Defining Epithelial-Mesenchymal Plasticity in Cancer Using Single-Cell Genomics

David Cook

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Doctorate in Philosophy (PhD) program in Cellular and Molecular Medicine

Department of Cellular and Molecular Medicine
Faculty of Medicine
University of Ottawa

Table of Contents

Abstract

Epithelial-mesenchymal plasticity (EMP) describes the interconversion of cells between epithelial and mesenchymal phenotypes. During the epithelial-mesenchymal transition (EMT), epithelial cells lose defining characteristics, such as stable cell-cell junctions, and gain the ability to migrate and invade through extracellular matrices. This plasticity contributes to tumour progression, promoting therapy resistance and immune cell evasion. Despite its importance, defining molecular features of this plasticity have largely remained elusive due to the limited scale of most studies. Here, I present my studies applying comparative single-cell genomics to map transcriptional changes associated with the EMT in diverse experimental conditions and EMP in tumours, I identify regulatory features associated with these dynamics, and explore opportunities to pharmacologically restrict them. This work provides critical steps towards building quantitative models of EMP, which will inform effective strategies to restrict these dynamics in cancer and improve patient prognosis.

List of Figures

List of Tables

List of Abbreviations

ATAC - Assay of transposase-accessible chromatin
ATP - Adenosine triphosphate
CAF - Cancer-associated fibroblasts
cDNA - Complementary DNA
CPM - Counts per million
CRISPR - Clustered regularly interspaced short palindromic repeats
CRISPRa - CRISPR activation
CRISPRi - CRISPR interference
DC - Dendritic cell
DMEM - Dulbecco's Modified Eagle Medium
E/M - Epithelial-mesenchymal
ECM - Extracellular matrix
EDTA - Ethylenediaminetetraacetic acid
EMP - Epithelial-mesenchymal plasticity
EMT - Epithelial-mesenchymal transition
FACS - Fluorescence-activated cell sorting
FBS - Fetal bovine serum
FGF - Fibroblast growth factor
GO - Gene Ontology
GSEA - Gene set enrichment analysis
HDAC - Histone deacetylase
MET - Mesenchymal-epithelial transition
MOI - Multiplicity of infection
Nec-5 - Necrostatin-5
NK - Natural killer
NSCLC - Non-small cell lung cancer
PBS - Phosphate-buffered saline
PC - Principal component
PCA - Principal Component Analysis
PCR - Polymerase chain reaction
pEMT - Partial epithelial-mesenchymal transition
PFI - Progression-free interval
qPCR - Quantitative polymerase chain reaction
RTK - Receptor tyrosine kinase
scRNA-seq - Single-cell RNA sequencing
sgRNA - Single-guide RNA
t-SNE - *t*-Distributed Stochastic Neighbor Embedding
TCGA - The Cancer Genome Atlas
TEMTIA - The EMT International Association
TGFB - Transforming growth factor beta
TME - Tumour microenvironment

TNBC - Triple-negative breast cancer
Treg - Regulatory T cell
UMAP - Uniform Manifold Approximation and Projection
UMI - Unique molecular index

Contributions

This dissertation primarily contains findings from the following studies:

1. Cook, DP., Vanderhyden, BC. (2020). Context specificity of the EMT transcriptional response. *Nature Communications.* 11(1):2142

2. Cook, DP., Vanderhyden, BC. (2021). Transcriptional census of epithelial-mesenchymal plasticity in cancer. Preprint on *bioRxiv* (doi: 10.1101/2021.03.05.434142)

Katayoun Sheikheleslamy and Caroline Vergette from StemCore (Ottawa Hospital Research Institute) processed scRNA-seq samples, performed library preparation, and sequencing. Fernando Ortiz (StemCore) assisted with flow cytometry analysis. Beyond this, I performed all experimental and computational components.

Funding Sources

- Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN 2018-0653-8; Recipient: Barbara Vanderhyden)

- Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Awards (Recipient: David Cook)

- Ontario Graduate Scholarship (Recipient: David Cook)

Thesis Advisory Committee Members

- Dr. Douglas Gray
- Dr. Marjorie Brand
- Dr. Ted Perkins

Complete publication list

1. **Cook, DP.***, Vanderhyden, BC. (2021). Transcriptional census of epithelial-mesenchymal plasticity in cancer. Preprint on *bioRxiv* (doi: 10.1101/2021.03.05.434142). Submitted to *Science Advances*
*Corresponding author

2. Lavictoire, SJ., Jomaa, D., Dont, A., Jardine, K., **Cook, DP.**, Lorimer, IAJ. (2021). Rac guanine nucleotide exchange factors promoting Lgl phosphorylation in glioblastoma. Preprint on bioRxiv (doi: 10.1101/2020.12.01.406538). *Journal of*

Biological Chemistry, In Review.

3. Al-Zahrani, KN., Abou-Hamad, J., Labrèche, C., Garland, B., Cook, DP., Sabourin, LA. (2021). AKT-mediated phoshorylation of Sox9 induces Sox10 transcription in a murine model of HER2-positive breast cancer. *Breast Cancer Research*, In Review.

4. Carter, LE.*, **Cook, DP.*,** McCloskey, CW., Dang, T., Collins, O., Gamwell, LF. Vanderhyden, BC. (2021). Transcriptional heterogeneity of stemness phenotypes in the ovarian epithelium. Preprint in bioRxiv (doi:10.1101/2020.06.10.145045). *Communications Biology*, In Press.
*Co-first authors

5. Hurskainen, KM., Mižíková I., **Cook, DP.,** Cyr-Depauw, C., Lesage, F., Helle, E., Renesme, L., Lithopoulous, MA., Jankov, RP., Vanderhyden, BC., Thébaud, B. (2021). Single-cell transcriptomic analysis of normal and impaired lung development in the mouse. Preprint in *bioRxiv* (doi:10.1101/868802). *Nature Communications*. 12(1):1565

6. Robichaud, S., Fairman, G., Vijithakumar, V., Mak, E., **Cook, DP**., Pelletier, A., Huard, S., Vanderhyden, BC., Figeys, D., Lavallée-Adam, M., Baetz, K., Ouimet, M. (2021). Identification of novel lipid droplet factors that regulate lipophagy and cholesterol efflux in macrophage foam cells. *Autophagy. 1-19*

7. **Cook, DP.*,** Steed, K., Read, C., Baysarowich, R., Redway, T., Robineau-Charette, P., Carnegie, J. (2020). Science Outreach: Six examples of programs that enrich the learning environments of students and educators. *HAPS Educator.*
***Corresponding author**

8. Karunakaran, D., Turner, AW., Duchez, A., Soubeyrand, S., Rasheed, A., Smyth, D., **Cook, DP**., Kandiah, JW., Pan, C., Geoffrion, M., Nikpay, M., Lee, R., Boytard, JW., Pan, C., Nguyen, M., Lau, P., Laakso, M., Ramkhelawon, B., Vanderhyden, BC., Liu, P., Berger, SB., Gough, PJ., Beal, AM., Bertin, J., Harper, M., Lusis, AJ., McPherson, R., Rayner, KJ. (2020). RIPK1 gene variants associate with obesity in humans and can be therapeutically silenced to reduce obesity in mice. *Nature Metabolism. 2(10):1113-1125*

9. Al-Zahrani, K., Abou-Hamad, J., **Cook, DP.,** Pryce, BR., Hodgins, JJ., Labrèche, C., Robineau-Charette, P., de Souza, CT., Bell, JC., Auer, RC., Ardolino, M., Vanderhyden, BC., Sabourin, LA. (2020). Loss of the Ste20-like kinase induces a basal/stem-like phenotype in HER2-positive breast cancers. *Oncogene*. 39(23):4592-4602

10. **Cook, DP**.*, Vanderhyden, BC*. (2020). Context specificity of the EMT transcriptional response. *Nature Communications.* 11(1):2142
*Co-corresponding author

11. Galpin, KJC.*, **Cook, DP.*,** Salemi, LM., Urowitz, S., Williams, C., Bell, JC., Brundage, MD., Vanderhyden, BC. (2020). Report on the Canadian Cancer Research Conference 2019. *Current Oncology*, 27(2):e226-e230.
*Co-first authors

12. **Cook, DP.*,** Vanderhyden, BC.* (2020). Cre-mediated deletion of SMARCA5 disrupts pluripotency in mouse embryonic stem cells. Preprint in *bioRxiv* (doi:10.1101/2020.04.02.022285).
    *Co-corresponding author

13. McCloskey, CW., **Cook, DP.,** Kelly, BS., Azzi, F., Allen, CH., Forsyth, A., Upham, J., Rayner, KJ., Gray, DA., Boyd, RW., Murugkar, S., Lo, B., Trudel, D., Senterman, MK., Vanderhyden, BC. (2019). Metformin abrogates age-associated ovarian fibrosis in women. *Clinical Cancer Research*. 26(3):632-642

14. Carter, LE., **Cook, DP.**, Collins O., Gamwell, LF., Dempster, HA., Wong, HW., McCloskey, CW., Garson, K., Vuong, NH., Vanderhyden, BC. (2019). COX2 is induced in the ovarian epithelium during ovulatory wound repair and promotes cell survival. *Biology of Reproduction.* 101(5):961-974

15. **Cook, DP.*,** Vanderhyden, BC. (2019). Ovarian cancer and the evolution of subtype classifications using transcriptional profiling. *Biology of Reproduction*. 101(3):645-658
    Invited review for New Horizons Anniversary Issue
    *Corresponding author

16. Carter, LE.*, **Cook, DP.*,** Vanderhyden, BC. (2018). Phenotypic Plasticity and the Origins and Progression of Ovarian Cancer. *The Ovary* (3rd Edition; Elsevier).
    *Co-first authors

17. Vuong, NH.*, **Cook, DP.***, Forrest, LA., Carter, LE., Robineau-Charette, P., Kofsky, JM., Hodgkinson, KM., Vanderhyden, BC. (2018). Single-cell RNA sequencing reveals transcriptional dynamics of estrogen-induced dysplasia in the ovarian surface epithelium. *PLOS Genetics*. 14(11):e1007788
    *Co-first authors

18. Al-Zahrani, K., **Cook, DP.,** Vanderhyden, BC., Sabourin, LA. (2018) Assessing the efficacy of androgen receptor and Sox10 as independent markers of the triple-negative breast cancer subtype by transcriptome profiling. *Oncotarget*. 9(70):33348-33359

19. Vlasschaert, C., **Cook, DP.**, Xia, X., Gray, D. (2017). The evolution and functional diversification of the deubiquitinating enzyme superfamily. *Genome Biol and Evol.* 9(3):558-573

Acknowledgements

Oh, Barb. It's hard to express the gratitude I have for you. I came up to you at the end of an undergrad class in 2012 to ask a question and accidentally stumbled into a position as an Honours student. I didn't know it at the time, but that small decision would shape my entire career. If I had instead decided to go get a coffee after this early morning class, I really doubt I would have ended up in research. I probably would have ended up in—ugh—medical school. You have always encouraged me to explore my scientific interests even if they were new territory for us. You embraced my ideas and provided unending support, but still challenged, pushed, and dug. You have crafted me into a scientist and sharpened my skills, all while being a caring and supportive mentor. Having been in science for a while now, I understand how rare and truly special this is and I really am forever appreciative.

Occasionally you're lucky enough to meet someone that makes you want to be the best person you can be. I was fortunate enough to find this person when I was 15 years old and ended up marrying her. Andrea, you've been my partner through this entire journey and have always been there for me when I needed it. Your drive and commitment inspire me to do the best I can do. I can't thank you enough for your non-stop support over the years.

Olivia, you're only 7 months old right now, but I didn't know I could feel so much love in such a short period of time. Be curious. Be brave. Never let anything hold you back.

To my kind and supportive parents, Mary-Catherine and Brian. Thank you for buying me a guitar instead of textbooks. Music was my first real passion and rather than pushing me into my studies, you gave me the freedom to explore and encouraged my creativity. The uncertainty of my future must have been terrifying. But this shaped who I am and I would not be the scientist I am today without it.

When I first joined the lab, I never expected to meet a group of friends that would become such an important support network in my life. Pascale, I'm so grateful for our friendship. We've supported each other through the ups and downs of science and I'm lucky to have had you there to push me to always do better, even if that meant arguing for an hour about the clarity of a sentence. To the entire Biohazard ultimate team, I don't think I could have made it through this experience without you. When I look back on my PhD, I will always remember losing ultimate games with smiles on our faces and our nights of a few too many drinks.

Chapter 1: Introduction

1.1 The epithelial-mesenchymal transition: an overview

From the 1960s to the late 1990s, Elizabeth Hay's pioneering research led to the first descriptions of the "epithelio-mesenchymal transformation", in which epithelial cells lose defining characteristics such as stable cell-cell junctions, and transition towards a mesenchymal state capable of migration and invasion through extracellular matrices (ECM) (Hay, 1995) (**Figure 1**). Hay's initial studies of this phenomenon focused on migratory cells during primitive streak formation, but this phenomenon plays a central role in many developmental processes, with nearly all organs forming through sequential transitions between epithelial and mesenchymal states (Thiery et al., 2009). In the years following this focus on embryonic development, this process has also been implicated in adult tissue homeostasis following injury (eg. wound healing), and in pathologies such as fibrosis and cancer (Kalluri and Weinberg, 2009). At the first meeting of the EMT International Association (TEMTIA) in 2003, the term "transformation" was replaced with "transition" to reflect the reversibility of this process (the mesenchymal-epithelial transition; MET) and to distinguish it from oncogenic transformation. Since these initial discoveries, the epithelial-mesenchymal transition (EMT) has perhaps become one of the most frequently studied phenomena in biology, with an increasing number of publications each year and over 10,000 studies in the last two years alone.

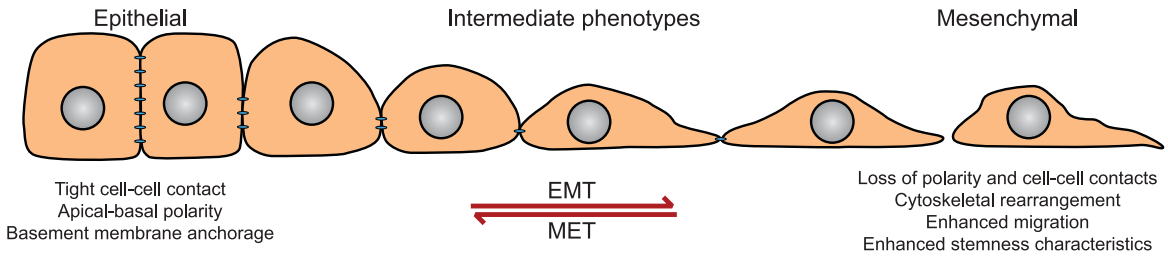

Figure 1. An overview of the epithelial-mesenchymal transition. Schematic showing the gradual morphological and functional changes as cells transition between epithelial and mesenchymal phenotypes.

1.2 Cellular properties associated with the EMT

As described above, at the most fundamental level, the EMT involves the loss of epithelial junctions, while gaining the capacity to migrate and invade through ECM. Underlying these functional changes are a dramatic transformation in the cell's repertoire of proteins on the plasma membrane and the cell's cytoskeleton.

One of the earliest described features of the EMT was the observation that the adherens junction protein E-cadherin (*CDH1*) was lost during the transition (Hay, 1995). The EMT is canonically associated with "cadherin switch", where E-cadherin-rich junctions are broken down and the mesenchymal-associated N-cadherin (*CDH2*) is expressed on the cell surface. Similar to E-cadherin, N-cadherin is connected internally to the cell's cytoskeleton, but has reduced junctional affinity for epithelial cells and increased affinity for fibroblast and endothelial cells (Yilmaz and Christofori, 2009).

The disruption of adherens junctions, along with desmosomes and tight junctions leads to the loss of apical-basal polarity (Yilmaz and Christofori, 2009). Cellular projections provide mesenchymal cells front-rear polarity that provides motility and the ability to invade through surrounding tissue. These projections are mediated through complex dynamics of the cells' cytoskeleton. Upon activation of the EMT, thin bundles of cortical actin comprising the epithelial cytoskeleton are reorganized into thick contractile filaments at the ventral surface that enable dynamic elongation and directional motility (Haynes et al., 2011; Yilmaz and Christofori, 2010). This remodelling enables the projection of actin-rich lamellipodia and filopodia to form from the cell and provide motility and sensory properties (Yilmaz and Christofori, 2009). Increased expression of the intermediate filament vimentin ensures integrity of cytoplasmic components during this transition (Goldman et al., 1996)

Beyond simple motility along a planar surface, mesenchymal cells are capable of penetrating and invading through three-dimensional ECM. Actin-rich invadopodia project from the cell body and penetrate through the basement membrane underlying the cells. This is facilitated by localized release of proteolytic matrix metalloproteinases (MMPs) and ADAM metalloproteinases (Jacob and Prekeris, 2015). The repertoire of integrin proteins expressed on the surface of the cell also changes, altering the cell's physical interactions with the ECM. Epithelial integrins, such as the laminin- and E-cadherin-binding α3β1 integrin, are repressed, while expression of mesenchymal-associated integrins is increased, including α5β1 and α1β1 integrins, which interact with fibronectin and type I collagen, respectively (Lamouille et al., 2014). These changes weaken the cell's association with the adjacent epithelium and enhance interactions with proteins comprising the ECM. Integrins are also coupled to intracellular signalling cascades that can further promote the EMT upon interaction.

1.3 Signal transduction pathways promoting the EMT response

Early knowledge of signals that promote the EMT came from studies of primitive streak formation and gastrulation during embryonic development. Nuclear localization of the Wnt signalling transcription factor β-catenin (*CTNNB1*) specifies cells that form the primitive streak, and signals from members of the transforming growth factor beta (TGFB) superfamily (eg. Nodal) induce gastrulation and, along with fibroblast growth factor (FGF), drive specification of the mesendoderm (Mohamed et al., 2004; Thiery et al., 2009). Beyond development, these pathways — particularly TGFB signalling — have been implicated in promoting EMT responses in a wide variety of contexts.

Regulation of the EMT response, however, is far from limited to these pathways. In 1987, a factor derived from fibroblasts was shown to promote marked changes in epithelial cell morphology and motility, leading to the colloquial nickname “scatter factor” (Stoker et al., 1987). This factor was later identified as hepatocyte growth factor (HGF), which signals through the receptor tyrosine kinase (RTK) c-MET that has since been reproducibly shown to promote an EMT (Jeon and Lee, 2017). Other RTKs have also been demonstrated to elicit this response, including EGFR, FGFR, PDGFR, VEGFR, and IGFR (Gonzalez and Medici, 2014; Lamouille et al., 2014). A recent genetic screen demonstrated that successful progression through an EMT response is dependent on activity of MEK — a downstream kinase of RTKs (McFaline-Figueroa et al., 2019).

In fact, one would be hard-pressed to identify a signalling pathway that has not been implicated in the EMT. Beyond those mentioned above, JNK (Alcorn et al., 2008), Semaphorin-plexin (Gurrapu and Tamagnone, 2019), Notch (Wang et al., 2010), Hippo (Yamaguchi and Taouk, 2020), Shh (Zhang et al., 2016a), NFkB (Taniguchi and Karin, 2018), BMP (Zhang et al., 2016b), and STAT (Jin, 2020) signalling have all been shown to promote the EMT. The cyclic AMP-dependent PKA was shown to promote an MET in cells (Pattabiraman et al., 2016) and has been a proposed therapeutic target to sensitize cancer cells to chemotherapy (Gooding and Schiemann, 2016), however other findings suggest that PKA is essential for HIF1A-mediated EMT (Shaikh et al., 2012).

Signalling responses to many environmental factors also promote EMT responses. Hypoxia is a strong inducer of the transition, thought to primarily be mediated through the transcription factor HIF1a (Tam et al., 2020). Acidity has also been shown to lead to an EMT in a variety of contexts (Andreucci et al., 2020; Riemann et al., 2019; Suzuki et al., 2014). Physical properties of the cells’ environment can promote the transition, including high

tissue stiffness (Broders-Bondon et al., 2018), mechanical stretch (Heise et al., 2011), and exposure to fluid flow (eg. ascites in the peritoneal cavity) (Rizvi et al., 2013). Even experimentally simulated microgravity may cause an EMT (Ranieri et al., 2017).

1.4 Transcription factor networks coordinating the EMT response

Many signal transduction pathways converge on the genome, forming connected networks of transcription factors that coordinate gene expression. The cell's transcriptional response to these signals is ultimately determined by its present state: the repertoire of proteins it is expressing, its metabolic status, the physical organization of its genome, variations in its genetic code, and likely additional layers of regulation that I have not included here. The role of transcription factors as direct mediators of gene expression, however, has garnered much attention: identifying "master transcription factors" of specific cell types and states has been a driving motivation of many studies.

Over the last 20 years, several transcription factors have emerged as core regulators of the EMT. SNAI1 (Snail) was first identified as a mediator of E-cadherin suppression in epithelial cells (Batlle et al., 2000; Cano et al., 2000) and its homolog SNAI2 (Slug) was implicated in the same function soon after (Bolós et al., 2003; Hajra et al., 2002). Around this time, ZEB1 and the basic helix-loop-helix (bHLH)-family factor TWIST1 were also implicated in the EMT through regulation of E-cadherin (Eger et al., 2005; Yang et al., 2004). These three families of factors are highly distinct in their size and structure, with their ability to induce an EMT being the only obvious commonality (Stemmler et al., 2019). While dozens of other transcription factors have since been implicated in the transition, including various members of the FOX, SOX, and KLF transcription factor families (Stemmler et al., 2019), the position of Snail, Slug, ZEB1, ZEB2, and TWIST1 as master regulators of the EMT has become

cemented in the field and far less is known about how other factors contribute to this process.

Perhaps due to their recognition as master regulators, these canonical EMT transcription factors are well studied and their expression is often used as a measure of EMT activity. With an abundance of studies, there are now many counter examples of EMT responses that occur independently from the activity of specific factors and even cases where the expression of certain factors is repressed (comprehensively reviewed in [Stemmler et al., 2019]). Given the many signals known to promote an EMT, little is known about how distinct combinations of transcription factors are regulated. Non-redundant functions of the canonical factors have been described (Stemmler et al., 2019), and while this may suggest that they promote nuances in the functional properties of cells undergoing the transition, it may also reflect context specificity that cannot directly be linked to specific cellular functions. For example, in the KPC mouse model of pancreatic cancer, ZEB1 is essential for metastasis and Snail and TWIST1 are dispensable (Zheng et al., 2015), but TWIST1 activation promotes metastasis in a model of skin cancer (Tsai et al., 2012). Given the diversity of experimental models and perturbations, it seems appropriate to approach generalization of these claims with caution.

1.5 Epithelial-mesenchymal plasticity in cancer

Considering the cellular changes associated with the EMT, its association with cancer — and particularly metastatic disease — is not surprising (**Figure 2**). Throughout tumours, cells with phenotypes spanning an epithelial-mesenchymal (E/M) continuum can be observed. Towards the periphery of tumours, for example, the epithelial architecture becomes progressively disorganized and cancer cells express high levels of mesenchymal-

associated genes (Gabbert et al., 1985; Puram et al., 2017). Historically, it was thought that intermediate phenotypes along the E/M continuum were transitory and phenotypic stability only existed for the polar phenotypes. Growing evidence suggests that stable cells with mixed properties of E/M states are quite prevalent, particularly in cancer. It is currently unclear if intermediate states have any unique molecular features or if they are simply a hybrid of the two phenotypes. Recently, there has been much interest in how hybrid states affect the functional properties of cells and contribute to tumour progression (Pastushenko and Blanpain, 2019). Broadly, this source of intratumoural heterogeneity is referred to as epithelial-mesenchymal plasticity (EMP) due to the dynamic interconversion between phenotypes along this continuum.

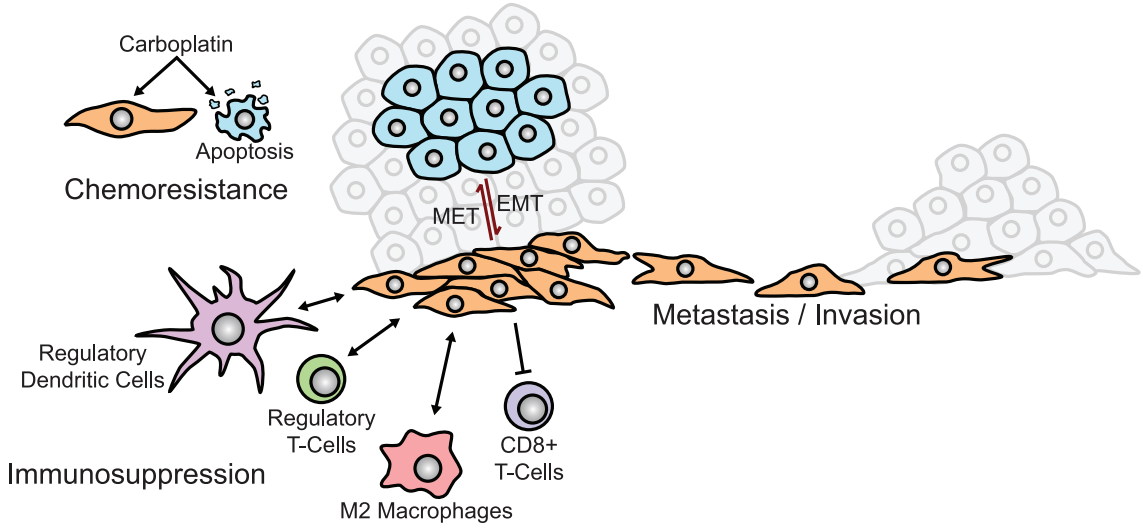

Figure 2. Epithelial-mesenchymal plasticity (EMP) in cancer. EMP is a prominent source of intratumoural heterogeneity. Mesenchymal phenotypes have been attributed with various properties that promote tumour progression, including metastasis, resistance to therapies, and immunosuppression.

1.5.1 Metastasis

The role of EMP in the metastatic cascade has drawn much interest over the last several decades. EMT could allow cells to dissociate from a primary tumour, invade through surrounding ECM, and facilitate intravasation into the bloodstream where the cell could

travel to distant sites, extravasate, and seed metastatic tumours. While fundamentally logical, establishing causal links has proven challenging due to the diverse molecular programs that can coordinate the EMT. There is ample evidence to suggest that experimental induction of an EMT in a tumour can promote dissemination. For example, in an experimental model of skin cancer, doxycycline-induced TWIST1 expression promoted metastasis, but only when induced locally and not systemically, suggesting that TWIST1 activation can lead to dissemination, but its repression (leading to an MET) was necessary for growth at the metastatic site (Tsai et al., 2012). However, many studies such as this have received the criticism that although experimental induction of an EMT may promote metastasis, it does not provide evidence that naturally occurring metastasis is dependent on this transition. Zheng *et al*. (2015) demonstrated that genetic deletion of *Snai1* and *Twist1* in the KPC mouse model of pancreatic cancer did not affect metastasis rate or size, suggesting that they are dispensable for the process, however in the same model, deletion of *Zeb1* reduces metastasis (Krebs et al., 2017). Lineage tracing experiments have also not offered a clear resolution: cells comprising metastases in models of breast and pancreatic cancer were shown to have never expressed certain mesenchymal markers (Fischer et al., 2015; Zheng et al., 2015), but when different markers were used, metastatic cells were positive for the label (Aiello et al., 2017; Ye et al., 2017).

1.5.2 Treatment resistance

EMP provides phenotypic properties that contribute to tumour progression beyond its possible role in metastasis. It is well established that cancer cells that adopt a more mesenchymal phenotype show resistance to a variety of therapies. In colon and ovarian cancer, cells made resistant to platinum-based chemotherapies through prolonged exposure gain a more mesenchymal morphology (Kajiyama et al., 2007; Yang et al., 2006). While these findings don't establish directionality to this relationship, *SNAI1* and *SNAI2*

overexpression confers resistance to paclitaxel, adriamycin, and radiation by preventing p53-mediated apoptosis (Kajita et al., 2004; Kurrey et al., 2009), while reduction of TWIST1 and overexpression of the negative EMT regulator miR-200c both sensitize cells to treatments (Cochrane et al., 2009; Li et al., 2009). Several mechanistic explanations for this resistance include increased expression of antiapoptotic proteins, increased levels of ATP-binding cassette transporters that enable drug efflux, and reduced proliferation rates associated with mesenchymal cells (Shibue and Weinberg, 2017). EMP can also provide resistance to various targeted therapies. In non-small cell lung cancer (NSCLC) with wild-type EGFR, expression of EMT-associated genes correlates with resistance to EGFR inhibition (Thomson et al., 2005). In another study of NSCLC cells, EMT-associated expression was associated with increased resistance to both EGFR and PI3K inhibitors (Byers et al., 2013).

1.5.3 Interaction with the tumour microenvironment

It seems likely that interactions with stromal cell types within the tumour microenvironment (TME) play an important role in mediating EMP. While cancer cells themselves can spontaneously exhibit non-genetic EMP in the absence of stromal cell types, these cell types express an abundance of EMT-inducing factors, including TGFB1, TNF, HGF, IL6, and EGF (Dongre and Weinberg, 2019). In head and neck squamous cell carcinoma, mesenchymal phenotypes exist in close proximity to cancer-associated fibroblasts (CAF) at the leading edge of tumours (Puram et al., 2017), and co-culture of cancer cells with CAFs (Shintani et al., 2016; Zhou et al., 2014), macrophages (Dehai et al., 2014; Kuwada et al., 2018), T cells (Wang et al., 2020), natural killer (NK) cells (Huergo-Zapico et al., 2018), and tumor-associated neutrophils (Li et al., 2019b) have all been shown to induce an EMT response.

This regulatory relationship is not one directional: cells undergoing an EMT can modulate both innate and adaptive immunity within the TME. Overexpression of the EMT transcription factor *Snai1* in the murine B16 melanoma model resulted in tumours with more immunosuppressive regulatory T cells (Tregs) and impairs dendritic cell (DC) activity, rendering them resistant to DC-based immunotherapy (Kudo-Saito et al., 2009). Consistent with this immunosuppressive phenotype, overexpression of *SNAI1* in MCF7 cells reduces T cell reactivity (Akalay et al., 2013). Using cell lines with variable E/M status collected from MMTV-PyMT murine mammary tumours, Dongre *et al.* (2017) demonstrated that tumours from mesenchymal lines were associated with reduced infiltration of CD8 T cells, increasing Tregs, and protumoural CD206+ macrophages. Epithelial tumours were susceptible to anti-CTLA4 immunotherapy, whereas mesenchymal tumours showed no response. In this model, deletion of the immunosuppressive factors CD73, CSF1, SPP1, LGALS3, and MASP1 in the cancer cells increased the infiltration of CD8 T cells without clearly affecting the mesenchymal state of the cells (Dongre et al., 2020).

Recent excitement around invigorating anti-tumour immunity through immunotherapies provides hope for the development of effective treatments. However, the ability of EMP to impair immune function and confer resistance to immunotherapies is problematic. This highlights the need to develop strategies to program cancer cell state and restrict plasticity. If EMP can be therapeutically blocked, it may be possible to halt tumour progression and sensitize cells to anti-tumour immunity.

1.6 What is the molecular basis of EMP?

Having discussed the functional and molecular features of the EMT, it could seem that we have a clear understanding of the phenomenon. The basic cellular definition is perhaps the

most agreed upon: morphological changes and increased motility. In fact, TEMTIA recently released guidelines on the definition of EMT/EMP and strongly suggest basing experimental conclusions on these properties (Yang et al., 2020). This recommendation, however, is not practical for retrospective study of tissues or analysis of molecular data. Decades of research on the molecular basis of the EMP has provided a wealth of information, but has not provided a clear understanding of its nature. Dozens-to-hundreds of extracellular factors can induce an EMT response, nearly all signalling pathways can regulate it, and while EMT transcription factors have been identified, examples of responses independent of each factor have been observed. From the existing literature, there is no clear consensus on the molecular definition of EMP. It is clear, however, that molecular features of EMP are variable across contexts, but the extent of this variation is unknown as we often compare findings from vastly different experimental models. A controlled comparative analysis of the molecular basis of EMP will be critical to update our conceptual model of this phenomenon and understand how molecular variability affects phenotypic properties of cells.

1.7 Single-cell genomics as a platform to define EMP

Advances in single-cell genomics hold the promise to generate molecular definitions of the EMT in a largely unbiased, quantitative manner. Methodologies in this field have matured at a remarkable pace (Svensson et al., 2018). In 2009, Tang *et al*. (2009) reported the first whole-transcriptome RNA-seq of an individual mouse blastomere. A mere 12 years later, it is now possible to generate atlases of data from millions of single cells (Cao et al., 2019, 2020). These advances have largely been enabled by new approaches to generate cDNA libraries from single cells. At the time of writing this, the most impactful advance has likely been the development of microfluidic devices where a stream of cells in aqueous solution are partitioned into droplets by an intersecting stream of oil generating millions of nanoliter-

scale reaction chambers (Klein et al., 2015; Macosko et al., 2015). Commercialization of this approach a few years later by 10x Genomics gave labs easy access to this methodology without the need to build custom lab equipment (Zheng et al., 2017).

In the simplest sense, the power of single-cell genomics is its ability to resolve heterogeneity in a sample. Bulk assays involving the lysis of thousands or millions of cells provide measurements representing the average profile of a sample. The problem, however, is that averages can often be misleading, and with complex populations, the average can be representative of no individual component. In a single-cell RNA-seq (scRNA-seq) experiment, global gene expression profiles of the dozens of cell types comprising a tissue can be measured and explored independently.

Beyond resolving individual cell types, single-cell assays capture variation in the state of individual cell types. Due to both stochastic and biological factors, the expression profiles of no two cells are the same. By sampling many individual cells, one can define the distribution of expression patterns that a given cell type occupies. In theory, biological sources of variation involve the coordinated expression of many genes due to modular regulation, allowing them to be discerned from stochastic noise. This approach can be expanded to explore cellular responses. Taking advantage of this non-uniformity, along with the asynchrony of cellular responses to stimuli, Trapnell *et al*. (2014) sampled differentiating myoblasts daily for 3 days and demonstrated that the expression profiles span a continuous developmental trajectory despite the discrete nature of sampling. Cells can be pseudo-temporally ordered along this continuum to resolve a response trajectory (**Figure 3**). Sources of heterogeneity can be identified from the data with various matrix factorization methods that can identify low-dimensional structure from high-dimensional data (Stein-O'Brien et al., 2018). In practice, these methods can take a cell-by-gene expression matrix

and decompose it into two matrices: a factor-by-gene matrix where each factor represents an independent, coordinated source of heterogeneity and is associated with weights for each gene, and a cell-by-factor matrix that provides the activity level of each factor.

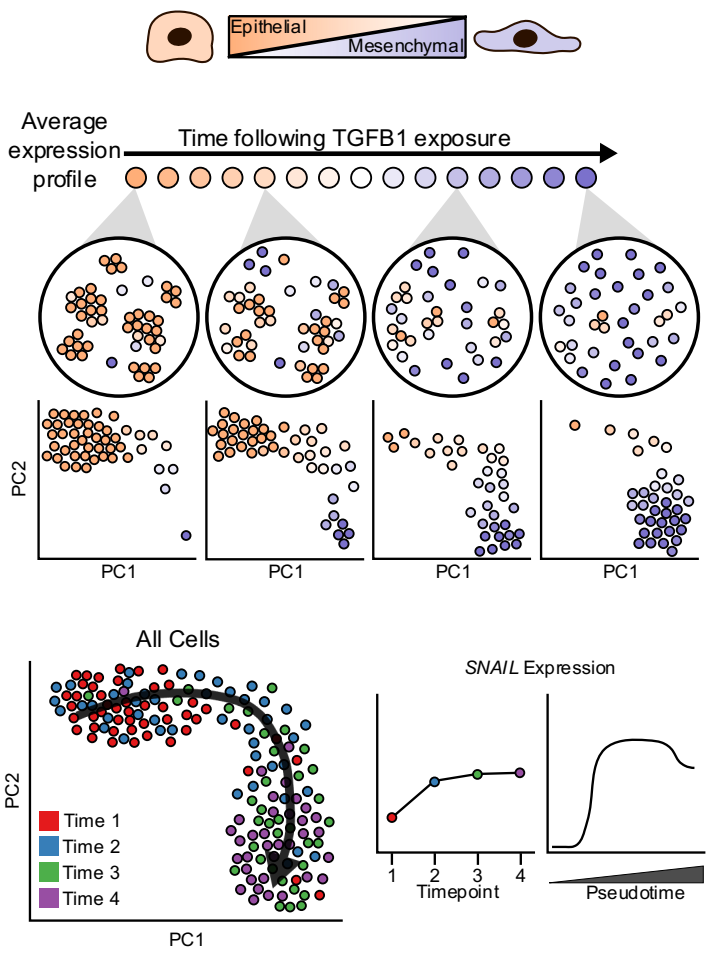

Figure 3. Using single-cell genomics to resolve EMT dynamics. As cells undergo an EMT, responses within an individual population of cells (eg. cell culture plates depicted in circles), responses are asynchronous. This intrasample heterogeneity is represented in scRNA-seq data. By sampling populations at discrete time points following EMT induction, a continuous EMT response trajectory can be inferred (bottom left schematic). By embedding individual cells along this trajectory rather than by discrete time point, continuous gene expression dynamics can be resolved.

1.8 Rationale and specific aims

Single-cell transcriptomics can provide the opportunity to define continuous transcriptional dynamics associated with the EMT response. Also, under the assumption that EMP is a prevalent source of intratumoural heterogeneity, scRNA-seq data from tumours should capture the gene expression patterns associated with it. To date, comparative analyses of EMP across contexts have been limited by the scale of studies. In these studies, **we aimed**

to apply single-cell genomics to understand molecular properties of EMP across contexts and determine if unifying molecular features exist. We approached this through three specific aims:

1. Compare transcriptional dynamics following experimental induction of the EMT in different contexts
2. Identify features of tumour-intrinsic EMP across diverse cancer types
3. Determine signalling dependencies associated with diverse EMP programs

Chapter 2: Materials and Methods

2.1 - Context specificity of the EMT transcriptional response

2.1.1 Cell Culture

A549, DU145, and MCF7 cells were obtained from ATCC (CCL-185, HTB-81, and HTB-22, respectively). OVCA420 cells were kindly provided by Dr. Gordon Mills (MD Anderson Cancer Center). All cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) with 4.5g/L glucose, L-glutamine, and sodium pyruvate (Corning, 10-013-CV), supplemented with 10% of fetal bovine serum (FBS) and cultured at 37^{o}C with 5% CO_2.

