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

Kaiyi Zhu1,2,3☯, Lingyi Cai1,2,3☯, Chenqian Cui2, Juan R. de los Toyos4, Dimitris Anastassiou1,2,3,5*

1 Department of Systems Biology, Columbia University, New York, New York, United States of America, 2 Department of Electrical Engineering, Columbia University, New York, New York, United States of America, 3 Center for Cancer Systems Therapeutics, Columbia University, New York, New York, United States of America, 4 Immunology Department, School of Medicine and Health Sciences, University of Oviedo, Oviedo, Spain, 5 Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York, United States of America

☯ These authors contributed equally to this work.
• Current address: Icahn School of Medicine at Mount Sinai, New York, New York, United States of America
* 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 poor prognosis, invasiveness, metastasis and resistance to therapy in multiple cancer types, characterized by a gene expression signature with prominent presence of genes COL11A1, THBS2 and INHBA. Identifying the underlying biological mechanisms responsible for their creation may facilitate the discovery of targets for potential pan-cancer therapeutics. Using a novel computational approach for single-cell gene expression data analysis identifying the dominant cell populations in a sequence of samples from patients at various stages, we conclude that these fibroblasts are produced by a pan-cancer cellular transition originating from a particular type of adipose-derived stromal cells naturally present in the stromal vascular fraction of normal adipose tissue, having a characteristic gene expression signature. Focusing on a rich pancreatic cancer dataset, we provide a detailed description of the continuous modification of the gene expression profiles of cells as they transition from APOD-expressing adipose-derived stromal cells to COL11A1-expressing cancer-associated fibroblasts, identifying the key genes that participate in this transition. These results also provide an explanation to the well-known fact that the adipose microenvironment contributes to cancer progression.
population for a particular cell type in the microenvironment of tumors, we observed that there is a remarkably continuous modification of this profile among patients, corresponding to a cellular transition. Specifically, we found that the starting point of this transition has a unique characteristic signature corresponding to cells that are naturally residing in normal adipose tissue. We also found that the endpoint of the transition has another characteristic signature corresponding to a particular type of cancer-associated fibroblasts with prominent expression of gene \textit{COL11A1}, which has been found strongly associated with invasiveness, metastasis and resistance to therapy in multiple cancer types. Our results provide an explanation to the well-known fact that the adipose tissue contributes to cancer progression, shedding light on the biological mechanism by which tumor cells interact with the adipose microenvironment. We provide a detailed description of the changing profile during the transition, identifying associated genes as potential targets for pan-cancer therapeutics inhibiting the underlying mechanism.

**Introduction**

This work investigates, using computational analysis of rich single-cell datasets from many patients, the nature and origin of a particular type of cancer-associated fibroblasts (CAFs) that has been found to be strongly associated with invasiveness, metastasis, resistance to therapy, and poor prognosis, in multiple cancer types. These fibroblasts can be identified by their characteristic signature with prominent presence of collagen \textit{COL11A1} and several other co-expressed genes such as \textit{THBS2} and \textit{INHBA}. There are indications that the generation of those CAFs is part of a universal biological process in cancer that plays essential roles in cancer progression. Therefore, the driving vision for this research has been that it may provide testable hypotheses for the development of pan-cancer therapeutics targeting the biological mechanisms responsible for the creation of those CAFs. As described below, to achieve this task we used both established techniques for studying the dynamic changes in gene expression of cells associated with lineages, such as trajectory inference, as well as complementary computational approaches with novel application in single-cell data analysis. These techniques allowed the precise identification of the expression profile of the origin of the underlying cellular transition as a particular cell type of adipose derived stromal/stem cells (ASCs). We also independently validated the presence of those ASCs as naturally occurring, by applying the same computational methods in other available datasets of normal adipose tissue. In the remaining part of this section we provide introductory information about the \textit{COL11A1}-expressing CAFs, explain the motivation for our choice of computational methods, and provide evidence for their advantages and unique capabilities analyzing the particular data sets that we used.

These CAFs were first identified in 2010 [1] by their cancer stage-associated signature. Specifically, a \textit{COL11A1}/\textit{INHBA}/\textit{THBS2}-expressing gene signature was found to be present only after a particular staging threshold, different in each cancer type, was reached. For example, it only appeared in ovarian cancer of at least stage III; in colon cancer of at least stage II; and in breast cancer of at least invasive stage I (but not in carcinoma in situ). We had observed the striking consistency of that 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. The breast cancer data compare invasive ductal carcinoma with 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 four genes \textit{COL11A1}, \textit{INHBA}, \textit{THBS2}, \textit{COL5A2} appear among the top 15 in all three sets ($P = 6 \times 10^{-23}$ by multi-set intersection test [5]). The actual $P$ value is much lower than that, because, in addition to the above overlap, ten additional genes (\textit{COL10A1}, \textit{COL1A1}, \textit{COL5A1}, \textit{FAP}, \textit{FBN1}, \textit{FN1}, \textit{LOX}, \textit{MFAP5}, \textit{POSTN}, \textit{SULF1}) appear among the top 15 in at least two of the three sets (and are highly ranked in all three sets anyway). This similarity demonstrates that the signature is well-defined and associated with a universal biological mechanism in cancer.

We had also found that gene \textit{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 in terms of the correlation of their expression with that of \textit{COL11A1}. Accordingly, we had identified a \textit{COL11A1}-correlated pan-cancer gene signature, listed in table 4 of [1], which we deposited in the Molecular Signatures Database (MSigDB). We had referred to those CAFs as MAFs (“metastasis-associated fibroblasts”), because their presence 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 “\textit{COL11A1}-expressing CAFs.”

Since then, many research results were published connecting one of the genes \textit{COL11A1}, \textit{INHBA}, \textit{THBS2} with poor prognosis, invasiveness, metastasis, or resistance to therapy, in various cancer types [6–15].

Furthermore, several designated tumor subtypes were identified in individual cancer types as a result of the presence of those pan-cancer CAFs. For example, the top 15 genes distinguishing the ovarian “mesenchymal subtype” according to [16] are \textit{POSTN}, \textit{COL11A1}, \textit{THBS2}, \textit{COL5A2}, \textit{ASPN}, \textit{FAP}, \textit{MMP13}, \textit{VCAN}, \textit{LUM}, \textit{COL10A1}, \textit{CTSK}, \textit{COMP}, \textit{CXCL14}, \textit{FABP4}, \textit{INHBA}. Similarly, the 24 characterizing genes of the “activated stroma subtype” of pancreatic cancer in Fig 2 of [17] are \textit{SPARC}, \textit{COL1A2}, \textit{COL3A1}, \textit{POSTN}, \textit{COL5A2}, \textit{COL1A1}, \textit{THBS2}, \textit{FN1}, \textit{COL10A1}, \textit{COL5A1}, \textit{SFRP2}, \textit{CDH11}, \textit{CTXRC1}, \textit{PNDC1}, \textit{SULF1}, \textit{FAP}, \textit{LUM}, \textit{COL11A1}, \textit{ITGA11}, \textit{MMP11}, \textit{INHBA}, \textit{VCAN}, \textit{GREM1}, \textit{COMP}. In both of these examples, these gene lists

### Table 1. Top 15 ranked genes in terms of fold change (FC) for three different cancer types revealing the signature of the \textit{COL11A1}-expressing cancer-associated fibroblasts.

| Rank | Breast | Ovarian | Pancreatic |
|------|--------|---------|------------|
|      | Gene   | FC      | Gene       | FC      | Gene     | Log2FC   |
| 1    | COL11A1| 6.5     | COL11A1    | 8.23    | INHBA    | 5.15     |
| 2    | COL10A1| 4.07    | COL1A1     | 5.67    | COL10A1  | 5        |
| 3    | MFAP5  | 3.73    | TIMP3      | 5.52    | POSTN    | 4.92     |
| 4    | LRRC15 | 3.61    | FN1        | 5.4     | SULF1    | 4.63     |
| 5    | INHBA  | 3.44    | INHBA      | 4.94    | COL8A1   | 4.6      |
| 6    | FBN1   | 3.43    | EFEMP1     | 4.86    | COL11A1  | 4.4      |
| 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      | 3.9      |
| 11   | COL5A1 | 3.08    | POSTN      | 3.97    | CSPG2    | 3.87     |
| 12   | THBS2  | 2.99    | COL5A1     | 3.95    | WISP1    | 3.8      |
| 13   | LAMB1  | 2.97    | THBS2      | 3.91    | FN1      | 3.69     |
| 14   | FAP    | 2.96    | FBN1       | 3.9     | COMP     | 3.53     |
| 15   | SPOCK  | 2.91    | FAP        | 3.84    | COL5A2   | 3.38     |

https://doi.org/10.1371/journal.pcbi.1009228.t001
are clearly due to the presence of the COL11A1/INHBA/THBS2-expressing CAFs and therefore these are not cancer-type specific subtype signatures.

