated proteins ($p < 0.01$), and ECM Regulators ($p < 0.01$) (Figure 3B). The increase in cumulative coverage is most noticeable with the ECM-affiliated proteins and ECM regulators, which are less abundant, and therefore would have benefitted the most according to our hypothesis (Figure 3A, 3B).

To rule out that the increase in the sequence coverage seen upon data aggregation was due to the increase in the number of datasets collated, we aggregated four datasets obtained using the standard 18-hour digestion method (*i.e.,* four files) and compared the average aggregated coverage to the average aggregated coverage obtained by integrating data from the four distinct digestion timepoints (Supplementary Figure 1C). This confirmed that the increased percentage of sequence coverage is indeed due to the benefit of the time-lapsed digestion method and the release of different peptides over time, and not the collation of datasets.

To evaluate the contribution and potential benefit of each timepoint to the increased coverage attained with the time-lapsed digestion protocol, we calculated the coverage with increasing extent of aggregation—30-min to 2-hour, 30-min to 4-hour, 30-min to 18-hour—and evaluated their correlations to one another (Supplementary Figure 2). While the differences between the cumulative aggregated coverage of 4-hour and that of the 18-hour were minimal, as denoted by Pearson's correlation coefficient ($r > 0.97$), the greatest differences were seen between the 30-min unaggregated coverages and the cumulative aggregated coverage of 18-hour ($r = 0.93$) (Supplementary Figure 2). Upon calculation of cumulative coverage, we found that the highest coverage was achieved for fibronectin (76 %, while fibronectin's coverage was of ~60% using an

18-h standard digestion protocol, *see* Table 2 and Supplementary Table 2), followed by perlecan (60 %, while fibronectin's coverage was of ~46% using an 18-h standard digestion protocol, *see* Table 1 and Supplementary Table 2).

To further identify the proteins that benefitted the most from the increase in coverage achieved by employing a time-lapsed digestion protocol, we calculated the ratio of aggregated coverage to standard digestion coverage, for each protein (Table 2). We found that type XI and type XVI collagens (Col11a1 and Col16a1), the proteoglycan prolargin (Prelp), and several fibulins (Fbln7, Efemp1, Fbln5) saw the largest increase in coverage. These changes in coverage were further plotted with respect to protein molecular weight, to evaluate whether a protein's size affected the observed coverage increase (Figure 3C, Supplementary Table 2). While a greater number of ECM glycoproteins saw an improvement in their coverage (~50 % of all identified ECM glycoproteins), the extent of improvement was similar to those of the proteoglycans, ECM-affiliated proteins, and ECM regulators (Figure 3C, Table 2). The overlap of the distribution with respect to size and coverage observed between the most and least affected proteins suggested no particular effects of these parameters on the increased coverage upon data aggregation (Figure 3C).

In conclusion, we demonstrate that, while performing time-lapsed tryptic digestion and aggregating data results in an only moderate increase in protein identification, it permits a more accurate estimation of relative protein abundance and leads to a significant increase in protein sequence coverage. Because the time-lapsed digestion protocol results in the release of peptides over time, we can also obtain information on protein digestibility and accessibility and then infer information about protein folding. The first draft of the MEF matrisome protein accessibility map provided here, thus represent a first step towards obtaining a detailed topographical map of ECM.

Identification of SNED1-dependent changes in ECM protein composition and abundance

Our next step was to test whether this new protocol could help us gain new knowledge on the ECM. We previously reported changes in the organization of the ECM of mammary tumors in presence of SNED1 which correlated with tumor's metastatic potential [39]. More recently, we have shown that SNED1 is incorporated in fibrillar structures within the ECM meshwork (Figure 4A, [40]), yet, we do not know how SNED1 assembled into an ECM and how it modulates ECM architecture. We thus sought to use the newly developed time-lapsed proteomic workflow

presented here to gain insight into the possible roles of this protein on ECM composition and ECM protein abundance and folding. To do so, we analyzed CDMs produced by *Sned1*^{KO} MEFs over-expressing GFP-tagged SNED1 and compared it to the CDMs produced *Sned1*^{KO} MEFs over-expressing GFP alone (control; *see above*). Quality-control analysis of the dataset obtained, on the model of what we presented above, revealed a similar pattern in terms of protein identification using single timepoint digestions or time-lapsed digestion (Supplementary Figure 3, Supplementary Tables 4 and 5).

As described above, we could define the composition of CDMs containing or not SNED1 as the list of proteins at the intersection of two biological replicates for each condition (Supplementary Table 6) and identified proteins whose detection changed as a function of SNED1's expression. Doing so revealed that the α3 chain of type IV collagen (Col4a3), the α1 chain of type VIII collagen (Col8a1), and Itih1 were not detected in the control CDMs, while S100a6, was only identified in SNED1-containing CDMs (Supplementary Table 6).

We next sought to evaluate whether the presence of SNED1 altered the abundance of certain ECM proteins. To do so, we calculated the NSAF for all proteins detected in SNED1-containing CDMs (Supplementary Figure 3G, Supplementary Table 7) and compared these values with those obtained for control CDMs (Figure 4B, Supplementary Table 7). We found that type I collagen (chains α1 and α2), the glycoproteins thrombospondin 1 and periostin (Postn), the proteoglycan biglycan (Bgn), and annexin A4 were more abundant in SNED1-containing CDMs. In contrast, the glycoproteins Mfap2, Tinagl1, Mgp, and Nid2 and the Anxa5, Serpinh1, were present in lower abundance, in SNED1-containing CDMs as compared to control.

Changes in ECM protein profiles upon expression of SNED1

To gain insight into the possible role of SNED1 in modulating ECM architecture, we exploited the coverage data obtained by proteomics. Indeed, protein digestibility by trypsin is directly linked to how accessible lysine and arginine residues are, and this is linked to protein folding and protein-protein interactions, two parameters that can alter exposure of these residues. To capture this, we defined a dissimilarity index reflecting the magnitude of the differences in coverage (Figure 4C, Table 3, *see Materials and Methods*). Using this index, we identified proteins with increased coverage in CDMs containing SNED1, including thrombospondins 1 and 2 (Thbs1 and Thbs2) and EGF containing fibulin extracellular matrix protein 2 (Efemp2, also called fibulin

11

4), while proteins with decreased coverage in SNED1-containing CDMs included of emilin1, Tinagl1, nidogen 2, galectin 9, and annexin A5 (Figure 4D).

Since an increase in protein abundance could result in an artificial increase in coverage, we next sought to identify proteins detected with increased coverage without differential abundance. With this approach, proteins such as Fbln1, Col4a1 can be considered as those whose increase or decrease in coverage, respectively, is independent of their relative abundance. On the other hand, proteins such as Col1a1, Col1a2, Fbln2, Fbn1, Col2a1 are identified with increased abundance without increase in coverage.

Altogether, our results support the conclusion that SNED1's expression modulates the composition and coverage of the ECM produced by fibroblasts. It will be interesting in the future to identify the mechanisms leading to these changes (*see Discussion*).

Sequential release of peptides provides insight on protein accessibility to trypsin

Perhaps the most significant innovation brought by the time-lapsed digestion protocol presented here is its spatio-temporal resolution. As discussed above, protein digestibility reflects the surface exposure of a protein, and is directly linked to protein folding and protein-protein interactions. Thus time-lapsed peptide mapping can identify protein regions readily exposed and accessible to trypsin (peptides matching these regions would be released at early digestion timepoints) and regions of proteins hard to access, either because of extensive post-translational modifications, such as O-glycosylations hindering accessibility, or because the engagement of protein-protein interactions that are not disrupted by the partial denaturation to which samples are submitted.

Digestion kinetics and topographical information of SNED1

As of today, we do not have any experimentally-validated information on the folding of SNED1 when incorporated in the ECM. To obtain topographical information on SNED1, we assessed the diversity of peptides generated over the course of the time-lapsed digestion. Overall, we found that the aggregated SNED1 coverage yielded 27.32 %, a significant increase from the 13.8 % coverage obtained using the 18h standard digestion protocol (Supplementary Table 4). It is also a substantial increase from the ~5% coverage obtained by aggregating data from MatrisomeDB, a database reporting proteomics data on insoluble ECMs from tissues [35], and an increase from the 24.13% coverage reported in ProteomicsDB, with the caveat that this database

12

compiles global proteomic datasets and that for the most part includes proteomic analysis of soluble protein samples [44]. We observed that the number of peptide spectrum matches (PSMs) was the highest for the 30-minute timepoint of the time-lapse trypsin digestion. The most peptides at this timepoint were in the CCP/Sushi domain (SMART SM000032) and in the first FN3 (SMART SM000060) domain, followed by peptides in the NIDO domain (SMART SM000539) (Figure 5A, Supplementary Table 8). Over time, the number of PSMs decreased, yet we found that each timepoint contributed unique peptides. In particular, we saw that a large number of peptides mapped to the second FN3 domain at the later digestion timepoints, with six and five PSMs for the 4-hour and 18-hour timepoints, respectively, which suggests that this domain gets exposed over time as the tryptic digestion progresses.

While the structure of SNED1 has not been resolved yet, we can turn to molecular modeling to gain insight into the 3-dimensional folding of the protein. Here, we retrieved the 3-D structure of SNED1 predicted by AlphaFold [45] (https://alphafold.ebi.ac.uk/entry/Q8TER0) and mapped, in a color-coded manner, the peptides identified at each timepoint of the time-lapsed digestion (Figure 5B, Supplementary Table 8, Supplementary File 2). While the folding of some regions of the protein is only predicted with low confidence and while it is likely that SNED1's folding is different when it is assembled in the ECM, this is a first attempt at overlaying experimental data obtained by time-lapsed proteomics on a theoretical model of an ECM protein.

Peptide mapping of nidogen 2 whose coverage decreased upon SNED1 presence

To illustrate how SNED1 presence in the ECM influenced the coverage of ECM proteins, we performed a similar peptide mapping for nidogen 2. We found that Nidogen 2 (Nid2) had a higher sequence coverage in the CDMs lacking SNED1 (Figure 4C, Figure 6A, Table 3). The G2F domain was identified with the highest number of peptides, with 45 in the SNED1-lacking CDM and 36 in the SNED1-containing CDM. This was followed by the NIDO domain which was identified with 33 spectra in the SNED1-lacking CDM and 21 spectra in the SNED1-containing CDM. The 3-D model of Nid2 predicted by AlphaFold was further used to overlay peptide identified at each timepoint of the time-lapsed digestion. This revealed the precise location of peptides detected differentially in presence or absence of SNED1 (Figure 6B, Supplementary Files 3 and 4). We can speculate that peptides released at later timepoints of the time-lapses digestion are part of protein domains not as readily accessible to tryptic digestion, and possibly because

13

these domains are involved in protein-protein interactions, protecting them from tryptic digestion. Interestingly, nidogen 2 was one of the proteins we predicted to interact with SNED1 by molecular modeling [40], supporting the possibility that protein-protein interaction between SNED1 and nidogen 2 is leading to the decreased coverage of nidogen 2 in presence of SNED1. The fact that the presence of SNED1 modulates trypsin accessibility of several ECM proteins, suggests that this novel protein of the ECM plays a key role in modulating the assembly and organization of the ECM.

