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

Fabio Sacher [a], Christian Feregrino [a], Patrick Tschopp [a]* and Collin Y. Ewald [b]*

a - ***Laboratory of Regulatory Evolution,*** *DUW Zoology, University of Basel, Basel CH-4051, Switzerland*
b - ***Laboratory of Extracellular Matrix Regeneration,*** *Institute of Translational Medicine, Department of Health Sciences and Technology, ETH Zürich, Schwerzenbach CH-8603, Switzerland*

Correspondence to Patrick Tschopp and Collin Y. Ewald: *patrick.tschopp@unibas.ch (P. Tschopp), collin-ewald@ethz.ch (C.Y. Ewald)*
https://doi.org/10.1016/j.mbplus.2021.100069

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 essential 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. Moreover, using cross-species analyses, we show that these cell type-specific signatures are evolutionarily conserved. 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 looked 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 [9–13]. In fact, transcription and translation of some ECM genes are even coupled to circadian rhythm, highlighting the dynamic nature of ECM composition [14]. Experimentally, ECM composition has so far been determined mostly by proteomics assays [15]. 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 [15,16]. 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 [15,16].

Importantly, it has been postulated that each cell type produces its own unique ECM [16–19]. 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 [19]. For instance, cellular status, including metabolic, healthy or pathologic, or aging, have been associated with distinct ECM expression patterns (*i.e.,* matreotypes) [16,19–23]. Furthermore, cancer-specific cell types can be identified based on their unique ECM composition [15,16,22,24]. 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 examined 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 mouse limbs, we find that the core matrisome has less predictive power in undifferentiated cells, early during development, but outcompetes transcription factors later in development and in more differentiated cell types. Intriguingly, these cell-type-specific core matrisome signatures appear to be conserved between homologous cell types of distantly related species. Consequently, we propose matreotype gene expression signatures as context-dependent proxies for identifying cell types.

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 [16,25–28]. 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) [29] chicken hind limb scRNA-seq data [30]. At this point of development, chicken limb progenitor cells have already differentiated into transcriptionally distinct tissue types [30], 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

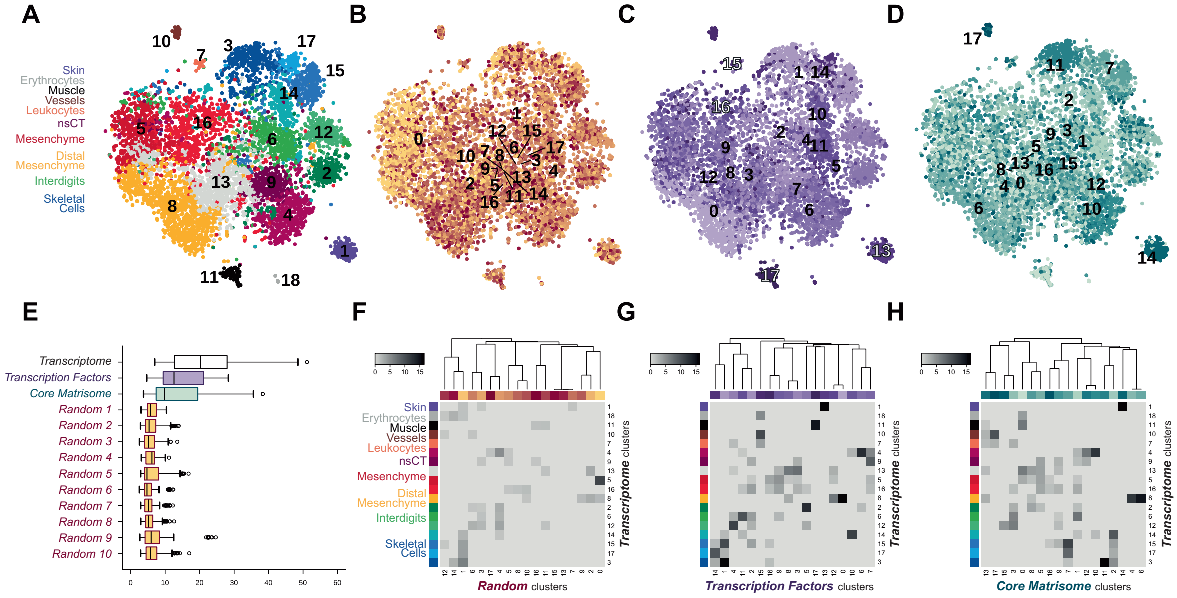

Fig. 1. Core matrisome and transcription factors re-capitulate entire transcriptome cell clusters (A) tSNE representation of 6823 HH29 chicken hindlimb autopod cells from Feregrino et al., 2019. Colors represent unsupervised clustering results based on the transcriptome (A), a randomly sampled set of 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 the square root of negative log 10 probability of cluster overlap by hypergeometric test between 'Transcriptome' and 'Random' clusters (F), 'Transcription Factor' (G), and 'Core Matrisome' (H). 'Transcriptome' clusters are grouped by tissue or cell type (nsCT = non-skeletal connective tissue). For (A–F) details, see Data Source File 1.

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 Factor' clusters clearly outperformed ten iterations of 'Random' 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 Factor' 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 subpopulation is reflected by distinct 'matreotypes' (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.

