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

Kaiyi Zhu[1,2,3,5,¶], Lingyi Cai[1,2,3,¶], Chenqian Cui[2] and 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]Current affiliation: Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

***** Corresponding author

E-mail: d.anastassiou@columbia.edu (DA)

[¶]These authors contributed equally to this work.

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.

Author summary

Computational analysis of rich gene expression data at the single-cell level from cancer biopsies can lead to biological discoveries about the nature of the disease. Using a computational methodology that identifies the gene expression profile of the dominant cell 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 *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.

1. Introduction

This research is a follow-up of our previous work in 2010 [1], in which we had identified a stage-associated signature of a particular type of cancer-associated fibroblasts (CAFs) by comparing the gene expression of lower stage with that of higher stage cancer samples. The signature, with prominent presence of genes *COL11A1*, *INHBA* and *THBS2*, was identical across cancer types and would only be present after a particular staging threshold, different in each cancer type, was reached. For example, the signature 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 never in 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 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 paper [3]); and the pancreatic data compare whole tumor tissue with normal pancreatic tissue (table 1 of the paper [4]).

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

	Breast		**Ovarian**		**Pancreatic**	
Rank	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

The four genes *COL11A1*, *INHBA*, *THBS2*, *COL5A2* appear among the top 15 in all three sets ($P = 6\times10^{-23}$ 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 similarity demonstrates that the signature is well-defined and associated with a universal biological mechanism in cancer. We had also found 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 in terms of the correlation of their expression with that of *COL11A1*. Accordingly, we had identified a *COL11A1*-correlated pan-cancer gene signature, listed in table 4 of [1], which we deposited in GSEA. We had referred to those 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."

Since then, many research results were published connecting one of the genes *COL11A1*, *INHBA*, *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 *POSTN*, *COL11A1*, *THBS2*, *COL5A2*, *ASPN*, *FAP*, *MMP13*, *VCAN*, *LUM*, *COL10A1*, *CTSK*, *COMP*, *CXCL14*, *FABP4*, *INHBA*. Similarly, the 24 characterizing genes of the "activated stroma subtype" of pancreatic cancer in fig. 2 of [17] are *SPARC*, *COL1A2*, *COL3A1*, *POSTN*, *COL5A2*, *COL1A1*, *THBS2*, *FN1*, *COL10A1*, *COL5A1*, *SFRP2*, *CDH11*, *CTHRC1*, *PNDC1*, *SULF1*, *FAP*, *LUM*, *COL11A1*, *ITGA11*, *MMP11*, *INHBA*, *VCAN*, *GREM1*, *COMP*. In both cases, the subtype-defining signature is clearly the *COL11A1*/*INHBA*/*THBS2*-expressing CAF signature.

In this work we show that the origin of those *COL11A1*-expressing CAFs is 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*. 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).

Our main computational methodology is attractor analysis (Materials and Methods), and is used in a novel manner for the analysis of rich single-cell RNA sequencing (scRNA-seq) data. The unsupervised attractor algorithm [18] 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 [19,20]. When applied on single cell data from a sample, it identifies the gene expression profiles of the dominant cell populations in the sample. Accordingly, we demonstrate that there is a continuous "ASC to *COL11A1*-expressing CAF transition."

We used publicly available scRNA-seq data sets from five different cancer types. We first focused on one of those datasets [21] from pancreatic ductal adenocarcinoma, because it is exceptionally rich, containing gene expression profiles from 57,530 cells in 24 tumor samples (T1-T24) and 11 normal control samples (N1-N11). Using attractor analysis, we established that (a)

the dominant fibroblastic population in each of the 11 normal samples is identical to that of the ASC type described above, and (b) that the gene expression profiles of the dominant fibroblastic population of the 24 tumor samples undergo a gradual change as the transition proceeds, starting from the state of the normal ASCs and passing through a continuum of intermediate states. We also validated these results by applying trajectory inference analysis. Finally, we validated our results in the other cancer types (head and neck, ovarian, lung, breast), suggesting the pan-cancer nature of the ASC to *COL11A1*-expressing CAF transition.

2. Results

2.1 Pancreatic Ductal Adenocarcinoma (PDAC) dataset

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). We used the general fibroblastic marker gene *LUM* as seed, and found that all samples yielded strong co-expression signatures involving many genes with big overlap among them. S1 Table 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 transition process, as described below.

2.1.1 Establishing the presence of *COL11A1*-expressing CAFs. Because it 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 with *COL11A1* and see if *INHBA* and *THBS2* are top ranked. Indeed, this happens in several tumor samples, as shown in Table 2 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.

Table 2. Ranked *COL11A1*-associated genes in five PDAC samples.