Peptide mapping of thrombospondin 1 whose coverage increased upon SNED1presence

In contrast with what we observed for nidogen 2, thrombospondin 1 saw its sequence coverage increased when SNED1 is present in the ECM (Figure 4C, Table 3, Figure 7A). We found that the thrombospondin N-terminal like domain contained the highest number of PSMs, 102 in SNED1-containing CDMs, compared to 78 counts in SNED1-lacking CDMs. To better capture differences, on the model of what has been proposed by Eckersley and others [Eckersley et al 2020], PSM counts were normalized by the amino-acid length of each domain (Figure 7A).

The 3-D model of Thbs1 predicted by AlphaFold was further used to overlay peptide identified at each timepoint of the time-lapsed digestion (Figure 7B, Supplementary Files 5 and 6). The resulting annotation showed a large overlap of peptide sequences identified across the different timepoints, revealing that most of the peptides were obtained with a short 30-minute digestion. Further analysis of the distribution of the peptides identified in SNED1-lacking CDMs with those identified in the SNED1-containing CDMs, revealed their precise topography (Figure 7B, white arrows). One hypothesis to explain how the presence of SNED1 in the ECM results in an increased in the sequence coverage of a protein, like thrombospondin 1, is if it directly or indirectly induced a conformational change on this protein, either by binding to it, or by modulating the forces exerted on it, which would, in turn, increase the exposure of this protein to trypsin. Future studies will be aimed at exploring this hypothesis.

DISCUSSION

In this study, we report the development of time-lapsed digestion coupled to mass spectrometry to enhance ECM protein identification and sequence coverage, and gain insight into

ECM protein folding in the context of an assembled ECM meshwork. By comparing MS output obtained from time-limited tryptic digestions, we found that shorter digestions are sufficient to identify the same number, if not more, than the gold standard overnight trypsin digestion protocol. We further demonstrate the significant benefit of conducting time-lapsed digestion and aggregating output in terms of protein identification and coverage over a standard overnight trypsin digestion protocol. Last, we highlight how time-lapsed proteomic data can be leveraged via 1-D and 3-D peptide mapping to gain insights into ECM protein folding and identification of possible sites of protein-protein interactions.

Modulation of the duration of the trypsin digestion is only one of the parameters of a multi-step proteomic workflow that can be modified. While we demonstrated its benefits, in particular in terms of sequence coverage, there is still a dark side to the matrisome that has yet to be illuminated. We were the first to propose and show that hard-to-digest ECM proteins would benefit from sequential multi-proteases digestion (in our case LysC + trypsin) [17], this combination is now broadly adopted. On the model of what has been developed to achieve higher coverage of the intracellular proteome [46–49], and based on the findings reported here, we propose that a sequential (or simultaneous) digestion with *multiple* proteases, beyond the use of the LysC-trypsin pair, could achieve a deeper matrisome coverage. However, this poses challenges including the optimization of reaction conditions to ensure protease compatibility and downstream computational analysis to support multiplexed digestion for accurate peptide identification.

The complex structure of the ECM meshwork mediated by protein-protein interactions that are dependent on proper protein folding, is crucial to its function. Conventional structural proteomic approaches such as cross-linking mass spectrometry (XL-MS, [50]) are not suitable to study ECM as crosslinking reagents will further harden already hard-to-digest ECM proteins. On the other hand, classical limited-proteolysis-coupled mass spectrometry method previously developed is devised to handle soluble proteins [37]. While the approach developed here can shed light on proteins' accessibility in the assembled state of the ECM, translated as the proteins' digestibility by trypsin, it is worth noting that it is applied here to samples that have been partially denatured. Our next step will be to attempt to work on native ECM samples to conduct a true topographical survey of the ECM using proteomics. In addition, in native conditions, protein

domains engaged in protein-protein interactions will be inaccessible and as a result resistant to proteolysis. Peptide mapping should thus reveal, more directly, a map of potential protein-protein interaction sites at the ECM scale.

As a proof-of-concept, we applied the newly devised time-lapsed digestion protocol to compare the ECM produced by *Sned1*^{KO} or SNED1-overexpressing mouse embryonic fibroblasts. The *Sned1* gene was clones nearly two decades ago [51], yet, until very recently, we did not know anything about the physiological or pathological roles of SNED1. We became interested in studying SNED1 after discovering its role as a promoter of mammary tumor metastasis [39]. More recently, we reported the generation of a *Sned1* knockout mouse model and its essential role during development, since *Sned1* knockout led to early neonatal lethality, likely due to craniofacial malformations and its involvement in neural crest cell fate [52]. Based on sequence analysis, we categorized SNED1 as part of the core matrisome [17,30], before we knew that it could incorporate in the ECM and constitute one of its structural components, something we recently demonstrated experimentally [40]. However, how SNED1 assembles in the ECM and how it governs different pathophysiological processes remain unknown. Early observations looking at the ECM of tumors expressing or not SNED1suggested that SNED1 may play a role in modulating the tumor ECM [39]. We thus sought to use the newly developed time-lapsed proteomic pipeline to test this. Here we showed that overexpression of SNED1 by mouse embryonic fibroblasts indeed modulated ECM composition, as observed when calculating the relative abundance ECM proteins in cell-derived matrices produced by *Sned1*^{KO} or SNED1-overexpressing cells, and ECM protein accessibility to trypsin. Interestingly, several proteins whose abundance and/or whose coverage changed as a function of SNED1 are also predicted to be possible SNED1 interactors. These include fibronectin, and nidogen 2. Future work will be aimed at understanding the mechanisms leading to these changes and determine if, and if so, how, they relate to the pathophysiological process in which we have implicated SNED1.

Our view of the ECM has significantly changed over the past decades, from an inert scaffold to a remarkably complex and dynamic signaling platform. We now know that the ECM is actively remodeled over time during the processes of development, tissue aging, wound healing, and disease progression [4,7,8,10,12]. Changes in cell-ECM interactions, microenvironmental cues (*e.g.,* tissue oxygenation levels), and environmental factors (*e.g.,* smoke, UV) induce

biochemical (*e.g.,* glycation [53]) or structural (*e.g.,* degradation, cross-linking) post-translational modifications, leading to alteration of the structure of individual ECM proteins and that of the ECM meshwork. Bottom-up proteomic and peptide mapping have contributed to appreciate and evaluate the extent of these changes for example during aging and photoaging of the skin [19–21,27,54], fibrosis [11,26,27], or cancer progression [55], but much more remains to be discovered. As demonstrated here, we propose that applying time-lapsed proteolysis coupled to mass spectrometry can contribute to unravel the compositional and structural complexity of the ECM, a necessary first step toward deciphering ECM-dependent pathophysiological processes and the development of ECM-targeting therapeutic strategies.

EXPERIMENTAL PROCEDURES

Cell culture and generation of cell-derived matrices (CDMs)

Mouse embryonic fibroblasts (MEFs) were previously isolated from embryos of $Sned1^{LacZ\text{-}Neo/LacZ\text{-}Neo}$ ($Sned1^{KO}$) mice [52]. MEFs were first immortalized and then transduced with retroviral constructs to achieve stable expression of either the SNED1_GFP fusion protein (SNED1_GFP) or GFP alone (control), as previously described [40]. For all experiments, cells were cultured in DMEM (Corning, #10-017-CV) supplemented with 10 % fetal bovine serum (Sigma), and 2 mM glutamine (Corning).

To obtain cell derived matrices (CDMs), cells were cultured as described by Franco-Barraza, et al. [41]. In brief, cells were plated in 6-well plates at 475,000 cells/well. Upon cell reaching confluency (within 24 to 36 hours post seeding), the culture medium was replaced with fresh medium supplemented with 0.05 µg/mL of ascorbic acid (Sigma). Half of the medium was then replaced with fresh medium containing 0.1 µg/mL of ascorbic acid, every other day, for 7 days following the first ascorbic acid treatment. After 7 days of ascorbic acid treatment, cells were washed twice with warm D-PBS containing calcium and magnesium (Corning), and the ECM produced by the cells was decellularized using warmed extraction buffer (0.5 % Triton X-100 and 20 mM NH_4OH in PBS) for 6 minutes. Decellularization efficiency was assessed by microscopic observation. The buffer containing cellular lysate was then removed and replaced with cold PBS. The obtained CDMs were cut in half using a scalpel and transferred from the cell culture plate to a conical tube. The CDMs were washed in PBS under constant inversion at 4 °C for 4 hours. The

17

PBS was replaced after a centrifugation at 4,000 rpm for 8 minutes at 4 °C, and the wash was repeated once for 4 hours and then for 12 hours. Last, CDMs were pelleted by centrifugation at 13,000 rpm for 2 minutes at 4 °C, and stored at -80 °C.

For immunofluorescence visualization, cells were cultured on a glass coverslip coated with crosslinked gelatin as described by Harris *et al.* [56]. After decellularization (*see above*), CDMs were fixed with 4 % paraformaldehyde, blocked with 1 % BSA in 0.05 % Triton X-100, and incubated first with a polyclonal rabbit anti-SNED1 antibody (used at a 1 μg/mL concentration, previously described [40]) and a monoclonal mouse anti-GFP antibody (Abcam #ab1218, used at a 1:500 dilution), and then with a goat anti-rabbit antibody conjugated to AlexaFluor568 (Thermo Fisher A11036, used at 1:1000) and a goat anti-mouse antibody conjugated to AlexaFluor647 (Thermo Fisher A21236, used at 1:1000) for visualization. Coverslips were mounted using Fluoromount-G (Southern Biotechnology) and imaged using Zeiss confocal microscope LSM880.

Protein solubilization and digestion into peptides

Each CDM pellet was resuspended in 100 mM NH_4HCO_3 solution containing 8 M urea and 10 mM dithiothreitol (DTT; Thermo Fisher) and incubated at 37 °C for 2 hours under agitation, for partial solubilization and reduction. Reduced thiol groups were alkylated using 40 mM iodoacetate (IAA; Thermo Fisher) for 30 minutes at room temperature in the dark. Urea concentration was brought to 2 M with 100 mM NH_4HCO_3 and samples were then de-glycosylated with PNGase F (New England Biolab), under agitation at 37 °C for 2 hours, and pre-digested with Lys-C (Thermo Fisher), under agitation at 37 °C for 2 hours, as previously described [17,38].

Single timepoint tryptic digestion

For single timepoint digestion (n = 4 for 2-hour and 18-hour timepoints; n = 2 for 30 min and 4-hour timepoints), trypsin (Thermo Scientific) was added, and tryptic digestion was performed under agitation at 37°C for 30 minutes, 2 hours, 4 hours, or 18 hours. Upon completion of each timepoint, samples were centrifuged at 13,000 rpm for 5 minutes to separate any undigested CDM material, and peptide-containing supernatants were transferred to a clean tube, acidified using 50% trifluoroacetic acid (TFA) and stored at -20 °C.

Time-lapsed tryptic digestion

For time-lapsed tryptic digestion, trypsin was added following the pre-digestion, and the tryptic digestion was continued under agitation at 37 °C (n = 2). After 30 minutes, samples were

centrifuged at 13,000 rpm for 5 minutes, and the supernatants were transferred to a clean tube, while the remaining proteins were resuspended in 2 M urea in 100 mM NH_4HCO_3, and the digestion continued at 37 °C with addition of equal amount of trypsin. After 2 hours since the initial trypsin addition (or 1.5 hour after addition of the second batch of trypsin), samples were centrifuged, and the process was repeated, and the 4-hour and 18-hour timepoint samples were obtained. Samples were acidified with 50% TFA upon timepoint completion and stored at -20 °C until further processing.