2.1.2 EMT time course experiments

For the OVCA420 pilot experiment, 25,000 cells were plated in wells of a 6-well plate. TGFB1 (10ng/mL; R&D Systems, #240-B-010) was added on a schedule such that all time points would complete at the same time for processing. Media was changed and replenished every 2 days to maintain TGFB1 concentrations. Cells were passaged as needed to ensure each condition was approximately 75% confluent at end point.

For each cell line of the multiplexed time course experiment, 10,000 cells were plated into each well of a 96-well plate according to the schematic in Figure 1a. The addition of recombinant TGFB1, EGF, and TNF were scheduled such that all timepoints completed at the same time for collection. Cells were treated with 10ng/mL TGFB1 (R&D Systems, #240-B-010), 30ng/mL EGF (Invitrogen, #PHG0311), or 10ng/mL TNF (Invitrogen, #PHC3015), which we had previously shown to promote morphological changes associated with EMT in preliminary experiments. Media was changed and fresh TGFB1, EGF, or TNF were added every two days to ensure relatively constant concentrations of these factors. Time course experiments also included treatment withdrawal samples, in which culture media was removed, cells were washed with phosphate-buffered saline (PBS), and replaced with fresh

media lacking EMT inducers. To avoid over-confluence throughout the experiments, cells were passaged as required, but not within the last two days of the time course to avoid artifacts at the time of collection. After the scheduled treatments, cells were immediately processed for scRNA-seq multiplexing.

The time course experiments were performed twice independently. Each time, the two time course replicates were performed in parallel, and on the second time through the experiment, two 10x libraries were generated for each plate replicate. Samples from the first replicate are labelled “Mix1” and “Mix2”, corresponding to the two plates running in parallel. Samples from the second replicate are labelled “Mix3a/b” and “Mix4a/b”.

2.1.3 Quantitative PCR (qPCR)

RNA was extracted using the RNeasy Mini Kit (Qiagen) and cDNA was synthesized using the OneStep RT-PCR Kit (Qiagen). The ABI 7500 FAST qRT-PCR machine (Applied Biosystems) was used for qPCR using the Taqman gene expression (Life Technologies) and SsoFast gene expression (Bio-rad) assays. *TBP* was used as an endogenous control for the assays. RQ (relative quantity) was determined using the cycling threshold for the gene of interest in control or untreated samples compared to the cycling threshold in experimental samples, calculated using the Applied Biosystems 7500 FAST v2.3 software.

2.1.4 Proliferation assay

OVCA420 cells ($5x10^4$/well) were seeded into 24-well tissue culture plates (Corning). Proliferation was assessed by counting viable cells based on trypan blue exclusion using the Vi-CELL XR cell viability analyzer (Beckman Coulter) for each day of the experiment.

2.1.5 Migration assay

OVCA420 cells (7x10^4/well) were plated into 96-well tissue culture plates (Essen ImageLock, Essen Instruments) with or without TGFB1 (10 ng/mL) 48 hours prior to scratch assay. The scratch assay was performed using the Wound Maker (Essen Instruments) and wound confluence was monitored every 2 hours with the Incucyte LiveCell Imaging System and software (Essen Instruments).

2.1.6 Invasion assay

OVCA420 cells were pre-treated with TGFB1 (10ng/mL) or a vehicle control (PBS + bovine serum albumin [BSA]) for four days prior to the assay. Cells were then plated at 5000 cells/well into the upper chamber of 24-well transwell chambers with 8um PET membranes coated with matrigel (Corning BioCoat Matrigel Invasion Chambers). Both top and bottom compartments of the chamber were filled with normal DMEM medium with 10% FBS. Invasive cells were assessed 24 hours after plating by staining the bottom membrane with crystal violet.

2.1.7 Multiplexing individual samples for scRNA-seq

Multiplexing was performed according to the MULTI-seq protocol (McGinnis et al., 2019), and reagents were kindly provided by Dr. Zev Gartner. Briefly, culture media was removed and cells in each well were washed with 1x PBS (Corning, #21-031-CV). Next, a lipid-modified DNA oligonucleotide (LMO) and a unique sample barcode oligonucleotide were added at 200nM to 0.05% trypsin with 0.53mM EDTA. This was added to each sample to be multiplexed, with each sample receiving a different sample barcode. Cells were incubated with this trypsin mixture for 5 minutes at 37^{o}C, and plates were gently mixed periodically. After 5 minutes, a common lipid-modified co-anchor was added to each well at 200nM to stabilize the membrane residence of the barcodes. Cells were incubated for an additional 5

minutes at 37°C with periodic mixing. After this labelling time, all cells were in suspension, lifted from the plate. The trypsin was then neutralized with cultured media, and the cells were mixed by pipetting to ensure a single cell suspension. Samples were then transferred to V-bottom 96-well plates, and pelleted at 400xg for 5 minutes. Barcode-containing media was removed, and the cells were then washed with PBS + 1% BSA. Washes were performed twice, and after the final wash, cells were resuspended in PBS + 1% BSA, pooled together, re-pelleted, and resuspended in PBS + 1% BSA. Viability and cell counts were then determined before preparation of the scRNA-seq libraries.

2.1.8 scRNA-seq library preparation and sequencing

Single-cell suspensions were processed using the 10x Genomics Single Cell 3' RNA-seq kit (v2 for time course experiments, v3 for kinase inhibition). Gene expression libraries were prepared according to the manufacturer's protocol. MULTI-seq barcode libraries were retrieved from the samples and libraries were prepared independently according to the MULTI-seq library preparation protocol (McGinnis et al., 2019). Briefly, barcode libraries are separated from the cDNA libraries during the first round of size selection in the 10x Genomics library preparation protocol and PCR-amplified prior to sequencing (McGinnis et al., 2019). Final libraries were sequenced on a NextSeq500 (Illumina). Expression libraries were sequenced so that time course libraries reached an approximate depth of 40,000-50,000 reads per cell (for the v2 scRNA-seq kit), and 20,000-25,000 reads per cell for the kinase inhibitor experiment (v3 scRNA-seq kit). For the time course data, we detected a median of 3649 genes and 17,330 unique molecular indices (UMIs) per cell, and for the kinase inhibitor screens, we detected a median of 2360 genes and 7634 UMIs.

2.1.9 Processing of raw sequencing reads

Raw sequencing reads from the gene expression libraries were processed using CellRanger (Zheng et al., 2017) v2.2.0 for the time course data, and v3.0.2 for the kinase inhibitor data. The GRCh38 build of the human genome was used for both. Except for explicitly setting --expect-cells=25000, default parameters were used for all samples. MULTI-seq barcode libraries were simply trimmed to 26bp (v2 kit) or 28bp (v3 kit) using Trimmomatic (Bolger et al., 2014) (v0.36) prior to demultiplexing.

2.1.10 Demultiplexing expression data with MULTI-seq barcode libraries

Demultiplexing was performed using the deMULTIplex R package (v1.0.2) (https://github.com/chris-mcginnis-ucsf/MULTI-seq). The key concepts for demultiplexing are described in McGinnis *et al.* (2019). Briefly, the tool takes the barcode sequencing reads and counts the number of times each of the 96 barcodes appears for each cell. Then, for each barcode, it assesses the distribution of counts in cells and determines an optimal quantile threshold to deem a cell positive for a given barcode. Cells positive for more than one barcode are classified as doublets and are removed. Only cells positive for a single barcode are retained for downstream analysis. As each barcode corresponds to a specific sample in the experiment, the sample annotations can then be added to all cells in the data set.

2.1.11 Data quality control and processing

Quality control was first performed independently on each 10x Genomic library, and all main processing steps were performed with Seurat v3.0.2 (Butler et al., 2018). Expression matrices for each sample were loaded into R as Seurat objects, only retaining cells with more than 200 genes detected. Cells with a high percentage of mitochondrial gene expression were also removed. We then subsetted the data, making independent Seurat

objects for each time course or kinase inhibition experiment (ie. for all independent cell line and EMT inducer combinations). Each condition was then processed independently with a standard workflow. We first removed genes detected in fewer than 1% of the cells for the given experiment. The expression values were then normalized with standard library size scaling and log-transformation. The top 3000 variable genes were detected using the vst selection method in Seurat. Expression values were scaled and the following technical factors were regressed out: percentage of mitochondrial reads, number of RNA molecules detected, cell cycle scores, and for the time course data, batch was also included. For initial exploration, PCA was run on the variable genes, but all UMAP embeddings included in figures are based on PCA run on genes used for pseudotemporal ordering of cells. UMAP embeddings were calculated from the first 30 principal components.

2.1.12 Pseudotemporal ordering of cells

Pseudotime models for each time course experiment were built using the R package psupertime v0.2.1 (Macnair and Claassen, 2019) on the top 3000 variable genes from each condition. Psupertime is based on ordinal logistic regression, taking scRNA-seq data with sequential labels and identifying a linear combination of genes that places the cells in the specified label order. To build the pseudotime model for each time course, we first omitted the treatment withdrawal samples. Because psupertime is based on regression, however, pseudotime values for new data can be calculated by simply performing matrix multiplication between the coefficient matrix of the pseudotime model and the expression matrix of the new data. We used this approach to calculate pseudotime values for both the treatment withdrawal samples of the time course experiment. We also used the time course models to calculate pseudotime values for the respective kinase inhibition experiments. As the range of pseudotime values can vary between conditions, we simply rescaled the values from 0-1 in cases where multiple models were compared in the same figure.

2.1.13 Differential expression analysis

For time course experiments, expression dynamics of each gene, or transcription factor regulon score, as a function of pseudotime was modelled using the generalized additive model function provided by the R package mgcv with the model exp ~ s(pseudotime, k=4) + batch, with the smoothing parameter estimation method set to restricted maximum likelihood (method="REML"). The number of basis functions (k) was chosen such that the residuals were randomly distributed. P-values associated with the smoothed pseudotime function for each gene were adjusted using the p.adjust() function in R with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). As many genes may significantly vary throughout pseudotime but have low effect sizes, we only evaluated significant genes (adjusted p-value < 0.05) that are also within the top 2000 variable genes of each time course experiment. While others may be biologically relevant, their signal in the data is often too low to assess reliably.

When assessing transcription factor activity and cytokine production, we were more generally interested in assessing the directionality of change over pseudotime, so in these cases, we used the same approach, but removed the smoothing function from the model. This allowed us to report the single coefficient associated with the pseudotime covariate, representing whether activity generally increased or decreased throughout the transition.

For the kinase inhibition experiment, we assessed the number of differentially expressed genes in cell lines treated with a kinase inhibitor, but no EMT inducer. For this, we again used the gam() function provided by the mgcv package with the model exp ~ inhibitor, setting the no-inhibitor controls as the intercept. We then quantified the number of genes with an adjusted p-value < 0.05.

2.1.14 Calculating smoothed expression trends

To calculate smoothed expression trends over pseudotime, we used models used for differential expression, but calculated the fit values for 200 evenly spaced pseudotime values ranging between the minimum and maximum pseudotime values.

2.1.15 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed using the R package fgsea (Korotkevich et al., 2021). Input genes were ranked either by their variance values after the variance stabilizing transformation (vst), computed by Seurat's FindVariableFeatures() function, or by adjusted p-value from the differential expression analysis. Reference gene sets were collected from the Molecular Signatures Database (MSigDB) v6.2.

2.1.16 Gene set scoring

Gene set scoring of the EMT hallmark gene set and the KEGG pathway "Cytokine-cytokine receptor interaction" was performed using the AddModuleScore() function provided by the Seurat package. Default parameters were used.

2.1.17 Transcription factor regulon scoring of single cells

Regulon scores for individual cells were computed using the SCENIC workflow (Aibar et al., 2017). Log-transformed expression values for each time course experiment were used as input into the command-line interface functions of pySCENIC. First, gene regulatory networks were computed using the grnboost2 method in the grn function. Next, enriched motifs were identified using the ctx function, providing the cisTarget v9 databases of regulatory features 500bp upstream, 5kb centered on the TSS, and 10kb centered on the TSS. Finally, individual cells were scored for motifs using the aucell function.

2.1.18 Identifying over-represented transcription factor motifs in gene lists

The R package RcisTarget (Aibar et al., 2017) was used to identify enriched transcription factor motifs associated with gene lists, using the cisTarget v9 transcription factor motif annotations and the hg19-tss-centered-10kb-10species.mc9nr database of motif rankings. To compare enrichment between two gene lists, we calculated the difference in normalized enrichment scores for motifs between the two lists and ranked motifs to identify uniquely enriched motifs.

2.1.19 ATAC-seq sample preparation and analysis

Assay of transposase-accessible chromatin (ATAC)-seq samples were prepared from OVCA420 cells treated with 10ng/mL of TGFB1 for 0, 1, 3, or 7 days, and the experiment was performed independently twice. Sample preparation was performed as described by Buenrostro *et al.* (2015). Briefly, nuclei were extracted from 50,000 cells per sample and chromatin was tagmented using the TDE1 transposase provided in the Nextera DNA Library Preparation Kit (Illumina). While the original protocol recommended 2.5µL of enzyme, we found that optimal tagmentation of these samples required 5µL of enzyme at 37°C for 30 minutes with gentle mixing. Finally, ATAC libraries were amplified and sequenced on a NextSeq500 150-cycle high output run, yielding approximately 50M reads per sample.

Raw reads were aligned to the hg38 build of the human genome using Bowtie2 (Langmead and Salzberg, 2012) and peaks were called using MACS2 (Zhang et al., 2008) with the following parameters: -q 0.01 --nomodel --shift -100 --extsize 200 -B --SPMR --broad. Differential motif accessibility was calculated using the R package chromVAR (Schep et al., 2017). Briefly, the summits of peaks from all samples were merged, and expanded to a 250bp window, centered on the summit. Motifs from the human_pwms_v2 list included with the package were mapped to the peaks using the matchMotifs() function and then

deviations across samples were computed. Significant deviations in motif accessibility were identified using the differentialDeviations() function.

2.1.20 Pooled CRISPR interference and activation screens

A549 and DU145 cells were transduced with lentivirus (with 0.8µg/mL polybrene) carrying either i) CRISPRi: UCOE-SFFV-dCas9-KRAB (Addgene: 85969) or ii) CRISPRa: dCas9-10xGCN4 (Addgene: 60903) and scFV-sfGFP-VP64 (Addgene: 60904). Transduced cells were purified by fluorescence-activated cell sorting (FACS) based on fluorescence of selectable proteins for each vector (CRISPRi: BFP+, CRISPRa: BFP+/GFP+). Design of single-guide RNAs (sgRNAs), library production, titering, and quality control was performed as a service by Millipore Sigma. Pooled sgRNA lentiviral libraries were transduced into cells at an MOI of 0.5 to minimize the number of cells with >1 integration. For CRISPRa experiments, cells were collected 9 days following infection. For CRISPRi experiments, cells were treated with TGFB1 (10ng/mL) or a vehicle control 3 days following transduction of the library and were cultured for 7 additional days. Single-cell suspensions were processed for scRNA-seq on the 10x Genomics Chromium platform (3' scRNA-seq v3 kit with 3' feature barcoding). Final libraries were sequenced on two NovaSeq S4 lanes (Illumina). Gene expression and sgRNA quantifications were calculated from raw sequencing files using CellRanger (10x Genomics).

2.1.21 Data and code availability

Raw sequencing files and processed UMI count matrices have been deposited in the NCBI Gene Expression Omnibus under the accession GSE147405. All code used to process data and generate figures is available on a public Github repository at https://github.com/dpcook/emt_dynamics.

2.2 - Transcriptional census of EMP in cancer

2.2.1 Preparation of scRNA-seq data

At the time of preparing these studies, we collected all published scRNA-seq data of solid, human tumours that had been generated using droplet-based scRNA-seq to avoid comparing data collected by vastly different technologies. Several data sets included matched normal tissue samples. In these cases, we removed the normal samples and only proceeded with the tumour samples. Raw UMI count matrices and cell metadata were collected from the following sources:

Cancer	Source	Accession	Pre-annotated?
Breast	Wu *et al.* (2020)	ENA accession PRJEB35405	Yes
Breast	Qian *et al.* (2020)	http://blueprint.lambrechtslab.org	Yes
Colorectal	Lee *et al.* (2020)	GEO Accession GSE144735 & GSE132465	Yes
Colorectal	Uhlitz *et al.* (2021)	Direct from authors	No
Colorectal	Qian *et al.* (2020)	http://blueprint.lambrechtslab.org	Yes
Gastric	Sathe *et al.* (2020)	https://dna-discovery.stanford.edu	No
Lung	Kim *et al.* (2020)	GEO Accession GSE131907	Yes
Lung	Lambrechts *et al.* (2018)	ArrayExpress Accessions E-MTAB-6149 & E-MTAB-6653	No
Lung	Qian *et al.* (2020	http://blueprint.lambrechtslab.org	Yes
Lung	Laughney *et al.* (2020)	GEO Accession GSE123904	Yes
Ovarian	Geistlinger *et al.* (2020)	GEO Accession GSE154600	No
Ovarian	Qian *et al.* (2020)	http://blueprint.lambrechtslab.org	Yes
Pancreatic	Steele *et al.* (2020)	GEO Accession GSE155698	No
Squamous cell carcinoma	Ji *et al.* (2020)	GEO Accession GSE144236	Yes

Uveal Melanoma	Durante *et al.* (2020)	GEO Accession GSE139829	No

Table 1. Sources of scRNA-seq data of solid tumours.

Initial quality control was performed independently for each data set using the R package Seurat v4.0 (Hao et al., 2020). Cells with fewer than 200 detected genes were removed and only genes detected in more than 3 cells were included in the analysis. Cells with a high percentage of mitochondrial transcripts were also removed. As the distribution of mitochondrial reads can vary significantly between sample preparations, we manually defined thresholds for each data set based on a heuristic of “trimming the skew” to result in an approximately normal distribution. However, thresholds were similar between datasets, ranging from 10-25%. All data sets were normalized using SCTransform (Hafemeister and Satija, 2019) and the percentage of mitochondrial reads was regressed out to remove any associated variation. The data was then processed with principal component analysis (PCA), a nearest neighbor graph was generated on the first 30 PCs and the data were clustered at a fairly low resolution (FindClusters, resolution = 0.2) using the Louvain algorithm. For visualizations presented in the figures, UMAP embeddings were generated from the first 30 PCs

For the majority of data sets, we found that non-malignant cell types from different tumours within the same data set often clustered together and coarse-grained cell type annotation did not require data integration. However, for 2 of the data sets, we found significant batch effects between samples. For these samples, we performed integration using the default pipeline with SCTransform normalization implemented in Seurat v4 (Hao et al., 2020). Following integration, the data was re-clustered from the aligned embeddings and we found that coarse cell type annotations based on canonical markers matched the clustering well.

2.2.2 Cell type annotation

Cell types were annotated largely based on the expression of canonical markers and supported by annotations provided by the original data source if provided. The identification of cancer cells was also supported by the trend that non-malignant cells typically cluster well between tumours, whereas the diversity of malignant cells causes them to frequently cluster separately from each other. Cancer cells were identified as clusters with clear expression of epithelial markers (eg. *EPCAM*, *KRT19*). For analyses including non-malignant cell types, we only aimed to achieve a coarse-grained annotation. We defined clusters associated with T cells (*CD3E*), B cells (*MS4A1*), NK cells (*NKG7, CD3E* negative), macrophages (*CD14*), endothelial cells (*CLDN5*), and fibroblasts (*COL1A1, ACTA2*). Clusters that were not clearly positive for one of the canonical markers were simply annotated as “Other”.

2.2.3 Identifying latent EMP expression programs with archetypal analysis

Samples with <100 annotated cancer cells were removed from the analysis. Multi-resolution archetypal analysis was performed independently on the cancer cells from all 160 tumours using ACTIONet v2.0.15 (Mohammadi et al., 2020) to decompose cells’ gene expression profiles into a small set of latent expression programs that are heterogeneously expressed throughout the population. Reduced kernel matrices were first computed with the reduce.ace() function implemented in the R package ACTIONet with the parameter reduced_dim=25. Given that each population represented a single cell type, ACTIONet was then run with the k_max=6 option to reduce the maximum depth of decompositions and with min_cells_per_arch=5 to prevent archetypes driven by a small number of cells.

Resulting archetype footprints (program activities) were correlated with gene set scores for 10 EMT gene sets. Clustering of the Pearson correlation coefficients allowed us to define EMP-associated programs as clusters with high correlation values. We then used linear

models to identify genes whose expression is associated with all latent programs. To only recover reliable changes, differential expression was limited to the top 2000 variable genes with a minimum detection frequency of 5% in each cancer cell population. Pearson residuals from SCTransform's model were used for differential expression, modelling each gene's residuals as a function of program activity and program-associated genes were defined as those with a Benjamini-Hochberg-corrected P-value of <0.05 and a model coefficient (effect size) of >0.5. For generating a conserved EMP signature, we first modified these thresholds slightly and only considered genes with a model coefficient of >2 in at least one EMP program.

2.2.4 Gene set scoring

Gene set scoring was performed using Seurat's AddModuleScore function using default options, which calculates the average expression of genes within a gene set relative to control genes of similar average expression levels.

2.2.5 Cell type specificity scoring

Specificity scores were calculated for EMP signature genes using the R package genesorteR v0.4.3 (Ibrahim and Kramann, 2019). Specificity scores represent the exclusivity of a gene to a given cluster and extent to which it is expressed (proportion of cells with detection greater than population median levels). Values range from 0-1, with 1 corresponding to genes that are exclusive to a given cluster and detected in all cells.

2.2.6 Pan-cancer TCGA analysis

Bulk RNA-seq profiles and clinical outcomes were accessed from https://gdc.cancer.gov/node/905, tumor purity estimates were acquired from Aran *et al.* (2015), and immune cell proportion estimates were accessed from Thorsson *et al.* (2019).

RNA-seq count data for each sample was normalized to counts per million (CPM) and log-transformed. EMP signature scores were calculated for each sample by computing the average gene-level Z-score of the genes from the 165 cancer cell-specific genes. The association of EMP signature activity with progression free interval was assessed using a Cox proportional hazards model, including tumour type, purity, age, and sex along with continuous EMP scores as covariates in the model. Changes in immune cell proportions were assessed by independently using linear models to model each cell type's predicted proportion as a function EMP signature score, including tumour type as a covariate. We used the Benjamini-Hochberg method to correct p-values to account for multiple comparisons.

2.2.7 Inferring EMP-associated signalling activity

The R package PROGENy v1.11.2 (Holland et al., 2020a, 2020b; Schubert et al., 2018) was used to infer the activity of 14 signalling pathways in each individual cell. For each tumour sample, Pearson residuals from the SCTransform model were used to calculate pathway activity with the progeny() function with the top=500 parameter set to use the top 500 genes of each pathway in PROGENy's model for the activity calculation. We used simple linear models to identify changes in signalling pathway activity with latent program activity.

2.2.8 Cell communication inference

The R package nichenetr (NicheNet; v1.0.0) (Browaeys et al., 2020) was used to infer cell communication within the tumour microenvironment that could contribute to a sample's specific EMP program. Given the diversity of EMP programs, this was performed on all 160 tumours independently. For each sample, cancer cells annotated by ACTIONet as the EMP-associated archetype were defined as the "receiver" population and all cell types were

considered as potential “sender” cells. The “gene set of interest” was defined as that sample’s specific EMP program and expressed genes were defined as those with a detection rate of at least 5%. For each sample, we considered top ligands as the top 20 ligands inferred to promote expression of the EMP-associated genes.

2.3 - Paracrine signalling dependencies of EMP

2.3.1 Kinase inhibitor screen

For each cell line, 10,000 cells were plated into four 96-well plates according to the schematic in Figure 34b. Cells were simultaneously treated with small molecule kinase inhibitors (listed in Figure 34b) and either 10ng/mL TGFB1, 30ng/mL EGF, or 10ng/mL of TNF. No-inhibitor and No-EMT-inducer controls were also included for all conditions. All inhibitors were used at a final concentration of 1μM (Cayman Chemical Kinase Screening Library, Item No. 10505, Batch No. 0537554). EMT inducers and kinase inhibitors were refreshed daily after replacing the culture media. After 7 days of treatment, all samples were immediately processed for scRNA-seq multiplexing, library preparation, and sequencing (see sections 2.1.7 and 2.1.8).

2.3.2 Assessing effects of small molecule inhibitors on EMP

ScRNA-seq data from epithelial cancer cell lines in control and drug-treated culture conditions were acquired from McFarland *et al.* (2020a) and Cook & Vanderhyden (2020). Initial data processing was performed identically to tumour samples and cell lines with fewer than 100 measured cells were removed. EMP programs were defined for each cell line using ACTIONet with only data from control cultured cells. Genes associated with each program were defined and used as a sample-specific EMP gene set for scoring both control

and drug-treated cells. We then used a linear model to compare the effects of MEK and TGFBR inhibition on EMP in each line.

2.3.4 Inferring transcription factor activity

Regulon scores for individual cells were computed using the SCENIC workflow (Aibar et al., 2017). Raw UMI counts from each tumour sample were used as input into the command-line interface functions of pySCENIC. First, gene regulatory networks were computed using the grnboost2 method in the grn function. Next, enriched motifs were identified using the ctx function, providing the cisTarget v9 databases of regulatory features 500bp upstream, 5kb centered on the TSS, and 10kb centered on the TSS. Finally, individual cells were scored for motifs using the aucell function. Using a simple linear model, resulting regulon scores for each cell were modelled as a function of the cells' archetype scores that correspond to the sample's EMP program. Only regulons with a significant association (Benjamini-Hochberg $p < 0.05$) with the EMP program were considered. The model coefficients were used to determine the direction and extent of a transcription factor's activity.

Chapter 3 - Results

3.1 Context specificity of the EMT transcriptional response

3.1.1 TGFB1 promotes a canonical EMT in OVCA420 ovarian cancer cells

As a preliminary trial to determine if scRNA-seq could be a valuable tool to measure EMT transcriptional responses, we first confirmed the EMT response of the OVCA420 ovarian cancer cell line following treatment with TGFB1 (10ng/mL). Following treatment, we observed progressive morphological changes consistent with an EMT: cell-cell junctions broke down, the cells flattened and extended cytoplasmic projections (**Figure 4A**). Expression of the epithelial *KRT19* decreased, while various mesenchymal markers increased (*VIM*, *CDH2*, *ZEB1*) (**Figure 4B**). After scratching a "wound" through confluent wells of cells, TGFB1-treated cells had increased motility, closing the wound at a faster rate (**Figure 4C**). This change is not attributable to increased proliferation as the rate of proliferation in TGFB1-treated cells is slower than control cells (**Figure 4D**). We also found that TGFB1-treated cells had an increased invasive capacity, moving through a matrigel-coated chamber while control cells failed to do so (**Figure 4E**). Together, these findings confirm that TGFB1 promotes a typical EMT response in OVCA420 cells.

3.1.2 Single-cell RNA sequencing resolves continuous transcriptional dynamics of the EMT

We next used scRNA-seq to measure the transcriptional changes following TGFB1 treatment of OVCA420 cells. We performed a time course experiment, processing cells following 1, 3, or 7 days following TGFB1 treatment along with an untreated control. Cell suspensions were processed with the 10x Genomics Chromium system, generating scRNA-seq data for 10,489 cells from the four time points. We sequenced the samples deeply to an average depth of over 100,000 reads per cell (median UMI/cell = 35,938) to ensure high coverage of the cells' transcriptome.

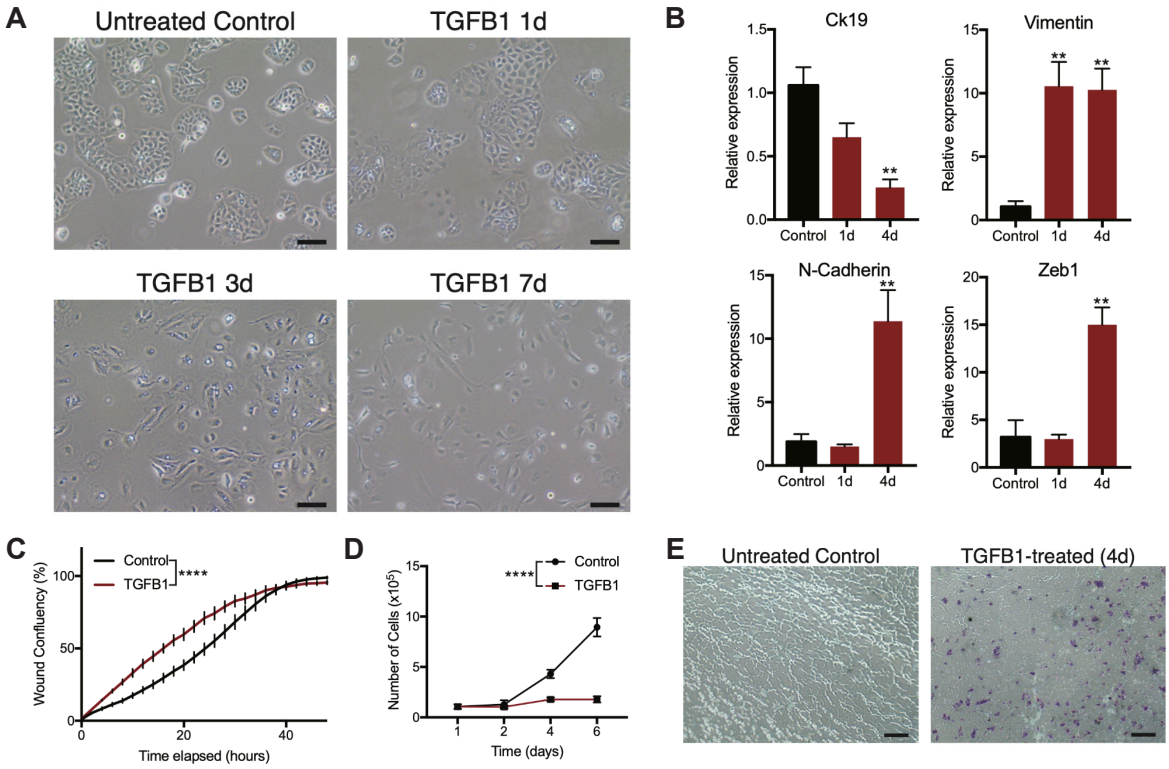

Figure 4. TGFB1 promotes a stereotypical EMT response in OVCA420 cells. A) Phase contrast images of OVCA420 cells following 0, 1, 3, or 7 days of TGFB1 treatment (10ng/mL). Scale bar=100μm. **B)** qPCR expression data of epithelial and mesenchymal markers after 1 and 4 days of TGFB1 treatment. ** $p < 0.01$ compared to control samples (ANOVA, Dunnett's post-test). Error bars represent SEM (n=3). **C)** Scratch-wound migration measuring wound closure over 48 hours. **** interaction $p < 0.0001$ (Two-way ANOVA). Error bars represent SEM (n=3). **D)** Comparing the growth rates of untreated cells against those that had previously been treated with TGFB1 for 4 days. **** interaction $p < 0.0001$ (Two-way ANOVA). **E)** Crystal violet-stained membranes from a matrigel-coated Boyden chamber. Positive staining highlights cells that had successfully invaded through the matrigel.

We performed PCA on the data and found that the cells from each time point formed a continuous trajectory along the first two principal components (PCs) (**Figure 5A**). Throughout treatment, cells progressed along PC1, whereas PC2 represents a transient response associated with early time points that resolves by 7 days following TGFB1 treatment (**Figure 5B**). To explore gene expression patterns associated with these two components, we assessed genes' loadings for each. Genes with positive loadings for PC1 (ie. higher expression at later time points) included several canonical EMT markers,

including *IL32*, *KRT17*, *SOX4*, and *TGFB1*, while several epithelial markers (eg. *KRT19*, *ELF3*) were among the genes with the most negative loadings (**Figure 5C**). We performed GSEA on genes ranked by their loadings and found that increases along PC1 involve activation of genes associated with EMT and reduced expression of oxidative metabolism (**Figure 5D**).

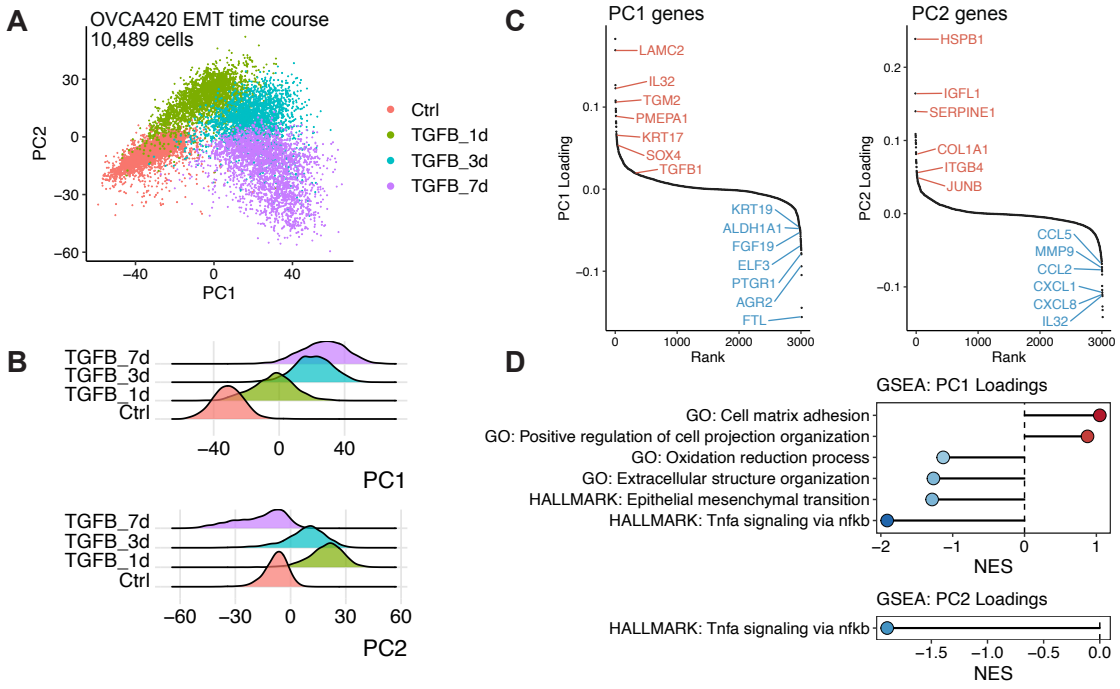

Figure 5. scRNA-seq resolves transcriptional dynamics of the TGFB1-induced EMT in OVCA420 cells. A) PCA embedding of scRNA-seq data from 10,489 cells following 0, 1, 3, or 7 days of TGFB1 treatment. Each point represents an individual cell. **B)** Density plots showing the distribution of cells from each time point along PC1 (top) and PC2 (bottom). **C)** Ranked PC1 (left) and PC2 (right) loadings for genes included in the PCA. Loadings represent the contribution of a gene's expression to a cell's position along that PC. Select genes are highlighted. **D)** Gene set enrichment analysis results of loading-ranked genes for each PC. All gene sets shown are significantly enriched (Benjamini-Hochberg-adjusted $p < 0.05$) and plots show the normalized enrichment score (NES) for each gene set. Negative NES values are associated with genes that reduce a cell's PC value, while positive values are the opposite.

Genes with a positive contribution to PC2 seem to represent a stress signature (*HSPB1, JUNB*), but negatively loaded genes, which are most highly expressed in cells after 7 days

of TGFB1 treatment, include a large number of cytokines and chemokines (*IL32*, *CXCL8*, *CXCL1*, *MM9*) (**Figure 5C**). This suggests that immunoregulatory features of EMT may only be associated with more extreme mesenchymal phenotypes that occur later in the response. Interestingly, GSEA of PC2 loadings only resulted in a single enriched term (TNFa signalling) that was associated with the negatively loaded genes (**Figure 5D**). However, we found that the acute response of cells following 1 day of treatment (higher PC2 scores) were associated with a dramatic reduction in the proportion of proliferating cells that is largely resolved by 7 days following treatment (**Figure 6**).

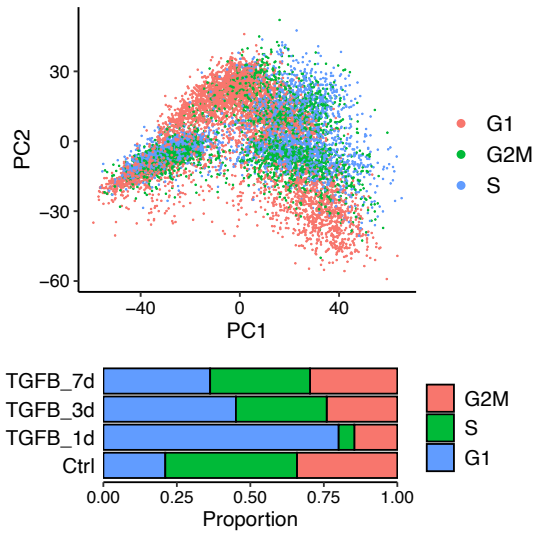

Figure 6. Cell cycle features of OVCA420 cells undergoing an EMT. Top: Same PCA as in Figure 5a, but colored by the inferred cell cycle stage of each cell. Cell cycle stages are predicted based on the relative expression of known S and G2M genes in each cell. Bottom: Plot showing the proportion of G1/S/G2M cells in each time point.

This pilot experiment highlighted the abilities of scRNA-seq to reveal the dynamics of the EMT response. While it provided valuable insight into this response in OVCA420 cells, it is unclear how these findings translate to different cancer types and different EMT inducers.