Rank	T23	MI	T11	MI	T6	MI	T15	MI	T18	MI
1	COL11A1	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	C5orf46	0.2753
6	COL1A2	0.3099	C1QTNF3	0.2155	TWIST1	0.1714	C1QTNF3	0.2514	PPAPDC1A	0.2668
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	IGFL2	0.1178	COL8A1	0.2325	IGFL2	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	0.2449	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	STMN2	0.1926
18	COL6A3	0.2339	INHBA	0.1696	P4HA3	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

2.1.2 Dominant fibroblastic populations in the tumor samples. Table 3 shows the ranked lists of genes for the attractors in each of the rearranged tumor samples. There is a remarkable continuity in the shown expression profiles. The samples at the right side of the table include *COL11A1* at increasingly higher ranks. On the other hand, we identified gene *APOD* (Apolipoprotein D) as particularly significant, as it is highly ranked on the left side and its rank gradually decreases as we go to the right. 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. Furthermore, all genes shown in Table 3 have high strengths and therefore are co-expressed inside the same individual cells in each sample.

Table 3. Rearranged PDAC tumor samples showing the continuously changing pattern of the signature profile.

Genes *APOD* and *COL11A1* are highlighted in green and red, respectively.

Rank	T2	T14	T19	T3	T10	T15	T18	T12	T6	T24	T1	T5	T4	T7	T22	T9	T11	T23	T21	T17	T8
1	LUM	DCN	SFRP2	MMP2	PDGFRA	SFRP2	DCN	MMP2	COL10A1	DCN	SFRP2	PDGFRA	COL10A1	CYP1B1	COL1A2	LUM	LUM	COL1A1	COL10A1	COL10A1	COL11A1
2	APOD	APOD	APOD	LUM	HTRA3	LUM	SFRP2	LUM	PDGFRA	LUM	VCAN	CYP1B1	SFRP2	SFRP2	PDGFRA	DCN	DCN	COL1A2	CTHRC1	CTHRC1	COL10A1
3	VCAN	LUM	LUM	APOD	DPT	DCN	LUM	PDGFRA	SFRP2	FBLN1	LUM	SFRP2	COL1A1	COL8A1	THBS2	RARRES2	CTHRC1	COL3A1	THBS2	COL11A1	CREB3L1
4	SFRP4	SFRP4	IGF1	DCN	APOD	VCAN	C3	CTHRC1	CYP1B1	VCAN	PDGFRA	SFRP4	MMP2	PDGFRA	MMP2	CTHRC1	SFRP2	COL6A3	GJB2	ISLR	RP11-400N13.3
5	SFRP2	TSHZ2	EFEMP1	FBLN1	MEG3	APOD	MMP2	ITGBL1	MMP2	SFRP4	COL1A2	DPT	LUM	COL10A1	COL1A1	SFRP2	COL10A1	LUM	SFRP2	MMP2	SFRP2
6	MMP2	HTRA3	PDGFRA	VCAN	OMD	FBLN1	APOD	EFEMP1	VCAN	COL1A2	EFEMP1	LUM	COL1A2	SFRP4	COL3A1	AEBP1	RARRES2	FN1	COL11A1	COL1A1	BASP1
7	RARRES1	FBLN1	OGN	FBLN5	ITGBL1	MMP2	EFEMP1	SFRP2	CTHRC1	MMP2	DCN	MEG3	CTHRC1	CTHRC1	ITGBL1	COL10A1	AEBP1	COL5A2	CCDC80	COL1A2	PDPN