Acidified peptides were desalted on C18 desalting columns (Thermo Scientific) following the manufacturer's instructions and eluted in 50 % acetonitrile (ACN) solution containing 0.1 % TFA. Peptides were lyophilized for 4 hours and reconstituted in 5 % ACN solution containing 0.1 % formic acid. For all samples, peptide concentration was measured using a colorimetric peptide quantification assay (Thermo Scientific). Finally, the concentration of the samples was adjusted to 250 ng/µL.

LC-MS/MS analysis

Data acquisition

1µg-equivalent of peptides from each sample was dissolved in 15uL 0.1% formic acid and injected and separated by a capillary C18 reverse-phase column of the mass-spectrometer-linked Ultimate 3000 HPLC system (Thermo Fisher) using a 90-minute gradient (5-85% acetonitrile in water) at a 300 nL/min flow rate. Mass spectrometry data were acquired on a Q Exactive HF mass spectrometry system (Thermo Fisher) with nanospray ESI source using data-dependent acquisition.

Database search

Mass spectrometry raw files were converted into .mzML files using msConvert [57], searched and filtered by MSFragger [58] (version 3.2) and Philosopher [59], respectively. The database search was performed against a custom proteome reference database generated by amending standard mouse reference proteome (UniProt Mus musculus C57BL/6J proteome last modified on March 7, 2021, and including 110,926 entries, 55,463 forward sequences and 55,463 decoy sequences) with human SNED1's protein sequence (UniProt Q8TER0). Trypsin was specified as cleavage enzyme allowing 1 missed cleavage, mass tolerance was set to 50 ppm for precursor ions and 20 ppm for fragment ions. The search included variable modifications of methionine oxidation, lysine

and proline hydroxylation [60], asparagine, and glutamine deamidation, glutamine to pyroglutamate, and fixed modification of carbamidomethyl cysteine. Peptide length was set within a range between 7 to 50 amino acids and false discovery rate (FDR) was set to 1% at the peptide/spectrum match (PSM), peptide, and protein levels.

Data analysis

We restricted our analysis to reviewed UniProt entries and annotated the dataset to identify ECM and ECM-associated proteins according to the matrisome nomenclature [17,30]. For protein identification, we further restricted our analysis to proteins detected with at least 2 peptides. Proteome coverage was calculated from mapping all identified peptides to custom proteome reference sequences described above.

Semi-quantitative analysis of protein abundance

For each protein, the normalized spectral abundance factor (NSAF) was calculated as described by Zybailov et al. [42]. In brief, NSAF is calculated as the number of spectral counts (SC) identifying a protein, divided by the protein's length (L), divided by the sum of SC/L for all proteins of the dataset [43].

Analysis of sequence coverage

To investigate the aggregated coverage at a given timepoint of the time-lapsed tryptic digestion, identified peptides from previous and current timepoints were pooled together prior to peptide mapping and coverage calculation (*e.g.*, aggregated coverage at 18-hour was calculated based on all identified peptides from the sequential 30-minutes 2-hour, 4h-hour and 18-hour digestions). Change in sequence coverage when comparing the 18-hour stand digestion method and the data aggregated from the time-lapsed digestion method was calculated as follows:

$$Sequence\ coverage\ improvement = \frac{Seq_{Aggregate}\ coverage - 18h\ coverage}{18h\ coverage} * 100$$

Dissimilarity index was calculated as follows:

$$Dissimilarity\ index = \frac{Euclidean\ distance}{Cosine\ Similarity} = \frac{\sqrt{\sum_{i=1}^{n}(Qi - Pi)^2}}{\frac{\boldsymbol{Q} \cdot \boldsymbol{P}}{||\boldsymbol{Q}||\ ||\boldsymbol{P}||}}$$

Where n = 4 (4 independent timepoints). Q and P represent the vectors consisting of 4 aggregated coverage in control or SNED1-overexpressing samples, respectively.

Statistical analysis

The data used for comparison between the coverages in proteins grouped by ECM categories was first checked for normality using the Shapiro-Wilk test [61]. Considering the non-normality of the data, the distribution of standard 18-hour digestion coverages and the aggregate coverage counterparts were compared using Wilcoxon signed-rank test, with p-values reported following Bonferroni adjustments for stringent results [62].

Data visualization

Proportional 2-circle Venn diagrams were generated using the Venn Diagram Generator: http://barc.wi.mit.edu/tools/venn/.

Visualization of the intersection of multiple datasets was performed using UpSetR [63] and the online application: https://gehlenborglab.shinyapps.io/upsetr/.

Visualization of the coverages and NSAF values was performed using Morpheus: https://software.broadinstitute.org/morpheus.

Peptide mapping on 1D representation of protein structures

The information of domain start and end positions were obtained from the Simple Modular Architecture Research Tool (SMART [64]). Each peptide was assigned the domains using a heuristic created in an Excel file that determined where the start and end position of the identified sequences belonged to (*see Supplementary File 1*). The outliers that were matched onto the non-domain regions of the protein were manually excluded in the count of spectra count ratio on the model of what has been described by Eckersley [65].

Peptide mapping on 3D representation of protein structures

The pdb files of AlphaFold prediction models were obtained for the proteins of interests [45]. Each pdb file was visualized and annotated using UCSF Chimera (https://www.cgl.ucsf.edu/chimera/) by selecting the sequences of interest to modify their colors manually onto the model.

Data availability

All scripts were written in Python 3.8 and are available on the Gao lab website at http://pepchem.org:35091/blackjack/sequential_sned1/tree/master.

Raw mass spectrometry data have been deposited to the ProteomeXchange Consortium [66] via the PRIDE partner repository [67] with the dataset identifier PXD030713. *The raw data will be made publicly available upon acceptance of the manuscript.*

CRediT AUTHOR STATEMENT

Fred Lee: Formal analysis, Investigation; Methodology; Visualization; Roles/Writing - original draft.

Xinhao Shao: Formal analysis, Investigation; Methodology; Software; Visualization; Roles/Writing - original draft.

Yu Gao: Conceptualization, Formal analysis, Funding acquisition, Investigation; Methodology; Resources; Software; Supervision; Validation; Roles/Writing - original draft.

Alexandra Naba: Conceptualization, Formal analysis, Funding acquisition, Investigation; Methodology; Project administration; Resources; Supervision; Validation; Roles/Writing - original draft.

ACKNOWLEDGMENTS

FUNDING SOURCES

CONFLICT OF INTEREST

AN has a sponsored research agreement with Boehringer-Ingelheim for work unrelated to the research presented in this study. The other authors declare that they have no conflict of interest.

REFERENCES

[1] K.M. Yamada, J.W. Collins, D.A. Cruz Walma, A.D. Doyle, S.G. Morales, J. Lu, K. Matsumoto, S.S. Nazari, R. Sekiguchi, Y. Shinsato, S. Wang, Extracellular matrix dynamics in cell migration, invasion and tissue morphogenesis, Int J Exp Pathol. 100 (2019) 144–152. https://doi.org/10.1111/iep.12329.

[2] B.J. Dzamba, D.W. DeSimone, Extracellular Matrix (ECM) and the Sculpting of Embryonic Tissues, Curr. Top. Dev. Biol. 130 (2018) 245–274. https://doi.org/10.1016/bs.ctdb.2018.03.006.

[3] N.K. Karamanos, A.D. Theocharis, Z. Piperigkou, D. Manou, A. Passi, S.S. Skandalis, D.H. Vynios, V. Orian-Rousseau, S. Ricard-Blum, C.E.H. Schmelzer, L. Duca, M. Durbee, N.A. Afratis, L. Troeberg, M. Franchi, V. Masola, M. Onisto, A Guide to the Composition and Functions of the Extracellular Matrix, FEBS J. (2021). https://doi.org/10.1111/febs.15776.

[4] D.A. Cruz Walma, K.M. Yamada, The extracellular matrix in development, Development. 147 (2020). https://doi.org/10.1242/dev.175596.

[5] G. Charras, E. Sahai, Physical influences of the extracellular environment on cell migration, Nat. Rev. Mol. Cell Biol. 15 (2014) 813–824. https://doi.org/10.1038/nrm3897.

[6] S.R. Lamandé, J.F. Bateman, Genetic Disorders of the Extracellular Matrix, Anat Rec (Hoboken). (2019). https://doi.org/10.1002/ar.24086.

[7] C. Bonnans, J. Chou, Z. Werb, Remodelling the extracellular matrix in development and disease, Nat. Rev. Mol. Cell Biol. 15 (2014) 786–801. https://doi.org/10.1038/nrm3904.

[8] M.W. Pickup, J.K. Mouw, V.M. Weaver, The extracellular matrix modulates the hallmarks of cancer, EMBO Rep. 15 (2014) 1243–1253. https://doi.org/10.15252/embr.201439246.

[9] T.R. Cox, The matrix in cancer, Nature Reviews Cancer. (2021) 1–22. https://doi.org/10.1038/s41568-020-00329-7.

[10] N.K. Karamanos, A.D. Theocharis, T. Neill, R.V. Iozzo, Matrix modeling and remodeling: A biological interplay regulating tissue homeostasis and diseases, Matrix Biology. 75–76 (2019) 1–11. https://doi.org/10.1016/j.matbio.2018.08.007.

[11] P. Pakshir, B. Hinz, The big five in fibrosis: Macrophages, myofibroblasts, matrix, mechanics, and miscommunication, Matrix Biol. 68–69 (2018) 81–93. https://doi.org/10.1016/j.matbio.2018.01.019.

[12] L. Moretti, J. Stalfort, T.H. Barker, D. Abebayehu, The interplay of fibroblasts, the extracellular matrix, and inflammation in scar formation, J Biol Chem. (2021) 101530. https://doi.org/10.1016/j.jbc.2021.101530.

[13] A. Naba, K.R. Clauser, H. Ding, C.A. Whittaker, S.A. Carr, R.O. Hynes, The extracellular matrix: Tools and insights for the "omics" era, Matrix Biology. 49 (2016) 10–24. https://doi.org/10.1016/j.matbio.2015.06.003.

[14] I.N. Taha, A. Naba, Exploring the extracellular matrix in health and disease using proteomics, Essays in Biochemistry. 63 (2019) 417–432. https://doi.org/10.1042/EBC20190001.

[15]   M.J. Randles, M.J. Humphries, R. Lennon, Proteomic definitions of basement membrane composition in health and disease, Matrix Biol. 57–58 (2017) 12–28. https://doi.org/10.1016/j.matbio.2016.08.006.

[16]   L. Krasny, P.H. Huang, Advances in the proteomic profiling of the matrisome and adhesome, Expert Rev Proteomics. 18 (2021) 781–794. https://doi.org/10.1080/14789450.2021.1984885.

[17]   A. Naba, K.R. Clauser, S. Hoersch, H. Liu, S.A. Carr, R.O. Hynes, The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices, Mol Cell Proteomics. 11 (2012) M111.014647. https://doi.org/10.1074/mcp.M111.014647.

