Single-cell analysis reveals the transformation of adipose-derived stromal cells into COL11A1-expressing cancer-associated fibroblasts

Kaiyi Zhu[1,2,3,5], Lingyi Cai[1,2,3,5], Chenqian Cui[2], Dimitris Anastassiou[1,2,3,4,*].

[1] Department of Systems Biology, Columbia University, New York, NY, USA.

[2] Department of Electrical Engineering, Columbia University, New York, NY, USA.

[3] Center for Cancer Systems Therapeutics, Columbia University, New York, NY, USA.

[4] Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA.

[5] These authors contributed equally to this work.

* Corresponding author: Dimitris Anastassiou.

Email: d.anastassiou@columbia.edu

Abstract

During the last ten years, many research results have been referring to a particular type of cancer-associated fibroblasts associated with invasiveness, metastasis and resistance to therapy, present in nearly identical form in many individual types of solid cancer. These fibroblasts are characterized by the expression of several genes, prominent among them COL11A1, THBS2, and INHBA, and their presence suggests the initiation of metastasis. Identifying the origin of this universal fibroblastic cell population and the underlying biological mechanisms responsible for their creation may help towards the identification of drug targets for pan-cancer metastasis-inhibiting therapeutics. This information, however, remains elusive. We have performed an extensive computational analysis of single-cell gene expression data from many cancer types, concluding that these fibroblasts are produced by a cancer-associated transformation of adipose-derived stromal cells naturally occurring in the stromal vascular fraction of the adipose microenvironment. Focusing on a rich pancreatic cancer dataset, we also provide a detailed description of the continuous modification of the gene expression profile of the fibroblastic population as they transition from APOD-expressing adipose-derived stromal cells to COL11A1-expressing cancer-associated fibroblasts, identifying the key genes that participate in this transformation.

Significance Statement

Cancer cells are known to often acquire increased invasiveness and metastatic potential in the presence of adipose tissue, but the underlying mechanisms are unclear. Using single-cell analysis and systems biology methods applied on biopsy samples, we show that there is a pan-cancer transformation of stromal cells naturally occurring in normal adipose tissue, into a particular type of cancer-associated fibroblasts expressing collagen COL11A1, at which point metastasis is imminent. This transformation is part of a universal biological process in cancer in

which malignant cells interact with the adipose microenvironment. Our work identifies details of this continuous transformation, including participating genes, with potential applications in developing pan-cancer metastasis-inhibiting therapeutics targeting such genes.

Introduction

This research is a follow-up of our previous work in 2010 (1), in which we had identified a cancer stage-associated signature by comparing the gene expression of lower stage with that of higher stage cancer samples. The signature was remarkably similar across cancer types and would only be present after a particular staging threshold, different in each cancer type, was reached. The genes of the signature, with prominent presence of collagen COL11A1, activin A (INHBA), and thrombospondin-2 (THBS2), would never be overexpressed in samples below that cancer-type-specific threshold, but they were overexpressed in a subset of profiled samples above that threshold. For example, the signature only appeared in ovarian cancer samples of at least stage III; in colon cancer samples of at least stage II; and in breast cancer samples of at least invasive stage I (but never in samples from carcinoma in situ).

We had observed the striking consistency of the signature across cancer types, which was obvious at that time from bulk microarray data. For example, **Table 1** shows the top 15 genes ranked in terms of fold change for three different cancer types (breast (2), ovarian (3), pancreatic (4)) using data provided in papers published independently in 2006, 2007, 2008, respectively. The breast cancer data compare invasive ductal carcinoma vs. ductal carcinoma in situ (supplementary data 3, “up in IDC” of the paper (2)); the ovarian cancer data compare metastatic tissue in the omentum with primary tumor (table 2 of the paper (3)); and the pancreatic data compare whole tumor tissue vs. normal pancreatic tissue (table 1 of the paper (4)). The four genes COL11A1, INHBA, THBS2, COL5A2 appear among the top 15 in all three sets ($P = 6x10^{-23}$ by hypergeometric test (5)). The actual P value is much lower than that, because, in addition to the above overlap, ten additional genes (COL10A1, COL1A1, COL5A1, FAP, FBN1, FN1, LOX, MFAP5, POSTN, SULF1) appear among the top 15 in at least two of the three sets (and are highly ranked in all three sets anyway). This demonstrates that the signature is well-defined and associated with an important universal biological mechanism in cancer. We had also found (1) that gene COL11A1 serves as a proxy of the full signature, in the sense that it is the only gene

from which all other genes of the signature are consistently top-ranked when listed in terms of the correlation of their expression with that of COL11A1.

Accordingly, we had identified at that time a COL11A1-correlated pan-cancer gene signature of activated fibroblasts, listed in table 4 of reference (1), which we deposited in GSEA. We had referred to those cancer associated fibroblasts (CAFs) as MAFs ("metastasis-associated fibroblasts"), because their presence at the invasive edge of the tumor suggests that metastasis is imminent. To avoid any inaccurate interpretation of the term as implying that such fibroblasts are markers of metastasis that has occurred already, here we refer to them as "COL11A1-expressing CAFs."

In this work we show, using techniques from systems biology, that these fibroblasts are adipose-derived, which explains the above stage association as resulting from the interaction of cancer cells with the adipose microenvironment: The underlying biological process occurs, under some circumstances, as part of an interaction when tumor cells encounter adipose microenvironment. Indeed, adipose tissue is encountered when ovarian cancer cells reach the omentum (stage III); after colon cancer has grown outside the colon (stage II); and in breast cancer from the beginning of the spread (stage I, but not in situ stage 0). Furthermore, this provides an explanation to the well-known fact that adipose tissue contributes to the development of cancer as a result of interaction with malignant cells (6–13).

Since then, numerous research results have been published regarding COL11A1 in many individual types of solid cancer, noting its association with poor prognosis, metastasis, and resistance to therapy (14–20). However, the origin and underlying mechanisms responsible for the generation of those pan-cancer fibroblasts have remained unclear. Single-cell technology provides a unique opportunity to further scrutinize the evolving expression of each gene in each individual fibroblast. In this follow-up work, we perform data analysis of single-cell RNA-seq

(scRNA-seq) data, demonstrating that there is a continuous transition in a path that starts from a particular type of adipose-derived stromal cells (ASCs) and leads to the COL11A-expressing CAFs, to which we refer as the “ASC to COL11A1-expressing CAF transformation.”

We show that the ASC cell population from which this transformation originates expresses a unique characteristic signature, which is rich in both fibroblastic markers such as LUM, DCN, and certain collagens, as well as adipose-related genes, such as APOD and PTGDS, and has stem-like properties, as evidenced by gene set enrichment analysis. We also show that this population is naturally present in the stromal vascular fraction (SVF) of normal adipose tissue. Some of those cells are gradually transformed as a result of interactions of tumor cells with the adipose microenvironment, so that they start expressing genes such as RARRES1 and SFRP4, eventually obtaining an activated fibroblastic phenotype expressing COL11A1, INHBA and THBS2.

Our computational analysis includes two different methods (**Materials and Methods**), validating each other: We applied attractor analysis (21, 22) in the various datasets, identifying the gene expression profile of the dominant fibroblastic population in individual samples. We also applied trajectory inference analysis using the Slingshot algorithm (23). We used publicly available scRNA-seq data sets (**Materials and Methods**) from five different cancer types. We gave particular emphasis on one of those datasets (24) from pancreatic ductal adenocarcinoma (PDAC), because it is particularly rich, containing gene expression profiles of 57,530 pancreatic cells taken from 24 PDAC tumor samples (T1-T24) and 11 normal control pancreatic samples (N1-N11). For example, using attractor analysis, we established that the dominant fibroblastic cell population in each of the eleven normal samples was of the ASC type described above. We also analyzed the attractor signatures of each of the tumor datasets, producing an observed continuous gene expression change of the dominant fibroblastic cell population as the biological mechanism proceeds. Using trajectory inference analysis, we validated the above results by

producing trajectories starting from the point representing the normal ASCs, passing through intermediate states, eventually reaching the final state of COL11A1-expressing CAFs.

