Matrix Biology Plus 23 (2024) 100156

Contents lists available at ScienceDirect

Matrix Biology Plus

journal homepage: www.sciencedirect.com/journal/matrix-biology-plus

Analysis of cancer cell line and tissue RNA sequencing data reveals an essential and dark matrisome

Joshua A. Rich [a,1], Yu Fan [b,1], Qingrong Chen [b], Daoud Meerzaman [b], William G. Stetler-Stevenson [a], David Peeney [a,*]

[a] Extracellular Matrix Pathology Section, Laboratory of Pathology, National Cancer Institute, National Institute of Health, Bethesda, MD, United States
[b] Computational Genomics and Bioinformatics Branch, Center for Biomedical Informatics & Information Technology, National Cancer Institute, National Institute of Health, Rockville, MD, United States

ARTICLE INFO

Keywords:
Matrisome
RNA sequencing
Cancer cell line encyclopedia
Genotype-tissue expression

ABSTRACT

Extracellular matrix remodeling is a hallmark of tissue development, homeostasis, and disease. The processes that mediate remodeling, and the consequences of such, are the topic of extensive focus in biomedical research. Cell culture methods represent a crucial tool utilized by those interested in matrisome function, the easiest of which are implemented with immortalized/cancer cell lines. These cell lines often form the foundations of a research proposal, or serve as vehicles of validation for other model systems. For these reasons, it is important to understand the complement of matrisome genes that are expressed when identifying appropriate cell culture models for hypothesis testing. To this end, we harvested bulk RNA sequencing data from the Cancer Cell Line Encyclopedia (CCLE) to assess matrisome gene expression in 1019 human cell lines. Our examination reveals that a large proportion of the matrisome is poorly represented in human cancer cell lines, with approximately 10% not expressed above threshold in any of the cell lines assayed. Conversely, we identify clusters of essential/common matrisome genes that are abundantly expressed in cell lines. To validate these observations against tissue data, we compared our findings with bulk RNA sequencing data from the Genotype-Tissue Expression (GTEx) portal and The Cancer Genome Atlas (TCGA) program. This comparison demonstrates general agreement between the "essential/common" and "dark/uncommon" matrisome across the three datasets, albeit with discordance observed in 59 matrisome genes between cell lines and tissues. Notably, all of the discordant genes are essential/common in tissues yet minimally expressed in cell lines, underscoring critical considerations for matrix biology researchers employing immortalized cell lines for their investigations.

Introduction

The extracellular matrix (ECM), a key feature of metazoan biology, comprises the network of secreted macromolecules providing structural and biochemical support in the extracellular space of organismal tissues. The ability of cells to influence the ECM and, in turn, be influenced by the ECM encapsulates the idea of "dynamic reciprocity" coined by Bissell et al. in 1982 [1]. Dysregulation of the ECM's crucial structural and regulatory properties underlie diverse pathologies, granting substantial clinical significance to the ECM's many components. Considering the importance of ECM in biological processes, great effort has been made to understand the ECM proteome, dubbed the "matrisome" by Naba and colleagues in 2012 [2].

In vitro cell culture models often form the basis of biomedical research, including studies of matrix biology. Rightly so, as the ease and relative simplicity of in vitro culture systems maintain accessibility in biomedical research, in large part due to the low financial burden and limited availability of tissue/animal models for human disease. However, it is well appreciated that in vitro models of biological systems, particularly those propagated in 2-D, do not sufficiently capture the

Abbreviations: CCLE, Cancer Cell Line Encyclopedia; DBSCAN, Density-Based Spatial Clustering of Applications with Noise; ECM, Extracellular Matrix; GTEx, Genotype-Tissue Expression (project); M, Median; MAD, Median Absolute Deviation; SCLC, Small Cell Lung Carcinoma; TCGA, The Cancer Genome Atlas (program); UMAP, Uniform Manifold Approximation and Projection.

* Corresponding author.
E-mail address: david.peeney@nih.gov (D. Peeney).
[1] Authors contributed equally.

https://doi.org/10.1016/j.mbplus.2024.100156
Received 5 April 2024; Received in revised form 18 June 2024; Accepted 24 June 2024
Available online 27 June 2024

complexity of tissue structure and function. Despite this, in vitro culture will continue to be an important component of a matrix biologists' repertoire, and for this reason it is important to understand which elements of the matrisome are expressed across common cell line models. To this end, we harvested bulk RNA sequencing data from the Cancer Cell Line Encyclopedia (CCLE) to assess the variability in matrisome gene expression across a selection of over 1000 common human cell lines. Somewhat surprisingly, we found that approximately 50 % (526) of the matrisome is minimally expressed across all cell lines, which we refer to as the dark/uncommon matrisome. A deeper look into the dark matrisome reveals that 113 matrisome genes are not expressed above threshold in any of the analyzed cell lines. On the contrary, our analysis also reveals a set of 192 matrisome genes that are consistently expressed across the majority of cell lines (above threshold in >75 % of cell lines), which we refer to as the essential/common matrisome. Despite this commonality between analyzed cell lines, the matrisome expression profile can be a defining trait, exemplified through a unique profile in small cell lung carcinoma cell lines. Comparison of these findings with tissue RNA sequencing data from the Genotype-Tissue Expression (GTEx) project and The Cancer Genome Atlas (TCGA) reveals a general agreement between datasets despite discordant expression in 59 genes between cell lines and tissues. The data described herein, summarized in Table S1, highlight important considerations for investigators interested in ECM research, particularly those that are concerned with broader matrisome behavior using cell line models.

