Extracellular matrix gene expression signatures as cell type and cell state identifiers

Fabio Sacher[1], Christian Feregrino[1], and Patrick Tschopp[1*], Collin Y. Ewald[2*]

1 Universität Basel, DUW Zoology, Basel CH-4051, Switzerland

2 Laboratory of Extracellular Matrix Regeneration, Institute of Translational Medicine, Department of Health Sciences and Technology, ETH Zürich, Schwerzenbach CH-8603, Switzerland

*Corresponding authors: collin-ewald@ethz.ch (CYE) and patrick.tschopp@unibas.ch (PT)

Keywords: matrisome, matreotype, cell type, scRNA-seq, limb development

Highlights:

- Cell types produce unique extracellular matrix compositions
- Dynamic extracellular matrix gene expression profiles hold predictive power for cell type and cell state identification

Abstract

Transcriptomic signatures based on cellular mRNA expression profiles can be used to categorize cell types and states. Yet whether different functional groups of genes perform better or worse in this process remains largely unexplored. Here we test the core matrisome - that is, all genes coding for structural proteins of the extracellular matrix - for its ability to delineate distinct cell types in embryonic single-cell RNA-sequencing (scRNA-seq) data. We show that even though expressed core matrisome genes correspond to less than 2% of an entire cellular transcriptome, their RNA expression levels suffice to recapitulate important aspects of cell type-specific clustering. Notably, using scRNA-seq data from the embryonic limb, we demonstrate that core matrisome gene expression outperforms random gene subsets of similar sizes and can match and exceed the predictive power of transcription factors. While transcription factor signatures generally perform better in predicting cell types at early stages of chicken and mouse limb development, *i.e.,* when cells are less differentiated, the information content of the core matrisome signature increases in more differentiated cells. Our findings suggest that each cell type produces its own unique extracellular matrix, or matreotype, which becomes progressively more refined and cell type-specific as embryonic tissues mature.

Introduction

How to define and identify different cell types remains a fundamental challenge in biology [1–4]. Cell types have traditionally been classified based on their morphology and function, by the tissues from where they were isolated, their ontogenetic origin, or their molecular signatures [3]. In recent years, gene expression data from single-cell transcriptomic studies (scRNA-seq) have been used to characterize and fine-tune different cell type classification systems [2,3,5].

Cellular fate and cell-type-specific gene expression programs are thought to be largely regulated by transcription factors and their corresponding *cis*-regulatory networks [2,4,6]. Accordingly, transcription factor expression profiles can be useful in identifying cell types from scRNA-seq data [2,7,8]. Yet other cellular properties can also vary dynamically, in a cell type-specific manner. Hence, we were looking for additional sets of putative 'biomarker' genes, to identify cell types and states.

The extracellular matrix (ECM) has traditionally been thought of as a static protein network surrounding cells and tissues. However, the ECM has recently emerged as a highly dynamic system [9–11]. In fact, transcription and translation of some ECM genes are even coupled to circadian rhythm, highlighting the dynamic nature of ECM composition [12]. Experimentally, ECM composition has so far been determined mostly by proteomics assays [13]. More recently, *in-silico* approaches have defined the 'matrisome' gene sets representing all genes either forming or remodeling the ECM, as present in a given species' genome [13,14]. The matrisome is divided into two main categories: the core matrisome

encompassing all proteins that form the actual ECM (collagens, glycoproteins, proteoglycans) and the matrisome-associated proteins that either bind to the ECM, remodel the ECM, or are secreted from the ECM [13,14].

Importantly, it has been postulated that each cell type produces its own unique ECM [14–17]. To capture this concept, we have recently defined the 'matreotype', an extracellular matrix signature associated with - or caused by - a given cellular identity or physiological status [17]. For instance, cellular status, including metabolic, healthy or pathologic, or aging have been associated with distinct ECM expression patterns (*i.e.,* matreotypes) [14,17–21]. Furthermore, cancer-specific cell types can be identified based on their unique ECM composition [13,14,20,22]. This indicates that ECM composition is plastic and adapts to cellular needs or status. Since this is a highly dynamic process, snapshots of unique ECM compositions are reflected in distinct matreotypes.

Based on this, we hypothesized that ECM gene expression is a dynamic parameter that could hold predictive value to function as a biomarker for cell type and state identification. To test our hypothesis, we re-analyzed publicly available scRNA-seq data and specifically examine ECM gene expression signatures. Unsupervised clustering of scRNA-seq data using the whole transcriptome - or highly variable genes therein - is a common strategy to classify cell types [2,3,5]. Here we use defined transcriptome subsets - namely, expressed transcription factors, core matrisome genes, and random transcriptome subsets of equal size - to re-cluster scRNA-seq data and evaluate the resulting clusters in comparison to the performance of the entire transcriptome. In embryonic data coming from chicken and

66 mouse limbs, we find that the core matrisome has less predictive power in undifferentiated
67 cells, early during development, but outcompetes transcription factors later in development
68 and in more differentiated cell types. Consequently, we propose matreotype gene
69 expression signatures as context-dependent proxies for identifying cell types.
70

Results

Defining the chicken core matrisome

The matrisome has been defined for humans (1027 genes), mice (1110 genes), zebrafish (1002 genes), planarian (256 genes), *Drosophila* (641 genes), and *C. elegans* (719 genes), where it corresponds to roughly 4% of their protein-coding genes [14,23–26]. In order to expand the number of model organisms amenable to 'matreotype' investigation, we first decided to define the chicken matrisome. Using the 1110 mouse and 1027 human matrisome gene lists to perform orthology and InterPro domain searches, we identified 631 and 656 chicken matrisome genes, respectively (Supplementary Fig. 1, Supplementary Table 1). In summary, we define the chicken matrisome with 217 core-matrisome genes and 443 associated-matrisome genes (Supplementary Table 1).

The chicken core matrisome as a molecular signature with cell-type specificity

To evaluate the cell type clustering performance of the 'chicken core matrisome', we re-analyzed embryonic stage HH29 (stage 29 Hamburger and Hamilton) [27] chicken hind limb scRNA-seq data [28]. At this point of development, chicken limb progenitor cells have already differentiated into transcriptionally distinct tissue types [28], which is reflected in the separation of our t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction and the superimposed, color-coded clustering information (Fig. 1A). We compared the cell type clustering of the core matrisome to the entire transcriptome and contrasted its performance with highly variably expressed transcription factors - representing a 'traditional cell type identifier' - and an equal number of randomly picked genes, to estimate baseline clustering. Of the 217 chicken core matrisome genes, 136 were

expressed in our limb scRNA-seq data (Data Source File 1). Accordingly, we picked 136 genes randomly, as well as the 136 most variably expressed transcription factors, chosen by maximum variance across all cells in the sample. With these three small subsets of genes - representing only 1.26% of all expressed genes -, we re-clustered our data using the Louvain-Jaccard algorithm. We adjusted the resolution to obtain the same number of clusters as for the entire transcriptome, and plotted the resulting clusters in an unsupervised manner onto a t-SNE plot calculated from the entire transcriptome (Fig. 1B-D). A qualitative inspection of the plots showed that the clusters resulting from a 'random gene set' did not clearly coincide with any clusters identified using the entire transcriptome, suggesting that they failed as transcriptional predictors for any given cell type (Fig. 1A, B). By contrast, 'transcription factor' clusters showed good correspondence to our whole transcriptome clustering (Fig. 1A, C). Intriguingly, we found that the 'core matrisome' was sufficient to identify several cell type clusters (Fig. 1A, D). For example, 'core matrisome' clusters m-7, m-11, m-14, and m-17 corresponded roughly to skeletal progenitors (t-15), joint progenitors (t-3), skin (t-1), and vessel (t-10) clusters, as identified by the entire transcriptome (Fig. 1A, D). Thus, these core matrisome-identified clusters largely reflected cell types of tissues that are embedded in collagen-rich ECMs.