To computationally investigate the origin of those CAFs, we reasoned that analysis of rich datasets from single-cell RNA sequencing (scRNA-seq) provides unique opportunities for tracking the trajectories of cell differentiation lineages. There are several single-cell trajectory inference methods [18] performing “trajectory inference analysis,” ordering cells along a trajectory based on similarities in expression patterns.

In particular, we identified one exceptionally rich dataset [19] from pancreatic ductal adenocarcinoma, containing gene expression profiles from 24 tumor samples and 11 normal control samples. We found that several among the 24 tumor samples contained populations of cells strongly co-expressing COL11A1, THBS2 and INHBA, while none of the normal samples contained such cells. We also observed that the prominence of this co-expression signature varied significantly among the tumor samples, having only hints of their presence in some of them, suggesting that the corresponding patients were at various stages of the generation of COL11A1-expressing CAFs. This provides an opportunity to perform additional complementary computational analysis by comparing the prevalent fibroblastic cell populations across the tumor samples, and comparing them with those in the normal samples.

Therefore, in this paper we also used attractor analysis (Materials and Methods) in a novel manner for the analysis of rich scRNA-seq data. The unsupervised attractor algorithm [20] iteratively finds co-expression signatures converging to “attractor metagenes” pointing to the core (“heart”) of co-expression. Each attractor metagene is defined by a ranked set of genes along with scores determining their corresponding strengths within the signature, so the top-ranked genes are the most representative of the signature. The attractor algorithm has previously been used successfully for identifying features useful for breast cancer prognosis [21,22]. When applied on single cell data from a sample, it identifies the gene expression profiles of the dominant cell populations in the sample, and the algorithm is designed to ensure that all the top-ranked genes are co-expressed in the same cells. The purpose of the attractor algorithm is not to classify cells into mutually exclusive subsets. Instead, it identifies the genes at the core of co-expression signatures representing cellular populations from single-cell data, and it provides information that cannot be deduced with traditional clustering methods (see Discussion).

When we applied the attractor algorithm separately in each of the normal samples, we identified a set of nearly identical attractor signatures, corresponding to a type of adipose-derived stromal/stem cells (ASCs) naturally present in the stromal vascular fraction (SVF) of normal adipose tissue, expressing a unique characteristic signature containing fibroblastic markers such as LUM and DCN as well as adipose-related genes, such as APOD, CFD and MGP.

When we applied the algorithm in each of the tumor samples, we found a set of signatures that were changing in a remarkably continuous manner across the samples, some of them being very similar to those of the normal samples, while others are similar to the COL11A1-based signature. This suggests that the signatures undergo a gradual change as the transition proceeds, starting from the state of the normal ASCs and passing through a continuum of intermediate states. These results were consistent with those found by applying trajectory inference analysis, but they provided additional significant information based on their unique capabilities. Accordingly, this method demonstrated that there is a continuous “ASC to COL11A1-expressing CAF transition.”

This finding explains the stage association of the COL11A1-expressing signature as resulting from the interaction of tumor cells with the 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).

Finally, we validated our results in other cancer types (head and neck, ovarian, lung, breast), suggesting the pan-cancer nature of the ASC to COL11A1-expressing CAF transition.

Results

ASC to COL11A1-expressing CAF transition identified in pancreatic ductal adenocarcinoma (PDAC)

The PDAC dataset [19] consists of 57,530 scRNA-seq profiles from 24 PDAC tumor samples (T1-T24) and 11 normal samples (N1-N11). To find the expression profile of the dominant fibroblastic population in each sample, we applied the attractor algorithm on the set of identified mesenchymal cells (Materials and Methods). All samples (11 normal and 23 tumor samples, excluding sample T20 as it did not contain identified fibroblasts) yielded strong co-expression signatures involving many genes with big overlap among them. Genes LUM, DCN, FBLN1, MMP2, SFRP2 and COL1A2 appear in the top 100 genes in at least 33 out of the 34 samples (S1 Table), revealing a strong similarity shared by all those fibroblastic expression profiles. This strong overlap is consistent with the continuous transition process, as described below.

Dominant fibroblastic population in the normal pancreatic samples is adipose-derived. There is a striking similarity among the attractor profiles (Materials and Methods) of the eleven normal pancreatic samples, indicating that they represent a stable and normally occurring cell population. Specifically, there are 12 genes commonly shared among the top 30 genes in the attractors of at least ten of the eleven normal samples (Table 2), of which four genes are shared among all the samples \( P = 3 \times 10^{-113} \) by multi-set intersection test [5]. In addition to fibroblastic markers, there are several strongly expressed adipose-related or stemness-related genes in the list, such as APOD, CXCL12, and DPT, revealing that they are ASCs. Consistently, Gene Set Enrichment Analysis (GSEA) of these 12 commonly shared genes identified the most significant enrichment (FDR \( q \) value = \( 2.16 \times 10^{-19} \)) in the "BOQUEST_-_STEM_CELL_UP" dataset of genes upregulated in stromal stem cells from adipose tissue versus the non-stem counterparts [23].

To investigate the nature of this ASC population, we referred to recent results from single-cell analysis of general human adipose tissue [24]. We applied the attractor algorithm on the dataset with the single-cell expression profiles of all 26,350 cells taken from the SVF of normal adipose tissue from 25 samples, and compared the identified attractor with the “consensus attractor” (Materials and Methods) of the 11 normal pancreatic samples, which represented the main state of the normal fibroblastic population (Table 3). There are 14 overlapping genes between the top 30 gene lists \( P = 10^{-33} \) by hypergeometric test), 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 naturally occurring cell population. Furthermore, excluding the general fibroblastic markers LUM and DCN, we found that gene APOD (Apolipoprotein D) has the highest average ranking in Table 3, and is top-ranked in the independently found SVF fibroblastic population of cluster VP4 (supplementary file 20) of [24]. Therefore, we selected APOD as the representative marker for the ASC population.

Establishing the presence of COL11A1-expressing CAFs in PDAC tumor samples. Because COL11A1 serves as proxy of the full signature [1], a reliable test for determining if a sample contains the COL11A1-expressing CAFs is to rank all genes in terms of their association, measured by mutual information (Materials and Methods), with COL11A1 and see if
INHBA and THBS2 are top ranked. Indeed, this happens in several tumor samples, as shown in Table 4 for some of them (T23, T11, T6, T15, T18). For each sample, the shown genes are co-expressed in the same cells, because of the high correlations in a single-cell dataset.

Dominant fibroblastic populations in the tumor PDAC samples exhibits a continuous transition from ASCs to COL11A1-expressing CAFs. Based on the selection of APOD as a representative marker for the ASC population as described previously, we rearranged the attractors of the PDAC tumor samples in terms of descending order of the rank of APOD (Table 5) from left to right. There is a remarkable continuity in the shown expression profiles. The samples at the right side of the table include COL11A1 at increasingly high ranks. The intermediate tumor samples shown in the middle have cells expressing genes that are top-ranked in both the lists on the left as well as on the right. In other words, these cells are in a genuine intermediate state, rather than being a mixture of distinct subtypes (see detailed discussion in Materials and Methods).