3.1.3 MULTI-seq enables comparative analysis of the EMT

To assess transcriptional dynamics of the EMT across a variety of contexts, we used MULTI-seq (McGinnis et al., 2019) to generate scRNA-seq data from twelve distinct EMT time course experiments. We assessed four different cancer cell lines capable of

undergoing an EMT (A549, lung; DU145, prostate; MCF7, breast; OVCA420, ovarian) and exposed each to known EMT-inducing factors: TGFB1, EGF, and TNF. These cell lines were chosen because they all have an epithelial morphology in culture, have been shown to undergo an EMT in previous studies (Dalmau et al., 2015; Dong et al., 2007; Kasai et al., 2005; Li et al., 2015; Lu et al., 2003; Osborne et al., 2014; Sun et al., 2018), and represent four distinct cancer types. The specific inducers were chosen as they all have been previously shown to promote an EMT in different cell lines—including those used in this study in most cases (Dalmau et al., 2015; Dong et al., 2007; Kasai et al., 2005; Li et al., 2015; Lu et al., 2003; Osborne et al., 2014; Sun et al., 2018)—and they initiate independent signalling pathways. In response to these factors, each cell line exhibited morphological changes, consistent with an EMT (**Figure 7**). Qualitatively, we observed that different inducers can promote different morphologies in the same cell line (eg. DU145 with TGFB1 or TNF), and some changes were modest in comparison to others (eg. MCF7 cells treated with EGF). Cells that fail to acquire the typical spindle-shaped morphology, however, does not preclude other EMT traits. For example, at higher doses, EGF has been shown to promote an EMT associated with a circular morphology (Devaraj and Bose, 2019). Ultimately, it is likely that these differences arise from subtleties in the expression programs initiated by each inducer, and particularly the different expression dynamics of various cytoskeletal and extracellular matrix proteins.

For each of the 12 conditions, samples were collected at 5 distinct time points from 8 hours to one week after treatment, and three additional time points from 8 hours to 3 days after the EMT-inducing stimulus had been removed (**Figure 8A**). The 3-day withdrawal time point was chosen based on preliminary data suggesting transcriptional reversion in as few as 3 days. In the aggregated data, expression profiles clustered dominantly by cell line, and after demultiplexing, the majority of cell line annotations (95.8% on average) were restricted to a

dominant cluster, demonstrating robust multiplexing (**Figure 8B,C**). In total, we annotated 58,088 single cells from across 576 samples, comprising six replicates of the twelve time course experiments (**Figure 8D**). Replicates were highly correlated, supporting the consistency of the experimental procedures and processing workflow.

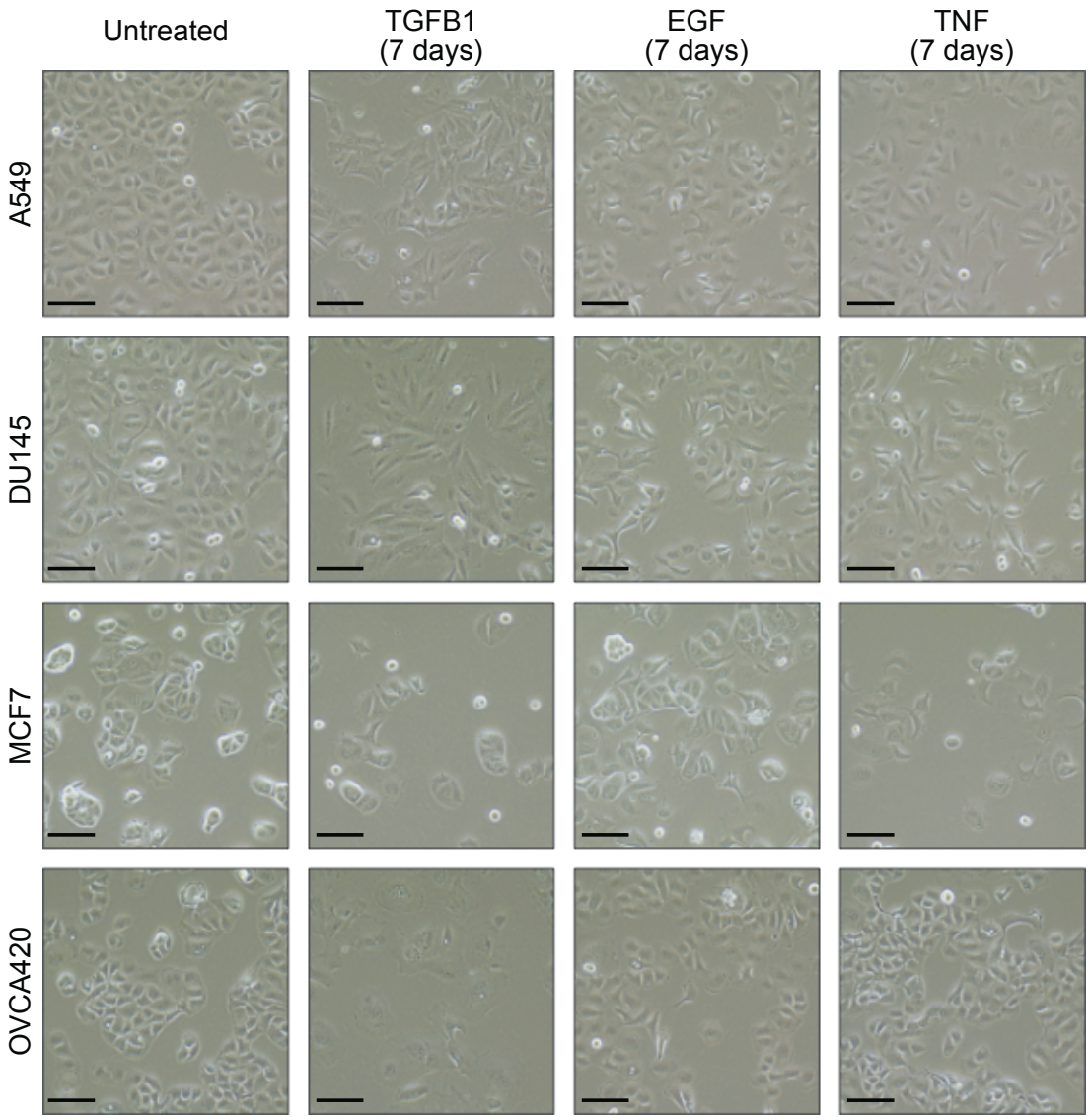

Figure 7. Morphology diverse cell types following EMT induction. Phase contrast images of A549, DU145, MCF7, and OVCA420 cells treated with TGFB1, EGF, or TNF. Displayed images are representative fields of all replicates. Scale bar = 100μm.

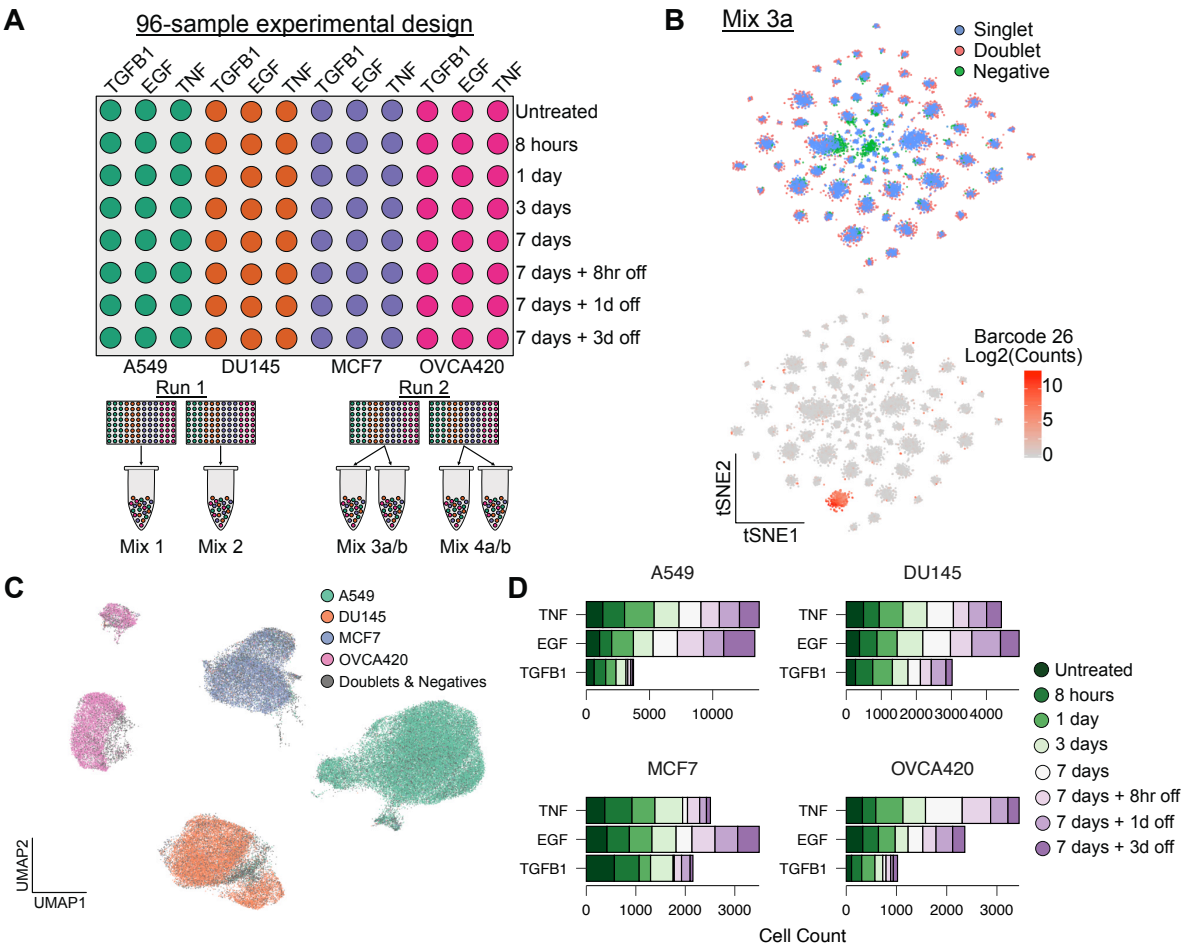

Figure 8. Multiplexed scRNA-seq profiling of twelve EMT time course experiments. A) Schematic of the 96-well experimental design for the 12 EMT time course experiments. **B)** *t*-SNE embeddings of the MULTI-seq barcode counts, demonstrating strong signal for demultiplexing. Each discrete group corresponds to a population of cells with similar barcode profiles and therefore likely belong to a common original sample. **C)** UMAP embeddings of scRNA-seq data from multiplexed EMT time course experiments. Cells are coloured by their annotation after demultiplexing. Individual cell lines are contained within discrete clusters, supporting the accuracy of samples annotations. **D)** Plot showing the number of cells captured for each condition of the 12 time course experiments.

3.1.4 Transcriptional dynamics of the EMT are highly context specific

We next assessed the temporal progression of each time course. In each case, time-dependent shifts in cells' expression profiles were evident, and withdrawal samples showed reversion back towards the untreated state (**Figure 9**). In each cell line, receptors for the three EMT inducers were detectable, explaining these dynamics (**Figure 10**). While the top

1000 variable genes for each time course showed some commonalities between cell lines, context-dependent changes were dominant (**Figure 11A**). GSEA of variance-ranked genes for each time course did, however, demonstrate enrichment for the MSigDB Hallmark EMT gene set in all conditions (Liberzon et al., 2015) (**Figure 11B**). This is consistent with the morphological changes we observed, and further supports that these changes are associated with an EMT response. The minimal overlap of top variable genes among conditions suggests that the specific EMT genes involved in the response may vary.

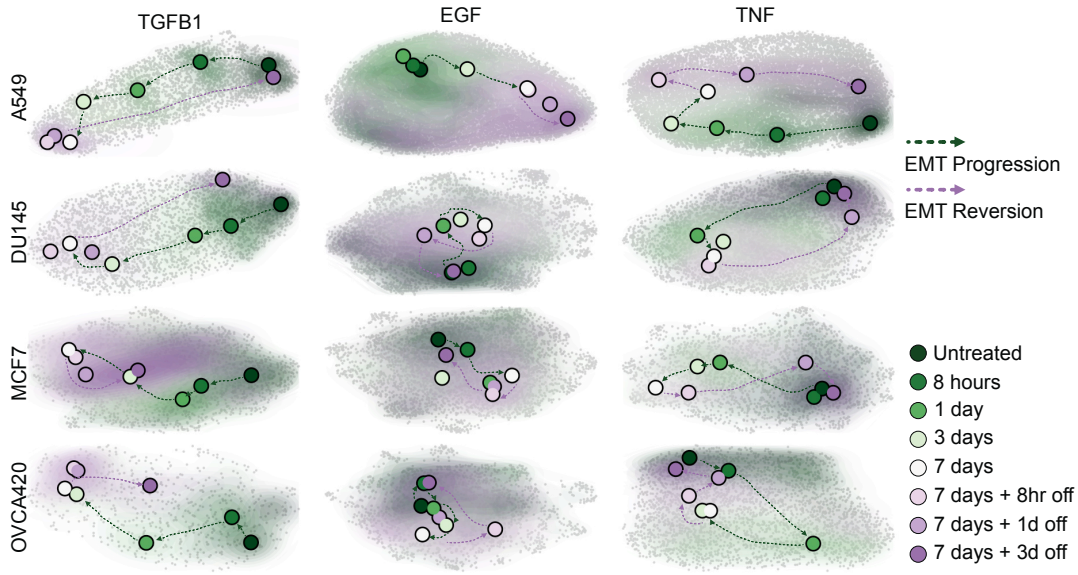

Figure 9. scRNA-seq resolves transcriptional responses throughout the EMT. UMAP embeddings of each of the 12 time course experiments. Grey dots correspond to individual cells, shaded regions represent the related sample density for each timepoint, and coloured dots correspond to the maxima of the density function for each time point.

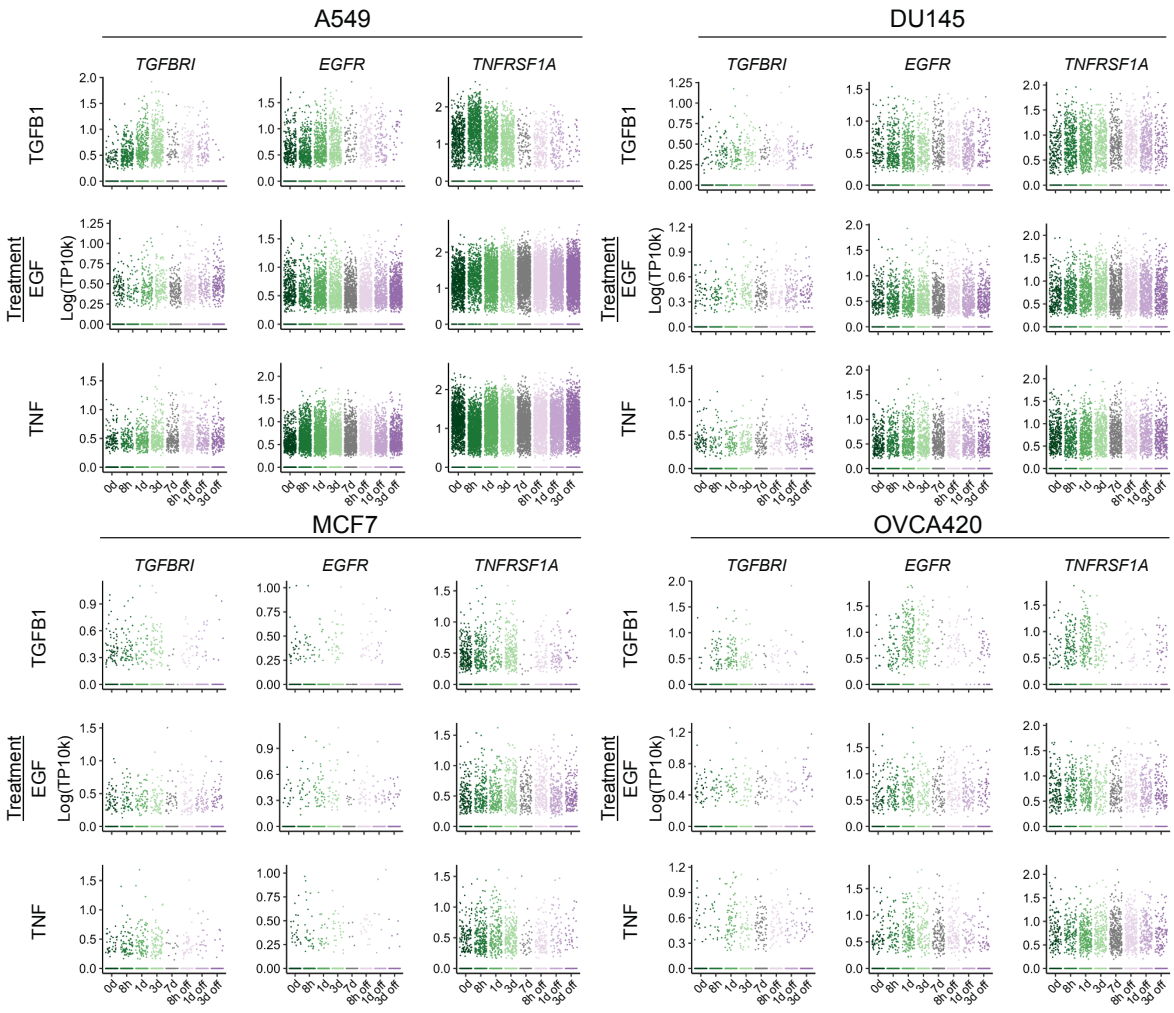

Figure 10. Expression of receptors for EMT inducers. Plots showing the expression levels of *TGFBR1*, *EGFR*, and *TNFRSF1A* in each time course experiment. Expression values are log-transformed counts per 10k molecules (TP10k). Each point corresponds to a single cell.

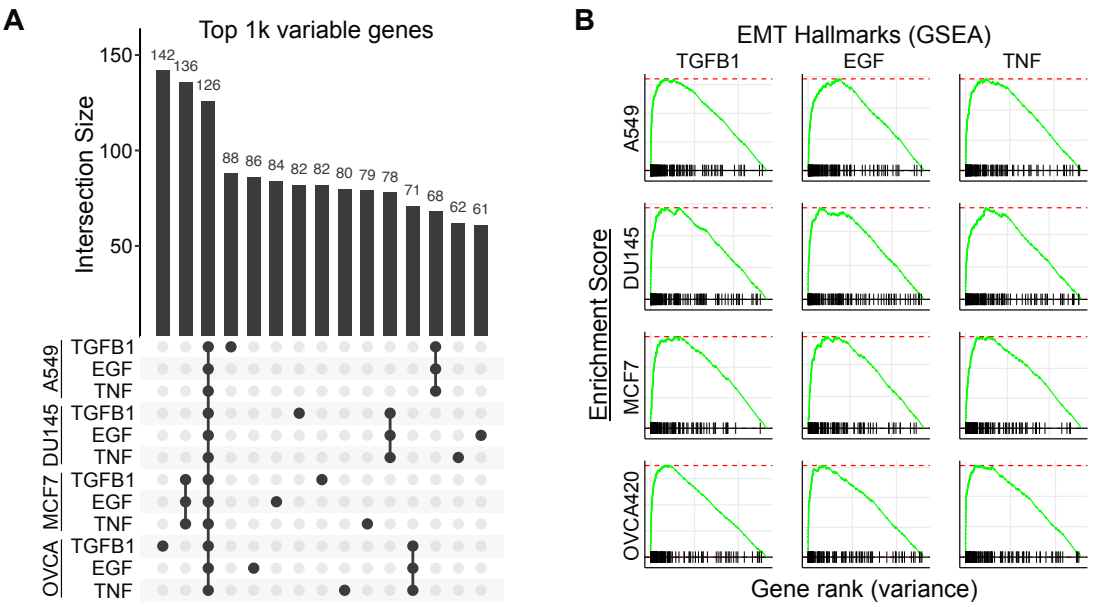

Figure 11. EMT-associated heterogeneity in scRNA-seq data. A) UpSet plot showing the intersections of the top 1000 variable genes of each time course experiment. The top portion of the plot shows the number of genes in common between the samples highlighted by black dots in the bottom portion. **B)** GSEA plots showing the enrichment for the EMT Hallmark genes in the variance-ranked genes for each time course. A skewed representation of the gene set at the left side of the ranking suggests EMT-associated genes are variably expressed across the cells of an individual time course experiment.

To specifically compare temporal dynamics of the EMT, we first pseudotemporally ordered the cells from each condition (**Figure 12A**). In each time course, cells progressively transitioned throughout the full 7 days of EMT induction, and withdrawal of the EMT stimulus led to a near-complete reversion after as few as three days (**Figure 12B**). We note that it is possible that the cells could have continued to transition following day 7. It will be important for future studies to assess the temporal limits of the EMT response. We then assessed gene expression dynamics throughout the pseudotemporal trajectories. In all cases, transitions were not simply linear processes of two opposing E/M expression programs. Rather, all involved combinations of discrete transcriptional events similar to our observations in OVCA420 cells (**Figure 13**). We found that each condition, with the

exception of A549 cells induced with EGF and OVCA420 treated with TNF, was associated with an average increase in the expression of the EMT hallmark gene set (Liberzon et al., 2015), with TGFB1 often producing the most potent effects (**Figure 14A**). GSEA revealed, however, that differentially expressed genes from these two conditions, along with all others, were enriched for the hallmark gene set (**Figure 14B**), but in these two specific conditions, several EMT hallmark genes are repressed, resulting in a net neutral EMT score (**Figure 15**).

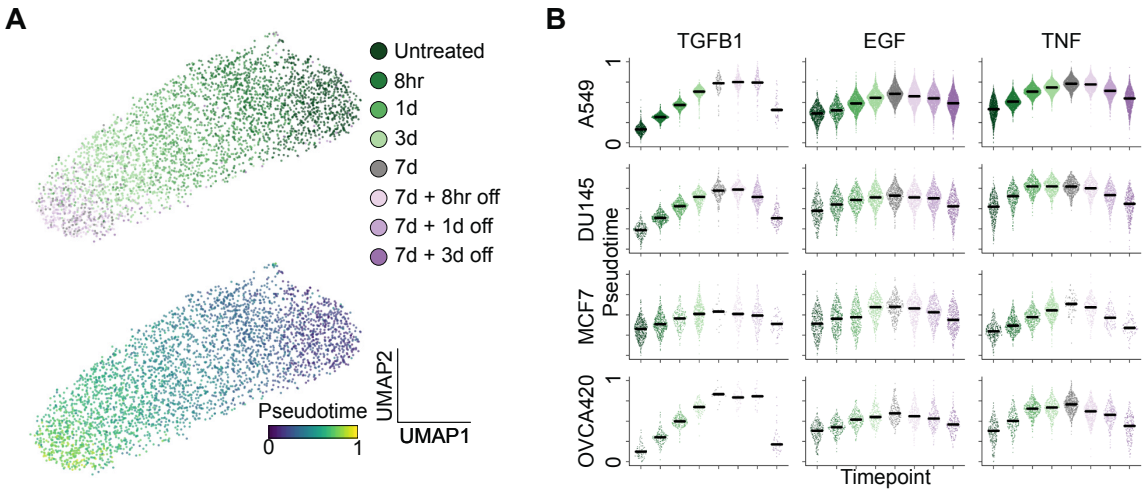

Figure 12. Pseudotemporal ordering of cells throughout the EMT. A) UMAP embeddings of A549 cells treated with TGFB1. Each point represents an individual cell and colours correspond to time point (top) or pseudotime value (bottom). Pseudotime corresponds to the position of each cell along the inferred continuous responses trajectory. **B)** Plot showing the distribution of pseudotime values across time points for all 12 conditions.

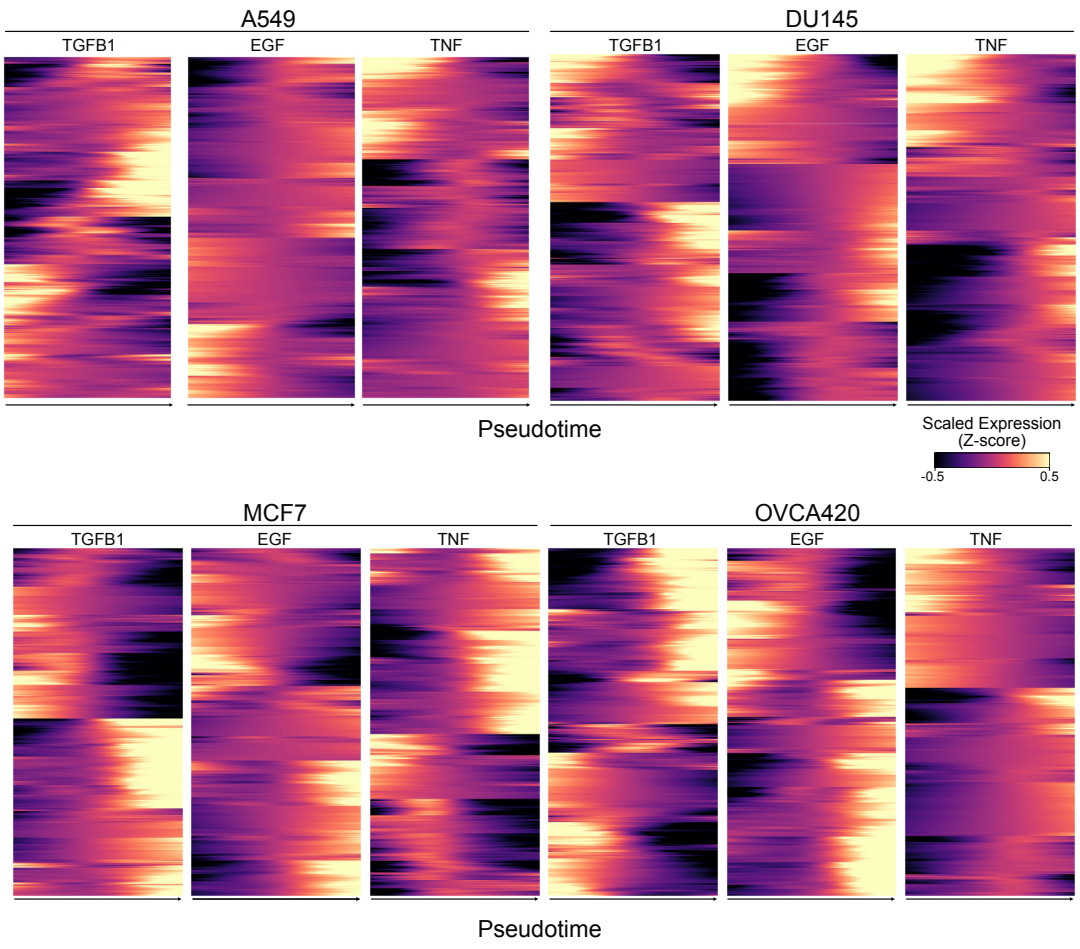

Figure 13. The EMT involves diverse transcriptional patterns. Heatmaps showing expression dynamics of genes differentially expressed in each time course experiment over pseudotime. The x-axis of the heatmaps correspond to EMT progression (pseudotime).

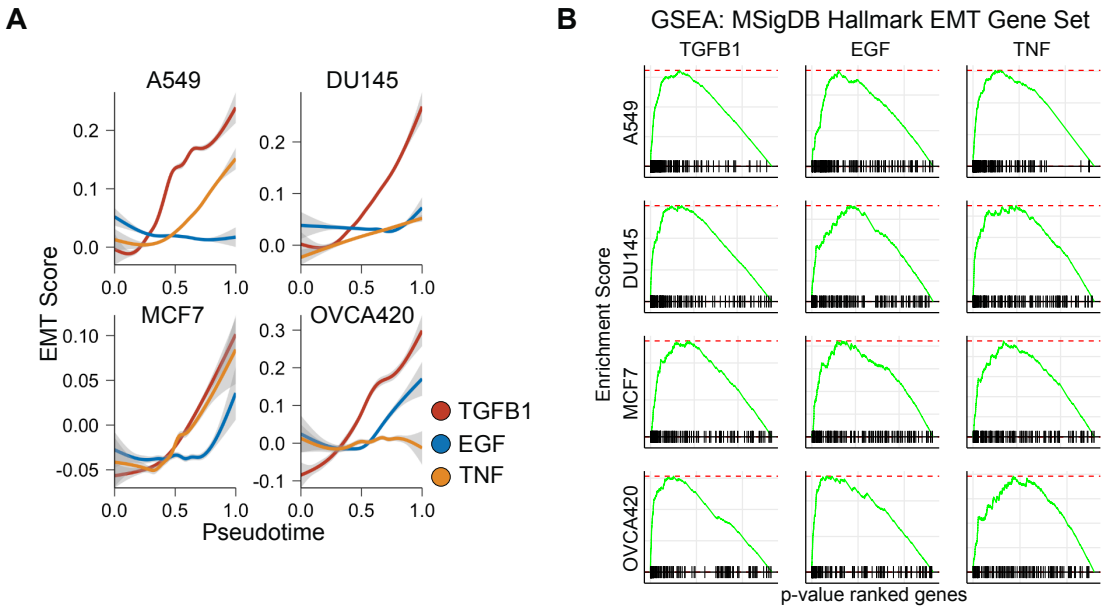

Figure 14. Inferred trajectories capture EMT-associated dynamics. A) Dynamics of the MSigDB Hallmark EMT gene set as a function of pseudotime. Gene set scores were calculated for each cell and scores were then modelled with a generalized additive model with a smoothing function. The solid line represents the model fit and the shaded bands for each line correspond to the standard error. **B)** GSEA plots showing the enrichment of the MSigDB EMT hallmark gene set in the most differentially expressed genes for each time course.

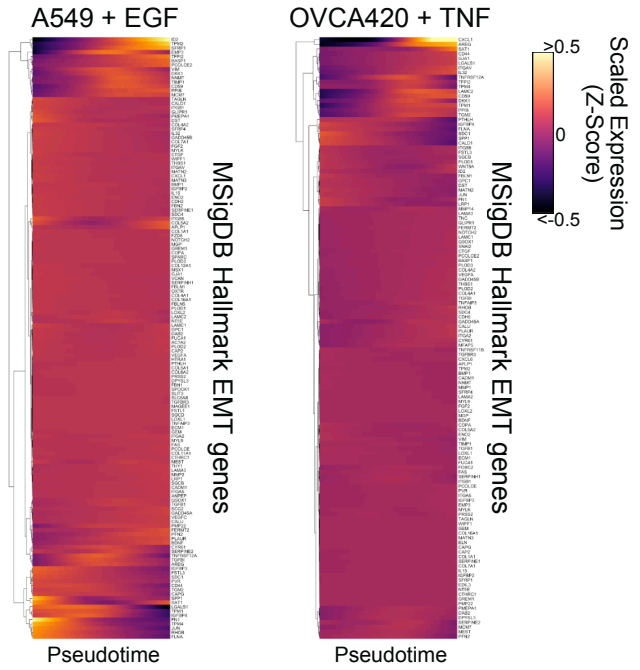

Figure 15. Hallmark gene set scores can obscure EMT-associated changes. Heatmaps showing the dynamics of genes from the MSigDB Hallmark EMT gene set in two conditions that showed no change in activity in the plots in Figure 14A. These conditions involve activation of various EMT genes, but also repress the expression of others, leading to a net neutral gene set score.

Surprisingly, responses of individual cell lines to different stimuli were more similar than the responses of different cell lines to the same stimulus, but importantly, all pairwise comparisons show very little overlap in their differentially expressed genes (average Jaccard index of 0.22; **Figure 16A**). Of all genes differentially expressed across conditions, the majority changed in as few as 1-2 conditions, suggesting that the global expression programs associated with the EMT are remarkably context-specific (**Figure 16B**). A small subset of canonical EMT genes, including *TGFB1*, *CD44*, and *FN1*, along with less-reported genes, such as *TGM2* and *PMEPA1*, were differentially expressed in most conditions. The majority of the MSigDB Hallmark EMT gene set was differentially expressed in only a small number of conditions, with only 49/200 hallmark genes being differentially expressed across the majority of conditions (**Figure 16C**). Extracellular matrix proteins, proteases, and integrins from the hallmark gene set are variably affected across conditions, which could explain the differences in morphological changes observed (**Figure 16C**). This may reflect that the hallmark genes were derived from various founder gene sets that may have been driven by fibroblast expression rather than an EMT (McCorry et al., 2018). Interestingly, however, many canonical EMT genes, including *SNAI1*, *CDH1* (E-cadherin), and *CDH2* (N-cadherin) differentially expressed in only a small number of conditions (**Figure 16B**).

To identify signatures that may not have been represented in the hallmark gene set, we took all genes that were differentially expressed in at least 8 (defined as two-thirds of our conditions as to not be too restrictive) of our experimental conditions and compiled our own gene set of 86 conserved upregulated genes and 17 downregulated genes (**Figure 16D**). While no gene represents a universal marker of the transition, this list contains those that were most frequently changed. Common epithelial-associated (downregulated) genes included various keratins (*KRT8, KRT18,* and *KRT19*), consistent with morphological changes and the loss of epithelial features. While the conserved mesenchymal-associated

(upregulated) genes contain several canonical EMT genes, many are not typically associated with the transition. These upregulated genes, however, do enrich for Gene Ontology (GO) Terms associated with typical EMT traits, including extracellular matrix disassembly (p = 5.0e-4) and organization (p = 3.7e-4), cell migration (p = 2.0e-3), and negative regulation of apoptosis (p = 4.4e-11) (**Figure 17A**). Regulatory regions of the 86 mesenchymal-associated genes are also enriched in binding sites for AP-1, MYC, MEF2, and KLF transcription factors (**Figure 17B**). These factors have all been implicated in the EMT and could represent conserved regulators of the transition (Bakiri et al., 2015; Cieślik et al., 2013; Limame et al., 2014; Su et al., 2016; Yu et al., 2014). We also confirmed that these 86 mesenchymal-associated genes have variable expression levels among cancer cells from individual human lung tumours and syngeneic mouse tumour models, as well as in scRNA-seq data of healthy epithelium from various mouse tissues (**Figure 18A**). Further, in each of these data sets, the 86 mesenchymal-associated genes are highly correlated (**Figure 18B**). Together, this suggests that this expression program is not simply an artifact of culture experiments, but are coexpressed *in vivo* and may contribute to an E/M heterogeneity program in these tissues.

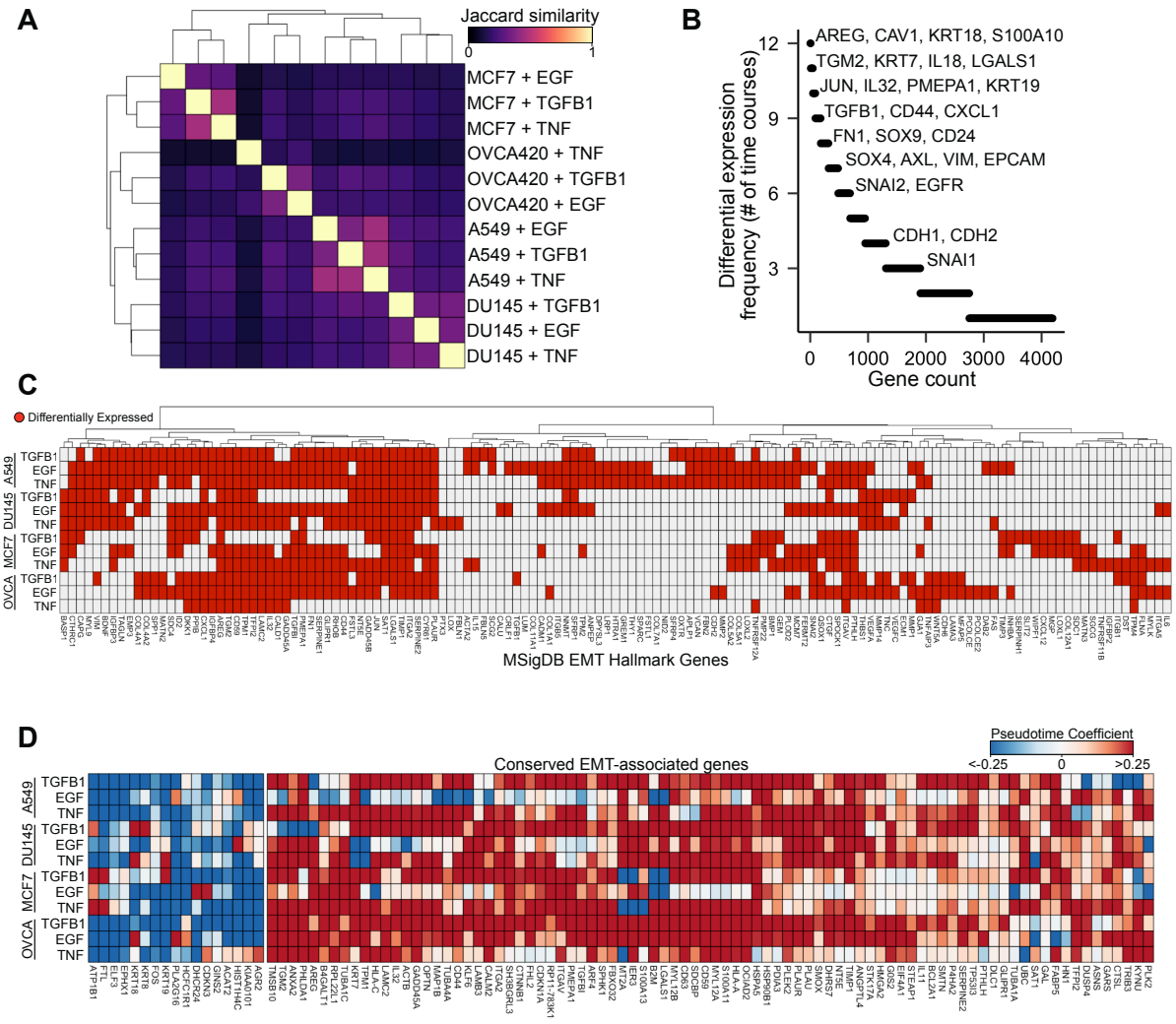

Figure 16. EMT responses are highly context specific. A) Clustered heatmap of all pairwise Jaccard similarity values for the differentially expressed genes in each condition. Similarity values cluster by cell line, suggesting cell line responses to distinct factors are more consistent than those induced with a common factor. **B)** Plot showing the number of time course experiments in which each gene is differentially expressed. C) Heatmap showing Hallmark genes differentially expressed in at least one condition. D) Heatmap showing EMT-associated expression changes of all genes that are differentially expressed in at least 8 time course experiments. The colormap corresponds to the pseudotime beta coefficient of the linear model for each gene.