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' gene clusters contributed almost uniformly to the different 'Transcriptome' clusters (Fig. 2A). The median percentage contribution of the single-largest 'Random' gene 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 'Transcription Factor' cluster tf-17. However, the same "muscle" cluster was only re-captured to 12% by the largest 'Random' cluster contributor r-3 (compare Fig. 2A to B, 'transcriptome' cluster t-11). Likewise, the 'Core Matrisome' gene subset also performed better than 'Random', with the 'Muscle' cluster represented to 66% by 'Core Matrisome' cluster m-0, or 'Skin' recaptured to 91% by cluster m-14 (Fig. 2C). However, when comparing the 'Transcription Factor' and 'Core Matrisome' clustering performances within the closely lineage-related lateral plate mesoderm-derived cell types, differences between the two gene subsets emerged. Lateral plate mesoderm-derived tissues in our sample included non-skeletal connective tissue (cl. t-4, t-9), undifferentiated mesenchyme (cl. t-13, t-5, t-16, t-8), interdigital mesenchyme (cl. t-2, t-6, t-12) and skeletal progenitors (cl. t-14, t-15, t-17, t-3). Amongst these, certain cell type clusters contributing to mesenchymal tissues were well defined by their 'Transcription Factor' signature, yet much less so by their 'Core Matrisome' expression status (e.g. compare cl. t-8, t-2, t-12, Fig. 2B and C). Again, some of these discrepancies might relate to the fact that 'Core Matrisome' signatures can also be indicative of different cell states, whereas 'Transcription Factor' profiles assign predominantly to cell types. However, cell types contributing to more differentiated tissues with complex ECM composition were equally well defined by both 'Transcription Factor' and 'Core Matrisome' gene expression signatures (e.g., cl. t-4, and t-14, t-15, t-17, t-3).

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 'Core 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 [31]. 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 [31]: 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), which represented the geometric mean of precision and recall. The closer each of these indices scores to 1, the more similar the respective gene subset clustering can be considered to the transcriptome 'Gold

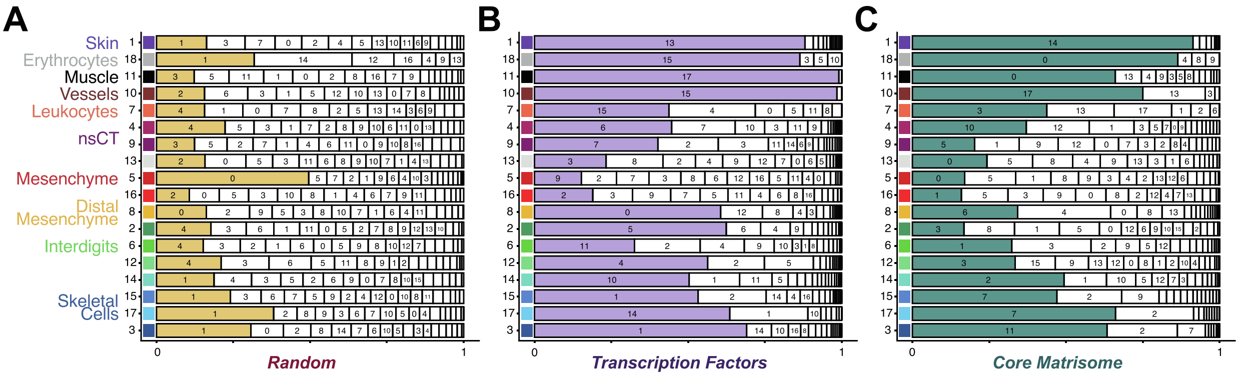

Fig. 2. Relative cluster contributions to transcriptome clusters (A) Relative contribution of 'Random' clusters to the 'Transcriptome' clusters ordered by size. Cluster IDs are indicated where possible, and the biggest contribution per transcriptome cluster is highlighted in color. (B) 'Transcription Factor' cluster contributions and (C) 'Core Matrisome' cluster contributions are indicated in the same manner. 'Transcriptome' clusters are grouped by cell type (nsCT = non-skeletal connective tissue).

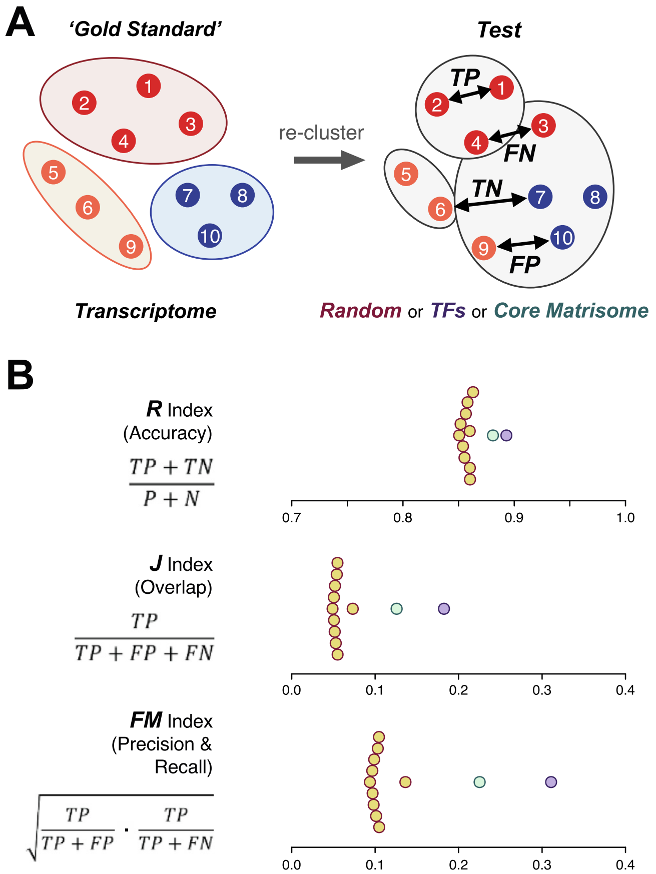

Fig. 3. Binary classification and re-cluster indices (A) Each pair of cells in a given 'Test' clustering – *i.e.,* 'Core Matrisome', 'Transcription Factor' or 'Random' – is classified based on their relationship to the 'Gold Standard' clustering, as calculated from the entire transcriptome. (B) Based 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 factor' (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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Standard' clustering. Regardless of the index used, the 'Core Matrisome' and 'Transcription Factor' gene subsets clearly outperformed ten iterations of 'Random' genes, with 'Transcription Factor' scoring slightly higher than 'Core Matrisome' genes (Fig. 3B).

