Systematic Analysis of Actively Transcribed Core Matrisome Genes Across Tissues and Cell Phenotypes

Tristen V. Tellman [a,1], Merve Dede [b,1], Vikram A. Aggarwal [c], Duncan Salmon [a], Alexandra Naba [d] and Mary C. Farach-Carson [a,c,e]

***a* - Department of Diagnostic & Biomedical Sciences,** *University of Texas Health Science Center at Houston School of Dentistry, 1941 East Road, BBS-4220, Houston, TX 77054, USA*

***b* - Department of Bioinformatics and Computational Biology,** *The University of Texas MD Anderson Cancer Center, P.O. Box 301402, Houston, TX 77230, USA*

***c* - Departments of BioSciences and Bioengineering,** *Rice University, 6100 Main St., Houston, TX 77005, USA*

***d* - Department of Physiology and Biophysics,** *University of Illinois at Chicago, 835 S. Wolcott, Rm E202 (MC901), Chicago, IL 60612, USA*

***e* - Center for Theoretical Biological Physics,** *Rice University, 6100 Main St., Houston, TX 77005, USA*

Corresponding to Mary C. Farach-Carson: *Department of Diagnostic & Biomedical Sciences, University of Texas Health Science Center at Houston School of Dentistry, 1941 East Road, BBS-4220, Houston, TX 77054, USA.*
Mary.C.FarachCarson@uth.tmc.edu.
https://doi.org/10.1016/j.matbio.2022.06.003

Abstract

The extracellular matrix (ECM) is a highly dynamic, well-organized acellular network of tissue-specific biomolecules, that can be divided into structural or core ECM proteins and ECM-associated proteins. The ECM serves as a blueprint for organ development and function and, when structurally altered through mutation, altered expression, or degradation, can lead to debilitating syndromes that often affect one tissue more than another. Cross-referencing the FANTOM5 SSTAR (Semantic catalog of Samples, Transcription initiation And Regulators) and the defined catalog of core matrisome ECM (glyco)proteins, we conducted a comprehensive analysis of 511 different human samples to annotate the context-specific transcription of the individual components of the defined matrisome. Relative log expression normalized SSTAR cap analysis gene expression peak data files were downloaded from the FANTOM5 online database and filtered to exclude all cell lines and diseased tissues. Promoter-level expression values were categorized further into eight core tissue systems and three major ECM categories: proteoglycans, glycoproteins, and collagens. Hierarchical clustering and correlation analyses were conducted to identify complex relationships in promoter-driven gene expression activity. Integration of the core matrisome and curated FANTOM5 SSTAR data creates a unique tool that provides insight into the promoter-level expression of ECM-encoding genes in a tissue- and cell-specific manner. Unbiased clustering of cap analysis gene expression peak data reveals unique ECM signatures within defined tissue systems. Correlation analysis among tissue systems exposes both positive and negative correlation of ECM promoters with varying levels of significance. This tool can be used to provide new insight into the relationships between ECM components and tissues and can inform future research on the ECM in human disease and development. We invite the matrix biology community to continue to explore and discuss this dataset as part of a larger and continuing conversation about the human ECM. An interactive web tool can be found at matrixpromoterome.github.io along with additional resources that can be found at dx.doi.org/10.6084/m9.figshare.19794481 (figures) and https://figshare.com/s/e18ecbc3ae5aaf919b78 (python notebook).

Introduction

The extracellular matrix (ECM) is a highly dynamic, well-organized acellular network of biomolecules that are assembled in a tissue-specific manner. The ECM lends overall function and form to tissues, aiding in the fine-tuning of cellular phenotype, adhesion, wound repair, mechanical transduction, development, differentiation, and, when disrupted, disease and dysregulated repair [1–4]. In the seminal matrisome paper published by the Hynes group in 2011, the ECM and its associated proteins were divided into two major groups (core matrisome and matrix-associated) based on *in silico* definitions [5,6]. The core matrisome was divided further into three major categories of ECM: proteoglycans, glycoproteins, and collagens [6–10]. Briefly, these categories can be defined by major features of the group wherein glycoproteins are macromolecules with covalently linked carbohydrates, or glycans, of varying lengths and degrees of branching, attached to a protein core. Proteoglycans, a subclass of glycoproteins, contain specific linear glycosaminoglycans attached to a protein core with repeating disaccharides that define them as chondroitin sulfate, heparin/heparan sulfate, dermatan sulfate, or keratan sulfate [11,12]. Collagens are the most abundant category, by percentage, of the ECM, and are characterized by their unique right handed three parallel polypeptide strand helical structure that can be either continuous or interrupted [13]. These definitions, as well as consensus structural elements and structural domain elements, are what helped define the original annotated matrisome.

While this original analysis provided a first-of-its-kind definition of the ECM and affiliated components, there remained the intriguing opportunity to overlay these individual ECM components with their relative tissue- and cell-level distributions. Several groups have conducted analysis on the matrisome using datasets comprised of single-cell RNA-sequencing and gene expression data from the Genotype-Tissue Expression, The Cancer Genome Atlas program, and the Gene Expression Omnibus, to name a few. These papers looked at the matrisome in tissue, age, sex, disease, and as signatures for cell typing in developing embryos, though none are able to provide a comprehensive overview of the matrisome at the promoter level in homeostatic tissues [14,15].

The SSTAR (Semantic catalog of Samples, Transcription initiation And Regulators) was released through the RIKEN FANTOM5 project and contains relative read values of cap analysis gene expression (CAGE) peak data as it relates to promoter-level activity, where unique cap identifiers are read to generate tags per million reference values that correspond to each active promoter [16]. CAGE peak is unique in its ability to capture the true active transcription of specific genes and providing promoter-level expression as it relates to individual samples. This data set was a collaborative effort from laboratories around the world, including our own, that submitted samples (tissues and cells) from various organ systems. The comprehensive database contains transcription start site data from ~1800 human samples, with detailed readings that can be interrogated either by gene symbol or name (https://fantom.gsc.riken.jp/5/sstar/Main_Page). We combined the information in the matrisome and FANTOM5 database to reveal the complexity and specificity of the components of the matrisome in the human body. Here we describe for the first time this useful analysis and discuss examples of the relationships that can be observed through correlation and clustering analysis of ECM-encoding genes and tissue/cell samples.

Results and Interpretation

Data reduction produced a data matrix of 261 genes with 511 samples

The original data set for this analysis was derived from cross-referencing the matrisome [8] and the FANTOM5 SSTAR databases [16] resulting in a matrix of 274 core matrisome genes x 891 tissue/cell samples as shown in **Figure 1**. The data was divided further into three major categories: proteoglycans, glycoproteins, and collagens containing 36, 182, and 43 genes, respectively. Matrisome genes not included in this analysis are AMELY, BSPH1, CDCP2, DSPP, NTN5, OTOG, OTOL1, POMZP3, SSPO, TECTA, ZP4, ZPLD1, and COL6A6 because the corresponding SSTAR data was missing from the FANTOM5 database. The final matrix promoter-level gene expression dataset included 261 ECM-encoding genes that were analyzed. We note here that AGRN was re-categorized from the original matrisome definition to a proteoglycan, consistent with what is now known about the protein. [9] COLXVIII and COLXV are considered proteoglycans, but for this analysis were left in the collagen category, consistent with the original matrisome assignments.