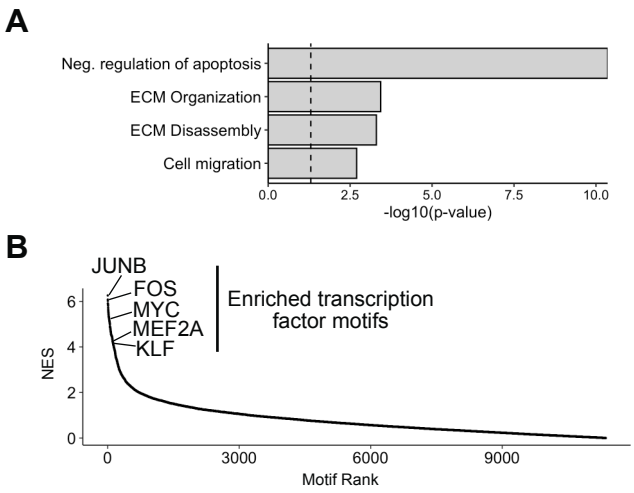

Figure 17. Characteristics of conserved EMT genes. A) Select GO Terms associated with the 86 genes commonly upregulated during the EMT. The dashed vertical line corresponds to p=0.05. P-values were calculated from a Fisher's exact test with a Benjamini-Hochberg false discovery rate correction for multiple comparisons. **B)** Normalized enrichment scores (NES) of transcription factor motifs enriched in the regulatory regions of the 86 mesenchymal-associated genes.

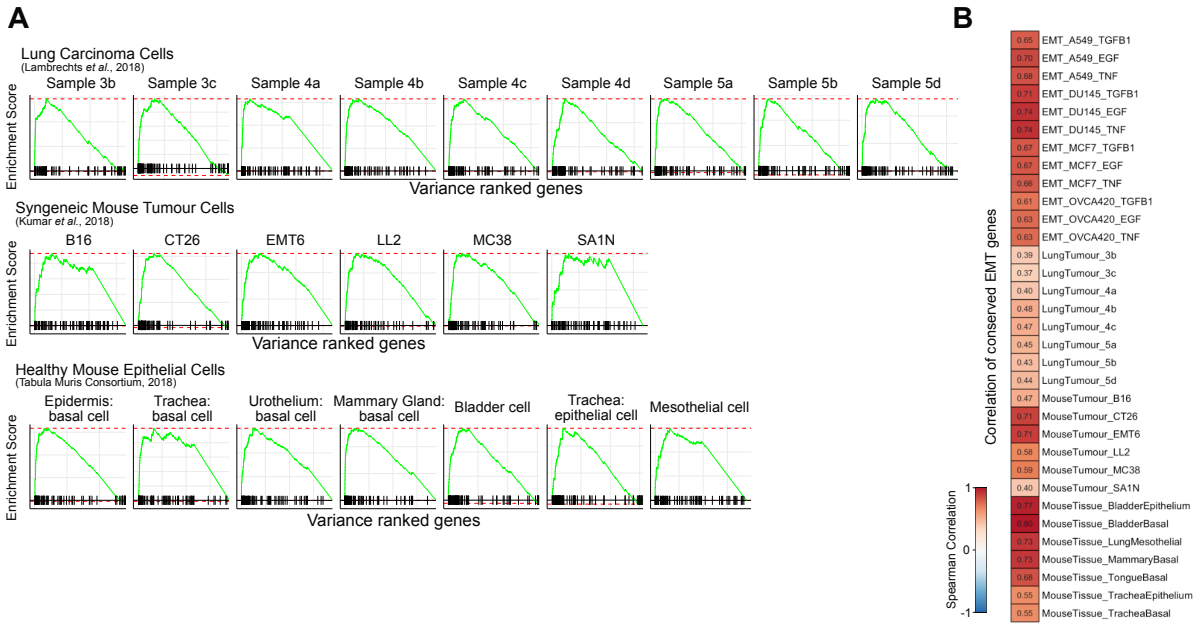

Figure 18. Conserved EMT-associated genes contribute to a coordinate heterogeneity program in epithelial tissues and tumours. A) GSEA plots of conserved EMT genes in the most variable genes from a variety of epithelial and carcinoma populations. **B)** The average correlation (Spearman) of the same conserved EMT genes and data sets as in (A).

3.1.5 The EMT can be coordinated by diverse transcription factor networks

While many of the most conserved EMT genes are regulated by shared regulatory factors (**Figure 17B**), these conserved genes only represent a small fraction of differentially expressed genes. We next sought to determine if the remainder of EMT-associated expression dynamics are driven by a common regulatory program that perhaps gives rise to distinct expression patterns due to cells' epigenetic or mutational profiles, for example. Across the experimental conditions we assessed, most canonical EMT transcription factors—other than *SNAI2*—were rarely differentially expressed (**Figure 19**). While in some cases (eg. *TWIST1*) the transcription factors were not detected, perhaps owing to insufficient sensitivity to lowly expressed genes, canonical EMT transcription factors were often readily detectable, but did not show dynamics throughout the EMT response (**Figure 20**). We scored each cell for the coexpression of transcription factors and their putative target genes (regulons) and identified those that showed differential activity throughout the EMT. We found that transcription factor activity is also remarkably context-specific, with most being restricted to 1-2 of our time course experiments (**Figure 21A**). Several factors were fairly well conserved, however. Consistent with our list of conserved genes, AP-1 (JUN, JUNB), the NFkB-associated RELB, ATF4, SOX4, and KLF6 regulons showed frequent activation, whereas ELF3 and MYBL2 activity often decreased (**Figure 21B**). These factors have all been previously implicated in the EMT, but are not typically considered canonical EMT regulators (Gondkar et al., 2019; Holian et al., 2008; Huber et al., 2004; Suzuki et al., 2010; Tiwari et al., 2013; Ward et al., 2018).

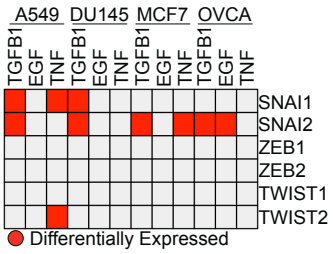

Figure 19. Involvement of canonical EMT transcription factors. Plot showing which time course experiments canonical EMT transcription factors are differentially expressed.

To assess the accuracy of these results, we used ATAC-seq to assess the accessibility of transcription factor motifs throughout the EMT and compared accessibility dynamics to the inferred regulon activity. For validation, we chose to assess the OVCA420 TGFB1 time course (**Figure 21C**). This was the smallest dataset in our scRNA-seq cohort, so we chose to validate the approach on the condition with the lowest power for inferring transcription factor activity. We found that in many cases, motif accessibility throughout the EMT mirrored regulon activity measured from scRNA-seq data alone (**Figure 21D**). This supports that the regulon activity inference provides an accurate representation of transcription factor activity throughout each of the conditions assessed.

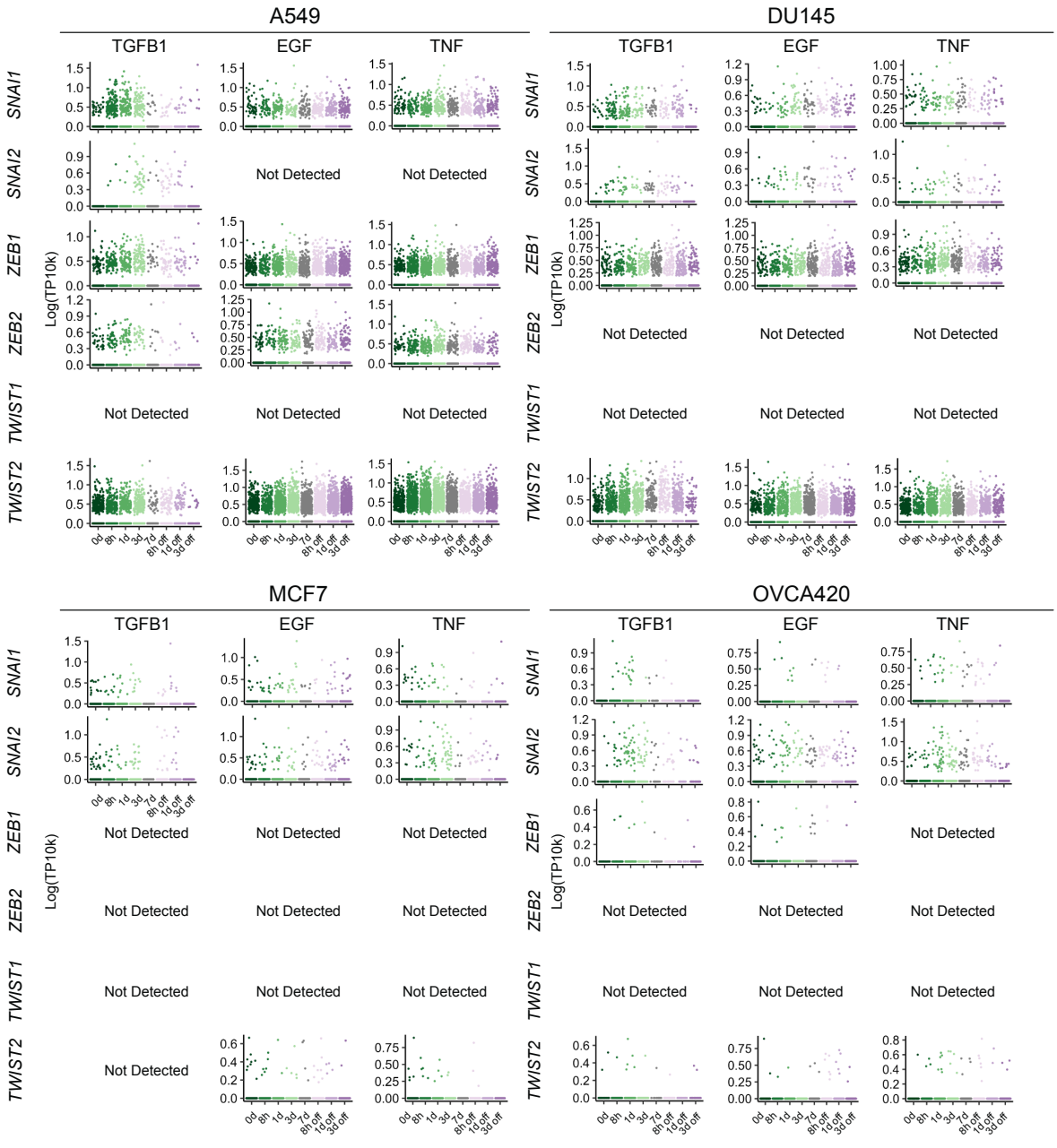

Figure 20. Expression of canonical EMT transcription factors across time course experiments. Plots showing expression levels of SNAI1, SNAI2, ZEB1, ZEB2, TWIST1, and TWIST2 in each time course experiment. Expression values are represented as log-transformed counts per 10k molecules (TP10k). Each dot corresponds to a single cell.

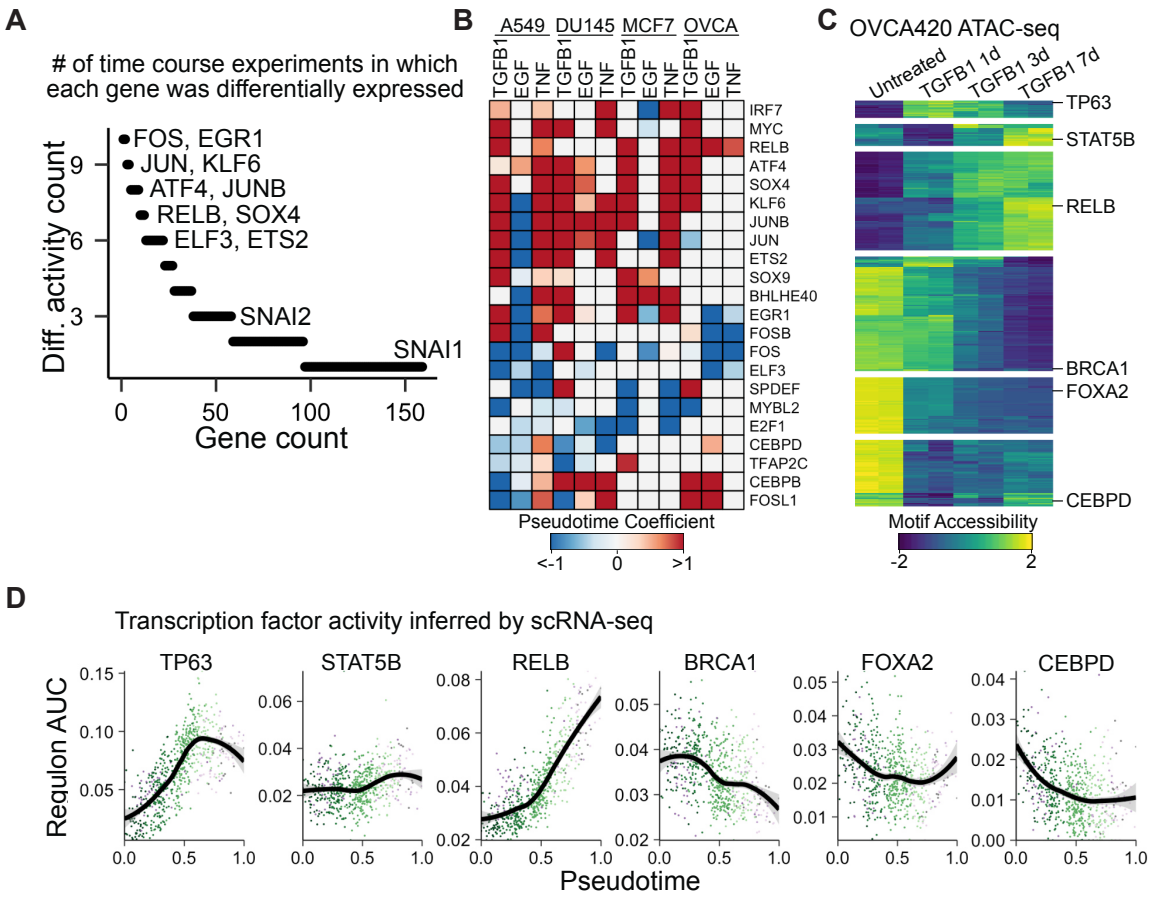

Figure 21. Inferring transcription factor activity throughout the EMT. A) Counts of how frequently various transcription factor regulons are differentially active among time course experiments. **B)** Heatmap showing EMT-associated changes of regulons that are differentially active in at least 6 time course experiments. The colormap corresponds to the pseudotime beta coefficient of a linear model for each regulon. **C)** Differential accessibility of transcription factor motifs from ATAC-seq data of OVCA420 cells treated with TGFB1 for 0, 1, 3, or 7 days. The colormap represents the accessibility Z-score for each transcription factor motif. Examples of transcription factors from each cluster are listed. **D)** Regulon activity score of the same transcription factors listed in (C) inferred from the OVCA420 TGFB1 time course experiment. Each dot represents a single cell, coloured by time point. The black line corresponds to the modelled trend from a generalized additive model. Shaded bands for each line correspond to the standard error for each model.

3.1.6 Perturbations of EMT-associated transcription factors

It is unclear to what extent the EMT response is dependent on these factors. They may also regulate independent modules of responsive genes or coordinate the EMT through a highly connected regulatory network. The ability of any individual factor to initiate an EMT

response is unknown. To better understand these regulatory relationships, we designed pooled screens to overexpress or knockdown the expression of 25 different transcription factors using the CRISPR interference (CRISPRi) and activation (CRISPRa) systems.

We first transduced A549 and DU145 cells with lentiviral constructs of the nuclease-dead Cas9 fused to either the repressive KRAB domain or the “SunTag” activation system (Tanenbaum et al., 2014) (**Figure 22**). We then designed independent pooled sgRNA libraries for the two systems targeting 25 different transcription factors (3 sgRNA/cell, 5 non-targeting controls) (**Figure 23A**). The sgRNAs were designed with a capture sequence within the stem-loop of the sgRNA scaffold, allowing them to be primed for cDNA synthesis by the 10x Genomics chemistry. These libraries were then transduced into A549 and DU145 cells in a pooled format at a low multiplicity of infection (MOI) of 0.5, processed for scRNA-seq, and the perturbation associated with each cell could be detected along with the cell’s gene expression profile (Replogle et al., 2020).

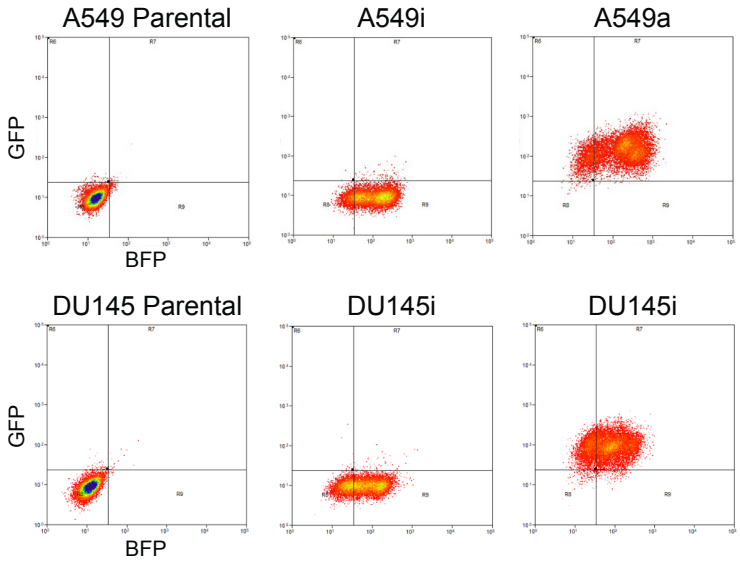

Figure 22. Generating cell lines for CRISPRi and CRISPRa. Flow cytometry data showing the distribution of BFP and GFP activity in A549 and DU145 cells. Parental lines had not been transduced. The CRISPRi system involves a BFP protein fused to the dead Cas9. The CRISPRa system involves two separate vectors: one with BFP fused to Cas9, another with GFP fused to the VP64 transcriptional activator.

We processed the A549 and DU145 lines for scRNA-seq 9 days following transduction with the CRISPRa library, approximating 7 days of transcription factor overexpression. For the

CRISPRi screen, we added TGFB1 or a vehicle control to the cells 3 days after transducing the sgRNA library and processed the cells for scRNA-seq after a week of TGFB1 treatment. For all conditions, culture media was changed daily to eliminate confounding effects of paracrine signalling.

The majority of guides were represented in the scRNA-seq data of each sample. After determining the perturbation associated with each cell, we assessed the effect of the sgRNA on the expression of its target gene. In each sample, dozens of guides were associated with a significant effect on their target gene (**Figure 23B**). However, when we performed differential expression to determine the global effects of these perturbations, we were unable to detect any other effects in targeted cells (**Figure 23C)**.

There are a limited number of explanations for this. First, it is possible that the perturbation of the transcription factors were not long enough to elicit a response. In all conditions, the sgRNA was expressed for approximately 7 days. This period was long enough to show effects on the target gene's transcript levels, suggesting that the sgRNA was active, but perhaps it was not sufficient for protein loss and subsequent changes to occur. Based on kinetics of RNA interference experiments, this seems unlikely, but is possible. Another explanation may be limited power due to an insufficient number of cells. The experiment yielded only 30-50% of the targeted number of cells per condition, causing many sgRNAs to only be present in several dozen cells. Combined with variability in the perturbation strength in each cell, any effects may have been too heterogeneous to test with such a limited number of cells. However, power analysis of similar perturbation screens have been performed and signature-level effects should be detectable with tens of cells (Dixit et al., 2016). Finally, it may also be possible that the perturbation effect was simply not strong enough to affect the transcription factor's function in any meaningful way.

Understanding the involvement of these non-canonical transcription factors in the EMT response will be informative, and similar large-scale perturbation screens are a convenient strategy to accomplish this. Unfortunately, due to time and funding constraints for this project, we were unable to pursue this further.

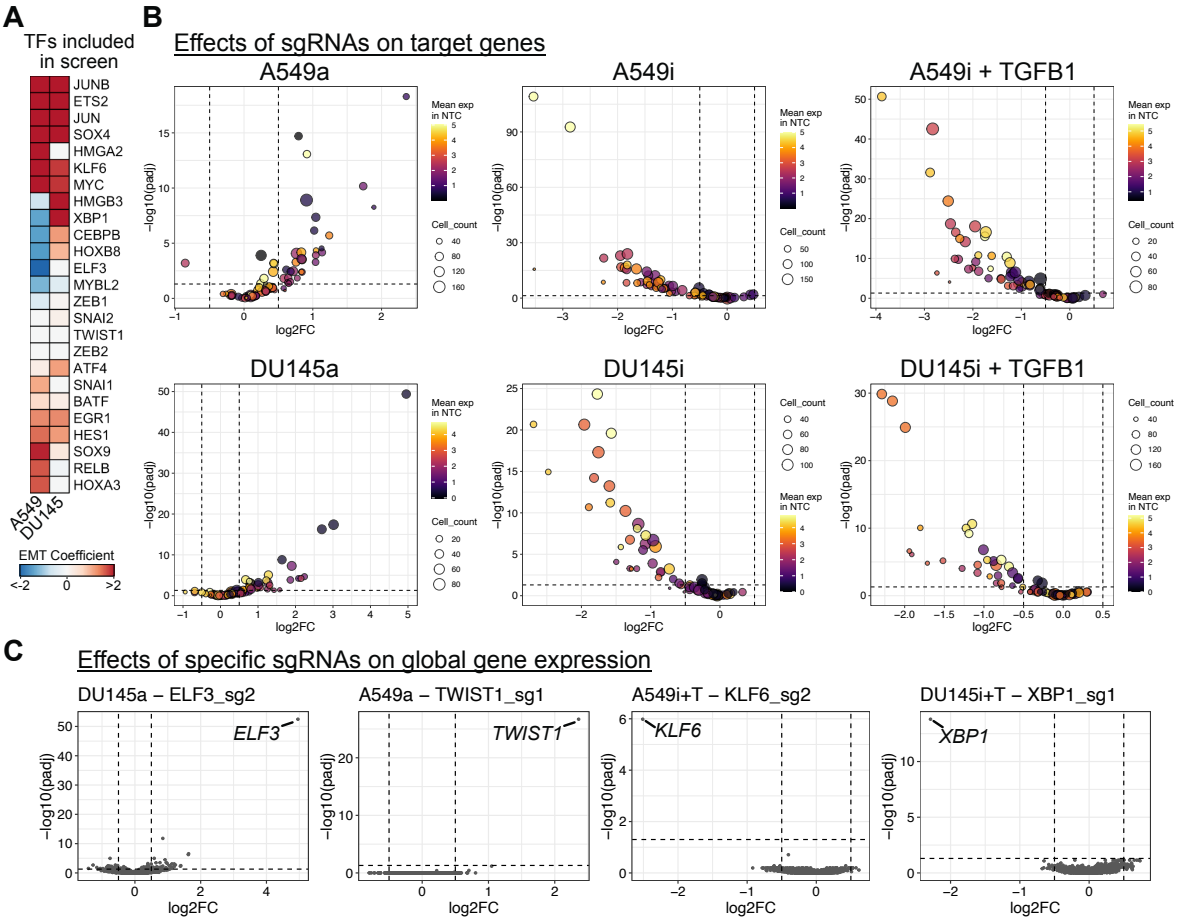

Figure 23. Summary of pooled CRISPR library screens. A) The pooled libraries comprise sgRNAs targeting 25 transcription factors (3 guides/target) involved in EMT responses. Heatmap shows the inferred transcription factor activity associated with the EMT response in A549 and DU145 cells following TGFB1 treatment. **B)** Plots showing the efficacy of sgRNAs in modulating the expression level of target factors. Each point represents the differential expression of the target transcription factor in cells with a targeting sgRNA relative to cells with non-targeting control guides. **C)** Global differential expression in cells with select sgRNAs that effectively modulated the target transcription factor.

3.1.7 Summary

Here, we have applied multiplexed scRNA-seq to perform a comparative analysis of EMT responses across distinct contexts and demonstrated extensive context specificity, with the majority of responsive genes being differentially expressed in as few as 1-2 conditions. Transcriptional dynamics associated with each condition are consistent with previously defined EMT gene sets, but the specific associations in distinct contexts are highly variable. We highlight that canonical EMT markers may be poor indicators of EMT, at least in scRNA-seq data. We have also defined an 86-gene list of the most conserved genes involved in EMT responses, which will be valuable for assessing new data.

3.2 Transcriptional census of EMP in cancer

3.2.1 A multi-cancer census of EMT-associated gene expression

If EMP is a consistent feature of solid tumours, gene expression patterns associated with it should be readily detectable in scRNA-seq data. To assess this, we first collected droplet-based scRNA-seq data from 12 studies of 8 different cancer types, including colorectal (Lee et al., 2020; Qian et al., 2020; Uhlitz et al., 2021), gastric (Sathe et al., 2020), lung (Kim et al., 2020; Lambrechts et al., 2018; Laughney et al., 2020; Qian et al., 2020), uveal melanoma (Durante et al., 2020), squamous cell (Ji et al., 2020), ovarian (Geistlinger et al., 2020; Qian et al., 2020), pancreatic (Steele et al., 2020), and breast (Qian et al., 2020). After removing samples with fewer than 100 malignant cells, the data comprises expression profiles of 182,198 cancer cells from 160 tumour samples (**Figure 24, Figure 25**).

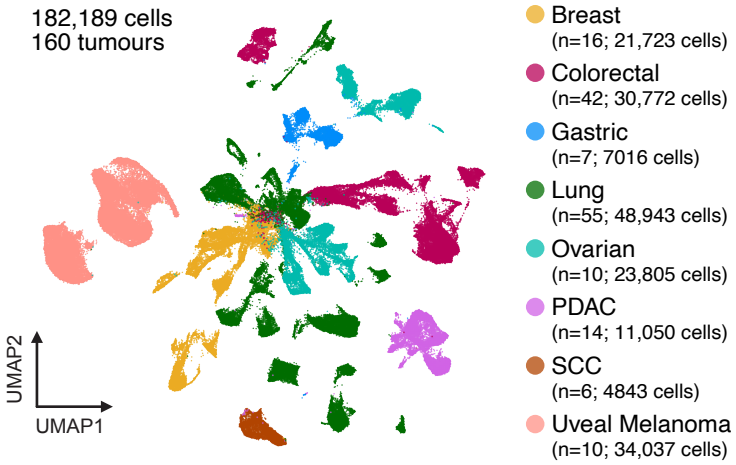

Figure 24. scRNA-seq of malignant cells from 160 tumours. UMAP embedding of scRNA-seq data 182,189 malignant cells from 160 tumours of 8 different cancer types, integrated from 12 different studies.

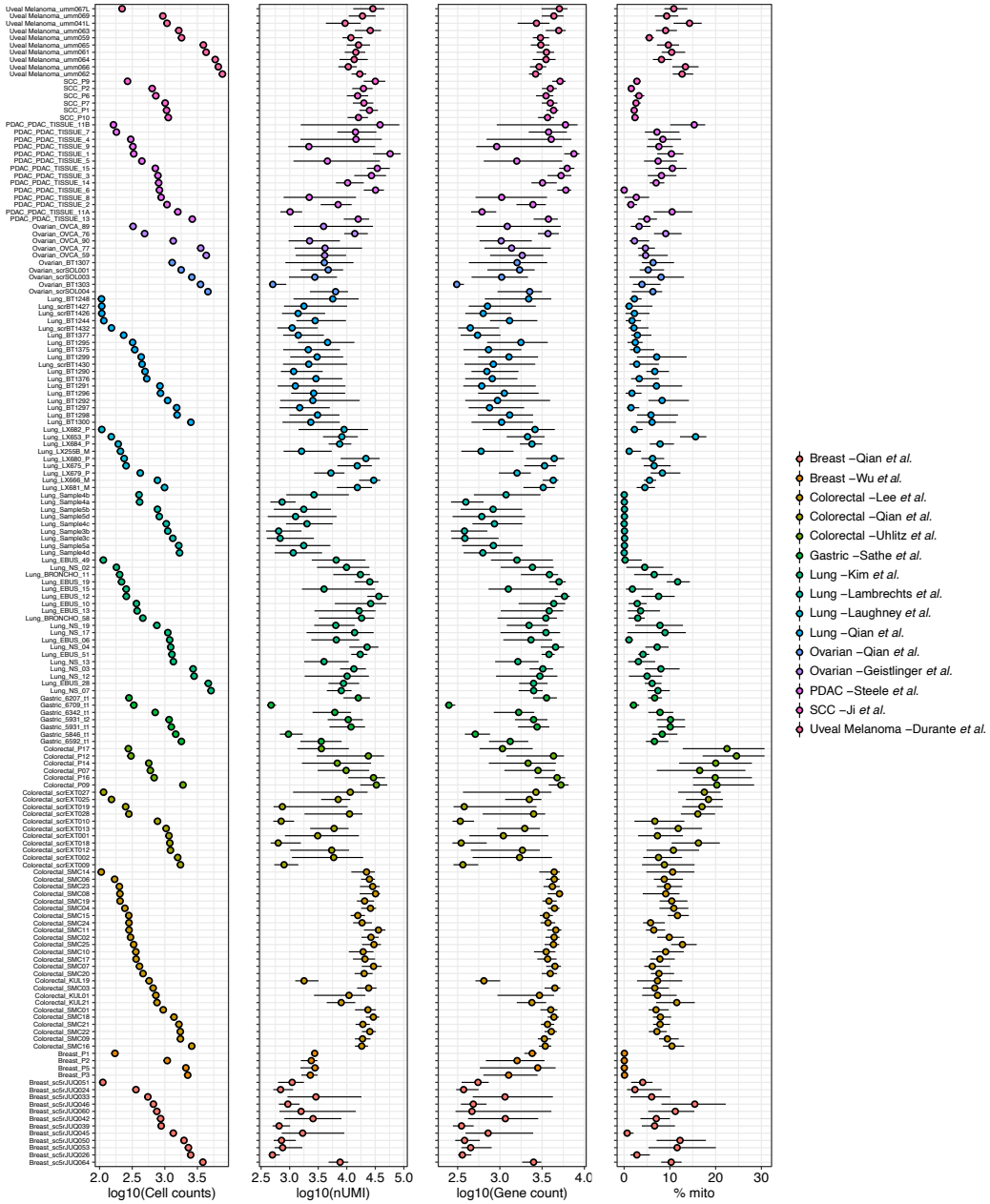

Figure 25. Quality control metrics for scRNA-seq data of malignant cells from 160 tumours. Plots showing the total number of cells, the distributions of unique transcript counts per cell, numbers of genes per cell, and percentage of mitochondrial reads per cell, respectively. Points represent the median value and the line spans the 25th-75th percentile.

To assess intratumoural heterogeneity of EMT-associated gene expression, we scored each cell for its relative expression of genes contained in 10 different EMT gene sets, including several curated sets [the “Epithelial-mesenchymal transition” GO Term, MSigDB Hallmark gene set (Liberzon et al., 2015), CancerSEA (Yuan et al., 2019), dbEMT1.0 (Zhao et al., 2015)] and others derived from individual studies [Taube (Taube et al., 2010), Puram (Puram et al., 2017), Cook (Cook and Vanderhyden, 2020), and Kinker I, II, III (Kinker et al., 2020)]. The composition of these gene sets is high variable, with over two-thirds of the genes being present in only a single gene set (**Figure 26A**). While it is possible these diverse gene sets could represent different subsets of a common underlying biological process, we correlated the gene set scores across all cells and found that this is not uniformly true. The 10 gene sets formed two groups of well-correlated scores, suggesting that they may represent distinct EMP programs **(Figure 26B**).

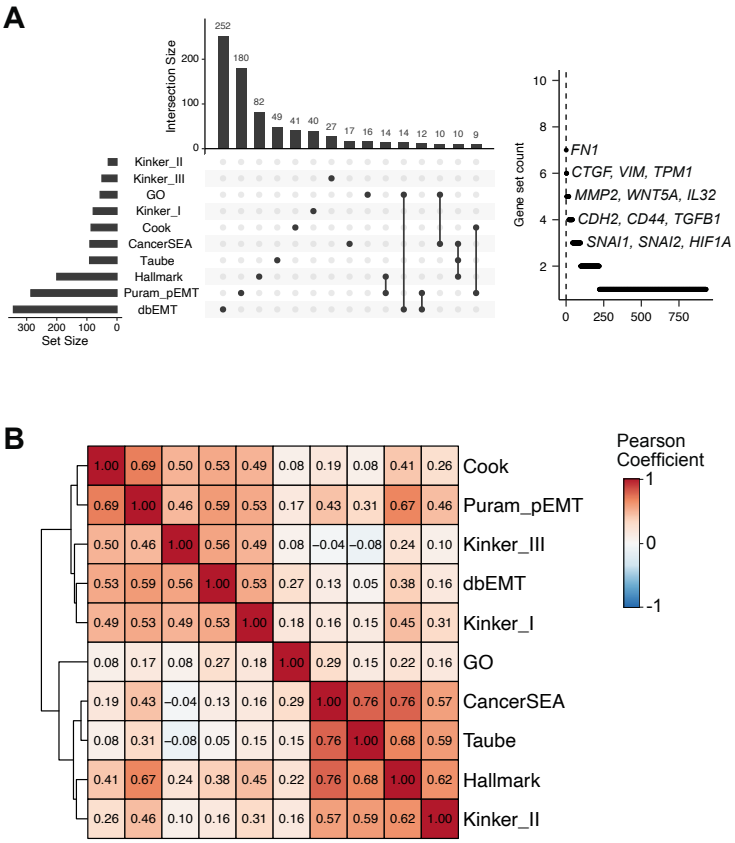

Figure 26. The composition of EMT gene sets is variable and may represent diverse processes. A) UpSet plot showing the intersection of genes comprising 10 EMT gene sets (left) and plot showing the total number of gene sets within which each gene occurs (right). **B)** Pearson correlation coefficients of gene set scores for each of the 10 EMT gene sets across the 160 tumours analyzed.

Gene set scores can provide biological insight into gene expression patterns, but they can be influenced by uninteresting features of the data. Specifically, they can be inflated by high expression of a small proportion of the set's genes that are not necessarily determinants of the queried biological process. Variation in scores across a population of cells may also reflect random fluctuations of the set's genes and not necessarily a coordinated activation of the process. Matrix factorization approaches have been applied to scRNA-seq data to learn coordinated expression programs heterogeneously expressed across a population. By learning these programs from the data itself, reliance on previously defined gene sets is restricted to only the interpretation of the programs.

We next sought to explore heterogeneously expressed programs from the 160 tumours. To identify these latent gene expression programs, we performed multi-resolution archetypal analysis on each tumour sample using the ACTIONet algorithm (Mohammadi et al., 2020), learning cell activity scores for distinct programs in each tumour (**Figure 27A**). This identified multiple programs in each sample, including expected sources of variation, such as cell cycle activity. To identify those associated with EMP, we correlated the cellular activity of each program learned by ACTIONet with the previous EMT gene set scores. All 160 tumours had programs well-correlated with these scores, suggesting intratumoural EMP is a ubiquitous feature of solid tumours and readily captured in scRNA-seq experiments (**Figure 27B,C**). Kinker *et al*. (2020) recently demonstrated recurrent heterogeneity of programs consistent with EMP in scRNA-seq data derived from monolayer cultures of cancer cell lines, but this is the first quantitative demonstration of this plasticity in solid tumours across multiple cancer types.

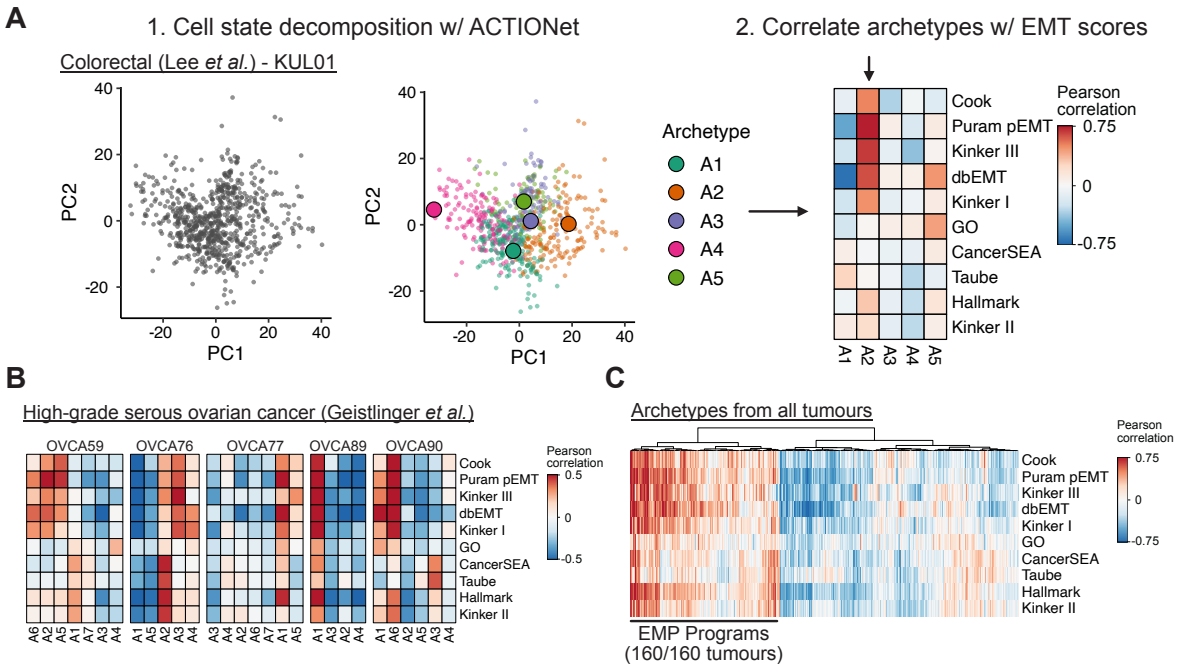

Figure 27. Using archetypal analysis to learn EMP-associated gene expression programs. A) Schematic representation of the analysis strategy to identify sample-specific EMP programs using archetypal analysis and correlating archetype scores with EMT gene set scores. **B)** Pearson correlation coefficients of archetype scores with EMT gene set scores of individual cells from 5 high-grade serous ovarian tumours. **C)** Hierarchically clustered heatmap of Pearson correlation coefficients of all archetype scores with EMT gene set scores from the 160 tumours analyzed.