Collectively, we demonstrated that both 'Transcription Factor' and 'Core Matrisome' genes can be used as cell type identifiers in scRNA-seq data. The extent to which this holds true, however, seems to depend on the tissue type to which the respective cell types contribute, differences in cell state, as well as the ontogenetic state of their differentiation.

Predictive power of 'Core Matrisome' signatures increases with developmental differentiation

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 [32]. In their scRNA-seq data sets, we found 244–254 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 an increase in cell type complexity from the earliest stage E11.5–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 Factors' and 'Core Matrisome' increased, *i.e.,* more cell pairs were correctly attributed together in a way reflective of

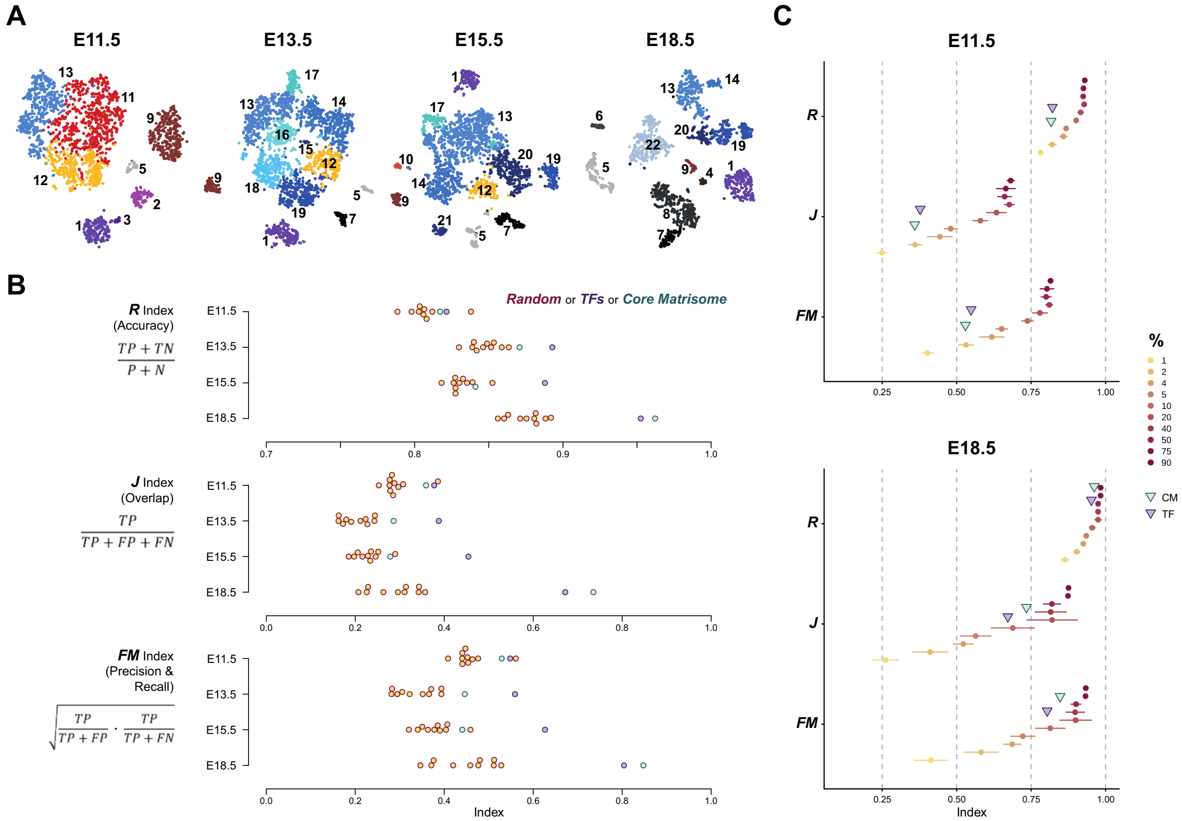

Fig. 4. Predictive power of the core matrisome during murine development (A) tSNE representations of four mouse embryo limb data sets of increasing developmental stages (E11.5 to E18.5, E: embryonic day). Shared colors and numbers indicate transcriptionally similar clusters (see 'Equivalency matrix', Data Source File 1). (B) Rand, Jaccard and Fowlkes-Mallows indices for all four stages. Colors as in Fig. 3B. (C) Indices of 'Core Matrisome' and 'Transcription Factor' clustering (triangles) in E11.5 and 18.5 compared to indices of 'Random' gene sets of increasing size (transcription factors and core matrisome genes were excluded from sampling). Gradient indicates the percentage of the remaining transcriptome represented by the 'Random' gene set. For each percentage, 5 gene sets were sampled independently and used for clustering.

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 remaining transcriptome (*i.e.*, with transcription factors and the core matrisome removed), 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 remaining transcriptome, each sampled and re-clustered 5 times – and plotted the spread of their performances in relation to the 'Transcription Factor' and 'Core Matrisome' gene subsets (Fig. 4C, D). At E11.5, both 'Transcription Factor' and 'Core Matrisome' gene subsets clusterings performed at about the rate of 2% of all genes, randomly selected from the remaining transcriptome (Fig. 4C). At E18.5, however, their predictive powers had increased to a level of more than 10% of the remaining 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 'Core Matrisome' genes subset outperformed 'Transcription Factors' in these indices, likely a reflection of the ongoing maturation of tissues with high ECM content.