Finally, we validated our results in other cancer types (head and neck, ovarian, lung, breast), suggesting that the ASC to COL11A1-expressing CAF transformation, involving the same participating genes, is present in all solid cancer types.

Results

Pancreatic Ductal Adenocarcinoma dataset (attractor method)

The PDAC dataset (24) consists of 57,530 single-cell RNA-seq profiles from 24 PDAC tumor samples (T1-T24) and 11 normal control samples (N1-N11). We first identified the mesenchymal cells in all these samples (**Materials and Methods**) resulting in 6,267 fibroblasts and 5,342 stellate cells.

To find the expression profile of the dominant fibroblastic population present in each sample, we applied the unsupervised attractor algorithm (21) on the set of mesenchymal cells for each sample (**Materials and Methods**). The algorithm is iterative and designed to converge to the core ("heart") of a co-expression signature, referred to as an "attractor metagene," and has been previously used successfully for identifying features useful for breast cancer prognosis (22, 25). Each attractor metagene is defined by a ranked set of genes along with their scores defining their corresponding strengths within the signature, therefore the top ranked genes are the most representative of the signature. Briefly, a "seed" gene is used for the first iteration of the algorithm and each subsequent iteration results in a metagene (a hypothetical gene whose expression levels are a weighted average of mutually associated genes), converging to an attractor metagene. If there exists a strong co-expression signature involving several genes including the seed gene, then the algorithm will converge to that signature. In that case, any among the

top-ranked genes of an attractor metagene will yield precisely the same attractor metagene if used as seed. In our case, we used the general fibroblastic gene LUM as seed, and indeed LUM remained ranked among the top positions in the converged results. We found that all samples yielded strong coexpression signatures involving many genes corresponding to their dominant fibroblastic population, with strong overlap among them.

Dataset S1 shows the top 100 genes for each of the 34 samples (11 normal and 23 tumor samples, excluding sample T20 as it did not contain identified fibroblasts). Genes LUM, DCN, FBLN1, MMP2, SFRP2 and COL1A2 appear in at least 33 out of the 34 samples, revealing a strong similarity shared by all those fibroblastic expression profiles. This strong overlap is consistent with the continuous transformation process, as described below.

Dominant fibroblastic population in the normal pancreatic samples is adipose-derived: There was a remarkable similarity among the attractor profiles of the eleven normal pancreatic samples which demonstrates that they represent a gene expression pattern of a stable and uniform, normally occurring cell population. Specifically, there are 18 genes shared among the top 50 genes in the attractors of at least ten of the eleven normal samples, listed by average rank: LUM, APOD, PTGDS, C7, DCN, DPT, SFRP2, FBLN1, CCDC80, FBLN5, C1S, CXCL12, C1R, PDGFRA, CFD, COL6A3, MGP, C3 ($P < 10^{-7}$ by permutation test). In addition to the canonical fibroblastic markers, there are several strongly expressed adipose-related or stemness-related genes in the list (such as apolipoprotein D (APOD), adipsin (CFD), stromal cell-derived factor 1 (CXCL12) (26), dermatopontin (DPT), and PDGFRA (27, 28), revealing that these are adipose-derived stromal/stem cells. Consistently, Gene Set Enrichment Analysis (GSEA) (29) of these 18 genes identified most significant enrichment (FDR q value = 5.38×10^{-29}) in the "BOQUEST_STEM_CELL_UP" dataset of genes upregulated in stromal stem cells from adipose tissue versus the non-stem counterparts (30). **Table 2** demonstrates the striking similarity among

the eleven normal sample attractor metagenes by listing the 20 top ranked genes for each sample and color coding for genes LUM, DCN, APOD, C7, and PTGDS.

To demonstrate that the ASC population is naturally occurring in the normal adipose tissue of all types, we referred to recent results from single-cell analysis of general human adipose tissue (31). We found that one of the identified clusters (P5, table s2 of the paper (31)) has eight genes with zero adjusted *P* value. Six among those eight genes are LUM, DCN, APOD, MGP, PTGDS, DPT, all of which are among the top-ranked genes in all eleven attractor signatures above. In addition to cluster P5, there were also additional identified clusters with similar composition, suggesting imperfect separation of cell subpopulations.

Therefore, to precisely identify the ASC subpopulation, we used the single-cell expression profiles of all 26,350 cells taken from the SVF of normal adipose tissue from the 25 samples of the dataset. The first column in Dataset S2 contains the top 30 ranked genes in the corresponding attractor signature. The second column contains the top 30 ranked genes of the "consensus attractor" (**Materials and Methods**) of the 11 normal pancreatic samples, representing the main state of the normal fibroblastic population. There are 14 overlapping genes in the two lists ($P = 10^{-37}$), and most of the non-highlighted genes in each column are still ranked highly in the other column. This extreme similarity of the two gene expression profiles indicates that they correspond to the same cell population. In other words

Dominant fibroblastic population in the tumor PDAC samples exhibits a continuous transition from the normal adipose-derived stromal cells to COL11A1-expressing metastasis associated fibroblasts: While retaining the same top-ranked fibroblastic marker genes, the attractor metagenes for the PDAC tumor samples have different rankings for other genes, and several new genes appear in the list (Dataset S1). These changes can be seen as representing a continuous transition of the gene expression profile of a cell type that is gradually

differentiated. Consistently, the attractor of one of the normal samples (N10) contains RARRES1 and THBS2, which are among the genes that appear in the attractors of nearly all tumor samples (see below). This suggests that these fibroblasts are in the initial stage of differentiation, and therefore sample N10, despite being diagnosed as normal, belongs to a patient in which some malignant cells have started interacting with the adipose microenvironment.

Specifically, in the tumor samples the rankings of the adipose related genes (such as APOD, PTGDS, C7, CFD, DPT) eventually dropped, while many genes of the final CAF signature (such as COL11A1, INHBA, THBS2, COL10A1) rose to the top (consistent with our subsequent validation using Trajectory Inference Analysis, as shown in Figs 1c,1d, 1e, referring to the same genes). The reduction of the adipose-related gene expression became more prominent in the strongly COL11A1-expressing samples, suggesting that these samples represent the final stage of differentiation.

Therefore, to investigate the continuous transition of ASCs to COL11A1-expressing CAFs, we partitioned the 34 pancreatic samples into three groups, considering whether the corresponding attractor includes COL11A1 among the top 100 genes. Group1 includes the eleven normal samples (N1 to N11). Group 2 includes twelve tumor samples not expressing COL11A1 (T1, T2, T3, T5, T10, T12, T13, T14, T16, T18, T19, T24). Group 3 includes eleven cancer samples expressing COL11A1 (T4, T6, T7, T8, T9, T11, T15, T17, T21, T22, T23). We then applied the consensus version of attractor finding algorithm (**Materials and Methods**) and identified the "consensus attractors" representing the main state of the fibroblasts for each of the above three sample groups, so we can observe the genes whose ranking changes significantly during the transformation (**Table 3**).

As expected from the previous analysis of all individual samples, there are many common genes demonstrating the continuity of the transition based on the similarity among the fibroblasts in all

groups (e.g., LUM, DCN, FBLN1, APOD, SFRP2 and COL6A3 appear among the top 20 genes in all three groups). On the other hand, the groups have distinct gene rankings. Group 1 (normal samples) contains many adipose-related genes, as mentioned previously. Group 3 contains, in addition to COL11A1, many among the other CAF genes, such as THBS2, INHBA, MMP11, AEBP1, MFAP5, COL10A1, VCAN and FAP, apparently representing the desmoplastic stroma present in pancreatic tumors.