The data set was filtered further to remove immortalized cells lines, diseased tissues, experimentally treated primary cells, and samples with missing values. The samples were categorized manually into systems based on their primary affiliation. Primary affiliation here is defined as the system in which a cell or tissue type is most closely phenotypically, rather than anatomically, related. The traditional physiological ten system nomenclature (cardiovascular, nervous and sensory, digestive, respiratory, renal, reproductive, endocrine, immune, musculoskeletal, and integumentary system) did not fit the matrix promoter-level expression data well (data not

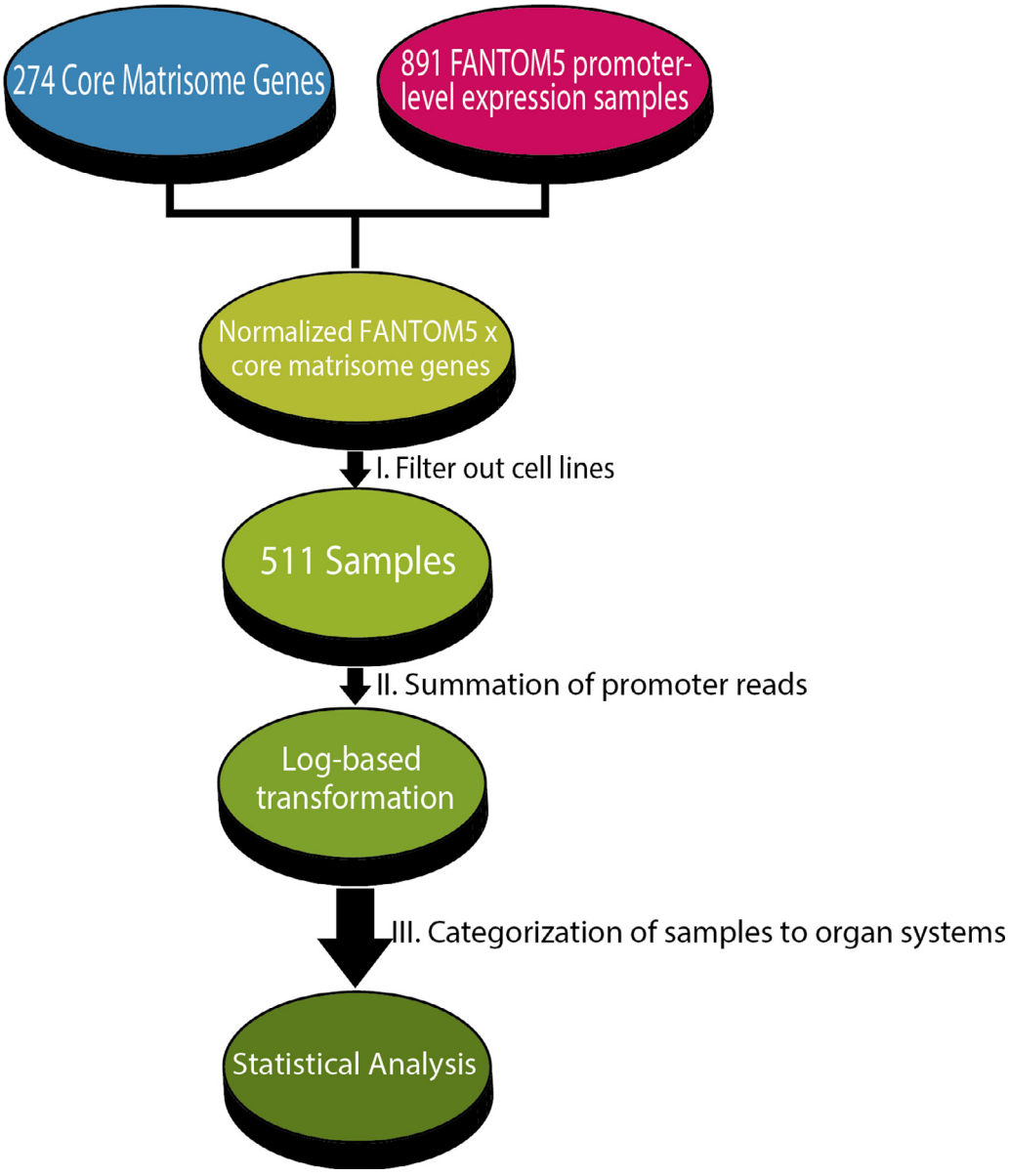

Figure 1. Workflows diagram of matrisome and FANTOM5 data reduction. Matrisome genes were cross-referenced with the FANTOM5 CAGE peak data to create the resulting analysis matrix. Further data reduction produced three major matrices with manual system annotations for each sample.

shown) so major tissue categories were defined for this analysis as: cardiovascular, connective, digestive, endothelial, epithelial, immune, nervous, and reproductive. As an example of how these categories fit the samples, adipocytes from the breast were assigned to the connective tissue system. We note here that some samples may adequately fit within multiple categories but are defined using additional information available in the FANTOM5 database. Any samples that could not fit broadly into these major tissue categories were eliminated from the analysis, resulting in a data set of 511 total samples.

Hierarchical clustering reveals unique tissue/cell-level clusters of gene promoter activity

Delineation of matrisome components into eight color-coded primary systems reveals unique clusters and broad trends in promoter-level expression. **(Figure 2)** Samples were hierarchically clustered, as described in the methods section, where average linkage representing the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and the Euclidean distance metric were used to put like genes in a spatially related order, as indicated by

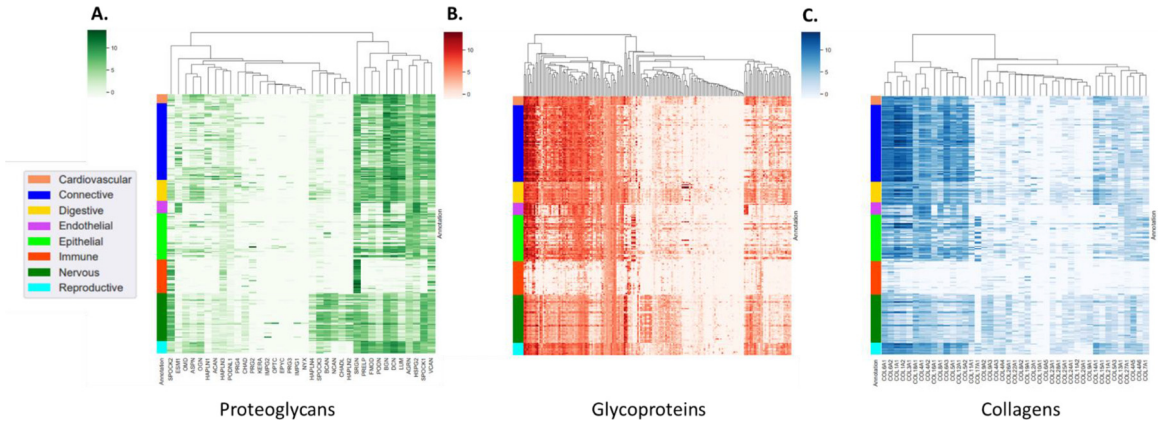

Figure 2. ClusterMaps of genes expressing proteoglycans, glycoproteins, and collagens with annotated tissue systems. Proteoglycans (green, A), glycoproteins (red, B), and collagens (blue, C) show unique patterns of expression through hierarchical clustering. Relative expression of gene promoters within hierarchical clustering maps divided into the previously defined matrisome categories.

hierarchical trees. **(Figure 2)** From this clustering, we could determine relative trends in expression between certain tissue/cell systems and their transcriptionally active ECM gene sets.

Among the proteoglycans, KERA, IMPG2, OPTC, EPYC, PRG3, IMPG1, and NYX showed little expression in almost all tissue/cell samples. HAPLN4, SPOCK3, BCAN, NCAN, CHADL, and HAPLN2 make up a unique region that is localized at higher levels to the nervous system. **(Figure 2A)** From the glycoproteins category, fibrinogens FGB, FGA, and FGG comprise one of the most highly expressed gene sets and are present in abundance in liver-derived samples, as expected. FNDC7 also is unique in this set for its near absence. **(Figure 2B) (**A fully annotated version of the glycoproteins correlation map can be found in **Supplementary Figure S10.**) Similar to what is seen for proteoglycans and glycoproteins, the immune system showed an overall lack of expression of ECM-encoding genes in the collagen data set. Distinct features in this ClusterMap include COL17A1 expression in the epithelial system and COL6A1, COL6A2, COL1A1, and COL1A2 consistently expressed in both the connective and reproductive systems. **(Figure 2C)**

Z-scores were used to enable statistical comparisons across both genes and samples, allowing for a more significant interpretation of the data. A positive (red) z-score indicates a sample with a greater relative tags per million value, as compared to the mean, which correlates to a more active promoter. A negative (blue) z-score indicates a sample that has a comparatively smaller tags per million value, indicating a lower level of transcriptional activation. Comparisons of z-scores across a gene, such as in **Supplementary Figure S1,** demonstrates the statistical significance of the differences between different samples/categories for a singular gene. Z-scores calculated across samples, such as in **Supplementary Figure S2**, demonstrates the relative distribution of transcriptionally active genes within a single sample.