3.2.2 Defining tumour-specific expression programs associated with EMP

To define the specific genes contributing to EMP in each tumour, we identified genes whose expression significantly changes as a function of the cells' activity of each program learned from ACTIONet. We previously showed that transcriptional responses of experimentally induced EMTs are highly context-specific, but it was unclear if the same diversity existed *in vivo*. Of the 3822 genes differentially expressed in at least one EMP program, the vast majority were associated with a small number of samples. We clustered differentially expressed genes based on their model coefficients for each EMP program and identified a group of 640 genes with a frequent positive association (**Figure 28A**). We further refined this signature by removing genes that were associated with fewer than 10 EMP programs or

also downregulated in more than 10 EMP programs, resulting in an EMP signature of 289 genes (**Figure 28B**).

EMP programs were associated with reduced expression of a group of 125 genes. Interestingly, epithelial genes were not enriched among this group. Rather, they strongly enrich for GO terms associated with cell cycle (p = 2.9e-73, Fisher's exact test). This is consistent with the proliferation-migration trade-off associated with a mesenchymal phenotype. We do note that several programs classified as EMP programs in multiple tumour types showed activation of these genes. The lack of expression patterns associated with loss of an epithelial phenotype further supports the growing evidence of hybrid E/M phenotypes being highly prevalent in cancer (Kröger et al., 2019; Pastushenko and Blanpain, 2019; Pastushenko et al., 2018; Puram et al., 2017).

Of the 289 genes positively associated with EMP, no individual gene was a perfect indicator of its activity, but as a collective, they represent a fairly consistent signature. They also enrich for GO Terms consistent with a mesenchymal phenotype, including cell motility, regulation of cell adhesion, and response to wounding (**Figure 28C**). Many of the canonical EMT genes are not included in this signature, including *CDH2*, *VIM*, *SNAI1*, *SNAI2*, and *ZEB1*, although these genes did have variable associations with EMP programs (**Figure 28D**). The signature did, however, include many genes that have previously been implicated in the EMT, including various transcription factors (*SOX4*, *KLF2/4/6/10, JUND*), integrins (*ITGB1/4/8*, *ITGA2/3*), secreted factors (*VEGFA*, *IL32*, *CXCL1*, *CXCL8*), and membrane proteins (*CD24*, *CD59*, *S100A6*, *CEACAM1*), and more (**Figure 28E**).

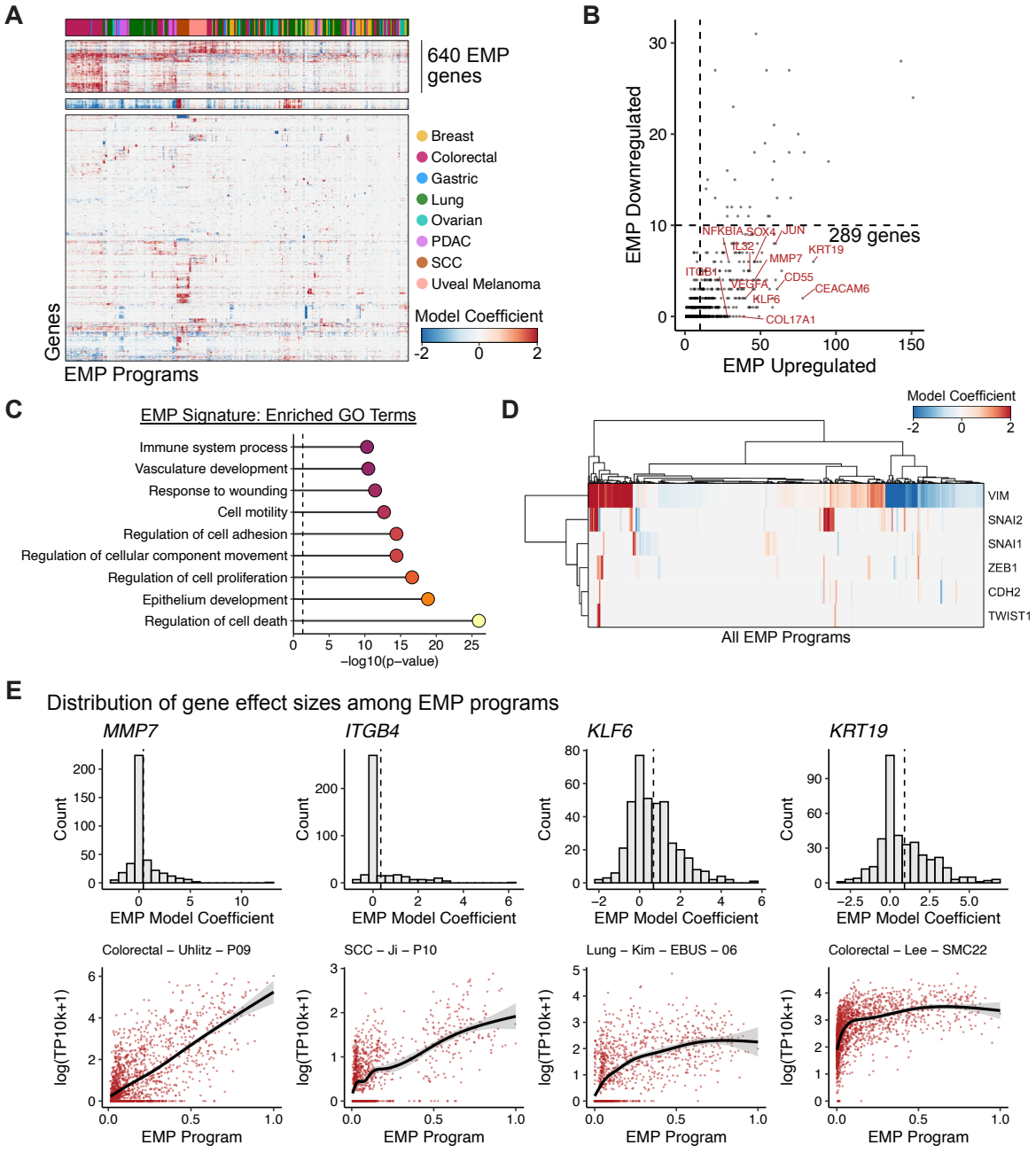

Figure 28. Defining a conserved EMP gene signature. A) Hierarchically clustered heatmap of EMP model coefficients for 3822 genes differentially expressed in at least one EMP program. **B)** Plot showing how frequently each of the 640 genes were down- and upregulated in EMP programs. **C)** GO terms significantly enriched in the 289 conserved EMP genes. *p*-values were calculated using Fisher's exact tests and adjusted using the Benjamini-Hochberg method. **D)** Heatmap of EMP model coefficients across all EMP programs. **E)** Examples of EMP-associated genes. Top: distribution of effect sizes for each gene with all EMP programs. Dashed line represents the mean value. Bottom plots show expression values in the EMP program with the highest effect size for that gene.

The stability and distribution of phenotypes along an epithelial-mesenchymal continuum has gained attention recently, with the relevance of hybrid phenotypes being contrasted to fully epithelial or mesenchymal cells (Aiello et al., 2018). Using gene set scores of the conserved signature as a relative measure of the cells' mesenchymal activity, we found that different tumours can have different average levels of mesenchymal expression, but also that the variance of scores can vary dramatically, suggesting that range of phenotypes represented within the tumour can vary (**Figure 29, Figure 30**). We also note that the majority of tumours don't have clear multimodal distributions that would be consistent with a model where various states along the phenotypic continuum have elevated stability. Rather, cells span the continuum, forming a distribution with most cells occupying intermediate states and tails spreading to more extreme phenotypes. These patterns are recapitulated when using gene set scores from the EMT-associated gene sets explored previously and the specific EMP-associated program learned from the tumours themselves (**Figure 30**).

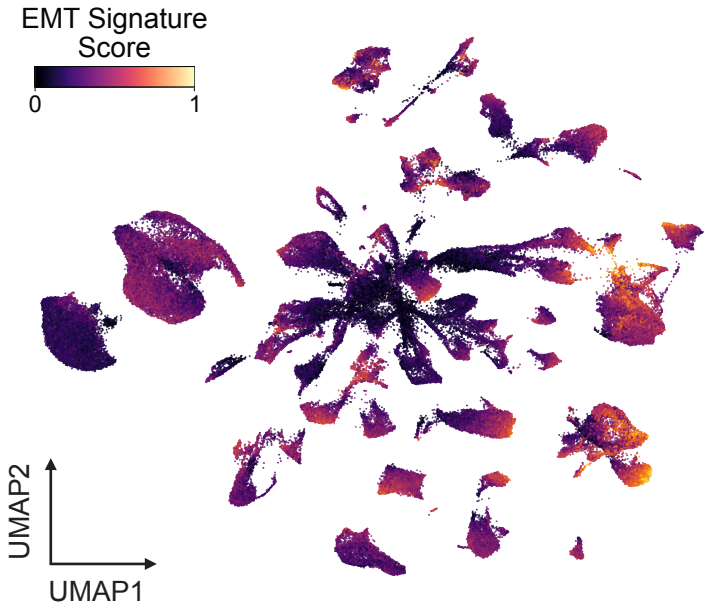

Figure 29. Intratumoural heterogeneity of a conserved EMP signature. UMAP embedding of malignant cells from the 160 tumours analyzed, coloured by a gene set score for the signature of 289 EMP-associated genes.

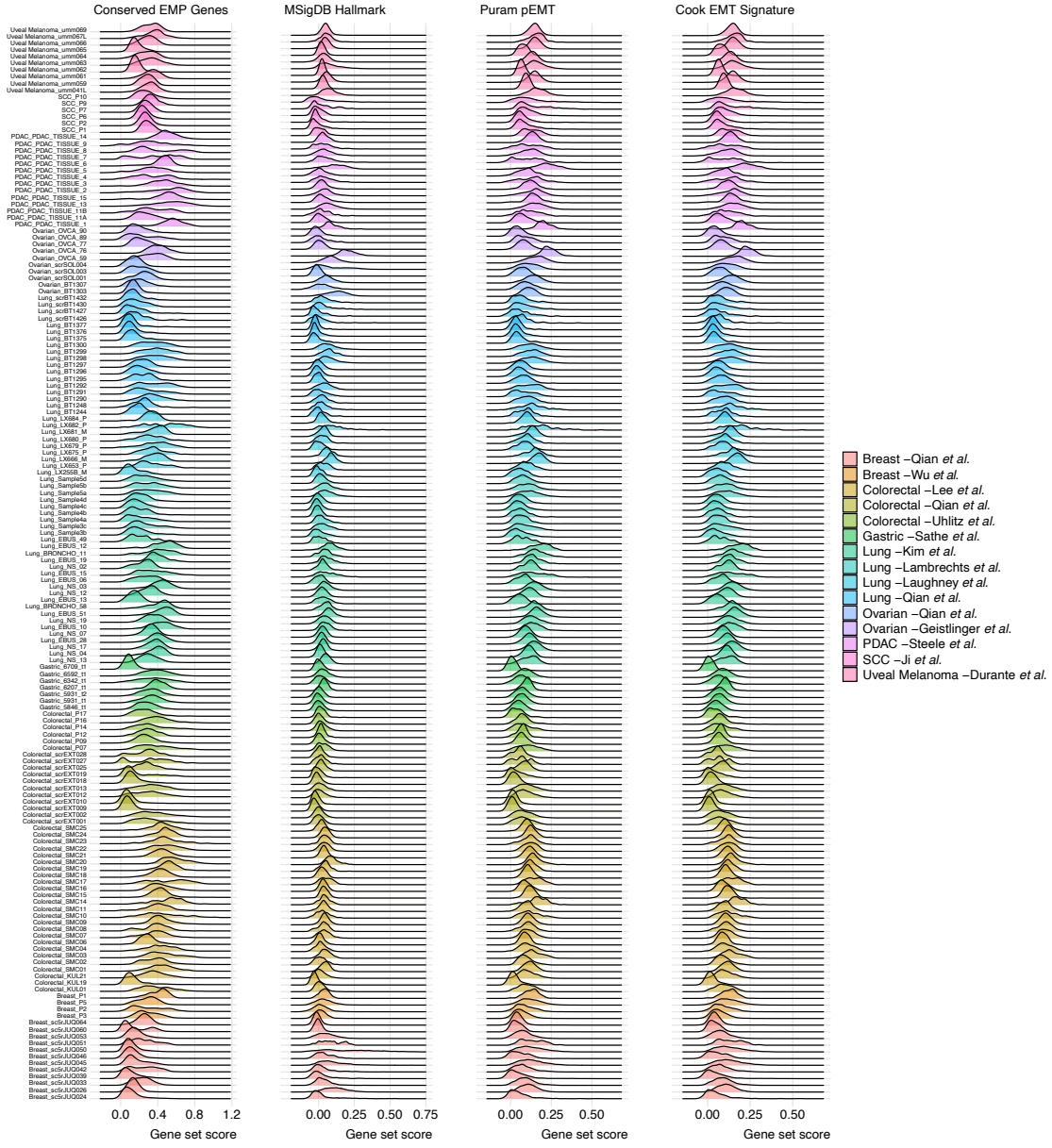

Figure 30. Distribution of EMP gene set scores. Density plots showing the distribution of gene set scores for the identified EMP signature (left) and three public EMT gene sets in malignant cells from all 160 tumours.

3.2.3 EMP is associated with poor prognosis an immunosuppressive TME

Due to their inherent similarities, the ability to distinguish fibroblast and EMP-specific expression patterns has been a challenge. Many EMT gene sets contain genes highly expressed in fibroblast populations and as a result, “mesenchymal” features of tumours

defined from bulk RNA-seq data have been found to often be associated with fibroblast content of the tumour rather than cancer cell plasticity (Isella et al., 2015; Izar et al., 2020; Puram et al., 2017). The choice of specific markers used to assess EMP in studies has also led to controversy (Yang et al., 2020). This confusion has made it challenging to draw conclusions about the involvement of EMP in tumour progression and clinical outcomes.

For each tumour sample, we calculated a cell type specificity score for each of the 289 conserved EMP genes and averaged these scores across tumours to get an overview of how specific the markers were to cancer cells (**Figure 31A**). Of the 289 genes, 165 were highly specific to cancer cells, whereas the remaining 124 were also expressed in fibroblasts, macrophages, and/or T cells. A signature of cancer cell-specific EMP genes could be valuable for generating EMP activity scores in scRNA-seq data, so we established a refined signature comprising the 165 genes highly specific to cancer cells. We scored individual cells comprising 22 colorectal tumours (Lee et al., 2020) for their activity of this signature and found heterogeneous activity among cancer cell populations, with effectively no activity in non-malignant cell types (**Figure 31B**).

Given this high specificity, the signature could be used as a measure of mesenchymal properties in bulk RNA-seq data without being confounded by expression from fibroblasts. We used the pan-cancer RNA-seq data from The Cancer Genome Atlas (TCGA) (Hoadley et al., 2018) to calculate a relative EMP score for all tumours. Modelling patients' progression-free interval (PFI) as a function of this signature activity, tumour type, age at diagnosis, and tumour purity, we found that EMP activity was associated with a reduced PFI (Cox hazard ratio: 1.74; $p = 0.025$) (**Figure 31C**). Using estimates of immune cell proportions across all TCGA samples (Thorsson et al., 2019), we also found high expression of this mesenchymal signature was associated with larger proportions of

inflammatory cell types (activated mast and dendritic cells), but also more immunosuppressive regulatory T cells and fewer effector cell types, including CD8 T cells and naive CD4 T cells (**Figure 31D**).

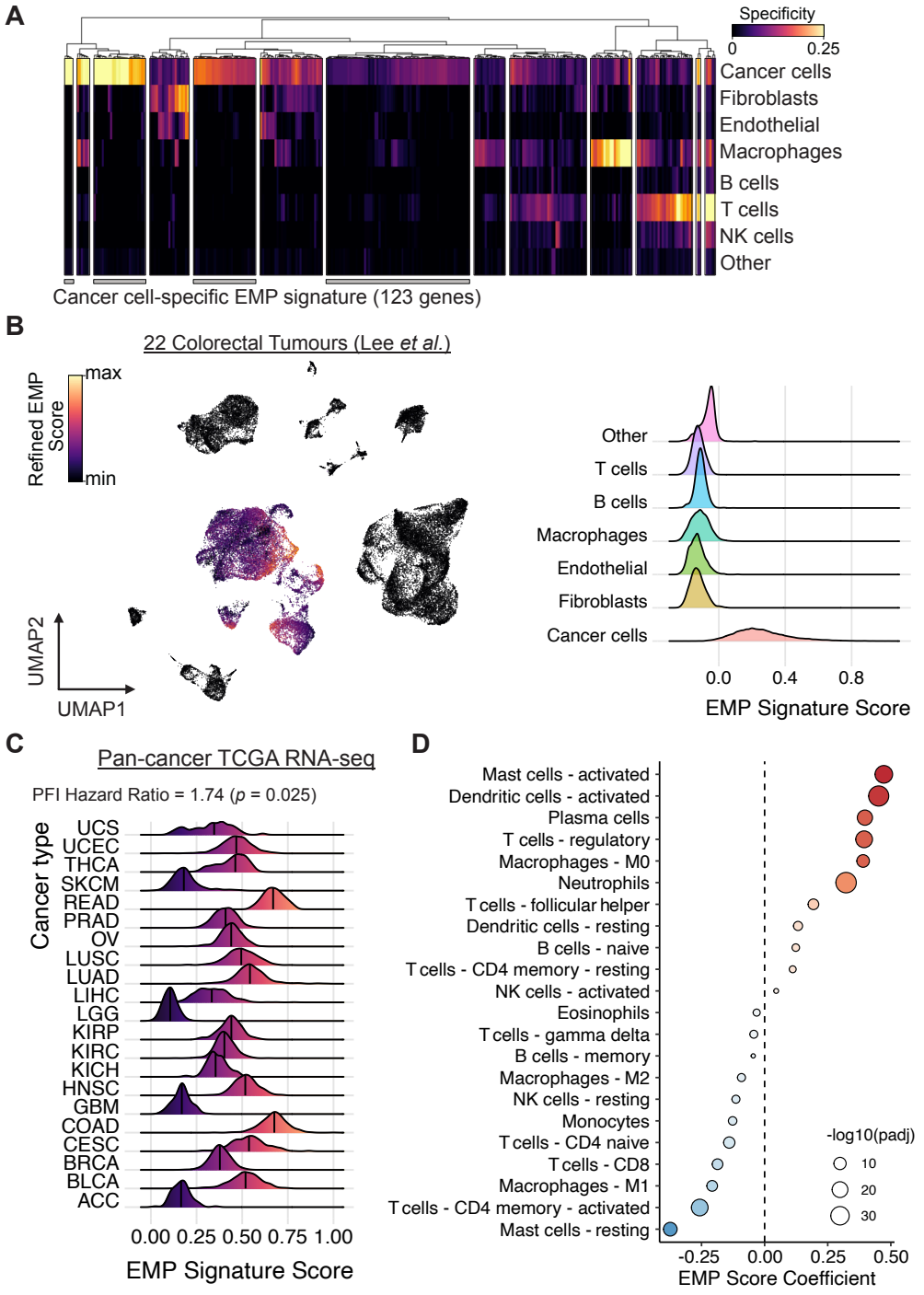

Figure 31. EMP is associated with worse progression-free survival and an immunosuppressive TME. A) Clustered heatmap of cell type specificity scores for each of the 289 EMP-associated genes. Clusters of genes with high cancer cell specificity were defined as a cancer cell-specific EMP signature. **B)** UMAP embedding (left) and density plot (right) showing the distribution of gene set scores for the cancer cell-specific EMP signature in all cell types from 22 colorectal tumours. **C)** Progression-free interval (PFI) hazard ratio and distribution of signature scores from the TCGA's pan-cancer bulk RNA-seq cohort. **D)** Changes in immune cell proportion estimates as a function of EMP signature scores from the TCGA's pan-cancer cohort.

3.2.4 Diverse paracrine signalling modulates EMP

While unifying molecular signatures are appealing, appreciating the diversity of EMP programs is critical as it may contribute to functional nuances of the phenotype. These programs may also have varying regulatory dependencies that would warrant different therapeutic approaches. Variation in cell state emerges from complex differences in the cells' microenvironment, including changes in oxygen concentration, nutrient availability, cytokine and growth factor levels, and juxtacrine interactions with adjacent cell types. These factors converge on signal transduction pathways and transcription factor networks, whose output is further tuned by epigenetic and mutational features of the cell. We next sought to use computational approaches to infer how regulation at these levels contribute to the patient-specific EMP programs we learned from scRNA-seq data.

We used PROGENy's signalling activity model (Holland et al., 2020a, 2020b; Schubert et al., 2018) to calculate relative activity scores for 14 signalling pathways and assessed their activity as a function of EMP program activity in each tumour sample. EMP programs were consistently associated with elevated TGFB, NFkB, and TNFa signalling, but some had distinctly high levels of either EGFR and MAPK signalling or JAK-STAT, Hypoxia, and p53 signalling (**Figure 32A**). This suggests that hypoxia-associated EMP may have distinct features from plasticity coordinated by MAPK/ERK signalling.

Differences in the activity of these pathways could be related to the cells' association with non-malignant cells of the TME, which secrete many of the factors that can promote EMP. Since the data sets we have used included matched gene expression profiles of stromal cells for all tumours, we inferred cell-cell communication for each tumour to identify ligands from the TME that could account for that sample's specific EMP expression program. Several cytokines were associated with EMP in many tumours, including TGFB1, IL6, IFNG, TNFSF15, and others (**Figure 32B**). Consistent with previous literature, macrophages and fibroblasts were dominant sources of many factors contributing to EMP (Dongre and Weinberg, 2019; Puram et al., 2017) (**Figure 32C**). Mesenchymal malignant cells also express various ligands predicted to promote EMP, suggesting that they may establish self-regulatory signalling loops within the TME.

To identify factors that may lead to the differences between the EMP programs with high EGFR/MAPK activity and those with dominant JAK-STAT/Hypoxia/p53, we compared how frequently ligands were associated with each set of programs (**Figure 32D,E**). Interestingly, several ligands expressed most highly in mesenchymal cells themselves were preferentially implicated in MAPK-associated EMP (eg. *YARS, NRG2, CCL28, LGALS3, BMP7*), whereas EMP programs with high STAT/Hypoxia/p53 signalling were associated with ligands from stromal populations (eg. *IFNG*,*HLA-A*, *ADAM17, CXCL12*). Together, this highlights that although malignant populations may have cell-autonomous mechanisms to promote EMP, the molecular features of this plasticity can be modulated by complex interactions with stromal populations in the TME.

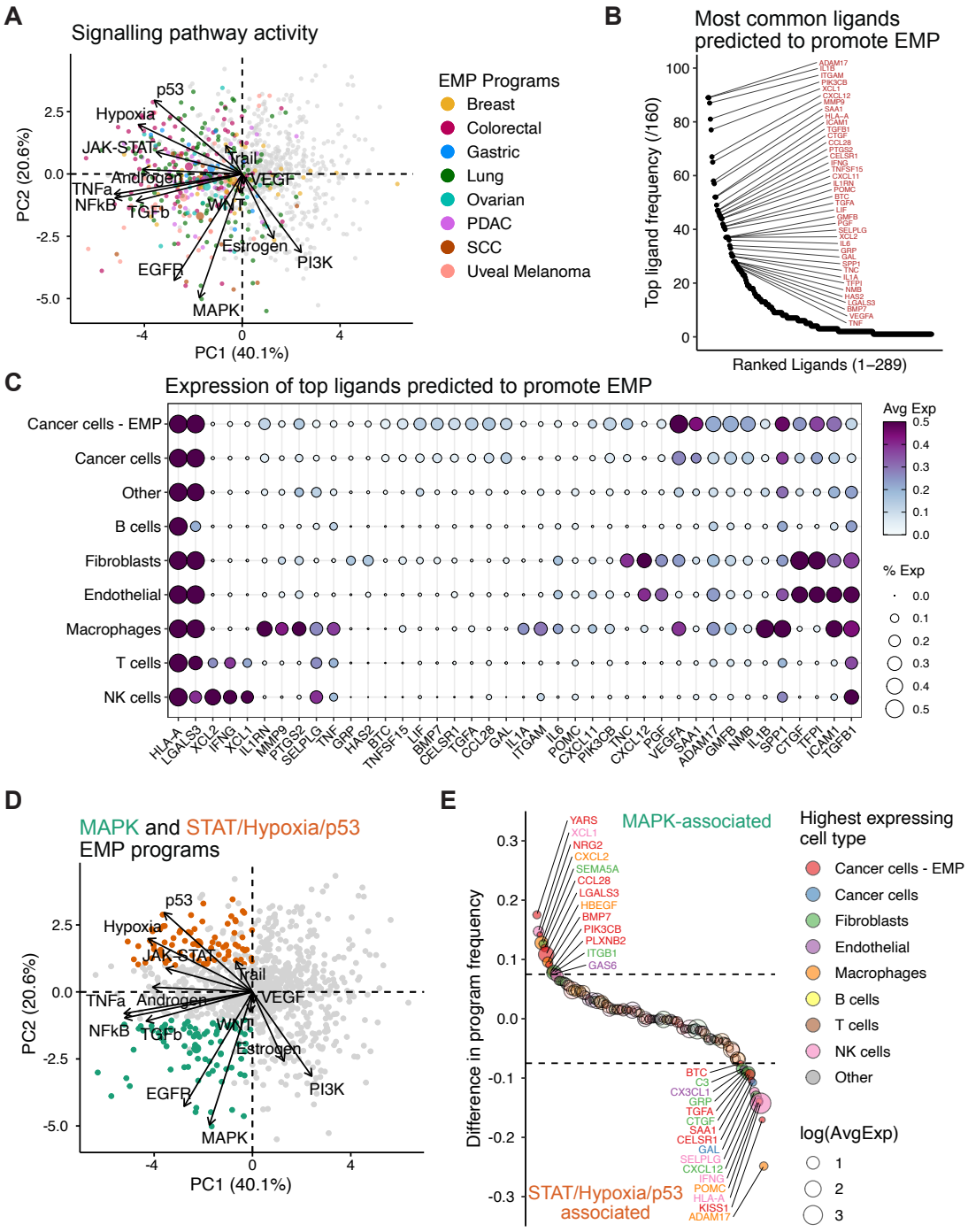
A
Signalling pathway activity
p53
Hypoxia
JAK-STAT
Trail
Androgen
VEGF
TNFa
NFkB
TGFb
WNT
Estrogen
PI3K
EGFR
MAPK
PC2 (20.6%)
PC1 (40.1%)
EMP Programs
Breast
Colorectal
Gastric
Lung
Ovarian
PDAC
SCC
Uveal Melanoma
B
Most common ligands predicted to promote EMP
Top ligand frequency (/160)
Ranked Ligands (1–289)
C
Expression of top ligands predicted to promote EMP
Cancer cells - EMP
Cancer cells
Other
B cells
Fibroblasts
Endothelial
Macrophages
T cells
NK cells
Avg Exp
% Exp
D
MAPK and STAT/Hypoxia/p53 EMP programs
E
MAPK-associated
STAT/Hypoxia/p53 associated
Difference in program frequency
Highest expressing cell type
Cancer cells - EMP
Cancer cells
Fibroblasts
Endothelial
Macrophages
B cells
T cells
NK cells
Other
log(AvgExp)

Figure 32. Regulatory mechanisms contributing to EMP expression programs. **A)** PCA of model coefficients for PROGENy's signalling pathway activity inferences for all expression programs. Grey dots represent programs not associated with EMP. **B)** Counts for the number of EMP programs in which each ligand was implicated as a top regulator. **C)** Expression of top ligands in each cell type. Values represent averages of all 160 tumours. **D)** Same as (A), but coloured to define STAT/Hypoxia/p53 (orange) and MAPK-associated (green) EMP programs. **E)** Ligands inferred to preferentially contribute to MAPK- or STAT/Hypoxia/p53-associated EMP. Values represent the difference in the proportion of programs in which each ligand was implicated.

3.2.5 EMP is frequently associated with non-canonical regulatory factors

As we had previously found that many non-canonical transcription factors were associated with experimentally induced EMTs, we next used a similar approach to identify factors that may be contributing to tumour-intrinsic EMP. We used the computational pipeline implemented in SCENIC to build gene regulatory networks associated with each sample, prune the network to only retain targets with DNA binding motifs of the specific regulator in upstream regulatory regions, and score each cell for their activities of the resulting regulons (Aibar et al., 2017). We then modelled regulon activity as a function of the cells' activity of the tumour-specific EMP program. While effectively all regulons were found to have both increased or decreased activity across EMP programs, many showed a clear bias (**Figure 33**). Of those whose activity primarily increases as the cells become more mesenchymal, we found many of the same factors that we observed in experimentally induced EMTs, including KLF6, KLF4, JUNB, SOX4, and RELB. Factors associated with reduced activity include MYC, MYBL2, and HES6 (**Figure 33)**.

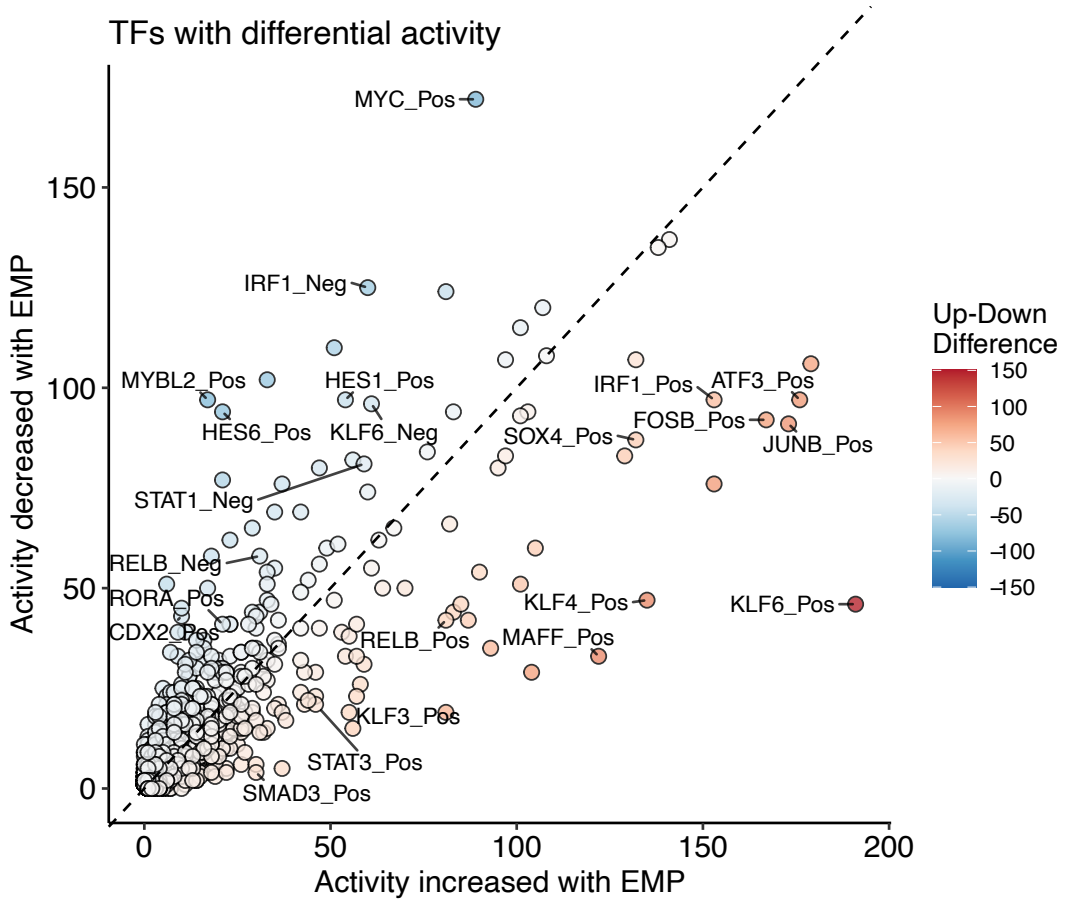

Figure 33. Transcription factor activity contributing to EMP in tumours. Transcription factor regulon activity was modelled as a function of EMP program activity for all programs identified. The plot shows how frequently regulons had increased or decreased activity with EMP activity. Repressive relationships ("_Neg") between transcription factors were also included. For example, decreased activity of a repressive regulon suggests that target genes are repressed with EMP activity, but given the repressive relationship with the transcription factor, the transcription factor's activity is increased. For this reason, positive and negative regulons have opposite trends in many cases.

3.2.6 Summary

EMP has long been appreciated as a prominent source of intratumoural heterogeneity that ultimately promotes tumour progression and hinders effective treatment. Here, we took advantage of its ubiquity to learn gene expression patterns associated with it in 160 tumours of different cancer types. We had previously shown that experimentally induced EMTs are

diverse and this analysis confirmed that this context specificity is not simply an experimental artifact. While EMP-associated expression was variable, we identified a set of 289 genes that were most frequently associated with the mesenchymal phenotype. We were able to refine this to include the 160 genes with high specificity to malignant populations, allowing us to define EMP activity in bulk RNA-seq data from tumours. Despite the diversity of the cells' gene expression, we found consistent patterns in signalling activity involving TGFB, MAPK-ERK, and STAT pathways.

3.3 Paracrine signalling dependencies of EMP

3.3.1 Kinase inhibitor screens reveal diverse dependencies on signalling activity

Paracrine signalling is another regulatory feature likely to coordinate the EMT across a population of cells (Dongre and Weinberg, 2019; Dongre et al., 2017; Scheel et al., 2011; Yao et al., 2019). In fact, we found that the expression of secreted factors spanning a variety of signalling pathways broadly increased in each of our 12 time course experiments (**Figure 34A**). Given this, we next established an experimental design to mechanistically assess the dependence of the EMT on multiple signalling pathways and compare these dependencies across contexts. We curated a selection of 22 small molecule inhibitors targeting a variety of kinases and treated cell lines alone for 7 days, or in combination with one of the three EMT inducers previously used (**Figure 34B**). Leveraging MULTI-seq to multiplex samples, we ultimately generated scRNA-seq profiles for 45,911 cells across the 384 distinct conditions.

From retrieved cell counts alone, drop-out patterns from cell line-dependent and -independent cytotoxic/cytostatic effects can be observed (**Figure 34C**). By determining the number of differentially expressed genes following treatment of control cells with these inhibitors also provides insight into the baseline effects of these compounds (**Figure 34D**). To assess the impact of these inhibitors on EMT progression, however, we calculated pseudotime values for the inhibited cells using the models built from corresponding time course experiments of the same cell line and EMT inducer (**Figure 35A**). From this, we could identify inhibitors that reduced cells' pseudotime values at 7 days compared to uninhibited controls, therefore dampening the EMT response. LY364947 (TGFBR1 inhibitor), for example, abrogated TGFB1-induced EMTs (**Figure 35A,B,C**), and Erlotinib and Gefitinib (EGFR inhibitors) consistently blocked the effects of EGF (**Figure 35A**).

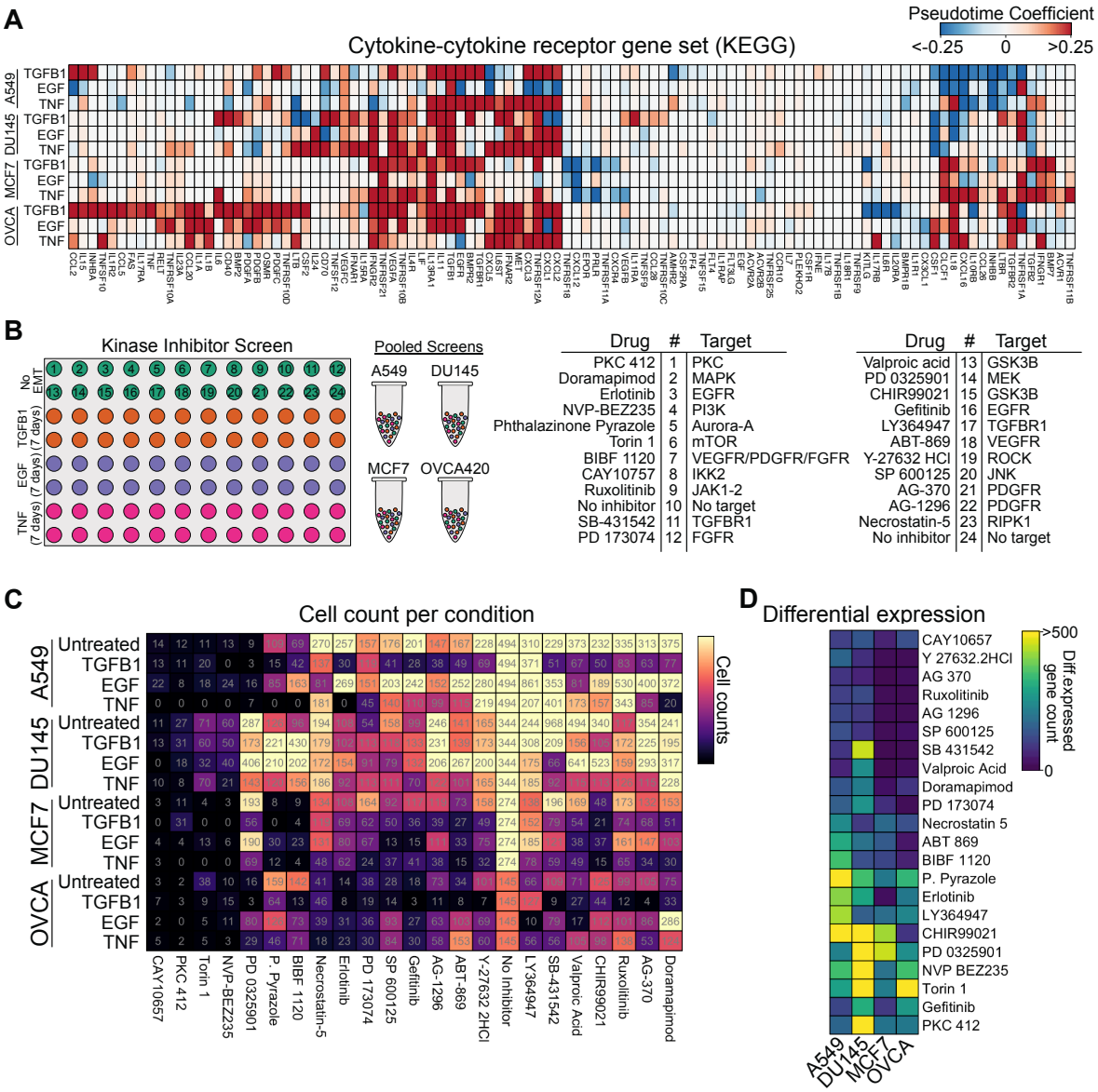

Figure 34. Kinase inhibitor screens to identify regulatory dependencies of the EMT. A) Heatmap showing EMT-associated changes of the individual genes of the KEGG Cytokine-cytokine receptor gene set, only listing those with a significant change in at least one time course experiment. **B)** Schematic of the 384-sample experimental design for the kinase inhibitor screen. **C)** Heatmap showing the number of cells annotated to each condition after demultiplexing the scRNA-seq data. **E)** Summary of the number of genes that are differentially expressed in each cell line exposed to the inhibitors without EMT induction, representing a baseline effect of each compound.

