

## University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

# University of Southampton

Faculty of Medicine

Cancer Sciences

**Characterising cancer-associated fibroblast heterogeneity in non-small cell lung  
cancer: relating molecular phenotype to function**

by

**Dr. Sara Waise BA MBBS**

ORCID ID 000-0003-0929-3121

Thesis for the degree of Doctor of Philosophy

May 2020

# University of Southampton

## Abstract

Faculty of Medicine

Cancer Sciences

Thesis for the degree of Doctor of Philosophy

Characterising cancer-associated fibroblast heterogeneity in non-small cell lung cancer:  
relating molecular phenotype to function

by

Dr. Sara Waise

Despite recent therapeutic advances, non-small cell lung cancer (NSCLC) remains the leading cause of cancer death worldwide. To improve survival outcomes in this disease, novel therapeutic approaches are required. Relative to other cancers, NSCLC show low tumour purity with high proportions of immune and stromal cell populations. Cancer-associated fibroblasts (CAFs) are the most common stromal cell type in a range of solid tumours, where they have a number of tumour-promoting effects and are frequently associated with poor outcome. To date, therapeutic interventions targeting CAFs have shown largely disappointing results: this may be due in part to a lack of understanding of the variation within the CAF population.

The aim of this work was to characterise the heterogeneity in the CAF population in NSCLC. First, we optimised our approach for the isolation of fibroblasts from primary lung tissues, determining that prolonged incubation with Collagenase is required. We next devised pipelines for the quality control of single-cell RNA sequencing data, identifying low-quality droplets and transcriptomic changes induced by prolonged enzymatic incubation. This approach was applied to 12 NSCLC and 6 patient-matched normal samples processed using the Drop-seq platform. Combining the resulting stromal cell data with those from a NSCLC dataset published during the course of this project identified 9 distinct populations, including 4 CAF groups. Two CAF populations showed overlap with the commonly-described “myofibroblastic” phenotype, and may have roles in the deposition and remodelling of extracellular matrix. Further functional characterisation requires *in vitro* work: as fibroblast culture is known to impact gene expression, we used the results of the above analyses to inform *in vitro* recapitulation of the identified *ex vivo* phenotypes. This approach allowed partial recreation of *ex vivo* gene expression profiles, although to an insufficient extent for preliminary functional analysis, and requires further refinement to allow accurate characterisation. Such studies should facilitate the development of more specific and successful stromal targeting strategies in NSCLC.





# Table of Contents

<b>Table of Contents .....</b>	<b>i</b>
<b>Table of Tables .....</b>	<b>vii</b>
<b>Table of Figures .....</b>	<b>ix</b>
<b>List of Accompanying Materials .....</b>	<b>xiii</b>
<b>Research Thesis: Declaration of Authorship .....</b>	<b>xv</b>
<b>Acknowledgements .....</b>	<b>xvii</b>
<b>Definitions and Abbreviations .....</b>	<b>xix</b>
<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1 Introduction.....	1
1.2 Fibroblast heterogeneity in normal tissues .....	3
1.3 Definitions and origins.....	4
1.4 $\alpha$ -SMA and the “myofibroblastic” CAF.....	5
1.5 Alternative CAF subtypes and markers .....	6
1.6 Overlap between markers .....	9
1.7 Population-based analysis of fibroblast gene expression profiles.....	10
1.8 Single-cell RNA sequencing for transcriptomic characterisation.....	12
1.9 CAF gene expression profiling by single-cell RNA sequencing.....	13
1.10 CAFs as a therapeutic target.....	14
1.11 CAFs in non-small cell lung cancer.....	16
1.12 Discussion .....	18
1.13 Aims and objectives.....	19
<b>Chapter 2 Methods.....</b>	<b>21</b>
2.1 Cell culture.....	21
2.1.1 Cell culture principles .....	21
2.1.2 Freezing cells .....	22
2.1.3 Defrosting cells for culture .....	23
2.1.4 Counting cells .....	23
2.1.5 <i>Mycoplasma</i> polymerase chain reaction (PCR) .....	23
2.1.6 Isolation and culture of primary cells .....	25

## Table of Contents

2.1.7	Coating culture plates.....	25
2.1.8	Three-dimensional cultures.....	25
2.1.9	Gel contraction assays.....	26
2.1.10	Migration assays.....	26
2.2	Protein analysis.....	27
2.2.1	Protein extraction and quantitation.....	27
2.2.2	Polyacrylamide gel electrophoresis.....	27
2.2.3	Western blotting.....	28
2.3	Real-time quantitative polymerase chain reaction (RT-PCR) .....	29
2.3.1	RNA extraction.....	29
2.3.2	cDNA synthesis .....	30
2.3.3	Quantitative real-time PCR.....	30
2.4	Assessment of metabolic activity .....	32
2.5	Tissue disaggregation .....	32
2.6	Fluorescence-activated cell sorting (FACS).....	33
2.7	Single-cell RNA sequencing (scRNA-seq) .....	35
2.7.1	Capturing single-cell transcriptomes .....	35
2.7.2	cDNA synthesis, PCR and sequencing.....	37
2.7.3	Sequence alignment, transcript identification and quantification .....	40
2.8	Bioinformatic analysis.....	40
2.8.1	Quality control.....	40
2.8.2	Dimensionality reduction .....	41
2.8.3	Clustering.....	42
2.8.4	Identification of marker genes .....	42
2.8.5	Cell type identification.....	42
2.8.6	Merging datasets.....	43
2.8.7	Gene set enrichment analysis (GSEA).....	43
2.8.8	Trajectory analysis.....	43
2.9	Histological processing and analysis.....	44
2.9.1	Immunohistochemical staining.....	44
2.9.2	Digital Pathology processing.....	45

2.10 Patient characteristics .....	45
2.11 Statistics.....	45
<b>Chapter 3 Results 1: Optimising primary tissue dissociation for single-cell analysis....</b>	<b>49</b>
3.1 Introduction.....	49
3.2 Optimising FACS analysis of fibroblasts .....	49
3.3 Tissue disaggregation using Collagenase P for 60 minutes is required to increase the number of fibroblasts isolated from tissue samples .....	51
3.4 Lung disaggregation with Collagenase P for sixty minutes is compatible with single- cell RNA sequencing using the drop-seq platform .....	56
3.5 Discussion .....	56
<b>Chapter 4 Results 2: Single-cell RNA sequencing and “lineage clustering” .....</b>	<b>59</b>
4.1 Introduction.....	59
4.2 Identifying distinct lineages with single-cell RNA sequencing.....	59
4.3 Optimising the analysis pipeline .....	62
4.3.1 Determining quality-control thresholds .....	62
4.3.2 Identifying low-quality droplets .....	65
4.3.3 Identifying disaggregation-associated gene expression changes.....	73
4.4 Single-cell RNA-seq identifies distinct cell lineages in primary tissue .....	75
4.5 Discussion .....	78
<b>Chapter 5 Results 3: Analysing lung fibroblast phenotypes and transdifferentiation mechanisms .....</b>	<b>81</b>
5.1 Introduction.....	81
5.2 Identifying stromal populations in normal and malignant tissues .....	82
5.2.1 Identifying stromal markers .....	82
5.2.2 Stromal cell filtering .....	83
5.3 Characterising fibroblast phenotypes in normal and malignant tissues .....	83
5.3.1 Sub-lineage clustering .....	84
5.3.2 Inferring stromal phenotypes from differential gene expression and gene set enrichment analysis.....	85
5.3.3 Histological validation of <i>ex vivo</i> fibroblast populations in fixed tissues.....	88

## Table of Contents

5.4	Fibroblast trajectory analysis.....	90
5.5	Discussion .....	94
<b>Chapter 6</b>	<b>Results 4: Recreating <i>ex vivo</i> fibroblast phenotypes .....</b>	<b>99</b>
6.1	Introduction.....	99
6.2	Optimising <i>in vitro</i> culture conditions for analysing fibroblast subtypes .....	99
6.2.1	<i>In vitro</i> culture alters transcriptomes .....	99
6.2.2	2D <i>in vitro</i> culture conditions impact fibroblast proliferation .....	101
6.2.3	3D <i>in vitro</i> culture significantly alters gene expression and can be used to skew fibroblast phenotypes towards previously described sub-types.....	101
6.3	Recapitulating <i>ex vivo</i> fibroblast phenotypes.....	105
6.3.1	Manipulation of <i>in vitro</i> culture conditions upregulates genes associated with <i>ex vivo</i> fibroblast populations.....	106
6.3.2	Alteration of culture conditions partially recreates <i>ex vivo</i> fibroblast transcriptomes .....	109
6.3.3	Treated cells show distinct differentiation trajectories.....	112
6.4	Characterising fibroblast phenotypes.....	113
6.4.1	Migration assays .....	113
6.4.2	Gel contraction assays .....	114
6.5	Discussion .....	115
<b>Chapter 7</b>	<b>Discussion .....</b>	<b>119</b>
<b>Appendix A 125</b>		
A.1	Summary of studies examining prognostic impact of CAF in NSCLC .....	125
A.2	Experimental induction of apoptosis did not yield sufficient cDNA for library generation and sequencing .....	125
A.3	Results of differential gene expression analysis for the identified stromal clusters .....	126
A.4	Gene set enrichment (GSEA) results for stromal clusters .....	126
A.5	Gene sets used for GSEA identified at literature review .....	141
A.6	Gene set enrichment results for trajectory analysis States.....	143
A.7	Transcription factor enrichment results for trajectory analysis States .....	145

A.8	Composition of TargetLung (Drop-seq) dataset by patient .....	149
A.9	Cells allocated to trajectory State by cluster .....	149
A.10	Culture on plastic relative to Matrigel shows less striking changes in gene expression than does culture on plastic relative to 3D .....	150
A.11	Cells allocated to cluster by culture condition using a random forest classifier ...	151
A.12	Treated cells allocated to trajectory States .....	151
A.13	Gene set enrichment analysis of treated cell trajectories .....	152
<b>List of References .....</b>		<b>153</b>



## Table of Tables

Table 1.1 Reported overlap between described fibroblast markers.....	9
Table 1.2 Results of previous CAF-targeting studies with a described clinical or survival outcome .....	15
Table 2.1 Origins of cells used in this study and their preferred growth media .....	21
Table 2.2 Growth media composition .....	22
Table 2.3 Reaction preparations for rounds 1 and 2 of Mycoplasma PCR .....	24
Table 2.4 Mycoplasma PCR primer sequences.....	24
Table 2.5 Mycoplasma PCR cycling conditions.....	24
Table 2.6 Three-dimensional gel composition (400 µl/gel).....	26
Table 2.7 SDS-PAGE gel composition for a total volume of 10 ml resolving gel and 5 ml stacking gel .....	28
Table 2.8 Running buffer and transfer buffer composition .....	28
Table 2.9 Antibodies used for protein detection .....	29
Table 2.10 Working concentrations of primers used in RT-PCR.....	31
Table 2.11 Reagent volumes for each RT-PCR reaction .....	31
Table 2.12 Concentrations of enzyme and DNase tested in tissue disaggregation optimisation	33
Table 2.13 FACS staining mastermix composition per $1 \times 10^6$ cells .....	35
Table 2.14 Cell lysis buffer composition (for 1 ml) .....	36
Table 2.15 Reverse transcription mix (for 200 µl; one sample) .....	38
Table 2.16 Exonuclease mix (for 200 µl; one sample).....	38
Table 2.17 PCR mixture (50 µl; one sample) .....	39
Table 2.18 Library PCR mixture composition (25 µl per sample) .....	39
Table 2.19 Primary antibodies used in multiplexed immunohistochemical staining .....	45

## Table of Tables

Table 2.20 Characteristics of scRNA-seq patients. LUSC: squamous cell carcinoma, LUAD:

adenocarcinoma.....47

Table 3.1 Composition of disaggregation enzymes used. Specific collagenase activity is given in

Wünsch units.....53

Table 4.1 Canonical cell lineage markers .....66

Table 4.2 Top enrichment statistic for each cluster .....76

Table 5.1 Top 10 “marker” genes identified by differential gene expression analysis, ranked by  
decreasing adjusted *p* value .....85

Table 5.2 Representative gene set enrichment analysis result for each stromal cluster .....86

Table 5.3 Genes upregulated in State 3 and by TGF- $\beta$  .....92

Table 6.1 Summary of key genes differentially expressed between fibroblasts cultured on plastic  
and *ex vivo* .....103

Table 6.2 Overlap between genes significantly ( $p < 0.001$ ) differentially expressed in both *in vitro*  
culture conditions and *ex vivo* counterpart trajectory State .....110

Table 6.3 Representative GSEA results for differentiation from State 5 to each terminal State113



## Table of Figures

Figure 1.1 The relationship between surface marker expression and prognostic effect for CAFs in NSCLC .....	16
Figure 2.1 Transwell® migration assay setup .....	27
Figure 2.2 Overview of the tissue disaggregation pipeline .....	33
Figure 2.3 The microfluidic device used in Drop-Seq (from Macosko <i>et al.</i> <sup>131</sup> ) .....	36
Figure 2.4 Drop-Seq experimental set-up (from Macosko <i>et al.</i> <sup>131</sup> ) .....	37
Figure 3.1 FACS gating strategy for identifying single live cells.....	50
Figure 3.2 CD90 is a more robust marker of lung fibroblasts than PDGFR- $\alpha$ either alone or in combination with PDGFR- $\beta$ .....	51
Figure 3.3 FACS panel validation and identification of positive staining thresholds.....	52
Figure 3.4 Disaggregation enzymes and incubation times have a significant impact on stromal cell isolation.....	54
Figure 3.5 Disaggregation enzyme and duration do not affect cell yield or viability .....	55
Figure 3.6 TrypLE increases the fraction of EpCAM-positive cells.....	55
Figure 3.7 Tissue disaggregation using the optimised protocol generates sufficient cDNA quantities (> 3000 pmol/l) for single-cell RNA sequencing with Drop-seq. BioAnalyzer trace showing Tagmented cDNA library quantification .....	56
Figure 4.1 Quality control metrics for lung cell lines.....	60
Figure 4.2 Top ten genes for principal components 1 (a) and 2 (b) .....	61
Figure 4.3 Drop-seq identifies distinct lung cell line lineages .....	61
Figure 4.4 Heatmap showing the top ten differentially expressed genes by cluster .....	62
Figure 4.5 Filtering data from primary lung tissues using parameters for cell lines identifies multiple clusters.....	63
Figure 4.6 Quality control metrics for whole lung samples.....	64
Figure 4.7 <i>Ex vivo</i> cells show a higher fraction of reads mapping to mitochondrial genes.....	64

## Table of Figures

Figure 4.8 Lung cell lines and <i>ex vivo</i> cells show differing library complexity and viability .....	65
Figure 4.9 Violin plot showing the number of genes (nGene) in cells of each type, using data without applied filters.....	67
Figure 4.10 Standardised quality-control metrics improve clustering quality of scRNA-seq data	68
Figure 4.11 Identifying quality-control metrics with potential to distinguish ‘low-quality’ and ‘cell’ groups .....	69
Figure 4.12 Implementation of quality-control metrics to remove low-quality droplets .....	70
Figure 4.13 The emptyDrops function removes fewer low-quality events than does the random forest classifier .....	72
Figure 4.14 Use of a random forest classifier improves clustering quality relative to other commonly-used methods .....	72
Figure 4.15 Longer tissue disaggregation enables detection of more cell types, with concomitant increases in disaggregation-associated gene expression changes .....	74
Figure 4.16 tSNE plot of quality-controlled data .....	75
Figure 4.17 All cell lineages are composed of cells from multiple patients .....	77
Figure 4.18 Feature plot showing expression of cell type marker genes across the dataset. Points (representing cells) coloured purple show upregulation of the specified gene	77
Figure 5.1 Lung fibroblast cell line markers are not specific for <i>ex vivo</i> fibroblasts.....	82
Figure 5.2 Expression of <i>COL1A2</i> , <i>COL3A1</i> and <i>DCN</i> by the “stromal cell” cluster allows exclusion of misclassified or low-quality cells .....	83
Figure 5.3 Analysis of stromal cells reveals the presence of nine distinct subtypes .....	84
Figure 5.4 tSNE plot showing filtered stromal cells labelled by assigned function. ....	88
Figure 5.5 Staining for periostin and serpin E1 identifies distinct stromal populations in lung adenocarcinoma.....	89
Figure 5.6 Staining for periostin and serpin E1 identifies distinct stromal populations in lung squamous cell carcinoma .....	90

Figure 5.7 Primary stromal cells show progression from normal to cancer-associated fibroblasts over pseudotime .....	91
Figure 5.8 Primary fibroblasts show trajectory-dependent upregulation of marker genes and gene sets .....	94
Figure 6.1 Cells analysed immediately following isolation “ <i>ex vivo</i> ” and following culture on plastic “ <i>in vivo</i> ” show differences in gene expression .....	100
Figure 6.2 Alteration of tissue culture surface impacts fibroblast proliferation .....	101
Figure 6.3 Culture on plastic relative to 3D leads to upregulation of genes differentially expressed by <i>ex vivo</i> CAF and <i>in vitro</i> fibroblasts.....	102
Figure 6.4 Changes in gene expression across culture substrates are not cell type- or patient-dependent.....	104
Figure 6.5 The observed changes in gene expression are maintained at a protein level.....	105
Figure 6.6 Primary stromal cells from the combined dataset (described in Section 5.3) show distinct differentiation trajectories. Trajectory plots coloured by (a) State and (b) cluster.....	106
Figure 6.7 Alteration of culture conditions leads to upregulation of genes differentially expressed by <i>ex vivo</i> stromal cells.....	108
Figure 6.8 Manipulation of culture conditions partially recreates the transcriptomes of <i>ex vivo</i> stromal cells .....	111
Figure 6.9 Transdifferentiated fibroblasts show distinct differentiation trajectories.....	112
Figure 6.10 Treated fibroblasts do not show differential migration towards serum .....	114
Figure 6.11 Transdifferentiated NOF show differential contractile capacity .....	114



## List of Accompanying Materials

Original data supporting this work are openly available from the University of Southampton repository at <https://doi.org/10.5258/SOTON/D1383> and at Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo>) as GSE126111. In addition, this work includes re-analysis of existing data that are publicly available from ArrayExpress at <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6149> and <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6653>.



## Research Thesis: Declaration of Authorship

Print name:	Dr. Sara Waise
Title of thesis:	Characterising cancer-associated fibroblast heterogeneity in non-small cell lung cancer: relating molecular phenotype to function

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Parts of this work have been published as:

An optimised tissue disaggregation and data processing pipeline for characterising fibroblast phenotypes using single-cell RNA sequencing. Waise, S., Parker, R., Rose-Zerilli, M.J.J., Layfield, D.M., Wood, O., West, J., Ottensmeier, C.H., Thomas, G.J.\* & Hanley, C.J.\* (2019). *Scientific Reports* 9 9580

An optimized method to isolate human fibroblasts from tissue *ex vivo* analysis. Waise, S., Parker, R., Rose-Zerilli, M.J.J., Layfield, D.M., Wood, O., West, J., Ottensmeier, C.H., Thomas, G.J.\* & Hanley, C.J.\* (2019). *Bio-protocol* 9 e3440

Signature:

Date:





## Acknowledgements

I would like to thank Professor Gareth Thomas, Professor Peter Johnson, Dr. Christopher Hanley and Professor Christian Ottensmeier for their supervision and guidance during this project. I am grateful to the University of Southampton Drop-seq community, in particular Rachel Parker, Lucy Kimbley, Jack Harrington, Dr. Paola Barragan and Dr. Sybil Jongen, for their assistance. I would also like to thank Dr. Serena Chee and the TargetLung clinical trials associates Benjamin Johnson, Carine Fixmer and Maria Lane, for facilitating access to primary patient samples. In addition, I am grateful to Cancer Research UK, the Medical Research Council and The Pathological Society of Great Britain and Ireland, for the funding support received for this project. I would like to thank Strephon Swemmer and Tas Waise for their support and advice. Above all, I am indebted to the patients enrolled in the TargetLung study, without whose generous consent this work would not have been possible.



## Definitions and Abbreviations

$\alpha$ -SMA	$\alpha$ -smooth muscle actin
AEC	3-amino-9-ethylcarbazole
BAM	Binary Alignment Map
BP	Biological processes
BSA	Bovine serum albumin
CAF	Cancer-associated fibroblast
CP/CGP	Canonical pathways/chemical and genetic perturbations
CTL	Control
CXCL12	C-X-C chemokine 12 (stromal-derived factor 1)
DAB	Diaminobenzidine
<i>DCN</i>	Decorin
DGE	Digital gene expression
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DFS	Disease-free survival
DSS	Disease-specific survival
ECM	Extracellular matrix
EMT	Epithelial-mesenchymal transition
EP	Expression programs
FACS	Fluorescence-activated cell sorting
FAP- $\alpha$	Fibroblast activation protein alpha
FCS	Foetal calf serum

## Definitions and Abbreviations

FDR	False discovery rate
FSP-1	Fibroblast-specific protein-1
FWER	Family-wise error rate
GSEA	Gene set enrichment analysis
HNSCC	Head and neck squamous carcinoma
HR	Hazard ratio
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MAD	Median absolute deviation
MMP-2	Matrix metalloprotease-2
MV	Multivariate
NES	Normalised enrichment score
nGene	Number of genes
nUMI	Number of unique molecular identifiers
NiF	Fibroblast from non-involved lung
NOX4	NAD(P)H oxidase 4
NSCLC	Non-small cell lung cancer
OR	Odds ratio
OS	Overall survival
PBS	Phosphate-buffered saline
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDAC	Pancreatic ductal adenocarcinoma

PDGFR	Platelet-derived growth factor receptor
PFS	Progression-free survival
QC	Quality control
RT	Reverse transcription
SAM	Sequence Alignment Map
SASP	Senescence-associated secretory phenotype
SCLC	Small cell lung cancer
scRNA-seq	Single-cell RNA sequencing
SEM	Standard error of the mean
Shh	Sonic hedgehog
STAMP	Single-cell transcriptome attached to microparticle
RT-PCR	Real-time PCR
TCP	Tissue culture plastic
TGF- $\beta_1$	Transforming growth factor beta 1
tSNE	t-distributed stochastic neighbour embedding
UV	Univariate
VSMC	Vascular smooth muscle cell



# Chapter 1 Introduction

## 1.1 Introduction

Lung carcinoma remains the leading cause of cancer death globally<sup>1</sup>. This disease is divided into two categories according to histological appearance: small cell (SCLC) and non-small cell (NSCLC), accounting for approximately 15% and 85% of cases, respectively<sup>2,3</sup>. NSCLC is further sub-classified into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma (representing 40%, 25-30% and 5-10% of all lung cancers<sup>3</sup>). Despite more recent therapeutic advances, there has been no significant improvement in survival: outcomes remain dismal, with a 10-year survival of only 5%<sup>4</sup>. This is due, at least in part, to the advanced disease stage at presentation in the majority of patients: surgical resection offers the best curative potential, but fewer than one-fifth of patients are deemed suitable for this intervention at the time of diagnosis<sup>2,4</sup>. Non-surgical management options appear to be ineffective at inducing long-term remission, with conventional chemotherapy regimens conferring a reported median overall survival of 8-10 months<sup>5-8</sup>; targeted and immunotherapies benefit only a subset of patients, and have yet to improve long-term outcomes in treated cohorts<sup>9</sup>.

Relative to other cancers, NSCLC tumours show a particularly low degree of purity, with high proportions of both stromal and immune cell populations<sup>10</sup>. The tumour microenvironment is now widely accepted to impact both tumour progression and response to therapy<sup>11,12</sup>. However, the main focus of research in this area to date has been profiling of the immune landscape and its associated implications for immunotherapies<sup>13-15</sup>. Stromal populations, including fibroblasts, remain less well-characterised.

Fibroblasts are cells of mesenchymal origin and are almost ubiquitous in human tissues. Under normal conditions, they are considered quiescent, with no notable transcriptomic or metabolic activity<sup>16</sup>. In response to tissue wounding, fibroblasts can transdifferentiate into myofibroblasts<sup>16</sup>; cells with a contractile,  $\alpha$ -SMA-positive phenotype. Myofibroblasts have pivotal roles in the formation of granulation tissue and completion of wound healing, synthesising and secreting extracellular matrix (ECM) components and remodelling enzymes<sup>17,18</sup>. Myofibroblasts are largely absent from normal tissues<sup>17,19</sup>, and at the resolution of wound healing, either revert to the quiescent fibroblast phenotype or are removed by apoptosis<sup>17</sup>. Failure of this process is associated with pathologies including scarring and tissue fibrosis<sup>16</sup>.

Fibroblasts are the most common cell type in the stroma of a number of solid tumours<sup>18,20-22</sup>, where they are referred to as cancer-associated fibroblasts (CAFs). A desmoplastic stroma, rich in

CAFs, is frequently associated with poor prognosis<sup>23,24</sup>. Similar to fibroblasts in tissue wounding, CAFs are key in ECM remodelling<sup>21</sup>: they are main source of ECM-degrading enzymes and connective tissue ECM components, including collagen<sup>20,22,25</sup>. CAFs are associated with a number of the hallmarks of malignancy<sup>22,26</sup>: through both direct and indirect effects, CAFs promote tumour growth, angiogenesis and chemoresistance<sup>21,27,28</sup>. They also have a number of immunomodulatory functions, including prevention of tumour penetration by immune cells and induction of T-cell hyporesponsiveness<sup>17,29,30</sup>.

Given these tumour-promoting effects, and their genetic stability relative to cancer cells<sup>22</sup>, it is unsurprising that fibroblasts are an attractive therapeutic target. However, clinical trials targeting CAFs have so far yielded disappointing results<sup>31,32</sup>. This may, in part, be due to variation within the fibroblast population: these cells are known to be heterogeneous in both normal and disease states<sup>16,17,21,33</sup>, and remain a poorly-characterised cell type. It is likely that a particular fibroblast population is composed of discrete phenotypes with distinct functions<sup>34,35</sup>, but it is not yet clear how many subtypes are present within a given tumour type or the nature of any functional differences between groups<sup>21,22,36</sup>. The phenotypic, functional and molecular variation seen in fibroblast populations may be attributable to a number of factors including diverse differentiation stimuli and cells of origin, or subtype- or context-dependent function<sup>16,28,37</sup>.

The definition of CAF sub-populations has been hampered by multiple factors. First, the terminology used in this area is variable. The terms “cancer-associated fibroblast”, “activated fibroblast” and “myofibroblast” are often used interchangeably in the literature. However, although CAFs are most commonly described as  $\alpha$ -SMA-positive “myofibroblastic” cells, not all CAFs have this phenotype (and indeed, not all “myofibroblastic” cells are CAFs). CAFs may also be referred to as “peritumoural” or “tumour-associated” fibroblasts, or more generally as “reactive stroma”<sup>17</sup>. Thus, it is not always evident whether authors are referring to CAFs as a whole, or specifically those with an  $\alpha$ -SMA-positive phenotype.

Secondly, there is no marker specific for CAFs<sup>16</sup>, and no single molecular marker that will reliably identify all CAFs<sup>28,38,39</sup>. There are also no well-defined markers to distinguish normal from cancer-associated fibroblasts<sup>20</sup>. These limitations have prevented specific isolation of CAFs (or any subpopulations thereof) by for example, flow cytometry, or accurate assessment of their spatial distributions and relationships with other cell types by immunohistochemical staining. Groups have often studied CAFs based on one or a few markers of interest<sup>38</sup>, although some have additionally examined overlap between multiple markers<sup>34,40,41</sup>.



Novel therapeutic approaches may improve survival in NSCLC: stromal targeting has the potential to augment the response to current therapies<sup>42,43</sup>. To date, the data regarding the prognostic impact of CAFs in NSCLC are conflicting<sup>44-46</sup>, and there have been few studies examining the phenotypic and functional heterogeneity in this population. Identification and characterisation of pro-tumorigenic CAF phenotypes will facilitate development of appropriate stromal targeting strategies.

## 1.2 Fibroblast heterogeneity in normal tissues

It is well-established that fibroblasts in normal tissues show considerable variability between, and even with, anatomical sites<sup>47</sup>. The skin is currently the best-characterised model of normal fibroblast heterogeneity, with three distinct populations described: those of the papillary dermis, those of the reticular dermis and those associated with hair follicles<sup>48,49</sup>. These populations show differential gene expression, origin lineage (in animal models) and function<sup>50,51</sup>. For example, fibroblasts in the papillary dermis have roles in follicle development, whereas those of the reticular dermis are key in initial dermal repair following injury<sup>50</sup>. In keeping with this, these populations show variation in their production of collagen and procollagen mRNAs<sup>49,52</sup>. Dermal fibroblasts from distinct anatomical sites show differences in their proliferative capacity, baseline expression of TGF- $\beta$  and contractile response to the same stimulus<sup>48,53</sup>.

Similar variation has been observed in lung fibroblasts. Fibroblasts of the proximal airways show distinct morphology, gene expression and synthetic and secretory profiles when compared to their distal counterparts<sup>54-56</sup>. Lung fibroblasts, along with those of the myometrium and orbit, have been divided into distinct groups based on expression of CD90 (Thy-1, a glycoprotein found on the surface of a number of cell types<sup>57</sup>). This stratifies fibroblasts as “myofibroblastic” (CD90-positive) or “lipofibroblastic” (CD90-negative): these populations differ in their morphology, secretory profile and response to cytokine stimulus<sup>47,57-60</sup>.

Surface marker expression has also been used to separate fibroblast populations in the myocardium and colon. Staining for fibroblast-specific protein-1 (FSP-1) and  $\alpha$ -SMA identifies two distinct groups of myocardial fibroblasts<sup>61-63</sup>; in the colon, distinct populations of fibroblasts expressing either  $\alpha$ -SMA or PDGFR- $\alpha$  have been described. Fibroblasts from both the myocardium and colon show heterogeneous morphology. The relationship between surface marker expression, morphology and function has yet to be clearly defined in these tissues<sup>19,64</sup>.

Although intra-organ fibroblast heterogeneity is well-described, relatively few studies have directly compared the variation between fibroblast populations in different organs. Both similarities and differences in transcriptome and protein expression have been observed between

the organs of the gastrointestinal system. In the series mentioned above, the distinct populations of  $\alpha$ -SMA- and PDGFR- $\alpha$ -positive populations present in the colon were also identified in the oesophagus, stomach and small intestine<sup>64</sup>. In contrast, sub-epithelial and sub-peritoneal fibroblasts show transcriptional differences in the stomach, ileum and colon, but not the duodenum or oesophagus<sup>65</sup>.

Human cardiac, dermal and pulmonary fibroblasts have been reported to show organ-specific proliferation rates and expression of key ECM genes (*e.g.* *MMP1*, encoding a matrix metalloprotease)<sup>66</sup>; fibroblasts from the human lung and nasopharynx show differential induction of protein synthesis in response to TGF- $\beta$ <sup>67</sup>. Together, the above observations may be reflective of functional differences; this remains to be fully explored.

### 1.3 Definitions and origins

Fibroblasts are usually defined by their spindle cell morphology and ability to adhere to tissue culture surfaces<sup>16,20,26,33</sup>. As a result of a paucity of markers (described above), fibroblasts are often defined by the absence of expression of other cell lineage markers<sup>19,48,68</sup>. In the context of malignancy, fibroblasts may be defined as all the non-neoplastic, non-vascular, non-epithelial, non-inflammatory cells within a tumour<sup>38</sup>.

Heterogeneity within the CAF population may be partly explained by their diverse origins<sup>20,65,69</sup>. CAF origin studies *in vitro* and in animal models indicate that only a fraction arise from a single cell type<sup>33</sup>. For example, using a panel of markers in *in vitro* breast cancer models, Rønnov-Jessen *et al.* observed that the majority of recruited CAFs were derived from the local fibroblast population. However, a smaller proportion originated from vascular smooth muscle cells and, to a lesser extent, pericytes<sup>70</sup>. Data from similar human studies remain sparse<sup>62</sup>, due to the difficulties associated with fate-mapping in humans. However, there is evidence that CAFs may arise from multiple sources and that this is site-dependent<sup>21,28,38</sup>. The majority of CAFs are usually derived from the resident fibroblast population<sup>26,28,33,71,72</sup>, but may arise from a number of other sources. In the liver and pancreas, stellate cells differentiate to CAFs<sup>17,22</sup> and may form the primary source for these cells<sup>30</sup>. Bone-marrow derived cells are also known to be a source of CAFs<sup>24,28,73</sup>: these cells can be identified in the stroma of mouse models of colorectal and gastric carcinomas following transplantation<sup>24,74</sup>. In the latter model, it was reported that at least 20% of CAFs originated from the bone marrow<sup>24</sup>. Further reported CAF differentiation mechanisms include epithelial-, endothelial- and mesenchymal-mesenchymal (*e.g.* pericytes and adipocytes) transition<sup>18,26,28,38,73</sup>. The relative contribution of each source to the CAF population is likely to vary

by tissue and has yet to be fully determined, but it is likely that the diverse origins of CAF contribute to the observed molecular heterogeneity.

The diverse range of CAF phenotypes may also be due to differences in differentiation stimuli. Although TGF- $\beta_1$  has traditionally been seen as the major driver of CAF differentiation, there is evidence that this process may also be mediated by other stimuli, including cellular senescence (discussed below in Section 1.4). Recent work from our group has shown that the differentiation of fibroblasts to myofibroblasts is dependent on generation of reactive oxygen species by the NAD(P)H oxidase 4 (NOX4) enzyme. Inhibition of NOX4 by genetic or pharmacological means reverted the  $\alpha$ -SMA CAF phenotype *in vitro*, and reduced the accumulation of “myofibroblastic” CAFs *in vivo*<sup>75</sup>. NOX4-dependent myofibroblast differentiation was observed across multiple cancer types and differentiation stimuli, suggesting that this may represent a common mechanism for “myofibroblastic” CAF differentiation. This is in keeping with previous data indicative of redox regulation of CAF differentiation: administration of anti-oxidants prevents CAF formation and abolishes the secretion of matrix metalloproteases<sup>76</sup>.

## 1.4 $\alpha$ -SMA and the “myofibroblastic” CAF

To date, the  $\alpha$ -SMA-positive, “activated myofibroblast” phenotype is the most commonly described<sup>17,26,39</sup>, with the majority of reported CAF functions attributable to this population<sup>16</sup>. “Myofibroblastic” CAFs have a contractile, smooth muscle-like phenotype<sup>19</sup>, secrete growth factors and ECM-remodelling enzymes<sup>18,77</sup> and are associated with a poor prognosis in a range of cancers<sup>23,24,39,78</sup>. This population promote many of the hallmarks of malignancy, including tumour invasion and metastasis, angiogenesis and immune evasion<sup>79-82</sup>. However, there is increasing evidence that there is context-dependent heterogeneity even within the  $\alpha$ -SMA-positive CAF population<sup>17,37</sup>. For example, this marker will identify the functionally distinct “myofibroblastic” and “senescent” subsets<sup>83</sup> and in human lung adenocarcinoma, the level of  $\alpha$ -SMA expression increases with the degree of malignancy<sup>84</sup>.

Senescence is the irreversible growth arrest of damaged or ageing cells<sup>85,86</sup>, and is believed to act as a barrier to malignancy in epithelial cells<sup>87</sup>. However, senescent CAFs (which are largely  $\alpha$ -SMA-positive) exhibit a characteristic secretome, known as the senescence-associated secretory phenotype (SASP): this has documented pro-tumorigenic and immunomodulatory functions<sup>85,86,88</sup>. Senescent CAFs support ECM remodelling<sup>86</sup>, promote carcinoma cell epithelial-mesenchymal transition (leading to a more motile, proteolytic phenotype<sup>76,85,88</sup>) and stimulate proliferation and invasion of carcinoma cells in culture, as well as tumour development *in vivo*<sup>38,87,88</sup>. Many components of the SASP are pro-inflammatory (*e.g.* IL-6, IL-8) and it has previously been

suggested that senescent CAFs can be considered pro-inflammatory cells<sup>86,88,89</sup>. However, the SASP is a diverse mixture of cytokines and chemokines, and it is also likely that some elements may in fact suppress the inflammatory response<sup>86</sup>.

Despite both populations showing  $\alpha$ -SMA expression and having tumour-promoting functions, “myofibroblastic” and “senescent” CAFs show transcriptomic differences, particularly in genes associated with the deposition and organisation of ECM<sup>83</sup>. Collagen fibre organisation correlates with poor prognosis, and thus these two populations show differential prognostic impact<sup>83</sup>.

Data from pancreatic ductal adenocarcinoma (PDAC) suggest that the functions of  $\alpha$ -SMA-positive CAFs are disease-dependent. Some human and murine PDAC studies suggest that, in contrast to the majority of solid tumours, high levels of stromal  $\alpha$ -SMA expression correlate with improved survival. Özdemir *et al.* observed that in mice, pancreatic  $\alpha$ -SMA depletion at either the *in situ* or invasive stages of PDAC resulted in reduced survival secondary to enhanced tumour invasion and suppression of immune surveillance<sup>41</sup>. These same authors reported that in human PDAC, low stromal  $\alpha$ -SMA staining correlated with shorter survival. Together, these results are indicative of a beneficial role for “myofibroblastic” CAFs in PDAC, and further support the notion that it is not possible to determine CAF function on the basis of  $\alpha$ -SMA expression alone.

### 1.5 Alternative CAF subtypes and markers

Not all CAFs are identified by  $\alpha$ -SMA<sup>22</sup>. The next most frequently-studied groups with respect to the relationship between surface marker and function are CAFs expressing fibroblast activation protein alpha (FAP- $\alpha$ ), fibroblast-specific protein 1 (FSP1) or members of the platelet-derived growth factor receptor (PDGFR) family<sup>22,39</sup>. Other commonly-described CAF markers include podoplanin, Thy-1 (CD90), periostin and NG2<sup>22,73</sup>. The extent to which the expression of these markers overlaps with that of  $\alpha$ -SMA is discussed below.

Part of the S-100 superfamily, FSP-1 is also known as S100A4<sup>22</sup>. FSP-1 is associated with the regulation of cell shape and motility, promotes angiogenesis and induces the release of cytokines (including granulocyte colony stimulating factor 1)<sup>90-93</sup>. Due to its expression by numerous cell types (including macrophages and malignant epithelial cells<sup>39</sup>), it is difficult to attribute the functions of FSP-1 to a specific population. In mouse models of breast cancer, reports of stromal FSP-1 depletion are somewhat conflicting, with both a reduction in tumour development and metastasis<sup>90</sup>, and a reduction in metastasis with no effect on the primary tumour<sup>93</sup>, described.

FSP-1-positive fibroblasts have immunomodulatory roles, stimulating the infiltration of T-cells and macrophages into tumours in mouse models, whereas FSP-1-negative fibroblasts do not<sup>93,94</sup>. The

effect of FSP-1-positive cells on macrophage recruitment is mediated at least in part through monocyte chemotactic protein-1 release: in a squamous skin carcinoma model, antibodies to this protein abolish the chemoattractant effect<sup>95</sup>. The FSP-1-positive fibroblasts in this series did not express high levels of  $\alpha$ -SMA, with the authors commenting that CAFs do not need to be “myofibroblastic” to be tumorigenic<sup>95</sup>.

FAP- $\alpha$  is a serine protease (belonging to the dipeptidyl peptidase family), although little is currently known about its substrate specificity<sup>39</sup>. It is not usually expressed in normal tissues outside wound healing<sup>96</sup>, but is highly expressed in CAFs in a range of cancers, including lung, breast and pancreas<sup>45</sup>. Expression of FAP- $\alpha$  is correlated with tumour recurrence and poor outcome in multiple tumours<sup>45</sup>, although has been associated with improved prognosis in breast cancer<sup>97</sup>, potentially indicative of disease-dependent CAF functions.

Similar to FSP-1-positive fibroblasts, FAP- $\alpha$ -positive fibroblasts have been demonstrated to have predominantly immunomodulatory functions. In non-small cell lung cancer patients, a high percentage of FAP- $\alpha$  expression in the stroma is associated with an increased neutrophil to lymphocyte ratio in peripheral blood, suggestive of an impaired T-cell response<sup>45</sup>. In keeping with this, in a murine PDAC model, FAP- $\alpha$ -positive CAFs were found to mediate T cell exclusion through secretion of CXCL12 (also known as stromal-derived factor 1<sup>98</sup>); CXCL12 and its receptor CXCR4 are associated with immune evasion and promotion of tumour growth and metastasis in a number of malignancies<sup>99</sup>. In this PDAC model, inhibition of CXCR4 induced T cell infiltration of tumours, acting synergistically with anti-PD-L1 therapy to induce cancer cell death<sup>98</sup>. FAP- $\alpha$ -positive CAFs may further modulate the immune response by recruiting myeloid-derived suppressor cells through STAT3/CCL2 signalling<sup>100</sup>, and can affect the response to therapy: the CXCL12/CXCR4 axis confers gemcitabine resistance to pancreatic carcinoma cells *in vitro*<sup>101</sup>.

The final group for which there is a frequently-described relationship between surface marker and function are CAFs positive for PDGFR- $\alpha$  or - $\beta$ . PDGFR- $\alpha$  is expressed by fibroblasts under normal conditions, during wound healing, and by up to 90% of these cells in the stroma of solid tumours<sup>29</sup>. These cells support tumour growth and angiogenesis, at least in part through the secretion of growth factors such as FGF-2 and FGF-7<sup>29,39,102</sup>, and may mediate chemoresistance<sup>103</sup>. PDGFR- $\alpha$ -positive CAFs are also pro-inflammatory, expressing a gene signature which includes pro-inflammatory cytokines and chemokines responsible for recruitment of neutrophils and macrophages<sup>29</sup>. However, similar to other CAF populations, the effects of PDGFR- $\alpha$ -positive CAFs appear to disease-dependent: high stromal PDGFR- $\alpha$  expression correlates with improved prognosis in non-small cell lung cancer<sup>104</sup>.

In normal tissues, PDGFR- $\beta$  is expressed by vascular smooth muscle cells and pericytes<sup>39</sup>, and is important in mesenchymal cell differentiation<sup>105</sup>. PDGFR- $\beta$ -positive CAFs promote formation of both tumours and reactive stroma<sup>39,106</sup>, and are associated with shorter survival in a number of malignancies, including breast, prostate, colorectal and serous ovarian carcinomas<sup>23,107</sup>. The glycoprotein STC1 appears to be a key mediator of pro-tumorigenic effects mediated through PDGFR- $\beta$  signalling: in an orthotopic model, co-injection of PDGFR- $\beta$ -positive *STC1*<sup>-/-</sup> fibroblasts reduced tumour cell proliferation and metastasis when compared with wild-type fibroblasts<sup>23</sup>. Interestingly, tumour cells co-injected with *STC1* <sup>+/+</sup> fibroblasts showed higher rates of epithelial-mesenchymal transition (EMT): this phenomenon may be at least partially responsible for the observed effect of STC1 on metastasis.

Podoplanin is a glycoprotein normally expressed by lymphatic endothelial cells, and podoplanin-positive CAFs are associated with increased lymphatic density<sup>108</sup>. The prognostic impact of podoplanin-positive CAFs again appears to vary with disease, being associated with improved outcome in colorectal, cervical and small cell lung carcinoma<sup>109-111</sup>, but with the opposite effect in head and neck squamous carcinoma and NSCLC<sup>108,112-115</sup>. There is a possibility that the roles of podoplanin-positive CAFs are subtype-dependent in non-small cell lung cancer, although the current evidence is not conclusive; there are reports that high stromal podoplanin expression correlates with shorter patient survival in adenocarcinoma<sup>108,114,115</sup>. The data regarding squamous cell carcinoma are less clear, with different groups reporting both a negative prognostic impact and no effect on survival<sup>108,113</sup>.

Podoplanin-positive fibroblasts have multiple documented pro-tumorigenic effects, including ECM remodelling and enhancing tumour formation, invasion and metastasis in NSCLC models<sup>116,117</sup>. This population also appear to mediate resistance to targeted tyrosine kinase therapies in this disease through direct stromal-tumour cell contact<sup>22,118</sup>. In small cell lung cancer, in keeping with their positive prognostic impact, podoplanin-positive CAFs significantly reduce the proliferation and viability of carcinoma cells in co-culture models compared to control CAFs<sup>111</sup>.

The surface marker CD90 has been used to identify and stratify functionally distinct fibroblast populations in a range of normal tissues, including the skin, lung, myometrium and orbit<sup>47,58,60</sup>. CD90-high prostate CAFs *in vitro* show higher expression of tumour-promoting genes (including TGF- $\beta$  and angiogenic factors) and increase epithelial cell expression of CXCR4 (known to mediate multiple pro-malignant effects as described above<sup>99</sup>), than do CD90-low CAFs<sup>119</sup>. However, other data regarding the roles of CD90-positive fibroblasts in the context of cancer are sparse.

## 1.6 Overlap between markers

Although a number of studies have commented on the overlap between fibroblast markers, any associations between the functions of these populations remain relatively poorly-described. The current data in this area are summarised in Table 1.1.

Marker 1	Marker 2	Overlap	Independent expression
$\alpha$ -SMA	FAP- $\alpha$	pancreas <sup>120</sup> , Lewis lung <sup>79</sup> , breast <sup>121</sup>	NSCLC <sup>44</sup>
$\alpha$ -SMA	FSP-1	breast, pancreas, skin <sup>34,40,95</sup>	
$\alpha$ -SMA	NG2	breast, pancreas <sup>34</sup>	
$\alpha$ -SMA	PDGFR- $\alpha$	squamous skin <sup>29</sup> , melanoma <sup>40</sup>	
$\alpha$ -SMA	PDGFR- $\beta$	breast, pancreas <sup>34,121</sup>	ovarian serous carcinoma <sup>107</sup>
$\alpha$ -SMA	Podoplanin	lung, breast, kidney <sup>108</sup>	
FAP- $\alpha$	IGFBP7	colorectum, NSCLC, pancreas, ovary, breast <sup>20</sup>	

Table 1.1 Reported overlap between described fibroblast markers

Expression of  $\alpha$ -SMA has been reported to show overlap with a number of other described fibroblast markers, including FAP- $\alpha$ , FSP-1 and PDGFR- $\alpha$ , in a variety of tissues. In keeping with the tissue-dependent functional heterogeneity described above, there is some evidence that surface marker overlap varies by site. For example, although expression of  $\alpha$ -SMA and PDGFR- $\beta$  show overlap in breast and pancreatic cancer models, these markers identify distinct populations in human ovarian serous carcinoma<sup>107</sup>. In a direct comparison in mouse models, 43.5% of the FSP-1-positive breast CAF population were also positive for  $\alpha$ -SMA, compared to only 10.9% in the pancreas<sup>34</sup>. Species-specific differences have also been described: although  $\alpha$ -SMA and FAP- $\alpha$  staining overlap in mouse lung cancer models, their expression is mutually exclusive in sections from human non-small cell lung cancer<sup>44,79</sup>.

In mouse squamous skin carcinoma, a subset of PDGFR- $\alpha$ -positive skin CAFs were found to co-express  $\alpha$ -SMA<sup>29</sup>. Staining for  $\alpha$ -SMA, PDGFR- $\alpha$  and FSP-1 in a murine melanoma model identified three fibroblast sub-populations with differential spatial distributions<sup>40</sup>. One group, found at the tumour core, were positive for FSP-1 only. The second group, located at the tumour edge, only expressed PDGFR- $\alpha$ . The third sub-population were FSP-1-positive but also expressed low levels of PDGFR- $\alpha$  and were found in the tumour's edge and surrounding capsule. All three groups showed expression of  $\alpha$ -SMA.

Together, these data further underline the difficulties in assigning CAF functions based on surface marker expression: although a number of groups have described overlap between the expression of  $\alpha$ -SMA and other fibroblast markers, few have assessed the differential spatial distributions or any variation in the functions of the identified subtypes. One such study identified four subtypes of human breast CAF based on their relative expression of six markers: CD29, FAP- $\alpha$ , FSP-1,  $\alpha$ -SMA, PDGFR- $\beta$  and CAV1<sup>121</sup>. These subsets showed differential spatial expression and accumulation across breast carcinoma subtypes. One population, expressing high levels of both FAP- $\alpha$  and  $\alpha$ -SMA, enhanced the differentiation of regulatory T cells through secretion of CXCL12 (the known immunomodulatory properties of FAP- $\alpha$ -positive CAF are discussed in Section 1.5), whereas tumours enriched for a second population ( $\alpha$ -SMA-high, FAP- $\alpha$ -negative) showed increased CD8<sup>+</sup> T cell infiltration.

Differential marker co-expression and function have also been reported in both murine and human PDAC. Öhlund *et al.* observed that although the majority of CAFs showed co-expression of FAP with low levels of  $\alpha$ -SMA, a subset expressed FAP and high levels of  $\alpha$ -SMA<sup>120</sup>. These CAFs showed distinct spatial distributions according to their  $\alpha$ -SMA expression:  $\alpha$ -SMA-high fibroblasts were seen adjacent to carcinoma cells, whereas those with low  $\alpha$ -SMA levels were located more distally. Furthermore, these sub-populations also appear to be functionally distinct.  $\alpha$ -SMA-high CAF showed behaviour typical of “myofibroblastic” CAFs, inducing desmoplasia *in vitro*. The  $\alpha$ -SMA-low population showed a more inflammatory phenotype, secreting IL-6 and other cytokines.

### 1.7 Population-based analysis of fibroblast gene expression profiles

Initial studies of CAF gene expression profiles focused largely on defining and examining the prognostic value of stromal signatures from *ex vivo* or laser capture microdissected samples. These studies showed a consistent link between stromal changes and poor survival in solid tumours, and provided insight into the inter-tumour heterogeneity of CAF gene expression profiles. For example, a CAF-specific gene expression signature has been reported in human skin



basal cell carcinoma: this same signature was absent from skin squamous cell, prostate, and colon carcinoma<sup>122</sup>.

Primary CAFs originating from HER2-positive breast cancers show significant differences in their gene expression profile when compared to those from oestrogen receptor-positive and triple-negative carcinomas<sup>69</sup>. In this series, the authors suggested that the upregulated genes (in particular, those in pathways associated with integrin signalling and the cytoskeleton) in the HER2-positive CAF population likely contributed to the enhanced cancer cell migration induced by this population *in vitro*, and could confer a negative prognostic impact<sup>69</sup>. Aside from this variation between subtypes, breast CAFs also show heterogeneity in gene expression between patients with the same subtype<sup>123</sup> and perhaps unsurprisingly, between different cohorts<sup>69,124</sup>.

A pro-inflammatory gene signature has been identified in CAFs in a murine model of squamous skin carcinogenesis<sup>29</sup>. This signature was also present in murine breast and pancreatic models, and in human squamous skin and pancreatic ductal adenocarcinoma tissues, but not in a mouse cervical carcinoma model<sup>29</sup>.

Data from these studies are indeed indicative of disease-specific variation in CAF gene expression. The evidence in this area is likely to be augmented by recent developments in techniques such as single-cell RNA sequencing (scRNA-seq), allowing profiling of fibroblast populations at an individual cell level<sup>125-127</sup>.

The studies noted above have used a variety of approaches for gene expression profiling; either microarray analysis of cultured or laser-captured stromal cells<sup>69,123,124</sup> or analysis of fluorescence-activated cell sorted populations (FACS)<sup>29</sup>. Each of these methods generates slightly different information: array analysis of laser-captured stroma can be used to analyse multiple separate areas of stroma from a single patient, allowing comparisons between different foci within one patient and between patients or different tumour types. However, data acquired using this technique are at a population rather than single-cell level, so distinct subtypes within a given area would not be identified. Use of cultured fibroblasts yields similar information and poses some of the same problems. In addition, comparison of different regions within the same tumour is less practical (requiring fibroblast cultures from multiple distinct areas), and such data will be impacted by the alteration in fibroblast phenotype caused by cell culture *in vitro*<sup>120</sup>.

Expression profiling of FACS-sorted stromal cells has also been used to give information at a population level<sup>29</sup>, and can also be used for isolation of individual cells for input to some scRNA-seq platforms (e.g. Smart-seq2<sup>128</sup>). However, use of FACS sorting depends on having a robust

marker which identifies the entire target population: as discussed previously, such a marker is currently lacking<sup>28,38,39</sup>.

### 1.8 Single-cell RNA sequencing for transcriptomic characterisation

Single-cell RNA sequencing is a relatively novel technology which allows characterisation of the heterogeneity within cell types for an individual tumour, as well as between patients within a cohort. In addition, scRNA-seq of whole tumours generates data for multiple different cell types, facilitating investigation of the relationships and interactions between these populations.

All methods for scRNA-seq require cell lysis for RNA isolation and reverse transcription, with amplification of the resulting cDNA to generate a sequencing library<sup>129,130</sup>. As a result of the minute starting quantities of RNA, these processes can result in significant technical variation<sup>129</sup>.

A number of scRNA-seq platforms are currently available; these differ in their sensitivity, accuracy and cost-effectiveness<sup>130</sup>. Of these, Smart-seq2 and droplet-based platforms have been particularly prominent in recent studies<sup>125-127</sup>. Smart-seq2 applies full-length sequencing to FACS-sorted individual cells<sup>128</sup>. This approach yields the highest number of genes *per* cell with high accuracy, although is one of the more expensive platforms currently available<sup>130</sup>.

Drop-seq, an open source droplet-based scRNA-seq platform, and its' commercial counterpart, the 10x Genomics Chromium pipeline, are in widespread use<sup>127,131-133</sup>. These technologies capture single cells in individual droplets, labelling each read with a droplet-specific barcode. This allows bioinformatic reconstruction of transcriptomes at a single cell level using short-read sequencing<sup>131</sup>. Drop-seq in particular provides a similar accuracy to Smart-seq2, although with a lower number of genes *per* cell than some other platforms. However, Drop-seq is among the most cost-effective of the scRNA-seq platforms, particularly when sequencing large numbers of cells<sup>130</sup>.

Regardless of the method used, scRNA-seq of primary tissues requires the generation of single-cell suspension. Both mechanical and enzymatic approaches may not accurately mirror the tissue of origin. Enzymatic methods can affect cell viability, yield and surface marker expression; mechanical disaggregation gives lower viability and yield, but does not alter surface epitopes<sup>134-136</sup>. Bacterial collagenases in particular have been implicated in alteration of surface marker expression, although the current data in this area are mixed. There have been reports of both superior preservation (relative to trypsin<sup>137</sup>), and indiscriminate cleavage of surface markers, leading to reduced surface epitope detection by FACS<sup>136</sup>. Furthermore, prolonged or over-vigorous disaggregation may alter gene expression profiles<sup>138</sup>. Generating a single-cell suspension

may therefore impact expression of both genes and surface markers, with the potential to influence downstream analysis and create artificial cell populations.