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

Homologous cell types show evolutionary conserved extracellular matrix gene expression profiles across species boundaries

Next, we asked to what extent the cell type information contained within the core matrisome would hold up over evolutionary timescales. The two lineages giving rise to the avian and mammalian clades have been separated for over 300 million years. Given the similar tissue composition of the two sets of samples used so far, we reasoned that a comparative mouse and chicken analysis would allow us to probe the information content of the core matrisome across amniotes. To do so, we first averaged the individual cellular normalized UMI counts on a cluster-by-cluster basis into so-called ‘pseudobulk’ count matrices, to approximate cell type-specific bulk gene expression levels (see Material and Methods). Using the ‘Transcriptome’ clustering of each sample as the common cell type delineator, we calculated pseudobulk count matrices for the four previously used gene sets – that is, ‘Transcriptome’, ‘Random’, ‘Transcription Factors’ and ‘Core Matrisome’. Taking into account only one-to-one orthologs expressed in all samples, and in both species, this resulted in pseudobulk count matrices spanning 62 cell type clusters (18 from chicken and 44 from mouse) and 7512 genes for the ‘Transcriptome’, and 106 genes for the ‘Core Matrisome’. We contrasted these 106 core matrisome genes with pseudobulk matrices of 106 transcription factor genes – ordered and selected by their sample-specific variance and ranksum – and 106 randomly picked genes. After combining chicken and mouse data sets, we transformed the four pseudobulk count matrices into gene specificity indices to resolve cross-platform and -experiment differences [33]. We calculated pairwise Spearman’s rank correlation coefficients for all cell-type-specific pseudobulks and performed unsupervised hierarchical clustering based on Euclidean distances. Both ‘Transcriptome’ and ‘Random’ heatmaps were dominated by a strong ‘species signal’, *i.e.,* transcriptional signatures of pseudobulks coming from either chicken or mouse samples showed strong negative correlation to pseudobulks from the other species. Moreover, pseudobulk correlation within a given species was uniformly high, without much differentiation between distinct cell types or developmental stages (Fig. 5A, B). Contrary to that, ‘Transcription Factor’ and ‘Core Matrisome’ pseudobulks showed much less pronounced species differences. In fact, certain homologous cell types – like, e.g., those contributing to skin, vessels or blood – even showed strong positive pseudobulk correlation across species boundaries (Fig. 5C, D). This became even more evident when plotting the distribution of Spearman’s rank correlation coefficients for all four gene sets, separated by inter- *versus* intra-species comparisons. On average, ‘Transcription Factor’ and ‘Core Matrisome’ pseudobulks showed less negative correlations when contrasting chicken and mouse samples, than it was the case for ‘Transcriptome’ and ‘Random’ pseudobulks (Fig. 5E). Lastly, the ability to delineate a given cell type using Spearman’s ρ within a single species was also superior using ‘Transcription Factor’ and ‘Core Matrisome’ pseudobulks, compared to the ‘Transcriptome’ or ‘Random’ gene subsets. Particularly, while comparisons of the same cell type resulted in a high Spearman’s ρ regardless of the gene subset used, non-related cell type pairs within the same species still had mostly strong positive correlations with the ‘Transcriptome’ or ‘Random’ pseudobulks. ‘Transcription Factor’ and ‘Core Matrisome’ pseudobulks, however, were better at differentiating non-related cell types in the same species, which on average manifested itself in a lower – or even negative – Spearman’s ρ (Fig. 5A–D, F).

Collectively, using cell-type-specific pseudobulks of homologous cell types in two distantly related species, the chicken and the mouse, we showed that ‘Transcription Factor’ and ‘Core Matrisome’ gene subsets suffer less from a ‘species signal’ than either the entire transcriptome or randomly picked genes, and that they are better at delineating cell type-specific transcriptional signatures within a species.

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. Moreover, core matrisome gene expression signatures are able to identify homologous cell types across amniotes, and can help to better delineate distinct cell types within a given species. The predictive power of the core 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 [18]. 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 (Figs. 1 and 2). The lack of a clear ‘Core 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