The effects of these inhibitors, however, weren't limited to blocking the direct signalling of the EMT-inducing factor. For example, TGFBR1 inhibition partially blocked EMT progression in a variety of conditions, including EGF-treated A549 and OVCA420 cells, and TNF-treated A549 and MCF7 cells (**Figure 35A**). This suggests that activation of paracrine TGFB1 signalling may be critical for EMT progression following a variety of initial stimuli, supporting previous work showing the dependence of the EMT on transcription-factor-activated TGFB1 autocrine loops (Gregory et al., 2011; Larocca et al., 2013; Yeh et al., 2018).

Effects of direct EGFR inhibition with Erlotinib and Gefitinib were largely restricted to EGF-treated EMT responses, but inhibition of its downstream kinase MEK (with PD 0325901) hindered the EMT response in TGFB1-treated A549 and MCF7 cells. Non-canonical TGFBR1 signalling through MEK/ERK has been previously reported (Principe et al., 2017; Xie et al., 2004), and two recent studies have proposed a MEK-dependent regulatory checkpoint in the EMT (Chen et al., 2020; McFaline-Figueroa et al., 2019). While our data for TGFB1-treated A549 and MCF7 cells are consistent with these findings, it also demonstrates that this checkpoint is not universal, even among other TGFB1-induced EMT responses, as DU145 and OVCA420 cells are not susceptible to MEK inhibition (**Figure 35A**).

3.3.2 A novel role for the kinase RIPK1 in the EMT response

Inhibition of RIPK1—a kinase involved in activating NFkB and necroptosis pathways—with Necrostatin-5 (Nec-5) blocked EMT progression in all of the same conditions as TGFBR1 inhibition. Nec-5-treated cells, however, consistently had higher pseudotime values than TGFBR1-inhibited cells, suggesting a partial EMT response (**Figure 35A,B**). To determine if the partial response corresponds to reduced magnitude of gene expression changes, or a selective inhibition of a subset of genes, we assessed expression levels of all genes

differentially expressed following TGFB1 treatment. In each case, RIPK1 inhibition only abrogated a subset of TGFB1-induced expression changes, producing a partial EMT response (**Figure 35C,D**). Importantly, we note that this partial response with RIPK1 inhibition is not due to a temporal block in EMT progression (i.e. preventing late EMT dynamics), as inhibition does not exclusively prevent late response genes. This suggests that the EMT involves multiple independent regulatory modules that can be perturbed without impacting others.

To our knowledge, no direct cross-talk between the TGFB1 signalling and RIPK1 has been documented, but loss of RIPK1 has been previously associated with an enhanced epithelial phenotype, reduced ERK1/2 phosphorylation, and reduced transcriptional activity of the AP-1 complex (Li et al., 2019a; Yonekawa et al., 2015). To determine if RIPK1 inhibition prevents the activation of AP-1 targets in our EMT models, we assessed the enrichment of transcription factor binding motifs in the promoters of genes that failed to change throughout the EMT in RIPK1-inhibited cells. We found that the AP-1 binding site (JUN/FOS, BACH2 motifs) was the most enriched in promoters of genes that failed to become upregulated in Nec-5 treated cells in response to TGFB1 (**Figure 35E**). Other enriched motifs include EGR1 and PAX4 binding sites. Both AP-1 and EGR1 can be activated through ERK1/2 signalling, providing a possible mechanistic link between RIPK1 and these transcriptional changes (Tarcic et al., 2012; Yonekawa et al., 2015). As ERK1/2 is a downstream effector of MEK, this may also explain the previously proposed MEK checkpoint of the EMT (Chen et al., 2020; McFaline-Figueroa et al., 2019). While it is still unclear how RIPK1 becomes activated, this regulatory axis is conserved in every condition we assessed that is also dependent on TGFB1 signalling (based on similarity to TGFBR1 inhibition), and may represent a common, though not universal, regulatory network of the EMT.

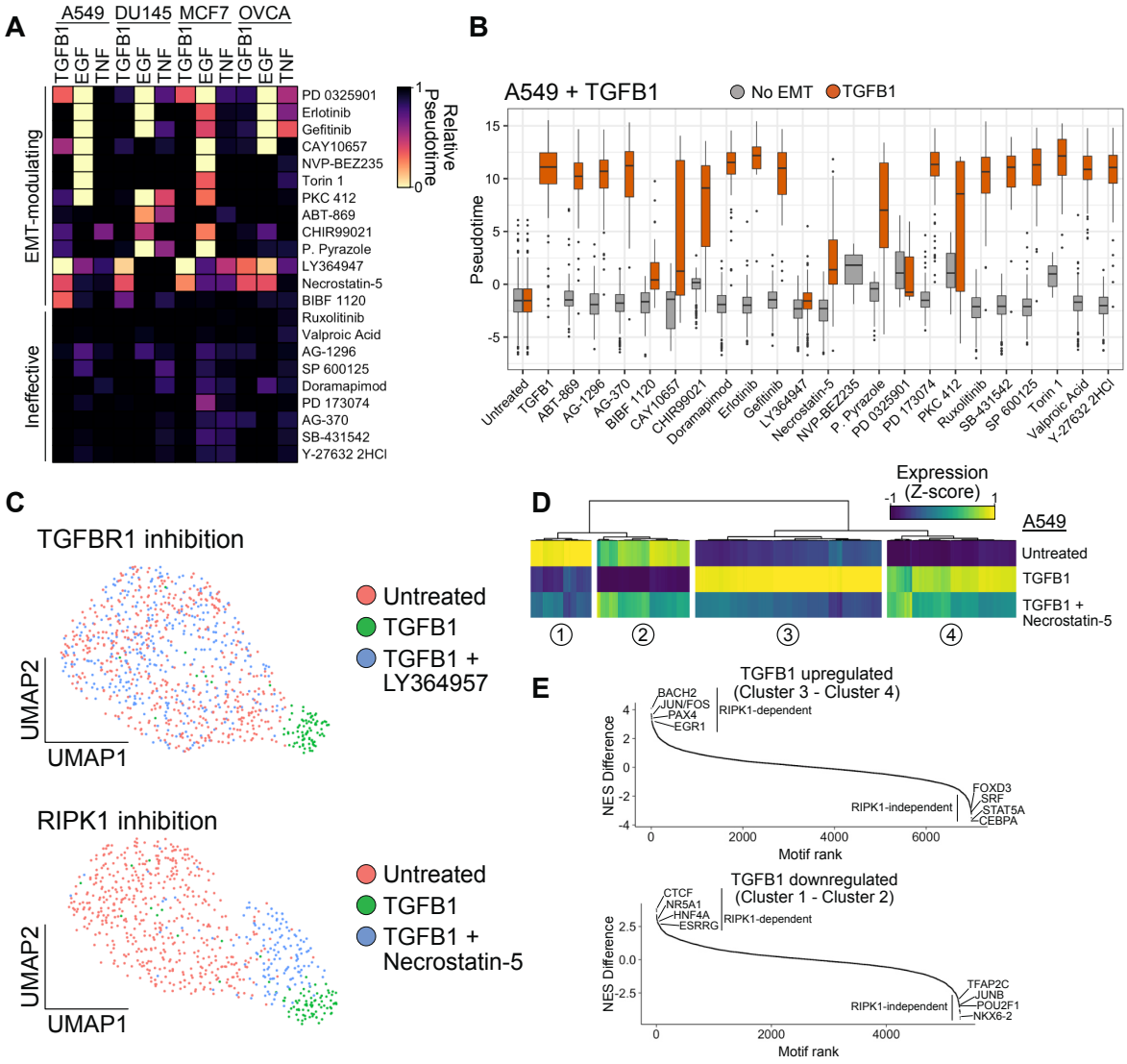

Figure 35. RIPK1 is a novel regulator of diverse EMT responses. A) Average pseudotime values calculated for each condition. Pseudotime values of 1 correspond to cells at the end of an uninhibited EMT response. Conditions with values <1 suggest impairment of the response. **B)** Boxplots showing the distribution of pseudotime values for A549 cells treated with the inhibitors alone (grey) or in combination with TGFB1 (orange). **C)** UMAP embeddings of untreated A549 cells with those had been treated with TGFB1 alone or in combination with the TGFBR1 inhibitor LY364947 (top), or the RIPK1 inhibitor Necrostatin-5 (bottom). **D)** Heatmap showing expression (Z-score) of genes differentially expressed in A549 cells by TGFB1 in control, TGFB1, and TGFB1 and Nec-5 inhibited cells. **E)** Difference in normalized enrichment scores for transcription factor targets in the genes that are successfully inhibited by Necrostatin-5 compared to those that are not. Positive values correspond to regulons that are enriched in Necrostatin-5-inhibited genes, whereas negative values represent those not affected by Necrostatin-5.

3.3.3 Inferred signalling activity predicts susceptibility to kinase inhibition

Diversity of EMP programs could introduce challenges for effectively preventing plasticity, but the dependence of EMP on factors from the cells' microenvironment suggests that the diversity likely arises from combinatorial effects from the relatively limited number of signal transduction pathways. Therefore, we suspect that diverse EMP programs may be susceptible to common pathway perturbations and rational treatments could be devised by inferring signalling activity associated with EMP in a given tumour.

To begin to test this prediction, we first explored the MIX-Seq data set comprising scRNA-seq profiles of over 100 cancer cell lines treated with various drugs, including the MEK inhibitor Trametinib (McFarland et al., 2020b). We used ACTIONet to define cell line-specific EMP programs from untreated expression profiles, identifying high-confidence EMP programs in 46 of the 99 lines we assessed (lines with >100 cells; **Figure 36A,B**). Many others had programs that correlated well with individual EMT gene sets and may represent EMP, but out of caution, we did not annotate them as such. We then inferred changes in signalling pathway activity associated with all cell state programs and found that, like in the tumour samples, TGFB/NFkB/TNFa activity was consistently higher in EMP programs, but programs could be distinguished by high EGFR/MAPK or STAT/Hypoxia/p53 signalling (**Figure 36C**).

To determine if MEK inhibition preferentially limits MAPK-associated EMP, we used the MIX-Seq dataset to assess the effects of trametinib on all EMP programs. Trametinib had diverse effects across cell lines and even increased EMP program activity in several, but in general, the programs inferred to be associated with high MAPK signalling had reduced activity following MEK inhibition (**Figure 36C,D**). Given the consistent association of EMP with TGFB1 and NFkB signalling, we would also predict that inhibition of these pathways

could restrict EMP. While TGFB1 and NFkB inhibitors were not included in the MIX-Seq data set, we assessed our own scRNA-seq data of four cancer cell lines (A549, DU145, MCF7, and OVCA420 without EMT induction) treated with the TGFBR1 inhibitor LY364947 (**Figure 37A**). Performing the same analysis on this data, we found that inhibition of TGFBR1 in cells cultured in control conditions led to repression of most EMP programs, proportional to the inferred level of TGFB1 activity associated with each line's intrinsic EMP program (**Figure 37B,C**).

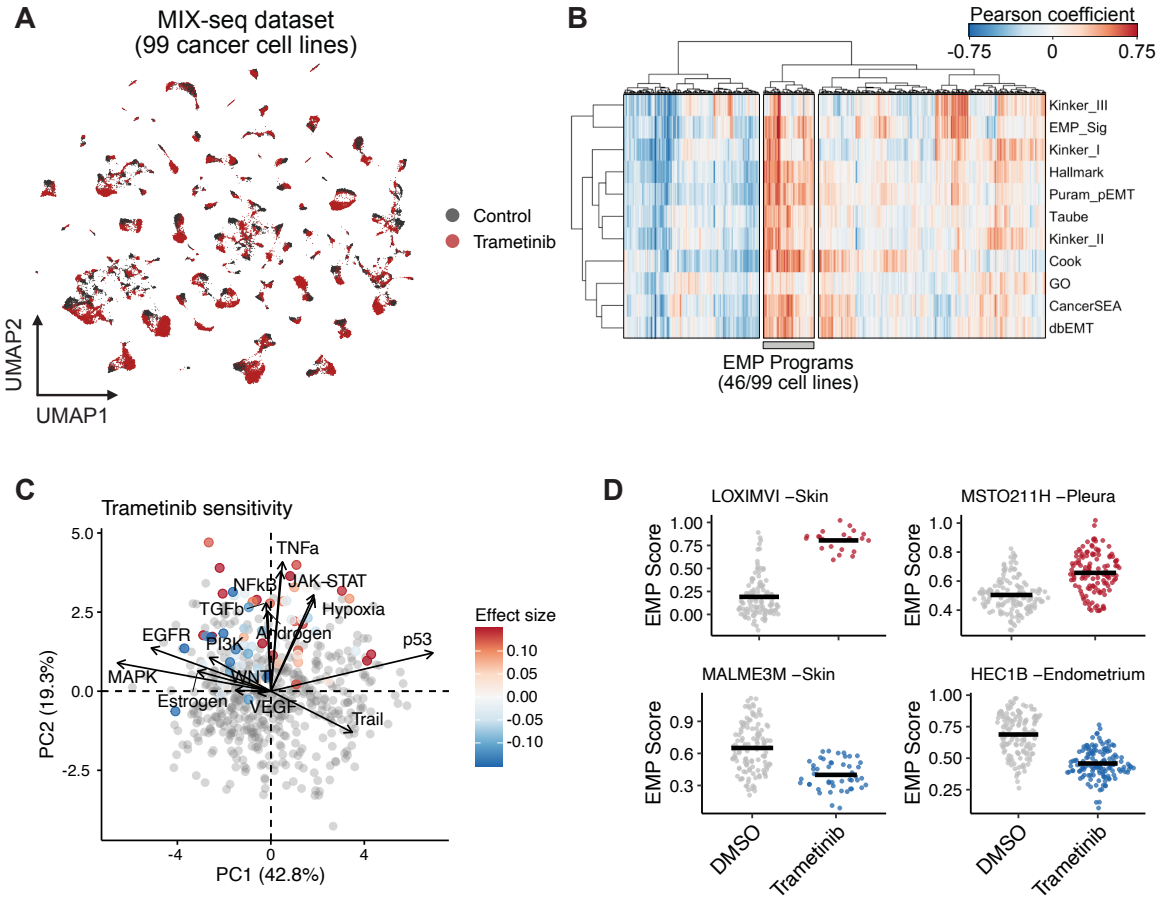

Figure 36. Assessing the effects of MEK inhibition on intrinsic EMP. A) UMAP embedding of 99 cancer cell lines from the MIX-seq data set. **B)** Pearson correlation coefficients of archetype activity scores and EMT gene set scores for all cells in the MIX-seq data set. **C)** PCA of model coefficients for PROGENy's signalling pathway activity inferences for all of the cell state programs. EMP programs are coloured by the effect trametinib has on its activity. **D)** Examples of cell lines whose specific EMP program is enhanced (top) or limited (bottom) by trametinib.

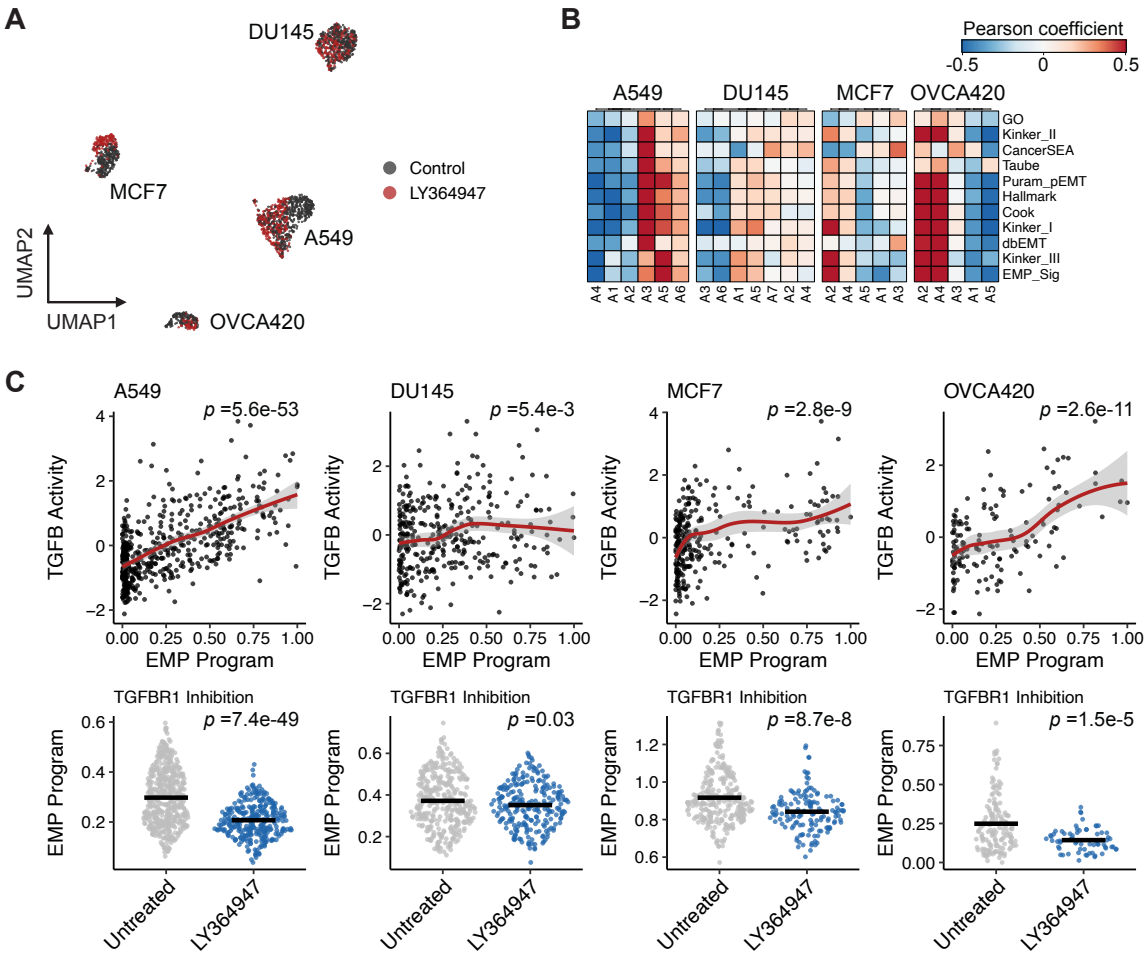

Figure 37. Assessing the effects of TGFBR1 inhibition on intrinsic EMP. A) UMAP embedding of control A549, DU145, MCF7, and OVCA420 cells from the kinase inhibitor screen. **B)** Correlation of each cell lines' archetypes with EMT gene set scores. **C)** Effect of TGFBR1 inhibition by LY364947 on sample-specific EMP programs of A549, DU145, MCF7, and OVCA420 cell lines. p-values were calculated from linear models for each condition and were all corrected with the Benjamini-Hochberg method.

3.3.4 Summary

The ability to therapeutically restrict plasticity in tumours could greatly improve the efficacy of existing treatment options. To identify regulatory dependencies associated with EMP, we used data from kinase inhibitor screens and assessed the ability of various inhibitors to restrict EMT responses and tumour-intrinsic EMP. From our previous analysis, we had predicted that TGFB signalling activity was often associated with EMP, whereas only a

subset of EMP programs were associated with ERK signalling. Inhibiting these pathways led to restriction of EMP-associated expression, approximately proportional to the inferred activity of the pathway in the sample. Together, this supports that diverse EMP programs may still have a core set of common dependencies and personalized strategies for restricting EMP can be discovered by inferring these regulatory mechanisms.

Chapter 4 - Discussion

4.1 A consensus molecular program for EMP does not exist

4.1.1 Molecular features of EMP

The EMT has long been associated with several canonical molecular events. These include the cadherin switch from E-cadherin to N-cadherin; the increased expression of the transcription factors Snail, Slug, Zeb1, Zeb2, and Twist1; and increased abundance of the intermediate filament Vimentin. However, co-expression patterns of these mesenchymal genes are inconsistent. For example, among a panel of non-small cell lung cancer cell lines, N-cadherin and Vimentin were only occasionally activated in the lines expressing high levels of ZEB1 (Gemmill et al., 2011). Beyond these limited canonical markers, global expression patterns also appear diverse. Microarray analysis of three cell lines exposed to a combination of TGFB1 and TNF alpha resulted in EMT responses with only 10-30% of differentially expressed genes shared between conditions (Peixoto et al., 2019). And in a single mammary epithelial cell line, TGFB1 treatment and a spontaneous EMT induction model resulted in responses with only an approximately 50% overlap in their differentially expressed genes (McFaline-Figueroa et al., 2019).

In an attempt to define unifying molecular features of EMP, we adopted quantitative approaches to systematically compare expression patterns associated with tumour-intrinsic EMP and experimentally induced EMTs. In both cases, we have found notable diversity in their associated gene expression programs. We found that canonical markers were involved quite infrequently. In some cases (eg. Snail expression in the time course data), this may simply reflect the detection limitations of current scRNA-seq methodologies (Lähnemann et al., 2020), but in many cases, these factors were readily detected, but not dynamically expressed.

While several genes were implicated in many of the contexts we explored, no gene was a perfect indicator of the transition and the majority of genes were involved in a limited number of contexts. From the signatures we derived from the most conserved genes in experimentally induced EMTs (86 genes) and tumour-intrinsic EMP (289 genes), 28 genes were present in both. Given their properties, several of these genes have clear functions in mesenchymal cells. TGFB1 is perhaps the most frequently studied inducer of the EMT (Xu et al., 2009). ITGA2 is an integrin that mediates adhesion to the ECM (Valery Adorno-Cruz, 2019). IL32 is an RGD motif-containing cytokine with high affinity for integrins and stimulates MAPK/ERK signalling (Wen et al., 2019). Some of the factors have been implicated in EMT, but their role is less understood. TIMP1 is an MMP inhibitor, which could presumably reduce the invasive capacity of cells, but has been shown to induce EMT through MMP-independent signalling (D'Angelo et al., 2014). KRT7 is an epithelial keratin, but it has been shown to activate TGFB signalling in ovarian cancer cells (An et al., 2021). PMEPA1 is a transmembrane protein that has been linked to non-canonical TGFB signalling and EMT in colorectal cancer (Zhang et al., 2019). SERPINE2 has been implicated in the EMT in a variety of contexts, but its precise function is not known (Zhang et al., 2020). EMT has been associated with reduced MHC class I expression (Terry et al., 2017), but HLA-A/C were in both signatures. Similarly, the basement membrane proteins LAMB3/C2 are typically associated with epithelial cells and yet were in both signatures (Scanlon et al., 2013). Perhaps the expression of these genes are feedback mechanisms to counter EMT responses or following post-transcriptional regulation of these proteins.

4.1.2 Regulatory constraints on EMT responses

A cell's dynamics are constrained at multiple levels, producing diverse responses to common stimuli. Responses to many extracellular signals are mediated through the receptors expressed on the cell's membrane and the cocktail of signalling factors in its

cytoplasm. The evolution of complex ligand-receptor families, the promiscuity of intracellular signalling proteins, and stoichiometric properties of their interactions makes the response to even a single ligand challenging to predict (Klumpe et al., 2020; Su et al., 2020). Signalling pathways often converge on transcription factors that bind chromatin at specific DNA motifs to regulate gene expression of hundreds to thousands of target genes. The ability of these factors to bind target regions is constrained by the position of nucleosomes along chromatin, the biochemical properties provided by histone modifications, and the three-dimensional organization of chromatin within the nucleus. Variation in DNA sequences at regulatory regions can affect the strength of a gene's expression (The GTEx Consortium, 2020), and mutations in the protein-coding regions can prevent it from being expressed at all. Presumably the diversity we see in EMP-associated dynamics is attributable to variation at each of these layers of regulation. However, it shouldn't be ignored that, while incomprehensibly complex, these regulatory networks have evolved for millennia to promote dynamics that lead to favourable outcomes. So in one regard, the complexity of regulation makes it unsurprising that EMP-associated dynamics are so diverse. But if EMP is such an important phenomenon, it may be expected that regulatory networks would have evolved well-defined attractor states for the mesenchymal state like has emerged for proliferation and apoptosis responses. It seems increasingly likely that, despite the common conceptual model, processes that have been defined as EMP are not simply a well-defined interconversion between distinct E/M phenotypes, but rather represent a more general, diverse plasticity. This will be discussed further in later sections.

4.1.3 Functional conservation among diverse gene expression programs

It is interesting to consider that in nearly all contexts, although the specific genes contributing to EMP were variable, the involved genes still enriched for gene sets associated with mesenchymal properties, such as cell motility, cytoskeleton remodelling, and ECM

degradation. Although limited by the unknown quality of these gene sets, it is reassuring to see that despite diversity of the gene expression programs, functional properties may be conserved. Additionally, this could suggest that unlike well-defined processes such as cell cycle progression, these processes may be more variable and nuanced. One could imagine distinct combinations of proteases, each capable of—broadly—degrading the ECM, or different repertoires of membrane proteins that facilitate motility on diverse substrates. Over decades of accumulating gene expression data associated with broadly defined phenomena, many gene sets have likely become conglomerates of genes involved in loosely related processes across distinct contexts. Perhaps this represents unappreciated nuances of these cellular properties, or that distinct baseline states require different dynamics to accomplish common tasks, or perhaps it is effectively redundant variation. With recent advances in genome-wide loss-of-function screening, it would be interesting to compare basic cellular properties across contexts. For example, basic cell migration assays could be used as a selective process in a pooled CRISPR library screen to identify critical mediators of migration.

4.1.4 A unifying molecular program for EMP does not exist

Collectively, our work suggests that a consensus molecular program does not underlie the EMT. This conclusion has been based on transcriptomic data, and so it could be argued that common features exist at post-transcriptional levels of regulation. While this may be true for certain features of the EMT, such as internalization of epithelial junction proteins (Aiello et al., 2018), it does not account for accumulating effectors of the mesenchymal phenotype (eg. ECM receptors, cytoskeleton proteins), which are absent from epithelial cells. It could also be argued that this is simply a phenomenon of EMP in cancer and not healthy systems. While it seems likely that development has evolved to ensure that the various EMTs responsible for organogenesis are reproducible, it would not be surprising if the EMT

involved in gastrulation is the same as the one involved in adult wound healing. Unfortunately, this suspicion is not yet supported by data, however distinct EMT responses have been observed in healthy mammary epithelial cells (McFaline-Figueroa et al., 2019).

4.2 Hybrid states are innumerable and polar states are undefinable

4.2.1 The partial EMT and hybrid phenotypes in cancer

Observations of EMTs without the full repertoire of canonical changes have often been reconciled as a partial EMT (pEMT). The supposed E/M hybrid phenotype resulting from a pEMT has received increased attention over the last several years. This is particularly the case in cancer, where the hybrid phenotypes have been reported to be highly prevalent and have unique properties from the polar phenotypes in their contribution to tumour progression (Pastushenko and Blanpain, 2019).

The EMT in cancer cells has frequently been shown to produce cells with properties of stem cells, loosely defined by the cells' tumour propagating potential (Mani et al., 2008; Morel et al., 2008). In various models, however, hybrid phenotypes co-expressing epithelial and mesenchymal markers have been shown to have higher stemness properties than polar phenotypes. In a mouse model of breast cancer, tumourigenicity was dependent on implantation of hybrid cells (Kröger et al., 2019). Tumour- and spheroid-initiating capacity of pancreatic cancer cells were higher in those co-expressing EPCAM and Vimentin than those positive for only one of the markers (Ruscetti et al., 2015).

While polar mesenchymal cells do seem to be more invasive (Ruscetti et al., 2015), hybrid phenotypes may have higher metastatic potential. Cells with a stable mesenchymal state through experimental over expression of EMT-associated transcription factors form fewer

metastases due to the inability to undergo an MET (Ocaña et al., 2012; Tsai et al., 2012). Intravenous injection of hybrid skin cancer cells, however, produces more lung metastasis than EPCAM- cells (Latil et al., 2017). In fact, polar mesenchymal cells may not be involved in metastasis at all: both invasive cells at the leading edge of head and neck squamous cell carcinoma tumours and a large proportion of circulating tumour cells in breast cancer patients co-express epithelial and mesenchymal markers (Puram et al., 2017; Yu et al., 2013).

This relevance in cancer led to much discussion about the stability and population structure of these states. Mathematical models based on stereotypical regulatory networks with mutually inhibitory feedback loops (eg. regulatory relationships between EMT transcription factors and the inhibitory miR-200 and miR-34). These models have predicted metastable hybrid states and cell state dynamics that give rise to multimodal distributions along an E/M continuum (Jolly et al., 2015). Pastuchenko et al. (2018) proposed six distinct hybrid states based on combinatorial expression of EPCAM, CD106, CD51, and CD61. We note, however, that scRNA-seq of sorted populations based on these markers show gene expression patterns that reflect a fairly continuous relationship between these phenotypes (Pastushenko et al., 2018).

In contrast to models predicting phenotypes with distinct stability and multimodality along the E/M continuum, our analysis of malignant cell gene expression from 160 tumours demonstrates the majority are associated with a monomodal and approximately Gaussian distribution of E/M phenotypes. This is true whether using EMT gene set scores, sample-specific archetype scores, or even looking at cells' distribution along components of variation (eg. along principal components). This suggests that either these models are inaccurate, or that EMP exists at a frequency lower than we can detect through the

measurement of hundreds-to-thousands of cells per sample, which seems unlikely. We also note that the tumours' average E/M state can vary and the extent of phenotypic variation within a tumour ranges widely. If distinct polar and hybrid states do not exist, it is intriguing to question if it is more relevant to a tumour's biology that its malignant cells are more mesenchymal on average, or if they span a wider range of E/M phenotypes.

4.2.2 The partial EMT is a poorly defined concept

The semantics of polar and hybrid E/M states are also concerning from a molecular perspective. We've demonstrated that an individual epithelial cancer cell line can undergo markedly different EMT responses when exposed to different inducing factors. For example, in A549 cells, both TGFB1 and TNF promote a mesenchymal morphology, yet only share approximately 50% of their differentially expressed genes. It could be argued that perhaps this intersecting portion of the response represents the true EMT response, while the remaining genes correspond to EMT-independent components. Given that there is no ground truth definition of EMT, this cannot be decidedly countered, however several stereotypical EMT markers show distinct differences: TGFB1 induces both *FN1* and *SNAI2*, whereas TNF does not.

This suggests that even if one could define a fully epithelial state (which we have not addressed here, but any cursory exploration of epithelial markers will show variability across epithelial cell types), EMT responses are numerous and it seems unlikely that this phenotypic variability can be reduced to a one-dimensional continuum. If E/M phenotypes span a multidimensional space, a single polar mesenchymal state cannot be defined. And considering that we have shown that cells span the continuous space of E/M phenotypes, the number of possible hybrid states co-expressing some mixture of E/M components are effectively innumerable and likely have diverse properties. It should be noted, however, that

it is very likely that various constraints limit the cell's ability to transverse much of this phenotypic space, but we are far from defining these limits. Therefore, I argue that perhaps beyond high-level functional properties of the cell, the notion of polar and hybrid states is an ill-defined oversimplification that likely fails to capture relevant properties of cells. If we continue through this -omics era with a reductionist lens, it seems likely that we will see an increasing number of studies reporting pEMTs and hybrid phenotypes to reconcile the complexity of this phenomenon. Moving forward, I believe that conceptual models of EMP need to revise the linear continuum and embrace this complexity.

4.3 Conserved signatures have utility for learning clinical features of EMP

4.3.1 Utility and limitations of gene sets for the interpretation of biological data

Cellular processes involve the coordinated action of many genes. As such, biologists have long been interested in curating lists of genes involved in specific processes (Ashburner et al., 2000). These gene sets provide researchers with a list of relevant genes to query if they are interested in a particular phenomenon, or allow them to take a list of differentially expressed genes and identify cellular processes that may be affected by these changes. These gene sets are invaluable for the interpretation of large-scale data sets, but their use should be approached with caution. Several limitations make it easy to draw misleading conclusions. Not all gene sets are of the same quality. For example, all GO terms are associated with evidence codes that range from direct experimental validation to inferred function from protein sequence. There is also an inherent incompleteness of gene sets as they are based on the current state of knowledge. Annotation bias is also problematic, where 58% of GO term annotations represent only 16% of human genes (Tomczak et al., 2018). Beyond these technical limitations, there are also biological considerations.

Frequently, only subsets of a gene set are contained in a list of differentially expressed genes, and while the number of genes may be higher than expected by chance, it's unclear if that specific subset is capable of affecting that cellular function. Gene pleiotropy can lead to hits in irrelevant gene sets.

Genes associated with heterogeneous cellular processes are particularly challenging to catalogue. Many EMT gene sets have been generated over the years, but given the increasing number of genes shown to affect E/M phenotypes, the quality of these gene sets has been questionable. The Molecular Signatures Database (MSigDB) has attempted to reconcile this issue by developing a series of 50 "Hallmark" gene sets (including EMT), refined from existing databases to reduce gene set redundancy and maximize relevancy of their genes (Liberzon et al., 2015). In our analysis, this hallmark gene set correlated well with EMT processes in our time course experiments, but when we explored which genes from the set were differentially expressed, there was inconsistent involvement across contexts and a large portion of the gene set was not differentially expressed in any condition. When identifying EMP programs in tumours, we assessed 10 different EMT gene sets that contained a total of 930 genes and found minimal overlap between gene sets. This variability of gene sets derived from different sources may highlight the diversity of EMT responses, but is concerning if trying to interpret new data. The purpose of gene sets is not to represent the critical genes for a given process, but rather to be indicators of the process with optimal sensitivity and specificity. We derived signatures from the genes most frequently involved in both experimental EMTs and tumour-intrinsic EMP, producing a gene set with high sensitivity for this diverse process. A common specificity issue with EMT gene sets is that fibroblasts express many of the same genes as mesenchymal cells, confounding the interpretation of bulk RNA-seq data (Isella et al., 2015; Izar et al., 2020; Puram et al., 2017). In our survey of EMP programs from 160 tumours, we had scRNA-seq data from

stromal components, allowing us to refine our EMP signature to only contain genes specifically expressed in malignant cells.

4.3.2 EMP is associated with poor prognosis and immunosuppression

We used this refined EMT gene set to assess clinical features from the pan-cancer TCGA database and demonstrated that higher gene set scores (ie. a more mesenchymal phenotype) were associated with reduced progression free survival and changes in immune cell proportions consistent with an exhausted, immunosuppressive environment. Previous analysis of molecular subtypes of various cancer types have suggested mesenchymal tumours have a worse prognosis, however it has been unclear to what extent this has been associated with the proportion of fibroblasts in the tumours. For example, in colorectal cancer, the EMT-associated CMS4 molecular subtype is associated with reduced survival (Guinney et al., 2015), but this mesenchymal signature has been directly attributed to stromal content (Isella et al., 2015). The same is true for ovarian cancer (Chen et al., 2018a; Izar et al., 2020).

The association of mesenchymal cells with immunosuppressive environments is not novel, but evidence is certainly still limited. Studies have often focused on a subset of immune cells in the TME, so our analysis provides valuable insight into the global immune landscape of the thousands of tumours we assessed. Dongre et al. (2017) found that murine breast tumours generated from mesenchymal cells were associated with an increased proportion of regulatory T cells and exhausted CD8 T cells, consistent with our findings. Expression of the EMT transcription factor Snail also has been attributed to resistance to checkpoint blockade, suggesting that EMP may limit the clinical efficacy of immunotherapies (Kudo-Saito et al., 2009), further suggesting that developing strategies to restrict this plasticity is critical.

4.4 Dependence of mesenchymal phenotypes on extracellular signals

For quite some time, I have been intrigued by the imbalance in the number of factors and conditions that promote an EMT compared to those that promote an epithelial phenotype. As discussed previously, effectively all signalling pathways and various environmental conditions have been implicated in the EMT. In 1997, Dr. Steven M. Frisch (West Virginia University) proposed the hypothesis that the epithelial cell is a default phenotype (Frisch, 1997). Simply in the absence of tissue-specific (or mesenchymal) factors, the epithelial phenotype will be promoted. The epithelial phenotype is the first to emerge in embryonic development, with blastomeres of the 8-cell embryo exhibiting epithelial features. Frisch also argues that epithelial promoters use ubiquitous transcription factors (eg. the E-cadherin promoter has binding sites for Sp1, NF1, and AP2). However, in the years since, several transcription factors have been associated with the epithelial phenotype, including ELF3 (Sengez et al., 2019; Yeung et al., 2017), OVOL1/2 (Roca et al., 2013; Watanabe et al., 2019), and GRHL factors (Sundararajan et al., 2020).