To quantify the separation among random genes-, transcription factors-, and core matrisome-based clusters, we plotted the distribution of all pairwise Euclidean distances, i.e. distances between all pairs of clusters, and compared them to the entire transcriptome result. Both the 'core matrisome' and the 'transcription factors' clusters clearly outperformed ten iterations of 'randomly-picked' genes subsets of equal size (Fig. 1E).

Moreover, using a hypergeometric test, we were able to demonstrate that the probability of cluster overlap - between the entire transcriptome clusters and the three subsets clusters - was substantially higher for 'core matrisome' and the 'transcription factors' clusters (Fig. 1F-H). For the 'core matrisome', this was particularly evident for clusters corresponding to cell types known to produce a complex ECM, such as skeletal cells or skin (Fig. 1G). Moreover, even within the same cell type, the matrisome seemed able to distinguish discrete cell states. For example, 'core matrisome' clusters m-4 and m-6 reconstituted 'transcriptome' cluster t-8, the distal mesenchyme, indicating that the highly proliferative state of this mesenchymal sub-population is reflected by a distinct 'matreotype' (Fig. 1G). Taken together, our re-clustering analysis of chicken limb scRNA-seq data - using only the expression status of either core matrisome genes, transcription factors, or a random control gene set - indicates the potential of core matrisome gene expression status as a cell type and cell state identifier.

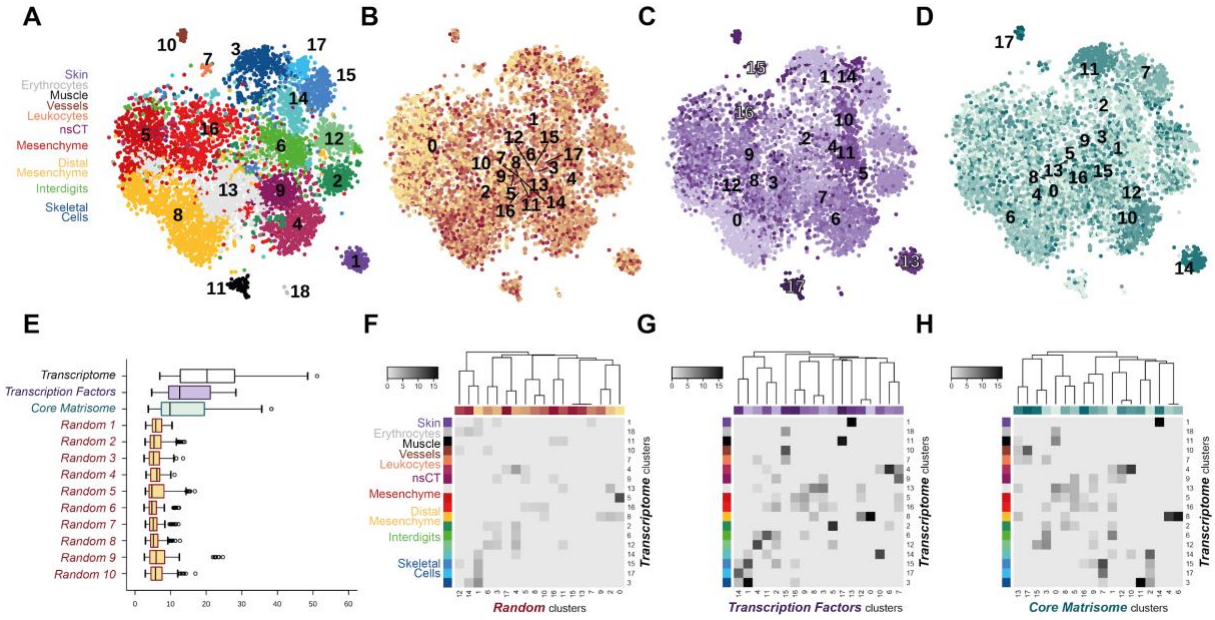


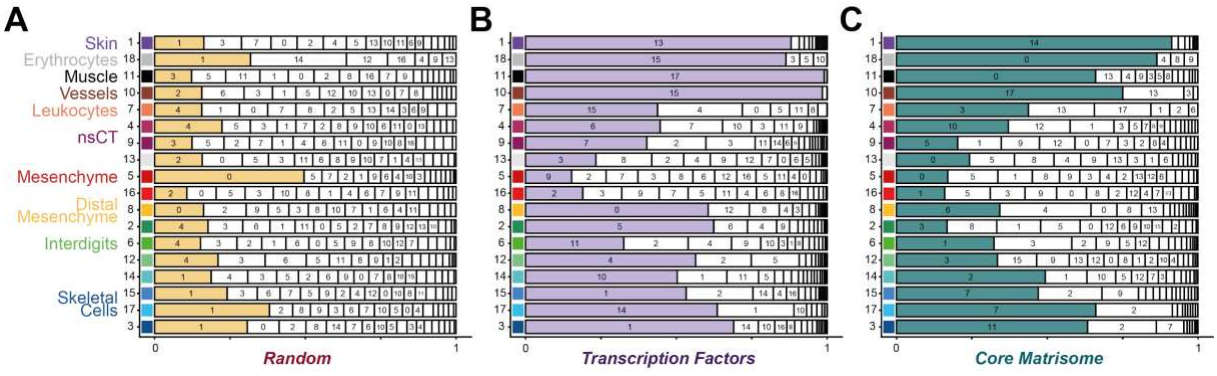
Figure 1. Core matrisome and transcription factors re-capitulate entire transcriptome cell clusters

(A) tSNE representation of 6823 HH29 chicken hindlimb autopod cells from Feregrino et al., 2019. Colors represent unsupervised clustering results based on the transcriptome (A), a randomly sampled set of genes (B), transcription factors (C), and the core matrisome (D). (E) Boxplot of Euclidean distances between clusters based on average expression of 2000 variably expressed genes, calculated for transcriptome, transcription factor, core matrisome, and 10 sets of random gene clusters. (F-H) Heatmap of square root of negative log 10 probability of cluster overlap by hypergeometric test between 'transcriptome' and 'random clusters' (F), 'transcription' factor' (G), and 'matrisome' (H). 'Transcriptome' clusters are grouped by tissue or cell type. For (A-F) details, see Data Source File 1.