Results

Matrisome gene expression in the Cancer Cell Line Encyclopedia

The Cancer Cell Line Encyclopedia (CCLE) is a collection of biological indices from 1019 common human cancer cell lines across many tissues and neoplasms. Fig. 1A illustrates the tissues of origin featured in this dataset, highlighting a significant presence of cell lines originating from lung and hematopoietic/lymphoid tissue (188 and 173 cell lines, respectively). Several tissues have poor representation (less than 10 cell lines), including the pleura, prostate, biliary tract, cervix, salivary gland, and small intestine. Furthermore, multiple major cell types/tissues are unrepresented in this dataset, including skeletal/cardiac muscle, testis, adipose, and vascular tissue. Assessment of the number of matrisome

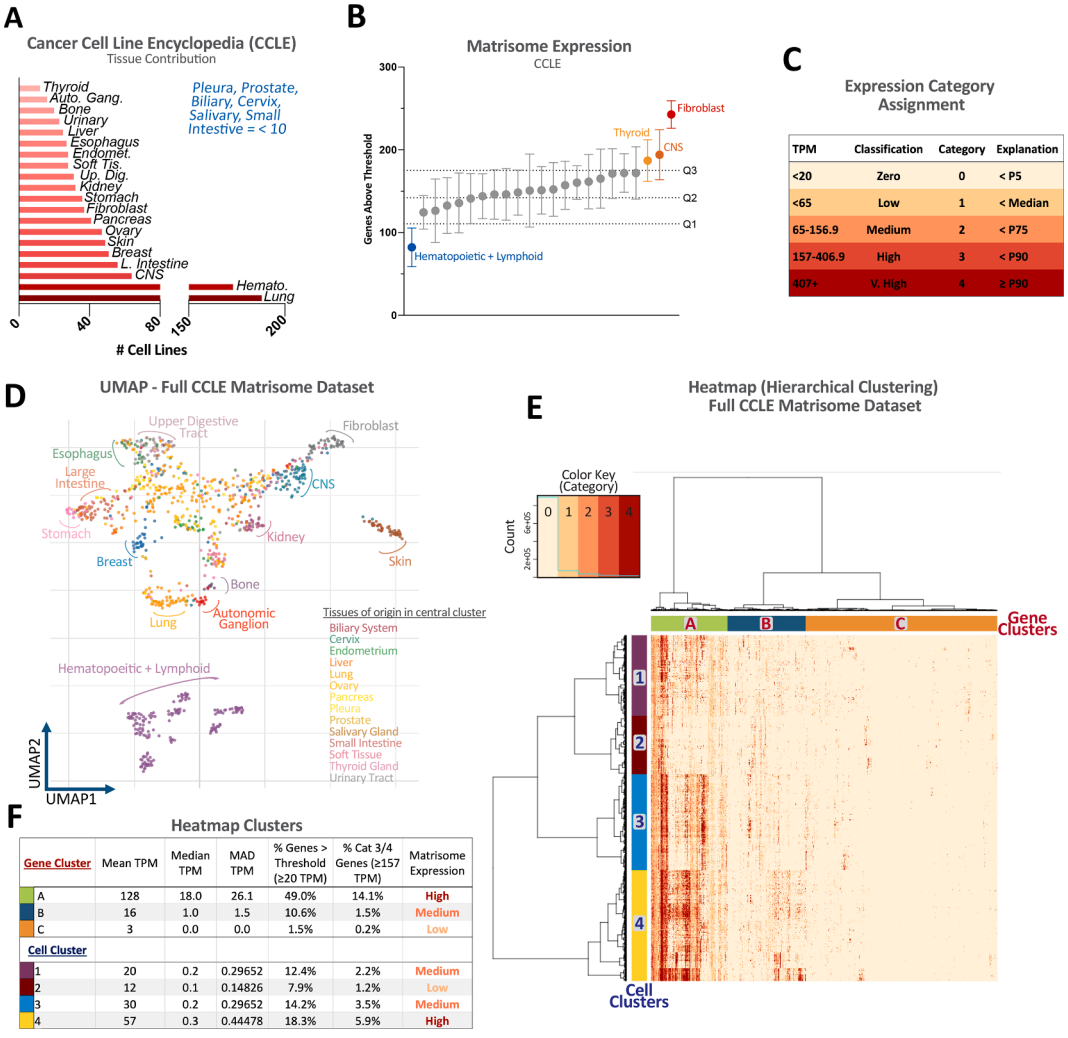

Fig. 1. Matrisome expression across all cell lines of the Cancer Cell Line Encyclopedia (CCLE). (A) Tissue contributions to the CCLE dataset. (B) Average number of matrisome genes expressed above threshold (≥20 TPM) across CCLE cell line tissues (only tissues with ≥8 cell lines), with labeled tissues residing outside of the interquartile range. Error bars = standard deviation. (C) Table describing the five expression categories, based on the distribution of data. P = Percentile. (D) UMAP of matrisome expression across the full CCLE dataset. (E) Heatmap of log2 expression data visualizing the 3 major gene clusters and 4 major cell clusters. (F) Basic statistics of 3 major matrisome gene clusters and 4 major cell line clusters from the CCLE dataset. *Abbreviations: (A + B) Auto. Gang = Autonomic Ganglion, CNS = Central Nervous System, Endomet. = Endometrium, Hemato. = Hematopoietic and Lymphoid Tissue, L. Intestine = Large Intestine, Soft Tis. = Soft Tissue, Up. Dig. = Upper Digestive Tract. (E) TPM = Transcripts per million, MAD = Median absolute deviation, Cat = Category.*