It is intriguing to consider the dependence of the EMT on extracellular signals. We have demonstrated that simply removing the cells' conditioned supernatant following an EMT is sufficient to promote reversion to an epithelial phenotype, indistinguishable from the cells' state prior to induction. This is consistent with the observation that EMTs are most often transient processes in development and tissue homeostasis (eg. epithelial and endothelial repair) and cells revert to an epithelial phenotype upon the loss of these cues (Derynck and Weinberg, 2019). Irreversible transitions have been proposed to contribute to fibrosis in the lung, liver, and kidney (Yang et al., 2020), but lineage tracing experiments to identify the source of fibrosis-associated myofibroblasts have not yielded conclusive evidence that they

are of an epithelial origin (Iwano et al., 2002; LeBleu et al., 2013). However, even an irreversible EMT in fibrosis may only be irreversible due to chronic inflammatory signals within a fibrotic niche that could support a mesenchymal phenotype long-term (Wynn, 2008). By inferring the activity of 14 signalling pathways within individual cells of 160 tumour samples, we were able to assess how signalling activity changes throughout EMP. With exception to PI3K-AKT signalling, all pathways showed increased activity—albeit to varying extents—in the cells with a more mesenchymal phenotype. Certainly additional pathways that were not included in the analysis may show different patterns, but this surprisingly supports the large body of research implicating these diverse pathways in the EMT. Further, it is consistent with the notion that the mesenchymal state in tumours may simply be a general phenotype reflective of adapting to environmental signals rather than one whose primary role is functional (ie. migrating and invading through tissue). This will be explored further in an upcoming section.

4.5 Developing rational strategies to restrict EMP

Therapeutic restriction of EMP holds the promise of reducing metastasis, sensitizing tumours to existing treatments, and reinvigorating anti-tumoural immunity. One could imagine several strategies for accomplishing this: interfering with the cues that initiate plasticity, therapeutically impairing transcriptional regulatory mechanisms (eg. inhibiting histone modifying proteins or transcription factors), or targeting effector proteins associated with a given cell state (eg. neutralizing immunosuppressive cytokines released from mesenchymal cells). The latter may be challenging due to the diversity of EMP phenotypes, but is perhaps the most direct strategy for preventing undesirable features of mesenchymal cells. While this diversity introduces challenges for effective therapeutics, our work suggests

that these diverse programs may have common dependencies that can be inferred from their molecular features and exploited therapeutically.

Given the potential impact, many groups have been interested in targeting EMP from various angles. Most preclinical studies have explored inhibition of various kinases that promote EMT (eg. TGFBRI/II, EGFR, HGFR, STAT3), although various approaches to impair functional components of mesenchymal cells have been proposed (eg. disrupting integrin interactions, MMP inhibition) (Yari Fontebasso, 2015). Several strategies have shown promising results. Inhibiting the RTK AXL with a small molecule drug (BGB324; Bemcentinib) in mouse models of pancreatic cancer reduces the expression of mesenchymal genes and reduces the frequency of immunosuppressive myeloid cells in tumours. Additionally, combination therapy of BGB324 with the chemotherapy gemcitabine led to reduced tumour growth and prolonged survival (Ludwig et al., 2018). Similar combination therapies are the basis of several clinical trials (pancreatic: phase II, NCT03649321; triple-negative breast cancer [TNBC]: phase II, NCT03184558). Another promising approach that is the basis of several clinical trials has been the neutralization of the EMT-promoting cytokine IL8, which impairs plasticity and promotes immune cell killing in TNBC (Dominguez et al., 2017) and sensitizes NSCLC cells to erlotinib and chemotherapy (Fernando et al., 2016). GlaxoSmithKline and Merck have produced a particularly interesting therapy—Bintrafusp alfa—involving the fusion of an anti-PDL1 antibody with the extracellular domain of the TGFBRII receptor, serving as a “trap” for free TGFB1. The compound is being tested in various clinical trials, including a phase II/III for biliary tract cancer (NCT04066491) and phase I for HER2+ breast cancer (NCT03620201). However, earlier this year, it was announced that the compound was unlikely to meet the primary endpoint (improved progression-free survival) in a phase III trial for NSCLC (NCT03631706).

Targeting gene regulatory machinery as a means of restricting plasticity is certainly less studied, but inhibitors of various classes of epigenetic regulators have been approved by the FDA (Bates, 2020). Their use in solid tumours is limited, but various preclinical evidence is promising. Inhibition of histone deacetylases (HDACs) has been shown by various groups to inhibit the EMT. Treatment of human pancreatic cancer cells with the class I HDAC inhibitor mocetinostat decreased mesenchymal gene expression and sensitized cells to gemcitabine both in vitro and in vivo (Meidhof et al., 2015). The generation of drug-resistant cell lines often leads to increased expression of mesenchymal genes, but treatment of resistant lines with 5-aza-2'-deoxycytidine to inhibit DNA methylation brings mesenchymal expression back to parental levels and sensitizes cells to various chemotherapies (Galle et al., 2020). Polycomb repressive complex 2 is an epigenetic regulator responsible for reducing gene expression of target loci through methylation of H3K27 and several studies have demonstrated that Inhibition of its methyltransferase EZH2 can impair the EMT (Perotti et al., 2019; Serresi et al., 2021; Zhao et al., 2019). Targeting gene regulatory machinery is an appealing strategy to prevent dynamics associated with plasticity, but given their general function, it is challenging to predict their effects on non-target cell types and tissues.

While many targeted therapies are being developed to restrict EMP, the general approach has been fairly crude. We have demonstrated that the signalling activity associated with EMP in tumours of a common cancer type are not necessarily consistent. EMP programs were consistently associated with TGFB and NFkB signalling, whereas some had notably high levels of either MAPK/EGFR or STAT/Hypoxia/p53 signalling. Leveraging transcriptomic data from various drug screens, we confirmed that EMP inhibition was directly related to the activity of each samples' specific signalling profile. Given this, STAT inhibition, for example, may give mixed results in a broad clinical trial of a specific cancer

type as it may only be effective in a subset of patients. Preclinical studies of various approaches have also often assessed EMP on the basis of several marker genes and the global effects on the mesenchymal gene expression program are unclear. For example, we have identified a novel dependency for the kinase RIPK1 in EMT responses from multiple inducers, however its inhibition only blocks the expression of about 50% TGFB1-induced genes in A549 cells. We observed similar partial responses for common targets, including TGFBRI and MEK. Furthermore, most trials have been based on either monotherapy of EMT-restricting compounds or simultaneous combination treatments. Rationally, for a combination therapy, it would be desirable for sensitization to precede the treatment of chemotherapy, for example. When considering the findings collectively, we believe there is a need for thorough preclinical studies performing deeper exploration of phenotypes after restricting EMP and understanding how these phenotypes evolve over time. Understanding how drugs and specific treatment regimens interact with these populations and implementing more personalized treatments could improve likelihood that this approach will yield positive results and improved patient outcome.

4.6 Phenotypic generalization as a framework for understanding EMP in cancer

4.6.1 A need to revise our conceptual model of EMP in cancer

Despite assessing the gene expression profiles of a dozen EMT responses and the intrinsic EMP programs of over 250 tumours and cell lines, and also determining the effects of signalling perturbations in over 100 conditions, it is still challenging to define the molecular basis of this plasticity. Perhaps the simplest conclusion is that, in fact, there is no unifying molecular basis because EMP is not a defined process like cell cycle or apoptosis. Fortunately, biological systems don't particularly care how we describe them and will

continue to behave as they do, but semantics are important to biologists. It doesn't serve the community well to say that "there is a great diversity of EMT phenotypic manifestations" (Yang et al., 2020), but then continually present EMP as a well-defined phenomenon, repeatedly listing its defining criteria in hundreds of review articles, and continually reconciling diversity as "partial EMTs". We should strive to modify our conceptual model of EMP, embrace its complexity, and adopt quantitative strategies to understand it. Here, I propose a framework based on Pareto optimality that suggests that the EMT in cancer broadly represents a transition from a *specialist* epithelial phenotype to a *generalist* phenotype as an adaptation to environmental heterogeneity.

4.6.2 Pareto optimality and the geometry of cellular phenotypes

Pareto optimality describes the trade-off behaviour of systems that must optimize the performance of multiple objectives (Shoval et al., 2012). Natural selection of a system with single fitness function evolves towards a single, optimal point in trait space. However, the fitness of many evolving systems (eg. species) depends on the performance of multiple tasks. An optimal phenotype for a given task is often not the optimal phenotype for another, resulting in a fitness trade-off (**Figure 38A**). Pareto optimality refers to the phenotypic configurations of a system where no individual fitness function could be improved without negatively affecting another. The Pareto front is the space that spans these configurations where optimality is achieved. For example, consider a system with two tasks: in trait space, each task would have a single point where the traits provide optimal fitness (referred to as archetypes). Changing traits away from these points will reduce fitness of that task. The Pareto front of this system would be a one-dimensional line in trait space connecting the archetypes for each task, representing the optimal trade off in performance. Any point not on this line could be improved at both tasks simultaneously by moving towards the front. In a system with three tasks, the Pareto front will be a triangular plane connecting the

archetypes, and with four tasks, the front will be a tetrahedron. Trait configurations near archetypes are specialist phenotypes at a given task, whereas those between archetypes are phenotypic generalists (**Figure 38A**).

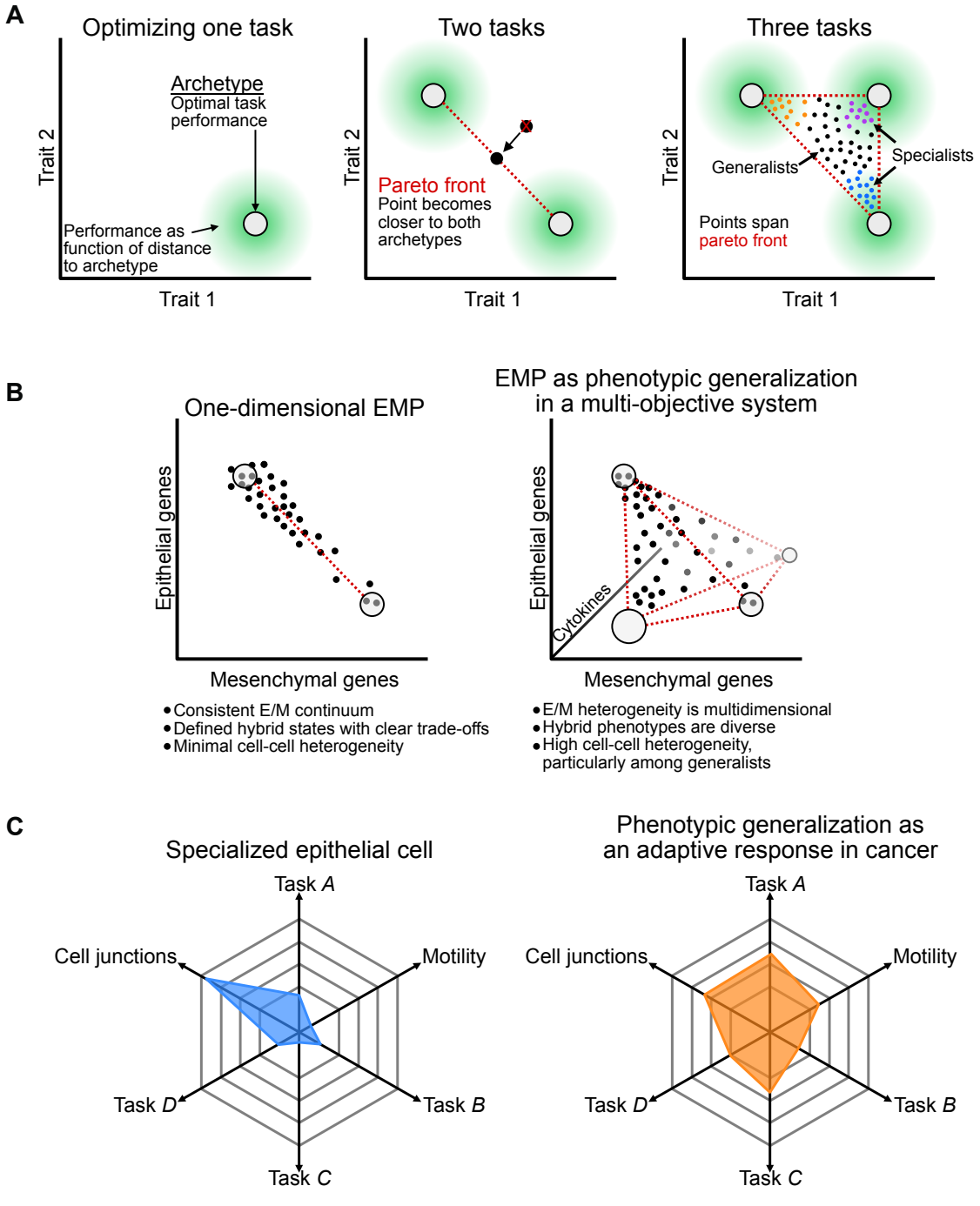

Figure 38. A framework based on Pareto optimality for understanding EMP in cancer. A) Pareto optimality suggests that systems evolve to minimize fitness trade-offs when optimizing for multiple objectives. Optimal states are defined by the Pareto front, which represents evolutionary constraints on the traits of a system. **B)** Contrasting a one-dimensional view of EMP with a multi-objective model where cells must optimize fitness at multiple tasks. In the latter model, EMP can emerge simply as the result of becoming less specialized at epithelial tasks, which can occur in response to heterogeneous environmental pressures. **C)** Comparing task performance of epithelial specialists that face little environmental pressure in homeostatic conditions to those of a malignant cell existing in a dynamic, heterogeneous environment.

Systems under evolutionary pressure tend to evolve towards the Pareto front. These patterns can be seen in the phenotypic diversity of various species, including Darwin's ground finches, leaf-cutter ants, and bats. For example, the distribution of head width and poison sac length measurements from leaf-cutter ants span a triangle whose three vertices represent archetypical gardening/nursing, foraging, and soldiering ants (Shoval et al., 2012). This principle is not limited to anatomical traits of an organism, but applies to any measurements that can represent the traits of a system. In fact, a cell's global gene expression profile provides a unique representation of its collection of traits. A transcriptome does not necessarily reveal how its translated proteins specifically contribute to traits, and it is certainly not a complete representation of the cell from a molecular perspective. However, it could be reasonably argued that any unique cellular state will also have a unique gene expression profile, even though this profile will not contain the entirety of relevant information about the state. Therefore, gene expression profiles of individual cells vary with differences in cellular state and it could be expected that gene regulatory systems have evolved to the geometric constraints of Pareto optimality.

Pareto optimality has been used as a framework to interpret bulk RNA-seq data (Hausser et al., 2019), but its use to understand the heterogeneity of individual cells has been limited. Adler et al. (2019) demonstrated that gene expression heterogeneity of both hepatocytes

and intestinal epithelial cells span polyhedrons whose vertices correspond to archetypal traits of these cells at specific locations in the tissue. For example, hepatocyte gene expression occupies a tetrahedron whose four archetypes each have maximal expression of genes involved in different types of metabolism. Heterogeneity within the tetrahedron corresponds to defined zonation patterns within liver lobules (Adler et al., 2019).

In cancer, diverse tasks are accomplished to promote tumour progression, from sustained proliferation, modified energetics, immune evasion, and tissue remodelling (Hanahan and Weinberg, 2000, 2011). It is unclear, however, to what extent these cancer hallmarks can be attributed to cancer cell-intrinsic properties and how different malignant cell states contribute to each. Advances in single-cell genomics have sparked interest in identifying cancer cell states (Kinker et al., 2020; Tirosh and Suvà, 2018), but most exploration has still viewed these states as discrete entities through partitioning malignant cells into sub-clusters. Trajectory inference methods have been used to assess phenotypic gradients, but these methods are built under the assumption that relevant heterogeneity can be reduced to a one-dimensional trajectory. If cancer cells must optimize for multiple tasks, however, their phenotypes would likely vary along multiple dimensions independently. For example, Baron et al. (2020) demonstrated that the gene expression profiles of cancer cells from a zebrafish model of melanoma span a triangular pattern in principal component space with vertices corresponding to mature melanocyte, neural crest, and stress-like states. Little is still known about continuous geometric properties of cancer cell states.

4.6.3 EMT in cancer is consistent with phenotypic generalization

Under homeostatic conditions, many cell types could be described as specialists. Organs comprise distinct cell types, each performing limited roles that ultimately contribute to organ function. Multicellular organisms have evolved in favour of a division of labour among

specialist cell types rather than a population of dynamic generalists. Tightly controlled developmental processes lead to the terminal differentiation that forms these specialist states, constraining gene expression programs through progressive epigenetic silencing (Atlasi and Stunnenberg, 2017) and establish cellular niches that support these states (Tadanori Mammoto, 2010). The primary task of a simple epithelial cell is barrier formation and these cells express high levels of junctional proteins to accomplish this (Giepmans and van Ijzendoorn, 2009). In certain tissues, however, specific subsets epithelial cells perform secretory roles and, consistent with the Pareto optimality framework, secretory cells do not retain comparable cell junctions. For example, desmosome expression is lost when acinar cells of the mammary epithelium become secretory, consistent with a phenotypic trade-off (Pitelka et al., 1973).

Epithelial cells have specialized phenotypes under homeostatic conditions, but these states are not necessarily fixed. Over the last several years, there has been increased recognition that many epithelial tissues exhibit remarkable plasticity in response to injury, resulting in interconversion of distinct epithelial phenotypes within the tissue. This has been observed in the airway (Purushothama Rao Tata, 2017), skin (Ito et al., 2007), pancreatic (Sapna Puri, 2010), stomach (Stange et al., 2013), and intestinal epithelium (Ayyaz et al., 2019). Tissue injury ultimately disrupts regulatory signals within the cells' microenvironment through the physical and chemical changes associated with inflammation, including accumulations of the various growth factors (eg. TGFB1, PDGF), cytokines (eg. interleukins, TNF), and chemokines (Chen et al., 2018b). Tumourigenesis similarly disrupts the cells' microenvironment. Tumours are structurally dynamic throughout their progression, affecting cell density and composition throughout the tissue. Many are also characterized by chronic inflammation, leading to continuous exposure to inflammatory factors (Coussens and Werb,

2002; Hanahan and Weinberg, 2011). As such, tumours have been referred to as “wounds that do not heal” (Dvorak, 1986).

We propose that in response to these environmental disruptions, epithelial cells become less specialized, adopting a generalist phenotype. Mechanistically, the convergence of the many environmental signals may override the phenotypic constraints established during terminal differentiation. Functionally, a generalist phenotype may offer a distinct survival advantage to the cells in these harsh environments by performing adequately at multiple tasks (a phenotypic jack-of-all-trades). Its co-activation of multiple traits may provide adaptability to further changes in the environment, allowing the cell to rapidly prioritize specific programs in response to selective pressures. This property is the relationship between generalists and the concept of plasticity: “generalist” describes the trade-off behaviour of sampling distinct functionalities, which enables the plasticity for these cells to respond to change.

A consequence of generalization is that it would involve reduced activity of epithelial traits in favour of activating alternative programs, presumably including canonical mesenchymal properties. Given this, we believe that EMT patterns observed in cancer are most often a reflection of generalization rather than a purposeful activation of mesenchymal traits (**Figure 38 B,C**). Because generalization involves the activation of multiple traits, this may also explain the diverse properties that have been attributed to mesenchymal cells in cancer, including increased drug resistance, tumour initiating properties, and release of angiogenic factors. These properties may represent functionality independent from canonical mesenchymal properties (ie. loss of polarity, increased motility). This is consistent with evidence that the EMT in cancer may be mechanistically independent from the tumour-initiating properties that are often associated with mesenchymal cells(Nieto et al., 2016).

4.6.4 Generalist phenotypes can be quantitatively defined

An appealing property of this generalist framework for EMP is that it has quantitative properties at cell and population levels that can be evaluated without dependence on the choice of select marker genes. Generalization involves the balancing of different cellular functions. Given this, individual cells could be expected to express multiple distinct expression programs at some level. If these functions are independent from one another, their regulation may also be independent. Several reasonable assumptions can be made based on this property. First, an individual generalist would likely express a larger number of unique genes than a specialist of any of the relevant functions. Interestingly, unique gene counts in individual cells have been used as a proxy for a cell's developmental potential and is the basis of the related CytoTRACE tool (Gulati et al., 2020). In our cohort of 160 tumour samples, we find that activity of a tumour's specific EMP program frequently correlates with the total number of expressed genes (**Figure 39a**). This may also be associated with several mechanistic features of individual cells. For example, a larger number of expressed genes would likely be associated with a higher proportion of accessible chromatin across the genome. Across a population of generalists, each cell may exhibit different trade-offs for each task. If these functions are independently regulated, cell-cell variation would exist along more dimensions than specialists of individual tasks. Due to this increased dimensionality, cell-cell heterogeneity would also be higher for generalists. We used ROGUE — an entropy-based statistic to assess population purity — and confirmed that expression patterns in cells activating intratumoural EMP programs were frequently associated with reduced entropy, consistent with increased heterogeneity (**Figure 39b**).

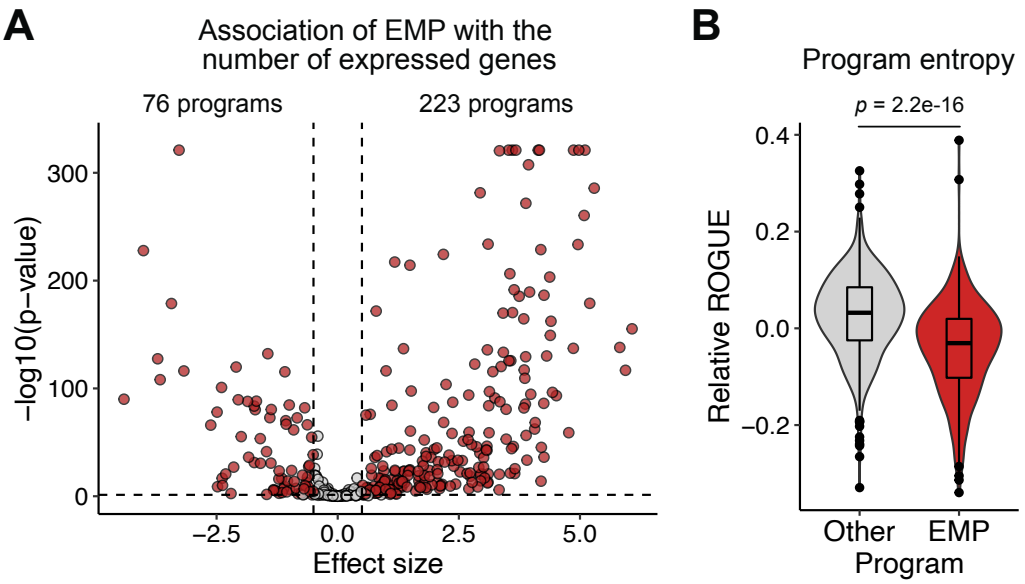

Figure 39. Features of EMP are consistent with phenotypic generalization. A) Plot showing the association of EMP activity scores with the total number of detectable genes in individual cells. The relationship was modelled with a linear model and the effect size corresponds to the model's beta coefficient. **B)** Distribution of ROGUE entropy values associated with different cell states. Relative scores were calculated by centering ROGUE scores within each individual sample to avoid confounding issues with differences in score distributions between samples.

4. 6.5 The specialist-generalist framework enables mechanism-independent prediction of cellular behaviours

Much of molecular biology research over the last several decades has focused on defining biochemical mechanisms, and while these mechanistic models can reveal therapeutic targets, they can also be overly simplistic and fail to cohesively represent cell behaviours. The specialist-generalist framework is a higher-level abstraction that fails to reveal biochemical underpinnings, but can provide valuable insight on cellular behaviour that cannot be easily gleaned from mechanistic models. For example, if EMP is associated with generalization, the model itself predicts that generalists are highly adaptable and therefore the resistance to both chemo- and targeted therapies is expected. Furthermore, the modular regulation of independent tasks performed by generalists would make it unlikely to identify individual treatments that simultaneously target these distinct functions.

The plasticity of generalists poses a challenge for cancer therapeutics. Historically, drug development has focused on targeting cancer dependencies to elicit lethal responses (Tsherniak et al., 2017). These treatments are often thwarted by adaptive responses (Boumahdi and de Sauvage, 2019). Ideally, treatments could be designed to first restrict adaptation through phenotypic specialization prior to targeting these dependencies. Generalists evolve in response to contrasting selection pressures associated with environmental heterogeneity (Bradshaw, 1965; Van Buskirk, 2002), but given the dynamic nature of tumours, it is unclear if it is feasible to stabilize the TME sufficiently to favour specialization. However, perhaps specific cell therapies could be engineered to accomplish this through release of specific cytokines or growth factors. Alternatively, there may exist relatively unintuitive strategies to steer cell fate towards specialization and reduce plasticity, such as modifying metabolite or nutrient availability within the TME (Baksh et al., 2020). As discussed in the previous section, restricting plasticity through specialization could be a promising approach to sensitize tumours to targeted therapies.

Together, we believe that this model provides a new perspective on EMP in cancer that is consistent with previous observations and reconciles both molecular and functional heterogeneity. Importantly, this conceptual model highlights that the primary cause for EMP in cancer is the environmental heterogeneity associated with the TME, explaining the ubiquity of EMP in solid tumours. It also suggests that the activation of mesenchymal-associated genes may simply be a general adaptive response and is not necessarily a purposeful activation of motility and invasive properties. Importantly, the diverse functionality that has been attributed to mesenchymal cells in cancer may occur from uncoupled regulatory mechanisms as part of the adaptive response. Adopting this conceptual model

provides a systems-level perspective of plasticity in cancer, enabling predictions about cell behaviour free from the focus on biochemical mechanisms.

Chapter 5 - Conclusion

Epithelial-mesenchymal plasticity contributes to tumour progression, promoting metastasis, therapy resistance, and immunosuppression. Despite being one of the most studied cellular processes, surprisingly little has been known about its defining molecular features. Rather, perhaps due to historical reasons, intense focus has been given to a small selection of features. An increasing number of studies reporting EMTs with conflicting molecular properties have challenged our understanding of this phenomenon and led to skepticism and uncertainty about its molecular basis (Yang et al., 2020).

We have attempted to reconcile these issues by performing comparative studies that assess EMP in a variety of conditions. Using single-cell genomics to track transcriptional dynamics throughout the EMT, we found striking variation between contexts. Individual populations could undergo multiple EMT responses and different cell types induced with the same factor shared little in common. This variability is not simply an experimental artifact. Using approaches to decompose gene expression profiles into heterogeneously expressed programs, we defined EMP-associated programs from 160 tumours spanning multiple cancer types. While a signature of 289 genes were frequently associated with these programs, no gene was a perfect indicator. Despite this diversity, we found unifying regulatory features associated with EMP programs. All were associated with higher activity of TGFB1 and NFkB signalling, and subsets of programs could be distinguished by higher activity of either STAT and hypoxia signalling or MAPK/ERK signalling. Pharmacological inhibition of these pathways led to reduced expression of EMP-associated genes and by performing a kinase inhibitor screen, we identified a novel role for the kinase RIPK1 in EMP.

We have performed the first large scale studies to define the molecular basis of EMP in cancer across contexts. While it may be unfortunate that unifying components of the cells' gene expression program are non-existent, we believe that this is an important feature to

illuminate. This work suggests that EMP in cancer may not be a well-defined process at all. Rather, we propose EMP may represent an adaptive response to environmental disruption within the TME, promoting phenotypic generalization. Diverse phenotypes emerge due to phenotypic trade-offs resulting from the distinct features of a given TME and the cancer cells responding to them.

Moving forward, we believe studies of more abstract, systems-level features of EMP will allow us to understand this phenomenon better, moving beyond a definition that leans on a core set of molecular markers. Understanding the distribution of phenotypes in tumours and how they change in response to chemical, environmental, and genetic perturbations will be invaluable. As we improve our conceptual model of EMP — or generalists and plasticity more broadly — we can hopefully develop rational strategies to reprogram cancer cell fate to improve treatment efficacy.

References

Adler, M., Kohanim, Y.K., Tendler, A., Mayo, A., and Alon, U. (2019). Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. Cell Systems *8*, 43–52.e5.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Aiello, N.M., Brabletz, T., Kang, Y., Nieto, M.A., Weinberg, R.A., and Stanger, B.Z. (2017). Upholding a role for EMT in pancreatic cancer metastasis. Nature *547*, E7–E8.

Aiello, N.M., Maddipati, R., Norgard, R.J., Balli, D., Li, J., Yuan, S., Yamazoe, T., Black, T., Sahmoud, A., Furth, E.E., et al. (2018). EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. Dev. Cell *45*, 681–695.

Akalay, I., Janji, B., Hasmim, M., Noman, M.Z., Thiery, J.P., Mami-Chouaib, F., and Chouaib, S. (2013). EMT impairs breast carcinoma cell susceptibility to CTL-mediated lysis through autophagy induction. Autophagy *9*, 1104–1106.

Alcorn, J.F., Guala, A.S., van der Velden, J., McElhinney, B., Irvin, C.G., Davis, R.J., and Janssen-Heininger, Y.M.W. (2008). Jun N-terminal kinase 1 regulates epithelial-to-mesenchymal transition induced by TGF-β1. J. Cell Sci. *121*, 1036–1045.

An, Q., Liu, T., Wang, M.-Y., Yang, Y.-J., Zhang, Z.-D., Liu, Z.-J., and Yang, B. (2021). KRT7 promotes epithelial-mesenchymal transition in ovarian cancer via the TGF-β/Smad2/3 signaling pathway. Oncol. Rep. *45*, 481.

Andreucci, E., Peppicelli, S., Ruzzolini, J., Bianchini, F., Biagioni, A., Papucci, L., Magnelli, L., Mazzanti, B., Stecca, B., and Calorini, L. (2020). The acidic tumor microenvironment drives a stem-like phenotype in melanoma cells. J. Mol. Med. *98*, 1431–1446.

Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity. Nat. Commun. *6*, 8971.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Michael Cherry, J., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Atlasi, Y., and Stunnenberg, H.G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. Nat. Rev. Genet. *18*, 643–658.

Ayyaz, A., Kumar, S., Sangiorgi, B., Ghoshal, B., Gosio, J., Ouladan, S., Fink, M., Barutcu, S., Trcka, D., Shen, J., et al. (2019). Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. Nature *569*, 121–125.

Bakiri, L., Macho-Maschler, S., Custic, I., Niemiec, J., Guío-Carrión, A., Hasenfuss, S.C., Eger, A., Müller, M., Beug, H., and Wagner, E.F. (2015). Fra-1/AP-1 induces EMT in mammary epithelial cells by modulating Zeb1/2 and TGFβ expression. Cell Death Differ. *22*,

336–350.

Baksh, S.C., Todorova, P.K., Gur-Cohen, S., Hurwitz, B., Ge, Y., Novak, J.S.S., Tierney, M.T., Cruz-Racelis, J.D., Fuchs, E., and Finley, L.W.S. (2020). Extracellular serine controls epidermal stem cell fate and tumour initiation. Nat. Cell Biol. *22*, 779–790.

Baron, M., Tagore, M., Hunter, M.V., Kim, I.S., Moncada, R., Yan, Y., Campbell, N.R., White, R.M., and Yanai, I. (2020). The Stress-Like Cancer Cell State Is a Consistent Component of Tumorigenesis. Cell Systems *11*, 536–546.

Bates, S.E. (2020). Epigenetic Therapies for Cancer. N. Engl. J. Med. *383*, 650–663.

Batlle, E., Sancho, E., Francí, C., Domínguez, D., Monfar, M., Baulida, J., and de Herreros, A.G. (2000). The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells. Nat. Cell Biol. *2*, 84–89.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) *57*, 289–300.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bolós, V., Peinado, H., Pérez-Moreno, M.A., Fraga, M.F., Esteller, M., and Cano, A. (2003). The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with Snail and E47 repressors. J. Cell Sci. *116*, 499–511.

Boumahdi, S., and de Sauvage, F.J. (2019). The great escape: tumour cell plasticity in resistance to targeted therapy. Nat. Rev. Drug Discov. *19*, 39–56.

Bradshaw, A.D. (1965). Evolutionary Significance of Phenotypic Plasticity in Plants. In Advances in Genetics, (Academic Press), pp. 115–155.

Broders-Bondon, F., Nguyen Ho-Bouldoires, T.H., Fernandez-Sanchez, M.-E., and Farge, E. (2018). Mechanotransduction in tumor progression: The dark side of the force. J. Cell Biol. *217*, 1571–1587.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat. Methods *17*, 159–162.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–9.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Byers, L.A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., Shen, L., Fan, Y., Giri, U., Tumula, P.K., et al. (2013). An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin. Cancer Res. *19*, 279–290.

Cano, A., Pérez-Moreno, M.A., Rodrigo, I., Locascio, A., Blanco, M.J., del Barrio, M.G., Portillo, F., and Nieto, M.A. (2000). The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nat. Cell Biol. *2*, 76–83.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721.

Chen, G.M., Kannan, L., Geistlinger, L., Kofia, V., Safikhani, Z., Gendoo, D.M.A., Parmigiani, G., Birrer, M., Haibe-Kains, B., and Waldron, L. (2018a). Consensus on Molecular Subtypes of High-grade Serous Ovarian Carcinoma. Clin. Cancer Res. *24*, 5037–5047.

Chen, L., Deng, H., Cui, H., Fang, J., Zuo, Z., Deng, J., Li, Y., Wang, X., and Zhao, L. (2018b). Inflammatory responses and inflammation-associated diseases in organs. Oncotarget *9*, 7204–7218.

Chen, W.S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., and Krishnaswamy, S. (2020). Uncovering axes of variation among single-cell cancer specimens. Nature Methods *17*, 302–310.

Cieślik, M., Hoang, S.A., Baranova, N., Chodaparambil, S., Kumar, M., Allison, D.F., Xu, X., Wamsley, J.J., Gray, L., Jones, D.R., et al. (2013). Epigenetic coordination of signaling pathways during the epithelial-mesenchymal transition. Epigenetics Chromatin *6*, 28.

Cochrane, D.R., Spoelstra, N.S., Howe, E.N., Nordeen, S.K., and Richer, J.K. (2009). MicroRNA-200c mitigates invasiveness and restores sensitivity to microtubule-targeting chemotherapeutic agents. Mol. Cancer Ther. *8*, 1055–1066.

Cook, D.P., and Vanderhyden, B.C. (2020). Context specificity of the EMT transcriptional response. Nat. Commun. *11*, 2142.

Coussens, L.M., and Werb, Z. (2002). Inflammation and cancer. Nature *420*, 860–867.

Dalmau, N., Jaumot, J., Tauler, R., and Bedia, C. (2015). Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells. Molecular BioSystems *11*, 3397–3406.

D'Angelo, R.C., Liu, X.W., Najy, A.J., Jung, Y.S., Won, J., Chai, K.X., Fridman, R., and Kim, H.R. (2014). TIMP-1 via TWIST1 induces EMT phenotypes in human breast epithelial cells. Mol. Cancer Res. *12*, 1324–1333.

Dehai, C., Bo, P., Qiang, T., Lihua, S., Fang, L., Shi, J., Jingyan, C., Yan, Y., Guangbin, W., and Zhenjun, Y. (2014). Enhanced invasion of lung adenocarcinoma cells after co-culture with THP-1-derived macrophages via the induction of EMT by IL-6. Immunol. Lett. *160*, 1–10.

Derynck, R., and Weinberg, R.A. (2019). EMT and Cancer: More Than Meets the Eye. Dev. Cell *49*, 313–316.

Devaraj, V., and Bose, B. (2019). Morphological State Transition Dynamics in EGF-Induced Epithelial to Mesenchymal Transition. J. Clin. Med. Res. *8*, 911.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychndhury, R., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. Cell *167*, 1853.

Dominguez, C., McCampbell, K.K., David, J.M., and Palena, C. (2017). Neutralization of IL-8 decreases tumor PMN-MDSCs and reduces mesenchymalization of claudin-low triple-negative breast cancer. JCI Insight *2*, e94296.

Dong, R., Wang, Q., He, X.L., Chu, Y.K., Lu, J.G., and Ma, Q.J. (2007). Role of nuclear factor kappa B and reactive oxygen species in the tumor necrosis factor-alpha-induced epithelial-mesenchymal transition of MCF-7 cells. Braz. J. Med. Biol. Res. *40*, 1071–1078.

Dongre, A., and Weinberg, R.A. (2019). New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. Nature Reviews Molecular Cell Biology *20*, 69–84.

Dongre, A., Rashidian, M., Reinhardt, F., Bagnato, A., Keckesova, Z., Ploegh, H.L., and Weinberg, R.A. (2017). Epithelial-to-mesenchymal Transition contributes to Immunosuppression in Breast Carcinomas. Cancer Res. *77*, 3982–3989.

Dongre, A., Rashidian, M., Eaton, E.N., Reinhardt, F., Thiru, P., Zagorulya, M., Nepal, S., Banaz, T., Martner, A., Spranger, S., et al. (2020). Direct and Indirect Regulators of Epithelial-Mesenchymal Transition (EMT)-mediated Immunosuppression in Breast Carcinomas. Cancer Discov. *11*, 1-20

Durante, M.A., Rodriguez, D.A., Kurtenbach, S., Kuznetsov, J.N., Sanchez, M.I., Decatur, C.L., Snyder, H., Feun, L.G., Livingstone, A.S., and Harbour, J.W. (2020). Single-cell analysis reveals new evolutionary complexity in uveal melanoma. Nat. Commun. *11*, 496.

Dvorak, H.F. (1986). Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. N. Engl. J. Med. *315*, 1650–1659.

Eger, A., Aigner, K., Sonderegger, S., Dampier, B., Oehler, S., Schreiber, M., Berx, G., Cano, A., Beug, H., and Foisner, R. (2005). DeltaEF1 is a transcriptional repressor of E-cadherin and regulates epithelial plasticity in breast cancer cells. Oncogene *24*, 2375–2385.

Fernando, R.I., Hamilton, D.H., Dominguez, C., David, J.M., McCampbell, K.K., and Palena, C. (2016). IL-8 signaling is involved in resistance of lung carcinoma cells to erlotinib. Oncotarget *7*, 42031–42044.

Fischer, K.R., Durrans, A., Lee, S., Sheng, J., Li, F., Wong, S.T.C., Choi, H., El Rayes, T., Ryu, S., Troeger, J., et al. (2015). Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. Nature *527*, 472–476.

Frisch, S.M. (1997). The epithelial cell default-phenotype hypothesis and its implications for cancer. Bioessays *19*, 705–709.

Gabbert, H., Wagner, R., Moll, R., and Gerharz, C.D. (1985). Tumor dedifferentiation: an important step in tumor invasion. Clin. Exp. Metastasis *3*, 257–279.

Galle, E., Thienpont, B., Cappuyns, S., Venken, T., Busschaert, P., Van Haele, M., Van Cutsem, E., Roskams, T., van Pelt, J., Verslype, C., et al. (2020). DNA methylation-driven EMT is a common mechanism of resistance to various therapeutic agents in cancer. Clin. Epigenetics *12*, 1–19.