To demonstrate this idea, the proteoglycan SRGN showed highest expression in the immune system, as indicated by a highly positive (red) z-score, as compared to other systems. **(Supplementary Figure S1)**. Within the immune system, proteoglycans SPOCK2, HAPLN3, and VCAN showed high levels of promoter-level expression as compared to other proteoglycans **(Supplementary Figure S2)**. In the glycoproteins group, SPARC, IGFBP7, and LAMB2 are least active in the immune system and relatively consistent in activity across the connective system. **(Supplementary Figure S3)** Across the nervous system. SPARCL1, SPP1, NELL2, IGFBP7, SPARC, IGFBP4, IGFBP5, and MGP were some of the most highly transcribed genes, while LAMA3, LAMB3, and LAMC2 appeared as major players in the epithelial system. **(Supplementary Figure S4)** In the category of collagens, COL6A2 was the most highly transcribed collagen in the immune samples. **(Supplementary Figure S6)** Connective tissues and cells dominated production of this matrisome category overall as compared to other systems. **(Supplementary Figure S5)**

Extracted cluster of interest from epithelial glycoproteins demonstrates tissue system-specific signatures that can unveil unforeseen expression patterns

ClusterMaps and z-score analysis provided high-level detail of promoter-level gene expression in all samples, so that relative trends could be assessed. Unique signatures were revealed through these analyses by employing hierarchical clustering methods and statistical analysis. Zooming in on one

cluster of interest, as an example, revealed the relative active transcription of genes encoding SPARCL1, SPP1, FGL2, EMILIN2, TNFAIP6, LAMA3, LAMB3, and LAMC2 in epithelial samples. **(Figure 3)** LAMA3, LAMB3, LAMC2, and SPP1 were actively transcribed in epithelial samples extracted in this region, while genes such as FGL2 were not. **(Supplementary Figure S7)**

While the relationship between LAMA3, LAMB3, and LAMC2 may seem obvious, other members of this cluster can provide new information into previously unknown ECM interactions in tissue/cell systems. In this example, the reason for this interaction in the dataset is not immediately obvious. SPARCL1 is not well described in epithelial cell literature but is found predominantly in neural tissues. In studies of SPARCL1 knockout mice, SPARCL1 was found to modulate the dermal ECM via regulation of decorin levels and collagen fibril assembly and functions to create intermediate states of adhesion in cells adjacent to the epithelium [17]. SPP1 is a member of a subgroup of ECM proteins known as matricellular proteins and is well-known for its involvement in the attachment of osteoclasts to the mineralized bone matrix [18–20]. Interestingly, this analysis shows kidney epithelial cells as one of the highest expressers of SPP1 (bone marrow is the other), consistent with the known disease states associated with its mutation [21]. EMILIN2 acts in various ways throughout the body, serving to induce angiogenesis in tumors via an epidermal growth factor receptor-dependent interaction [22], acting in a pro-apoptotic role[23], and forming an association with elastin-microfibrils [24]. TNFAIP6 is a hyaluronate-binding protein closely related to CD44 and is implicated in cell-cell and cell-matrix interactions during

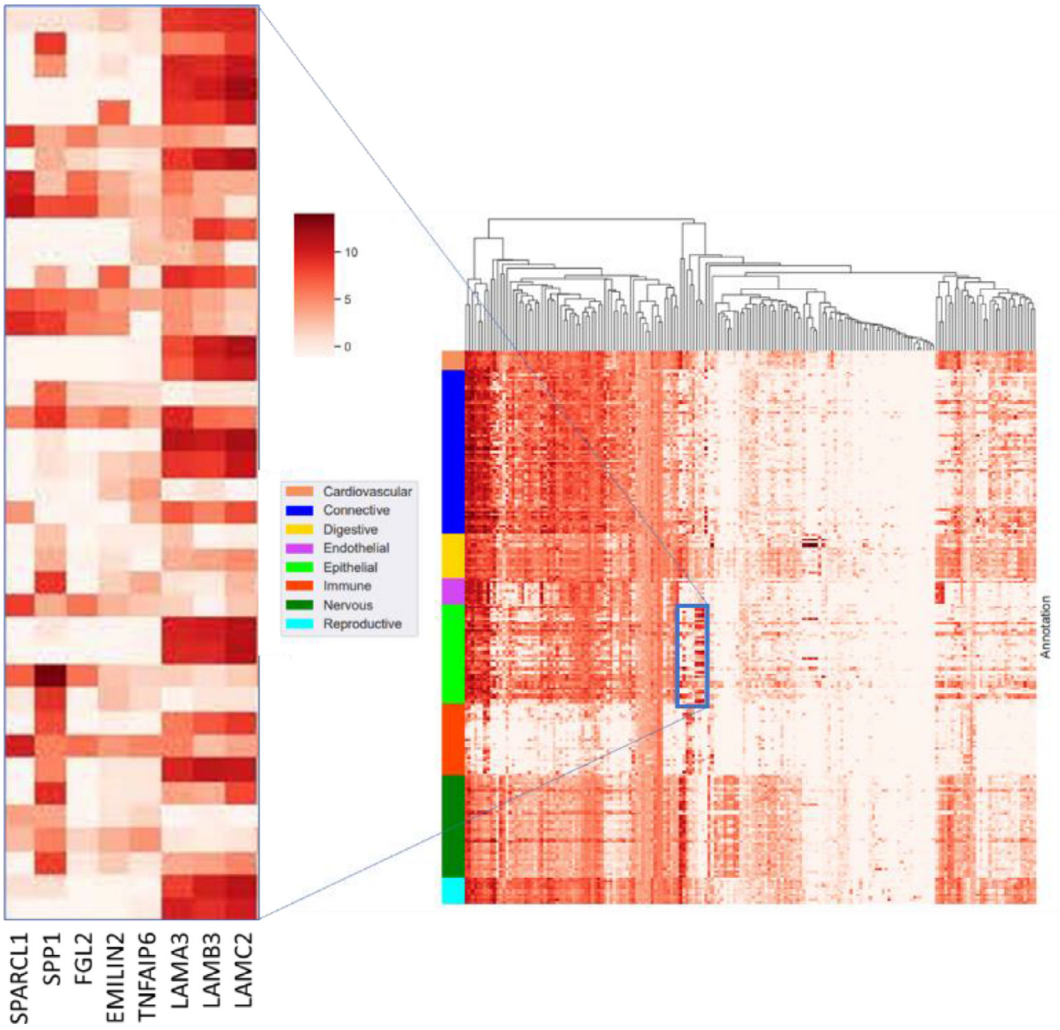

Figure 3. Extracted cluster of interest from glycoproteins category reveals ECM-encoding gene expression patterns in epithelial samples consistent with type I hemidesmosomes. The extracted cluster contains SPARCL1, SPP1, FGL2, EMILIN2, TNFAIP6, LAMA3, LAMB3, LAMC2.

inflammation and tumorigenesis [25]. This protein may also play an essential role in endometrium during the proliferative and secretory phases of the menstrual cycle [26]. LAMA3, LAMB3, LAMC2 are subunits of laminin 332 (commonly known as laminin 5) in the epithelial basement membrane and play an integral role in the formation of epithelial anchoring complexes including type I hemidesmosomes [27,28]. Mutations in laminin 332 result in junctional epidermolysis bullosa-Herlitz. where, in 80% of cases, LAMB3 is affected [29]. While this extracted group of proteins plays, in some capacity, a role in epithelial cell anchorage and movement, their interconnected relationship has not been well defined. This analysis offers potential opportunities such as this one to uncover novel interactions in ECM gene expression in unexpected tissue locations.

Correlation of matrisome components across tissue types

Correlation maps of the matrisome components across all tissue types can reveal unique interactions that exist among these ECM molecules. Positive correlation (red) indicates actively transcribed genes that behave in a similar way while negative correlation (blue) indicates a high probability that these promoters exist independently of one another. Once again, for this analysis the promoter-level expression data was divided into the three matrisome categories of proteoglycans, glycoproteins, and collagens.

One interesting example of a positive correlation cluster in the proteoglycans group is among OPTC, HAPLN2, IMPG2, NCAN, HAPLN4, and IMPG1. SRGN, another proteocglycan, was very negatively correlated with molecules such as KERA and ASPN. **(Figure 4A)** Taken together, we can appreciate that OPTC, HAPLN2, IMPG2, NCAN, HAPLN4, and IMPG1 are likely to be actively transcribed or not transcribed in unison while negatively correlated SRGN and KERA would be less likely to show similar connections in terms of promoter activity.