Further demonstration of the continuity of the transition. As an additional confirmation of the continuity of the transition (as opposed to the presence of a mixture of distinct

Table 2. Top 30 genes of the identified attractors for each pancreatic normal sample (N1-N11). 12 commonly shared genes in at least ten of the eleven normal samples are highlighted.

| Rank | N1  | N2  | N3  | N4  | N5  | N6  | N7  | N8  | N9  | N10 | N11 |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1    | DCN | LUM | LUM | C7  | APOD| LUM | PTGD5| C7  | DCN | MMP2| LUM |
| 2    | LUM | DCN | FBLN1| FBLN5| DPT | DCN | APOD| LUM | APOD| DCN |
| 3    | C7  | C7  | C7  | APOD| DPT | LUM | FBLN5| FBLN1| DCN | C7  | LUM |
| 4    | FBLN1| FBLN1| PTGD5| DCN | PTGD5| DPT | APOD| FBLN1| APOD| FBLN1| EFEMP1|
| 5    | MGP | C1S | APOD| APOD| CXL12D| DPT | C7  | FBLN1| APOD| CT1K| CED |
| 6    | C1S | MGP | DPT | PTGD5| LUM | ABCA8| ADH1B| SFRP2| SFRP2| SFRP2| APOD |
| 7    | CCDC80| C1S | PDGFRA| FBLN1| COL6A3| C3 | DPT | PTGD5| SERPIN1| PLTP| MGP |
| 8    | PTGD5| DPT | APOD| C1R | PTGD5| APOD| COL6A3| CCDC80| PTGD5| MGST1| SERPIN1|
| 9    | DPT | CCDC80| SPF2| SPF2| SPF2| DCN | SPF2| SPF2| SPF2| SPF2| SPF2|
| 10   | C1R | PTGD5| DCN | SPFX| CCDC80| C1S | SPFRA| C1S | C1S | FBLN1| C3 |
| 11   | APOD| FBLN5| CXL12D| FMO2| CFD | C7  | CXL12D| CXL12D| SEPP1| SPON2| ADH1B|
| 12   | SEPP1| SEPP1| C1R | SEPP1| SPF2| SPF2| MRC2| PTGD5| SCN7A| DCN | SPFRA|
| 13   | FBLN5| COL1A2| COL6A3| CXL12D| FGF7| SPF2| SPF2| SPF2| SPF2| SPF2| SPF2|
| 14   | CXL12D| SPF2| ADH1B| CYR61| SPF2| FBLN5| MEG3| C1R | OLFM13| CXL12D| C1S |
| 15   | EFEMP1| SPFX| SPF2| SPF2| SPF2| MARCS| SPF2| SPF2| SPF2| SPF2| SPF2|
| 16   | COL1A2| SERPIN1| SPF2| CFD | CLEC11A| LRP1| CXL12D| OLFM13| CFD | C1R | COL6A3| C1R |
| 17   | SERPIN1| OLFM13| LAM1A2| PDGFRA| FMO2| CFD | C7  | CXL12D| CXL12D| SVEP1| COL6A3| PTN|
| 18   | ALDH1A1| COL6A3| C3  | C3  | NR2F1| NR2F1| MGP | DCN | SPF2 | MGP | COL6C11|
| 19   | CFD | MEG3 | FBLN5| C1S | TXNB | CCDC80| SPF2| EFEMP1| ALDH1A1| PDGFRA| GPMB |
| 20   | COL6A3| C1R | ABCA8| ABCA8| DCN | MRC2 | MRC2| SPF2| SPF2| EGF1| COL6A3| ST3|
| 21   | EMP1 | MFAP4| LR1P| CCDC80| LOX | COL1A2| FBLN5| PDGFRA| COL6A2| CYGB | DPT |
| 22   | PCOLEC| RARRES2| SLT2| S1N| PTN | C1R | CFD | C3  | DPT | CST3 | ARSK | SPF4 |
| 23   | C3  | PCOLEC| CFD | SERPIN1| IGF1| SVEP1| COL1A2| COL6A3| STHG1L| COL6A2| COL6A2|
| 24   | SPF2| PCOLEC| CFD | S1N| SPF2| SVEP1| HEG1| SMO2C| ABCA8| ADH1B| COL6C14| COL6A3|
| 25   | SERPIN1| COL12D| COL1A2| CFD | RP1-572C15.6| GSN | SPF2| NEGR1| C3  | BNP1| SMO2C |
| 26   | ANXA1 | FG7F | BOC | LAM1B| C3  | COL6A2| ACRV1| COL6A2| COL6A2| COL6A2| COL6A2|
| 27   | CYR61| PDGFRA| FSTL1| FTL| ADAMTS3| CFD | TIMP2| BOC | MMP2| GPC3 | FMO2 |
| 28   | CST3 | COL6A3| SVEP1| ANTX2R| STK17B| OLFML3| LAM12| OLFML3| PCOLEC| TMEM67| RP1-572C15.6|
| 29   | RARRES2| ALDH1A1| ABCA9| COL6A3| EMP1| PDGFRA| DAB2| EMP1| IGFL1| C1R | PCOLEC|
| 30   | PDGFRA| SPRY1| CYR61| MGP | MPZL1| PCOLEC| NR2F1| LAM12| ABCA8| PLXDC1| SEPP1|
fibroblastic subtypes), Fig 1 shows scatter plots for genes APOD and COL11A1, color-coded for the expression of fibroblastic marker LUM, of the mesenchymal cells in two fibroblast-rich samples T11 and T23. The presence of cells covering the full range from the upper-left to the bottom-right sides of the plots, including the intermediate stages in which cells co-express both markers, demonstrates the presence in each sample of cells representing the continuously varying transition from ASCs to COL11A1-expressing CAFs.

Table 3. Comparison of the attractors (top 30 genes) identified in the SVF of normal adipose tissue (Dataset 1) and in the normal pancreatic samples (Dataset 2). Common genes are highlighted in yellow.

| Rank | Dataset 1 | Dataset 2 | Rank (cont'd) | Dataset 1 | Dataset 2 |
|------|-----------|-----------|---------------|-----------|-----------|
| 1    | DCN       | LUM       | 16            | FOS       | PDGFRA    |
| 2    | LUM       | DCN       | 17            | MGST1     | SRPX      |
| 3    | APOD      | FBLN1     | 18            | COL1A2    | COL6A3    |
| 4    | CFD       | C7        | 19            | COL6A3    | ADH1B     |
| 5    | CXCL14    | APOD      | 20            | LAPTM4A   | CFD       |
| 6    | MGP       | PTGDS     | 21            | CXCL12    | OLFM3     |
| 7    | SERPINF1  | SFRP2     | 22            | WISP2     | SERPINF1  |
| 8    | GSN       | C1S       | 23            | SRPX      | MMP2      |
| 9    | GPX3      | CCDC80    | 24            | JUN       | CST3      |
| 10   | MFAP4     | MGP       | 25            | MMP2      | SEPP1     |
| 11   | PLAC9     | DPT       | 26            | COL6A2    | ABCA8     |
| 12   | S100A13   | CXCL12    | 27            | C1S       | COL1A2    |
| 13   | IGFBP6    | C1R       | 28            | CCDC80    | LAMB1     |
| 14   | DPT       | FBLN5     | 29            | EGR1      | SVEP1     |
| 15   | MFAP5     | C3        | 30            | PCOLCE    | MEG3      |

Table 4. Ranked COL11A1-associated genes in five PDAC samples. MI = Mutual Information.

| Rank | T23 | MI  | T11 | MI  | T6  | MI  | T15 | MI  | T18 | MI  |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1    | COL11A1 | 1   | COL11A1 | 1  | COL11A1 | 1   | COL11A1 | 1  |
| 2    | COL10A1 | 0.3603 | CTHRC1 | 0.2434 | MFAP5 | 0.2353 | MFAP5 | 0.3198 | MFAP5 | 0.3408 |
| 3    | COL12A1 | 0.3383 | MFAP5 | 0.2357 | FNDC1 | 0.1997 | GJB2 | 0.2583 | SUGCT | 0.3379 |
| 4    | COL1A1 | 0.3187 | COL12A1 | 0.2345 | NTM | 0.1912 | COL10A1 | 0.2580 | COL10A1 | 0.2899 |
| 5    | THBS2 | 0.3167 | COL10A1 | 0.2238 | COL8A1 | 0.1877 | INHBA | 0.2561 | CSorf46 | 0.2753 |
| 6    | COL1A2 | 0.3099 | C1QTNF3 | 0.2155 | TWIST1 | 0.1714 | C1QTNF3 | 0.2514 | PPAPDC1A | 0.2686 |
| 7    | COL5A2 | 0.3003 | THBS2 | 0.2123 | COL10A1 | 0.1619 | MATN3 | 0.2505 | NTM | 0.2649 |
| 8    | CTHRC1 | 0.2854 | COL1A2 | 0.2045 | THBS2 | 0.1559 | FNDC1 | 0.2503 | COL8A1 | 0.2534 |
| 9    | FN1 | 0.2781 | COL8A1 | 0.2018 | ITGA11 | 0.1556 | COL8A2 | 0.2411 | INHBA | 0.2430 |
| 10   | COL3A1 | 0.2770 | AEBP1 | 0.2000 | PPAPDC1A | 0.1305 | COL1A1 | 0.2399 | FNDC1 | 0.2264 |
| 11   | INHBA | 0.2746 | LUM | 0.1989 | DIO2 | 0.1298 | COL12A1 | 0.2351 | COL12A1 | 0.2194 |
| 12   | AEBP1 | 0.2688 | COL1A1 | 0.1985 | JGFL2 | 0.1178 | COL8A1 | 0.2325 | JGFL2 | 0.2153 |
| 13   | COL5A1 | 0.2626 | FNDC1 | 0.1963 | SUGCT | 0.1170 | THBS2 | 0.2292 | THBS2 | 0.2094 |
| 14   | VCAN | 0.2457 | SFRP2 | 0.1955 | ADAM12 | 0.1165 | NTM | 0.2257 | CTHRC1 | 0.2026 |
| 15   | MFAP5 | 0.2419 | GJB2 | 0.1879 | C1QTNF3 | 0.1165 | COL1A2 | 0.2220 | SULF1 | 0.2015 |
| 16   | MMP11 | 0.2360 | MATN3 | 0.1817 | ITGBL1 | 0.1109 | GREM1 | 0.2156 | COMP | 0.1926 |
| 17   | COL8A1 | 0.2357 | COL3A1 | 0.1740 | GREM1 | 0.1018 | FN1 | 0.2146 | STM2 | 0.1926 |
| 18   | COL6A3 | 0.2339 | INHBA | 0.1696 | PHA3 | 0.1008 | IGFL2 | 0.2141 | WNT2 | 0.1925 |
| 19   | POSTN | 0.2316 | DCN | 0.1692 | INHBA | 0.1002 | CXCL14 | 0.2112 | MMP11 | 0.1919 |
| 20   | MFAP2 | 0.2275 | CTGF | 0.1691 | COL5A1 | 0.0983 | ITGBL1 | 0.2048 | SPOCK1 | 0.1878 |
Table 5. Rearranged PDAC tumor samples showing the continuously changing pattern of the signature profile. Columns are sorted based on APOD rankings. Genes APOD and COL11A1 are highlighted in green and red, respectively.