genes expressed in the cell lines of each tissue reveals that fibroblast cell lines are predominant sources of matrisome gene expression, and most cell lines expressing between 10 and 20 % of the matrisome (Fig. 1B). Central nervous system (CNS) and thyroid cell lines are also high matrisome expressing cell types, whereas those of a hematopoietic/lymphoid lineage exhibit low matrisome expression (Fig. 1B). For simplified visualization, we assigned matrisome expression values into 5 categories (Fig. 1C). Category 0 is defined as zero expression, which represents approximately the 5th percentile of matrisome gene expression across the whole CCLE dataset. Ascending categories are defined by the median, 75th, and 90th percentiles of the whole dataset (Fig. 1C). Uniform Manifold Approximation and Projection (UMAP) analysis of the full CCLE matrisome dataset reveals modest visual clustering of data dependent on the tissue of origin, with hematopoietic/lymphoid and skin cell lines forming distinct groups (Fig. 1D). Combining UMAP with DBSCAN (Density-Based Spatial Clustering of Applications with Noise) or k-means clustering further signifies the uniqueness of hematopoietic/lymphoid cell lines, whereas skin cell lines form a unique cluster with DBSCAN, but not with k-means clustering (Fig. S1). Plotting the expression data through a heatmap with hierarchical clustering, colored by the assigned category, reveals 3 major matrisome clusters and 4 prominent cell type clusters that can each be defined by high, medium, and low matrisome gene expression (Fig. 1E and F). In gene cluster A, defined as matrisome high, 49 % of the values across all cell lines are above threshold. Furthermore, 14.1 % of these values are category 3/4 (defined as high or very high expression), Fig. 1F. Cell line and gene cluster IDs are defined in full in Table S1, Tab 1/2.