In the glycoproteins gene set **(Figure 4B)**, several correlation clusters stood out during data analysis. These new correlations can provide clues into previously unappreciated gene clusters that may account for common phenotypes or identify negative regulators of cell behavior. One such cluster is that of positive correlation between MATN4, KCP, RSPO1, AMELX, VWA3B, ZP1, IGSF10, ELSPBP1, and FNDC8. In terms of negative correlation, we saw LRG1 whose expression is strikingly anti-correlative with MATN3, RSPO4, LAMA1, and NTNG1, indicating these proteins exhibit distinct promoter-level behaviors from each other. To aid in visualization of the negative correlation between LRG1 and other glycoproteins, we extracted this region and included it in **Supplementary Figure S8.** LRG1 positively correlated with MATN1, DMBT1, IGFBP1, VWA5B2, FGG, FGA, FGL1, FGB, TINAG, OIT2, IGFALS, and VTN of the glycoprotein group, meaning it behaves in a similar manner to these genes. Once again, this region was extracted from the large glycoprotein matrix to provide more insight into these positive LRG1 interactions. **(Supplementary Figure S9)**

From the positive cluster, MATN4 is one of the most ubiquitously expressed matrilins in the human body and can found in epithelial, muscle, and nervous tissue as well as connective tissue of internal organs [30,31]. MATN4 is crucial to maintaining the stability of articular cartilage and interacts with various proteins to interconnect and stabilize these macromolecular networks [32]. RSPO1 and AMELX also demonstrate unique roles in the context of bone biology. RSPO1 has a bone anabolic effect, enhancing

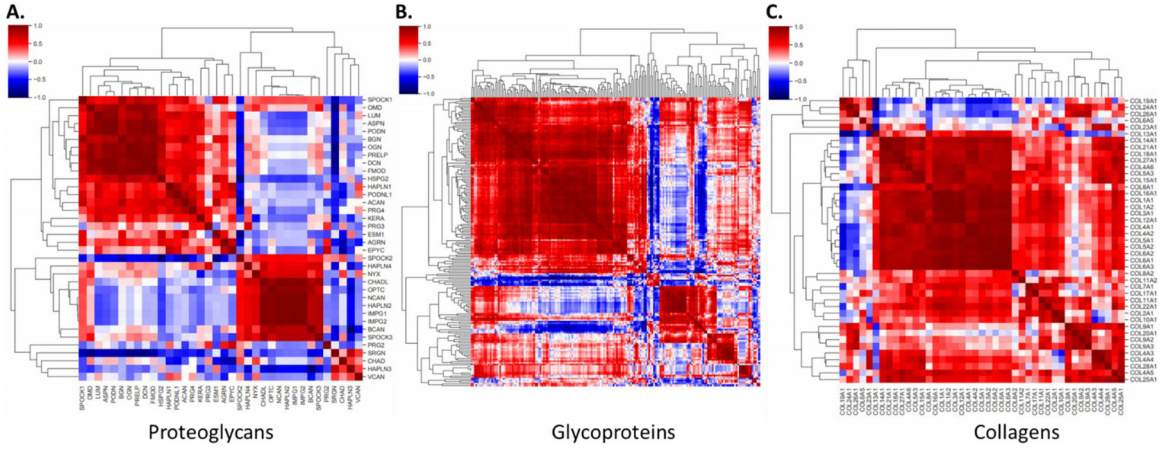

Figure 4. Correlation of matrisome components in all tissue types. Proteoglycans (A), glycoproteins (B), and collagens (C) show positive (red) and negative correlation (blue) in promoter-level gene expression among various ECM constituents.

osteogenic markers and osteoprotegerin expression [33–36] as well as inducing and enhancing osteoblast differentiation. AMELX is a protein secreted by ameloblasts to create tooth enamel [37]. KCP is a potent paracrine enhancer of bone morphogenetic protein signaling in embryonic brain and kidney samples [38], increases expression in human failing hearts [39], and aids in high fat diet-induced obesity [40]. While there is no directly described role of KCP in bone, clustering of this protein near other bone-affecting genes supports its ability to interact with and signal in bone via its known bone morphogenetic protein signaling capacity. Additionally, VWA3B is known for its effect, when mutated, in spinocerebellar ataxia [41] but has no known role in bone, meaning it could face a similar fate as KCP within this bone-associated cluster. Taken together, this analysis offers new insight into the potential role and relationship of these proteins in a new context, the connective system.

ZP1, IGSF10, and ELSPBP1 all play crucial roles in human reproduction and clustered quite readily with the lesser known FNDC8. ZP1 is expressed in the zona matrix of secondary and antral follicles, ovulated oocytes, atretic follicles, and degenerating intravascular oocytes in the female reproductive tract but also binds to capacitated spermatozoa and induces acrosomal exocytosis [42]. IGSF10 mutations caused delayed puberty and hypogonadism [1] while also exerting an important effect on breast cancer tumorigenesis [43]. ELSPBP1 known primary role is in its binding to already dead spermatozoa through epididymosomes [44–47]. The roll of FNDC8 is not well described in literature, but hierarchical clustering in this region could be indicative of its potential role in these areas, where its positive correlation indicates similar behaviors to locally clustered genes. Additionally, RSPO1, discussed in the previous paragraph in the context of bone, has two hormone-related roles in the body. RSPO1 can play a role in sex determination via Wnt4/β-catenin signaling during ovarian development [48,49] and is required for normal development of the mammary gland [50]. From the more negatively glycoprotein cluster, LRG1 can function in signal transduction, cell proliferation, cell migration, cell invasion, cell adhesion, cell survival, and cell apoptosis [51], roles which we believe would make it more likely to positively correlate with other genes, making this negative correlation even more striking.

Collagens traditionally have been thought of in the context of deposition by fibroblasts in connective tissue, but this analysis provides insight into other cell and tissue types that express collagen family members [52]. COL19A1 was unique in this dataset for its large negative correlation with other members of the collagen family. **(Figure 4C)** Interestingly, in mouse studies of Col19a1, mRNA can be found in all tissues except liver. In the adult mouse however, Col19a1 mRNA was largely limited to the brain, which may explain its limited correlation with other matrix components [53]. COL5A1, COL5A2, COL3A1, COL16A1, COL1A1, and COL1A2 represent a positively correlated cluster in this family, indicative of similar trends in promoter-level expression across the human body. COL5A1, COL5A2, COL3A1, COL1A1, and COL1A2 are all fibril-forming collagens, while COL16A1 is a fibril-associated collagen [10]. COL2A1 and COL3A1 are present in hyaline cartilage [54], COL1A1, COL1A2, and COL3A1 are present in skin, while COL1A1, COL1A2, COL5A1, and COL5A2 are found in the cornea [55]. COL16A1 is a component of microfibrils containing fibrillin-1 in skin alongside COL2A1, COL2A2, COL11A1, and COL11A2 [56]. While all of these collagens appear to be associated with a broad array of tissues, this analysis offers a better look at how these genes may interact at the promoter level across different tissues and cells where they are actively transcribed.

Extracted correlation map of proteoglycans in epithelial samples

Interesting patterns emerged when individual systems were extracted from the data and correlation analyses were performed. In **Figure 5,** a correlation map was utilized to provide insight into the expression relationships of proteoglycans in the epithelial system alone. Pairs such as HAPLN2 and SPOCK3 and BCAN and HAPLN4 showed positive correlation between genes. Larger clusters such as PRELP, FMOD and PODN; BGN, SRGN, DCN, and LUM; CHADL, CHAD, and KERA also exhibited positive correlation where positive correlation (red) means two genes are likely to behave in the same way. OPTC and HSPG2, NYX and HAPLN3, and PODNL3, BCAN, and HAPLN4 were all negatively correlated ECM-encoding genes. This negative correlation indicates that this set of promoters have an opposite relationship, with one being more active when the other is less active. Taken together, these clustering regions indicate regions of either similar or dissimilar behavior within ECM promoters of the epithelial system, showing a system-specific signature of gene behavior as opposed to a signature across all tissues/cell systems. This type of extraction allows for a clearer idea of how one system may function in comparison to all systems in the body.