| Rank | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 | T11 | T12 | T13 | T14 | T15 | T16 | T17 | T18 | T19 | T20 | T21 | T22 | T23 | T24 |
|------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1    | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD |
| 2    | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD |
| 3    | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD |
| 4    | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD |
| 5    | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD | IL15 | LCN | DCN | WIF1 | RTP2 | METP2 | IGFBP3 | APOD |
| 6    | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD | VEGF | DCN | CTS | IL15 | RTP2 | METP2 | IGFBP3 | APOD |

The role of adipose-derived stromal cells in cancer invasiveness.

https://doi.org/10.1371/journal.pcbi.1009228.t005
To further investigate the continuous transition, we partitioned the 34 pancreatic samples into three groups. Group 1 includes the eleven normal samples (N1 to N11). For tumor samples, we divided the rearranged samples in Table 5 into two groups (Group 2 and Group 3). Group 2 contains all samples to the left of and including T22, so that \( \text{APOD} \) is ranked before \( \text{COL11A1} \) in the attractors of that Group, representing a relatively earlier stage of this transition. We then applied the consensus version of the attractor finding algorithm (Materials and Methods) and identified the signatures representing the main state of the fibroblasts for each of the above three sample groups (Table 6). Although there are many shared genes, the groups have distinct gene rankings. Group 1 (normal samples) contains many adipose-related genes, consistent with Table 2. Group 3 contains, in addition to \( \text{COL11A1} \), many among the other CAF genes, such as \( \text{THBS2} \), \( \text{INHBA} \), \( \text{AEBP1} \), \( \text{MFAP5} \) and \( \text{COL10A1} \). Group 2 displays an intermediate state, including markers of both ASCs as well as CAFs.

To find potential critical genes at the initiation phase of the cellular transition, we focused on the first tumor samples (with highest \( \text{APOD} \) ranking) in Table 5, so we can compare them with those of the normal ASCs. We observe that gene \( \text{SFRP4} \) stands out, as it appears for the first time remarkably among the top genes in all the first samples T2, T13, T14, T19, ranked 4th, 6th, 4th 8th, respectively. This suggests that the Wnt pathway is involved in the initiation of the cellular transition, because \( \text{SFRP4} \) is a Wnt pathway regulator whose expression has been found associated with various cancer types [25,26]. Interestingly, \( \text{SFRP4} \) disappears from the list of the attractors, indicating that it is downregulated in the final stage of the transition.

It is also known that gene \( \text{RARRES1} \) (aka TIG1) plays an important role in regulating the proliferation and differentiation of ASCs [27]. Consistently, Table 6 reveals that \( \text{RARRES1} \)
appears for the first time in the attractors of the initial tumor samples. Just like SFRP4, RARRES1 is downregulated in the final stage, related to the fact that it has been suggested as a tumor suppressor [28,29].

We also performed differential expression (DE) analysis comparing the normal samples with the first samples (T2, T13, T14, T19) of Table 6 (Materials and Methods; S2 Table). The results of such DE analysis represent the full population of fibroblasts and not necessarily reflect the expression changes in the particular cells undergoing the ASC to COL11A1 expressing CAF transition. Gene CFD was found to be most downregulated, consistent with the expected downregulation of adipose-related genes as they differentiate into fibroblasts. Genes SFRP4 and RARRES1 are upregulated consistent with their appearance in the attractors.

On the other hand, the top upregulated gene is phospholipase A2 group IIA (PLA2G2A), which is not among the top genes of any attractors we identified, indicating that it is not expressed by cells undergoing the ASC to COL11A1-expressing CAF transition. It probably still plays, however, an important related parallel role and many previous studies referred to its effects on prognosis of multiple cancer types [30–32]. The PLA2G2A protein is a member of a family of enzymes catalyzing the hydrolysis of phospholipids into free fatty acid. We hypothesize that this process leads to fatty acid oxidation, which may facilitate metastatic progression. Indeed, it has been recognized that fatty acid oxidation is associated with the final COL11A1-expressing stage of the transition [33]. These results suggest that lipid metabolic reprogramming plays an important role in the metastasis-associated biological mechanism [34], by potentially providing energy for the metastasizing tumor cells.

Validation with trajectory inference. We independently applied trajectory inference (TI) analysis on the PDAC fibroblasts by using the Slingshot [35] method in an unsupervised manner. We first performed unsupervised clustering on the identified fibroblasts (Materials and Methods), resulting in four subgroups X1, X2, X3, X4 (S1A Fig) with the top differentially expressed genes shown in S1B Fig. One of these clusters (X4) was discarded from further TI analysis, because it mainly expressed the IL1 CAF marker HAS1 (Hyaluronan Synthase 1), which is not expressed by either ASCs or COL11A1-expressing CAFs (and does not appear at

Table 6 Top 30 genes of the consensus attractors for three different PDAC sample groups. Group1: normal samples; Group 3: T11, T21, T23, T9, T16, T17, T8; Group 2: other tumor samples.

| Rank | Group1 | Group2 | Group3 | Rank (cont’d) | Group1 | Group2 | Group3 |
|------|--------|--------|--------|--------------|--------|--------|--------|
| 1    | LUM    | LUM    | COL1A1 | 16           | PDGFR A| CYP1B1 | MMP2   |
| 2    | DCN    | SFRP2  | COL1A2 | 17           | SRPX   | FBLN5  | DCN    |
| 3    | FBLN1  | APOD   | COL3A1 | 18           | COL6A3 | MEG3   | SFRP2  |
| 4    | C7     | SFRP4  | FN1    | 19           | ADH1B  | COL1A1 | TMSB10 |
| 5    | APOD   | MMP2   | COL5A2 | 20           | CFD    | C3     | POSTN  |
| 6    | PTGDS  | VCAN   | COL5A1 | 21           | OLEML3 | RARRES1| MXRA5  |
| 7    | SFRP2  | PDGFR A| COL6A3 | 22           | SERPINF1| CDC80  | COL6A2 |
| 8    | C1S    | FBLN1  | COL1A1 | 23           | MMP2   | MOXD1  | ISLR   |
| 9    | CCDC80 | DCN    | CTHRC1 | 24           | CST3   | PLXDC2 | AEBP1  |
| 10   | MGP    | EFEMP1 | THBS2  | 25           | SEPP1  | HTRA3  | MEG3   |
| 11   | DPT    | CTHRC1 | VCAN   | 26           | ABCA8  | COL10A1| MAFP5  |
| 12   | CXCL12 | ISLR   | COL10A1| 27           | COL1A2 | COL8A1 | SERPINH1|
| 13   | C1R    | COL6A3 | LUM    | 28           | LAMB1  | ITGBL1 | MMP14  |
| 14   | FBLN5  | COL1A2 | SPARC  | 29           | SVEP1  | OMD    | MAFP2  |
| 15   | C3     | CTSK   | COL12A1| 30           | MEG3   | PTGDS  | INHBA  |

https://doi.org/10.1371/journal.pcbi.1009228.t006

The role of adipose-derived stromal cells in cancer invasiveness

PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1009228 July 20, 2021 10 / 22
all in S1 Table), and contained only 3% of fibroblasts resulting almost exclusively from patient T11 (S1C Fig).