8	C3	MMP2	SFRP4	PDGFRA	PAPPA	COL6A3	MFAP4	FBLN5	LUM	SFRP2	CCDC80	EFEMP1	DCN	APOD	COL10A1	NBL1	NBL1	VCAN	NBL1	COL3A1	BNC2
9	MEG3	COL6A3	VCAN	SFRP2	MRC2	COL1A1	FBLN1	VCAN	SFRP4	C1R	ISLR	VCAN	CTSK	MMP2	CTHRC1	MMP2	CTSK	COL5A1	DCN	AEBP1	C5orf46
10	HTRA3	VCAN	CTSK	C3	LSAMP	COL1A2	SFRP4	APOD	APOD	C1S	SFRP4	IGF1	MFAP2	PLXDC2	LUM	CTSK	VCAN	THBS2	AEBP1	MMP11	PLXDC2
11	FBLN1	GPC3	COL1A2	MGP	CYP1B1	ISLR	CCDC80	COL8A1	PLXDC2	APOD	COL6A3	FBLN5	APOD	VCAN	SFRP2	THBS2	CTGF	SFRP2	INHBA	THBS2	SPOCK1
12	MGP	CTGF	STEAP1	EFEMP1	COL10A1	COL10A1	RARRES1	STEAP1	COL8A1	CTSK	CYP1B1	SERPINE2	MATN3	BNC2	EFEMP1	FBLN1	COL8A1	CTHRC1	LUM	COL12A1	ADM
13	DCN	SFRP2	MMP2	OMD	COL8A1	CTHRC1	C1S	COL1A2	FBLN1	CCDC80	COL1A1	FBLN1	MEG3	MRC2	MRC2	VCAN	THBS2	MMP2	FBLN1	HTRA1	MMP2
14	CYP1B1	C1S	CYP1B1	RP11-572C15.6	PDPN	CCDC80	VCAN	PTGDS	OMD	ISLR	APOD	SCN7A	ISLR	DPYSL3	COL6A3	CCDC80	MMP2	COL10A1	MMP2	MMP14	ARL4C
15	COL1A2	OMD	PTGDS	CCDC80	CXCL14	SFRP4	PTGDS	ISLR	THBS2	C3	CLDN11	LTBP2	FBLN1	COL1A2	PDPN	COL1A1	COL11A1	COL11A1	COL6A3	SFRP2	MEG3
16	MOXD1	SPON2	FBLN1	IGF1	MMP23B	CTSK	C1R	LXN	FAP	EFEMP1	CTSK	APOD	COL11A1	FBLN5	VCAN	S100A6	INHBA	AEBP1	OMD	SULF1	GJA1
17	PTGDS	C3	RARRES1	TSHZ2	ABCA9	COL3A1	CTSK	OLFML3	MFAP2	MGP	FBLN1	PTGDS	COL3A1	CREB3L1	DPYSL3	COL8A1	C1QTNF3	COL12A1	MEG3	LUM	VCAN
18	FBLN5	F2R	DCN	ITM2A	LUM	S100A10	PDGFRA	THBS2	COL6A3	TSHZ2	MMP2	ISLR	SFRP4	COL3A1	LOX	TMSB10	MATN3	DCN	COL8A1	SDC1	FIBIN
19	PDGFRA	ANKH	COL3A1	SFRP4	PDGFRL	THBS2	SERPINF1	FBLN1	RARRES1	PDGFRA	COL3A1	PODN	CXCL14	OMD	GJA1	C1S	SFRP4	SPARC	ISLR	DCN	COL1A2
20	COL6A3	CTSK	MFAP5	RARRES1	STXBP6	HTRA1	MOXD1	MGST1	EFEMP1	OMD	PTGDS	DCN	ITGBL1	RARRES1	APOD	COL1A2	HTRA1	TMSB10	MMP11	MFAP5	ZFHX4
21	CTHRC1	C1R	FBLN5	COL8A1	SVEP1	SEMA3C	GPNMB	MEG3	COL1A2	FBLN5	FBLN2	MMP2	COL6A3	PODN	FAP	COL3A1	ITGBL1	MMP14	MFAP5	VCAN	MFAP2
22	FAP	IGFBP3	COL1A1	OGN	BICC1	FBLN2	RARRES2	PDPN	ANKH	COL1A1	CTHRC1	MGP	VCAN	LSAMP	MXRA5	HTRA1	MFAP5	SDC1	PPAPDC1A	COL6A3	MME
23	F2R	MOXD1	ISLR	C7	ABCA6	LRP1	FBLN5	MFAP2	FNDC1	OGN	C3	C7	IGFL2	THBS2	PODN	CD99	CXCL14	POSTN	CTSK	GJB2	MFAP5
24	ISLR	CTHRC1	C3	DPT	MFAP2	MRC2	ISLR	DPT	SPON2	SERPINF1	PLXDC2	MGST1	RARRES2	SVEP1	CXCL14	ISLR	GJB2	FBLN1	MXRA5	GREM1	RAB3B
25	TIMP1	MEG3	MEG3	BICC1	BNC2	PDPN	ITGBL1	C3	PDPN	CTGF	RARRES1	SPOCK1	FNDC1	ITGBL1	COL8A1	SERPINF1	IGFBP3	INHBA	FNDC1	TIMP2	ITGBL1
26	C7	PDGFRA	CTHRC1	CTHRC1	WNT5A	MXRA5	COL10A1	MRC2	CTSK	FBLN2	RP11-572C15.6	CCDC80	COL5A1	PTGDS	SFRP4	TSC22D3	CCDC80	SERPINH1	COL1A1	COL5A2	GJB2