[18]   E. Sonpho, F.G. Mann, M. Levy, E.J. Ross, C. Guerrero-Hernández, L. Florens, A. Saraf, V. Doddihal, P. Ounjai, A. Sánchez Alvarado, Decellularization Enables Characterization and Functional Analysis of Extracellular Matrix in Planarian Regeneration, Mol Cell Proteomics. 20 (2021) 100137. https://doi.org/10.1016/j.mcpro.2021.100137.

[19]   I. Angelidis, L.M. Simon, I.E. Fernandez, M. Strunz, C.H. Mayr, F.R. Greiffo, G. Tsitsiridis, M. Ansari, E. Graf, T.-M. Strom, M. Nagendran, T. Desai, O. Eickelberg, M. Mann, F.J. Theis, H.B. Schiller, An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics, Nat Commun. 10 (2019) 963. https://doi.org/10.1038/s41467-019-08831-9.

[20]   M.C. McCabe, R.C. Hill, K. Calderone, Y. Cui, Y. Yan, T. Quan, G.J. Fisher, K.C. Hansen, Alterations in extracellular matrix composition during aging and photoaging of the skin, Matrix Biol Plus. 8 (2020) 100041. https://doi.org/10.1016/j.mbplus.2020.100041.

[21]   A. Eckersley, M. Ozols, P. Chen, V. Tam, J.A. Hoyland, A. Trafford, D. Chan, M.J. Sherratt, Peptide Location Fingerprinting Reveals Tissue Region-Specific Differences in Protein Structures in an Ageing Human Organ, Int J Mol Sci. 22 (2021) 10408. https://doi.org/10.3390/ijms221910408.

[22]   M. Randles, F. Lausecker, Q. Kong, H. Suleiman, G. Reid, M. Kolatsi-Joannou, P. Tian, S. Falcone, B. Davenport, P. Potter, T. Van Agtmael, J. Norman, D. Long, M. Humphries, J. Miner, R. Lennon, Identification of an Altered Matrix Signature in Kidney Aging and Disease, J Am Soc Nephrol. (2021) ASN.2020101442. https://doi.org/10.1681/ASN.2020101442.

[23]   J. Barallobre-Barreiro, B. Loeys, M. Mayr, M. Rienks, A. Verstraeten, J.C. Kovacic, Extracellular Matrix in Vascular Disease, Part 2/4: JACC Focus Seminar, J Am Coll Cardiol. 75 (2020) 2189–2203. https://doi.org/10.1016/j.jacc.2020.03.018.

[24]   A.M. Socovich, A. Naba, The cancer matrisome: From comprehensive characterization to biomarker discovery, Seminars in Cell & Developmental Biology. 89 (2019) 157–166. https://doi.org/10.1016/j.semcdb.2018.06.005.

[25]   G.E. Arteel, A. Naba, The liver matrisome – looking beyond collagens, JHEPReport. 2 (2020). https://doi.org/10.1016/j.jhep.2020.100115.

[26]   H.B. Schiller, I.E. Fernandez, G. Burgstaller, C. Schaab, R.A. Scheltema, T. Schwarzmayr, T.M. Strom, O. Eickelberg, M. Mann, Time- and compartment-resolved proteome profiling of

the extracellular niche in lung injury and repair, Molecular Systems Biology. 11 (2015) 819. https://doi.org/10.15252/msb.20156123.

[27]   J. Merl-Pham, T. Basak, L. Knüppel, D. Ramanujam, M. Athanason, J. Behr, S. Engelhardt, O. Eickelberg, S.M. Hauck, R. Vanacore, C.A. Staab-Weijnitz, Quantitative proteomic profiling of extracellular matrix and site-specific collagen post-translational modifications in an in vitro model of lung fibrosis, Matrix Biology Plus. 1 (2019) 100005. https://doi.org/10.1016/j.mbplus.2019.04.002.

[28]   J.V. Olsen, S.-E. Ong, M. Mann, Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues *, Molecular & Cellular Proteomics. 3 (2004) 608–614. https://doi.org/10.1074/mcp.T400003-MCP200.

[29]   T. Basak, L. Vega-Montoto, L.J. Zimmerman, D.L. Tabb, B.G. Hudson, R.M. Vanacore, Comprehensive Characterization of Glycosylation and Hydroxylation of Basement Membrane Collagen IV by High-Resolution Mass Spectrometry, J. Proteome Res. 15 (2016) 245–258. https://doi.org/10.1021/acs.jproteome.5b00767.

[30]   R.O. Hynes, A. Naba, Overview of the Matrisome—An Inventory of Extracellular Matrix Constituents and Functions, Cold Spring Harb Perspect Biol. 4 (2012) a004903. https://doi.org/10.1101/cshperspect.a004903.

[31]   M.C. McCabe, L.R. Schmitt, R.C. Hill, M. Dzieciatkowska, M. Maslanka, W.F. Daamen, T.H. van Kuppevelt, D.J. Hof, K.C. Hansen, Evaluation and refinement of sample preparation methods for extracellular matrix proteome coverage, Molecular & Cellular Proteomics. 0 (2021). https://doi.org/10.1016/j.mcpro.2021.100079.

[32]   L. Krasny, A. Paul, P. Wai, B.A. Howard, R.C. Natrajan, P.H. Huang, Comparative proteomic assessment of matrisome enrichment methodologies, Biochemical Journal. 473 (2016) 3979–3995. https://doi.org/10.1042/BCJ20160686.

[33]   R. Aebersold, J.N. Agar, I.J. Amster, M.S. Baker, C.R. Bertozzi, E.S. Boja, C.E. Costello, B.F. Cravatt, C. Fenselau, B.A. Garcia, Y. Ge, J. Gunawardena, R.C. Hendrickson, P.J. Hergenrother, C.G. Huber, A.R. Ivanov, O.N. Jensen, M.C. Jewett, N.L. Kelleher, L.L. Kiessling, N.J. Krogan, M.R. Larsen, J.A. Loo, R.R. Ogorzalek Loo, E. Lundberg, M.J. MacCoss, P. Mallick, V.K. Mootha, M. Mrksich, T.W. Muir, S.M. Patrie, J.J. Pesavento, S.J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schlüter, S. Sechi, S.A. Slavoff, L.M. Smith, M.P. Snyder, P.M. Thomas, M. Uhlén, J.E. Van Eyk, M. Vidal, D.R. Walt, F.M. White, E.R. Williams, T. Wohlschlager, V.H. Wysocki, N.A. Yates, N.L. Young, B. Zhang, How many human proteoforms are there?, Nature Chemical Biology. 14 (2018) 206–214. https://doi.org/10.1038/nchembio.2576.

[34]   K.E. Burnum-Johnson, T.P. Conrads, R.R. Drake, A.E. Herr, R. Iyengar, R.T. Kelly, E. Lundberg, M.J. MacCoss, A. Naba, G.P. Nolan, P.A. Pevzner, K.D. Rodland, S. Sechi, N. Slavov, J.M. Spraggins, J.E. Van Eyk, M. Vidal, C. Vogel, D.R. Walt, N.L. Kelleher, D.R. Walt, N.L. Kelleher, New views of old proteins: clarifying the enigmatic proteome, ArXiv:2108.07660 [q-Bio]. (2021). http://arxiv.org/abs/2108.07660 (accessed October 6, 2021).

[35]   X. Shao, I.N. Taha, K.R. Clauser, Y. Gao, A. Naba, MatrisomeDB: the ECM-protein knowledge database, Nucleic Acids Res. 48 (2020) D1136–D1144. https://doi.org/10.1093/nar/gkz849.

[36]   S.J. Hubbard, The structural aspects of limited proteolysis of native proteins, Biochim Biophys Acta. 1382 (1998) 191–206. https://doi.org/10.1016/s0167-4838(97)00175-1.

[37]   N. de Souza, P. Picotti, Mass spectrometry analysis of the structural proteome, Curr Opin Struct Biol. 60 (2020) 57–65. https://doi.org/10.1016/j.sbi.2019.10.006.

[38]   A. Naba, K.R. Clauser, R.O. Hynes, Enrichment of extracellular matrix proteins from tissues and digestion into peptides for mass spectrometry analysis., J Vis Exp. 101 (2015) e53057. https://doi.org/10.3791/53057.

[39]   A. Naba, K.R. Clauser, J.M. Lamar, S.A. Carr, R.O. Hynes, Extracellular matrix signatures of human mammary carcinoma identify novel metastasis promoters, ELife. 3 (2014) e01308. https://doi.org/10.7554/eLife.01308.

[40]   S.D. Vallet, M.N. Davis, A. Barque, A.H. Thahab, S. Ricard-Blum, A. Naba, Computational and experimental characterization of the novel ECM glycoprotein SNED1 and prediction of its interactome, Biochemical Journal. (2021). https://doi.org/10.1042/BCJ20200675.

[41]   Franco-Barraza Janusz, Beacham Dorothy A., Amatangelo Michael D., Cukierman Edna, Preparation of Extracellular Matrices Produced by Cultured and Primary Fibroblasts, Current Protocols in Cell Biology. (2016) 10.9.1-10.9.34. https://doi.org/10.1002/cpcb.2@10.1002/(ISSN)1934-2616.ExtracellularMatrix.

[42]   B. Zybailov, A.L. Mosley, M.E. Sardiu, M.K. Coleman, L. Florens, M.P. Washburn, Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces c erevisiae*, J. Proteome Res. 5 (2006) 2339–2347. https://doi.org/10.1021/pr060161n.

[43]   W. Zhu, J.W. Smith, C.-M. Huang, Mass Spectrometry-Based Label-Free Quantitative Proteomics, Journal of Biomedicine and Biotechnology. 2010 (2009) e840518. https://doi.org/10.1155/2010/840518.

[44]   P. Samaras, T. Schmidt, M. Frejno, S. Gessulat, M. Reinecke, A. Jarzab, J. Zecha, J. Mergner, P. Giansanti, H.-C. Ehrlich, S. Aiche, J. Rank, H. Kienegger, H. Krcmar, B. Kuster, M. Wilhelm, ProteomicsDB: a multi-omics and multi-organism resource for life science research, Nucleic Acids Research. 48 (2020) D1153–D1163. https://doi.org/10.1093/nar/gkz974.

[45]   J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[46]   L. Tsiatsiani, A.J.R. Heck, Proteomics beyond trypsin, The FEBS Journal. 282 (2015) 2612–2626. https://doi.org/10.1111/febs.13287.

[47]   T. Dau, G. Bartolomucci, J. Rappsilber, Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin, Anal. Chem. 92 (2020) 9523–9527. https://doi.org/10.1021/acs.analchem.0c00478.

[48]   P. Giansanti, L. Tsiatsiani, T.Y. Low, A.J.R. Heck, Six alternative proteases for mass spectrometry–based proteomics beyond trypsin, Nat Protoc. 11 (2016) 993–1006. https://doi.org/10.1038/nprot.2016.057.

[49]   D. Morsa, D. Baiwir, R. La Rocca, T.A. Zimmerman, E. Hanozin, E. Grifnée, R. Longuespée, M.-A. Meuwis, N. Smargiasso, E.D. Pauw, G. Mazzucchelli, Multi-Enzymatic Limited Digestion: The Next-Generation Sequencing for Proteomics?, J. Proteome Res. 18 (2019) 2501–2513. https://doi.org/10.1021/acs.jproteome.9b00044.