As seen from the list of top differentially expressed genes of each cluster, 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 CED. Consistently, S2A and S2B Fig show the single trajectory path resulting from T1 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 transition. The orderings of patient groups and sample identity (S2C and S2D Fig) are also consistent with our previous findings based on attractor analysis. S3 Table 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 transition.

We then used a generalized additive model (GAM) fit to pseudotime-ordered expression data to visualize the trend of gene expressions (Fig 2A).

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 2B), while the expression of COL11A1-associated genes gradually increased (Fig 2C). 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 SFRP4 and RARRES1 (Fig 2D) increased consistently at the beginning and then decreased after reaching a peak, suggesting that they may play important roles in the differentiation path.

Validation in other cancer types

Next, we validated the ASC to COL11A1-expressing CAF transition 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 cancer (HNSCC) [36], ovarian cancer [37], lung cancer [38] and breast cancer [39].

The COL11A1-expressing CAF signature has been confirmed to be a pan-cancer signature [40–42]. 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 transition, and to also confirm that some samples are at an intermediate stage, expressing genes such as SFRP4, RARRES1 and THBS2, in addition to the core ASC genes, demonstrating that they are at an intermediate stage of the transition.

Head and neck squamous cell carcinoma. For the HNSCC dataset, the authors of the paper presenting the data [36] reported that the cancer-associated fibroblasts in the dataset can be partitioned into two subsets, which they name CAF1 and CAF2. In S5 Table 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 S5 Table 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 S5 Table 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 transition.

In our independent analysis, we performed clustering identifying 1,026 fibroblasts from all available cells (S3A Fig; 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 (S4 Table), 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 (S5 Table) 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 transition.

**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 2. Trajectory analysis of PDAC.** A. GAM fit to pseudotime ordered expression data to visualize the trend of gene expressions. B. Expression of adipose-related genes along the transition lineage. The x axis shows the cell orders and the y axis shows the normalized read count. C. Expression of COL11A1-associated genes along the transition lineage. D. Expression of RARRES1 and SFRP4 genes along the transition lineage.

https://doi.org/10.1371/journal.pcbi.1009228.g002
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. S5 Table 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 transition, 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 (S3C Fig, S7 Table). The presence of the transition is evident by the attractors identified in the mesenchymal cells for patients 4 and 3 (S5 Table). 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) (S3D Fig, S8 Table). 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 (S5 Table), as the former contains C1S, C1R, CXCL12, PTGDS, C3, while the latter contains THBS2, COL11A1, COL10A1 at top-ranked positions.

### Potential therapeutic targets inhibiting the invasiveness-associated transition

This work provides opportunities for identifying therapeutic targets inhibiting the cellular transition. For example, targeting of gene MFAP5 was recently found to enhance chemosensitivity in ovarian and pancreatic cancers [43]. Specifically, the author states that “MFAP5 blockade suppresses fibrosis through downregulating of fibrosis-related genes such as COL11A1.” Consistently, we found MFAP5 as one of the most highly associated genes with COL11A1 (Table 4).

As mentioned earlier, genes SFRP4 and RARRES1 are transiently expressed in Group 2 of Table 6, suggesting that they can be investigated for inhibiting the cellular transition.

Of particular interest as potential drivers are noncoding RNAs due to their typical regulatory role. Because the expression of these genes is not accurately captured by scRNA-seq technology, we did a thorough analysis of the full set of The Cancer Genome Atlas (TCGA) pan-cancer data. For the RNA sequencing and miRNA sequencing dataset of each cancer type, we removed the genes in which more than 50% of the samples have zero counts. Then we performed quantile normalization using the limma package [44] (v3.40.6) on log2 transformed counts. In each of the 33 cancer types, we ranked all protein-coding genes in terms of the association (using the metric of mutual information) of their expression with that of gene COL11A1. We excluded the 11 cancer types (LGG, SKCM, SARC, LAML, PCPG, GBM, TGCT, THYM, ACC, UVM, UCS) in which neither THBS2 nor INHBA was among the 50 top-ranked genes, because of the absence of significant amounts of COL11A1-expressing CAFs in those samples (1st sheet in S9 Table). In each of the remaining 22 cancer types, we then ranked all long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) in terms of their association
with COL11A1 (2nd and 3rd sheets in S9 Table). Finally, we did pan-cancer sorting of all lncRNAs and miRNAs in terms of the median rank of all lncRNAs and miRNAs (4th sheet in S9 Table).

We found that LINC01614 represents a particularly promising therapeutic target. It had a perfect score of 1 in the pan-cancer sorting list, being strikingly the top-ranked gene in 14 (BRCA, UCEC, KIRC, HNSC, LUAD, LUAD, LUSC, OV, STAD, ESCA, PAAD, MESO, DLBC, CHOL) out of the 22 cancer types (2nd sheet in S9 Table). In fact, the association of LINC01614 was even higher than that of marker protein-coding gene INHBA. The pan-cancer consensus ranking of protein-coding genes in terms of LINC01614 (5th sheet in S9 Table) corresponds precisely to the COL11A1-expressing CAF signature. These rankings, in which marker genes unique to the original and intermediate stages are missing, indicate that LINC01614 is involved in the very final stage of the creation of the COL11A1-expressing CAFs. Therefore, we hypothesize that therapeutics targeting LINC01614 specifically in patients’ CAFs may inhibit the final metastasis-facilitating stage of the transition.

We also found that the three top-ranked miRNAs were miR-199a-1, miR-199b, miR-199a-2. The associated miR-214 is also very highly ranked (3rd sheet in S9 Table).

**Discussion**

Our results indicate that the cancer invasiveness-associated COL11A1-expressing CAFs are produced as a result of the interaction of tumor cells with the adipose microenvironment. Therefore, one contribution of our work is that it provides a potential explanation to the well-known fact that adipose tissue contributes to the development and progression of cancer [45–47].

Another contribution is that it precisely identifies the ASC population, as evidenced by the consistent presence of its marker genes among the top-ranked attractor genes in each of the eleven columns of Table 2. The identification of those particular marker genes (APOD prominent among them) cannot be due to chance, because these were eleven totally independent unbiased experiments, and also because the attractor algorithm applied on the SVF of normal adipose tissue in another independent dataset identified precisely the same genes. This finding could not have been achieved with traditional methods.

There is consensus agreement that CAFs are a promising potential target for optimizing therapeutic strategies against cancer, but such developments are restricted by our current limitations in our understanding of the origin of CAFs and heterogeneity in CAF function [48]. Therefore, there is an urgent need to enhance our understanding of those matters. Our results provide clarity on one important particular component (out of several) of the heterogeneous fibroblast tumor microenvironment. To avoid potential erroneous conclusions after applying bioinformatics algorithms, single-cell data analysis provides an unprecedented capability to validate results, including those resulting from the attractor algorithm, by “seeing” individual cells in color-coded scatter plots, such as the one shown in Fig 1, observing and confirming the presence or absence of distinct populations characterized by the combined presence of particular marker genes.

In particular, there are several published papers relying on the application of clustering algorithms following dimensionality reduction on the particular datasets they use, and concluding that there exist a number of distinct and mutually exclusive CAF subpopulations. These reported fibroblastic subpopulations occasionally have gene expression profiles that are conflicting with each other in significant ways among these publications. Examples include the hC1 and hC0 clusters in [49], the C9 and C10 clusters in [42], the CAF2 and CAF1 clusters in
[36], the iCAF and myCAF clusters in [50,51] and the iCAF an mCAF clusters in [52]. A review of such results in pancreatic cancer appears in [53].

As an example of conflicting results, the “iCAFs” identified in [52] have significant differences from those identified in other papers and are, in fact, identical to the normal ASCs (Fig 3B of [52]) identified in this paper, as evidenced by the list of its differentially expressed genes (PTGDS, LUM, CFD, FBLN1, APOD, DCN, CXCL14, SFRP2, MMP2, all of which appear in Table 3), further validating the ASC signature. Therefore, this identified cluster contains mainly normal cells at the origin of the transition, which should not even be called CAFs.