27	PHLDA3	FBLN5	MOXD1	PODN	CST3	OMD	CYP1B1	PTGIS	MFAP5	COL6A3	THBS2	SLC19A2	OMD	FAP	SEMA3C	FTL	CD99	MXRA5	SDC1	FAP	COL3A1
28	OMD	ITM2A	COL6A3	COL6A3	SFRP2	ITGBL1	C7	MOXD1	HTRA3	PODN	C7	HTRA3	CST4	LUM	LRP1	MFAP2	C1S	HTRA1	VCAN	MFAP2	NTM
29	FBLN2	COL1A2	MGP	CXCL14	MOXD1	FBN1	PLXDC2	HSD11B1	LRP1	LTBP2	LAMA2	MOXD1	INHBA	FBLN1	NTM	ANXA2	LOXL1	MMP11	GREM1	CTSK	PDLIM4
30	SCN7A	PTGDS	C7	COL1A2	ZFHX4	RARRES2	CTHRC1	SFRP4	COL1A1	ITGBL1	OLFML3	ITGBL1	GJB2	SULF1	C3	LAPTM4A	FIBIN	ISLR	PDPN	COL5A1	CMTM8
31	EFEMP1	RARRES1	CILP	ISLR	RARRES1	PLXDC2	DPT	LSAMP	DIO2	C7	MFAP4	CTSK	IGFBP3	INHBA	COL11A1	NNMT	FBLN1	MEG3	FBLN2	PPAPDC1A	TANC2
32	COL10A1	OLFML3	COL8A1	MEG3	BOC	CXCL14	CLU	PLXDC2	DCN	PLXDC2	SLIT2	FBLN2	HTRA3	SEMA3C	UNC5B	C1R	PALLD	TIMP2	MFAP2	FN1	NT5E
33	SERPINF1	OGN	THBS2	CTSK	PODN	PDGFRA	NPC2	MFAP4	ALDH1A3	MOXD1	IGF1	CTHRC1	CST1	LOX	LOXL1	RPL27A	SDC1	FSTL1	C1QTNF3	TMEM158	TENM3
34	BNC2	ITGBL1	MRC2	HSD11B1	TMEM119	MATN3	MEG3	OGN	LOX	TMEM119	LRP1	SVEP1	FBLN2	PDPN	COL5A2	INHBA	MFAP2	COL6A2	COL5A2	POSTN	EPDR1
35	CTSK	PTCH1	NR2F1	CTGF	PTGIS	COL8A1	RP11-572C15.6	COL6A3	FBLN2	COL8A1	PDLIM3	MXRA5	MFAP5	IGF1	SCARA3	NUPR1	ANXA2	CTSK	CDH11	ANTXR1	MYH10
36	MRC2	COL8A1	ITGBL1	C1S	OGN	NBL1	MGP	SPON2	OLFML3	LRP1	LTBP2	STEAP1	MXRA5	LAMP5	COL8A2	LGALS1	APOD	MFAP2	LOX	PLAU	LOX
37	TSHZ2	BOC	TSHZ2	COL10A1	FAP	TMEM119	COL6A3	COL1A1	TMEM119	RARRES1	FBLN5	NEGR1	THBS2	FBLN2	PTGDS	COL6A3	OMD	MFAP5	COL3A1	INHBA	COL8A1
38	LRP1	SERPINF1	STEAP2	SERPINE2	EFEMP1	HTRA3	LTBP2	LTBP2	MEG3	BOC	MGP	C3	CTGF	OGN	CCDC80	COL11A1	ISLR	GAS1	PDGFRA	GJA1	EVA1A
39	SLIT2	MGP	MFAP2	SERPINF1	LAMA2	FAP	AEBP1	PODN	FGF7	CYBRD1	MEG3	RARRES1	MRC2	PTGFRN	ALDH1A3	OMD	SLC6A6	LRRC15	F13A1	LGALS1	MXRA5
40	COL1A1	IGFBP6	PDGFRL	LTBP2	GSTM5	MEG3	PDPN	DCN	TMSB10	LXN	BICC1	RP11-572C15.6	GJA1	GAS7	FGFR1	CD55	CYR61	COL8A2	FRMD6	PTK7	BICC1
41	SVEP1	TIMP1	PDPN	NEGR1	IGF1	FNDC1	COL1A1	CTSK	CXCL14	LOX	MRC2	PLXDC2	FAP	CXCL14	BOC	PDPN	COL1A2	FBLN2	APOD	CD99	C1orf198
42	THBS2	MFAP4	PLXDC2	MOXD1	F2R	NTM	COL8A1	SPOCK1	ITGBL1	HTRA3	FGFR1	LRP1	SPON2	COL11A1	PLXDC2	LOXL1	MMP11	GREM1	COL1A2	NBL1	INHBA
43	TMEM119	CLEC11A	OMD	C1R	F3	DPYSL3	S100A13	COL3A1	IGFBP3	RP11-572C15.6	ABL2	ABI3BP	BICC1	FNDC1	MFAP2	FIBIN	MEG3	APOD	DIO2	NTM	PDGFC
44	IGFBP3	INHBA	MMP23B	FGF7	MMP2	SLC6A6	OGN	RARRES1	CDH11	CYP1B1	MXRA5	FAP	EMP1	HTRA3	MFAP5	NTM	PLXDC2	COL8A1	CD55	RARRES2	FBLN2
45	C1S	MXRA8	LXN	COL1A1	SFRP4	CYP1B1	MMP23B	SVEP1	IGF1	CTHRC1	RGS2	COL6A3	FGF7	MOXD1	TMSB10	S100A10	FBLN2	CD99	HTRA1	FBLN1	B4GALT1