[50]   M.L. Mendes, L. Fischer, Z.A. Chen, M. Barbon, F.J. O'Reilly, S.H. Giese, M. Bohlke-Schneider, A. Belsom, T. Dau, C.W. Combe, M. Graham, M.R. Eisele, W. Baumeister, C. Speck, J. Rappsilber, An integrated workflow for crosslinking mass spectrometry, Mol Syst Biol. 15 (2019). https://doi.org/10.15252/msb.20198994.

[51]   C. Leimeister, N. Schumacher, H. Diez, M. Gessler, Cloning and expression analysis of the mouse stroma marker Snep encoding a novel nidogen domain protein, Dev. Dyn. 230 (2004) 371–377. https://doi.org/10.1002/dvdy.20056.

[52]   A. Barqué, K. Jan, E. De La Fuente, C.L. Nicholas, R.O. Hynes, A. Naba, Knockout of the gene encoding the extracellular matrix protein SNED1 results in early neonatal lethality and craniofacial malformations, Dev Dyn. 250 (2021) 274–294. https://doi.org/10.1002/dvdy.258.

[53]   A.K. Pastino, T.M. Greco, R.A. Mathias, I.M. Cristea, J.E. Schwarzbauer, Stimulatory effects of advanced glycation endproducts (AGEs) on fibronectin matrix assembly, Matrix Biol. 59 (2017) 39–53. https://doi.org/10.1016/j.matbio.2016.07.003.

[54]   A. Eckersley, M. Ozols, R. O'Cualain, E.-J. Keevill, A. Foster, S. Pilkington, D. Knight, C.E.M. Griffiths, R.E.B. Watson, M.J. Sherratt, Proteomic fingerprints of damage in extracellular matrix assemblies, Matrix Biol Plus. 5 (2020) 100027. https://doi.org/10.1016/j.mbplus.2020.100027.

[55]   C. Tian, Y. Huang, K.R. Clauser, S. Rickelt, A.N. Lau, S.A. Carr, M.G. Vander Heiden, R.O. Hynes, Suppression of pancreatic ductal adenocarcinoma growth and metastasis by fibrillar collagens produced selectively by tumor cells, Nat Commun. 12 (2021) 2328. https://doi.org/10.1038/s41467-021-22490-9.

[56]   G.M. Harris, I. Raitman, J.E. Schwarzbauer, Chapter 5 - Cell-derived decellularized extracellular matrices, in: R.P. Mecham (Ed.), Methods in Cell Biology, Academic Press, 2018: pp. 97–114. https://doi.org/10.1016/bs.mcb.2017.08.007.

[57]   M.C. Chambers, B. Maclean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T.

Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L. Tabb, P. Mallick, A cross-platform toolkit for mass spectrometry and proteomics, Nat Biotechnol. 30 (2012) 918–920. https://doi.org/10.1038/nbt.2377.

[58]   A.T. Kong, F.V. Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics, Nat Methods. 14 (2017) 513–520. https://doi.org/10.1038/nmeth.4256.

[59]   F. da Veiga Leprevost, S.E. Haynes, D.M. Avtonomov, H.-Y. Chang, A.K. Shanmugam, D. Mellacheruvu, A.T. Kong, A.I. Nesvizhskii, Philosopher: a versatile toolkit for shotgun proteomics data analysis, Nat Methods. 17 (2020) 869–870. https://doi.org/10.1038/s41592-020-0912-y.

[60]   A. Naba, O.M.T. Pearce, A. Del Rosario, D. Ma, H. Ding, V. Rajeeve, P.R. Cutillas, F.R. Balkwill, R.O. Hynes, Characterization of the extracellular matrix of normal and diseased tissues using proteomics., J. Proteome Res. 16 (2017) 3083–3091. https://doi.org/10.1021/acs.jproteome.7b00191.

[61]   S.S. Shapiro, M.B. Wilk, An Analysis of Variance Test for Normality (Complete Samples), Biometrika. 52 (1965) 591–611. https://doi.org/10.2307/2333709.

[62]   F. Wilcoxon, Individual Comparisons by Ranking Methods, Biometrics Bulletin. 1 (1945) 80–83. https://doi.org/10.2307/3001968.

[63]   J.R. Conway, A. Lex, N. Gehlenborg, UpSetR: an R package for the visualization of intersecting sets and their properties, Bioinformatics. 33 (2017) 2938–2940. https://doi.org/10.1093/bioinformatics/btx364.

[64]   I. Letunic, S. Khedkar, P. Bork, SMART: recent updates, new developments and status in 2020, Nucleic Acids Research. 49 (2021) D458–D460. https://doi.org/10.1093/nar/gkaa937.

[65]   A. Eckersley, K.T. Mellody, S. Pilkington, C.E.M. Griffiths, R.E.B. Watson, R. O'Cualain, C. Baldock, D. Knight, M.J. Sherratt, Structural and compositional diversity of fibrillin microfibrils in human tissues, Journal of Biological Chemistry. 293 (2018) 5117–5133. https://doi.org/10.1074/jbc.RA117.001483.

[66]   E.W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D.S. Campbell, M. Bernal-Llinares, S. Okuda, S. Kawano, R.L. Moritz, J.J. Carver, M. Wang, Y. Ishihama, N. Bandeira, H. Hermjakob, J.A. Vizcaíno, The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition, Nucleic Acids Res. 45 (2017) D1100–D1106. https://doi.org/10.1093/nar/gkw936.

[67]   Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, Ş. Yılmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A.F. Jarnuczak, T. Ternent, A. Brazma, J.A. Vizcaíno, The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucleic Acids Res. 47 (2019) D442–D450. https://doi.org/10.1093/nar/gky1106.

Table 1. Average sequence coverage achieved using different digestion durations for the different matrisome protein categories.

	Average % sequence coverage			
	30 min	**2h**	**4h**	**18h**
ECM Glycoproteins	21.5%	25.2%	22.9%	15.9%
Collagens	9.7%	8.2%	10.0%	9.0%
Proteoglycans	34.8%	35.1%	37.0%	26.5%
ECM-affiliated	14.7%	14.7%	12.8%	8.0%
ECM Regulators	18.0%	19.9%	19.4%	10.3%
Secreted Factors	12.7%	11.2%	9.0%	11.2%

29

Table 2. Average sequence coverage achieved using an 18-hour or a time-lapsed tryptic digestion

Matrisome Category	Gene Symbol	UniProt ID	Average % coverage (18h)	Average % coverage (Seq_Aggregate)	Increase % coverage
ECM Glycoproteins	Fbln7	Q501P1	1.70	9.55	460.00
	Efemp1	Q8BPB5	1.32	6.29	376.92
	Fbln5	Q9WVH9	2.12	9.38	342.11
	Aebp1	Q640N1	3.32	13.83	316.00
	Lamb2	Q61292	4.45	15.20	241.88
	Thbs2	Q03350	4.48	12.93	188.57
	Matn2	O08746	5.54	14.85	167.92
	Tinagl1	Q99JR5	21.24	53.22	150.51
	Tgfbi	P82198	22.18	55.05	148.18
	Dpt	Q9QZZ6	4.23	10.45	147.06
	Efemp2	Q9WVJ9	12.64	29.68	134.82
	Lama5	Q61001	15.17	34.45	127.13
	Postn	Q62009	22.43	50.42	124.73
	Igsf10	Q3V1M1	8.21	17.14	108.69
	Vtn	P29788	0.84	1.67	100.00
	Ltbp1	Q8CG19	17.38	34.05	95.97
	Lamb1	P02469	9.13	16.80	84.05
	Nid1	P10493	29.08	52.37	80.11
	Fbn2	Q61555	16.48	29.12	76.72
	Ecm1	Q61508	6.89	11.36	64.94
	Fbln1	Q08879	4.75	7.80	64.18
	Pcolce	Q61398	16.03	25.00	56.00
	Emilin2	Q8K482	24.58	37.20	51.33
	Thbs1	P35441	27.52	41.62	51.24
	Nid2	O88322	31.75	47.15	48.48
	Mfap5	Q9QZJ6	14.02	20.73	47.83
	Tnc	Q80YX1	36.16	53.22	47.18
	Emilin1	Q99K41	43.07	57.08	32.53
	Mfap2	P55002	15.85	20.77	31.03
	Fbn1	Q61554	48.69	60.08	23.37
	Fbln2	P37889	48.12	58.44	21.45
	Fn1	P11276	64.43	76.06	18.05
Collagens	Col11a1	Q61245	0.25	1.88	655.56
	Col16a1	Q8BLX7	0.60	4.53	652.63
	Col4a3	Q9QZS0	0.30	0.60	100.00
	Col4a1	P02463	2.43	4.16	71.60
	Col6a2	Q02788	25.58	42.36	65.60
	Col18a1	P39061	7.19	11.87	65.10
	Col6a1	Q04857	29.71	48.39	62.89
	Col5a1	O88207	4.35	6.88	58.13
	Col3a1	P08121	18.44	28.04	52.04
	Col1a2	Q01149	28.94	42.75	47.73
	Col12a1	Q60847	24.28	35.37	45.68
	Col1a1	P11087	30.18	41.05	36.03
	Col5a2	Q3U962	13.89	17.03	22.60
	Col4a2	P08122	5.77	6.30	9.14
	Col8a1	Q00780	1.41	1.41	0.00
	Col2a1	P28481	3.83	3.73	-2.63
Proteo-glycans	Prelp	Q9JK53	1.98	14.42	626.67
	Aspn	Q99MQ4	13.27	38.20	187.88
	Bgn	P28653	41.06	58.13	41.58
	Hspg2	Q05793	43.15	60.83	40.98

Matrisome Category	Gene Symbol	UniProt ID	Average % coverage (18h)	Average % coverage (Seq_Aggregate)	Increase % coverage
ECM-affiliated Proteins	Anxa1	P10107	3.61	19.22	432.00
	Lgals3	P16110	5.49	17.80	224.14
	Anxa5	P48036	7.52	24.29	222.92
	Lman1	Q9D0F3	7.06	20.02	183.56
	Lgals9	O08573	13.03	29.89	129.35
	Anxa4	P97429	18.81	37.62	100.00
	Colec12	Q8K4Q8	10.38	20.15	94.16
	Lgals1	P16045	29.63	48.15	62.50
	Grem1	O70326	7.07	10.05	42.31
	Anxa2	P07356	21.09	23.89	13.29
	Anxa11	P97384	1.59	1.59	0.00
ECM Regulators	Lox	P28301	6.93	24.82	257.89
	Itih2	Q61703	12.32	30.07	144.21
	Serpinh1	P19324	14.63	34.65	136.89
	Ctsb	P10605	17.11	36.14	111.21
	Itih1	Q61702	0.39	0.77	100.00
	Loxl4	Q924C6	9.58	18.23	90.34
	Loxl2	P58022	17.07	31.64	85.28
	P4ha1	Q60715	16.01	28.65	78.95
	Tgm2	P21981	27.55	49.13	78.31
	Htra1	Q9R118	9.90	16.77	69.47
	Loxl1	P97873	10.46	13.01	24.41
	Adamtsl4	Q80T21	10.47	11.63	11.06
	Itih3	Q61704	7.09	6.86	-3.17
Secreted Factors	S100a13	P97352	5.61	20.41	263.64
	S100a11	P50543	13.78	27.55	100.00
	S100a6	P14069	7.87	8.99	14.29