Some populations, such as stromal and epithelial cells have been relatively under-represented in previous scRNA-seq datasets<sup>127</sup>. Fibroblasts, embedded within ECM, are particularly difficult to isolate. In order to characterise accurately *ex vivo* fibroblast phenotypes, optimisation of the approach used to isolate these cells from primary tissues is needed. The effects of this process on cellular transcriptomes will also require characterisation.

## 1.9 CAF gene expression profiling by single-cell RNA sequencing

The platforms detailed above have been used to identify fibroblast populations in a number of tissues including melanoma<sup>125</sup>, head and neck carcinoma<sup>126</sup> and, in a dataset published during the course of this project, NSCLC<sup>127</sup>. Fibroblasts could be grouped independently of their tumour or patient of origin in all three datasets.

In head and neck squamous carcinoma (HNSCC), three fibroblast populations were identified. One showed expression of classical myofibroblast markers (*e.g.* *ACTA2*); a second, “activated CAF”, population showed increased expression of ECM genes such as *FAP* and *PDPN* (this group was further sub-divided into two phenotypes based on expression of *e.g.* ECM genes). The third group was not enriched for myofibroblast or CAF markers, and was therefore designated “non-activated resting” fibroblasts<sup>126</sup>.

Single-cell RNA sequencing of fibroblasts from 8 NSCLC patients revealed 5 distinct subtypes<sup>127</sup>. These populations showed differential expression of multiple collagen genes, *e.g.* *COL4A1* and *COL10A1*, suggestive of functional differences. One cluster showed increased expression of a myogenic transcription factor (*MEF2C*), and downregulation of an inhibitor of myogenesis (*MSC*), and was therefore labelled as a “myogenic” phenotype. A second population showed upregulation of transcription factors promoting ECM genes, and was therefore assigned an “extracellular matrix” phenotype. The remaining three clusters did not receive putative functional labels. The five fibroblast clusters showed differential enrichment for a number of gene sets, including those associated with TGF- $\beta$  signalling, angiogenesis and hypoxia: this may indicate differences in differentiation stimuli, function and metabolism<sup>127</sup>.

In melanoma, Tirosh *et al.* inferred relationships between the CAF population and both immune cells and malignant melanoma cells, identifying CAF genes with an effect on T cell accumulation (*e.g.* *CXCL12*, discussed above), and melanoma cell expression profiles associated with CAF abundance<sup>125</sup>. Interactions between CAF and malignant cells were also apparent in HNSCC: CAF

expressed a significantly higher number of tumour receptor ligands than other cell types, such as T cells and macrophages; tumour cells located adjacent to CAFs showed a distinct “partial EMT” gene expression program<sup>126</sup>. However, phenotypic characterisation of CAF in the above studies was largely performed *in silico*, without any *in vitro* functional assessment.

### 1.10 CAFs as a therapeutic target

As discussed in the preceding sections, CAF promote a number of the hallmarks of malignancy and are genetically stable relative to carcinoma cells: this population are therefore an attractive therapeutic target. Although data from pre-clinical animal models have shown some therapeutic promise, the results from clinical trials of CAF targeting in humans have been disappointing. The results of pre-clinical and clinical CAF-targeting studies with associated outcomes are listed in Table 1.2.

Most CAF-targeting studies to date have focused on FAP- $\alpha$ , and the results are largely in keeping with the known immunomodulatory effects of this population. For example, direct ablation of FAP- $\alpha$ -positive cells in mouse models resulted in immunological control of tumours (an effect mediated by interferon- $\gamma$  and TNF- $\alpha$ ), although whether this translates into a significant survival benefit was not reported<sup>79</sup>. As discussed previously, the expression of FAP- $\alpha$  is widespread in adult tissues, and targeting of this marker is not without issue. Non-selective depletion in animal models results in cachexia (weight loss has also been reported in murine  $\alpha$ -SMA targeting<sup>41</sup>) and fatal bone toxicity, with limited anti-tumour effects<sup>139,140</sup>. These difficulties have translated into disappointing outcomes in human stromal targeting. In phase II clinical trials in metastatic colorectal cancer, inhibition of FAP did not induce remission (the majority of patients showed disease progression) or affect metastasis<sup>31,32</sup>.

Relatively few groups have reported clinical or survival outcomes for other stromal targets. Depletion of stromal  $\alpha$ -SMA in a murine PDAC appeared to impair the immune response to tumour<sup>41</sup>. However, PDA is an extremely CAF-rich tumour, and is possible that this disease represents a unique scenario: in contrast to most other solid tumours, there are reports that high  $\alpha$ -SMA expression correlates with improved survival<sup>41,146</sup>. Targeting of both PDGFR- $\beta$ - and FSP-1-positive stromal cells has been associated with improved experimental outcomes in animal models<sup>90,94,106</sup>. Other groups have sought to target pro-fibrotic or anti-apoptotic signalling by CAF. Inhibition of sonic hedgehog (Shh) signalling (a modulator of “myofibroblastic” CAF differentiation<sup>147</sup>) and anti-apoptotic signalling in CAF improve survival in pre-clinical mouse models<sup>143,145</sup>.

Outcome measure	Target	Model	Result
Reduce tumour growth	FAP- $\alpha$	Murine lung, colon <sup>141</sup>	Success
	PDGFR- $\beta$	Murine colon xenograft <sup>106</sup>	Success
	FSP-1	Murine breast <sup>94</sup>	Success
	NOX4	Murine HNSCC, lung <sup>75</sup>	Success
Improve response to therapy	FAP- $\alpha$	Murine breast <sup>142</sup>	Success
	Shh signalling	Murine PDAC <sup>143</sup>	Success
Induce remission or prolong survival	FAP- $\alpha$	Murine breast, colon <sup>144</sup>	Success
	<b>FAP-<math>\alpha</math></b>	<b>Human colon<sup>32</sup></b>	<b>Failure</b>
	$\alpha$ -SMA	Murine PDAC <sup>41</sup>	Failure
	FSP-1	Murine breast <sup>90</sup>	Success
	Anti-apoptotic signalling	Murine cholangiocarcinoma <sup>145</sup>	Success
Reduce metastasis	<b>FAP-<math>\alpha</math></b>	<b>Human colon<sup>31</sup></b>	<b>Failure</b>
Improve immune response	FAP- $\alpha$	Murine lung, PDAC <sup>79</sup>	Success

Table 1.2 Results of previous CAF-targeting studies with a described clinical or survival outcome

Shh: sonic hedgehog, HNSCC: head and neck squamous carcinoma, NOX4: NAD(P)H oxidase 4. Entries highlighted in bold denote clinical trials in man

Recent work from our group has shown that targeting of NOF-CAF transdifferentiation mechanisms yields successful outcomes<sup>75</sup>. We identified that the generation of reactive oxygen species through the NOX4 enzyme is a key mediator of myofibroblastic CAF differentiation in multiple human carcinomas. Pharmacological inhibition using an anti-fibrotic compound or genetic targeting of this enzyme reduced tumour growth in murine models of lung and head and neck squamous carcinoma<sup>75</sup>.

Although stromal targeting in clinical trials has been disappointing, there is yet hope in this area. The current data suggest that targeting stromal surface markers would likely be more successful as adjunctive therapy (for example, FAP- $\alpha$  depletion or inhibition may be expected to improve the

efficacy of immunotherapy) rather than in isolation. Furthermore, disruption of CAF transdifferentiation or survival signalling, even in established tumours, shows promising results. Characterisation of the CAF subtypes present within a given cancer should facilitate more stromal specific targeting.

1.11 CAFs in non-small cell lung cancer

Although CAFs confer a negative prognostic effect in a range of malignancies<sup>75</sup>, the picture in NSCLC is less established. The prognostic impact of CAF seems to vary with the marker examined and, in some cases, is contradictory (*e.g.* FAP<sup>44,45</sup>; Fig. 1.1, further data in Table A.1): this in itself may imply functional diversity within this population.

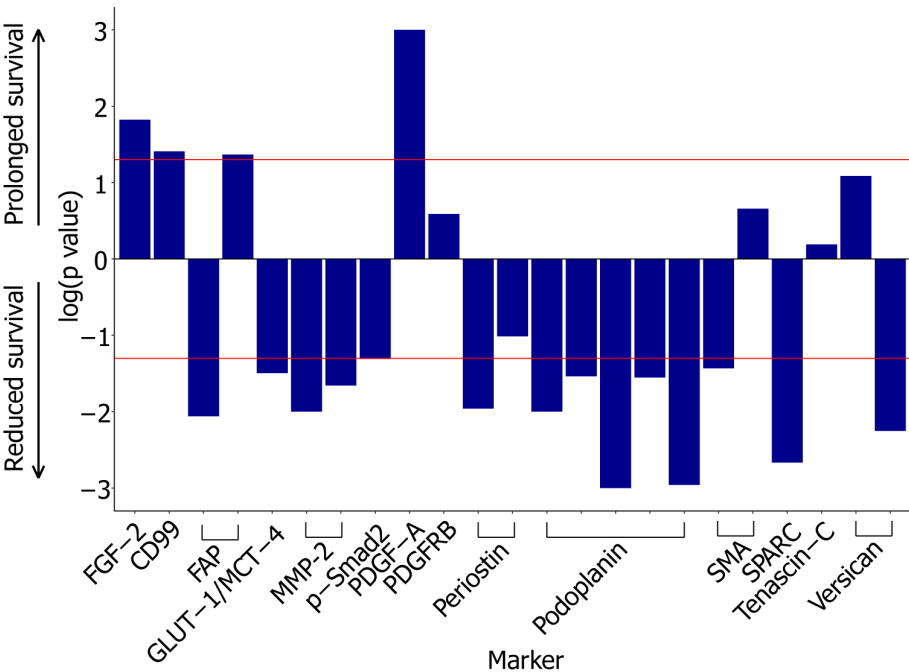


Figure 1.1 The relationship between surface marker expression and prognostic effect for CAFs in NSCLC

The impact on prognosis is expressed as log(*p* value) for the given survival statistic; significance thresholds of *p* < 0.05 are indicated in red. Full survival data are listed in Table A1.1

High expression of podoplanin (the most frequently-studied CAF surface marker in NSCLC), is consistently reported as having a significant negative prognostic impact<sup>108,113-115,148</sup>. High stromal matrix metalloprotease-2 (MMP-2) expression correlates with reduced survival<sup>149</sup>, although one study reported that this effect was subtype-specific, holding true for squamous cell carcinoma only<sup>150</sup>.

Although high  $\alpha$ -SMA expression is associated with shorter survival across all NSCLC<sup>151</sup>, the effect of  $\alpha$ -SMA is less well-studied in comparison to other tissues; a non-significant correlation with improved outcome has been reported in squamous cell carcinoma<sup>44</sup>. Subtype-dependent associations with survival have also been reported for podoplanin, FAP, GLUT1 and MCT-4, and versican<sup>44,108,152,153</sup>; however, some authors have examined marker effects in one subtype in isolation<sup>113-115,148</sup> and others consider all NSCLC subtypes together (*e.g.* Edlund *et al.*, 2012, Liao *et al.*, 2013). Further complicating interpretation and comparison of these findings, the current data in this area encompass a variety of survival analysis approaches and metrics (*e.g.* overall survival, disease-specific survival and disease-free survival), and cohort sizes vary, with some comprising relatively few patients.

Relatively few studies have assessed the co-expression of surface markers in NSCLC CAF. Staining for both CD99 and podoplanin overlaps with that for  $\alpha$ -SMA<sup>108,154</sup>; there is no correlation between  $\alpha$ -SMA and FAP expression in squamous cell carcinoma<sup>44</sup>. In keeping with other tissues, FAP- $\alpha$ -CAF appear to have immunomodulatory roles in the lung: high expression levels were associated with an elevated neutrophil to lymphocyte ratio (proposed as an indicator of an impaired immune response) in a clinical cohort<sup>45</sup>. Depletion of CAF expressing FAP- $\alpha$  or enzymatic FAP- $\alpha$  inhibition is associated with reduced tumour growth and angiogenesis, and induction of immune-mediated tumour cell death<sup>79,141</sup>. Targeting of stromal PDGFR $\alpha$  in a murine lung carcinoma model also inhibits tumour growth and angiogenesis (independent of carcinoma cell PDGFR $\alpha$  expression), implying a role for this population in these processes<sup>103</sup>. Finally,  $\alpha$ -SMA-positive CAF appear to modulate therapeutic response in lung cancer: this population reduce tumour cell response to a MEK inhibitor *in vitro*<sup>155</sup>.

There are multiple other reports of CAF in NSCLC reducing tumour cell response to therapies, although these have not been correlated with expression of particular surface markers. Lung CAF show differential modulation of response to chemotherapy *in vitro*, inducing anti-apoptotic effects in carcinoma cell lines after treatment with paclitaxel, but not cisplatin<sup>43</sup>. These cells appear to induce resistance to tyrosine kinase inhibition in *EGFR*-mutant cell lines through a variety of mechanisms, including Shh-mediated induction of EMT<sup>156</sup>, upregulation of HGF<sup>157</sup> and increased Aurora-A kinase<sup>42</sup>. Consistent with pro-tumorigenic functions in other cancers, lung CAF increase tumour growth, motility and invasion<sup>158,159</sup> and promote angiogenesis<sup>160</sup> and recruitment of monocytes in culture<sup>161</sup>.

There are some gene expression profiling data suggestive of functionally-distinct CAF populations in NSCLC: recently, Hao *et al.* categorised CAF as “high desmoplasia” or “low desmoplasia” depending on extent of stromal response in tumour of origin<sup>162</sup>. Microarray expression profiling

revealed thirteen genes showing differential expression between the two groups. These transcriptomic differences were suggestive of functional heterogeneity: *ST8SIA2*, the gene showing the highest differential upregulation in “high desmoplasia” CAFs, was associated with increased tumour cell invasion *in vitro*. Single-cell transcriptomic analysis of primary human NSCLC tissues has identified five distinct fibroblast populations (discussed in Section 1.9)<sup>127</sup>. These populations showed differential expression of, for example, multiple collagen genes, which may indicate functional differences. However, such characterisation, and associated histological validation, are yet to be performed.

### 1.12 Discussion

Despite recent therapeutic advances, lung carcinoma remains the leading cause of cancer death worldwide<sup>1</sup>. Non-small cell lung cancer, which accounts for 85% of cases, shows a high abundance of CAFs<sup>2,3,10</sup>. CAFs are a heterogeneous population associated with a number of the hallmarks of cancer<sup>22,26</sup>. Despite their frequency in solid tumour stroma, they remain poorly-characterised, with a lack of data linking phenotype to function. The lack of CAF definition at a sub-population level is a current barrier to therapeutic targeting: functional characterisation is needed to facilitate the development of more refined targeting strategies.

Although a high proportion of cancer-associated fibroblasts confers a poor prognosis in a number of solid tumours<sup>23,24</sup>, the picture in NSCLC is mixed: data linking fibroblast surface marker to outcome show variable results depending on the marker examined<sup>104,113,154,163</sup>. No single surface marker will reliably identify all CAF<sup>28,39</sup>; many groups examining CAF function, particularly in NSCLC, have focused on expression of one marker alone. Although there are some data linking surface marker to function in this disease<sup>45,79,103,141,156</sup>, few have examined the co-expression of markers, or comprehensively defined fibroblast phenotypes. It appears possible that fibroblast populations show NSCLC subtype-dependent prognostic impact<sup>44,108,113-115,148,150,152,153</sup>, although few studies have examined this directly.

Given their roles in promoting the hallmarks of malignancy, and their genomic stability relative to carcinoma cells, it is unsurprising that CAFs are an attractive therapeutic target. Although some results from pre-clinical animal models have shown promise<sup>79,106,142,144</sup>, this has failed to translate into survival benefit in clinical trials in man<sup>31,32</sup>. This is likely, at least in part, due to the poor characterisation of CAF thus far, rendering previous targeting strategies non-specific.

To date, there have been relatively few gene expression profiling studies of CAF in NSCLC. Recent microarray expression profiling of CAF from “high desmoplasia” and “low desmoplasia” NSCLC tumours identified transcriptomic differences suggestive of functional heterogeneity<sup>162</sup>. Some of



these findings were validated *in vitro*, although as these experiments were performed at a bulk level, it is not clear whether “high desmoplasia” and “low desmoplasia” CAFs represent two distinct functional entities, or are composed of multiple sub-populations.

Data from scRNA-seq of NSCLC has indicated the presence of multiple fibroblast populations<sup>127</sup>. These groups show differential expression of genes and gene sets, indicating likely functional differences: in this series, some populations were assigned putative phenotypes based on their gene expression profiles. Examination of such differences in these data was somewhat limited, and performed exclusively *in silico*, with no *in vitro* functional characterisation of the identified populations.

Current therapeutic strategies in NSCLC are ineffective in the majority of patients, leading to poor survival outcomes. CAFs remain an attractive therapeutic target, despite the disappointing results of previous clinical trials. Characterisation of CAF subtypes using single-cell RNA sequencing will facilitate identification of pro-malignant subtypes and the molecular mechanisms regulating their accumulation: this will allow identification of potential stromal targeting approaches.

### 1.13 Aims and objectives

The aim of this work is to characterise the phenotypic and functional heterogeneity in CAF in NSCLC. To identify and characterise distinct CAF subpopulations, we have defined three objectives. First, we shall identify the CAF groups present in NSCLC, using Drop-seq to enable scRNA-seq. The resulting data will be interrogated using bioinformatic analysis packages in R to perform dimensionality reduction and clustering and identify distinct CAF subtypes. As part of this, we shall optimise our approach for the isolation of fibroblasts from primary lung tissues and the quality control of scRNA-seq data. Secondly, we shall use this data to identify potential functions and differentiation stimuli for each subpopulation. The scRNA-seq data will be used to generate gene expression profiles for each CAF subgroup. Gene set enrichment and trajectory analyses will then be performed using these genes to determine the putative roles and molecular drivers for these populations. Finally, we shall perform *in vitro* functional characterisation of the identified phenotypes. The trajectory analysis findings will inform *in vitro* recapitulation of the *ex vivo* phenotypes. These fibroblasts will be used in functional assays to determine populations with pro-tumorigenic potential and identify possible future therapeutic targets.



## Chapter 2      Methods

### 2.1      Cell culture

#### 2.1.1      Cell culture principles

All routine cell culture was performed in class II mycoplasma-clean laminar flow hoods. Cells (Table 2.1) were cultured in their preferred growth medium (Table 2.2). Cells were routinely cultured in 75 or 175 cm<sup>2</sup> flasks (Corning) at 5% CO<sub>2</sub> and 37 °C. Fibroblast and adenocarcinoma cell growth medium was changed twice weekly.

For adherent cells, when the cell monolayer reached 100% confluence, cells were washed with phosphate-buffered saline (PBS) before detachment using trypsin-EDTA (Sigma-Aldrich) and collection in fresh growth medium. For maintenance of cells in culture, fibroblasts and adenocarcinoma cells were split 1:4 and 1:5, respectively. Absence of *Mycoplasma* contamination in fibroblast cell lines was routinely verified using *Mycoplasma* PCR (Section 2.1.5).

Cell name	Source	Growth medium
H441	Lung adenocarcinoma cell line	10% DMEM
HFFF2	Human foetal foreskin fibroblast cell line	10% DMEM
HUVEC	Human umbilical vein endothelial cell line	10% DMEM
IMR-90	Human foetal lung fibroblast cell line	10% DMEM
NiF and CAF	Primary lung fibroblasts; see Section 2.1.6	“Complete” DMEM

Table 2.1 Origins of cells used in this study and their preferred growth media

DMEM, Dulbecco’s modified Eagle medium; NiF, fibroblasts from non-involved lung; CAF, fibroblasts isolated from tumour samples

Growth medium	Reagent	Supplier
“Serum-free” DMEM	DMEM	Sigma
	L-glutamine (2 mM)	Sigma
“Empty” DMEM	DMEM	Sigma
	L-glutamine (2 mM)	Sigma
	1% penicillin-streptomycin	Sigma
“Low serum” DMEM	DMEM	Sigma
	1% FCS	Biosera
	L-glutamine (2 mM)	Sigma
“Complete” DMEM	DMEM	Sigma
	10% FCS	Biosera
	L-glutamine (2 mM)	Sigma
	1% penicillin-streptomycin	Sigma
10% DMEM	DMEM	Sigma
	10% FCS	Sigma
	L-glutamine (2 mM)	Sigma

Table 2.2 Growth media composition

FCS: foetal calf serum

### 2.1.2 Freezing cells

Cell stocks were maintained by freezing down for long-term storage. To preserve cellular integrity, cell lines and primary fibroblasts were frozen in 10% DMEM and 100% FCS, respectively, with 10% dimethyl sulfoxide (DMSO; Sigma-Aldrich), a cryoprotectant. Cells were detached in the usual manner when reaching 100% confluency, before being spun at 1500 rpm for 5 minutes. The resulting pellet was re-suspended in the appropriate preparation for freezing, aliquoted into cryovials and transferred to a Nalgene® Mr. Frosty Cryo Freezing Container. This encases vials in isopropanol, providing a rate of cooling very

close to 1 °C/minute. Following freezing and short-term storage at -80 °C, cells were transferred to liquid nitrogen (-196 °C) for longer-term storage.

### **2.1.3 Defrosting cells for culture**

To maintain maximum viability, cells were thawed as rapidly as possible: cells were defrosted in a water bath at 37 °C before being added to 10 ml pre-warmed medium and spun at 1500 rpm for 5 minutes, to remove the DMSO. The supernatant was discarded and the remaining pellet re-suspended in the appropriate volume of pre-warmed medium. This suspension was then plated to a T25 tissue culture flask (Corning).

### **2.1.4 Counting cells**

For primary fibroblast culture, single-cell RNA sequencing (Section 2.7) and functional assays, cells were counted using a CASY Cell Counter (Roche). Ten microlitres of cells (either detached as in Section 2.1 or the single-cell suspension generated in Section 2.5) were added to 10 ml of CASYton (an isotonic buffer; OMNI Life Science). To improve accuracy with the smaller cell numbers, cell counting for migration assays (Section 2.1.10) was performed at a 20X dilution. The CASY Cell Counter exposes cells passing through a measuring pore to a low-voltage electrical field. A resistance signal is generated based on cell size and conductivity: live cells have an intact plasma membrane, and give a higher resistance value. Dead or dying cells have increased membrane permeability and therefore give a lower resistance signal. The CASY Cell Counter is thus able to calculate cell concentration, volume and viability.

### **2.1.5 *Mycoplasma* polymerase chain reaction (PCR)**

To avoid changes in cell behaviour resulting from *Mycoplasma* contamination, cell lines were confirmed *Mycoplasma*-free prior to use. Between 10 and 15 ml supernatant was collected from confluent cells grown in antibiotic-free medium for at least two weeks, and centrifuged at 1500 rpm for 10 minutes. The excess resulting supernatant was discarded, with the remaining volume collected and spun at 13000 rpm for 5 minutes. The excess supernatant was again discarded, and the pellet re-suspended in the remaining volume (~200 µl) for use in the *Mycoplasma* PCR. This process amplifies sequences in the 16S-23S spacer region in RNA operons, common to 14 species of *Mycoplasma*.

## Chapter 2

Each reaction was prepared on ice, with both negative (RNase-free water) and positive (a 5PT cell line supernatant, previously confirmed positive for *Mycoplasma*) controls. Two rounds of PCR were performed, with the reaction mixtures prepared as in Table 2.3 (primers given in Table 2.4), and the cycling conditions as given in Table 2.5. Ten microlitres of the product from the second PCR round was run, with a 100 bp ladder (Promega), in TAE buffer (for 1 litre: 242 g Tris-Base, 57.1 ml glacial acetic acid, 18.6 g EDTA, prepared as a 50x solution) for 1 hour at 120 V on a 1% agarose gel with RedSafe DNA Stain (diluted 1:20 000; Life Technologies) added. Bands were visualised using a GelDoc-It™ Imaging System (UVP, LLC).

Round 1: band at 720bp	Round 2: band at 145bp
Sample DNA/Supernatant (1µl) 1 µl	Round 1 PCR Product 1 µl
Forward Primer 1 (10pmol/µl) 0.1 µl	Forward Primer 2 (10pmol/µl) 0.1 µl
Reverse Primer (10pmol/µl) 0.1 µl	Reverse Primer (10pmol/µl) 0.1 µl
Formamide 0.3 µl	Master Mix* 18.8 µl
Master Mix* 18.5 µl	

Table 2.3 Reaction preparations for rounds 1 and 2 of Mycoplasma PCR

\*Master Mix: MegaMix-Blue (Microzone)

Primer	Sequence
Forward 1	ACT CCT ACG GGA GGC AGC AGT A
Forward 2	CTT AAA GGA ATT GAC GGG AAC CCG
Reverse	TGC ACC ATC TGT CAC TCT GTT AAC CTC

Table 2.4 Mycoplasma PCR primer sequences

Cycling conditions
95°C x 30s 35 cycles of: 95°C x 30s 55°C x 30s 72°C x 1 min 72°C x 1 min

Table 2.5 Mycoplasma PCR cycling conditions

### **2.1.6 Isolation and culture of primary cells**

Samples of tumour and non-involved lung were obtained from patients enrolled in the TargetLung Study undergoing tumour resection at Southampton General Hospital (Section 2.10). Following generation of a single-cell suspension from the tissue (Section 2.5), cells were plated to the largest possible surface area at 100 000 cells/cm<sup>2</sup>. Cells were incubated at 37 °C for 2 hours to allow fibroblasts to adhere to the flask surface before 3 PBS washes to remove non-adherent cells.

### **2.1.7 Coating culture plates**

Six- and 96-well plates were coated with either gelatin (Sigma), Matrigel® Basement Membrane Matrix (Corning) or left uncoated. To coat plates with Matrigel®, 500 µl or 25 µl (for 6- and 96-well plates respectively) of Matrigel® at 16 ng/µl<sup>164</sup> were added to plates and incubated on a rocker at 4 °C for 1 hour. Plates were coated with the same respective volumes of 0.1% gelatin before incubation at 37 °C for 30 minutes.

### **2.1.8 Three-dimensional cultures**

Gel mixes were prepared as in Table 2.6. Confluent primary fibroblasts were detached, counted and diluted to 10x10<sup>6</sup> cells/ml. One hundred microlitres of this suspension were added to 400 µl of the gel master mix. Four hundred microlitres of the resulting mixture were plated to 1 well of a 24-well plate. Gels were incubated at 37 °C for 30 minutes to allow setting, before 1 ml of “low serum” DMEM was added to each well. Gels were detached the following day if not already spontaneously detached. Plates were incubated at 37 °C for 1 week, with growth medium changed three times.

On day 7, 100 µl of Collagenase IV stock (10 mg/ml; Sigma-Aldrich) were added to each well. Plates were incubated with agitation at 37 °C for 1 hour. The gel-enzyme mixture was pipetted to disaggregate any remaining fragments before transfer to an Eppendorf. The suspension was centrifuged at 13000 rpm for 15 seconds. The resulting pellet was washed with PBS and stored at -20°C for RNA extraction and use in real-time PCR (RT-PCR; Section 2.3).

Reagent	Collagen-Matrigel® gel (µl)
Collagen (3 mg/ml)	250
Matrigel®	195
10x DMEM	50
0.1 M NaOH	50
FCS	5

Table 2.6 Three-dimensional gel composition (400 µl/gel)

### 2.1.9 Gel contraction assays

Primary fibroblasts were washed with once with PBS, detached using 500 µl trypsin-EDTA, and collected in “serum-free” DMEM. Cells were then diluted to  $2 \times 10^5$  cells/100 µl, with 100 µl of this suspension added for each 900 µl gel. Collagen (Corning) was used at a concentration of 1.5 mg/ml; water was added to yield a total volume of 900 µl *per* gel. Gels were prepared on ice by the stepwise addition of water, collagen and cells to 100 µl 10x DMEM (Thermo Fisher). Nine hundred microliters of the resulting mixture were added in triplicate to wells of a 24-well plate. Gels were set by incubation for 1 hour at 37 °C before addition of 1 ml “serum-free” DMEM. Gels were monitored for contraction and weighed at 48 hours.

### 2.1.10 Migration assays

Cell migration assays were performed using Transwell® inserts (polycarbonate filters, 8µm pore size; Corning). For each condition, 200 µl of attractant was applied to triplicate wells of a 24-well plate (Corning), and a Transwell® culture insert added. The experimental setup is represented in Figure 2.1. Plates were incubated at 37 °C for 1 hour. A serum-negative control was included in each experiment.



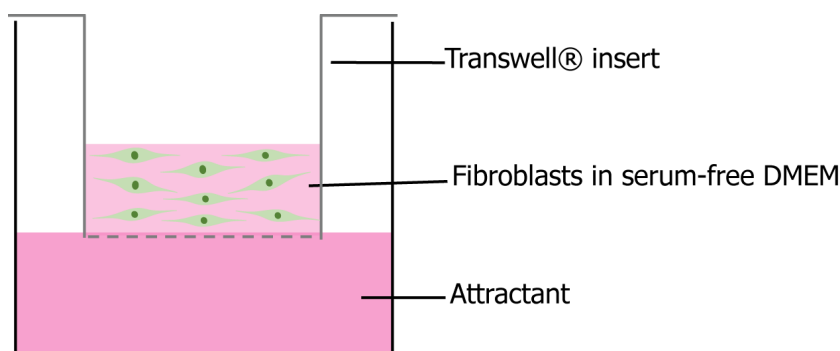


Figure 2.1 Transwell® migration assay setup

Primary fibroblasts were detached using 500  $\mu$ l trypsin-EDTA before collection in “serum-free” DMEM. This suspension was spun at 1500 rpm for 3 minutes, and the resulting supernatant discarded. The resulting pellet was re-suspended in 2.5 ml ‘serum-free’ medium, before dilution to  $50 \times 10^4$  cells/ml. 100  $\mu$ l of this suspension were added to the upper chamber of each Transwell insert. Plates were incubated overnight at 37 °C. The following day, conditioned medium was removed and the wells washed with 500  $\mu$ l PBS. Plates were incubated with 500  $\mu$ l trypsin-EDTA per well at 37 °C for 1 hour, before cell counting using the CASY Cell Counter (Section 2.1.4).

## 2.2 Protein analysis

### 2.2.1 Protein extraction and quantitation

Cells were washed once with PBS before lysis on ice using fibroblast lysis buffer (25 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.5% Triton-X, 20 mM  $\text{NH}_4\text{OH}$ ) with 1% protease inhibitor cocktail (Set 1; Calbiochem, Merck). Cells were incubated with agitation at 4 °C for 30 minutes, before scraping and collection on ice. To remove insoluble cellular components, lysates were centrifuged at 13000 rpm at 4 °C for 15 minutes. Total protein quantitation was then performed using the DC<sup>TM</sup> Protein Assay Kit (Bio-Rad) according to the manufacturer’s instructions. Sample absorbance at 750 nm was measured using the VarioSkan Flash plate reader (Thermo Fisher).

### 2.2.2 Polyacrylamide gel electrophoresis

Samples were diluted to equal protein concentrations across experimental conditions and prepared in 1x Laemmli buffer (from a 5x concentrated stock; 625mM Tris pH 6.8, 10% SDS, 25% glycerol, 0.015% bromophenol blue, 15%  $\beta$ -mercaptoethanol) to reduce the proteins

## Chapter 2

and facilitate loading. Samples were then boiled at 95 °C for 5 minutes to ensure protein denaturation, and 30-40 µg were loaded into 8% SDS-PAGE gels (Table 2.7). Gels were run in a Mini-PROTEAN Tetra Cell (Bio-Rad) in running buffer (Table 2.8) at 100 V for 30 minutes, then 150 V for 90 minutes.

Reagent	8% Resolving gel	Stacking gel
Acrylamide/bis-acrylamide (30%/0.8%)	2.6ml	670µl
ddH <sub>2</sub> O	4.6ml	3.59ml
1.5M Tris-HCl (pH 8.8)	2.6ml	N/A
1M Tris-HCl (pH 6.8)	N/A	630µl
10% SDS	100µl	50µl
10% APS	100µl	50µl
TEMED	10µl	10µl

Table 2.7 SDS-PAGE gel composition for a total volume of 10 ml resolving gel and 5 ml stacking gel

Reagent	Mass for running buffer (g)	Mass for transfer buffer (g)
Trizma-Base	75g	145g
Glycine	360g	725g
SDS	25g	25g

Table 2.8 Running buffer and transfer buffer composition

Both prepared as a 5X solution in 5 litres ddH<sub>2</sub>O

### 2.2.3 Western blotting

Proteins were transferred to a methanol-activated polyvinylidene (PVDF) membrane (Millipore) in transfer buffer (Table 2.8) at 15 V overnight using the Bio-Rad Mini-PROTEAN Tetra Cell. Membranes were then blocked in 5% milk (Marvel) in TBS-T (10 mM Tris-HCl pH7.5, 150 mM NaCl, 0.05% Tween® 20) for 1 hour at room temperature, before incubation with the relevant primary antibody (in PBS with 5% BSA and 0.5% Tween; Table 2.9) for 2 hours at room temperature. Membranes underwent 3 5-minute washes, and were then

incubated with the appropriate secondary antibody for 1 hour at room temperature. Following 3 five-minute washes in TBS-T, proteins were visualised using Supersignal West Pico or Femto Chemiluminescent Substrate (Thermo Fisher Scientific). This was applied to membranes, and the signal detected using the ChemiDoc-It Imaging System (UVP, LLC). Relative densitometry was performed using the Fiji software package.

Antibody	Molecular weight (kDa)	Working dilution	Manufacturer
FN-EDA	262	1:1000	Millipore
Col I	139	1:5000	Abcam
Palladin	110	1:1000	Novus Biologicals
Hsc-70	70	1:1000	Santa Cruz
HSP-47	47	1:5000	Abcam
SMA	42	1:1000	Sigma
Polyclonal rabbit anti- mouse	N/A	1:5000	Dako, Agilent
Polyclonal swine anti-rabbit	N/A	1:5000	Dako, Agilent

Table 2.9 Antibodies used for protein detection

Target proteins are upregulated in myofibroblasts relative to normal fibroblasts

## 2.3 Real-time quantitative polymerase chain reaction (RT-PCR)

### 2.3.1 RNA extraction

For harvesting, cell monolayers were detached as described in Section 2.1. The resulting suspension was spun at 13000 rpm for 15 seconds, before being washed once with PBS and spun again. Following aspiration of the supernatant, cell pellets underwent RNA extraction using the Reliaprep<sup>TM</sup> RNA Cell Miniprep System (Promega), including optional DNase digestion steps, in accordance with the manufacturer's instructions. Following extraction, samples were kept on ice, and RNA quantitation performed using the NanoDrop Spectrophotometer (Thermo Fisher Scientific).

### **2.3.2 cDNA synthesis**

One microgram of RNA was reverse transcribed in a 20  $\mu$ l reaction using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems), according to the kit's protocol. Thermal cycling conditions were those given in the manufacturer's instructions (25 °C for 10 minutes, 37 °C for 2 hours, 85°C for 5 minutes, hold at 4 °C). The resulting cDNA was diluted to 2 ng/ $\mu$ l with ddH<sub>2</sub>O and stored at -80 °C.

### **2.3.3 Quantitative real-time PCR**

Real-time PCR was performed using either Power SYBR® Green Real-Time PCR Master Mix (Thermo Fisher Scientific) or TaqMan® Gene Expression Assays (Thermo Fisher Scientific). Power SYBR® Green Real-Time PCR Master Mix was used in conjunction with the 7500 Real Time PCR System (Applied Biosystems). SYBR® Green I is a fluorescent dye that intercalates between nucleotides, binding the DNA complexes formed during PCR. With increasing cycle number and cDNA amplification, the fluorescence generated by SYBR® Green I increases. The  $C_t$ , or threshold cycle, is the intersection between the cDNA amplification curve and the fluorescence threshold line; higher target sequence concentrations will result in higher fluorescence at a lower cycle number, and thus give a smaller  $C_t$  value. The PCR system calculates  $C_t$  values for the target cDNA sequences relative to specified housekeeping genes ( $\beta$ -actin for cell lines, and a combination of  $\beta$ -actin,  $\beta_2$ -microglobulin and GAPDH for primary fibroblasts, to account for varied expression of different housekeeping genes between samples). The  $C_t$  of the target is then compared to the  $C_t$  of the housekeeping gene (or genes) to determine the relative concentration of RNA present in each sample.

Real-time PCR was performed by adding 5  $\mu$ l of cDNA (2 ng/ $\mu$ l) with a SYBR® Green–primer mixture (Table 2.10 and Table 2.11) to a 96-well plate (STARLAB [UK] Ltd). All samples were plated in duplicate or triplicate before running on the 7500 Real Time PCR System (50 °C for 2 minutes, 95 °C for 10 minutes, then 40 cycles of 95 °C for 15 second and 60 °C for 1 minute], followed by 95 °C for 30 seconds and 60 °C for 15 seconds).

Product	Working concentration ( $\mu\text{M}$ )
ACTA2	0.05
ACTB	0.1
B2M	0.1
COL1A1	0.05
COL3A1	0.1
CTGF	0.2
GAPDH	0.05
FN	0.05
MMP2	0.1

Table 2.10 Working concentrations of primers used in RT-PCR

Reagent	Volume( $\mu\text{l}$ )/reaction (primer/ddH <sub>2</sub> O)
Power SYBR® Green Real-Time PCR Master Mix	12.5
0.2 $\mu\text{M}$ primer	0.5/6.5
0.1 $\mu\text{M}$ primer	0.25/7
0.05 $\mu\text{M}$ primer	0.125/7.25

Table 2.11 Reagent volumes for each RT-PCR reaction

Similar to cDNA quantification using SYBR® Green, TaqMan® Gene Expression Assays require a target-specific primer. However, TaqMan® also uses a sequence-specific probe with a fluorescent reporter at the 5' end and a quencher at the 3' end. When in close proximity, the quencher prevents fluorescent emission from the reporter. During the extension phase of PCR, the Taq polymerase enzyme liberates the reporter from the probe, separating it from the quencher, resulting in emission of measurable fluorescence. This increases with cycle number and cDNA amplification, and is used to determine the relative RNA concentration of samples in the same manner as SYBR® Green.

For real-time PCR, 2  $\mu\text{l}$  of cDNA (1 ng/ $\mu\text{l}$ ) were added with 10  $\mu\text{l}$  of TaqMan Fast Advanced Master Mix (Thermo Fisher Scientific), 1  $\mu\text{l}$  of TaqMan Gene Expression Assay and 7  $\mu\text{l}$  of

nuclease-free water to a MicroAmp™ Fast Optical 96-well plate (Applied Biosystems).

TaqMan® Assays were used with the QuantStudio 7 Flex Real-Time PCR system according to the manufacturer's guidelines (50 °C for 2 minutes, 95 °C for 20 seconds, 95 for 20 seconds, then 40 cycles of 95 °C for 1 second and 60 °C for 20 seconds).

For use with the Biomark platform (Fluidigm), cDNA was first pre-amplified by addition of 62.5 ng to 5 µl PreAmp Supermix (Kapa Biosystems) and 2.5 µl pooled TaqMan® Assays. Amplification was performed according to the manufacturer's guidelines (16 cycles of 95 °C for 3 minutes, 95 °C for 15 seconds then 48 °C for 4 minutes). TaqMan Gene Expression Assays were diluted 1:2 with Assay Loading Reagent (Fluidigm). For each reaction, 2.25 µl pre-amplified cDNA was added to 2.5 µl TaqMan Fast Advanced Master Mix and 0.25 µl Sample Loading Reagent (Fluidigm). Assays and sample mix were loaded to a 48.48 Dynamic Array integrated fluid circuit and the GE 48x48 Fast v1 protocol run, in accordance with the manufacturer's instructions.

### **2.4 Assessment of metabolic activity**

Metabolic activity was evaluated using the CellTiter 96™ AQueous Nonradioactive Cell Proliferation Assay Kit (MTS; Promega). This assay contains MTS, a tetrazolium compound. In metabolically active cells, MTS is reduced by NADPH or NADH to form a coloured product. Assays were prepared according to the manufacturer's instructions. Reagents were diluted 1:6 in "low serum" DMEM and 100 µl of this solution was added to each well. Following incubation at 37 °C for 2 hours, absorbance at 490 nm was measured using the VarioSkan Flash plate reader. To account for background absorbance, the MTS assay reagents were plated in triplicate.

### **2.5 Tissue disaggregation**

The tissue disaggregation protocol was performed as summarised in Figure 2.2. Samples were received fresh from patients with non-small cell lung cancer enrolled in the TargetLung study, undergoing surgical lung resection at Southampton General Hospital. Samples of tumour and non-involved lung were transported in 5 ml "empty" DMEM on ice. Following a five-minute wash in PBS with Amphotericin B (Gibco) to remove excess blood, samples were incised 10-15 times to relax the tissue. Samples were then added to 5 ml "complete" DMEM with DNase and enzyme (Table 2.12).

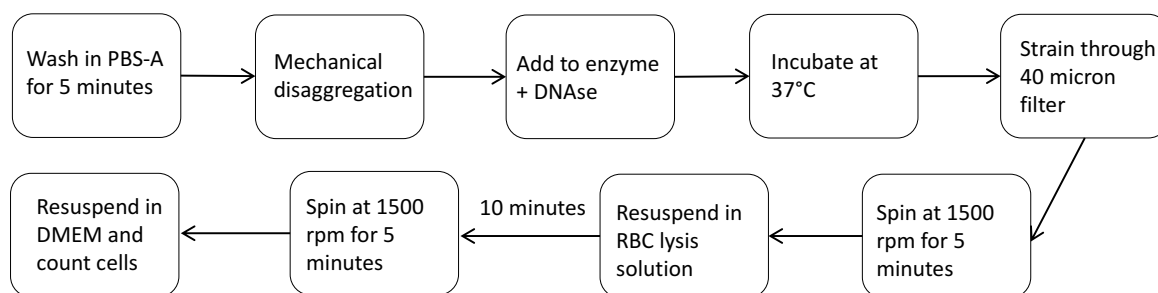


Figure 2.2 Overview of the tissue disaggregation pipeline

Reagent	Working concentration (U/ml)	Manufacturer
DNase	0.4	Sigma
Collagenase P	3	Sigma
Liberase DL/TL	0.25/1	Sigma/Roche
Liberase TM	0.25	Sigma

Table 2.12 Concentrations of enzyme and DNase tested in tissue disaggregation optimisation

Samples were agitated at 37 °C for 15 or 60 minutes, with sequential pipetting (25 ml, 10 ml and 5 ml pipettes) at 15, 30 and 60 minutes to promote tissue disaggregation. The resulting suspension was strained through a 40 µm-pore filter (Falcon, Corning), washed with 10 ml “empty” DMEM, and centrifuged at 1500 rpm for 5 minutes. The resulting pellet was re-suspended in 2 ml red cell lysis buffer (for an assumed  $1 \times 10^6$  cells; BioLegend) and incubated at 4 °C for 10 minutes. “Complete” DMEM (10 ml) was added, and the samples centrifuged at 1500 rpm for 5 minutes. The resulting pellet was re-suspended in 1 ml of either cell suspension buffer (90% ddH<sub>2</sub>O, 9% Optiprep (Sigma), 1% PBS with 0.1% BSA) for use in Drop-Seq, or “complete” DMEM for fluorescence-activated cell sorting. Any cells remaining were plated for isolation of primary cells, as described in Section 2.1.6.

## 2.6 Fluorescence-activated cell sorting (FACS)

Cell lines (Section 2.1.1) were used for initial antibody optimisation. FACS analysis was performed using  $1 \times 10^6$  cells diluted in 1 ml 10% DMEM. In order to determine the relative proportions of cell types in the single-cell suspension,  $2 \times 10^5$  cells were re-suspended in 1 ml “complete” DMEM.

## Chapter 2

All cells were first incubated with viability dye mastermix, containing 5  $\mu$ l of viability dye (7-AAD, 50  $\mu$ g/ml; BioLegend) and 105  $\mu$ l of FACS buffer *per* sample, at room temperature for 10 minutes. Samples were washed with FACS buffer, before a second, 30-minute, incubation with the staining mastermix (Table 2.13) or equivalent volume of FACS buffer for negative controls. Following two further washes with FACS buffer, cells were re-suspended in FACS buffer; 1 ml per  $1 \times 10^6$  cells.

Stained cells were analysed by flow cytometry with the FACS Canto (BD Biosciences). For all analyses, gating was performed as follows: 7-AAD vs. forward scatter area for live cells, forward scatter area vs. forward scatter height to exclude cell doublets, and side scatter area vs. forward scatter area to exclude debris. Gating to identify populations positive for each antibody was performed on populations using forward scatter area vs. the relevant antibody (see Figure 3.3). Negative populations were taken forward for further gating, meaning that EpCAM-positive populations were negative for CD45, CD31-positive populations were negative for CD45 and EpCAM, and CD90-positive populations were negative for CD45, EpCAM and CD31. For initial antibody selection and optimisation using the cell lines, histogram overlays of control and stained cells were generated. All data were analysed using the FlowJo software package (version 10.2; FlowJo, LLC).



Reagent	Working concentration (µg/ml)	Clone identifier	Fluorophore	Manufacturer	Staining mastermix (µl)
Anti-CD45 antibody	10	H130	FITC	BioLegend	5
Anti-EpCAM antibody	10	9C4	Pacific blue	BioLegend	5
Anti-CD90 antibody	5	RPA-T4	APC	BioLegend	5
Anti-PDGFR-α antibody	5	16A1	PE	BioLegend	5
Anti-PDGFR-β antibody	20	18.A2	PE	BioLegend	5
Anti-CD31 antibody	5	WM59	PE	BioLegend	5
FACS buffer	-	-	-	-	90

Table 2.13 FACS staining mastermix composition per  $1 \times 10^6$  cells

## 2.7 Single-cell RNA sequencing (scRNA-seq)

### 2.7.1 Capturing single-cell transcriptomes

A single cell suspension was generated from resection specimens as described in Section 2.5, with the resulting pellet re-suspended in cell suspension buffer. Cells were counted and diluted to 100 cells/µl: 150 000 cells were then loaded into a 3 ml Luer tip syringe (Henke Sass Wolf). Droplet Generation Oil (Bio-Rad) was loaded into a 5ml Luer tip syringe (Henke Sass Wolf). Finally, barcoded beads (Chemgenes, MA, USA) were re-suspended in cell lysis buffer (Table 2.14) to give a final concentration of 100 000 beads/ml. Beads were then loaded, with a magnetic bead, into a 3 ml Luer tip syringe. A Multi Stirrus<sup>TM</sup> magnetic stirrer

## Chapter 2

(set to a speed of 20, approximately 300 rpm; V & P Scientific, Inc.) was used to maintain the suspension and even distribution of beads in the lysis buffer.

Reagent	Manufacturer	Volume ( $\mu$ l)
H <sub>2</sub> O	Millipore	500
20% Ficoll PM-400	Sigma	300
20% Sarkosyl	Sigma	10
0.5 M EDTA	Ambion	40
2 M Tris pH 7.5	Sigma	100
1 M DTT	Sigma	50

Table 2.14 Cell lysis buffer composition (for 1 ml)

Syringes were loaded into syringe pumps (Centre for Hybrid Biodevices, University of Southampton) and attached to the microfluidic device (Figure 2.3) by plastic tubing (experimental setup shown in Figure 2.4). The device was placed on a microscope stage (Centre for Hybrid Biodevices, University of Southampton), allowing visualisation with GTK+ UVC Viewer and Pd-extended software (both open source) on a Raspberry Pi device.

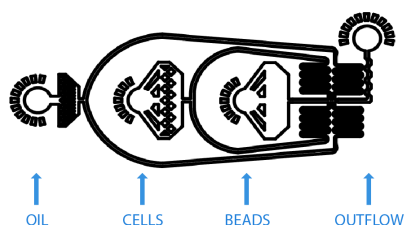


Figure 2.3 The microfluidic device used in Drop-Seq (from Macosko *et al.*<sup>131</sup>)

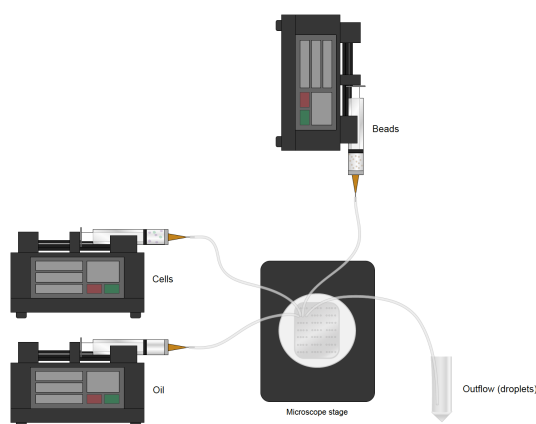


Figure 2.4 Drop-Seq experimental set-up (from Macosko *et al.*<sup>131</sup>)

Syringe pumps were started in the order: cells, beads, oil. This prevents bead solution (cell lysis buffer) from flowing back into the cell channel and lysing cells prior to droplet generation. Flow rates were set at 15 000  $\mu\text{l}/\text{hour}$  for oil and 4 000  $\mu\text{l}/\text{hour}$  for cells and beads. After outflow stabilisation (approximately 10-40 seconds), 0.2  $\mu\text{l}$  of droplets were collected to a Fuchs-Rosenthal haemocytometer chamber (NanoEnTek). Droplets were visualised using a microscope to confirm uniform droplet size, and that the rate of doublets (droplets containing two beads) was within an acceptable range (<5%). Approximately one millilitre of aqueous flow was collected. Excess oil was removed and 30 ml of 6X SSC buffer (Gibco, Life Technologies) and 1 ml perfluoro-octanol (Sigma-Aldrich) were added.

Droplets were broken (by 3-4 vigorous vertical shakes) to allow collection of the barcoded beads and bound mRNA. The suspension was spun for 1 minute at 1000  $g$  and the supernatant removed. To remove remaining oil, 30 ml of 6X SSC was added. Following a few seconds' pause to allow oil to precipitate, the supernatant was transferred to a new tube before being spun at 1000  $g$  for 1 minute. The supernatant was removed, leaving the precipitated beads. Beads were washed twice with 1 ml 6X SSC and once with 300  $\mu\text{l}$  5X reverse transcription (RT) buffer (Thermo Scientific).

### 2.7.2 cDNA synthesis, PCR and sequencing

Reverse transcription was performed to generate cDNA from the RNA hybridised to the bead primers. Two hundred microlitres of RT mix (Table 2.15) were added to each sample. Samples were incubated with agitation at room temperature for 30 minutes, then at 42  $^{\circ}\text{C}$  for 90 minutes. Samples were washed once with TE-SDS (10 mM Tris pH 8.0, 1 mM EDTA, 0.5% SDS),

## Chapter 2

once with TE-TW (10 mM Tris pH 8.0, 1 mM EDTA, 0.01% Tween-20), and once with 1 ml 10 mM Tris pH 8.0.

Reagent	Volume (μl)	Manufacturer
H <sub>2</sub> O	75	Millipore
5x RT buffer	40	Thermo Scientific
20% Ficoll PM-400	40	Sigma
10 mM dNTPs	20	Clontech
RNase inhibitor	5	Lucigen
50 μM Template Switch Oligo	10	Eurogentech
H- RTase	10	Thermo Scientific

Table 2.15 Reverse transcription mix (for 200 μl; one sample)

To remove bead primers which did not bind an RNA molecule, exonuclease I treatment was then performed. Each sample was incubated with 200 μl of exonuclease mix (Table 2.16) for 45 minutes at 37 °C, before washing once with 1 ml TE-SDS, twice with 1 ml TE-TW and once with 1 ml H<sub>2</sub>O.

Reagent	Volume (μl)	Manufacturer
Exo I buffer (10x)	20	BioLabs
H <sub>2</sub> O	170	Millipore
Exo I	10	BioLabs

Table 2.16 Exonuclease mix (for 200 μl; one sample)

Beads were re-suspended in 1 ml H<sub>2</sub>O before counting using a Fuchs-Rosenthal haemocytometer chamber. To generate a cDNA library, 2000 beads (to yield approximately 100 STAMPs; single-cell transcriptomes attached to microparticles) were added to each PCR tube. One hundred microlitres of PCR mix (Table 2.17) were added to each tube before PCR (95 °C for 3 minutes, 4 cycles of 98 °C for 20 seconds, 65 °C for 45 seconds, 72 °C for 3 minutes then 10 cycles of 98 °C for 20 seconds, 67 °C for 20 seconds, 72 °C for 3 minutes then 72 °C for 5 minutes and hold at 4 °C).

Reagent	Volume (μl)	Manufacturer
Water	25.6	Millipore
PCR primer	0.4	Integrated DNA Technologies
HiFi Hotstart Readymix	25	Kapa Biosystems

Table 2.17 PCR mixture (50 μl; one sample)

The cDNA library was purified according to the manufacturer's instructions using 30 μl AMPure XP beads (Beckman Coulter) *per* PCR tube. PCR amplicons bind these magnetic beads. When PCR tubes are placed in a magnetic plate, the amplicons bound to beads are separated from contaminants, which are removed by washing with ethanol. Purified samples were eluted in 13 μl H<sub>2</sub>O. Library quantification was performed using a BioAnalyzer High Sensitivity Chip (Agilent Genomics) according to the manufacturer's instructions.

Samples were then tagmented using the NextEra XT Library Prep Kit (Illumina). This process fragments and labels cDNA to generate a sequencing library, and allows multiplexing of samples at sequencing. For each sample, 600 pg of purified cDNA were made up to a total volume of 5 ml with H<sub>2</sub>O. Nextera TD buffer (5 μl) and Amplicon Tagment enzyme (5 μl) were added to each tube. The resulting mixture was incubated at 55 °C for 5 minutes. Five microlitres of neutralisation buffer were added before incubation at room temperature for 5 minutes. PCR mix (Table 2.18) was added to each tube before PCR was performed (95 °C for 30 seconds, 12 cycles of 95 °C for 10 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds then 72 °C for 5 minutes and hold at 4 °C).

Reagent	Volume (μl)	Manufacturer
Nextera PCR Mix	15	Illumina
H <sub>2</sub> O	8	Millipore
10 μM New-P5-SMART PCR hybrid oligo	1	Illumina
10 μM Nextera N70X oligo	1	Illumina

Table 2.18 Library PCR mixture composition (25 μl per sample)

The tagged library was purified and quantified as for the cDNA library. For sequencing, 20  $\mu$ l of a 3 nM library pool was prepared. To denature samples, 10  $\mu$ l of the library pool were incubated with 10  $\mu$ l of 0.2M NaOH for 5 minutes. Denaturing was halted by the addition of 980  $\mu$ l of HT1 buffer (Illumina). One hundred microlitres of this solution were added to a further 1.2 ml HT1 buffer to generate a 2.3 pM library for sequencing. Sample libraries were loaded to a NextSeq 500 sequencer (Illumina) for paired-end read sequencing.