Open-Access Data Through Web Application Deployment

Modern web frameworks make published data easily accessible to the larger scientific community. The Dart/Flutter Software Development Kit is an open-source Google project designed for cross-

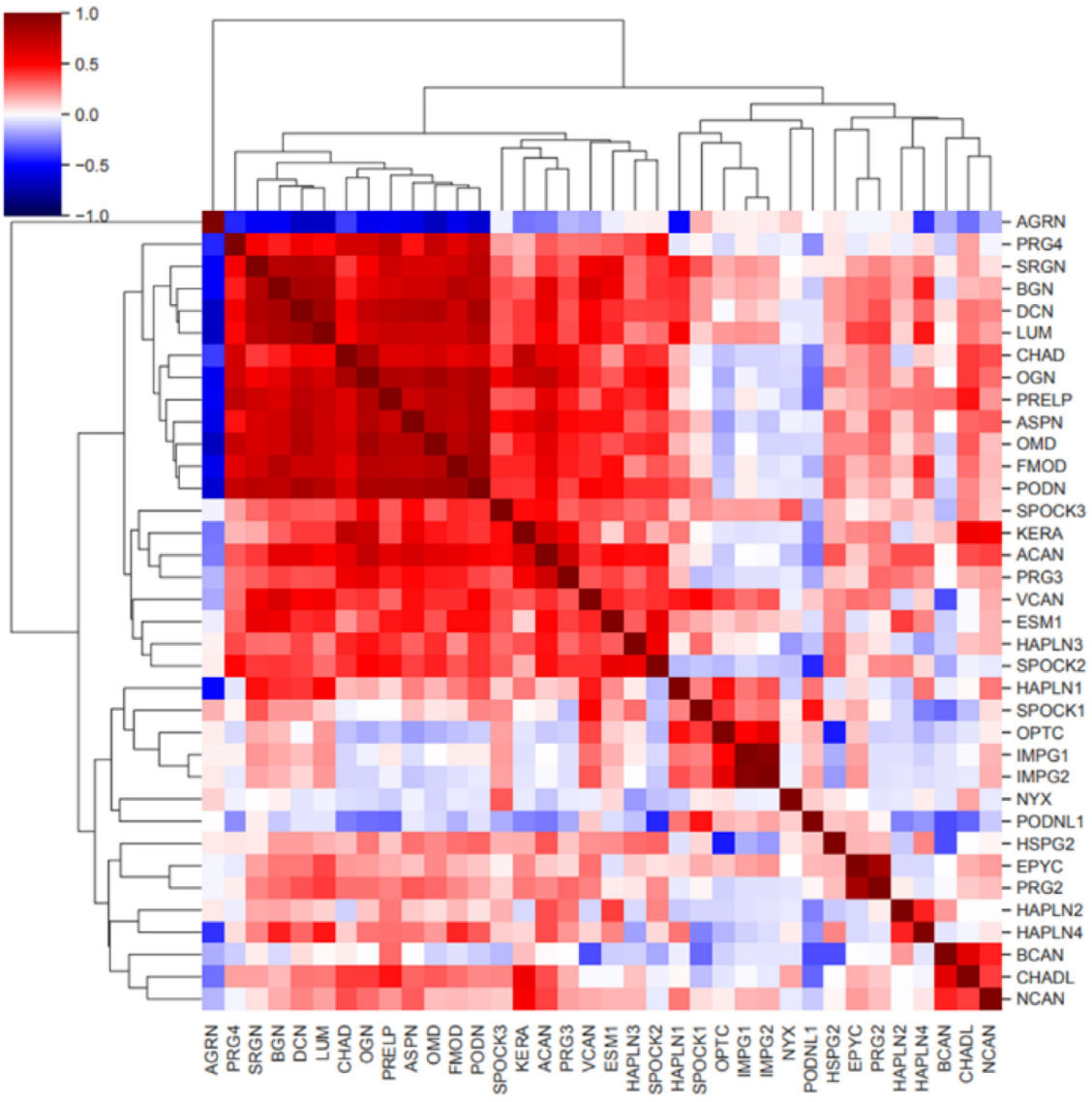

Figure 5. Correlation map of proteoglycans in the epithelial system.

platform applications, which was used to develop a GitHub pages website (matrixpromoterome.github.io) to share the log-transformed promoter activity data. Widgets for data queries, interactive graphing, and downloads allow for more granular insight into a matrix-promoting gene of interest's promoter activity in a particular tissue or cell type. Furthermore, high resolution images can be found at dx.doi.org/10.6084/m9.figshare.19794481. The code used for this analysis can also be found as a Python notebook in the figshare repository at https://figshare.com/s/e18ecbc3ae5aaf919b78.

Discussion

Since it was published in 2011, the concept of the matrisome has remained a landmark work in the field of matrix biology. For a decade, this analysis has provided major insight into the ECM and shifted perceptions of links between structure and function. With the analysis presented in this paper, we further establish the importance of the matrisome and offer a new tool to interrogate the dynamics of human matrix biology at the gene promoter level. Our analysis is unique in the comprehensive nature of the data from the FANTOM5 database as it bypasses previous limitations of sample size and consistent experimental value acquisition. Additionally, this analysis examines specific promoter-level gene expression as opposed to relative protein levels, indicative of what is being actively transcribed in cells and tissues. It is important to note that this data is uninformative about steady state RNA levels or post-transcriptional regulation, revealing only the amount of active transcription occurring in the system. Additionally, this dataset is internally calibrated,

as all biological samples were shipped to Riken under prescribe conditions after which they were prepared and analyzed at the same site, eliminating any technical differences that could have arisen from protocol deviations at the contributing sites. We use the epithelial system as an example of how this information can be utilized and recognize the value of this type of analysis in other systems as well.

In this analysis, immortalized cell lines were excluded from the data set due to existing research on the consequences of long-term cell culture on human cell lines. In the premier FANTOM5 paper, in-depth analysis on primary tissues/cells vs immortalized cell lines shows that cell lines are far more likely to cluster with themselves than with their derived tissue, indicative of the changes that occur in these cell lineages over time [57]. As one specific example, studies have shown that the cultured cells adapt to their culture environment (e.g. plastic) to express non-physiological levels of ECM-binding proteins such as the vitronectin receptor [58]. While this text has a tissue-centric organization, it is important to remember that this analysis includes immune cells as well as epithelial cells regardless of tissue organ, endothelial cells regardless of tissue origin, and some specialized cells that each have their own unique signature imbedded in the analysis. In the clustermaps in **Figure 2**, each row represents each unique cell sample, grouped into their respective tissue systems. We mention briefly in the results section the difficulty of fitting our dataset to the traditional ten physiological system nomenclature. Given our observations with these datasets, the authors argue that the traditional ten system nomenclature that is commonly used may need to be reevaluated as gene expression data at the single cell level becomes common. As an example, an epithelial cell in the breast is more like an epithelial cell in the colon than it is like an adipocyte in the breast, where anatomical constriction confounds the data rather than clarifying it.

Correlation analysis of the ECM as performed here offers new unbiased insights into tissue locations and potential interactions where ECM genes are being actively transcribed. Such information can inform future interrogations of the roles of specific ECM constituents in tissue formation, repair, and disease. To our knowledge, this is the first data to reveal the transcription activity of the matrisome, with earlier analyses rather focused on steady state levels of proteins and mRNAs. Gaining an understanding of this transcription provides an opportunity to consider tissue- and cell-level differences as they relate to control of transcription that can determine the types of matrix that cells produce. We unveiled examples of well-established relationships among ECM components while also revealing potentially new tissue-level interactions. While we focused here only on normal cells and tissues, this analysis can provide insights into the potential involvement of various ECM molecules in human developmental disorders. In analyses where tissue specimens were grouped, it is important to recognize the potential limitations of correlation. Considering different tissue types individually can reveal interactions that are more predominant in some systems over others, explaining why certain mutations cause changes only in some tissues. This analysis can be paired with results of mouse gene knockout strategies where often unpredicted phenotypes are seen, examples of which include osteopontin and small leucine-rich proteoglycan deficiencies [59,60]

For this analysis we focused only on the interactions of the core matrisome, but the importance of matrix-associated genes could be considered in future analyses. Many disease states are associated not just with the matrix itself, but also with those molecules that interact with and modify the matrix environment. For example, levels of various matrix metalloproteinases can support the aggressive metastasis of various cancers [61,62]. Understanding how these modifiers interact with the core matrisome in normal vs disease states could help identify new targets for novel therapeutics to prevent dysregulation of proteins.