Table 3. Sequence coverage dissimilarity index

Gene Symbol	Matrisome Category	% Coverage Dissimilarity Index	Gene Symbol	Matrisome Category	% Coverage Dissimilarity Index
Anxa1	ECM-affiliated Proteins	31.98	Col18a1	Collagens	-0.47
Anxa5	ECM-affiliated Proteins	30.66	Itih3	ECM Regulators	-0.78
Nid2	ECM Glycoproteins	28.41	Angptl2	Secreted Factors	-0.92
Tinagl1	ECM Glycoproteins	21.55	Cd109	ECM Regulators	-0.96
Mgp	ECM Glycoproteins	21.15	Ctsa	ECM Regulators	-1.99
Lgals9	ECM-affiliated Proteins	17.50	Fbln7	ECM Glycoproteins	-2.21
Lamb2	ECM Glycoproteins	15.39	Col12a1	Collagens	-2.34
Serpinh1	ECM Regulators	14.70	Lamb1	ECM Glycoproteins	-2.61
Mfap2	ECM Glycoproteins	14.12	Ctsz	ECM Regulators	-2.61
S100a13	Secreted Factors	13.12	Col3a1	Collagens	-2.64
Colec12	ECM-affiliated Proteins	12.99	Col11a1	Collagens	-2.72
Ctsb	ECM Regulators	12.80	Anxa2	ECM-affiliated Proteins	-3.34
Tgm2	ECM Regulators	11.05	Col2a1	Collagens	-3.37
Pcolce	ECM Glycoproteins	8.03	Fbn1	ECM Glycoproteins	-4.27
Lgals3	ECM-affiliated Proteins	7.99	Col6a2	Collagens	-4.40
Emilin1	ECM Glycoproteins	7.86	Anxa11	ECM-affiliated Proteins	-4.48
Col6a1	Collagens	7.47	Adamtsl4	ECM Regulators	-4.54
Ltbp1	ECM Glycoproteins	7.08	Fbln2	ECM Glycoproteins	-4.90
Itih2	ECM Regulators	6.38	Emilin2	ECM Glycoproteins	-5.23
Loxl4	ECM Regulators	6.12	Col1a2	Collagens	-6.16
Bgn	Proteoglycans	5.84	Col5a1	Collagens	-6.55
Lgals8	ECM-affiliated Proteins	5.11	Aebp1	ECM Glycoproteins	-6.72
Matn2	ECM Glycoproteins	4.87	Col1a1	Collagens	-6.79
Igsf10	ECM Glycoproteins	4.62	Lox	ECM Regulators	-7.95
Col4a2	Collagens	4.15	Col5a2	Collagens	-8.72
Col4a1	Collagens	3.98	Lgals1	ECM-affiliated Proteins	-9.02
Dpt	ECM Glycoproteins	3.46	Tnc	ECM Glycoproteins	-10.02
Hspg2	Proteoglycans	3.38	S100a6	Secreted Factors	-11.57
Anxa7	ECM-affiliated Proteins	3.34	P4ha1	ECM Regulators	-11.68
Fn1	ECM Glycoproteins	3.06	P4ha2	ECM Regulators	-11.92
Grem1	ECM-affiliated Proteins	3.02	Fbln1	ECM Glycoproteins	-12.36
Ecm1	ECM Glycoproteins	2.68	S100a11	Secreted Factors	-12.43
Lman1	ECM-affiliated Proteins	2.53	Nid1	ECM Glycoproteins	-12.98
Gpc4	ECM-affiliated Proteins	2.39	Postn	ECM Glycoproteins	-14.23
Loxl2	ECM Regulators	2.31	Fbln5	ECM Glycoproteins	-14.62
Prelp	Proteoglycans	2.23	Aspn	Proteoglycans	-14.83
Col8a1	Collagens	2.02	Tgfbi	ECM Glycoproteins	-16.76
Lama5	ECM Glycoproteins	1.67	Anxa4	ECM-affiliated Proteins	-16.88
Itih1	ECM Regulators	1.18	Loxl1	ECM Regulators	-19.18
Col16a1	Collagens	1.10	Htra1	ECM Regulators	-19.83
Loxl3	ECM Regulators	0.69	Fbn2	ECM Glycoproteins	-20.10
Col4a3	Collagens	0.32	Efemp1	ECM Glycoproteins	-21.94
Vtn	ECM Glycoproteins	0.00	Thbs1	ECM Glycoproteins	-23.27
Mfap5	ECM Glycoproteins	0.00	Efemp2	ECM Glycoproteins	-25.50
Timp3	ECM Regulators	0.00	Thbs2	ECM Glycoproteins	-29.27

Decreased coverage in SNED1-containing CDMs   *Increased coverage in SNED1-containing CDMs*

FIGURE LEGENDS

Figure 1. Identification of ECM proteins using time-restricted proteolysis protocol

A. Schematic representation of the experimental pipeline including the production of cell-derived matrix (CDM) by immortalized mouse embryonic fibroblasts (MEFs) over the course of 8 days and the initial steps of sample preparation prior to LC-MS/MS analysis.

B. Schematic representation of the time-limited proteolysis protocol using single timepoint digestions of defined duration (30 minutes, 2 hours, 4 hours, or 18 hours).

C. Bar graph depicts, for each matrisome category, the number of ECM proteins identified in two biological replicates (intersection), using single timepoint digestions (*see also Supplementary Table 2*).

D. Venn diagram shows the overlap between the proteins identified using different trypsin digestion durations. For each timepoint, the protein set is defined as the intersection of proteins identified in two biological replicates.

Figure 2. Identification of ECM proteins using time-lapsed trypsin digestion

A. Schematic representation of the sequential trypsin digestion protocol to achieve time-lapsed proteolysis.

B. Bar graph depicts, for each matrisome category, the number of ECM proteins identified in two biological replicates using a time-lapsed digestion protocol. Aggregation of the data from all 4 timepoints of the time-lapsed digestion is also provided. For this, the average number of proteins identified across 2 biological replicates, is obtained first by defining proteins identified at any of the four digestion timepoints for each replicate, and then averaging these numbers (*see also Supplementary Tables 2 and 3*).

C. Venn diagram shows the overlap between the proteins identified at each timepoint of the time-lapsed digestion protocol. For each timepoint, the protein set is defined as the intersection of proteins identified in two biological replicates.

D. Matrisome proteins identified at each of the four sequential timepoints or the standard 18-hour single timepoint digestions are compared, and the intersections are visualized using an UpSet plot (*see also Supplementary Tables 2 and 3*).

E. Venn Diagram illustrates the comparison between matrisome proteins identified using the standard 18-hour digestion protocol and upon aggregation of the data from the 4 sequential digestion timepoints, both lists being defined as the ensemble of matrisome proteins identified in two biological replicates.

F. Heatmap illustrates the normalized spectral abundance factor (NSAF) obtained for each matrisome protein using the time-lapse trypsin digestion protocol or the standard single timepoint (18-hour) digestion protocol.

Figure 3. Aggregation of data from sequential tryptic digestion timepoints results in enhanced sequence coverage

A. Each line illustrates the cumulative average sequence coverage percentage at each timepoint for a given matrisome protein (*see also Supplementary Table 2*).

B. Violin plots compare, for each matrisome category, the distribution of average protein sequence coverage obtained with a standard 18-hour digestion protocol or by aggregating data from the 4 timepoints of the time-lapsed digestion. Median and interquartile range are indicated by the white dots and grey bars, respectively. Wilcoxon signed-rank test was used to compare the distribution, and p-value was reported after Bonferroni adjustments were made (*$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$, ns: not significant).

C. Scatter plot represents the average aggregated percentage of sequence coverage for each protein detected (y-axis) as a function of protein molecular weight (kDa, x-axis). The diameter of each dot represents the fold-change in coverage as compared to the standard 18-hour single timepoint digestion method, and the proteins with changes above 90th and below 10th percentile are annotated (*see Table 2 and Supplementary Table 2*).

Figure 4. The ECM protein SNED1 modulates MEF matrisome composition and coverage

A. Immunofluorescence images of decellularized CDMs produced by immortalized mouse embryonic fibroblasts overexpressing GFP alone (control, top panels) or SNED1_GFP (bottom panels) co-stained with anti-SNED1 (orange, left panels) and anti-GFP antibodies (light blue, middle panels), producing an overlay (right panels). Scale bars: 50 μm.

B. Heatmap represents the aggregated normalized spectral abundance factor (NSAF) of proteins identified in CDMs produced by control or *SNED1_GFP*-overexpressing MEFs. Difference in

abundance is calculated by subtracting the relative abundance of proteins of control CDMs from those of SNED1-containing CDMs (*see Supplementary Table 4*). Proteins are sorted in descending order of differential abundance.

C. A custom dissimilarity index was calculated to represent changes in protein sequence coverage in the aggregated dataset of the 4 timepoints of the time-lapsed digestion of control and SNED1-containing CDMs. Proteins depicted by dots toward the top of the boxes show matrisome proteins whose sequence coverage is lower when SNED1 is present, while proteins depicted by dots toward the bottom show the opposite.

D. Scatter plot represents the relation between the average sequence coverage of all matrisome proteins detected in both control and SNED1-containing CDMs. Proteins found above the diagonal (*e.g.,* Thbs1, Thbs2, Efemp2) are proteins whose coverage increase in SNED1-containing CDMs, while proteins found below the diagonal (*e.g.,* Nid2, Tinagl1) are proteins whose sequence coverage is decreased in SNED1-containing CDMs (*see Table 3 and Supplementary Table 4*). Proteins with dissimilarity index above 95th and below 5th percentile are annotated.

Figure 5. Time-lapsed peptide mapping of SNED1

A. Mapping of peptides detected at each of the timepoints of the time-lapsed digestion on schematic representation of the domain-based organization of SNED1 (UniProt Q8TER0). Each peptide is mapped using the start and end position of each peptide, based on information provided by the Simple Modular Architecture Research Tool (SMART, http://smart.embl-heidelberg.de/). The color scale from light to dark grey represents, for each domain, the number of peptide-spectrum matches (PSMs). Peptide sequences depicted are the ones uniquely identified at a given timepoint. Bolded sequences correspond to peptides detected in two biological replicates.

B. 3-D mapping of detected peptides on the AlphaFold-predicted model of SNED1 (https://alphafold.ebi.ac.uk/entry/Q8TER0). Peptides are color-coded to indicate at which digestion timepoints they were detected. For the aggregate illustration the timepoint was assigned to the timepoint of first appearance.

Figure 6. Nidogen-2 sequence coverage decreases upon SNED1 overexpression

A. Representation of peptide mapping of nidogen-2 (Nid2, UniProt O88322). Upper panel represents the mapping of Nid2 peptides detected in control CDM samples; lower panel represents

34

the mapping of Nid2 peptides detected in SNED1-containing CDM samples. The color scale from light to dark grey represents, for each domain, the number of peptide-spectrum matches (PSMs). The line plot illustrates the difference in numbers of PSM detected in control or SNED1-containing CDMs and normalized by domain length.