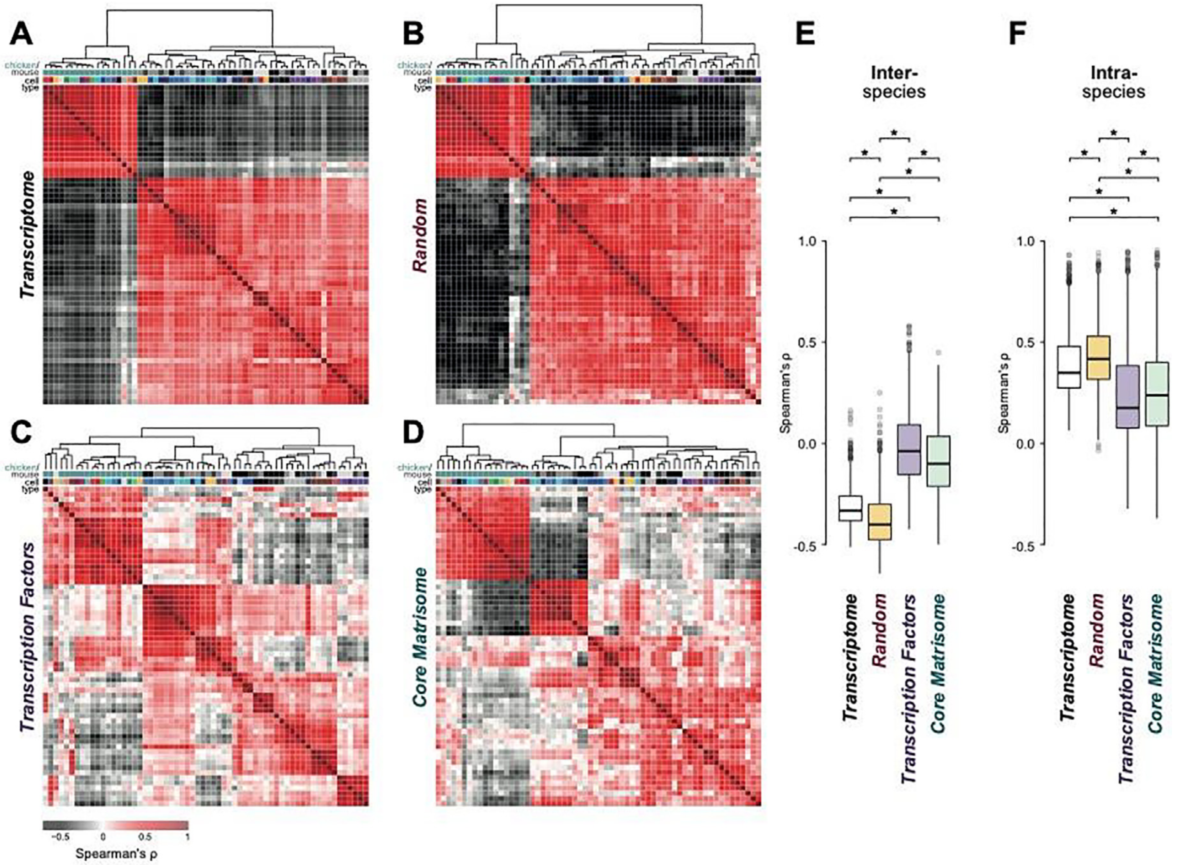

Fig. 5. Gene subset correlations of cell type-specific pseudobulks across species (A-D) Heatmaps of pairwise pseudobulk Spearman's rank correlations of 'Gold Standard' transcriptome clusters across both species and all time points. Pseudobulk expression is scaled by mean expression across all samples, the so-called gene specificity index (GSI) [33]. Pseudobulk of all expressed (in each sample, and species) orthologs (A, n = 7512), 'Core Matrisome' genes (B, n = 106), 'Transcription factors' ordered by variance and ranksum (C, n = 106), and a 'Random' gene subset (D, n = 106, matrisome genes and TFs excluded). The first color bar indicates species and stage of pseudobulk (green = chick, HH29; grey to black = mouse, E11.5 to E18.5), second bar the cell type (colors according to Fig. 1A and Fig. 4A). (E–F) Boxplots of inter-species (E) and intra-species (F) correlation values for all four heatmaps. Significance calculated by pairwise Wilcoxon rank sum test (*: *P* less than 0.05). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

type, each with its own putatively distinct 'matreotype'. Moreover, the overall predictive power of the core matrisome increases when comparing cell populations in ECM-rich tissues at progressively later stages of development (Fig. 4).

Previous work using scRNA-seq to determine molecular changes during adipogenesis (day 1–7) *in vitro* found that at day 3 the cell clustering was mainly driven by ECM genes, and at day 7 the core matrisome was one of the top ten most differentially expressed gene ontology terms [34]. Beyond development, planarian scRNA-seq revealed that muscles produce most of the matrisome, and inhibiting one key matrisome gene (hemicentin) resulted in severe epidermal ruffling and displacement of cells during homeostatic tissue turnover, suggesting an important role for tissue regeneration [28]. Furthermore, in healthy human lumbar discs, the core matrisome can be used to distinguish primary annulus fibrosus and nucleus pulposus cells based on 90 out of the 274 core matrisome genes being differentially expressed in the opposite direction using scRNA-seq [35]. Similarly, 115 matrisome genes are characteristically expressed in the six cell types that make up the human cutaneous neurofibroma microenvironment [36]. Beyond cell type distinction in tissues, a differential expression of matrisome genes can be observed when cells change their state from a healthy to a diseased cell. For instance, differential expression of matrisome genes was one main characteristic of reprogramming from normal fibroblasts, pericytes, and endothelial cells into tumor cells [37]. During cancer progression, deregulation of matrisome genes is a crucial step observed in early and late metastasis [38]. Moreover, core matrisome genes help identify circulating tumor cells in the blood using scRNA-seq [39,40]. These results support our conclusion that matrisome gene expression can serve as a key signature to determine individual cell types, as well as cell states.

Intriguingly, using a comparative evolutionary approach, *i.e.,* by correlating transcriptional pseudobulk signatures from both chicken and mouse, we found that matrisome signatures can even outperform the entire transcriptome at identifying cell types across species boundaries (Fig. 5). Namely, with the exception of blood, whole transcriptome pseudobulks showed a generally strong negative correlation between chicken and mouse samples. This effect was much less pronounced for matrisome pseudobulks and, importantly, several cell types even showed positive correlations to their homologous mouse counterparts. While this may seem counterintuitive at first glance, it most likely is a consequence of so-called concerted transcriptome evolution, which – over such extensive evolutionary timescales – can result in a pronounced 'species signal' at the level of the whole transcriptome [41]. In this scenario, gene regulatory interactions shared amongst all developing characters – or cell types, here – will change in unison, once mutations in these circuiteries occur. Accordingly, if stabilizing selection on the expression levels of a particular dominant gene subset (e.g., housekeeping genes) is weak or near-neutral, this may result in drift and the distortion of the overall transcriptome correlation between species, especially over long evolutionary distances [42–44]. This distortion effect can potentially manifest itself even within a single species, as evidenced by the overall higher correlation of the entire transcriptome between non-related cell types (Fig. 5A). Signatures of gene subsets whose expression levels are under stabilizing selection, however, will maintain their information content, even for comparisons between distantly related species [42,45]. Moreover, if directional selection was involved in shaping the signatures of these gene subsets in homologous cell or tissue types, this can result in a better separation of such pesubulks even within a single species. We observe this effect for both 'Transcription Factor' pseudbulks (Fig. 5C), in line with previous suggestions from others [2,4], as well as for the core matrisome (Fig. 5D).