Because COL11A1 can be seen as a proxy of the final stage of the signature (1), we also ranked all genes in terms of their association with COL11A1 in each of the Group 3 patients using the measure of mutual information (**Materials and Methods**). Dataset S3 shows the top 20 genes for each of them, in which we observe that genes MFAP5 and COL10A1 appear in all of them, in other words their expression further increases, along with COL11A1, during the final stage.

Group 2 displays an intermediate stage, including many markers of both ASCs as well as CAFs. To find the most unique and representative gene(s) for Group 2, we focused on the top 20 genes of its attractor and calculated the ranking increment when compared with the rankings of the same genes in the attractors of the other two groups. We considered a ranking increase of greater than 50 as significant. As a result, we found that RARRES1 is the only gene satisfying the criterion in both comparisons.

Gene RARRES1 (aka TIG1) appears in the beginning of the ASC to COL11A1-expressing CAF transformation process and in all samples of Group 2. This is consistent with the known fact that it plays an important, though not yet clear, role in regulating proliferation and differentiation of adipose tissue-derived mesenchymal stem cells (32). On the other hand, the expression of RARRES1 is reduced in Group 3, disappearing completely in the final stages of differentiation

when COL11A1 is expressed strongly (consistent with suggestions that it is a tumor suppressor (33, 34)).

Differential expression analysis on fibroblasts: We also performed differential expression (DE) analysis on the classified fibroblasts between the three sample groups defined above (**Materials and Methods**; see complete DE gene lists in Dataset S4). Importantly, the results of such analysis represent the full population of fibroblasts and not necessarily the expression changes in the particular cells undergoing the ASC to COL11A1-expressing CAF transformation.

For the 'Group1 to Group 2' stage, genes CFD and DPT are the most downregulated genes, consistent with the downregulation of adipose-related genes. On the other hand, the top up regulated gene is phospholipase A2 group IIA (PLA2G2A), whose encoded protein is a member of a family of enzymes catalyzing the hydrolysis of phospholipids into free fatty acid. At the same time, other extracellular matrix (ECM) genes are also strongly up regulated, as expected.

For the 'Group 2 to Group 3' stage, remarkably PLA2G2A again shows as the top DE gene, however this time it is down-regulated. Similarly, the tumor suppressor RARRES1, which was significantly up-regulated in the early stage becomes down-regulated in this stage. On the other hand, the most significantly up regulated genes, following the top gene COL11A1, were matrix metalloproteinase (MMP11), collagens of type X and XII (COL10A1 and COL12A1), thrombospondin 2 (THBS2), growth factor binding protein 3 (IGFBP3), while CFD, APOD and C7 continued decreasing their expression levels. Thus, this stage represents the final formation of COL11A1 expressing CAFs.

The top differentially expressed gene PLA2G2A is not among the top genes of any attractors we identified and is expressed by less than half of cells, even in Group 2 in which it appears. Therefore, it is expressed by cells in Group 2 that are not among the ones undergoing the ASC to

COL11A1-expressing CAF transformation, although it probably still plays an important related role (**Discussion**) and many previous studies referred to its effects on prognosis of multiple cancer types (35–37).

Pancreatic Ductal Adenocarcinoma dataset (Trajectory Inference)

We independently applied trajectory inference (TI) analysis on the PDAC fibroblasts by using the Slingshot (23) method, to examine the cell states. We first performed unsupervised clustering on the identified fibroblasts (**Materials and Methods**), resulting in four subgroups X1, X2, X3, X4 (*SI Appendix*, Fig. S1a) with the top differentially expressed genes shown in *SI Appendix*, Fig. S1b. One of these clusters (X4) was discarded from further TI analysis, because it mainly expressed the IL1 CAF marker HAS1 (38), which is not expressed by either ASCs or COL11A1-expressing CAFs (and does not appear at all in Dataset S1), and contained only 3% of fibroblasts resulting almost exclusively from patient T11 (*SI Appendix*, Fig. S1c).

As seen from the list of top differentially expressed genes, X1 contains CAF genes top ranked (including MMP11, COL11A1, THBS2, INHBA), X2 has RARRES1 at the top, and X3 has ASC genes top ranked, including DPT, C7, CXCL12 and CFD. Consistently, Figs 1a and 1b show the single trajectory path resulting from TI analysis, where X3 is the starting point and X1 is the end point of the trajectory, while X2 (highly expressing RARRES1), is an intermediate point, thus validating the continuous ASC to COL11A1-expressing CAF transformation.

Dataset S5 shows the top 100 genes with zero *P* value, ranked by their variances, resulting from pseudotime-based differential gene expression analysis (**Materials and Methods**). We can clearly identify as top-ranked several ASC genes, as well as CAF genes, while some general fibroblastic markers, such as DCN, are missing, consistent with the continuity of the ASC to

COL11A1-expressing CAF transformation. We then used a generalized additive model (GAM) fit to pseudotime ordered expression data to visualize the trend of gene expressions (Fig. 1c).

There was a prominent difference between adipose-related genes and COL11A1-associated genes. The expression of the adipose-related genes steadily fell across the process (Fig. 1d), while, conversely, the expression of COL11A1-associated genes gradually increased (Fig. 1e). There is a significant negative correlation between these two groups of genes, e.g., COL11A1 (the last among those genes to increase its expression) was exclusively overexpressed in the mature CAFs, which did not express C7. Of particular interest, genes RARRES1 and SFRP4 (Fig. 1f) increased consistently at the beginning and then decreased after reaching a peak, suggesting that they play important roles in the differentiation path.

Validation in other cancer types

Next, we validated the ASC to COL11A1-expressing CAF transformation hypothesis in other solid cancer types. Although we could not find currently available datasets as rich as the PDAC dataset, we selected those containing a large (at least 100) number of fibroblasts and separately analyzed each of them, obtaining consistent results. Specifically, we used four scRNA-seq datasets from head and neck squamous cell carcinoma (HNSCC) (39), ovarian cancer (40), lung cancer (41) and breast cancer (42).

The COL11A1-expressing CAF signature has been confirmed to be a pan-cancer signature (16, 43–45) (**Table 1**). Therefore, the most important validation task would be to confirm the existence of the APOD/CFD/CXCL12/MGP/PTGDS-expressing ASCs as the starting point of the transformation, and to also confirm that some samples are at an intermediate stage, expressing

genes such as RARRES1, SFRP4 and THBS2, in addition to the core ASC genes, demonstrating that they are at an intermediate stage of the transformation.

Head and neck squamous cell carcinoma: For the HNSCC dataset, the authors of the paper presenting the data (39) reported that the cancer-associated fibroblasts in the dataset can be partitioned into two subsets, which they name CAF1 and CAF2. In table s5 of that paper, the top three differentially expressed genes of the CAF2 group are CFD, APOD and CXCL12, while the full gene list for CAF2 presented in the same table s5 also includes genes MGP, C3, C7, DPT, PTGDS. This strongly suggests that the partitioning used in the paper was influenced by the presence of an ASC cell subpopulation, identical, or at least very similar to, those discovered in the PDAC. Similarly, the list of differentially expressed genes for CAF1 in table s5 includes genes INHBA, THBS2, CTHRC1, POSTN, MMP11, COL5A2, COL12A1, suggesting that the identified CAF1 subpopulation was influenced by the presence of differentiated CAFs, which would eventually express COL11A1. Finally, gene RARRES1 also appears among the list of CAF2 genes, suggesting that it was captured among cells which had started the process of ASC to COL11A1-expressing CAF transformation.

In our independent analysis, we performed clustering identifying 1,026 fibroblasts from all available cells (Fig. 2a; **Materials and Methods**). There were two fibroblastic clusters (X7 and X9) expressing CAF-associated genes (COL11A1, COL12A1, MMP11, INHBA, THBS2, COL10A1, COL8A1, FN1) and ASC associated genes (APOD, C7, PTGDS), respectively (Dataset S6), which confirmed the presence of these two populations in HNSCC.