Similarly, a recent single-cell data analysis [54] identified two clusters “touching” each other in a UMAP plot (Fig 2A of [54]), C0 and C3, which are precisely the two endpoints of the ASC to COL11A1-expressing CAF transition. Indeed, as identified in Table S6-1 of [54], C0 cluster has the marker genes APOD, PTGDS, C7, C3, MGP . . . which the attractor algorithm had identified and validated in this paper. On the other hand, the marker genes of cluster C3 are precisely those of the COL11A1-expressing CAFs, in which all three genes COL11A1, INHBA and THBS2 are top-ranked (because the metastatic process was already underway). Importantly, as shown in Fig 2B of [54], the ASC marker genes APOD and PTGDS (top ranked in C0 and unrelated to CAFs) are significantly expressed even in the COL11A1-expressing cluster C3 of the paper, providing further evidence of the presence of intermediate states consistent with the transition—and the separating line between C0 and C3 in the diagram is not generated by any biologically reliable manner, consistent with the continuity.

On the other hand, our work is consistent with, and complementary to the results of [49] focusing on the immunotherapy response, in which the presence of the “TGF-beta CAFs” was inferred by an 11-gene signature consisting of MMP11, COL11A1, C1QTNF3, CTHRC1, COL12A1, COL10A1, COL5A2, THBS2, AEBP1, LRRC15, ITGA11. This population apparently represents the COL11A1-expressing CAF endpoint of the transition, and gene LRRC15 was selected as the representative gene based on the fact that it was found to be the most differentially expressed gene between CAFs and normal tissue fibroblasts in mouse models. Indeed, LRRC15 is a key member of the COL11A1-expressing CAF signature (Table 4 of [1]) and we also found that COL11A1 is the highest associated gene to LRRC15 in the Group 3 PDAC patients.

In our work we used a detailed gene association-based scrutiny of all our results, including numerous color-coded scatter plots, rather than blindly accepting clustering results. We believe that this nontraditional computational methodology, when used on rich single-cell data, represents a paradigm shift in which systems biology alone can be trusted, by itself, for producing reliable results. We hope that our results will give rise to testable hypotheses that could eventually lead to the development of pan-cancer therapeutics inhibiting the ASC to COL11A1-expressing CAF transition.

Materials and methods

Datasets availability

The pancreatic dataset [39] was downloaded from the Genome Sequence Archive with accession number CRA001160. The four validation datasets of other cancer types are also publicly available: HNSCC [36] (GSE103322), ovarian [37] (GSE118828), lung cancer [38] (E-MTAB-6149 and E-MTAB-6653), breast cancer [39] (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 selected the Seurat R toolkit [55] for data processing and cell identification. Seurat implements the entire clustering workflow and has an advantage in speed and scalability to analyze large datasets [56]. We applied the Seurat (v3.1.4) 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, and RGS5, ACTA2, PDGFRB and ADIRF 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 [57]. We used a spline based estimator [58] to estimate MI values and normalized so the maximum possible value is 1. The MI value is clipped to zero if the Pearson correlation between the two variables is negative. The details of the estimation method are described in the paper introducing the attractor algorithm [20]. We used the getMI or getAllMIWz function implemented in the cafr R package with parameter negateMI = TRUE.

Attractor-based analysis

The attractor algorithm was first proposed for identifying co-expression signatures from bulk expression values in samples [20]. In this study, we use the attractor algorithm for the first time for the purpose of scrutinizing cell populations in single-cell data. Compared to conventional single-cell methods, the attractor algorithm features the unique capability of discovering precise profiles of cell populations, which other methods cannot achieve (see Discussion).

Briefly, the 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. Using the expression vector corresponding to a seed gene as input, the algorithm converges to an “attractor” in the form of a list of ranked genes, together with scores (ranging from 0 to 1) for each of these genes measuring the strength of the membership of that gene in the signature. It has a characteristic property that using different “attractee” genes belonging to a co-expression signature as seeds leads to the identical attractor.

The attractor algorithm had previously been used to find co-expression signatures in bulk gene expression data, in which case a converged attractor could represent a mixture of contributions from distinct cell subpopulations. When using single-cell data, however, the
characteristic genes of particular distinct subpopulations will have high expression values only in the cells from those subpopulations and low values in other cells. These genes will have pairwise positive and large correlations, and therefore they will be highly ranked in attractor signatures representing such individual subpopulations. On the other hand, two characteristic marker genes belonging to two different distinct subpopulations will have reverse-associated expression values across those cells, which will contribute negatively to the overall correlation between these two genes. Only if two genes are co-expressed across individual cells will they appear highly ranked in the same attractor.

For single dataset, we applied the attractor finding algorithm using the findAttractor function implemented in the cafr (v0.312) R package [20] with the general fibroblastic marker gene LUM as seed. Identical results in all samples will be found, with very rare exceptions, if other fibroblastic markers, such as DCN, are used. The exponent (a) was set to different values for scRNA-seq datasets profiled from different protocols. 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. To find the consensus attractor for multiple datasets, we applied the consensus version of the attractor finding algorithm as described in [59]. 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 [35] method for TI analysis, based on its robustness and suggestions made by the benchmarking pipeline dynverse [18]. 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, where the number of clusters in the Mclust function was set to 3 based on the fact that there were three clusters in our previous Seurat clustering results. The final step of lineage inference analysis used the slingshot wrapper function in an unsupervised manner. 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 for overlapping genes from different sets.* We applied the hypergeometric test for evaluating the significance of genes shared by different sets. If there are two sets to compare, we used the phyper R function. If there are more than two sets to compare, we employed the multi-set intersection test [5] by applying the cpsets function implemented in the SuperExactTest R package. Regarding the background universe size of genes, we used the total number of genes analyzed in the specific expression matrix. In the case of comparing sets coming from different studies, we used 20,000 as the universe size.

*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
The role of adipose-derived stromal cells in cancer invasiveness

\( P \) 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.

Supporting information

S1 Fig. Overview of the PDAC fibroblasts. A. 6,267 fibroblasts originated from 11 control pancreases and 23 tumor samples were petitioned into four groups X1-X4. Fractions of the fibroblasts were: 45%, 38%, 14%, and 3%. B. Table showing the top 20 DE genes for each cluster. C. Bar plots presenting the numbers of cells captured for each cluster.

(TIF)

S2 Fig. Trajectory analysis of 6,075 fibroblasts in PDAC dataset. A. 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. Color-coded trajectory analysis of fibroblasts for group information. D. Color-coded trajectory analysis of fibroblasts for sample identity.

(TIF)

S3 Fig. 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.

(TIF)

S1 Table. LUM-seeded attractors (top 100 genes) identified in each PDAC sample.

(XLSX)

S2 Table. Differentially expressed genes comparing normal pancreatic samples against four PDAC samples at the initial phase of the transition.

(XLSX)

S3 Table. Top 100 genes of temporally expressed genes on the pseudotime variable.

(XLSX)

S4 Table. Differentially expressed genes among different clusters of HNSCC dataset.

(XLSX)

S5 Table. LUM-seeded attractors (top 100 genes) from validating samples of other cancer types.

(XLSX)

S6 Table. Differentially expressed genes among different clusters of ovarian cancer dataset.

(XLSX)

S7 Table. Differentially expressed genes among different clusters of mesenchymal cells from lung cancer dataset.

(XLSX)

S8 Table. Differentially expressed genes among different clusters of stromal cells from breast cancer dataset.

(XLSX)

S9 Table. Ranked genes lists in terms of their association (mutual information) with gene \( COL11A \) by using the full set of pan-cancer TCGA bulk RNA-seq data.

(XLSX)
Acknowledgments

The noncoding RNA-related results presented here are in whole based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author Contributions

Conceptualization: Kaiyi Zhu, Lingyi Cai, Juan R. de los Toyos, Dimitris Anastassiou.

Data curation: Kaiyi Zhu, Lingyi Cai, Chenqian Cui.

Formal analysis: Kaiyi Zhu, Lingyi Cai, Chenqian Cui, Dimitris Anastassiou.

Funding acquisition: Dimitris Anastassiou.

Investigation: Kaiyi Zhu, Lingyi Cai, Chenqian Cui, Juan R. de los Toyos, Dimitris Anastassiou.