This raises the question of why the core 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 [46]. However, it is less appreciated that each cell type can synthesize its own ECM that entails it with unique physical properties [17,47,48]. For instance, placing primary preadipocytes into decellularized ECMs derived from subcutaneous, visceral, or brown adipose tissue influences the preadipocytes' terminal differentiation [49]. 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 [50]. 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 [51,52]. Similarly, placing tumor cells into an embryonic ECM, reprograms them to non-tumorigenic cells [53]. Hence, there is 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 changes matrisome gene expression (*i.e.,* altering the matreotype), prevented tumor growth in xenograft mouse models [54]. 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 [55]. 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 [56]. Matreotypes reflecting chronological ages have been recently used to predict drugs that promote healthy aging [57]. 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 this regard, the fact that we observed evolutionary conserved matreotypes, in homologous cell types of mice and chicken highlights the potential of model organisms research. Identification of conserved orthologous ECM-based genotype-phenotype interactions might thus inform human biology or delineate novel ECM-related drug targets promoting healthy aging [57,58].

In summary, with our scRNA-seq analyses, we provide evidence for a previously postulated concept, namely that 'each cell type produces its unique ECM' [17,19]. 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, even across species boundaries.

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/; [16]. 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 [30] and the mouse embryonic limb [32]; 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 [30]. 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.

Dimensionality reduction

For all chicken and mouse seurat objects, principal components analysis was performed on all expressed genes, and significant components were selected as such, if they were located outside of the Marchenko Pastur distribution [59]. The same criterion for significance was applied on all principal component analysis on 'Core Matrisome', 'Transcritpion Factor' and 'Random' subset genes. The cells were visualized with the dimensionality reduction algorithm tSNE [60]. 'Core Matrisome', 'Transcription Factor' and 'Random' clusters for all datasets were represented on the same tSNEs generated from the 'Transcriptome' principal components. To define a 'Gold Standard' of scRNAseq-based cell type clustering, k-nearest neighbour (kNN) graphs and Jaccard indices of overlap between a cell and its neighbours were used to create shared nearest neighbour (SNN) graphs, with the Seurat 'FindNeighbors' function using all expressed 'Transcriptome' genes. Clusters of cells were then defined by 'FindClusters', by applying Louvain modularity optimization algorithms on SNN graphs. As the number of clusters can be influenced by the resolution parameters, please refer to the supplementary data for detailed parameters of significant dimensions and resolutions used in clustering for all samples. For 'Core Matrisome'-based clustering, all expressed core matrisome genes were considered for clustering. For 'Random'-based clustering, for ten iterations, randomly picked genes from the whole transcriptome were used such that they matched the number of expressed core matrisome genes, as well as resulting in the same number of individual clusters as the 'Transcriptome' and 'Core Matrisome' clustering. The core matrisome genes and transcription factors were excluded from the sampling. After ordering the transcription factors by expression variability, the set of top transcription factors matching the size of the expressed core matrisome was used to recluster the cells.

Cluster cell-type annotation

Differentially expressed genes between mouse clusters with a minimum natural log fold change of 0.25 were identified using a Wilcoxon rank sum test, and were then used to assign putative cell-type identities of each cluster. Only genes expressed in at least 25% of cells in one of the two populations were considered. For all clusters, all and the top five differentially genes per cluster can be found in the Data Source File 1. Chicken clusters had been previously annotated [30].

Distance Boxplots

To assess cluster-to-cluster proximity of 'Transcriptome'-, 'Core Matrisome'-, 'Random'-, and 'Transcription Factor'-based clustering approaches, Euclidean pairwise distances between each cluster were calculated on the averaged scaled expression per cluster of the top 2000 variably expressed genes. The same 2000 genes were used to compare all three clustering approaches.

Hypergeometric test

Probabilities of overlap between clusters were calculated with 'phyper'. The hypergeometric test takes into account the size of the reference cluster ('Transcriptome') m, the size of the test cluster ('Core Matrisome', 'Random', and 'Transcription Factor') 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' [61] was used to plot square root negative log 10 transformed probabilities.

Visualizing cluster contributions

Barplots were created with 'ggplot2' [62].

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 represents 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 'core matrisome' and 'transcription factors' against 'random' subsets of increasing size

The information content of the 'Core 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 remaining E11.5 and E18.5 chicken transcriptomes (*i.e.*, with transcription factors and the core matrisome removed). For each percentage, five 'Random' subsets resulting in the same number of clusters as the 'Core Matrisome' were sampled, clustered, and indices were calculated. The core matrisome genes and transcription factors were excluded from this sampling.

Spearman's rank correlation of 'Transcriptome'-based cluster pseudobulk expression

'Transcriptome'-based cluster pseudobulk expression for each of the four gene sets ('Transcriptome', 'Random', 'Transcription Factor', 'Core Matrisome') was calculated with Seurat's 'AverageExpression' function using default settings. The transcriptome gene set consists of all one-to-one orthologs that were expressed in all samples (n = 7512). The 'Core Matrisome' subset of those orthologs consisted of 106 genes. Expressed and orthologous transcription factors were ranked for each sample by variance. Then, an increasing rank sum was used to order them across samples and the top 106 TFs were selected. A 'Random' subset of 106 genes was selected from all expressed orthologs, excluding matrisome genes and transcription factors. After fusing all sample pseudobulks into a table, each gene's pseudobulk expression was divided by its average pseudobulk expression across samples and clusters to calculate the gene specificity index (GSI) [33]. Then pairwise Spearman's rank correlation coefficients for all GSI pseudobulks were calculated and plotted as a heatmap using pheatmap (Kolde (2019) https://CRAN.R-project.org/package=pheatmap), clustering rows and columns with hierarchical clustering based on euclidean distances.

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.

Acknowledgments

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mbplus.2021.100069.