Among the individual patients, we found that the most prominent case is sample HNSCC28, which contains a rich set of cells undergoing differentiation. Applying the attractor finding algorithm on the fibroblasts of that sample (Dataset S7) resulted in genes LUM, APOD, COL6A3, PDGFRA, DCN, and CFD being among the top-ranked, revealing that it represents an ASC

population. Furthermore, the presence of genes THBS2, MFAP5 and VCAN in the same attractor reveals that these cells have already started undergoing the transformation.

Ovarian cancer: For the ovarian dataset, the clustering results showed two clusters (X6 and X9) expressing COL11A1-associated genes and ASC-associated genes, respectively (Fig. 2b, Dataset S8; **Materials and Methods**). Among the individual patients, we found that the ones validating our hypotheses most are HG2F and LG2, both of whose datasets, consistently, contain cells from the fatty omental tissue. Dataset S7 includes the corresponding two attractors identified in the cells of each patient. Among the top ranked genes for HG2F are DCN, LUM, C1S, C7, and C3, but also RARRES1, suggesting that they represent fibroblasts undergoing the transformation, while the LG2-based attractor contains highly ranked all three genes COL11A1, INHBA, THBS2.

Lung cancer: The dataset contains a large number (> 50,000) of cells, but we only classified ~2% (=1,346) among them as mesenchymal cells, including fibroblasts and pericytes (**Materials and Methods**). Among those cells, there were two fibroblastic clusters (X1 and X2) expressing related genes (COL11A1, COL12A1, MMP11, INHBA, THBS2, COL10A1, COL8A1, FN1) and ASC related genes (APOD, C7, PTGDS), respectively (Fig. 2c, Dataset S9). The presence of the transformation is evident by the attractors identified in the mesenchymal cells for patients 4 and 3 (Dataset S7). The former prominently contains genes CFD, PTGDS and C7, while the latter contains THBS2, COL10A1 and INHBA.

Breast cancer: The size of the breast cancer dataset is small (~1,500 cells in total), and 169 cells among them were classified as mesenchymal (**Materials and Methods**). By further clustering these cells, we identified ASCs (X1) and COL11A1-expressing CAFs (X3) (Fig. 2d, Dataset S10). ASC related genes (APOD, MFAP4, CFD) were identified in X1, while CAF-related genes (COL10A1, COL11A1, MMP11, INHBA, FN1, THBS2, AEBP1, COL12A1) are among the top 15 of X3. Patients PT089 and PT039 contain the highest proportions (>50%) of the ASC and

COL11A1-expressing CAF subpopulations, respectively, and we found consistent results in their attractors (Dataset S7), as the former contains C1S, C1R, CXCL12, PTGDS, C3, while the latter contains THBS2, COL11A1, COL10A1, at top-ranked positions.

Discussion

Our work sheds light on the mechanisms by which adipose tissue contributes to the development of cancer as a result of interaction with malignant cells (6–13). Using techniques from systems biology, we have established that there is a cancer-associated transformation starting from cells belonging to a population of adipose-derived stromal cells co-expressing both fibroblastic genes such as LUM and DCN, as well as adipose-related genes such as APOD, PTGDS, CFD and C7. The final end-point of this transformation is another cell population expressing a precise pan-cancer signature (shown in **Table 1**) of COL11A1/INHBA/THBS2 expressing fibroblasts, which we had previously (1) identified, resulting only after malignant cells have encountered adipose tissue. We also provide a detailed description of the gradual modification of the gene expression profile, concluding that the cells undergoing this transformation exhibit a transient overexpression of some genes, such as RARRES1 and SFRP4.

This transformation is part of a complex biological process of interaction with the adipose tumor microenvironment resulting in extracellular matrix remodeling. This process is naturally accompanied by modifications in other cell populations, including tumor cells, pericytes as well as other types of fibroblasts. For example, tumor cells become increasingly invasive in the process, as they undergo epithelial-to-mesenchymal transition (EMT). It appears that TGF-beta signaling plays an important role in the transformation, as we had speculated by the prominent presence of related genes INHBA and THBS2 (1), and as also proposed by various studies (14, 46, 47).

Our work is related to previous research efforts referring to cell populations named “adipose derived mesenchymal stem cells,” “adipose derived fibroblasts,” “cancer-associated adipocytes,” etc. For example, it has been found that adipose-derived fibroblasts (ADFs) promote tumor progression (48) following further de-differentiation of cancer-associated adipocytes (CAAs) (49). As shown in our results, however, the adipose-derived stromal cell population undergoing the transformation to the COL11A1-expression cancer-associated fibroblasts is naturally present in the SVF of normal adipose tissue.

Furthermore, there are many published results related to fibroblast heterogeneity in cancer, in most cases using computational clustering methods following dimensionality reduction, or choice of arbitrary marker genes (such as αSMA), each time leading to the designation of a number of mutually exclusive cell subpopulations. Although such subpopulations derived from different publications often have similarities with each other, they often have significant differences, leading to confusion. Each of those clustering results only represents an approximation of biological reality, sometimes mixing together or isolating subpopulations. For example, several previous analyses resulted in conflicting results for identified clusters having some similarities with the endpoints of the transformation, notably the C9 and C10 clusters in (45), the CAF2 and CAF1 clusters in (39), and the iCAF and myCAF clusters in (50, 51).

To address this issue and reach definitive conclusions, we used, in addition to the above methodology, a detailed gene association-based scrutiny of all our results, including numerous color-coded scatter plots (not included in the paper), observing and confirming the presence of cells expressing particular combinations of genes, rather than blindly accepting clustering results. We also used the attractor algorithm, which is specifically designed to converge to the core of the dominant expression profile of cell subpopulations. Our conclusions confirmed the stable presence of the two “endpoints” of what we call ASC to COL11A1-expressing CAF transformation.

The stable presence of the final pan-cancer COL11A1-expressing population had already been established in previous studies and illustrated in Table 1. The fact that gene COL11A1 uniquely serves as the best proxy the pan-cancer signature (1) further indicates that such cells represent the final endpoint of the transformation. Interestingly, we found that αSMA is not necessarily significantly expressed in all such cells, which indicates that it should not be used as a marker gene for that signature.

In this paper, we provide evidence of the fact that the ASC subpopulation serving as the initial endpoint of the transformation is stable and consists of stem-like adipose-derived stromal cells naturally occurring in the SVF of normal adipose tissue (including the peripancreatic adipose tissue). We believe that our paper is the first one that identifies the precise profile signature of this important cell population, which, as mentioned previously, has a big overlap in GSEA with the BOQUEST_STEM_CELL_UP set of genes up-regulated in freshly isolated stromal cells from adipose tissue versus the non-stem counterparts. This ASC population in the adipose microenvironment represents the starting point of an interaction with tumor cells, leading to the transformation of some ASCs into a particular type of cancer-associated fibroblasts, as part of an important pan-cancer biological mechanism. Indeed, we only found this ASC population in tissues that are inside (as in the omentum and the breast), or in direct proximity with, adipose tissue, but not in other tissues, such as in datasets from normal colon and ovary, consistent with the fact that the COL11A1-expressing CAF signature is never encountered in stage I colon cancer patients, or in lower than stage III ovarian cancer patients.

We also found that PLA2G2A is the single most differentially overexpressed gene in some fibroblast population (though not in those undergoing the transformation itself) in the intermediate stage of the process, and then it is also the most differentially underexpressed at the final stage. This gene is a phospholipase catalyzing the hydrolysis of phospholipids into fatty acids. On the

other hand, it has been recognized that fatty acid oxidation (FAO) is associated with the final COL11A1-expressing stage of the transformation (52). These results suggest that lipid metabolic reprogramming plays an important role in the metastasis-associated biological mechanism (53). We hope that our results will give rise to testable hypotheses that could eventually lead to the development of pan-cancer metastasis-inhibiting therapeutics by targeting potential causal biological mechanisms or driver genes identified in this paper, involved in the ASC to COL11A1-expressing CAF transformation. For example, although RARRES1 is considered to be a tumor suppressor (33, 34), and this is consistent with its silencing at the final stage, its increased expression is actually associated with the initiation of the transformation of the ASCs. Similarly, another gene whose expression is increased at the initial stage is SFRP4, which is a Wnt pathway regulator whose expression has been found associated with various cancer types (54–57).