2.1.3 Dominant fibroblastic population in the normal samples. Table 4 shows the attractor metagenes of the 11 normal samples by listing the 20 top-ranked genes for each sample, in which commonly shared genes are color-coded. There is a striking similarity among them, indicating that they represent a stable and 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 ($P < 10^{-7}$ by permutation test). In addition to fibroblastic markers, there are several strongly expressed adipose-related or stemness-related genes in the list, such as *APOD*, *CFD*, *CXCL12* and *DPT*, revealing that they are ASCs. Consistently, Gene Set Enrichment Analysis 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 [22].

Table 4. Top 20 genes of the identified attractors for each pancreatic normal sample (N1-N11).

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

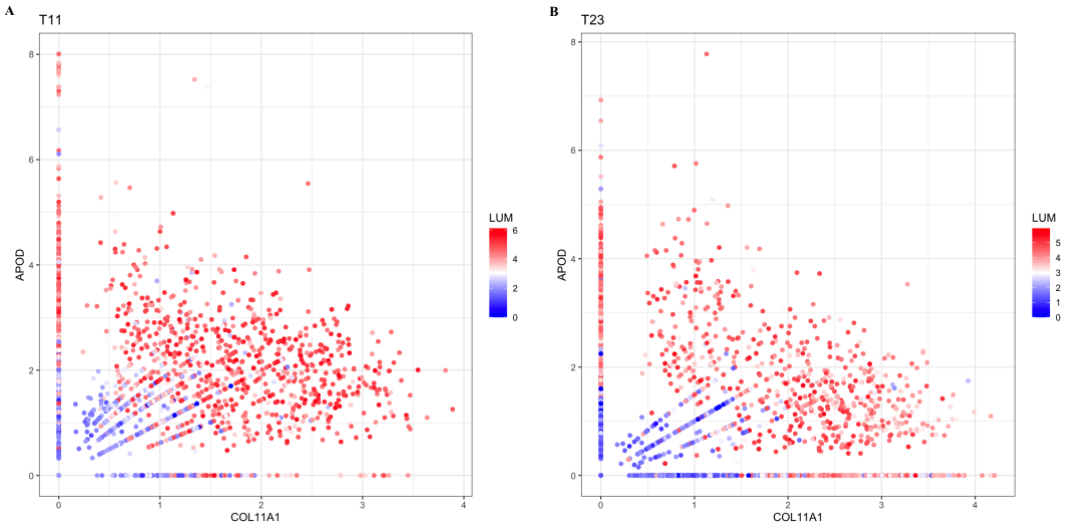
To investigate the nature of this ASC population, we referred to recent results from single-cell analysis of general human adipose tissue [23]. We used the dataset with the single-cell expression profiles of all 26,350 cells taken from the SVF of normal adipose tissue from 25 samples. The first column in Table 5 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 naturally occurring cell population.

Table 5. 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	OLFLM3
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

2.1.4 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 fibroblastic subtypes), Fig 1 shows scatter plots, color-coded for the expression of fibroblastic marker *LUM*, for the fibroblast-rich samples T11 and T23, demonstrating the presence in each sample of cells representing the full range of the continuously varying transition from *APOD* to *COL11A1*-expressing cells.

Fig 1. Scatter plots for fibroblast-rich samples for patients T11 and T23.



To further investigate the continuous transition, we partitioned the 34 pancreatic samples into three groups, considering whether the corresponding attractor includes *COL11A1* among the top 100 genes (S1 Table). Group 1 includes the eleven normal samples (N1 to N11). Group 2 includes twelve tumor samples in which *COL11A1* is not included (T1, T2, T3, T5, T10, T12, T13, T14, T16, T18, T19, T24). Group 3 includes eleven cancer samples in which it is included (T4, T6, T7, T8, T9, T11, T15, T17, T21, T22, T23). 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), so we can observe the genes whose ranking changes significantly during the transition.