Received 11 March 2021;
Accepted 6 May 2021;
Available online 21 May 2021

Keywords:
Matrisome;
Matreotype;
Cell type;
scRNA-seq;
Limb development

† Present address: Helmholtz Institute for pharmaceutical sciences Campus E8 1, 66123 Saarbrücken, Germany.

Abbreviations:
ADAMTS, a disintegrin and metalloproteinase with thrombospondin motifs; AS, aortic valve stenosis; BMP, bone morphogenetic protein; CVD, cardiovascular disease; CKD, chronic kidney disease; CP, C-propeptide; CUB, complement, Uegf, BMP-1; DMD, Duchenne muscular dystrophy; ECM, extracellular matrix; EGF, epidermal growth factor; eGFR, estimated glomerular filtration rate; ELISA, enzyme-linked immunosorbent assay; HDL, high-density lipoprotein; HSC, hepatic stellate cell; HTS, hypertrophic scar; IPF, idiopathic pulmonary fibrosis; LDL, low-density lipoprotein; MI, myocardial infarction; MMP, matrix metalloproteinase; mTLD, mammalian tolloid; mTLL, mammalian tolloid-like; NASH, nonalcoholic steatohepatitis; NTR, netrin; PABPN1, poly(A)-binding protein nuclear 1; OPMD, oculopharyngeal muscular dystrophy; PCP, procollagen C-proteinase; PCPE, procollagen C-proteinase enhancer; PNP, procollagen N-proteinase; SPC, subtilisin proprotein convertase; TIMP, tissue inhibitor of metalloproteinases; TGF-β, transforming growth-factor β; TSPN, thrombospondin-like N-terminal

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]. McKinley, K.L., Castillo-Azofeifa, D., Klein, O.D., (2020). Tools and concepts for interrogating and defining cellular identity. *Cell Stem Cell*, **26**, 632–656. https://doi.org/10.1016/j.stem.2020.03.015.

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

[5]. Crow, M., Gillis, J., (2019). Single cell RNA-sequencing: replicability of cell types. *Curr. Opin. Neurobiol.*, **56**, 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]. Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., Szalai, B., Saez-Rodriguez, J., (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.*, **21**, 36. https://doi.org/10.1186/s13059-020-1949-z.

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

[9]. Hynes, R.O., (2009). The extracellular matrix: not just pretty fibrils. *Science*, **326**, 1216–1219. https://doi.org/10.1126/science:1176009.

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

[11]. Bonnans, C., Chou, J., Werb, Z., (2014). Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biology.*, **15** (12), 786–801. https://doi.org/10.1038/nrm3904.

[12]. Neill, T., Kapoor, A., Xie, C., Buraschi, S., Iozzo, R.V., (2021). A functional outside-in signaling network of proteoglycans and matrix molecules regulating autophagy. *Matrix Biol.*,. https://doi.org/10.1016/j.matbio.2021.04.001.

[13]. Iozzo, R.V., Theocharis, A.D., Neill, T., Karamanos, N.K., (2020). Complexity of matrix phenotypes. *Matrix Biol. Plus*, **6–7**, 100038. https://doi.org/10.1016/j.mbplus.2020.100038.

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

[15]. 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. Proteom.: MCP. 11 (2012) M111.014647. https://doi.org/10.1074/mcp.m111.014647.

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

[17]. Frantz, C., Stewart, K.M., Weaver, V.M., (2010). The extracellular matrix at a glance. *J. Cell Sci.*, **123**, 4195–4200. https://doi.org/10.1242/jcs.023820.

[18]. Walma, D.A.C., Yamada, K.M., (2020). The extracellular matrix in development. *Development.*, **147**, dev175596. https://doi.org/10.1242/dev.175596.

[19]. Ewald, C., (2020). The matrisome during aging and longevity: a systems-level approach toward defining

matreotypes promoting healthy aging. *Gerontology*, **66**, 266–274. https://doi.org/10.1159/000504295.

[20]. Chapman, M.A., Mukund, K., Subramaniam, S., Brenner, D., Lieber, R.L., (2017). Three distinct cell populations express extracellular matrix proteins and increase in number during skeletal muscle fibrosis. *Am. J. Physiol.-Cell Ph.*, **312**, C131–C143. https://doi.org/10.1152/ajpcell.00226.2016.

[21]. Hiebert, P., Wietecha, M.S., Cangkrama, M., Haertel, E., Mavrogonatou, E., Stumpe, M., Steenbock, H., Grossi, S., Beer, H.-D., Angel, P., Brinckmann, J., Kletsas, D., Dengjel, J., Werner, S., (2018). Nrf2-mediated fibroblast reprogramming drives cellular senescence by targeting the matrisome. *Dev. Cell*, **46**, 145–161.e10. https://doi.org/10.1016/j.devcel.2018.06.012.

[22]. I.N. Taha, A. Naba, Exploring the extracellular matrix in health and disease using proteomics, Essays Biochem. 63 (2019) 417–432. https://doi.org/10.1042/ebc20190001.

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

[24]. Socovich, A.M., Naba, A., (2019). The cancer matrisome: From comprehensive characterization to biomarker discovery. *Semin. Cell Dev. Biol.*, **89**, 157–166. https://doi.org/10.1016/j.semcdb.2018.06.005.

[25]. Nauroy,P.,Hughes,S.,Naba,A.,Ruggiero,F.,(2018).Thein-silico zebrafish matrisome: A new tool to study extracellular matrix gene and protein functions. *Matrix Biol.*, **65**, 5–13. https://doi.org/10.1016/j.matbio.2017.07.001.

[26]. Teuscher, A.C., Jongsma, E., Davis, M.N., Statzer, C., Gebauer, J.M., Naba, A., Ewald, C.Y., (2019). The in-silico characterization of the Caenorhabditis elegans matrisome and proposal of a novel collagen classification. *Matrix Biol. Plus*,, 1–13. https://doi.org/10.1016/j.mbplus.2018.11.001.