Materials and Methods

Datasets availability

The processed count matrix of the PDAC dataset (24) was downloaded from the Genome Sequence Archive with accession number CRA001160. The four validation datasets of other cancer types are also publicly available: HNSCC (39) (GSE103322), ovarian (40) (GSE118828), lung cancer (41) (E-MTAB-6149 and E-MTAB-6653), breast cancer (42) (GSE118389). We excluded samples from lymph nodes. The numbers of patients included in these datasets are 35 (PDAC), 18 (HNSCC), 9 (ovarian), 5 (lung), and 6 (breast).

Data processing and cell identification

We applied the *Seurat* (v3.1.4) (58) R package to process the gene expression matrix and characterize the cell type identity for each scRNA-seq dataset. The count matrix was normalized and log transformed by using the *NormalizeData* function. We selected the 2,000 most variable

features and then performed principal component analysis (PCA) followed by applying an unsupervised graph-based clustering approach. We used default parameter settings in all the above steps except that the resolution parameter in the *FindCluster* function is set to 1.0 to increase the granularity of downstream clustering. To identify differentially expressed genes for each cluster, we used the *FindMarkers* function. To characterize the identity of mesenchymal cells in each dataset, we made use of the expression of known markers: LUM, DCN, COL1A1 for fibroblasts, RGS5, ACTA2 for pericytes.

For the smaller-size datasets (ovarian, breast), we performed clustering once on all cells for mesenchymal cell identification. For datasets of larger size (PDAC, HNSCC, lung), we applied 'two-step clustering' to ensure accuracy: The first step was initial clustering within individual samples. Then we excluded samples with very few (< 20) detected fibroblasts and pooled the mesenchymal cells of the remaining samples together for a second clustering, which resulted in the final set of mesenchymal cells for the dataset. For the PDAC dataset, we included an additional step to remove low-quality cells, by retaining cells for which at least one of the corresponding markers had expression levels ≥ 3.

Mutual information

Mutual information (MI) is a general measure of the association between two random variables (59). We used a spline based estimator (60) to estimate MI values and normalized so the maximum possible value is 1. The details of the estimation method are described in the paper introducing the attractor algorithm (21). We used the *getMI* or *getAllMIWz* function implemented in the *cafr* R package with parameter *negateMI* = TRUE.

Attractor-based analysis

For single dataset, we applied the attractor finding algorithm using the *findAttractor* function implemented in the *cafr* (v0.312) R package (21) with gene LUM as seed. The exponent (*a*) was

set to different values for scRNA-seq datasets profiled from different protocols (61). For the analysis of UMI-based (e.g. 10x) and full-length-based (e.g. Smart-seq2) datasets, we used $a = 3$ and $a = 5$, respectively. Briefly, the attractor finding algorithm iteratively finds mutually associated genes from an expression matrix, converging to the core of the co-expression mechanism. The association measure used is the normalized mutual information (as described above), which captures the general relationships (including nonlinear effects) between variables. To find the consensus attractor for multiple datasets, we applied the consensus version of the attractor finding algorithm as described in (62). In the consensus version, the association measures between genes are evaluated as the weighted median of the corresponding measures taken from the individual datasets. The weights are proportional to the number of samples included in each individual dataset in log scale.

Trajectory inference (TI) analysis

We selected the *Slingshot* (23) method for TI analysis, based on its robustness and suggestions made by the benchmarking pipeline *dynverse* (63). We used the raw counts as input and followed the Slingshot lineage analysis workflow (v1.4.0). To begin this process, Slingshot chose robustly expressed genes if it has at least 10 cells that have at least 1 read for each. After gene filtering, we proceeded to full quantile normalization. Following diffusion map dimensionality reduction, Gaussian mixture modelling was performed to classify cells. The final step of lineage inference analysis used the slingshot wrapper function. A cluster based minimum spanning tree was subjected to describe the lineage. After analyzing the global lineage structure, we fitted a generalized additive model (GAM) for pseudotime and computed *P* values. Genes were ranked by *P* values and variances. After running Slingshot, we identified genes whose expression values significantly vary over the derived pseudotime by using a GAM, allowing us to detect non-linear patterns in gene expression.

Statistical analysis

P value evaluation by permutation test

The significance of the consistency of N different attractors of size M was evaluated as follows. We randomly selected M genes out of the 24,005 genes to generate a random gene set, and we created N such random gene sets. Each time, we counted the number of genes common in all N gene sets, and we repeated this process ten million times. In these ten million experiments, it never occurred that there is one or more gene common in all N gene sets. Therefore, the corresponding *P* value is less than 10^{-7}.

Differential expression analysis

We used a Wilcoxon Rank Sum test by applying the *FindMarkers* function in *Seurat* to identify the differentially expressed (DE) genes between fibroblasts of different groups. DE genes with |log fold-change| > 0.25 and Bonferroni adjusted *P* value < 0.1 are considered as significant. The positive and negative DE genes are ranked separately in terms of the absolute values of their log fold-change.

References

1. H. Kim, J. Watkinson, V. Varadan, D. Anastassiou, Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med Genomics* **3**, 51 (2010).
2. C. S. Schuetz, *et al.*, Progression-Specific Genes Identified by Expression Profiling of Matched Ductal Carcinomas In situ and Invasive Breast Tumors, Combining Laser Capture Microdissection and Oligonucleotide Microarray Analysis. *Cancer Res* **66**, 5278–5286 (2006).
3. E. Bignotti, *et al.*, Gene expression profile of ovarian serous papillary carcinomas: identification of metastasis-associated genes. *American Journal of Obstetrics & Gynecology* **196**, 245.e1-245.e11 (2007).
4. L. Badea, V. Herlea, S. O. Dima, T. Dumitrascu, I. Popescu, Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology* **55**, 2016–2027 (2008).
5. M. Wang, Y. Zhao, B. Zhang, Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* **5**, 16923 (2015).
6. E. Lengyel, L. Makowski, J. DiGiovanni, M. G. Kolonin, Cancer as a Matter of Fat: The Crosstalk between Adipose Tissue and Tumors. *Trends in Cancer* **4**, 374–384 (2018).
7. Y. Cao, Adipocyte and lipid metabolism in cancer drug resistance. *J Clin Invest* **129**, 3006–3017 (2019).
8. A. J. Cozzo, A. M. Fuller, L. Makowski, Contribution of Adipose Tissue to Development of Cancer. *Compr Physiol* **8**, 237–282 (2017).
9. Y.-X. Wang, *et al.*, Friend or foe: Multiple roles of adipose tissue in cancer formation and progression. *Journal of Cellular Physiology* **234**, 21436–21449 (2019).

10. F. Su, S. Ahn, A. Saha, J. DiGiovanni, M. G. Kolonin, Adipose stromal cell targeting suppresses prostate cancer epithelial-mesenchymal transition and chemoresistance. *Oncogene* **38**, 1979–1988 (2019).
11. C. Himbert, *et al.*, Signals from the Adipose Microenvironment and the Obesity–Cancer Link—A Systematic Review. *Cancer Prev Res* **10**, 494–506 (2017).
12. D. F. Quail, A. J. Dannenberg, The obese adipose tissue microenvironment in cancer development and progression. *Nature Reviews Endocrinology* **15**, 139–154 (2019).
13. K. Eckel-Mahan, A. Ribas Latre, M. G. Kolonin, Adipose Stromal Cell Expansion and Exhaustion: Mechanisms and Consequences. *Cells* **9**, 863 (2020).
14. Y.-H. Wu, T.-H. Chang, Y.-F. Huang, H.-D. Huang, C.-Y. Chou, COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene* **33**, 3432–3440 (2014).
15. C. García-Pravia, *et al.*, Overexpression of COL11A1 by Cancer-Associated Fibroblasts: Clinical Relevance of a Stromal Marker in Pancreatic Cancer. *PLOS ONE* **8**, e78327 (2013).
16. F. Vázquez-Villa, *et al.*, COL11A1/(pro)collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. *Tumor Biol.* **36**, 2213–2222 (2015).
17. A. Li, J. Li, J. Lin, W. Zhuo, J. Si, COL11A1 is overexpressed in gastric cancer tissues and regulates proliferation, migration and invasion of HGC-27 gastric cancer cells in vitro. *Oncol. Rep.* **37**, 333–340 (2017).
18. L. Shen, *et al.*, COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. *Oncol. Rep.* **36**, 877–885 (2016).
19. H. Fischer, R. Stenling, C. Rubio, A. Lindblom, Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis* **22**, 875–878 (2001).