Clustering performance and cell type identification by transcription factors and the core matrisome

To further assess the potential of such limited gene subsets to reliably identify cell types from scRNA-seq data, we next sought to quantify their ability to recreate our entire transcriptome cluster composition. We did this on a cluster-by-cluster as well as on a cell-by-cell basis. We first plotted - ordered by percentage - the respective cellular contributions of individual gene subset clusters to the 18 entire transcriptome clusters. As expected, 'random genes' clusters contributed almost uniformly to the different 'transcriptome' clusters (Fig. 2A). The median percentage contribution of the single-largest 'random genes' clusters - highlighted in yellow - was 17%, again reflective of that gene subset's low information content regarding cell type identification. Certain 'transcription factor' clusters, however, contributed more than 90% of a given 'entire transcriptome' cluster (Fig. 2B). For example, 'transcriptome' cluster t-11, *i.e.,* "muscle", was represented to 99% by

156 'transcription factor' cluster tf-17. However, the same "muscle" cluster was only re-
157 captured to 12% by the largest 'random' cluster contributor r-3 (compare Fig. 2A to B,
158 'transcriptome' cluster t-11). Likewise, the 'matrisome' gene subset also performed better
159 than 'random', with the 'Muscle' cluster represented to 66% by 'matrisome' cluster m-0, or
160 'Skin' recaptured to 91% by cluster m-14 (Fig. 2C). However, when comparing the
161 'transcription factor' and 'matrisome' clustering performances within the closely lineage-
162 related lateral plate mesoderm-derived cell types, differences between the two gene subsets
163 emerged. Lateral plate mesoderm-derived tissues in our sample included non-skeletal
164 connective tissue (cl. t-4, t-9), undifferentiated mesenchyme (cl. t-13, t-5, t-16, t-8),
165 interdigital mesenchyme (cl. t-2, t-6, t-12) and skeletal progenitors (cl. t-14, t-15, t-17, t-3).
166 Amongst these, certain cell type clusters contributing to mesenchymal tissues were well
167 defined by their 'transcription factor' signature, yet much less so by their 'matrisome'
168 expression status (e.g. compare cl. t-8, t-2, t-12, Fig. 2B and C). Again, some of these
169 discrepancies might relate to the fact that 'matrisome' signatures can also be indicative of
170 different cell states, whereas 'transcription factors' profiles assign predominantly to cell
171 types. However, cell types contributing to more differentiated tissues with complex ECM
172 composition were equally well defined by both 'transcription factor' and 'matrisome' gene
173 expression signatures (*e.g.,* cl. t-4, and t-14, t-15, t-17, t-3).



174

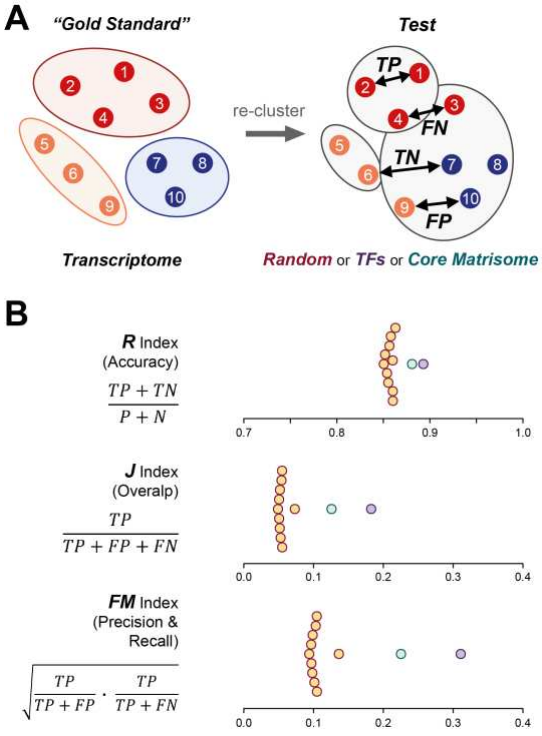
Figure 2. Relative cluster contributions to transcriptome clusters

(A) Relative contribution of 'random' clusters to the 'transcriptome' clusters ordered by size. Cluster IDs are indicated where possible and the biggest contribution per transcriptome cluster is highlighted in color. (B) 'transcription factor' cluster contributions and (C) 'matrisome' cluster contributions are indicated in the same manner. 'transcriptome' clusters are grouped by cell type.

The core matrisome predicts preferentially ECM-rich cell types in early development

To quantify the ability of all our three gene-input-subsets - 'random', 'transcription factor', and 'matrisome' - to correctly predict cluster membership of our "Gold Standard" transcriptome clustering, we decided to use a binary classification scheme based on pairs of cells being in the same or different clusters [29]. Each pair of cells was classified as either "true positive" (TP: two cells are in the same cluster regardless of the input data used), "true negative" (TN: two cells are in different clusters regardless of input data), "false positive" (FP: two cells are in the same cluster although they are in different clusters in the "Gold Standard"), and "false negative" (FN: two cells are in different clusters although they are in the same cluster in the "Gold Standard") (Fig. 3A). Based on the cumulative numbers of TP, TN, FP, and FN of these binary cell pair classifications, we then calculated three different indices commonly used to compare different clustering algorithms [29]: the Rand index, also known as accuracy (*R* index), which measured the percentage of correct classifications; the Jaccard index of overlap (*J* index), which was calculated as the intersection of the two sets divided by the union of the two sets; and the Fowlkes-Mallows index (*FM* index),

197 which represented the geometric mean of precision and recall. The closer each of these
198 indices scores to 1, the more similar the respective gene subset clustering can be considered
199 to the transcriptome "Gold Standard" clustering. Regardless of the index used, the 'core
200 matrisome' and 'transcription factor' gene subsets clearly outperformed ten iterations of
201 'random genes', with 'transcription factors' scoring slightly higher than 'matrisome' genes
202 (Fig. 3B).
203 Collectively, we demonstrate that both 'transcription factors' and 'matrisome' genes can be
204 used as cell type identifiers in scRNA-seq data. The extent to which this holds true,
205 however, seems to depend on the tissue type to which the respective cell types contribute,
206 differences in cell state, as well as the ontogenetic state of their differentiation.



207

208 **Figure 3: Binary classification and re-cluster indices**

209 (A) Each pair of cells in a given 'Test' clustering - i.e. 'Core Matrisome', 'Transcription
210 Factor' or 'random' - is classified based on their relationship to the "Gold Standard"

clustering, as calculated from the entire transcriptome. (B) Based on those binary classifications the quality of each 'Test' clustering is measured with the three indices following Kafieh and Mehridehnavi, 2013. 'Test' clustering indices are calculated for transcription factors (purple), core matrisome (green), and 10 random gene set clusterings (yellow). TP: true positive, TN: true negative, FP: false positive, FN: false negative, R: Rand index, J: Jaccard index, FM: Fowlkes-Mallows index.