B. 3-D mapping of detected peptides on the AlphaFold-predicted model of nidogen-2 (Nid2; https://alphafold.ebi.ac.uk/entry/O88322). Upper panel represents the 3-D mapping of Nid2 peptides detected in control CDMs; lower panel represents the 3-D mapping of Nid2 peptides detected in SNED1-containing CDMs. Each peptide color-coded to indicate the timepoint at which it was detected. For the aggregate illustration, the timepoint was assigned to the timepoint of first appearance.

Figure 7. Thrombospondin-1 sequence coverage is increased upon SNED1 presence in the ECM

A. Representation of peptide mapping of thrombospondin-1 (Thbs1, UniProt P35441). Upper panel represents the mapping of Thbs1 peptides detected in control CDM samples; lower panel represents the mapping of Thbs1 peptides detected in SNED1-containing CDM samples. The color scale from light to dark grey represents, for each domain, the number of peptide-spectrum matches (PSMs). The line plot illustrates the difference in numbers of PSM detected in control or SNED1-containing CDMs normalized by domain length.

B. 3-D mapping of detected peptides on the AlphaFold-predicted model of thrombospondin-1 (Thbs1; https://alphafold.ebi.ac.uk/entry/P35441). Upper panel represents the 3-D mapping of Thbs1 peptides detected in control CDMs; lower panel represents the 3-D mapping of Thbs1 peptides detected in SNED1-containing CDMs. Each peptide is color-coded to indicate the timepoint at which it was detected. For the aggregate illustration, the timepoint was assigned to the timepoint of first appearance.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Time-lapsed tryptic digestion increases overall sequence coverage

A. Each line of the line plots represents, for a given matrisome protein, the average sequence coverage calculated at each timepoint of the time-lapse digestion. Each graph represents the data for a given matrisome category.

B. Violin plot represents the distribution of average sequence coverage between the standard 18-hour digestion protocol (blue), the coverage obtained at the 2-hour timepoint of the time-lapsed digestion protocol (orange), and the coverage obtained by aggregating data from the 4 timepoints of the sequential digestion (grey). The median is represented by the dotted line and the interquartile range is indicated by the colored dotted lines. One-way ANOVA was performed to compare the groups, and the p-value was reported (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).

C. To rule out that the increase in the sequence coverage percentages seen upon aggregation of data from the different digestion timepoints was due to the simple increase in the number of datasets collated, we aggregated four datasets obtained using the standard 18-hour digestion method (*i.e.,* four files) and compared the average aggregated coverage (left violin plots) to the average aggregated coverage obtained by integrating data from the four distinct digestion timepoints (right violin plots). We confirm that the increased percentage of sequence coverage is indeed due to the benefit of the time-lapsed digestion method.

Supplementary Figure 2. Time-lapsed tryptic digestion increases overall sequence coverage

Histograms represent the distribution of average coverage percentage of all ECM proteins at each timepoint of the time-lapsed trypsin digestion. Scatter plots compare the log2-transformed average sequence coverage of all identified matrisome proteins at different timepoints of the time-lapsed digestion. Proteins aligned along the diagonal indicate a similar coverage between the two methods.

Supplementary Figure 3. Proteomic profiling of the composition of the ECM produced by SNED1_GFP-overexpressing MEFs

A. Bar graph depicts, for each matrisome category, the number of matrisome proteins identified in two biological replicates (intersection), in CDMs produced by MEFs overexpressing

36

SNED1_GFP, using different trypsin digestion durations either individually (left) or sequentially (right) (*see also Supplementary Table 4*).

B. Venn diagram shows the overlap between matrisome proteins identified at each timepoints of the sequential digestion. For each timepoint, the protein set is defined as the ensemble of proteins identified in two biological replicates.

C. Matrisome proteins identified at each of the four sequential timepoints or the standard 18-hour single timepoint digestions are compared, and the intersections are visualized using an UpSet plot (*see also Supplementary Tables 4 and 5*).

D. Venn diagram illustrates the comparison between matrisome proteins identified using a standard 18-hour digestion protocol or by aggregating data from the 4 sequential digestion timepoints of the time-lapsed protocol.

E. Each line illustrates, for a given matrisome protein, its cumulative average sequence coverage percentage at each timepoint of the time-lapsed digestion (*see also Supplementary Table 3*).

F. Violin plots compare, for each matrisome category, the distribution of average protein sequence coverage obtained with a standard 18-hour digestion protocol or by aggregating data from the 4 timepoints of the time-lapsed digestion. Median and interquartile range are indicated by the white dots and black bars, respectively. Wilcoxon signed-rank test was used to compare the distribution, and p-value was reported after Bonferroni adjustments were made (*$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$, ns: not significant).

G. Heatmap represents the normalized spectral abundance factor (NSAF) obtained for each matrisome protein produced by SNED1-overexpressing MEFs and detected using the time-lapse trypsin digestion protocol or the standard single timepoint (18-hour) digestion protocol.

H. Histogram represents the distribution of average coverage percentage of all matrisome proteins at each timepoint of the time-lapsed trypsin digestion detected in CDMs produced by SNED1-overexpressing MEFs. Scatter plots compare the log2-transformed average sequence coverage of all identified matrisome proteins at different timepoints of the time-lapsed digestion. Proteins aligned along the diagonal indicate a similar coverage between the two methods.

37

SUPPLEMENTARY TABLES

Supplementary Table 1. Sample list and MS metrics

Supplementary Table 2. Proteomic profiling of cell-derived matrices produced by immortalized *Sned1^{KO}* mouse embryonic fibroblasts (MEFs) overexpressing GFP alone

Supplementary Table 3. Overview of matrisome proteins identified CDMs made by immortalized *Sned1^{KO}* mouse embryonic fibroblasts overexpressing GFP alone

Supplementary Table 4. Proteomic profiling of cell-derived matrices produced by immortalized *Sned1^{KO}* mouse embryonic fibroblasts overexpressing GFP-tagged SNED1

Supplementary Table 5. Overview of matrisome proteins identified CDMs made by immortalized *Sned1^{KO}* mouse embryonic fibroblasts overexpressing GFP-tagged SNED1

Supplementary Table 6. Comparison of matrisome proteins identified in CDMs lacking- and containing SNED1

Supplementary Table 7. Semi-quantitative analysis using normalized spectral abundance factor (NSAF)

Supplementary Table 8. Complete peptide report

SUPPLEMENTARY FILES

Supplementary File 1. Example of a peptide mapping heuristic (.xlsx)

Supplementary File 2. Time-resolved peptide mapping on the 3-D AlphaFold model of SNED1. Video shows a full rotation around the y-axis (.mp4)

Supplementary File 3. Time-resolved mapping of peptides identified in control CDMs on the 3-D AlphaFold model of nidogen 2. Video shows a full rotation around the y-axis (.mp4)

Supplementary File 4. Time-resolved mapping of peptides identified in SNED1_GFP-containing CDMs on the 3-D AlphaFold model of nidogen 2. Video shows a full rotation around the y-axis (.mp4)

Supplementary File 5. Time-resolved mapping of peptides identified in control CDMs on the 3-D AlphaFold model of thrombospondin 1. Video shows a full rotation around the y-axis (.mp4)

Supplementary File 6. Time-resolved mapping of peptides identified in SNED1_GFP-containing CDMs on the 3-D AlphaFold model of thrombospondin 1. Video shows a full rotation around the y-axis (.mp4)
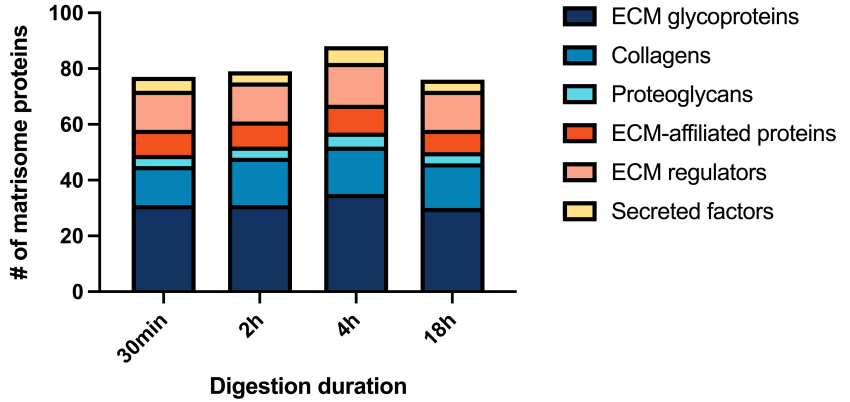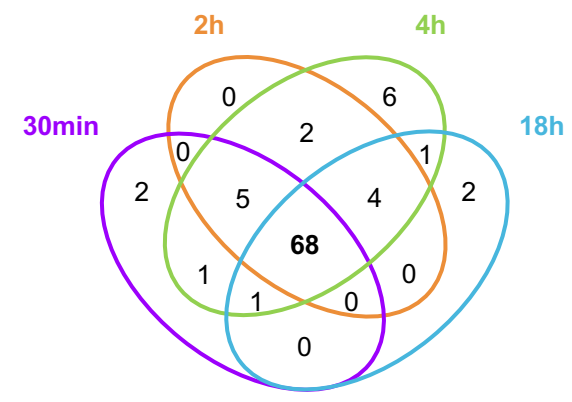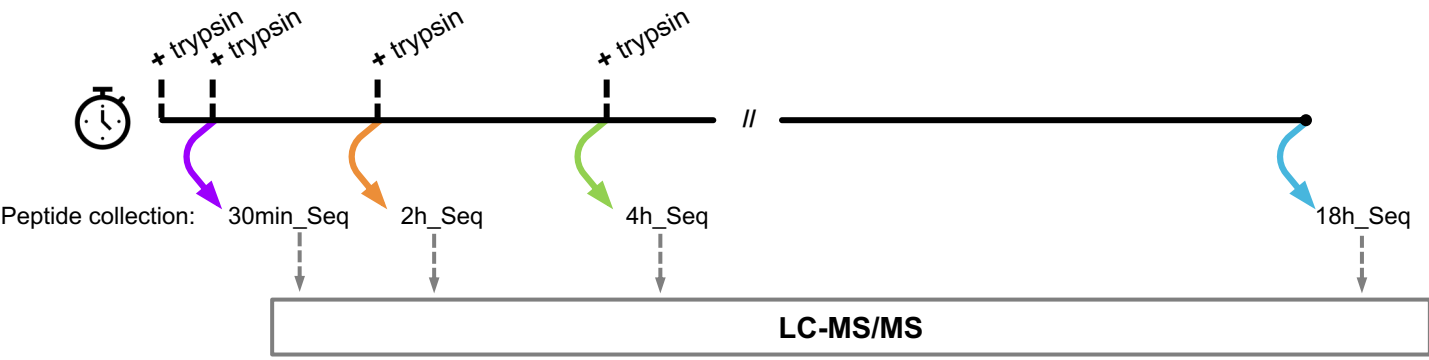
Figure 1

A.

iMEFs

ECM deposition — *8 days*

Decellularization

Cell-derived matrix (CDM)

Sned1^{LacZ-Neo/LacZ-Neo} + GFP (control)
or
Sned1^{LacZ-Neo/LacZ-Neo} + *SNED1_GFP*

- Partial solubilization + Reduction
- Deglycosylation
- LysC pre-digestion

B. Single time-point tryptic digestions

+ trypsin

30min
2h
4h
18h *Standard protocol*

LC-MS/MS

C.