Table 6. Top 50 genes of the consensus attractors for three different PDAC sample groups. Group1: normal samples; Group 2: tumor samples in which the dominant fibroblastic population does not include *COL11A1*; Group 3: tumor samples in which the dominant fibroblastic population includes *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

Although there are many shared genes, the groups have distinct gene rankings. Group 1 (normal samples) contains many adipose-related genes, consistent with Table 4. Group 3 contains, in addition to *COL11A1*, many among the other CAF genes, such as *THBS2*, *INHBA*, *MMP11*, *AEBP1*, *MFAP5*, *COL10A1* and *VCAN*.

Group 2 displays an intermediate state, including 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 transition 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 [24]. 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 [25,26]).

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 S2 Table). 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 transition.

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 upregulated 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 genes are also strongly upregulated, 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 upregulated 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.

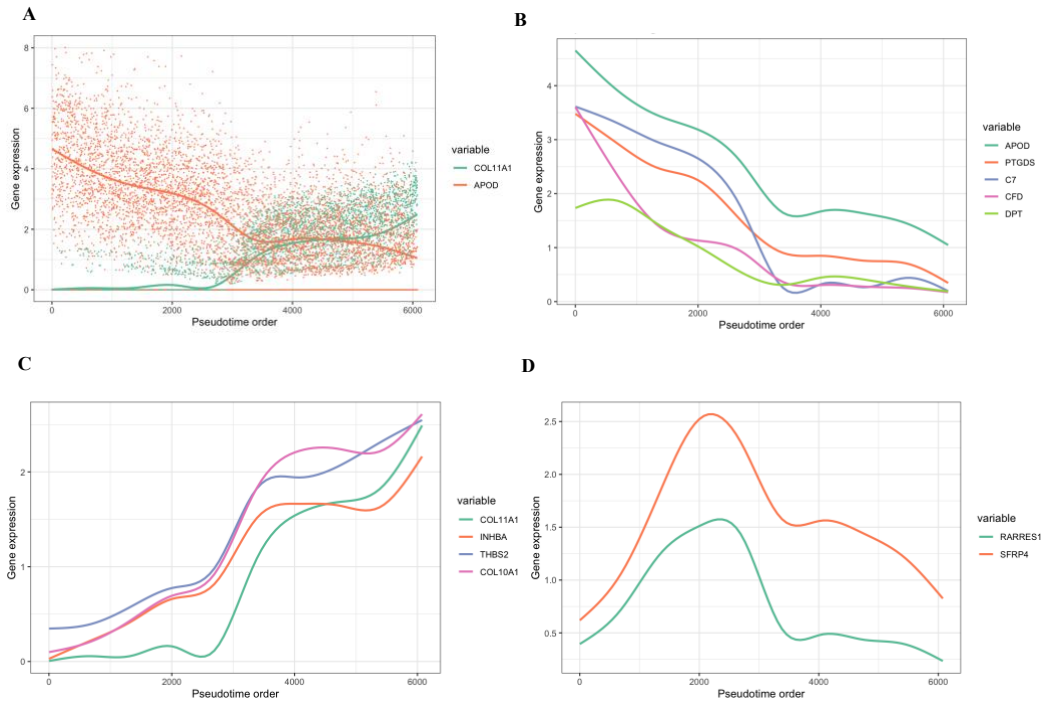
The top differentially expressed gene *PLA2G2A* is not among the top genes of any attractors we identified and is expressed by less than half of cells, even in Group 2 in which it appears. Therefore, it is expressed by cells in Group 2 that are not among the ones undergoing the ASC to *COL11A1*-expressing CAF transition, although it probably still plays an important related role and many previous studies referred to its effects on prognosis of multiple cancer types [27–29]. 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 is associated with the final *COL11A1*-expressing stage of the transition [30]. These results suggest that lipid metabolic reprogramming plays an important role in the metastasis-associated biological mechanism [31], providing energy for the metastasizing tumor cells.

2.1.5 Trajectory Inference. We independently applied trajectory inference analysis on the PDAC fibroblasts by using the Slingshot [32] method. 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 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, 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, S2A Fig and S2B Fig 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 transition. 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).

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.



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 *RARRES1* and *SFRP4* (Fig 2D) increased consistently at the beginning and then decreased after reaching a peak, suggesting that they play important roles in the differentiation path.

2.2 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) [33], ovarian cancer [34], lung cancer [35] and breast cancer [36].