20. L. Tang, Y. Lei, Y. Liu, B. Tang, S. Yang, The Expression of 7 Key Genes are Capable of Predicting Distant Metastasis of Colorectal Cancer to Liver or to Lung. *J Dig Dis* (2020) https:/doi.org/10.1111/1751-2980.12936.
21. W.-Y. Cheng, T.-H. Ou Yang, D. Anastassiou, Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput. Biol.* **9**, e1002920 (2013).
22. W.-Y. Cheng, T.-H. O. Yang, D. Anastassiou, Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. *Science Translational Medicine* **5**, 181ra50-181ra50 (2013).
23. K. Street, *et al.*, Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
24. J. Peng, *et al.*, Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Research* **29**, 725–738 (2019).
25. N. McCarthy, Rising to the challenge. *Nature Reviews Cancer* **13**, 378–378 (2013).
26. A. Saha, *et al.*, Proinflammatory CXCL12–CXCR4/CXCR7 Signaling Axis Drives Myc-Induced Prostate Cancer in Obese Mice. *Cancer Res* **77**, 5158–5168 (2017).
27. C. Dani, A. Pfeifer, The complexity of PDGFR signaling: regulation of adipose progenitor maintenance and adipocyte-myofibroblast transition. *Stem Cell Investig* **4** (2017).
28. C. Sun, W. L. Berry, L. E. Olson, PDGFRα controls the balance of stromal and adipogenic cells during adipose tissue organogenesis. *Development* **144**, 83–94 (2017).
29. A. Liberzon, *et al.*, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
30. A. C. Boquest, *et al.*, Isolation and Transcription Profiling of Purified Uncultured Human Stromal Stem Cells: Alteration of Gene Expression after In Vitro Cell Culture. *Mol Biol Cell* **16**, 1131–1141 (2005).
31. J. Vijay, *et al.*, Single-cell analysis of human adipose tissue identifies depot- and disease-specific cell types. *Nature Metabolism* **2**, 97–109 (2020).

32. S. Ohnishi, *et al.*, Involvement of tazarotene-induced gene 1 in proliferation and differentiation of human adipose tissue-derived mesenchymal stem cells. *Cell Prolif* **42**, 309–316 (2009).
33. C. Jing, *et al.*, Tazarotene-induced gene 1 (TIG1) expression in prostate carcinomas and its relationship to tumorigenicity. *J. Natl. Cancer Inst.* **94**, 482–490 (2002).
34. E. E. Oldridge, *et al.*, Retinoic acid represses invasion and stem cell phenotype by induction of the metastasis suppressors RARRES1 and LXN. *Oncogenesis* **2**, e45–e45 (2013).
35. M. Kashiwagi, *et al.*, Group II and IV phospholipase A2 are produced in human pancreatic cancer cells and influence prognosis. *Gut* **45**, 605–612 (1999).
36. A. Buhmeida, *et al.*, PLA2 (group IIA phospholipase A2) as a prognostic determinant in stage II colorectal carcinoma. *Annals of Oncology* **20**, 1230–1235 (2009).
37. H. Cai, *et al.*, Elevated Phospholipase A2 Activities in Plasma Samples from Multiple Cancers. *PLOS ONE* **8**, e57081 (2013).
38. C. X. Dominguez, *et al.*, Single-cell RNA sequencing reveals stromal evolution into LRRC15+ myofibroblasts as a determinant of patient response to cancer immunotherapy. *Cancer Discov* (2019) https:/doi.org/10.1158/2159-8290.CD-19-0644 (May 26, 2020).
39. S. V. Puram, *et al.*, Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
40. A. J. Shih, *et al.*, Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. *PLOS ONE* **13**, e0206785 (2018).
41. D. Lambrechts, *et al.*, Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine* **24**, 1277–1289 (2018).
42. M. Karaayvaz, *et al.*, Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications* **9**, 3588 (2018).
43. D. Jia, *et al.*, A COL11A1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets. *Cancer Lett.* **382**, 203–214 (2016).

44. Q. Wan, *et al.*, BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database (Oxford)* **2015** (2015).
45. J. Qian, *et al.*, A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Research*, 1–18 (2020).
46. J. A. Galván, *et al.*, Validation of COL11A1/procollagen 11A1 expression in TGF-β1-activated immortalised human mesenchymal cells and in stromal cells of human colon adenocarcinoma. *BMC Cancer* **14**, 867 (2014).
47. D.-J. Cheon, *et al.*, A collagen-remodeling gene signature regulated by TGFβ signaling is associated with metastasis and poor survival in serous ovarian cancer. *Clin Cancer Res* **20**, 711–723 (2014).
48. L. Bochet, *et al.*, Adipocyte-Derived Fibroblasts Promote Tumor Progression and Contribute to the Desmoplastic Reaction in Breast Cancer. *Cancer Res* **73**, 5657–5668 (2013).
49. B. Dirat, *et al.*, Cancer-Associated Adipocytes Exhibit an Activated Phenotype and Contribute to Breast Cancer Invasion. *Cancer Res* **71**, 2455–2465 (2011).
50. D. Öhlund, *et al.*, Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. *J Exp Med* **214**, 579–596 (2017).
51. E. Elyada, *et al.*, Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer Discov* **9**, 1102–1123 (2019).
52. S. Nallanthighal, *et al.*, Inhibition of collagen XI alpha 1-induced fatty acid oxidation triggers apoptotic cell death in cisplatin-resistant ovarian cancer. *Cell Death & Disease* **11**, 1–12 (2020).
53. N. Koundouros, G. Poulogiannis, Reprogramming of fatty acid metabolism in cancer. *British Journal of Cancer* **122**, 4–22 (2020).

54. T. Constantinou, *et al.*, SFRP-4 abrogates Wnt-3a-induced β-catenin and Akt/PKB signalling and reverses a Wnt-3a-imposed inhibition of in vitro mammary differentiation. *J Mol Signal* **3**, 10 (2008).
55. E. Sandsmark, *et al.*, SFRP4 gene expression is increased in aggressive prostate cancer. *Scientific Reports* **7**, 14276 (2017).
56. D. Huang, *et al.*, SFRP4 was overexpressed in colorectal carcinoma. *J. Cancer Res. Clin. Oncol.* **136**, 395–401 (2010).
57. C. E. Ford, *et al.*, The Wnt Gatekeeper SFRP4 Modulates EMT, Cell Migration and Downstream Wnt Signalling in Serous Ovarian Cancer Cells. *PLOS ONE* **8**, e54362 (2013).
58. T. Stuart, *et al.*, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
59. J. A. T. Thomas M. Cover, *Elements Of Information Theory 2nd Ed* (2006) (February 19, 2020).
60. C. O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* **5**, 118 (2004).
61. C. Ziegenhain, *et al.*, Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631-643.e4 (2017).
62. K. Zhu, T.-H. Ou Yang, V. Dorie, T. Zheng, D. Anastassiou, Meta-analysis of expression and methylation signatures indicates a stress-related epigenetic mechanism in multiple neuropsychiatric disorders. *Translational Psychiatry* **9**, 1–12 (2019).
63. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37**, 547–554 (2019).