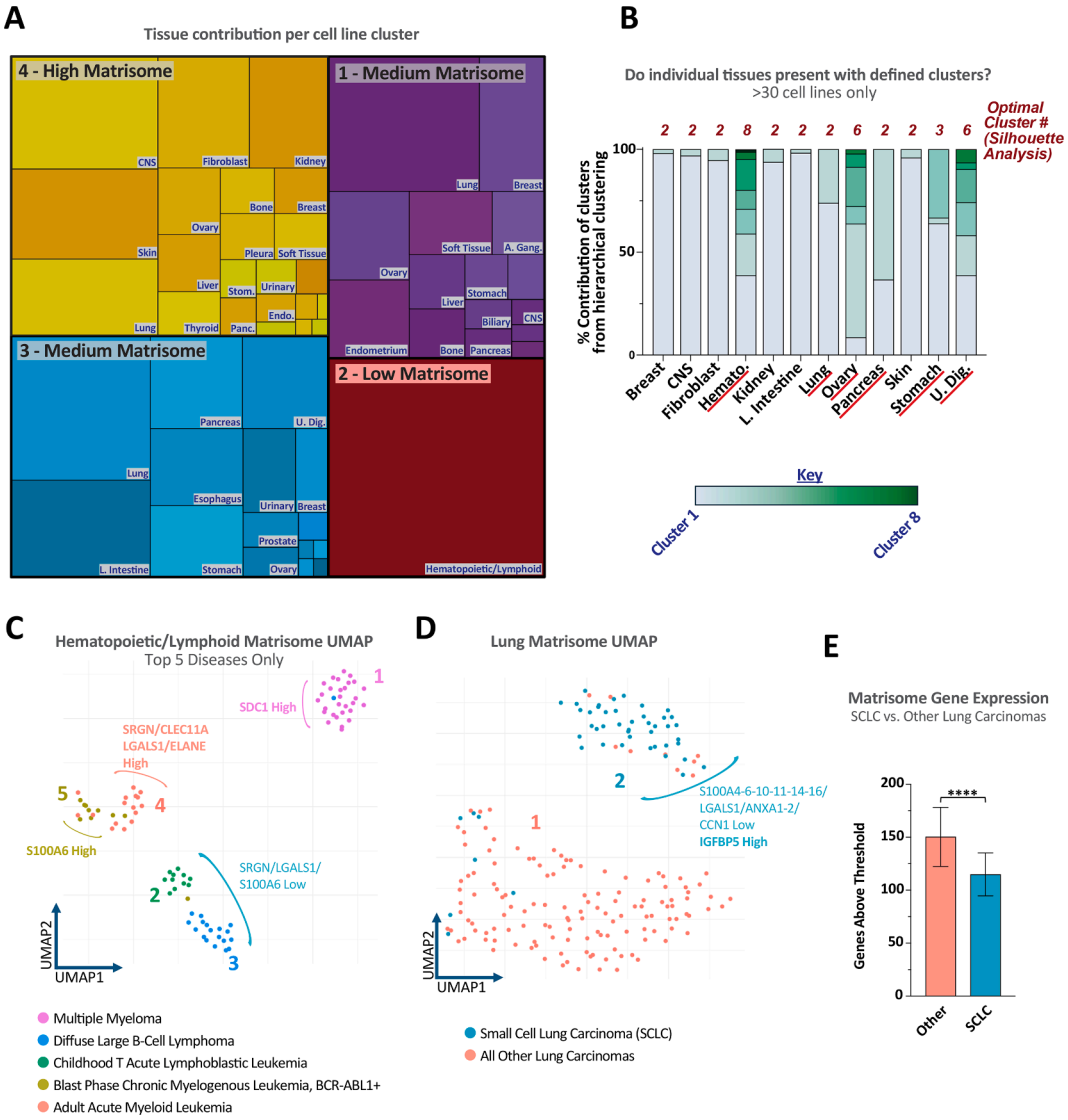

Fig. 2. Exploration of the major cell line clusters from the CCLE matrisome expression dataset. (A) Treemap depicting the tissue contributions to each cell line cluster. (B) Silhouette analysis of log2 transformed expression data reveals the optimal cluster number for each tissue subset of the CCLE data. The percentage contribution of these clusters to the subsetted tissue data is plotted on the bar graph. Tissues with defined cluster groups (<90 % within a single cluster) are underlined in red. (C) UMAP analysis reveals unique clusters in hematopoietic/lymphoid cell lines that group via disease model (top 5 disease models only). Cluster numbers are labeled. (D) UMAP analysis reveals unique clusters in lung cell lines, with unique clustering for small cell lung carcinomas (SCLC). Cluster numbers are labeled. (E) Comparison of lung cell line matrisome gene expression between cluster 1 (all other lung carcinomas) and cluster 2 (SCLC). Error bars = standard deviation, **** = $p < 0.0001$. *Abbreviations: (A + B) A. Gang = Autonomic Ganglion, CNS = Central Nervous System, Endo. = Endometrium, Stom. = Stomach, L. Intestine = Large Intestine, Panc. = Pancreas, U. Dig. = Upper Digestive Tract.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

CCLE cell lines cluster into high, medium, and low matrisome gene expression

Analysis of the four major cell line clusters derived from hierarchical clustering analysis reveals that hematopoietic/lymphoid derived cell lines are the major constituents of cluster 2 (low matrisome expression), with only one additional bone tissue-derived cell line joining this group (SK-PN-DW; primitive neuroectodermal tumor). The remaining three clusters are made up of largely adherent cell lines derived from solid tumors. In some instances, tissues of origin are well represented across the three clusters (lung, breast, ovary). In other cases, tissues of origin display clear cluster preference such as CNS/skin/fibroblast/kidney (cluster 4; high matrisome), and gastrointestinal tissues (esophagus/upper digestive tract/large intestine) within cluster three (medium matrisome), Fig. 2A. Following this, we sought to determine whether any of these tissues of origin formed distinct groups based on their neoplastic subtype, utilizing the existing metadata from the CCLE dataset. To streamline the process, we performed silhouette analysis to determine the optimal cluster number for each dataset and combined this with hierarchical clustering (Euclidean distance with Ward linkage) to assign data points to the optimal cluster numbers per tissue (Fig. 2B). This analysis reveals that half of the analyzed tissues of origin (those with >30 cell lines for practical cluster analysis) are unlikely to produce well-defined clusters, since >90 % of the data points (cell lines) reside in an individual cluster. The remaining half are good candidates for further study; hematopoietic/lymphoid, lung, ovary, pancreas, stomach, and upper digestive tract. Each of these tissue subsets were subjected to UMAP analysis with hierarchical clustering or DBSCAN for the visual interpretation of disease-based clustering within the data. Despite displaying a ubiquitously low matrisome gene expression profile, hematopoietic and lymphoid tissue did display a disease specific hierarchical clustering (showing only the top 5 represented disease subtypes), Fig. 2C. DBSCAN of the same data revealed 3 clusters, visualized by UMAP in Fig. S2A. Differential expression analysis between these clusters amplifies the low matrisome gene expression in hematopoietic/lymphoid cell lines, revealing few matrisome genes that exhibit significant differences in expression (Fig. 2C, Table S1-Tab 3). Lung-derived cell lines display a broader profile of matrisome expression, and UMAP analysis combined with hierarchical clustering or DBSCAN reveals a distinct clustering between small cell lung carcinoma (SCLC) cell lines and other lung carcinoma cell lines (Figs. 2D, S2B). Differential expression between these two clusters reveals a largely reduced expression of matrisome genes in SCLC cell lines, with 32 down-regulated and 2 up-regulated matrisome genes (Fig. 2D, E, Table S1-Tab 4). The remaining four tissues of origin (ovary, pancreas, stomach, upper digestive tract) did not exhibit visually clear clusters in UMAP, and derived only 1 cluster with DBSCAN (Fig. S2C–F).

Analysis of CCLE and Genotype-Tissue Expression (GTEx) matrisome expression reveals an essential and dark matrisome

To validate whether the observed patterns of matrisome expression in cancer cell lines constitutes a cell line-specific phenomenon, or is a consistent feature of human cell biology, we compared CCLE matrisome transcriptome data to that of the Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) projects. The (GTEx) dataset is a collection of postmortem tissue from non-diseased tissues derived from almost 1000 donors [3], and TCGA contains transcriptome data from over 10,000 tumors. To assess whether these datasets are compatible for analysis, we harvested lung only data from each and compared matrisome expression and distribution to illustrate that tissues (GTEx and TCGA) generally express a higher proportion of matrisome genes than cell lines (Fig. 3A/3B). Hierarchical clustering of each data source produces 4 major clusters, that we refer to as essential (high), common, uncommon, and dark cluster categories (Fig. 3C). Comparison of the essential/common and dark/uncommon clusters between the datasets