Methodology: Kaiyi Zhu, Lingyi Cai, Juan R. de los Toyos, Dimitris Anastassiou.

Project administration: Dimitris Anastassiou.

Resources: Kaiyi Zhu, Lingyi Cai.

Software: Kaiyi Zhu, Lingyi Cai, Chenqian Cui, Dimitris Anastassiou.

Supervision: Dimitris Anastassiou.

Validation: Kaiyi Zhu, Lingyi Cai, Dimitris Anastassiou.

Visualization: Kaiyi Zhu, Lingyi Cai, Dimitris Anastassiou.

Writing – original draft: Kaiyi Zhu, Lingyi Cai, Juan R. de los Toyos, Dimitris Anastassiou.

Writing – review & editing: Kaiyi Zhu, Lingyi Cai, Chenqian Cui, Juan R. de los Toyos, Dimitris Anastassiou.

References

1. Kim H, Watkinson J, Varadan V, Anastassiou D. Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. BMC Genomics. 2010 Nov 3; 3(1):51. https://doi.org/10.1186/1755-8794-3-51 PMID: 21047417

2. Schuetz CS, Bonin M, Clare SE, Nieselt K, Sotlar K, Walter M, 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. 2006 May 15; 66(10):5278–86. https://doi.org/10.1158/0008-5472.CAN-05-4610 PMID: 16707453

3. Bignotti E, Tassi RA, Calza S, Ravaghi A, Bandiera E, Rossi E, et al. Gene expression profile of ovarian serous papillary carcinomas: identification of metastasis-associated genes. Am J Obstet Gynecol. 2007 Mar; 196(3):245.e1-11. https://doi.org/10.1016/j.ajog.2006.10.874 PMID: 17346539

4. Badea L, Herlea V, Dima SO, Dumitrascu T, Popescu I. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. Hepatogastroenterology. 2008 Dec; 55(88):2016–27. PMID: 19260470

5. Wang M, Zhao Y, Zhang B. Efficient Test and Visualization of Multi-Set Intersections. Scientific Reports. 2015 Nov 25; 5(1):16923. https://doi.org/10.1038/srep16923 PMID: 26603754

6. Shen L, Yang M, Lin Q, Zhang Z, Zhu B, Miao C. COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. Oncology Reports. 2016 Aug 1; 36(2):877–85. https://doi.org/10.3892/or.2016.4869 PMID: 27373316

7. García-Pravia C, Galván JA, Gutiérrez-Corral N, Solar-García L, García-Pérez E, García-Ocaña M, et al. Overexpression of COL11A1 by Cancer-Associated Fibroblasts: Clinical Relevance of a Stromal Marker in Pancreatic Cancer. PLOS ONE. 2013 Oct 23; 8(10):e78327. https://doi.org/10.1371/journal.pone.0078327 PMID: 24194920
8. Wu Y-H, Chang T-H, Huang Y-F, Chen C-C, Chou C-Y. COL11A1 confers chemoresistance on ovarian cancer cells through the activation of Akt/c/EBPβ pathway and PDK1 stabilization. Oncotarget. 2015 Jun 10; 6(27):23748–63. https://doi.org/10.18632/oncotarget.4250 PMID: 26087191

9. Chen P-C, Tang C-H, Lin L-W, Tsai C-H, Chu C-Y, Lin T-H, et al. Thrombospondin-2 promotes prostate cancer bone metastasis by the up-regulation of matrix metalloproteinase-2 through down-regulating miR-376c expression. Journal of Hematology & Oncology. 2017 Jan 25; 10(1):33. https://doi.org/10.1186/s13045-017-0390-6 PMID: 28122633

10. Seder CW, Hartojo W, Lin L, Silvers AL, Wang Z, Thomas DG, et al. INHBA Overexpression Promotes Cell Proliferation and May Be Epigenetically Regulated in Esophageal Adenocarcinoma. Journal of Thoracic Oncology. 2009 Apr 1; 4(4):455–62. https://doi.org/10.1097/JTO.0b013e31819c791a PMID: 19240652

11. Wang Q, Wen Y-G, Li D-P, Xia J, Zhou C-Z, Yan D-W, et al. Upregulated INHBA expression is associated with poor survival in gastric cancer. Med Oncol. 2012 Mar 1; 29(1):77–83. https://doi.org/10.1007/s12032-010-9766-y PMID: 21132402

12. Tu H, Li J, Lin L, Wang L. COL11A1 Was Involved in Cell Proliferation, Apoptosis and Migration in Non-Small Cell Lung Cancer Cells. Journal of Investigative Surgery. 2020 Nov 5; 0(0):1–6. https://doi.org/10.1080/08941939.2019.167289 PMID: 33148075

13. Wang X, Zhang L, Li H, Sun W, Zhang H, Lei M. THBS2 is a Potential Prognostic Biomarker in Colorectal Cancer. Scientific Reports. 2016 Sep 16; 6(1):33366. https://doi.org/10.1038/srep33366 PMID: 27632935

14. Li X, Yu W, Liang C, Xu Y, Zhang M, Ding X, et al. INHBA is a prognostic predictor for patients with colon adenocarcinoma. BMC Cancer. 2020 Apr 15; 20(1):305. https://doi.org/10.1186/s12885-020-06743-2 PMID: 32293338

15. Seder CW, Hartojo W, Lin L, Silvers AL, Wang Z, Thomas DG, et al. Upregulated INHBA Expression May Promote Cell Proliferation and Is Associated with Poor Survival in Lung Adenocarcinoma. Neoplasia. 2009 Apr; 11(4):388–96. https://doi.org/10.1593/neo.81582 PMID: 19308293

16. Verhaak RGW, Tamayo P, Yang J-Y, Hubbard D, Zhang H, Creighton CJ, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J Clin Invest. 2013 Jan; 123(1):517–25. https://doi.org/10.1172/JCI65833 PMID: 23257362

17. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nature Genetics. 2015 Oct; 47(10):1168–78. https://doi.org/10.1038/ng.3398 PMID: 26343385

18. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nature Biotechnology. 2019 May; 37(5):547–54. https://doi.org/10.1038/s41587-019-0071-9 PMID: 30936559

19. Peng J, Sun B-F, Chen C-Y, Zhou J-Y, Chen Y-S, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Research. 2019 Sep; 29(9):725–38. https://doi.org/10.1038/s41422-019-0195-z PMID: 31732937

20. Cheng W-Y, Yang T-HO, Anastassiou D. Biomolecular Events in Cancer Revealed by Attractor Metagenes. PLOS Computational Biology. 2013 Feb 21; 9(2):e1002920. https://doi.org/10.1371/journal.pcbi.1002920 PMID: 23468603

21. Cheng W-Y, Ou Yang T-H, Anastassiou D. Development of a prognostic model for breast cancer survival in an open challenge environment. Sci Transl Med. 2013 Apr 17; 5(181):181ra50. https://doi.org/10.1126/scitranslmed.3005974 PMID: 23592962

22. McCarthy N. Rising to the challenge. Nature Reviews Cancer. 2013 Jun; 13(6):378–378. https://doi.org/10.1038/nrc3530 PMID: 23604208

23. Boquest AC, Shahdadfar A, Frensdal K, Sigurjonsson O, Tunheim SH, Collas P, et al. Isolation and Transcription Profiling of Purified Uncultured Human Stromal Stem Cells: Alteration of Gene Expression after In Vitro Cell Culture. MBoC. 2005 Jan 5; 16(3):725–38. https://doi.org/10.1093/mbc/e04-10-0949 PMID: 15655089

24. Vijay J, Gauthier M-F, Biswell RL, Louiselle DA, Johnston JJ, Cheung WA, et al. Single-cell analysis of human adipose tissue identifies depot- and disease-specific cell types. Nature Metabolism. 2020 Jan; 2(1):97–109. https://doi.org/10.1038/s41597-019-0152-6 PMID: 32066997

25. J D, Am S, J W, R F, N Z, A C, et al. Expression of secreted frizzled-related protein 4 (SFRP4) in primary serous ovarian tumours. Eur J Gynaecol Oncol. 2009 Jan 1; 30(2):133–41. PMID: 19480240

26. Sandsmark E, Andersen MK, Boﬁn AM, Bertilsson H, Drables F, Batenh TF, et al. SFRP4 gene expression is increased in aggressive prostate cancer. Scientiﬁc Reports. 2017 Oct 27; 7(1):14276. https://doi.org/10.1038/s41598-017-14622-3 PMID: 29079735
27. Ohnishi S, Okabe K, Obata H, Otani K, Ishikane S, Ogino H, et al. Involvement of tazarotene-induced gene 1 in proliferation and differentiation of human adipose tissue-derived mesenchymal stem cells. Cell Proliferation. 2009; 42(3):309-16. https://doi.org/10.1111/j.1365-2184.2008.00592.x PMID: 19250291