The *COL11A1*-expressing CAF signature has been confirmed to be a pan-cancer signature [37–39]. 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 *RARRES1*, *SFRP4* and *THBS2*, in addition to the core ASC genes, demonstrating that they are at an intermediate stage of the transition.

2.2.1 Head and neck squamous cell carcinoma. For the HNSCC dataset, the authors of the paper presenting the data [33] 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 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.

2.2.2 Ovarian cancer. For the ovarian dataset, the clustering results showed two clusters (X6 and X9) expressing *COL11A1*-associated genes and ASC-associated genes, respectively (S3B Fig, S6 Table; 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. 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*.

2.2.3 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*.

2.2.4 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.

2.3 Potential therapeutic targets

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 [40]. 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 2).

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. Consistently, gene *RARRES1* (aka *TIG1*) was found to play an important, though not yet clear, role in regulating proliferation and differentiation of adipose tissue derived mesenchymal stem cells [24], while *SFRP4* is a Wnt pathway regulator whose expression has been found associated with various cancer types [41,42].

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 [43] (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 top-ranked (top1) 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*. 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).

3. Discussion

One contribution of our work is that it provides an explanation to the fact that adipose tissue contributes to the development and progression of cancer as a result of its interaction with malignant cells [44–46].

Another contribution is that it sheds light on one important mechanism within the heterogeneous tumor microenvironment, which also contains fibroblasts that do not participate in the transition.

There are several published results, often conflicting with each other, identifying various distinct subtypes of fibroblasts, resulting from computational clustering methods following dimensionality reduction, or using particular marker genes (such as *αSMA*, which, interestingly we found that is not consistently expressed in the *COL11A1*-expressing CAFs, therefore it should not be used as its marker), which represent an approximation of biological reality. Examples include the hC1 and hC0 clusters in [47], the C9 and C10 clusters in [39], the CAF2 and CAF1 clusters in [33], the iCAF and myCAF clusters in [48,49] and the iCAF an mCAF clusters in [50]. A review of such results in pancreatic cancer appears in [51].

As an example of conflicting results, the "iCAFs" identified in [50] are identical to the normal ASCs (fig. 3b of [50]) 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 5, 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. On the other hand, the iCAFs identified in [49] also have significant overlap in their marker composition with the ASCs (such as *CFD*, *DPT* and *CXCL12*), while omitting essential ASC markers such as *APOD*, and including additional genes.

On the other hand, our work is consistent with, and complementary to the results of [47] 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, observing and confirming the presence of cells expressing particular combinations of genes, 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.

4. Materials and methods

4.1 Datasets availability

The pancreatic dataset [21] was downloaded from the Genome Sequence Archive with accession number CRA001160. The four validation datasets of other cancer types are also publicly available: HNSCC [33] (GSE103322), ovarian [34] (GSE118828), lung cancer [35] (E-MTAB-6149 and E-MTAB-6653), breast cancer [36] (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).

4.2 Data processing and cell identification

We applied the Seurat (v3.1.4) [52] 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, and *RGS5* 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.

4.3 Mutual information

Mutual information (MI) is a general measure of the association between two random variables [53]. We used a spline based estimator [54] 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 [18]. We used the getMI or getAllMIWz function implemented in the cafr R package with parameter negateMI = TRUE.

4.4 Attractor-based analysis

For single dataset, we applied the attractor finding algorithm using the findAttractor function implemented in the cafr (v0.312) R package [18] with gene *LUM* as seed. 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. 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 [55]. 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.

4.5 Trajectory inference (TI) analysis

We selected the Slingshot [32] method for TI analysis, based on its robustness and suggestions made by the benchmarking pipeline dynverse [56]. 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.

4.6 Statistical analysis

4.6.1 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}.

4.6.2 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.

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.

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 Med Genomics. 2010 Nov 3;3:51.

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.

3. Bignotti E, Tassi RA, Calza S, Ravaggi A, Bandiera E, Rossi E, et al. Gene expression profile of ovarian serous papillary carcinomas: identification of metastasis-associated genes. American Journal of Obstetrics & Gynecology. 2007 Mar 1;196(3):245.e1-245.e11.

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.

5. Wang M, Zhao Y, Zhang B. Efficient Test and Visualization of Multi-Set Intersections. Scientific Reports. 2015 Nov 25;5(1):16923.

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. Oncol Rep. 2016 Aug;36(2):877–85.

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.

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 Sep 15;6(27):23748–63.

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. J Hematol Oncol [Internet]. 2017 Jan 25 ;10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5264454/

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. J Thorac Oncol. 2009 Apr;4(4):455–62.