Figures and Tables

Figure 1

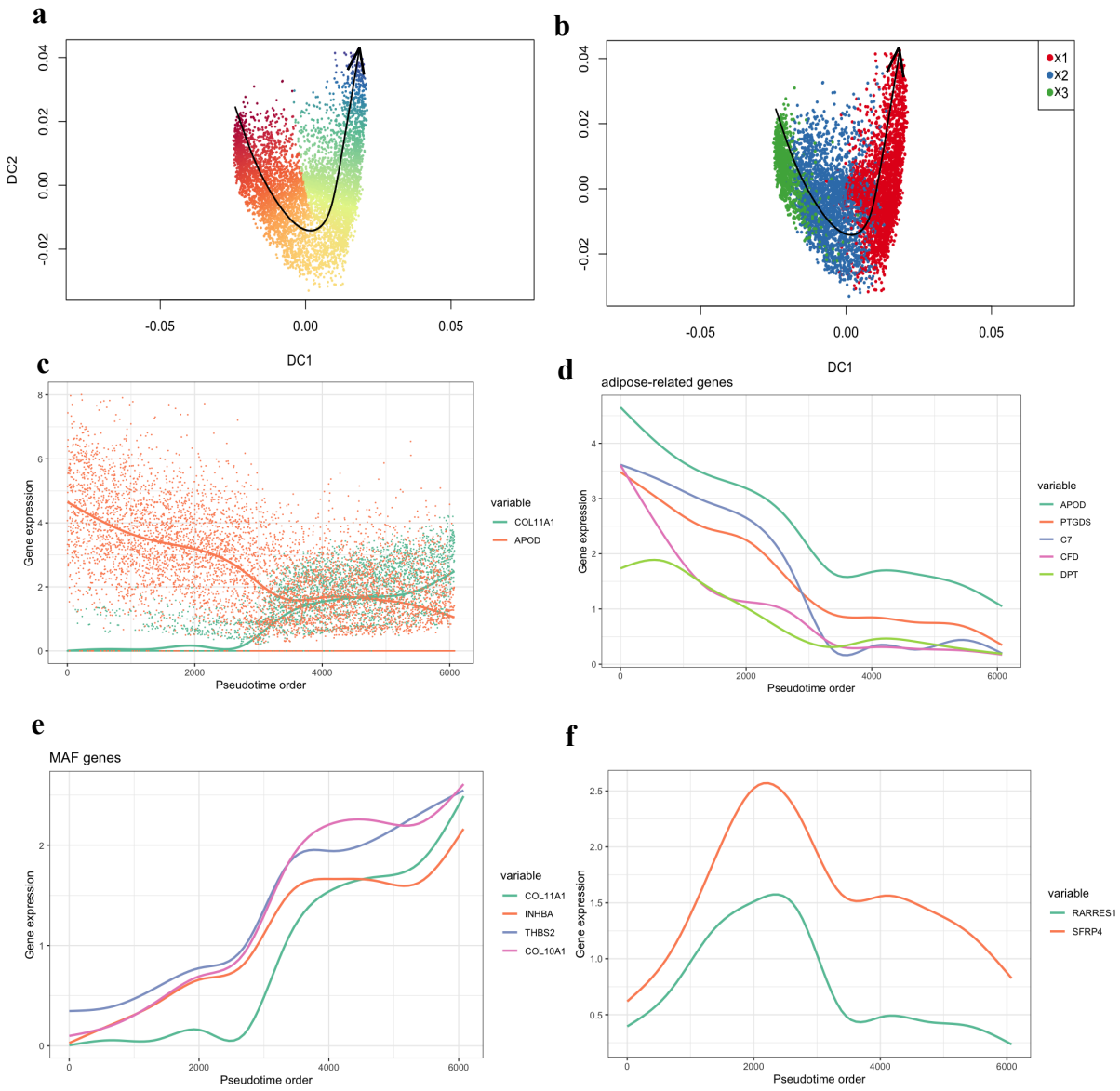

Figure 1. Trajectory analysis of PDAC. (a) Trajectory analysis of 6,075 fibroblasts. Colors-coded for pseudotime changing, red presenting the beginning of differentiation and blue presenting the end. (b) Color-coded trajectory analysis of fibroblasts for annotated three clusters. (c) GAM fit to pseudotime-ordered expression data to visualize the trend of gene expressions. (d) Expression of adipose-related genes along the transformation lineage. The x axis shows the cell orders and the y axis shows the normalized read count. (e) Expression of COL11A1-associate genes along the transformation lineage. (f) Expression of RARRES1 and SFRP4 genes along the transformation lineage.

Figure 2

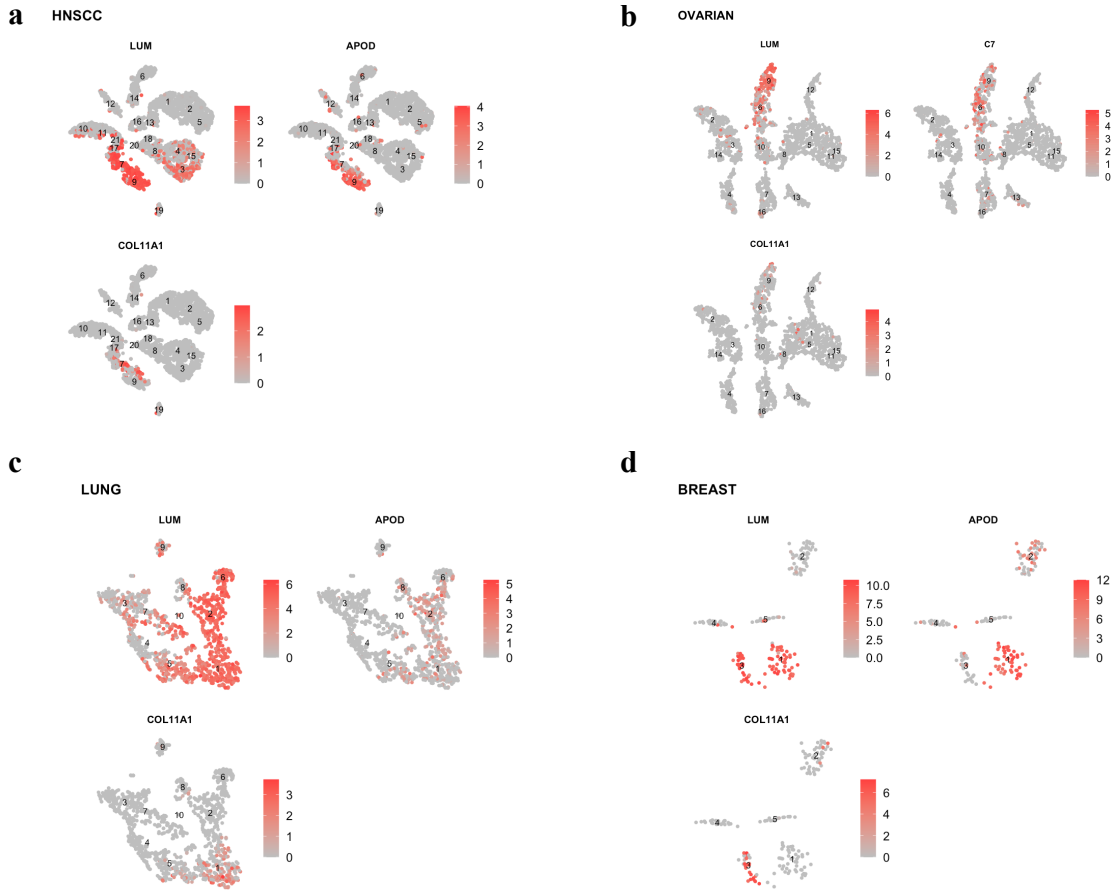

Figure 2. Unsupervised clustering of four datasets from HNSCC, ovarian cancer, lung cancer and breast cancer. (a) t-SNE embedding of the whole HNSCC dataset. (b) t-SNE embedding of the whole ovarian cancer dataset. (c) t-SNE embedding of the mesenchymal cells from lung cancer dataset. (d) t-SNE embedding of the mesenchymal cells from breast cancer dataset.