### 2.7.3 Sequence alignment, transcript identification and quantification

Raw reads from the sequencer were converted to a sorted, unmapped SAM file using Picard IlluminaBasecallsToSam. Reads were filtered to remove all read pairs with a quality score of less than ten. The second of the paired reads was trimmed at the 5' end to remove adapter sequences, and at the 3' end to remove polyA tails. Reads were aligned to the human genome (hg19) using STAR<sup>165</sup>, then sorted, converged and merged to form a BAM file. Reads were tagged with the suffix -GE to allow data extraction. Digital mRNA transcript counting was performed using the DigitalExpression program<sup>131</sup>, creating a digital gene expression matrix (DGE; contains one measurement per gene per cell) for downstream analysis.

## 2.8 Bioinformatic analysis

### 2.8.1 Quality control

Bioinformatic analysis of the DGE was performed using the Seurat package<sup>166</sup> in R. Initial quality control was performed to remove dead cells and doublets (cells which adhere together or are co-isolated). Cells that are damaged or broken (equating to dead or dying cells) are most frequently identified bioinformatically by their upregulation of mitochondrially-encoded genes<sup>167,168</sup>. This is believed to be indicative of loss of cytoplasmic content: cytoplasmic mRNAs will be lost from broken cells at a greater rate than those of the mitochondria, which are retained within two mitochondrial membranes<sup>167</sup>. Therefore, to remove dead or dying cells, events with a high outlier level of mitochondrial DNA percentage were filtered and excluded from downstream analysis. Doublets were identified using a graph showing the correlation between the number of genes identified *per cell* (nGene) and the number of unique molecular identifiers *per cell* (nUMI). Outliers on this plot were then filtered and removed from subsequent analysis.

As an additional metric for the identification of low-quality cells, we calculated *per*-droplet estimates of ambient RNA capture. These calculations were based on the assumption that genes most highly expressed across each individual sample are, by probability, those most likely to account for prominent components of ambient RNA. Therefore, low-quality or empty droplets will show enrichment for these genes and a low total number of genes (nGene). To calculate these estimates, we first estimated a lower threshold for nGene, based on the distribution of nGene across all events. This was taken as 2.5 median absolute deviations (MADs) below the median for  $\log_{10}(\text{nGene})^{169}$  (64 genes in this dataset). Using the raw DGE matrix for each sequenced sample, the top sixty-four highly expressed genes (by total number of reads) were identified across all cells as an “ambient” RNA signature. The percentage of total genes detected (Ambient.RNA.genes) or reads (Ambient.RNA.counts) per cell composed of this signature was calculated, and used as a relative measure of ambient RNA incorporation.

### 2.8.2 Dimensionality reduction

In Seurat, the most highly variable genes within the dataset are calculated and used for downstream analysis. Variable genes are identified by calculating the average expression and dispersion for each gene. Genes are then assigned to a bin, and a z-score for dispersion within each bin is calculated. Thresholds for dispersion and expression are set by identifying outliers on the dispersion-expression plot.

Next, to improve the performance of further analysis, linear dimensionality reduction (principal component analysis; PCA) is carried out on the highly variable genes. PCA projection is also performed: this function scores each gene in the dataset based on its correlations with the identified principal components. This can be used to identify genes that correlate with cellular heterogeneity, but may have been excluded during selection of variable genes.

Seurat clusters cells based on their principal component scores, so it is necessary to select which principal components are included in this stage of analysis. This is performed using the “jackstraw” function, which calculates *p* values for each of the principal components. This allows identification of the true dimensionality of the dataset. Principal components with a *p* value of less than 0.001 were included in downstream analysis.

### 2.8.3 Clustering

Clustering was performed in Seurat, which uses a graph-based approach. The package constructs a  $k$ -nearest neighbours graph and refines this using the shared local neighbourhood overlap between cells. To group cells iteratively, SLM (a modularity optimisation technique) is applied. In Seurat, the clustering function contains a resolution parameter, which allows alteration of the granularity of the downstream clustering (a larger resolution value results in more clusters). The optimum resolution (*i.e.* the resolution generating the highest average silhouette width; see below) was calculated using a *for* loop in R.

Non-linear dimensional reduction (tSNE; t-distributed stochastic neighbour embedding) is used to enable visualisation of the clusters. This technique groups cells in similar neighbourhoods in high-dimensional space together in low-dimensional space: cells clustered as above co-localise on the tSNE plot. Clustering quality was assessed using the “silhouette” function in the Cluster package<sup>170</sup> in R. This provides a measure (the average silhouette width) of how similar each event is to the other cells in its assigned cluster: using a scale of 0 to 1, a high value indicates that a cell is appropriately clustered.

### 2.8.4 Identification of marker genes

Differential gene expression was used to identify the markers that define each cluster. Seurat compares each cluster to all other cells in the dataset, identifying both positive and negative markers. The function includes the “min.pct” and “thresh.use” parameters. The former allows setting of thresholds for the minimum percentage at which a gene must be detected in either of the two cell groups being tested. The ‘thresh.use’ argument specifies the minimum average difference in expression of a gene needed between the two groups. Differential expression was assessed using the likelihood-ratio test for single-cell gene expression<sup>171</sup>.

### 2.8.5 Cell type identification

Cell types were assigned to clusters using the ToppFun gene set enrichment tool<sup>172</sup>. Cluster marker genes with a  $p$  value of less than 0.001 were compared to gene sets from the Immunological Genome<sup>173</sup> and LungGENS<sup>174</sup> projects.



### 2.8.6 Merging datasets

Canonical correlation analysis (a form of dimensionality reduction) was applied using Seurat, using the highly variable genes common to both datasets. The results of this are the canonical correlation vectors, which project both datasets into the maximally correlated subspace. Similar to principal component analysis (Section 2.8.2), it is necessary to select which canonical correlation vectors to include in downstream analysis. This was performed using the `MetageneBicorPlot` Seurat function<sup>175</sup>, as the “JackStraw” function used for principal components is currently not compatible with canonical correlation vectors. The chosen vectors are then used to align the gene expression values of the two datasets, allowing direct comparisons of the integrated data.

### 2.8.7 Gene set enrichment analysis (GSEA)

Gene set enrichment analysis was performed with the Broad Institute’s GSEA program<sup>176,177</sup> using differentially expressed genes ranked by decreasing average difference. This technique examines specified genes to determine whether there is statistically significant enrichment of genes associated with, for example, specific biological processes or pathways. The generated output includes values for both the normalised enrichment score (NES) and false discovery rate (FDR)  $q$ -value. The normalised enrichment score represents an enrichment score corrected for *e.g.* differences in the gene set size; the FDR  $q$ -value represents the probability that a given gene set enrichment result is a false-positive finding. An FDR  $q$ -value  $< 0.2$  was taken as significant.

The GSEA program was also used to perform leading edge analysis. This function analyses the input gene list to identify the core genes responsible for a gene set’s enrichment score. When applied to a list of cluster marker genes, leading edge analysis identifies the genes associated with the highest number of enriched gene sets.

### 2.8.8 Trajectory analysis

Trajectory analysis was performed using the Monocle package in R<sup>178</sup>. The Monocle algorithm constructs single-cell trajectories and plots the progression of cells through a given biological process (“pseudotime”; for example, the transition from normal to cancer-associated fibroblast), assigning each cell to a branch (or “State”) depending on its progression. Differential gene expression analysis was then performed using the in-built Monocle

“differentialGeneTest” function to identify which genes define the differentiation of fibroblasts to distinct States.

## **2.9 Histological processing and analysis**

### **2.9.1 Immunohistochemical staining**

Immunohistochemical staining of sections from TargetLung patients was performed by Maria Machado using a previously-described multiplexed protocol<sup>179</sup> with optimisation by Maria Machado and Dr. Chris Hanley (University of Southampton). Four micrometre sections of formalin-fixed paraffin-embedded sections were mounted on Superfrost slides (ThermoFisher) and baked for 60 minutes at 60 °C. Deparaffinisation, rehydration, antigen retrieval and immunohistochemical staining were performed using the PT Link Autostainer (Dako) pre-defined program. Antigen retrieval for all antibodies was performed using the EnVision FLEX Target Retrieval Solution, High pH (Dako).

Sections were incubated with primary antibody (Table 2.19) for 20 minutes (with the exception of pan-cytokeratin, which was incubated for 30 minutes). Endogenous peroxidase activity was blocked using the Envision FLEX Peroxidase-Blocking reagent (Dako). EnVision FLEX HRP detection reagent (Dako) was used for secondary amplification and enzymatic conjugation. Chromogenic visualisation was performed using haematoxylin counterstaining and 2 5-minute washes in either diaminobenzidine (DAB, for AE1/AE3 staining) or 3-amino-9-ethylcarbazole (AEC, for periostin, serpin E1 and SMA staining). Following staining for cytokeratin, sections were sequentially stained for  $\alpha$ -SMA, periostin and serpin E1. Antigen retrieval was re-performed between each staining iteration, along with removal of the labile AEC staining and the previous round of antibodies using the following set of organic solvents: 50% ethanol, 2 minutes; 100% ethanol, 2 minutes; 100% xylene, 2 minutes; 100% ethanol, 2 minutes; 50% ethanol, 2 minutes. This process was repeated between each of the antibodies.

Antibody	Working dilution	Clone ID	Manufacturer
Anti-pan-cytokeratin	Pre-diluted	AE1/AE3	Dako
Anti-periostin	1:1000	Polyclonal	abcam
Anti-serpin E1	1:50	Polyclonal	Sigma
Anti- $\alpha$ -SMA	Pre-diluted	1A4	Dako

Table 2.19 Primary antibodies used in multiplexed immunohistochemical staining

### 2.9.2 Digital Pathology processing

Stained slides were scanned at 20X with the ZEISS Axio Scan.Z1, using ZEN 2 software (ZEISS). A pre-defined scan profile was used for immunohistochemical staining. Pseudo-immunofluorescence images were created using a macro in the Fiji software package, written by Dr. Chris Hanley (University of Southampton).

## 2.10 Patient characteristics

Fresh lung tissue was received from treatment-naïve patients with resectable disease enrolled in the TargetLung study (approved by NRES Committee South Central: Hampshire A, REC number 14/SC/0186) undergoing surgery at Southampton General Hospital. Patients eligible for recruitment to TargetLung are those aged 18 or over undergoing a diagnostic procedure for suspected primary lung pathology (including cancer and fibrosis) and able to give written informed consent for prospective tissue sampling. Patient characteristics are listed in Table 2.20. All patients underwent *ALK* and *EGFR* mutation status testing.

## 2.11 Statistics

Continuous data with a normal distribution were analysed using the unpaired two-tailed *t*-test. Categorical data were analysed using the Mann-Whitney *U* test. Where values are expressed as a fold change of the control, analysis was carried out with Welch's *t*-test. Comparisons across multiple categories were performed with the ordinary one-way ANOVA. Statistical analysis was performed in Prism (GraphPad) and R. Figures were prepared in Prism (GraphPad), R and Illustrator (Adobe).

## Chapter 2

Unless otherwise stated, graphs show mean values  $\pm$  the standard error of the mean (S.E.M.). Significance values are denoted as follows:  $p \geq 0.05$ : ns,  $0.01 \leq p < 0.05$ : \*,  $0.001 \leq p < 0.01$ : \*\*,  $0.0001 \leq p < 0.001$ : \*\*\*.

Age	Gender	Subtype	TNM stage			Stage	Smoking			Mutation status
			T	N	M		Status	When stopped	Pack years	
76	F	LUSC	3	0	0	2B	Ex-smoker	Not available	75	Negative
58	M	LUSC	3	2	0	3A	Current	N/A	20	Negative
79	M	LUSC	1a	0	0	1A	Ex-smoker	5 years	45	Negative
53	F	LUAD	2a	0	0	1B	Current	N/A	70	Negative
79	M	LUAD	2a	0	0	1B	Never	N/A	N/A	<i>EGFR</i> exon 19 deletion
63	F	LUAD	3	0	0	2B	Ex-smoker	15 years	Unknown	Negative
64	M	LUSC	1b	0	0	1A	Ex-smoker	8 years	Unknown	Negative
87	M	LUAD	3	2	0	3A	Ex-smoker	40 years	Unknown	<i>ALK</i> translocation
84	M	LUSC	1b	0	0	1A	Current	N/A	16	Negative
59	F	LUSC	2a	0	0	1B	Current	N/A	40	Negative
70	F	LUAD	1b	0	0	1A	Ex-smoker	40 years	Unknown	<i>EGFR</i> missense at codon 858, exon 21
69	F	LUSC	2b	0	0	2A	Unknown	N/A	N/A	Negative

Table 2.20 Characteristics of scRNA-seq patients. LUSC: squamous cell carcinoma, LUAD: adenocarcinoma



## Chapter 3      Results 1: Optimising primary tissue dissociation for single-cell analysis

### 3.1      Introduction

Fibroblasts are most commonly isolated on tissue culture plastic using serum to stimulate outgrowth from tissue, before expansion *in vitro* prior to analysis<sup>180,181</sup>. This is a well-described and reliable method of generating fibroblast cultures<sup>47,182,183</sup>. However, *in vitro* culture has been shown to alter fibroblast phenotype<sup>120</sup> and whether the functional differences described *in vitro* recapitulate *in vivo* phenotypes remains to be determined<sup>134,183</sup>.

Single-cell RNA sequencing represents a significant technological advance in gene expression profiling from human tissue samples, and could provide a valuable tool for characterising fibroblast phenotypes. To use this technology, tissue samples must be reduced to a single-cell suspension, preferably enriched with the cells of interest to reduce the costs associated with unnecessary sequencing. Enzymatic tissue disaggregation, alone or in combination with mechanical methods, has been routinely used to generate single-cell suspensions of a variety of immune cell populations<sup>135,136,184-186</sup>. Unlike immune cells, fibroblasts are typically embedded in a dense extracellular matrix, and it has not yet been determined whether previously-described approaches would be as effective for fibroblast isolation. Standardised disaggregation protocols for human solid tissues, including lung, are lacking.

Here, we first optimise a fluorescence-activated cell sorting (FACS) gating strategy and identify the most robust surface marker for lung fibroblasts. We then compare previously-described approaches for tissue disaggregation, determining the optimal method for stromal cell isolation from primary human lung tissue.

### 3.2      Optimising FACS analysis of fibroblasts

Fibroblasts are known to be heterogeneous with respect to surface marker expression: no single marker will define all fibroblasts. First, optimisation of gating for live cells was performed using cell lines, as described in Section 2.6 (Figure 3.1).

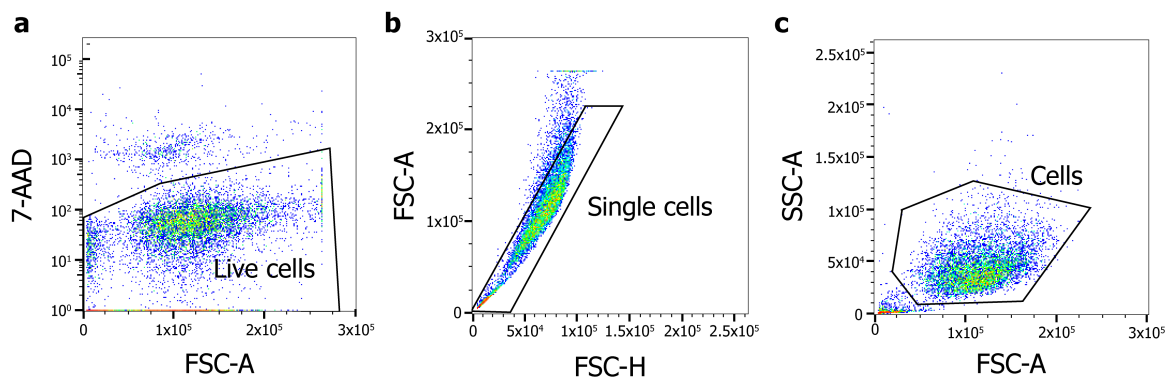


Figure 3.1 FACS gating strategy for identifying single live cells.

Scatter plots showing: (a) live cell gating, 7-AAD vs. FSC-A; (b) gating for single cells, FSC-A vs. FSC-H; (c) gating to exclude cell debris, SSC-A vs. FSC-A

The expression of three previously-described fibroblast surface markers (PDGFR- $\alpha$  and - $\beta$ , and CD90) by foetal lung (IMR-90) and skin (HFFF2) fibroblasts and a lung cancer cell line (H441) was then compared (Figure 3.2). Platelet-derived growth factor alpha (PDGFR- $\alpha$ ) has been described as a robust marker of fibroblasts<sup>29</sup>, and is expressed by up to 90% of fibroblasts in solid tumours<sup>187</sup>. CD90 (Thy-1) is a glycosylphosphatidyl-inositol-linked cell surface glycoprotein that is expressed by the majority of normal lung fibroblasts<sup>188</sup>. CD90 was found to be a robust marker of lung fibroblasts with greater sensitivity and specificity than PDGFR- $\alpha$  alone or in combination with PDGFR- $\beta$  and was therefore used in all further analyses (CD90: 99.2% positivity with 1175-fold increase in mean fluorescence intensity [MFI] compared to the negative control; PDGFR- $\alpha$ : 4.1% positivity with 4.6-fold increase in MFI; and PDGFR- $\alpha$  with - $\beta$ : 9.2% positivity with 19-fold increase in MFI; Figure 3.2). It is noteworthy that HFFF2 cells showed heterogeneous expression of CD90. This may suggest that although positive CD90 staining is a robust method to identify some fibroblast populations by FACS, this marker may not be expressed ubiquitously by all fibroblasts.



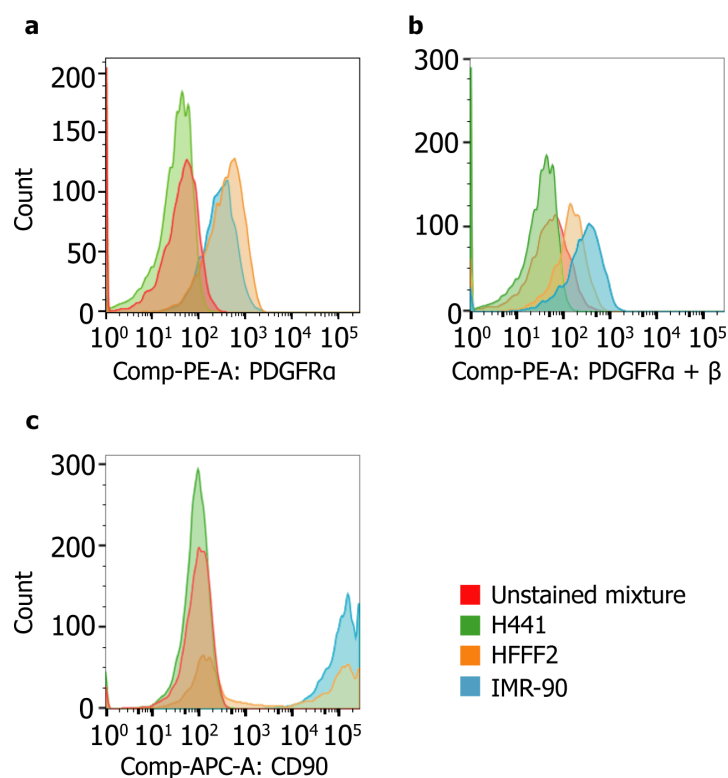


Figure 3.2 CD90 is a more robust marker of lung fibroblasts than PDGFR- $\alpha$  either alone or in combination with PDGFR- $\beta$

Histograms showing positivity of cell lines and the unstained mixture for: (a) PDGFR- $\alpha$ ; (b) PDGFR- $\alpha$  and - $\beta$  and (c) CD90

### 3.3 Tissue disaggregation using Collagenase P for 60 minutes is required to increase the number of fibroblasts isolated from tissue samples

Immune, epithelial and endothelial cells were identified using the well-described markers CD45<sup>189</sup>, EpCAM<sup>190</sup> and CD31<sup>181</sup>, respectively. The gating strategy was set using the following positive controls: Liberase-treated normal lung tissue (immune cells), H441 adenocarcinoma cells (EpCAM), HUVECs (CD31) and IMR-90 cells (CD90). These gating thresholds were applied to all future analyses (Figure 3.3).

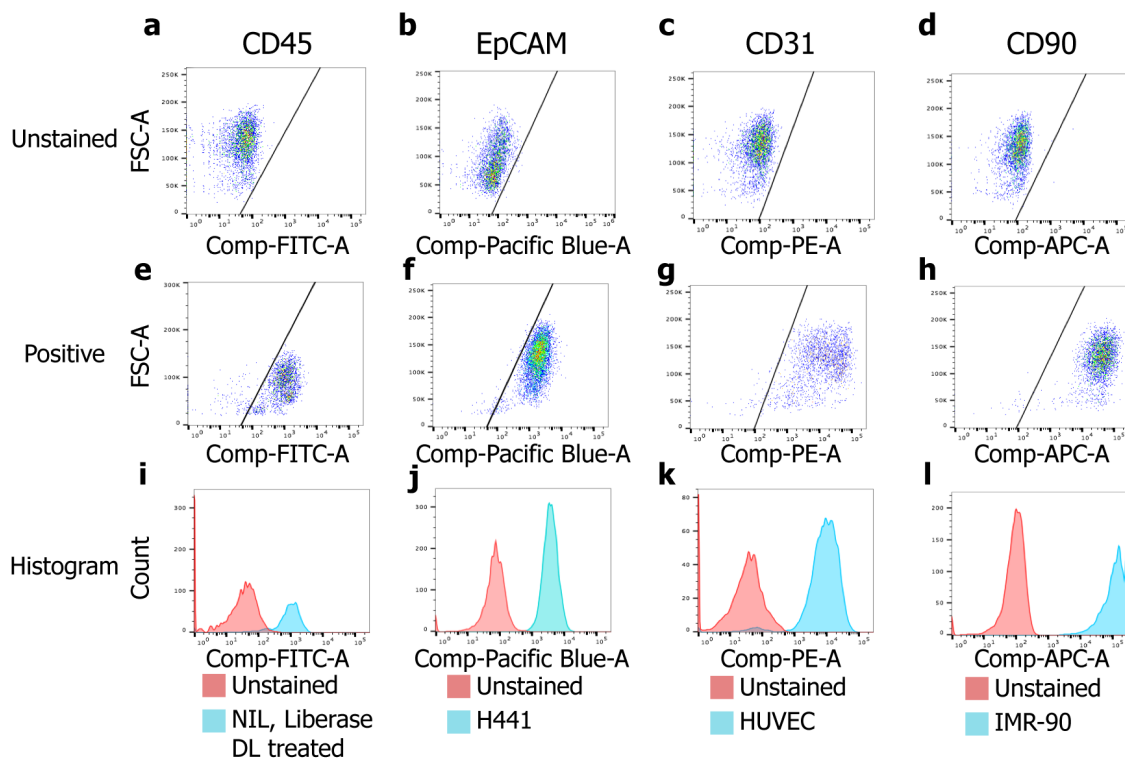


Figure 3.3 FACS panel validation and identification of positive staining thresholds

Scatter plots showing positivity of: (a-d) unstained populations, (e-h) positive control populations. (i-l) histograms showing positivity of unstained and positive control populations. Left to right: CD45, immune cells; EpCAM, epithelial cells; CD31, endothelial cells; CD90, fibroblasts

Matched tumour-normal samples were disaggregated as described in Section 2.5. FACS was then used to examine the effect of different protease cocktails (Collagenase P, Liberase DL, TL and TM) on fibroblast isolation from lung tissues. To compare directly different conditions in the same sample by FACS analysis, approximately two hundred thousand cells were needed. This required a minimum of one hundred milligrams of tissue per condition. Sufficient tissue was received from non-involved lung only, due to the smaller sizes of tumour samples received.

The effect of both the duration of tissue digestion (15 vs. 60 minutes) and the choice of enzymatic preparation (Collagenase P, Liberase DL/TL and Liberase TM; Table 3.1) were examined (Figure 3.4). Shorter disaggregation times (15 minutes) and lower protease-strength enzymes (Liberase DL, TL and TM) were insufficient to isolate stromal cells; instead yielding high proportions of immune cells (CD45+). In contrast, tissue digestion with Collagenase P for 60 minutes resulted in a greater diversity of cell types isolated (Figure 3.4a), with a significantly lower proportion of CD45+ cells and a significantly higher proportion of CD45-EpCAM-CD31-CD90+ cells (*i.e.* fibroblasts; Figure 3.4b and c). There was no significant change in the fractions of CD45-EpCAM+ (epithelial)

or CD45-EpCAM-CD31+ (endothelial) cells (Figure 3.4d and e). Dissociation enzyme and duration did not significantly affect cell yield or viability (Figure 3.5a and b).

Enzyme	Description; specific activity at working concentration
Collagenase P	Collagenase P, multiple enzyme activities; 3 U/ml
Liberase DL	Collagenase I & II, low concentration Dispase; 0.25 U/ml
Liberase TL	Collagenase I & II, low concentration Thermolysin; 0.25 U/ml
Liberase TM	Collagenase & II, medium concentration Thermolysin; 0.25 U/ml

Table 3.1 Composition of disaggregation enzymes used. Specific collagenase activity is given in Wünsch units

This analysis also revealed that epithelial cells represented a surprisingly low proportion of the cells isolated. We hypothesised that this was due to the tight intercellular interactions between these cells, with resulting cell aggregates removed either during filtration or at FACS analysis. To test this, following Collagenase P incubation for 60 minutes, single cell suspensions were additionally incubated with TrypLE (Thermo Fisher) at 37 °C for 10 minutes. This yielded a significantly higher proportion of epithelial cells (Figure 3.6a), accompanied by a reduced immune cell fraction (Figure 3.6b). The percentages of endothelial cells and fibroblasts were not significantly affected (Figure 3.6c and d). Dissociation using Collagenase P for sixty minutes with the additional TrypLE step was selected for use in all further tissue disaggregation. Together, these data show that the enzymes and incubation periods used to disaggregate tissue samples can have a significant impact on the proportions of different cell types isolated.

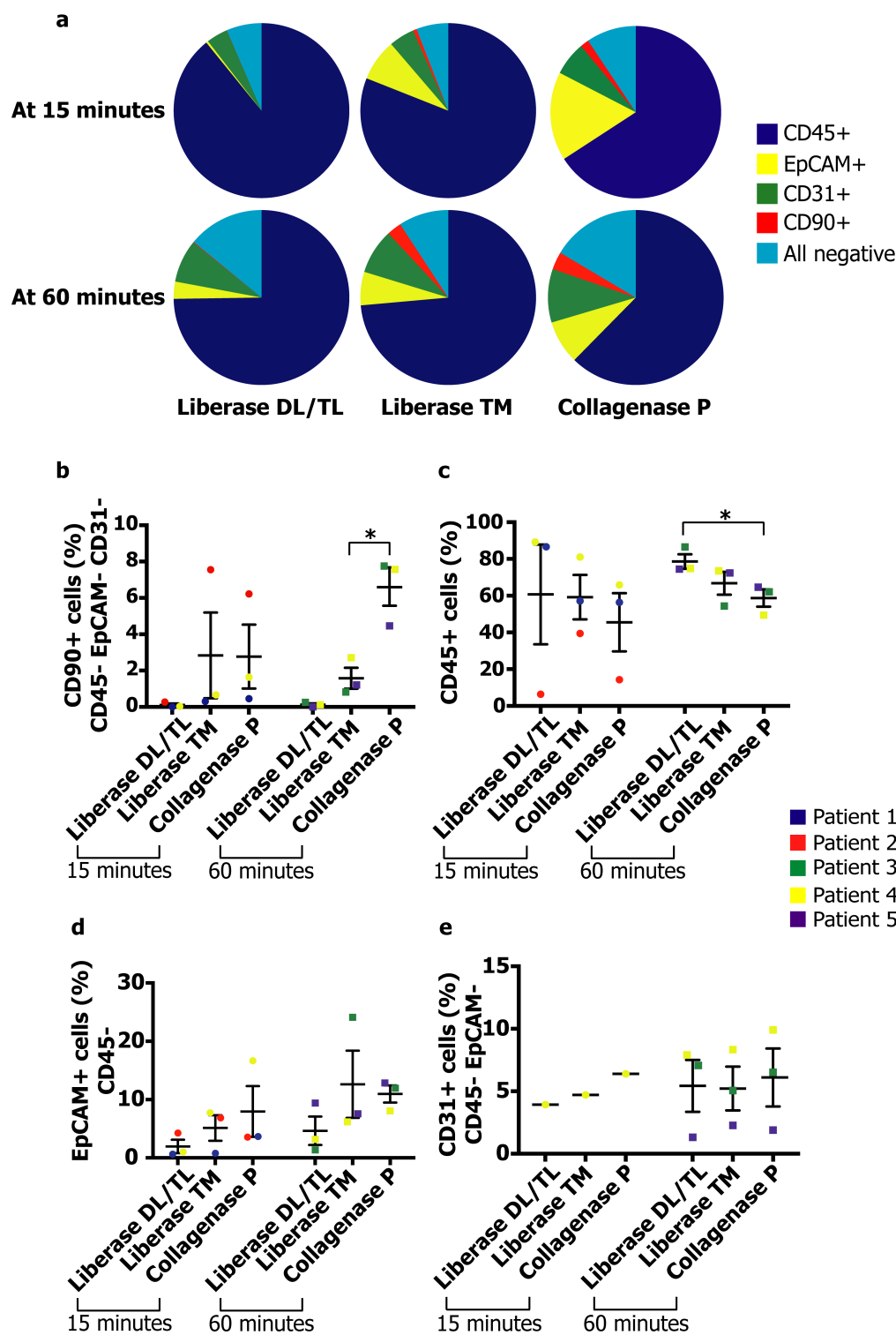


Figure 3.4 Disaggregation enzymes and incubation times have a significant impact on stromal cell isolation

(a) Representative pie charts for each disaggregation time and enzyme cocktail. Dot plots showing cell-type fractions isolated by different disaggregation procedures across human patient samples (n=5): (b) Fibroblasts (CD45-EpCAM-CD31-CD90+), (c) Immune cells (CD45+), (d) epithelial cells (CD45-EpCAM+) and (e) endothelial cells (CD45-EpCAM-CD31+). \* $p < 0.05$ , unpaired  $t$  test

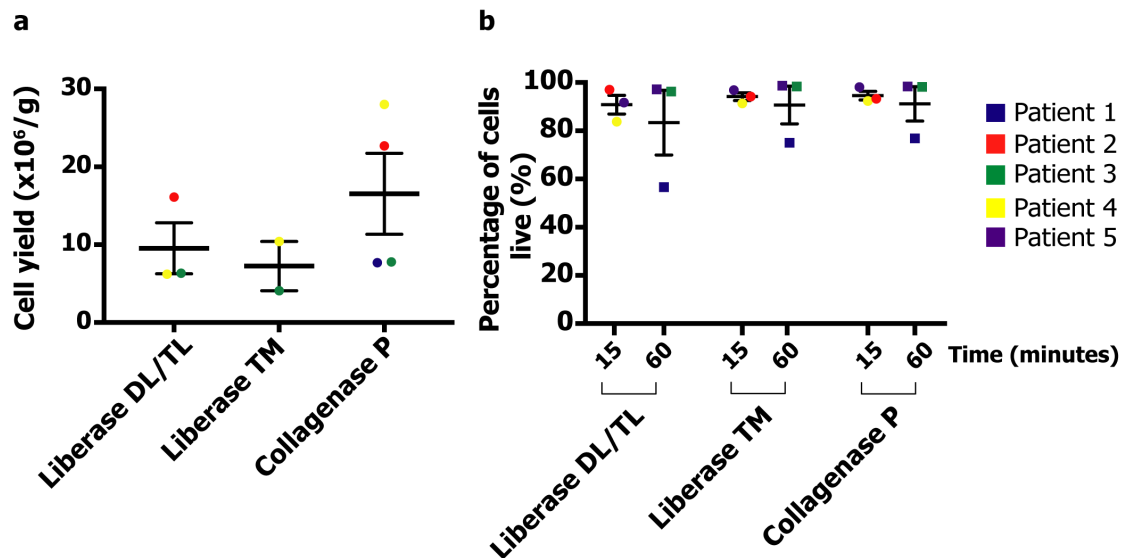


Figure 3.5 Disaggregation enzyme and duration do not affect cell yield or viability

Dot plots showing (a) cell yield and (b) cell viability percentage for different disaggregation procedures across human patient samples (n=5)

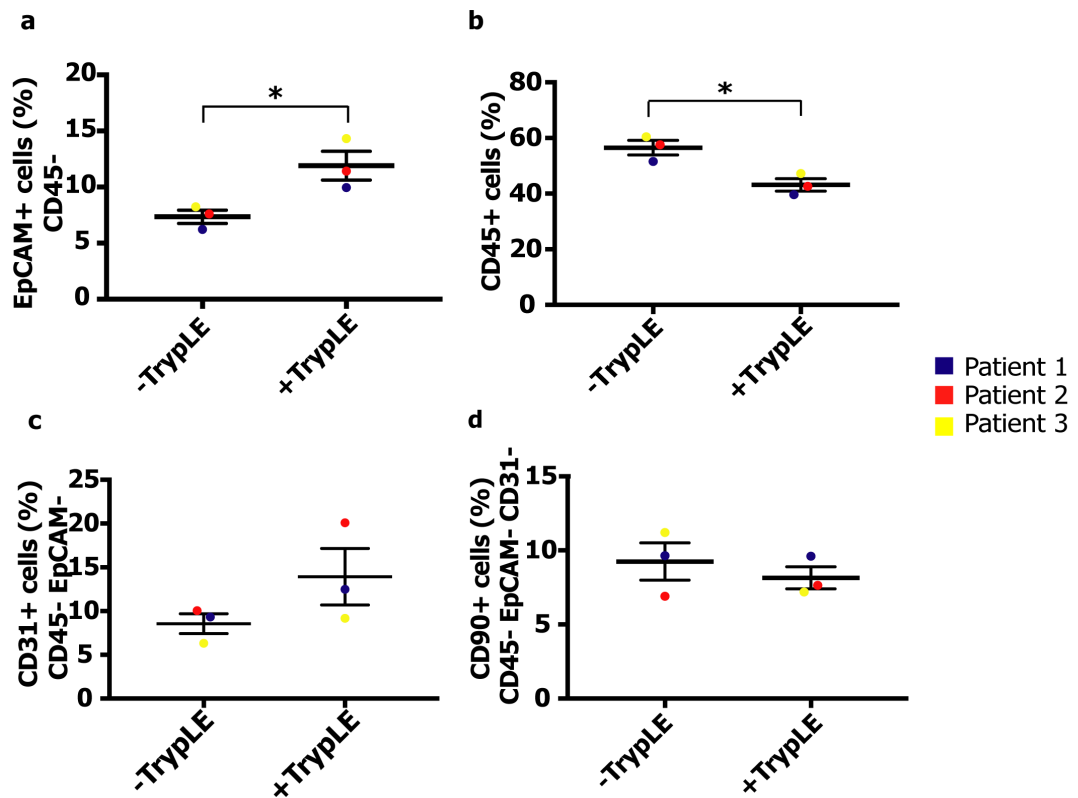


Figure 3.6 TrypLE increases the fraction of EpCAM-positive cells

Dotplots showing the percentages of cell types generated by tissue disaggregation for sixty minutes with Collagenase P, with and without TrypLE (n=3): (a) EpCAM, (b) CD90, (c) CD45 and (d) CD31. \* $p < 0.05$ , unpaired  $t$  test

3.4 Lung disaggregation with Collagenase P for sixty minutes is compatible with single-cell RNA sequencing using the drop-seq platform

To determine the suitability of the optimised disaggregation protocol for examination of gene expression profiles with single-cell RNA sequencing (scRNA-seq), we processed primary human lung tissue (a lung involved by granulomatous inflammation) and captured single-cell transcriptomes using Drop-seq<sup>131</sup>. Quantification and fragment analysis of PCR-amplified cDNA confirmed that the quantities of full-length transcripts generated were sufficient for sequencing (Figure 3.7).

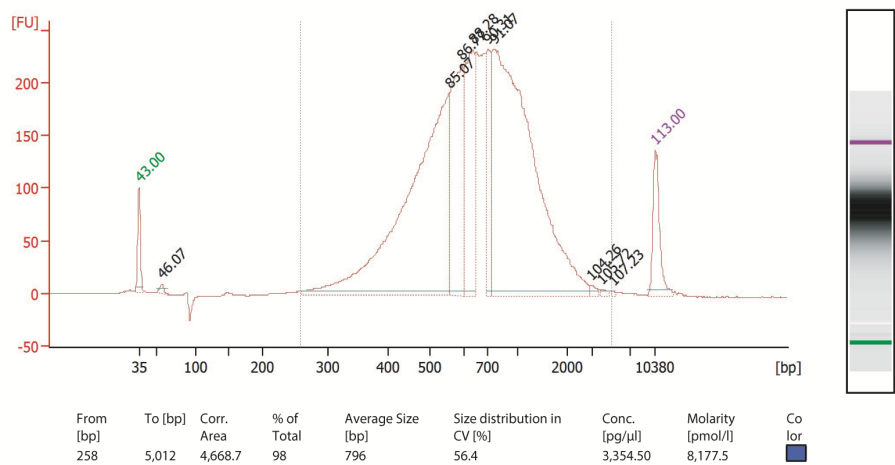


Figure 3.7 Tissue disaggregation using the optimised protocol generates sufficient cDNA quantities (> 3000 pmol/l) for single-cell RNA sequencing with Drop-seq. BioAnalyzer trace showing Tagmented cDNA library quantification

3.5 Discussion

Comparing previously-described fibroblast markers, we established that CD90 is a robust marker of lung fibroblasts. We used this marker to identify the proportion of fibroblasts generated by different disaggregation protocols, and observed that disaggregation using Collagenase P for sixty minutes yielded the highest fraction. Adding TrypLE to this approach significantly increased the proportion of epithelial cells. Quantification of the tagmented library generated using this method confirmed that this generates cDNA of sufficient quality and quantity for analysis using Drop-seq.

The described analysis demonstrates that the proportions of cell types isolated from tissues are significantly impacted by the choice of disaggregation method: this has implications for profiling tissues using this approach with, for example, single-cell RNA sequencing. There are, however, a number of limitations to this analysis. First, there is no “ground truth” with which to compare the fractions of isolated cell types, in order to determine which approach most accurately reflects the proportion of cells types seen *in vivo*. Evaluation of which method generated the highest fraction of CD90-positive cells was performed relative to the other tested approaches.

Secondly, optimisation of the disaggregation process was performed on tissue from non-involved lung only, due to the smaller mass of tumour samples available. Although disaggregation with Collagenase P for one hour generated the highest fraction of CD90-positive cells from normal tissues, there is a possibility this finding would not extrapolate to tumour samples: some differences in cellular composition are to be expected when comparing normal and malignant tissues. For example, a subset of non-small cell lung carcinomas show significant desmoplasia (a stromal reaction rich in CAF)<sup>162</sup>, and would therefore be expected to contain a high proportion of fibroblasts. In contrast, normal lung parenchyma would comprise relatively low levels of fibroblasts, given that these cells are primarily located in the adventitia of the airways and vasculature<sup>191</sup>.

Thirdly, enzymatic disaggregation has previously been shown to impact surface marker expression through indiscriminate protease activity<sup>135-137</sup>. This may limit the use of this technique for cell type identification by FACS (in addition to other downstream applications), if the markers of interest are susceptible to proteolytic cleavage by the Collagenase P enzyme cocktail. However, it is unlikely that such changes in surface marker expression will be reflected by a corresponding change in mRNA, or that this would significantly impact single-cell RNA sequencing or longer-term tissue culture.

The described heterogeneity in the fibroblast population extends to surface marker expression: there is no single marker that will reliably identify all fibroblasts. We assessed the suitability of PDGFR- $\alpha$ , - $\beta$  and CD90 (Thy-1) as use of positive surface markers for fibroblasts. CD90 was highly sensitive marker of lung fibroblasts, outperforming PDGFR- $\alpha$  and  $\beta$ , and was therefore used to identify fibroblasts from tissue samples. It is of note, however, that CD90 was expressed by fewer than half of skin fibroblasts. Furthermore, the lung cell line used for marker validation was of foetal origin: it is possible that the above approach overestimates the fraction of CD90-positive cells, should human adult fibroblasts show heterogeneous expression of this marker (as seen in animal models<sup>188,192</sup>). It is therefore likely that some fibroblasts will not be identified using the FACS panel described.

Comparison of different disaggregation durations and enzymatic cocktails showed that extended collagenase incubation times are required to release fibroblasts from tissue samples, whereas non-adherent cells (such as immune cells) are more readily and rapidly isolated by enzymatic disaggregation. Although prolonged incubation with bacterial collagenases has been reported to cause cellular damage through proteolysis<sup>134</sup>, we did not observe any significant differences in cell viability between the conditions investigated here. This protocol may be modified to isolate other cell types: we hypothesised that EpCAM-positive cells were underrepresented at FACS as a result of the removal of cellular aggregates during processing and analysis. Addition of TrypLE (an enzyme used for the dissociation of adherent cells, with a lower impact on cell viability than trypsin<sup>193</sup>) generates a significantly higher fraction of epithelial cells without affecting the proportions of stromal cells.

As mentioned previously, no single marker will identify all fibroblasts: of the markers tested, CD90 was the most sensitive and specific for lung fibroblasts, and was therefore used to identify this population in disaggregated primary lung tissues. Disaggregation using Collagenase P for one hour generated the highest proportion of CD90-positive cells relative to the other approaches examined, whilst also maintaining isolation of a range of cell types for analysis by scRNA-seq.



## Chapter 4 Results 2: Single-cell RNA sequencing and “lineage clustering”

### 4.1 Introduction

Surface marker expression in fibroblasts is heterogeneous: no single marker will identify all fibroblasts. In addition, the markers currently used (such as  $\alpha$ -smooth muscle actin,  $\alpha$ -SMA; platelet-derived growth factor, PDGFR; and fibroblast activation protein- $\alpha$ , FAP- $\alpha$ ) are also expressed by other cell types<sup>39</sup>. In order to characterise accurately fibroblast subtypes and investigate functional differences, precise phenotyping at a single-cell resolution is required.

Droplet-barcoded single-cell RNA sequencing (Drop-seq) is a relatively cost-effective platform giving broad coverage of a range of cell types<sup>130</sup>. Use of Drop-seq enables identification of different cell populations following sequencing, without the need for identification or sorting upfront. This technology has been used successfully to characterise cell populations in diverse tissues including the retina, thymus and kidney<sup>131-133</sup>. Bioinformatic analysis of the single-cell RNA sequencing data generated was primarily performed using the Seurat R package<sup>175</sup>. This package facilitates dimensionality reduction and clustering using a shared nearest neighbour-based algorithm<sup>194</sup>.

Here, we optimise a single-cell RNA sequencing analysis pipeline using lung cell lines, confirming that this approach is able to identify distinct cell lineages. We define a generalisable method for quality-control of single-cell RNA sequencing data, and assess the impact of enzymatic tissue disaggregation on gene expression. We then use this approach to analyse single-cell suspensions generated from primary tissues as described in the previous chapter. Having identified multiple distinct cell types, we then refine the identification of stromal cells within the dataset, excluding low-quality or misclassified cells from downstream analysis.

### 4.2 Identifying distinct lineages with single-cell RNA sequencing

An initial optimisation experiment was performed using a mixture of lung carcinoma and fibroblast cell lines (H441 and IMR-90; Section 2.1.1) to confirm that the Drop-Seq platform and subsequent analysis were able to separate and identify cells of different lineages. Sequencing data was aligned to the genome and a digital gene expression matrix generated as described in Section

2.7.3. A total of two hundred and one cells were sequenced. Cells that were outliers for the number of unique molecular identifiers (nUMI; indicative of low-quality cells) were identified using violin plots (Figure 4.1a). Outliers for number of genes expressed (nGene; low values suggestive of poor-quality cells, high values indicative of cell doublets) were identified on an nGene vs. nUMI scatter plot (Figure 4.1b).

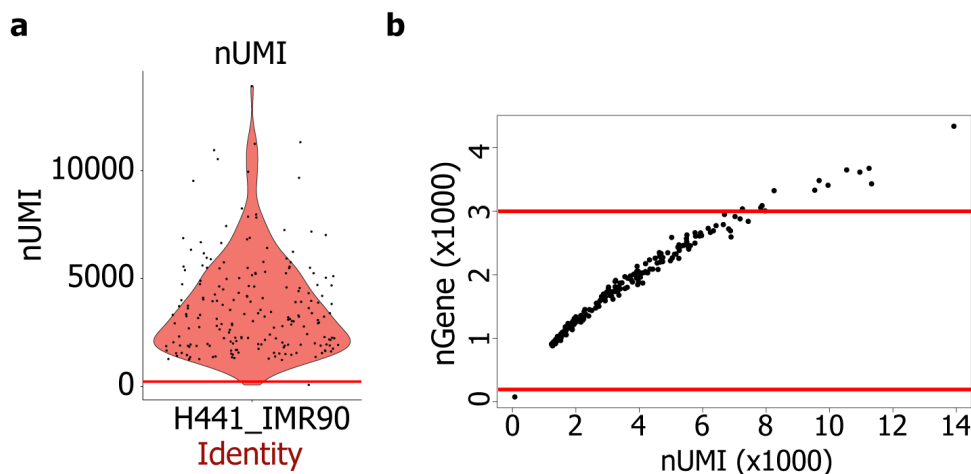


Figure 4.1 Quality control metrics for lung cell lines

(a) Violin plot of nUMI, (b) Plot of nGene vs. nUMI. Solid red lines represent exclusion thresholds for each parameter

Using these quality control metrics, cells with fewer than 200 unique molecular identifiers (nUMI) or where the number of genes (nGene) was either greater than 3000 or fewer than 100 were excluded. nUMI and the lower nGene threshold were set as *per* previous studies in this area<sup>127</sup>; the upper nGene threshold was selected based on the nGene-nUMI plot for our data. The most highly variable genes in the dataset were identified as described in Section 2.8.2.

To correct for confounding variables, nUMI and nGene were regressed against each gene. Linear dimensionality reduction of the variable genes was performed by principal component analysis (PCA; Section 2.8.2). Jackstraw analysis was used to identify significant principal components for downstream analysis<sup>166</sup>. In this dataset, principal components 1 and 2 were identified as significant ( $p < 0.001$ ). Principal component 1 (PC1) was defined by known epithelial (*EPCAM*, *KRT18*) and fibroblast (*COL1A2*, *VIM*) markers. The genes showing variation across principle component 2 (PC2) included multiple genes associated with cell replication and division (*e.g.* *CCNB1*, *CENPF*; in keeping with the higher replication rate of H441 compared to IMR-90 cells) and genes associated with fibroblast function (*e.g.* *LUM* and *ADAMTS1*, encoding an extracellular matrix protein and metalloprotease, respectively). Heatmaps showing the top ten genes for these principal components are shown in Figure 4.2.

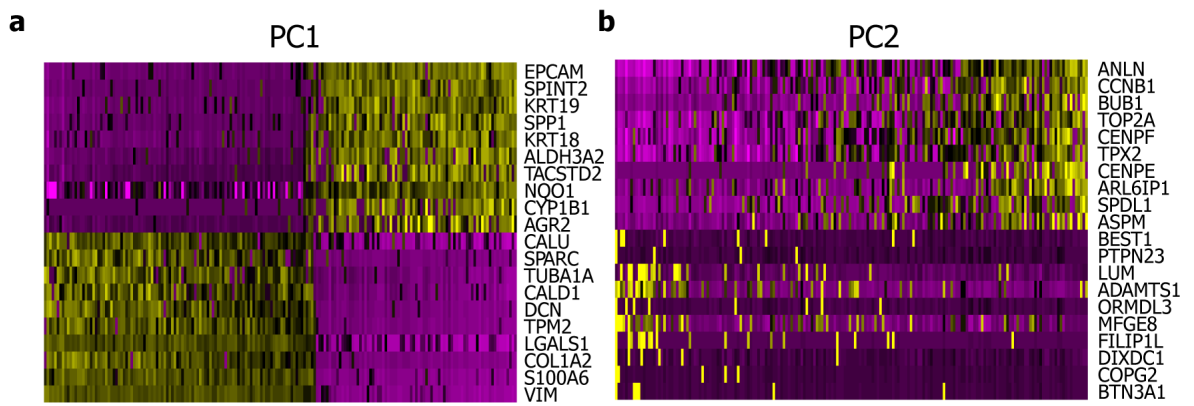


Figure 4.2 Top ten genes for principal components 1 (a) and 2 (b)

Genes (vertical axis) and cells (horizontal axis) are ordered by PCA score. Yellow indicates upregulation of the given gene by cells; purple indicates downregulation

Clustering was performed using these significant principal components (Sections 2.8.2 and 2.8.3). Non-linear dimensionality reduction was performed using tSNE to allow visualisation of the clustered data in two dimensions. The resulting tSNE plot, showing two transcriptomically distinct cell clusters, is shown in Figure 4.3a. These two clusters clearly represented the H441 and IMR-90 cells, as demonstrated by their expression of the epithelial and mesenchymal marker genes *EPCAM* and *VIM* respectively (Figure 4.3b and c). Figure 4.4 shows a heatmap of the top differentially expressed genes for each cluster, which include additional well-described epithelial (e.g. cytokeratins *KRT18* and *KRT19*) and mesenchymal (e.g. *COL1A2* and *S100A6*) markers.

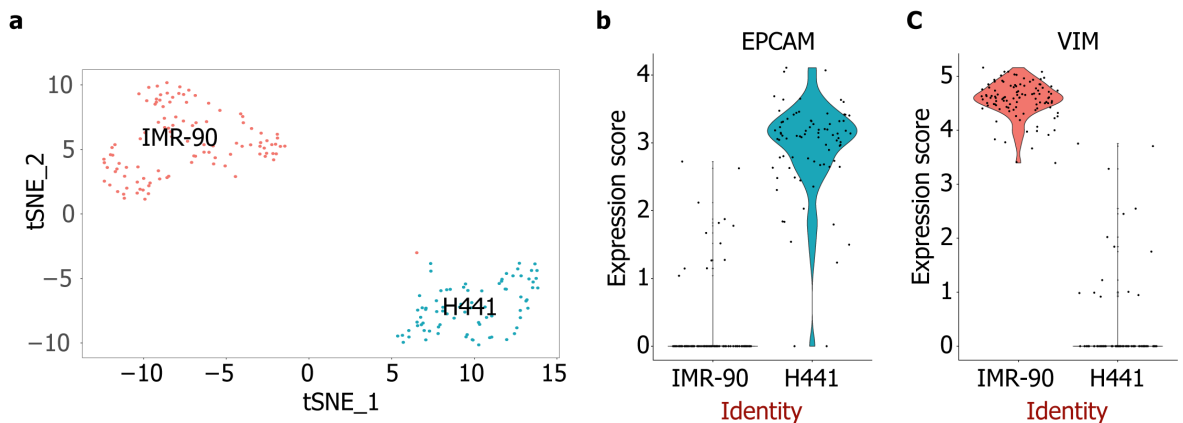


Figure 4.3 Drop-seq identifies distinct lung cell line lineages

(a) tSNE plot of whole dataset showing two distinct clusters. Each point represents a single cell, groups of cells with similar transcriptomes are referred to as a “cluster”; clusters are distinguished by colour and labelled by cell type. Violin plots showing expression of (b) *EPCAM* and (c) *VIM* across the clusters

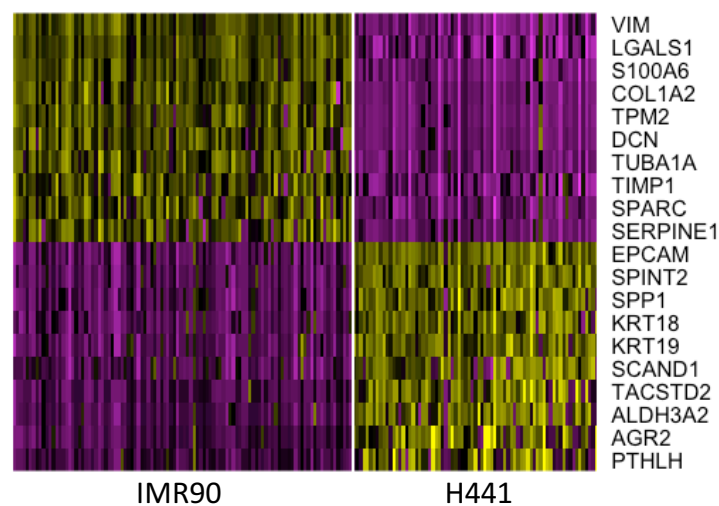


Figure 4.4 Heatmap showing the top ten differentially expressed genes by cluster

Heatmap showing the top ten differentially expressed genes by cluster. Genes (vertical axis) and cells (horizontal axis) are ordered by PCA score. Yellow indicates upregulation of the given gene by cells, purple indicates downregulation

4.3 Optimising the analysis pipeline

The Drop-Seq platform (outlined in Section 2.7) was used to process the single-cell suspension generated by tissue disaggregation of primary lung tissues (described in Sections 2.5 and 3.4). In total, 11 600 cells from 12 tumour (7 squamous cell carcinoma and 5 adenocarcinoma) and 6 patient-matched normal samples were sequenced.

4.3.1 Determining quality-control thresholds

Variable gene analysis was initially performed using the quality control parameters described in the previous section. The resulting tSNE plot is shown in Figure 4.5.

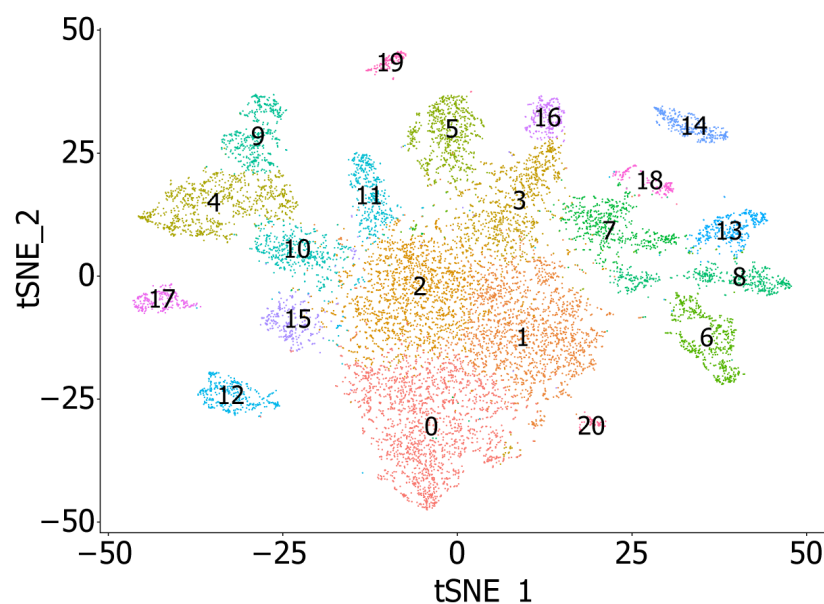


Figure 4.5 Filtering data from primary lung tissues using parameters for cell lines identifies multiple clusters

tSNE plot showing data from 11 600 primary cells. Each point represents a single cell, groups of cells with similar transcriptomes are referred to as a “cluster”; clusters are distinguished by colour and labelled by cell type

This analysis identified multiple clusters. However, in contrast to the cell lines dataset, where the two cell types were clearly distinguishable from each other (Figure 4.3), these clusters were ill-defined and indistinct, with poor separation. To investigate this further, and determine whether it was possible to refine this clustering, we examined the impact of potential sources of noise within this primary tissue dataset compared to the cell lines dataset.