Predictive power of 'core matrisome' signature depends on developmental differentiation and is evolutionarily conserved amongst vertebrates

To determine the effect of developmental progression on the cell type-predictive powers of the matrisome, we next focused our attention on an embryonic scRNA-seq times series. We incorporated a previously published time series of the developing mouse hind limb by Kelly and colleagues into our analysis [30]. In their scRNA-seq data sets, we found 244 to 254 out of the total 274 mouse core-matrisome genes expressed. Initial clustering of single-cell transcriptomes showed - as expected - similar tissue composition as in our chicken hindlimb data, as well as as an increase in cell type complexity from the earliest stage E11.5 to E18.5 (Fig. 4A, Data Source File 1).

Using the clusters identified by the entire transcriptome as a benchmark, we then re-clustered the data using either 'random', 'transcription factor' or 'core matrisome' gene sets of equal size, and compared their performances using the previously introduced indices. Over the course of the sampled developmental time window, the predictive powers of both 'transcription factor' and 'core matrisome' increased, *i.e.,* more cell pairs were correctly attributed together in a way reflective of the entire transcriptome cell type clustering (Fig.

4B). To quantify this effect and relate it to the predictive powers of different fractions of 'random' genes from the entire transcriptome, we focused our attention on the two temporal extremes of the time series, E11.5 and E18.5. We randomly sampled increasing numbers of 'random' genes - from 1% to 90% of the entire transcriptome, each sampled and re-clustered 5 times - and plotted the spread of their performances in relation to the 'transcription factor' and 'matrisome' gene subsets (Fig. 4C, D). At E11.5, both 'transcription factor' and 'matrisome' gene subsets clusterings performed at about the rate of 2% of all genes, randomly selected from the transcriptome (Fig. 4C). At E18.5, however, their predictive powers had increased to a level of more than 10% of the entire transcriptome (Fig. 4D). This is noteworthy, as the number of expressed 'core matrisome' genes at that stage corresponds to only 1.28% of the entire transcriptome. Moreover, at these later stages, the 'matrisome' genes subset outperformed 'transcription factors' in these indices, likely a reflection of the ongoing maturation of tissues with high ECM content.

We concluded, as tissues and their ECM mature, the expression status of the core matrisome becomes progressively better at delineating cell types.

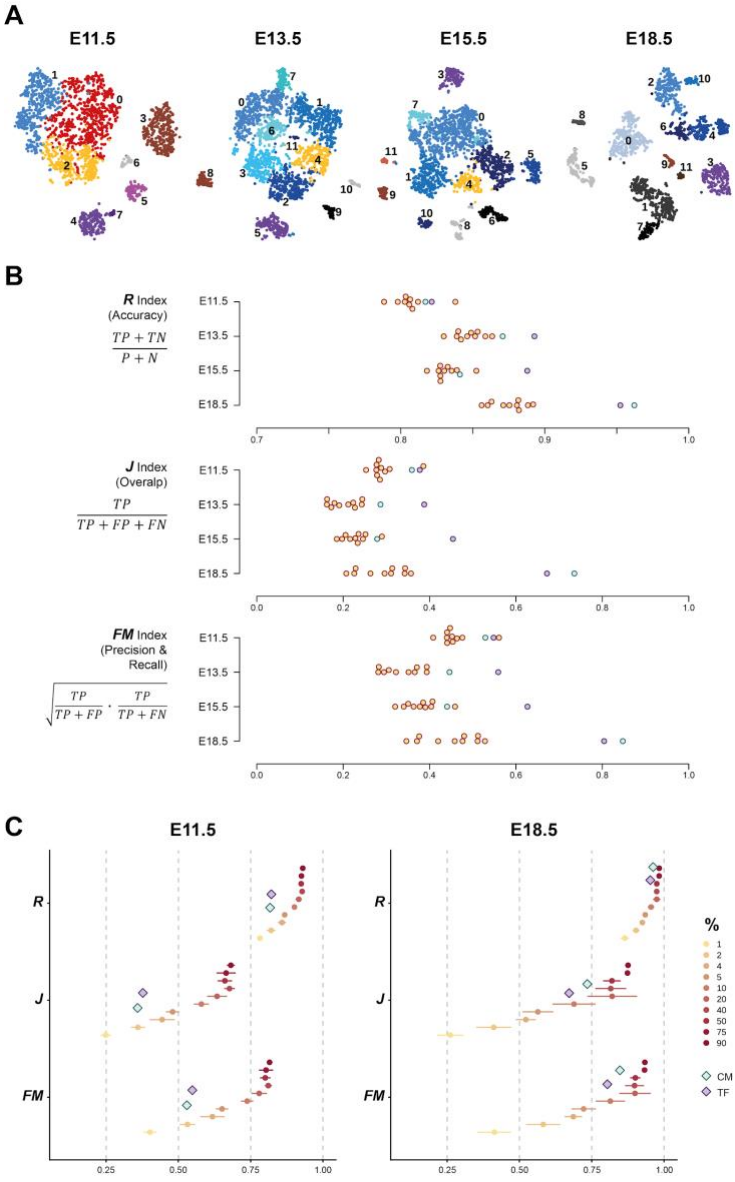


Figure 4. Predictive power of the matrisome during murine development

(A) tSNE representations of four mouse embryo limb data sets of increasing developmental stages (E11.5 to E18.5, E: embryonic day). Colors and numbers indicate 'transcriptome' clusters. Cluster annotation in Data source File 1. (B) Rand, Jaccard and Fowlkes-Mallows indices for all four stages. Colors as in Fig. 3B. (C) Indices of 'matrisome' and 'transcription factor' clustering (triangles) in E11.5 and 18.5 compared to indices of random gene sets of increasing size. Gradient indicates the percentage of the whole transcriptome represented

257 by the random gene set. For each percentage, 5 gene sets were sampled independently and

258 used for clustering.

Discussion

Understanding the molecular parameters that define different cell types and states is fundamental to developmental and regenerative biology. Here we show that the expression status of a small subset of genes, the core matrisome, can suffice to identify cell types and states in the developing chick and mouse limb. Even though it corresponds to less than 2% of the entire transcriptome, we demonstrate that core matrisome expression encodes enough information to cluster scRNA-seq data according to cell types, and cell states. The predictive power of the matrisome increases with developmental time, and can even outperform transcription factors in more differentiated cell and tissue types with high ECM content.