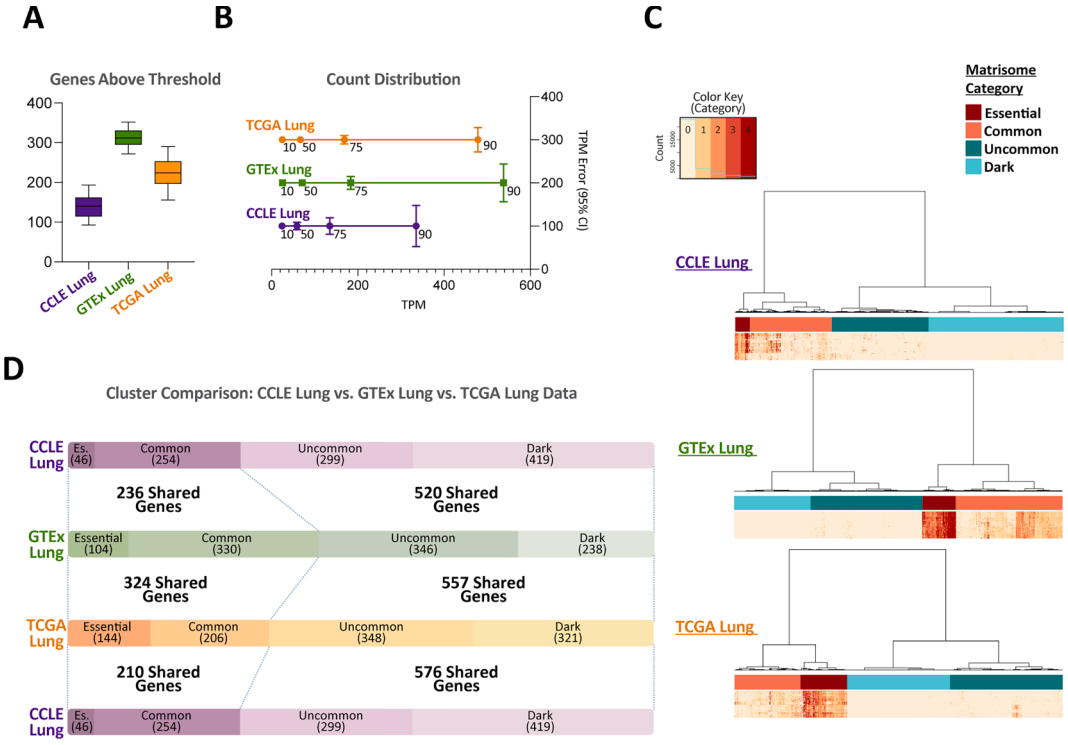

Fig. 3. Comparison of lung/lung-origin matrisome gene expression profiles in CCLE, GTEx, and TCGA datasets. (A) Comparison of the number of genes above threshold (20 TPM) and (B) distribution between each dataset. (C) Heatmaps of log2 expression data depicting matrisome gene expression independently for each dataset. Heatmaps are compressed for conciseness and pattern illustrations purposes. (D) Comparison of the shared genes between the essential/common matrisome and dark/uncommon matrisome, determined through hierarchical clustering.

reveals a strong agreement in the defined matrisome categories (Fig. 3D), suggesting that a subsequent comparison between each data source will be fair.

For simplicity in comparison with the CCLE dataset, and to acquire individual values for matrisome expression in each tissue, we calculated the median (M) and median absolute deviation (MAD) for each GTEx and TCGA tissue. The two values were summarized to account for variability in gene expression while favoring above median expression values (M + MAD). Hierarchical clustering reveals four clear gene groups separated by matrisome expression levels, which we define as essential, common, uncommon, and dark clusters (Fig. 4A/B). The constituents of these categories are defined in Table S1-Tab 5/6. Reduction of the TCGA data two 2-dimensions using UMAP reveals visually prominent clustering of data points based on tissue, unlike that of the CCLE data (Fig. 4C). Tissues contain higher above threshold counts for matrisome genes than the CCLE data, which can be attributed to higher sample complexity (many cell types per tissue) and count metrics that favor higher expression values in tissue data (M + MAD)