The nUMI counts and nGene thresholds were assessed as previously (Figure 4.6). Compared to cell lines, cells from primary tissues showed a wider range of nUMI counts and number of genes per cell (Figure 4.6a). In comparison to *ex vivo* dataset, the data from cell lines showed higher complexity (a measure of the number of unique transcripts within the sequenced library), with a median nUMI of 3088 vs. 910 in the primary lung data. The *ex vivo* data also showed a lower median nGene than the cell lines, although with a greater range (Figure 4.6b). To avoid exclusion of cells containing potentially useful information in this dataset, a higher upper threshold for nGene would be required.

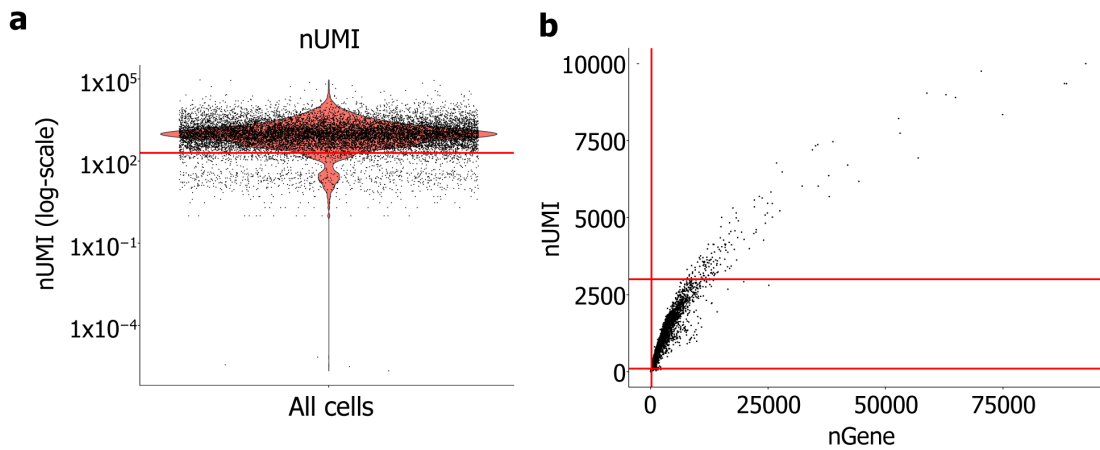


Figure 4.6 Quality control metrics for whole lung samples

(a) Violin plot of nUMI, (b) Plot of nGene vs. nUMI. Each point represents a single cell; solid red lines represent exclusion thresholds based on cell line data

To assess the difference in viability between cell lines and primary tissues, the proportion of reads mapping to mitochondrial genes was examined. This metric (“percent.mito”) is used as a measure of cell death<sup>167,168</sup> (Section 2.8.1). Attempts to validate this metric were unsuccessful: experimental induction of apoptosis in primary cell cultures yielded cDNA concentrations insufficient for sequencing (Section A.2). Primary cells had a much higher fraction of reads mapping to mitochondrial genes (representing dead or dying cells; Figure 4.7). To prevent inclusion of reads from dead cells in downstream analysis, cells for which more than 20% of reads mapped to mitochondrial genes were excluded (removing outliers with a particularly high percentage of mitochondrial reads).

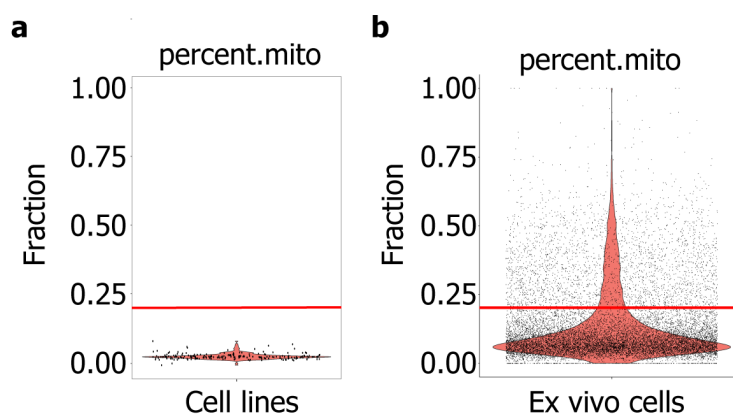


Figure 4.7 *Ex vivo* cells show a higher fraction of reads mapping to mitochondrial genes

Violin plots showing mitochondrial gene fractions from (a) H441 and IMR90 cell lines and (b) whole lung tumours. Each point represents a cell; the solid red line indicates the 20% maximum threshold used for filtering *ex vivo* cells

These quality control metrics were overlaid onto tSNE plots to demonstrate the impact of excluding these low-quality cells on nUMI, nGene and percent.mito (Figure 4.8). This clearly demonstrates that these lower quality cells cluster together, making up a large proportion of the cells at the interfaces between the more poorly-defined clusters. Identifying and removing these cells should therefore enable refined clustering of the different cell types.

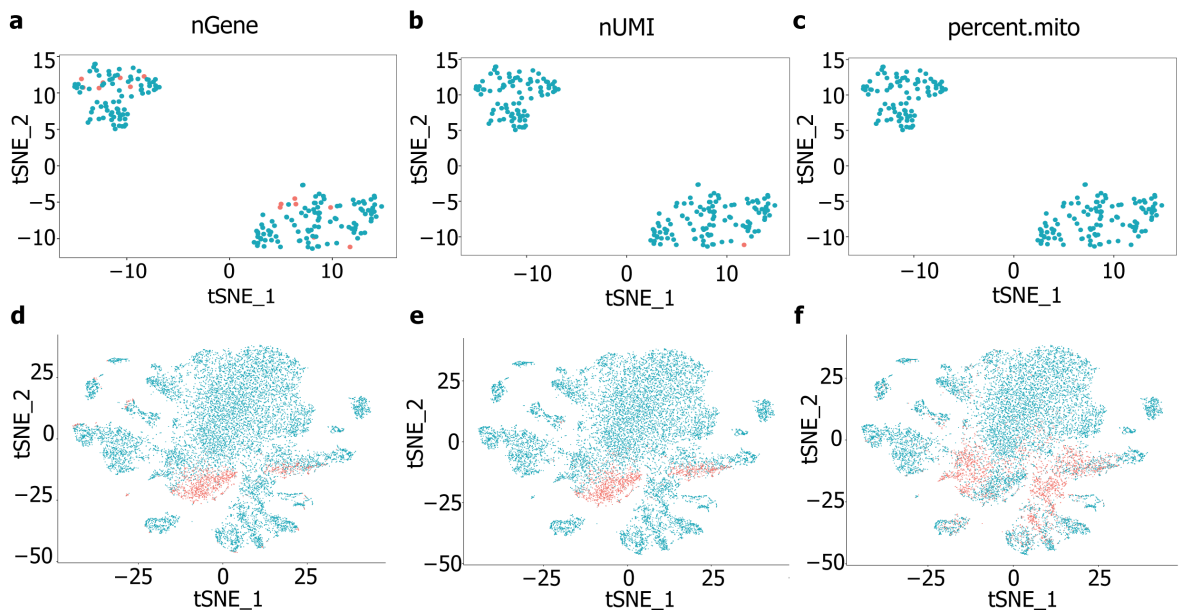


Figure 4.8 Lung cell lines and *ex vivo* cells show differing library complexity and viability

tSNE plots for nGene, nUMI and percent.mito. (a-c): H441 and IMR-90 cell lines, (d-f): whole primary lung samples. Each point represents a cell; cells coloured red are removed using the respective filter (nUMI < 200; nGene < 100 or > 3000; percent.mito > 20%)

#### 4.3.2 Identifying low-quality droplets

Identification of low-quality events (such as sequenced cell fragments, or cells with an insufficient sequencing depth) is one of the essential quality-control steps prior to downstream analysis of scRNA-seq data. Droplet-based platforms have become increasingly popular as a result of their relative affordability and commercial availability. However, identification of low-quality events in the resulting data can be challenging, as these platforms typically yield lower sequencing depths. In particular, it may be difficult to distinguish low-quality droplets from true cells with a low nGene (*e.g.* lymphocytes<sup>127</sup>). We first evaluated the variation in nGene across cell types, using a subset of the unfiltered data. Cell type identification was based on expression of canonical markers (Table 4.1). As no such currently marker exists for fibroblasts, CD90 (which we found to

be a robust markers of cell line lung fibroblasts; Section 3.2) was initially used to identify fibroblast clusters in the primary lung dataset. Differential gene expression analysis (Section 2.8.4) performed on the whole dataset identified *DCN* (encoding decorin, a collagen-binding protein which regulates fibrillogenesis<sup>195</sup>) as the gene showing greatest differential expression by fibroblasts in our data (average log fold change 3.35). In contrast, *CD90* had an average log fold change of 1.45; *DCN* was therefore selected as a fibroblast marker for downstream analysis. The resulting violin plot is shown in Figure 4.9. The number of genes *per* cell varies considerably across cell types: for example, T cells and plasma cells show a particularly low nGene in comparison to other cell types.

Cell type	Marker
T cells	<i>TRBC2</i>
CD8+ T cells	<i>CD8A</i>
NK cells	<i>NKG7</i>
Myeloid cells	<i>LYZ</i>
Mast cells	<i>KIT</i>
B cells	<i>MS4A1 (CD20)</i>
Fibroblasts	<i>DCN</i>
Epithelial cells	<i>KRT18</i>
Endothelial cells	<i>EPAS1</i>

Table 4.1 Canonical cell lineage markers



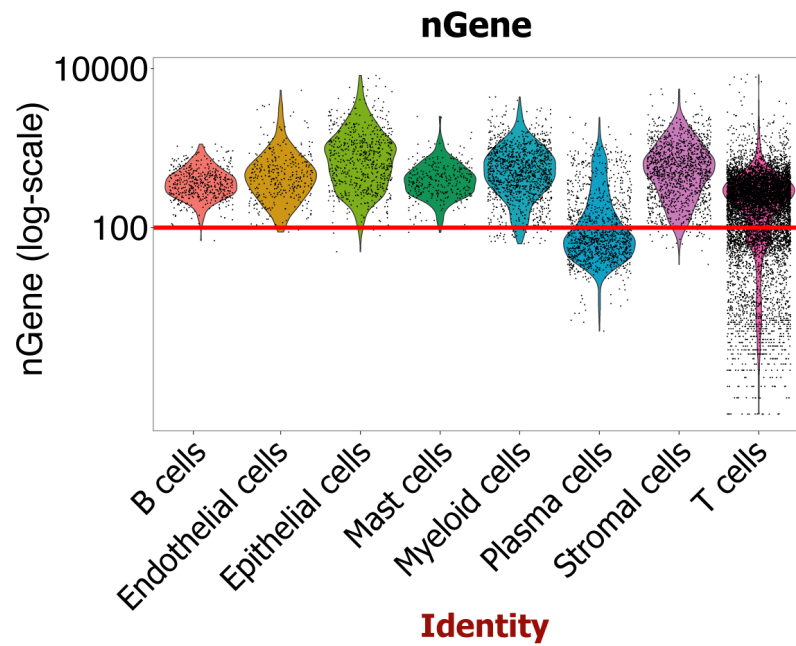


Figure 4.9 Violin plot showing the number of genes (nGene) in cells of each type, using data without applied filters

Each black point represents a cell; the red line represents the lower threshold of 100

The approaches used to perform filtering for low-quality events have not previously been standardised, and vary considerably between studies<sup>125,127</sup>. Variability in quality-control parameters can be inherent to the platform used. For example, SMART-seq2 affords a greater depth of sequencing<sup>130</sup>, and the lowest acceptable number of genes *per* cell (nGene) will therefore be higher than for data generated using droplet-based technologies (*e.g.* 10X and Drop-Seq)<sup>125,127</sup>.

Given the lack of standardisation, we sought to develop a generalisable approach for the pre-processing of droplet-based scRNA-seq data, with the aim of improving filtering and thus cell type identification by unsupervised clustering. First, we assessed baseline clustering quality (calculating the average silhouette width, a commonly-used measure of clustering quality; Section 2.8.3) using unfiltered data. This identified 17 clusters, with an average silhouette width of 0.45 (Figure 4.10a). We also applied quality control (QC) thresholds described in previous studies (removing cells with more than 15000 or fewer than 200 nUMIs, over 5000 or below 100 genes, or over 20% of reads mapping to mitochondrial genes)<sup>125-127</sup>, finding that this improved clustering quality (average silhouette width 0.51; Figure 4.10b).

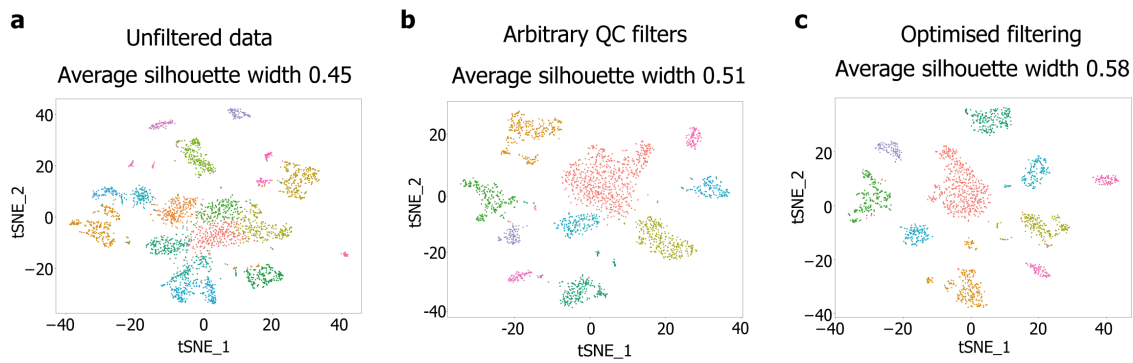


Figure 4.10 Standardised quality-control metrics improve clustering quality of scRNA-seq data

tSNE plots showing principal component-based clustering and average silhouette width for: **(a)** unfiltered data, **(b)** data filtered using widely-used quality-control metrics and **(c)** the optimised approach described here. Each point represents an individual cell, groups of cells with similar transcriptomes are referred to as a “cluster” and distinguished by colour

To determine the optimal approach, we examined the variation of a range of QC metrics between events assumed to be low-quality (nGene less than 100) and those likely to represent true cells (nGene between the median and 2.5 MAD above the median; Figure 4.11a). QC metrics comprised: nUMI, the percentage of reads mapping to mitochondrial (percent.mito) or ribosomal genes (percent.ribo), the ratio between these two values, and housekeeping gene expression. We additionally defined an algorithm to estimate the contribution of both reads and genes detected due to encapsulation of “ambient” RNA (Section 2.8.1). Each of these metrics showed a significant difference between the low-quality droplets and true cells (Bonferroni corrected  $p < 0.0001$ , Figure 4.11b-h).

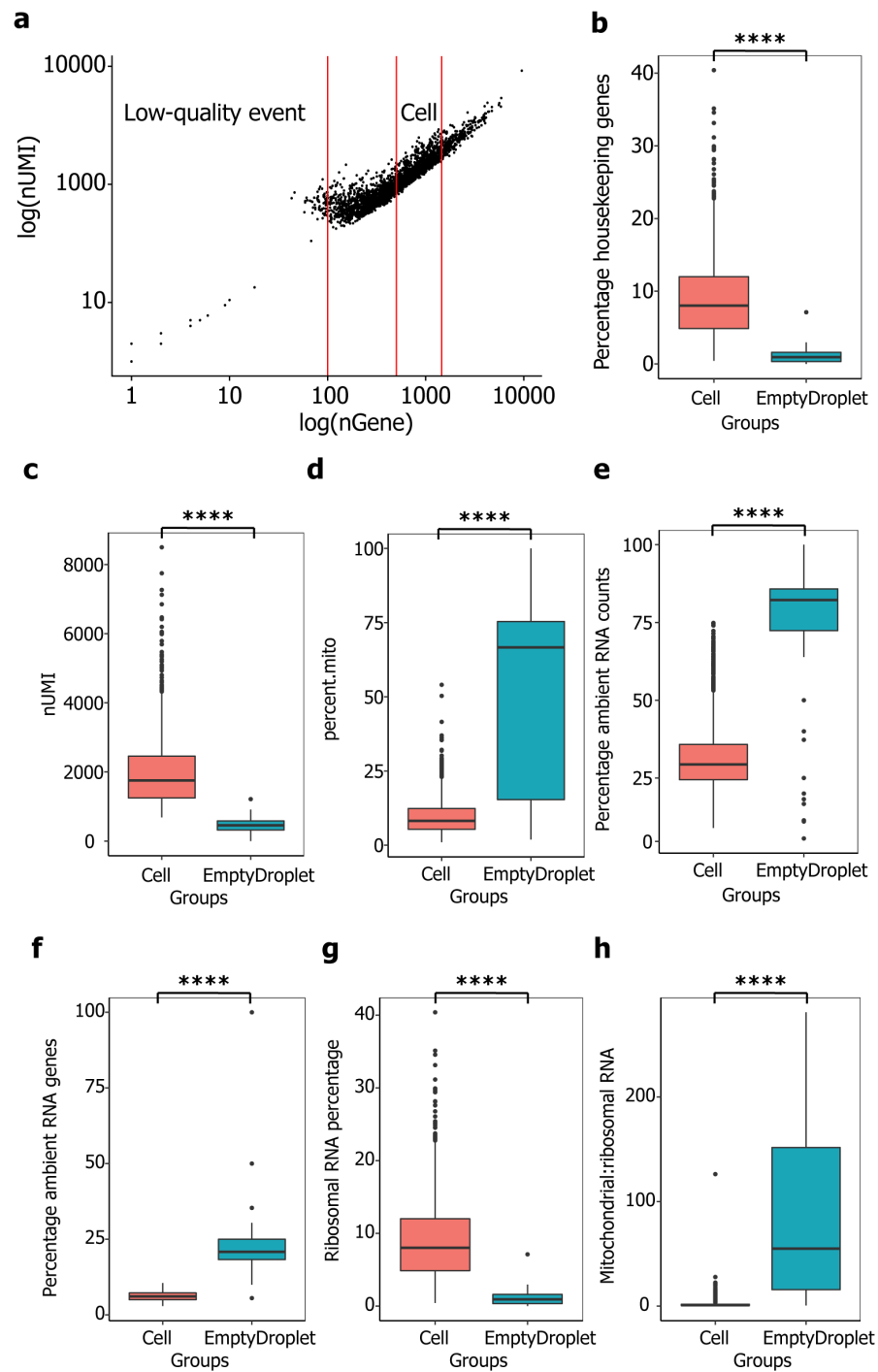


Figure 4.11 Identifying quality-control metrics with potential to distinguish “low-quality” and “cell” groups

(a) Plot of  $\log(\text{nGene})$  vs.  $\log(\text{nUMI})$  showing events classified as low-quality or cells. (b-h) Boxplots showing differences between the low-quality and cell groups for: (b): percentage of reads mapping to housekeeping genes, (c) nUMI, (d) percentage of reads mapping to mitochondrial genes, (e) percentage of counts associated with the ambient RNA signature, (f) ambient RNA gene percentage, (g) percentage of reads mapping to ribosomal genes and (h) ratio between percentage of reads mapping to mitochondrial or ribosomal genes. \*\*\*\* $p < 0.0001$ , unpaired two-tailed  $t$ -test

To harness the predictive power of each variable, we trained a machine learning (random forest<sup>196</sup>) model to distinguish between the low-quality droplet and true cell groups. The data was divided at random into “training” and “test” datasets (in a ratio of 3 to 1). When applying the trained classifier to the “test” dataset, “cells” were detected at a sensitivity and specificity of 100%. Analysing the relative importance of each metric in the classifier<sup>197</sup> showed that the ratio between the number of “ambient” genes detected and total number of genes detected (Ambient.RNA.genes) was the most important variable in distinguishing between cells and low-quality droplets (Figure 4.12a). Highlighting the droplets identified as low-quality by the classifier, on a log(nUMI) vs. log(nGene) plot (Figure 4.12b), illustrates that this machine learning approach to filtering provides greater sensitivity than filtering using hard thresholds for individual QC metrics.

Cell-cell doublets were identified on an nUMI vs. nGene plot (Figure 4.12c) and removed from downstream analysis. Dead or dying cells can also impact clustering and analysis; we therefore removed cells with a percent.mito greater than 2.5 MAD above the median (Figure 4.12d). This combined approach further improved the quality of clustering (average silhouette width 0.58; Figure 4.10d).

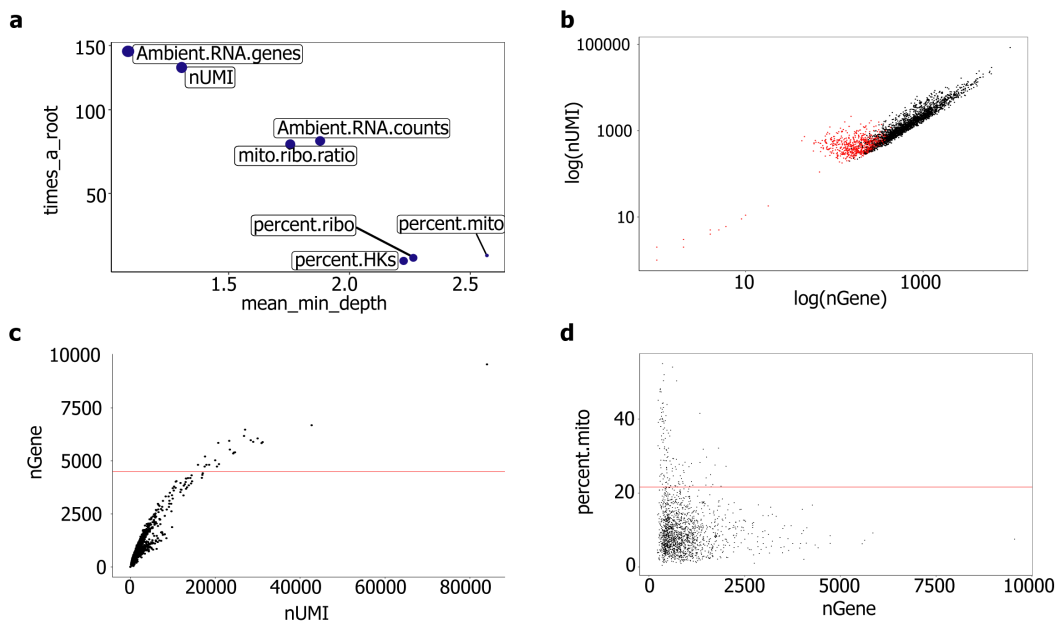


Figure 4.12 Implementation of quality-control metrics to remove low-quality droplets

(a) Multi-way importance plot showing the relative importance of each metric in distinguishing between low-quality events and droplets, (b) Plot of log(nGene) vs. log(nUMI). Data points identified as low-quality events are highlighted in red. (c) Plot of nUMI vs. nGene. The red line indicates the upper nGene threshold (defined as 2.5 MAD above the median) and (d) Plot of nGene vs. percent.mito. The red line indicates the upper threshold (defined as 2.5 MAD above the median)

This approach compares favourably with previously-described methods for low-quality cell removal: DecontX and emptyDrops, functions in the *celda*<sup>198</sup> and *DropletUtils*<sup>199</sup> R packages, respectively. DecontX is based on a principle similar to that of our “ambient” RNA score: namely, that the top differentially expressed genes for each cluster will be expressed at low levels in other clusters as a result of their contaminating the cell suspension. The algorithm scales the observed data to remove reads associated with contaminating RNA.

One drawback of this technique is that it requires a vector of cluster labels (*i.e.* upfront labelling of which cells belong to which cluster): this will be affected by the resolution used for clustering. A higher resolution will yield a higher number of clusters; the algorithm incorporates the top differentially expressed genes from each cluster. Therefore, increasing the number of clusters will increase the number of genes included in the contaminating signature, removing more counts from cells. Using the cluster labels assigned at the resolution that generated the highest average silhouette width (0.7), applying this approach to our data did not improve clustering quality (0.44 vs. 0.45 for unfiltered data). Using the cluster labels assigned by an arbitrary low resolution (0.1; to reduce the number of genes included in the contaminating signature) improves clustering quality to a level similar to using previously-described arbitrary pre-processing steps (0.52; see above).

The emptyDrops function identifies low-quality droplets by creating a gene expression profile for “empty” cells (defined by the authors as  $nUMI < 100$ <sup>199</sup>). Events whose gene expression profile does not differ significantly from this signature are designated empty cells and removed from the gene expression matrix. Identification of events classified as low quality by emptyDrops is shown in Figure 4.13a. Comparing this to the cells removed by our approach (Figure 4.13b), it is clear that emptyDrops labels far fewer events as low quality than does our random forest classifier: application of this function to our data did not improve clustering quality (0.46 vs. 0.45 for unfiltered data). The maximum average silhouette width achieved using each of the above quality-control approaches are summarised in Figure 4.14.

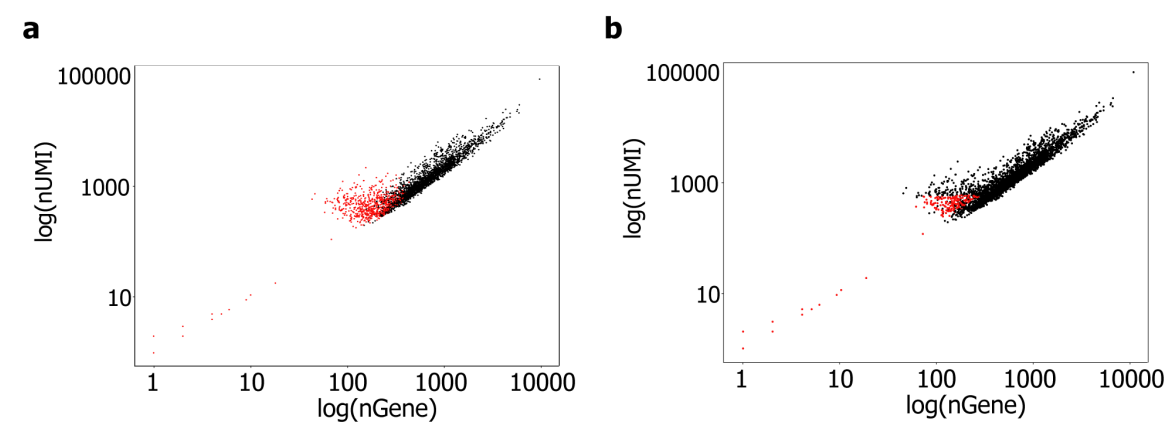


Figure 4.13 The emptyDrops function removes fewer low-quality events than does the random forest classifier

Dot plots showing plots of  $\log(nGene)$  vs.  $\log(nUMI)$  for (a) the optimised approach described above and (b) emptyDrops. Data points identified as low-quality events are highlighted in red

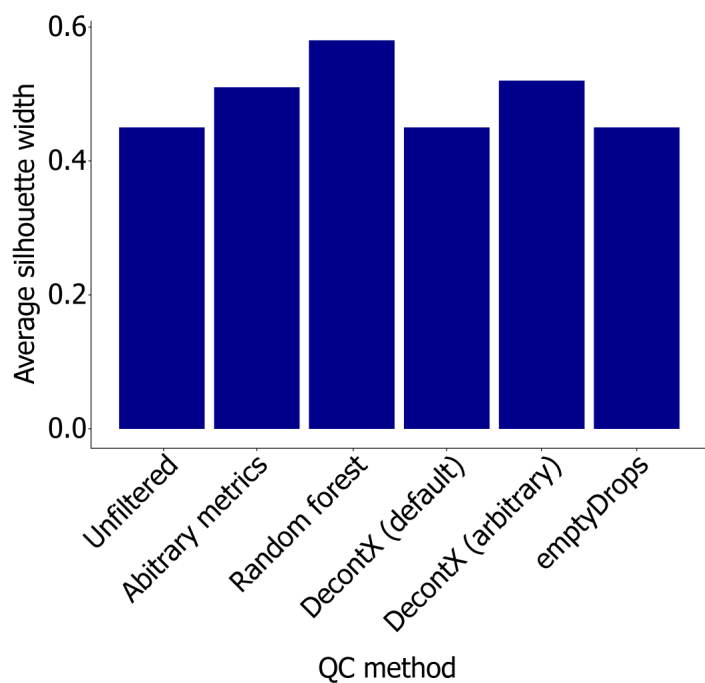


Figure 4.14 Use of a random forest classifier improves clustering quality relative to other commonly-used methods

Bar plot showing maximum average silhouette width for each of the examined approaches. The DecontX function<sup>198</sup> was assessed both using the default settings and using an arbitrary low resolution

### 4.3.3 Identifying disaggregation-associated gene expression changes

Analysis of the scRNA-seq data confirmed that disaggregation for 60 minutes yields a greater fraction of stromal cells (as well as more even coverage of other cell types) than disaggregation for 15 minutes (Figure 4.15a). However, previous authors have shown that enzymatic disaggregation can lead to changes in gene expression, describing a disaggregation-associated gene signature (derived using murine stem cells)<sup>138</sup>. We assessed the effect of applying this signature to our data: as expected, its expression was increased in samples disaggregated for 60 minutes compared to those processed for 15 minutes (Figure 4.15c). However, we found that this signature did not appear to impact cell clustering, and was not a prominent feature of any individual cluster (Figure 4.15e).

As the above gene signature was derived from murine experiments, and includes markers for human cell types (*e.g.* *DCN*, expressed by stromal cells in a number of datasets, including our own<sup>51,80,122,200</sup>), we sought to refine this signature based on our analysis. We therefore cross-referenced the genes in this signature with those upregulated in our data ( $p < 0.001$ , average  $\log_{10}(\text{fold change}) > 1$ ) following disaggregation for 60 minutes vs. 15 minutes, excluding those identified as cell type markers. This identified a list of eleven genes which were consistently upregulated following extended disaggregation, and were not upregulated by a particular cell type (Figure 4.15b). We applied this list to our data as a refined disaggregation signature.

The distribution of expression of this signature in our dataset is shown in Figure 4.15d and f. Clusters 0, 6 and 9 (representing T cells, B cells and mast cells) show higher expression, indicating that some cell types may be more sensitive to extended enzymatic disaggregation than others. However, none of the principal components used for clustering correlated with this gene signature ( $p > 0.23$ ): it is therefore unlikely that disaggregation-induced changes in gene expression have impacted cell clustering. Removal of cells with high expression (greater than 2.5 MAD above the median) of this signature did not improve clustering quality (maximum average silhouette width 0.56 vs. 0.58 prior to filtering).

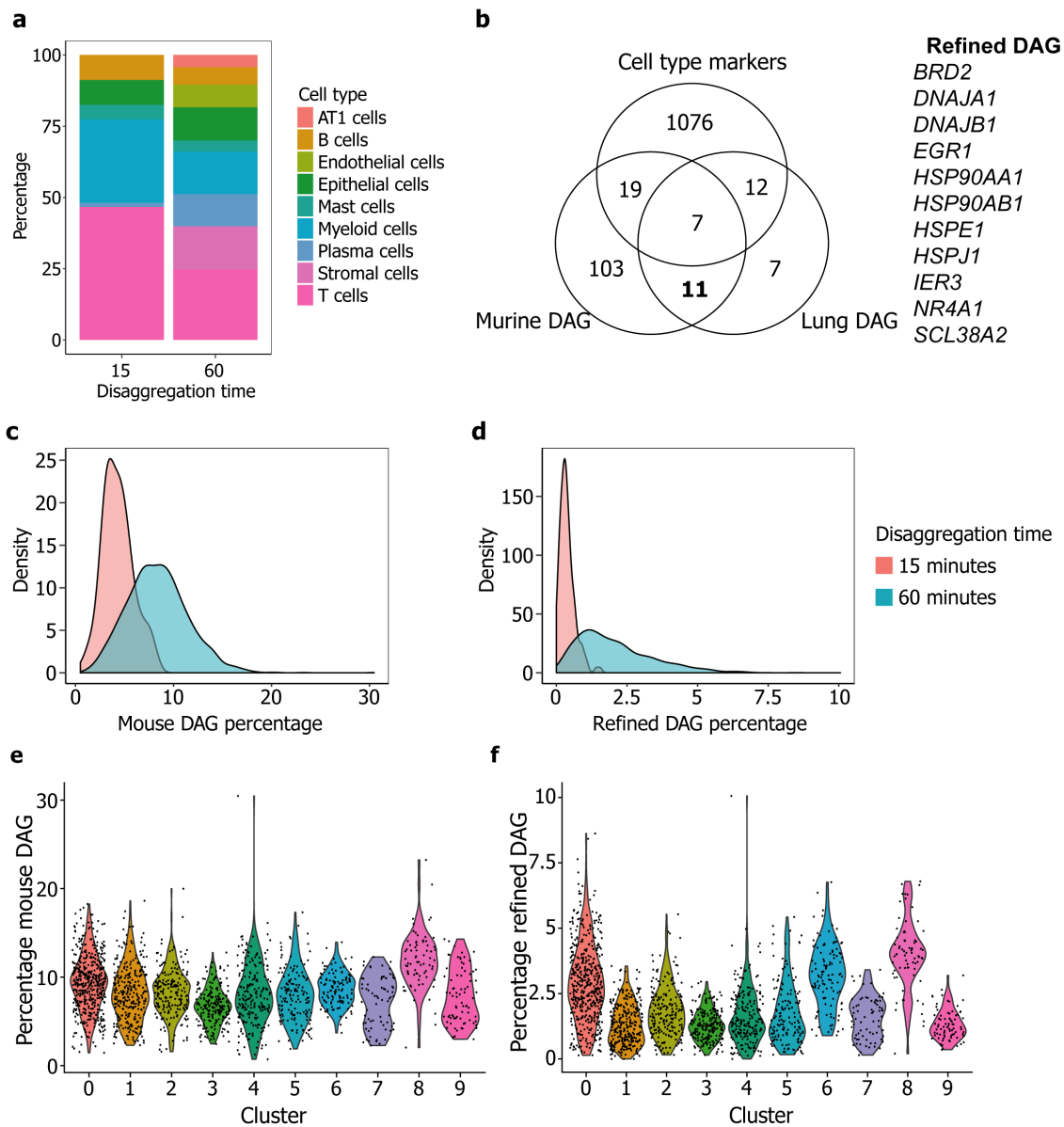


Figure 4.15 Longer tissue disaggregation enables detection of more cell types, with concomitant increases in disaggregation-associated gene expression changes

(a) Stacked barplot showing cell type fractions generated by disaggregation for 15 and 60 minutes. (b) Venn diagram showing overlap between murine disaggregation-associated signature<sup>138</sup> (Murine DAG), human disaggregation-associated signature (Lung DAG) and cell type markers. The 11 genes comprising the refined signature (highlighted in bold) are shown on the right. (c) and (d) Density plots showing percentage expression of the murine disaggregation signature (c) and the refined disaggregation signature (d), in human lung tissue disaggregated for 15 and 60 minutes. Expression of the murine and refined disaggregation-associated gene signature is not a prominent feature of any one cluster (cell type). Violin plot showing expression of (e) the murine disaggregation-associated signature and (f) the refined disaggregation-associated signature across clusters



#### 4.4 Single-cell RNA-seq identifies distinct cell lineages in primary tissue

The quality-control approach described above was then expanded to the whole dataset. Cells with more than 3000 genes (to exclude doublets) or more than 18% of genes mapping to mitochondrial reads (2.5 MAD above the median for this dataset; representing dead or dying cells) were excluded from further analysis. Following removal of empty droplets and accounting for disaggregation-associated gene expression changes, variable gene analysis was performed as described in Section 4.2. Principal component analysis was performed on the 1143 variable genes identified. Subsequent JackStraw analysis showed that principal components 1 to 38 and 43 were significant ( $p < 0.001$ ). Cluster analysis based on these significant principal components identified thirty-three clusters. The resulting tSNE plot for the whole dataset is shown in Figure 4.16.

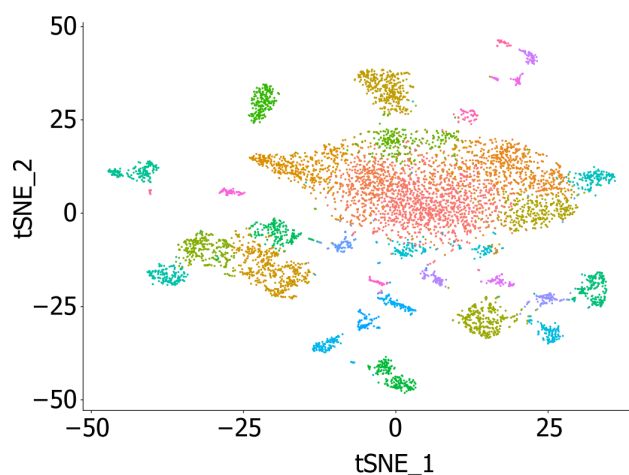


Figure 4.16 tSNE plot of quality-controlled data

Quality-controlled data from primary lung tissues identifies multiple clusters. tSNE showing data from 11 600 primary cells. Each point represents a single cell, groups of cells with similar transcriptomes are referred to as a “cluster”; clusters are distinguished by colour and labelled by cell type

Cluster marker identification was performed with the Seurat FindAllMarkers function, using genes expressed in at least 25% of cells in a cluster (further details in Section 2.8.4). Genes encoding heat shock or ribosomal proteins were excluded for more accurate gene set enrichment analysis. Using this analysis, *DCN* was identified as a stroma-specific marker in this dataset. The differentially expressed genes were then cross-referenced with a co-expression atlas (section 2.8.5) to identify cell types: transcriptomic profiles were compared to those from the LungGENS<sup>174</sup> and Immunological Genome<sup>173</sup> projects using the ToppFun tool<sup>172</sup>. The most significant ToppFun result for each cluster is given in Table 4.1. The majority of identified cell lineages were composed of cells from most patients (Figure 4.17).

Cluster	Assigned name	ToppFun enrichment statistics			
		Cell type	Bonferroni-corrected $p$ value	Genes from input	Genes in annotation
0	T cells	Gamma delta T cells	$2.647 \times 10^{-6}$	6	410
1	Alpha beta T cells	Alpha beta T cells	$2.708 \times 10^{-15}$	9	75
2	CD8+ T cells	Alpha beta T cells (CD8+)	$1.187 \times 10^{-10}$	10	359
3	Foxp3+ T cells	Alpha beta T cells (Foxp3+)	$1.374 \times 10^{-22}$	13	73
4	Myeloid cells	Myeloid cells (GN)	$6.914 \times 10^{-46}$	43	418
5	B cells	B cells	$2.122 \times 10^{-18}$	14	326
6	CD4+ T cells	Alpha beta T cells (CD4+)	$1.412 \times 10^{-14}$	9	78
7	Stromal cells	Stromal cells	$8.234 \times 10^{-70}$	65	453
8	Myeloid cells	Myeloid cells (GN)	$4.396 \times 10^{-51}$	47	418
9	T cells	Alpha beta T cells	$1.989 \times 10^{-4}$	3	57
10	Mast cells	Myeloid cells	$1.967 \times 10^{-6}$	10	401
11	Epithelial cells	Epithelial cells	$6.912 \times 10^{-22}$	29	444
12	Myeloid cells	Myeloid cells (GN)	$4.915 \times 10^{-18}$	20	418
13	Stromal cells	Stromal cells	$7.936 \times 10^{-113}$	94	438
14	Vascular endothelial cells	Endothelial cells (vascular)	$9.593 \times 10^{-83}$	69	459
15	Myeloid cells	Myeloid cells (GN)	$2.371 \times 10^{-39}$	38	418
16	NK cells	NK cells	$2.140 \times 10^{-9}$	11	402
17	Red blood cells	No results	N/A	N/A	N/A
18	Stromal cells	Stromal cells	$4.824 \times 10^{-85}$	72	438
19	Epithelial cells	Epithelial cells	$2.376 \times 10^{-43}$	38	260
20	Epithelial cells	Epithelial cells	$6.392 \times 10^{-19}$	24	444
21	Epithelial cells	Epithelial cells	$2.517 \times 10^{-18}$	25	444
22	CD8+ T cells	Alpha beta T cells (CD8+)	$1.174 \times 10^{-27}$	21	410
23	Stromal cells	Stromal cells	$5.282 \times 10^{-67}$	57	453
24	Stromal cells	Stromal cells	$2.532 \times 10^{-40}$	38	445
25	B cells	B cells	$2.067 \times 10^{-5}$	7	380
26	Stromal cells	Stromal cells	$3.522 \times 10^{-61}$	47	453
27	B cells	B cells	$2.103 \times 10^{-4}$	4	54
28	B cells	B cells	$5.664 \times 10^{-6}$	5	84
29	Epithelial cells	Myeloid cells (DC)	$7.587 \times 10^{-14}$	15	452
30	Myeloid cells	Myeloid cells (DC)	$2.120 \times 10^{-6}$	4	78
31	B cells	B cells	$5.747 \times 10^{-5}$	6	339
32	Lymphatic endothelial cells	Endothelial cells (lymphatic)	$8.423 \times 10^{-27}$	13	91

Table 4.2 Top enrichment statistic for each cluster

Quoted enrichment statistics are the most significant corresponding result from either the Immunological Genome Project (immune cell clusters) or LungGENS dataset (all other cell types)

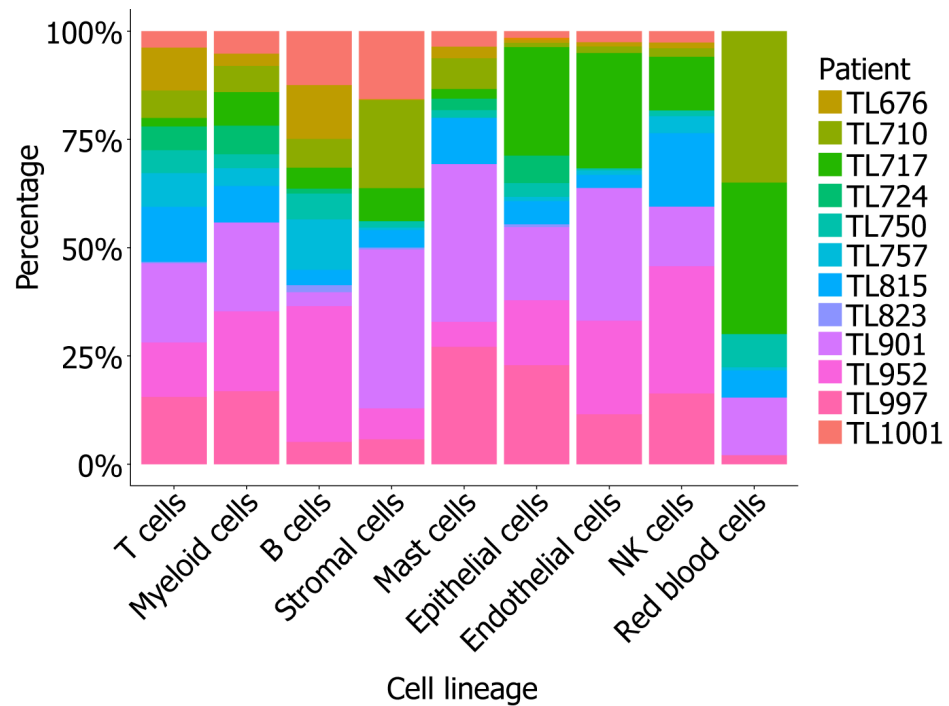


Figure 4.17 All cell lineages are composed of cells from multiple patients

Stacked barplot showing proportion contribution of each patient sample to cell lineages

The Immunological Genome project contains more information regarding differential expression profiles of immune cell subtypes, allowing more accurate identification of these populations. This database was therefore used as a reference for clusters composed of immune cells. To further confirm cell type identification, the expression of canonical cell lineage markers (in Table 4.1 above) was evaluated across the entire dataset using the Seurat FeaturePlot tool (Figure 4.18).

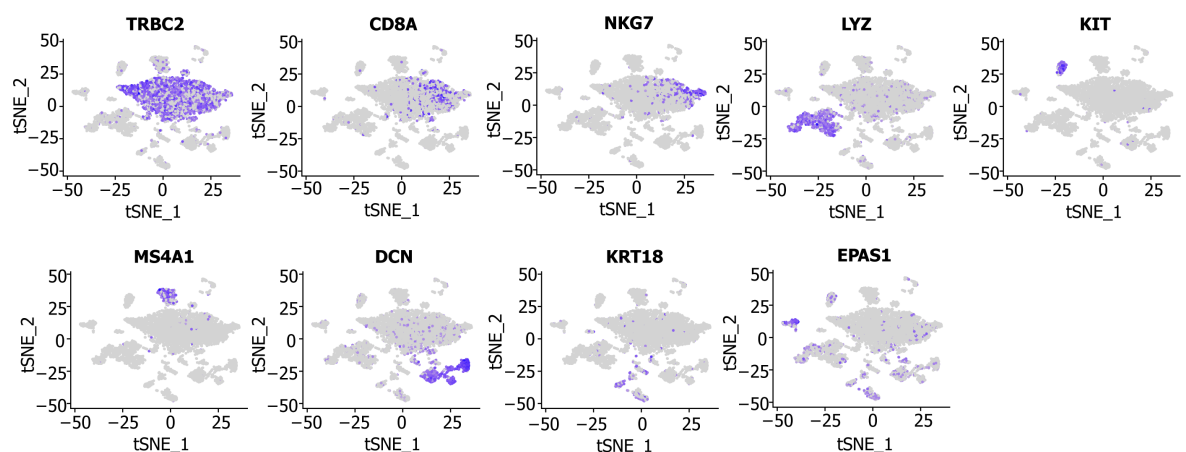


Figure 4.18 Feature plot showing expression of cell type marker genes across the dataset. Points (representing cells) coloured purple show upregulation of the specified gene

The majority of clusters showed canonical gene expression consistent with their cell type identified using the co-expression atlas. Clusters 0 and 9 (T cells), 16 (NK cells), 25, 27, 28 and 31 (B cells) and 30 (myeloid cells) had relatively high (*i.e.* less significant) enrichment statistic *p* value, but expressed markers in keeping with these cell types. In two clusters, the most significant marker genes suggested a different cell type assignment that was likely to be more accurate than that identified through co-expression atlas enrichment. Cluster 10 was identified as “myeloid cells”. However, genes associated with mast cell function (*KIT*, *CPA3*, *TPSAB1*, *HDC*) were the most significantly upregulated by this cluster and it was therefore labelled as mast cells (a myeloid cell subtype not represented in the co-expression atlas). Although the most significant enrichment result for cluster 29 was “myeloid cells”, the most common result in the top 10 were epithelial cell annotations, and multiple keratin genes (*e.g.* *KRT5* and *KRT6A*) were among the top differentially-expressed genes. The population were therefore labelled as having an epithelial origin. Finally, no results were identified for cluster 17. The top differentially-expressed genes for this cluster included genes encoding haemoglobin subunits (*HBA1*, *HBA2* and *HBB*), and this group was therefore labelled as red blood cells (a cell type also not included in the co-expression atlas).

## 4.5 Discussion

Droplet-barcoded (Drop-seq) single-cell RNA sequencing is a cost-effective method for profiling large numbers of cells. We used lung cell lines to perform initial optimisation of this platform, confirming its suitability for distinguishing between different lineages and defining quality control metrics. The same approach was then used for transcriptomic profiling of *ex vivo* cells from twelve lung tumour samples. However, cells from primary lung tissues and cell lines showed considerable differences: *ex vivo* cells have a lower library complexity and median nGene (although with a higher range) and higher numbers of non-viable cells. Application of quality control filters designed for cell lines did not give clear, well-defined cell clusters.

Other groups have previously used arbitrary, platform-dependent filtering thresholds<sup>125-127</sup>. We therefore developed a standardised approach for the identification of low-quality droplets using a random forest classifier. This method improved clustering quality compared to the use of arbitrary metrics, and gave a greater improvement in clustering quality than other similar approaches such as the DecontX and emptyDrops functions. It has previously been shown that extended disaggregation with collagenase can impact gene expression<sup>138</sup>. We refined an existing disaggregation-associated gene expression signature to assess the impact of this process on our data. Application of this refined signature to quality-controlled data did not further increase clustering quality. However, some cell types (*e.g.* T cells) appeared to be affected to a greater extent than others. Thus, the potential impact of extended enzymatic disaggregation on cellular

transcriptomes should be taken into consideration, particularly where specific clusters appear to be defined by high expression of a disaggregation-associated signature.

This optimised processing pipeline was applied to data from 12 tumour samples; the resulting analysis identified 33 distinct clusters. These were labelled using a co-expression atlas together with the top differentially expressed genes for each group, which enabled confident classification as previously-described cell types for the majority of clusters. However, our analysis has highlighted some difficulties in cell type assignment where the cluster markers showed less significant enrichment in the co-expression atlas. In some cases, this may be indicative of a cell type's absence from the reference database. For example, the top differentially-expressed genes for cluster 10 were consistent with mast cells. However, the most significant co-expression atlas result was "Myeloid cells", from the ImmGen database. This reflects the fact that mast cells were not included in the ImmGen or LungGENS datasets.

The most significant result for cluster 29 was also "Myeloid cells": more specifically, for Langerhans cells. The most common result in the top 10 results (by significance) was for epithelial cells. The top differentially-expressed genes included those encoding keratins, and those with reported expression in keratinising epithelial cells such as *SFN* (encoding stratifin)<sup>201</sup>. Langerhans cells, a specialised subgroup of dendritic cells, have documented expression of epithelial markers<sup>202</sup>. This population normally reside in the skin epidermis, but may be found in the lung under pathological conditions (Langerhans cell histiocytosis; LCH)<sup>203,204</sup>. However, cluster 29 did not show expression of the markers classical for this phenomenon (*i.e.* *CD1A*, *CD207*, *S100*<sup>204</sup>), and none of the patients associated with this study had a histological diagnosis of LCH. On balance, due to the differential upregulation of epithelial-expressed genes, this population was deemed most likely to represent epithelial cells of malignant origin (as it originated largely from tumour). However, the top co-expression atlas results for this population also included stromal and, as described above, myeloid cell results. In keeping with this, the top differentially-expressed genes for this cluster contained some genes consistent with these lineages *e.g.* *S100A6* (abundantly expressed by fibroblasts<sup>205</sup>) and *PRXL2A* (expressed by monocytes<sup>206</sup>). Of note, this population also shows differential expression of *SOX2*, a regulator of transcription in stem cells<sup>207</sup>; it is therefore also possible that this group could represent a stem cell population.



## Chapter 5      Results 3: Analysing lung fibroblast phenotypes and transdifferentiation mechanisms

### 5.1      Introduction

CAF are associated with a number of the hallmarks of malignancy, facilitating tumour invasion and metastasis, immune evasion and angiogenesis<sup>17,21,22,26-28,30</sup>. There are a paucity of data relating CAF phenotype to function: It is not yet clear whether the different effects of CAF are mediated by discrete phenotypes<sup>21</sup>. Single-cell RNA sequencing allows examination of cellular heterogeneity at a transcriptomic level: this method has identified multiple CAF populations in a number of tissues, including the lung<sup>125-127</sup>. Characterisation of this heterogeneous population will facilitate identification of pro-malignant CAF populations and allow the development of refined stromal targeting strategies.

The scRNA-Seq data generated from Drop-Seq provide a valuable resource to analyse fibroblast phenotypes in NSCLC using bioinformatic tools. Differential gene expression analysis of the identified clusters determines “marker” genes for each population. Gene set enrichment analysis (GSEA) using these genes enables preliminary characterisation of the identified populations’ potential functions and facilitates the identification of upregulated genes associated with specific biological processes and pathways<sup>176</sup>. scRNA-Seq data also provide valuable insights into the mechanisms of cellular differentiation, identifying cells that are in transition between states using trajectory analysis<sup>178</sup>. Applying these analytical approaches will reveal genes that may represent novel therapeutic targets for specific CAF sub-populations.

A dataset containing approximately 100 000 primary human lung cells analysed using the 10X platform was published during the course of this project<sup>127</sup>. Here, we combine the stromal cells from this dataset with the Drop-seq data (Section 4.4), performing unsupervised clustering of stromal cells to identify sub-populations. Examination of this amalgamated dataset increases the number of cells available for analysis, and provides the opportunity to confirm the identified stromal populations are present in multiple datasets. Using these data, we examine differential gene set enrichment to identify potential functions for the sub-populations identified. We then perform trajectory analysis to infer the molecular mechanisms that regulate differentiation between these phenotypes.

## 5.2 Identifying stromal populations in normal and malignant tissues

### 5.2.1 Identifying stromal markers

The genes identified as markers of the IMR-90 lung fibroblast cell line (Section 4.2) were assessed for suitability as *ex vivo* stromal cell markers in the Drop-seq data, using the Seurat FeaturePlot tool (Figure 5.1). This allowed visualisation of the expression of these genes across the whole primary lung dataset. Expression of some IMR-90 marker genes appears specific to stromal cells. *DCN* and *COL1A2* were specifically expressed by the majority of cells in the stromal cluster. Expression of *SERPINE1* and *SPARC* was also relatively specific to stromal cells, but were only expressed by a subset of these cells. The remaining genes, including *VIM* (encoding vimentin, widely used as marker of mesenchymal cells<sup>18</sup>) and *S100A6* (which is abundantly expressed by fibroblasts<sup>205</sup>), show diffuse expression across multiple cell types within this dataset.