These findings make sense with regards to developmental progression and tissue maturation. During ontogenetic development, transcription factors are thought to guide early differentiation trajectories and eventually specify terminally differentiated cell types [2,4,6]. At later stages of development, the ECM becomes increasingly important, instructing stem cell differentiation and regulating cell and tissue shape, morphogenetic movements, and organogenesis [16]. This holds especially true for tissues with complex ECM composition or high ECM turnover. Consistent with this, we found that in our chicken limb data, skin cells, muscle and skeletal progenitors clusters segregate especially well using core matrisome expression alone (Fig. 1 and 2). The lack of a clear 'matrisome'-based clustering for some of the other mesenchymal cell populations may indicate a less specialized extracellular matrix, or, alternatively, the presence of different cell states within a cell type, each with its own putatively distinct 'matreotype'. Moreover, the overall

282 predictive power of the matrisome increases when comparing cell populations in ECM-rich
283 tissues at progressively later stages of development (Fig. 4).
284
285 Previous work using scRNA-seq to determine molecular changes during adipogenesis (day
286 1-7) *in vitro* found that at day 3 the cell clustering was mainly driven by ECM genes, and
287 at day 7 the core matrisome was one of the top ten most differentially expressed gene
288 ontology terms [31]. Beyond development, planarian scRNA-seq revealed that muscles
289 produce most of the matrisome, and inhibiting one key matrisome gene (hemicentin)
290 resulted in severe epidermal ruffling and displacement of cells during homeostatic tissue
291 turnover, suggesting an important role for tissue regeneration [26]. Furthermore, in healthy
292 human lumbar discs, the core matrisome can be used to distinguish primary annulus
293 fibrosus and nucleus pulposus cells based on 90 out of the 274 core matrisome genes being
294 differentially expressed in the opposite direction using scRNA-seq [32]. Similarly, 115
295 matrisome genes are characteristically expressed in the six cell types that make up the
296 human cutaneous neurofibroma microenvironment [33]. Beyond cell type distinction in
297 tissues, a differential expression of matrisome genes can be observed when cells change
298 their state from a healthy to a diseased cell. For instance, differential expression of
299 matrisome genes was one main characteristic of reprogramming from normal fibroblasts,
300 pericytes, and endothelial cells into tumor cells [34]. During cancer progression,
301 deregulation of matrisome genes is a crucial step observed in early and late metastasis [35].
302 Moreover, core matrisome genes help identify circulating tumor cells in the blood using
303 scRNA-seq [36,37]. These results support our conclusion that matrisome gene expression
304 can serve as a key signature to determine individual cell types, as well as cell states.

This raises the question of why the matrisome is such a good predictor of cell type and state. It is well known that cells can be distinguished based on cell surface receptors [38]. However, it is less appreciated that each cell type can synthesize its own ECM that entails it with unique physical properties [15,39,40]. For instance, placing primary preadipocytes into decellularized ECMs derived from subcutaneous, visceral, or brown adipose tissue influences the preadipocytes' terminal differentiation [41]. Hence, the physical properties of ECM seem to be able to dictate cellular fate and drive stem cell differentiation into neurons, muscle, or bone cells [42]. Besides providing instructive cues during development, ECMs can also change cellular status. Placing senescent cells or aged stem cells in a "younger ECM" rejuvenates these old cells to regain proliferative capacities or stem cell potential, respectively [43,44]. Similarly, placing tumor cells into an embryonic ECM reprograms to non-tumorigenic cells [45]. Hence, there is an intrinsic crosstalk between the ECM and the cells it encapsulates. ECMs, or niches, are made and adapted according to the respective cellular needs or states. Disrupting the crosstalk between cancer and cancer-associated fibroblasts, for instance, by a small molecule that inhibits chromatin remodeling and change matrisome gene expression (*i.e.,* altering the matreotype), prevented tumor growth in xenograft mouse models [46]. Although we lack a current understanding of these underlying molecular crosstalk, these snapshots of ECM compositions - or matreotypes - clearly can reflect distinct cellular properties.

Accordingly, since matreotypes mirror cellular status, they also hold potentially promising prognostic value. For instance, 43 out of the 274 core matrisome genes are significantly

upregulated across multiple cancer types, and 9 ECM genes predicted cancer outcome [47]. Another classifier similar to the matreotype concept is termed tumor matrisome index, which is based on 29 matrisome genes, reliably predicts low- and high-risk groups and chemotherapy responses for small cell lung cancer patients [48]. Matreotypes reflecting chronological age have been recently used to predict drugs that promote healthy aging [49]. Therefore, defining matreotypes has translational value for future biomedical research. Moreover, identifying different subpopulations of a given cell type will be critical to overcome the problem of cellular heterogeneity and aid personalized medical applications.

In summary, with our scRNA-seq analyses, we provided evidence for a previously postulated concept, namely that 'each cell type produces its unique ECM' [15,17]. While the best molecular proxies for cell-type identification continue to be discussed [1–3], we made the unexpected discovery that expressed core matrisome genes - corresponding to less than 2% of a typical transcriptome - hold enough information to re-cluster scRNA-seq data as well as transcription factor signatures. For more mature cells, the core matrisome embodied substantial predictive value to identify cell types and states. Hence, future work on defining matreotypes of different cell types and states might inform diagnostics and personalized medicine.

Materials and Methods

Matrisome gene lists

Curated matrisome gene lists for mouse and human are available on 'The Matrisome Project' (http://matrisome.org/; [14]. To create a matrisome list for chicken, a union of the human and mouse matrisome lists was used to define chick one-to-one orthologs in the ENSEMBL Galgal5.0 annotation.

Single-cell RNA-sequencing data

Previously published single-cell RNA sequencing (scRNA-seq) datasets sampling the chicken embryonic limb [28] and the mouse embryonic limb [30]; stages E11.5 to E18.5) were used for all analyses. The raw data is accessible at Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), under accession numbers GSE130439 (chicken) and GSE142425 (mouse).

Data pre-processing

Raw UMI count tables were used to initiate 'Seurat objects' for all mouse samples in R, using package Seurat v3.1.4. Next, low quality cells and outliers were filtered out. Chicken cells with an UMI count higher than 4 times the average UMI count, less than 20 percent of the median UMI count or more than 10 percent mitochondrial or ribosomal content were removed [28]. Mouse cells expressing less than 200, more than 6000 genes, or more than 10 percent mitochondrial RNA were removed. All expressed genes were considered. Normalization, identifying the top 2000 variable genes and scaling of the data was applied with Seurat's built-in functions.

370

371 ## Dimensionality reduction

372 For all chicken and mouse, Seurat objects, principal components analysis was performed on
373 all expressed genes, and significant components were selected as such, if they were located
374 outside of the Marchenko Pastur distribution [50]. The same criterion for significance was
375 applied on all principal component analysis on core matrisome and random subset genes.
376 The cells were visualized with the dimensionality reduction algorithm tSNE [51].
377 'matrisome' and 'random' clusters for all datasets were represented on the same tSNEs
378 generated from the 'transcriptome' principal components. To define a 'gold standard' of
379 scRNAseq-based cell type clustering, k-nearest neighbour (kNN) graphs and Jaccard indices
380 of overlap between a cell and its neighbours were used to create shared nearest neighbour
381 (SNN) graphs, with the Seurat 'FindNeighbors' function using all expressed 'transcriptome'
382 genes. Clusters of cells were then defined by 'FindClusters', by applying Louvain modularity
383 optimization algorithms on SNN graphs. As the number of clusters can be influenced by the
384 resolution parameters, please refer to the supplementary data for detailed parameters of
385 significant dimensions and resolutions used in clustering for all samples. For 'matrisome'-
386 based clustering, all expressed core matrisome genes were considered for clustering. For
387 'random'-based clustering, for ten iterations, randomly picked genes from the whole
388 transcriptome were used such that they matched the number of expressed core matrisome
389 genes, as well as resulting in the same number of individual clusters as the 'transcriptome'
390 and 'matrisome' clustering. The core matrisome genes and transcription factors were
391 excluded from this sampling. After ordering the transcription factors by expression