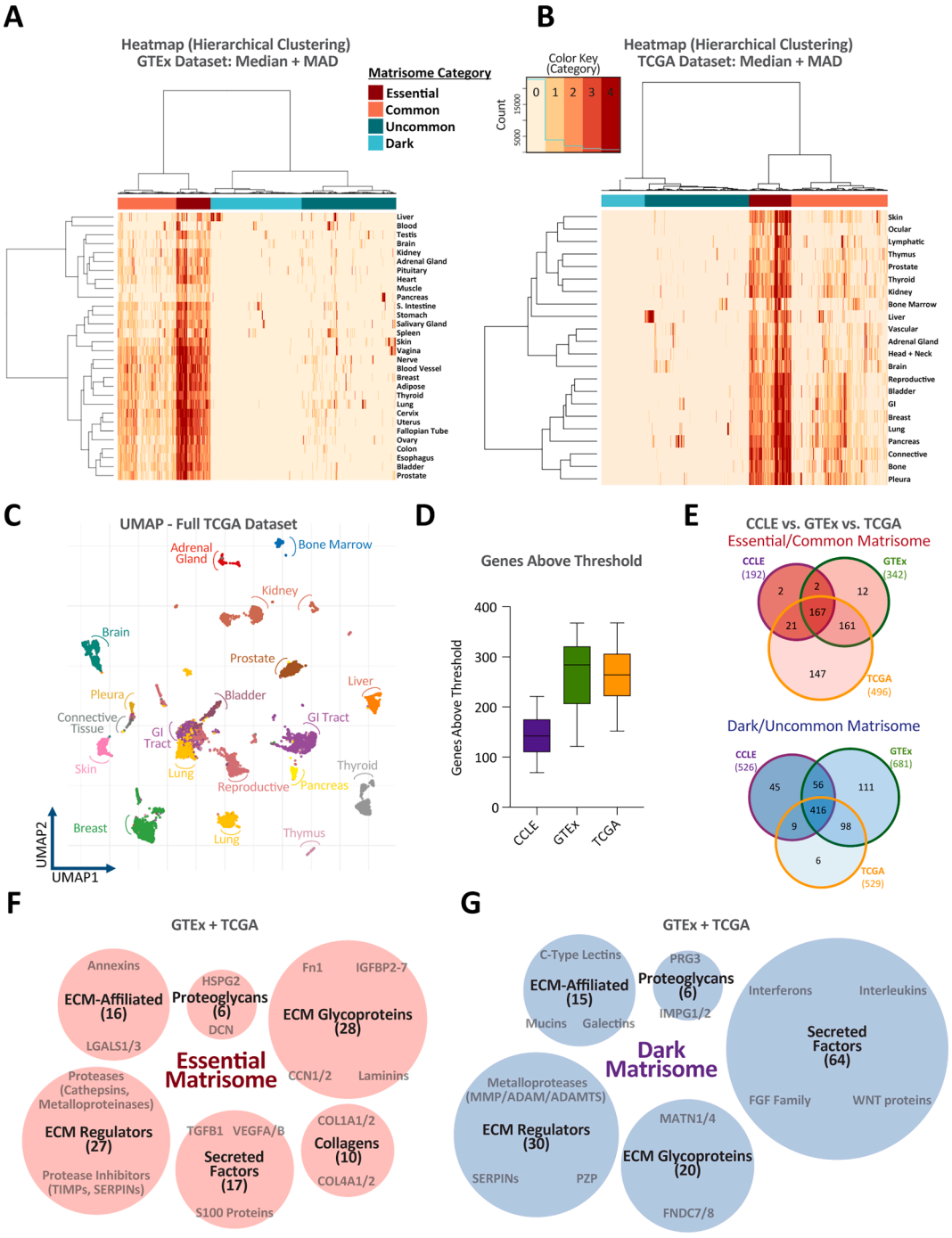

Fig. 4. The essential and dark matrisome of the Cancer Cell Line Encyclopedia (CCLE), Genotype-Tissue Expression (GTEx) Portal, and The Cancer Genome Atlas (TCGA). (A/B) Heatmap of log2 expression data depicting matrisome gene expression from 30 tissues defined in the GTEx portal and 22 tissues from TCGA, with colors depicting expression using the same categories as defined in Fig. 1C. (C) UMAP analysis of the full TCGA data, colored by defined tissue. (D) Comparison of the number of genes above threshold (20 TPM) for each dataset. (E) Venn diagrams illustrating that the essential/common matrisome and the dark/uncommon matrisome between the CCLE, GTEx, and TCGA are comparable. (F/G) Bubble maps defining the essential and dark matrisome shared across the CCLE, GTEx, and TCGA datasets, with notable example genes/gene families included. *Abbreviations: GI = Gastrointestinal Tract.*

(Fig. 4D). To apply the same matrisome definitions to the CCLE data, we harvested genes that constitute CCLE gene cluster A (Matrisome High; Fig. 1E/F) and excluded 32 genes that displayed zero counts in over 75 % of the analyzed cell lines. Genes with above threshold expression in 75 % of cell lines were classified as essential, and the remaining genes categorized as common (Table S1-Tab 7). The dark matrisome was classified as any gene with 99 % of counts below threshold (20 TPM), and uncommon defined as genes with 95 % of values below threshold (Table S1-Tab 8). A total of 113 gene members of the dark matrisome do not exhibit above threshold expression in any of the 1019 cell lines in the CCLE. Comparison of the essential/common and dark/uncommon matrisome profiles reveals significant cross-over between all datasets with 167 and 416 shared matrisome genes, respectively (Fig. 4E, Table S1-Tab 9/10). The combined essential and dark matrisome is summarized in Fig. 4F–G, with a selection of notable genes highlighted. Annexins are prominent constituents of the essential matrisome, with Annexin A7 (ANXA7) being expressed above threshold in all cell lines of the CCLE. The dark matrisome consists of many factors that are considered pro-inflammatory, including multiple members of the interferon family that are not expressed above threshold in any of the cell lines analyzed. Pathway analysis reiterates the inflammatory nature of the dark matrisome, with the top pathways associated with responses to pathogens and disease, and upstream regulators that include NFκB and interleukins (Table 1, Table S1-Tabs 11/12). Parallel analyses with the essential/common matrisome reveal pathways that are involved in extracellular matrix organization/turnover, insulin-like growth factor transport, and fibrosis (Table 1, Table S1-Tabs 13/14). Between the combinations of data sources, there is a level of discordance in gene classification (at least two levels between classes, i.e. dark matrisome versus common matrisome). GTEx and TCGA data displays discordance in 29 genes, all of which exhibit enriched expression in tumor tissues of TCGA. Between the CCLE and tissue (GTEx and TCGA) datasets there are 59 discordant genes. In each case, inflammatory mediators (chemokines, cytokines, complement components) and collagens represent predominant constitutents, summarized in Table 2 and Table S1-Tab 15.