The matrix promoter-level expression dataset is a novel example of the power of large data analytics and how it can inform more basic biological questions. This analysis is unique in its specific interrogation of the promoter-level expression of the matrisome genes in the context of normal, homeostatic tissues/cells and provides a unique opportunity to interrogate not just context-specific expression but the relationships that may exist between these ECM genes. This analysis provides insights that can be leveraged by both matrix biologists and clinicians alike, as they seek to better understand the role of the ECM in human biology. We offer this analysis to the matrix community at large as a resource for exploring the ECM and invite members of the community to dig into this data further via the publicly available annotated raw dataset that can be found at http://farachcarsonwulab.com/. We believe this analysis functions as a living document and this is simply the first iteration of the matrix promoter-level gene expression data set that we invite the community to explore and discuss as part of a larger conversation.

Experimental Procedures

Data Processing

Relative log expression (RLE) normalized SSTAR CAGE peak data files annotated with hg19human reference genome assembly were downloaded from

the FANTOM5 online database for both phase 1 and 2 using FANTOM5 Table Extraction Tool (TET) found at https://fantom.gsc.riken.jp/5/tet/#!/search/hg19.cage_peak_counts_ann_decoded.osc.txt.gz. The data set was filtered to exclude all cell lines, diseased tissues and samples with missing values resulting in a total of 511 samples across all tissue systems (3 samples were removed due to missing values (nan s). Genes encoding the core matrisome were extracted and utilized for this analysis. First, the individual CAGE peak tags per million expression values of each gene from the sample set were summed to obtain a gene level value. A pseudocount of 0.5 was added across the data set before logarithm transformation in order to prevent undefined values where data values are zero prior to downstream statistical analysis. Samples were further categorized into 8 core tissue systems and genes were annotated into 3 major ECM categories: proteoglycans, glycoproteins, and collagens. CAGE peak data for each category was gathered for all 511 samples in a matrix.

Data Analysis

All analysis was executed using Python version 3.6 [63] using multiple packages including pandas [64], NumPy [65], the sklearn.metrics, sklearn.utils, in SciKits [66], Matplotlib [67], seaborn [68], and scipy [69]. The seaborn.clustermap function of the seaborn package was used to perform hierarchical clustering of the heatmaps for each gene category using the average linkage representing the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method and the Euclidean distance metric to calculate the distance between each new cluster and the remaining cluster. The ClusterMaps were annotated with color bars indicating the core tissue systems of the samples. To demonstrate if the differences of CAGE expression across ECM-encoding genes and the differences across samples were statistically significant, we converted the expression values to z-scores. We did this by applying z-score transformation to the log2 transformed data using the z-score statistical function from the stats module of the scipy package in Python. To investigate the co-expression patterns of genes in each ECM category and determine complex relationships in promoter activity, correlation analyses were conducted. For these analyses, the log transformed data was used and the Pearson correlations were calculated by evaluating pairwise correlations of all gene pairs after excluding all null values. To ensure the reproducibility of our results, all code used for this analysis is available as a python notebook in the figshare repository https://figshare.com/s/e18ecbc3ae5aaf919b78.

Author Contributions

T.V.T. conceptualization, investigation, methodology, validation, visualization, writing – original draft, writing – review & editing; M.D. data curation, formal data analysis, investigation, methodology, software, validation, visualization, writing – original draft, writing – review & editing; V.A.A. visualization, writing – review & editing; D.S. methodology, visualization, writing – review& editing; A.N. conceptualization, writing – review & editing; M.C.F.C. conceptualization, supervision, visualization, writing – review & editing

Funding

This funding was supported by the National Institute of Dental & Craniofacial Research [National Institutes of Health R01 DE022969]; National Institute of Dental & Craniofacial Research [National Institutes of Health R56 DE026530]; the UTHealth Innovation for Cancer Prevention Research Training Program Pre- and Post-doctoral Fellowship, Houston, TX [Cancer Prevention and Research Institute of Texas RP160015 & RP210042]; the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dr. John J. Kopchick Fellowship, Houston, TX; the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Larry Deaven Ph.D. Fellowship in Biomedical Sciences, Houston, TX; the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Schissler Foundation Fellowship.

Conflict of Interest

The authors, T.V.T., M.D., V.A.A., D.S., A.N. and M.C.F-C., declare no conflicts of interest.

Acknowledgements

We would like to acknowledge the RIKEN consortium and thank them for making the FANTOM5 data publicly available for the scientific community.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.matbio.2022.06.003.

Received 31 January 2022;
Received in revised form 20 May 2022;

Accepted 13 June 2022
Available online 14 June 2022

Keywords:
matrisome;
tissue level analysis;
promoter-level expression;
cap analysis gene expression;
clustermaps

Abbreviations:
ECM, extracellular matrix; SSTAR, Semantic catalog of Samples, Transcription initiation And Regulators; RLE, relative log expression; CAGE, cap analysis gene expression; UPGMA, Unweighted Pair Group Method with Arithmetic Mean

[1]These authors contributed equally to this work

References

[1] S.R. Howard, L. Guasti, G. Ruiz-Babot, A. Mancini, A. David, H.L. Storr, L.A. Metherell, M.J. Sternberg, C.P. Cabrera, H.R. Warren, M.R. Barnes, R. Quinton, N. de Roux, J. Young, A. Guiochon-Mantel, K. Wehkalampi, V. André, Y. Gothilf, A. Cariboni, L. Dunkel, IGSF10 mutations dysregulate gonadotropin-releasing hormone neuronal migration resulting in delayed puberty, EMBO Mol. Med. 8 (2016) 626, doi: 10.15252/EMMM.201606250.

[2] 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. Durbeej, N.A. Afratis, L. Troeberg, M. Franchi, V. Masola, M. Onisto, A guide to the composition and functions of the extracellular matrix, FEBS J 288 (2021) 6850–6912, doi: 10.1111/FEBS.15776.

[3] J.H. Hughes, J.M. Ewy, J. Chen, S.Y. Wong, K.M. Tharp, A. Stahl, S. Kumar, Transcriptomic analysis reveals that BMP4 sensitizes glioblastoma tumor-initiating cells to mechanical cues, Matrix Biol 85−86 (2020) 112–127, doi: 10.1016/J.MATBIO.2019.06.002.

[4] W.P. Daley, S.B. Peters, M. Larsen, Extracellular matrix dynamics in development and regenerative medicine, J. Cell Sci. 121 (2008) 255–264, doi: 10.1242/JCS.006064.

[5] J.M. Gebauer, A. Naba, The Matrisome of Model Organisms: From In-Silico Prediction to Big-Data Annotation, (2020) 17−42. https://doi.org/10.1007/978-3-030-58330-9_2.

[6] A. Naba, K.R. Clauser, H. Ding, C.A. Whittaker, S.A. Carr, R.O. Hynes, The extracellular matrix: Tools and insights for the “omics” era, Matrix Biol (2015), doi: 10.1016/j.matbio.2015.06.003.

[7] R.O. Hynes, A. Naba, Overview of the matrisome−an inventory of extracellular matrix constituents and functions, Cold Spring Harb. Perspect. Biol. 4 (2012), doi: 10.1101/CSHPERSPECT.A004903.

[8] 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 (2011), doi: 10.1074/MCP.M111.014647.

[9] R.V. Iozzo, L. Schaefer, Proteoglycan form and function: A comprehensive nomenclature of proteoglycans, Matrix Biol 42 (2015) 11–55, doi: 10.1016/J.MATBIO.2015.02.003.

[10] S. Ricard-Blum, The Collagen Family, Cold Spring Harb. Perspect. Biol. 3 (2011) 1–19, doi: 10.1101/CSHPERSPECT.A004978.

[11] M. Ly, T.N. Laremore, R.J. Linhardt, Proteoglycomics: Recent Progress and Future Challenges, Https://Home.Liebertpub.Com/Omi. 14 (2010) 389−399. https://doi.org/10.1089/OMI.2009.0123.

[12] J.M. Tarbell, L.M. Cancel, The glycocalyx and its significance in human medicine, J. Intern. Med. 280 (2016) 97–113, doi: 10.1111/JOIM.12465.

[13] M.D. Shoulders, R.T. Raines, Collagen Structure and Stability, Annu. Rev. Biochem. 78 (2009) 929, doi: 10.1146/ANNUREV.BIOCHEM.77.032207.120833.

[14] F. Sacher, C. Feregrino, P. Tschopp, C.Y. Ewald, Extracellular matrix gene expression signatures as cell type and cell state identifiers, Matrix Biol. Plus. 10 (2021) 100069, doi: 10.1016/J.MBPLUS.2021.100069.

[15] T.O. Nieuwenhuis, A.Z. Rosenberg, M.N. McCall, M.K. Halushka, Tissue, age, sex, and disease patterns of matrisome expression in GTEx transcriptome data, Sci. Reports 11 (2021) 1–14 2021 111, doi: 10.1038/s41598-021-00943-x.