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

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

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

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

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

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

[33]. Tosches, M.A., Yamawaki, T.M., Naumann, R.K., Jacobi, A.A., Tushev, G., Laurent, G., (2018). Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science*, **360**, eaar4237. https://doi.org/10.1126/science:aar4237.

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

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

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

[37]. Sathe, A., Grimes, S.M., Lau, B.T., Chen, J., Suarez, C., Huang, R.J., Poultsides, G., Ji, H.P., (2020). Single-cell genomic characterization reveals the cellular reprogramming of the gastric tumor microenvironment. *Clin. Cancer Res.*, **26**, 2640–2653. https://doi.org/10.1158/1078-0432.CCR-19-3231.

[38]. 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.

[39]. Ting, D., Wittner, B., Ligorio, M., Vincent Jordan, N., Shah, A., Miyamoto, D., Aceto, N., Bersani, F., Brannigan, B., Xega, K., Ciciliano, J., Zhu, H., MacKenzie, O., Trautwein, J., Arora, K., Shahid, M., Ellis, H., Qu, N.a., Bardeesy, N., Rivera, M., Deshpande, V., Ferrone, C., Kapur, R., Ramaswamy, S., Shioda, T., Toner, M., Maheswaran, S., Haber, D., (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.*, **8**, 1905–1918. https://doi.org/10.1016/j.celrep.2014.08.029.

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

[41]. Musser, J.M., Wagner, G.P., (2015). Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called “species signal”. *J. Exp. Zool. Part B: Mol. Dev. Evol.*, **324**, 588–604. https://doi.org/10.1002/jez.b.22636.

[42]. Chen, J., Swofford, R., Johnson, J., Cummings, B.B., Rogel, N., Lindblad-Toh, K., Haerty, W., Palma, F.d., Regev, A., (2019). A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.*, **29**, 53–63. https://doi.org/10.1101/gr.237636.118.

[43]. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., Pääbo, S., David Botstein, (2004). A neutral model of

transcriptome evolution. *PLoS Biol.*, **2**, e132. https://doi.org/10.1371/journal.pbio.0020132.

[44]. Tschopp, P., Tabin, C.J., (2017). Deep homology in the age of next-generation sequencing. *Philos. Trans. R. Soc. B Biol. Sci.*, **372**, 20150475. https://doi.org/10.1098/rstb.2015.0475.

[45]. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., Kaessmann, H., (2011). The evolution of gene expression levels in mammalian organs. *Nature*, **478** (369), 343–348. https://doi.org/10.1038/nature10532.

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

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

[48]. Yue, B., (2014). Biology of the extracellular matrix. *J. Glaucoma*, **23**, S20–S23. https://doi.org/10.1097/ijg.0000000000000108.

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

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

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

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

[53]. Hendrix, M.J.C., Seftor, E.A., Seftor, R.E.B., Kasemeier-Kulesa, J., Kulesa, P.M., Postovit, L.-M., (2007). Reprogramming metastatic tumour cells with embryonic microenvironments. *Nat. Rev. Cancer*, **7** (4), 246–255. https://doi.org/10.1038/nrc2108.

[54]. Honselmann, K.C., Finetti, P., Birnbaum, D.J., Monsalve, C.S., Wellner, U.F., Begg, S.K.S., Nakagawa, A., Hank, T., Li, A., Goldsworthy, M.A., Sharma, H., Bertucci, F., Birnbaum, D., Tai, E., Ligorio, M., Ting, D.T., Schilling, O., Biniossek, M.L., Bronsert, P., Ferrone, C.R., Keck, T., Mino-Kenudson, M., Lillemoe, K.D., Warshaw, A.L., Fernández-del Castillo, C., Liss, A.S., (2020). Neoplastic-stromal cell cross-talk regulates matrisome expression in pancreatic cancer. *Mol. Cancer Res.*, **18**, 1889–1902. https://doi.org/10.1158/1541-7786.MCR-20-0439.

[55]. Yuzhalin, A.E., Urbonas, T., Silva, M.A., Muschel, R.J., Gordon-Weeks, A.N., (2018). A core matrisome gene signature predicts cancer outcome. *Br. J. Cancer*, **118**, 435–440. https://doi.org/10.1038/bjc.2017.458.

[56]. Lim, S.B., Tan, S.J., Lim, W.-T., Lim, C.T., (2017). An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. *Nat. Commun.*, **8**, 1734. https://doi.org/10.1038/s41467-017-01430-6.

[57]. C. Statzer, E. Jongsma, S.X. Liu, A. Dakhovnik, F. Wandrey, P. Mozharovskyi, F. Zülli, C.Y. Ewald, Youthful and age-related matreotypes predict drugs promoting longevity, BioRxiv. (2021). https://doi.org/https://doi.org/10.1101/2021.01.26.428242.

[58]. Statzer, C., Ewald, C.Y., (2020). The extracellular matrix phenome across species. *Matrix Biol. Plus*, **8**, 100039. https://doi.org/10.1016/j.mbplus.2020.100039.

[59]. Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., McCarroll, S.A., Cepko, C.L., Regev, A., Sanes, J.R., (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, **166**, 1308–1323.e30. https://doi.org/10.1016/j.cell.2016.07.054.

[60]. L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. (2008) 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html.

[61]. Zhao, S., Guo, Y., Sheng, Q., Shyr, Y., (2014). Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinf.*, **15** (Suppl 10), P16. https://doi.org/10.1186/1471-2105-15-S10-P16.

[62]. C. Ginestet, ggplot2: Elegant graphics for data analysis: book reviews, J. R. Stat. Soc. Ser. Stat. Soc. 174 (2011) 245–246. https://doi.org/10.1111/j.1467-985x.2010.00676_9.x.