Geistlinger, L., Oh, S., Ramos, M., Schiffer, L., LaRue, R.S., Henzler, C.M., Munro, S.A., Daughters, C., Nelson, A.C., Winterhoff, B.J., et al. (2020). Multiomic Analysis of Subtype Evolution and Heterogeneity in High-Grade Serous Ovarian Carcinoma. Cancer Res. *80*, 4335–4345.

Gemmill, R.M., Roche, J., Potiron, V.A., Nasarrea, P., Mitas, M., Coldren, C.D., Helfrich, B.A., Garrett-Mayer, E., Bunn, P.A., and Drabkin, H.A. (2011). ZEB1-responsive genes in non-small cell lung cancer. Cancer Lett. *300*, 66–78.

Giepmans, B.N.G., and van Ijzendoorn, S.C.D. (2009). Epithelial cell–cell junctions and plasma membrane domains. Biochimica et Biophysica Acta (BBA) - Biomembranes *1788*, 820–831.

Goldman, R.D., Khuon, S., Chou, Y.H., Opal, P., and Steinert, P.M. (1996). The function of intermediate filaments in cell shape and cytoskeletal integrity. J. Cell Biol. *134*, 971–983.

Gondkar, K., Patel, K., Krishnappa, S., Patil, A., Nair, B., Sundaram, G.M., Zea, T.T., and Kumar, P. (2019). E74 like ETS transcription factor 3 (ELF3) is a negative regulator of epithelial- mesenchymal transition in bladder carcinoma. Cancer Biomark. *25*, 223–232.

Gonzalez, D.M., and Medici, D. (2014). Signaling mechanisms of the epithelial-mesenchymal transition. Sci. Signal. *7*, re8.

Gooding, A.J., and Schiemann, W.P. (2016). Harnessing protein kinase A activation to induce mesenchymal-epithelial programs to eliminate chemoresistant, tumor-initiating breast cancer cells. Transl. Cancer Res. *5*, S226–S232.

Gregory, P.A., Bracken, C.P., Smith, E., Bert, A.G., Wright, J.A., Roslan, S., Morris, M., Wyatt, L., Farshid, G., Lim, Y.-Y., et al. (2011). An autocrine TGF-beta/ZEB/miR-200 signaling network regulates establishment and maintenance of epithelial-mesenchymal transition. Mol. Biol. Cell *22*, 1686–1698.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The Consensus Molecular Subtypes of Colorectal Cancer. Nat. Med. *21*, 1350.

Gulati, G.S., Sikandar, S.S., Wesche, D.J., Manjunath, A., Bharadwaj, A., Berger, M.J., Ilagan, F., Kuo, A.H., Hsieh, R.W., Cai, S., et al. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. Science *367*, 405–411.

Gurrapu, S., and Tamagnone, L. (2019). Semaphorins as Regulators of Phenotypic Plasticity and Functional Reprogramming of Cancer Cells. Trends Mol. Med. *25*, 303–314.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Hajra, K.M., Chen, D.Y.-S., and Fearon, E.R. (2002). The SLUG zinc-finger protein represses E-cadherin in breast cancer. Cancer Res. *62*, 1613–1618.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zagar, M., et al. (2020). Integrated analysis of multimodal single-cell data.

Hausser, J., Szekely, P., Bar, N., Zimmer, A., Sheftel, H., Caldas, C., and Alon, U. (2019). Tumor diversity and the trade-off between universal cancer tasks. Nat. Commun. *10*, 1–13.

Hay, E.D. (1995). An overview of epithelio-mesenchymal transformation. Acta Anat. *154*, 8–20.

Haynes, J., Srivastava, J., Madson, N., Wittmann, T., and Barber, D.L. (2011). Dynamic actin remodeling during epithelial-mesenchymal transition depends on increased moesin expression. Mol. Biol. Cell *22*, 4750–4764.

Heise, R.L., Stober, V., Cheluvaraju, C., Hollingsworth, J.W., and Garantziotis, S. (2011). Mechanical Stretch Induces Epithelial-Mesenchymal Transition in Alveolar Epithelia via Hyaluronan Activation of Innate Immunity. J. Biol. Chem. *286*, 17435–17444.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291–304.e6.

Holian, J., Qi, W., Kelly, D.J., Zhang, Y., Mreich, E., Pollock, C.A., and Chen, X.-M. (2008). Role of Kruppel-like factor 6 in transforming growth factor-beta1-induced epithelial-mesenchymal transition of proximal tubule cells. Am. J. Physiol. Renal Physiol. *295*, F1388–F1396.

Holland, C.H., Szalai, B., and Saez-Rodriguez, J. (2020a). Transfer of regulatory knowledge from human to mouse for functional genomics analysis. Biochim. Biophys. Acta Gene Regul. Mech. *1863*, 194431.

Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., et al. (2020b). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. *21*, 36.

Huber, M.A., Azoitei, N., Baumann, B., Grünert, S., Sommer, A., Pehamberger, H., Kraut, N., Beug, H., and Wirth, T. (2004). NF-κB is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression. Journal of Clinical Investigation *114*, 569–581.

Huergo-Zapico, L., Parodi, M., Cantoni, C., Lavarello, C., Fernández-Martínez, J.L., Petretto, A., DeAndrés-Galiana, E.J., Balsamo, M., López-Soto, A., Pietra, G., et al. (2018). NK-cell Editing Mediates Epithelial-to-Mesenchymal Transition via Phenotypic and Proteomic Changes in Melanoma Cell Lines. Cancer Res. *78*, 3913–3925.

Ibrahim, M.M., and Kramann, R. (2019). genesorteR: Feature Ranking in Clustered Single Cell Data. bioRxiv https://doi.org/10.1101/676379.

Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., et al. (2015). Stromal contribution to the colorectal cancer transcriptome. Nat. Genet. *47*, 312–319.

Ito, M., Yang, Z., Andl, T., Cui, C., Kim, N., Millar, S.E., and Cotsarelis, G. (2007). Wnt-dependent de novo hair follicle regeneration in adult mouse skin after wounding. Nature *447*, 316–320.

Iwano, M., Plieth, D., Danoff, T.M., Xue, C., Okada, H., and Neilson, E.G. (2002). Evidence that fibroblasts derive from epithelium during tissue fibrosis. J. Clin. Invest. *110*, 341–350.

Izar, B., Tirosh, I., Stover, E.H., Wakiro, I., Cuoco, M.S., Alter, I., Rodman, C., Leeson, R., Su, M.-J., Shah, P., et al. (2020). A single-cell landscape of high-grade serous ovarian cancer. Nat. Med. *26*, 1271–1279.

Jacob, A., and Prekeris, R. (2015). The regulation of MMP targeting to invadopodia during cancer metastasis. Front Cell Dev Biol *3*, 4.

Jeon, H.-M., and Lee, J. (2017). MET: roles in epithelial-mesenchymal transition and cancer stemness. Annals of Translational Medicine *5*, 5–5.

Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George, B.M., Mollbrink, A., Bergenstråhle, J., et al. (2020). Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell *182*, 1661–1662.

Jin, W. (2020). Role of JAK/STAT3 Signaling in the Regulation of Metastasis, the Transition of Cancer Stem Cells, and Chemoresistance of Cancer by Epithelial-Mesenchymal Transition. Cells *9*, 217.

Jolly, M.K., Boareto, M., Huang, B., Jia, D., Lu, M., Ben-Jacob, E., Onuchic, J.N., and Levine, H. (2015). Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. Front. Oncol. *5*, 155.

Kajita, M., McClinic, K.N., and Wade, P.A. (2004). Aberrant expression of the transcription factors snail and slug alters the response to genotoxic stress. Mol. Cell. Biol. *24*, 7559–7566.

Kajiyama, H., Shibata, K., Terauchi, M., Yamashita, M., Ino, K., Nawa, A., and Kikkawa, F. (2007). Chemoresistance to paclitaxel induces epithelial-mesenchymal transition and enhances metastatic potential for epithelial ovarian carcinoma cells. Int. J. Oncol. *31*, 277–283.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. J. Clin. Invest. *119*, 1420–1428.

Kasai, H., Allen, J.T., Mason, R.M., Kamimura, T., and Zhang, Z. (2005). TGF-β1 induces human alveolar epithelial to mesenchymal cell transition (EMT). Respir. Res. *6*, 56.

Kim, N., Kim, H.K., Lee, K., Hong, Y., Cho, J.H., Choi, J.W., Lee, J.-I., Suh, Y.-L., Ku, B.M., Eum, H.H., et al. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat. Commun. *11*, 2285.

Kinker, G.S., Greenwald, A.C., Tal, R., Orlova, Z., Cuoco, M.S., McFarland, J.M., Warren, A., Rodman, C., Roth, J.A., Bender, S.A., et al. (2020). Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. Nat. Genet. *52*, 1208–1218.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Klumpe, H., Langley, M.A., Linton, J.M., Su, C.J., Antebi, Y.E., and Elowitz, M.B. (2020). The context-dependent, combinatorial logic of BMP signaling. bioRxiv doi: 2020.12.08.416503.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. bioRxiv doi: 10.1101/060012.

Krebs, A.M., Mitschke, J., Lasierra Losada, M., Schmalhofer, O., Boerries, M., Busch, H., Boettcher, M., Mougiakakos, D., Reichardt, W., Bronsert, P., et al. (2017). The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. Nat. Cell Biol. *19*, 518–529.

Kröger, C., Afeyan, A., Mraz, J., Eaton, E.N., Reinhardt, F., Khodor, Y.L., Thiru, P., Bierie, B., Ye, X., Burge, C.B., et al. (2019). Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. Proc. Natl. Acad. Sci. U. S. A. *116*, 7353–7362.

Kudo-Saito, C., Shirako, H., Takeuchi, T., and Kawakami, Y. (2009). Cancer metastasis is accelerated through immunosuppression during Snail-induced EMT of cancer cells. Cancer Cell *15*, 195–206.

Kurrey, N.K., Jalgaonkar, S.P., Joglekar, A.V., Ghanate, A.D., Chaskar, P.D., Doiphode, R.Y., and Bapat, S.A. (2009). Snail and slug mediate radioresistance and chemoresistance by antagonizing p53-mediated apoptosis and acquiring a stem-like phenotype in ovarian cancer cells. Stem Cells *27*, 2059–2068.

Kuwada, K., Kagawa, S., Yoshida, R., Sakamoto, S., Ito, A., Watanabe, M., Ieda, T., Kuroda, S., Kikuchi, S., Tazawa, H., et al. (2018). The epithelial-to-mesenchymal transition induced by tumor-associated macrophages confers chemoresistance in peritoneally disseminated pancreatic cancer. J. Exp. Clin. Cancer Res. *37*, 307.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 1–35.

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., Van den Eynde, K., et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med. *24*, 1277–1289.

Lamouille, S., Xu, J., and Derynck, R. (2014). Molecular mechanisms of epithelial–mesenchymal transition. Nat. Rev. Mol. Cell Biol. *15*, 178–196.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Larocca, C., Cohen, J.R., Fernando, R.I., Huang, B., Hamilton, D.H., and Palena, C. (2013).

An autocrine loop between TGF-β1 and the transcription factor brachyury controls the transition of human carcinoma cells into a mesenchymal phenotype. Mol. Cancer Ther. *12*, 1805–1815.

Latil, M., Nassar, D., Beck, B., Boumahdi, S., Wang, L., Brisebarre, A., Dubois, C., Nkusi, E., Lenglez, S., Checinska, A., et al. (2017). Cell-Type-Specific Chromatin States Differentially Prime Squamous Cell Carcinoma Tumor-Initiating Cells for Epithelial to Mesenchymal Transition. Cell Stem Cell *20*, 191–204.

Laughney, A.M., Hu, J., Campbell, N.R., Bakhoum, S.F., Setty, M., Lavallée, V.-P., Xie, Y., Masilionis, I., Carr, A.J., Kottapalli, S., et al. (2020). Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat. Med. *26*, 259–269.

LeBleu, V.S., Taduri, G., O'Connell, J., Teng, Y., Cooke, V.G., Woda, C., Sugimoto, H., and Kalluri, R. (2013). Origin and function of myofibroblasts in kidney fibrosis. Nat. Med. *19*, 1047–1053.

Lee, H.-O., Hong, Y., Etlioglu, H.E., Cho, Y.B., Pomella, V., Van den Bosch, B., Vanhecke, J., Verbandt, S., Hong, H., Min, J.-W., et al. (2020). Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat. Genet. *52*, 594–603.

Li, C.-Z., Lin, Y.-X., Huang, T.-C., Pan, J.-Y., and Wang, G.-X. (2019a). Receptor-Interacting Protein Kinase 1 Promotes Cholangiocarcinoma Proliferation And Lymphangiogenesis Through The Activation Protein 1 Pathway. Onco. Targets. Ther. *12*, 9029–9040.

Li, L., Qi, L., Liang, Z., Song, W., Liu, Y., Wang, Y., Sun, B., Zhang, B., and Cao, W. (2015). Transforming growth factor-β1 induces EMT by the transactivation of epidermal growth factor signaling through HA/CD44 in lung and breast cancer cells. Int. J. Mol. Med. *36*, 113.

Li, Q.-Q., Xu, J.-D., Wang, W.-J., Cao, X.-X., Chen, Q., Tang, F., Chen, Z.-Q., Liu, X.-P., and Xu, Z.-D. (2009). Twist1-Mediated Adriamycin-Induced Epithelial-Mesenchymal Transition Relates to Multidrug Resistance and Invasive Potential in Breast Cancer Cells. Clinical Cancer Research *15*, 2657–2665.

Li, S., Cong, X., Gao, H., Lan, X., Li, Z., Wang, W., Song, S., Wang, Y., Li, C., Zhang, H., et al. (2019b). Tumor-associated neutrophils induce EMT by IL-17a to promote migration and invasion in gastric cancer cells. J. Exp. Clin. Cancer Res. *38*, 1–13.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417–425.

Limame, R., Op de Beeck, K., Lardon, F., De Wever, O., and Pauwels, P. (2014). Krüppel-like factors in cancer progression: three fingers on the steering wheel. Oncotarget *5*, 29–48.

Lu, Z., Ghosh, S., Wang, Z., and Hunter, T. (2003). Downregulation of caveolin-1 function by EGF leads to the loss of E-cadherin, increased transcriptional activity of beta-catenin, and enhanced tumor cell invasion. Cancer Cell *4*, 499–515.

Ludwig, K.F., Du, W., Sorrelle, N.B., Wnuk-Lipinska, K., Topalovski, M., Toombs, J.E., Cruz, V.H., Yabuuchi, S., Rajeshkumar, N.V., Maitra, A., et al. (2018). Small molecule inhibition of Axl targets tumor immune suppression and enhances chemotherapy in pancreatic cancer.

Cancer Res. *78*, 246.

Macnair, W., and Claassen, M. (2019). psupertime: supervised pseudotime inference for single cell RNA-seq data with sequential labels. bioRxiv doi: 10.1101/622001.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Mani, S.A., Guo, W., Liao, M.J., Eaton, E.N., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M., et al. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell *133*, 704–715.

McCorry, A.M., Loughrey, M.B., Longley, D.B., Lawler, M., and Dunne, P.D. (2018). Epithelial-to-mesenchymal transition signature assessment in colorectal cancer quantifies tumour stromal content rather than true transition. J. Pathol. *246*, 422–426.

McFaline-Figueroa, J.L., Hill, A.J., Qiu, X., Jackson, D., Shendure, J., and Trapnell, C. (2019). A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. Nat. Genet. *51*, 1389–1398.

McFarland, J.M., Paolella, B.R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W.N., Jones, A., Chambers, E., et al. (2020a). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat. Commun. *11*, 4296.

McFarland, J.M., Paolella, B.R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W.N., Jones, A., Chambers, E., et al. (2020b). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat. Commun. *11*, 4296.

McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626.

Meidhof, S., Brabletz, S., Lehmann, W., Preca, B., Mock, K., Ruh, M., Schüler, J., Berthold, M., Weber, A., Burk, U., et al. (2015). ZEB 1-associated drug resistance in cancer cells is reversed by the class I HDAC inhibitor mocetinostat. EMBO Molecular Medicine *7*, 831–847.

Mohamed, O.A., Clarke, H.J., and Dufort, D. (2004). Beta-catenin signaling marks the prospective site of primitive streak formation in the mouse embryo. Dev. Dyn. *231*, 416–424.

Mohammadi, S., Davila-Velderrain, J., and Kellis, M. (2020). A multiresolution framework to characterize single-cell state landscapes. Nat. Commun. *11*, 5399.

Morel, A.-P., Lièvre, M., Thomas, C., Hinkal, G., Ansieau, S., and Puisieux, A. (2008). Generation of Breast Cancer Stem Cells through Epithelial-Mesenchymal Transition. PLoS One *3*, e2888.

Nieto, M.A., Huang, R.Y., Jackson, R.A., and Thiery, J.P. (2016). EMT: 2016. Cell *166*, 21–45.

Ocaña, O.H., Córcoles, R., Fabra, A., Moreno-Bueno, G., Acloque, H., Vega, S., Barrallo-

Gimeno, A., Cano, A., and Nieto, M.A. (2012). Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. Cancer Cell *22*, 709–724.

Osborne, L.D., Li, G.Z., How, T., Tim O'Brien, E., Blobe, G.C., Superfine, R., and Mythreye, K. (2014). TGF-β regulates LARG and GEF-H1 during EMT to affect stiffening response to force and cell invasion. Mol. Biol. Cell *25*, 3528.

Pastushenko, I., and Blanpain, C. (2019). EMT Transition States during Tumor Progression and Metastasis. Trends in Cell Biology *29*, 212–226.

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi, S., Van Keymeulen, A., Brown, D., Moers, V., Lemaire, S., et al. (2018). Identification of the tumour transition states occurring during EMT. Nature *556*, 463–468.

Pattabiraman, D.R., Bierie, B., Kober, K.I., Thiru, P., Krall, J.A., Zill, C., Reinhardt, F., Tam, W.L., and Weinberg, R.A. (2016). Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability. Science *351*, aad3680.

Peixoto, P., Etcheverry, A., Aubry, M., Missey, A., Lachat, C., Perrard, J., Hendrick, E., Delage-Mourroux, R., Mosser, J., Borg, C., et al. (2019). EMT is associated with an epigenetic signature of ECM remodeling genes. Cell Death Dis. *10*, 205.

Perotti, V., Baldassari, P., Molla, A., Nicolini, G., Bersani, I., Grazia, G., Benigni, F., Maurichi, A., Santinami, M., Anichini, A., et al. (2019). An actionable axis linking NFATc2 to EZH2 controls the EMT-like program of melanoma cells. Oncogene *38*, 4384–4396.

Pitelka, D.R., Hamamoto, S.T., Duafala, J.G., and Nemanic, M.K. (1973). Cell contacts in the mouse mammary gland. I. Normal gland in postnatal development and the secretory cycle. J. Cell Biol. *56*, 797–818.

Principe, D.R., Diaz, A.M., Torres, C., Mangan, R.J., DeCant, B., McKinney, R., Tsao, M.-S., Lowy, A., Munshi, H.G., Jung, B., et al. (2017). TGFβ engages MEK/ERK to differentially regulate benign and malignant pancreas cell function. Oncogene *36*, 4336–4348.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611–1624.e24.

Purushothama Rao Tata, J.R. (2017). Plasticity in the lung: making and breaking cell identity. Development *144*, 755.

Qian, J., Olbrecht, S., Boeckx, B., Vos, H., Laoui, D., Etlioglu, E., Wauters, E., Pomella, V., Verbandt, S., Busschaert, P., et al. (2020). A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res. *30*, 745–762.

Ranieri, D., Proietti, S., Dinicola, S., Masiello, M.G., Rosato, B., Ricci, G., Cucina, A., Catizone, A., Bizzarri, M., and Torrisi, M.R. (2017). Simulated microgravity triggers epithelial mesenchymal transition in human keratinocytes. Sci. Rep. *7*, 1–10.

Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol. *38*, 954–961.

Riemann, A., Rauschner, M., Gießelmann, M., Reime, S., Haupt, V., and Thews, O. (2019). Extracellular Acidosis Modulates the Expression of Epithelial-Mesenchymal Transition (EMT) Markers and Adhesion of Epithelial and Tumor Cells. Neoplasia *21*, 450–458.

Rizvi, I., Gurkan, U.A., Tasoglu, S., Alagic, N., Celli, J.P., Mensah, L.B., Mai, Z., Demirci, U., and Hasan, T. (2013). Flow induces epithelial-mesenchymal transition, cellular heterogeneity and biomarker modulation in 3D ovarian cancer nodules. Proc. Natl. Acad. Sci. U. S. A. *110*, E1974–E1983.

Roca, H., Hernandez, J., Weidner, S., McEachin, R.C., Fuller, D., Sud, S., Schumann, T., Wilkinson, J.E., Zaslavsky, A., Li, H., et al. (2013). Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. PLoS One *8*, e76773.

Ruscetti, M., Quach, B., Dadashian, E.L., Mulholland, D.J., and Wu, H. (2015). Tracking and Functional Characterization of Epithelial–Mesenchymal Transition and Mesenchymal Tumor Cells during Prostate Cancer Metastasis. Cancer Res. *75*, 2749–2759.

Sapna Puri, M.H. (2010). Cellular Plasticity within the Pancreas— Lessons Learned from Development. Dev. Cell *18*, 342.

Sathe, A., Grimes, S.M., Lau, B.T., Chen, J., Suarez, C., Huang, R.J., Poultsides, G., and Ji, H.P. (2020). Single-Cell Genomic Characterization Reveals the Cellular Reprogramming of the Gastric Tumor Microenvironment. Clin. Cancer Res. *26*, 2640–2653.

Scanlon, C.S., Van Tubergen, E.A., Inglehart, R.C., and D'Silva, N.J. (2013). Biomarkers of Epithelial-Mesenchymal Transition in Squamous Cell Carcinoma. J. Dent. Res. *92*, 114.

Scheel, C., Eaton, E.N., Li, S.H.-J., Chaffer, C.L., Reinhardt, F., Kah, K.-J., Bell, G., Guo, W., Rubin, J., Richardson, A.L., et al. (2011). Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. Cell *145*, 926–940.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat. Commun. *9*, 20.

Sengez, B., Aygün, I., Shehwana, H., Toyran, N., Tercan Avci, S., Konu, O., Stemmler, M.P., and Alotaibi, H. (2019). The Transcription Factor Elf3 Is Essential for a Successful Mesenchymal to Epithelial Transition. Cells *8*, 858.

Serresi, M., Kertalli, S., Li, L., Schmitt, M.J., Dramaretska, Y., Wierikx, J., Hulsman, D., and Gargiulo, G. (2021). Functional antagonism of chromatin modulators regulates epithelial-mesenchymal transition. Science Advances *7*, eabd7974.

Shaikh, D., Zhou, Q., Chen, T., Ibe, J.C.F., Usha Raj, J., and Zhou, G. (2012). cAMP-dependent protein kinase is essential for hypoxia-mediated epithelial–mesenchymal transition, migration, and invasion in lung cancer cells. Cell. Signal. *24*, 2396–2406.

Shibue, T., and Weinberg, R.A. (2017). EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. Nat. Rev. Clin. Oncol. *14*, 611–629.

Shintani, Y., Fujiwara, A., Kimura, T., Kawamura, T., Funaki, S., Minami, M., and Okumura, M. (2016). IL-6 Secreted from Cancer-Associated Fibroblasts Mediates Chemoresistance in NSCLC by Increasing Epithelial-Mesenchymal Transition Signaling. J. Thorac. Oncol. *11*, 1482–1492.

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K., and Alon, U. (2012). Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. Science *336*, 1157–1160.

Stange, D.E., Koo, B.K., Huch, M., Sibbel, G., Basak, O., Lyubimova, A., Kujala, P., Bartfeld, S., Koster, J., Geahlen, J.H., et al. (2013). Differentiated Troy+ chief cells act as reserve stem cells to generate all lineages of the stomach epithelium. Cell *155*, 357–368.

Steele, N.G., Carpenter, E.S., Kemp, S.B., Sirihorachai, V.R., Stephanie The, Delrosario, L., Lazarus, J., Amir, E.-A.D., Gunchick, V., Espinoza, C., et al. (2020). Multimodal mapping of the tumor and peripheral blood immune landscape in human pancreatic cancer. Nature Cancer *1*, 1097–1112.

Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. *34*, 790–805.

Stemmler, M.P., Eccles, R.L., Brabletz, S., and Brabletz, T. (2019). Non-redundant functions of EMT transcription factors. Nat. Cell Biol. *21*, 102–112.

Stoker, M., Gherardi, E., Perryman, M., and Gray, J. (1987). Scatter factor is a fibroblast-derived modulator of epithelial cell mobility. Nature *327*, 239–242.

Su, C.J., Murugan, A., Linton, J.M., Yeluri, A., Bois, J., Klumpe, H., Antebi, Y.E., and Elowitz, M.B. (2020). Ligand-receptor promiscuity enables cellular addressing. bioRxiv doi: 2020.12.08.412643.

Su, L., Luo, Y., Yang, Z., Yang, J., Yao, C., Cheng, F., Shan, J., Chen, J., Li, F., Liu, L., et al. (2016). MEF2D Transduces Microenvironment Stimuli to ZEB1 to Promote Epithelial-Mesenchymal Transition and Metastasis in Colorectal Cancer. Cancer Res. *76*, 5054–5067.

Sun, Y., Schaar, A., Sukumaran, P., Dhasarathy, A., and Singh, B.B. (2018). TGFβ-induced epithelial-to-mesenchymal transition in prostate cancer cells is mediated via TRPM7 expression. Molecular Carcinogenesis *57*, 752–761.

Sundararajan, V., Pang, Q.Y., Choolani, M., and Huang, R.Y.-J. (2020). Spotlight on the Granules (Grainyhead-Like Proteins) – From an Evolutionary Conserved Controller of Epithelial Trait to Pioneering the Chromatin Landscape. Front. Mol. Biosci. *7*, 213.

Suzuki, A., Maeda, T., Baba, Y., Shimamura, K., and Kato, Y. (2014). Acidic extracellular pH promotes epithelial mesenchymal transition in Lewis lung carcinoma model. Cancer Cell Int. *14*, 129.

Suzuki, T., Osumi, N., and Wakamatsu, Y. (2010). Stabilization of ATF4 protein is required for the regulation of epithelial-mesenchymal transition of the avian neural crest. Dev. Biol. *344*, 658–668.

Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-

cell RNA-seq in the past decade. Nat. Protoc. *13*, 599–604.

Tadanori Mammoto, D.E.I. (2010). Mechanical control of tissue and organ development. Development *137*, 1407.

Tam, S.Y., Wu, V.W.C., and Law, H.K.W. (2020). Hypoxia-Induced Epithelial-Mesenchymal Transition in Cancers: HIF-1α and Beyond. Front. Oncol. *10*, 486.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell *159*, 635–646.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Taniguchi, K., and Karin, M. (2018). NF-κB, inflammation, immunity and cancer: coming of age. Nat. Rev. Immunol. *18*, 309–324.

Tarcic, G., Avraham, R., Pines, G., Amit, I., Shay, T., Lu, Y., Zwang, Y., Katz, M., Ben-Chetrit, N., Jacob-Hirsch, J., et al. (2012). EGR1 and the ERK-ERF axis drive mammary cell migration in response to EGF. FASEB J. *26*, 1582–1592.

Taube, J.H., Herschkowitz, J.I., Komurov, K., Zhou, A.Y., Gupta, S., Yang, J., Hartwell, K., Onder, T.T., Gupta, P.B., Evans, K.W., et al. (2010). Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proc. Natl. Acad. Sci. U. S. A. *107*, 15449–15454.

Terry, S., Savagner, P., Ortiz-Cuaran, S., Mahjoubi, L., Saintigny, P., Thiery, J., and Chouaib, S. (2017). New insights into the role of EMT in tumor immune escape. Mol. Oncol. *11*, 824.

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Thiery, J.P., Acloque, H., Huang, R.Y.J., and Nieto, M.A. (2009). Epithelial-mesenchymal transitions in development and disease. Cell *139*, 871–890.

Thomson, S., Buck, E., Petti, F., Griffin, G., Brown, E., Ramnarine, N., Iwata, K.K., Gibson, N., and Haley, J.D. (2005). Epithelial to mesenchymal transition is a determinant of sensitivity of non-small-cell lung carcinoma cell lines and xenografts to epidermal growth factor receptor inhibition. Cancer Res. *65*, 9455–9462.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2019). The Immune Landscape of Cancer. Immunity *51*, 411–412.

Tirosh, I., and Suvà, M.L. (2018). Deciphering Human Tumor Biology by Single-Cell Expression Profiling. Annual Review of Cancer Biology *3*, 151–166.

Tiwari, N., Tiwari, V.K., Waldmeier, L., Balwierz, P.J., Arnold, P., Pachkov, M., Meyer-Schaller, N., Schübeler, D., van Nimwegen, E., and Christofori, G. (2013). Sox4 is a master regulator of epithelial-mesenchymal transition by controlling Ezh2 expression and epigenetic

reprogramming. Cancer Cell *23*, 768–783.

Tomczak, A., Mortensen, J.M., Winnenburg, R., Liu, C., Alessi, D.T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N.H., et al. (2018). Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci. Rep. *8*, 1–10.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. Nat. Biotechnol. *32*, 381.

Tsai, J.H., Donaher, J.L., Murphy, D.A., Chau, S., and Yang, J. (2012). Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer Cell *22*. 725-36.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*. 564-76.

Uhlitz, F., Bischoff, P., Peidli, S., Sieber, A., Obermayer, B., Blanc, E., Trinks, A., Lüthen, M., Ruchiy, Y., Sell, T., et al. (2021). Mitogen-activated protein kinase activity drives cell trajectories in colorectal cancer. bioRxiv doi: 10.1101/2020.01.10.901579.

Valery Adorno-Cruz, H.L. (2019). Regulation and functions of integrin α2 in cell adhesion and disease. Genes & Diseases *6*, 16.

Van Buskirk, J. (2002). A comparative test of the adaptive plasticity hypothesis: relationships between habitat and phenotype in anuran larvae. Am. Nat. *160*, 87–102.

Wang, J.-J., Siu, M.K.-Y., Jiang, Y.-X., Chan, D.W., Cheung, A.N.-Y., Ngan, H.Y.-S., and Chan, K.K.-L. (2020). Infiltration of T cells promotes the metastasis of ovarian cancer cells via the modulation of metastasis-related genes and PD-L1 expression. Cancer Immunol. Immunother. *69*, 2275–2289.

Wang, Z., Li, Y., Kong, D., and Sarkar, F.H. (2010). The Role of Notch Signaling Pathway in Epithelial-Mesenchymal Transition (EMT) During Development and Tumor Aggressiveness. Curr. Drug Targets *11*, 745.

Ward, C., Volpe, G., Cauchy, P., Ptasinska, A., Almaghrabi, R., Blakemore, D., Nafria, M., Kestner, D., Frampton, J., Murphy, G., et al. (2018). Fine-Tuning Mybl2 Is Required for Proper Mesenchymal-to-Epithelial Transition during Somatic Reprogramming. Cell Rep. *24*, 1496–1511.e8.

Watanabe, K., Liu, Y., Noguchi, S., Murray, M., Chang, J.-C., Kishima, M., Nishimura, H., Hashimoto, K., Minoda, A., and Suzuki, H. (2019). OVOL2 induces mesenchymal-to-epithelial transition in fibroblasts and enhances cell-state reprogramming towards epithelial lineages. Sci. Rep. *9*, 1–12.

Wen, S., Hou, Y., Fu, L., Xi, L., Yang, D., Zhao, M., Qin, Y., Sun, K., Teng, Y., and Liu, M. (2019). Cancer-associated fibroblast (CAF)-derived IL32 promotes breast cancer cell invasion and metastasis via integrin β3-p38 MAPK signalling. Cancer Lett. *442*, 320–332.

Wynn, T.A. (2008). Cellular and molecular mechanisms of fibrosis. J. Pathol. *214*, 199.

Xie, L., Law, B.K., Chytil, A.M., Brown, K.A., Aakre, M.E., and Moses, H.L. (2004). Activation of the Erk Pathway Is Required for TGF-β1-Induced EMT In Vitro. Neoplasia *6*, 603–610.

Xu, J., Lamouille, S., and Derynck, R. (2009). TGF-beta-induced epithelial to mesenchymal transition. Cell Res. *19*, 156–172.

Yamaguchi, H., and Taouk, G.M. (2020). A Potential Role of YAP/TAZ in the Interplay Between Metastasis and Metabolic Alterations. Front. Oncol. *10*, 928.

Yang, A.D., Fan, F., Camp, E.R., van Buren, G., Liu, W., Somcio, R., Gray, M.J., Cheng, H., Hoff, P.M., and Ellis, L.M. (2006). Chronic oxaliplatin resistance induces epithelial-to-mesenchymal transition in colorectal cancer cell lines. Clin. Cancer Res. *12*, 4147–4153.

Yang, J., Mani, S.A., Donaher, J.L., Ramaswamy, S., Itzykson, R.A., Come, C., Savagner, P., Gitelman, I., Richardson, A., and Weinberg, R.A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. Cell *117*, 927–939.

Yang, J., Antin, P., Berx, G., Blanpain, C., Brabletz, T., Bronner, M., Campbell, K., Cano, A., Casanova, J., Christofori, G., et al. (2020). Guidelines and definitions for research on epithelial–mesenchymal transition. Nat. Rev. Mol. Cell Biol. *21*, 341–352.

Yao, L., Conforti, F., Hill, C., Bell, J., Drawater, L., Li, J., Liu, D., Xiong, H., Alzetani, A., Chee, S.J., et al. (2019). Paracrine signalling during ZEB1-mediated epithelial–mesenchymal transition augments local myofibroblast differentiation in lung fibrosis. Cell Death & Differentiation *26*, 943–957.

Yari Fontebasso, S.M.D. (2015). Drug Development for Metastasis Prevention. Crit. Rev. Oncog. *20*, 449.

Ye, X., Brabletz, T., Kang, Y., Longmore, G.D., Nieto, M.A., Stanger, B.Z., Yang, J., and Weinberg, R.A. (2017). Upholding a role for EMT in breast cancer metastasis. Nature *547*, E1–E3.

Yeh, H.-W., Hsu, E.-C., Lee, S.-S., Lang, Y.-D., Lin, Y.-C., Chang, C.-Y., Lee, S.-Y., Gu, D.-L., Shih, J.-H., Ho, C.-M., et al. (2018). PSPC1 mediates TGF-β1 autocrine signalling and Smad2/3 target switching to promote EMT, stemness and metastasis. Nat. Cell Biol. *20*, 479–491.

Yeung, T.-L., Leung, C.S., Wong, K.-K., Gutierrez-Hartmann, A., Kwong, J., Gershenson, D.M., and Mok, S.C. (2017). ELF3 is a negative regulator of epithelial-mesenchymal transition in ovarian cancer cells. Oncotarget *8*, 16951.

Yilmaz, M., and Christofori, G. (2009). EMT, the cytoskeleton, and cancer cell invasion. Cancer Metastasis Rev. *28*, 15–33.

Yilmaz, M., and Christofori, G. (2010). Mechanisms of motility in metastasizing cells. Mol. Cancer Res. *8*, 629–642.

Yonekawa, T., Gamez, G., Kim, J., Tan, A.C., Thorburn, J., Gump, J., Thorburn, A., and Morgan, M.J. (2015). RIP1 negatively regulates basal autophagic flux through TFEB to control sensitivity to apoptosis. EMBO Rep. *16*, 700–708.

Yu, M., Bardia, A., Wittner, B.S., Stott, S.L., Smas, M.E., Ting, D.T., Isakoff, S.J., Ciciliano, J.C., Wells, M.N., Shah, A.M., et al. (2013). Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. Science *339*, 580–584.

Yu, W., Huang, C., Wang, Q., Huang, T., Ding, Y., Ma, C., Ma, H., and Chen, W. (2014). MEF2 transcription factors promotes EMT and invasiveness of hepatocellular carcinoma through TGF-β1 autoregulation circuitry. Tumour Biol. *35*, 10943–10951.

Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., Xu, L., Luo, T., Yan, H., Long, Z., et al. (2019). CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. *47*, D900–D908.

Zhang, J., Tian, X.-J., and Xing, J. (2016a). Signal Transduction Pathways of EMT Induced by TGF-β, SHH, and WNT and Their Crosstalks. J. Clin. Med. Res. *5*, 41.

Zhang, J., Furong, A., Tongyang, H., and Liu, G.Z. (2020). SERPINE2 promotes esophageal squamous cell carcinoma metastasis by activating BMP4. Cancer Lett. *469*, 390–398.

Zhang, L., Ye, Y., Long, X., Xiao, P., Ren, X., and Yu, J. (2016b). BMP signaling and its paradoxical effects in tumorigenesis and dissemination. Oncotarget *7*, 78206.

Zhang, L., Wang, X., Lai, C., Zhang, H., and Lai, M. (2019). PMEPA1 induces EMT via a non-canonical TGF-β signalling in colorectal cancer. J. Cell. Mol. Med. *23*, 3603–3615.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, M., Kong, L., Liu, Y., and Qu, H. (2015). dbEMT: an epithelial-mesenchymal transition associated gene resource. Sci. Rep. *5*, 11459.

Zhao, M., Hu, X., Xu, Y., Wu, C., Chen, J., Ren, Y., Kong, L., Sun, S., Zhang, L., Jin, R., et al. (2019). Targeting of EZH2 inhibits epithelial-mesenchymal transition in head and neck squamous cell carcinoma via regulating the STAT3/VEGFR2 axis. Int. J. Oncol. *55*, 1165–1175.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zheng, X., Carstens, J.L., Kim, J., Scheible, M., Kaye, J., Sugimoto, H., Wu, C.-C., LeBleu, V.S., and Kalluri, R. (2015). Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. Nature *527*, 525–530.

Zhou, B., Chen, W.-L., Wang, Y.-Y., Lin, Z.-Y., Zhang, D.-M., Fan, S., and Li, J.-S. (2014). A role for cancer-associated fibroblasts in inducing the epithelial-to-mesenchymal transition in human tongue squamous cell carcinoma. J. Oral Pathol. Med. *43*, 585–592.