Table 1. Top 15 ranked genes in terms of fold change (FC) for three different cancer types revealing the metastasis-associated fibroblastic signature. Shown are the rankings, reported by the authors, for breast, ovarian and pancreatic cancers. We eliminated multiple entries of the same gene (keeping the one that appears first) and dashes. Genes shared in all three cancer types are highlighted in green, while genes appearing twice are highlighted in yellow. Paste table below the title and we kept the original scales.

	Breast		**Ovarian**		**Pancreatic**	
Rank	Gene	FC	Gene	FC	Gene	Log2FC
1	COL11A1	6.50	COL11A1	8.23	INHBA	5.15
2	COL10A1	4.07	COL1A1	5.67	COL10A1	5.00
3	MFAP5	3.73	TIMP3	5.52	POSTN	4.92
4	LRRC15	3.61	FN1	5.40	SULF1	4.63
5	INHBA	3.44	INHBA	4.94	COL8A1	4.60
6	FBN1	3.43	EFEMP1	4.86	COL11A1	4.40
7	SULF1	3.35	DSPG3	4.36	CTHRC1	4.38
8	GREM1	3.35	COL5A2	4.07	COL1A1	4.12
9	COL5A2	3.22	LOX	4.03	THBS2	3.97
10	LOX	3.22	MFAP5	4.01	HNT	3.90
11	COL5A1	3.08	POSTN	3.97	CSPG2	3.87
12	THBS2	2.99	COL5A1	3.95	WISP1	3.80
13	LAMB1	2.97	THBS2	3.91	FN1	3.69
14	FAP	2.96	FBN1	3.90	COMP	3.53
15	SPOCK	2.91	FAP	3.84	COL5A2	3.38

Table 2. Top 20 genes of the identified attractors for each PDAC normal sample (N1-N11). Commonly shared genes including LUM, DCN, APOD, C7 and PTGDS are color-coded. Paste table below the title.

Rank	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
1	DCN	LUM	LUM	C7	APOD	LUM	PTGDS	C7	DCN	MMP2	LUM
2	LUM	DCN	FBLN1	FBLN5	DPT	DCN	APOD	LUM	LUM	APOD	DCN
3	C7	C7	C7	LUM	FBLN5	FBLN1	LUM	DCN	C7	LUM	FBLN1
4	FBLN1	FBLN1	PTGDS	DCN	PDGFRA	ADH1B	FBLN1	APOD	FBLN1	EFEMP1	SFRP2
5	MGP	APOD	C1S	APOD	CXCL12	DPT	C7	FBLN1	APOD	CTSK	CFD
6	C1S	MGP	DPT	PTGDS	LUM	ABCA8	ADH1B	SFRP2	SFRP2	SFRP2	APOD
7	CCDC80	C1S	PDGFRA	FBLN1	COL6A3	C3	DPT	PTGDS	SERPINF1	PLTP	MGP
8	PTGDS	DPT	APOD	C1R	PTGDS	APOD	COL6A3	CCDC80	PTGDS	MGST1	SERPINF1
9	DPT	CCDC80	SFRP2	DPT	C7	MMP2	EFEMP1	FBLN5	GSN	LSP1	CCDC80
10	C1R	PTGDS	DCN	SRPX	CCDC80	C1S	PDGFRA	C1S	C1S	FBLN1	C3
11	APOD	FBLN5	CXCL12	FMO2	CFD	C7	CXCL12	CXCL12	SEPP1	SPON2	ADH1B
12	SEPP1	SEPP1	C1R	SEPP1	MRC2	PTGDS	SCN7A	C3	CCDC80	PTGDS	PTGDS
13	FBLN5	COL1A2	COL6A3	CXCL12	FGF7	SFRP2	MMP2	CFD	DPT	SVEP1	C7
14	CXCL12	SFRP2	ADH1B	CYR61	SFRP2	FBLN5	MEG3	C1R	OLFML3	CXCL12	C1S
15	EFEMP1	SRPX	SPON2	SFRP2	MARCKS	C1R	C1S	MGP	FBLN5	SCN7A	CST3
16	COL1A2	SERPINF1	CFD	CLEC11A	LRP1	CXCL12	OLFML3	CFH	C1R	COL6A3	C1R
17	SFRP2	OLFML3	LAMA2	PDGFRA	FMO2	CST3	SVEP1	COL6A3	PTN	CCDC80	CXCL14
18	ALDH1A1	CST3	C3	NR2F1	NR2F1	MGP	DCN	SRPX	MGP	COLEC11	MMP2
19	CFD	MEG3	FBLN5	C1S	TNXB	CCDC80	SFRP2	EFEMP1	ALDH1A1	PDGFRA	GPNMB
20	COL6A3	C1R	ABCA8	ABCA8	DCN	MRC2	MRC2	SEPP1	PDGFRA	HBP1	S100A4

Table 3. Top 50 genes of the consensus attractors for three different PDAC sample groups. Group1: normal samples; Group 2: tumor samples not expressing COL11A1; Group 3: tumor samples expressing COL11A1.

Rank	Group1	Group2	Group3	Rank (cont'd)	Group1	Group2	Group3
1	LUM	LUM	COL1A1	**26**	ABCA8	PLXDC2	NTM
2	DCN	DCN	COL10A1	**27**	COL1A2	COL1A1	NBL1
3	FBLN1	APOD	COL1A2	**28**	LAMB1	OGN	ITGBL1
4	C7	SFRP4	CTHRC1	**29**	SVEP1	C1S	FN1
5	APOD	SFRP2	SFRP2	**30**	MEG3	PTGDS	MFAP2
6	PTGDS	VCAN	COL3A1	**31**	PCOLCE	ITGBL1	RARRES2
7	SFRP2	FBLN1	MMP2	**32**	NEGR1	TSHZ2	GJB2
8	C1S	MMP2	LUM	**33**	FGF7	PODN	MXRA5
9	CCDC80	PDGFRA	VCAN	**34**	LRP1	SERPINF1	LOX
10	MGP	EFEMP1	THBS2	**35**	CYR61	OMD	MEG3
11	DPT	RARRES1	COL6A3	**36**	ALDH1A1	COL10A1	MRC2
12	CXCL12	FBLN5	DCN	**37**	EFEMP1	HSD11B1	HTRA1
13	C1R	ISLR	COL11A1	**38**	FMO2	IGF1	PDPN
14	FBLN5	COL6A3	APOD	**39**	CFH	RP11-572C15.6	FBLN2
15	C3	C3	SFRP4	**40**	PLTP	LTBP2	C1QTNF3
16	PDGFRA	CCDC80	COL5A2	**41**	SLIT2	FBLN2	TMSB10
17	SRPX	CTSK	FAP	**42**	MRC2	COL3A1	POSTN
18	COL6A3	CTHRC1	FBLN1	**43**	SCN7A	STEAP1	CDH11
19	ADH1B	CYP1B1	COL8A1	**44**	LAMA2	LXN	MMP14
20	CFD	MEG3	COL5A1	**45**	TIMP1	BOC	CCDC80
21	OLFML3	COL1A2	AEBP1	**46**	COL6A2	MRC2	SULF1
22	SERPINF1	MOXD1	MFAP5	**47**	CXCL14	LRP1	TMEM119
23	MMP2	C7	INHBA	**48**	LTBP4	THBS2	CXCL14
24	CST3	MGP	CTSK	**49**	EMP1	MFAP4	FNDC1
25	SEPP1	COL8A1	ISLR	**50**	GSN	FAP	IGFBP3