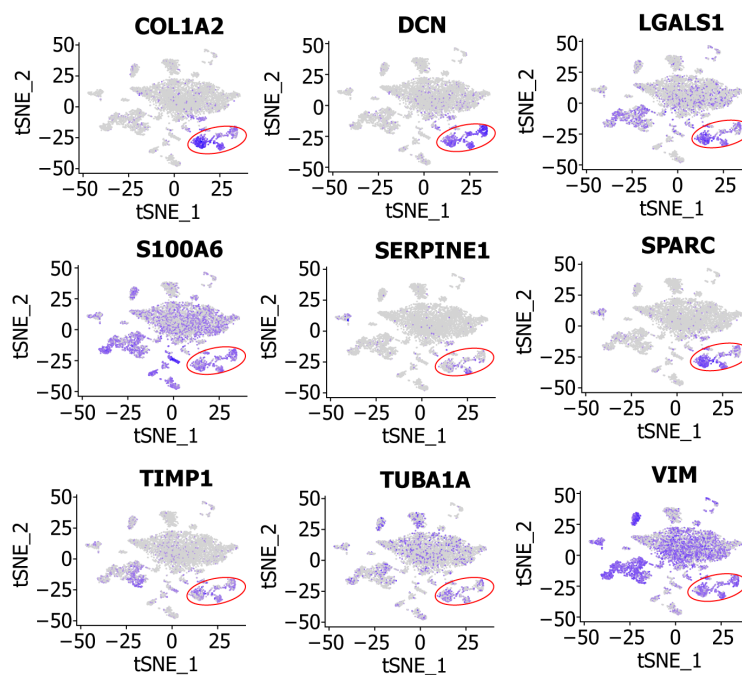


Figure 5.1 Lung fibroblast cell line markers are not specific for *ex vivo* fibroblasts

tSNE plot showing expression of genes upregulated in IMR-90 cells. Points (representing cells) coloured purple show upregulation of the specified gene. Stromal cell populations are encircled in red.

Given the relative lack of specificity of some of these genes, we devised a stromal gene signature for the Drop-seq dataset. This signature (comprising *COL1A2*, *DCN*, *COL3A1*) was created by cross-referencing the top three differentially-expressed genes for this group with the genes in the most significant ToppFun enrichment result. This signature was then used to filter and refine the stromal cell cluster for downstream analysis.



### 5.2.2 Stromal cell filtering

Cells belonging to the cluster identified as “Stromal cells” in the initial cell lineage identification (Section 4.4) were taken forward for further analysis. The raw data for these cells were extracted from the main dataset and processed as detailed in Section 2.8. A score for the stromal gene signature was calculated for each of the cells in this cluster (violin plots showing expression of the individual genes and signature across the stromal cluster are shown in Figure 5.2). To exclude cells which were either low quality or unlikely to represent true stromal cells, cells with a score of less than one were excluded. The remaining cells were taken forward for further analysis.

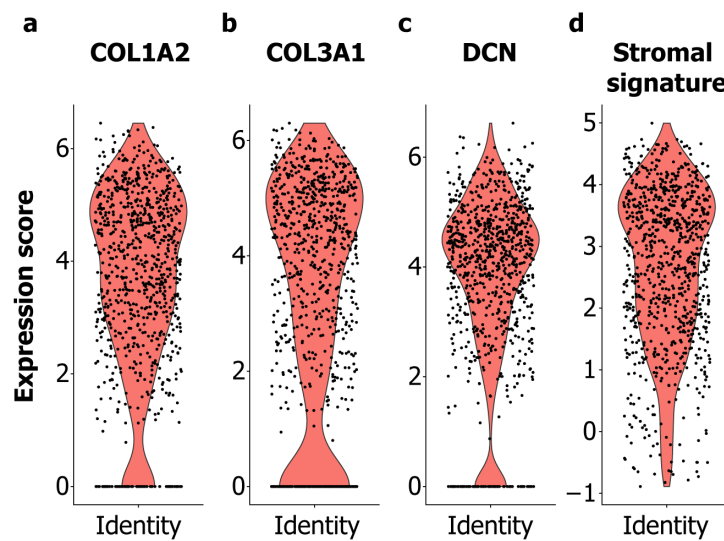


Figure 5.2 Expression of *COL1A2*, *COL3A1* and *DCN* by the “stromal cell” cluster allows exclusion of misclassified or low-quality cells

Violin plots showing the level of expression of genes across the cluster, where each cell is represented by a point: (a) *COL1A2*, (b) *COL3A1*, (c) *DCN* and (d) stromal gene signature

### 5.3 Characterising fibroblast phenotypes in normal and malignant tissues

A random forest classifier<sup>196</sup>, trained on the genes differentially expressed by the filtered Drop-seq stromal population (“TargetLung”; Section 4.4), was used to identify stromal cells in the NSCLC scRNA-seq dataset described above (“Lambrechts *et al.*”). These two stromal datasets were combined using canonical correlation analysis (Section 2.8.6), increasing the number of cells and samples used for the identification of stromal cell sub-populations and analysis of phenotypes.

5.3.1 Sub-lineage clustering

The combined stromal dataset (containing two thousand one hundred and forty-seven cells) was used for cluster identification as described previously (Section 4.4). This analysis showed the presence of nine distinct stromal subtypes within the dataset (Figure 5.3a). The majority of cells in clusters 4, 5 and 7 originated from normal samples (64%, 73% and 64%, respectively), whereas the remaining populations were composed predominately of stromal cells from tumour (Figure 5.3b). All clusters contained cells from both datasets (Figure 5.3c), although the relative contribution of each patient to each population varied across clusters (Figure 5.3d).

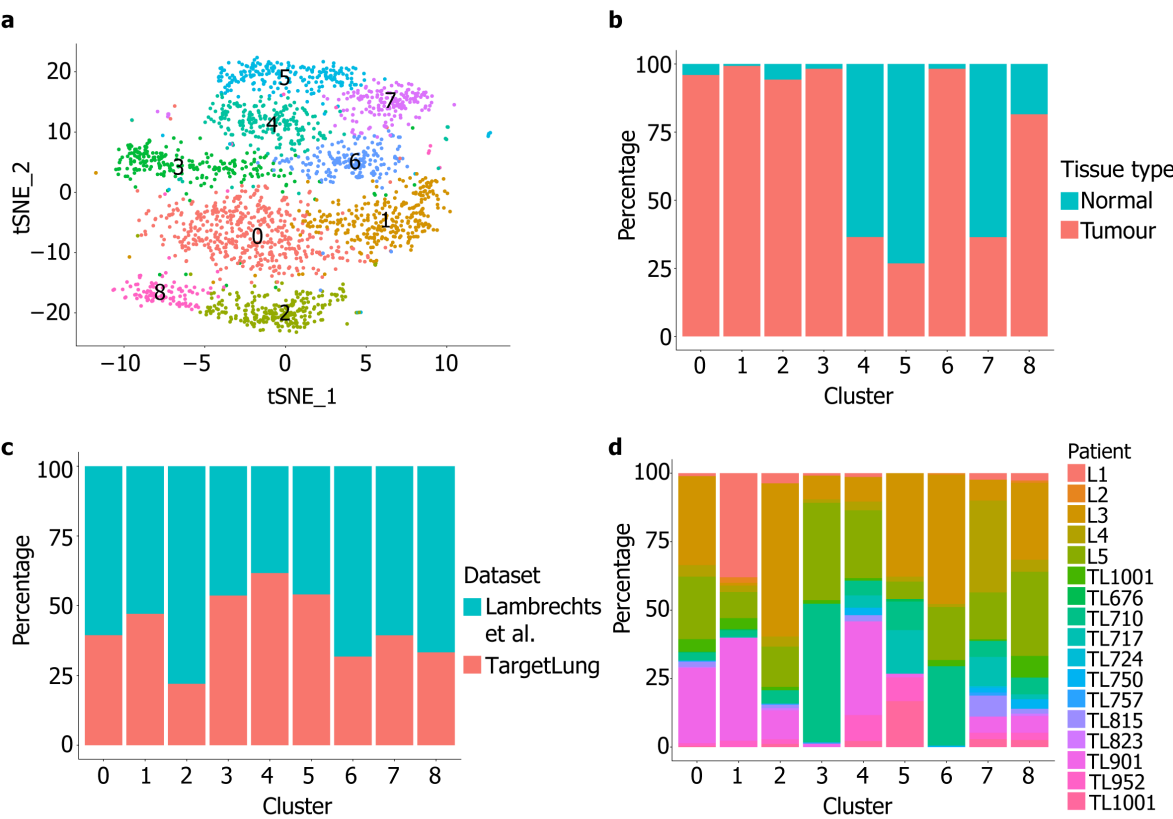


Figure 5.3 Analysis of stromal cells reveals the presence of nine distinct subtypes

(a) tSNE plot showing filtered stromal cells by assigned cluster number. Stacked bar plots showing cluster composition by (b) tissue type, (c) dataset and (d) patient of origin.

### 5.3.2 Inferring stromal phenotypes from differential gene expression and gene set enrichment analysis

To determine robust markers for each population, differential gene expression analysis (Section 2.8.4) was performed on each dataset separately. Genes upregulated in both datasets were taken as cluster markers. The top ten differentially expressed genes are shown in Table 5.1 (more extensive results are listed in Table A.2). We then used these marker genes to characterise the stromal sub-populations by performing gene set enrichment analysis for the biological processes and curated genes sets in the MSigDB database<sup>176</sup> (Section 2.8.7). The top ten results by significance for each of these analyses are given in Section A.4. A representative gene set enrichment plot, along with the genes common to the highest number of gene sets (as identified by leading edge analysis) for each cluster are shown in Table 5.2.

Cluster 0 (CAF)	Cluster 1 (CAF)	Cluster 2 (Pericytes)	Cluster 3 (CAF)	Cluster 4 (NOF)
<b>POSTN</b> FN1 COL11A1 INHBA COL1A2 COL10A1 CTHRC1 COL5A2 SULF1 SPARC	<b>MMP1</b> POSTN MMP11 MMP3 DIO2 CTSK FAP CTHRC1 INHBA TDO2	<b>RGS5</b> NDUFA4L2 HIGD1B MCAM GJA4 COX4I2 KCNJ8 ANGPT2 COL4A1 CCDC102B	<b>IGF1</b> SERPINE1 APOE UBC DUSP1 CXCL2 ZFP36 TNFAIP3 KLF6 PLIN2	<b>PI16</b> SCARA5 CFD MFAP5 PCOLCE2 IGFBP6 GSN CHRD1 DCN ADH1B
Cluster 5 (NOF)	Cluster 6 (CAF)	Cluster 7 (NOF)	Cluster 8 (VSMCs)	
<b>IL6</b> KDM6B NAMPT HAS1 ADAMTS4 MT1A ICAM1 SFTPC DDX21 CCL2	<b>IGF1</b> CXCL12 PTGDS SERPINF1 IGFBP4 RARRES2 PLA2G2A TPT1 KIAA1234L APOE	<b>NPNT</b> SCN7A LIMCH1 ADAMTS8 A2M INMT BMP5 MAMDC2 PLEKHH2 TCF21	<b>MYH11</b> PPP1R14A TAGLN ADIRF ACTA2 NTRK3 PTMA IGFBP7 TINAGL1 MCAM	

Table 5.1 Top 10 “marker” genes identified by differential gene expression analysis, ranked by decreasing adjusted *p* value

The top marker gene, which will be used to refer to each cluster, is highlighted in bold.

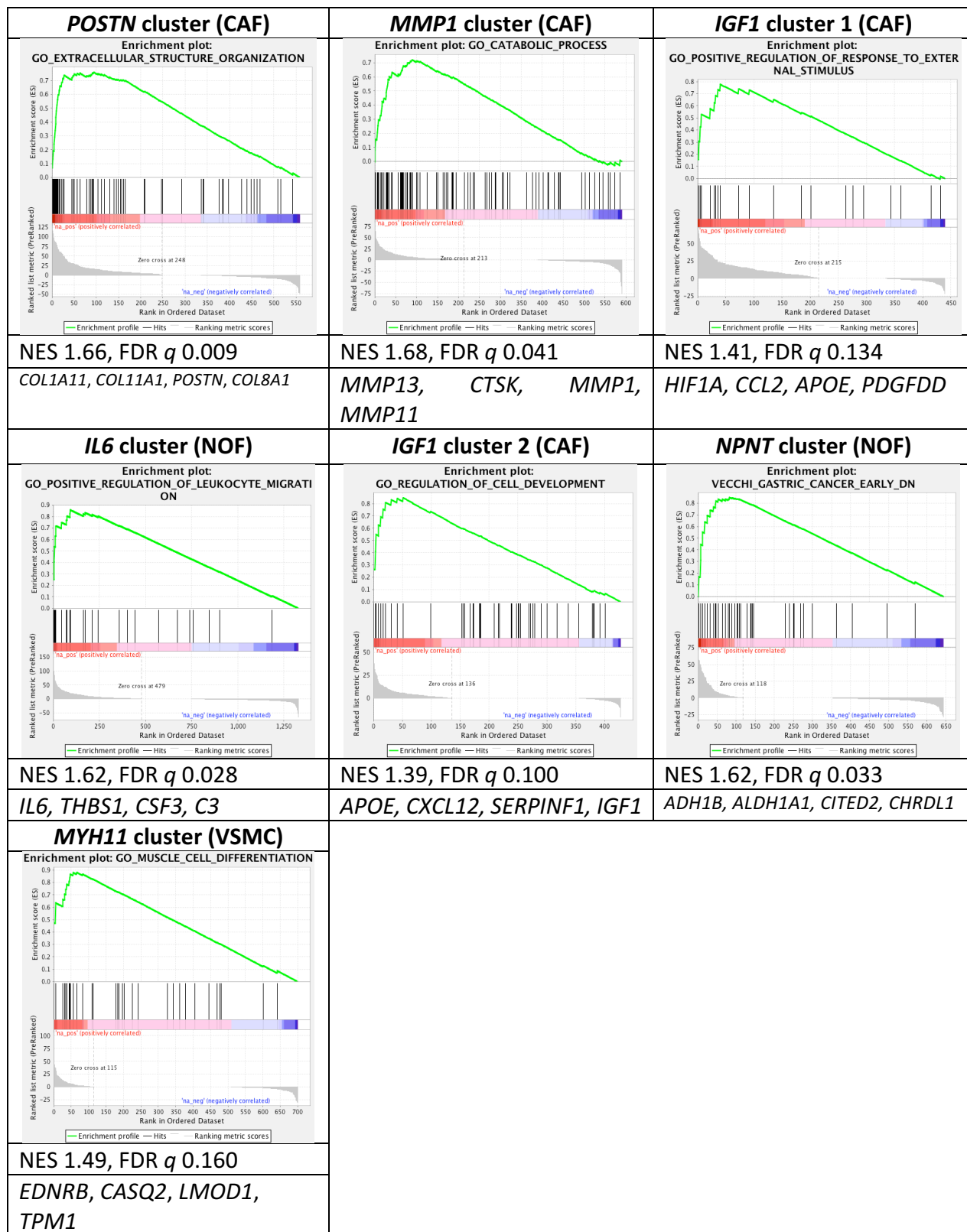


Table 5.2 Representative gene set enrichment analysis result for each stromal cluster

Clusters are identified by their highest differentially expressed gene. Associated enrichment statistics and top leading edge genes are given beneath each plot. NES, Normalised Enrichment Score. FDR  $q$ , false discovery rate

The top differentially expressed genes for the *POSTN* cluster, a CAF-predominant cluster, include multiple collagens and other extracellular matrix genes (e.g. *FN1*, *SPARC*; Table 5.1). An extracellular structure organisation signature was among the most significantly-upregulated gene sets for this population (Table A.4). Together, these features are consistent with a “fibrogenic” CAF phenotype<sup>16,76</sup>.

The most significantly-enriched biological processes for the *MMP1* cluster were associated with catabolic processes (Table 5.2 and Table A.6). This population also showed differential upregulation of a number of genes encoding protease enzymes (*MMP1*, *MMP11*, *MMP3*, *CTSK*, *FAP*<sup>39,208,209</sup>). Taking these features into account, this population was labelled a “catabolic” phenotype.

Two CAF clusters were marked by upregulation of *IGF1* (insulin-like growth factor 1), a gene with reported stromal expression and a role in mediation of crosstalk between stromal and carcinoma cells<sup>210</sup>. The first of these groups also showed upregulation of *UBC* and *DUSP1*, both of which may be induced in response to, and have protective functions against, cellular stress<sup>211,212</sup>, and was enriched for gene sets associated with regulation of the response to external stimulus (Table 5.2). Together, these results would be in keeping with a “stress response” phenotype, inducing expression of protective genes in response to external cellular stressors.

The second CAF cluster defined by *IGF1* expression also expressed other known stromal genes e.g. *CXCL12* (encoding stromal-cell derived factor 1, shown to have pro-tumorigenic and immunomodulatory functions<sup>98,213</sup>). This population was enriched for multiple gene sets associated with cell development, differentiation and growth (Table 5.2, Table A.12), consistent with the described functions of the IGF signalling axis<sup>214</sup>. *IGFBP4* (encoding an IGF binding protein<sup>215</sup>) was among the top differentially-expressed genes for this group. It is therefore possible that this group represents an “IGF signalling” phenotype, mediating the effects of the IGF signalling cascade.

Cluster 5 (composed predominantly of normal fibroblasts) was most significantly marked by expression of *IL6* and shows enrichment for multiple inflammatory response signatures (e.g. “positive regulation of leucocyte migration”), as well as targets of NF-κB (HINATA\_NFKB\_TARGETS\_FIBROBLASTS\_UP; NES 1.61, FDR  $q$  0.010). This is in keeping with a previously-described “inflammatory” fibroblast subtype<sup>29,120</sup>. The *NPNT* NOF-predominant cluster shows enrichment for gene sets associated with normal tissue or early stage malignancy (Table A.13). This population did not show enrichment for any biological processes at FDR  $q$  < 0.2.

Of the top differentially expressed genes for cluster 8, *MYH11*, *PP1R14A*, *TAGLN* and *ACTA2* are associated with smooth muscle differentiation and contractility<sup>216,217</sup>. In keeping with this, this population was enriched for muscle differentiation and contraction biological processes gene sets (Table 5.2 and Table A.14); this phenotype is therefore likely to represent vascular smooth muscle cells (VSMCs).

There were no results at FDR  $q < 0.2$  for clusters 2 and 4. Therefore, manual interrogation of the marker genes for these populations was performed. The top differentially expressed gene for cluster 2 was *RGS5*, a well-described pericyte marker<sup>218</sup>. The NOF-predominant cluster 4, marked by *PI16*, showed differential expression of both *DCN* and *CFD*. *DCN* is present in normal connective tissue<sup>219</sup>; the protein encoded by *CFD* is reported to be expressed by senescent fibroblasts<sup>220</sup>. However, this population did not show enrichment for any senescent signatures, or components of the senescence-associated secretory phenotype<sup>221</sup>. A tSNE plot labelled with the refined stromal population identities is shown in Figure 5.4.

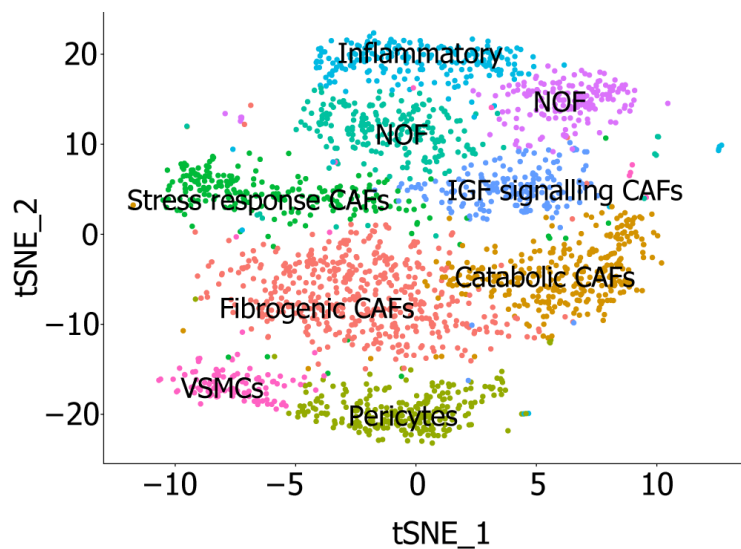


Figure 5.4 tSNE plot showing filtered stromal cells labelled by assigned function.

Each point represents a single cell, groups of cells with similar transcriptomes are referred to as a “cluster”; clusters are distinguished by colour and labelled by cell type

### 5.3.3 Histological validation of *ex vivo* fibroblast populations in fixed tissues

In the TargetLung (Drop-seq) dataset, almost all of cells in the “stress response” cluster, and the majority of the cells labelled “fibrogenic” and “catabolic” CAF, originated from single patients. Ninety-nine percent of cells in the “stress response” group were isolated from a patient with squamous cell carcinoma; a patient with adenocarcinoma contributed 86% of the total cells in the

“fibrogenic” and “catabolic” CAF clusters (Section A.8). Immunohistochemical staining was used to confirm the validity of the two of the markers identified in the previous section: *POSTN* (differentially-expressed by the “fibrogenic” and “catabolic” CAF populations, encoding periostin) and *SERPINE1* (upregulated by the “stress response” CAFs, encoding serpin E1), and to examine the differential spatial distributions of the identified CAF subtypes.

Representative sections from each patient were stained for SMA (as the most commonly-used CAF marker<sup>17,26,39</sup>), periostin and serpin E1 (Figure 5.5 and Figure 5.6). Periostin and serpin E1 staining identified spatially distinct populations. Consistent with the scRNA-Seq data, both patients had predominant staining for one marker, with little or no expression of the second. Sections from adenocarcinoma showed high levels of periostin expression in stromal regions containing SMA-positive CAFs. In contrast, sections from squamous cell carcinoma showed staining for serpine E1 with evidence of co-expression with SMA, and minimal periostin staining.

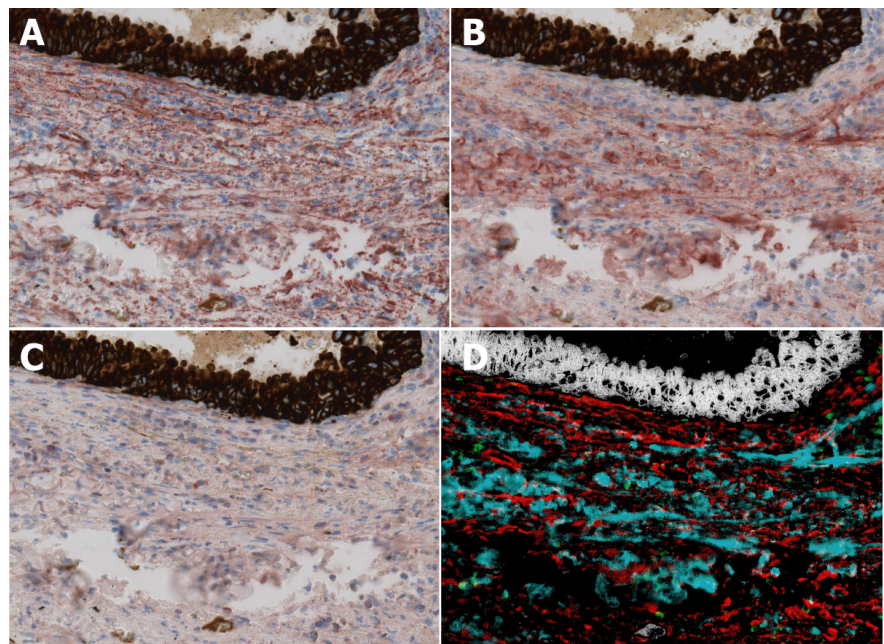


Figure 5.5 Staining for periostin and serpin E1 identifies distinct stromal populations in lung adenocarcinoma

Carcinoma cells are highlighted in brown by a pan-cytokeratin antibody. Secondary staining (in red) for: (A) SMA, (B) Periostin and (C) Serpin E1. (D): composite image showing pan-cytokeratin (white), SMA (red), periostin (cyan) and SERPINE1 (green)



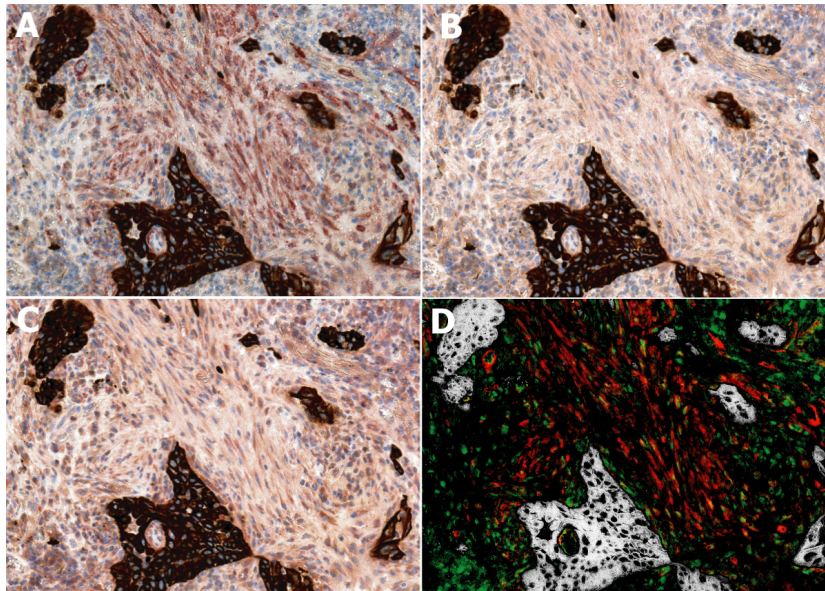


Figure 5.6 Staining for periostin and serpin E1 identifies distinct stromal populations in lung squamous cell carcinoma

Carcinoma cells are highlighted in brown by a pan-cytokeratin antibody. Secondary staining (in red) for: (A) SMA, (B) Periostin and (C) Serpin E1. (D): composite image showing pan-cytokeratin (white), SMA (red), periostin (cyan) and serpin E1 (green). Areas where SMA and Serpin E1 co-localise are yellow

## 5.4 Fibroblast trajectory analysis

Trajectory analysis was performed using the Monocle package in R as described in Section 2.8.8. The genes differentially expressed between fibroblast clusters were used as input genes for ordering. To facilitate interpretation, Pseudotime 0 was set to the State composed of the highest fraction of fibroblasts from normal tissue (State 2). The resulting cell trajectory plots are shown in Figure 5.7.



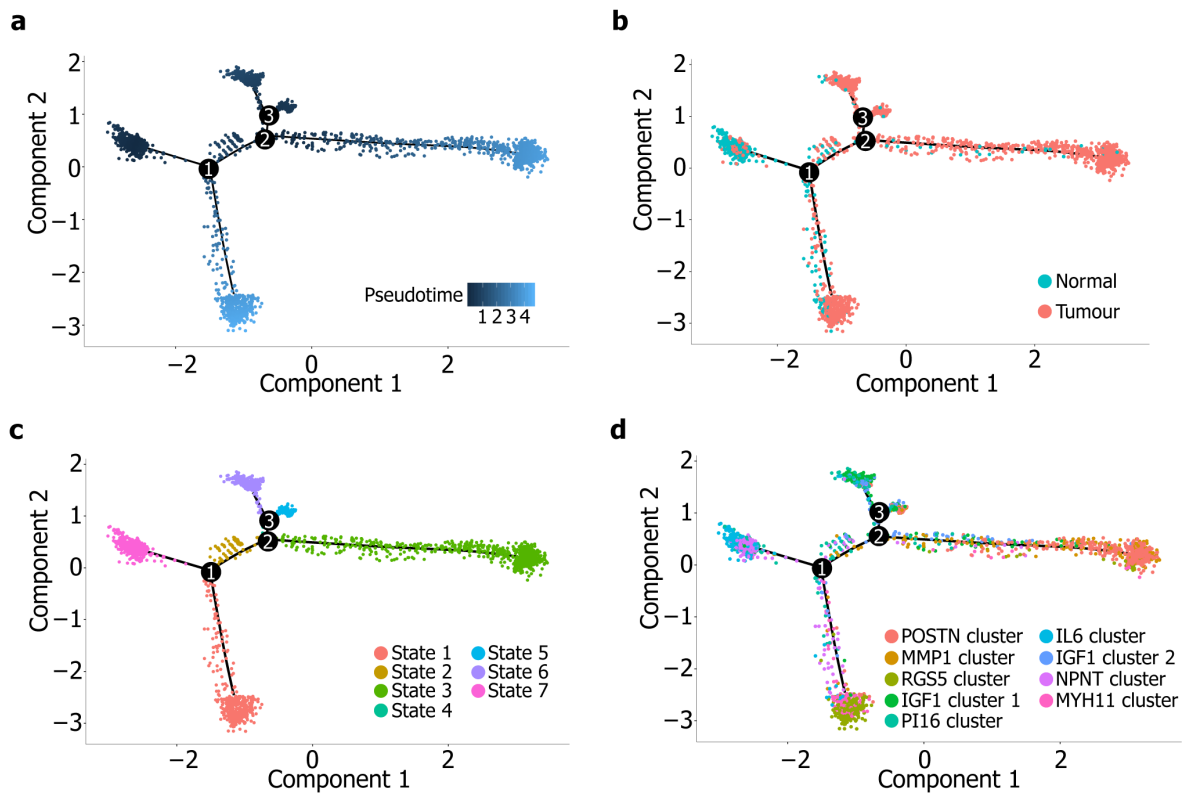


Figure 5.7 Primary stromal cells show progression from normal to cancer-associated fibroblasts over pseudotime

Trajectory plots coloured by: (a) Pseudotime, (b) Tissue type, (c) State, (d) cluster.

Each point represents a single cell

Trajectory analysis showed a transition over pseudotime from normal to both cancer-associated fibroblasts (State 3) and pericytes and VSMCs (State 1; Figure 5.7). State 3 primarily contained cells from the *POSTN* and *MMP1* clusters, whereas State 1 was composed of cells from the *RGS5* and *MYH11* clusters (Figure 5d). Both NOF and CAF populations contributed to State 2; States 5 and 6 were comprised primarily of CAF subtypes. State 7 was largely composed of cells from the *IL6* NOF cluster.

To investigate the genes differentially expressed between States, stromal cells were subsetted into groups by State. Identification of differentially-expressed genes was performed using the Monocle `differentialGeneTest` function (Section 2.8.8). GSEA was performed with these genes to identify processes or pathways associated with differentiation to the distinct trajectories observed in our analysis. The biological processes and curated gene sets from MSigDB (as in Section 5.4), and a list of gene sets identified at literature review (Section A.5) were used for GSEA. The expression of transcription factors in the Enrichr ENCODE\_TF\_ChIP-seq\_2015 and Transcription Factor PPIs libraries<sup>222</sup> was assessed using the `enrichR` tool<sup>223</sup> in R. Where trajectories did not show significant upregulation of signalling pathways, manual interrogation of

the differentially-expressed was assessed. The changes in expression in selected representative genes for each key trajectory (based on the marker genes of the constituent clusters; Table 5.1) were also examined. A summary of these changes is shown in Figure 5.8 below.

Differentiation to State 3 (containing “fibrogenic” and “catabolic” CAF) is associated with upregulation of a number of ECM genes (*e.g.* *COL1A1*, *MMP1* and *POSTN*). This trajectory shows differential expression of transcription factors which induce “fibrogenic” and “myofibroblastic” fibroblast phenotypes (*FLI1*, *MYOD* and *YAP1*)<sup>224-226</sup>, and is enriched for fibroblast gene sets from both head and neck carcinoma<sup>126</sup> and pulmonary fibrosis<sup>227</sup> (*PURAM\_CAF1* and *IPF* vs. normal, respectively; Section A.6). The head and neck signature was ascribed to one of two CAF populations in a scRNA-seq dataset (the functions of this population were not examined in this study)<sup>126</sup>. The “fibrogenic” CAF population was enriched for genes associated with response to TGF-β (Section A.4): expression of these genes also increased with differentiation to State 3 (Figure 5.8a), suggestive of a potential differentiation stimulus for this trajectory. The genes which are both differentially upregulated in State 3 (at  $p < 0.001$ ) and present in the TGF-β signature are shown in Table 5.3.

State 3 genes upregulated by TGF-β		
<i>CADM1</i>	<i>COMP</i>	<i>NTM</i>
<i>COL11A1</i>	<i>DDIT4</i>	<i>POSTN</i>
<i>COL1A1</i>	<i>ELN</i>	<i>PRSS23</i>
<i>COL4A1</i>	<i>FN1</i>	<i>SLIT3</i>
<i>COL4A2</i>	<i>INHBA</i>	<i>SULF1</i>
<i>COL5A1</i>	<i>ITGA11</i>	<i>THBS2</i>
<i>COL5A1</i>	<i>KIF26B</i>	<i>VCAN</i>
<i>COL8A2</i>	<i>MMP11</i>	

Table 5.3 Genes upregulated in State 3 and by TGF-β

The 22 genes differentially upregulated (at  $p < 0.001$ ) in State 3 which are also present in the signature for genes upregulated by TGF-β<sup>83</sup>

Ninety-six percent of cells in State 5, and 98% of cells in State 6, were derived from a mixture of CAF populations (Section A.9). These states show a common differentiation trajectory (State 4), before diverging to distinct pathways (Figure 5.7c). State 5 showed downregulation of *SIRT3*: reduced expression of this transcription factor is mediated *via* TGF-β signalling<sup>228</sup>. In keeping with this, State 5 was significantly enriched for TGF-β-associated genes (Mellone TGF-β upregulated, NES 1.40, FDR  $q$  0.146)<sup>83</sup>, and a set of genes upregulated in primary lung CAFs compared to NOFs

(Navab primary CAF vs. NOF, NES 1.40, FDR  $q$  0.099)<sup>229</sup>. In addition, this trajectory was associated with increased expression of Ras signalling genes<sup>230</sup> (GO\_REGULATION\_OF\_RAS\_PROTEIN\_SIGNAL\_TRANSDUCTION, NES 1.46, FDR  $q$  0.163). Sixty-four percent of the cells in State 6 originated from the “stress response” CAF population (Section A.9). This trajectory was enriched for the GO\_SINGLE\_ORGANISM\_CATABOLIC\_PROCESS gene set (NES 1.47, FDR  $q$  0.160). This is reflective of the origin in the cells in this trajectory: two of the top ten biological processes upregulated by the “stress response” CAF are for catabolic processes (Table A.6).

Neither State 1 (composed of pericytes and VSMCs) nor State 7 (normal fibroblasts) were significantly enriched for any of the tested gene sets. Therefore, for State 1, changes in expression of genes and gene sets upregulated by the component cell populations were assessed. VSMCs were enriched for gene sets associated with smooth muscle function (Table A.14); State 1 showed a non-significant increase in such processes (*e.g.* Figure 5.8c). This is likely to reflect enrichment of these gene sets in VSMCs but not pericytes (which did not show changes in any of the examined gene sets; Section 5.3.2), leading to no significant overall changes. The expression of *ACTA2* (a commonly-used marker of both “myofibroblastic” CAF and smooth muscle cells<sup>16</sup>) increased with differentiation to State 1 ( $p < 0.0001$ , Figure 5.8d). However, given the lack of significant changes in gene set expression, the differentiation stimulus for this trajectory remains unclear.

State 7 was composed of normal fibroblasts, mostly of the “inflammatory” phenotype (Figure 5.7d). As differentiation to this trajectory was not associated with significant upregulation of any tested gene sets, the expression of *IL6* and targets of NF- $\kappa$ B (differentially upregulated by the “inflammatory” phenotype; Table 5.1) were assessed. The cells in this trajectory showed upregulation of *IL6* ( $p < 0.0001$ ). The *IL6* phenotype cells at the distal end of this trajectory were enriched for the HINATA\_NFKB\_TARGETS\_FIBROBLASTS\_UP signature<sup>231</sup> (Figure 5.7d, Figure 5.8e and f), although the trajectory as a whole was not (NES 0.625, FDR  $q$  1.000). However, State 7 did show increased expression of *RELA*, a subunit, and a key member, of the NF- $\kappa$ B family<sup>232,233</sup>. This indicates that differentiation to this state is, some extent, but not fully, driven by NF- $\kappa$ B: it is possible that NF- $\kappa$ B signalling is particularly important in regulating progression from the *PI16* to the *IL6* phenotype at the distal end of this trajectory.

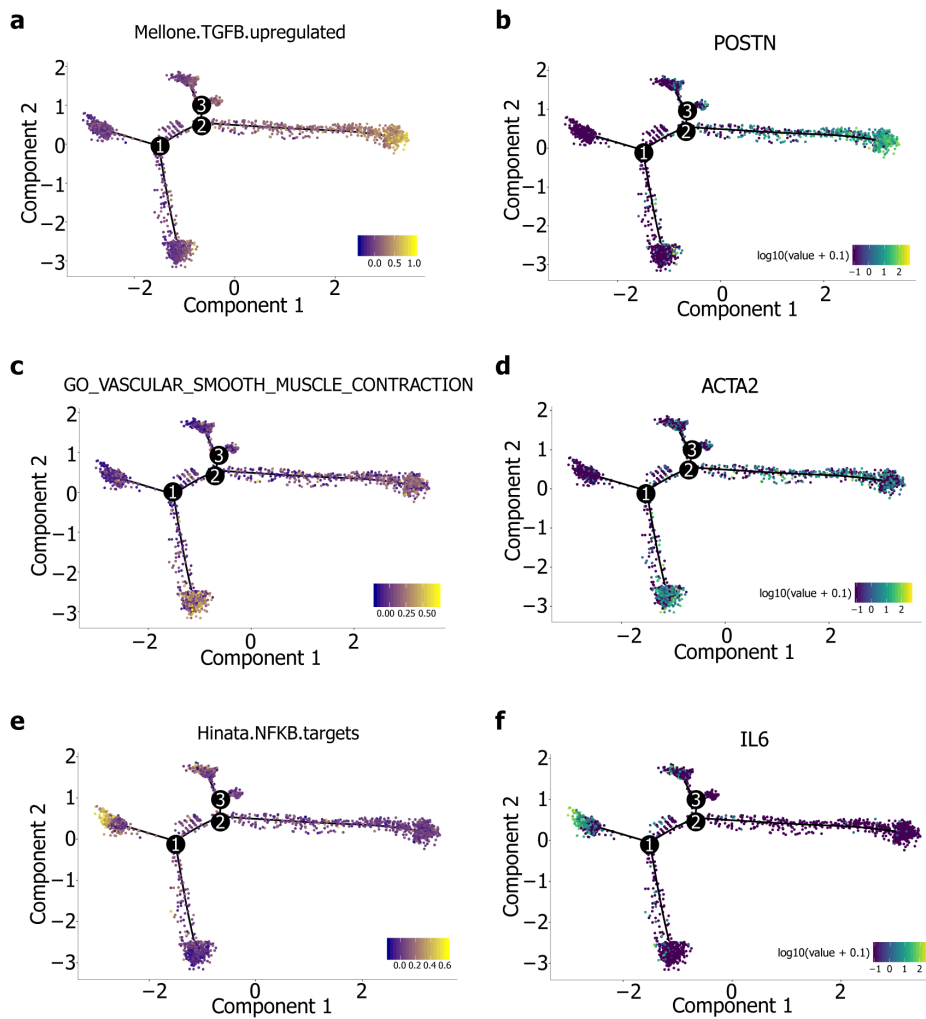


Figure 5.8 Primary fibroblasts show trajectory-dependent upregulation of marker genes and gene sets

Trajectory plots showing changes across trajectories in (a) genes upregulated by TGF- $\beta^{83}$ , (b) *POSTN*, (c) vascular smooth muscle contraction<sup>176</sup>, (d) *ACTA2*, (e) the targets of NF- $\kappa$ B<sup>231</sup> and (f) *IL6*

## 5.5 Discussion

To minimise the impact of misclassified cells on downstream analysis, we created a stromal gene signature. Cross-referencing the genes showing the highest differential expression by this cluster with those present in the most significant co-expression atlas result gave three genes: *COL1A2*, *COL3A1* and *DCN*. Cells belonging to the “Stromal cell” cluster in the larger dataset with a low expression of these three genes in combination were not taken forward. Such cell type-specific signatures may also prove useful for refining cell types with a lower nGene, such as plasma cells and T cells.

Clustering using the combined stromal populations identified in the Drop-Seq and Lambrechts *et al.*<sup>127</sup> datasets revealed 9 stromal populations: 6 originating predominantly from tumour samples (4 CAF, 1 pericyte and 1 VSMC population) and 3 largely derived from normal tissues. Of the CAF populations, one showed differential expression of *POSTN* (encoding periostin) and multiple fibrillar collagen genes. Periostin and fibrillar collagens synthesised by CAF are implicated in promoting carcinoma cell invasion, through activation of the PI3kinase-Akt pathway and enhanced tissue stiffness, respectively<sup>78,234</sup>. Accordingly, this cluster was enriched for signatures associated with cancer invasiveness<sup>235,236</sup>. A second (*MMP1*) CAF phenotype showed upregulation of multiple matrix metalloprotease genes and catabolic process gene sets<sup>176</sup>. In the context of malignancy, matrix metalloproteases enhance tumour cell migration and invasion through, for example, degradation of the ECM and production of mitogenic factors<sup>237</sup>. In keeping with this, the *MMP1* population also showed enrichment of gene sets associated with cancer invasiveness<sup>235,236</sup>. The *MMP1* population may thus represent a “catabolic” phenotype, facilitating the invasion and metastasis of malignant cells. Both generation and remodelling of fibrous tumour stroma have previously been described as functions of the “myofibroblastic” CAF phenotype<sup>17,18,77</sup>. However, in this dataset, it appears that these functions may be mediated through different populations: the “fibrogenic” and “catabolic” phenotypes could represent subgroups of what has previously been described as one “myofibroblastic” population.

The remaining two CAF phenotypes both showed differential upregulation of *IGF1*. The first *IGF1* population was labelled “stress response” CAF, due to its expression of genes and gene sets associated with regulating the response to external stressors<sup>211,212,230</sup>. This population also showed upregulation of genes with previously-described stromal expression, such as *SERPINE1*<sup>200</sup>. Stromal expression of the protein encoded by this gene (serpin E1) is associated with tumour cell proliferation, invasion and metastasis<sup>238,239</sup>. Serpin E1 and periostin show differential spatial distribution; the expression of these proteins appears to differ between adenocarcinoma and squamous cell carcinoma, indicating that there may be NSCLC subtype-dependent accumulation of the distinct CAF phenotypes. The second *IGF1* population was enriched for multiple gene sets associated with cell development, differentiation and growth, consistent with the functions of the IGF signalling axis<sup>214</sup>. Together, these findings may be indicative of an “IGF signalling” phenotype: IGF1 has previously been shown to mediate tumour-stroma crosstalk in a pancreatic carcinoma model, enhancing cellular proliferation and mediating resistance to apoptosis<sup>210</sup>.

Trajectory analysis revealed a common differentiation pathway for the “fibrogenic” and “catabolic” CAF populations, associated with upregulation of TGF- $\beta$  target genes<sup>83</sup>. This cytokine has typically been seen as the main driver of the “myofibroblastic” CAF phenotype<sup>77</sup> (with which the “fibrogenic” and “catabolic” CAF populations show functional overlap<sup>17,18,77</sup>), and could

represent a potential differentiation mechanism for this trajectory. However, these genes account for only 3.4% (22/645) of the total upregulated in this trajectory: differentiation to this State may also be determined by other pathways. This State shows non-significant upregulation of multiple senescence-associated gene sets (e.g. FRIDMAN SENESENCE, NES 1.26, FDR  $q$  0.352; Genetically unstable vs. normal OSCC, NES 1.28, FDR  $q$  0.355<sup>240,241</sup>); it is possible that differentiation to this trajectory is governed by both TGF- $\beta$  signalling and induction of cellular senescence.

Trajectory states 5 and 6 were composed largely of mixed CAF phenotypes. These states shared a common differentiation trajectory (State 4) before diverging to distinct pathways (Figure 5.7c). Both state 5 and 6 showed upregulation of TGF- $\beta$  target genes, although the increase in State 6 was non-significant; State 5 was additionally enriched for a Ras signalling signature. It is possible that differentiation to all CAF phenotypes identified in this dataset may be driven by TGF- $\beta$ : TGF- $\beta$  signalling and senescence may result in differentiation to State 3, combined TGF- $\beta$  and Ras signalling to State 5, and TGF- $\beta$  in isolation to State 6. There are currently a scarcity of data examining the combined effects of Ras and TGF- $\beta$  in fibroblasts: this is a potential avenue for future work.

The remaining populations derived largely from tumour were identified as pericytes and vascular smooth muscle cells. The pericyte group did not show significant enrichment for any of the examined gene sets, and were labelled based on their differential expression of *RGS5* (a marker of this cell type<sup>218</sup>). The vascular smooth muscle cell population showed upregulation of both genes and gene sets associated with smooth muscle cell differentiation and contractility<sup>216,217</sup>. The common differentiation trajectory shared by these populations (Figure 5.7d) showed non-significant enrichment for gene sets associated with smooth muscle function, likely reflecting upregulation of these processes in VSMCs, but not pericytes. Fibroblasts are known to form smooth muscle-like cells following vascular injury<sup>242</sup>. However, although pericytes have the potential to differentiate to CAF<sup>26</sup>, evidence for the reverse phenomenon is sparse. It is therefore possible that the pericyte/VSMC trajectory shows no apparent differentiation mechanism because the pericytes in this dataset do not arise from the fibroblast population.

Of the populations predominantly originating from normal tissue, the *IL6* phenotype showed differential gene expression and gene set enrichment consistent with the previously-described “inflammatory” fibroblast phenotype<sup>29,120</sup>. These cells were present almost exclusively (96%) in State 7 (Figure 5.7d). Although the *IL6* population were significantly enriched for NF- $\kappa$ B, State 7 as a whole was not. This indicates that differentiation to this trajectory must, in part, be mediated through other mechanisms. These stimuli are not currently clear: this state shows non-significant

enrichment for a variety of pathways and processes including TGF- $\beta$ , inflammatory and angiogenic signatures.

The second NOF subtype was marked by expression of *NPNT*, a marker of a “matrix fibroblast” population in a murine pulmonary fibrosis single-cell atlas<sup>243</sup>. The *PI16* subtype showed differential upregulation of *CFD*. Data describing the expression and function of *CFD* in fibroblasts are sparse, although other elements of the complement cascade are differentially expressed by CAF in a melanoma<sup>125</sup>. One study has described upregulation of *CFD* in senescent skin fibroblasts, where it induces expression of *MMP1* (which degrades type I collagen)<sup>220</sup>. This population also showed expression of other genes involved in both synthesis and degradation of ECM proteins, including *DCN*, *MFAP5* and *PCOLCE2*<sup>244-246</sup>; it is therefore possible that this phenotype is concerned with the maintenance of the tumour stroma.

Here, we describe nine distinct stromal cell populations in a combined NSCLC scRNA-seq dataset. Of the four CAF phenotypes, the “fibrogenic” and “catabolic” populations share expression and enrichment of some genes and gene sets with the “myofibroblastic” CAF phenotype. Other clusters also show features consistent with previously-described phenotypes and functions: for example, the *IL6* cluster is enriched for the signalling pathway known to drive differentiation to the “inflammatory” fibroblast phenotype. The “fibrogenic” and “catabolic” CAF appear to share a common differentiation trajectory, as do the pericyte and VSMC populations. The remaining clusters show distinct differentiation pathways according to whether they originate from normal or tumour tissue. Identification of putative differentiation stimuli will facilitate recreation of the *ex vivo* phenotypes for *in vitro* functional assays.





## Chapter 6 Results 4: Recreating *ex vivo* fibroblast phenotypes

### 6.1 Introduction

Fibroblasts are most commonly isolated on tissue culture plastic using serum to stimulate outgrowth from tissue, followed by expansion *in vitro* prior to analysis<sup>180,181</sup>; this is a well-described and reliable method of generating fibroblast cultures<sup>47,182,183</sup>. However, alteration of culture substrate has been shown to skew fibroblast phenotypes<sup>120</sup>. Whether *ex vivo* phenotypes are maintained in culture, and whether the functional differences described *in vitro* hold true for *in vivo* phenotypes has yet to be determined<sup>134,183</sup>.

Gene set enrichment analysis (GSEA) of the genes differentially expressed across differentiation pathways allows identification of potential differentiation stimuli for the separate trajectories described in the previous chapter<sup>176,178</sup>. Using gene expression profiling by RT-PCR and transcriptomic analysis with Drop-seq, it is possible to assess the extent to which manipulation of culture conditions can recreate *ex vivo* fibroblast phenotypes: adequate recapitulation of *ex vivo* fibroblast phenotypes will be necessary for *in vitro* functional characterisation. Here, we examine the impact of *in vitro* culture on fibroblast transcriptomes and function, assessing how culture conditions can be used to skew fibroblast phenotypes.

### 6.2 Optimising *in vitro* culture conditions for analysing fibroblast subtypes

#### 6.2.1 *In vitro* culture alters transcriptomes

Recent studies have shown that the gene expression profiles of cancer-associated fibroblasts (CAFs) from both head and neck squamous carcinoma (HNSCC) and melanoma are significantly altered by culture *in vitro*<sup>125,126</sup>. To determine whether this holds true for the lung, the gene expression profiles of primary fibroblasts isolated from tumour and non-involved tissue were compared with those of *ex vivo* fibroblasts.

Normal and cancer-associated fibroblasts were isolated from tissue samples and expanded in culture for one passage. These cells were then cultured in “low-serum” (1%; Section 2.1) DMEM and analysed using single-cell RNA sequencing. These cultured cells (“*in vitro*”) were compared

with fibroblasts from the main dataset (“*ex vivo*”). *In vitro* samples showed a similar number of genes *per* cell as *ex vivo* samples, thus initial filtering parameters remained the same. Principle components 1:15 were identified as significant, based on 3970 variable genes.

We found that fibroblasts isolated from tumour or non-involved tissues cluster separately from *ex vivo* cells following culture *in vitro* for one passage (Figure 6.1a). Gene expression by these cells more closely resembles IMR-90 cells than *ex vivo* fibroblasts (Figure 6.1b), showing that culture on plastic surfaces causes significant transcriptomic changes in primary lung fibroblasts. Furthermore, the genes differentially expressed before and after culture are consistent with those described previously in CAF from similar experiments in melanoma and HNSCC<sup>125,126</sup> (Figure 6.1c). These, together with the list of the top ten differentially expressed genes in our dataset and *DCN* (identified as fibroblast-specific in our data), were used to create a panel to allow analysis of fibroblast phenotypes by RT-PCR (Section 2.3).

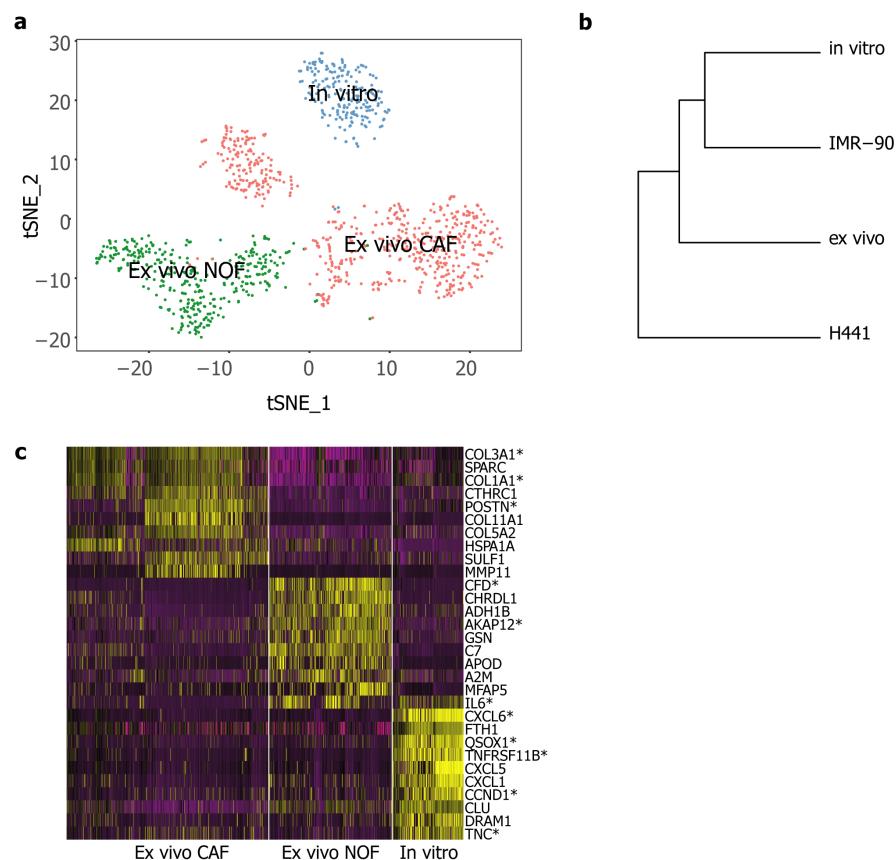


Figure 6.1 Cells analysed immediately following isolation “*ex vivo*” and following culture on plastic “*in vitro*” show differences in gene expression

(a) tSNE plot showing clustering of *ex vivo* and *in vitro* cells. (b) Dendrogram for lung adenocarcinoma (H441) and fibroblast (IMR-90) cell lines and primary lung cells. (c) Heatmap of the top 10 genes differentially expressed by each group. Genes marked with an asterisk were included in the RT-PCR panel

### 6.2.2 2D *in vitro* culture conditions impact fibroblast proliferation

Given the variations observed in gene expression between *ex vivo* and *in vitro* fibroblasts the effect of *in vitro* culture on fibroblast phenotypes was investigated. Primary lung fibroblasts were seeded at 50 000/ml to 6-well tissue culture plates. Plates were either coated with substrates with the potential to alter cell adherence (either 0.1% gelatin or Matrigel® Matrix) or left uncoated. These cells were harvested at 3 and 7 days for assessment of proliferation and metabolic activity (Section 2.4). Normal primary fibroblasts showed similar fold-change in growth relative to plastic when cultured on both gelatin and Matrigel (Figure 6.2a). CAF cultured on gelatin or Matrigel showed a significant increase in cell number when compared with plastic at 3 days. This difference persisted at 7 days, but was no longer significant, likely because the cells had reached confluence prior to this timepoint (Figure 6.2b).

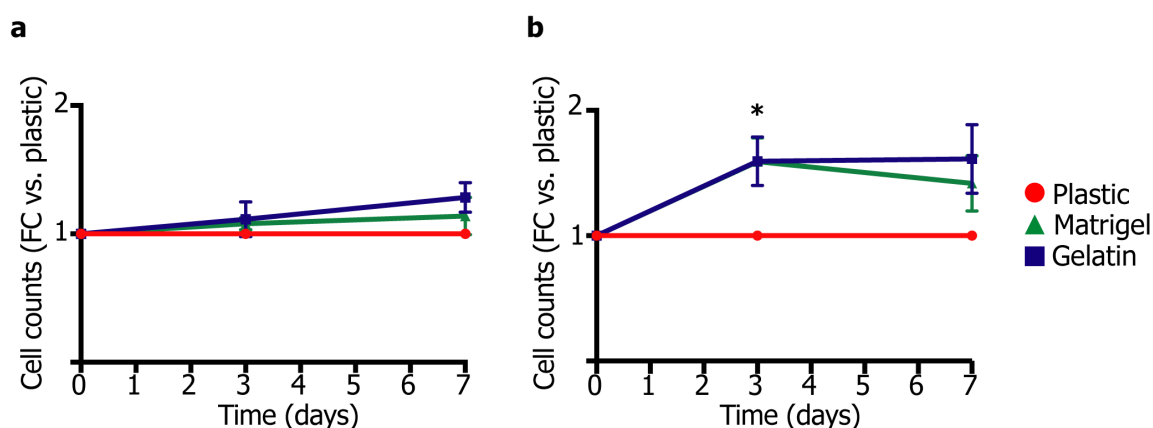


Figure 6.2 Alteration of tissue culture surface impacts fibroblast proliferation

Cell counts in primary (a) normal and (b) cancer-associated fibroblasts when cultured on plastic, gelatin or Matrigel surfaces (n = 5). Counts expressed as fold change (FC) relative to plastic. \* $p < 0.05$ , Welch's  $t$  test

### 6.2.3 3D *in vitro* culture significantly alters gene expression and can be used to skew fibroblast phenotypes towards previously described sub-types

In order to determine whether *ex vivo* fibroblast phenotypes can be recapitulated by manipulation of *in vitro* conditions, primary cells were cultured either on plastic, Matrigel® or in collagen/Matrigel® Matrix 3D gels. Cells were harvested after five days and any changes in gene expression measured using real-time PCR (data for 3D gels in Figure 6.3). Culturing fibroblasts on Matrigel-coated compared to uncoated tissue culture plastic caused similar changes to gene expression as uncoated plastic relative to culture in 3D, albeit of a lower magnitude (Section A.10).