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;29(1):77–83.

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. J Invest Surg. 2020 Nov 5;1–6.

13. Wang X, Zhang L, Li H, Sun W, Zhang H, Lai M. THBS2 is a Potential Prognostic Biomarker in Colorectal Cancer. Scientific Reports. 2016 Sep 16;6(1):33366.

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.

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 1;11(4):388–96.

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 2;123(1):517–25.

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.

18. Cheng W-Y, Ou Yang T-H, Anastassiou D. Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput Biol. 2013;9(2):e1002920.

19. Cheng W-Y, Ou Yang T-H, Anastassiou D. Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. Science Translational Medicine. 2013 Apr 17;5(181):181ra50-181ra50.

20. McCarthy N. Rising to the challenge. Nature Reviews Cancer. 2013 Jun;13(6):378–378.

21. 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.

22. Boquest AC, Shahdadfar A, Frønsdal 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. Mol Biol Cell. 2005 Mar;16(3):1131–41.

23. 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.

24. 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 Prolif. 2009 Feb 24;42(3):309–16.

25. 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. J Natl Cancer Inst. 2002 Apr 3;94(7):482–90.

26. 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.

27. Kashiwagi M, Friess 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.

28. 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.

29. 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.

30. Nallanthighal S, Rada M, Heiserman JP, Cha J, Sage J, Zhou B, et al. Inhibition of collagen XI alpha 1-induced fatty acid oxidation triggers apoptotic cell death in cisplatin-resistant ovarian cancer. Cell Death & Disease. 2020 Apr 20;11(4):1–12.

31. Koundouros N, Poulogiannis G. Reprogramming of fatty acid metabolism in cancer. British Journal of Cancer. 2020 Jan;122(1):4–22.

32. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018 Jun 19;19(1):477.

33. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017 Dec 14;171(7):1611-1624.e24.

34. 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.

35. Lambrechts D, Wauters E, Boeckx B, Aibar 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.

36. 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.

37. Vázquez-Villa F, García-Ocaña M, Galván JA, García-Martínez J, García-Pravia C, Menéndez-Rodríguez P, et al. COL11A1/(pro)collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. Tumor Biol. 2015 Apr 1;36(4):2213–22.

38. Jia D, Liu Z, Deng N, Tan TZ, Huang RY-J, Taylor-Harding B, et al. A COL11A1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets. Cancer Lett. 2016 28;382(2):203–14.

39. 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 Jun 19;1–18.

40. Yeung T-L, Leung CS, Yip K-P, Sheng J, Vien L, Bover LC, et al. Anticancer Immunotherapy by MFAP5 Blockade Inhibits Fibrosis and Enhances Chemosensitivity in Ovarian and Pancreatic Cancer. Clin Cancer Res. 2019 Nov 1;25(21):6417–28.

41. Drake J, Shearwood AM, White J, Friis R, Zeps N, Charles A, et al. Expression of secreted frizzled-related protein 4 (SFRP4) in primary serous ovarian tumours. Eur J Gynaecol Oncol. 2009;30(2):133–41.

42. Sandsmark E, Andersen MK, Bofin AM, Bertilsson H, Drabløs F, Bathen TF, et al. SFRP4 gene expression is increased in aggressive prostate cancer. Scientific Reports. 2017 Oct 27;7(1):14276.

43. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

44. Quail DF, Dannenberg AJ. The obese adipose tissue microenvironment in cancer development and progression. Nature Reviews Endocrinology. 2019 Mar;15(3):139–54.

45. Lengyel E, Makowski L, DiGiovanni J, Kolonin MG. Cancer as a Matter of Fat: The Crosstalk between Adipose Tissue and Tumors. Trends in Cancer. 2018 May 1;4(5):374–84.

46. Cozzo AJ, Fuller AM, Makowski L. Contribution of Adipose Tissue to Development of Cancer. Compr Physiol. 2017 Dec 12;8(1):237–82.

47. Dominguez CX, Müller S, Keerthivasan S, Koeppen 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.

48. Öhlund D, Handly-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.

49. 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 Aug 1;9(8):1102–23.

50. 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.

51. Helms E, Onate MK, Sherman MH. Fibroblast Heterogeneity in the Pancreatic Tumor Microenvironment. Cancer Discov. 2020 May 1;10(5):648–56.

52. 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):1888-1902.e21.

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

54. 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;5:118.

55. 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.

56. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nature Biotechnology. 2019 May;37(5):547–54.

Supplementary Material

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.

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.

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.

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

S2 Table. Differentially expressed genes of between the fibroblasts from different PDAC sample groups.

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

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

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

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

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

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

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.