392 variability, the set of top transcription factors matching the size of the expressed core
393 matrisome was used to recluster the cells.
394
395 **Cluster cell type annotation**
396 Differentially expressed genes between mouse clusters with a minimum natural log fold
397 change of 0.25 were identified using a Wilcoxon rank sum test, and were then used to assign
398 putative celltype identities of each cluster. Only genes expressed in at least 25% of cells in
399 one of the two populations were considered. For all clusters, all and the top five
400 differentially genes per cluster can be found in the Data Source File 1. Chicken clusters had
401 been previously annotated (Feregrino et al., 2019).
402
403 **Distance Boxplots**
404 To assess cluster-to-cluster proximity of 'transcriptome'-, 'matrisome'-, 'random'-, and
405 'transcription factor'- based clustering approaches, Euclidean pairwise distances between
406 each cluster were calculated on the averaged scaled expression per cluster of the top 2000
407 variably expressed genes. The same 2000 genes were used to compare all three clustering
408 approaches.
409
410 **Hypergeometric test**
411 Probabilities of overlap between clusters were calculated with 'phyper'. The
412 hypergeometric test takes into account the size of the reference cluster ('transcriptome') *m*,
413 the size of the test cluster ('matrisome', 'random', and 'transcription factors') *k*, the number

of non-tested cells (total number of cells $N-m$) and the size of the overlap x to calculate the probability of the overlap to occur at random. Probabilities were calculated for overlaps between all clusters. Probabilities equal to zero were replaced with the smallest non-zero probability to prevent infinite values after transformation, and probabilities bigger than 0.05 were set to 1 for plot aesthetics. 'Heatmap3' [52] was used to plot square root negative log 10 transformed probabilities.

Visualizing cluster contributions

Barplots were created with 'ggplot2' [53].

Indices

The 'Rand index', also known as 'accuracy', was calculated as following:

$$R = \frac{TP + TN}{TP + FP + FN + TN}$$

It measures the percentage of correct classifications.

The 'Jaccard index of overlap' is calculated as intersection over union. It does not take the TN into account, which represent the most classifications and might be confounding in the Rand index:

$$J = \frac{TP}{TP + FP + FN}$$

At last, the 'Fowlkes-Mallows index' is the geometric mean of precision and recall. Precision measures how many positive pairs (cells within the same cluster in the test clustering) are true positives (cells within the same cluster in the gold standard). Recall is the percentage of true positives identified by all actual positives:

$$FM = \sqrt{\frac{TP}{TP + FP} * \frac{TP}{TP + FN}}$$

All indices range from 0 (no correct classification by the test clustering) to 1 (identical clustering by the test clustering).

Comparing 'matrisome' and 'transcription factors' against 'random' subsets of increasing size

The information content of the 'matrisome' and the 'transcription factor' subset was compared to 'random' subsets containing 1, 2, 4, 5, 10, 20, 40, 50, 75, and 90 percent of the E11. 5 and E18.5 chicken transcriptomes. For each percentage, five 'random' subsets resulting in the same number of clusters as the 'matrisome' were sampled, clustered, and indices were calculated.

Author contributions

FS, CF, PT and CYE designed the study. FS performed all analyses with help from CF. All authors interpreted the data. FS, PT, and CYE wrote the manuscript with comments from CF.

Author Information

The authors have no competing interests to declare. Correspondence should be addressed to C. Y. E. and P.T.

Acknowledgment

We thank Christian Beisel and the joint Genomics Facility Basel for help with single-cell sequencing, and members of both groups for useful discussions. All calculations were performed at sciCORE (http://scicore.unibas.ch/), the scientific computing center at the University of Basel. This work was supported by funding from the Swiss National Science Foundation PP00P3_163898 to CYE, and from the Swiss 3R Competence Centre (3RCC grant OC-2018-005), the Swiss National Science Foundation (SNSF project grant 310030_189242), and the University of Basel to PT.

References

[1] What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?, Cell Syst. 4 (2017) 255–259. https://doi.org/10.1016/j.cels.2017.03.006.

[2] B. Xia, I. Yanai, A periodic table of cell types, Development. 146 (2019) dev169854. https://doi.org/10.1242/dev.169854.

[3] K.L. McKinley, D. Castillo-Azofeifa, O.D. Klein, Tools and Concepts for Interrogating and Defining Cellular Identity, Cell Stem Cell. 26 (2020) 632–656. https://doi.org/10.1016/j.stem.2020.03.015.

[4] D. Arendt, J.M. Musser, C.V.H. Baker, A. Bergman, C. Cepko, D.H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, M.D. Laubichler, G.P. Wagner, The origin and evolution of cell types, Nat Rev Genet. 17 (2016) 744–757. https://doi.org/10.1038/nrg.2016.127.

[5] M. Crow, J. Gillis, Single cell RNA-sequencing: replicability of cell types., Curr Opin Neurobiol. 56 (2019) 69–77. https://doi.org/10.1016/j.conb.2018.12.002.

[6] The Regulatory Genome, (2006). https://doi.org/10.1016/b978-0-12-088563-3.x5018-4.

[7] C.H. Holland, J. Tanevski, J. Perales-Patón, J. Gleixner, M.P. Kumar, E. Mereu, B.A. Joughin, O. Stegle, D.A. Lauffenburger, H. Heyn, B. Szalai, J. Saez-Rodriguez, Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data, Genome Biol. 21 (2020) 36. https://doi.org/10.1186/s13059-020-1949-z.

[8] F.B. Ardakani, K. Kattler, T. Heinen, F. Schmidt, D. Feuerborn, G. Gasparoni, K. Lepikhov, P. Nell, J. Hengstler, J. Walter, M.H. Schulz, Prediction of single-cell gene expression for transcription factor analysis., Gigascience. 9 (2020) giaa113-. https://doi.org/10.1093/gigascience/giaa113.

[9] R.O. Hynes, The Extracellular Matrix: Not Just Pretty Fibrils, Science. 326 (2009) 1216–1219. https://doi.org/10.1126/science.1176009.

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

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

[12] J. Chang, R. Garva, A. Pickard, C.-Y.C. Yeung, V. Mallikarjun, J. Swift, D.F. Holmes, B. Calverley, Y. Lu, A. Adamson, H. Raymond-Hayling, O. Jensen, T. Shearer, Q.J. Meng, K.E. Kadler, Circadian control of the secretory pathway maintains collagen homeostasis., Nat Cell Biol. 22 (2020) 74–86. https://doi.org/10.1038/s41556-019-0441-z.