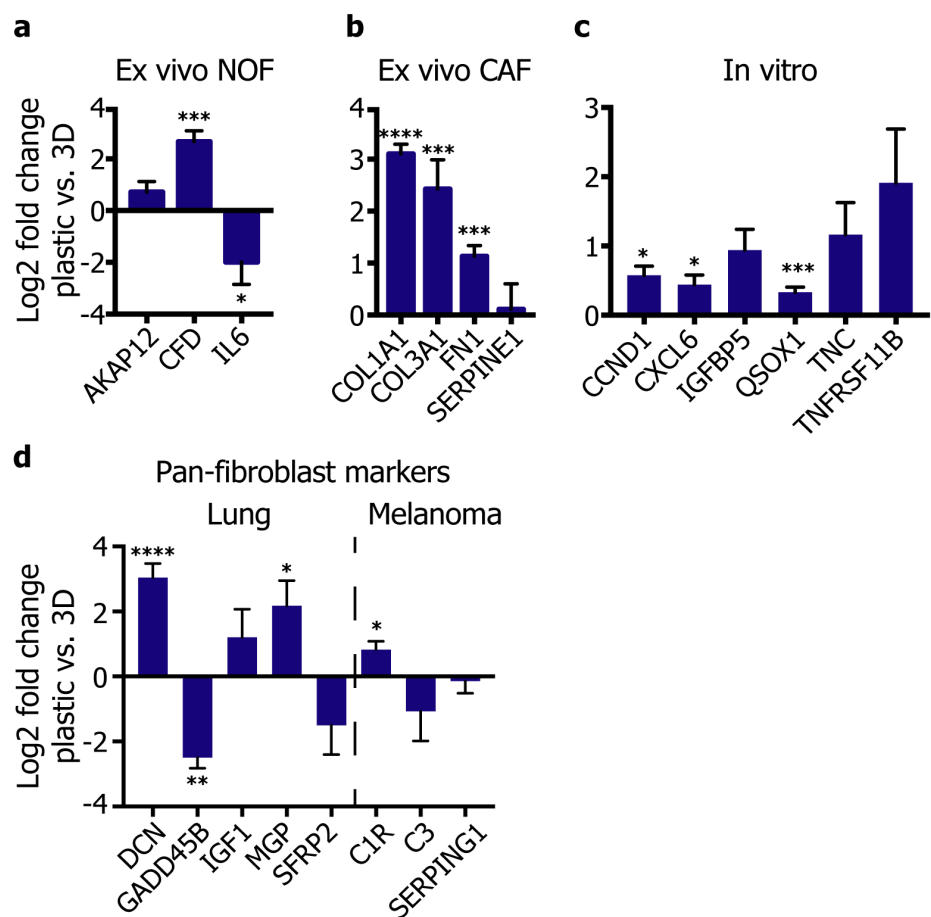


Figure 6.3 Culture on plastic relative to 3D leads to upregulation of genes differentially expressed by *ex vivo* CAF and *in vitro* fibroblasts

Bar charts showing fold change in expression for genes differentially expressed by: (a) *ex vivo* NOF, (b) *ex vivo* CAF, (c) *in vitro* fibroblasts, (d) all *ex vivo* fibroblast populations in both our dataset and melanoma<sup>125</sup>. Values are expressed as log2(fold change plastic:3D). \* $p < 0.05$ , \*\*\* $p < 0.001$ , \*\*\*\* $p < 0.0001$ , Welch's  $t$  test

The effect of culture conditions on the expression of *ex vivo* NOF markers was variable: although *CFD* was upregulated in fibroblasts grown on plastic, expression of *IL6* (a marker of “inflammatory” fibroblasts<sup>120</sup>) was increased by 3D culture (Figure 6.3a). Perhaps unsurprisingly, *in vitro* fibroblast markers were upregulated by culture on plastic, although this increase was statistically significant for *CCND1*, *CXCL6* and *QSOX1* only (Figure 6.3c). Expression of the *ex vivo* CAF markers *COL1A1* and *COL3A1* (markers of myofibroblast transdifferentiation<sup>16,37</sup>) was significantly increased in cells cultured on plastic (Figure 6.3b). Alteration of culture conditions also had a variable impact on genes expressed across *ex vivo* fibroblast populations (“pan-fibroblast markers”; Figure 6.3d): whereas *DCN*, *MGP* and *C1R* were upregulated by growth on plastic, expression of *GADD45B* was increased by culture in 3D. The changes in some of these key genes (those used as markers for *ex vivo* and *in vitro* phenotypes) are summarised in Table 6.1 below.

Gene	Average log(fold change)	Adjusted <i>p</i> value
<i>COL1A1</i>	1.597949	$2.599801 \times 10^{-15}$
<i>COL3A1</i>	1.811002	$3.936564 \times 10^{-20}$
<i>IL6</i>	-1.513206	$1.884241 \times 10^{-13}$
<i>CFD</i>	-2.411358	$2.647318 \times 10^{-13}$
<i>MGP</i>	-2.91536	$2.201023 \times 10^{-25}$

Table 6.1 Summary of key genes differentially expressed between fibroblasts cultured on plastic and *ex vivo*

Genes with a positive average log(fold change) are upregulated in cells cultured on plastic; those with a negative value are increased by culture in 3D

None of the above genes showed differential expression when comparing cells by tissue of origin (normal or tumour) or by patient rather than by culture condition (Figure 6.4). The changes in expression of *COL1A1* and *IL6* (selected as representative markers for *ex vivo* NOF and CAF populations) were found to be maintained at the protein level (Figure 6.5).

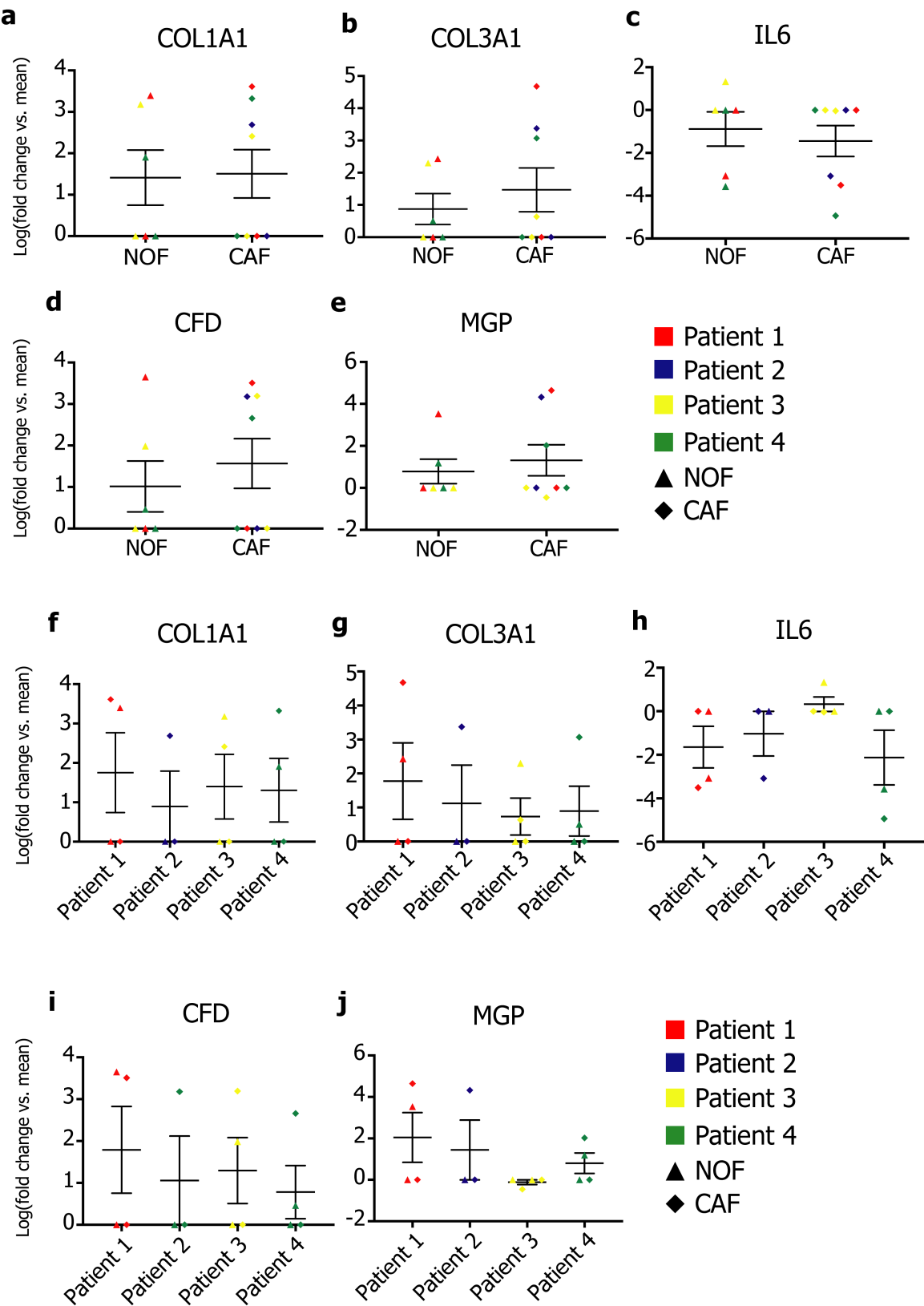


Figure 6.4 Changes in gene expression across culture substrates are not cell type- or patient-dependent

Dot plots showing changes in gene expression in cells cultured on plastic and in 3D by cell type (a-e) and across patients (f-j). Gene expression levels are expressed at the log2 of the fold change relative to the mean expression across all samples (n = 4)

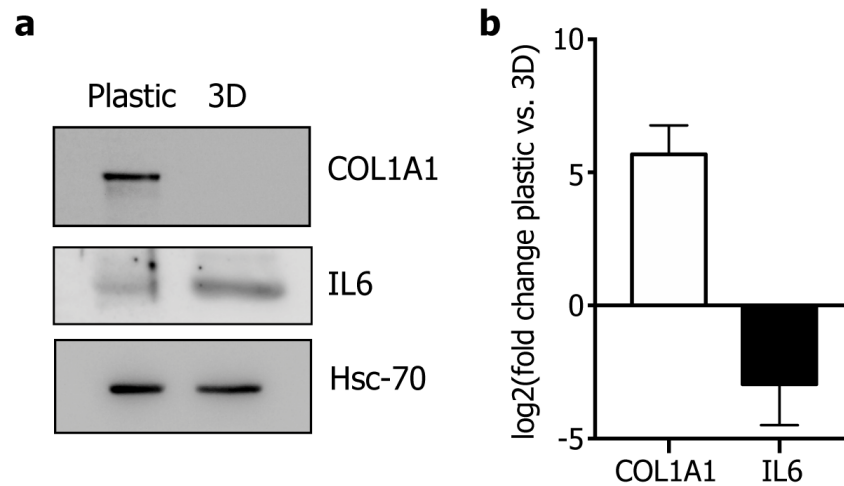


Figure 6.5 The observed changes in gene expression are maintained at a protein level

(a) Western blot for expression of COL1A1, Hsc-70 (loading control) and IL6. (b) Bar chart showing log(fold change) in measured optical density for plastic relative to 3D for COL1A1 and IL6. Graphs show mean values with the standard deviation for technical replicates (n=3)

### 6.3 Recapitulating *ex vivo* fibroblast phenotypes

Having shown that manipulating *in vitro* culture conditions can alter the expression of genes associated with known fibroblast subtypes, we sought to define the culture conditions required to recapitulate the phenotypes of the different *ex vivo* CAF differentiation trajectories. Trajectory analysis (Section 2.8.8, Figure 6.6) identified pathway enrichment associated with the distinct trajectories. Differentiation to State 3 (*POSTN* and *MMP1* CAF phenotypes) was characterised by enrichment of genes upregulated by TGF- $\beta$  treatment (a well-described driver of differentiation to the “myofibroblastic” phenotype<sup>77,247,248</sup>). Differentiation to State 7 (*IL6* fibroblasts) was characterised by enrichment of inflammatory processes and the targets of NF- $\kappa$ B (a driver of the “inflammatory” fibroblast phenotype<sup>29</sup>).

Based on this analysis we hypothesised that the differentiation of primary lung fibroblasts could be induced using recombinant protein stimulation *in vitro*, using TGF- $\beta$  to induce differentiation to State 3 and IL-1 $\beta$  (an upstream activator of NF- $\kappa$ B signalling<sup>249</sup>) to induce differentiation to State 7. It is well-documented that in addition to growth factor or cytokine stimulation, culture surface can have a significant impact on fibroblast phenotypes: increased substrate stiffness generates an “activated myofibroblast” phenotype<sup>250</sup>. This effect may go some way to explaining the differences in gene expression between *ex vivo* and *in vitro* fibroblasts (Figure 6.1). We therefore

compared the impact of TGF- $\beta$  or IL-1 $\beta$  treatment stimulation on fibroblasts cultures on tissue culture plastic (TCP) and low elastic modulus culture plates (2 kPa, similar to that of lung tissue<sup>251</sup>). As mentioned previously, fibroblasts are most commonly cultured on plastic substrates: this condition (TCP) was included to facilitate comparisons with previously-published findings. Cells cultured on low elastic modulus plates without additional treatment were included as a control. Plates were coated with either collagen (100  $\mu$ g/ml, Corning) or Matrigel, as previously.

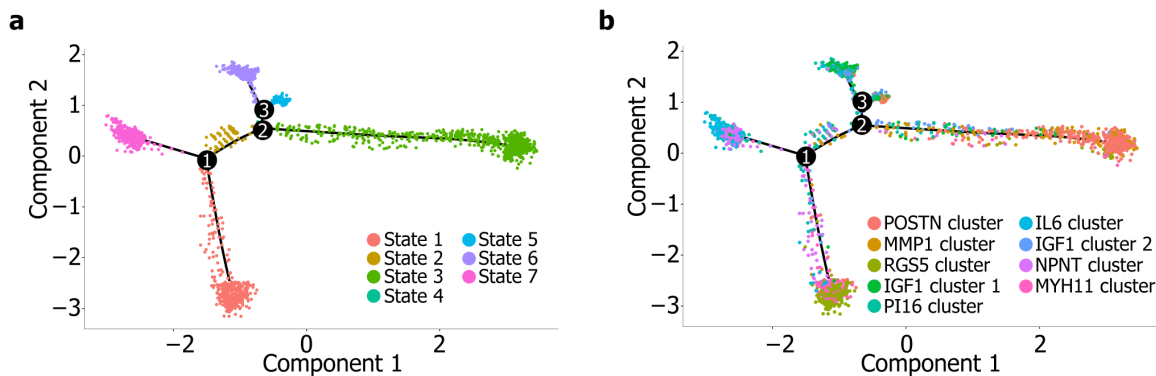


Figure 6.6 Primary stromal cells from the combined dataset (described in Section 5.3) show distinct differentiation trajectories. Trajectory plots coloured by (a) State and (b) cluster

### 6.3.1 Manipulation of *in vitro* culture conditions upregulates genes associated with *ex vivo* fibroblast populations

Cells treated as above were collected at seventy-two hours and underwent RNA extraction and RT-PCR analysis as *per* Sections 2.2 and 2.3. A panel of genes indicative of differentiation to distinct fibroblast clusters or trajectory states was generated using the differentially expressed genes listed in Table 5.1 and Figure 6.1c. RNA concentrations were calculated as the fold-change in gene expression levels relative to control (2 kPa CTL). RT-PCR results are shown in Figure 6.7.

Differentiation to State 3 (largely composed of *POSTN* and *MMP* CAF) was characterised by an increase in the expression of *POSTN* and *COL1A1* (the latter also differentially upregulated in all CAF relative to NOF *ex vivo*; Figure 6.1). *COL1A1* was upregulated relative to the control across substrates in both untreated and TGF- $\beta$ -treated samples (Figure 6.7a). Expression of *POSTN* was significantly increased in TGF- $\beta$ -treated cells cultured on collagen-coated low elastic modulus plates only, with a significant reduction in expression on collagen-coated tissue culture plastic (Figure 6.7b).

States 5 and 6 were predominantly composed of cells from multiple CAF populations. Over half of cells in State 5 originated from the “stress response” phenotype, also identified by *IGF1*



expression. Sixty-four percent of cells in State 6 were from the “IGF signalling” population, also marked by *IGF1*. *IGF1* was upregulated in TGF- $\beta$ -treated samples across all substrates (Figure 6.7c); expression of *SERPINE1* (also upregulated in State 5) was significantly increased in culture of TGF- $\beta$ -treated cells on collagen-coated low elastic modulus plates (Figure 6.7d).

State 7 was composed almost exclusively of NOF populations, most frequently the *IL6* “inflammatory” phenotype. This trajectory showed differential expression of *IL6* and *CCL2* (as did the “inflammatory” population). *IL6* was upregulated in IL-1 $\beta$ -treated samples across all substrates (Figure 6.7e). Expression of *CCL2* was increased in IL-1 $\beta$ -treated cells cultured on collagen-coated low elastic modulus plates or tissue culture plastic (Figure 6.7f).

CAF in trajectory State 3 are characterised by upregulation of *POSTN* (Figure 6.6). As expression of this gene was only upregulated in cells cultured with TGF- $\beta$  on collagen-coated low elastic modulus surfaces, these conditions were selected for generation of this phenotype (acknowledging that this may also induce a phenotype showing some overlap with CAF in State 5, which showed significant enrichment for genes upregulated by TGF- $\beta$  treatment). Treatment with IL-1 $\beta$  and culture on collagen-coated low elastic modulus plates and tissue culture plastic resulted in upregulation of both *IL6* and *CCL2*, markers of the “inflammatory” fibroblast phenotype and trajectory State 7. As culture on tissue culture plastic also led to increased expression of *ACTA2* (associated with VSMCs and the classical “myofibroblastic” CAF phenotype; Figure 6.7g), culture on low elastic modulus plates was chosen to best recreate this phenotype.

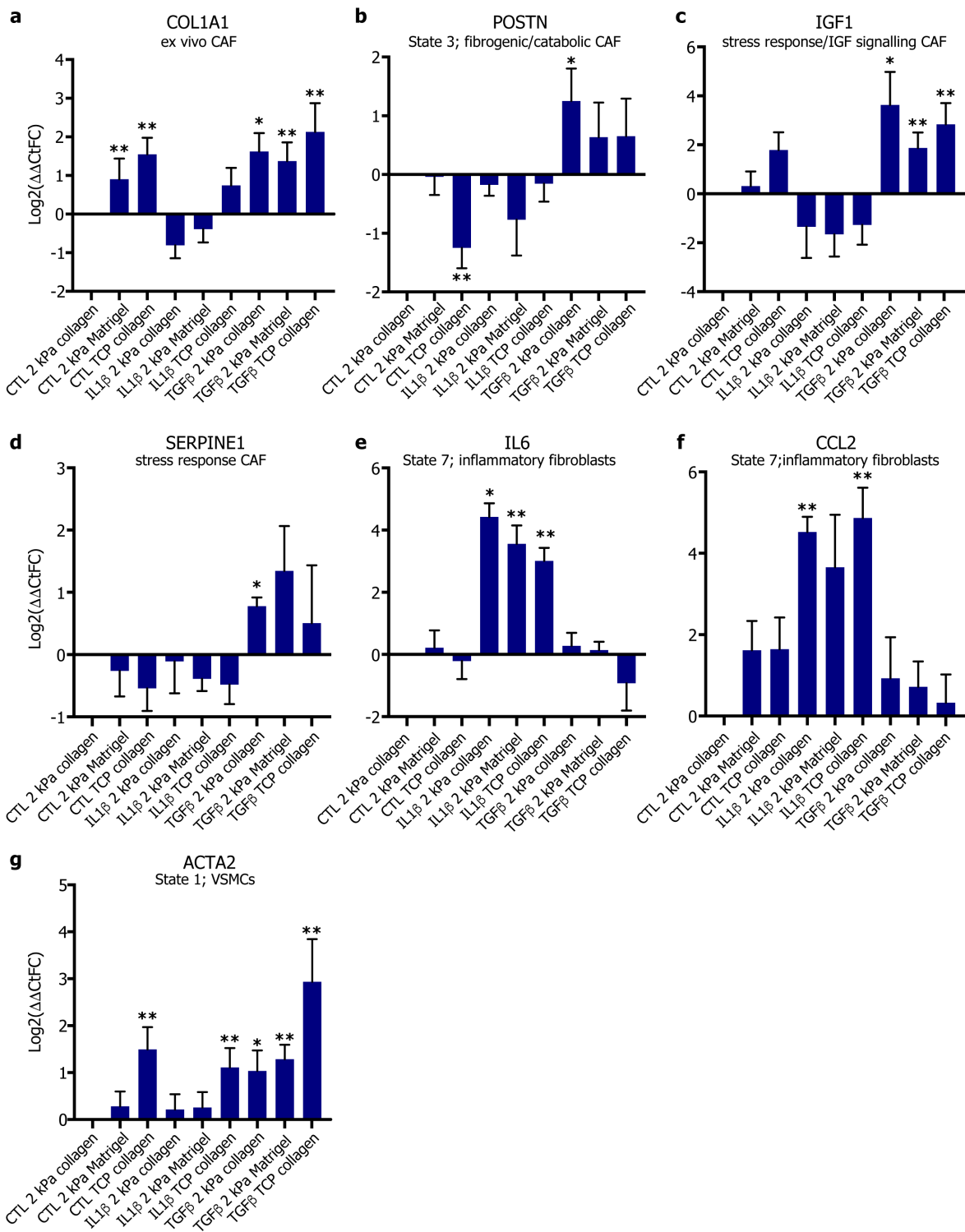


Figure 6.7 Alteration of culture conditions leads to upregulation of genes differentially expressed by *ex vivo* stromal cells

Bar charts showing fold change in expression for (a) *COL1A1*, (b) *POSTN*, (c) *IGF1*, (d) *SERPINE1*, (e) *IL6*, (f) *CCL2* and (g) *ACTA2* for either control (CTL), TGF- $\beta$ -treated or IL-1 $\beta$ -treated cells on low elastic modulus (2 kPa) and tissue culture plastic (TCP) plates. Values are expressed as log2-fold change relative to control (CTL 2 kPa collagen). \* $p < 0.05$ , \*\* $p < 0.01$ , Welch's  $t$ -test

### 6.3.2 Alteration of culture conditions partially recreates *ex vivo* fibroblast transcriptomes

To determine whether use of these conditions accurately recapitulated *ex vivo* phenotypes, treated cells were processed using the Drop-seq and bioinformatic pipeline described in Sections 2.7 and 2.8. The shared differentially expressed genes for the IL-1 $\beta$ , TGF- $\beta$  and TCP culture conditions and the comparable *ex vivo* populations are given in Table 6.2.

Culture on low elastic modulus plates with IL-1 $\beta$  treatment (IL-1 $\beta$  2 kPa) led to significant ( $p < 0.001$ ) upregulation of 10.5% of the genes differentially expressed by State 7 *ex vivo* cells (Table 6.2). *IL6* and *CCL2*, used as markers of trajectory State 7, were upregulated in cells grown in these conditions (echoing the gene expression pattern seen at RT-PCR; Section 6.2.1). As discussed previously, NF- $\kappa$ B is believed to be the main driver of the “inflammatory” fibroblast phenotype seen in State 7<sup>29</sup>. Both *IL6* and *CCL2* are included in a gene signature characterising the targets of NF- $\kappa$ B in fibroblasts<sup>231</sup>: these, and the other NF- $\kappa$ B targets upregulated by IL-1 $\beta$  2 kPa culture, are highlighted in bold in Table 6.2.

Seven percent of genes differentially expressed by cells in State 3 were upregulated by growth on low elastic modulus plates with TGF- $\beta$  treatment (TGF- $\beta$  2 kPa; Table 6.2). The increases in *COL1A1* and *POSTN* induced by these culture conditions at RT-PCR (Figure 6.7) were not seen in the transcriptomic data. *COL11A1*, *PRSS23*, *SULF* and *VCAN* (part of a set of genes upregulated by TGF- $\beta$  treatment<sup>83</sup>) were differentially upregulated in both State 3 and TGF- $\beta$  2 kPa conditions. However, this condition did not increase the expression of the multiple fibrillar collagens seen in State 3 (e.g. *COL1A2*, *COL3A1* and *COL5A2*). Cells grown on tissue culture plastic upregulated 5% of the genes differentially expressed in State 3 (TCP; Table 6.2). Among these were multiple genes associated with “myofibroblastic” differentiation and functions, including *ITGB1*, *TIMP1* and *TNC*<sup>93,252-254</sup>.

IL-1 $\beta$ 2 kPa and State 7; <i>IL6</i> ("inflammatory")		TGF- $\beta$ 2 kPa and State 3 ( <i>POSTN/MMP1</i> CAF)		TCP and State 3 ( <i>POSTN/MMP1</i> CAF)	
<b><i>CXCL1</i></b>	<i>BIRC3</i>	<i>AEBP1</i>	<i>COL6A3</i>	<i>MT2A</i>	<i>TPM1</i>
<b><i>IL8</i></b>	<i>GAPDH</i>	<i>ALDH1A1</i>	<i>CTHRC1</i>	<i>TAGLN</i>	<i>CHN1</i>
<i>CXCL3</i>	<i>NAMPTL</i>	<i>RARRES2</i>	<i>MMP14</i>	<i>TPM2</i>	<i>CYR61</i>
<b><i>SOD2</i></b>	<i>CH25H</i>	<i>AGT</i>	<i>RGCC</i>	<i>SFRP1</i>	<i>ASPN</i>
<b><i>CXCL2</i></b>	<i>COL3A1</i>	<i>SEPP1</i>	<i>TGM2</i>	<i>THY1</i>	<i>FILIP1L</i>
<i>TNFAIP6</i>	<i>TNFRSF11B</i>	<i>GREM1</i>	<i>ARL4C</i>	<i>ITGB1</i>	<i>KRT7</i>
<b><i>CCL2</i></b>	<i>ANXA1</i>	<i>C10orf10</i>	<i>TXNIP</i>	<i>S100A10</i>	<i>SULF1</i>
<i>NAMPT</i>	<b><i>IER3</i></b>	<i>CTSK</i>	<i>CST3</i>	<i>TIMP1</i>	
<b><i>RND3</i></b>	<i>SPARC</i>	<i>KIAA119</i>	<i>PLIN2</i>	<i>IGFBP6</i>	
<i>CA12</i>	<i>IL31</i>	<i>COL11A1</i>	<i>ANTXR1</i>	<i>ADIRF</i>	
<i>GOS2</i>	<i>CLU</i>	<i>HIST1H4C</i>	<i>AKAP12</i>	<i>IGFBP3</i>	
<i>WTAP</i>	<i>EMP1</i>	<i>PTGDS</i>	<i>SNHG6</i>	<i>UACA</i>	
<i>ZC3H12A</i>		<i>LMCD1</i>	<i>SNHG5</i>	<i>TNC</i>	
<i>ACSL4</i>		<i>PODN</i>	<i>BGN</i>	<i>LMO7</i>	
<i>FST</i>		<i>FBX032</i>	<i>MFAP4</i>	<i>ITGBL1</i>	
<b><i>IL6</i></b>		<i>C1QTNF5</i>	<i>VMP1</i>	<i>TNFRSF11B</i>	
<i>FTH1</i>		<i>PRSS23</i>	<i>SULF1</i>	<i>WNT5A</i>	
<i>PHDLA1</i>		<i>SAT1</i>	<i>MGP</i>	<i>MT1E</i>	
<b><i>NFKBIA</i></b>		<i>FGF7</i>	<i>VCAN</i>	<i>SERPINE1</i>	
<i>PPAP2B</i>		<i>GADD45B</i>	<i>GDF15</i>	<i>PTX3</i>	
<i>ALDOA</i>		<i>ISLR</i>	<i>SPON2</i>	<i>ADAMTS1</i>	
<i>SERPINE2</i>		<i>ANGPTL2</i>	<i>TIMP1</i>	<i>MYL12B</i>	
<i>CAV1</i>		<i>SCN7A</i>	<i>CHI3L1</i>	<i>RP11-48O20.4</i>	

Table 6.2 Overlap between genes significantly ( $p < 0.001$ ) differentially expressed in both *in vitro* culture conditions and *ex vivo* counterpart trajectory State

The IL-1 $\beta$ -treated genes highlighted in bold are targets of NF $\kappa$ B<sup>231</sup>

A random forest classifier was trained on the stromal cell dataset described in the previous chapter, using the genes differentially expressed by each cluster in both the Drop-seq and Lambrechts *et al.*<sup>127</sup> datasets. The trained classifier was used to assign *in vitro* skewed cells to the clusters identified in the stromal cell dataset. The results of this are summarised in Figure 6.8 below (full results are given in Section A1.11).

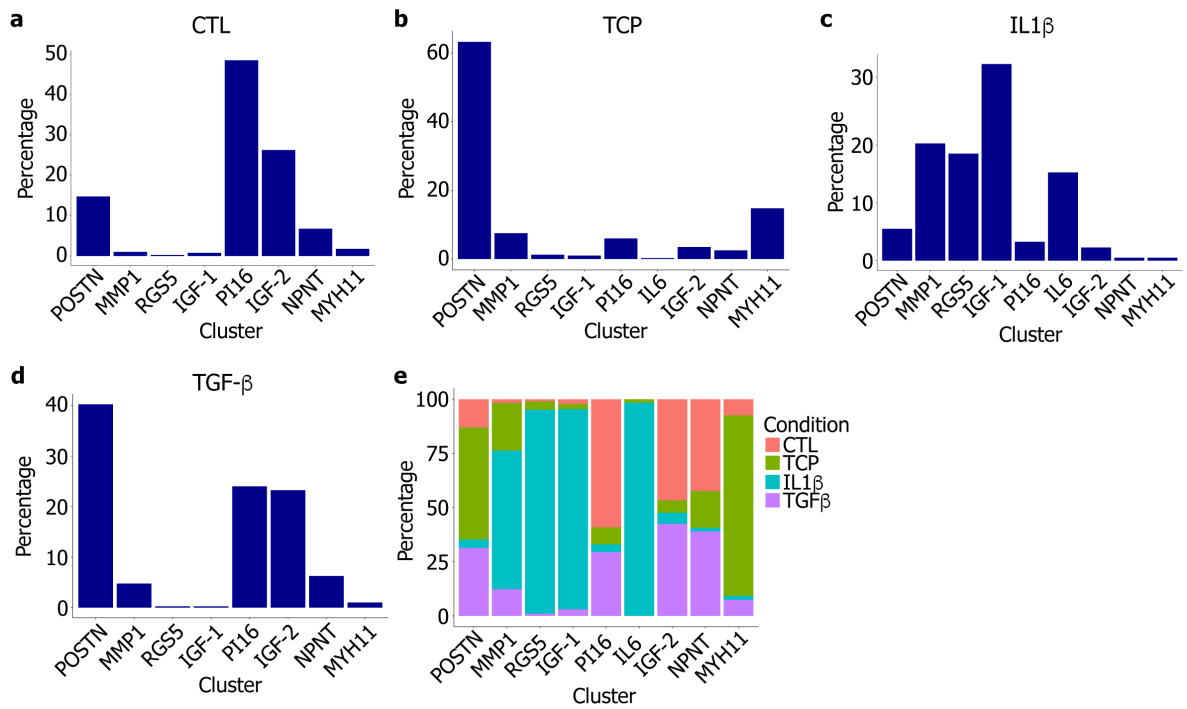


Figure 6.8 Manipulation of culture conditions partially recreates the transcriptomes of *ex vivo* stromal cells

Stacked barplots showing composition of (a-d) culture conditions assigned by stromal cell clusters and (e) stromal cell subtype by culture condition

All *ex vivo* stromal cell clusters were represented in the sequenced library; each culture condition generated multiple phenotypes (Figure 6.8a-d). Although under control conditions (untreated cells cultured on low elastic modules plates; CTL) a number of cells were classified as *POSTN* or *IGF1* cluster 2 CAF, cells were most frequently assigned to the *PI16* NOF subtype (49%; Figure 6.8a). The *PI16* phenotype is the most frequently-occurring cell type at differentiation pseudotime zero. It therefore appears that culture under control conditions is at least partially successful in recreating a normal fibroblast phenotype.

Treatment with IL-1β yielded a range of phenotypes, most commonly that of the *IGF1*-1 cluster (Figure 6.8c). Ninety-eight percent of cells assigned to the *IL6* cluster, and over 90% of cells assigned to the *RGS5* and *IGF1* cluster 1 subtypes, originated from the IL-1β-treated population (Figure 6.8e). As hypothesised, cells treated with TGF-β were predominantly (40%) assigned to the *POSTN* cluster (Figure 6.8d). However, this was not sufficient to convert all cells to this phenotype: TGF-β treatment also generated cells assigned to CAF (*IGF1* cluster 2) and NOF (*PI16* cluster) subtypes. Interestingly, 68% of cells cultured in 10% FCS on TCP were assigned to the *POSTN* CAF cluster (Figure 6.8b).

### 6.3.3 Treated cells show distinct differentiation trajectories

Trajectory analysis (Section 2.8.8) was performed to assess both the differences and similarities in gene set enrichment between the transdifferentiated *in vitro* cells and the *ex vivo* data (Section 5.3). TGF- $\beta$ -treated cells largely shared the same State as control (CTL) cells, although a minority were allocated to a trajectory with a subset of cells cultured in 10% FCS on tissue culture plastic (TCP, State 3; Figure 6.9). The majority of cells from TCP and IL-1 $\beta$ -treated samples were assigned to distinct States (4 and 1, respectively). Counts of cells assigned to each trajectory are given in Section A1.12.

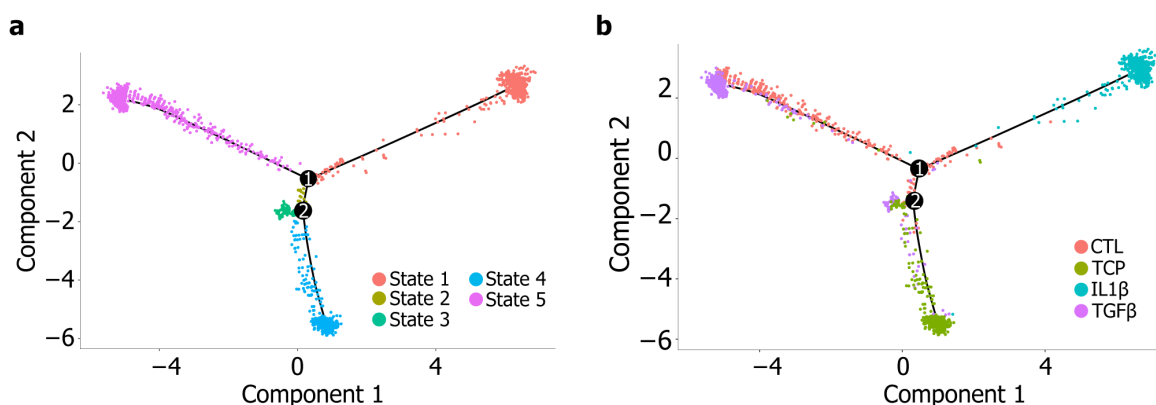


Figure 6.9 Transdifferentiated fibroblasts show distinct differentiation trajectories

Trajectory plots coloured by (a) State and (b) sample of origin. Points represent individual cells

Gene set enrichment analysis was performed to characterise the biological processes and curated gene sets upregulated by each trajectory (summarised in Table 6.3; further results are given in Section A.13). Differentiation to State 1 (IL-1 $\beta$  trajectory) was, similar to the *ex vivo* IL6 cluster, associated with an increase in the expression of inflammatory response gene sets (including the targets of NF- $\kappa$ B). States 3 and 5 both showed significant enrichment of one gene set only. State 3, comprising cells from TGF $\beta$  and TCP conditions, was enriched for a gene set upregulated in untreated cells vs. those treated with interferon gamma. This was largely a result of the differential expression of *ITGB1* (included in the signature) in this trajectory. State 5, also composed of both TGF- $\beta$ -treated cells and those grown on TCP, showed increased expression of a catabolic signature as a result of the differential expression of a number of ribosomal proteins by this trajectory. Leading edge analysis was not possible for States 3 and 5, as this requires multiple gene sets for comparison.

State 1 (IL-1 $\beta$ )	State 3 (TCP/TGF $\beta$ )	State 5 (TCP/TGF $\beta$ )
NES 1.637, FDR $q$ 0.025	NES 1.287, FDR $q$ 0.079	NES 1.399, FDR $q$ 0.157
<i>CXCL1</i> , <i>CXCL12</i> , <i>CCL2</i> , <i>CCL8</i>	N/A	N/A

Table 6.3 Representative GSEA results for differentiation from State 5 to each terminal State

Associated enrichment statistics and top leading edge genes are given beneath each plot. NES, Normalised Enrichment Score. FDR  $q$ , false discovery rate

## 6.4 Characterising fibroblast phenotypes

### 6.4.1 Migration assays

CAFs are known to show motile properties<sup>94,200</sup>; we therefore sought to determine whether treated cells showed differential migratory capacities. Treated fibroblasts were plated for migration assays as described in Section 2.1.10, using serum as a chemoattractant. Results are shown in Figure 6.10; cell counts are expressed as the log<sub>2</sub> of the fold change relative to the control sample. There were no significant differences in migration between transdifferentiated cells and control.

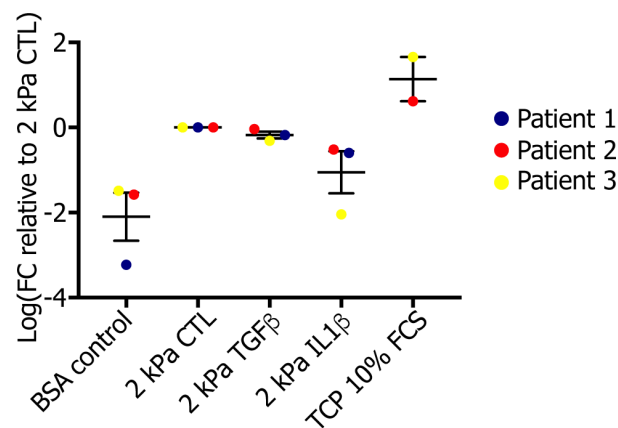


Figure 6.10 Treated fibroblasts do not show differential migration towards serum

Dot plot showing migrated cell counts for transdifferentiated fibroblasts to serum. Values are expressed as the log2 of the fold change relative to the control (2 kPa CTL) sample (n = 3)

6.4.2 Gel contraction assays

To determine whether the differentiated fibroblast populations show differential contractile properties, treated cells were plated to gel contraction assays as *per* Section 2.1.9; results are shown in Figure 6.11. Cells grown in 10% serum on tissue culture plastic showed the greatest reduction in gel mass relative to control; TGF-β and IL-1β-treated cells appeared similar to the control (CTL).

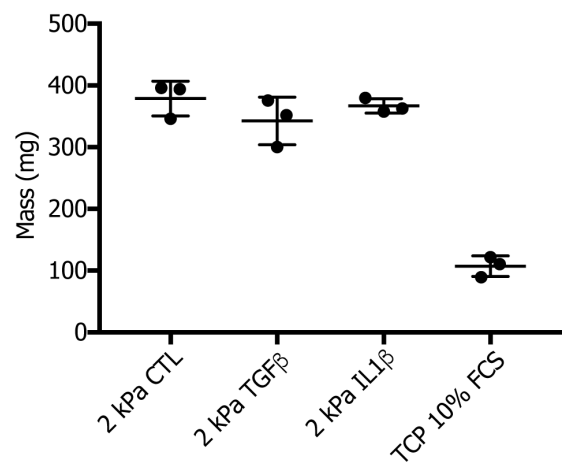


Figure 6.11 Transdifferentiated NOF show differential contractile capacity

Dot plot showing mass of gel following forty-eight hours in culture for control cells (CTL), TGF-β- and IL-1β-treated cells cultured on low elastic modulus plates (2 kPa) and those cultured on tissue culture plastic (TCP). Graphs show values with standard deviation for technical replicates (n = 3)



## 6.5 Discussion

Any functional characterisation of the identified fibroblast subtypes would initially necessitate culture *in vitro*: this process can also further understanding of *in vivo* fibroblast activation pathways. We assessed the impact of such techniques on the gene expression of primary fibroblasts, showing that this results in a significant alteration in their transcriptome. We then examined how manipulation of *in vitro* culture conditions can alter fibroblast proliferation rate and expression of genes and proteins.

Fibroblasts cultured on uncoated plastic surfaces and in 3D Matrigel gels showed differential expression of genes associated with the previously-described “myofibroblastic” and “inflammatory” fibroblast subtypes<sup>16,37,120</sup>. However, culture in 3D alone was not sufficient to recover the *ex vivo* fibroblast gene expression profile, indicating that this approach required further refinement. The effectiveness of culture on low elastic modulus plates with either TGF- $\beta$  or IL-1 $\beta$  treatment (based on the trajectory results in the previous chapter) in recreating *ex vivo* phenotypes was initially assessed using RT-PCR. Based on expression of select markers these conditions recreated the gene expression profiles of the *ex vivo* State 3 and State 7 cells, respectively. However, analysis of treated cells using Drop-seq indicated that this method is only partially successful in recreating *ex vivo* transcriptomes.

Treatment with IL-1 $\beta$  resulted in differential upregulation of a number of the targets of NF- $\kappa$ B<sup>231</sup>, considered to be the main driver of differentiation to the “inflammatory” phenotype<sup>29</sup> (from which the greatest proportion of *ex vivo* cells in State 7 originated). IL-1 $\beta$  treatment also generated a proportion of cells with a transcriptome similar to that of “inflammatory” fibroblasts. However, this condition also yielded a range of both NOF and CAF phenotypes, most commonly that of the “stress response” (*IGF1-1*) cluster. Although this population was characterised by expression of stress response genes, it also showed upregulation of a subset of inflammatory genes expressed by the *IL6* cluster (e.g. *CCL2*, *CXCL2*). This suggests that refinement of these *in vitro* conditions is required to prevent triggering a stress response when inducing the “inflammatory” phenotype.

Treatment of cells with TGF- $\beta$  alone appears insufficient to fully recreate the phenotypes of *ex vivo* cells in State 3. State 3 was largely composed of cells from the “fibrogenic” and “catabolic” populations. At scRNA-seq analysis, TGF- $\beta$ -treated cells did not show upregulation of key genes which are either associated with differentiation to State 3, or likely to be of functional importance in these phenotypes (e.g. *POSTN*, *FN1*, fibrillary collagens<sup>37,72,78,255</sup>). This lack of *POSTN* upregulation is in contrast to the RT-PCR results. There are potentially a number of reasons for this: RT-PCR is performed on total RNA, whereas Drop-seq uses captured mRNA; alternatively,

*POSTN* may show low mRNA expression level, which would make it more subject to dropout. It is also possible that the observed differences are due to different passage fibroblasts being used for the RT-PCR and Drop-seq analyses: further work directly comparing fibroblasts from the same passage would assist in elucidating the reasons behind these differences. Cells treated with TGF- $\beta$  also showed upregulation of *IGF1* and *SERPINE1* at RT-PCR (associated with separate CAF subtypes and trajectories). In keeping with this, TGF- $\beta$  treatment yielded a range of populations at scRNA-seq (although most frequently the *POSTN* phenotype), and does not generate State 3 cells specifically.

To date, the majority of *in vitro* fibroblast functional assays have been performed using culture on plastic substrates. At scRNA-seq analysis, this approach yielded predominantly *POSTN* (“fibrogenic”) phenotype cells. Cells grown on plastic also appear to show increased contractility but no difference in migration relative to the control. Together, these features are more consistent with a “myofibroblastic” CAF phenotype, and could be induced as a result of the rigidity of the culture substrate<sup>247,256</sup>. Indeed, it is known that traditional *in vitro* fibroblast culture using stiff substrates will yield an “activated” (CAF-like) phenotype<sup>257-259</sup>: the current data in this field are reflective of this.

Functional differences between the skewed fibroblast phenotypes were examined using migration and gel contraction assays: all results require experimental repeats for validation. The differentiated populations showed differential contractile capacities, although no differences in migration were observed. However, neither of the recreated phenotypes represent any one *ex vivo* fibroblast population or differentiation trajectory. Refinement of the skewing protocol to recreate more faithfully the *ex vivo* fibroblast phenotypes is needed, before repetition of these experiments to determine functional differences between the populations.

This work has shown some potential for using recombinant protein stimulation to re-create different fibroblast sub-populations *in vitro*. An alternative approach could be the use of shRNA or overexpression vectors to achieve longer-term modulation of gene expression<sup>260</sup>, which may generate more stable phenotypes. Alternatively, a more comprehensive recapitulation of the 3D microenvironment that generates these phenotypes *in vivo* may be required to accurately recreate these phenotypes. Spheroid or organoid cultures mimic *in vivo* architecture and cell-cell interactions in a variety of tissues, including the lung<sup>261,262</sup>. Such techniques have also been shown to recreate *in vivo* gene expression in, for example, melanoma<sup>263</sup>; it is therefore possible that the incorporation of stromal cells into these models could facilitate recreation of fibroblast phenotypes more akin to those seen *in vivo*.

*In vitro* functional characterisation will require faithful recapitulation of the *ex vivo* fibroblast phenotypes. Using differentiation stimuli identified by trajectory analysis in the previous chapter partially recreates *ex vivo* gene expression profiles, but requires refinement to improve accuracy. Preliminary analysis of the skewed populations indicates that these groups may have differential functional properties. More extensive characterisation will identify pro-malignant subtypes and differentiation pathways for potential future therapeutic targeting.



## Chapter 7 Discussion

An increasing body of evidence indicates that fibroblasts are a heterogeneous population in both normal and disease states<sup>16,17,21,33</sup>; the aim of this project was to characterise this variation, and to link the molecular phenotypes of any distinct subpopulations to their functions.

To ensure accurate characterisation of *in vivo* phenotypes across the entire stromal cell population in NSCLC and normal lung, we have addressed a number of technical challenges. We have developed novel protocols for the processing of human tissue samples to maximise the yield of fibroblasts, and for enhanced quality control of scRNA-seq data generated from these samples. Applying this approach to a larger cohort of NSCLC samples has identified nine stromal sub-populations, which are likely to have distinct functional roles within the tumour microenvironment. We have also used this data to identify key molecular mechanisms which regulate transdifferentiation between these sub-populations. Finally, we have examined both the potential and limitations associated with examining these stromal phenotypes *in vitro*.

An initial challenge in this project was to determine the optimal approach to quantify extraction of fibroblasts from tissue samples. As there is no single marker that will reliably identify all fibroblasts, we assessed the suitability of previously-described fibroblast markers to identify lung fibroblasts. We found CD90 to be a highly sensitive and robust marker in comparison to PDGFR- $\alpha$  (previously described as a marker for fibroblast isolation across multiple tissue types<sup>264</sup>). It is of note, however, that CD90 was expressed by fewer than half of skin fibroblasts. It is therefore likely that this approach will not effectively identify fibroblasts in all tissues, and that some fibroblasts will not be identified using the FACS panel described.

The majority of solid tissue disaggregation protocols have focused on immune cell isolation<sup>135,136,184-186</sup>. We compared different disaggregation durations and enzymatic cocktails, finding that extended Collagenase incubation times were required to release fibroblasts from tissue samples. In contrast, non-adherent cells (such as immune cells) were more rapidly isolated by enzymatic disaggregation. This analysis demonstrated that isolating cells from tissue samples using such techniques may not accurately reflect the cell type proportions present in the primary tissue. This impacts the application of FACS-based analysis of single cell suspensions to quantify variation in the fractions of different cell types present within tissues. This also has significant implications for applications such as the scRNA-seq performed in this study: disaggregation protocols should be optimised to maximise the extraction of the population of interest. This will ensure that the examined sample of these cells

is the most accurate possible representation of the entire population. It is also noteworthy that even following extended Collagenase incubation times, the fraction of extracted epithelial cells was surprisingly low given their abundance in tissue sections. We hypothesised that cell-cell adhesions resulted in their loss during the filtration steps required to generate a single cell suspension, preventing their isolation. We showed that addition of TrypLE to the disaggregation protocol addressed this, generating a significantly higher fraction of epithelial cells. This also demonstrated that this protocol may be further refined or optimised to enrich for different cell types for future studies.

Droplet-barcoded (Drop-seq) single-cell RNA sequencing is a cost-effective method for profiling large numbers of cells. Initial optimisation using cell lines confirmed that this technology was successfully able to distinguish between different cell lineages. The described disaggregation approach was then used for transcriptomic profiling of *ex vivo* fibroblasts using Drop-seq. Comparison of data from lung cell lines and primary *ex vivo* cells indicated the need for quality control metrics tailored to sample type and sequencing platform.

We developed a standardised approach for the removal of low-quality events from scRNA-seq data; this improves clustering quality compared to the use of previously-described quality-control metrics<sup>125-127,198,199</sup>. Enzymatic dissociation has been reported to impact gene expression<sup>138</sup>.

Disaggregation-associated changes in gene expression have the potential to result in assignment of cells to the incorrect lineage, and to generate erroneous clusters composed of cells with high expression of these disaggregation-associated genes. We therefore refined an existing disaggregation-induced gene signature to assess the impact this has on the Drop-seq data. Clustering quality was not further improved by application of the refined signature, although it appears that some immune cell types are differentially impacted by enzymatic disaggregation.

Applying this optimised processing pipeline to data from 12 tumour samples identified 33 distinct clusters. This will provide a valuable resource for other researchers, both as a “test” dataset for validation of findings made in other data series, and as a tool to allow interrogation of other cell types. Clusters were labelled using a co-expression atlas together with the top differentially expressed genes for each group. This method provides an unbiased approach for assigning scRNA-seq clusters to different cell types, and is likely to be more robust than labelling of clusters based on expression of one or two differentially upregulated genes. However, this process relies on the inclusion of a reference gene set for each of the cell types under investigation. For some cell types (*e.g.* mast cells) such signatures were absent: identification of these populations required additional refinement before their final cell type identity was assigned.

To further minimise the impact of low-quality or misclassified cells on downstream analysis, we created a stromal gene signature. Cross-referencing the top differentially expressed genes for this cluster with those present in the most significant co-expression atlas result gave three genes: *COL1A2*, *COL3A1* and *DCN*. Cells belonging to the “Stromal cell” cluster in the larger dataset with a low expression of these three genes in combination were not taken forward. Such cell type-specific signatures may also prove useful for refining cell types with a lower nGene, such as plasma cells and T cells.

This refined stromal data was used to identify stromal cells in a NSCLC scRNA-seq dataset published during the course of this project<sup>127</sup>. Analysis of these combined stromal populations identified 9 distinct stromal clusters: 4 CAF, 3 NOF, 1 pericyte and 1 VSMC population. Of the CAF populations, two (marked by *POSTN* and *MMP1* expression) showed expression of genes and enrichment of gene sets suggestive of “fibrogenic” and “catabolic” phenotypes. It is possible that these populations represent subgroups of the “myofibroblastic” phenotype, which has traditionally been seen as responsible for the generation and remodelling of tumour stroma<sup>17,18,77</sup>. The remaining CAF clusters are marked by *IGF1* upregulation and were labelled “stress response” and “IGF signalling” CAF due to their expression of genes involved in the response to cellular stress, and mediation of IGF signalling, respectively. Of the clusters predominantly originating from normal tissue, one, marked by *IL6*, was consistent with the ‘inflammatory’ fibroblast phenotype described previously<sup>29,120,231</sup>. The remaining two populations were marked by *NPNT* and *PI16* expression, and may represent “matrix” and quiescent fibroblast populations, respectively.

Immunohistochemical validation of staining for periostin (the product of *POSTN*) and serpin E1 confirmed that these proteins appear to have differential spatial expression and presence across NSCLC subtypes. Serpin E1 showed co-expression with  $\alpha$ -SMA in some stromal cells. Staining for periostin was largely extracellular, and was found in similar regions to the cytoplasmic  $\alpha$ -SMA staining. The selection of markers for the two CAF groups requires refinement: periostin in particular extensively stains the ECM, rendering identification of intracellular staining challenging, and staining for serpin E1 is relatively weak. Staining for intracellular or cell surface epitopes, or *in situ* hybridisation (using antibodies to mRNA) may facilitate more accurate spatial characterisation and mapping. Spatial mapping for each of the identified populations may provide an additional layer of functional information: for example, CAFs are known to have immunomodulatory roles<sup>79,98</sup>. Assessment of the spatial relationships between CAFs and cytotoxic T cells may suggest a mechanism by which CAFs impair the immune response to tumour *e.g.* by immune cell exclusion from the tumour or inducing T cell exhaustion.

Trajectory analysis of the stromal cell data indicated distinct differentiation pathways for the identified populations. The “fibrogenic” and “catabolic” phenotypes shared a common trajectory; the other CAF trajectories contained multiple phenotypes. All CAF trajectories showed some degree of TGF- $\beta$  dependence, consistent with previous observations describing the importance of this cytokine in fibroblast activation<sup>75,248,265</sup>. It is possible that this represents a common CAF phenotype driver, with differentiation to distinct phenotypes dependent on secondary stimuli (such as Ras signalling or cellular senescence). Differentiation to the trajectory containing “inflammatory” fibroblasts appears at least in part to be mediated by NF- $\kappa$ B signalling, although the other drivers of this trajectory are not clear.