Discussion

Classification and characterization of the matrisome is maintained through the MatrisomeDB, a powerful proteomic atlas for ECM researchers [4]. The proteomic approaches that form the basis of this platform have revealed that over 100 matrisome proteins can be detected within specific tissues, healthy and diseased [2,5]. Despite being the gold standard in ECM research, proteomics approaches are hindered by matrisome complexity and insolubility, compounding the well-established nuances in proteomics such as missing values [6]. RNA sequencing, applied here, is less troubled by analogous nuances, making it an ideal complimentary tool for exploratory analyses into the matrisome. Our work describes tissue- and disease-specific clusters of matrisome expression patterning, revealing a highly expressed essential matrisome and a minimally expressed dark matrisome, both of which are largely conserved between bulk tissue samples and individual cell lines.

Our data supports the idea that much of the dark matrisome may represent an inducible matrisome, genes that require unique biological stimuli such as damage-associated molecular patterns (DAMPs), pathogen-associated molecular patterns (PAMP), or reactive oxygen species. This hypothesis is supported by pathway analysis and upstream regulator analysis of the dark matrisome, the latter of which highlighting toll-like receptor (TLR) function, cytokine signaling, and inflammatory transcription factor activity. This largely inducible nature implies the matrisome is a highly responsive cellular compartment that is readily rewired in response to molecular signals, further evidence that the extracellular compartment is more than just a scaffold that anchors cells and tissues in place. Noteworthy members of the dark matrisome (across tissues and cell lines) include many cytokines of the interferon and interleukin families, numerous growth factors (fibroblast growth factors, TGFß superfamily), twelve members of the metalloproteinase family and five C-type lectins. Interrogation of their coverage in the MatrisomeDB reveals proteomic detection across a range of tissues [7,8].

In interpreting these results, it is crucial to recognize fundamental differences between the utilized CCLE, GTEx, and TCGA datasets. It is unsurprising that tissue displays higher numbers of above-threshold matrisome expression, considering the many cell types present within an individual tissue sample. In contrast, the CCLE is comprised of immortalized cell lines that display one or many hallmarks of cancer. This adds a strong confounding factor to these analyses and may contribute to some of the unique matrisome expression patterns (i.e., essential or dark matrisome). Furthermore, certain tissue types are overrepresented in the CCLE such as lung and hematopoietic/lymphoid tissue, whereas others are poorly covered. In the same manner, specialized tissues such as dental tissue, ocular tissue, and gametes are unrepresented across all datasets, Such limitations, however, do not negate the potential utility of our described cell line clusters, nor do they invalidate the clear presence of an extensive dark matrisome in both immortalized cell lines and tissue. Similar pitfalls in representation render the matrisome definition itself imperfect or at least very nuanced. The matrisome includes proteins such as MFAP1, a molecule herein described as a component of the essential matrisome that also displays prominent nuclear localization, which may represent its dominant isoform. Similar reasoning applies to the cathepsins, which are classically lysosomal proteases. These occasional subtleties do not compromise the utility of the matrisome as concept; rather, they invite further investigation into

Table 1
A summary of pathway analysis and upstream regulator analysis performed using Ingenuity Pathway Analysis on the essential/common matrisome and the dark matrisome gene lists.

	Essential/Common Matrisome	Dark Matrisome
Top Unique Pathways	ECM organization	Pathogen induced cytokine storm
	Collagen biosynthesis and modifying enzymes	Airway pathology in COPD
	Hepatic fibrosis / hepatic stellate cell activation	Cytokine communication between immune cells
	Collagen degradation	Osteoblasts in rheumatoid arthritis
	Regulation of IGF transport and uptake by IGFBPs	Wound Healing Signaling Pathway
Top Upstream Regulators	AGT	NFkB
	TGFB1	MYF6
	CCR2	IL17A
	AHR	IL22
	TNF	EBF3

Table 2
Comparison of Essential/Common and Dark/Uncommon matrisome genes between CCLE, GTEx, and TCGA datasets reveals some discordance between healthy versus cancerous (GTEx vs. TCGA), and cell line versus tissue (CCLE vs. GTEx + TCGA).

Discordance			
GTEx vs. **TCGA** (29)	Enriched in?	**CCLE** vs. **Tissue** (59)	Enriched in?
CCL Chemokines (5)	TCGA	CCL Chemokines (5)	Tissue
CXCL Chemokines (2)	TCGA	Collagens (4)	Tissue
Metalloproteinases (2)	TCGA	Complement C1Q Family (4)	Tissue
COL10A1/COL11A1	TCGA	TNFSF proteins (3)	Tissue
IL1B/IL24	TCGA	SERPIN proteins (3)	Tissue
SDC1	TCGA	CXCL Chemokines (2)	Tissue
S100A1	TCGA	S100 proteins (2)	Tissue

the complexities of the ECM and its individual components.

This possibility of using matrisome transcriptomics to resolve and even define different populations of cells is supported by previous single cell RNA sequencing findings indicating that core matrisome expression patterns, or matreotypes, can be a defining feature of cell types and even cell states [9]. Others have performed a pan-cancer analysis of the matrisome using TCGA, describing that matrisome expression can convey prognostic significance [10]. Furthermore, this study described a series of matrisome-controlling transcriptional master regulators that likely control major aspects of the tumor-associated matrix. Similarly, a mass spectrometry-based atlas of ECM protein expression in mouse organs uncovered tissue-specific ECM expression signatures; 75 % of the top 100 detected ECM proteins are described as essential/common in our tissue transcriptomic analysis (data supplied directly from Dr. Kirk Hansen) [11]. Our results from the CCLE indicate far less defined clustering by cell type versus tissue (with the exception of hematopoietic and lymphoid cells), suggesting that the specificity of matrisome expression signatures may be degraded in 2-D culture conditions. Still, the tissue and disease process clusters we uncovered support the potential utility of transcriptional data in understanding tissue- and disease-characteristic differences in matrisome expression. At a transcriptional level, the matrisome masquerades as a reasonably stable compartment between cell types and tissues, somewhat contrary to proteomic data. Mass changes in the landscape and constitution of the extracellular matrix are deeply intwined in disease pathology, and therefore further characterization of the pathology/aging-related changes in matrisome transcriptional profile may unveil new perspectives and research avenues.