# of matrisome proteins

Digestion duration: 30min, 2h, 4h, 18h

- ECM glycoproteins
- Collagens
- Proteoglycans
- ECM-affiliated proteins
- ECM regulators
- Secreted factors

D.

2h   4h
30min   18h

0 2 6
0   1
2   2
5   4
68
1   0
1   0
0

Figure 2

A. Time-lapsed (sequential) tryptic digestion

Figure 3
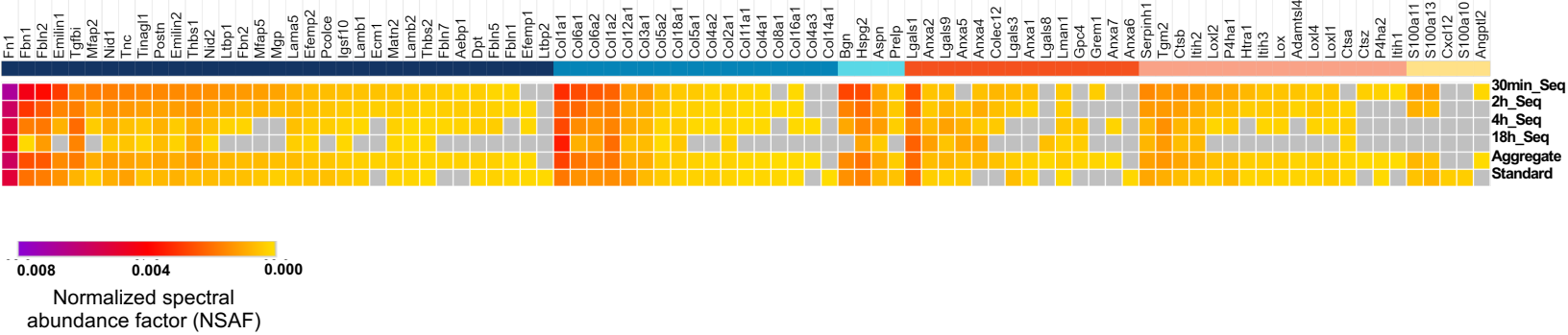
A.

B.

C.

Figure 4

A.

B.

C.

D.

Figure 5

A.

SNED1 30min_Seq

CQCPAGFGGPTCETAQSPCDTK

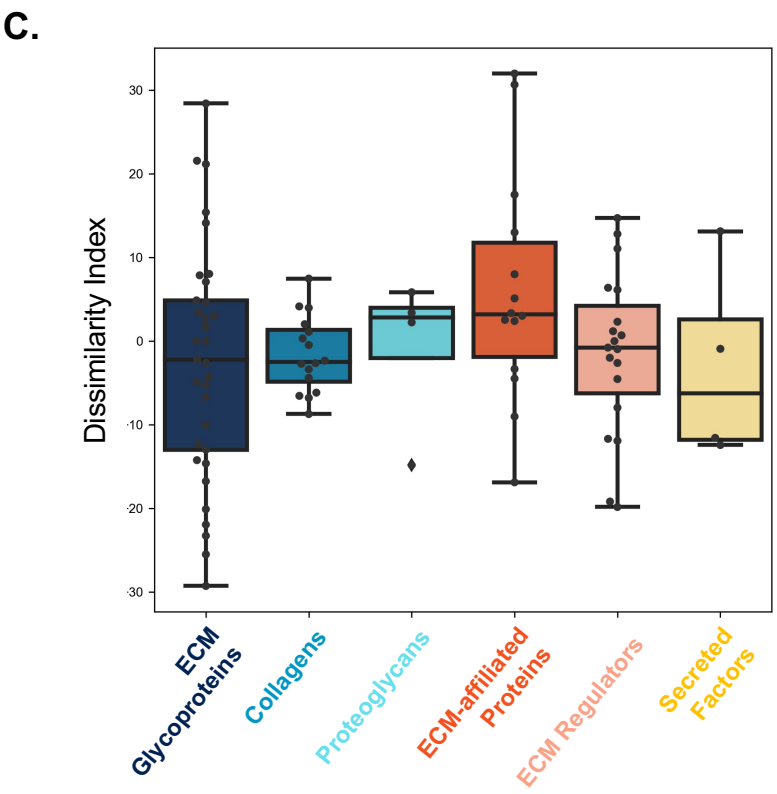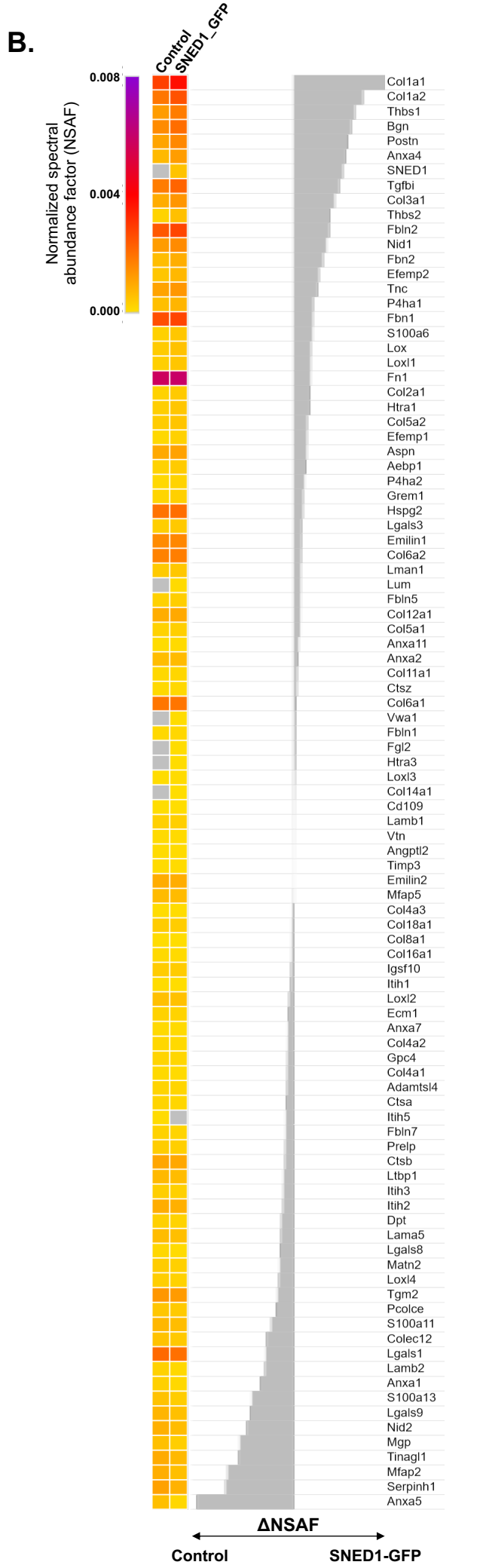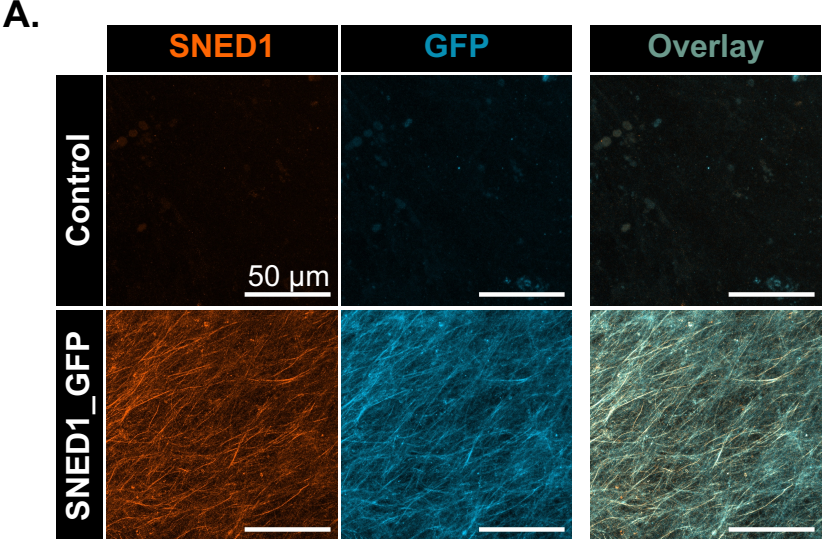LGAVALYACDRGYSLSAPSR

MERVEESGVSISWNPPNGPAAR, QMLDGYAVTYVSSDGSYR

SRYVPNGK

1 — 1413

ERVPDDCECR

HCEIAPSPCFR

RCQAEVDCGPPEEVK

EAGGEYHCSCPYR

VAGYLCLCSTGYEGAHCELERDECR

RVSLALQLPEHGSK

YQLSVIAVQSTELGPQHSEPAHLYIITSPR

SNED1 2h_Seq

QDDGGSGLRPLSVPFPFFGAEHSGLYVNNNGIISFLK

SVDRFTFR

VTATSAHVVWDAPTPGSLLEAYVINVTTSQSTK

1 — 1413

SNED1 4h_Seq

EATDPAMLR

HPEALR**DQATDVDR**

VSIRHPEALR

1 — 1413

SNED1 18h_Seq

YTIQLTTLSGLR

1 — 1413

Protein Domains

Signal Peptide   NIDO   EGF   EGF-Ca^{2+} FoIn   CCP/Sushi   FN3

Coverage

1 - 3 PSMs   7 - 9 PSMs

4 - 6 PSMs   > 9 PSMs

B.

30min_Seq 17.8% | 2h_Seq 13.1% | 4h_Seq 5.9% | 18h_Seq 3.0% | Aggregate 27.32%

Figure 6

A.

Nid2 coverage in control CDM — 1 ... 1403 — *Aggregate Coverage 47 %*

Δ PSMs / Domain length (SNED1_GFP – GFP)

Legend:
- 6 - 10 PSMs
- 11 - 15 PSMs
- 16 - 20 PSMs
- > 20 PSMs

Nid2 coverage in SNED1_GFP CDM — 1 ... 1403 — *Aggregate coverage 33 %*

Protein Domains
- Signal Peptide
- NIDO
- EGF
- EGF-Ca^{2+}
- G2F
- TY
- LLY

B.

Nid2 in control CDM

Nid2 in SNED1_GFP CDM

30min_Seq   2h_Seq   4h_Seq   18h_Seq   Aggregate

Figure 7

A.

Thbs1 coverage in control CDM — 1 ... 1170 — *Aggregate coverage 42 %*

ΔPSM / Domain length (SNED1_GFP – GFP)

- 1 - 5 PSMs
- 6 - 10 PSMs
- 11 - 15 PSMs
- > 15 PSMs

Thbs1 coverage in SNED1_GFP CDM — 1 ... 1170 — *Aggregate coverage 53 %*

Protein Domains

- Signal Peptide
- Thrombospondin N-terminal like
- VWC
- TSP1
- EGF
- Low complexity region

B.

Thbs1 in control CDM

Thbs1 in SNED1_GFP CDM

30min_Seq   2h_Seq   4h_Seq   18h_Seq   Aggregate