28. Jing C, El-Ghany MA, Beesley C, Foster CS, Rudland PS, Smith P, et al. Tazarotene-Induced Gene 1 (TIG1) Expression in Prostate Carcinomas and Its Relationship to Tumorigenicity. JNCI: Journal of the National Cancer Institute. 2002 Apr 3; 94(7):482-90. https://doi.org/10.1093/jnci/94.7.482 PMID: 11929948

29. Oldridge EE, Walker HF, Stower MJ, Simms MS, Mann VM, Collins AT, et al. Retinoic acid represses invasion and stem cell phenotype by induction of the metastasis suppressors RARRES1 and LXN. Oncogenesis. 2013 Apr; 2(4):e45-e45. https://doi.org/10.1038/oncosci.2013.6 PMID: 23588494

30. Kashwagi M, Friesse H, Uhl W, Berberat P, Abou-Shady M, Martignoni M, et al. Group II and IV phospholipase A2 are produced in human pancreatic cancer cells and influence prognosis. Gut. 1999 Oct; 45(4):605-12. https://doi.org/10.1136/gut.45.4.605 PMID: 10486373

31. Buhmeida A, Bendardaf R, Hilska M, Laine J, Collan Y, Laato M, et al. PLA2 (group IIA phospholipase A2) as a prognostic determinant in stage II colorectal carcinoma. Annals of Oncology. 2009 Jul 1; 20(7):1230-5. https://doi.org/10.1093/annonc/mdn783 PMID: 19276399

32. Cai H, Chiorean EG, Chiorean MV, Rex DK, Robb BW, Hahn NM, et al. Elevated Phospholipase A2 Activities in Plasma Samples from Multiple Cancers. PLOS ONE. 2013 Feb 22; 8(2):e57081. https://doi.org/10.1371/journal.pone.0057081 PMID: 23411159

33. Nallanthighal S, Rada M, Heiserman JP, Cha J, Sage J, Zhou J, et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. PLOS ONE. 2018 Nov 1; 13(11):e0206785. https://doi.org/10.1371/journal.pone.0206785 PMID: 30383866

34. Lambrecht D, Wauters E, Boeckx B, Adar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nature Medicine. 2018 Aug; 24(8):1277-89. https://doi.org/10.1038/s41591-018-0096-5 PMID: 29914354

35. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nature Communications. 2018 Sep 4; 9(1):3588. https://doi.org/10.1038/s41467-018-06052-0 PMID: 30189524

36. Shih AJ, Menzin A, Whyte J, Lovecchio J, Liew A, Khalili H, et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. PLOS ONE. 2018 Nov 1; 13(11):e0206785. https://doi.org/10.1371/journal.pone.0206785 PMID: 30383866

37. Oldridge EE, Walker HF, Stower MJ, Simms MS, Mann VM, Collins AT, et al. Retinoic acid represses invasion and stem cell phenotype by induction of the metastasis suppressors RARRES1 and LXN. Oncogenesis. 2013 Apr; 2(4):e45-e45. https://doi.org/10.1038/oncosci.2013.6 PMID: 23588494

38. Vázquez-Villa F, García-Ócaña M, Galván JA, García-Martínez J, García-Pravia C, Menéndez-Rodríguez P, et al. COL11A1 (procollagen 11A1) expression in proliferating pancreatic cancer cells and its relationship to tumorigenicity. JNCI: Journal of the National Cancer Institute. 2002 Apr 3; 94(7):482-90. https://doi.org/10.1093/jnci/94.7.482 PMID: 11929948

39. Jia D, Liu Z, Deng N, Tan TZ, Huang HY-J, Taylor-Harding B, et al. A COL11A1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets. Cancer Letters. 2016 Nov 28; 382(2):203-14. https://doi.org/10.1016/j.canlet.2016.09.001 PMID: 27690969

40. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Research. 2020 Sep; 30(9):745-62. https://doi.org/10.1038/s41422-020-0355-0 PMID: 32561858

41. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analysis for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 Apr 20; 43(7):e47-e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792
45. Quail DF, Dannenberg AJ. The obese adipose tissue microenvironment in cancer development and progression. Nature Reviews Endocrinology. 2019 Mar; 15(3):139–54. https://doi.org/10.1038/s41574-018-0126-x PMID: 30459447

46. Lengyel E, Makowski L, DiGiovanni J, Kolonin MG. Cancer as a Matter of Fat: The Crosstalk between Adipose Tissue and Tumors. Trends Cancer. 2018; 4(5):374–84. https://doi.org/10.1016/j.trecan.2018.03.004 PMID: 29709261

47. Cozzo AJ, Fuller AM, Makowski L. Contribution of Adipose Tissue to Development of Cancer. Compr Physiol. 2017 12; 8(1):237–82. https://doi.org/10.1002/cphy.c170008 PMID: 29357128

48. Sahai E, Astsaturov I, Cukierman E, DeNardo DG, Egeblad M, Evans RM, et al. A framework for advancing our understanding of cancer-associated fibroblasts. Nature Reviews Cancer. 2020 Mar; 20(3):174–86. https://doi.org/10.1038/s41574-019-0238-x PMID: 31980749

49. Dominguez CX, Müller S, Keerthivasan S, Koepchen H, Hung J, Gierke S, et al. Single-Cell RNA Sequencing Reveals Stromal Evolution into LRRC15+ Myofibroblasts as a Determinant of Patient Response to Cancer Immunotherapy. Cancer Discov. 2020 Feb 1; 10(2):232–53. https://doi.org/10.1158/2159-8290.CD-19-0644 PMID: 31699795

50. Öhlund D, Handy-Santana A, Biffi G, Elyada E, Almeida AS, Ponz-Sarvise M, et al. Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. J Exp Med. 2017 Mar 6; 214(3):579–96. https://doi.org/10.1084/jem.20162024 PMID: 28232471

51. Elyada E, Bolisetty M, Laise P, Flynn WF, Courtois ET, Burkhart RA, et al. Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. Cancer Discov. 2019; 9(8):1102–23. https://doi.org/10.1158/2159-8290.CD-19-0094 PMID: 31197017

52. Chen Z, Zhou L, Liu L, Hou Y, Xiong M, Yang Y, et al. Single-cell RNA sequencing highlights the role of inflammatory cancer-associated fibroblasts in bladder urothelial carcinoma. Nature Communications. 2020 Oct 8; 11(1):5077. https://doi.org/10.1038/s41467-020-18916-5 PMID: 33033240

53. Helms E, Onate MK, Sherman MH. Fibroblast Heterogeneity in the Pancreatic Tumor Microenvironment. Cancer Discov. 2020 May 1; 10(5):648–56. https://doi.org/10.1158/2159-8290.CD-19-1353 PMID: 32014869

54. Wang Y, Liang Y, Xu H, Zhang X, Mao T, Cui J, et al. Single-cell analysis of pancreatic ductal adenocarcinoma identifies a novel fibroblast subtype associated with poor prognosis but better immunotherapy response. Cell Discov. 2021 May 25; 7(1):1–17. https://doi.org/10.1038/s41421-020-00236-2 PMID: 33390590

55. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13; 177(7):1886–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031 PMID: 31178118

56. Kiselev YV, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics. 2019 May; 20(5):273–82. https://doi.org/10.1038/s41576-018-0088-9 PMID: 30617241

57. Thomas M. Cover JAT. Elements Of Information Theory 2nd Ed [Internet]. 2006 [cited 2021 Apr 6]. Available from: http://archive.org/details/ElementsOfInformationTheory2ndEd

58. Daub CO, Steuer R, Selbig J, Kloska S. Estimating mutual information using B-spline functions—an improved similarity measure for analysing gene expression data. BMC Bioinformatics. 2004 Aug 31; 5(1):116. https://doi.org/10.1186/1471-2105-5-116 PMID: 15339346

59. Zhu K, Ou Yang T-H, Dorie V, Zheng T, Anastassiou D. Meta-analysis of expression and methylation signatures indicates a stress-related epigenetic mechanism in multiple neuropsychiatric disorders. Translational Psychiatry. 2019 Jan 22; 9(1):1–12. https://doi.org/10.1038/s41398-018-0355-8 PMID: 30664621