[13] A. Naba, K.R. Clauser, S. Hoersch, H. Liu, S.A. Carr, R.O. Hynes, The matrisome: in

502 silico definition and in vivo characterization by proteomics of normal and tumor
503 extracellular matrices., Molecular & Cellular Proteomics : MCP. 11 (2012) M111.014647.
504 https://doi.org/10.1074/mcp.m111.014647.
505 [14] A. Naba, K.R. Clauser, H. Ding, C.A. Whittaker, S.A. Carr, R.O. Hynes, The
506 extracellular matrix: Tools and insights for the “omics” era., Matrix Biology : Journal of the
507 International Society for Matrix Biology. 49 (2016) 10–24.
508 https://doi.org/10.1016/j.matbio.2015.06.003.
509 [15] C. Frantz, K.M. Stewart, V.M. Weaver, The extracellular matrix at a glance., Journal
510 of Cell Science. 123 (2010) 4195–4200. https://doi.org/10.1242/jcs.023820.
511 [16] D.A.C. Walma, K.M. Yamada, The extracellular matrix in development,
512 Development. 147 (2020) dev175596. https://doi.org/10.1242/dev.175596.
513 [17] C.Y. Ewald, The Matrisome during Aging and Longevity: A Systems-Level Approach
514 toward Defining Matreotypes Promoting Healthy Aging., Gerontology. (2019) 1–9.
515 https://doi.org/10.1159/000504295.
516 [18] M.A. Chapman, K. Mukund, S. Subramaniam, D. Brenner, R.L. Lieber, Three distinct
517 cell populations express extracellular matrix proteins and increase in number during
518 skeletal muscle fibrosis, Am J Physiol-Cell Ph. 312 (2017) C131–C143.
519 https://doi.org/10.1152/ajpcell.00226.2016.
520 [19] P. Hiebert, M.S. Wietecha, M. Cangkrama, E. Haertel, E. Mavrogonatou, M. Stumpe,
521 H. Steenbock, S. Grossi, H.-D. Beer, P. Angel, J. Brinckmann, D. Kletsas, J. Dengjel, S.
522 Werner, Nrf2-Mediated Fibroblast Reprogramming Drives Cellular Senescence by
523 Targeting the Matrisome., Dev Cell. 46 (2018) 145-161.e10.
524 https://doi.org/10.1016/j.devcel.2018.06.012.
525 [20] I.N. Taha, A. Naba, Exploring the extracellular matrix in health and disease using
526 proteomics., Essays in Biochemistry. 63 (2019) 417–432.
527 https://doi.org/10.1042/ebc20190001.
528 [21] A. Naba, O.M.T. Pearce, A.D. Rosario, D. Ma, H. Ding, V. Rajeeve, P.R. Cutillas, F.R.
529 Balkwill, R.O. Hynes, Characterization of the Extracellular Matrix of Normal and
530 Diseased Tissues Using Proteomics, J Proteome Res. 16 (2017) 3083–3091.
531 https://doi.org/10.1021/acs.jproteome.7b00191.
532 [22] A.M. Socovich, A. Naba, The cancer matrisome: From comprehensive
533 characterization to biomarker discovery., Seminars in Cell & Developmental Biology. 89
534 (2019) 157–166. https://doi.org/10.1016/j.semcdb.2018.06.005.
535 [23] P. Nauroy, S. Hughes, A. Naba, F. Ruggiero, The in-silico zebrafish matrisome: A new
536 tool to study extracellular matrix gene and protein functions., Matrix Biology : Journal of
537 the International Society for Matrix Biology. 65 (2018) 5–13.
538 https://doi.org/10.1016/j.matbio.2017.07.001.
539 [24] A.C. Teuscher, E. Jongsma, M.N. Davis, C. Statzer, J.M. Gebauer, A. Naba, C.Y.
540 Ewald, The in-silico characterization of the Caenorhabditis elegans matrisome and

proposal of a novel collagen classification, Matrix Biology Plus. (2019) 1–13. https://doi.org/10.1016/j.mbplus.2018.11.001.

[25] M.N. Davis, S. Horne-Badovinac, A. Naba, In-silico definition of the Drosophila melanogaster matrisome, Matrix Biology Plus. 4 (2019) 100015. https://doi.org/10.1016/j.mbplus.2019.100015.

[26] L.E. Cote, E. Simental, P.W. Reddien, Muscle functions as a connective tissue and source of extracellular matrix in planarians, Nat Commun. 10 (2019) 1592. https://doi.org/10.1038/s41467-019-09539-6.

[27] V. Hamburger, H.L. Hamilton, A series of normal stages in the development of the chick embryo, J Morphol. 88 (1951) 49–92. https://doi.org/10.1002/jmor.1050880104.

[28] C. Feregrino, F. Sacher, O. Parnas, P. Tschopp, A single-cell transcriptomic atlas of the developing chicken limb., Bmc Genomics. 20 (2019) 401. https://doi.org/10.1186/s12864-019-5802-2.

[29] R. Kafieh, A. Mehridehnavi, A comprehensive comparison of different clustering methods for reliability analysis of microarray data., J Medical Signals Sensors. 3 (2013) 22–30.

[30] N.H. Kelly, N.P.T. Huynh, F. Guilak, Single cell RNA-sequencing reveals cellular heterogeneity and trajectories of lineage specification during murine embryonic limb development, Matrix Biol. 89 (2020) 1–10. https://doi.org/10.1016/j.matbio.2019.12.004.

[31] A.K. Ramirez, S.N. Dankel, B. Rastegarpanah, W. Cai, R. Xue, M. Crovella, Y.-H. Tseng, C.R. Kahn, S. Kasif, Single-cell transcriptional networks in differentiating preadipocytes suggest drivers associated with tissue heterogeneity, Nat Commun. 11 (2020) 2117. https://doi.org/10.1038/s41467-020-16019-9.

[32] L.M. Fernandes, N.M. Khan, C.M. Trochez, M. Duan, M.E. Diaz-Hernandez, S.M. Presciutti, G. Gibson, H. Drissi, Single-cell RNA-seq identifies unique transcriptional landscapes of human nucleus pulposus and annulus fibrosus cells., Sci Rep-Uk. 10 (2020) 15263. https://doi.org/10.1038/s41598-020-72261-7.

[33] J.-P. Brosseau, A.A. Sathe, Y. Wang, T. Nguyen, D.A. Glass, C. Xing, L.Q. Le, Human cutaneous neurofibroma matrisome revealed by single-cell RNA sequencing., Acta Neuropathologica Commun. 9 (2021) 11. https://doi.org/10.1186/s40478-020-01103-4.

[34] A. Sathe, S.M. Grimes, B.T. Lau, J. Chen, C. Suarez, R.J. Huang, G. Poultsides, H.P. Ji, Single-Cell Genomic Characterization Reveals the Cellular Reprogramming of the Gastric Tumor Microenvironment, Clin Cancer Res. 26 (2020) 2640–2653. https://doi.org/10.1158/1078-0432.ccr-19-3231.