[16] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C.J. Mungall, E. Arner, J.K. Baillie, N. Bertin, H. Bono, M. de Hoon, A.D. Diehl, E. Dimont, T.C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T.F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E.A. Schultes, P.A. ‘t Hoen, Z. Tatum, M. Thompson, T. Toyoda, D.W. Wright, C.O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A.R. Forrest, H. Kawaji, Gateways to the FANTOM5 promoter level mammalian expression atlas, Genome Biol. 161 (16) (2015) 1–14 2015, doi: 10.1186/S13059-014-0560-6.

[17] M.M. Sullivan, T.H. Barker, S.E. Funk, A. Karchin, N.S. Seo, M. Höök, J. Sanders, B. Starcher, T.N. Wight, P. Puolakkainen, E.H. Sage, Matricellular Hevin Regulates Decorin Production and Collagen Assembly *, J. Biol. Chem. 281 (2006) 27621–27632, doi: 10.1074/JBC.M510507200.

[18] F.P. Ross, J. Chappel, J.I. Alvarezs, D. Sanderll, W.T. Butler, M.C. Farach-Carson11, K.A. Mintz, P.G. Robey, S.L. Teitelbaums, D.A. Chereshll, Interactions between the bone matrix proteins osteopontin and bone sialoprotein and the osteoclast integrin alpha v beta 3 potentiate bone resorption, J. Biol. Chem. 268 (1993) 9901–9907, doi: 10.1016/S0021-9258(18)98430-9.

[19] F.P. Reinholt, K. Hultenby, A. Oldberg, D. Heinegård, Osteopontin−a possible anchor of osteoclasts to bone, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 4473, doi: 10.1073/PNAS.87.12.4473.

[20] J.E. Murphy-Ullrich, E.H. Sage, Revisiting the matricellular concept, Matrix Biol 37 (2014) 1–14, doi: 10.1016/J.MATBIO.2014.07.005.

[21] B. Kaleta, The role of osteopontin in kidney diseases, Inflamm. Res. 682 (68) (2018) 93–102 2018, doi: 10.1007/S00011-018-1200-5.

[22] A. Paulitti, E. Andreuzzi, D. Bizzotto, R. Pellicani, G. Tarticchio, S. Marastoni, C. Pastrello, I. Jurisica, G. Ligresti, F. Bucciotti, R. Doliana, R. Colladel,

P. Braghetta, E. Poletto, A. Di Silvestre, G. Bressan, A. Colombatti, P. Bonaldo, M. Mongiat, The ablation of the matricellular protein EMILIN2 causes defective vascularization due to impaired EGFR-dependent IL-8 production affecting tumor growth, Oncogene 3725 (37) (2018) 3399–3414 2018, doi: 10.1038/s41388-017-0107-x.

[23] A. Colombatti, P. Spessotto, R. Doliana, M. Mongiat, G.M. Bressan, G. Esposito, The EMILIN/Multimerin Family, Front. Immunol. (2012) 93, doi: 10.3389/FIMMU.2011.00093.

[24] G.M. Bressan, D. Daga-Gordini, A. Colombatti, I. Castellani, V. Marigo, D. Volpin, Emilin, a Component of Elastic Fibers Preferentially Located at the Elastin-Microfibrils Interface, 2022 (n.d.).

[25] T. Lee, H. Wisniewski, J. Vilcek, A novel secretory tumor necrosis factor-inducible protein (TSG-6) is a member of the family of hyaluronate binding proteins, closely related to the adhesion receptor CD44, J. Cell Biol. 116 (1992) 545–557, doi: 10.1083/JCB.116.2.545.

[26] E. Capp, C.M. Milner, J. Williams, L. Hauck, J. Jauckus, T. Strowitzki, A. Germeyer, Modulation of tumor necrosis factor-stimulated gene-6 (TSG-6) expression in human endometrium, Arch. Gynecol. Obstet. 289 (2014) 893–901, doi: 10.1007/S00404-013-3080-9.

[27] P. Rousselle, K. Beck, Laminin 332 processing impacts cellular behavior, Cell Adh. Migr. 7 (2013) 122–134, doi: 10.4161/cam.23132.

[28] H. Schneider, C. Mühle, F. Pacho, Biological function of laminin-5 and pathogenic impact of its deficiency, Eur. J. Cell Biol. 86 (2007) 701–717, doi: 10.1016/J.EJCB.2006.07.004.

[29] D. Kiritsi, C. Has, L. Bruckner-Tuderman, Laminin 332 in junctional epidermolysis bullosa, Cell Adh. Migr. 7 (2013) 135, doi: 10.4161/CAM.22418.

[30] A.R. Klatt, M. Paulsson, R. Wagener, Expression of matrilins during maturation of mouse skeletal tissues, Matrix Biol 21 (2002) 289–296, doi: 10.1016/S0945-053X(02)00006-9.

[31] R. Wagener, B. Kobbe, M. Paulsson, Matrilin-4, a new member of the matrilin family of extracellular matrix proteins, FEBS Lett 436 (1998) 123–127, doi: 10.1016/S0014-5793(98)01111-9.

[32] P. Li, L. Fleischhauer, C. Nicolae, C. Prein, Z. Farkas, M.M. Saller, W.C. Prall, R. Wagener, J. Heilig, A. Niehoff, H. Clausen-Schaumann, P. Alberton, A. Aszodi, Mice Lacking the Matrilin Family of Extracellular Matrix Proteins Develop Mild Skeletal Abnormalities and Are Susceptible to Age-Associated Osteoarthritis, Int. J. Mol. Sci. 21 (2020) 666, doi: 10.3390/IJMS21020666.

[33] H. Wang, T.A. Brennan, E. Russell, J.-H. Kim, K.P. Egan, Q. Chen, C. Israelite, D.C. Schultz, F.B. Johnson, R.J. Pignolo, R-spondin 1 promotes vibration-induced bone formation in mouse models of osteoporosis, J. Mol. Med. 9112 (91) (2013) 1421–1429 2013, doi: 10.1007/S00109-013-1068-3.

[34] G. Krönke, S. Uderhardt, K.-A. Kim, M. Stock, C. Scholtysek, M.M. Zaiss, C. Surmann-Schmitt, J. Luther, J. Katzenbeisser, J.-P. David, S. Abdollahi-Roodsaz, K. Tran, J.M. Bright, M.E. Binnerts, A. Akhmetshina, C. Böhm, J.H. Distler, L.A.B. Joosten, G. Schett, A. Abo, R-spondin 1 protects against inflammatory bone damage during murine arthritis by modulating the Wnt pathway, Arthritis Rheum 62 (2010) 2303–2312, doi: 10.1002/ART.27496.

[35] A. Sharma, B. Choi, J. Park, D. Lee, J. Lee, H. Kim, J. Yoon, D. Song, J. Nam, S. Lee, Rspo 1 promotes osteoblast differentiation via Wnt signaling pathway, Indian J. Biochem. Biophys. 50 (2013) 19–25 https://pubmed.ncbi.nlm.nih.gov/23617070/ accessed August 19, 2021.

[36] W. Lu, K.-A. Kim, J. Liu, A. Abo, X. Feng, X. Cao, Y. Li, R-spondin1 synergizes with Wnt3A in inducing osteoblast differentiation and osteoprotegerin expression, FEBS Lett 582 (2008) 643–650, doi: 10.1016/J.FEBSLET.2008.01.035.

[37] C.W. Gibson, The Amelogenin Proteins and Enamel Development in Humans and Mice, J. Oral Biosci. 53 (2011) 248, doi: 10.2330/JORALBIOSCI.53.248.

[38] J. Lin, S. Patel, X. Cheng, E. Cho, I. Levitan, M. Ullenbruch, S. Phan, J. Park, G. Dressler, Kielin/chordin-like protein, a novel enhancer of BMP signaling, attenuates renal fibrotic disease, Nat. Med. 11 (2005) 387–393, doi: 10.1038/NM1217.

[39] J. Ye, Z. Wang, M. Wang, Y. Xu, T. Zeng, D. Ye, J. Liu, H. Jiang, Y. Lin, J. Wan, Increased kielin/chordin-like protein levels are associated with the severity of heart failure, Clin. Chim. Acta. 486 (2018) 381–386, doi: 10.1016/J.CCA.2018.08.033.