A number of studies have demonstrated that mechanical changes to tissue culture substrates can impact fibroblast phenotypes<sup>266-269</sup>, yet the vast majority of research analysing these cells is still carried out on plastic. Our results show that culturing fibroblasts on plastic causes a significant shift in transcriptome compared to that seen immediately following isolation, consistent with previous studies<sup>125,126</sup>. In the scRNA-seq data, in keeping with other observations in this area<sup>257-259</sup>, culture of fibroblasts on a rigid substrate appears to generate an activated “myofibroblastic” phenotype. This has implications for both planning of future work and the interpretation of previous studies of fibroblast function. In the majority of studies to date, CAFs are described as having tumour-promoting functions *in vitro*<sup>16</sup>. However, some data (for example, immunohistochemical analyses of patient tissues) indicate that CAFs may have differential prognostic impact<sup>44,104,149,270</sup>. This emphasises the need to profile these cells directly from tissues, without culture *in vitro*, to characterise accurately *in vivo* phenotypes. Using genes identified by both differential gene expression and trajectory analysis, we created a panel to allow assessment of *in vitro* fibroblast phenotypes by RT-PCR. This identified that culture on low elastic modulus substrates with either IL-1 $\beta$  or TGF- $\beta$  treatment recreated some of the *ex vivo* phenotypes. However, more extensive transcriptomic analysis with scRNA-seq revealed that this approach was only partially successful, and further refinement of differentiation stimuli is needed.

Although Drop-seq is a cost-effective platform for profiling larger numbers of cells, this approach gives a lower number of genes *per cell* than, for example, SMART-seq2<sup>130</sup>. The latter platform requires upfront FACS of target populations (and therefore, well-defined surface markers). Prior to this work, such markers were lacking for fibroblasts in NSCLC. Use of the markers identified in the scRNA-seq data presented here could allow transcriptomic analysis of the identified populations at greater resolution (*i.e.* more genes *per cell*). Higher resolution data may allow more detailed differential gene expression analysis. This in turn could facilitate more informative gene set



enrichment and trajectory analysis and allow better *in vitro* recreation of phenotypes and functional analysis.

CAFs are the most common stromal cell type, and highly prognostic in, a range of solid tumours<sup>37</sup>.

Targeting of pre-malignant CAF subtypes is an attractive prospect, but hampered by a lack of phenotypic characterisation. Here, we have used scRNA-seq to confirm that CAFs are a heterogeneous cell type, with multiple, apparently functionally distinct, sub-populations.

Bioinformatic deconvolution can be used to examine which of the identified fibroblast phenotypes would be appropriate therapeutic targets<sup>271</sup>. These techniques generate a gene expression signature for each cell type in a user's dataset. The presence of these signatures can then be evaluated in other datasets *e.g.* The Cancer Genome Atlas<sup>272</sup>, giving the relative proportions of each cell type for each patient in the database. These values can then be correlated with a variety of clinical parameters, and could identify fibroblast populations with a negative prognostic impact.

Identification of potential therapeutic targets will also require further functional characterisation. Trajectory analysis using the scRNA-seq data has given some indications of potential differentiation pathways to recreate *ex vivo* fibroblast phenotypes *in vitro*: these driver mechanisms may also represent therapeutic targets. However, the lack of significant GSEA results for trajectories makes determination of the precise drivers of each phenotype difficult: using alternative approaches to gene set enrichment, differential gene expression and trajectory analysis<sup>273-275</sup> may provide further information.



## Appendix A

### A.1 Summary of studies examining prognostic impact of CAF in NSCLC

The existing data regarding the prognostic impact of CAF in NSCLC are summarised in Table A.1.

### A.2 Experimental induction of apoptosis did not yield sufficient cDNA for library generation and sequencing

The fraction of reads in a sample which map to mitochondrial genes is used as a measure of cell death<sup>167</sup>. We attempted to validate this experimentally through scRNA-seq of apoptotic cells. H<sub>2</sub>O<sub>2</sub> is a commonly-used method for the experimental induction of apoptosis<sup>276</sup>. An initial optimisation experiment was performed to determine the optimal treatment length: cultured fibroblasts were treated with 50 mM H<sub>2</sub>O<sub>2</sub> for either 1 or 2 hours. Viability analysis by FACS revealed the fractions of live cells and 1 and 2 hours to be 27.7% and 1.73%, respectively. The live cell proportion at two hours was felt to be insufficient to allow subsequent comparison with the dead population. The dead cell fraction at 1 hour was inadequate to facilitate characterisation by Drop-seq; therefore, an intermediate timepoint of 90 minutes was selected.

Confluent cultured fibroblasts from one patient were treated with 50 mM H<sub>2</sub>O<sub>2</sub> for 90 minutes to induce apoptosis. Treated cells were then processed using the Drop-seq platform as *per* Section 2.7. BioAnalyzer quantification of the resulting cDNA library revealed a library concentration of 2.62 pg/μl (Fig. A.1). Generation of a Tagmented library for sequencing requires a minimum concentration of 100 pg/μl, therefore further analysis of this sample was not possible. Although widely used to induce apoptosis, H<sub>2</sub>O<sub>2</sub> may also lead to RNA damage and degradation<sup>277</sup>. It is therefore likely that although treatment successfully initiates cell death, it also results in sufficient mRNA damage to render analysis by scRNA-seq impossible. An alternative method for apoptosis would be required for further attempts at validation: previous reports confirming use of mitochondrial gene fraction as an appropriate proxy for cell death used digitonin<sup>168</sup>.

## Appendix A

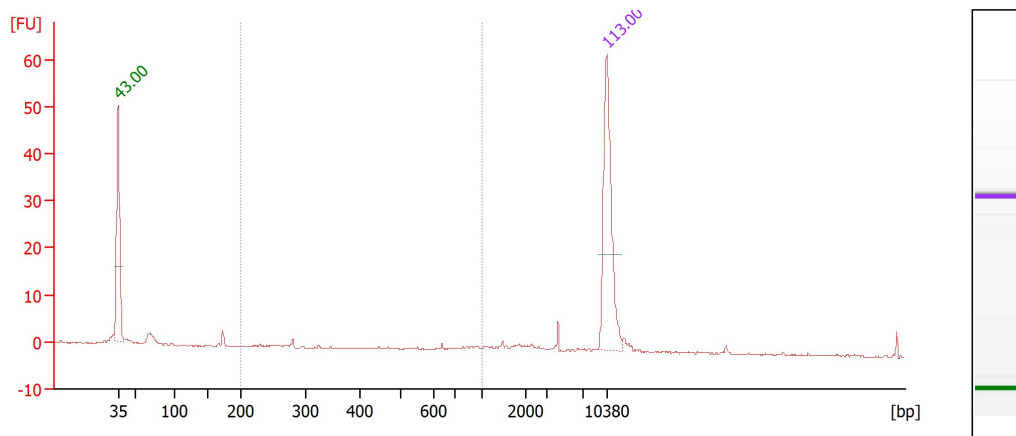


Figure A.1 Induction of apoptosis using  $\text{H}_2\text{O}_2$  does not generate sufficient cDNA for single-cell RNA sequencing with Drop-seq. BioAnalyzer trace showing cDNA library quantification

### A.3 Results of differential gene expression analysis for the identified stromal clusters

Differential gene expression analysis was performed as described in Section 2.8.4. The top twenty-five genes for each stromal cluster (by decreasing adjusted  $p$  value) are listed in Table A.2.

### A.4 Gene set enrichment (GSEA) results for stromal clusters

The full statistics for the top ten most enriched gene sets for each stromal cluster (summarised in section 5.3.2) are given in Tables A.3-14. ES, enrichment score; NES, normalised enrichment score; NOM  $p$ -val, nominative (not corrected for gene set size or multiple testing)  $p$  value; FDR, false discovery rate; FWER, family-wise error rate.

Marker	Survival effect	p value	Directional log(p value)	Analysis type	Survival measure	Other	Reference
CD99	Increased	0.039	1.409	MV	OR		Edlund <i>et al.</i> (2012) <sup>154</sup>
FAP	Increased	0.043	1.367	UV	HR	In LUAD, not LUSC	Kilvaer <i>et al.</i> (2015) <sup>44</sup>
FAP	Decreased	0.087	-2.060	-	OS		Liao <i>et al.</i> (2013) <sup>45</sup>
FGF-2	Increased	0.015	1.824	MV	DSS		Donnem <i>et al.</i> (2009) <sup>278</sup>
GLUT1-/MCT-4	Decreased	0.032	-1.495	-	DSS	In LUAD, not LUSC	Meijer <i>et al.</i> (2012) <sup>152</sup>
MMP-2	Decreased	0.01	-2	MV	OS, DFS		Leinonen <i>et al.</i> (2008) <sup>149</sup>
MMP-2	Decreased	0.022	-1.658	MV	HR	In LUSC, not LUAD	Ishikawa <i>et al.</i> (2004) <sup>150</sup>
PDGF-A	Increased	0.001	3	MV	DSS		Donnem <i>et al.</i> (2008) <sup>104</sup>
PDFRB	Increased	0.258	0.588	UV	OR		Edlund <i>et al.</i> (2012) <sup>154</sup>
Periostin	Decreased	0.011	-1.958	MV	HR		Hong <i>et al.</i> (2013) <sup>270</sup>
Periostin	Decreased	0.097	-1.013	UV	OR		Edlund <i>et al.</i> (2012) <sup>154</sup>
Podoplanin	Decreased	0.029	-1.537	MV	HR	LUAD, LUSC not examined	Ito <i>et al.</i> (2012) <sup>114</sup>
Podoplanin	Decreased	0.0011	-2.959	MV	OS	LUSC, LUAD not examined	Ono <i>et al.</i> (2013) <sup>113</sup>
Podoplanin	Decreased	0.001	-3	UV	OS	LUAD, LUSC not examined	Kawase <i>et al.</i> (2008) <sup>148</sup>
Podoplanin	Decreased	0.028	-1.552	UV	PFS	LUAD, LUSC not examined	Koriyama <i>et al.</i> (2014) <sup>115</sup>
Podoplanin	Decreased	0.01	-2	MV	OS	In LUAD, not LUSC	Kitano <i>et al.</i> (2010) <sup>108</sup>
p-Smad2	Decreased	0.049	-1.310	UV	OS		Chen <i>et al.</i> (2014) <sup>46</sup>
SPARC	Decreased	0.007	-2.155	MV	OS		Kurtul <i>et al.</i> (2014) <sup>163</sup>
SPARC	Decreased	0.308	-0.511	UV	OR		Edlund <i>et al.</i> (2012) <sup>154</sup>
SMA	Increased	0.22	0.658	UV	-	In LUSC, not LUAD	Kilvaer <i>et al.</i> (2015) <sup>44</sup>
SMA	Decreased	0.037	-1.431	MV	OS		Chen <i>et al.</i> (2014) <sup>151</sup>
Tenascin-C	Increased	0.646	0.190	UV	OR		Edlund <i>et al.</i> (2012) <sup>154</sup>
Versican	Decreased	0.0056	-2.252	UV	DFS	In LUAD, not LUSC	Pirinen <i>et al.</i> (2005) <sup>153</sup>
Versican	Increased	0.082	-1.086	UV	OR		Edlund <i>et al.</i> (2012) <sup>154</sup>

Table A.1 Prognostic impact of CAF markers in NSCLC. Positive directional log(p value) indicates positive prognostic impact; negative directional log(p value) indicates negative prognostic impact and is to three decimal places where applicable. MV, multivariate; UV, univariate; OR, odds ratio; HR, hazard ratio; OS, overall survival; DSS, disease-specific survival; DFS, disease-free survival; PFS, progression-free survival; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma

Appendix A

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
<i>POSTN</i>	<i>MMP1</i>	<i>RGS5</i>	<i>IGF1</i>	<i>PI16</i>	<i>IL6</i>	<i>IGF1</i>	<i>NPNT</i>	<i>MYH11</i>
<i>FN1</i>	<i>POSTN</i>	<i>NDUFA4L2</i>	<i>SERPINE1</i>	<i>SCARA5</i>	<i>KDM6B</i>	<i>CXCL12</i>	<i>SCN7A</i>	<i>PPP1R14A</i>
<i>COL11A1</i>	<i>MMP11</i>	<i>HIGD1B</i>	<i>APOE</i>	<i>CFD</i>	<i>NAMPT</i>	<i>PTGDS</i>	<i>LIMCH1</i>	<i>TAGLN</i>
<i>INHBA</i>	<i>MMP3</i>	<i>MCAM</i>	<i>UBC</i>	<i>MFAP5</i>	<i>HAS1</i>	<i>SERPINF1</i>	<i>ADAMTS8</i>	<i>ADIRF</i>
<i>COL1A2</i>	<i>DIO2</i>	<i>GJA4</i>	<i>DUSP1</i>	<i>PCOLCE2</i>	<i>ADAMTS4</i>	<i>IGFBP4</i>	<i>A2M</i>	<i>ACTA2</i>
<i>COL10A1</i>	<i>CTSK</i>	<i>COX4I2</i>	<i>CXCL2</i>	<i>IGFBP6</i>	<i>MT1A</i>	<i>RARRES1</i>	<i>INMT</i>	<i>NTRK3</i>
<i>CTHRC1</i>	<i>FAP</i>	<i>KCNJ8</i>	<i>ZFP36</i>	<i>GSN</i>	<i>ICAM1</i>	<i>PLA2G2A</i>	<i>BMP5</i>	<i>PTMA</i>
<i>COL5A2</i>	<i>CTHRC1</i>	<i>ANGPT2</i>	<i>TNFAIP3</i>	<i>CHRD1</i>	<i>SFTPC</i>	<i>TPT1</i>	<i>MAMDC2</i>	<i>IGFBP7</i>
<i>SULF1</i>	<i>INHBA</i>	<i>COL4A1</i>	<i>KLF6</i>	<i>DCN</i>	<i>DDX21</i>	<i>KIAA1324L</i>	<i>PLEKHH2</i>	<i>TINAGL1</i>
<i>SPARC</i>	<i>TDO2</i>	<i>CCDC102B</i>	<i>PLIN2</i>	<i>ADH1B</i>	<i>CCL2</i>	<i>APOE</i>	<i>TCF21</i>	<i>MCAM</i>
<i>COL8A1</i>	<i>MMP14</i>	<i>P2RY14</i>	<i>SGK1</i>	<i>AKAP12</i>	<i>C7</i>	<i>C3</i>	<i>MACF1</i>	<i>SYNPO2</i>
<i>COL6A3</i>	<i>CTSB</i>	<i>LAMC3</i>	<i>EFEMP1</i>	<i>OGN</i>	<i>ELL2</i>	<i>FGF7</i>	<i>FIGF</i>	<i>EFHD1</i>
<i>COL1A1</i>	<i>TGFBI</i>	<i>COL4A2</i>	<i>SOD2</i>	<i>CCDC80</i>	<i>MT2A</i>	<i>EFEMP1</i>	<i>MFAP4</i>	<i>PARM1</i>
<i>COL3A1</i>	<i>HLA-B</i>	<i>APOLD1</i>	<i>MMP19</i>	<i>LEPR</i>	<i>HAS2</i>	<i>COL1A1</i>	<i>FMO2</i>	<i>CCL21</i>
<i>TPM1</i>	<i>LOXL2</i>	<i>ESAM</i>	<i>BDKRB1</i>	<i>APOD</i>	<i>SFTPA1</i>	<i>MMP2</i>	<i>MYH10</i>	<i>CYFIP2</i>
<i>COL12A1</i>	<i>MME</i>	<i>EGFLAM</i>	<i>C3</i>	<i>NFIA</i>	<i>NR4A3</i>		<i>GDF10</i>	<i>MEF2C</i>
<i>LGALS1</i>	<i>GPR68</i>	<i>GJC1</i>	<i>KRT18</i>	<i>FBLN1</i>	<i>CHD1</i>		<i>PRELP</i>	<i>LMOD1</i>
<i>ACTA2</i>	<i>GAPDH</i>	<i>NOTCH3</i>	<i>LIF</i>	<i>CLU</i>	<i>CRISPLD2</i>		<i>ROBO2</i>	
<i>NTM</i>	<i>FTH1</i>	<i>IGFBP7</i>	<i>ALDH1A3</i>	<i>CST3</i>	<i>RDH10</i>		<i>CCBE1</i>	
<i>COL5A1</i>		<i>PLXDC1</i>	<i>GEM</i>	<i>PODN</i>	<i>SOD2</i>		<i>GPX3</i>	
<i>ITGBL1</i>		<i>LPL</i>	<i>NFKBIA</i>	<i>C3</i>	<i>CSF3</i>		<i>SEPP1</i>	
<i>NUAK1</i>		<i>TINAGL1</i>	<i>IER3</i>	<i>MGP</i>	<i>MAFF</i>		<i>DST</i>	
<i>TAGLN</i>		<i>GUCY1A2</i>	<i>LXN</i>	<i>ALDH1A1</i>	<i>RGS2</i>		<i>ADH1B</i>	
<i>VCAN</i>		<i>CLMN</i>	<i>PTX3</i>	<i>C16orf89</i>	<i>REL</i>		<i>RGCC</i>	
<i>TPM2</i>		<i>COL18A1</i>	<i>WTAP</i>	<i>MGST1</i>	<i>MT1M</i>		<i>TIMP3</i>	

Table A.2 Top twenty-five differentially expressed genes by decreasing adjusted *p* value for each stromal cluster

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP	34	0.93	1.93	0	0	0	33
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_DN	31	0.93	1.91	0	0	0	33
ANASTASSIOU_MULTICANCER_INVASIVENESS_SIGNATURE	46	0.86	1.83	0	0	0	54
VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_UP	57	0.82	1.78	0	0	0	72
DODD_NASOPHARYNGEAL_CARCINOMA_DN	65	0.81	1.77	0	0	0	61
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_UP	34	0.83	1.76	0	0	0	68
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	98	0.78	1.74	0	0	0	95
POOLA_INVASIVE_BREAST_CANCER_UP	27	0.85	1.73	0	0	0	54
PID_AVB3_INTEGRIN_PATHWAY	21	0.87	1.72	0	0	0.001	52
WANG_SMARCE1_TARGETS_UP	56	0.79	1.71	0	0.001	0.001	87

Table A.3 Top 10 GSEA curated gene sets results for cluster 0 (*POSTN* cluster)

# Appendix A

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
GO_SENSORY_ORGAN_DEVELOPMENT	22	0.85	1.69	0	0.012	0.014	16
GO_MULTICELLULAR_ORGANISM_METABOLIC_PROCESS	22	0.84	1.69	0	0.006	0.014	52
GO_SKELETAL_SYSTEM_DEVELOPMENT	52	0.78	1.69	0	0.004	0.014	64
GO_MULTICELLULAR_ORGANISMAL_MACROMOLECULE_METABOLIC_PROCESS	21	0.85	1.67	0	0.007	0.031	52
GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	60	0.76	1.66	0	0.009	0.050	94
GO_SINGLE_ORGANISM_CATABOLIC_PROCESS	46	0.75	1.62	0	0.024	0.148	55
GO_RESPONSE_TO_TRANSFORMING_GROWTH_FACTOR_BETA	20	0.80	1.57	0	0.069	0.421	10
GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION	25	0.78	1.57	0.002	0.064	0.438	83
GO_BIOLOGICAL_ADHESION	80	0.71	1.56	0	0.067	0.496	119
GO_FOREBRAIN_DEVELOPMENT	15	0.83	1.55	0.004	0.066	0.530	10

Table A.4 Top 10 GSEA biological processes gene set results for cluster 0 (*POSTN* cluster)



NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
DODD_NASOPHARYNGEAL_CARINOMA_DN	72	0.74	1.68	0	0.051	0.053	89
POOLA_INVASIVE_BREAST_CANCER_UP	30	0.81	1.65	0	0.075	0.148	42
ANASTASSIOU_MULTICANCER_INVASIVENESS_SIGNATURE	33	0.78	1.62	0	0.116	0.316	108
GRADE_COLON_CANCER_UP	39	0.76	1.60	0	0.115	0.396	89
CROMER_TUMORIGENESIS_UP	21	0.84	1.59	0.001	0.152	0.569	44
PUJANA_BRCA1_PCC_NETWORK	56	0.71	1.58	0	0.137	0.595	103
NUYTEN_EZH2_TARGETS_UP	90	0.67	1.57	0	0.158	0.709	78
MANALO_HYPOXIA_UP	34	0.76	1.57	0.004	0.141	0.716	85
WINTER_HYPOXIA_METAGENE	46	0.73	1.56	0.003	0.143	0.770	89
SANA_TNF_SIGNALING_UP	17	0.85	1.56	0.001	0.143	0.797	43

Table A.5 Top 10 GSEA curated gene sets results for cluster 1 (*MMP1* cluster)

# Appendix A

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
GO_CATABOLIC_PROCESS	84	0.72	1.68	0	0.041	0.038	89
GO_SINGLE_ORGANISM_CATABOLIC_PROCESS	59	0.74	1.64	0	0.067	0.119	89
GO_MULTICELLULAR_ORGANISM_METABOLIC_PROCESS	22	0.85	.64	0	0.044	0.119	77
GO_MULTICELLULAR_ORGANISMAL_MACROMOLECULE_METABOLIC_PROCESS	22	0.85	1.62	0	0.057	0.188	77
GO_ORGANONITROGEN_COMPOUND_METABOLIC_PROCESS	86	0.69	1.61	0	0.066	0.259	91
GO_BLOOD_VESSEL_MORPHOGENESIS	36	0.74	1.56	0.009	0.201	0.660	44
GO_NITROGEN_COMPOUND_TRANSPORT	18	0.84	1.55	0.001	0.186	0.691	68
GO_PROTEOLYSIS	53	0.71	1.54	0.003	0.208	0.781	66
GO_NUCLEOBASE_CONTAINING_SMALL_MOLECULE_METABOLIC_PROCESS	26	0.77	1.54	0.009	0.193	0.794	91
GO_PEPTIDE_METABOLIC_PROCESS	27	0.77	1.53	0.008	0.208	0.854	88

Table A.6 Top 10 GSEA biological processes gene set results for cluster 1 (*MMP1* cluster)

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
ONDER_CDH1_SIGNALING_VIA_CTNNB1	15	0.85	1.50	0.001	0.168	0.159	31
LENAOUR_DENDRITIC_CELL_MATURATION_DN	21	0.79	1.44	0.001	0.348	0.523	23
OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_UP	36	0.76	1.44	0	0.252	0.556	80
SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_DN	24	0.78	1.44	0	0.200	0.573	65
BERENJENO_TRANSFORMED_BY_RHOA_UP	33	0.76	1.42	0.001	0.211	0.665	95
REACTOME_GPCR_DOWNSTREAM_SIGNALING	15	0.80	1.41	0.003	0.220	0.738	32
RUTELLA_RESPONSE_TO_HGF_VS_CSF2RB_AND_IL4_UP	39	0.74	1.41	0	0.192	0.742	75
RUTELLA_RESPONSE_TO_CSF2RB_AND_IL4_DN	29	0.75	1.41	0	0.180	0.760	75
NAGASHIMA_EGF_SIGNALING_UP	25	0.77	1.41	0	0.170	0.786	95
AMIT_EGF_RESPONSE_40_HELA	16	0.78	1.40	0.009	0.184	0.849	106

Table A.7 Top 10 GSEA curated gene set results for cluster 3 (*IGF1* cluster 1)

# Appendix A

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
GO_NEGATIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	17	0.81	1.45	0.002	0.350	0.302	42
GO_CIRCULATORY_SYSTEM_PROCESS	20	0.80	1.45	0.000	0.179	0.307	40
GO_AMEBOIDAL_TYPE_CELL_MIGRATION	16	0.82	1.45	0.001	0.133	0.335	59
GO_CELLULAR_CATABOLIC_PROCESS	27	0.78	1.44	0.000	0.107	0.350	25
GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION	18	0.80	1.43	0.002	0.127	0.473	73
GO_NEGATIVE_REGULATION_OF_KINASE_ACTIVITY	15	0.80	1.42	0.007	0.119	0.519	42
GO_POSITIVE_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	21	0.78	1.41	0.001	0.134	0.609	40
GO_MACROMOLECULE_CATABOLIC_PROCESS	27	0.76	1.41	0.008	0.119	0.616	17
GO_G_PROTEIN_COUPLED_RECEPTOR_SIGNALING_PATHWAY	22	0.76	1.40	0.006	0.134	0.728	76
GO_REGULATION_OF_SYSTEM_PROCESS	21	0.74	1.39	0.003	0.119	0.749	40

Table A.8 Top 10 GSEA biological processes gene sets for cluster 3 (*IGF1* cluster 1)

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
BROWNE_HCMV_INFECTION_16HR_UP	23	0.92	1.72	0	0	0	62
SEKI_INFLAMMATORY_RESPONSE_LPS_UP	33	0.87	1.72	0	4.53x10 <sup>-4</sup>	0.001	135
PHONG_TNF_RESPONSE_VIA_P38_PARTIAL	56	0.84	1.71	0	3.02x10 <sup>-4</sup>	0.001	122
ALTEMEIER_RESPONSE_TO_LPS_WITH_MECHANICAL_VENTILATION	41	0.84	1.70	0	4.49x10 <sup>-4</sup>	0.002	202
ZHOU_INFLAMMATORY_RESPONSE_LIVE_UP	97	0.80	1.70	0	3.59x10 <sup>-4</sup>	0.002	148
GHANDHI_DIRECT_IRRADIATION_UP	33	0.85	1.68	0	0.001	0.008	112
OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_UP	69	0.81	1.68	0	0.001	0.008	134
CHEN_HOXA5_TARGETS_9HR_UP	46	0.82	1.68	0	0.001	0.010	201
NAGASHIMA_NRG1_SIGNALING_UP	67	0.81	1.67	0	0.001	0.012	205
BUYTAERT_PHOTODYNAMIC_THERAPY_STRESS_UP	132	0.77	1.66	0	0.002	0.025	207

Table A.9 Top 10 GSEA curated gene set results for cluster 5 (*IL6* cluster)

# Appendix A

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION	27	0.87	1.68	0	0.013	0.013	50
GO_POSITIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	97	0.78	1.66	0	0.017	0.034	158
GO_POSITIVE_REGULATION_OF_DNA_METABOLIC_PROCESS	17	0.88	1.63	0	0.025	0.025	90
GO_POSITIVE_REGULATION_OF_LEUKOCYTE_MIGRATION	24	0.86	1.63	0	0.028	0.028	92
GO_CELLULAR_RESPONSE_TO_BIOTIC_STIMULUS	28	0.83	1.62	0	0.025	0.025	113
GO_INFLAMMATORY_RESPONSE	61	0.78	1.62	0	0.026	0.026	113
GO_REGULATION_OF_LEUKOCYTE_MIGRATION	31	0.82	1.61	0	0.030	0.030	179
GO_REGULATION_OF_PROTEIN_LOCALIZATION_TO_NUCLEUS	27	0.84	1.61	0	0.027	0.027	92
GO_RESPONSE_TO_INTERLEUKIN_1	25	0.83	1.61	0.001	0.026	0.026	113
GO_RESPONSE_TO_MOLECULE_OF_BACTERIAL_ORIGIN	52	0.78	1.60	1	0.025	0.025	113

Table A.10 Top 10 GSEA biological process gene sets for cluster 5 (*IL6* cluster)

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
WELCSH_BRCA1_TARGETS_UP	18	0.94	1.44	0	0.203	0.224	28
AMIT_EGF_RESPONSE_480_HELA	21	0.91	1.42	0.001	0.152	0.318	20
TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_8D_UP	19	0.91	1.41	0.001	0.144	0.417	20
GRAESSMANN_RESPONSE_TO_MC_AND_SERUM_DEPRIVATION_UP	16	0.91	1.40	0.006	0.147	0.522	31
FARMER_BREAST_CANCER_BASAL_VS_LUTRAL	19	0.89	1.39	0.006	0.189	0.710	14
MARTORIATI_MDM4_TARGETS_FETAL_LIVER_DN	22	0.88	1.38	0.003	0.160	0.716	17
BOQUEST_STEM_CELL_UP	40	0.83	1.38	0	0.179	0.809	39
RUTELLA_RESPONSE_TO_HGF_VS-CSF2RB_AND_IL4_UP	35	0.84	1.37	0.004	0.189	0.855	15
MARTINEZ_RB1_TARGETS_DN	26	0.85	1.37	0.003	0.178	0.876	13
HATADA_METHYLATED_IN_LUNG_CANCER_UP	22	0.86	1.36	0.003	0.200	0.907	24

Table A.11 Top 10 GSEA curated gene set results for cluster 6 (*IGF1* cluster 2)

# Appendix A

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	19	0.94	1.45	0.001	0.100	0.092	13
GO_POSITIVE_REGULATION_OF_NEURON_DIFFERENTIATION	16	0.95	1.45	0	0.057	0.104	12
GO_REGULATION_OF_VASCULATURE_DEVELOPMENT	23	0.91	1.45	0	0.039	0.106	27
GO_REGULATION_OF_DEVELOPMENTAL_GROWTH	17	0.92	1.43	0.001	0.046	0.166	15
GO_CELLULAR_RESPONSE_TO_ACID_CHEMICAL	15	0.93	1.42	0.003	0.052	0.221	31
GO_REGULATION_OF_NEURON_DIFFERENTIATION	27	0.87	1.41	0.003	0.066	0.316	21
GO_REGULATION_OF_CELL_DEVELOPMENT	43	0.85	1.39	0.001	0.100	0.493	51
GO_POSITIVE_REGULATION_OF_GROWTH	16	0.90	1.39	0.005	0.101	0.546	15
GO_NEGATIVE_REGULATION_OF_CELL_PROLIFERATION	41	0.84	1.39	0.003	0.101	0.599	9
GO_REGULATION_OF_ION_TRANSPORT	19	0.88	1.38	0.007	0.099	0.629	13

Table A.12 Top 10 GSEA biological processes gene sets for cluster 6 (*IGF1* cluster 2)



NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
VECCHI_GASTRIC_CANCER_EARLY_DN	39	0.85	1.62	0	0.033	0.044	83
DELYS_THYROID_CANCER_DN	41	0.85	1.60	0	0.034	0.088	84
SABATES_COLORECTAL_ADENOMA_DN	27	0.88	1.59	0	0.042	0.159	46
SMID_BREAST_CANCER_NORMAL_LIKE_UP	61	0.78	1.56	0	0.063	0.300	82
FARMER_BREAST_CANCER_APOCRINE_VS_BASAL	20	0.91	1.56	0	0.136	0.349	12
RIGGI_EWING_SARCOMA_PROGENITOR_UP	46	0.79	1.53	0.004	0.183	0.680	78
SMID_BREAST_CANCER_BASAL_DN	54	0.77	1.51	0.004	0.172	0.829	61
LE_EGR2_TARGETS_DN	16	0.89	1.51	0.002	0.225	0.852	28
NAKAYAMA_SOFT_TISSUE_TUMORS_PCA2_DN	22	0.86	1.49	0.006	0.240	0.940	24
BOQUEST_STEM_CELL_CULTURED_VS_FRESH_UP	96	0.72	1.48	0.002	0.274	0.965	85

Table A.13 Top 10 GSEA curated gene set results for cluster 7 (*NPNT* cluster)

# Appendix A

NAME	SIZE	ES	NES	NOM <i>p</i> -val	FDR <i>q</i> -val	FWER <i>p</i> -val	RANK AT MAX
GO_MUSCLE_SYSTEM_PROCESS	36	0.92	1.61	0	0.021	0.029	56
GO_REGULATION_OF_SYSTEM_PROCESS	42	0.89	1.59	0.001	0.019	0.051	56
GO_MUSCLE_CONTRACTION	27	0.92	1.54	0	0.082	0.282	56
GO_REGULATION_OF_ANATOMICAL_STRUCTURE_SIZE	35	0.87	1.52	0.006	0.118	0.470	56
GO_ION_TRANSMEMBRANE_TRANSPORT	32	0.88	1.50	0.005	0.114	0.604	73
GO_POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	32	0.86	1.49	0.005	0.159	0.732	53
GO_MUSCLE_CELL_DIFFERENTIATION	32	0.88	1.49	0.002	0.160	0.786	66
GO_TRANSMEMBRANE_TRANSPORT	46	0.82	1.47	0.011	0.195	0.889	75
GO_CATION_TRANSPORT	37	0.84	1.47	0.013	0.202	0.924	73
GO_REGULATION_OF_BLOOD_CIRCULATION	21	0.90	1.47	0.067	0.211	0.943	54

Table A.14 Top 10 GSEA biological processes gene sets for cluster 8 (*MYH11* cluster)

## A.5 Gene sets used for GSEA identified at literature review

A literature review was performed to identify gene sets upregulated by stromal cells in disease states (*e.g.* fibrosis) or in response to stimulus (*e.g.* TGF- $\beta$ ).

A summary of these is given in Table A.15.

Gene set	Description
GSE107677 IPF vs. normal <sup>279</sup>	Array profiling of cultured fibroblasts from IPF and normal lungs
GSE17978 IPF vs. normal <sup>227</sup>	Array profiling of cultured non-fibroblasts from IPF and normal lungs
GSE66616 CAF vs. NOF <sup>161</sup>	Array profiling of cultured NSCLC fibroblasts
GSE22863 tumour stroma vs. normal <sup>229</sup>	Array profiling of laser captured tumour stroma and normal lung
GSE10547 IR vs. control <sup>280</sup>	Array profiling of irradiated lung cell lines
GSE81850 pancreas CAF vs. NOF 2D <sup>281</sup>	Array profiling of bioengineered pancreatic microtissues
GSE47616 IFNB skin vs. non-treated <sup>282</sup>	Array profiling of interferon-beta-treated cultured primary skin fibroblasts
GSE79621 TGFB skin vs. untreated <sup>283</sup>	Array profiling of TGF- $\beta$ -treated cultured skin fibroblasts
GSE64192 TGFB colon vs. untreated <sup>279</sup>	Array profiling of primary human colon fibroblasts treated with TGF- $\beta$
GSE12493 scleroderma vs. CTL <sup>284</sup>	Array profiling of primary dermal fibroblasts from scleroderma and control patients
GSE68685 IL1B gingiva vs. CTL <sup>285</sup>	Array profiling of isolated gingival fibroblasts treated with IL-1 $\beta$
Erez pro-inflammatory <sup>29</sup>	Expression profiling of murine skin squamous dysplasia vs. control
GSE3920 IFNa vs. INFY treated <sup>286</sup>	Array profiling of fibroblasts treated with interferon alpha and gamma
GSE3920 Untreated vs. IFNa <sup>286</sup>	Array profiling of fibroblasts treated with interferon alpha and untreated controls
GSE3920 Untreated vs. INFY <sup>286</sup>	Array profiling of fibroblasts treated with interferon gamma and untreated controls
Mellone IR <sup>83</sup>	Bulk RNA profiling of irradiated foetal human skin fibroblasts
Mellone TGFB treated <sup>83</sup>	Bulk RNA profiling of TGF- $\beta$ -treated foetal human skin fibroblasts
Mellone TGFB upregulated <sup>83</sup>	Bulk RNA profiling of TGF- $\beta$ -treated foetal human skin fibroblasts
Navab cell line CAF vs. cell line NOF <sup>229</sup>	Array profiling of matched NSCLC CAF and NOF cell lines
Navab primary CAF vs. NOF <sup>229</sup>	Array profiling of primary NSCLC CAF and NOF
Navab CAF NSCLC prognostic gene signature <sup>229</sup>	Signature identified at array profiling of NSCLC NOF and CAF
OSCC genetically unstable vs. normal <sup>240</sup>	Array profiling of CAF from genetically unstable OSCC vs. normal fibroblasts
OSCC genetically unstable vs. stable <sup>240</sup>	Array profiling of CAF from genetically unstable OSCC vs. genetically stable OSCC

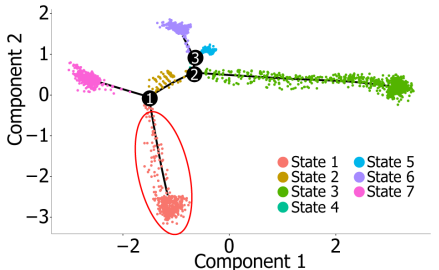
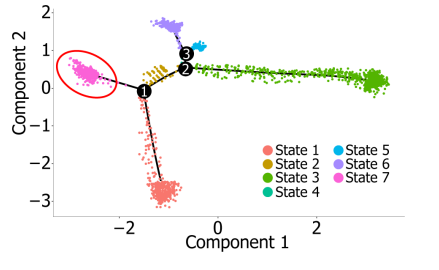
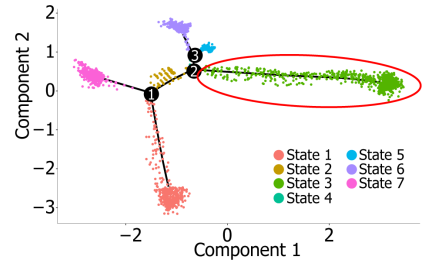
## Appendix A

Gene set	Description
Öhlund MF <sup>120</sup>	Bulk RNA sequencing of cultured “myofibroblastic” pancreatic CAF
Öhlund iCAF <sup>120</sup>	Bulk RNA sequencing of cultured “inflammatory” pancreatic CAF
Lambrechts_Fibroblast 1 <sup>127</sup>	DEGs for fibroblast cluster 1 in a NSCLC scRNA-seq dataset
Lambrechts_Fibroblast 2 <sup>127</sup>	DEGs for fibroblast cluster 2 in a NSCLC scRNA-seq dataset
Lambrechts_Fibroblast 4 <sup>127</sup>	DEGs for fibroblast cluster 4 in a NSCLC scRNA-seq dataset
Lambrechts_Fibroblast 5 <sup>127</sup>	DEGs for fibroblast cluster 5 in a NSCLC scRNA-seq dataset
Lambrechts_Fibroblast 6 <sup>127</sup>	DEGs for fibroblast cluster 6 in a NSCLC scRNA-seq dataset
Lambrechts_Fibroblast 7 <sup>127</sup>	DEGs for fibroblast cluster 7 in a NSCLC scRNA-seq dataset
Puram_CAF1 <sup>126</sup>	DEGs for CAF subpopulation 1 in a HNSCC scRNA-seq dataset
Puram_CAF2 <sup>126</sup>	DEGs for CAF subpopulation 2 in a HNSCC scRNA-seq dataset
Tirosh_CAF <sup>125</sup>	DEGs for CAF in a melanoma scRNA-seq dataset

Table A.15 Fibroblast expression programs identified at literature review. DEGs, differentially expressed genes; HNSCC, head and neck squamous carcinoma; OSCC, oral squamous cell carcinoma

## A.6 Gene set enrichment results for trajectory analysis States

Differential gene expression and trajectory analysis were performed as described in section 2.8 The top ten gene set enrichment analysis for each identified group are in Table A.16. CP/CGP, canonical pathways/chemical and genetic perturbations; BP, biological processes; EP, expression programs

	Trajectory 1	Trajectory 2	Trajectory 3
			
CP/CGP	N/A	N/A	DODD_NASOPHARYNGEAL_CARCCINOMA_DN
BP	N/A	N/A	N/A
EP	N/A	N/A	PURAM_CAF1 GSE17978 IPF VS. NORMAL

# Appendix A

	Trajectory 4	Trajectory 5
CP/CGP	N/A	N/A
BP	GO_SINGLE_ORGANISM_CATABOLIC_PROCESS	GO_REGULATION_OF_RAS_PROTEIN_SIGNAL_TRANSDUCTION
EP	N/A	GSE17978 IPF VS. NORMAL MELLONE TGFB UPREGULATED

Table A.16 Gene set enrichment results for *ex vivo* stromal cell Monocle trajectories. Gene sets analysed are those upregulated in the encircled population relative to State 2 (pseudotime 0). Results shown are significant at FDR  $q < 0.2$

## A.7 Transcription factor enrichment results for trajectory analysis States

Assessment of transcription factor expression was performed using the enrichR tool<sup>223</sup>. The top ten results by significance are listed in Tables A.17-A.20.

Pathway	<i>P</i> adj	NES
WRNIP1	0.08788707	2.087929
ASH2L	0.08788707	2.266078
KRT7	0.08788707	2.094949
CBX3	0.08788707	2.445000
PPARGC1A	0.08788707	-1.943834
TCF3	0.08788707	2.172725
FLI1	0.08788707	-2.049138
POU5F1	0.08788707	-1.786912
RBPJL	0.08788707	-2.031737
MYCN	0.13601128	1.939251
YAP1	0.13601128	1.953101
ETS1	0.17500378	1.944080
IRF1	0.17500378	1.852675

Table A.17 Significant ( $p$  adj < 0.2) enrichment results for State 3 in the Transcription\_Factor\_PPis database. *P* adj, adjusted  $p$  value; NES, normalised enrichment score

Pathway	<i>P</i> adj	NES
MYOD1_C2C12_mm9	0.006358111	2.159653
CREB1_H1-hESC_hg19	0.006358111	1.557745
SRF_MCF-7_hg19	0.006358111	1.533353
RFX5_IMR-90_hg19	0.006358111	1.837130
CTCF_keratinocyte_hg19	0.006358111	1.847638
FOXA1_HepG2_hg19	0.006358111	1.641582
STAT1_K562_hg19	0.006358111	1.698972
POLR2AphosphoS5_Panc1_hg19	0.006358111	1.863985
ELF1_SK-N-SH_hg19	0.006358111	1.858216
RELA_GM12892_hg19	0.006358111	1.963819
IRF3_GM12878_hg19	0.006358111	1.747078

Table A.18 Significant ( $p$  adj < 0.2) enrichment results for State 3 in the ENCODE\_TF\_ChIP-seq\_2015 database. *P* adj, adjusted  $p$  value; NES, normalised enrichment score



Pathway	<i>P</i> adj	NES
PHC1	0.08643081	-2.050836
PPARGC1A	0.08643081	-2.070627
AIRE	0.08643081	-1.893110
PML	0.08643081	-2.053267
USF1	0.08643081	-2.128759
SMARCA4	0.08643081	-2.136245
HNF1A	0.08643081	-2.131612
CCNT2	0.10897798	-1.922234
SIRT3	0.10897798	-1.823280
GADD45B	0.10897798	-1.958565
EMX2	0.10897798	-1.897324
THAP11	0.11361852	-1.768349
NFYA	0.11361852	-2.008986
NOTCH1	0.11361852	-1.769467
ILF3	0.11361852	-1.945654
ILF2	0.11361852	-1.777374
KDM5B	0.13095844	-1.739158
STAT5A	0.13095844	-1.799797
SP3	0.15773128	-1.745969
IGF1R	0.19290355	-1.738107

Table A.19 Significant ( $p$  adj < 0.2) enrichment results for State 5 in the ENCODE\_TF\_ChIP-seq\_2015 database. *P* adj, adjusted  $p$  value; NES, normalised enrichment score

Pathway	<i>P</i> adj	NES
STAT1_HeLa-S3_hg19	0.1235241	1.889329
CREB1_K562_hg19	0.1235241	1.824764
RELA_GM19099_hg19	0.1235241	1.762704
CTCF_GM10266_hg19	0.1235241	1.854769
SREBF2_GM12878_hg19	0.1235241	1.804687
RFX5_HepG2_hg19	0.1235241	1.877827
ABPA_A549_hg19	0.1235241	1.770017
BACH1_K562_hg19	0.1235241	1.774732
POLR2AphosphoS5_HL-60_hg19	0.1235241	1.844661
ZKSCAN1_HeLa-S3_hg19	0.1562620	1.723296

Table A1.20 Significant ( $p$  adj < 0.2) enrichment results for State 7 in the ENCODE\_TF\_ChIP-seq\_2015 database. *P* adj, adjusted  $p$  value; NES, normalised enrichment score

### A.8 Composition of TargetLung (Drop-seq) dataset by patient

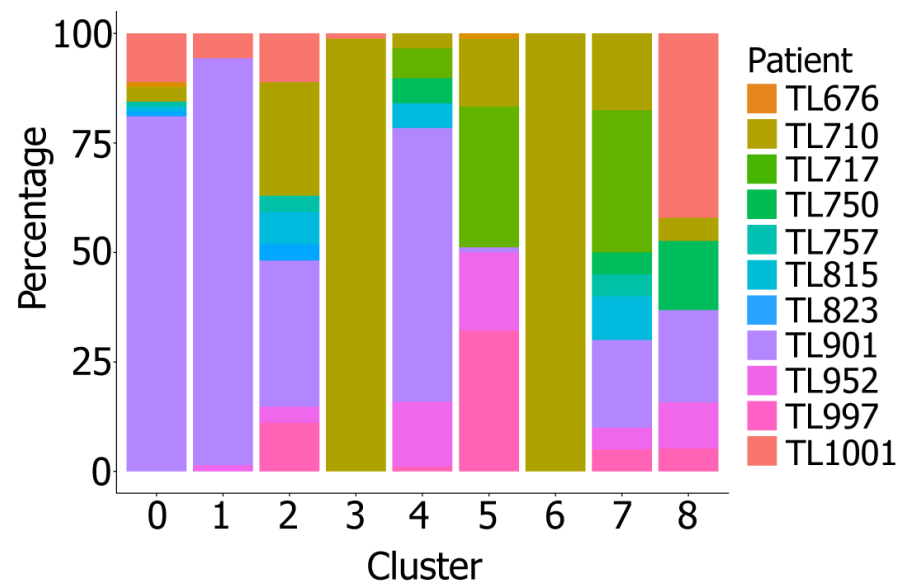


Figure A.2 Stacked barplot showing the composition of clusters in the TargetLung (Drop-seq) stromal cells by patient of origin

### A.9 Cells allocated to trajectory State by cluster

	<i>POSTN</i>	<i>MMP1</i>	<i>RGS5</i>	<i>IGF1-1</i>	<i>PI16</i>	<i>IL6</i>	<i>IGF1-2</i>	<i>NPNT</i>	<i>MYH11</i>
State 1	1	1	168	0	3	3	0	22	54
State 2	0	6	0	11	18	0	0	4	0
State 3	249	149	3	7	2	1	17	0	4
State 4	0	0	0	0	2	0	1	0	0
State 5	3	12	0	11	2	0	28	0	0
State 6	0	1	0	108	19	2	39	0	0
State 7	0	1	0	1	74	151	0	83	0

Table A.21 Number of cells allocated to each trajectory State by cluster

**A.10 Culture on plastic relative to Matrigel shows less striking changes in gene expression than does culture on plastic relative to 3D**

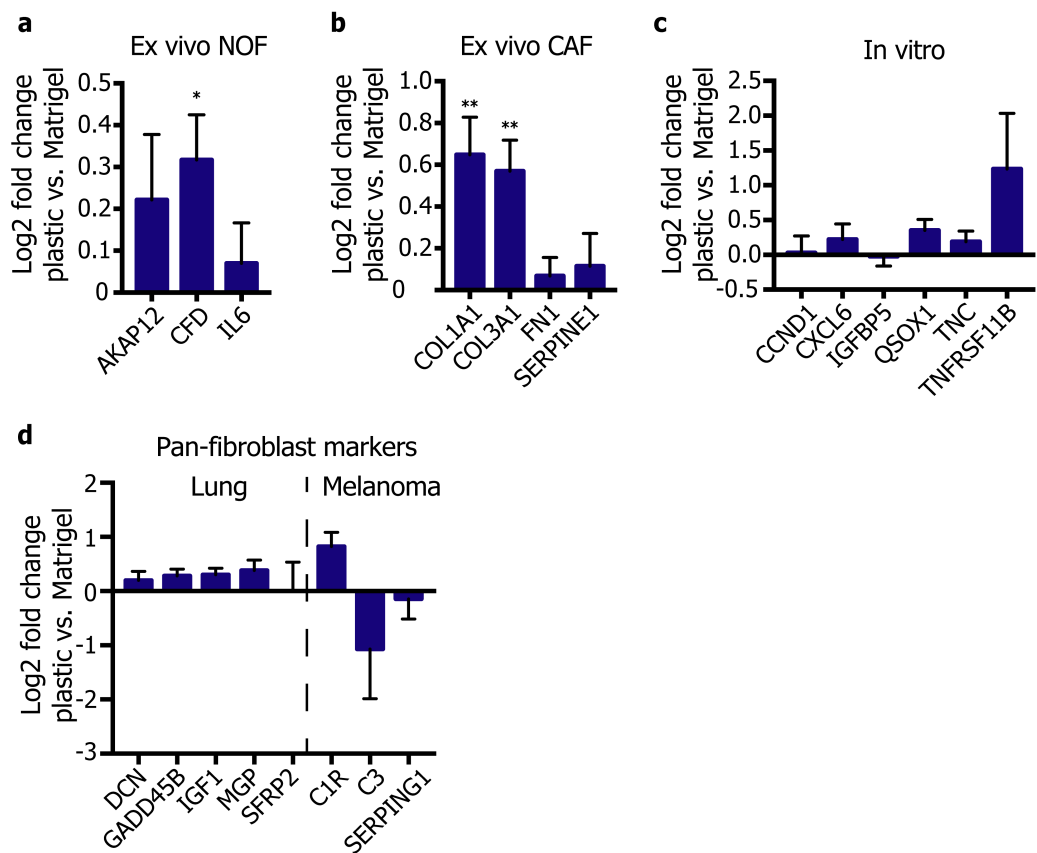


Figure A.3 Culture on plastic relative to Matrigel leads to upregulation of some genes differentially expressed by *ex vivo* fibroblasts. Bar charts showing fold change in expression for genes differentially expressed by: (a) *ex vivo* NOF, (b) *ex vivo* CAF, (c) *in vitro* fibroblasts, (d) all *ex vivo* fibroblast populations in both our dataset and melanoma<sup>125</sup>. Values are expressed as log2-fold change plastic:Matrigel. \* $p < 0.05$ , \*\*  $p < 0.01$ ; Welch's  $t$  test

### A.11 Cells allocated to cluster by culture condition using a random forest classifier

	CTL	TCP 10% FCS	IL-1 $\beta$	TGF- $\beta$
<b>POSTN</b>	59	253	22	161
<b>MMP1</b>	4	30	81	19
<b>RGS5</b>	1	5	74	1
<b>IGF1-1</b>	3	4	136	1
<b>PI16</b>	194	24	13	96
<b>IL6</b>	0	1	61	0
<b>IGF-2</b>	105	14	9	93
<b>NPNT</b>	27	10	2	25
<b>MYH11</b>	7	59	2	4

Table A.22 Allocation of differentiated cells to *ex vivo* clusters by the random forest classifier

### A.12 Treated cells allocated to trajectory States

	CTL	FCS	IL-1 $\beta$	TGF- $\beta$
<b>State 1</b>	47	2	398	2
<b>State 2</b>	19	0	0	2
<b>State 3</b>	11	23	0	50
<b>State 4</b>	5	368	1	28
<b>State 5</b>	318	7	1	317

Table A.23 Allocation of differentiated fibroblast to each differentiation trajectory

### A.13 Gene set enrichment analysis of treated cell trajectories

The results of the GSEA for the differentiated fibroblast trajectories are given in Table A.24.

	Trajectory 1	Trajectory 2
CP/CGP	WANG_SMARCE1_TARGETS_DN KEGG_CYTOKINE_CYTOKINE_RECEPTOR INTERACTION GHANDI_BYSTANDER_IRRADIATION_UP HINATA_NFKB_TARGETS_FIBROBLAST_UP REACTOME_G_ALPHA_I_SIGNALLING_EVENT S GRAHAM_CML_DIVIDING_VS_NORMAL_QUE ISCENT_DN ZHANG_RESPONSE_TO_IKK_INHIBITOR_AND _TNF_UP GAURNIER_PSMD4_TARGETS REACTOME_GPCR_LIGAND_BINDING BROWNE_HCMV_INFECTION_24_HR_UP	N/A
BP	GO_INFLAMMATORY_RESPONSE	N/A
EP	HINATA NFKB TARGETS UNTREATED VS. IFN $\gamma$ CHI VSMC	UNTREATED VS. IFN $\gamma$
	Trajectory 3	Trajectory 4
CP/CGP	N/A	N/A
BP	N/A	GO_CELLULAR_CATABOLIC_PROCESS
EP	N/A	N/A

Table A.24 Gene set enrichment results for differentiated stromal cell Monocle trajectories. Gene sets analysed are those upregulated in the encircled population relative to State 2 (pseudotime 0). Results shown are significant at FDR  $q < 0.2$ . CP/CGP, canonical pathways/chemical and genetic perturbations; BP, biological processes; EP, expression programs

## List of References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2018;68(6):394-424.
2. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clinic proceedings*. 2008;83(5):584-594.
3. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. *Translational lung cancer research*. 2016;5(3):288-300.
4. Cancer Research UK. 2019; <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer>. Accessed July, 2019.
5. Kelly K, Crowley J, Bunn PA, Jr., et al. Randomized phase III trial of paclitaxel plus carboplatin versus vinorelbine plus cisplatin in the treatment of patients with advanced non--small-cell lung cancer: a Southwest Oncology Group trial. *Journal of Clinical Oncology*. 2001;19(13):3210-3218.
6. Scagliotti GV, De Marinis F, Rinaldi M, et al. Phase III randomized trial comparing three platinum-based doublets in advanced non-small-cell lung cancer. *Journal of Clinical Oncology*. 2002;20(21):4285-4291.
7. Schiller JH, Harrington D, Belani CP, et al. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *New England Journal of Medicine*. 2002;346(2):92-98.
8. Fossella F, Pereira JR, von Pawel J, et al. Randomized, multinational, phase III study of docetaxel plus platinum combinations versus vinorelbine plus cisplatin for advanced non-small-cell lung cancer: the TAX 326 study group. *Journal of Clinical Oncology*. 2003;21(16):3016-3024.
9. Osmani L, Askin F, Gabrielson E, Li QK. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. *Seminars in Cancer Biology*. 2018;52:103-109.
10. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nature Communications*. 2015;6:8971.
11. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-674.
12. Junttila MR, de Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*. 2013;501(7467):346-354.
13. Lavin Y, Kobayashi S, Leader A, et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell*. 2017;169(4):750-765.e717.
14. Wen T, Aronow BJ, Rochman Y, et al. Single-cell RNA sequencing identifies inflammatory tissue T cells in eosinophilic esophagitis. *The Journal of Clinical Investigation*. 2019;129(5):2014-2028.
15. Azizi E, Carr AJ, Plitas G, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*. 2018;174(5):1293-1308.e1236.
16. Kalluri R. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer*. 2016;16(9):582-598.
17. Desmoulière A, Guyot C, Gabbiani G. The stroma reaction myofibroblast: a key player in the control of tumor cell behavior. *The International Journal of Developmental Biology*. 2004;48(5-6):509-517.
18. Kalluri R, Zeisberg M. Fibroblasts in cancer. *Nature Reviews Cancer*. 2006;6(5):392-401.
19. Roulis M, Flavell RA. Fibroblasts and myofibroblasts of the intestinal lamina propria in physiology and disease. *Differentiation*. 2016;92(3):116-131.