[35] S. Mitra, K. Tiwari, R. Podicheti, T. Pandhiri, D.B. Rusch, A. Bonetto, C. Zhang, A.K. Mitra, Transcriptome Profiling Reveals Matrisome Alteration as a Key Feature of Ovarian Cancer Progression., Cancers. 11 (2019) 1513. https://doi.org/10.3390/cancers11101513.

[36] D.T. Ting, B.S. Wittner, M. Ligorio, N. Vincent Jordan, A.M. Shah, D.T. Miyamoto, N. Aceto, F. Bersani, B.W. Brannigan, K. Xega, J.C. Ciciliano, H. Zhu, O.C. MacKenzie, J.

Trautwein, K.S. Arora, M. Shahid, H.L. Ellis, N. Qu, N. Bardeesy, M.N. Rivera, V. Deshpande, C.R. Ferrone, R. Kapur, S. Ramaswamy, T. Shioda, M. Toner, S. Maheswaran, D.A. Haber, Single-Cell RNA Sequencing Identifies Extracellular Matrix Gene Expression by Pancreatic Circulating Tumor Cells, Cell Reports. 8 (2014) 1905–1918. https://doi.org/10.1016/j.celrep.2014.08.029.

[37] S.B. Lim, T. Yeo, W.D. Lee, A.A.S. Bhagat, S.J. Tan, D.S.W. Tan, W.-T. Lim, C.T. Lim, Addressing cellular heterogeneity in tumor and circulation for refined prognostication., P Natl Acad Sci Usa. 116 (2019) 17957–17962. https://doi.org/10.1073/pnas.1907904116.

[38] D. Bausch-Fluck, U. Goldmann, S. Müller, M. van Oostrum, M. Müller, O.T. Schubert, B. Wollscheid, The in silico human surfaceome., P Natl Acad Sci Usa. 115 (2018) E10988–E10997. https://doi.org/10.1073/pnas.1808790115.

[39] T.J. McKee, G. Perlman, M. Morris, S.V. Komarova, Extracellular matrix composition of connective tissues: a systematic review and meta-analysis, Sci Rep-Uk. 9 (2019) 10542. https://doi.org/10.1038/s41598-019-46896-0.

[40] B. Yue, Biology of the Extracellular Matrix, J Glaucoma. 23 (2014) S20–S23. https://doi.org/10.1097/ijg.0000000000000108.

[41] G. Grandl, S. Müller, H. Moest, C. Moser, B. Wollscheid, C. Wolfrum, Depot specific differences in the adipogenic potential of precursors are mediated by collagenous extracellular matrix and Flotillin 2 dependent signaling, Mol Metab. 5 (2016) 937–947. https://doi.org/10.1016/j.molmet.2016.07.008.

[42] A. Kumar, J.K. Placone, A.J. Engler, Understanding the extracellular forces that determine cell fate and maintenance, Development. 144 (2017) 4261–4270. https://doi.org/10.1242/dev.158469.

[43] H.R. Choi, K.A. Cho, H.T. Kang, J.B. Lee, M. Kaeberlein, Y. Suh, I.K. Chung, S.C. Park, Restoration of senescent human diploid fibroblasts by modulation of the extracellular matrix., Aging Cell. 10 (2011) 148–157. https://doi.org/10.1111/j.1474-9726.2010.00654.x.

[44] Y. Sun, W. Li, Z. Lu, R. Chen, J. Ling, Q. Ran, R.L. Jilka, X.-D. Chen, Rescuing replication and osteogenesis of aged mesenchymal stem cells by exposure to a young extracellular matrix., The FASEB Journal. 25 (2011) 1474–1485. https://doi.org/10.1096/fj.10-161497.

[45] M.J.C. Hendrix, E.A. Seftor, R.E.B. Seftor, J. Kasemeier-Kulesa, P.M. Kulesa, L.-M. Postovit, Reprogramming metastatic tumour cells with embryonic microenvironments., Nature Reviews. Cancer. 7 (2007) 246–255. https://doi.org/10.1038/nrc2108.

[46] K.C. Honselmann, P. Finetti, D.J. Birnbaum, C.S. Monsalve, U.F. Wellner, S.K.S. Begg, A. Nakagawa, T. Hank, A. Li, M.A. Goldsworthy, H. Sharma, F. Bertucci, D. Birnbaum, E. Tai, M. Ligorio, D.T. Ting, O. Schilling, M.L. Biniossek, P. Bronsert, C.R. Ferrone, T. Keck, M. Mino-Kenudson, K.D. Lillemoe, A.L. Warshaw, C.F. Castillo, A.S. Liss, Neoplastic–Stromal Cell Cross-talk Regulates Matrisome Expression in Pancreatic

619 Cancer, Mol Cancer Res. 18 (2020) 1889–1902. https://doi.org/10.1158/1541-7786.mcr-20-
620 0439.
621 [47] A.E. Yuzhalin, T. Urbonas, M.A. Silva, R.J. Muschel, A.N. Gordon-Weeks, A core
622 matrisome gene signature predicts cancer outcome., Brit J Cancer. 118 (2018) 435–440.
623 https://doi.org/10.1038/bjc.2017.458.
624 [48] S.B. LIM, S.J. TAN, W.-T. LIM, C.T. LIM, An extracellular matrix-related prognostic
625 and predictive indicator for early-stage non-small cell lung cancer, Nat Commun. 8 (2017)
626 1734. https://doi.org/10.1038/s41467-017-01430-6.
627 [49] C. Statzer, E. Jongsma, S.X. Liu, A. Dakhovnik, F. Wandrey, P. Mozharovskyi, F.
628 Zülli, C.Y. Ewald, Youthful and age-related matreotypes predict drugs promoting
629 longevity, BioRxiv. (2021). https://doi.org/https://doi.org/10.1101/2021.01.26.428242.
630 [50] K. Shekhar, S.W. Lapan, I.E. Whitney, N.M. Tran, E.Z. Macosko, M. Kowalczyk, X.
631 Adiconis, J.Z. Levin, J. Nemesh, M. Goldman, S.A. McCarroll, C.L. Cepko, A. Regev, J.R.
632 Sanes, Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell
633 Transcriptomics., Cell. 166 (2016) 1308-1323.e30.
634 https://doi.org/10.1016/j.cell.2016.07.054.
635 [51] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine
636 Learning Research. (2008) 2579–2605.
637 https://www.jmlr.org/papers/v9/vandermaaten08a.html.
638 [52] S. Zhao, Y. Guo, Q. Sheng, Y. Shyr, Heatmap3: an improved heatmap package with
639 more powerful and convenient features, Bmc Bioinformatics. 15 (2014) P16.
640 https://doi.org/10.1186/1471-2105-15-s10-p16.
641 [53] C. Ginestet, ggplot2: Elegant Graphics for Data Analysis: Book Reviews, J Royal
642 Statistical Soc Ser Statistics Soc. 174 (2011) 245–246. https://doi.org/10.1111/j.1467-
643 985x.2010.00676_9.x.