Our analysis provides important considerations for researchers looking to probe molecular pathways that are heavily influenced by matrisome composition. The highly limited expression of matrisome members needs to be deliberated when using simple cell line models in molecular screens or during investigations into matrisome protein function. Treatment of in vitro conditions to tease out dark matrisome expression concurrent with the context/disease of interest will be a critical step in attaining the most biologically relevant model systems for experiments probing ECM function.

Methods

RNA sequencing data resources and preprocessing

Cancer Cell Line Encyclopedia (CCLE) preprocessed expression values (TPM) and metadata [12] were downloaded from the European Bioinformatics Institute web portal (www.ebi.ac.uk/gxa/home). The data was filtered for matrisome genes [4] and untransformed or log2 transformed data were used for downstream analysis. Genotype-Tissue Expression (GTEx) data were downloaded directly as processed TPM counts through the GTEx portal (www.gtexportal.org) [3], then filtered for matrisome genes. The Cancer Genome Atlas (TCGA) program RNA-seq log2 transformed TPM (Transcripts Per Kilobase Million) values were retrieved from UCSC Xena Xenabrowser.net (dataset ID: TcgaTargetGtex_rsem_gene_tpm). Associated clinical metadata were downloaded from Genomic Data Commons data portal (https://portal.gdc.cancer.gov/). After removing one retired sample, and linking case submitter identifications with valid records, there were 10,534 samples with both TPM expression values and tissue/organ of origin information. For each tissue, we calculated the median TPM and the median absolute deviation. We then summarized these values to assign an individual count per gene in each tissue, allowing us to account for variability in gene expression while favoring above median expression values. Untransformed or log2 transformed data were used for downstream analysis.

Software and analyses

Most statistical analyses were performed using R Studio (Build 351), with code used supplied in Supplementary File 1. For heatmap cluster analysis, we utilized Euclidean distance with a Ward linkage metric due to its ease in cluster interpretation. Optimal clusters were visually determined (through hierarchical clustering or Uniform Manifold Approximation and Projection; UMAP) or through silhouette analysis and elbow plots z. Additional cluster analysis was performed using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or k-means clustering. Pathway Analysis and Upstream Regulator Analysis were performed using Ingenuity Pathway Analysis (Qiagen).

CRediT authorship contribution statement

Joshua A. Rich: Writing – review & editing, Writing – original draft, Conceptualization. **Yu Fan:** Writing – review & editing, Formal analysis, Data curation. **Qingrong Chen:** Writing – review & editing, Formal analysis, Data curation. **Daoud Meerzaman:** Supervision, Resources. **William G. Stetler-Stevenson:** Writing – review & editing, Funding acquisition. **David Peeney:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code is included as a Supplementary File.

Acknowledgement

This research was supported by the Intramural Research Program of the NIH (W.G.S.S. Project ID ZIA SC 009179).

Contributions

DP and JR conceived the study. DP, YF, JR performed data harvest and analysis. All authors assisted in interpretation of data and technical support. The manuscript was written by DP and JR. All authors reviewed the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mbplus.2024.100156.

References

[1] M.J. Bissell, H.G. Hall, G. Parry, How does the extracellular matrix direct gene expression? J. Theor. Biol. 99 (1) (1982) 31–68.
[2] A. Naba, et al., The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices, Mol. Cell. Proteomics 11 (4) (2012). M111 014647.
[3] G.T. Consortium, The Genotype-Tissue Expression (GTEx) project, Nat. Genet. 45 (6) (2013) 580–585.
[4] X. Shao, et al., MatrisomeDB 2.0: 2023 updates to the ECM-protein knowledge database, Nucl. Acids Res. 51(D1) (2023) D1519–D1530.
[5] A. Naba, et al., Characterization of the extracellular matrix of normal and diseased tissues using proteomics, J. Proteome Res. 16 (8) (2017) 3083–3091.
[6] R.C. Poulos, et al., Strategies to enable large-scale proteomics for reproducible research, Nat. Commun. 11 (1) (2020) 3793.

[7] J. Barallobre-Barreiro, et al., Cartilage-like composition of keloid scar extracellular matrix suggests fibroblast mis-differentiation in disease, Matrix Biol. Plus 4 (2019) 100016.
[8] A.M. Moreira, et al., Proteomic identification of a gastric tumor ECM signature associated with cancer progression, Front. Mol. Biosci. 9 (2022) 818552.
[9] 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.
[10] V. Izzi, et al., Pan-Cancer analysis of the expression and regulation of matrisome genes across 32 tumor types, Matrix Biol. Plus 1 (2019) 100004.
[11] M.C. McCabe, A.J. Saviola, K.C. Hansen, Mass spectrometry-based atlas of extracellular matrix proteins across 25 mouse organs, J. Proteome Res. 22 (3) (2023) 790–801.
[12] M. Ghandi, et al., Next-generation characterization of the Cancer Cell Line Encyclopedia, Nature 569 (7757) (2019) 503–508.