## List of References

20. Rupp C, Scherzer M, Rudisch A, et al. IGFBP7, a novel tumor stroma marker, with growth-promoting effects in colon cancer through a paracrine tumor–stroma interaction. *Oncogene*. 2014;34:815.
21. Servais C, Erez N. From sentinel cells to inflammatory culprits: cancer-associated fibroblasts in tumour-related inflammation. *The Journal of Pathology*. 2013;229(2):198-207.
22. Ishii G, Ochiai A, Neri S. Phenotypic and functional heterogeneity of cancer-associated fibroblast within the tumor microenvironment. *Advanced Drug Delivery Reviews*. 2016;99, Part B:186-196.
23. Peña C, Céspedes MV, Lindh MB, et al. STC1 Expression By Cancer-Associated Fibroblasts Drives Metastasis of Colorectal Cancer. *Cancer Research*. 2013;73(4):1287.
24. Quante M, Tu SP, Tomita H, et al. Bone marrow-derived myofibroblasts contribute to the mesenchymal stem cell niche and promote tumor growth. *Cancer cell*. 2011;19(2):257-272.
25. Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. *Nature*. 2004;432(7015):332-337.
26. Tao L, Huang G, Song H, Chen Y, Chen L. Cancer associated fibroblasts: An essential role in the tumor microenvironment. *Oncology Letters*. 2017;14(3):2611-2620.
27. Orimo A, Weinberg RA. Heterogeneity of stromal fibroblasts in tumors. *Journal of Cancer Biology and Therapeutics*. 2007;6.
28. Yazdani S, Bansal R, Prakash J. Drug targeting to myofibroblasts: Implications for fibrosis and cancer. *Advanced Drug Delivery Reviews*. 2017.
29. Erez N, Truitt M, Olson P, Arron ST, Hanahan D. Cancer-Associated Fibroblasts Are Activated in Incipient Neoplasia to Orchestrate Tumor-Promoting Inflammation in an NF-kappaB-Dependent Manner. *Cancer Cell*. 2010;17(2):135-147.
30. Nazareth MR, Broderick L, Simpson-Abelson MR, Kelleher RJ, Yokota SJ, Bankert RB. Characterization of Human Lung Tumor-Associated Fibroblasts and Their Ability to Modulate the Activation of Tumor-Associated T Cells. *The Journal of Immunology*. 2007;178(9):5552.
31. Narra K, Mullins SR, Lee HO, et al. Phase II trial of single agent Val-boroPro (Talabostat) inhibiting Fibroblast Activation Protein in patients with metastatic colorectal cancer. *Journal of Cancer Biology and Therapeutics*. 2007;6(11):1691-1699.
32. Hofheinz RD, al-Batran SE, Hartmann F, et al. Stromal Antigen Targeting by a Humanised Monoclonal Antibody: An Early Phase II Trial of Sibrotuzumab in Patients with Metastatic Colorectal Cancer. *Oncology Research and Treatment*. 2003;26(1):44-48.
33. Anderberg C, Pietras K. On the origin of cancer-associated fibroblasts. *Cell Cycle*. 2009;8(10):1461-1465.
34. Sugimoto H, Mundel TM, Kieran MW, Kalluri R. Identification of fibroblast heterogeneity in the tumor microenvironment. *Journal of Cancer Biology and Therapeutics*. 2006;5:1640-1646.
35. Witowski J, Kawka E, Rudolf A, Jörres A. New developments in peritoneal fibroblast biology: implications for inflammation and fibrosis in peritoneal dialysis. *BioMed Research International*. 2015;2015:134708.
36. Herrera M, Islam AB, Herrera A, et al. Functional heterogeneity of cancer-associated fibroblasts from human colon tumors shows specific prognostic gene expression signature. *Clinical Cancer Research*. 2013;19(21):5914-5926.
37. Hanley CJ, Noble F, Ward M, et al. A subset of myofibroblastic cancer-associated fibroblasts regulate collagen fiber elongation, which is prognostic in multiple cancers. *Oncotarget*. 2016;7(5):6159-6174.
38. Öhlund D, Elyada E, Tuveson D. Fibroblast heterogeneity in the cancer wound. *The Journal of Experimental Medicine*. 2014;211(8):1503-1523.
39. Cortez E, Roswall P, Pietras K. Functional subsets of mesenchymal cell types in the tumor microenvironment. *Seminars in Cancer Biology*. 2014;25:3-9.



40. Anderberg C, Li H, Fredriksson L, et al. Paracrine Signaling by Platelet-Derived Growth Factor-CC Promotes Tumor Growth by Recruitment of Cancer-Associated Fibroblasts. *Cancer Research*. 2009;69(1):369-378.
41. Özdemir Berna C, Pentcheva-Hoang T, Carstens Julianne L, et al. Depletion of Carcinoma-Associated Fibroblasts and Fibrosis Induces Immunosuppression and Accelerates Pancreas Cancer with Reduced Survival. *Cancer Cell*. 2017;25(6):719-734.
42. Chen J, Lu H, Zhou W, et al. AURKA upregulation plays a role in fibroblast-reduced gefitinib sensitivity in the NSCLC cell line HCC827. *Oncology Reports*. 2015;33(4):1860-1866.
43. Bartling B, Hofmann HS, Silber RE, Simm A. Differential impact of fibroblasts on the efficient cell death of lung cancer cells induced by paclitaxel and cisplatin. *Cancer Biology and Therapeutics*. 2008;7(8):1250-1261.
44. Kilvaer TK, Khanehkenari MR, Hellevik T, et al. Cancer Associated Fibroblasts in Stage I-IIIa NSCLC: Prognostic Impact and Their Correlations with Tumor Molecular Markers. *PLoS One*. 2015;10(8):e0134965.
45. Liao Y, Ni Y, He R, Liu W, Du J. Clinical implications of fibroblast activation protein- $\alpha$  in non-small cell lung cancer after curative resection: a new predictor for prognosis. *Journal of Cancer Research and Clinical Oncology*. 2013;139(9):1523-1528.
46. Chen Y, Xing P, Zou L, Zhang Y, Li F, Lu X. High p-Smad2 expression in stromal fibroblasts predicts poor survival in patients with clinical stage I to IIIa non-small cell lung cancer. *World Journal of Surgical Oncology*. 2014;12:328.
47. Koumas L, Smith TJ, Feldon S, Blumberg N, Phipps RP. Thy-1 expression in human fibroblast subsets defines myofibroblastic or lipofibroblastic phenotypes. *The American Journal of Pathology*. 2003;163(4):1291-1300.
48. Sriram G, Bigliardi PL, Bigliardi-Qi M. Fibroblast heterogeneity and its implications for engineering organotypic skin models in vitro. *European Journal of Cell Biology*. 2015;94(11):483-512.
49. Sorrell JM, Caplan AI. Fibroblast heterogeneity: more than skin deep. *Journal of Cell Science*. 2004;117(5):667-675.
50. Driskell RR, Watt FM. Understanding fibroblast heterogeneity in the skin. *Trends in Cell Biology*. 2015;25(2):92-99.
51. Hiraoka C, Toki F, Shiraishi K, et al. Two clonal types of human skin fibroblasts with different potentials for proliferation and tissue remodeling ability. *Journal of Dermatological Science*. 2016;82(2):84-94.
52. Ali-Bahar M, Bauer B, Tredget EE, Ghahary A. Dermal fibroblasts from different layers of human skin are heterogeneous in expression of collagenase and types I and III procollagen mRNA. *Wound Repair and Regeneration*. 2004;12(2):175-182.
53. Li H, Roos JC, Rose GE, Bailly M, Ezra DG. Eyelid and Sternum Fibroblasts Differ in Their Contraction Potential and Responses to Inflammatory Cytokines. *Plastic and Reconstructive Surgery - Global Open*. 2015;3(7):e448.
54. Kotaru C, Schoonover KJ, Trudeau JB, et al. Regional fibroblast heterogeneity in the lung: implications for remodeling. *American Journal of Respiratory and Critical Care Medicine*. 2006;173(11):1208-1215.
55. Preobrazhenska O, Wright JL, Churg A. Regional heterogeneity in murine lung fibroblasts from normal mice or mice exposed once to cigarette smoke. *PLoS One*. 2012;7(6):e39761.
56. Breen E, Falco VM, Absher M, Cutroneo KR. Subpopulations of rat lung fibroblasts with different amounts of type I and type III collagen mRNAs. *Journal of Biological Chemistry*. 1990;265(11):6286-6290.
57. Smith TJ, Koumas L, Gagnon A, et al. Orbital fibroblast heterogeneity may determine the clinical presentation of thyroid-associated ophthalmopathy. *The Journal of Clinical Endocrinology & Metabolism*. 2002;87(1):385-392.
58. Sempowski GD, Borrello MA, Blieden TM, Barth RK, Phipps RP. Fibroblast heterogeneity in the healing wound. *Wound Repair and Regeneration*. 1995;3(2):120-131.

## List of References

59. Sempowski GD, Derdak S, Phipps RP. Interleukin-4 and interferon- $\gamma$  discordantly regulate collagen biosynthesis by functionally distinct lung fibroblast subsets. *Journal of Cellular Physiology*. 1996;167(2):290-296.
60. Koumas L, King AE, Critchley HO, Kelly RW, Phipps RP. Fibroblast heterogeneity: existence of functionally distinct Thy 1(+) and Thy 1(-) human female reproductive tract fibroblasts. *The American Journal of Pathology*. 2001;159(3):925-935.
61. Alexeyenko A, Alkasalias T, Pavlova T, et al. Confrontation of fibroblasts with cancer cells in vitro: gene network analysis of transcriptome changes and differential capacity to inhibit tumor growth. *Journal of Experimental & Clinical Cancer Research*. 2015;34(1):62.
62. Zeisberg EM, Kalluri R. Origins of Cardiac Fibroblasts. *Circulation Research*. 2010;107(11):1304-1312.
63. Zeisberg EM, Tarnavski O, Zeisberg M, et al. Endothelial-to-mesenchymal transition contributes to cardiac fibrosis. *Nature Medicine*. 2007;13(8):952-961.
64. Kurahashi M, Nakano Y, Peri LE, Townsend JB, Ward SM, Sanders KM. A novel population of subepithelial platelet-derived growth factor receptor  $\alpha$ -positive cells in the mouse and human colon. *American Journal of Physiology - Gastrointestinal and Liver Physiology*. 2013;304(9):G823-834.
65. Higuchi Y, Kojima M, Ishii G, Aoyagi K, Sasaki H, Ochiai A. Gastrointestinal Fibroblasts Have Specialized, Diverse Transcriptional Phenotypes: A Comprehensive Gene Expression Analysis of Human Fibroblasts. *PLoS ONE*. 2015;10(6):e0129241.
66. Lindner D, Zietsch C, Becher PM, et al. Differential expression of matrix metalloproteases in human fibroblasts with different origins. *Biochemistry Research International*. 2012;2012:875742.
67. Nonaka M, Pawankar R, Fukumoto A, Yagi T. Heterogeneous response of nasal and lung fibroblasts to transforming growth factor-beta 1. *Clinical & Experimental Allergy*. 2008;38(5):812-821.
68. Chang HY, Chi JT, Dudoit S, et al. Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proceedings of the National Academy of Sciences of the United States of America*. 2002;99(20):12877-12882.
69. Tchou J, Kossenkova AV, Chang L, et al. Human breast cancer associated fibroblasts exhibit subtype specific gene expression profiles. *BMC Medical Genomics*. 2012;5(1):1-13.
70. Rønnevig-Jessen L, Petersen OW, Kotliansky VE, Bissell MJ. The origin of the myofibroblasts in breast cancer. Recapitulation of tumor environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells. *The Journal of Clinical Investigation*. 1995;95(2):859-873.
71. Strutz F, Zeisberg M. Renal Fibroblasts and Myofibroblasts in Chronic Kidney Disease. *Journal of the American Society of Nephrology*. 2006;17(11):2992-2998.
72. Allen M, Louise Jones J. Jekyll and Hyde: the role of the microenvironment on the progression of cancer. *The Journal of Pathology*. 2011;223(2):162-176.
73. Powell DW, Pinchuk IV, Saada JI, Chen X, Mifflin RC. Mesenchymal Cells of the Intestinal Lamina Propria. *Annual review of physiology*. 2011;73:213-237.
74. Shinagawa K, Kitadai Y, Tanaka M, et al. Mesenchymal stem cells enhance growth and metastasis of colon cancer. *International Journal of Cancer*. 2010;127(10):2323-2333.
75. Hanley CJ, Mellone M, Ford K, et al. Targeting the Myofibroblastic Cancer-Associated Fibroblast Phenotype Through Inhibition of NOX4. *JNCI: Journal of the National Cancer Institute*. 2018;110(1):109-120.
76. Cirri P, Chiarugi P. Cancer-associated-fibroblasts and tumour cells: a diabolic liaison driving cancer progression. *Cancer and Metastasis Reviews*. 2012;31:195-208.
77. De Wever O, Demetter P, Mareel M, Bracke M. Stromal myofibroblasts are drivers of invasive cancer growth. *International Journal of Cancer*. 2008;123(10):2229-2238.
78. Underwood TJ, Hayden AL, Derouet M, et al. Cancer-associated fibroblasts predict poor outcome and promote periostin-dependent invasion in oesophageal adenocarcinoma. *The Journal of Pathology*. 2015;235(3):466-477.

79. Kraman M, Bambrough PJ, Arnold JN, et al. Suppression of antitumor immunity by stromal cells expressing fibroblast activation protein- $\alpha$ . *Science*. 2010;330(6005):827-830.
80. Torres S, Bartolome RA, Mendes M, et al. Proteome profiling of cancer-associated fibroblasts identifies novel proinflammatory signatures and prognostic markers for colorectal cancer. *Clin Cancer Res*. 2013;19(21):6006-6019.
81. Brentnall TA. Arousal of cancer-associated stromal fibroblasts: palladin-activated fibroblasts promote tumor invasion. *Cell Adhesion & Migration*. 2012;6(6):488-494.
82. Orimo A, Gupta PB, Sgroi DC, et al. Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell*. 2005;121(3):335-348.
83. Mellone M, Hanley CJ, Thirdborough S, et al. Induction of fibroblast senescence generates a non-fibrogenic myofibroblast phenotype that differentially impacts on cancer prognosis. *Aging (Albany NY)*. 2017;9(1):114-131.
84. Du H, Chen D, Zhou Y, Han Z, Che G. Fibroblast phenotypes in different lung diseases. *Journal of Cardiothoracic Surgery*. 2014;9:147.
85. Herranz N, Gallage S, Mellone M, et al. mTOR regulates MAPKAPK2 translation to control the senescence-associated secretory phenotype. *Nature Cell Biology*. 2015;17(9):1205-1217.
86. Coppé JP, Desprez PY, Krtolica A, Campisi J. The senescence-associated secretory phenotype: the dark side of tumor suppression. *Annual Review of Pathology*. 2010;5:99-118.
87. Hassona Y, Cirillo N, Lim KP, et al. Progression of genotype-specific oral cancer leads to senescence of cancer-associated fibroblasts and is mediated by oxidative stress and TGF- $\beta$ . *Carcinogenesis*. 2013;34(6):1286-1295.
88. Laberge RM, Sun Y, Orjalo AV, et al. MTOR regulates the pro-tumorigenic senescence-associated secretory phenotype by promoting IL1A translation. *Nature Cell Biology*. 2015;17(8):1049-1061.
89. Alimbetov D, Davis T, Brook AJ, et al. Suppression of the senescence-associated secretory phenotype (SASP) in human fibroblasts using small molecule inhibitors of p38 MAP kinase and MK2. *Biogerontology*. 2016;17(2):305-315.
90. Grum-Schwensen B, Klingelhofer J, Berg CH, et al. Suppression of Tumor Development and Metastasis Formation in Mice Lacking the S100A4 Gene. *Cancer Research*. 2005;65(9):3772.
91. Okada H, Inoue T, Kanno Y, et al. Selective depletion of fibroblasts preserves morphology and the functional integrity of peritoneum in transgenic mice with peritoneal fibrosing syndrome<sup>11</sup>Jared Grantham, M.D., served as guest editor for this paper. *Kidney International*. 2003;64(5):1722-1732.
92. Grum-Schwensen B, Klingelhöfer J, Grigorian M, et al. Lung Metastasis Fails in MMTV-PyMT Oncomice Lacking S100A4 Due to a T-Cell Deficiency in Primary Tumors. *Cancer Research*. 2010;70(3):936.
93. O'Connell JT, Sugimoto H, Cooke VG, et al. VEGF-A and Tenascin-C produced by S100A4(+) stromal cells are important for metastatic colonization. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(38):16002-16007.
94. Grum-Schwensen B, Klingelhöfer J, Beck M, et al. S100A4-neutralizing antibody suppresses spontaneous tumor progression, pre-metastatic niche formation and alters T-cell polarization balance. *BMC Cancer*. 2015;15(1):44.
95. Zhang J, Chen L, Xiao M, Wang C, Qin Z. FSP1+ Fibroblasts Promote Skin Carcinogenesis by Maintaining MCP-1-Mediated Macrophage Infiltration and Chronic Inflammation. *The American Journal of Pathology*. 2011;178(1):382-390.
96. Levy MT, McCaughan GW, Abbott CA, et al. Fibroblast activation protein: a cell surface dipeptidyl peptidase and gelatinase expressed by stellate cells at the tissue remodelling interface in human cirrhosis. *Hepatology*. 1999;29.
97. Ariga N, Sato E, Ohuchi N, Nagura H, Ohtani H. Stromal expression of fibroblast activation protein/seprase, a cell membrane serine proteinase and gelatinase, is associated with

- longer survival in patients with invasive ductal carcinoma of breast. *International Journal of Cancer*. 2001;95(1):67-72.
98. Feig C, Jones JO, Kraman M, et al. Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2013;110(50):20212-20217.
99. Popple A, Durrant LG, Spendlove I, et al. The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. *British Journal of Cancer*. 2012;106(7):1306-1313.
100. Yang X, Lin Y, Shi Y, et al. FAP Promotes Immunosuppression by Cancer-Associated Fibroblasts in the Tumor Microenvironment via STAT3–CCL2 Signaling. *Cancer Research*. 2016;76(14):4124.
101. Singh S, Srivastava SK, Bhardwaj A, Owen LB, Singh AP. CXCL12–CXCR4 signalling axis confers gemcitabine resistance to pancreatic cancer cells: a novel target for therapy. *British Journal Of Cancer*. 2010;103(11):1671.
102. Pietras K, Pahler J, Bergers G, Hanahan D. Functions of paracrine PDGF signaling in the proangiogenic tumor stroma revealed by pharmacological targeting. *PLoS Medicine*. 2008;5.
103. Gerber DE, Gupta P, Dellinger MT, et al. Stromal platelet-derived growth factor receptor  $\alpha$  (PDGFR $\alpha$ ) provides a therapeutic target independent of tumor cell PDGFR $\alpha$  expression in lung cancer xenografts. *Molecular Cancer Therapeutics*. 2012;11(11):2473-2482.
104. Donnem T, Al-Saad S, Al-Shibli K, Andersen S, Busund LT, Bremnes RM. Prognostic impact of platelet-derived growth factors in non-small cell lung cancer tumor and stromal cells. *Journal of Thoracic Oncology*. 2008;3(9):963-970.
105. Hewitt KJ, Shamis Y, Knight E, et al. PDGFR $\beta$  expression and function in fibroblasts derived from pluripotent cells is linked to DNA demethylation. *Journal of Cell Science*. 2012;125(9):2276.
106. Kitadai Y, Sasaki T, Kuwai T, Nakamura T, Bucana CD, Fidler IJ. Targeting the Expression of Platelet-Derived Growth Factor Receptor by Reactive Stroma Inhibits Growth and Metastasis of Human Colon Carcinoma. *The American Journal of Pathology*. 2006;169(6):2054-2065.
107. Corvigno S, Wisman GB, Mezheyski A, et al. Markers of fibroblast-rich tumor stroma and perivascular cells in serous ovarian cancer: inter- and intra-patient heterogeneity and impact on survival. *Oncotarget*. 2016;7(14):18573-18584.
108. Kitano H, Kageyama S, Hewitt SM, et al. Podoplanin expression in cancerous stroma induces lymphangiogenesis and predicts lymphatic spread and patient survival. *Archives of Pathology & Laboratory Medicine*. 2010;134(10):1520-1527.
109. Yamanashi T, Nakanishi Y, Fujii G, et al. Podoplanin Expression Identified in Stromal Fibroblasts as a Favorable Prognostic Marker in Patients with Colorectal Carcinoma. *Oncology*. 2009;77(1):53-62.
110. Dumoff KL, Chu CS, Harris EE, et al. Low podoplanin expression in pretreatment biopsy material predicts poor prognosis in advanced-stage squamous cell carcinoma of the uterine cervix treated by primary radiation. *Modern Pathology*. 2006;19:708.
111. Takahashi A, Ishii G, Neri S, et al. Podoplanin-expressing cancer-associated fibroblasts inhibit small cell lung cancer growth. *Oncotarget*. 2015;6(11):9531-9541.
112. Yuan P, Temam S, El-Naggar A, et al. Overexpression of podoplanin in oral cancer and its association with poor clinical outcome. *Cancer*. 2006;107(3):563-569.
113. Ono S, Ishii G, Nagai K, et al. Podoplanin-Positive Cancer-Associated Fibroblasts Could Have Prognostic Value Independent of Cancer Cell Phenotype in Stage I Lung Squamous Cell Carcinoma: Usefulness of Combining Analysis of Both Cancer Cell Phenotype and Cancer-Associated Fibroblast Phenotype. *Chest*. 2013;143(4):963-970.
114. Ito M, Ishii G, Nagai K, Maeda R, Nakano Y, Ochiai A. Prognostic Impact of Cancer-Associated Stromal Cells in Patients With Stage I Lung Adenocarcinoma. *Chest*. 2012;142(1):151-158.

115. Koriyama H, Ishii G, Yoh K, et al. Presence of podoplanin-positive cancer-associated fibroblasts in surgically resected primary lung adenocarcinoma predicts a shorter progression-free survival period in patients with recurrences who received platinum-based chemotherapy. *Journal of Cancer Research and Clinical Oncology*. 2015;141(7):1163-1170.
116. Hoshino A, Ishii G, Ito T, et al. Podoplanin-Positive Fibroblasts Enhance Lung Adenocarcinoma Tumor Formation: Podoplanin in Fibroblast Functions for Tumor Progression. *Cancer Research*. 2011;71(14):4769.
117. Neri S, Hashimoto H, Kii H, et al. Cancer cell invasion driven by extracellular matrix remodeling is dependent on the properties of cancer-associated fibroblasts. *Journal of Cancer Research and Clinical Oncology*. 2016;142(2):437-446.
118. Yoshida T, Ishii G, Goto K, et al. Podoplanin-Positive Cancer-Associated Fibroblasts in the Tumor Microenvironment Induce Primary Resistance to EGFR-TKIs in Lung Adenocarcinoma with EGFR Mutation. *Clinical Cancer Research*. 2015;21(3):642.
119. Zhao H, Peehl Donna M. Tumor - promoting phenotype of CD90hi prostate cancer - associated fibroblasts. *The Prostate*. 2009;69(9):991-1000.
120. Öhlund D, Handly-Santana A, Biffi G, et al. Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. *The Journal of Experimental Medicine*. 2017;214(3):579-596.
121. Costa A, Kieffer Y, Scholer-Dahirel A, et al. Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer. *Cancer Cell*. 2018;33(3):463-479.e410.
122. Micke P, Kappert K, Ohshima M, et al. In Situ Identification of Genes Regulated Specifically in Fibroblasts of Human Basal Cell Carcinoma. *Journal of Investigative Dermatology*. 2007;127(6):1516-1523.
123. Bauer M, Su G, Casper C, He R, Rehrauer W, Friedl A. Heterogeneity of gene expression in stromal fibroblasts of human breast carcinomas and normal breast. *Oncogene*. 2010;29.
124. Peng Q, Zhao L, Hou Y, et al. Biological Characteristics and Genetic Heterogeneity between Carcinoma-Associated Fibroblasts and Their Paired Normal Fibroblasts in Human Breast Cancer. *PLoS ONE*. 2013;8(4):e60321.
125. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189-196.
126. Puram SV, Tirosh I, Parkhi AS, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*. 2017;171(7):1611-1624.
127. Lambrechts D, Wauters E, Boeckx B, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine*. 2018;28(8):1277-1289.
128. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*. 2014;9(1):171.
129. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*. 2015;58(4):610-620.
130. Ziegenhain C, Vieth B, Parekh S, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*. 2017;65(4):631-643.
131. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214.
132. Magella B, Adam M, Potter AS, et al. Cross-platform single cell analysis of kidney development shows stromal cells express Gdnf. *Developmental Biology*. 2018;434(1):36-47.
133. Kernfeld EM, Genga RMJ, Neherin K, Magaletta ME, Xu P, Maehr R. A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. *Immunity*. 2018;48(6):1258-1270.e1256.
134. Rittié L, Fisher GJ. Isolation and Culture of Skin Fibroblasts. In: Varga J, Brenner DA, Phan SH, eds. *Fibrosis Research: Methods and Protocols*. Totowa, NJ: Humana Press; 2005:83-98.

# List of References

135. Grange C, Letourneau J, Forget MA, et al. Phenotypic characterization and functional analysis of human tumor immune infiltration after mechanical and enzymatic disaggregation. *Journal of Immunological Methods*. 2011;372(1-2):119-126.
136. Quatromoni JG, Singhal S, Bhojnagarwala P, Hancock WW, Albelda SM, Eruslanov E. An optimized disaggregation method for human lung tumors that preserves the phenotype and function of the immune cells. *Journal of Leukocyte Biology*. 2015;97(1):201-209.
137. Gray DHD, Chidgey AP, Boyd RL. Analysis of thymic stromal cell populations using flow cytometry. *Journal of Immunological Methods*. 2002;260(1):15-28.
138. van den Brink SC, Sage F, Vértessy Á, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*. 2017;14:935-936.
139. Roberts EW, Deonaraine A, Jones JO, et al. Depletion of stromal cells expressing fibroblast activation protein- $\alpha$  from skeletal muscle and bone marrow results in cachexia and anemia. *The Journal of Experimental Medicine*. 2013;210(6):1137.
140. Tran E, Chinnasamy D, Yu Z, et al. Immune targeting of fibroblast activation protein triggers recognition of multipotent bone marrow stromal cells and cachexia. *The Journal of Experimental Medicine*. 2013;210(6):1125-1135.
141. Santos AM, Jung J, Aziz N, Kissil JL, Pure E. Targeting fibroblast activation protein inhibits tumor stromagenesis and growth in mice. *Journal of Clinical Investigation*. 2009;119(12):3613-3625.
142. Liao D, Luo Y, Markowitz D, Xiang R, Reisfeld RA. Cancer Associated Fibroblasts Promote Tumor Growth and Metastasis by Modulating the Tumor Immune Microenvironment in a 4T1 Murine Breast Cancer Model. *PLoS ONE*. 2009;4(11):e7965.
143. Olive KP, Jacobetz MA, Davidson CJ, et al. Inhibition of Hedgehog Signaling Enhances Delivery of Chemotherapy in a Mouse Model of Pancreatic Cancer. *Science*. 2009;324(5933):1457.
144. Loeffler M, Krüger JA, Niethammer AG, Reisfeld RA. Targeting tumor-associated fibroblasts improves cancer chemotherapy by increasing intratumoral drug uptake. *The Journal of Clinical Investigation*. 2006;116(7):1955-1962.
145. Mertens JC, Fingas CD, Christensen JD, et al. Therapeutic Effects of Deleting Cancer-Associated Fibroblasts in Cholangiocarcinoma. *Cancer Research*. 2013;73(2):897.
146. Wang W-Q, Liu L, Xu H-X, et al. Intratumoral  $\alpha$ -SMA Enhances the Prognostic Potency of CD34 Associated with Maintenance of Microvessel Integrity in Hepatocellular Carcinoma and Pancreatic Cancer. *PLOS ONE*. 2013;8(8):e71189.
147. Bailey JM, Swanson BJ, Hamada T, et al. Sonic Hedgehog Promotes Desmoplasia in Pancreatic Cancer. *Clinical Cancer Research*. 2008;14(19):5995.
148. Kawase A, Ishii G, Nagai K, et al. Podoplanin expression by cancer associated fibroblasts predicts poor prognosis of lung adenocarcinoma. *International Journal of Cancer*. 2008;123(5):1053-1059.
149. Leinonen T, Pirinen R, Bohm J, Johansson R, Kosma VM. Increased expression of matrix metalloproteinase-2 (MMP-2) predicts tumour recurrence and unfavourable outcome in non-small cell lung cancer. *Histology and Histopathology*. 2008;23(6):693-700.
150. Ishikawa S, Takenaka K, Yanagihara K, et al. Matrix metalloproteinase-2 status in stromal fibroblasts, not in tumor cells, is a significant prognostic factor in non-small-cell lung cancer. *Clinical Cancer Research*. 2004;10(19):6579-6585.
151. Chen Y, Zou L, Zhang Y, et al. Transforming growth factor-beta1 and alpha-smooth muscle actin in stromal fibroblasts are associated with a poor prognosis in patients with clinical stage I-IIIa nonsmall cell lung cancer after curative resection. *Tumour Biology*. 2014;35(7):6707-6713.
152. Meijer TW, Schuurbijs OC, Kaanders JH, et al. Differences in metabolism between adenocarcinoma and squamous cell non-small cell lung carcinomas: spatial distribution and prognostic value of GLUT1 and MCT4. *Lung Cancer*. 2012;76(3):316-323.
153. Pirinen R, Leinonen T, Bohm J, et al. Versican in nonsmall cell lung cancer: relation to hyaluronan, clinicopathologic factors, and prognosis. *Human Pathology*. 2005;36(1):44-50.

154. Edlund K, Lindskog C, Saito A, et al. CD99 is a novel prognostic stromal marker in non-small cell lung cancer. *International Journal of Cancer*. 2012;131(10):2264-2273.
155. Eberlein C, Rooney C, Ross SJ, Farren M, Weir HM, Barry ST. E-Cadherin and EpCAM expression by NSCLC tumour cells associate with normal fibroblast activation through a pathway initiated by integrin  $\alpha\text{v}\beta 6$  and maintained through TGF $\beta$  signalling. *Oncogene*. 2015;34(6):704-716.
156. Choe C, Shin YS, Kim C, et al. Crosstalk with cancer-associated fibroblasts induces resistance of non-small cell lung cancer cells to epidermal growth factor receptor tyrosine kinase inhibition. *OncoTargets and Therapy*. 2015;8:3665-3678.
157. Wang W, Li Q, Yamada T, et al. Crosstalk to stromal fibroblasts induces resistance of lung cancer to epidermal growth factor receptor tyrosine kinase inhibitors. *Clinical Cancer Research*. 2009;15(21):6630-6638.
158. Yamauchi Y, Izumi Y, Asakura K, et al. Lewis lung carcinoma progression is facilitated by TIG-3 fibroblast cells. *Anticancer Research*. 2013;33(9):3791-3798.
159. Choe C, Shin YS, Kim SH, et al. Tumor-stromal interactions with direct cell contacts enhance motility of non-small cell lung cancer cells through the hedgehog signaling pathway. *Anticancer Research*. 2013;33(9):3715-3723.
160. Anderson IC, Mari SE, Broderick RJ, Mari BP, Shipp MA. The Angiogenic Factor Interleukin 8 Is Induced in Non-Small Cell Lung Cancer/Pulmonary Fibroblast Cocultures. *Cancer Research*. 2000;60(2):269.
161. Rudisch A, Dewhurst MR, Horga LG, et al. High EMT Signature Score of Invasive Non-Small Cell Lung Cancer (NSCLC) Cells Correlates with NF $\kappa$ B Driven Colony-Stimulating Factor 2 (CSF2/GM-CSF) Secretion by Neighboring Stromal Fibroblasts. *PLoS One*. 2015;10(4):e0124283.
162. Hao J, Zeltz C, Pintilie M, et al. Characterization of Distinct Populations of Carcinoma-Associated Fibroblasts from Non-Small Cell Lung Carcinoma Reveals a Role for ST8SIA2 in Cancer Cell Invasion. *Neoplasia*. 2019;21(5):482-493.
163. Kurtul N, Eroglu C, Unal D, et al. Prognostic value of SPARC expression in unresectable NSCLC treated with concurrent chemoradiotherapy. *Asian Pacific Journal of Cancer Prevention*. 2014;15(20):8911-8916.
164. <http://stemcells.rpi.edu/wp-content/uploads/Matrigel-Plate-Coating-v2.pdf>.
165. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. *Current protocols in bioinformatics* 2015;51:11.14.11-11.14.19.
166. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*. 2015;33(5):495-502.
167. Ilicic T, Kim JK, Kolodziejczyk AA, et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*. 2016;17(1):29.
168. Removal of Dead Cells from Single Cell Suspensions Improves Performance for 10X Genomics Single Cell Applications.  
[https://assets.ctfassets.net/an68im79xiti/4tVumiyINGgAeoCg8SiWGG/1cf0888200d668142612c8d3f3679cf4/CG000130\\_10x\\_Technical\\_Note\\_DeadCell\\_Removal\\_RevA.pdf](https://assets.ctfassets.net/an68im79xiti/4tVumiyINGgAeoCg8SiWGG/1cf0888200d668142612c8d3f3679cf4/CG000130_10x_Technical_Note_DeadCell_Removal_RevA.pdf).
169. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 2013;49(4):764-766.
170. Maechler M, Rousseeuw P., Struyf A., Hubert M., Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.9. 2019.
171. McDavid A, Finak G, Chattopadhyay PK, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*. 2013;29(4):461-467.
172. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*. 2009;37(Web Server issue):W305-W311.
173. Heng TS, Painter MW. The Immunological Genome Project: networks of gene expression in immune cells. *Nature Immunology*. 2008;9(10):1091-1094.

## List of References

174. Du Y, Guo M, Whitsett JA, Xu Y. 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax*. 2015;70(11):1092-1094.
175. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. 2018;36:411.
176. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(43):15545-15550.
177. Mootha VK, Lindgren CM, Eriksson K-F, et al. PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*. 2003;34(3):267-273.
178. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*. 2014;32:381.
179. Clarke J, Panwar B, Madrigal A, et al. Single-cell transcriptomic analysis of tissue-resident memory T cells in human lung cancer. *The Journal of Experimental Medicine*. 2019;216(9):2128-2149.
180. Barkauskas CE, Crouce MJ, Rackley CR, et al. Type 2 alveolar cells are stem cells in adult lung. *The Journal of Clinical Investigation*. 2013;123(7):3025-3036.
181. Mackay LS, Dodd S, Dougall IG, et al. Isolation and characterisation of human pulmonary microvascular endothelial cells from patients with severe emphysema. *Respiratory Research*. 2013;14(1):23-23.
182. Comhair SAA, Xu W, Mavrikakis L, Aldred MA, Asosingh K, Erzurum SC. Human Primary Lung Endothelial Cells in Culture. *American Journal of Respiratory Cell and Molecular Biology*. 2012;46(6):723-730.
183. Lurton J, Rose TM, Raghu G, Narayanan AS. Isolation of a Gene Product Expressed by a Subpopulation of Human Lung Fibroblasts by Differential Display. *American Journal of Respiratory Cell and Molecular Biology*. 1999;20(2):327-331.
184. Ganesan A-P, Clarke J, Wood O, et al. Tissue-resident memory features are linked to the magnitude of cytotoxic T cell responses in human lung cancer. *Nature Immunology*. 2017;18:940.
185. Holt PG, Robinson BW, Reid M, et al. Extraction of immune and inflammatory cells from human lung parenchyma: evaluation of an enzymatic digestion procedure. *Clinical and Experimental Immunology*. 1986;66(1):188-200.
186. Perrot I, Blanchard D, Freymond N, et al. Dendritic cells infiltrating human non-small cell lung cancer are blocked at immature stage. *The Journal of Immunology*. 2007;178(5):2763-2769.
187. Micke P, Ostman A. Tumour-stroma interaction: cancer-associated fibroblasts as novel targets in anti-cancer therapy? *Lung Cancer*. 2004;45(S2):S163-175.
188. Sanders YY, Kumbala P, Hagood JS. Enhanced Myofibroblastic Differentiation and Survival in Thy-1(-) Lung Fibroblasts. *American Journal of Respiratory Cell and Molecular Biology*. 2007;36(2):226-235.
189. Donovan JA, Koretzky GA. CD45 and the immune response. *Journal of the American Society of Nephrology*. 1993;4(4):976-985.
190. Trzpis M, McLaughlin PMJ, de Leij LMFH, Harmsen MC. Epithelial Cell Adhesion Molecule. *The American Journal of Pathology*. 2007;171(2):386-395.
191. Phan SH. Biology of fibroblasts and myofibroblasts. *Proceedings of the American Thoracic Society*. 2008;5(3):334-337.
192. Phipps RP, Penney DP, Keng P, et al. Characterization of two major populations of lung fibroblasts: distinguishing morphology and discordant display of Thy 1 and class II MHC. *American Journal of Respiratory Cell and Molecular Biology*. 1989;1(1):65-74.



193. Ellerström C, Strehl R, Noaksson K, Hyllner J, Semb H. Facilitated Expansion of Human Embryonic Stem Cells by Single-Cell Enzymatic Dissociation. *STEM CELLS*. 2007;25(7):1690-1696.
194. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*. 2013;86(11):471.
195. Buraschi S, Neill T, Goyal A, et al. Decorin causes autophagy in endothelial cells via Peg3. *Proceedings of the National Academy of Sciences of the United States of America*. 2013;110(28):E2582-E2591.
196. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2:18-22.
197. Paluszyńska A. Understanding random forests with randomForestExplainer. 2017; <https://cran.r-project.org/package=randomForestExplainer>.
198. Campbell J, Corbett S, Koga Y, Wang Z. celda: CELLular Latent Dirichlet Allocation. R package version 1.0.2. 2019.
199. Lun ATL, Riesenfeld S, Andrews T, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome biology*. 2019;20(1):63-63.
200. Costea DE, Hills A, Osman AH, et al. Identification of two distinct carcinoma-associated fibroblast subtypes with differential tumor-promoting abilities in oral squamous cell carcinoma. *Cancer Research*. 2013;73(13):3888-3901.
201. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(24):13790-13795.
202. Su Q, Igyártó BZ. Keratinocytes Share Gene Expression Fingerprint with Epidermal Langerhans Cells via mRNA Transfer. *Journal of Investigative Dermatology*. 2019;139(11):2313-2323.e2318.
203. Rajesh A, Wise L, Hibma M. The role of Langerhans cells in pathologies of the skin. *Immunology & Cell Biology*. 2019;97(8):700-713.
204. Radzikowska E. Pulmonary Langerhans' cell histiocytosis in adults. *Advances in Respiratory Medicine*. 2017;85(5):277-289.
205. Leśniak W, Słomnicki ŁP, Kuźnicki J. Epigenetic Control of the S100A6 (Calcyclin) Gene Expression. *Journal of Investigative Dermatology*. 2007;127(10):2307-2314.
206. Yu Y, Zhang C, Zhou G, et al. Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. *Genome research*. 2001;11(8):1392-1403.
207. Zhang S, Cui W. Sox2, a key factor in the regulation of pluripotency and neural differentiation. *World journal of stem cells*. 2014;6(3):305-311.
208. Nagase H, Visse R, Murphy G. Structure and function of matrix metalloproteinases and TIMPs. *Cardiovascular Research*. 2006;69(3):562-573.
209. Novinec M, Lenarcic B. Cathepsin K: a unique collagenolytic cysteine peptidase. *Biological chemistry*. 2013;394(9):1163-1179.
210. Tape Christopher J, Ling S, Dimitriadis M, et al. Oncogenic KRAS Regulates Tumor Cell Signaling via Stromal Reciprocation. *Cell*. 2016;165(4):910-920.
211. Ryu K-Y, Maehr R, Gilchrist CA, et al. The mouse polyubiquitin gene UbC is essential for fetal liver development, cell-cycle progression and stress tolerance. *The EMBO journal*. 2007;26(11):2693-2706.
212. Hammer M, Mages J, Dietrich H, et al. Dual specificity phosphatase 1 (DUSP1) regulates a subset of LPS-induced genes and protects mice from lethal endotoxin shock. *The Journal of experimental medicine*. 2006;203(1):15-20.
213. Sugihara H, Ishimoto T, Yasuda T, et al. Cancer-associated fibroblast-derived CXCL12 causes tumor progression in adenocarcinoma of the esophagogastric junction. *Medical Oncology*. 2015;32(6):618.

## List of References

214. Denduluri SK, Idowu O, Wang Z, et al. Insulin-like growth factor (IGF) signaling in tumorigenesis and the development of cancer drug resistance. *Genes & Diseases*. 2015;2(1):13-25.
215. Hjortebjerg R. IGFBP-4 and PAPP-A in normal physiology and disease. *Growth hormone & IGF research*. 2018;41:7-22.
216. Keravnou A, Bashiardes E, Michailidou K, Soteriou M, Moushi A, Cariolou M. Novel variants in the ACTA2 and MYH11 genes in a Cypriot family with thoracic aortic aneurysms: a case report. *BMC Medical Genetics*. 2018;19(1):208.
217. Chakraborty R, Saddouk Fatima Z, Carrao Ana C, Krause Diane S, Greif Daniel M, Martin Kathleen A. Promoters to Study Vascular Smooth Muscle. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2019;39(4):603-612.
218. Bondjers C, Kalén M, Hellström M, et al. Transcription Profiling of Platelet-Derived Growth Factor-B-Deficient Mouse Embryos Identifies RGS5 as a Novel Marker for Pericytes and Vascular Smooth Muscle Cells. *The American Journal of Pathology*. 2003;162(3):721-729.
219. Bozoky B, Savchenko A, Guven H, Ponten F, Klein G, Szekely L. Decreased decorin expression in the tumor microenvironment. *Cancer Medicine*. 2014;3(3):485-491.
220. Ezure T, Sugahara M, Amano S. Senescent dermal fibroblasts negatively influence fibroblast extracellular matrix-related gene expression partly via secretion of complement factor D. *BioFactors*. 2019;45(4):556-562.
221. Coppé J-P, Patil CK, Rodier F, et al. Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. *PLoS Biology*. 2008;6(12):2853-2868.
222. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016;44(W1):W90-W97.
223. Jawaid W. enrichR: Provides an R Interface to 'Enrichr'. 2017; <https://cran.r-project.org/package=enrichR>.
224. Blum R, Dynlacht BD. The role of MyoD1 and histone modifications in the activation of muscle enhancers. *Epigenetics*. 2013;8(8):778-784.
225. Takahashi T, Asano Y, Sugawara K, et al. Epithelial Fli1 deficiency drives systemic autoimmunity and fibrosis: Possible roles in scleroderma. *Journal of Experimental Medicine*. 2017;214(4):1129-1151.
226. Piersma B, de Rond S, Werker PM, et al. YAP1 Is a Driver of Myofibroblast Differentiation in Normal and Diseased Fibroblasts. *The American journal of pathology*. 2015;185(12):3326-3337.
227. Emblom-Callahan MC, Chhina MK, Shlobin OA, et al. Genomic phenotype of non-cultured pulmonary fibroblasts in idiopathic pulmonary fibrosis. *Genomics*. 96(3):134-145.
228. Akamata K, Wei J, Bhattacharyya M, et al. SIRT3 is attenuated in systemic sclerosis skin and lungs, and its pharmacologic activation mitigates organ fibrosis. *Oncotarget*. 2016;7(43):69321-69336.
229. Navab R, Strumpf D, Bandarchi B, et al. Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(17):7160-7165.
230. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740.
231. Hinata K, Gervin AM, Jennifer Zhang Y, Khavari PA. Divergent gene regulation and growth effects by NF-κB in epithelial and mesenchymal cells of human skin. *Oncogene*. 2003;22(13):1955-1964.
232. Lesina M, Wörmann SM, Morton J, et al. RelA regulates CXCL1/CXCR2-dependent oncogene-induced senescence in murine Kras-driven pancreatic carcinogenesis. *The Journal of Clinical Investigation*. 2016;126(8):2919-2932.
233. Narang D, Chen W, Ricci CG, Komives EA. RelA-Containing NFκB Dimers Have Strikingly Different DNA-Binding Cavities in the Absence of DNA. *Journal of Molecular Biology*. 2018;430(10):1510-1520.

234. Nissen NI, Karsdal M, Willumsen N. Collagens and Cancer associated fibroblasts in the reactive stroma and its relation to Cancer biology. *Journal of Experimental & Clinical Cancer Research*. 2019;38(1):115.
235. Anastassiou D, Rumjantseva V, Cheng W, et al. Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. *BMC Cancer*. 2011;11:529.
236. Poola I, DeWitty RL, Marshalleck JJ, Bhatnagar R, Abraham J, Leffall LD. Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis. *Nature Medicine*. 2005;11(5):481-483.
237. Hojilla CV, Mohammed FF, Khokha R. Matrix metalloproteinases and their tissue inhibitors direct cell fate during cancer development. *British Journal of Cancer*. 2003;89(10):1817-1821.
238. Chen H, Peng H, Liu W, et al. Silencing of plasminogen activator inhibitor-1 suppresses colorectal cancer progression and liver metastasis. *Surgery*. 2015;158(6):1704-1713.
239. Pavon MA, Arroyo-Solera I, Cespedes MV, Casanova I, Leon X, Mangues R. uPA/uPAR and SERPINE1 in head and neck cancer: role in tumor resistance, metastasis, prognosis and therapy. *Oncotarget*. 2016;7(35):57351-57366.
240. Lim KP, Cirillo N, Hassona Y, et al. Fibroblast gene expression profile reflects the stage of tumour progression in oral squamous cell carcinoma. *The Journal of Pathology*. 2011;223(4):459-469.
241. Fridman AL, Tainsky MA. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. *Oncogene*. 2008;27:5975.
242. Wang G, Jacquet L, Karamariti E, Xu Q. Origin and differentiation of vascular smooth muscle cells. *The Journal of Physiology*. 2015;593(14):3013-3030.
243. Xie T, Wang Y, Deng N, et al. Single-Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis. *Cell Reports*. 2018;22(13):3625-3640.
244. Li Q, Zhang Y, Jiang Q. MFAP5 suppression inhibits migration/invasion, regulates cell cycle and induces apoptosis via promoting ROS production in cervical cancer. *Biochemical and Biophysical Research Communications*. 2018;507(1):51-58.
245. Steiglit BM, Keene DR, Greenspan DS. PCOLCE2 Encodes a Functional Procollagen C-Proteinase Enhancer (PCPE2) That Is a Collagen-binding Protein Differing in Distribution of Expression and Post-translational Modification from the Previously Described PCPE1. *Journal of Biological Chemistry*. 2002;277(51):49820-49830.
246. Bhide VM, Laschinger CA, Arora PD, et al. Collagen Phagocytosis by Fibroblasts Is Regulated by Decorin. *Journal of Biological Chemistry*. 2005;280(24):23103-23113.
247. Hinz B, Phan SH, Thannickal VJ, Galli A, Bochaton-Piallat ML, Gabbiani G. The myofibroblast: one function, multiple origins. *The American Journal of Pathology*. 2007;170(6):1807-1816.
248. Lewis MP, Lygoe KA, Nystrom ML, et al. Tumour-derived TGF- $\beta$ 1 modulates myofibroblast differentiation and promotes HGF/SF-dependent invasion of squamous carcinoma cells. *British Journal of Cancer*. 2004;90(4):822-832.
249. Hayden MS, Ghosh S. Signaling to NF-kappaB. *Genes Dev*. 2004;18(18):2195-2224.
250. Hinz B, Phan SH, Thannickal VJ, et al. Recent Developments in Myofibroblast Biology: Paradigms for Connective Tissue Remodeling. *The American Journal of Pathology*. 2012;180(4):1340-1355.
251. Goss BC, McGee KP, Ehman EC, Manduca A, Ehman RL. Magnetic resonance elastography of the lung: Technical feasibility. *Magnetic Resonance in Medicine*. 2006;56(5):1060-1066.
252. Rout UK, Saed GM, Diamond MP. Expression pattern and regulation of genes differ between fibroblasts of adhesion and normal human peritoneum. *Reproductive Biology and Endocrinology*. 2005;3(1):1.
253. Song T, Dou C, Jia Y, Tu K, Zheng X. TIMP-1 activated carcinoma-associated fibroblasts inhibit tumor apoptosis by activating SDF1/CXCR4 signaling in hepatocellular carcinoma. *Oncotarget*. 2015;6(14):12061-12079.

## List of References

254. Sampson N, Zenzmaier C, Heitz M, et al. Stromal insulin-like growth factor binding protein 3 (IGFBP3) is elevated in the diseased human prostate and promotes ex vivo fibroblast-to-myofibroblast differentiation. *Endocrinology*. 154(8):2586-2599.
255. Torr EE, Ngam CR, Bernau K, Tomasini-Johansson B, Acton B, Sandbo N. Myofibroblasts Exhibit Enhanced Fibronectin Assembly That Is Intrinsic to Their Contractile Phenotype. *Journal of Biological Chemistry*. 2015;290(11):6951-6961.
256. Räsänen K, Vaheri A. Activation of fibroblasts in cancer stroma. *Experimental Cell Research*. 2010;316(17):2713-2722.
257. Avery D, Govindaraju P, Jacob M, Todd L, Monslow J, Puré E. Extracellular matrix directs phenotypic heterogeneity of activated fibroblasts. *Matrix Biology*. 2018;67:90-106.
258. Arora PD, Narani N, McCulloch CAG. The Compliance of Collagen Gels Regulates Transforming Growth Factor- $\beta$  Induction of  $\alpha$ -Smooth Muscle Actin in Fibroblasts. *The American Journal of Pathology*. 1999;154(3):871-882.
259. Hinz B, Dugina V, Ballestrem C, Wehrle-Haller B, Chaponnier C.  $\alpha$ -Smooth Muscle Actin Is Crucial for Focal Adhesion Maturation in Myofibroblasts. *Molecular Biology of the Cell*. 2003;14(6):2508-2519.
260. Moore CB, Guthrie EH, Huang MT-H, Taxman DJ. Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown. *Methods in molecular biology (Clifton, NJ)*. 2010;629:141-158.
261. Costa EC, Moreira AF, de Melo-Diogo D, Gaspar VM, Carvalho MP, Correia IJ. 3D tumor spheroids: an overview on the tools and techniques used for their analysis. *Biotechnology Advances*. 2016;34(8):1427-1441.
262. Shamir ER, Ewald AJ. Three-dimensional organotypic culture: experimental models of mammalian biology and disease. *Nature Reviews Molecular Cell Biology*. 2014;15(10):647-664.
263. Ghosh S, Spagnoli GC, Martin I, et al. Three-dimensional culture of melanoma cells profoundly affects gene expression profile: A high density oligonucleotide array study. *Journal of Cellular Physiology*. 2005;204(2):522-531.
264. Sharon Y, Alon L, Glanz S, Servais C, Erez N. Isolation of normal and cancer-associated fibroblasts from fresh tissues by Fluorescence Activated Cell Sorting (FACS). *Journal of Visualized Experiments*. 2013(71):e4425.
265. Kojima Y, Acar A, Eaton EN, et al. Autocrine TGF- $\beta$  and stromal cell-derived factor-1 (SDF-1) signaling drives the evolution of tumor-promoting mammary stromal myofibroblasts. *Proceedings of the National Academy of Sciences of the United States of America*. 2010;107(46):20009-20014.
266. Hinz B. Masters and servants of the force: The role of matrix adhesions in myofibroblast force perception and transmission. *European Journal of Cell Biology*. 2006;85(3):175-181.
267. Achterberg VF, Buscemi L, Diekmann H, et al. The Nano-Scale Mechanical Properties of the Extracellular Matrix Regulate Dermal Fibroblast Function. *Journal of Investigative Dermatology*. 2014;134(7):1862-1872.
268. Kessler D, Dethlefsen S, Haase I, et al. Fibroblasts in Mechanically Stressed Collagen Lattices Assume a "Synthetic" Phenotype. *Journal of Biological Chemistry*. 2001;276(39):36575-36585.
269. Balestrini JL, Chaudhry S, Sarrazy V, Koehler A, Hinz B. The mechanical memory of lung myofibroblasts. *Integrative Biology*. 2012;4(4):410-421.
270. Hong LZ, Wei XW, Chen JF, Shi Y. Overexpression of periostin predicts poor prognosis in non-small cell lung cancer. *Oncology Letters*. 2013;6(6):1595-1603.
271. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*. 2015;12(5):453-457.
272. The Cancer Genome Atlas. <https://cancergenome.nih.gov/>.
273. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. 2009;4(1):44-57.
274. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*. 2019;20(1):40.

275. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*. 2019;37(5):547-554.
276. Xiang J, Wan C, Guo R, Guo D. Is Hydrogen Peroxide a Suitable Apoptosis Inducer for All Cell Types? *BioMed Research International*. 2016;2016:7343965-7343965.
277. Kong Q, Lin C-LG. Oxidative damage to RNA: mechanisms, consequences, and diseases. *Cellular and Molecular Life Sciences* 2010;67(11):1817-1829.
278. Donnem T, Al-Shibli K, Al-Saad S, Busund LT, Bremnes RM. Prognostic impact of fibroblast growth factor 2 in non-small cell lung cancer: coexpression with VEGFR-3 and PDGF-B predicts poor survival. *Journal of Thoracic Oncology*. 2009;4(5):578-585.
279. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*. 2002;30(1):207-210.
280. Kumaraswamy S, Chinnaiyan P, Shankavaram UT, Lu X, Camphausen K, Tofilon PJ. Radiation-induced gene translation profiles reveal tumor type and cancer-specific components. *Cancer Research*. 2008;68(10):3819-3826.
281. Brancato V, Comunanza V, Imparato G, et al. Bioengineered tumoral microtissues recapitulate desmoplastic reaction of pancreatic cancer. *Acta Biomaterialia*. 2017;49:152-166.
282. Fang F, Ooka K, Sun X, et al. A synthetic Toll-like receptor 3 ligand mitigates profibrotic fibroblast responses by inducing autocrine IFN signaling. *The Journal of Immunology*. 2013;191(6):2956-2966.
283. Bhattacharyya S, Wang W, Morales-Nebreda L, et al. Tenascin-C drives persistence of organ fibrosis. *Nature Communications*. 2016;7:11703.
284. Sargent JL, Milano A, Bhattacharyya S, et al. A TGFbeta-responsive gene signature is associated with a subset of diffuse scleroderma with increased disease severity. *Journal of Investigative Dermatology*. 2010;130(3):694-705.
285. Williams RC, Skelton AJ, Todryk SM, Rowan AD, Preshaw PM, Taylor JJ. Leptin and Pro-Inflammatory Stimuli Synergistically Upregulate MMP-1 and MMP-3 Secretion in Human Gingival Fibroblasts. *PLoS One*. 2016;11(2):e0148024.
286. Indraccolo S, Pfeffer U, Minuzzo S, et al. Identification of genes selectively regulated by IFNs in endothelial cells. *The Journal of Immunology*. 2007;178(2):1122-1135.