[40] A. Soofi, K.I. Wolf, M.P. Emont, N. Qi, G. Martinez-Santibanez, E. Grimley, W. Ostwani, G.R. Dressler, The kielin/chordin-like protein (KCP) attenuates high-fat diet-induced obesity and metabolic syndrome in mice, J. Biol. Chem. 292 (2017) 9051–9062, doi: 10.1074/JBC.M116.771428.

[41] T. Kawarai, A. Tajima, Y. Kuroda, N. Saji, A. Orlacchio, H. Terasawa, H. Shimizu, Y. Kita, Y. Izumi, T. Mitsui, I. Imoto, R. Kaji, A homozygous mutation of VWA3B causes cerebellar ataxia with intellectual disability, J. Neurol. Neurosurg. Psychiatry. 87 (2016) 656–662, doi: 10.1136/JNNP-2014-309828.

[42] A. Ganguly, A. Bukovsky, R. Sharma, P. Bansal, B. Bhandari, S. Gupta, In humans, zona pellucida glycoprotein-1 binds to spermatozoa and induces acrosomal exocytosis, Hum. Reprod. 25 (2010) 1643–1656, doi: 10.1093/HUMREP/DEQ105.

[43] M. Wang, M. Dai, Y. Wu, Z. Yi, Y. Li, G. Ren, Immunoglobulin superfamily member 10 is a novel prognostic biomarker for breast cancer, PeerJ 8 (2020), doi: 10.7717/PEERJ.10128.

[44] A. Saalmann, S. Mu, È. Nz, K. Ellerbrock, R. Ivell, C. Kirchhoff, Novel Sperm-Binding Proteins of Epididymal Origin Contain Four Fibronectin Type II-Modules, Mol. Reprod. Dev. 58 (2001) 88–100.

[45] O. D'Amours, G. Frenette, L.-J. Bordeleau, N. Allard, P. Leclerc, P. Blondin, R. Sullivan, Epididymosomes Transfer Epididymal Sperm Binding Protein 1 (ELSPBP1) to Dead Spermatozoa During Epididymal Transit in Bovine, Biol. Reprod. 87 (2012) 94–95, doi: 10.1095/BIOLREPROD.112.100990.

[46] L. Candenas, R. Chianese, Exosome Composition and Seminal Plasma Proteome: A Promising Source of Biomarkers of Male Infertility, Int. J. Mol. Sci. 21 (2020) 7022 2020217022, doi: 10.3390/IJMS21197022.

[47] M. Ekhlasi-Hundrieser, B. Schäfer, U. Philipp, H. Kuiper, T. Leeb, M. Mehta, C. Kirchhoff, E. Töpfer-Petersen, Sperm-binding fibronectin type II-module proteins are genetically linked and functionally related, Gene 392 (2007) 253–265, doi: 10.1016/J.GENE.2007.01.002.

[48] P. Parma, O. Radi, V. Vidal, M.C. Chaboissier, E. Dellambra, S. Valentini, L. Guerra, A. Schedl, G. Camerino, R-spondin1 is essential in sex determination, skin differentiation and malignancy, Nat. Genet. 3811 (38) (2006) 1304–1309 2006, doi: 10.1038/ng1907.

[49] K. Tomizuka, K. Horikoshi, R. Kitada, Y. Sugawara, Y. Iba, A. Kojima, A. Yoshitome, K. Yamawaki, M. Amagai, A. Inoue, T. Oshima, M. Kakitani, R-spondin1 plays an

essential role in ovarian development through positively regulating Wnt-4 signaling, Hum. Mol. Genet. 17 (2008) 1278–1291, doi: 10.1093/HMG/DDN036.

[50] S. Chadi, L. Buscara, C. Pechoux, J. Costa, J. Laubier, M.C. Chaboissier, E. Pailhoux, J.L. Vilotte, E. Chanat, F. Le Provost, R-spondin1 is required for normal epithelial morphogenesis during mammary gland development, Biochem. Biophys. Res. Commun. 390 (2009) 1040–1043, doi: 10.1016/J.BBRC.2009.10.104.

[51] A. Zhang, H. Fang, J. Chen, L. He, Y. Chen, Role of VEGF-A and LRG1 in Abnormal Angiogenesis Associated With Diabetic Nephropathy, Front. Physiol. (2020) 1064, doi: 10.3389/FPHYS.2020.01064.

[52] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Fibroblasts and Their Transformations: The Connective-Tissue Cell Family, (2002). https://www.ncbi.nlm.nih.gov/books/NBK26889/(accessed November 30, 2021).

[53] H. Sumiyoshi, K. Inoguchi, M. Khaleduzzaman, Y. Ninomiya, H. Yoshioka, Ubiquitous Expression of the α1(XIX) Collagen Gene (Col19a1) during Mouse Embryogenesis Becomes Restricted to a Few Tissues in the Adult Organism *, J. Biol. Chem. 272 (1997) 17104–17111, doi: 10.1074/JBC.272.27.17104.

[54] J.-J. Wu, M.A. Weis, L.S. Kim, D.R. Eyre, Type III Collagen, a Fibril Network Modifier in Articular Cartilage, J. Biol. Chem. 285 (2010) 18537, doi: 10.1074/JBC.M110.112904.

[55] P. Bruckner, Suprastructures of extracellular matrices: paradigms of functions controlled by aggregates rather than molecules, Cell Tissue Res 3391 (339) (2009) 7–18 2009, doi: 10.1007/S00441-009-0864-0.

[56] A. Kassner, U. Hansen, N. Miosge, D. Reinhardt, T. Aigner, L. Bruckner-Tuderman, P. Bruckner, S. Grässel, Discrete integration of collagen XVI into tissue-specific collagen fibrils or beaded microfibrils, Matrix Biol 22 (2003) 131–143, doi: 10.1016/S0945-053X(03)00008-8.

[57] T. FANTOM Consortium, the Riken Pmi, A promoter-level mammalian expression atlas, (2014). https://doi.org/10.1038/nature13182.

[58] M. Horton, Vitronectin receptor: tissue specific expression or adaptation to culture? Int. J. Exp. Pathol. 71 (1990) 741 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2001983/ accessed January 24, 2022.

[59] L. Ameye, M.F. Young, Mice deficient in small leucine-rich proteoglycans: novel in vivo models for osteoporosis, osteoarthritis, Ehlers-Danlos syndrome, muscular dystrophy, and corneal diseases, Glycobiology 12 (2002) 107–116.

[60] S.R. Rittling, D.T. Denhardt, Osteopontin function in pathology: lessons from osteopontin-deficient mice, Exp. Nephrol. 7 (1999) 103–113, doi: 10.1159/000020591.

[61] K. Kessenbrock, V. Plaks, Z. Werb, Matrix Metalloproteinases: Regulators of the Tumor Microenvironment, Cell 141 (2010) 52–67, doi: 10.1016/J.CELL.2010.03.015.

[62] N.M. Hooper, Y. Itoh, H. Nagase, Matrix metalloproteinases in cancer, Essays Biochem 38 (2002) 21–36, doi: 10.1042/BSE0380021.

[63] G. Van Rossum, F.l. Drake, Python 3 Reference Manual, Scotts Valley, CA, 2009 http://citebay.com/how-to-cite/python/ accessed August 19, 2021.

[64] J. Reback, jbrockmendel, W. McKinney, J. Van den Bossche, T. Augspurger, P. Cloud, S. Hawkins, gfyoung, Sinhrks, M. Roeschke, A. Klein, T. Petersen, J. Tratner, C. She, W. Ayd, P. Hoefler, S. Naveh, M. Garcia, J. Schendel, A. Hayden, D. Saxton, R. Shadrach, M.E. Gorelli, F. Li, V. Jancauskas, attack68, A. McMaster, P. Battiston, S. Seabold, K. Dong, pandas-dev/pandas: Pandas 1.3.2, (2021). https://doi.org/10.5281/ZENODO.5203279.

[65] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nat. 585 (2020) (2020) 357–362 5857825, doi: 10.1038/s41586-020-2649-2.

[66] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A.C. Müller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, (n.d.). https://github.com/scikit-learn (accessed August 19, 2021).

[67] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, doi: 10.1109/MCSE.2007.55.

[68] M.L. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021, doi: 10.21105/JOSS.03021.

[69] P. Virtanen, R. Gommers, T.E. Oliphant, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 173 (17) (2020) 261–272 2020, doi: 10.1038/s41592-019-0686-2.